PhD-FSTC-2017-06
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defence held on 20/01/2017 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Shaman NARAYANASAMY
Born 13 January 1985 in Klang, (Malaysia)

DEVELOPMENT OF AN INTEGRATED OMICS *IN SILICO* WORKFLOW AND ITS APPLICATION FOR STUDYING BACTERIA-PHAGE INTERACTIONS IN A MODEL MICROBIAL COMMUNITY

Dissertation defence committee

Dr. Paul Wilmes, dissertation supervisor
Assistant Professor, Université du Luxembourg

Dr. Reinhard Schneider, Chairman
Université du Luxembourg

Dr. Jorge Gonçalves, Vice Chairman
Professor, Université du Luxembourg

Dr. Anders Andersson
Associate Professor, KTH Royal Institute of Technology

Dr. Rohan Williams
National University of Singapore

Development of an integrated omics *in silico* workflow and its application for studying bacteria-phage interactions in a model microbial community

A dissertation
by
Shaman Narayanasamy

Completed in the
Eco-Systems Biology Group, Luxembourg Centre for Systems Biomedicine

To obtain the degree of
DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG
EN *BIOLOGIE*

Dissertation Defence Committee:

Supervisor:	Asst. Prof. Dr. Paul Wilmes
Chair of committee:	Dr. Reinhard Schneider
Vice chair of comittee:	Prof. Dr. Jorge Gonçalves
Committee members:	Asst. Prof. Dr. Anders Andersson
	Dr. Rohan Williams

2016

For my parents.
For my sisters.
In loving memory of my dogs Faith, Gollum, Baboo, Gulab.

Declaration

I hereby declare that this dissertation has been written only by the undersigned and without any assistance from third parties. Furthermore, I confirm that no sources have been used in the preparation of this thesis other than those indicated herein.

Shaman Narayanasamy
Belval, Luxembourg
February 10, 2017

ACKNOWLEDGEMENTS

I would first and foremost like to thank my direct supervisor Asst. Prof. Dr. Paul Wilmes for providing me with the opportunity to work under his supervision and in collaboration with his esteemed Eco-Systems Biology group. This work would not be possible without your wholehearted guidance and advice. I've learned a great deal about my topic throughout my time with you. We've managed to push the limits of our work to a level beyond my expectations. Thanks for allowing, not only me, but every single member of the group to formulate and expand upon our own ideas. Also, thanks for providing us with the best possible working environment and infrastructure to conduct our work! In my humble opinion, your open-mindedness was the foundation that brought this group to one of the top in this field of study. Although, this did not come by easily. I still remember the days when we as a group had difficulties getting our papers accepted. Now, we realized that it was simply because we were simply ahead of the rest. And this is thanks to your vision which has been cultivated within me and, I'm sure within the rest of the group. One thing I always had the utmost confidence is that if I put in the effort, you will always be there to back me up. Thanks for always having my back! I sincerely hope that my contribution represents all that you, and the group stands for and will further accelerate our work to further push the limits of what is possible, or rather achieve the IMPossible.

Asst. Prof. Dr. Anders Andersson and Dr. Rohan Williams, thanks for taking the time off your schedules part take in my PhD defence. It is a great honour to have you both in my jury. I would like to thank my CET committee members Prof. Dr. Jorge Goncalves and Dr. Reinhard Schneider who were always supportive of my work and for their guidance throughout my PhD studies. I am also honoured to have you both as chairpersons of my PhD defense. Thanks to Prof. Dr. Serge Haan and his team in the doctoral school for always doing their best in supporting the PhD students. Thank you to both the LCSB and the University of Luxembourg support staff for their help. Special thanks to the institute director, Prof. Dr. Rudi Balling for being the pillar of this institute and making it grow to what it is today. Your enthusiasm has always been and will always be inspiring and infectious!

My fellow ESBers, past and present, I would like to first thank all of you for creating an awesome working environment. Dr. Emilie Muller, there are no words that could express my gratitude towards you. This work would not have been possible without your guidance and support. It was the utmost joy working with you. You're the best! Dr. Anna Heints-Buschart, thanks very much for always supporting us PhD students.

Your problem solving abilities are parallel to none! Laura Lebrun, thanks you so much for your tireless and consistent effort in obtaining our samples and providing support in the lab. This work would definitely not be possible without you. Malte Herold, thanks for your supporting me in my work. Please bear the torch for the LAOers! Anne Kaysen, thanks for helping me along in my work and for being the number one user of IMP! Linda Wampach, thanks for all your graphical skills you brought to the group. Your puns make a hard day at the office more bearable! Joelle, Joanna, Kacy, Audrey, Janine, Luise and Mark, thanks for making the group a great one to work in, even if we don't necessary coincide in terms of work, I am grateful for all your support! Former members, Cedric, Abdul, Bimal, Dil, Pranjul and Mahesh. Thanks for all the good times! Finally, I would like to thank the coffee machine and my laptop for not breaking throughout my PhD, despite heavy usage.

Yohan Jarosz, thanks for your enthusiasm and your positive attitude in times of difficulty. You brought the work within this thesis to another level. You the man! To our number one collaborator, Dr. Patrick May, working with you has truly been a valuable learning experience! Thanks for all your support and guidance. Dr. Nicolas Pinel, thanks for setting up the foundation of this work. Also, thanks to the rest of the R3 team and the HPC team for all their support during this work.

I thank Fonds National de la Recherche for funding my PhD. Thanks to the University of Luxembourg, the Luxembourg Centre for Systems Biomedicine and Doctoral School in Systems and Molecular Biomedicine for their support throughout my PhD.

To my former supervisor from University of Helsinki, Dr. Jaarko Salojarvi, you've always been and still are an inspiration to me. Prof. Kaarina Sivonen, thanks for introducing me to the word of microbiology which sparked the interest in my current work. Dr. Hao Wang, thanks for your supervision during my master's thesis. University of Helsinki provided the foundation necessary to carry out the work.

To all my good friends in Luxembourg, Susana, Shankari, Aishwarya, Raman, Dheeraj, Zuogong, Anestis, Hector and Sevgin. You are my family over here. Thanks for always being around. Special thanks to my fellow UNIjammers for making my experience in Luxembourg something I would never have imagined! My friends in Helsinki, my home away from home, thank you. And last but not least, to my friends back in Malaysia, your encouragement is greatly valued. I'll never forget where I come from and I will always be a Klang boy. Satu hati sampai mati!

To my extended family back in Malaysia: Chinnavva, Susi pinni, Prema pinni, Vengkat chinnaya, Uncle Rajan, Mani mava, Ananthi At the, Karishma, Neelam, Snegha, Tharisan, Ram, Raj, Hamsa, Sarvin, Suresh mava, Aunty Evis, Nastenca, Yascinta, Sabrina, Berenth. Thanks for keeping Amma company when all of us are not around and when times were rough for her. I am truly indebted to all of you for all that you have done!

Finally, to my dearest sisters Kaarjel and Sai; thanks for being irritatingly over achieving siblings, and making me feel the need to push myself to fit in with you both! You are the best sisters anyone could ask for. Thanks very much for always being there. Love you both! Amma and Papa, thanks for always believing in me. Thanks for being an inspiration and a motivation. I am the person I am today because of your love and the values you instilled within me. And for that, I am eternally grateful to you both.

ABSTRACT

Microbial communities are ubiquitous and dynamic systems that inhabit a multitude of environments. They underpin natural as well as biotechnological processes, and are also implicated in human health. The elucidation and understanding of these structurally and functionally complex microbial systems using a broad spectrum of toolkits ranging from *in situ* sampling, high-throughput data generation ("omics"), bioinformatic analyses, computational modelling and laboratory experiments is the aim of the emerging discipline of Eco-Systems Biology. Integrated workflows which allow the systematic investigation of microbial consortia are being developed. However, *in silico* methods for analysing multi-omic data sets are so far typically lab-specific, applied *ad hoc*, limited in terms of their reproducibility by different research groups and sub-optimal in the amount of data actually being exploited. To address these limitations, the present work initially focused on the development of the Integrated Meta-omic Pipeline (IMP), a large-scale reference-independent bioinformatic analyses pipeline for the integrated analysis of coupled metagenomic and metatranscriptomic data. IMP is an elaborate pipeline that incorporates robust read preprocessing, iterative co-assembly, analyses of microbial community structure and function, automated binning as well as genomic signature-based visualizations. The IMP-based data integration strategy greatly enhances overall data usage, output volume and quality as demonstrated using relevant use-cases. Finally, IMP is encapsulated within a user-friendly implementation using Python while relying on Docker for reproducibility. The IMP pipeline was then applied to a longitudinal multi-omic dataset derived from a model microbial community from an activated sludge biological wastewater treatment plant with the explicit aim of following bacteria-phage interaction dynamics using information from the CRISPR-Cas system. This work provides a multi-omic perspective of community-level CRISPR dynamics, namely changes in CRISPR repeat and spacer complements over time, demonstrating that these are heterogeneous, dynamic and transcribed genomic regions. Population-level analysis of two lipid accumulating bacterial species associated with 158 putative bacteriophage sequences enabled the observation of phage-host population dynamics. Several putatively identified bacteriophages were found to occur at much higher abundances compared to other phages and these specific peaks usually do not overlap with other putative phages. In addition, there were several RNA-based CRISPR targets that were found to occur in high abundances. In summary, the present work describes the development of a new bioinformatic pipeline for the analysis of coupled metagenomic and metatranscriptomic datasets derived from microbial communities and its application to a study focused on the dynamics of bacteria-virus interactions. Finally, this work demonstrates the power of integrated multi-omic investigation of microbial consortia towards the conversion of high-throughput next-generation sequencing data into new insights.

SCIENTIFIC OUTPUT

Major parts of this thesis are based upon work that has either been published, is currently under peer-review and/or ready for submission with the candidate as the first author. In addition, the candidate has also co-authored several publications of which minor parts are incorporated in the thesis. The full list of scientific outputs is listed in sections below and the original manuscripts are provided in the **Appendix A**.

Publications in peer-review journals

- **Shaman Narayanasamy**, Emilie E.L. Muller, Abdul R. Sheik, Paul Wilmes (2015). Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. *Microbial Biotechnology* **8**: 363-368. [**Appendix A.1**]

- **Shaman Narayanasamy**[†], Yohan Jarosz[†], Emilie E.L. Muller, Anna Heintz-Buschart, Malte Herold, Anne Kaysen, Cédric C. Laczny, Nicolàs Pinel, Patrick May, Paul Wilmes (2016) IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biology* **17**: 260. [**Appendix A.2**]

- Emilie E.L. Muller, Nicolás Pinel, Cédric C. Laczny, Michael R. Hoopmann, **Shaman Narayanasamy**, Laura A. Lebrun, Hugo Roume, Jake Lin, Patrick May, Nathan D. Hicks, Anna Heintz-Buschart, Linda Wampach, Cindy M. Liu, Lance B. Price, John D. Gillece, Cédric Guignard, Jim M. Schupp, Nikos Vlassis, Nitin S. Baliga, Robert L. Moritz, Paul S. Keim, Paul Wilmes (2014). Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nature Communications* **5**: 5603. [**Appendix A.3**]

- Hugo Roume, Anna Heintz-Buschart, Emilie E.L. Muller, Patrick May, Venkata P. Satagopam, Cédric C. Laczny, **Shaman Narayanasamy**, Laura A. Lebrun, Michael R. Hoopmann, Jim M. Schupp, John D. Gillece, Nathan D. Hicks, David M. Engelthaler, Thomas Sauter, Paul S. Keim, Robert L. Moritz, Paul Wilmes (2015). Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *NPJ Biofilms and Microbiomes* **1**: 15007. [**Appendix A.4**]

[†] Co-first author

Submissions in peer-review journals

- Linda Wampach, Anna Heintz-Buschart, Angela Hogan, Emilie E.L. Muller, **Shaman Narayanasamy**, Cédric C. Laczny, Luisa W. Hugerth, Lutz Bindl, Jean Bottu, Anders F. Andersson, Carine de Beaufort, Paul Wilmes (submitted). Colonization and succession within the human gut microbiome by archaea, bacteria and microeukaryotes during the first year of life. *Frontiers in Microbiology* [**Appendix A.5**]
- Anne Kaysen, Anna Heintz-Buschart, Emilie E. L. Muller, **Shaman Narayanasamy**, Linda Wampach, Cédric C. Laczny, Katharina Franke, Jörg Bittenbring, Jochen G. Schneider, Paul Wilmes (submitted). Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic stem cell transplantation. *Journal of Experimental & Clinical Cancer Research* [**Appendix A.6**]

Publication in non-peer review platform

- **Shaman Narayanasamy**[†], Yohan Jarosz[†], Emilie E.L. Muller, Cédric C. Laczny, Malte Herold, Anne Kaysen, Anna Heintz-Buschart, Nicolàs Pinel, Patrick May, Paul Wilmes (2016). IMP: a pipeline for reproducible metagenomic and metatranscriptomic analyses. *BioRxiv*. [**Appendix A.8**]

Manuscripts in preparation

- Emilie E.L. Muller[†], **Shaman Narayanasamy**[†], Myriam Zeimes, Laura A. Lebrun, Nathan D. Hicks, John D. Gillece, James M. Schupp, Paul Keim, Paul Wilmes (in preparation). First draft genome sequence of a strain belonging to the *Zoogloea* genus and its gene expression *in situ*. [**Appendix A.7**]
- **Shaman Narayanasamy**, Emilie E. L. Muller, Laura A. Lebrun, Nathan D. Hicks, John D. Gillece, James M. Schupp, Paul S. Keim, Paul Wilmes (in preparation). The dynamics of bacteriophages and bacterial host populations within oleaginous microbial consortia within wastewater treatment plants. [**Chapter 3**]

Oral presentations in scientific conferences, symposia and workshops

- Dynamic changes in the CRISPR-spacer complement and targeted bacteriophages within a natural microbial community resolved using time-resolved metagenomics and metatranscriptomics (2014). *Opening the Microbial World With Metagenomics*. Helsinki, Finland.
- Metagenomic and metatranscriptomic analyses of CRISPR-*Cas* dynamics within a microbial community (2015). *Life Sciences PhD days*. Belval, Luxembourg.
- Integrated omics provides unprecedented insights into microbial community structure and function (2016) *European Space Agency workshop for Micro-Ecological Life Support Systems Alternative (MELiSSA)*. Lausanne, Switzerland.

[†] Co-first author

- Metagenomic and metatranscriptomic analyses of CRISPR-*Cas* dynamics within a mixed microbial community (2016). *16th Conference of the International Society of Microbial Ecology*. Montreal, Canada.
- IMP: A reproducible pipeline for reference-independent integrated metagenomic and metatranscriptomic analyses (2016). *LCSB Minisymposium on Lab Automation* Belval, Luxembourg.
- IMP: The Integrated Meta-omic Pipeline: a tale of analysis and automation towards reproducible research results (2016). *1st RSG Luxembourg Congress & 2nd BeNeLuxFr Symposium* Belval, Luxembourg.

Poster presentations in scientific conferences, symposia and workshops

- A dynamic population-level model of antiviral defense mechanisms (2013). *Life Science PhD Days*. Luxembourg, Luxembourg.
- Eco-systems biology of microbial communities: Integration of biomolecular information from unique samples (2013). *EMBL Symposium: New Approaches and Concepts in Microbiology*. Heidelberg, Germany.
- Eco-systems biology of microbial communities: Integration of biomolecular information from unique samples (2013). *2nd Symposium of Systems Biomedicine*. Belval, Luxembourg.

CONTENTS

Abstract **iii**

Scientific output **iv**

Contents **vii**

List of Figures **x**

List of Tables **xii**

1 Integrated omics for the characterization of microbial community structure, function and dynamics **1**

1.1 Microbial communities 2
1.2 Model microbial community 3
1.3 Bacteriophage - bacterial host interactions 9
  1.3.1 Phage infection mechanisms 10
  1.3.2 The CRISPR-*Cas* mechanism 10
1.4 Eco-Systems Biology 13
  1.4.1 Eco-Systems Biology for the study of phage-host interactions 13
  1.4.2 Biomolecular extraction 14
  1.4.3 Multi-omic measurements 16
  1.4.4 Meta-omic NGS data analysis 19
  1.4.5 Multi-omic analyses of meta-omic data 23
1.5 Objectives of this work 30

2 A pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses **31**

2.1 Abstract 32
2.2 Background 32

2.3 Methods . . . 32
2.3.1 Details of the IMP implementation and workflow . . . 33
2.3.2 Iterative single-omic assemblies . . . 37
2.3.3 Execution of pipelines . . . 38
2.3.4 Data usage assessment . . . 38
2.3.5 Assembly assessment and comparison . . . 39
2.3.6 Analysis of contigs assembled from MT data . . . 39
2.3.7 Analysis of subsets of contigs . . . 39
2.3.8 Computational platforms . . . 39
2.3.9 Availability of data and material . . . 40
2.4 Results . . . 41
2.4.1 Overview of the IMP implementation and workflow . . . 41
2.4.2 Assessment and benchmarking . . . 46
2.4.3 Use-cases of integrated metagenomic and metatranscriptomic analyses in IMP . . . 53
2.5 Discussion . . . 57
2.6 Conclusion . . . 59

3 The dynamics of bacteriophages and bacterial host populations within the model system 61
3.1 Abstract . . . 62
3.2 Background . . . 62
3.3 Methods and material . . . 63
3.3.1 Sampling and strain collection . . . 63
3.3.2 Extraction of biomolecules . . . 64
3.3.3 Metagenome and metatranscriptome sequencing . . . 64
3.3.4 Bioinformatic analyses . . . 65
3.4 Results . . . 68
3.4.1 Large-scale analyses using IMP . . . 69
3.4.2 Community-level analysis of CRISPR elements and protospacers . . . 71
3.4.3 Population-level analysis of CRISPR elements . . . 76
3.4.4 Putative bacteriophage sequences . . . 79
3.4.5 Bacteriophage and host dynamics . . . 85
3.5 Discussion . . . 91
3.6 Conclusion . . . 95

4 General conclusions and outlook 96
4.1 Integrated omics: From data to associations . . . 97
4.2 Extending the functionality of IMP . . . 99
4.2.1 Updates with state-of-the-art tools . . . 100
4.2.2 Integration of reference-based analysis . . . 101
4.2.3 Extension to multi-sample analyses . . . 102
4.2.4 Metaproteomic analyses engine . . . 102

4.2.5 Keeping up with technological advancements 103
4.2.6 Standardized benchmarking for integrated omics 103
4.3 Bacterial-phage interactions 104
4.3.1 Moving beyond associations and hypotheses 105
4.3.2 From Eco-Systems Biology to applications 106

References **107**

Glossary **131**

Appendices **134**

Appendix A Article manuscripts **135**
A.1 Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. 136
A.2 IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. 143
A.3 Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. 166
A.4 Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. 177
A.5 Colonization and succession within the human gut microbiome by archaea, bacteria and microeukaryotes during the first year of life 190
A.6 Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic stem cell transplantation. 235
A.7 First draft genome sequence of a strain belonging to the *Zoogloea* genus and its gene expression *in situ*. 292
A.8 IMP: a pipeline for reproducible metagenomic and metatranscriptomic analyses. 317

Appendix B Additional figures **318**
B.1 Supplementary figures for **Chapter 3** 319

Appendix C Additional tables **320**
C.1 Supplementary tables for **Chapter 3** 320

Appendix D Additional files **326**
D.1 Additional file 2.1 326
D.2 Additional file 2.2 326
D.3 Additional file 2.3 326

LIST OF FIGURES

1.1 Schematic representation of an activated sludge based biological wastewater treatment plant process . . . 6
1.2 The path from large-scale integrated omics to hypothesis testing and biotechnological application in the context of biological wastewater treatment . . . 7
1.3 Biofuel production from wastewater sludge . . . 8
1.4 Structure, life cycle and dynamics of bacteriophages . . . 10
1.5 Mode of action of type II CRISPR-*Cas* systems . . . 12
1.6 Concomitant extraction of biomolecules from a single unique microbial community sample and their downstream high-throughput measurement techniques . . . 15
1.7 Next generation sequencing protocols and chemistries . . . 18
1.8 Simplified workflow for reference-independent metagenomic and/or metatranscriptomic analyses . . . 22

2.1 Schematic overview of the IMP pipeline. . . . 44
2.2 Example output from the IMP analysis of a human microbiome dataset (HF1) . . . 45
2.3 Assessment of data usage and output generated from co-assemblies compared to single-omic assemblies . . . 50
2.4 Assessment of the IMP-based iterative co-assemblies in comparison to MOCAT- and MetAMOS-based co-assemblies. . . . 53
2.5 Metagenomic and metatranscriptomic data integration of a human fecal microbiome . . . 56

3.1 Simplified schema for the study of phage-host interaction . . . 69
3.2 Summary of input and output of IMP analyses of the LAMPs time-series datasets . . . 71
3.3 Summary of CRISPR elements detected using different methods. . . . 72
3.4 Community-wide dynamics of CRISPR elements . . . 74
3.5 Summary of CRISPR element information . . . 75
3.6 *M. parvicella* population-level dynamics of CRISPR elements . . . 78

3.7 LCSB005 population-level dynamics of CRISPR elements 79
3.8 Dynamics of LCSB005 host and associated bacteriophages 86
3.9 Dynamics of *M. parvicella* host and associated bacteriophages 87
3.10 Dynamics of *M. parvicella* host and associated RIGes. 90

B.1 Microscopy photo of bacterial strain LCSB005 319

LIST OF TABLES

1.1 Multi-omic studies of microbial communities 25

2.1 Statistics of iterative assemblies performed on MG and MT datasets 48
2.2 Mapping statistics for human microbiome samples 51
2.3 Contigs with a likely viral/bacteriophage origin/function reconstructed from the metatranscriptomic data. 54

3.1 Summary of CRISPR elements detected using different methods 72
3.2 Summary statistics of CRISPR elements. 76
3.3 Number of CRISPR repeats and flanks (from metagenomic and metatranscriptomic data) associated to draft genomes of lipid accumulating bacterial species. 77
3.4 Summary of putative phages of the *M. parvicella* and LCSB005 populations 81

4.1 The development of IMP 100

C.1 IMP time series analyses summary 321
C.2 Summary statistics of LAMP community CRISPR elements based different time points . . . 323
C.3 Summary statistics of *Candidatus* Microthrix parivicella Bio17-1 population CRISPR elements based different time points 324
C.4 Summary statistics of LCSB005 population CRISPR elements based different time points . . 325

CHAPTER 1

INTEGRATED OMICS FOR THE CHARACTERIZATION OF MICROBIAL COMMUNITY STRUCTURE, FUNCTION AND DYNAMICS

A major part of this chapter was adapted and modified from the following first-author peer-review publications:

Shaman Narayanasamy, Emilie E.L. Muller, Abdul R. Sheik, Paul Wilmes (2015). Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. *Microbial Biotechnology* **8**: 363-368. [**Appendix A.1**]

Shaman Narayanasamy†, Yohan Jarosz†, Emilie E.L. Muller, Anna Heintz-Buschart, Malte Herold, Anne Kaysen, Cédric C. Laczny, Nicolàs Pinel, Patrick May, Paul Wilmes (2016) IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biology* **17**: 260. [**Appendix A.2**]

---

†Co-first author

1.1 Microbial communities

Naturally occurring microbial communities (or consortia) are ubiquitous in the environment and underpin important biomedical, biotechnological and natural processes. For instance, the human microbiome (microbial communities in and on the human body) plays an important role in human health [Turnbaugh *et al.*, 2007; Greenhalgh *et al.*, 2016]; activated sludge microbial communities within biological wastewater treatment plants are important for the remediation of communal wastewater prior its release into the environment [Daims *et al.*, 2006]; and marine microbial communities are believed to be the main photosynthetic oxygen producers [Arrigo, 2005]. Given the importance of microbial communities, it is essential for the scientific community to better understand these important components of nature in their natural environments.

This work utilizes the terms microorganisms and microbes interchangeably. While these terms may carry a general definition and are used in various contexts, within this work these terms encompass a broad range of microbial taxa including, but not limited to, bacteria, archaea, protozoa, micro-eukaryotes and viruses. A collection of microbial cells of the same species/subtype present in the same place and at the same time, is referred to as a population. In general, microbes rarely ever exist naturally as isolated populations, but rather as mixtures of different microbial populations. These mixtures of microbial populations are referred to as microbial communities (or mixed microbial communities), which may have emergent properties, i.e. properties of the constituent populations do not sum to the properties of the entire community, and thus cannot be predicted by studying individual populations separately [Odum and Barrett, 1971]. The complexities of microbial communities vary a lot from one system to another. For example, acid mine drainage biofilms represent relatively simple communities, with low diversity and dominance by specific taxa [Denef *et al.*, 2010]. On the other end of the spectrum, soil microbial communities exhibit far more complex structures, with up to thousands of different microbial populations which undergo rapid changes of the community due to rapid environmental fluctuations [Mocali and Benedetti, 2010]. In between these two extremes, there are microbial communities such as those present within biological wastewater treatment plants which exhibit important characteristics of both low and high complexity microbial communities [Sheik *et al.*, 2014; Narayanasamy *et al.*, 2015; Muller *et al.*, 2014a]. Such communities therefore represent good model systems for microbial ecology [Daims *et al.*, 2006].

Co-existing microbial populations within natural microbial communities are usually present in differing abundances (i.e. differing community structures) and undergo constant change over time (i.e. community dynamics). These complex structures and dynamics result from constant adaptation of the community to environmental fluctuations which include physical (temperature, pH) and chemical (substrate availability) changes [Muller *et al.*, 2013; Narayanasamy *et al.*, 2015]. Furthermore, populations within a microbial community are constantly interacting with each other (e.g. predation, competition, mutualism, antagonism, etc.), further affecting the overall dynamics of the community.

Understanding microbial community structure (i.e. what are the members of the community) is of general interest. However, more recently interest has also focused on deciphering the function/phenotype of different microbial populations within communities to elucidate what the different members of the community are doing. This is under the assumption that different populations within a given microbial community are believed to carry out specific functions or roles, thus contributing to the collective phenotype of the community [Muller *et al.*, 2013; Narayanasamy *et al.*, 2015]. Given the aforementioned characteristics (i.e.

the complexity and dynamics) of microbial communities, they may be viewed as omnipresent highly complex systems, yet elusive components of the environment.

The field of microbiology and molecular biology have advanced greatly over the past years due to the emergence of cutting-edge technologies that enable high-resolution and high-throughput molecular measurements [Muller *et al.*, 2013; Segata *et al.*, 2013]. Thus, to complement classical tools, techniques and strategies of microbiology based on strain cultivation within controlled lab conditions [Stewart, 2012], the scientific community has moved towards the direct study of microbial communities within their natural environments. Studying microbial consortia by application of high-throughput, high resolution molecular measurements ("meta-omics") provides the opportunity to discover novel organisms and functionalities (genes), which may not be possible with classical microbiological methods, due to the unculturability of most naturally occurring microbial taxa under standard laboratory conditions [Staley and Konopka, 1985; Amann *et al.*, 1995; Stewart, 2012]. However, it is important to define microbial communities that will serve as models for fundamental understanding of microbial communities (i.e. complexity, interactions and dynamics) as well as communities that play an important role, either in nature, biotechnological processes and/or human health.

1.2 Model microbial community

This present work leveraged a model microbial community found within biological wastewater treatment (BWWT) plants for extensive study. These communities are biotechnologically relevant due to their influence on the wastewater treatment process, which is in turn important for the environment. In particular, this work will focus on microbial populations that accumulate lipids that are present in floating sludge islets that occur at the air-water interface of the anoxic tanks. The lipid accumulating phenotype of these organisms represent a potential resource for renewable energy production from wastewater. More importantly for the present work though, is the fact that this system is well suited for fundamental understanding of characteristics and dynamics of natural occurring microbial community.

Direct discharge of organic (e.g., carbohydrates, fats, proteins, organic solvents) and inorganic (e.g. phosphate, nitrate, metallic ions) compounds into natural water bodies may lead to severe perturbations of ecosystems as they can either serve as nutrient and stimulate growth of heterotrophic organisms leading to a reduction in dissolved oxygen or be toxic towards the native organisms [Conley *et al.*, 2009; Roume *et al.*, 2013b]. Therefore, BWWT relies on naturally-occurring microbial community-driven remediation of municipal and/or industrial wastewater, before its release into the environment. Since its conception about a century ago by E. Arden and W.T. Lockett, BWWT plants, including the standard activated sludge process and other ancillary processes, has become a widespread process that is present in most of the developed world. For instance, in 2013 Luxembourg had 109 BWWT plants that handled approximately 95 % of the total wastewater [Roume *et al.*, 2013b]. While the overall procedure seems rather simplified, BWWT is a complex process at the interface of engineering, biology and biochemistry, which is not completely understood to date [Wang and Pereira, 1987]. Conventional BWWT plants are made up of a combination of physical, chemical and biological stages that remove solids, organic matter and nutrients from wastewater (**Figure 1.1**). In summary, the objectives of wastewater treatment include: i) minimizing the release of organic compounds into natural water bodies to reduce the bloom of heterotrophic organisms and thereby reducing overall oxygen demand, ii) oxidization of ammonia to reduce toxicity and its deoxygenation effects and iii) reduction of eutrophic substances, such as phosphate [Mara and Horan, 2003; Conley *et al.*, 2009; Roume *et al.*, 2013b].

Traditional BWWT plants consist of three stages including: i) physical treatments which removes suspended solids from the wastewater, ii) primary treatment to remove settleable organic and inorganic solids via sedimentation as well as grease and oil removal by skimming and iii) secondary treatment involving the removal/reduction of organic matter in the wastewater using an aerobic biological treatment processes, i.e. the activated sludge process (**Figure 1.1**). This process relies on naturally occurring microbial communities to reduce organic compound availability in the wastewater [Wagner and Loy, 2002]. These organic compounds are mainly assimilated into microbial biomass (carbon sources) or are oxidized and released as carbon dioxide. In essence, wastewater treatment relies on the digestion of the energy-rich C-C bonds by microorganisms and its transformation into microbial biomass as a means of removing these compounds from the wastewater.

Although the activated sludge process is one of the most widely used biotechnological processes in the world, it is known to be highly energy- and resource-consuming (i.e. water pumping, air bubbling). Yet, BWWT processes hold great potential for future sustainable production of various commodities, including energy, from wastewater as well as from other mixed substrates, further expanding on their original function of wastewater treatment [Sheik *et al.*, 2014; Muller *et al.*, 2014a]. Indeed, BWWT plants host diverse and dynamic microbial communities, which in turn contain microbial species that possess varied metabolic capabilities over changing environmental conditions, e.g. microorganisms accumulating various storage compounds of biotechnological importance, thus making it a reservoir for potentially useful novel microbial species [Sheik *et al.*, 2014; Muller *et al.*, 2014a; Narayanasamy *et al.*, 2015]. Consequently, BWWT plants represent a readily available resource (and facility) for production of biofuels, with relatively low cost of modification to already existing structures. An approximated 226 prokaryotes were identified within various BWWT microbial communities. However, information and detailed study of these potentially useful microorganisms remain limited, with only 72 draft genomes reported so far, out of the total 226 identified organisms [McIlroy *et al.*, 2015].

The model microbial system subject of the present work is represented by microbial communities occurring within floating sludge islets of an anoxic tank of a BWWT plant (**Figures 1.1** to **1.3**) [Sheik *et al.*, 2014; Muller *et al.*, 2014b]. The anoxic tank (**Figure 1.1**) is part of the activated sludge process, more specifically, within the secondary treatment of the BWWT process that promotes denitrification, i.e. reduction of nitrate (NO_3) to nitrogen gas (N_2) by heterotrophic bacteria (i.e. bacteria that requires organic carbon for growth). The removal of nitrogen from wastewater is achieved by limiting dissolved oxygen (O_2) levels, such that heterotrophic bacteria are forced to consume nitrate for energy, instead of oxygen, which is then released as gaseous dinitrogen. The water surface of these anoxic tanks tend to accumulate foamy sludge islets (**Figure 1.3**), which contain lipid accumulating microbial populations (LAMPs) whereby the most notable is *Candidatus* Microthrix parvicella (also referred to as *M. parvicella*), a filamentous lipid accumulating organism which is highly dominant (up to 30 % relative abundance) in the system [Blackall *et al.*, 1996; Muller *et al.*, 2012; McIlroy *et al.*, 2013]. Consequently, its lipid accumulating properties are of pronounced interest from a biotechnological perspective, most specifically for lipid-based biofuel production from wastewater [Muller *et al.*, 2014a; Sheik *et al.*, 2016]. The floating sludge islets could be easily collected through surface skimming, compared to other by-products of BWWT plants. Therefore, it is of great interest to maximize the abundance of LAMPs, such as *M. parvicella*, through systematic manipulation of this specific microbial community for consistent and optimal production of biofuels from BWWT plants [Sheik *et al.*, 2014].

In addition to being a resource for the production of high added value compounds, the LAMPs are also well-suited for fundamental studies aimed at obtaining generalizable understanding and knowledge with regards to the ecology of microbial consortia. LAMPs exist within a fluctuating environment (water/air temperature, pH, oxygen and nutrient concentrations). However, these fluctuations are almost always within well-defined/-controlled physical and chemical boundaries [Daims *et al.*, 2006; Sheik *et al.*, 2014; Muller *et al.*, 2014a]. Overall, LAMPs represent a unique combination of a highly fluctuating, yet relatively well-controlled environment, which is rare in most natural ecosystems. More importantly, physico-chemical parameters, such as temperature, pH, oxygen and nutrient concentrations are routinely monitored and recorded. Such detailed monitoring allows the establishment of causal links between the influence of certain environmental effects on microbial community structure and/or function when coupled to temporal sampling. As such, this system also represents a convenient and virtually unlimited (high reproducibility) source of spatially and temporally resolved samples (**Figures 1.2** and **1.3**). Obtaining temporal and/or spatial samples from other microbial habitats, e.g. the marine environment, acid mine drainage biofilms, the human gastrointestinal tract, etc. would be rather challenging or in some cases, near impossible.

While being highly dynamic, LAMPs maintain a medium to high range of diversity/complexity with an alpha(α)-diversity of approximately 600, representing an important intermediary step/model between communities of lower diversity, e.g. acid mine drainage biofilms [Denef *et al.*, 2010], and complex communities, such as those from soil environments [Mocali and Benedetti, 2010]. In addition, LAMPs also exhibit a baseline stability over time, such that there is temporal succession of repeatedly few quantitatively (up to 30 % relative abundance) dominant populations [Muller *et al.*, 2014b,a; Roume *et al.*, 2015]. Overall, the model community demonstrates high dynamics, while retaining important and interesting hallmarks of other microbial communities including, for example, quantitative dominance of specific taxa (a characteristic of acid mine drainage biofilm communities) and rapid stochastic environmental fluctuations (a characteristic of soil environments). Microbial consortia from BWWT plants, including LAMPs, are very amenable to experimental validation at differing scales, ranging from laboratory-scale bioreactors to full-scale plants, thus providing the facility of conducting controlled experiments **Figure 1.2**. Overall, LAMPs exhibit important characteristics and properties rendering it an ideal model for microbial ecology [Daims *et al.*, 2006], and more specifically eco-systematic omic studies in line with a discovery-driven planning approach [Muller *et al.*, 2013], facilitating hypothesis formulation and verification in rapid succession (**Figure 1.2**).

In conclusion, the present work leverages a model community that is interesting from a biotechnological perspective of renewable biofuel production while being a representative system for studying microbial communities in general.

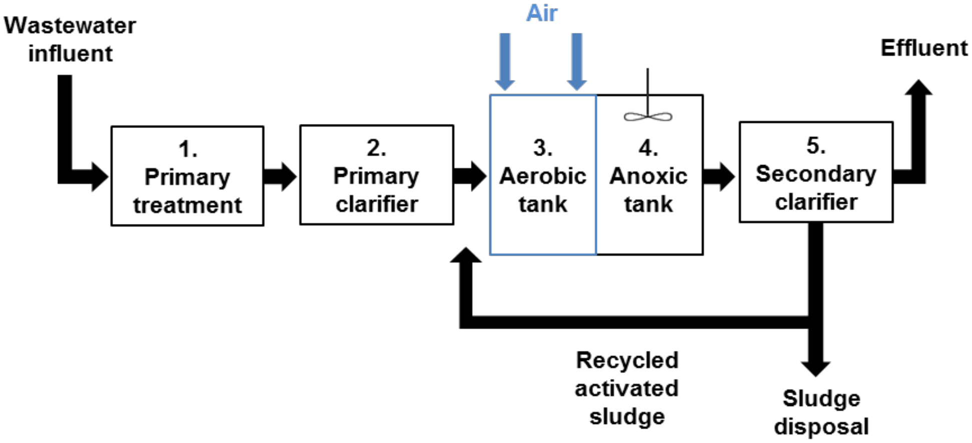


Figure 1.1: Schematic representation of an activated sludge based biological wastewater treatment plant process. Primary treatment consists of screening and grit removal in order to remove large-sized floating solids, while the primary clarifier is used to remove settling solids. The pre-treated wastewater is then mixed with microbial biomass present in the activated sludge and iteratively pumped into the aerobic tank where aerators enable its agitation and oxygenation and then into the anoxic tank. The activated sludge flocs are decanting in the secondary clarifier: the majority of this biomass is recycled to the beginning of the activated sludge process and the rest is either disposed or further used for methane production through anaerobic digestion. The treated wastewater effluent is then released into the environment (Adapted from Zeimes [2015]).

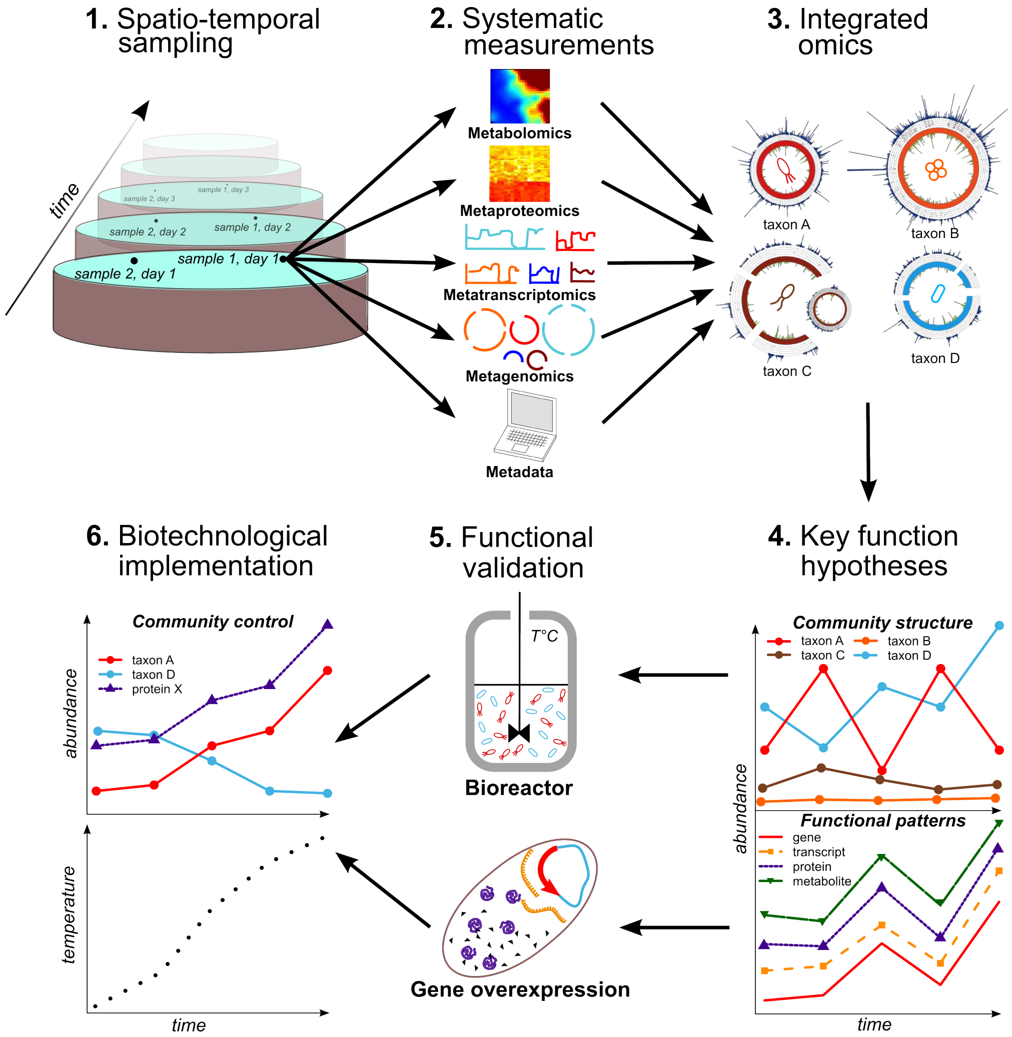


Figure 1.2: The path from large-scale integrated omics to hypothesis testing and biotechnological application in the context of biological wastewater treatment. Step 1; spatially and temporal resolved samples from BWWT plants. Step 2; sequential isolation of high-quality genomic deoxyribonucleic acid (DNA), ribonucleic acid (RNA), small RNA, proteins and metabolites from a single, undivided sample for subsequent systematic multi-omic measurements. Also including physico-chemical records. Step 3; multi-omic data integration and analysis for a multi-level snapshots of microbial community structure and function *in situ*. Step 4; statistical and mathematical modelling. Step 5; testing through targeted laboratory and/or *in situ* perturbation experiments followed by additional omic measurements. Step 6; control of microbial community structure and/or function (Adapted from Narayanasamy *et al.* [2015]).

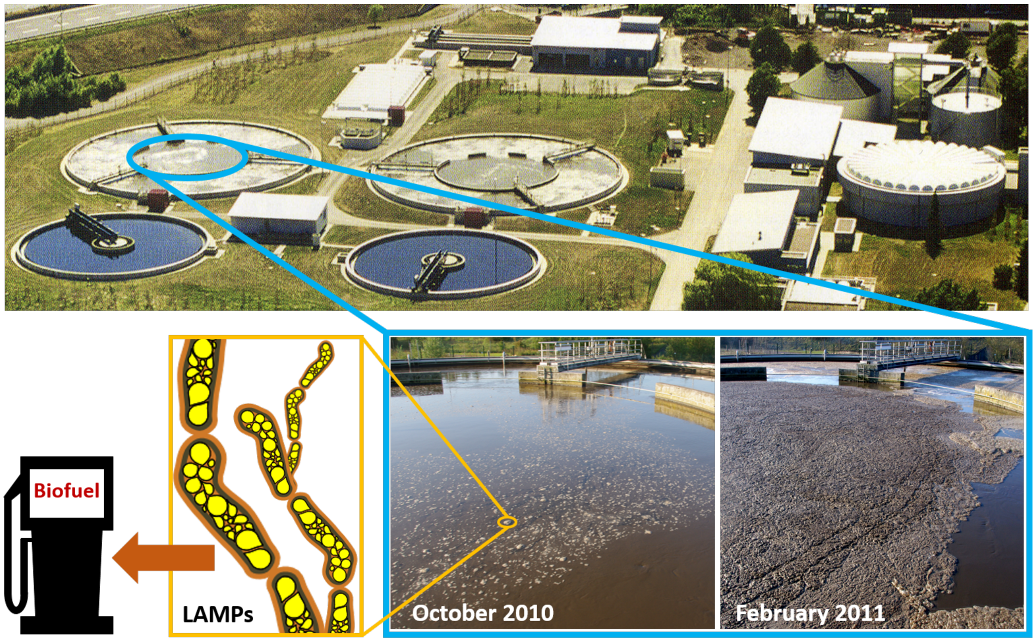


Figure 1.3: Biofuel production from wastewater sludge. Aerial photograph of the Schifflange biological wastewater treatment plant, Esch-sur-Alzette, Luxembourg (49°30’ 48.29” N; 6°1’ 4.53” E) operated by *Syndicat Intercommunal à Vocation Ecologique*. The “anoxic tank number 1” is highlighted by the blue circle and the corresponding photos, from that tank in autumn and winter show variable content of sludge in the different seasons. The sludge islets (e.g. highlighted in yellow) contain lipid accumulating microbial populations (LAMPs), which are potential biofuel producers (Courtesy of E.E.L Muller).

1.3 Bacteriophage - bacterial host interactions

Viruses are known to be the most abundant and diverse biological entities on the planet, inhabiting almost every environment, with an estimated range of 10^{30} to 10^{32} of total viral particles on Earth [Marcó *et al.*, 2012]. They are believed to be responsible for the lysis of up to 50 % of prokaryotic cells, thereby increasing the bioavailability of carbon and overall playing an important role in the carbon cycle [Breitbart and Rohwer, 2005]. It is important to note that viruses (and all relevant subclasses of virus), are referred to as biological entities/components within the scope of this work [Rybicki, 1990; Raoult and Forterre, 2008; Koonin and Starokadomskyy, 2016].

Bacteriophages are a subclass of viruses that infect and replicate specifically within bacterial cells (also referred to as phages throughout this work), which are believed to play an essential role in microbial communities by shaping their structure and influencing their dynamics (**Figure 1.4**) [Samson *et al.*, 2013]. Accordingly, studies have shown the involvement of phages within simple communities, such as acid mine drainage biofilms [Andersson and Banfield, 2008] and more complex microbial communities such as the: i) marine microbiome [Wommack and Colwell, 2000; Suttle, 2007; Sheik *et al.*, 2014], ii) human gastrointestinal tract microbiome [Stern *et al.*, 2012], iii) laboratory scale sludge bioreactors [Kunin *et al.*, 2008], and iv) full scale BWWT plants [Yasunori *et al.*, 2002].

Given the capability of phages to lyse bacterial cells (**Figure 1.4**), they have been suggested as a viable microbial community control strategy in various biomedical and biotechnological processes that rely on microbial communities [Withey *et al.*, 2005; Jassim *et al.*, 2016]. The idea itself could be traced back to as early as 1962 [Claeys, 1962], while there were also documented cases of the application of phage therapy [Withey *et al.*, 2005; Jassim *et al.*, 2016]. However, the inconsistent results of phage treatment coupled with the emergence of antibiotics, brought about the decline of phage therapy [Withey *et al.*, 2005; Jassim *et al.*, 2016]. More recently, the interest towards phage-based treatments has resurfaced, including its possible application as a control strategy for BWWT process [Withey *et al.*, 2005; Jassim *et al.*, 2016].

In principle, phage treatment could be used to mitigate common issues that plague BWWT plants, such as: i) foaming of activated sludge (i.e. anoxic tank floating sludge islets; **Section 1.2**), ii) sludge de-waterability and digestibility, iii) removal of pathogenic bacterial strains or iv) reduce strains that compete with functionally important/useful bacterial populations [Withey *et al.*, 2005; Jassim *et al.*, 2016]. In order to apply such strategies, it is essential to understand the role of bacteriophages in shaping BWWT plant communities, such as LAMPs [Withey *et al.*, 2005; Jassim *et al.*, 2016].

Despite the abundance and diversity of bacteriophages, information with regards these biological entities are relatively sparse compared to their bacterial host counterparts with approximately 2,200 viral genomes versus more than 45,000 bacterial genomes in publicly available databases [Reddy *et al.*, 2015; Paez-Espino *et al.*, 2016]. This gap in information is due to several reasons including, but not limited to: i) large fraction of their host populations cannot be cultivated, and thus preventing the culturing of the associated bacteriophages (**Section 1.1**), ii) the absence of marker genes for bacteriophages, such as the 16S rRNA genes for bacteria, create a challenge in classifying phage genomic material (**Sections 1.1** and **1.4.3**) [Roux *et al.*, 2011] and iii) some phages integrate their genomes within bacterial host genomes, hindering conclusive identification of phage genomes. Finally, given that a majority of bacterial species within BWWT remain unclassified **Section 1.2**, this translates to a sparse number of associated bacteriophages identified within LAMPs [Kunin *et al.*, 2008].

Fortunately, the advent of high-throughput omic datasets **Section 1.4.3** opens up new opportunities to study bacteriophages unlike previous efforts [Pride *et al.*, 2012; Reyes *et al.*, 2012; Wommack *et al.*, 2012; Shirley *et al.*, 2015; Paez-Espino *et al.*, 2016]. For instance, metagenomics (**Section 1.4.3**) provides access to all genomic (DNA) components within a given microbial community, including bacteriophages [Edwards *et al.*, 2015]. Specific techniques, including the use of information from bacterial antiviral defence mechanisms to associate bacteriophages and their host populations [Edwards and Rohwer, 2005; Andersson and Banfield, 2008; Stern *et al.*, 2012; Wommack *et al.*, 2012; Emerson *et al.*, 2013b,a; Edwards *et al.*, 2015].

1.3.1 Phage infection mechanisms

Phages are known to exhibit two types of life cycles (**Figure 1.4**). The first type is known as a lytic life cycle (**Figure 1.4**), which constitutes immediate replication of phages leading to lysis of host cells. However, certain bacteriophages follow a lysogenic life cycle (**Figure 1.4**) by being dormant within the host cells via integration of their genetic material into the host genomes as prophages. They then replicate as a lytic phage when the conditions are suitable. Overall, phages are obligate parasites which require successful infection of a host in order to replicate [Edwards *et al.*, 2015]. However, bacterial hosts are able to fend off the infections of phages through an arsenal of defence mechanisms [Labrie *et al.*, 2010; Edwards *et al.*, 2015]. Bacterial defense against phages include, but are not limited to: i) mutation of membrane receptors, ii) lipopolysaccharide coating of bacterial membrane, iii) restriction-modification of DNA and iv) CRISPR-*Cas* system [Labrie *et al.*, 2010]. As a consequence of these defense mechanisms, phages are under selective pressure to counter its hosts' defences [Samson *et al.*, 2013; Edwards *et al.*, 2015]. Consequently, bacterial hosts and phages are locked in a constant evolutionary arms race driven mainly by phage-bacteria interactions, which in turn drive microbial community dynamics [Samson *et al.*, 2013].

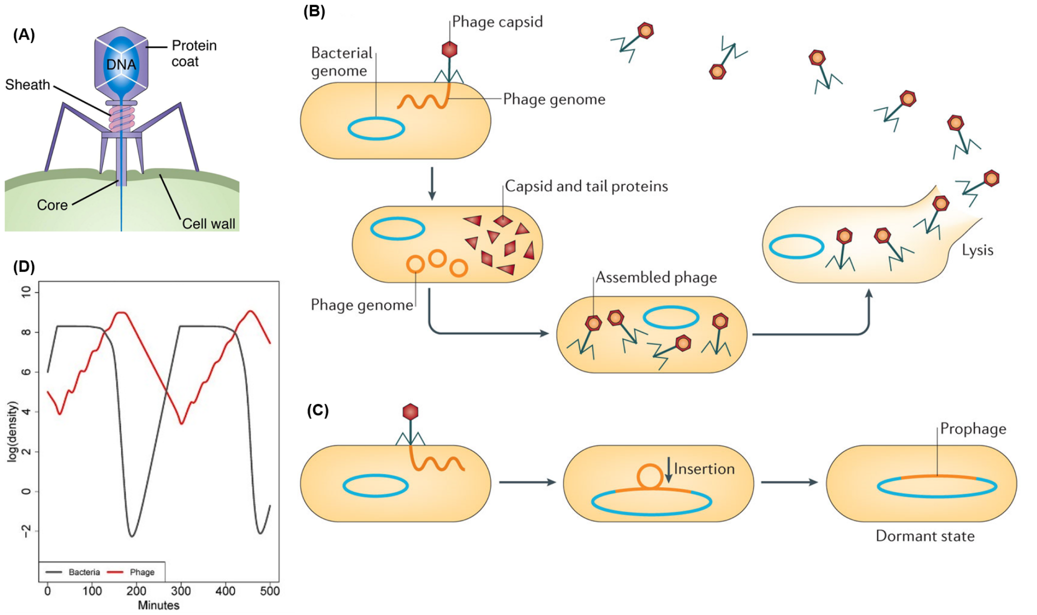


Figure 1.4: Structure, life cycle and dynamics of bacteriophages. **(A)** Structure of a phage (Adapted from Gelbart and Knobler [2008]). **(B)** Lytic phage life cycle involves: attachment to the bacterial cell; injection/adsorption of genetic material; replication of phage genome and generation of phage components; assembly of new phages; bacterial cell lysis and phage release (Adapted from Feiner *et al.* [2015]). **(C)** Lysogenic phage life cycle involves: attachment of phage to bacteria; injection/adsorption of genetic material; integration/insertion of phage genome into host genome; dormant state of phage as a prophage. The prophage exits the dormant stage and replicates as a lytic phage when conditions are favourable (Adapted from Feiner *et al.* [2015]). **(D)** Bacteriophage and bacterial host dynamics (Adapted from Bull *et al.* [2014]).

1.3.2 The CRISPR-*Cas* mechanism

While there are multiple ways of linking bacteriophage(s) with their host populations [Edwards *et al.*, 2015], this work primarily focuses on the CRISPR-*Cas* system as a means of associating of bacteriophages and bacterial hosts. The "clustered regularly interspaced palindromic repeats" or CRISPRs, are a class of sequences present within prokaryotic genomes that include distinct short repeat sequences, interspaced by short unique sequences [Barrangou and van der Oost, 2013; Rath *et al.*, 2015; Amitai and Sorek, 2016]. These sequences were first identified and described by Yoshizumi Ishino and colleagues upon accidentally locating these regions within the *E. coli* K12 strain [Ishino *et al.*, 1987]. Such regions were also later found in other prokaryotic species, such as *Haloferax mediterranei*, *Streptococcus pyogenes*, *Anabaena* sp. PCC 7120 and *Mycobacterium tuberculosis* [Jansen *et al.*, 2002]. The term "CRISPR" was coined by Jensen and colleagues [Mojica *et al.*, 2000; Jansen *et al.*, 2002]. In addition, they also identified CRISPR-associated (*cas*) genes (which translate to *Cas* proteins/enzymes) located adjacent (or in close proximity) to CRISPR genomic regions, and thereby suggesting the functional relationships between those genes and the CRISPR genomic regions. Since then, a large number of studies followed suit and deciphered the mechanism of the system as a memory-based immune system against invasive foreign genetic elements, such as bacteriophages and plasmids [Pourcel *et al.*, 2005; Kunin *et al.*, 2008; Shah *et al.*, 2013; Zhang *et al.*, 2013, 2014]. This form of defense came to be known as the CRISPR-*Cas* system and is estimated to exist within ~40 % of bacteria and ~90 % of archaea [Godde and Bickerton, 2006; Kunin *et al.*, 2007; Karginov and Hannon, 2010]. The simplified mechanism of the CRISPR-*Cas* system is represented in **Figure 1.5**.

The CRISPR genomic region within a prokaryotic genome is made up of multiple elements (**Figure 1.5**). The first element is a direct repeat sequence of approximately 24 to 47 bp, which occurs multiple times within a given CRISPR region [Ishino *et al.*, 1987; Jansen *et al.*, 2002; Datsenko *et al.*, 2012]. CRISPR *loci* are generally almost palindromic in nature implying that these regions could form hairpin structures upon transcription (and post-transcriptional processing) and are thereby well conserved within different prokaryotic clades [Kunin *et al.*, 2007]. Direct repeats are separated (or interspaced) by unique sequences known as spacers (**Figure 1.5**) [Ishino *et al.*, 1987; Jansen *et al.*, 2002; Kunin *et al.*, 2007; Datsenko *et al.*, 2012]. Unlike the direct repeat sequences, spacer sequencers were shown to be highly heterogeneous and dynamic, with constant additions, deletions and replacements of spacers within the CRISPR regions [Pourcel *et al.*, 2005]. This causes spacers to be highly heterogeneous within populations of a single prokaryotic species [Pourcel *et al.*, 2005]. Last but not least, functional CRISPR regions were shown to include an AT-rich leader sequence upstream [Bult *et al.*, 1996; Klenk *et al.*, 1997; Jansen *et al.*, 2002; Karginov and Hannon, 2010].

The simplified mechanism of the CRISPR-*Cas* system can be separated into three main/general stages: i) adaptation, ii) CRISPR-RNA (crRNA) biogenesis and iii) interference (**Figure 1.5**). The adaptation stage involves *Cas* proteins/enzymes, that detect and sample short fragments of sequences from foreign invasive elements. These short fragments are known as protospacers, i.e. the original elements of CRISPR spacers. Protospacers are recognized by these *Cas* proteins through specific sequence motifs, known as the protospacer adjacent motifs (PAM) [Marraffini and Sontheimer, 2010; Shah *et al.*, 2013]. Upon detection, the *Cas* proteins cleave a protospacer at two ends and incorporates the cleaved fragment into the CRISPR array of the host genome, usually within the leading end (i.e. the first repeat sequence from the AT-rich flanking leader sequence) [Marraffini and Sontheimer, 2010; Datsenko *et al.*, 2012]. As such, the spacer information is conserved in the next generation of bacteria, storing the history of previous infections [Marraffini and Sontheimer, 2010]. The crRNA biogenesis involves transcription of the CRISPR genomic regions to generate an unprocessed crRNA (pre-crRNA) [van Rij and Andino, 2006; Labrie *et al.*, 2010; Marraffini and Sontheimer, 2010]. The pre-crRNA is then further processed by splicing, such that individual spacer sequences are accompanied by a single repeat sequence, forming the processed crRNA [Marraffini and Sontheimer, 2010]. In the interference phase, these crRNAs form a complex with *Cas* proteins to detect and interfere with foreign genetic elements via splicing/inhibition [Marraffini and Sontheimer, 2010]. More specifically, the repeat region within the crRNA folds into a hairpin structure which is used to bond with the *Cas* proteins, while the spacer is utilized as a guide to target invasive genetic elements via complementary binding [Nishimasu *et al.*, 2014]. While the general mechanism of the CRISPR-*Cas* system is relatively simple, specific CRISPR-*Cas* mechanisms can be further classified into types I, II and III. This classification is based on their participating *cas* genes (or *Cas* enzymes) and could be further divided into various subtypes [Makarova *et al.*, 2011]. Furthermore, the catalogue of *cas* genes are continually expanding based on updated knowledge and new data [Zhang *et al.*, 2014]. Complementary to the discovery of new *Cas* enzymes, there are also many recent studies uncovering novel CRISPR-*Cas* mechanisms within prokaryotes.

The CRISPR-*Cas* system was rapidly translated into biotechnological application through the development of a genome editing tool based on the type II CRISPR-*Cas* system, mediated by *Cas9* nuclease protein/enzyme [Wiedenheft *et al.*, 2011; Jinek *et al.*, 2012; Sashital *et al.*, 2012; Selle and Barrangou, 2015]. Yin and collaborators applied the CRISPR-*Cas9* based genome editing tool to successfully correct a mutation within a mouse model of a human liver disease [Yin *et al.*, 2014].

Given that CRISPR *loci*, more specifically the CRISPR spacers, represent the history of previous infections, multiple studies effectively leveraged this information to identify concomitant phages, and thus representing an important method of tracking host-phage interactions within microbial consortia [Stern *et al.*, 2012; Biswas *et al.*, 2013; Zhang *et al.*, 2013; Edwards *et al.*, 2015; Paez-Espino *et al.*, 2016]. In summary, the aforementioned studies utilize CRISPR information and leverage omic data to expand the available genomic resource with regards to phages, and thereby advancing the field [Paez-Espino *et al.*, 2016] (**Section 1.3.1**). Despite the wealth of information delivered by these studies, there is still a gap elucidating the long- and/or short- term dynamics of bacteriophages and bacterial hosts within their natural environment, and thereby understanding the overall influence of phages within microbial consortia [Labrie *et al.*, 2010; Samson *et al.*, 2013].

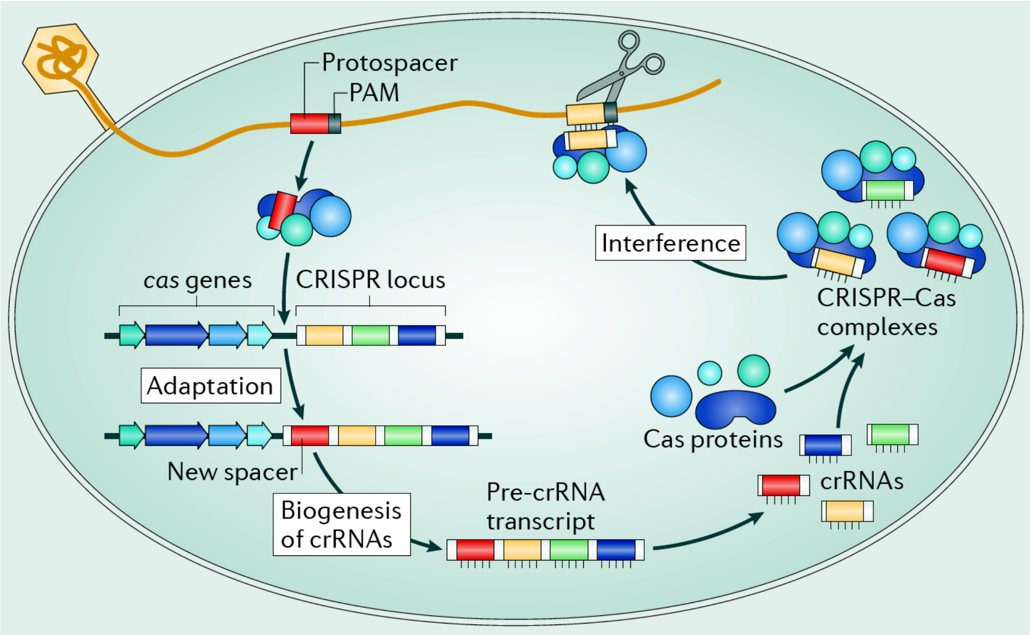


Figure 1.5: Mode of action of type II CRISPR-*Cas* systems. The direct repeats of the "clustered regularly interspaced palindromic repeats" (CRISPR) *locus* are separated by short stretches of non-repetitive (unique) DNA called spacers, which are acquired from the invading DNA of viruses or plasmids in a process known as adaptation, during which an additional repeat is also duplicated. The CRISPR *locus* is transcribed as a long primary pre- CRISPR-RNA (crRNA) transcript, which is processed to produce a collection of short crRNAs (a process referred to as biogenesis of crRNA). Each crRNA contains segments of a repeat and a full spacer and, in conjunction with a set of *Cas* proteins, forms the core of CRISPR-*Cas* complexes. These complexes act as a surveillance system and provide immunity against ensuing infections by phages or plasmids encoding DNA complementary to the crRNA. On recognition of a matching target sequence, the plasmid or viral DNA is cleaved in a sequence-specific manner (known as interference). The nucleotide sequence of the spacer must be highly similar to a region of the viral genome or plasmid (known as the protospacer) for the CRISPR-*Cas* complex to inhibit replication of these foreign genetic element. In type I and II CRISPR-*Cas* systems, a conserved sequence motif adjacent to the protospacer, known as the protospacer-adjacent motif (PAM), is needed for spacer acquisition and interference (Adapted from Samson *et al.* [2013]).

1.4 Eco-Systems Biology

It is important to address the concept of Systems Biology as a prerequisite to Eco-Systems Biology [Zengler, 2009; Zengler and Palsson, 2012]. Systems Biology involves the study of multiple biological components, including but not limited to, biomolecules, cells, tissues, organs and entire organisms within a biological system. The field of Systems Biology emerged due to the highly complex and dynamic nature of living systems which cannot be predicted and/or elucidated by looking at individual components/parts of a given biological system [Shahzad and Loor, 2012] . Consequently, Systems Biology is a highly inter-disciplinary field that combines various methodologies spanning from high-throughput molecular measurements, bioinformatic analyses, laboratory experiments and mathematical modelling. There are two generalized study designs/approaches defined under the umbrella of Systems Biology which includes the "top-down" and "bottoms-up" approach. Top-down systems biology characterizes biological components of a particular system using large-scale omic datasets followed by subsequent generation of mathematical models of the system. Those generated models may aid in uncovering new insights into the biological system in question [Zengler, 2009; Shahzad and Loor, 2012]. On the contrary, bottom-up systems biology begins with a detailed models of a specific biological system on the basis of its molecular properties and are usually followed by targeted measurements. These measurements stem from either isolation, cultivation and/or various single-cell techniques [Zengler, 2009]. Despite being regarded as isolated approaches, these approaches should not be viewed as separate, but rather should applied in an integrated manner to elucidate biological systems [Zengler, 2009].

Molecular Eco-Systems Biology (hereafter referred to as Eco-Systems Biology) applies similar principles and methodologies of Systems Biology within microbial systems, such as the complex microbial communities described in **Section 1.1** [Raes and Bork, 2008; Zengler, 2009; Zengler and Palsson, 2012]. Systematically obtained *in situ* time- and space-resolved datasets will allow deconvolution of structure-function relationships by identifying key community members and key community functions [Raes and Bork, 2008; Zengler and Palsson, 2012; Muller *et al.*, 2013; Narayanasamy *et al.*, 2015]. Knowledge garnered from such studies offers the potential to discover novel microorganisms and biological functionalities within the framework of Eco-Systems Biology [Albertsen *et al.*, 2013a; Muller *et al.*, 2014a; Roume *et al.*, 2015; Heintz-Buschart *et al.*, 2016; Laczny *et al.*, 2016]. In general, such insights may enable the control of microbial communities either through interventions for improvement/optimization of biomedical treatments and/or biotechnological processes [Muller *et al.*, 2013].

1.4.1 Eco-Systems Biology for the study of phage-host interactions

The application of Eco-Systems Biology can be extended to the study of bacteriophage and host interactions, especially given the possible application of bacteriophages in controlling microbial communities (**Section 1.3**).

More specifically, the information derived from microbial community samples *in situ* allows access to more information than compared to classical culture/isolate based methods. Accordingly, the current work combined the advantage of three separate avenues to effectively study phage and host interactions and dynamics. First, the facility to perform time-series sampling of microbial communities *in situ*, such as the LAMPs (**Section 1.2**), will enable the study of host and phage interactions within a natural system on a longitudinal scale. Second, the ability to mine for phage-related information from microbial community derived data, as highlighted in **Section 1.3**. Finally, using the CRISPR-*Cas* system as a valuable information source of phage-host interaction (**Section 1.3**).

The characteristic of microbial communities (**Section 1.1**) render standard microbiology-based methods (i.e. originally designed for pure isolate culture systems) ineffective [Muller *et al.*, 2013; Roume *et al.*, 2013b,a]. It is therefore absolutely essential to apply specialized non-culture based systematic approaches for the study of microbial systems. Eco-Systems Biology is an integrative framework that encompasses a wide array of specialized methods/techniques/analysis including: i) concomitantly extracted biomolecules ii) systematic high-throughput omic data measurements, iii) integration and analysis to the omic data, iv) experimental validation and ultimately v) the control of microbial systems (**Figure 1.2**). Accordingly, the following sections describe these aforementioned methods in detail, focusing primarily on the integration of the different omic data types. Briefly, the described methods enable the extraction of the information necessary for this work. These include, but are not limited to: i) genome sequences of both host and phage populations, ii) their predicted genes and corresponding functional annotations, iii) association of bacteriophages and their hosts (i.e. using CRISPR information) as well as iv) transcribed components, such as genes and CRISPR RNA.

1.4.2 Biomolecular extraction

The biomolecular extraction protocol designed by Roume and colleagues allows the sequential isolation of high-quality genomic deoxyribonucleic acid (DNA), ribonucleic acid (RNA), small RNA, proteins and metabolites from a single, undivided sample for subsequent systematic multi-omic measurements (**Figure 1.2**; step 2 and **Figure 1.6**) [Roume *et al.*, 2013b,a]. Importantly, this eliminates the need for subsampling the heterogeneous biomass and, therefore reduces the noise arising from incongruous omics data in the subsequent downstream integration and analysis steps (**Figure 1.2**; step 3 and **Figure 1.6**) [Muller *et al.*, 2013; Roume *et al.*, 2013b,a]. Biomolecular isolations obtained from the aforementioned methodologies are subjected to high-throughput measurements, resulting in omic data derived from a single unique sample to fulfill the premise of downstream integrated omic analysis (**Figures 1.2** and **1.6**) [Muller *et al.*, 2014a; Roume *et al.*, 2015; Heintz-Buschart *et al.*, 2016].

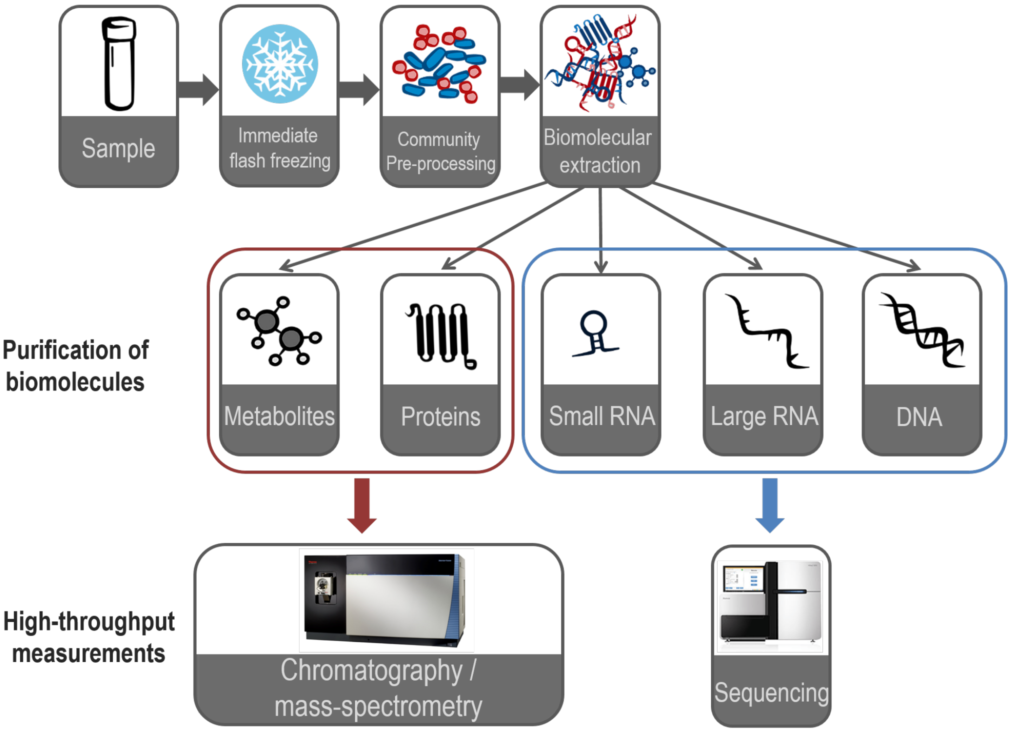


Figure 1.6: Concomitant extraction of biomolecules from a single unique microbial community sample and their downstream high-throughput measurement techniques (Courtesy of L. Wampach and A. Kaysen).

1.4.3 Multi-omic measurements

Studies within the context of Eco-Systems Biology were made possible mainly through the advent and availability of the high-throughput, high-resolution molecular measurements (referred to as omic data), applied to microbial consortia that were derived *in situ* [Raes and Bork, 2008; Zengler, 2009; Zengler and Palsson, 2012; Muller *et al.*, 2013]. Omic data involves the collective analysis (characterization and quantification) of certain features of a family/class of biomolecules (i.e. DNA, RNA, proteins or metabolites (**Figures 1.2** and **1.6**). The application of omic measurements to microbial communities results in meta-omic data, whereby the suffix "meta" (in the scope of this work) implies measurements/data derived from mixed microbial communities [Muller *et al.*, 2013; Segata *et al.*, 2013]. More specifically, meta-omic datasets (metagenomic, metatranscriptomic, metaproteomics and (meta-)metabolomics) enables high-resolution molecular-level studies of such microbial systems on a much larger scale compared with previous efforts [Muller *et al.*, 2013].

Metagenomics

The concept of DNA sequencing was introduced by Fredrick Sanger and colleagues in 1975 when they proposed a chemistry that combines the use of polymerase chain reaction, inhibition/termination of DNA polymerase activity and labelled fluorescence dyes [Sanger and Coulson, 1975; Sanger *et al.*, 1977]. This chemistry was further developed to achieve more rapid and accurate sequencing method, which resulted in the first sequenced genome of the phi-X174 bacteriophage [Sanger *et al.*, 1977]. Sanger sequencing was the only available method of sequencing until the emergence of next-generation sequencing (NGS) technologies, which enabled large-scale and deep sequencing of DNA fractions with relatively lower cost [Liu *et al.*, 2012]. Currently, there are multiple NGS technologies/platforms available, whereby each of these platforms employ a specific chemistries in deciphering DNA sequences (**Figure 1.7**) [Blow, 2008; Met; Quail *et al.*, 2012]. NGS technologies can be further divided into "second-generation sequencing" technologies, also called massive parallel sequencing, such as Illumina [Bentley *et al.*, 2008], Roche 454 [Margulies *et al.*, 2005] and SoLiD [McKernan *et al.*, 2009] and the more recent "third generation sequencing" technologies, such as Pacific Biosciences [Eid *et al.*, 2009] and Oxford Nanopore [Manrao *et al.*, 2012]. The clonal amplification step of DNA molecules to produce DNA colonies (**Figure 1.7**) is the main difference between second and third generation sequencing methods, whereby this step is absent in the latter, culminating in the concept of single molecule sequencing [Blow, 2008; Eid *et al.*, 2009; Manrao *et al.*, 2012]. It is important to note that the clonal amplification steps are necessary in generating the large volumes (throughput) in second-generation sequencing methods, which is not possible with third-generation methods.

NGS platforms are unable to read genome-sized (or long) DNA molecules, due to the current limitations of all NGS technologies. Therefore, the general protocol of genome sequencing first involves a preparation step of the DNA samples, such that they can be loaded onto NGS platforms (**Figure 1.7**), usually by random fragmentation of multiple copies of a genome (for isolate genomic samples) to generate shorter DNA fragments that would be readable by the NGS platforms. This overall preparatory procedure prior to sequencing is widely known as a whole genome shotgun (WGS) procedure and result in WGS libraries. More specifically, the term "shotgun" in WGS is used due to the aforementioned fragmentation process which is akin to the quasi-random firing pattern of a shotgun. It is important to note that WGS library preparation protocols vary for different sequencing technologies (**Figure 1.7**) [Met; Liu *et al.*, 2012]. WGS libraries are processed by an NGS instrument/machine (i.e. sequencer) to yield *in silico* representations of the biological DNA molecules, known as sequencing reads [Blow, 2008; Met; Quail *et al.*, 2012]. The Illumina sequencing platform is notably the most applied sequencing technology due to its ability to generate the highest-throughput (i.e. largest number of bases/reads per sequencing run) with relatively low cost [Liu *et al.*, 2012].

It is also important to highlight that DNA sequencing may also be carried out using a targeted approach, which typically refers to the sequencing of a known DNA *locus*, selected either for the encoded function or, more often for its phylogenetic/taxonomic information, using primer-based amplification. In the specific context of microbial communities, high-throughput ribosomal RNA (rRNA) gene amplicon sequencing (usually 16S rRNA gene sequencing) facilitates the preliminary characterization of microbial community composition and structure [Segata *et al.*, 2013]. However, such targeted amplicon sequencing will not be classified as metagenomic (MG) data within the scope of this work. Rather, this work defines MG data to be the result of a WGS sequencing procedure applied on bulk microbial community-derived DNA samples. Beyond targeted sequencing datasets, MG data is arguably the most commonly generated high-throughput dataset for microbial community studies, with 37,239 datasets publicly available on NCBI sequence read archive (SRA) [Leinonen *et al.*, 2011], as of 24 October 2016. MG data provides information on the community structure, (i.e. which microbial community members are present) as well as a prediction of gene functions (i.e. the functional potential) [Muller *et al.*, 2013; Vanwonterghem *et al.*, 2014]. Within the scope of this work, it is important to highlight that metagenomic sequencing entails the indiscriminate sequencing of all DNA molecules within a sample which also includes viral (or phage) DNA genomes. Hence, MG data was previously shown as a large resource that can be used to mine for viral sequences, which far surpasses the what would be able to be achieved with classical microbiology methods [Paez-Espino *et al.*, 2016].

Metatranscriptomics

Given that the RNA molecular structure is analogous to DNA, it can also be subjected to NGS with additional laboratory processing protocols. It is important to note that NGS platforms are only able to sequence DNA molecules. Therefore, the RNA samples serve as template to synthesize reverse transcribed DNA (i.e. complementary DNA - cDNA) before undergoing NGS. Similar to the DNA samples, RNA samples can also undergo targeted sequencing (as described in **Section 1.4.3**) or whole transcriptome shotgun (WTS) sequencing, i.e. random shotgun sequencing of bulk RNA samples equivalent to WGS.

In the context of this work, WTS performed on bulk RNA samples derived from microbial communities are considered as metatranscriptomic (MT) data. As of 24 October 2016, there are 397 MT datasets available on the NCBI SRA [Leinonen *et al.*, 2011], which is relatively little compared to MG data (37,239). In addition, rRNA depletion (often partial depletion) of MT samples enables deeper sequencing of mRNA and thus providing better access to functional readouts from MT data and other interesting RNA-based components, such as RNA viral genomes. Functional expression can be characterized using MT data, and by extension provides a snapshot of which members of the community are most active and the genes that they are expressing (and the quantity of expression). Similar to MG data, it also provides access to RNA sequences derived from viruses or bacteriophages, including both genes and RNA genomes. Finally, although this work focuses on the analysis of MG and MT data, it is important to note that additional function-based omic datasets, such as metaproteomic and (meta-)metabolomic datasets are crucial to fully understand the actual functional capacity of microbial communities (**Figures 1.2** and **1.2**).

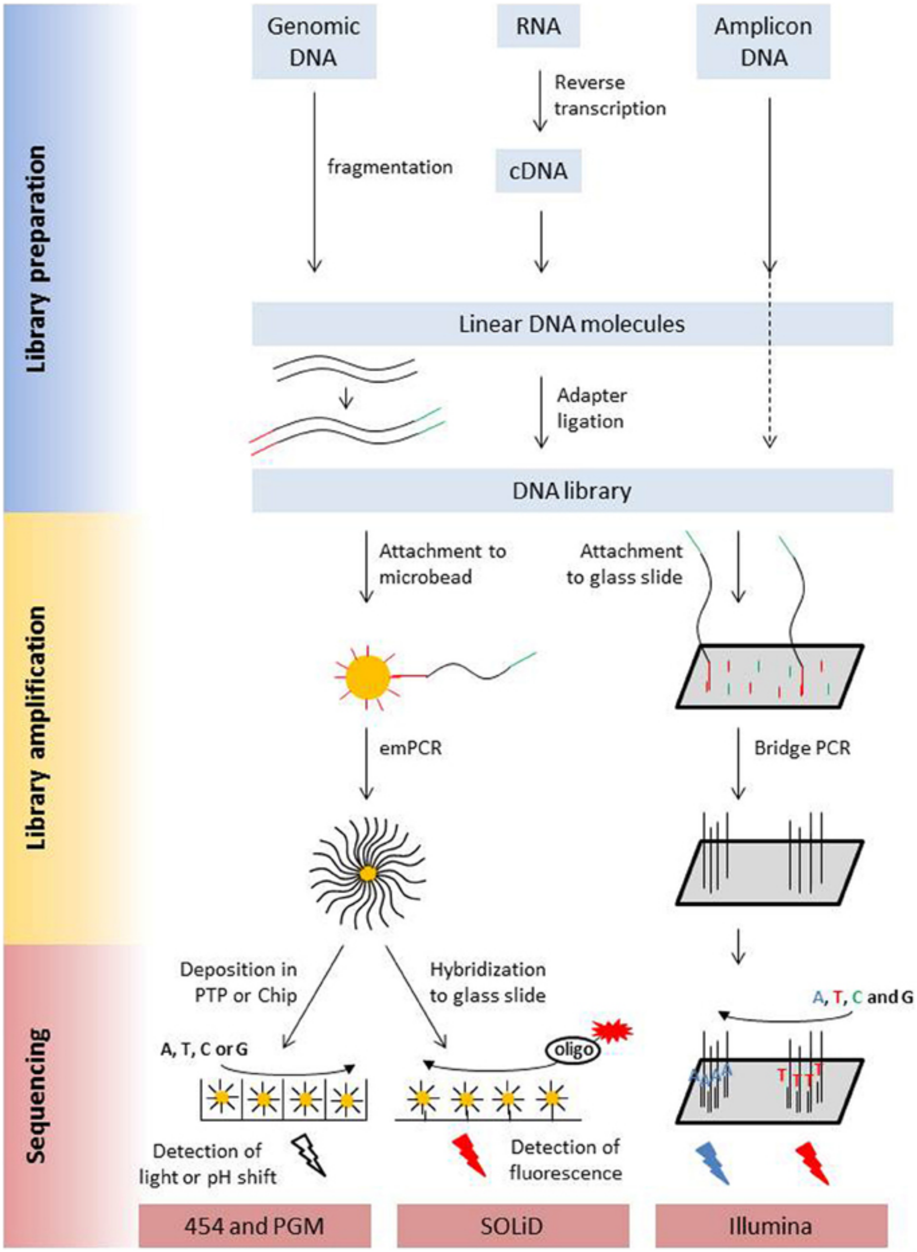


Figure 1.7: Next generation sequencing protocols and chemistries. Different types of starting molecules are converted into double-stranded DNA molecules that are flanked by adapters. Adapters are sequencing platform-specific artificial DNA molecules which are introduced to the biological DNA via ligation. They immobilize the biological DNA to surfaces that contain sequences complementary to the adapters. These surfaces include either beads (454/SoLiD/PGM) or a flow cell (Illumina). DNA molecules attached to these surfaces are amplified prior to sequencing. Clonal amplicons are spatially separated on the pico-titer plates (454), glass slides (SoLiD) or chips (Illumina). Sequencing chemistries involve the detection of nucleotides incorporated via complementarity (A-T, G-C) to the immobilized (and cloned) template DNA strand. Detection may be achieved either through usage of labelled nucleotides, light reaction (photon release) or proton release upon nucleotide incorporation. Labelled nucleotides based chemistries include ligation based processes with fluorescently labeled oligonucleotides of known sequence (SOLiD) or a sequencing by synthesis process (Illumina). During Illumina sequencing, four differently labeled nucleotides are flushed over the flow cell in multiple cycles, depending on the desired read length. During 454 and Ion PGM sequencing unlabeled nucleotides are flushed in a sequential order over the flow cell. Incorporation is detected via a coupled light reaction (454) or the detection of proton release during nucleotide incorporation (Adapted from [Knief, 2014]).

1.4.4 Meta-omic NGS data analysis

This work focuses mainly on the analysis of high-throughput MG and MT datasets generated from NGS platforms. Given that both these datasets comprise NGS reads, the general analysis procedure is largely similar, using sequence based methods (i.e. alignment- and *k*mer-based methodologies). Accordingly, computational solutions for MG and MT data analyses can be broadly classified into reference-dependent or reference-independent (*de novo* assembly-based) methods [Segata *et al.*, 2013]. In addition, concomitant (or coupled) MG and MT dataset, especially such as those produced via the protocol described in **Section 1.4.2**, are complementary and suitable for integrated analyses. However, there is presently a lack of standardized tools to perform integrated analysis of coupled MG and MT data. To that end, this work particularly focuses on reference-independent methodologies for the integrated analysis of coupled MG and MT data.

Reference-based analyses

NGS-based reference-dependent methods rely on detecting similarity between NGS reads to reference databases, such as a compendium of isolate genomes, gene catalogues and/or existing MG data. There are multiple ways of performing reference-based analyses, whereby the most widely used method is the alignment of NGS reads to reference databases using NGS read alignment tools. Modern NGS read alignment tools, such as the Burrows-Wheels Aligner (BWA [Li and Durbin, 2009]) and Bowtie2 [Langmead *et al.*, 2009], can rapidly align massive numbers of NGS reads through the indexing a given databases using the Burrows-Wheels transformation [Burrows and Wheeler, 1994; Langmead *et al.*, 2009; Li and Durbin, 2009]. In essence, alignment to a set of isolate genome sequences will provide information about the community structure while the alignment to an annotated gene catalogue will provide information about the community genetic potential. Moreover, the advent of pseudo alignment methodologies, such as Kallisto [Bray *et al.*, 2016] show promise in the application of microbial community based NGS data analyses [Schaeffer *et al.*, 2015; Teo and Neretti, 2016].

In addition to the aforementioned classical alignment methods, more recent methods, such as Kraken [Wood and Salzberg, 2014], Diamond [Buchfink *et al.*, 2015] and Kaiju [Menzel *et al.*, 2016], are currently available for rapid taxonomic assignments of NGS reads. On the other hand, functional information is obtained through gene annotation, which is a two-step process: i) genes are predicted from NGS reads using tools, such as MetaGeneMark [Trimble *et al.*, 2012] and ii) the predicted genes are annotated *in silico* using searches against genes with known functions using either sequence alignment based programs, such as the classical Basic Local Alignment Search Tool (blast [Pruitt *et al.*, 2002; Johnson *et al.*, 2008]) and/or Hidden Markov model (HMM) profile based searches [Eddy, 1996, 1998; Finn *et al.*, 2011]. Last but not least, there are user-friendly web based pipelines, such as the Metagenomic - Rapid Annotation using Subsystem Technology (MG-RAST) server [Meyer *et al.*, 2008] that performs both taxonomic assignments and gene annotation in a single workflow, directly using NGS reads. At present, most MT data analyses typically involve reference-based [Leimena *et al.*, 2013; Martinez *et al.*, 2016; Westreich *et al.*, 2016] or MG-dependent analysis workflows [Franzosa *et al.*, 2014; Bremges *et al.*, 2015; Satinsky *et al.*, 2015].

Reference-based methods provide a means for rapid analysis of MG and MT data. However, the quality of the analyses are highly dependent on the information contained within the selected reference databases. Therefore, a major drawback of such methods are the large number of NGS reads from uncultured species, divergent strains and/or unclassified genes that cannot be aligned (unmappable) due to their dissimilarity from the reference databases, and thus not considered during data analysis, thereby resulting in the loss of potentially useful information. This fact can be highlighted, based on analyses of MG data from the human gastrointestinal tract microbiome (arguably the best characterized microbial community in terms of culture-derived isolate genomes), approximately 43 % of the data are typically not mappable to the available isolate genomes [Sunagawa *et al.*, 2013]. Overall, reference-based approaches by themselves exhibit limitations, which may result in the omission of potentially useful information.

Reference-independent analyses

Conversely, reference-independent methodologies involve *de novo* assembly of the short sequencing reads into longer contiguous sequences (i.e. contigs). A simplified schema for reference-independent analysis is shown in **Figure 1.8**.

NGS reads are typically preprocessed (**Figure 1.8**) prior to *de novo* assembly to: i) remove low quality bases within reads, ii) low quality reads, iii) artificially introduced sequencing adapters and iv) other potentially unwanted sequences (i.e. human derived NGS reads from human microbiome MG data and rRNA sequences from MT data). Preprocessing was shown to increase the quality of contigs obtained in the downstream *de novo* assembly [Mende *et al.*, 2012]. *De novo* assemblies can be performed using two generalized *de novo* assembly algorithms including the overlap layout consensus (OLC) and the de Bruijn graph (DBG) [Li *et al.*, 2012]. The OLC method is a classical *de novo* assembly method developed to assemble data from the Sanger sequencing platform [Sanger *et al.*, 1977; Staden, 1979]. The method can be broadly categorized by three major steps: i) overlap - all reads are aligned against each other to determine overlapping regions ii) layout - a graph is formulated based on the overlapping regions of the reads and iii) consensus - a consensus is generated based on the overall agreement/similarity of overlapping reads [Staden, 1979; Li *et al.*, 2012]. Widely used OLC-based assemblers include Celera Assembler [Myers *et al.*, 2000], Newbler [Margulies *et al.*, 2005], Cap3 [Huang and Madan, 1999] and Phrap [de la Bastide *et al.*, 2007]. On the other hand, DBG [de Bruijn and van der Woude, 1946] assemblers work in three stages: i) defining all *k*mers (i.e. stretches of nucleotides of length *k*) within the collection of NGS reads, ii) the construction of the DBG based on the identified *k*mers and iii) inferring assembled sequence from the DBG [Idury and Waterman, 1995; Pevzner *et al.*, 2001; Compeau *et al.*, 2011]. Commonly used single-genome DBG assemblers include Velvet [Zerbino and Birney, 2008], ABySS [Simpson *et al.*, 2009] and SOAPdenovo [Luo *et al.*, 2012]. In addition, performing assemblies over multiple *k*mer sizes, such as those carried out by IDBA, were shown to improve quality of the assembly [Peng *et al.*, 2010]. At present, DBG assemblers are more widely used for genome assemblies due to their capability of handling the massive number of short NGS reads generated from the Illumina platform [Li *et al.*, 2012].

However, characteristics of microbial communities (described in **Section 1.1**), in particular the multiple genomes stemming of different constituent populations, occurring differing abundances, and thus result in differing sequencing depths within the generated MG data. Consequently, the aforementioned assembly programs do not take into account the characteristics of MG data, as they were originally designed for isolate genome assemblies. Fortunately, a wide array of MG-specific assemblers, such as MetaVelvet [Namiki *et al.*, 2012], IDBA-UD [Peng *et al.*, 2012] and MEGAHIT [Li *et al.*, 2015] have been developed to account for the uneven sequencing depth of MG data. In particular, given sufficient sequencing depth, current *de novo* MG assemblers are highly effective for medium complexity communities, such as those present within BWWT plant microbial communities [Segata *et al.*, 2013; Muller *et al.*, 2014a]. It was also shown that sequential use of DBG assemblers, such as IDBA-UD and MEGAHIT and OLC assemblers, such as Cap3, result in improved MG assemblies [Deng *et al.*, 2015; Lai *et al.*, 2015].

Contigs generated from *de novo* assemblies are subjected to various analyses to obtain meaningful information from these stretches of nucleotide sequences. These procedures include, but are not limited to the annotation of the sequences based on known taxonomy and/or gene functions (**Figure 1.8**). It has been demonstrated that the assembly of NGS reads into longer contigs greatly improves the taxonomic assignments and annotation of genes, as opposed to their direct identification from NGS reads [Nalbantoglu *et al.*, 2011; Celaj *et al.*, 2014]. The same tools described in **Section 1.4.4** may be applied at the contig level, for taxonomic assignment and gene annotation. To that end, automated reference-independent bioinformatic pipelines have so far been mainly developed for MG data. These include MOCAT [Kultima *et al.*, 2012] and MetAMOS [Treangen *et al.*, 2013] which incorporate the entire process of MG data analysis, ranging from preprocessing of sequencing reads, *de novo* assembly and post-assembly analysis (read alignment, taxonomic classification, gene annotation, etc., **Figure 1.8**). MOCAT has been used in large-scale studies, such as those within the MetaHIT Consortium [Qin *et al.*, 2010; Li *et al.*, 2014], while MetAMOS is a flexible pipeline which allows customizable workflows [Treangen *et al.*, 2013].

Similar to MG data, a comparative study by Celaj *et al.* [2014], has recently demonstrated that reference-independent approaches for MT data analysis are extremely useful when using either specialized MT assemblers (e.g. IDBA-MT [Leung *et al.*, 2013; Celaj *et al.*, 2014]), MG assemblers (e.g. IDBA-UD [Peng *et al.*, 2012; Leung *et al.*, 2013, 2014] and MetaVelvet [Namiki *et al.*, 2012; Celaj *et al.*, 2014]) or even single-species transcriptome assemblers (e.g. Trinity [Grabherr *et al.*, 2011; Celaj *et al.*, 2014]). All the aforementioned assemblers are able to handle the uneven sequencing depths, which is a common characteristic of MG and MT data. Importantly, the assembly of MT data may result in the assembly of novel genes (as in MG data) while providing access to additional components, such as RNA viruses, which would not be possible with a reference-based method and/or MG-only method. These include, but are not limited to RNA viruses/phages and/or genes that are lowly abundant on a genomic (or MG) level, but are highly expressed within a given microbial community, thus enhancing the overall information gain.

While this work mainly focuses on the analysis of MG and MT data, it is worth noting that the quality of metaproteomic data analysis is highly dependent on the underlying database used for the peptide searches. It was previously shown that amino acid sequences predicted from MG data improves overall detection of peptides from concomitant metaproteomic data, further highlighting the advantages of reference-independent MG and MT analysis methods in generating customized amino acid sequence databases for downstream proteomic analyses [Ram *et al.*, 2005; Heintz-Buschart *et al.*, 2016].

Despite the highlighted advantages of reference-independent analysis methods, *de novo* assembly-based methods are often not preferred due to high computing requirement and long runtimes. However, these issues are currently mitigated by the development of the rapid and memory efficient *de novo* MG assemblers such MEGAHIT [Li *et al.*, 2015, 2016].

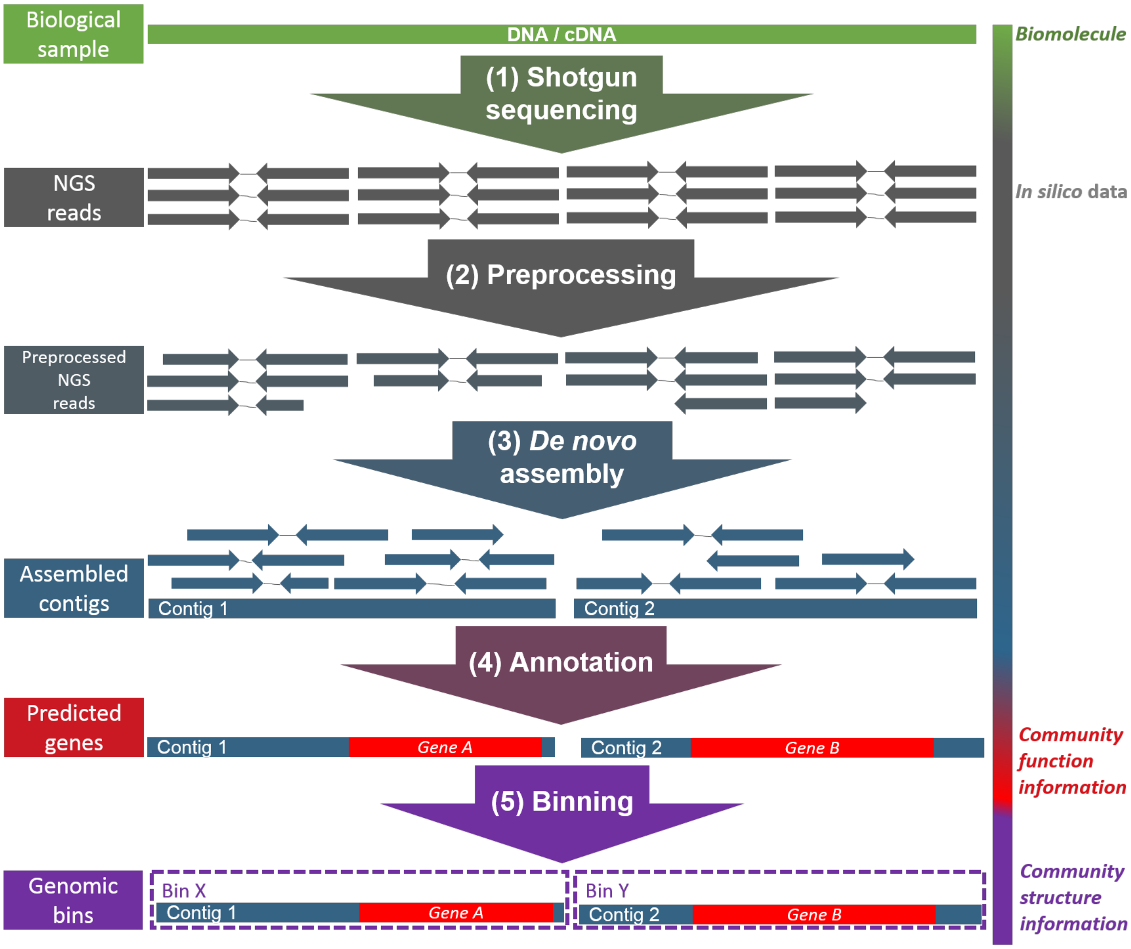


Figure 1.8: Simplified workflow for reference-independent metagenomic and/or metatranscriptomic analyses.The boxes on the left represent the data end product of each step.The bar on the right shows transition from biomolecules to *in silico* data, which is in turn converted to information about community structure and/or function. The process begins with a DNA or RNA (complementary DNA-cDNA) biological sample, which is followed by multiple steps including: **(1)** Shotgun next-generation sequencing (NGS) to generate *in silico* representations of the biomolecules. **(2)** Preprocessing of NGS reads to remove low quality bases and/or artificially introduced sequences, such as sequencing adapters. **(3)** *De novo* assembly involves the use of programs to align/overlaps reads against each other to generate longer contiguous sequences, i.e. contigs, representing a consensus of all the overlapping reads. **(4)** Annotation is a process of predicting gene sequences that occur within the contigs and assigning functions to those genes based on similarity to known genes (not shown in figure). **(5)** Binning is a method of separating/clustering the assembled contigs into genomic bins which should ideally represent a single population-level genome of an organism. Binning is exclusively applied to MG assemblies.

Binning

The process of grouping/clustering contigs generated by MG assemblies is described as "binning" (**Figure 1.8**). Binning may also be applied on the read level, although this is not typically performed due to the short read length of NGS technologies (**Section 1.4.3**). The clusters obtained via binning are assumed to represent genomes of single microbial populations, i.e. population-level genomes [Laczny, 2015; Laczny *et al.*, 2016].

Binning of MG-derived sequences (i.e. sequencing reads or contigs) may be carried out in a supervised or unsupervised manner. Supervised binning methods are analogous to reference-based methods (refer to **Section 1.4.4**), whereby either MG reads or assembled contigs are clustered based on alignment against genomes of known organisms, to determine population-level genomes [Laczny, 2015]. On the other hand, unsupervised binning methods cluster sequences based on nucleotide signature and/or abundance information [Laczny, 2015]. Unsupervised binning methods utilize assembled MG-contigs as input. These methods require the input sequences to be of a certain length, in order to be effective [Laczny, 2015; Alneberg *et al.*, 2014; Nielsen *et al.*, 2014]. Despite the aforementioned prerequisite of a *de novo* assembly, unsupervised binning methods enable the resolution and/or retrieval of population-level genomes from hitherto undescribed taxa, consequently resulting in the recovery of putatively novel genes, thereby allowing more of the data to be mapped and exploited for analysis [Segata *et al.*, 2013; Treangen *et al.*, 2013; Narayanasamy *et al.*, 2015; Hugerth *et al.*, 2015; Albertsen *et al.*, 2013a; Muller *et al.*, 2014a; Laczny *et al.*, 2016]. Furthermore, unsupervised binning methods are highly complementary to the current work, which focuses on reference-independent analysis of microbial community datasets.

In light of the advantages of unsupervised binning methods, recent years have seen the development and application of a wide range of automated [Imelfort *et al.*, 2014; Nielsen *et al.*, 2014; Wu *et al.*, 2014; Eren *et al.*, 2015; Kang *et al.*, 2015; Heintz-Buschart *et al.*, 2016; Alneberg *et al.*, 2014; Dick *et al.*, 2009] and manual [Dick *et al.*, 2009; Albertsen *et al.*, 2013a; Laczny *et al.*, 2014, 2015, 2016; Eren *et al.*, 2015] binning tools and/or methods. Some of the first unsupervised binning methods used nucleotide signature as a means of clustering MG sequences [Dick *et al.*, 2009; Laczny *et al.*, 2014, 2015]. More recently, methods utilize abundance information for clustering MG sequences [Albertsen *et al.*, 2013a]. Abundance information can be estimated through the mapping of reads to the assembled contigs [Laczny, 2015; Albertsen *et al.*, 2013a]. The current state-of-art binning methods are able to utilize both nucleotide signature and abundance information for the clustering of MG sequences [Heintz-Buschart *et al.*, 2016; Laczny *et al.*, 2015; Wu *et al.*, 2014; Kang *et al.*, 2015]. Furthermore, the lowered cost of generating sequencing data promotes sequencing of multiple MG samples from a given microbial community. These include multiple replicates, spatial samples, time-series-based samples and/or large cohorts. Several binning methods leverage the abundance information from multiple samples to further enhance the clustering of the sequences [Nielsen *et al.*, 2014; Imelfort *et al.*, 2014; Kang *et al.*, 2015; Wu *et al.*, 2014; Alneberg *et al.*, 2014; Eren *et al.*, 2015].

Overall, unsupervised binning in combination with *de novo* metagenomic assemblies bypass the need for culture-dependent methods to access potentially novel microbial taxa and functionalities (**Section 1.1**) [Narayanasamy *et al.*, 2015]. Genomic information derived from these methods are vital for the meaningful interpretation of additional functional omic data [Narayanasamy *et al.*, 2015; Muller *et al.*, 2013; Segata *et al.*, 2013; Heintz-Buschart *et al.*, 2016].

1.4.5 Multi-omic analyses of meta-omic data

Multi-omic analyses have already been applied to provide novel insights into microbial community structure and function in various different ecosystems. Some of them include studies of the human gut microbiome [Franzosa *et al.*, 2014], aquatic microbial communities from the Amazon river [Satinsky *et al.*, 2015], soil microbial communities [Hultman *et al.*, 2015; Beulig *et al.*, 2016], production-scale biogas plants [Bremges *et al.*, 2015], hydrothermal vents [Urich *et al.*, 2014] and microbial communities from biological wastewater treatment plants [Muller *et al.*, 2014a; Roume *et al.*, 2015]. These studies employed differing ways for analysing the data including reference-based approaches [Franzosa *et al.*, 2014; Urich *et al.*, 2014; Satinsky *et al.*, 2015], MG assembly-based approaches [Bremges *et al.*, 2015; Hultman *et al.*, 2015], MT assembly-based approaches [Urich *et al.*, 2014], and integrated analyses of the meta-omic data [Muller *et al.*, 2014a; Urich *et al.*, 2014; Roume *et al.*, 2015; Heintz-Buschart *et al.*, 2016]. An extensive list of studies that leveraged on multi-omic data sets are listed in **Table 1.1**.

Although high-throughput MG and MT data allow deep profiling of microbial communities given the relatively low cost of generating sequencing data, existing sequence-based approaches do have some important limitations. Given the availability of omic technologies and their falling costs (in particular for metagenomics and metatranscriptomics), fully integrated multi-omic analyses should be applied routinely in the study of microbial consortia for greater effectiveness. For instance, despite this wealth of information, current MG assemblies and analysis schemes, MG (and MT) data resulting from the use of current short-read NGS technologies and assembly approaches do not allow the comprehensive resolution of microdiversity, e.g. genetic heterogeneity of microbial populations [Wilmes *et al.*, 2009]. Furthermore, RNAseq technologies are subject to biases stemming from the extensive, yet compulsory pre-processing steps [Lahens *et al.*, 2014], thereby affecting the resulting MT data.

Integrated omic based analyses are currently gaining momentum towards providing enhanced understanding of community structure, function and dynamics *in situ*. This is evident based on studies that further extended the advantages of multi-omic analyses by integrating different omic datasets [Muller *et al.*, 2014a; Roume *et al.*, 2015; Heintz-Buschart *et al.*, 2016]. The backbone of these aforementioned studies are the *de novo* co-assemblies of MG and MT data which promises higher quality compared with conventional *de novo* MG assemblies, due to the ability to reconstruct and resolve genomic complements of low abundance (i.e. low MG coverage) yet highly active populations (i.e. high MT coverage for expressed genes; Muller *et al.* [2014a]; Roume *et al.* [2015]; Heintz-Buschart *et al.* [2016]; Zengler and Palsson [2012]). Such co-assemblies allow high-quality population-level genomic reconstructions after the application of binning/classification methods, such as those developed for a single sample [Wu *et al.*, 2014; Laczny *et al.*, 2015] or for spatio-temporally resolved samples [Albertsen *et al.*, 2013a; Alneberg *et al.*, 2014; Nielsen *et al.*, 2014; Kang *et al.*, 2015]. Furthermore, co-assemblies of MG and MT data allow the resolution of genetic variations with higher confidence through replication and highlights their potential relative importance, thereby allowing more detailed short-term evolutionary inferences regarding specific populations and while increasing sensitivity for downstream metaproteomic analysis [Muller *et al.*, 2014a; Roume *et al.*, 2015; Heintz-Buschart *et al.*, 2016]. It is important to highlight that despite the increasing number of studies applying multi-omics to study microbial consortia (**Table 1.1**), there is yet to be a study that leverages this data to perform a detailed study of bacteriophage and host interactions.

Table 1.1: Multi-omic studies of microbial communities.

Reference	Environments	Meta-omics				Comments
		MG	MT	MP	MM	
Lim *et al.* [2013]	Cystic fibrosis patients airways	Roche 454	Roche 454	-	-	• Reference-based approaches • Virome + microbiome
Franzosa *et al.* [2014]	Human gut microbiome	Illumina HiSeq	Illumina HiSeq	-	-	• Reference-based approaches
Heintz-Buschart *et al.* [2016]	Human gut microbiome	Illumina HiSeq	Illumina HiSeq	LC-MS	-	• DNA and protein co-extracted • Integrated data analyses of the meta-omic data (MG + MT co-assembly, peptides database derived from the assembly)
Erickson *et al.* [2012]	Crohn‘s deseased human gasrointestinal tract	Roche 454	-	2D-LC-MS/MS	-	• Mixed approaches (reference-based for short fragments or reads that failed to assemble) • Integrated data analyses of the meta-omic data (MG + MT co-assembly, peptides database derived from the assembly)
He *et al.* [2013]	Termites gut	Roche 454	Illumina GAIIx	-	-	• Integrated data analyses of the meta-omic data (MG + MT co-assembly) • 16S rRNA gene amplicon sequencing was also done
Zhang *et al.* [2015]	Mouse gut	Illumina HiSeq	-	-	^{1}H NMR	• Reference-based approaches (MG-RAST for DNA and Human Metabolome Database for metabolites)
Hua *et al.* [2015]	Acid mine drainage	Illumina HiSeq	Illumina HiSeq	-	-	• Integrated data analyses of the meta-omic data (MG + MT assembly merging with CD-hit)
Denef *et al.* [2010]	Acid mine drainage	ABI PRISM 3730	-	nano-2D-LC-MS/MS	-	• Analyses of 2 omic layers produced in 2 previous studies
Wilmes *et al.* [2010]	Acid mine drainage	-	-	nano-2D-LC-MS/MS	RPC and NPC	• Integrated metaproteomic and (meta-)metabolomic data
Goltsman *et al.* [2009]	Acid mine drainage	ABI PRISM 3730	-	nano-2D-LC-MS/MS	-	• Integrated MG and MP analyses
Bertin *et al.* [2011]	Acid mine drainage	ABI3730	-	PAGE-LC-MS/MS	-	• Integrated data analyses of the meta-omic data (peptides database derived from the assembly)
Rogers *et al.* [2013]	Subglacial antarctic lakes	Roche 454	Roche 454	-	-	• DNA and RNA co-extracted • Reference-based approaches (MG-RAST, Galaxy, Batch Mega-BLAST)
Nolla-Ardèvol *et al.* [2015]	Bioreactor seed with lake sediments	Ion Torrent PMG	Ion Torrent PMG	-	-	• MG assembly-based approaches, assembly binning

Continued on next page...

Reference	Environments	Meta-omics				Comments
		MG	MT	MP	MM	
Glass *et al.* [2014]	Lake mats	Illumina Hi Seq	Illumina Hi Seq	-	-	• MG assembly-based approaches
Hawley *et al.* [2014]	Fjords	ABI PRISM 3730	-	LC-MS/MS	-	• Integrated data analyses of the meta-omic data (peptides database derived from the assembly)
Satinsky *et al.* [2015]	Amazon river	Illumina HiSeq	Illumina HiSeq	-	-	• All filtration and stabilization was completed within 30 min of water collection • Reference-based approaches (RefSeq Protein database or RAPSearch2
Satinsky *et al.* [2014]	Amazon river plum	Illumina GAIIx, HiSeq, or MiSeq	Illumina GAIIx, HiSeq, or MiSeq	-	-	• All filtration and stabilization was completed within 30 min of water collection • MT on rRNA depleted fraction and MT on poly-A tailed mRNA (eukaryotic nuclear mRNA) • Reference-based approaches (RefSeq Protein database or RAPSearch2)
Grob *et al.* [2015]	Seawater	Roche 454	-	LC-MS/MS	-	• DNA and protein co-extracted • Mixed approaches (MG assembly, peptide also tested against NCBInr database) • Integrated data analyses of the meta-omic data (peptides database derived from the assembly) • DNA-SIP and protein-SIP
Andrade *et al.* [2015]	Seawater	Illumina MiSeq	-	-	GC-FID (lipidomics)	• MG assembly-based approaches
Gilbert *et al.* [2010]	Seawater	Roche 454	Roche 454	-	-	• All filtration and stabilization was completed within 30 min of water collection • Reference-based approaches (MG-RAST)
Wemheuer *et al.* [2015]	Seawater	Roche 454, Illumina GAIIx	Roche 454, Illumina GAIIx	-	-	• DNA and RNA co-extracted • Mixed approaches (15 reference genomes + MG assembly)
Mason *et al.* [2012]	Seawater	Illumina GAIIx	Illumina GAIIx	-	GC-MS	• MG and MT assembly-based approaches • Hydrocarbon analysis • Single cell WGS

Continued on next page...

Reference	Environments	Meta-omics				Comments
		MG	**MT**	**MP**	**MM**	
Martínez *et al.* [2013]	Microcosms of seawater	Roche 454	Roche 454	-	-	• Reference-based approaches (NCBI-nr database, KEGG database)
Wu *et al.* [2013]	Deep Sea	ABI 373, Roche 454	ABI 373, Roche 454	-	-	• Reference-based approaches (MG-RAST)
Shi *et al.* [2011]	Deep sea	Roche, GS20	Roche, GS20	-	-	• Reference-based approaches (NCBI-nr, SEED, GOS protein cluster and reference genomes)
Bargiela *et al.* [2015]	Sea sediment	-	-	1D-PAGE-LC-MS/MS	GC-FID and LC-Q-TOF-MS	• Integrated metaproteomic and (meta-)metabolomic analyses
Stokke *et al.* [2012]	Deep sea sediment	Roche 454	-	NanoLC-LTQ	-	• Integrated data analyses of the meta-omic data (peptides database derived from the assembly)
Kimes *et al.* [2013]	Deep sea sediment	Roche 454	-	-	GC-MS and LC-MS	• Reference-based approaches (MG-RAST)
Urich *et al.* [2014]	Hydrothermal vents microbial mats	Roche 454	Roche 454	1D SDS-PAGE then LTQ-Orbitrap-MS/MS	-	• The time from sampling until processing was about 90 min • DNA and RNA co-extracted • Reference-based approaches, MT assembly-based approaches or rRNA
Hultman *et al.* [2015]	Soil (permafrost)	Illumina GAII	Illumina HiSeq	2D-LC- MS/MS	-	• MG assembly-based approaches • Integrated data analyses of the meta-genomic and metaproteomic data (peptides database derived from the assembly), but not metatranscriptomic
Beulig *et al.* [2016]	Soil microbial communities	Illumina HiSeq 2500	Illumina HiSeq 2500			• rRNA-based analysis using BLASTN and MEGAN • Protein prediction from NGS MT reads and annotation of predicted amino acid sequences
Butterfield *et al.* [2016]	Soil microbial communities	Illumina	-	LC-MS/MS	LC-MS	• Integrated data analyses of the meta-omic data (peptides database derived from the assembly)
Liu *et al.* [2012]	Sponge	Roche 454	-	1D-PAGE-LC-MS/MS	-	• MG assembly-based approaches • Integrated data analyses of the meta-omic data (peptides database derived from the assembly)
Kleiner *et al.* [2012]	Gutless worm symbionts	-	-	1D-PAGE-LC-MS/MS and 2D-LC-MS/MS	GC-MS, LC-MS, and ^{1}H-NMR	• Symbionts metagenome obtained in a previous study • Integrated data analyses of the meta-omic data (peptides database derived from the assembly)

Continued on next page...

Reference	Environments	Meta-omics				Comments
		MG	MT	MP	MM	
Aylward *et al.* [2012]	Ant garden	Roche 454	-	LC-MS/MS	-	• Mixed approaches (reference genomes from the same environment + MG assembly)
Delmotte *et al.* [2009]	Phyllosphere	Roche 454	-	1D-PAGE-LC-MS/MS	-	• Mixed approaches (MG assembly +RefSeq or RefSeq only)
Knief *et al.* [2012]	Phyllosphere and rhizosphere	Roche 454	-	1D-PAGE-LC-MS/MS	-	• DNA and protein co-extracted • Mixed approaches (MG assembly + Uniref100)
D'haeseleer *et al.* [2013]	Compost	Roche 454 and Illumina	-	1D- LC-MS/MS and 2D- LC-MS/MS	-	• Mixed approaches (MG assembly + reference genomes from the same environment + CAZy and FOLy databases)
Bremges *et al.* [2015]	Production-scale biogas plants	Illumina GAII	Illumina GAII	-	-	• MG assembly-based approaches
Hanreich *et al.* [2013]	Biogas plant	Roche 454	-	LC-MS/MS	-	• Mixed approaches (MG assembly, peptide also tested against NCBInr database)
Ortseifen *et al.* [2016]	Biogas plant	Illumina HiSeq	-	2D-SDS-PAGE then MS/MS	-	• Mixed approaches (MG assembly, peptide also tested against TrEMB)
Xia *et al.* [2014]	Bioreactor (sludge + cellulose)	Illumina Hiseq	Illumina Hiseq	-	-	• MG assembly-based approaches
Muller *et al.* [2014b]	Biological wastewater treatment plants(LAMPs)	Illumina GAII	Illumina GAII	LC-MS/MS	GC-MS/MS	• DNA, RNA, proteins and metabolites co-extracted • Integrated analyses of the meta-omic data (DNA and RNA co-assembled, peptides database derived from the assembly) • Metabolites quantified by GC-MS/MS are lipids, protein and carbohydrate were quantified by colorimetric tests
Roume *et al.* [2015]	Biological wastewater treatment plants (LAMPs)	Illumina GAII	Illumina GAII	LC-MS/MS	-	• DNA, RNA, proteins and metabolites co-extracted • Integrated analyses of the meta-omic data (DNA and RNA co-assembled, peptides database derived from the assembly)
Yu and Zhang [2012]	Biological wastewater treatment plants (activated sludge)	Illumina Hi-seq	Illumina Hi-seq	-	-	• Reference-based approaches (MG-RAST)
Wilmes *et al.* [2008]	Laboratory-scale EBPR	Sanger	-	2D-LC-MS/MS	-	• Metagenomic data generated and published in another study • Integrated data analyses of the meta-omic data (peptides database derived from the assembly)
Barr *et al.* [2016]	Laboratory-scale EBPR	Illumina HiSeq	-	LC-MS/MS	-	• Integrated data analyses of the meta-omic data (peptides database derived from the MG assembly)

Continued on next page...

Reference	Environments	Meta-omics				Comments
		MG	MT	MP	MM	
Albertsen *et al.* [2013b]	EBPR	Illumina GAIIx	-	SDS-PAGE and nanoLC-QTOF MS	-	• Integrated data analyses of the meta-omic data (peptides database derived from the MG assembly) • MP focused on extracellular proteins and polysaccharide forming proteins

^{1}**H NMR**: proton nuclear magnetic resonance
1D: one-dimensional
2D: two-dimensional
ABI: Applied Biosystems
FID: flame Ionization Detector
GA: Genome Analyzer
GC: gas-chromatography
LC: liquid-chromatography
MG: metagenomic
MT: metatranscriptomic
MP: metaproteomic
MM: (meta-)metabolomic
MS: mass-spectrometry
NPC: normal-phase chromatography
PAGE: polyacrylamide gel electrophoresis
PMG: Personal Genome Machine
RPC: reverse-phase chromatography
SDS: sodium dodecyl sulfate
TOF: time of flight

1.5 Objectives of this work

There is presently a need for the application of multi-omic analyses for the study of phage and host interactions, especially with the provision of MG data (DNA phages) and to a certain extent MT data (RNA phages) that enables access to the phage genomic complements. The extended use of MT data provides an additional view in terms of expression activity in both hosts and phages, especially with regards to bacterial antiviral defence and replication of phages, while providing the possibility to assemble RNA-based invasive genetic elements. The identification of a suitable natural model microbial community which enables time-series sampling have made longitudinal sample sets available, thereby enabling the study of phage-host interactions and dynamics. This work has for objectives to: i) develop a standardized method for effective integrated analysis of MG and MT data and ii) studying phage-host interactions in a natural community using integrated multi-omics.

Although the studies that applied multi-omic (and integrated) analysis methodologies clearly demonstrated the advantages of such analyses by providing deeper insights into structure and function of microbial consortia from multiple environments (described in **Section 1.4.5**), there still lacks standardized and reproducible dry-lab workflows for integrating and analysing the multi-omic data. Such standardized approaches are required to compare results between different samples (i.e. cohorts, spatial- and time-resolved), studies and microbial systems of study. In addition, integrative analysis of MG and MT data, specifically the *de novo* co-assemblies of yet to be formally evaluated, benchmarked and documented to demonstrate the benefits of integrated omic analysis. Therefore, the first part of this work focuses on highlighting the benefits of integrated analysis of MG and MT datasets, and consequently the development of a standardized and reproducible reference-independent bioinformatic pipeline for integrated omic analyses. Extensive benchmarking was performed on the output from our method, comparing it to existing methods. The pipeline was mainly applied to the model microbial system.

MG datasets greatly increase the resource and information with regards to bacteriophages (**Section 1.3**). However, there is still lack an understanding of the dynamics of bacteriophages and their associated host populations within a naturally occurring microbial system. Therefore, the second part of this work focused on elucidating phage and host dynamics within a naturally occurring model microbial system, using an unprecedented temporal-based, multi-omic datasets. The output obtained using the integrated multi-omic pipeline, described in the first part of this work, was supplemented with specialized analysis to extract information with regards to the CRISPR-*Cas* system and bacteriophages. The CRISPR information was used to provide a broad summary of community-level CRISPR dynamics over time. The analysis then focused on the population-level analysis through the identification of two bacterial host populations and their corresponding 158 putative phage populations. The dynamics of the phage and host populations were then described by following phage and host abundances over time. Furthermore, this work extended the use of MT data to investigate putative RNA-based invasive elements, in addition to observing the expression of bacterial host *cas* genes. In summary, this work involved the development of a new bioinformatics pipeline for the integration of metagenomics and metatranscriptomic data and its subsequent application to the study of phage-host dynamics.

CHAPTER 2

A PIPELINE FOR REPRODUCIBLE REFERENCE-INDEPENDENT INTEGRATED METAGENOMIC AND METATRANSCRIPTOMIC ANALYSES

This chapter describes a bioinformatic pipeline designed specifically for integrated omics of coupled metagenomic and metatranscriptomic datasets. Moreover, the initial conception and workflow design was based the involvement of the author in an integrated multi-omic study of the LAMPs. Accordingly, these includes concepts and material from the following published first- and co-author peer-reviewed publications:

Shaman Narayanasamy†, Yohan Jarosz†, Emilie E.L. Muller, Anna Heintz-Buschart, Malte Herold, Anne Kaysen, Cédric C. Laczny, Nicolàs Pinel, Patrick May, Paul Wilmes (2016) IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biology* **17**: 260. [**Appendix A.2**]

Emilie E.L. Muller, Nicolás Pinel, Cédric C. Laczny, Michael R. Hoopmann, **Shaman Narayanasamy**, Laura A. Lebrun, Hugo Roume, Jake Lin, Patrick May, Nathan D. Hicks, Anna Heintz-Buschart, Linda Wampach, Cindy M. Liu, Lance B. Price, John D. Gillece, Cédric Guignard, Jim M. Schupp, Nikkos Vlassis, Nitin S. Baliga, Robert L. Moritz, Paul S. Keim, Paul Wilmes (2014). Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nature Communications* **5**: 5603. [**Appendix A.3**]

---

† Co-first author

2.1 Abstract

Existing workflows for the analysis of multi-omic microbiome datasets are lab-specific and often result in sub-optimal data usage. Here we present IMP, a reproducible and modular pipeline for the integrated and reference-independent analysis of coupled metagenomic and metatranscriptomic data. IMP incorporates robust read preprocessing, iterative co-assembly, analyses of microbial community structure and function, automated binning as well as genomic signature-based visualizations. The IMP-based data integration strategy enhances data usage, output volume and output quality as demonstrated using relevant use-cases. Finally, IMP is encapsulated within a user-friendly implementation using Python and Docker. IMP is available at `http://r3lab.uni.lu/web/imp/` (MIT license).

2.2 Background

The previous chapter summarized the various studies that applied multi omic data analyses, demonstrating that such studies are becoming more prevalent (**Section 1.4.5**). Due to the absence of established tools/workflows to handle multi-omic datasets, most of the aforementioned studies utilized non-standardized, *ad hoc* analyses, mostly consisting of custom workflows, thereby creating a challenge in reproducing the analyses [Treangen *et al.*, 2013; Belmann *et al.*, 2015; Di Tommaso *et al.*, 2015; Kenall *et al.*, 2015]. Given that the lack of reproducible bioinformatic workflows is not limited to those used for the multi-omic analysis of microbial consortia [Treangen *et al.*, 2013; Belmann *et al.*, 2015; Di Tommaso *et al.*, 2015; Kenall *et al.*, 2015], several approaches have recently been developed with the explicit aim of enhancing software reproducibility. These include a wide range of tools for constructing bioinformatic workflows [Köster and Rahmann, 2012; Amstutz *et al.*, 2016; Leipzig, 2016] as well as containerizing bioinformatic tools/pipelines using Docker [Belmann *et al.*, 2015; Bremges *et al.*, 2015; Di Tommaso *et al.*, 2015; Leipzig, 2016]. Here, we present IMP, the Integrated Meta-omic Pipeline, the first open source *de novo* assembly-based pipeline which performs standardized, automated, flexible and reproducible large-scale integrated analysis of combined multi-omic (MG and MT) datasets. IMP incorporates robust read preprocessing, iterative co-assembly of metagenomic and metatranscriptomic data, analyses of microbial community structure and function, automated binning as well as genomic signature-based visualizations. We demonstrate the functionalities of IMP by presenting the results obtained on an exemplary data set. IMP was evaluated using datasets from ten different microbial communities derived from three distinct environments as well as a simulated mock microbial community dataset. We compare the assembly and data integration measures of IMP against standard MG analysis strategies (reference-based and reference-independent) to demonstrate that IMP vastly improves overall data usage. Additionally, we benchmark our assembly procedure against available MG analysis pipelines to show that IMP consistently produces high-quality assemblies across all the processed datasets. Finally, we describe a number of particular use cases which highlight

2.3 Methods

The details of the IMP workflow, implementation and customizability is described in further detail. We also describe the additional analyses carried out for assessment and benchmarking of IMP.

2.3.1 Details of the IMP implementation and workflow

The details of the IMP workflow, implementation and customizability is A Python (ver. 3) wrapper script was implemented for user-friendly execution of IMP via the command line. The full list of dependencies, parameters (see below) and documentation are available on the IMP website (`http://r3lab.uni.lu/web/imp/doc.html`). Although IMP was designed specifically for integrated analysis of MG and MT data, it can also be used for single MG or MT analyses, as an additional functionality.

Reproducibility

IMP is implemented around a Docker container that runs the Ubuntu 14.04 operating system, with all relevant dependencies. Five mounting points are defined for the Docker container with the -v option: i) input directory, ii) output directory, iii) database directory, iv) code directory, and v) configuration file directory. Environment variables are defined using the -e parameter, including: i) paired MG data, ii) paired MT data, and iii) configuration file. The latest IMP Docker image will be downloaded and installed automatically upon launching the command, but users may also launch specific versions based on tags or use modified/customized versions of their local code base (documentation at `http://r3lab.uni.lu/web/imp/doc.html`).

Automation and modularity

Automation of the workflow is achieved using Snakemake 3.4.2 [Köster and Rahmann, 2012; Köster, 2014], a Python-based make language implemented specifically for building reproducible bioinformatic workflows and pipelines. Snakemake is inherently modular and thus allows various features to be implemented within IMP including the options of: i) executing specific/selected steps within the pipeline, ii) check-pointing, i.e. resuming analysis from a point of possible interruption/termination, iii) analysis of single-omic datasets (MG or MT). For more details regarding the functionalities of IMP, please refer to the documentation of IMP (`http://r3lab.uni.lu/web/imp/doc.html`).

Input data

The input to IMP includes MG and/or MT FASTQ paired files, i.e. pairs-1 and pairs-2 are in individual files. The required arguments for the IMP wrapper script are metagenomic paired-end reads ("m" options) and/or metatranscriptomic paired-end reads ("-t" option) with the specified output folder ("o" option). Users may customize the command with the options and flags described in the documentation (`http://r3lab.uni.lu/web/imp/doc.html`) and in **Section 2.3.1**.

Trimming and quality filtering

Trimmomatic 0.32 [Bolger *et al.*, 2014] is used to perform trimming and quality filtering of MG and MT Illumina paired-end reads, using the following parameters: ILLUMINACLIP:TruSeq3-PE.fa:2:30:10; LEADING:20; TRAILING:20; SLIDINGWINDOW:1:3; MAXINFO:40:0.5; MINLEN:40. The parameters may be tuned via the command line or within the IMP config file. The output from this step includes retained paired-end and single-end reads (mate discarded) which are all used for downstream processes. These parameters are configurable in the IMP config file (**Section 2.3.1**).

Ribosomal RNA filtering

SortMeRNA 2.0 [Kopylova *et al.*, 2012] is used for filtering rRNA from the MT data. The process is applied on FASTQ files for both paired- and single-end reads generated from the trimming and quality filtering step. Paired-end FASTQ files are interleaved prior to running SortMeRNA. If one of the mates within the paired-end read is classified as an rRNA sequence, then the entire pair is filtered out. After running SortMeRNA, the interleaved paired-end output is split into two separate paired-end FASTQ files. The filtered sequences (without rRNA reads) are used for the downstream processes. All available databases provided within SortMeRNA are used for filtering and the maximum memory usage parameter is set to 4 GB (option: -m 4000) which can be adjusted in the IMP config file (**Section 2.3.1**).

Read mapping

The read mapping procedure is performed using the bwa mem aligner [Li and Durbin, 2009] with settings: -v 1 (verbose output level), -M (Picard compatibility) introducing an automated samtools header using the -R option [Li and Durbin, 2009]. Paired- and single-end reads are mapped separately, and the resulting alignments are merged (using samtools merge [Li *et al.*, 2009]). Read mapping is performed at various steps in the workflow including: i) screening for host or contaminant sequences (**Section 2.3.1**), ii) recruitment of unmapped reads within the IMP-based iterative co-assembly (**Section 2.3.1**), and iii) mapping of preprocessed MG and MT reads to the final contigs. The memory usage is configurable in the IMP config file (**Section 2.3.1**).

Extracting unmapped reads

The extraction of unmapped reads (paired- and single-end) begins by mapping reads to a given reference sequence (**Section 2.3.1**). The resulting alignment file (BAM format) is used as input for the extraction of unmapped reads. A set of paired-end reads are considered unmappable if both or either one of the mates do not map to the given reference. The unmapped reads are converted from BAM to FASTQ format using samtools [Li and Durbin, 2009] and BEDtools 2.17.0 - bamToFastq utility [Quinlan and Hall, 2010]. Similarly, unmapped single-end reads are also extracted from the alignment information.

Screening host or contaminant sequences

By default, the host/contaminant sequence screening is performed by mapping both paired- and single-end reads (**Section 2.3.1**) onto the human genome version 38 (`http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/`), followed by extraction of unmapped reads (**Section 2.3.1**). Within the IMP command line, users are provided with the option of: i) excluding this procedure with the “no-filtering” flag, ii) using other sequence(s) for screening by providing the FASTA file (or URL) using “screen” option or iii) specifying it in the configuration file (**Section 2.3.1**).

Parameters of the IMP-based iterative co-assembly

The IMP-based iterative co-assembly implements MEGAHIT 1.0.3 [Li *et al.*, 2015] as the MT assembler while IDBA-UD 1.1.1 [Peng *et al.*, 2012] is used as the default co-assembler (MG & MT), with MEGAHIT [Li *et al.*, 2015] as an alternative option for the co-assembler (specified by the “-a” option of the IMP command line). All *de novo* assemblies are performed on *k*mers ranging from 25-mers to 99-mers, with an incremental step of four. Accordingly, the command line parameters for IDBA-UD are –mink 25 –maxk 99 –step 4 –similar 0.98 –pre-correction [Peng *et al.*, 2012]. Similarly, the command line parameters for MEGAHIT are –k-min 25 –k-max 99 –k-step 4, except for the MT assemblies which are performed with an additional –no-bubble option to prevent merging of bubbles within the assembly graph [Li *et al.*, 2015]. Furthermore, contigs generated from the MT assembly are used as “long read” input within the -l flag of IDBA-UD or -r flag of MEGAHIT [Peng *et al.*, 2012; Li *et al.*, 2015]. Kmer ranges for the IDBA-UD and MEGAHIT can be adjusted/specified in the configuration file (**Section 2.3.1**). Cap3 is used to reduce the redundancy and improve contiguity of the assemblies using a minimum alignment identity of 98 % (-p 0.98) with a minimum overlap of 100 bases (-o 100), which are adjustable in the configuration file (**Section 2.3.1**). Finally, the extraction of reads that are unmappable to the initial MT assembly and initial co-assembly is described in **Section 2.3.1**.

Annotation and assembly quality assessment

Prokka 1.11 [Seemann, 2014] with the –metagenome setting is used to perform functional annotation. The default blast and HMM databases of Prokka are used for the functional annotation. Custom databases may be provided by the user (refer to **Sections 2.3.1** and **2.3.1** for details).

MetaQUAST 3.1 [Mikheenko *et al.*, 2015] is used to perform taxonomic annotation of contigs with the maximum number of downloadable reference genomes set to 20 (–max-ref-number 20). In addition, MetaQUAST provides various assembly statistics. The maximum number of downloadable reference genome can be changed in the IMP config file Customization and further development for details.

Depth of coverage

Contig- and gene-wise depth of coverage values are calculated (per base) using BEDtools 2.17.0 [Quinlan and Hall, 2010] and aggregated (by average) using awk, adapted from the CONCOCT code [Alneberg *et al.*, 2014] (script: map-bowtie2-markduplicates.sh, `https://github.com/BinPro/CONCOCT`) and is non-configurable.

Variant calling

The variant calling procedure is performed using Samtools 0.1.19 [Li and Durbin, 2009] (mpileup tool) and Platypus 0.8.1 [Rimmer *et al.*, 2014], each using their respective default settings and are non-configurable. The input is the merged paired- and single-end read alignment (BAM) against the final assembly FASTA file (**Section 2.3.1**). The output files from both the methods are indexed using tabix and compressed using gzip. No filtering is applied to the variant calls, so that users may access all the information and filter them according to their requirements. The output from samtools mpileup is used for the augmented VizBin visualization.

Non-linear dimensionality reduction of genomic signatures (NLDR-GS)

VizBin [Laczny *et al.*, 2015] performs NLDR-GS onto contigs $\geq$ 1kb, using default settings, to obtain 2D embeddings. Parameters can be modified in the IMP config file (**Section 2.3.1**).

Automated binning

Automated binning of the assembled contigs is performed using MaxBin 2.0. Default setting are applied and paired-end reads are provided as input for abundance estimation [Wu *et al.*, 2014]. The sequence length cut-off is set to be same as VizBin (**Section 2.3.1**) and is customizable using the config file (**Section 2.3.1**).

Visualization and reporting

IMP compiles the multiple summaries and visualizations into a HTML report. FASTQC [Patel and Jain, 2012] is used to visualize the quality and quantity of reads before and after preprocessing. MetaQUAST [Mikheenko *et al.*, 2015] is used to report assembly quality and taxonomic associations of contigs. A custom script is used to generate KEGG-based [Kanehisa and Goto, 2000] functional Krona plots by running KronaTools [Ondov *et al.*, 2011] (script: genes.to.kronaTable.py, GitHub URL: `https://github.com/EnvGen/metagenomics-workshop`). Additionally, VizBin output (2D embeddings) is integrated with the information derived from the IMP analyses, using a custom R script for analysis and visualization of the augmented maps. The R workspace image is saved such that users are able to access it for further analyses. All the steps executed within an IMP run including parameters and runtimes are summarized in the form of a workflow diagram and a log-file. The visualization script is not configurable.

Output

The output generated by IMP includes a multitude of large files. Paired- and single-end FASTQ files of preprocessed MG and MT reads are provided such that the user may employ them for additional downstream analyses. The output of the IMP-based iterative co-assembly consists of a FASTA file, while the alignments/mapping of MG and MT preprocessed reads to the final co-assembly are also provided as a binary alignment format (BAM), such that users may use these for further processing. Predicted genes and their respective annotations are provided in the various formats produced by Prokka [Seemann, 2014]. Assembly quality statistics and taxonomic annotations of contigs are provided as per the output of MetaQUAST [Mikheenko *et al.*, 2015]. Two-dimensional embeddings from the NLDR-GS are provided such that they can be exported to and further curated using VizBin [Laczny *et al.*, 2015]. Additionally, abundance and expression information is represented by contig- and gene-level average depth of coverage values. MG and MT genomic variant information (VCF format), including both SNPs and INDELs (insertions and deletions), is also provided. The results of the automated binning using MaxBin 2.0 [Wu *et al.*, 2014] are provided in a folder which contains the default output from the program (i.e. fasta files of bins and summary files).

The HTML reports, e.g. **Additional file 2.1**: HTML S1 & S2 compiles various summaries and visualizations including: i) augmented VizBin maps, ii) MG- and MT-level functional Krona charts [Ondov *et al.*, 2011], iii) detailed schematics of the steps carried out within the IMP run, iv) list of parameters and commands, and v) additional reports (FASTQC report [Patel and Jain, 2012], MetaQUAST report [Mikheenko *et al.*, 2015]). Please refer to the documentation of IMP for a detailed list and description of the output (`http://r3lab.uni.lu/web/imp/doc.html`).

Databases

The IMP database folder (db) contains required databases required for IMP analysis. The folder contains the following subfolders and files with their specific content:

i) adapters folder –sequencing adapter sequences. Default version contains all sequences provided by Trimmomatic version 0.32 [Bolger *et al.*, 2014]

ii) cm, genus, hmm and kingdom folders – contains databases provided by Prokka 1.11 [Seemann, 2014]. Additional databases may be added into the corresponding folders as per the instructions in the Prokka documentation (`https://github.com/tseemann/prokka#databases`)

iii) sortmerna folder - contains all the databases provided in SortMeRNA 2.0 [Kopylova *et al.*, 2012]. Additional databases may be added into the corresponding folders as per the instructions in the SortMeRNA documentation (`http://bioinfo.lifl.fr/RNA/sortmerna/code/SortMeRNA-user-manual-v2.0.pdf`)

iv) ec2pathways.txt - enzyme commission (EC) number mapping of amino acid sequences to pathways

v) pathways2hierarchy.txt - pathway hierarchies used to generated for KEGG-based functional Krona plot (**Section 2.3.1**)

Customization and further development

Additional advanced parameters can be specified via the IMP command line including specifying a custom configuration file (-c option) and/or specifying a custom database folders (-d option). Threads (–threads) and memory allocation (–memcore and –memtotal) can be adjusted via the command line and the configuration file. The IMP launcher script provides a flag (–enter) to launch the Docker container interactively and the option to specify the path to the customized source code folder (-s option). These commands are provided for development and testing purposes (described on the IMP website and documentation: `http://r3lab.uni.lu/web/imp/doc.html`). Further customization is possible using a custom configuration file (JSON format). The customizable options within the JSON file are specified in individual subsections within section **Section 2.3.1**. Finally, the open source implementation of IMP allows users to customize the Docker image and source code of IMP according to their requirements.

2.3.2 Iterative single-omic assemblies

In order to determine the opportune number of iterations within the IMP-based iterative co-assembly strategy an initial assembly was performed using IMP preprocessed MG reads with IDBA-UD [Peng *et al.*, 2012]. Cap3 [Huang and Madan, 1999] was used to further collapse the contigs and reduce the redundancy of the assembly. This initial assembly was followed by a total of three assembly iterations, whereby each iteration was made up of four separate steps: i) extraction of reads unmappable to the previous assembly (using the procedure described in **Section 2.3.1**), ii) assembly of unmapped reads using IDBA-UD [Peng *et al.*, 2012], iii) merging/collapsing the contigs from the previous assembly using cap3 [Huang and Madan, 1999], and iv) evaluation of the merged assembly using MetaQUAST [Mikheenko *et al.*, 2015]. The assembly was evaluated in terms of the per-iteration increase in mappable reads, assembly length, numbers of contigs $\geq$ 1 kb, and numbers of unique genes.

Similar iterative assemblies were also performed for MT data using MEGAHIT [Li *et al.*, 2015] except, CD-HIT-EST [Fu *et al.*, 2012] was used to collapse the contigs at $\geq$ 95 % identity (-c 0.95) while MetaGeneMark [Zhu *et al.*, 2010] was used to predict genes. The parameters and settings of the other programs were the same as those defined in **Section 2.3.1**.

The aforementioned procedures were applied to all the datasets analyzed within this article. The merged contig sets (non-redundant) from the first iteration of both the MG and MT iterative assemblies were selected to represent the IMP single-omics assemblies (IMP_MG and IMP_MT) and were compared against co-assemblies.

2.3.3 Execution of pipelines

MetAMOS ver. 1.5rc3 was executed using default settings [Treangen *et al.*, 2013]. MG data was provided as input for single-omic assemblies (MetAMOS_MG) while MG and MT data was provided as input for multi-omic co-assemblies (MetAMOS_MGMT). All computations using MetAMOS were set to use eight computing cores (-p 8).

MOCAT ver. 1.3 (MOCAT.pl) [Kultima *et al.*, 2012] was executed using default settings. Paired-end MG data was provided as input for single-omic assemblies (MOCAT_MG) while paired-end MG and MT data was provided as input for multi-omic co-assemblies (MOCAT_MGMT). All computations using MOCAT were set to use eight computing cores (-cpus 8). Paired-end reads were first preprocessed using the read_trim_filter step of MOCAT (-rtf). For the human fecal microbiome datasets (HF1-5), the preprocessed paired- and single-end reads were additionally screened for human genome derived sequences (-s hg19). The resulting reads were afterwards the assembled with default parameters (-gp assembly –r hg19) using SOAPdenovo.

IMP ver. 1.4 was executed for each dataset using different assemblers for the co-assembly step: i) default setting using IDBA-UD, and ii) MEGAHIT (-a megahit). Additionally, the analysis of human fecal microbiome datasets (HF1-5) included the preprocessing step of filtering human genome sequences, which was omitted for the wastewater sludge datasets (WW1-4) and the biogas (BG) reactor dataset. Illumina TruSeq2 adapter trimming was used for wastewater datasets preprocessing, since the information was available. Computation was performed using eight computing cores (–threads 8), 32 GB memory per core (–memcore 32) and total memory of 256 GB (–memtotal 256 GB). The customized parameters were specified in the IMP configuration file (exact configurations available in **Additional file 2.1**: HTML S1 & S2). The analysis of the CAMI datasets were carried using the MEGAHIT assembler option (-a megahit), while the other options remained as default settings.

In addition, IMP was also used on a small scale dataset to evaluate performance of increasing the number of threads from 1 to 32 and recording the runtime (time command). IMP was launched on the AWS cloud computing platform running the MEGAHIT as the assembler (-a megahit) with 16 threads (–threads 16) and 122 GB of memory (–memtotal 122).

2.3.4 Data usage assessment

Preprocessed paired-end and single-end MG and MT reads from IMP were mapped (**Section 2.3.1**) onto the IMP-based iterative co-assemblies and IMP_MG assembly. Similarly, preprocessed paired-end and single-end MG and MT reads from MOCAT were mapped onto the MOCAT co-assembly (MOCAT_MGMT) and the MOCAT single-omic MG assembly (MOCAT_MG). MetAMOS does not retain single-end reads, therefore, preprocessed MG and MT paired-end reads from MetAMOS were mapped onto the MetAMOS co-assembly (MetAMOS_MGMT) and MetAMOS single-omic MG assembly (MetAMOS_MG).

Preprocessed MG and MT reads from the human fecal datasets (HF1-5) were mapped using the same parameters described in **Section 2.3.1** to the IGC reference database [Li *et al.*, 2014] for evaluation of a reference-based approach. Alignment files of MG and MT reads mapping to the IMP-based iterative co-assemblies and the aforementioned alignments to the IGC reference database were used to report the fractions of properly paired reads mapping in either IMP-based iterative co-assembly, IGC reference database, or both. These fractions were then averaged across all the human fecal datasets (HF1-5).

2.3.5 Assembly assessment and comparison

Assemblies were assessed and compared using MetaQUAST by providing contigs (FASTA format) from of all different (single- and multi-omic) assemblies of the same dataset as input [Mikheenko *et al.*, 2015]. The gene calling function (-f) was utilized to obtain the number of genes which were predicted from the various assemblies. An additional parameter within MetaQUAST was used for ground truth assessment of the simulated mock (SM) community assemblies by providing the list of 73 FASTA format reference genomes (-R). The CPM measure was computed based on the information derived from the results of MetaQUAST [Mikheenko *et al.*, 2015]. In order to be consistent with the reported values (i.e. N50 length), the CPM measures reported within this article are based on alignments of 500 bp and above, unlike the 1kb cut-off used in the original work [Deng *et al.*, 2015]. Prodigal was also used for gene prediction to obtain the number of complete and incomplete genes [Hyatt *et al.*, 2010].

2.3.6 Analysis of contigs assembled from MT data

A list of contigs with no MG depth of coverage together with additional information on these contigs (contig length, annotation, MT depth of coverage) was retrieved using the R workspace image which is provided as part IMP output (**Sections 2.3.1** and **2.3.1**). The sequences of these contigs were extracted and subjected to a blastn search on NCBI to determine their potential origin. Furthermore, contigs with length $\geq$ 1kb, average depth of coverage $\geq$ 20 bases and containing genes encoding known virus/bacteriophage functions were extracted.

2.3.7 Analysis of subsets of contigs

Subsets of contigs were identified by visual inspection of augmented VizBin maps generated by IMP. Detailed inspection of contig-level MT to MG depth of coverage ratios was carried out using the R workspace provided as part of IMP output (**Sections 2.3.1** and **2.3.1**). The alignment information of contigs to isolate genomes provided by MetaQUAST [Mikheenko *et al.*, 2015] were used to highlight subsets of contigs aligning to genomes of *Escherichia coli* P12B strain (*E. coli*) and *Collinsella intestinalis* DSM 13280 (*C. intestinalis*).

An additional reference-based analysis of MetaQUAST [Mikheenko *et al.*, 2015] was carried out for all the human fecal microbiome assemblies (HF1-5) by providing the genomes of *Escherichia coli* P12B and *Collinsella intestinalis* DSM 13280 as reference (flag: -R), to assess the recovery fraction of the aforementioned genomes within the different assemblies.

2.3.8 Computational platforms

IMP and MetAMOS were executed on a Dell R820 machine with 32 Intel(R) Xeon(R) CPU E5-4640 @ 2.40GHz physical computing cores (64 virtual), 1024 TB of DDR3 RAM (32 GB per core) with Debian 7 Wheezy as the operating system. MOCAT, IMP single-omic assemblies and additional analyses were performed on the Gaia cluster of the University of Luxembourg HPC platform (Varrette et al., 2014). IMP was executed on the Amazon Web Services (AWS) cloud computing platform using EC2 R3 type (memory optimized) model r3.4xlarge instance with 16 compute cores, 122 GB memory and 320 GB of storage space running a virtual Amazon Machine Image (AMI) Ubuntu ver. 16.04 operating system.

2.3.9 Availability of data and material

All the data, software and source code related to this manuscript are publicly available.

Coupled metagenomic and metatranscriptomic datasets

The published human fecal microbiome datasets (MG and MT) were obtained from NCBI Bioproject PRJNA188481 (`https://www.ncbi.nlm.nih.gov/bioproject/PRJNA188481`). They include samples from individuals: X310763260, X311245214, X316192082, X316701492, and X317690558 [Franzosa *et al.*, 2014], designated within this article as HF1-5, respectively. Only samples labeled as “Whole” (samples preserved by flash-freezing) were selected for analysis [Franzosa *et al.*, 2014].

The published wastewater sludge microbial community datasets (MG and MT) were obtained from NCBI Bioproject with the accession code PRJNA230567 (`https://www.ncbi.nlm.nih.gov/bioproject/PRJNA230567`). These include samples A02, D32, D36 and D49, designated within this article as WW1-4, respectively [Muller *et al.*, 2014a].

The published biogas reactor microbial community data set (MG and MT) was obtained from the European Nucleotide Archive (ENA) project PRJEB8813 (`http://www.ebi.ac.uk/ena/data/view/PRJEB8813`) and was designated within this article as BG [Bremges *et al.*, 2015].

Simulated coupled metagenomic and metatranscriptomic dataset

The simulated MT data was obtained upon request from the original authors [Celaj *et al.*, 2014]. A complementary metagenome was simulated using the same set of 73 bacterial genomes used for the aforementioned simulated MT [Celaj *et al.*, 2014]. Simulated reads were obtained using the NeSSM MG simulator (default settings) [Jia *et al.*, 2013]. The simulated mock community is designated as SM within this article [Jia *et al.*, 2013]. The simulated data along with the corresponding reference genomes used to generate the MG data is made available via LCSB WebDav (`https://webdav-r3lab.uni.lu/public/R3lab/IMP/datasets/`) and is archived on Zenodo (`http://doi.org/10.5281/zenodo.160261`).

CAMI simulated community metagenomic datasets

The medium complexity CAMI simulated MG data and the corresponding gold standard assembly were obtained from the CAMI website: `http://www.cami-challenge.org`.

Test dataset for runtime assessment

A subset of ~5 % of the WW1 MG and MT dataset (**Section 2.3.9**) was selected and used as the data to perform runtime assessments. This dataset could be used to test IMP on regular platforms such as laptops and desktops. It is made available via the LCSB R3 WebDav (`https://webdav-r3lab.uni.lu/public/R3lab/IMP/datasets/`) and is archived on Zenodo (`http://doi.org/10.5281/zenodo.160708`).

Software and source code

IMP is available under the MIT license, on the LCSB R3 website: `http://r3lab.uni.lu/web/imp/` which contains necessary information related to IMP. These includes links to the Docker images on the LCSB R3 WebDav (`https://webdav-r3lab.uni.lu/public/R3lab/IMP/dist/`) and is archived on Zenodo (`http://doi.org/10.5281/zenodo.160263`). Source code is available on LCSB R3 GitLab (`https://git-r3lab.uni.lu/IMP/IMP`), GitHub (`https://github.com/shaman-narayanasamy/IMP`) and is archived on Zenodo (`http://doi.org/10.5281/zenodo.160703`). Scripts and commands for additional analyses performed specifically within this manuscript are available on LCSB R3 GitLab (`https://git-r3lab.uni.lu/IMP/IMP_manuscript_analysis`) and on GitHub (`https://github.com/shaman-narayanasamy/IMP_manuscript_analysis`). Frozen pages containing all necessary material related to this article are available at: `http://r3lab.uni.lu/frozen/imp/`.

2.4 Results

2.4.1 Overview of the IMP implementation and workflow

IMP leverages Docker for reproducibility and deployment. The interfacing with Docker is facilitated through a user-friendly Python wrapper script (**Section 2.3.1**). As such, Python and Docker are the only prerequisites for the pipeline, enabling an easy installation and execution process. Workflow implementation and automation is achieved using Snakemake [Köster and Rahmann, 2012; Köster, 2014]. The IMP workflow can be broadly divided into five major parts: i) preprocessing, ii) assembly iii) automated binning, iv) analysis and v) reporting (**Figure 2.1**).

The preprocessing and filtering of sequencing reads is essential for the removal of low quality bases/reads and potentially unwanted sequences, prior to assembly and analysis. The input to IMP consists of MG and MT data (the latter preferably depleted of ribosomal RNA prior to sequencing) paired-end reads in FASTQ format (**Section 2.3.1**). MG and MT reads are preprocessed independently of each other. This involves an initial quality control step (**Figure 2.1** and **Section 2.3.1**) [Bolger *et al.*, 2014] followed by an optional screening for host/contaminant sequences, whereby the default screening is performed against the human genome while other host genome/contaminant sequences may also be used (**Figure 2.1** and **Section 2.3.1**). *In silico* rRNA sequence depletion is exclusively applied to MT data (**Figure 2.1** and **Section 2.3.1**).

The customized assembly procedure of IMP starts with an initial assembly of preprocessed MT reads to generate an initial set of MT contigs (**Additional file 2.2**: Figure S1). MT reads unmappable to the initial set of MT contigs undergo a biological applications of the IMP workflow. second round of assembly. The process of assembling unused reads, i.e. MG or MT reads unmappable to the previously assembled contigs, is henceforth referred to as "iterative assembly". The assembly of MT reads is performed first as transcribed regions are covered much more deeply and evenly in MT data. The resulting MT-based contigs represent high-quality scaffolds for the subsequent co-assembly with MG data overall leading to enhanced assemblies [Muller *et al.*, 2014a]. Importantly, the combined set of MT contigs from the initial and iterative assemblies are used as scaffolds to enhance the subsequent assembly with the MG data. MT data is assembled using the MEGAHIT *de novo* assembler using the appropriate option to prevent the merging of bubbles within the de Bruijn assembly graph [Qin *et al.*, 2010; Li *et al.*, 2015]. Subsequently, all preprocessed MT and MG reads, together with the generated MT contigs are used as input to perform a first co-assembly, producing a first set of co-assembled contigs. The MG and MT reads unmappable to this first set of co-assembled contigs then undergo an additional iterative co-assembly step. IMP implements two assembler options for the *de novo* co-assembly, namely IDBA-UD or MEGAHIT. The contigs resulting from the co-assembly procedure undergo a subsequent assembly refinement step by a contig-level assembly using the cap3 *de novo* assembler [Huang and Madan, 1999]. This aligns highly similar contigs against each other, thus reducing overall redundancy by collapsing shorter contigs into longer contigs and/or improving contiguity by extending contigs via overlapping contig ends (**Additional file 2.2**: Figure S1). This step produces the final set of contigs. Preprocessed MG and MT reads are mapped then back against the final contig set and the resulting alignment information is used in the various downstream analysis procedures (**Figure 2.1**). In summary, IMP employs four measures for the *de novo* assembly of preprocessed MG and MT reads including: i) iterative assemblies of unmappable reads, ii) use of MT contigs to scaffold the downstream assembly of MG data, iii) co-assembly of MG and MT data, and iv) assembly refinement by contig-level assembly. The entire *de novo* assembly procedure of IMP is henceforth referred to as the "IMP-based iterative co-assembly" (**Additional file 2.2**: Figure S1).

Contigs from the IMP-based iterative co-assembly undergo quality assessment as well as taxonomic annotation [Mikheenko *et al.*, 2015] followed by gene prediction and functional annotation [Seemann, 2014] (**Figure 2.1** and **Section 2.3.1**). MaxBin 2.0 [Wu *et al.*, 2014], an automated binning procedure (**Figure 2.1** and **Section 2.3.1**) which performs automated binning on assemblies produced from single datasets, was chosen as the *de facto* binning procedure in IMP. Experimental designs involving single coupled MG and MT datasets are currently the norm. However, IMP's flexibility does not forego the implementation of multi-sample binning algorithms such as CONCOCT [Alneberg *et al.*, 2014], MetaBAT [Kang *et al.*, 2015] and canopy clustering [Nielsen *et al.*, 2014] as experimental designs evolve in the future.

Non-linear dimensionality reduction of the contigs' genomic signatures (**Figure 2.1** and **Section 2.3.1**) is performed using the Barnes-Hut Stochastic Neighborhood Embedding (BH-SNE) algorithm allowing visualization of the data as two-dimensional scatter plots henceforth referred to as VizBin maps [Laczny *et al.*, 2014, 2015]. Further analysis steps include, but are not limited to, calculations of the contig- and gene-level depths of coverage (**Section 2.3.1**) as well as the calling of genomic variants (variant calling is performed using two distinct variant callers; **Section 2.3.1**). The information from these analyses are condensed and integrated into the generated VizBin maps to produce augmented visualizations (**Section 2.3.1**). These visualizations and various summaries of the output are compiled into a HTML report (examples of the HTML reports in **Additional file 2.1**).

Exemplary output of IMP (using the default IDBA-UD assembler) based on a human fecal microbiome dataset is summarized in **Figure 2.2**. The IMP output includes taxonomic (**Figure 2.2**A) and functional (**Figure 2.2**B & **Figure 2.2**C) overviews. The representation of gene abundances at the MG and MT levels enables comparison of potential (**Figure 2.2**B) and actual expression (**Figure 2.2**C) for specific functional gene categories (see Krona charts within **Additional file 2.1**: HTML S1). IMP provides augmented VizBin maps [Laczny *et al.*, 2014, 2015] including for example, variant densities (**Figure 2.2**D) as well as MT to MG depth of coverage ratios (**Figure 2.2**E). These visualizations may aid users in highlighting subsets of contigs based on certain characteristics of interest, i.e. population heterogeneity/homogeneity, low/high transcriptional activity, etc. Although an automated binning method [Wu *et al.*, 2014] is incorporated within IMP (**Figure 2.2**F), the output is also compatible with and may be exported to other manual/interactive binning tools such as VizBin [Laczny *et al.*, 2015] and Anvi'o [Eren *et al.*, 2015] for additional manual curation. Please refer to **Additional file 2.1** for additional examples.

The modular design (**Section 2.3.1**) and open source nature of IMP allow for customization of the pipeline to suit specific user-defined analysis requirements (**Section 2.3.1**). As an additional feature, IMP also allows single-omic MG or MT analyses (**Section 2.3.1**). Detailed parameters for the processes implemented in IMP are described in the **Section 2.3.1** and examples of detailed workflow schematics are provided in **Additional file 2.1**: HTML S1 & S2.

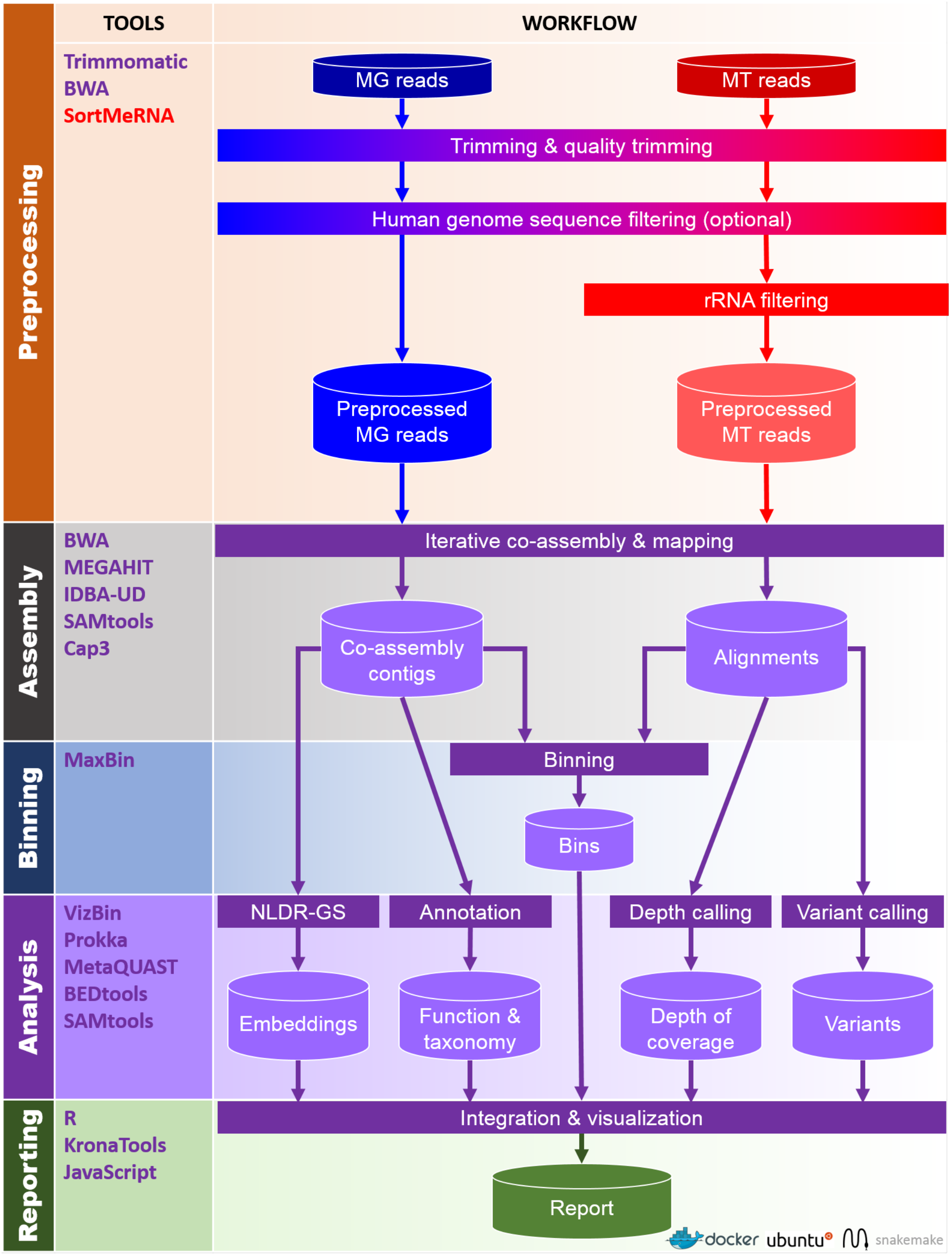


Figure 2.1: Schematic overview of the IMP pipeline Cylinders represent input and output while rectangles represent processes. Arrows indicate the flow between input, processes and output. MG: Metagenomic data, MT: Metatranscriptomic data, rRNA: ribosomal RNA, NLDR-GS: genomic signature non-linear dimensionality reduction. Processes, input and output specific to MG and MT data are labeled in blue and red, respectively. Processes and output that involve usage of both MG and MT data, are represented in purple. A detailed illustration of the "iterative co-assembly" is available in Figure S1 in **Additional file 2.2**.

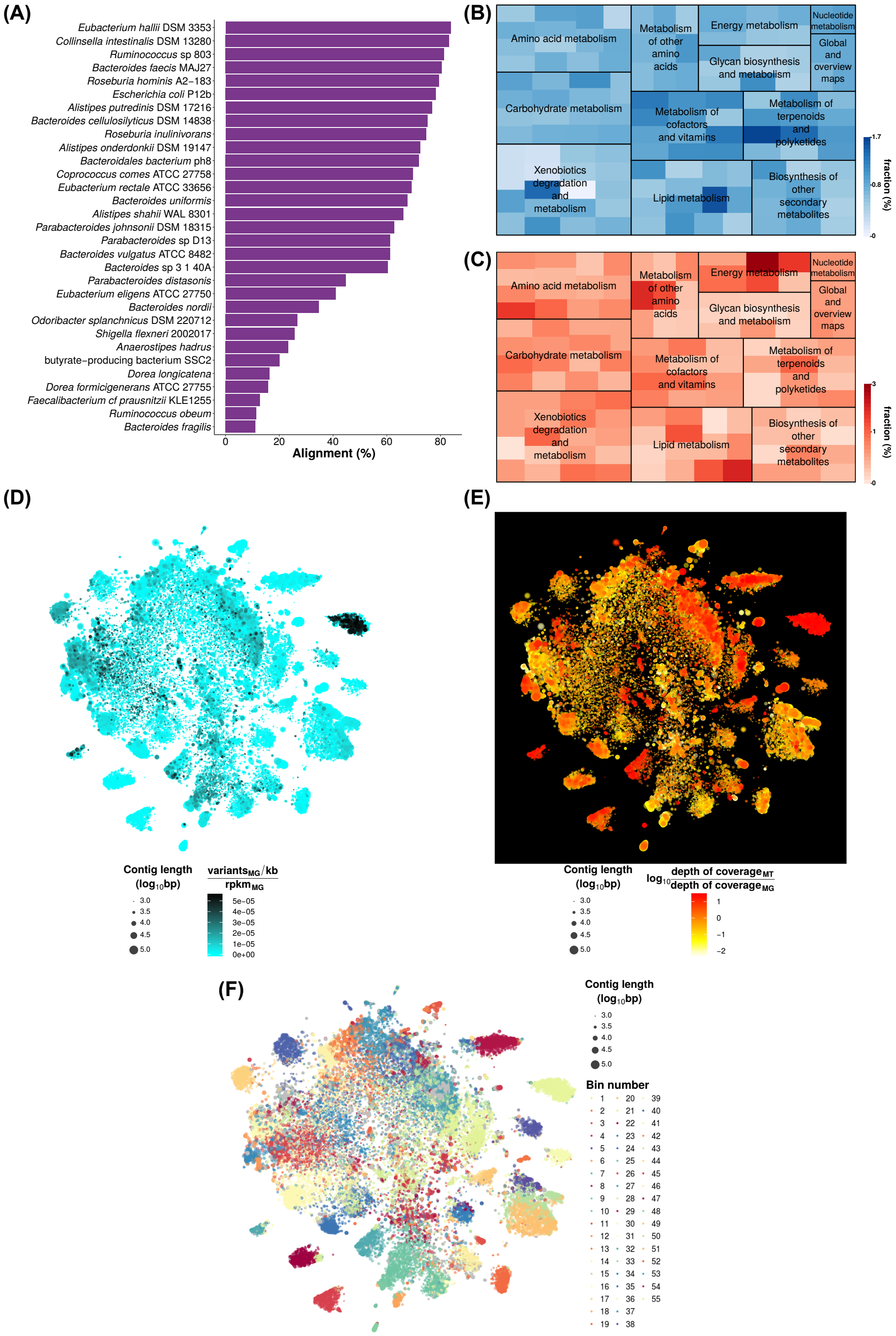


Figure 2.2: Example output from the IMP analysis of a human microbiome dataset (HF1). **(A)** Taxonomic overview based on the alignment of contigs to the most closely related genomes present in the NCBI genome database (see also **Additional file 2.1**: HTML S1), abundances of predicted genes (based on average depths of coverage) of various KEGG Ontology categories represented both at the **(B)** MG and **(C)** MT levels (see also Krona charts within **Additional file 2.1**: HTML S1). Augmented VizBin maps of contigs ≥ 1kb, representing **(D)** contig-level MG variant densities, (E) contig-level ratios of MT to MG average depth of coverage and **(F)** bins generated by the automated binning procedure. Additional examples are available in **Additional file 2.1**.

2.4.2 Assessment and benchmarking

IMP was applied to ten published coupled MG and MT datasets, derived from three types of microbial systems, including five human fecal microbiome samples (HF1, HF2, HF3, HF4, HF5) [Franzosa *et al.*, 2014], four wastewater sludge microbial communities (WW1, WW2, WW3, WW4) [Muller *et al.*, 2014b; Roume *et al.*, 2015] and one microbial community from a production-scale biogas (BG) plant [Bremges *et al.*, 2015]. In addition, a simulated mock (SM) community dataset based on 73 bacterial genomes [Celaj *et al.*, 2014], comprising both MG and MT was generated to serve as a means for ground truth-based assessment of IMP (details in **Section 2.3.9**). The SM dataset was devised given the absence of a standardized benchmarking dataset for coupled MG and MT data (this does solely exist for MG data as part of the CAMI initiative: `http://www.cami-challenge.org`).

Analysis with IMP was carried out with the two available *de novo* assembler options for the co-assembly step (**Figure 2.1** and **Additional file 2.2**: Figure S1), namely the default IDBA-UD assembler [Peng *et al.*, 2012] (hereafter referred to as IMP) and the optional MEGAHIT assembler [Li *et al.*, 2015] (henceforth referred to as IMP-megahit). IMP was quantitatively assessed based on resource requirement and analytical capabilities. The analytical capabilities of IMP were evaluated based on data usage, output volume and output quality. Accordingly, we assessed the advantages of the iterative assembly procedure as well as the overall data integration strategy.

Resource requirement and runtimes

IMP is an extensive pipeline that utilizes both MG and MT data within a reference-independent (assembly-based) analysis framework which renders it resource- and time-intensive. Therefore, we aimed to assess the required computational resource and runtimes of IMP.

All IMP-based runs on all datasets were performed on eight compute cores with 32 GB RAM per core and 1024 GB of total memory (**Section 2.3.8**). IMP runtimes ranged from approximately 23 hours (HF1) to 234 hours (BG) and the IMP-megahit runtimes ranged from approximately 21 hours (HF1) up to 281 hours (BG). IMP was also executed on the Amazon cloud computing (AWS) infrastructure, using the HF1 dataset on a machine with 16 cores (**Section 2.3.8**) whereby the run lasted approximately 13 hours (refer to **Additional file 2.2**: Note S1 for more details). The analysis of IMP resulted in an increase in additional data of around 1.2-3.6 times the original input (**Additional file 2.2**: Table S1). Therefore, users should account for the disc space for both the final output and intermediate (temporary) files generated during an IMP run. Detailed runtimes and data generated for all the processed data sets are reported in **Additional file 2.3**: Table S1.

We further evaluated the effect of increasing resources using a small scale test dataset (**Section 2.3.9**). The tests demonstrated that reduced runtimes are possible by allocating more threads to IMP-megahit (**Additional file 2.3**: Table S2). However, no apparent speed-up is achieved beyond allocation of eight threads, suggesting that this would be the optimal number of threads for this particular test dataset. Contrastingly, no speed up was observed with additional memory allocation (**Additional file 2.3**: Table S3). Apart from the resources, runtime may also be affected by the input size, the underlying complexity of the dataset and/or behavior of individual tools within IMP.

Data usage - iterative assembly

De novo assemblies of MG data alone usually result in a large fraction of reads that are unmappable to the assembled contigs and therefore remain unused, thereby leading to suboptimal data usage [Muller *et al.*, 2014b; Schürch *et al.*, 2014; Reyes *et al.*, 2015; Hitch and Creevey, 2016]. Previous studies have assembled sets of unmappable reads iteratively to successfully obtain additional contigs, leading to an overall increase in the number of predicted genes which in turn results in improved data usage [Muller *et al.*, 2014b; Schürch *et al.*, 2014; Reyes *et al.*, 2015; Hitch and Creevey, 2016]. Therefore, IMP uses an iterative assembly strategy to maximize NGS read usage. In order to evaluate the best iterative assembly approach for application within the IMP-based iterative co-assembly strategy, we attempted to determine the opportune number of assembly iterations in relation to assembly quality metrics and computational resources/runtimes.

The evaluation of the iterative assembly strategy was applied to MG and MT datasets. For both omic data types, it involved an "initial assembly" which is defined as the *de novo* assembly of all preprocessed reads. Additional iterations of assembly were then conducted using the reads that remained unmappable to the generated set of contigs (**Section 2.3.2** for details and parameters). The evaluation of the iterative assembly procedure was carried out based on the gain of additional contigs, cumulative contig length (bp), numbers of genes and numbers of reads mappable to contigs. Table 1 shows the results of the evaluation for four representative data sets and **Additional file 2.3**: Table S4 shows the detailed results of the application of the approach to eleven datasets. In all the datasets evaluated, all iterations (1 to 3) after the initial assembly lead to an increase in total length of the assembly and numbers of mappable reads (Table 1, **Additional file 2.3**: Table S4). However, there was a notable decline in the number of additional contigs and predicted genes beyond the first iteration. Specifically, the first iteration of the MG assembly yielded up to 1.6 % additional predicted genes while the equivalent on the MT data yielded up to 9 % additional predicted genes (**Additional file 2.3**: Table S4). Considering the small increase (< 1 %) in the number of additional contigs and predicted genes beyond the first assembly iteration on one hand and the extended runtimes required to perform additional assembly iterations on the other hand, a generalized single iteration assembly approach was retained and implemented within the IMP-based iterative co-assembly (**Figure 2.2** and **Additional file 2.2**: Figure S1). This approach aims to maximize data usage without drastically extending runtimes.

Despite being developed specifically for the analysis of coupled MG and MT datasets, the iterative assembly can also be used for single omic datasets. To assess IMP's performance on MG datasets, it was applied to the simulated MG datasets from the CAMI challenge (`http://www.cami-challenge.org`) and the results are shown in **Additional file 2.2**: Figure S2. IMP-based MG assembly using the MEGAHIT assembler on the CAMI dataset outperforms well-established MG pipelines such as MOCAT in all measures. In addition, IMP-based iterative assemblies also exhibit comparable performance to the gold standard assembly with regards to contigs $\geq$ 1kb and number of predicted genes (`http://www.cami-challenge.org`). Detailed results of the CAMI assemblies are available in **Additional file 2.3**: Table S5. However, as no MT and/or coupled MG and MT datasets so far exist for the CAMI challenge, the full capabilities of IMP could not be assessed in relation to this initiative.

Table 2.1: Statistics of iterative assemblies performed on MG and MT datasets.

Dataset	Iteration	MG iterative assembly				MT iterative assembly			
		Number of contigs (≥ 1kb)	**Cumulative length of assembled contigs (bp)**	**Number of predicted genes**	**Number of mapped reads**	**Number of contigs (all)**	**Cumulative length of assembled contigs (bp)**	**Number of predicted genes**	**Number of mapped reads**
SM	Initial assembly	29063	182673343	186939	18977716	13436	8994518	13946	822718
	1	16	483336	329	9515	1286	502535	1272	16038
	2	6	213094	126	3425	48	18460	49	656
	3	1	86711	47	1536	0	0	0	0
HF1	Initial assembly	27028	145938650	154760	20715368	40989	45300233	66249	17525586
	1	15	966872	274	39839	2471	969614	2238	329400
	2	-1	26822	5	1276	26	10315	24	45642
	3	0	4855	0	172	3	1640	6	54788
WW1	Initial assembly	14815	77059275	81060	6513708	45118	22525759	49859	8423603
	1	28	3146390	1136	73511	2115	723904	1589	529441
	2	2	175634	114	4031	250	82048	201	13335
	3	1	30032	16	572	31	10280	18	65866
BG	Initial assembly	105282	545494441	593688	109949931	47628	27493690	60566	3754432
	1	417	10998269	3902	456821	3956	1397409	3061	130131
	2	5	335313	219	21647	717	250223	754	12766
	3	7	79022	20	2511	24	9060	22	5827

Results for all datasets available in **Additional file 2.2**: Table S2

Data usage - multi-omic iterative co-assembly

In order to assess the advantages of integrated multi-omic co-assemblies of MG and MT data, IMP-based iterative co-assemblies (IMP and IMP-megahit) were compared against MG-only based assemblies which include single-omic iterative MG assemblies generated using IMP (referred to as IMP_MG) and standard MG assemblies by MOCAT (hereafter referred to as MOCAT_MG) and MetAMOS (hereafter referred to as MetAMOS_MG). Furthermore, the available reads from the human fecal microbiome dataset (preprocessed with IMP) were mapped to the MetaHIT Integrated Gene Catalog (IGC) reference database [Li *et al.*, 2014], to compare the data usage of the different assembly procedures against a reference-dependent approach.

IMP-based iterative co-assemblies consistently recruited larger fractions of properly paired MG (**Figure 2.3**A) and/or MT (**Figure 2.3**B) reads, compared to single-omic assemblies. The resulting assemblies also produced larger numbers of contigs $\geq$ 1kb (**Figure 2.3**C), predicted non-redundant unique genes (**Figure 2.3**D) and, even more important, complete genes as predicted with start and stop codon by Prodigal [Hyatt *et al.*, 2010] (**Additional file 2.3**: Table S5). Using the reference genomes from the SM data as ground truth, IMP-based iterative co-assemblies resulted in up to 25.7 % additional recovery of the reference genomes, compared to the single-omic MG assemblies (**Additional file 2.3**: Table S5).

IMP-based iterative co-assemblies of the human fecal microbiome datasets (HF1-5) allowed recruitment of comparable fractions of properly paired MG reads and an overall larger fraction of properly paired MT reads compared to those mapping to the IGC reference database (Table 2). The total fraction (union) of MG or MT reads mapping to either IMP-based iterative co-assemblies and/or the IGC reference database was higher than 90 %, thus demonstrating that the IMP-based iterative co-assemblies allow at least 10 % of additional data to be mapped when using these assemblies in addition to the IGC reference database. In summary, the complementary use of *de novo* co-assembly of MG and MT datasets in combination with iterative assemblies enhances overall MG and MT data usage and thereby significantly increases the yield of useable information, especially when combined with comprehensive reference catalogs such as the IGC reference database.

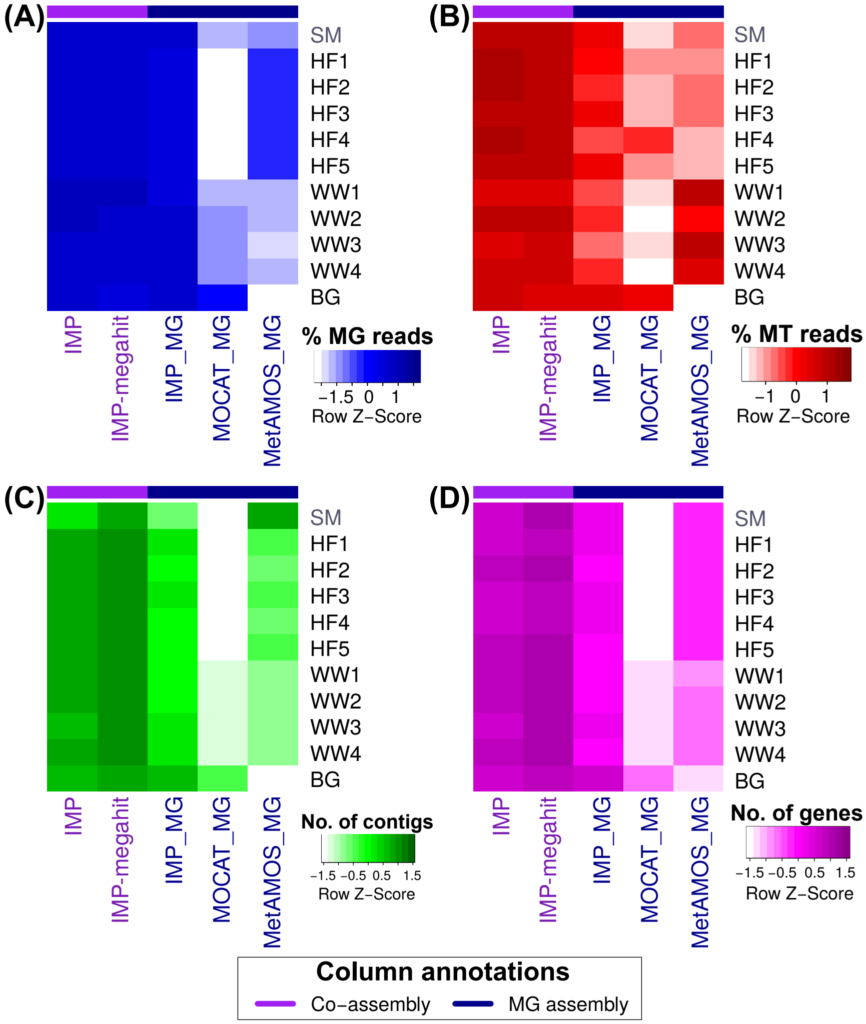


Figure 2.3: Assessment of data usage and output generated from co-assemblies compared to single-omic assemblies. Heat maps show **(A)** fractions of properly mapped MG read pairs, **(B)** fractions of properly mapped MT read pairs, **(C)** numbers of contigs $\geq$ 1kb, and **(D)** numbers of unique predicted genes. IMP and IMP-megahit represent integrated multi-omic MG and MT iterative co-assemblies while IMP_MG, MOCAT_MG and MetAMOS_MG represent single-omic MG assemblies. All numbers were row Z-score normalized for visualization. Detailed results available in **Additional file 2.3**: Table S5.

Table 2.2: Mapping statistics for human microbiome samples.

Reference	Average MG pairs mapping (%)	Average MT pairs mapping (%)
IGC	70.91	53.57
IMP	70.25	86.21
IMP-megahit	70.62	83.33
IMP_MG	68.08	58.54
MetAMOS_MG	57.31	37.34
MOCAT_MG	36.73	36.68
IMP + IGC	92.66	95.77
IMP-megahit + IGC	92.80	93.24

Average fractions (%) of properly paired reads from the human microbiome datasets (HF1-5) mapping to various references including IMP-based iterative co-assemblies (IMP and IMP-megahit) and single-omic co-assemblies (IMP_MG, MetAMOS_MG and MOCAT_MG) as well as the IGC reference database. IMP + IGC and IMP-megahit + IGC reports the total number of properly paired reads mapping to IMP-based iterative co-assemblies and/or the IGC reference database. Refer to **Additional file 2.3**: Table S3, for detailed information.

Assembly quality- multi-omic iterative co-assembly

In order to compare the quality of the IMP-based iterative co-assembly procedure to simple co-assemblies, we compared the IMP-based iterative co-assemblies against co-assemblies generated using MetAMOS [Treangen *et al.*, 2013] (henceforth referred to as MetAMOS_MGMT) and MOCAT [Kultima *et al.*, 2012] (henceforth referred to as MOCAT_MGMT). Although MetAMOS and MOCAT were developed for MG data analysis, we extended their use for obtaining MG and MT co-assemblies by including both MG and MT read libraries as input (**Section 2.3.3**). The assemblies were assessed based on contiguity (N50 length), data usage (MG and MT reads mapped) and output volume (number of contigs above 1kb and number of genes; **Additional file 2.3**: Table S5). Only the SM dataset allowed for ground truth-based assessment by means of aligning the generated *de novo* assembly contigs to the original 73 bacterial genomes used to simulate the data set (**Section 2.3.9**) [Celaj *et al.*, 2014; Mikheenko *et al.*, 2015]. This allowed the comparison of two additional quality metrics, i.e. the recovered genome fraction and the composite performance metric (CPM) proposed by Deng *et al.* [2015].

Assessments based on real datasets demonstrate comparable performance between IMP and IMP-megahit while both outperform MetAMOS_MGMT and MOCAT_MGMT in all measures (**Figure 2.4**A - C). The ground truth assessment using the SM dataset shows that IMP-based iterative co-assemblies are effective in recovering the largest fraction of the original reference genomes while achieving a higher CPM score compared to co-assemblies from the other pipelines. Misassembled (chimeric) contigs are a legitimate concern within extensive *de novo* assembly procedures such as the IMP-based iterative co-assembly. It has been previously demonstrated that highly contiguous assemblies (represented by high N50 lengths), tend to contain higher absolute numbers of misassembled contigs compared to highly fragmented assemblies, thereby misrepresenting the actual quality of assemblies [Mende *et al.*, 2012; Deng *et al.*, 2015; Lai *et al.*, 2015]. Therefore, the CPM score was devised as it represents a normalized measure reflecting both contiguity and accuracy for a given assembly [Deng *et al.*, 2015]. Based on the CPM score, both IMP and IMP-megahit yield assemblies that balance high contiguity with accuracy and thereby outperform the other methods (**Figure 2.4**C & D). In summary, cumulative measures of numbers of contigs $\geq$ 1kb, N50 lengths, numbers of unique genes, recovered genome fractions (%) and CPM scores (the latter two were only calculated for the SM dataset) as well as the mean fractions (%) of mappable MG and MT reads, show that the IMP-based iterative co-assemblies (IMP and IMP-megahit) clearly outperform all other available methods (**Figure 2.4**E; **Additional file 2.3**: Table S5).

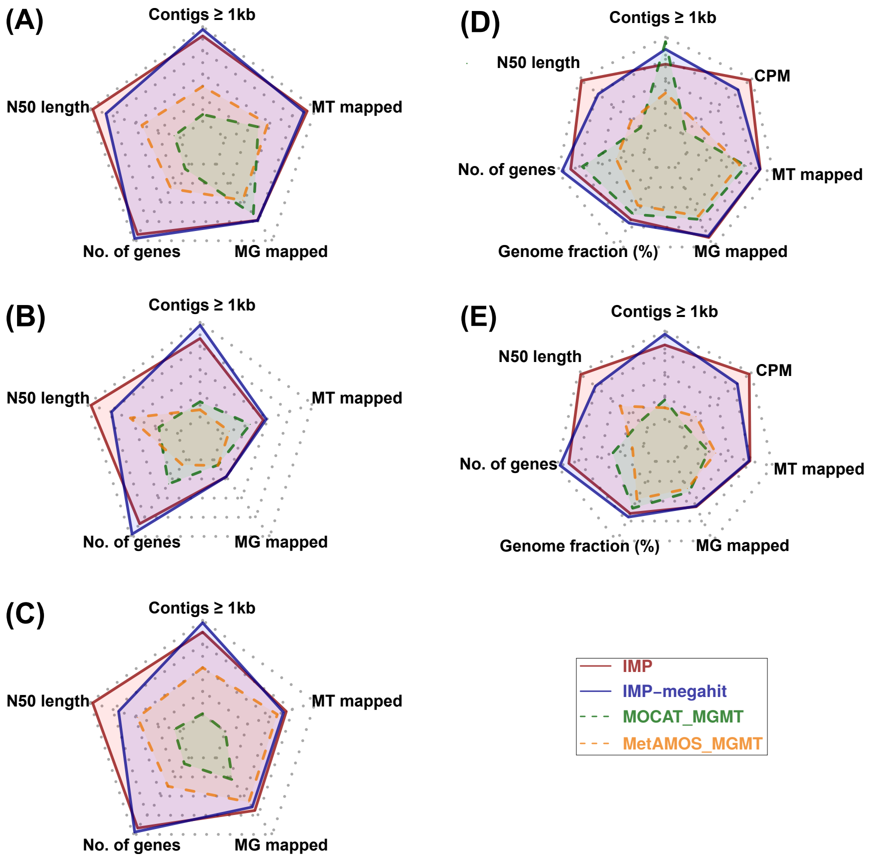


Figure 2.4: Assessment of the IMP-based iterative co-assemblies in comparison to MOCAT- and MetAMOS-based co-assemblies. Radar charts summarizing the characteristics of the co-assemblies generated using IMP, MetAMOS and MOCAT pipelines on: **(A)** human fecal microbiome, **(B)** wastewater sludge community, **(C)** biogas reactor, **(D)** simulated mock community. IMP co-assemblies were performed with two *de novo* assembler options, IDBA_UD and MEGAHIT, whereas MetAMOS and MOCAT were executed using default settings. Assessment metrics within the radar charts include, number of contigs $\geq$ 1kb, N50 length (contiguity, cut-off 500bp), number of predicted genes (unique) and fraction of properly mapped MG and MT read pairs. N50 statistics are reported using a 500bp cut-off. Additional ground truth assessments for simulated mock dataset included recovered genome fractions (%) and the composite performance metric (CPM) score with a cut-off of [Deng *et al.*, 2015]. **(E)** Summary radar chart reflecting the cumulative measures and mean fraction of properly mapped MG and MT read pairs from all analyzed 11 datasets while incorporating ground truth based measures from the simulated mock dataset. Higher values within the radar charts (furthest from center) represent better performance. Detailed information on the assembly assessments is available in **Additional file 2.3**: Table S5.

2.4.3 Use-cases of integrated metagenomic and metatranscriptomic analyses in IMP

The integration of MG and MT data provides unique opportunities for uncovering community- or population-specific traits, which cannot be resolved from MG or MT data alone. Here we provide two examples of insights gained through the direct inspection of results provided by IMP.

Tailored preprocessing and filtering of MG and MT data

The preprocessing of the datasets HF1-5 included filtering of human-derived sequences, while the same step was not necessary for the non-human derived datasets, WW1-4 and BG. MT data analyzed within this article included RNA extracts which were not subjected to wet-lab rRNA depletion, i.e. BG (Bremges et al., 2015), and samples which were treated with wet-lab rRNA removal kits (namely HF1-5 [Franzosa *et al.*, 2014] and WW1-4 [Muller *et al.*, 2014b]). Overall, the removal of rRNA pairs from the MT data showed a large variation, ranging from as low as 0.51 % (HF5) to 60.91 % (BG), demonstrating that wet-lab methods vary in terms of effectiveness and highlighting the need for such MT-specific filtering procedures (**Additional file 2.2**: Note S2 and **Additional file 2.3**: Table S6).

Identification of RNA viruses

To identify differences in the information content of MG and MT complements, the contigs generated using IMP were inspected with respect to coverage by MG and MT reads (**Additional file 2.3**: Table S7). In two exemplary datasets HF1 and WW1, a small fraction of the contigs resulted exclusively from MT data (**Additional file 2.3**: Table S7). Longer contigs ($\geq$ 1 kb) composed exclusively of MT reads and annotated with known viral/bacteriophage genes were retained for further inspection (Table 3; complete list contigs in **Additional file 2.3**: Table S8 & S9). A subsequent sequence similarity search against the NCBI NR nucleotide database [Pruitt *et al.*, 2002] of these candidate contigs revealed that the longer contigs represent almost complete genomes of RNA viruses (**Additional file 2.3**: Table S10 & S11). This demonstrates that the incorporation of MT data and its contrasting to the MG data allows the identification and recovery of nearly complete RNA viral genomes, thereby allowing their detailed future study in a range of microbial ecosystems.

Table 2.3: Contigs with a likely viral/bacteriophage origin/function reconstructed from the metatranscriptomic data.

Sample	Contig ID*	Contig length	Average contig depth of coverage	Gene product	Average gene depth of coverage
HF1	contig_34	6468	20927	Virus coat protein (TMV like)	30668
				Viral movement protein (MP)	26043
				RNA dependent RNA polymerase	22578
				Viral methyltransferase	18817
	contig_13948	2074	46	RNA dependent RNA polymerase	41
				Viral movement protein (MP)	56
WW2	contig_6405	4062	46	Tombusvirus p33	43
				Viral RNA dependent RNA polymerase	42
				Viral coat protein (S domain)	36
	contig_7409	3217	21	Viral RNA dependent RNA polymerase	18
				Viral coat protein (S domain)	21
	contig_7872	2955	77	hypothetical protein	112
				Phage maturation protein	103

*Contigs of $\geq$ 1kb and average depth of coverage $\geq$ 20 were selected.

Identification of populations with apparent high transcriptional activity

To further demonstrate the unique analytical capabilities of IMP, we aimed to identify microbial populations with a high transcriptional activity in the HF1 human fecal microbiome sample. Average depth of coverage at the contig- and gene-level is a common measure used to evaluate the abundance of microbial populations within communities [Albertsen *et al.*, 2013a; Alneberg *et al.*, 2014; Muller *et al.*, 2014b]. The IMP-based integrative analysis of MG and MT data further extends this measure by calculation of average MT to MG depth of coverage ratios, which provide information on transcriptional activity and which can be visualized using augmented VizBin maps [Laczny *et al.*, 2015].

In our example, one particular cluster of contigs within the augmented VizBin maps exhibited high MT to

MG depth of coverage ratios (**Additional file 2.2**: Figure S3). The subset of contigs within this cluster aligned to the genome of the *Escherichia coli* P12B strain (henceforth referred to as *E. coli*). For comparison, we also identified a subset, which was highly abundant at the MG level (lower MT to MG ratio), which aligned to the genome of *Collinsella intestinalis* DSM 13280 strain (henceforth referred to as *C. intestinalis*). Based on these observations, we highlighted the subsets of these contigs in an augmented VizBin map (**Figure 2.5**A). The *C. intestinalis* and *E. coli* subsets are mainly represented by clear peripheral clusters which exhibit consistent intra-cluster MT to MG depth of coverage ratios (**Figure 2.5**A). The subsets were manually inspected in terms of their distribution of average MG and MT depths of coverage and were compared against the corresponding distributions for all contigs. The MG-based average depths of coverage of the contigs from the entire community exhibited a bell-shape like distribution, with a clear peak (**Figure 2.5**B). In contrast, MT depths of coverage exhibited more spread, with a relatively low mean (compared to MG distribution) and no clear peak (**Figure 2.5**B). The *C. intestinalis* subset displays similar distributions to that of the entire community, whereas the *E. coli* subset clearly exhibits unusually high MT-based and low MG-based depths of coverage (**Figure 2.5**B). Further inspection of the individual omic datasets revealed that the *E. coli* subset was not covered by the MG contigs, while approximately 80 % of the *E. coli* genome was recoverable from a single-omic MT assembly (**Figure 2.5**C). In contrast, the *C. intestinalis* subset demonstrated genomic recovery in all co-assemblies (IMP, IMP-megahit, MOCAT_MGMT, MetAMOS_MGMT) and the single-omic MG assemblies (IMP_MG, MOCAT_MG, MetAMOS_MG; **Figure 2.5**C).

As noted by the authors of the original study by Franzosa *et al.* [2014], the cDNA conversion protocol used to produce the MT data is known to introduce approximately 1-2 % of *E. coli* genomic DNA into the cDNA as contamination which is then reflected in the MT data. According to our analyses, 0.12 % of MG reads and 1.95 % of MT reads derived from this sample could be mapped onto the *E. coli* contigs which is consistent with the numbers quoted by Franzosa *et al.* [2014].

Consistent recovery of the *E. coli* genome was also observed across all other assemblies of the human fecal microbiome datasets (HF2-5) which included their respective MT data (**Additional file 2.2**: Figure S4 and **Additional file 2.3**: Table S12). The integrative analyses of MG and MT data within IMP enables users to efficiently highlight notable cases such as this, and to further investigate inconsistencies and/or interesting characteristics within these multi-omic datasets.

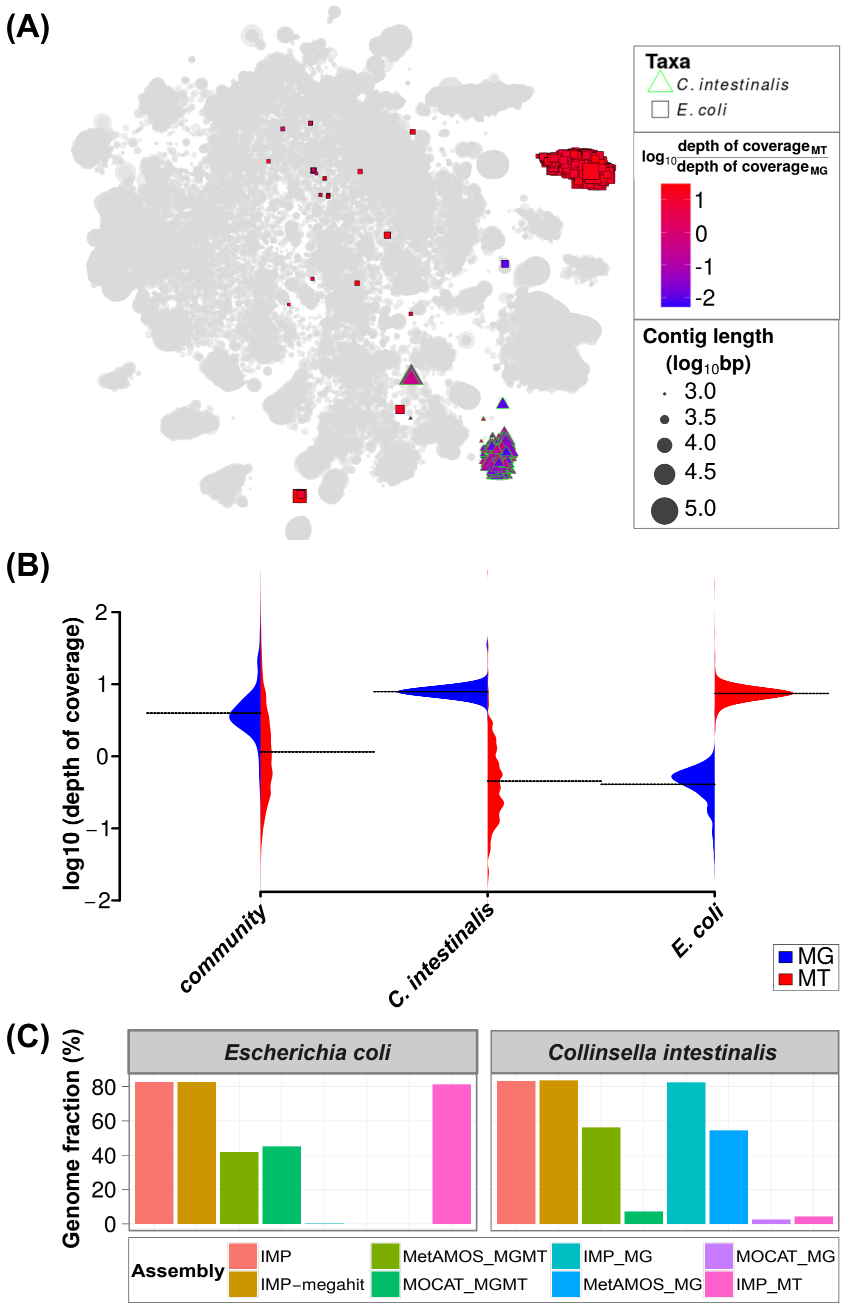


Figure 2.5: Metagenomic and metatranscriptomic data integration of a human fecal microbiome. **(A)** Augmented VizBin map highlighting contig subsets with sequences that are most similar to *Escherichia coli* P12b and *Collinsella intestinalis* DSM 13280 genomes. **(B)** Beanplots representing the densities of metagenomic (MG) and metatranscriptomic (MT) average contig-level depth of coverage for the entire microbial community and two subsets (population-level genomes) of interest. The dotted lines represent the mean. **(C)** Recovered portion of genomes of the aforementioned taxa based on different single-omic assemblies and multi-omic co-assemblies (**Additional file 2.3**: Table S5).

2.5 Discussion

The microbiome analysis workflow of IMP is unique in that it allows the integrated analysis of MG and MT data. To the best of our knowledge, IMP represents the only pipeline that spans the preprocessing of NGS reads to the binning of the assembled contigs, in addition to being the first automated pipeline for reproducible reference-independent metagenomic and metatranscriptomic data analysis. Although existing pipelines such as MetAMOS or MOCAT may be applied to perform co-assemblies of MG and MT data [Roume *et al.*, 2015], these tools do not include specific steps for the two data types in their pre- and post-assembly procedures, which is important given the disparate nature of these datasets. The use of Docker promotes reproducibility and sharing thereby allowing researchers to precisely replicate the IMP workflow with relative ease and with minimal impact on overall performance of the employed bioinformatic tools [Belmann *et al.*, 2015; Bremges *et al.*, 2015; Di Tommaso *et al.*, 2015; Leipzig, 2016]. Furthermore, static websites will be created and associated with every new version of IMP (Docker image), such that users will be able to download and launch specific versions of the pipeline to reproduce the work of others. Thereby, IMP enables standardized comparative studies between datasets from different labs, studies and environments. The open source nature of IMP encourages a community-driven effort to contribute to and further improve the pipeline. Snakemake allows the seamless integration of Python code and shell (bash) commands and the use of make scripting style, which are arguably some of the most widely used bioinformatic scripting languages which support parallel processing and the ability to interoperate with various tools and/or web services [Köster and Rahmann, 2012; Köster, 2014]. Thus, users will be able to customize and enhance the features of the IMP according to their analysis requirements with minimal training/learning.

Quality control of NGS data prior to *de novo* assemblies has been shown to increase the quality of downstream assembly and analyses (predicted genes) [Mende *et al.*, 2012]. In addition to standard preprocessing procedures (i.e. removal low quality reads, trimming of adapter sequences and removal), IMP incorporates additional tailored and customizable filtering procedures which account for the different sample and/or omic data types. For instance, the removal of host-derived sequences in the context of human microbiomes is required for protecting the privacy of study subjects. The MT-specific *in silico* rRNA removal procedure yielded varying fractions of rRNA reads between the different MT datasets despite the previous depletion of rRNA (**Section 2.4.3**) indicating that improvements in wet-lab protocols are necessary. Given that rRNA sequences are known to be highly similar, they are removed in IMP in order to mitigate any possible misassemblies resulting from such reads and/or regions [Salzberg and Yorke, 2005; Mariano *et al.*, 2016]. In summary, IMP is designed to perform stringent and standardized preprocessing of MG and MT data in a data-specific way thereby enabling efficient data usage and resulting in high-quality output.

It is common practice that MG and MT reads are mapped against a reference (e.g. genes, genomes and/or MG assemblies) [Franzosa *et al.*, 2014; Bremges *et al.*, 2015; Hultman *et al.*, 2015] prior to subsequent data interpretation. However, these standard practices lead to suboptimal usage of the original data. IMP enhances overall data usage through its specifically tailored iterative co-assembly procedure which involves four measures to achieve better data usage and yield overall larger volumes of output (i.e. a larger number of contigs $\geq$ 1kb and predicted unique and complete genes):

i) The iterative assembly procedure leads to increases in data usage and output volume in each additional iterative assembly step (**Section 2.4.2**). The exclusion of mappable reads in each iteration of the assembly, serves as a means of partitioning the data thereby reducing the complexity of the data and overall resulting in a higher cumulative volume of output [Mende *et al.*, 2012; Hitch and Creevey, 2016; Hug *et al.*, 2016].

ii) The initial assembly of MT-based contigs enhances the overall assembly, as transcribed regions are covered much more deeply and evenly in MT data, resulting in better assemblies for these regions [Muller *et al.*, 2014b]. The MT-based contigs represent high-quality scaffolds for the subsequent co-assembly with MG data.

iii) The co-assembly of MG and MT data allows the integration of these two data types while resulting in a larger number of contigs and predicted complete genes against which in turn a substantially higher fraction of reads can be mapped (**Section 2.4.2**). Furthermore, the analyses of the human fecal microbiome datasets (HF1-5) demonstrate that the numbers of MG reads mapping to the IMP-based iterative co-assemblies for each sample are comparable to the numbers of reads mapping to the comprehensive IGC reference database (Table 2). Previously, only fractions of 74 %-81 % of metagenomic reads mapping to the IGC have been reported [Li *et al.*, 2014]. However, such numbers have yet to be reported for MT data, in which case we observe lower mapping rates to the IGC reference database (35.5 %-70.5 %) compared to IMP-based assemblies (**Additional file 2.3**: Table S3). This may be attributed to the fact that the IGC reference database was generated from MG-based assemblies only, thus creating a bias [Li *et al.*, 2014]. Moreover, an excess of 90 % of MG and MT reads from the human fecal datasets (HF1-5) are mappable to either the IGC reference database and/or IMP-based iterative co-assemblies, emphasizing that a combined reference-based and IMP-based integrated-omics approach vastly improves data usage (Table 2). Although large fractions of MG and/or MT reads can be mapped to the IGC, a significant advantage of using a *de novo* reference-independent approach lies within the fact that reads can be linked to genes within their respective genomic context and microbial populations of origin. Exploiting the maximal amount of information is especially relevant for microbial communities with small sample sizes and which lack comprehensive references such as the IGC reference database.

iv) The assembly refinement step via a contig-level assembly with cap3 improves the quality of the assemblies by reducing redundancy and increasing contiguity by collapsing and merging contigs (**Section 2.4.2**). Consequently, our results support the described notion that the sequential use of multi-*k*mer-based de Bruijn graph assemblers, such as IDBA-UD and MEGAHIT, with overlap-layout-consensus assemblers, such as cap3, result in improved metagenomic assemblies [Deng *et al.*, 2015; Lai *et al.*, 2015] but importantly also extend this to MG and MT co-assemblies.

When compared to commonly used assembly strategies, the IMP-based iterative co-assemblies consisted of a larger output volume while maintaining a relatively high quality of the generated contigs. High-quality assemblies yield higher quality taxonomic information and gene annotations while longer contigs ($\geq$ 1kb) are a prerequisite for unsupervised population-level genome reconstruction [Albertsen *et al.*, 2013a; Laczny *et al.*, 2015, 2016] and subsequent multi-omics data integration [Muller *et al.*, 2014b; Roume *et al.*, 2015; Heintz-Buschart *et al.*, 2016]. Throughout all the different comparative analyses which we performed, IMP performed more consistently across all the different datasets when compared to existing methods, thereby emphasizing the overall stability and broad range of applicability of the method (**Section 2.4.2**).

Integrated analyses of MG and MT data with IMP provide the opportunity for analyses that are not possible based on MG data alone, such as the detection of RNA viruses (**Section 2.4.3**) and the identification of transcriptionally active populations (**Section 2.4.3**). The predicted/annotated genes may be used for further analyses and integration of additional omic datasets, most notably metaproteomic data [Muller *et al.*, 2014b; Roume *et al.*, 2015; Heintz-Buschart *et al.*, 2016]. Furthermore, the higher number of complete genes improves the downstream functional analysis, because the read counts per gene will be much more accurate when having full length transcript sequences and increase the probability to identify peptides. More specifically, the large number of predicted genes may enhance the usage of generated metaproteomic data, allowing more peptides, and thus proteins to be identified.

2.6 Conclusion

IMP represents the first self-contained and standardized pipeline developed to leverage the advantages of integrating MG and MT data for large-scale analyses of microbial community structure and function *in situ* [Muller *et al.*, 2013; Narayanasamy *et al.*, 2015]. IMP performs all the necessary large-scale bioinformatic analyses including preprocessing, assembly, binning (automated) and analyses within an automated, reproducible and user-friendly pipeline. In addition, IMP vastly enhances data usage to produce high-volume and high-quality output. Since its conception, several versions of IMP were released, which include new features, enhanced analytical and usability improvements (**Table 4.1**). The continuous improvement is a testament to the customizability and flexibility of the pipeline, in addition to its emphasis on reproducibility. A notable example would be the integration of automated binning into IMP version 1.4 as an added feature. This represents a vital implementation with regards to reference independent analysis of microbial communities. In addition, the IMP command line and installation procedure was continually enhanced for better user experience.

From an analytical perspective, IMP has been applied to various microbial communities from a multitude of environments, as highlighted within the current chapter. Out of these communities, one of the most discussed instances included datasets derived from human fecal (HF1-5) microbiome samples, which is a proxy for the microbial communities contained within the human gastrointestinal tract (GIT) [Greenhalgh *et al.*, 2016]. The microbial communities within the human GIT microbiome tend to exist in equilibrium/balance. However, disruption of this balance may lead to microbial dysbiosis which has been associated to a range of different diseases including but not limited to, colorectal cancer [Dulal and Keku, 2014; Vogtmann and Goedert, 2016], type I diabetes mellitus [Heintz-Buschart *et al.*, 2016] and Parkinson's disease [Hawkes *et al.*, 2007; Keshavarzian *et al.*, 2015]. Last but not least, the present challenge involves understanding if dysbiosis within the GIT microbiome may be cause or consequence of particular diseases.

Given the biomedical importance of this microbial community, IMP was applied to human GIT microbiomes derived from cancer patients undergoing allogeneic hematopoietic stem cell transplantation, which is an effective treatment for several hematologic malignancies. However, certain cases of such treatments result in adverse outcomes, most notably graft-versus-host disease. Furthermore, allogeneic hematopoietic stem cell transplantation is an intense treatment which is known to greatly impact the human GIT microbiome [Taur *et al.*, 2012]. Previous studies suggested potential links between the drastic changes in the GIT microbiome and graft-versus-host disease [Jenq *et al.*, 2012; Biagi *et al.*, 2015].

Specifically, an early version of IMP (ver. 1.1) was used for detailed analyses of coupled MG and MT datasets from one patient, before and after allogeneic hematopoietic stem cell transplantation for acute myeloid leukaemia. This particular patient developed severe graft-versus-host disease, leading to his/her death nine months after the transplantation. IMP analyses from pre- and post-treatment samples of this patient revealed a drastically decreased bacterial diversity after treatment. Furthermore, a large number of the remaining bacterial populations post-treatment were of strains that encoded higher numbers and higher expression levels for antibiotic resistance genes, thereby demonstrating the long-term effect of the treatment on the GIT microbial community. Overall, the identification of these antibiotic resistant bacterial populations possible within data derived from cancer patient GIT-derived is analogous to the use case highlighted in **Section 2.4.3**, which involved the identification of bacterial populations with unique characteristics (**Appendix A.6**).

Section 2.4.3, demonstrated that IMP is effective for the recovery of complete or nearly complete known viral genomes from microbial community samples. This particular capability of IMP was leveraged upon in the analysis of temporal data sets of the LAMPs (the model system) and is described in detail within the next chapter.

In summary, the analytical capabilities of IMP were successfully applied to both published (new analysis) and new data sets (i.e. cancer patient GIT microbiome and LAMP-derived datasets) to effectively convert these large datasets into information that could be used for detailed downstream analyses. More importantly, this work demonstrates that specific use cases (**Section 2.4.3**) such as finding bacterial populations with unique characteristics (**Section 2.4.3**) or detecting bacteriophage sequences (**Section 2.4.3**) are independent of the datasets used, i.e.: i) identification of bacterial populations with unique characteristics (based on abundance and expression of antibiotic resistance genes) has been possible in cancer patient GIT-derived data (**Appendix A.6**) and ii) identification of putatively novel bacteriophage sequences and putative RNA-based invasive genetic elements, which will be demonstrated in more detail in **Chapter 3** using other LAMP-derived datasets. Finally, the combination of these unique capabilities of IMP, open development and reproducibility should promote the general paradigm of reproducible integrated multi-omic research within the scientific community working on mixed microbial communities.

CHAPTER 3

THE DYNAMICS OF BACTERIOPHAGES AND BACTERIAL HOST POPULATIONS WITHIN THE MODEL SYSTEM

The material within this chapter is foreseen to be submitted for publication in a peer-reviewed journal.

3.1 Abstract

There is presently great interest towards uncovering phage-host interactions and dynamics within microbial communities *in situ*. Accordingly, this chapter describes phage-host dynamics through time-resolved concomitant metagenomic and metatranscriptomic datasets derived from a model microbial system. Large-scale integrated-omic analysis in combination with specialized tools for the extraction of CRISPR-*Cas* information and phage sequences enabled the association of phage and host populations to follow phage-host dynamics. The CRISPR information derived from the *in situ* datasets revealed variability within CRISPR repeats, spacers and flanking regions across the time-series. The high number of identified CRISPR spacers shows CRISPRs to be heterogeneous and dynamic genomic regions, while the high representation of CRISPR repeats within metatranscriptomic data affirms CRISPRs to be transcribed genomic regions, within this system. Population-level analyses focused on the two host populations, i.e. the dominant *M. parvicella* population and a lowly abundant novel taxon termed LCSB005, revealed 150 and eight putative phages associated with these populations, respectively. The observation of phage-host dynamics demonstrated that certain putative phages tend to occur in peaks of high abundances. The abundances of most of these putative phages are not necessarily in sync with their associated hosts, while specific high abundance peaks of particular putative phages do not overlap with other putative phages. A separate analysis showed the high abundances of MT-based contigs that contained protospacers associated with the *M. parvicella* host. In line with these observations, the *M. parvicella* population exhibited high expression of an endoribonuclease *Cas2* gene throughout the entire time-series compared to other *cas* gene types. In conclusion, this study demonstrates the use of an unprecedented time-series integrated multi-omic study to observe phage and host dynamics within a naturally occurring microbial community.

3.2 Background

Recent studies have addressed the lack of information with regards to bacteriophages (and viruses in general) by leveraging data derived from microbial consortia, namely MG information, for the identification of previously unknown viruses [Andersson and Banfield, 2008; Roux *et al.*, 2011, 2015a,b; Paez-Espino *et al.*, 2016]. Some of these aforementioned studies have relied upon publicly available MG datasets to expand the previously sparse knowledge with regards to bacteriophages and bacteriophage-host interactions [Roux *et al.*, 2011, 2015a,b; Paez-Espino *et al.*, 2016]. There have also been efforts aimed at deciphering the dynamics of phage and hosts through laboratory-based co-culture experiments [Cairns *et al.*, 2009], controlled *in vivo* experiments [Reyes *et al.*, 2013; Paez-Espino *et al.*, 2015] and in naturally occurring microbial communities [Andersson and Banfield, 2008; Parsons *et al.*, 2012; Reyes *et al.*, 2012; Stern *et al.*, 2012]. However, there is a relatively low number of studies that follow phage-host dynamics within a time-series setting [Paez-Espino *et al.*, 2015], especially within naturally occurring microbial communities [Parsons *et al.*, 2012]. This is mainly due to the limited number of natural microbial systems that allow time series based studies. In parallel, multi-omic studies are becoming more common due to their superiority compared to single-omic based analyses. However, to the best of our knowledge, there is presently a scarcity of such multi-omic studies aiming at deciphering phage-host interactions (**Table 1.1**). In that light, this work combines the advantages of a time-series and multi-omic data sets to study the model system of LAMPs within a BWWT plant to follow phage-host population dynamics. In order to perform this study, 53 samples of LAMPs, spanning approximately one year and seven months, were subjected to the generation of paired (concomitant) MG and MT datasets. These datasets then underwent sample-wise large-scale integrated-omic analysis using the IMP pipeline (**Chapter 2**). The output produced by IMP was further analyzed to resolve phage-host dynamics within the microbial system of LAMPs (**Figure 3.1**).

Specifically, this work leveraged information contained within CRISPR genomic regions as records of bacteriophage infection histories (or CRISPR *loci*) and more importantly, to formulate associations between bacterial host and bacteriophage populations. Previous studies have demonstrated that information from the CRISPR genomic regions within bacterial genomes (i.e. CRISPR spacer sequences) could be used to associate bacteriophages to specific bacterial populations [Andersson and Banfield, 2008; He and Deem, 2010; Stern *et al.*, 2012; Zhang *et al.*, 2013; Edwards *et al.*, 2015; Paez-Espino *et al.*, 2016]. Given the information contained within the CRISPR regions, and the general interest of the field towards the CRISPR-*Cas* system (**Section 1.3.2**), several tools were designed to extract CRISPR information from isolate prokaryotic genomes and metagenomes [Bland *et al.*, 2007; Edgar, 2007; Skennerton *et al.*, 2013]. In particular, PILER-CR [Edgar, 2007] and the CRISPR recognition tool (CRT) [Bland *et al.*, 2007] were developed to search for CRISPR sequences within isolate genome sequences. Furthermore, metaCRT, which was an extension of the aforementioned CRT [Bland *et al.*, 2007], extracts CRISPR sequences from assembled MG contigs. However, CRISPRs are semi-repetitive genomic regions that either elude, or result in low-quality *de novo* assembly reconstructions [Skennerton *et al.*, 2013]. Moreover, CRISPR sequences are heterogeneous, such that single bacterial species may contain different compositions of CRISPR spacer information [He and Deem, 2010]. Therefore, consensus-based *de novo* assemblies may result in the dilution of the heterogeneity of the CRISPR *loci*, and thereby reduce the overall information availability [Skennerton *et al.*, 2013]. In order to mitigate this issue, CRASS was developed to perform *k*mer based searches of CRISPR information directly from short MG NGS reads (e.g. Illumina paired-end reads) and promises minimal loss of CRISPR sequence heterogeneity information [Skennerton *et al.*, 2013]. In summary, leveraging on both type of tools will increase CRISPR related information from NGS datasets derived from microbial communities (**Figure 3.1**).

This work first describes community-wide dynamics with regards to the CRISPR-based information. CRISPR-based information is defined as "CRISPR elements" in the context of this work and comprise CRISPR repeats, spacers and in certain cases, flanking regions. On the other hand, protospacers are defined as either the original sequence of which the spacers were possibly derived from and/or the targets of spacers, given the mechanism of the CRISPR-*Cas* system (**Section 1.3.2**, [Amitai and Sorek, 2016]). Consequently, contigs containing any protospacers are referred to as protospacers-containing contigs. We then focus on the analyses of specific bacterial populations which contain at least one CRISPR *loci* and accompanying *cas* genes (i.e. CRISPR operon [Jansen *et al.*, 2002; Amitai and Sorek, 2016]). This was achieved by supplementing the present study with information from isolate genomes of lipid accumulating bacterial species (i.e. subset of LAMPs). The CRISPR information was then used to associate these bacterial populations to their putative phages. These putative phages were further analyzed using specialized bacteriophage sequence prediction tools. The abundance patterns of the associated phages and their hosts were following and particular cases of phage-host dynamics were highlighted.

3.3 Methods and material

3.3.1 Sampling and strain collection

Five individual sludge islets were sampled at 53 representative time points from the surface of anoxic tank number one of a biological wastewater treatment plant treating communal effluents (Schifflange, Esch-sur-Alzette, Luxembourg; 49°30'48.29"N; 6°1'4.53"E). Each 'islet' sample is independently collected, transferred into a sterile tube, snap frozen on site and maintained at -80 °C until further processing, and thus represent five biological replicates. The sampling time-series includes two initial sampling dates (4 October 2010 and 25 January 2011) followed by a higher frequency sampling phase beginning on 23 March 2011, which has been carried out until the present day. Samples spanning from 4 October 2010 to 3 May 2012 were selected for downstream processing and analyses (described in sections below). The interval between two sampling dates typically span from 6 to 10 days, with several exceptions where sampling could not be carried out on certain periods due to: i) no surface sludge islets (very often due to heavy precipitations which leads to their dispersion) and ii) WWTP maintenance. In summary, samples used within this work spanned exactly one year and seven months (i.e. 578 days including the start and end date).

In addition, 85 isolate cultures of lipid accumulating bacterial strains were derived from the sludge islets sampled from the same anoxic tank described above. The isolation protocol is described in **Appendix A.4** [Roume *et al.*, 2015].

3.3.2 Extraction of biomolecules

A single biological replicate from all sampling dates between (and inclusive of) 4 October 2010 and 3 May 2012 (**Section 3.3.1**) was randomly selected for high-resolution omic measurements. All biomolecular fractions were obtained using a biomolecular extraction framework that enables recovery of high-quality biomolecular fractions (DNA, RNA, proteins, polar and non-polar metabolites from the biomass as well as from the extracellular compartment) from unique undivided single samples [Roume *et al.*, 2013b,b]. For biomacromolecular purification, we used the AllPrep DNA/RNA/Protein Mini kit (Qiagen) on a batch of randomly selected samples. Resulting biomolecular fractions comprising genomic DNA, RNA, proteins and small molecules were subjected to high-throughput measurement techniques after stringent quality control.

3.3.3 Metagenome and metatranscriptome sequencing

DNA library preparation

The purified DNA fractions (**Section 3.3.2**) from the selected samples suspended in an elution buffer (pH 8.0) were used to prepare a paired-end library with the AMPure XP/Size Select Buffer Protocol [Kozarewa *et al.*, 2009], modified to allow for size selection of fragments using the double solid phase reversible immobilization procedure [Rodrigue *et al.*, 2010]. Size selection yielded metagenomic library fragments with a mean size of 450 bp. All enzymatic steps in the protocol were performed using the Kapa Library Preparation Kit (Kapa Biosystems) with the addition of 1M PCR-grade betaine in the PCR reaction to aid in the amplification of high G+C percentage content templates. The resulting DNA library was subjected to Illumina sequencing (described below). Finally, the purified DNA samples from the isolate genomes were prepared based on a previously described protocol, as highlighted in **Appendix A.4** [Roume *et al.*, 2015].

RNA library preparation

Following RNA purification (**Section 3.3.2**) from selected samples, RNA fractions were ethanol precipitated, overlaid with RNAlater solution (Ambion) and stored at -80 °C. Before sequencing library preparation, the RNA pellet was rinsed twice in 80 % ethanol and twice in 100 % ethanol to remove any excess of RNAlater solution. The pellet was then left on ice to dry. After ethanol evaporation, the RNA pellets were re-suspended in 1mM sodium citrate buffer at pH 6.4. Ribosomal RNAs were depleted using the Ribo-Zero Meta-Bacteria rRNA Removal Kit (Epicentre) according to the manufacturer's instructions. Transcriptome libraries were subsequently prepared using the ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicentre) according to the manufacturer's instructions. The resulting cDNA was subjected to Illumina sequencing (described below).

Next-generation sequencing

DNA and cDNA fractions from the microbial community samples were sequenced on an Illumina Genome Analyser (GA) IIx sequencer, as per described in previous work [Muller *et al.*, 2014b; Roume *et al.*, 2015]. DNA and cDNA samples from all the sampling dates, with exception of the first two initial samples, were randomized prior to sequencing to reduce possible batch effects. Selected samples from the two initial sampling dates (4 October 2010 and 25 January 2011) and three time series sampling dates (5 October 2011, 12 October 2011 and 11 January 2012) were included in previously published work [Muller *et al.*, 2014b; Roume *et al.*, 2015].

The purified DNA from the 85 isolate cultures (described in **Section 3.3.1**) were sequenced on an Illumina HiSeq Genome Analyzer IIx as previously described for the published draft genome of *Rhodococcus* sp. strain LCSB065 of **Appendix A.4** [Roume *et al.*, 2015].

All Illumina GAIIx and HiSeq reads produced were of 100 bp length with an insert size of approximately 300bp spanning the two pairs. Sequencing was performed at TGen North (AZ, USA).

3.3.4 Bioinformatic analyses

Large-scale integrated-omic analysis

Large-scale integrated MG and MT data (Illumina NGS reads) analyses was performed on all the time-series datasets using IMP ver. 1.3. The Truseq2 adapter trimming was carried out in the NGS read preprocessing step. The step for filtering reads of human origin was omitted from the preprocessing. The MEGAHIT *de novo* assembler [Li *et al.*, 2015] was selected for co-assembly of MG and MT data. All other parameters of IMP were retained as the default. The number of threads was set to 8, total memory was set to 256 GB and memory-per-core was set to 32 GB.

Isolate genome assembly

Illumina NGS reads from the DNA samples of isolate genome LCSB005 underwent *de novo* assembly using SPAdes. The genome of *Candidatus* Microthrix parvicella Bio17-1 was obtained from NCBI (Bioproject:

PRJNA174686).

Identification of CRISPR elements

CRASS [Skennerton *et al.*, 2013] was applied to the IMP-preprocessed MG and MT paired-end and single-end reads (**Figure 3.1**), using 12 threads while retaining all other parameters as default. Accordingly, CRASS extracts information of CRISPR spacers, repeats and flanking regions. MetaCRT, i.e. the metagenome version of CRT [Bland *et al.*, 2007] was applied on the IMP-based MT-contigs and co-assembly (**Figure 3.1**) using default parameters. MetaCRT extracts information on CRISPR spacers and repeats, but does not extract CRISPR flanking regions. The FASTA header information of all sequences (CRISPR repeats and spacers) extracted by metaCRT were edited to append sample information for use in downstream analyses.

All the CRISPR elements, i.e. spacers, repeats and flanking regions were clustered using CD-HIT-EST [Fu *et al.*, 2012] to reduce the redundancy of the information. Accordingly, CRISPR spacers were clustered based 0.99 % sequence identity (-c 0.99), covering the entire length of the both compared sequence (-aL 1 -aS 1 -s 1). CRISPR flanking regions were collapsed using 99 % sequence identity (-c 0.99) with at least 97.5 % coverage of both the compared sequence (-aL 0.975 -aS 0.975 -s 0.975). The clustering parameters of CRISPR repeats were determined by manually observing the clustering of known CRISPR repeats belonging to a single CRISPR *loci* of *Microthrix* sp. Accordingly, the sequence identity was first set to 99 % (-c 0.99) and the sequence coverage was set to 100 % (-s 1). Both these parameters were reduced by 5 % in each iteration (manually) until all the CRISPR repeats from single CRISPR *loci* (of *M. parvicella*) were observed to be clustered into a single cluster. At the end of these iterations, the parameters to cluster the CRISPR repeats were set to 80 % sequence identity (-c 0.8), covering the length of at least 75 % of the shorter sequence (-s 0.75). The FASTA headers of all the sequences were left unchanged (-d 0). 12 threads were used for all CD-HIT-EST based analyses. The clustering procedure of the CRISPR elements yielded non-redundant sequences CRISPR repeats, spacers and flanking regions.

Identification of protospacers and protospacers-containing contigs

A nucleotide blast (blastn [Johnson *et al.*, 2008]) search was performed for the non-redundant CRISPR spacers (i.e. defined as the blastn query sequences within the -query parameter) against all the IMP-based co-assembled contigs (i.e. defined as the blastn reference database within the -db parameter) derived from the time-series datasets. Matches based on the aforementioned search were defined as protospacers. The parameters described for the initial blastn search of the CRISPRtarget software [Biswas *et al.*, 2013; Edwards *et al.*, 2015] were replicated for this task. Accordingly, the parameters used included: i) the blastn-short task setting (-task 'blastn-short'), ii) mismatch penalty of 1 (-penalty -1) iii) a gap opening penalty of 10 (-gapopen 10) and iv) dust filtering turned off (-dust “no”). Similarly, the non-redundant CRISPR repeats and flanking regions were searched against all the IMP co-assembly contigs using blastn, with the blastn-short task settings (-task 'blastn-short'), retaining all other parameters as default.

Self-matches of CRISPR spacers were removed/filtered in two ways: i) by removing any IMP co-assembled contigs found to match to any CRISPR repeat sequence in the blastn search and ii) by removing all IMP co-assembled contigs identified by metaCRT to be carrying CRISPR sequences (i.e. based on metaCRT results). Accordingly, the remaining CRISPR spacer matches post-filtering are defined as protospacers and the respective contigs that contain these protospacers were defined protospacer-containing contigs and retained for further analyses.

Identification of candidate bacterial host populations

CRISPR repeat elements were used to link the CRISPR sequences (i.e. associated repeats, spacers and flanking regions) to bacterial populations by performing a blastn search against a collection of 86 isolate draft genomes, of which 85 were derived from floating sludge islet containing LAMPs (described in **Sections 3.3.1** and **3.3.3**) with the exception of the draft genome of *Candidatus* Microthrix parvicella Bio17-1, which was obtained from previously published work [Muller *et al.*, 2012]. The two bacterial genomes containing the highest number of CRISPR repeat matches were retained for further analyses.

The selected isolate genomes were scanned with metaCRT [Bland *et al.*, 2007] to detect the presence of CRISPR *loci* within the genomes. Additionally, the genes within those genomes were annotated, based on function using Prokka ver. 1.11 [Seemann, 2014]. The presence of *cas*genes within those genomes were manually inspected.

Identification of putative phage contigs

Protospacer-containing contigs of $\geq$ 1 kb, which were linked to the two bacterial isolate draft genomes (via CRISPR repeats), were selected for further processing and analysis. In order to avoid false detection of bacteriophage contigs, further removal of CRISPR spacer self-matches was performed. Accordingly, all population-associated protospacer-containing contigs were: i) re-analyzed with metaCRT and ii) blast [Johnson *et al.*, 2008] searched (using default parameters) against their respective associated host genomes. Protospacer-containing contigs encoding any CRISPR sequences (i.e. based on metaCRT [Bland *et al.*, 2007] output) and/or showing high similarity to their respective associated host genomes (based on blastn search) were excluded from further downstream analyses. The remaining host-associated protospacer-containing contigs were analyzed using VIRSorter [Roux *et al.*, 2015a] to predict putative phage contigs. All the host-associated protospacers containing contigs, including those not predicted by VIRSorter as putative phage sequences, were retained for further downstream analyses.

Identification of putative RNA-based invasive genetic elements

Putative RNA-based invasive genetic elements (RIGes) were identified based on the following criteria: i) contig is not represented on the MG level, ii) entire length of contig is covered by MT reads (i.e. from end to end and not just intragenic regions).

Estimating the abundance of putative phages and corresponding bacterial hosts

IMP preprocessed MG and MT reads from all the time-series samples were mapped against the two bacterial genomes of interest and all putative bacteriophage contigs. The average depth of coverage of MG (and MT) data was computed to represent abundances of the host and phage populations.

Observation of phage and host dynamics

Putative phage contigs associated with *M. parvicella* with a mean depth of coverage below one were excluded from further analyses. Putative phages associated with LCSB005 were retained for further downstream analyses, due to their overall low abundance.

Gene expression analyses

The DEseq2 package [Anders and Huber, 2010], within the R statistical software was used to normalize the *M. parvicella* gene-level MT depth of coverage. An expression matrix was formulated using the "DESeqDataSetFromMatrix" function and normalized using the "estimateSizeFactors" function. Normalization was carried out based on depth of coverage values of all predicted genes within the *M. parvicella* genome, but only the normalized counts of the *cas* genes are reported in the analyses. This analysis was not performed on the LCSB005 population due to the relatively low abundance of the population throughout the time-series.

Statistical analysis and visualization of data

The R statistical software was used for all statistical analysis (including those described in **Sections 3.3.4** to **3.3.4**) and data visualization (R ggplot2 package). Summaries of CRISPR elements were merged within R. An artificial coverage value of 1 is added to CRISPR spacers detected using metaCRT, since this information is not provided by the software. This is under the assumption that if a spacer was detected within the assembled contigs, there should be a coverage of at least 1.

Computational platforms

IMP ver. 1.3 was executed on a Dell R820 machine with 32 Intel(R) Xeon(R) CPU E5-4640 @ 2.40GHz physical computing cores (64 virtual), 1024 TB of DDR3 RAM (32 GB per core) with Debian 7 Wheezy as the operating system. Additional large-scale analysis outside the scope of IMP was performed on the Gaia cluster of the University of Luxembourg HPC platform [Varrette *et al.*, 2014].

3.4 Results

A simplified schema of the applied workflow is represented in **Figure 1.1**. A total of 53 samples of sludge islets derived from an anoxic tank of a BWWT plant were selected for extensive biomolecular extraction, systematic high-throughput measurements and large-scale bioinformatic analyses. These samples represent a time-series of one year and seven months. In most cases, the interval between two sampling events was approximately one week (i.e. mean 8 days, standard deviation 16 days) apart, with the exception of the two initial samples and several other samples within the time-series (described in **Section 3.3.1**).

These collected samples underwent a comprehensive biomolecular extraction that yields all types of biomolecules (DNA, RNA, protein and metabolites), which were then subjected to high-throughput measurements. This study focuses specifically on the data derived from the DNA and RNA fractions, namely MG and MT data. The generated MG and MT data underwent integrated omic analyses with IMP. The output from IMP, which was required for the extraction of information with regards to CRISPR genomic regions was retained for further downstream analyses. These included the: i) prepocessed MG and MT reads (paired-end and singleton sequences), ii) MT contigs, iii) co-assembly contigs and, to a certain extent, iv) gene annotation information. The results of their analysis are used to describe the community-level dynamics of elements within CRISPR genomic regions (repeats and spacers) over time. Information from 86 draft genomes of bacterial species that exhibit lipid accumulating phenotypes (**Figure 3.1**) were used to focus the analysis on two specific lipid accumulating microbial populations (LAMPs).

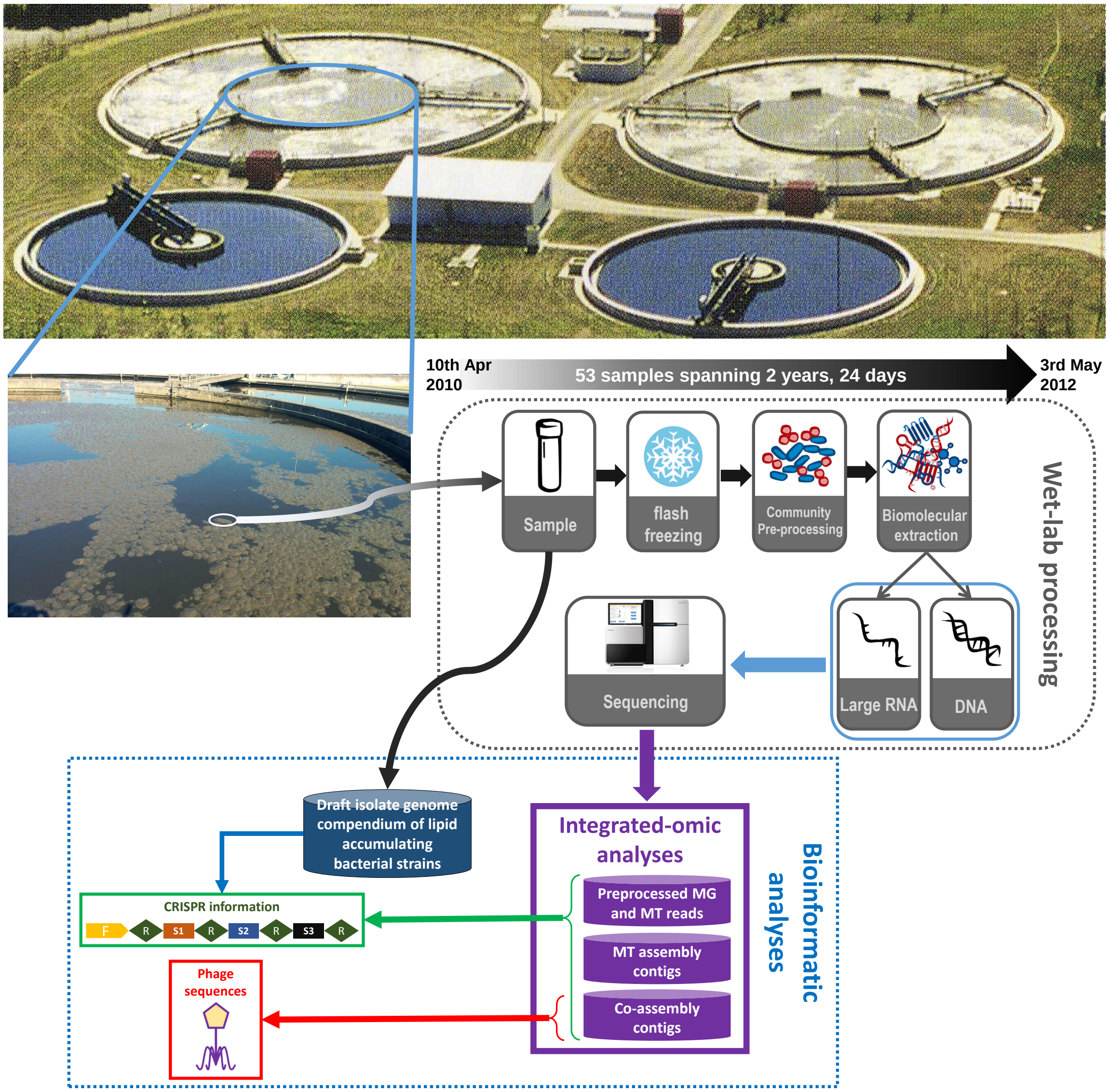


Figure 3.1: Simplified schema for the study of phage-host interaction. Samples of floating sludge islets, containing LAMPs are collected from anoxic tank number 1 of the Schifflange biological wastewater treatment plant, Esch-sur-Alzette, Luxembourg (49°30’ 48.29” N; 6°1’ 4.53” E). Selected samples were subjected to a comprehensive biomolecular extraction. The DNA and RNA fractions undergo high-throughput sequencing to generate metagenomic (MG) and metatranscriptomic (MT) data, respectively. These datasets undergo integrated omic analyses using IMP. The preprocessed MG and MT reads, MT contigs and co-assembly contigs generated by IMP are used for the extraction CRISPR information, i.e. CRISPR-repeats (labelled R) and -spacers (labelled S1…S3). CRISPR repeat sequences are used to link identified CRISPR sequences to draft isolate genomes of lipid accumulating bacterial strains. CRIPSR spacer sequences are used to detect bacteriophage (phage) derived contigs from the co-assembly contigs. The complementary information from the CRISPR-repeats and -spacers are used to associate bacterial populations to the phage contigs.

3.4.1 Large-scale analyses using IMP

Coupled MG and MT dataset (concomitant extractions) from each time point underwent sample-wise (i.e. datasets from each sample processed separately) large-scale integrated omic processing and analyses using IMP. NGS of DNA and RNA (cDNA) of all the time-series samples yielded a total of ~1.5 × 10^9 and ~1.75

$\times$ 10^9 paired-end reads, respectively (3.26 $\times$ 10^9 in total). The processing and analysis using IMP yields large volumes of output and information, including preprocessed NGS reads, assembled contigs (MT and co-assembly contigs), gene annotation, taxonomic assignments of contigs, variant calls and population-level genomic bins (**Chapter 2** and **Sections 2.3.1** and **2.4.1**). However, within the scope of this work, only the IMP: i) preprocessed MG reads ii) preprocessed MT reads, iii) MT contigs and iv) co-assembly contigs from the IMP output were utilized for further analysis. The relevant information from IMP input and output used within this work is summarized in **Figure 3.2**. Overall, the preprocessing of MG reads retained ~1.36 x 10^9 paired-end reads and ~9.54 $\times$ 10^8 singleton reads (for which the mate discarded). Similarly, the preprocessing of MT reads retained a total of ~8.26 x 10^8 paired-end reads and ~1.04 x 10^8 single-end reads. A large fraction of the MT reads were removed during the rRNA filtering step of IMP (**Figure 3.2**). Preprocessed MG and MT reads were retained for the downstream IMP-based iterative co-assembly procedure and other analyses steps. The iterative assembly of MT reads carried out within IMP yielded a total of ~5.34 $\times$ 10^6 MT contigs (mean 100,733; standard deviation 27,534 per sample), while the final co-assembly of MG and MT data from IMP produced a total of ~2.1 x 10^7 contigs (mean 414,872; standard deviation 79,092 per sample). The preprocessed NGS reads, MT contigs and co-assembly contigs were used for the extraction of CRISPR elements. In addition, the co-assembled contigs were used for the extraction of putative phage sequences/genomes.

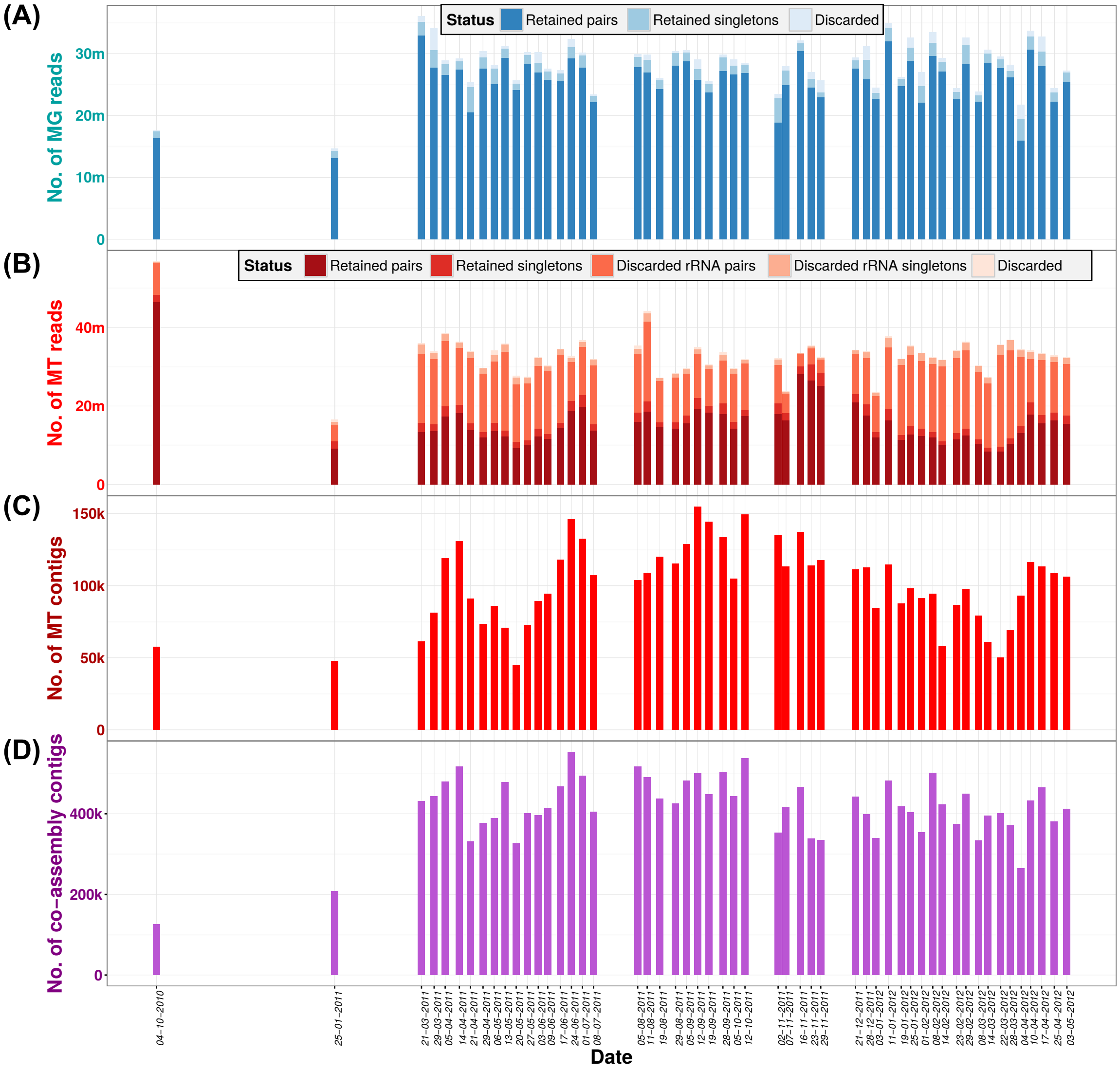


Figure 3.2: Summary of input and output of IMP analyses of the LAMPs time-series datasets. **(A)** Summary of the preprocessing of metagenomic (MG) data using IMP. **(B)** Summary of the preprocessing of metatranscriptomic (MT) data using IMP. **(C)** Summary of MT contigs generated by IMP. **(D)** Summary of co-assembled contigs generated by IMP. The x-axis represent the exact sampling dates. Please refer to **Table C.1** for detailed information.

3.4.2 Community-level analysis of CRISPR elements and protospacers

The output from IMP (**Section 3.4.1**) was used as input for CRASS [Skennerton *et al.*, 2013] and metaCRT [Bland *et al.*, 2007] to extract CRISPR information. CRASS was used to extract CRISPR information directly from the unassembled NGS reads (i.e. IMP preprocessed MG and MT paired and singleton reads). On the other hand, metaCRT was used to detect CRISPRs in the assembled contigs (i.e. IMP-based MT-assembled contigs and co-assembled contigs). Within the scope of this work, information content within CRISPR genomic regions include CRISPR repeats, spacers and flanking regions and are collectively referred to as CRISPR elements. **Figure 3.3** summarizes the number of CRISPR elements detected by CRASS and

Table 3.1: Summary of CRISPR elements detected by different methods.

CRISPR element	MetaCRT	CRASS
repeat	76,857 (86.54 %)	11,950 (13.46 %)
spacer	168,465 (32.16 %)	355,411 (67.84 %)
flanking regions	NA	16,730

metaCRT, respectively. MetaCRT is more effective at extracting CRISPR repeat information, thereby allowing the detection of 76,857 (86.5 %) non-unique repeats across datasets from the entire time-series, compared to the 11,950 (13.46 %) detected by CRASS (**Table 3.1**). On the contrary, CRASS is more effective in extracting CRISPR spacer information such that it detected approximately 68 % non-unique CRISPR spacers (**Table 3.1**). In addition, CRASS also provides information of CRISPR flanking regions, which are regions upstream (from the first repeat-spacer occurrence in the CRISPR *locus*) and downstream (from the last repeat-spacer occurrence within the CRISPR *locus*) [Skennerton *et al.*, 2013]. Overall, both methods are complementary to each other and allowed the maximization of CRISPR information from the data.

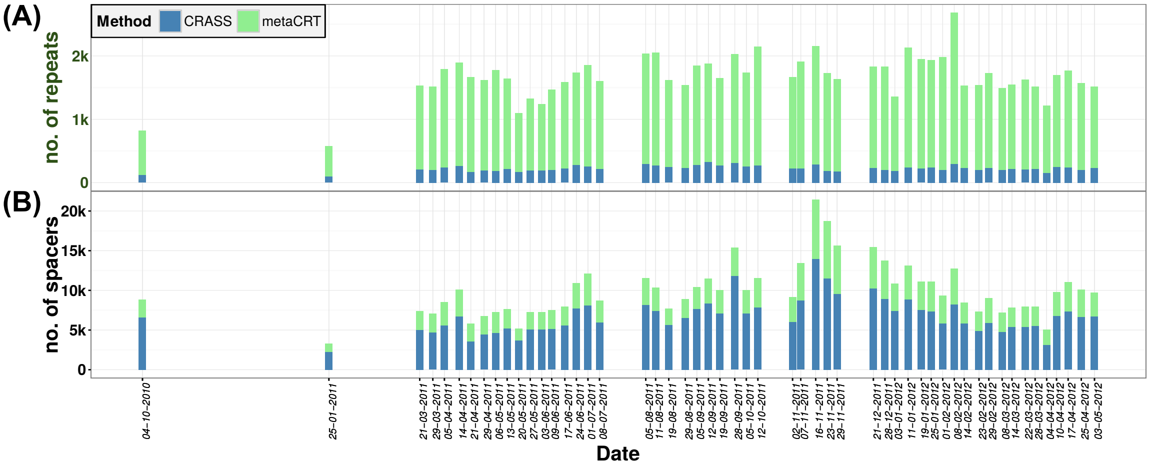


Figure 3.3: CRISPR detection summary using different methods. **(A)** CRISPR repeat detection. **(B)** CRISPR spacer detection. Colours of the bars represent detection on the different omic levels. The labels in the x-axis represent the exact sampling dates.

CRISPRs are known to be highly dynamic and heterogeneous genetic regions, especially with regards to the CRISPR spacers, which undergo constant change through either deletion of older spacers and/or insertion of new spacers [Amitai and Sorek, 2016]. Using a time-resolved dataset, the changes to the different CRISPR elements within community are observable over time. The results indicate that number of different CRISPR elements detected vary over time, with changes to either CRISPR repeats (**Figure 3.4**A), spacers (**Figure 3.4**B) and/or flanking sequences (**Figure 3.4**C). CRISPR spacers show the highest occurrences in terms of absolute numbers, followed by repeat sequences. Furthermore, the present work leveraged the availability of coupled MG and MT datasets to detect these CRISPR elements. Consequently, the different CRISPR elements could be separated based on detection on a particular omic level. CRISPR repeats are covered well in both MG and MT complements. On the other hand, CRISPR spacers and flanking regions appear to be more prominent within the MG data (**Figure 3.4**). Overall, a total of 88,807 (non-unique) CRISPR repeats sequences were detected over the 53 time points with a mean length of 32 bp (standard deviation 7.87) (**Figure 3.5**), while the shortest and longest repeats were 20 and 77 bp in length, respectively. On the other hand, there were a total of 523,876 CRISPR spacers (non-unique) detected over the 53 time points, with a mean length of 33 bp (standard deviation 5.04) (**Figure 3.5**), whilst the shortest and longest spacers were 11 and 119 bp in length. Additionally, there were a total of 16,730 flanking regions detected by CRASS, which span from 11 to 98 base pairs (mean 39, standard deviation 14.55) in length (**Figure 3.5**). Based on the % G+C, CRISPR repeats demonstrate a slight skew (based on density) towards approximately 30 % G+C, while CRISPR spacers and flanking regions demonstrate an bell shaped distribution around the centered around 50 % G+C (**Figure 3.5**).

In order to detect the occurrence frequencies of the different CRISPR elements across the time-series, non-redundant sets of CRISPR repeats, spacers and flanks were generated using CD-HIT-EST [Fu *et al.*, 2012]. CRIPSR repeats were clustered at a threshold of 80 % sequence identity, covering at least 75 % of the shorter sequence. This threshold was tuned based on the observed (via manual inspection) convergence/clustering of CRISPR repeats from a single CRISPR *locus*. Upon clustering, the number of non-redundant set of CRISPR repeats were 8,101 for which > 50 % of them occurred only in a single time point within the entire time-series, while there were seven CRISPR repeats that occurred in all 53 time points (**Figure 3.5**). On the other hand, CRISPR spacers were clustered on a more stringent criteria, i.e. based on 99 % sequence identity, considering the entire length of the sequences. This is mainly due to their requirement of being highly specific to their target (protospacers), for them to effectively function as guide RNA molecules [Marraffini and Sontheimer, 2010; Amitai and Sorek, 2016]. Clustering of the spacers yielded 176,763 non-redundant spacers, for which 71.4 % of them occur within a single time point within the entire time-series with only two spacer occurring in almost the entire time-series (**Figure 3.5**). Similarly, only one flanking sequences occurred in all the time points.

Protospacers are complementary sequences of CRISPR spacers which represent either the origin of the CRISPR spacer sequences and/or targets for inhibition/splicing of invasive genetic elements [Marraffini and Sontheimer, 2010; Amitai and Sorek, 2016]. A blastn [Johnson *et al.*, 2008] search was performed on the non-redundant set of CRISPR spacers, obtained from all the time-series datasets against all IMP-based iterative co-assembled contigs. The blastn parameters of CRISPRtarget were used for this task [Biswas *et al.*, 2013]. In this work, a protospacer is defined as the part of the contig either matching or showing high similarity (based on identity and query sequence coverage with regards to the CRISPR spacer sequence) to a CRISPR spacer sequence. In order to avoid possible self-matching of CRISPR-spacers, only protospacers occurring within contigs that do not contain any CRISPR repeats were considered for further analyses.

A total of ~5.9×10^6 protospacers were detected within this analysis (**Figure 3.4**). In general, the number of protospacers remain relatively constant throughout the time-series. Interestingly, the dataset that yielded the highest number of CRISPR spacers (16 November 2011) did not necessarily yield many of protospacers targets (**Figure 3.4**). On the other end of the spectrum, the dataset from 28 September 2011, yielded the highest number of protospacers (508,357) protospacers, from 15,387 CRISPR spacers (**Figure 3.4**). Detailed inspection of CRISPR spacers and their corresponding protospacer targets revealed one particular CRISPR spacer with 20,900 apparent protospacers targets, while the mean and median of number of protospacers targets per spacer is approximately 63 and 12 respectively (**Table 3.2**). In summary, this analysis demonstrates

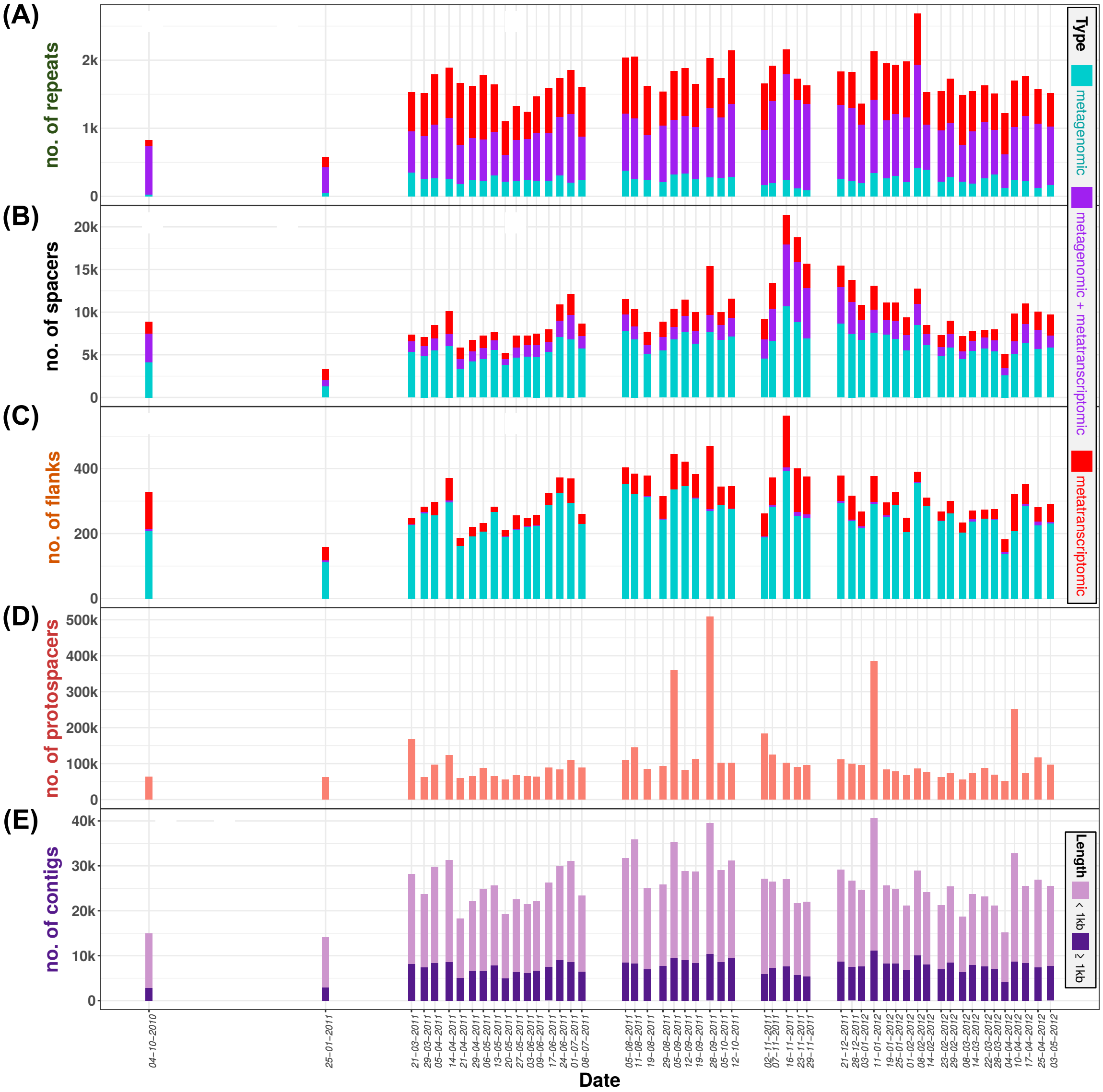


Figure 3.4: Community-wide dynamics of CRISPR elements. **(A)** Number of repeat sequences per time point. **(B)** Number of spacer sequences per time point. **(C)** Number of flanking sequences. **(D)** Number of protospacers detected. **(E)** Number of protospacer-containing contigs. Colours of the bars represent detection on the different omic levels. The labels in the x-axis represent the exact sampling dates. Please refer to **Table C.2** for detailed information.

that single CRISPR spacers are able to target multiple protospacers in distinct genomic locations.

A contig that contains at least one protospacer is hereby defined as protospacer-containing contig. Based on this definition, all the protospacers detected within this analysis could be traced back to a total of ~1.37 $\times\ 10^6$ protospacer-containing contigs, with a mean of 25,836 (standard deviation 5483.54) protospacer-containing contigs per time-point (**Figure 3.4**). Overall, a high number of detected protospacers within one dataset did not translate into a large number of protospacers-containing contigs (i.e. 28 September 2011 dataset, **Figure 3.4**). Detailed inspection of protospacers-containing contigs revealed the occurrence of 1,081 protospacers within a single contig, which was the highest number detected within this analysis

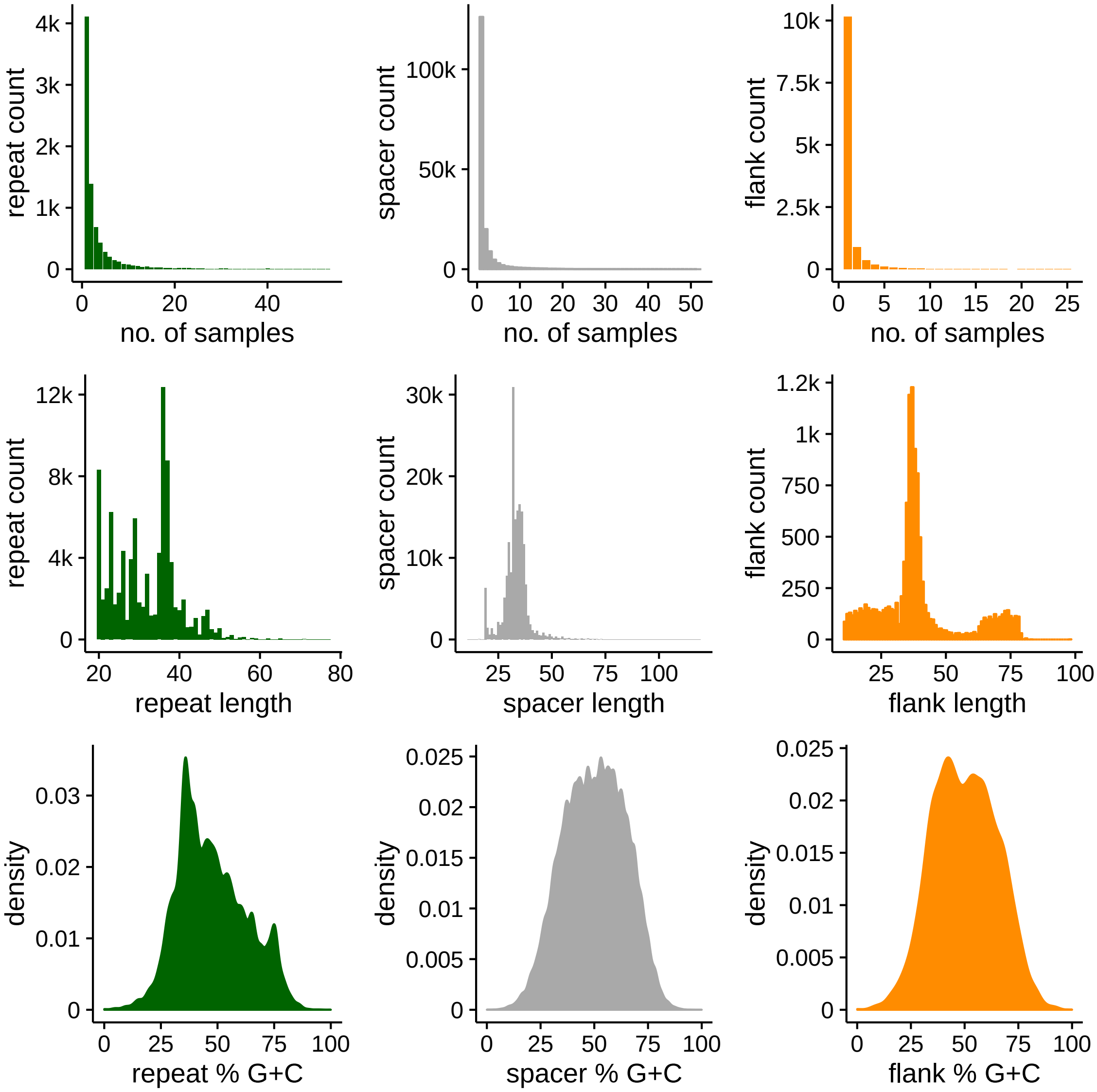


Figure 3.5: Summary of CRISPR element information. From top to bottom, each row represents a different measure, i.e. occurrence, length and % G+C. From left to right, columns (and colours) represent the different CRISPR elements, i.e. repeat, spacer and flanking sequence. Occurrence is based on the number of non-redundant CRISPR elements.

(**Figure 3.2**). It is also important to note that the large number of observed protospacers, relative to the lower number of protospacer-containing contigs could also be attributed to the redundancy among the contigs. In summary, we show that a single contig may carry multiple protospacers. Finally, these protospacers containing contigs represent putative invasive genetic elements that are targeted by CRISPR-*Cas* systems within bacterial populations, and thereby include putative phage-derived contigs.

Table 3.2: Summary statistics of CRISPR elements.

CRISPR element	Measure	Min	Q1	Median	Mean	Q3	Max	Std. dev
repeat	length (bp)	20	25	32	32	37	77	7.87
spacer	length (bp)	11	32	33	33	36	119	5.04
flank	length (bp)	11	34	37	39	41	98	14.55
repeat	% G+C	0	36.11	45.65	47.56	57.89	100.00	15.03
spacer	% G+C	0	37.50	47.06	48.40	59.38	100.00	14.29
flank	% G+C	0	37.21	47.37	48.62	59.46	100.00	14.96
repeat	occurrence (sample count)	1	1	1	3.65	3	53	6.24
spacer	occurrence (sample count)	1	1	1	2.36	2	52	4.26
flank	occurrence (sample count)	1	1	1	1.39	1	25	1.38
spacer	protospacer (count)	1	3	12	62.57	44	20,900	194.07
protospacer-containing contig	protospacer (count)	1	1	1	4.28	3	1,081	15.15

3.4.3 Population-level analysis of CRISPR elements

This work leveraged a compendium of 86 draft isolate genomes of lipid accumulating bacterial strains to focus the analysis on specific bacterial populations. CRISPR repeats and flanking regions extracted from the entire community were searched (blastn [Johnson *et al.*, 2008]) against all draft isolate genomes, such that identified CRISPR repeats could be conclusively linked to single bacterial populations. Out of the 86 draft isolate genomes, only 16 draft genomes contained at least one associated CRISPR repeat sequence from the entire collection of CRISPR repeats identified in the large-scale multi-omic analyses (**Table 3.3**). Given the limited number of bacterial populations that could be identified, two specific strains that contained the highest number of associated CRISPR repeats and flanking regions for were chosen for detailed population-level analyses (**Table 3.3**). These strains include: i) the *Candidatus* Microthrix parvicella Bio17-1 (hereafter referred to as *M. parvicella*) [Muller *et al.*, 2012], a filamentous lipid accumulating bacteria that is known to be dominant within BWWT plants, including the present system of study [Blackall *et al.*, 1996] and ii) a novel lipid accumulating organism referred to as LCSB005 (**Figure B.1**), that occurs in low abundance within the model system (refer to **Section 3.4.5** for inferred abundances and **Figure B.1** for microscopy images).

These selected draft genomes were scanned for CRISPR *loci* (using metaCRT [Bland *et al.*, 2007]) and CRISPR-associated genes (using Prokka [Seemann, 2014]), to identify CRISPR-*Cas* systems within these bacterial strains. Accordingly, there were four CRISPR *loci* detected within the *M. parvicella* genome, whereby CRISPR repeats within different *loci* were different from each other. However, only one contained more than 20 consecutive repeat-spacer pairs (89 total). Moreover, this particular CRISPR *locus*, (based on metaCRT results [Bland *et al.*, 2007]), is accompanied by six *cas* genes upstream (within positions 473,221 - 482,572 of scaffold 2, based on the annotation by Prokka [Seemann, 2014]). Similarly, seven CRISPR *loci* were detected within the isolate genome of LCSB005 (based on metaCRT [Bland *et al.*, 2007] and Prokka [Seemann, 2014]). However, unlike *M. parvicella* the seven CRISPR *loci* detected within the genome of LCSB005 shared four CRISPR repeat types. In addition, there were at least four of these CRISPR *loci* that were located in close proximity to (upstream/downstream) clusters of *cas* genes, of which three contained

Table 3.3: Summary of the number of CRISPR repeats and flanks detected within draft genomes of lipid accumulating bacterial species.

Isolate genome	No. of CRISPR repeats	No. of CRISPR flanks
**Candidatus* Microthrix parvicella Bio17-1	199	4720
LCSB005	79	2
LCSB403	35	1
LCSB408	29	29
LCSB454A	14	0
LCSB357A	6	6
LCSB589	4	2
LCSB455	3	0
LCSB556	3	1
LCSB663	3	0
LCSB252A	2	0
LCSB462B	2	1
LCSB406	1	0
LCSB541	1	1
LCSB565	1	8
LCSB660	1	0

*Published draft genomes

approximately 40 CRISPR repeat-spacer pairs.

CRISPR elements associated to *M. parvicella* were found in all the datasets throughout the time-series, with the exception of the dataset from 4 October 2010. These datasets show that the *M. parvicella* CRISPR repeats are consistently covered by both MG and MT data, demonstrating transcription of these regions throughout all samples within time-series, with the exception of the first two initial time points (**Figure 3.6**). In addition, a large number of CRISPR spacers and flanking regions are also observed for the *M. parvicella* population (**Figure 3.6**). Due to the large number of CRISPR spacers associated to *M. parvicella*, there is an accompanying large number of protospacer information, and therefore large number of associated protospacer-containing contigs (**Figure 3.6**). Interestingly, similar abundance patterns are observed in the CRISPR elements, protospacers and protospacers-containing contigs associated with the *M. parvicella* population, which was contrary to the observations within the community-wide analysis (**Section 3.4.2**).

On the other hand, there was a sparse occurrence of CRISPR elements associated to LCSB005, appearing only between datasets 5 September 2011 and 14 February 2012 (**Figure 3.7**A & B). However, despite the relatively small amount of CRISPR information associated to this population, protospacers and protospacer-

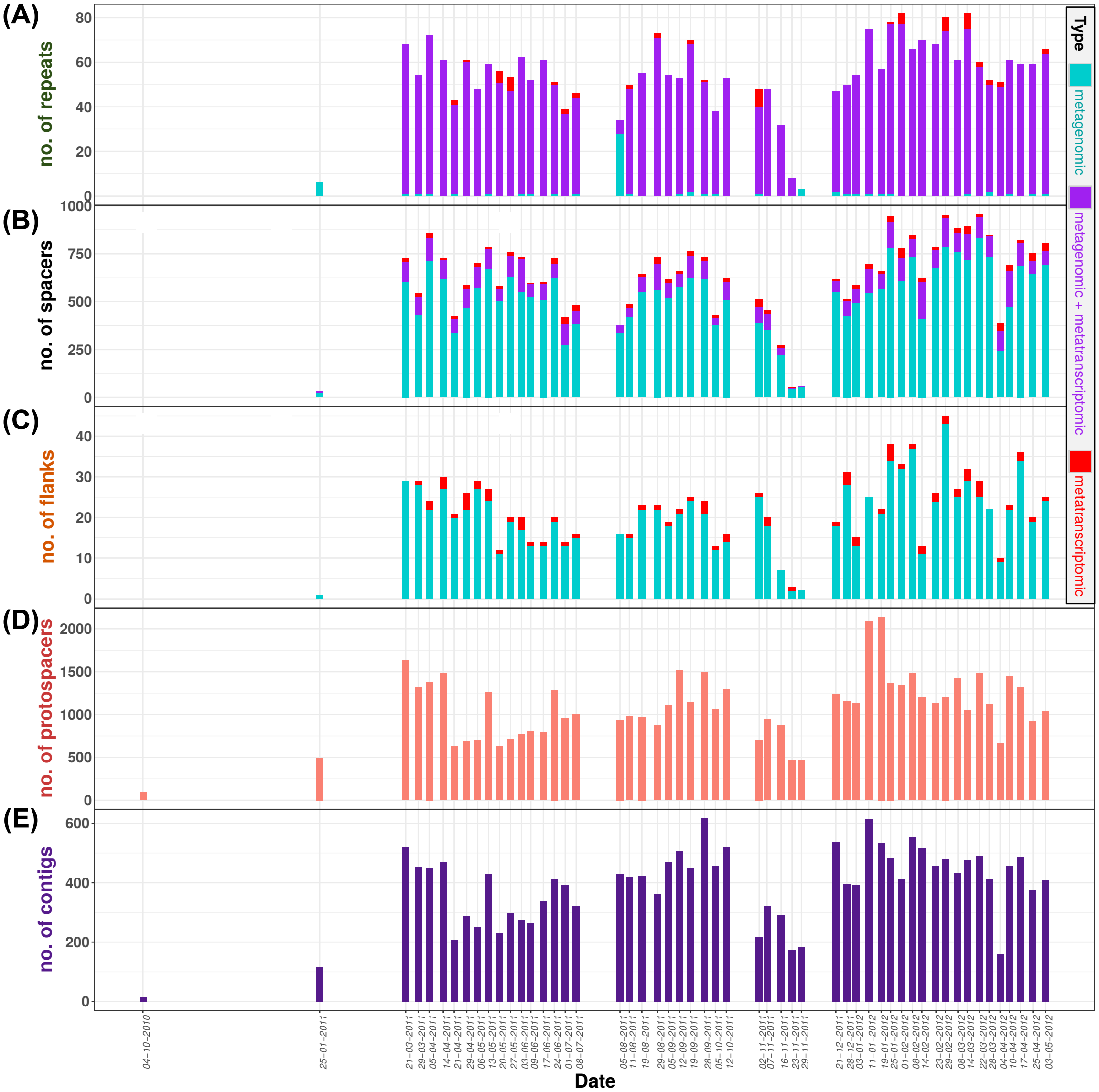


Figure 3.6: ***M. parvicella*** **population-level dynamics of CRISPR elements** The x-axis represent the exact sampling dates. **(A)** Number of repeat sequences per time point. **(B)** Number of spacer sequences per time point. **(C)** Number of flanking sequences. **(D)** Number of protospacers detected. **(E)**Number of protospacer-containing contigs. Colours of the bars represent detection on the different omic levels. The labels in the x-axis represent the exact sampling dates. Please refer to **Table C.3** for detailed information.

containing contigs could be detected in samples from all time points, with exception of the 4 October 2010 dataset.

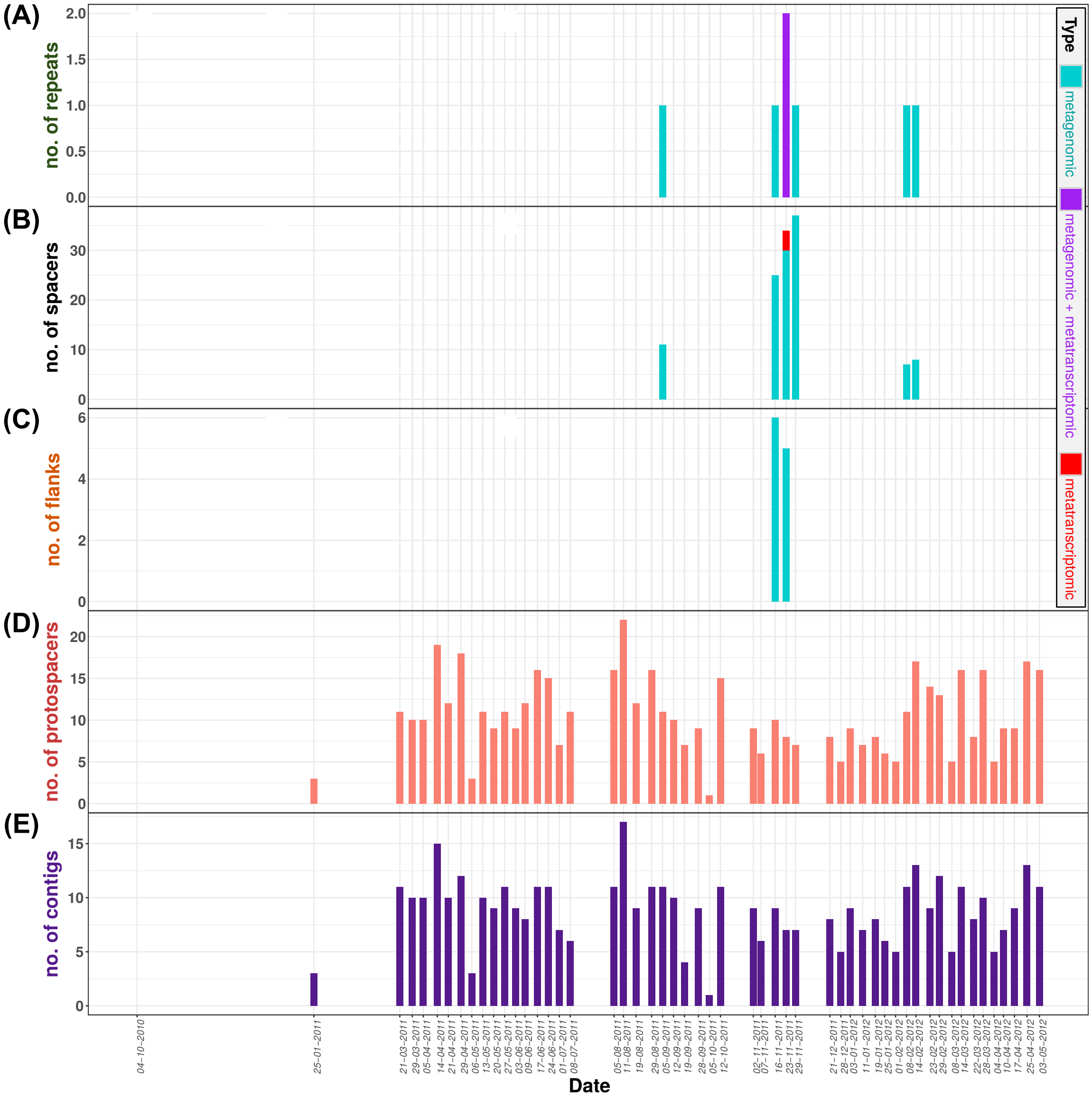


Figure 3.7: LCSB005 population-level dynamics of CRISPR elements The x-axis represent the exact sampling dates. **(A)** Number of repeat sequences per time point. **(B)** Number of spacer sequences per time point. **(C)** Number of flanking sequences. **(D)** Number of protospacers detected. **(E)** Number of protospacer-containing contigs. Colours of the bars represent detection on the different omic levels. The labels in the x-axis represent the exact sampling dates. Please refer to **Table C.4** for detailed information.

3.4.4 Putative bacteriophage sequences

The protospacer-containing contigs linked to the defined populations were selected for further analyses, namely using a specialized viral sequence annotation and identification tool [Roux *et al.*, 2015a]. A total of 20,632 protospacer-containing contigs associated to the *M. parvicella* population yielded 150 putative phage contigs (**Table 3.4**). Similarly, of the 4,232 protospacer-containing contigs identified for the LCSB005 population, only 8 of them were predicted to be phage derived sequences, whereby most of them were predicted at a low confidence (**Table 3.4**) [Roux *et al.*, 2015a].

Out of the 150 predicted phage sequences (using VIRSorter [Roux *et al.*, 2015a]) associated with the *M. parvicella* population, five sequences were selected for further analyses (**Table 3.4**). This selection was based on: i) high-confidence predictions by VIRSorter (i.e. category 1 or 2) [Roux *et al.*, 2015a], ii) presence of predicted genes of which at least one of them should be annotated by "phage hallmark genes" (defined by VIRSorter [Roux *et al.*, 2015a]), iii) length above 2kb, iv) contain $\geq$ 10 protospacers and v) exhibits a mean depth of coverage of $\geq$ 1 throughout the entire time-series. The putative *M. parvicella*-associated phages, selected based on the aforementioned criteria, are hereby referred to as "Putative phages M1 - 5" (**Table 3.4**). Other notable examples, that were not considered for further analyses, include the two putative phage sequences "D36_N_contig_120762" and "D36_N_contig_368349". These sequences contained a particularly high number of protospacers (306 and 252 respectively), and lengths > 10 kb, while the latter was annotated with a known phage gene. However, these contigs were filtered out in the early stages as they did not exhibit sufficient depth of coverage throughout the time-series (i.e. mean depth of coverage < 1).

Similarly, putative phage sequences associated with the LCSB005 population were also selected for further inspection and analyses. Nevertheless, given the sparse number of protospacer-containing contigs associated with the LCSB005 population and the overall low confidence predictions [Roux *et al.*, 2015a], a lower selection stringency was applied for the putative phages of LCSB005 resulting in two putative phage contigs being selected, based on: i) predicted genes within contigs, ii) lengths of at least 2kb, iii) with at least one protospacer occurring within contigs and iv) demonstrating high-depth of coverage in at least one sample within the entire time-series. The putative LCSB005 associated phages, selected based on the aforementioned criteria are hereby referred to as "Putative phage L1 and L2" (**Table 3.4**).

Table 3.4: Summary of putative phages of the *M. parvicella* and LCSB005 populations.

Bacterial host	Contig ID	*No. of predicted genes	*Category	*No. of phage hallmark genes	Contig length	Protospacer count	Pearsons correlation
M. parvicella	^{M1}D05_G7_contig_328367	5	1	1	3,614	15	0.450
M. parvicella	^{M2}D05_N_contig_128863	4	3	NA	1,042	13	0.556
M. parvicella	^{M3}D38_N_contig_103085	3	1	2	2,850	21	0.175
M. parvicella	^{M4}D22_N_contig_30013	38	2	3	27,414	12	0.500
M. parvicella	^{M5}D38_N_contig_87203	19	2	3	15,136	17	0.394
M. parvicella	A01_C9_contig_103133	4	2	1	2,630	36	0.359
M. parvicella	A01_C9_contig_12327	4	2	NA	2,049	5	0.420
M. parvicella	A02_N_contig_144959	4	2	NA	2,478	12	0.541
M. parvicella	D01_C26_contig_16003	12	2	1	4,512	1	0.547
M. parvicella	D01_C26_contig_268387	9	3	NA	4,263	1	0.493
M. parvicella	D01_N_contig_207458	35	3	NA	19,684	1	NA
M. parvicella	D01_N_contig_231221	9	1	3	6,474	10	NA
M. parvicella	D01_N_contig_255653	4	2	1	1,345	1	NA
M. parvicella	D01_N_contig_88940	13	2	2	8,757	11	NA
M. parvicella	D02_N_contig_123069	4	2	2	2,760	7	NA
M. parvicella	D02_N_contig_136526	4	2	1	3,143	2	NA
M. parvicella	D02_N_contig_160919	4	2	1	2,331	3	NA
M. parvicella	D02_N_contig_219755	3	2	NA	2,082	7	NA
M. parvicella	D02_N_contig_298847	11	2	3	6,494	1	NA
M. parvicella	D02_N_contig_371485	4	2	1	1,982	1	NA
M. parvicella	D02_N_contig_406646	8	3	NA	7,201	1	0.020
M. parvicella	D02_N_contig_89284	10	3	NA	4,065	1	NA
M. parvicella	D03_G6_contig_20381	9	2	4	9,945	2	NA
M. parvicella	D04_N_contig_162234	8	2	1	5,735	4	NA
M. parvicella	D04_O11_contig_119585	4	2	1	4,265	1	NA
M. parvicella	D04_O11_contig_181244	9	1	1	4,601	1	NA
M. parvicella	D05_N_contig_14033	9	3	NA	8,931	1	NA
M. parvicella	D05_N_contig_215666	4	2	1	1,328	14	NA
M. parvicella	D06_N_contig_156375	6	1	3	5,539	1	NA
M. parvicella	D06_N_contig_16457	8	3	NA	2,511	1	NA
M. parvicella	D06_N_contig_242074	14	2	2	10,915	2	-0.074
M. parvicella	D06_N_contig_94627	6	2	1	2,770	3	NA
M. parvicella	D07_N_contig_115938	5	2	2	3,862	1	NA
M. parvicella	D08_N_contig_145494	4	2	1	2,224	7	NA
M. parvicella	D08_N_contig_236819	4	2	1	1,818	1	NA
M. parvicella	D08_N_contig_356358	18	2	2	16,169	2	0.192
M. parvicella	D09_N_contig_16765	8	3	NA	3,169	27	0.372
M. parvicella	D09_N_contig_51151	4	1	1	2,590	13	0.310
M. parvicella	D10_N_contig_365665	11	2	2	6,645	2	NA
M. parvicella	D11_N_contig_151098	11	3	NA	11,364	5	NA
M. parvicella	D11_N_contig_173798	33	2	2	31,193	3	0.541
M. parvicella	D11_N_contig_174238	6	2	NA	2,663	1	NA
M. parvicella	D11_N_contig_178860	5	2	NA	4,058	4	NA
M. parvicella	D11_N_contig_264492	19	2	1	12,008	21	NA

Continued on next page...

Bacterial host	Contig ID	*No. of predicted genes	*Category	*No. of phage hallmark genes	Contig length	Protospacer count	Pearsons correlation
M. parvicella	D11_N_contig_298528	31	2	4	22,759	1	NA
M. parvicella	D11_N_contig_363192	5	2	1	3,000	1	0.330
M. parvicella	D12_N_contig_178981	7	3	NA	1,819	7	0.311
M. parvicella	D12_N_contig_26657	12	2	3	7,475	1	NA
M. parvicella	D12_N_contig_71630	9	2	1	5,877	1	NA
M. parvicella	D13_C22_contig_263218	16	3	NA	6,257	9	NA
M. parvicella	D13_C22_contig_366062	14	2	4	8,356	1	NA
M. parvicella	D13_N_contig_270482	5	2	NA	2,114	6	NA
M. parvicella	D14_N_contig_241938	5	1	1	5,303	1	NA
M. parvicella	D14_N_contig_311745	3	2	NA	2,234	2	0.430
M. parvicella	D14_N_contig_336166	4	3	NA	1,680	5	NA
M. parvicella	D15_N_contig_260057	4	2	1	1,234	16	0.220
M. parvicella	D16_N_contig_166285	27	2	4	19,308	1	NA
M. parvicella	D16_N_contig_223292	12	2	2	6,306	4	NA
M. parvicella	D16_N_contig_27648	5	2	1	2,656	2	NA
M. parvicella	D16_N_contig_314416	8	3	NA	8,410	5	NA
M. parvicella	D16_N_contig_65141	5	2	NA	14,867	1	NA
M. parvicella	D16_O7_contig_16386	10	1	3	6,577	1	NA
M. parvicella	D16_O7_contig_67733	14	3	NA	5,303	1	0.652
M. parvicella	D17_E15_contig_295127	12	3	NA	13,299	8	NA
M. parvicella	D19_N_contig_34607	8	2	1	4,605	16	NA
M. parvicella	D21_N_contig_108796	6	2	1	5,664	1	NA
M. parvicella	D21_N_contig_131834	8	3	NA	6,696	10	NA
M. parvicella	D21_N_contig_147221	12	1	4	7,420	1	NA
M. parvicella	D22_N_contig_15454	17	2	3	10,082	4	0.443
M. parvicella	D22_N_contig_185897	8	1	3	6,389	1	NA
M. parvicella	D22_N_contig_292395	8	1	3	4,489	1	0.450
M. parvicella	D22_N_contig_331165	13	2	2	8,138	11	NA
M. parvicella	D22_N_contig_399720	5	2	NA	5,241	1	NA
M. parvicella	D22_N_contig_41970	6	1	1	3,979	6	NA
M. parvicella	D22_N_contig_54652	5	2	2	2,605	3	NA
M. parvicella	D23_N_contig_218246	10	3	NA	5,253	1	NA
M. parvicella	D23_N_contig_68934	9	3	NA	5,836	1	NA
M. parvicella	D24_N_contig_423179	8	3	NA	5,389	10	NA
M. parvicella	D24_N_contig_82309	15	2	2	10,130	10	NA
M. parvicella	D25_N_contig_107837	8	3	NA	2,064	1	NA
M. parvicella	D26_C8_contig_158686	10	3	NA	5,397	1	NA
M. parvicella	D26_C8_contig_346103	12	3	NA	6,775	5	NA
M. parvicella	D27_E21_contig_194415	9	2	2	6,115	6	NA
M. parvicella	D27_N_contig_111812	6	2	1	2,748	1	NA
M. parvicella	D28_N_contig_337078	19	3	NA	18,840	1	0.452
M. parvicella	D28_N_contig_411613	4	3	NA	1,227	6	NA
M. parvicella	D29_N_contig_186885	41	2	1	28,969	1	NA
M. parvicella	D31_N_contig_29920	3	1	1	1,505	5	0.324
M. parvicella	D31_N_contig_363355	14	3	NA	6,873	1	0.138
M. parvicella	D32_N_contig_373777	12	3	NA	12,847	1	NA

Continued on next page...

Bacterial host	Contig ID	*No. of predicted genes	*Category	*No. of phage hallmark genes	Contig length	Protospacer count	Pearsons correlation
M. parvicella	D32_N_contig_407296	5	1	2	4,086	6	NA
M. parvicella	D33_N_contig_146635	3	1	2	3,760	1	NA
M. parvicella	D33_N_contig_324078	4	2	1	1,604	1	NA
M. parvicella	D33_N_contig_357926	6	1	1	3,420	1	NA
M. parvicella	D34_N_contig_179663	8	3	NA	5,300	10	NA
M. parvicella	D35_E21_contig_25150	15	2	2	10,923	11	NA
M. parvicella	D35_G11_contig_240772	22	3	NA	8,403	10	NA
M. parvicella	D35_N_contig_13435	4	2	1	2,268	23	0.332
M. parvicella	D36_N_contig_120762	16	3	NA	13,780	306	NA
M. parvicella	D36_N_contig_211305	5	2	1	2,206	3	NA
M. parvicella	D36_N_contig_213511	6	1	2	4,601	9	NA
M. parvicella	D36_N_contig_247937	12	3	NA	7,231	2	NA
M. parvicella	D36_N_contig_368349	19	2	1	14,085	252	NA
M. parvicella	D36_N_contig_385758	7	2	1	3,589	1	NA
M. parvicella	D36_N_contig_91600	8	2	1	5,133	5	NA
M. parvicella	D37_G11_contig_349205	4	2	1	2,153	2	NA
M. parvicella	D37_N_contig_304426	4	2	1	2,280	1	NA
M. parvicella	D37_N_contig_39500	6	2	1	1,980	9	NA
M. parvicella	D37_N_contig_55719	5	2	NA	2,378	6	NA
M. parvicella	D37_N_contig_73727	8	2	1	3,218	1	NA
M. parvicella	D38_N_contig_278556	8	2	NA	3,608	8	NA
M. parvicella	D38_N_contig_45838	5	2	NA	4,203	9	0.399
M. parvicella	D39_N_contig_177400	9	2	NA	10,172	3	0.193
M. parvicella	D39_N_contig_19448	5	1	3	4,261	1	NA
M. parvicella	D39_N_contig_307595	5	2	1	2,738	4	NA
M. parvicella	D39_N_contig_432421	8	2	2	4,199	1	NA
M. parvicella	D39_N_contig_46287	12	3	NA	7,132	4	NA
M. parvicella	D39_N_contig_70910	8	2	1	4,600	1	NA
M. parvicella	D39_N_contig_81757	4	1	2	3,608	1	NA
M. parvicella	D39_N_contig_91257	17	2	3	10,873	9	0.045
M. parvicella	D40_G4_contig_115508	15	3	NA	8,091	4	NA
M. parvicella	D40_G4_contig_118584	12	3	NA	4,454	2	NA
M. parvicella	D40_G4_contig_177971	7	2	1	6,453	1	NA
M. parvicella	D40_N_contig_293294	6	2	1	3,161	1	NA
M. parvicella	D40_N_contig_307536	4	1	1	2,050	4	NA
M. parvicella	D41_N_contig_114234	14	3	NA	6,945	1	NA
M. parvicella	D41_N_contig_212174	5	2	1	2,631	3	NA
M. parvicella	D41_N_contig_256737	12	2	4	8,294	1	NA
M. parvicella	D41_N_contig_72662	5	3	NA	1,097	3	NA
M. parvicella	D41_N_contig_76453	13	3	NA	8,874	4	0.402
M. parvicella	D42_N_contig_302940	7	2	1	4,407	10	NA
M. parvicella	D43_N_contig_323359	17	2	4	11,206	1	NA
M. parvicella	D45_N_contig_74156	16	3	NA	9,522	1	NA
M. parvicella	D46_N_contig_313024	6	2	NA	5,789	1	0.032
M. parvicella	D47_P30_contig_66499	27	3	NA	32,430	4	NA
M. parvicella	D48_N_contig_194520	6	2	1	4,342	9	NA

Continued on next page...

Bacterial host	Contig ID	*No. of predicted genes	*Category	*No. of phage hallmark genes	Contig length	Protospacer count	Pearsons correlation
M. parvicella	D48_N_contig_2013	6	2	NA	4,166	2	-0.221
M. parvicella	D48_N_contig_27942	9	1	3	5,720	1	NA
M. parvicella	D48_N_contig_342272	13	2	3	7,567	1	NA
M. parvicella	D48_N_contig_39235	6	2	1	2,878	4	NA
M. parvicella	D48_N_contig_56215	4	2	1	2,225	1	NA
M. parvicella	D49_C8_contig_40357	9	3	NA	4,662	2	NA
M. parvicella	D49_N_contig_369172	12	2	1	6,605	1	NA
M. parvicella	D50_E20_contig_298577	20	3	NA	10,689	1	NA
M. parvicella	D50_N_contig_205844	8	3	NA	2,319	5	NA
M. parvicella	D50_N_contig_210277	3	2	NA	1,507	6	NA
M. parvicella	D50_N_contig_222790	11	2	3	7,143	1	NA
M. parvicella	D50_N_contig_25642	9	2	1	5,523	4	NA
M. parvicella	D50_N_contig_51583	4	2	1	1,769	6	NA
M. parvicella	D51_N_contig_274619	9	2	1	5,645	1	NA
LCSB005	[L1]D31_E31_contig_80738	29	3	NA	14,967	73	0.348
LCSB005	[L2]D48_N_contig_290930	4	2	NA	2,876	1	-0.145
LCSB005	D01_N_contig_128142	17	3	NA	11,769	1	-0.115
LCSB005	D04_N_contig_97006	9	3	NA	3,533	1	-0.098
LCSB005	D07_N_contig_4877	25	3	NA	19,269	1	-0.109
LCSB005	D11_N_contig_66232	24	3	NA	18,698	1	-0.110
LCSB005	D17_N_contig_231006	9	3	NA	4,800	1	NA
LCSB005	D43_N_contig_361977	23	3	NA	12,830	3	-0.145

[M1] Putative phage M1

[M2] Putative phage M2

[M3] Putative phage M3

[M4] Putative phage M4

[M5] Putative phage M5

[L1] Putative phage L1

[L2] Putative phage L2

3.4.5 Bacteriophage and host dynamics

The hosts and putative phages were quantified by mapping the IMP-preprocessed MG reads, followed by calling of the average contig-level depth of coverage, which was subsequently used as a proxy for inferring the abundances. The abundance of the LCSB005 host and its putative phages (i.e. putative phage L1 and L2), are represented in **Figure 3.8**. There were no observable patterns of co-abundances between the putative phages and the host. However, a large spike in abundance of putative phage L1 is observed in the 29 November 2011 data point (> 150 average contig-level depth of coverage) that far surpasses the abundance of the host at that particular time point (i.e. approximately 6 average contig-level depth of coverage). Additionally, there was a general low abundance of this particular putative phage within samples from the rest of the time-series. On the other hand, putative phage L2 demonstrates peak occurrences in certain points in the time-series, decoupling its abundance trends to the one of its putative host.

The similar analysis applied to all selected putative phages of *M. parvicella* population (Putative phages M1-5) demonstrated similar trends of sharp increases in bacteriophage abundances (as with putative phages of LCSB005) (**Figure 3.9**). However, unlike the case of LCSB005, these putative phages never exceed the abundance of their host, while putative phages M1 and M2 seem to occur throughout the entire time-series, without diminishing. In addition, putative phage M1 demonstrates higher abundance throughout the entire time-series compared to other putative phages associated with the *M. parvicella* population. Another clearly observable trend is the drop in phage abundance occurring in tandem with the drop in host abundance. There are also observed increases of putative phages M2 and M3 following increasing of the *M. parvicella* host population. Finally, the putative phages do not seem to peak at the same time points, i.e. different phages peak at different time points, although the absolute abundance of putative phage M1 is almost always higher than other putative phages (M2-5).

The *de novo* assembly of MT data, carried out by IMP provides the possibility of assembling MT-based sequences. In line with the recent discovery of bacterial and RNA phage interaction via the CRISPR-*Cas* system [Abudayyeh *et al.*, 2016], we went on to inspect the possibility of CRISPR-*Cas* interaction with RNA-based invasive genetic elements using the MT-based protospacers-containing contigs. More specifically, these contigs are well-represented on the MT level (i.e. average contig-level depth of coverage), but are not represented on the MG level. In addition, only contigs that are fully covered by MT reads (i.e. from end to end, and not just intragenic regions) were retained for further analyses. Such contigs were only identified within the *M. parvicella* associated protospacers-containing contigs and were not detected for LCSB005 associated protospacers-contigs. These MT-based protospacer-containing contigs revealed a relatively higher contig-level depth of coverage compared to the putative DNA-based phages (**Figure 3.10**A), demonstrating that such MT-based protospacers-containing contigs, despite being lower in richness, occur in higher abundance. Furthermore, using the MT data, the expression of *cas* genes within the *M. parvicella* genome could be inspected. Interestingly, the *cas* gene annotated as an endoribonuclease *Cas2* demonstrated the highest expression (based on unnormalized and normalized depths of coverage **Figure 3.10**B), throughout the time-series. As the name indicates, this particular enzyme is believed to cleave RNA molecules and was described to form a complex with the *Cas1* protein to mediate spacer acquisition [Nuñez *et al.*, 2014].

Since none of these MT-based protospacers-containing contigs were predicted as phages by VIRSorter (**Table 3.4**), they are hereafter defined as a putative RNA-based invasive genetic element(s), abbreviated

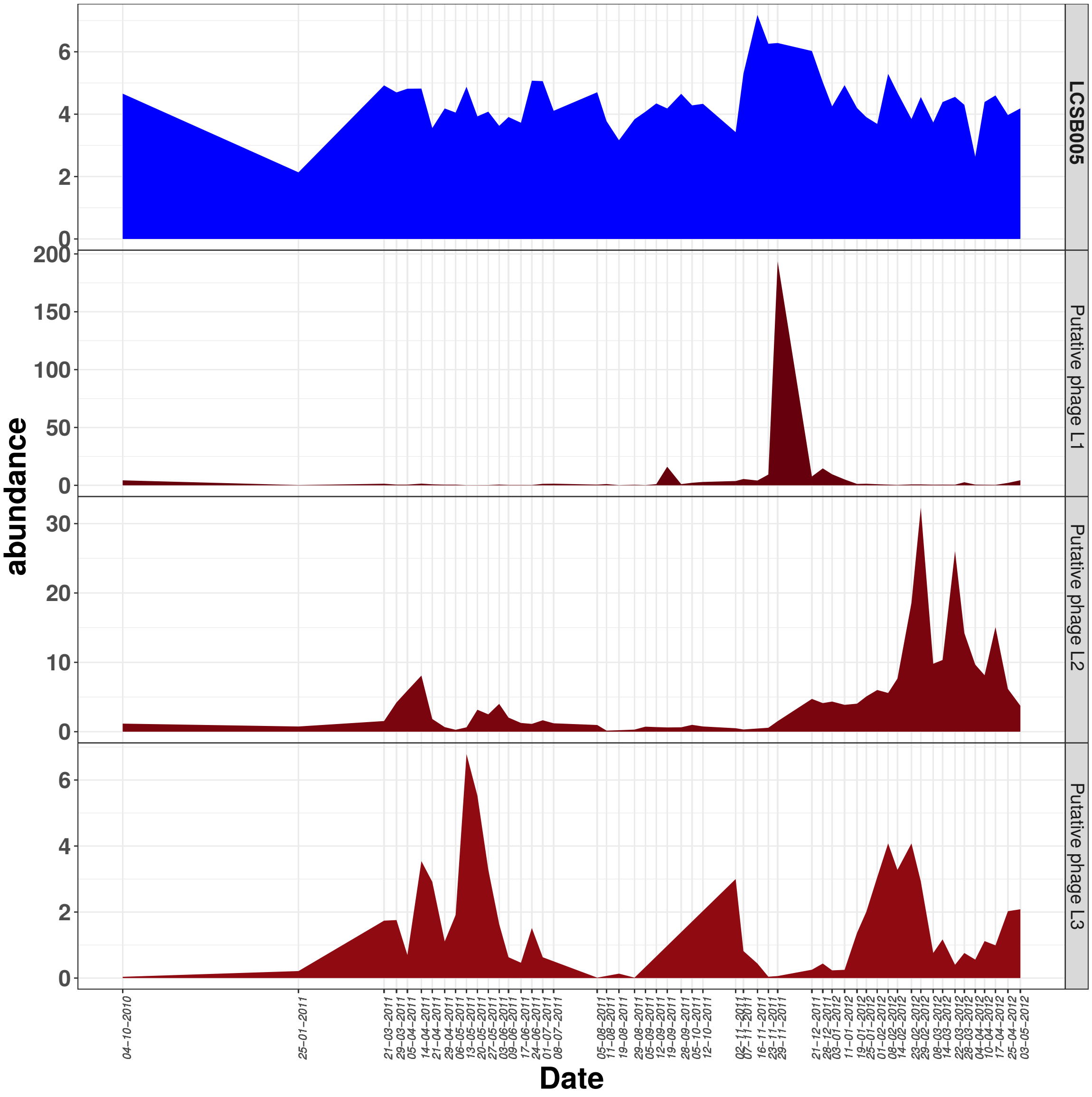


Figure 3.8: Dynamics of LCSB005 host and associated bacteriophages. The labels in the x-axis represent the exact sampling dates.

to RIGe. Accordingly, six of the highly abundant (throughout the entire time-series) putative RIGes were observed in relation to the abundance of the *M. parvicella* host population. RIGe M8 demonstrated relatively high peaks in time points 4 October 2010 (depth of coverage ~500) and 16 November 2011 (depth of coverage ~600), while being relatively low in abundance at other time points (**Figure 3.10**). Moreover, in the latter time point, the expression of the aforementioned endoribonuclease *Cas2* seemed to peak in correspondence to the high abundance of RIGe M8 (**Figure 3.10**). This event was then followed by a sharp drop in the usually dominant (i.e. quantitatively abundant) *M. parvicella* population. Another interesting observation about this RIGe M8 is that it was present in high abundance (depth of coverage ~300) during the low abundance phase of the *M. parvicella* host (between 7 November 2011 and 21 December 2011), unlike other observed phages

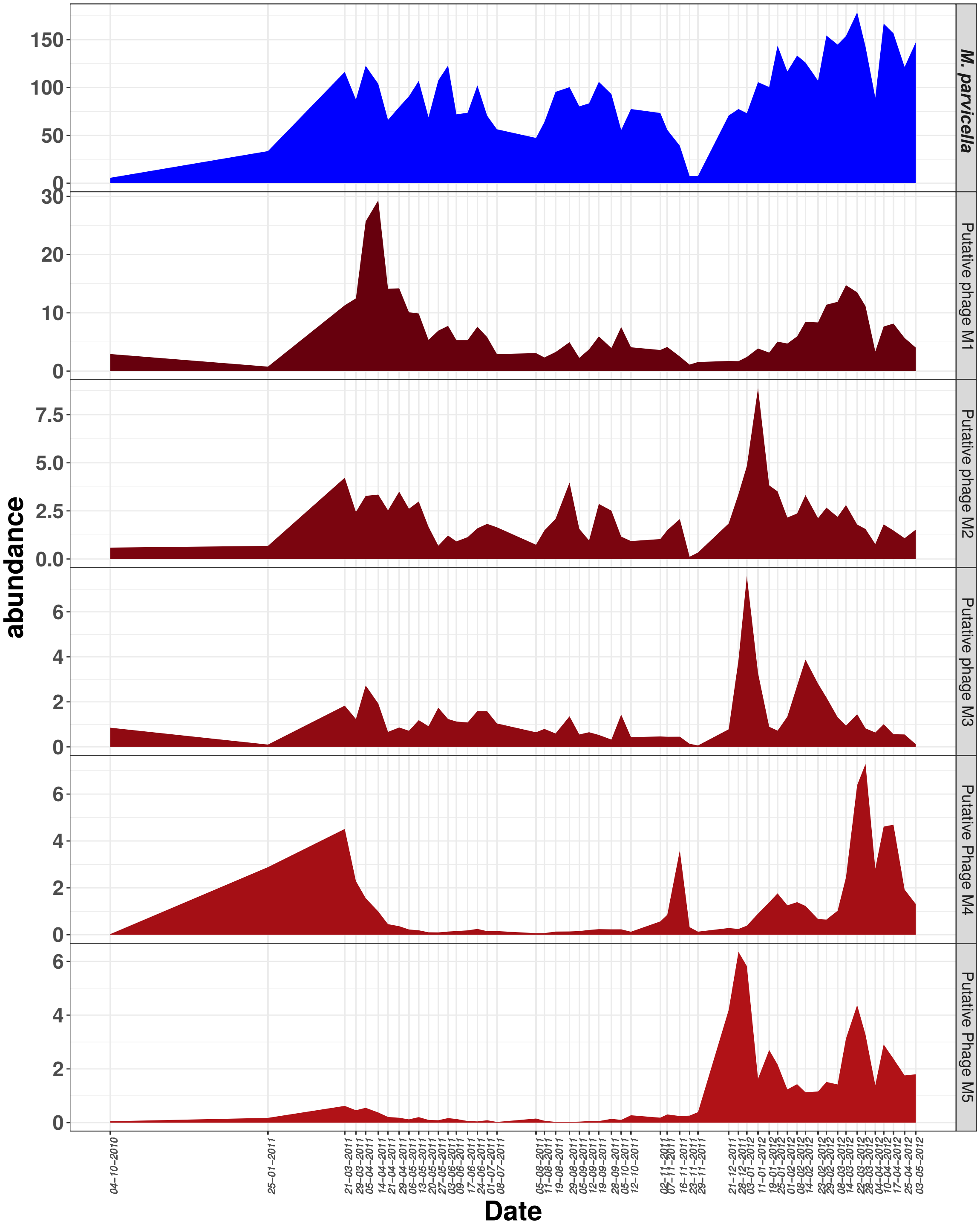
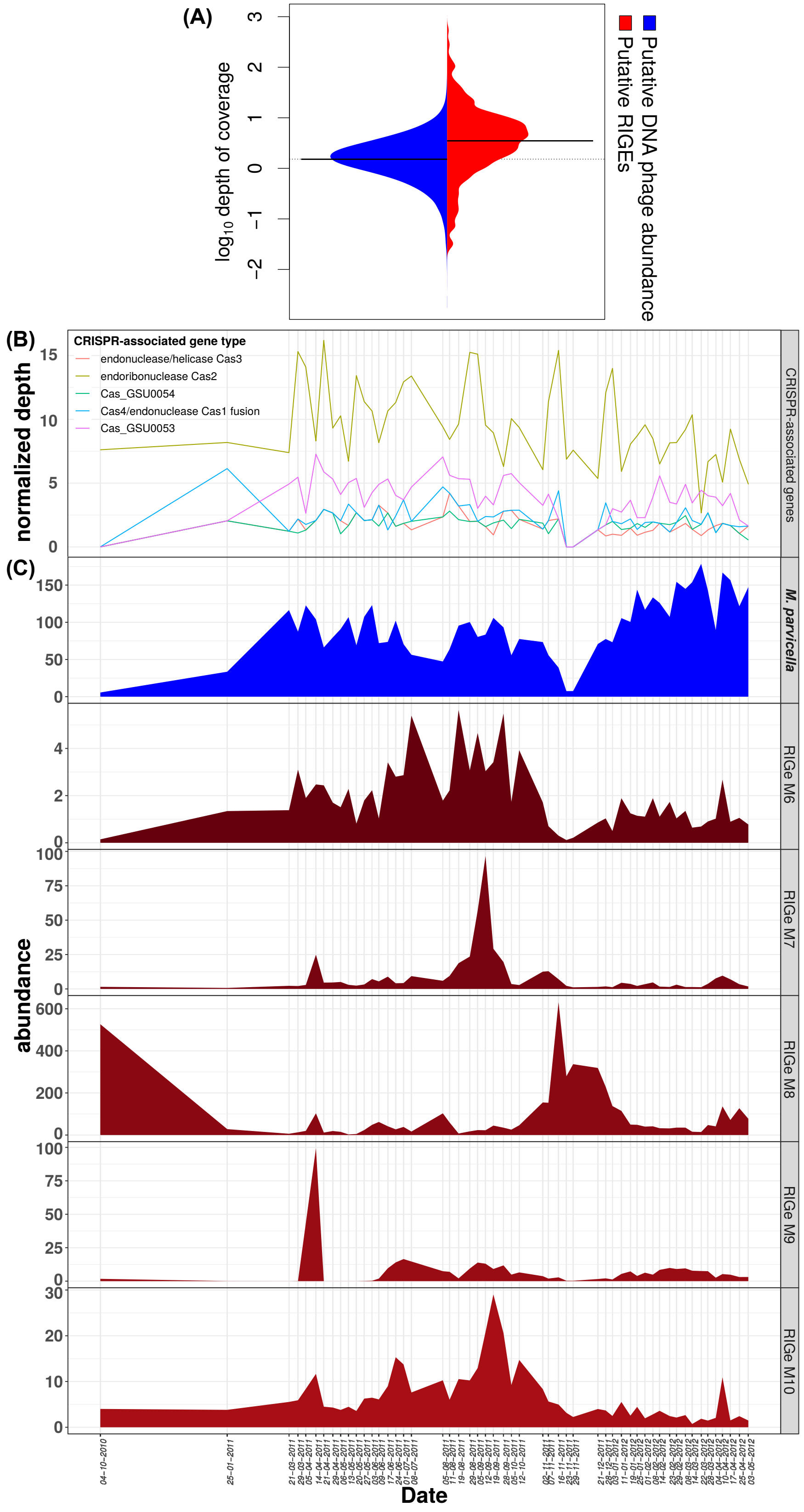


Figure 3.9: Dynamics of *M. parvicella* host and associated bacteriophages. The labels in the x-axis represent the exact sampling dates.

and RIGes (**Figures 3.9** and **3.10**) which tend to drop in abundance with the *M. parvicella* host population. This may suggest that these RIGes may not be selectively “infecting” the *M. parvicella* population, but also other host populations present within the community.



(See legend on next page)

Figure 3.10: Dynamics of *M. parvicella* host and associated RIGes. (A) Beanplot representing the densities of metagenomic (MG) and metatranscriptomic (MT) of contig-level depth of coverage of protospacer-containing contigs associated to *M. parvicella*. Solid lines represent the mean depth of coverage. Dotted line represents mean depth of coverage across both MG and MT based abundances. **(B)** Expression levels of different *cas* genes found within *M. parvicella* genome. **(C)** Abundance of *M. parvicella* bacterial host and associated putative RNA invasive genetic element (RIGe). The labels in the x-axis for figures **(B)** and (C) represent the exact sampling dates.

3.5 Discussion

The present study highlights the importance of using multiple methods to extract CRISPR information. CRISPR *loci* represent highly repetitive regions within archaeal and bacterial genomes. Modern *de novo* assemblers are unable to resolve repetitive regions, especially given the short read lengths produced by NGS technologies. Therefore, *de novo* assemblies may not be able to effectively resolve CRISPR regions, leading to a potential loss of information. MetaCRT detects CRISPRs from longer sequences, i.e. assembled contigs. It demonstrated effectiveness in the recovery of CRISPR repeats, this is possible due to the fact that repeat sequences are more abundant within the data, due to the usage of MG and MT data which may increase the coverage or even improve the assembly of CRISPR regions. However, metaCRT appears to exhibit less sensitivity in terms of CRISPR spacer detection. This could be due to the inherent heterogeneity of CRISPR spacers, such that a consensus-based method (i.e. *de novo* assembly) would result in the loss of spacer information, especially regarding less abundant and/or rare spacers. To that end, this work demonstrates that CRASS, a read-level CRISPR sequence mining program, is able to detect a larger number spacers compared to metaCRT. This read-level resolution of CRISPR spacers is important to extract the maximum amount of information from the data in order to detect putative phage contigs. In addition, CRASS also provides information on the CRISPR flanking regions (leader and downstream sequence), which enables the linking of relevant populations. Both length values reported for CRISPR repeats and spacers (**Figure 3.5**) were longer than the lengths reported in previous studies [Jansen *et al.*, 2002; Haft *et al.*, 2005; Amitai and Sorek, 2016]. This may be due to spurious detection by the software as these long spacers did not exhibit a high level of identity (based on blastn) to their protospacers (data not shown).

The present study combines longitudinal data and dual-omic datasets (MG and MT) to result in multiple advantages: i) the temporal nature of the study allows the resolution of CRISPR dynamics which are shown to vary significantly across time, ii) the MT data confirms that these CRISPR *loci* are transcribed over time and iii) the MT data allows for the detection of RIGes. Specifically, there is an apparent difference in the expression of CRISPR *loci* of different bacterial populations, namely the CRISPRs from the *M. parvicella* population appear to be constitutively expressed throughout most of the samples, while CRISPRs from the LCSB005 population appear to be intermittently expressed. However, this may be due to the overall abundance of *M. parvicella* within the community such that the sequencing is able to measure these characteristics, while the lower abundance of LCSB005 may result in the apparent expression of the CRISPR *loci* in a single time point. Overall, this supports the notion that CRISPR *loci* are highly dynamic and heterogeneous genomic regions [Deveau *et al.*, 2008; Amitai and Sorek, 2016; Silas *et al.*, 2016].

In addition, changes in the community structure (i.e. changes in constituent populations) can affect the CRISPR content within the overall community. This is observed with regards to the reduction in the population size for *M. parvicella* in time points 23 and 29 November 2011 (**Figure 3.9**) resulted in the reduction of CRISPR elements, protospacers and protospacers-containing contigs associated to *M. parvicella* (**Figure 3.6**). Furthermore, this apparent collapse in the *M. parvicella* population coincides with an overall increased population abundance and CRISPR element abundances associated with LCSB005 (**Figure 3.7**). In addition, the time-resolved data highlights the occurrence frequency of different CRISPR elements. For instance, there was approximately half of the non-redundant CRISPR repeats appearing only once in the entire time-series, similar to LCSB005. This may be attributed to organisms that occur in low abundance and in one/very few samples of the time-series. CRISPR repeats that occur in all/almost all the time points will probably belong to microbial population that exhibit quantitative dominance throughout the entire time-series, such as *M. parvicella* (**Figure 3.6**). On the other hand, CRISPR spacers that occur in almost all time points (which is rare in this dataset, **Figure 3.5**) are likely to represent conserved spacers, possibly due to the frequent occurrence of a particular invasive genetic element, thus pressuring the host to select/retain these particular spacers. Overall, all the CRISPR spacers are highly unique and specific compared to CRISPR repeats, which are known to be conserved within specific prokaryotic clades, as demonstrated here by assigning CRISPR repeats to specific species (**Table 3.3**) [Jansen *et al.*, 2002].

CRISPR repeats are well represented in both MG and MT datasets, thereby demonstrating activity of the CRISPR anti-viral defence mechanism through time. The availability of MT data reaffirms the notion that CRISPR *loci* are indeed expressed/transcribed genomic regions. However, this pattern of MT coverage is not observed in CRISPR spacers, despite being supposedly transcribed along with the repeat sequences. CRISPR spacers are typically considered highly heterogeneous, such that a single bacterial population (strain within a given species) may carry completely different spacers (intra-species variability) within their CRISPR regions. On top of that, these spacers are also highly dynamic, with the constant insertion of new spacers and/or removal of old spacers. The transcribed crRNA sequences are further processed to generate shorter post-processed CRISPR spacer-repeat pairs which work together with *Cas* enzymes to silence/inhibit invasive genetic elements [Barrangou *et al.*, 2007; Amitai and Sorek, 2016]. Therefore, from a technical perspective, the majority of NGS reads used within this work are generally longer than the processed CRISPR spacer-repeat sequences. Furthermore, most of the data stems from paired-end reads, thus limiting the overall detection of shorter sequences. The issue of CRISPR spacer detection within MT data could be circumvented by lowering the length threshold for the retained MT preprocessed NGS reads such that shorter reads could be used for downstream analyses. Following this, the threshold CRASS could also be adjusted to allow extraction of shorter CRISPR sequences. More specifically, CRASS provides an option to extract so-called “singleton spacers” [Skennerton *et al.*, 2013]. Singleton spacers would represent a single CRISPR spacer-repeat pairing, that should be more reflective of the nature of post-transcriptionally processed CRISPRs [Marraffini and Sontheimer, 2010; Amitai and Sorek, 2016]. In summary, it is likely that the apparent “dilution” of spacer representation on the MT level, may be traced back to this overall biological heterogeneity of the CRISPR spacers, in addition to the technological limitations of measuring the short final products of the post-processed CRISPR-RNA sequences, which in theory could be bypassed by using specific analysis parameters.

The presented data also shows that CRISPR spacers are highly sensitive to protospacers, suggesting that a single spacer may have matches to multiple targets. Although, this is arguably dependent on the parameters used for the blast search [Biswas *et al.*, 2013], and can be attributed to spurious matches. However, it was previously described that CRISPR-*Cas* systems allows non-identical spacer-protospacer matches (or priming [Fineran and Charpentier, 2012; Amitai and Sorek, 2016]). It was further suggested that this mechanism could be effective with up to 13 mismatches, which cannot be replicated using the parameters of current computational approaches [Edwards *et al.*, 2015]. Consequently, it is likely that there are much larger number of protospacer-containing contigs (i.e. putative invasive genetic elements) found within the data which could not be resolved due to the stringent criteria of the current analyses.

The community level analyses is rich in information and is a potential reservoir of putative phage sequences. However, it does not allow the observation of phage and host dynamics, due to the absence of associations, i.e. there are no defined hosts. To that end, this work leveraged draft genome sequences from isolated lipid accumulating bacterial strains. The provision of these genomes enabled the linking of CRISPR repeats to the corresponding genomes. From there, one could match the complementary spacer and repeat information in order to formulate associations between these defined bacterial populations and their putative phages. It is estimated that half bacterial species encode the CRISPR-*Cas* system within their genomes [Amitai and Sorek, 2016]. Although this work was able to identify bacterial species with corresponding CRISPRs, the overall number was still limited. It is possible that the genomes within the isolate genome compendium do not containing any CRISPR regions, or that they occur low abundances, such that there were insufficient for necessary signal, with the latter case to be more likely. However, having reference isolate genomes enabled the analysis to shift from a community-level perspective, to a population-level study. Such associations could be further improved using population-level genomic reconstructions through initial binning of the metagenomic data. However, genomic bins would need to be of high-quality (long contigs) to formulate confident associations. More specifically, CRISPR-containing contigs would need to be correctly assigned to such bins. As previously discussed, CRISPR regions are usually not well assembled within *de novo* MG assemblies [Skennerton *et al.*, 2013], thus resulting in short contigs or contigs that do not contain sufficient flanking regions. With regards to nucleotide signature-based binning methods, short contigs result in a reduced signal (signature) for effective bin assignment by such methods and/or while contigs with limited flanking regions could result in a misrepresented nucleotide signature and thereby erroneous binning of those CRISPR-containing contigs. Collectively, these would cause CRISPR-containing contigs to be either unassigned or incorrectly assigned to these genomic bins. There are several ways this issue could be bypassed including: i) close inspection and curation of generated genomic bins, especially with regards to the assignment of CRISPR sequences, ii) relying on an abundance based binning method instead of nucleotide signature based method and iii) leveraging on the flanking region information provided by CRASS for conclusive linking to high-quality contigs that do not contain CRISPRs (**Sections 3.4.2** and **3.4.3**).

Despite the vast number of protospacer-containing contigs associated with the defined bacterial hosts, a majority of those sequences could not be conclusively defined as bacteriophage populations by state-of-the-art tools. This could, first and foremost highlight the under-representation of known bacteriophage (or virus) genes. A second reason might be due to the fact that these sequences originate from other invasive genetic elements such as plasmids and transposons (or generally mobile genetic elements). This implies that it should be possible to scan these sequences for the presence of plasmids or transposons, to further expand the annotation of these sequences. However, the present study is focused on deciphering the dynamic relationships between phages and bacterial hosts.

This work described several different notable cases with regards to phage and host dynamics. The most obvious is the two types of populations that were well resolvable, one being highly abundant and dominant (*M. parvicella*), while the other being relatively low in abundance (LCSB005). The dominance of *M. parvicella* population within this community leads to a high amount of CRISPR information (i.e. repeats, spacers and flanking sequences), associated with this specific population. It is important to note that the flanking sequences detected by CRASS are based on read-level resolution thereby resulting in the apparent high number of such flanking sequences associated to the *M. parvicella* population (**Table 3.3**). Despite this, significant spikes in phages related to the lowly abundant bacterial strain, LCSB005 were observed (**Figure 3.8**). This is possibly due to the increased abundance (compared to the other time points) of the LCSB005 population from time point 16 November 2011 to 21 December 2011, suggesting a possible kill the winner scenario. However, this was not accompanied with confident prediction of phage sequences and/or a high number of protospacers. In general, the study of lowly abundant bacterial populations remains challenging, especially from a large-scale bioinformatic approach. The problem may be compounded if the sequencing is bias towards higher abundant populations such as the *M. parvicella*. Overall, this work showed the complementary use of *in situ* derived data and classical microbiology derived isolate data could potentially result in higher quality output and information.

The occurrence of the *M. parvicella* as a quantitatively dominant and stable population within this community served as an advantage in this work. In particular, the abundance of CRISPR information provided a large reservoir of protospacer-containing contigs, and thereby the identification of putative bacteriophage contigs. More importantly, a high number of protospacers, or rather CRISPR spacer targets, may represent the high rate of interactions occurring between the phage and host populations. Compared to the LCSB005 population, for which a total of eight predicted bacteriophages have been identified, there was a large number (150 in total) of predicted phages associated to *M. parvicella*. This supports the idea of a "kill the winner" scenario, such that the highly abundant bacterial populations within a community are targeted by bacteriophages. Specifically, bacteriophages have a higher chance of infecting and replicating within highly abundant host population that are constantly present as opposed to lowly abundant host populations. It has been suggested that phages and their host populations tend to correlate with each other, in terms of abundance [Edwards *et al.*, 2015]. However, there were no particularly high correlations observed in this study (**Table 3.4**), suggesting that the application of lagged correlations might be more effective in deciphering phage-host associations with higher confidence [Edwards *et al.*, 2015]. While correlation-based methods are highly suitable for time-series datasets, the current work may not provide the necessary resolution for such a study. The present datasets were taken on an almost weekly basis, while lytic phage infection (from adsorption to lysis) was suggested to occur within minutes to hours [Shao and Wang, 2008].

This work also identified the RIGes within the data, which are MT-based contigs that contained protospacers, suggesting possible interactions between these components and the *M. parvicella* population. Furthermore, these observations show the: i) high abundance of RIGes, which were reconstructed exclusively from MT data, ii) the constitutively high expression of an endoribonuclease *Cas2* gene within the host populations and iii) the observed high abundance phase of the RIGe M8 coinciding high expression of endoribonuclease *Cas2* within the host (**Figure 3.10**). However, these RIGes would require further inspection due to their lack of gene annotation (i.e. mostly annotated with hypothetical proteins). In that light, they may represent highly expressed transcripts that were not detectable on the MG (DNA) level and may stem from non-phage invasive elements such as plasmids. Furthermore, the endoribonuclease *Cas2* enzyme, although structurally well classified [Nuñez *et al.*, 2014], still lacks conclusive evidence in terms of influence on protospacer acquisition [Nuñez *et al.*, 2014, 2015; Wang *et al.*, 2015]. On the contrary, there are recent observations of the CRISPR-*Cas* system interacting with RNA sequences. These includes the interaction of bacteria with RNA phages [Abudayyeh *et al.*, 2016] and modification of the *Cas9* enzyme to target RNA sequences [Price *et al.*, 2015]. Indeed, the aforementioned studies relied on laboratory-based co-cultures of phages and bacteria, as deciphering such interactions within a natural systems remain challenging. This is mainly due to the sparse knowledge and resource with regards to RNA phages, thus limiting the annotation and detection of such sequences. Despite the present challenges, this work does raise interesting questions with regards to the possible interaction of the CRISPR-*Cas* system with RIGes, for which validation through laboratory based experiments would be absolutely essential in confirming the interaction of RIGes with bacterial hosts.

As an outlook, time-series based information and associations garnered from this study would be suitable for the generation of mathematical models. These models could potentially translate this information obtain more knowledge about the influence and roles of bacteriophages on microbial communities. The knowledge generated could hold the key for the manipulation of LAMPs through the addition of bacteriophages, for the maximization of lipid accumulation and thus optimal biofuel production.

3.6 Conclusion

This study demonstrated the application of the developed large-scale integrated omic analysis pipeline (IMP, **Chapter 2**) for the study of biological components of interest, which in this case represents bacteriophages and their associated hosts. This study leveraged multiple tools in order to maximize the amount of information with regards to the CRISPR elements (i.e. repeats, spacers and flanking regions). The community-level analyses of the CRISPR elements revealed that CRISPR regions are: i) highly dynamic with fluctuations over time ii) transcribed, as demonstrated by the use of MT data, and iii) highly heterogeneous, with the presence of a large number of CRISPR spacers. The work focused on two bacterial populations, namely the highly abundant *M. parvicella* population, and the low abundant LCSB005 population. Accordingly, a total of 150 putative phages were predicted for the *M. parvicella* population while a total of eight putative phages were predicted for LCSB005. We then selected five phages associated with *M. parvicella* and two phages associated with LCSB005 to observe phage-host dynamics. It also demonstrated that certain phages are more dominant compared to others, occurring in higher abundances relative to other phages. We also observed that phages seem to demonstrate peak occurrences of large abundances within certain time points. In most cases, the reduction of host abundance is accompanied by the reduction of associated phage abundances, with several exceptions. It is also observed that putative phages associated to the *M. parvicella* population do not demonstrate overlapping peaks of abundances. Further inspection of MT-based protospacers-containing contigs associated to the *M. parvicella* population suggests the possible interactions with RNA-based elements (RIGes). However, these cases should be thoroughly inspected to obtain more conclusive evidence.

In summary, the combination of an unprecedented multi-omic time-series dataset, a solid foundation of large-scale integrated omic analysis that resulted in high-quality assemblies in complementary with high-quality draft isolate genomes enabled the study of phage-host interactions unlike previous efforts, revealing different dynamical patterns of phages in relation to their hosts, including the detection of putative RIGes, which are rarely observed within a natural system. The information and associations garnered from this study should be validated via laboratory methods and translated into mathematical models to further elucidate the roles and influence of bacteriophages in microbial communities. Such knowledge would bring us a step closer towards manipulation of LAMPs using bacteriophages [Withey *et al.*, 2005; Jassim *et al.*, 2016] for the optimal production of biofuels [Sheik *et al.*, 2014; Muller *et al.*, 2014a]. More specifically, phages may be used to target and reduce the abundance of bacterial species that may compete with LAMPs, such as *M. parvicella*. Strategies such as these may allow control of LAMPs to make BWWT plants as a consistent and abundant source for biodiesel production.

CHAPTER 4

GENERAL CONCLUSIONS AND OUTLOOK

Parts of this chapter was adapted and modified from the following first-author peer-review publication:

Shaman Narayanasamy, Emilie E.L. Muller, Abdul R. Sheik, Paul Wilmes (2015). Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. *Microbial Biotechnology* **8**: 363-368. [**Appendix A.1**]

4.1 Integrated omics: From data to associations

Culture-independent methodologies have overcome the limitations of classical microbiological methods (**Section 1.1**). Specifically, the sampling of microbial consortia *in situ* combined with state-of-the-art wet-lab biomolecular extraction methodologies (**Section 1.4.2**) and systematic high-throughput measurements (**Section 1.4.3**) enables access to information not obtainable using culture-based methods. In addition, the falling cost of NGS sequencing has enabled deep characterization of the metagenome and the metatranscriptome (**Sections 1.4.3** and **1.4.3**). However, data-driven reference-independent methodologies (i.e. those not reliant on reference genomes) are very important to realize the full potential of the high-throughput MG and MT data sets as such methods enable the generation of hypotheses towards the discovery of novel microorganisms as well as functionalities (i.e. genes or combination of genes; **Section 1.4.4**). Given the relative ease of generating NGS-based datasets, such as MG and MT data, multi-omic studies of microbial communities are becoming more and more prevalent (**Section 1.4.5**). However, until the present work there was a clear lack in standardized workflows for the integrated analysis of these data types, resulting in the development of multiple *ad hoc* analysis methodologies, which made reproducing the work of others a major challenge (**Section 1.4.5** and **Chapter 2**).

To further illustrate this, time-resolved samples from the chosen model microbial community (LAMPs) (**Section 1.2**) were subject to two reference-independent integrated multi-omic studies that spanned multiple omic data sets [Muller *et al.*, 2014a; Roume *et al.*, 2015]. The study by Roume *et al.* [2015] applied available methods, more specifically the MOCAT pipeline [Kultima *et al.*, 2012]. Although MOCAT was intended for single-omic MG data analyses, in this case, it was successfully applied for integrated omic analysis [Roume *et al.*, 2015]. On the other hand, the study by Muller *et al.* [2014b] applied integrated analysis using a customized *ad hoc* analysis workflow. Indeed, both these studies resulted in findings that would not be possible with single-omic based analyses, i.e. identification of keystone genes and species [Roume *et al.*, 2015] and resolution of niche breadth of different populations [Muller *et al.*, 2014b] within the LAMPs. The utilization of MOCAT in the first study [Roume *et al.*, 2015] demonstrates the application of a user-friendly and convenient pipeline that could be easily installed and applied for the analyses of MG data, thus promoting reproducibility and standardization. However, this work clearly showed that MOCAT performs suboptimally in terms of data usage and assembly quality largely due to the fact that it was originally designed for single-omic analyses and that it incorporates older *de novo* assemblers within its workflow. The latter study [Muller *et al.*, 2014b] solved these issues applying an *ad hoc* bioinformatic workflow that included an optimized analysis for enhanced data usage and produces high-quality assemblies through the customized *de novo* MG and MT co-assemblies [Muller *et al.*, 2014b], in addition to using the latest tools [Peng *et al.*, 2012, 2013]. Yet, *ad hoc* workflows such as these would be challenging to reproduce or replicated within other labs as they are usually not automated, thereby hindering standardization of microbial community related studies.

The present work aimed to combine the advantages of the two highlighted integrated-omic studies, namely being able to consistently generate high-quality output as in the *ad hoc* integrated omic analysis workflow [Muller *et al.*, 2014a] and replicating the reproducibility and user-friendly features of existing MG analyses pipelines such as MOCAT [Kultima *et al.*, 2012; Roume *et al.*, 2015]. This work achieved the first objective through the development of the integrated meta-omic pipeline, IMP (**Appendix A.2** and **A.8**). Here my work focused heavily on performing benchmarking and assessments on the measures applied within IMP, more specifically, highlighting the: i) need for separate preprocessing of MG and MT data and ii) the advantages of the combined use of MG and MT data within an extensive co-assembly procedure (**Chapter 2**). This represented in an important formal evaluation for assessing the effectiveness of the various measures applied within the integrated-omic workflow, which have remained largely absent so far despite the recent increase of multi-omic studies of microbial consortia (**Section 1.4.5**). Taking into account the very detailed evaluation, IMP is highly effective in: i) optimizing/maximizing overall data usage, ii) maximising the overall output volume while, iii) generating high-quality output. Overall, the integrated omics approach described by within this work, demonstrated an important transition from systematic measurements to reproducible *in silico* analyses to generate the necessary information for detailed studies of microbial communities, described in **Appendix A.6** and **Chapter 3**.

The second part of this work demonstrated the real power of the integrated omics approach within the Eco-Systems Biology framework, through analysis of temporal datasets (**Figure 1.1**; steps 1 to 4;[Muller *et al.*, 2013; Zarraonaindia *et al.*, 2013]). Such a study was only realizable due to the unique characteristics of the model system (LAMPs), that enabled convenient long terms time-series sampling (**Section 1.2**). In addition, this system offered the opportunity to study the dynamics of bacterial and bacteriophage species using information from the CRISPR-*Cas* system (**Chapter 3**). More specifically, the output generated by IMP, i.e. high-quality preprocessed reads and contigs enabled the extraction of CRISPR sequences and putative phage contigs. Unlike previous studies, the facility of a temporal dataset enabled further observation of CRISPR dynamics within this bacterial community, namely repeat, spacer and protospacer (spacer-complement) dynamics. This work demonstrated that the CRISPR spacers are highly heterogeneous and dynamic elements on the genomic (MG) level (**Sections 3.4.2** and **3.4.3**). On the other hand, the availability of MT datasets demonstrated the CRISPRs to be transcribed genomic regions, mainly thorough the observation of the CRISPR repeat dynamics. This case clearly demonstrates that the MG and MT datasets are more effective at covering different aspects of the CRISPR information, i.e. MG data for CRISPR spacers and MT data for CRISPR repeats, further emphasizing the necessity of multi-omic datasets for the study of microbial communities. This work also leveraged on known isolate genomes to conclusively link bacterial taxa (using CRISPR repeats and flanking regions) to bacteriophages (using CRISPR spacers), representing an effective approach of complementary application of culture-independent and -dependent methodologies. This part of the work yielded putatively novel bacteriophages associated to the dominant lipid accumulating bacterial species of *M. parvicella*, which contained one main large CRISPR region, with accompanying *cas* genes upstream (**Section 3.4.3**) for which 3,956 (non-redundant) spacers were identified. These spacers could be linked to protospacers within 158 putative phage contigs, for which the dynamics of seven putative phages were highlighted within this work (**Section 3.4.5**). Furthermore, the availability of the time-series data set enabled the observation of phage-host dynamics. Overall, this part of the work represented the usage of information to decipher associations and dynamics between different components within the system, namely bacteriophages and bacterial hosts.

In summary, the overall objectives of this work were achieved through: i) the development of multi-omic integrated pipeline for reproducible analyses of coupled MG and MT datasets and ii) the eventual application of this pipeline for a detailed of bacteria-phage dynamics within the model microbial system of LAMPs. As an outlook, this analysis should be followed with the generation of mathematical models and further experimental validation. This will further aid the understanding of the roles and influences of bacteriophages within microbial communities.

4.2 Extending the functionality of IMP

IMP is currently in version 1.4 thereby demonstrating that it is a software that has undergone a large number of changes, since it's conception (**Table 4.1**). More specifically, the 1.4 release tag was due to the inclusion of the automated binning step and further improvement to the command line interface. IMP is relatively extensive compared to other available tools, such that it is able to perform either integrated metagenomic and metatranscriptomic analysis or single omic analysis and is currently the only pipeline which incorporates a binning method (**Chapter 2**). IMP is an open source software, which enables customization by any of the users. This open source nature would hopefully translate into community wide development of the IMP to include more tools and functionality. Updates and modifications to the pipeline are possible due to the modular implementation of IMP through Snakemake, which provides a facility to easily add new steps or modify existing steps (**Sections 2.3.1** and **2.3.1**). The adding of new tools is facilitated through the use of Docker, such that tools need to be install only once and stored as a Docker image, making reproducibility more convenient (**Section 2.3.1**). Here I discuss new features and improvements that could be further incorporated into IMP as described in the sections below.

Table 4.1: The development of IMP.

IMP version	Release date	Implementation
Template integrated-omic workflow	14 November 2014	• Pipeline constructed using shell scripts and was applied to the study by Muller *et al.* [2014a] • Semi-automated
Initial developmental version	15 February 2015	• Workflow constructed by wrapping bash scripts with Snakemake • New analyses steps • All software and dependencies wrapped in Docker • Trinity [Grabherr *et al.*, 2011] replaces IDBA-tran [Peng *et al.*, 2013] as the MT assembler
1.1.1	30 September 2015	• Complete migration of workflow to Snakemake (bash scripts deprecated) • Implementation of Python wrapper script • Update software and dependencies
1.2.1	10 February 2016	• Enhancements of the iterative co-assembly procedure • MEGAHIT [Li *et al.*, 2016, 2015] replaces Trinity [Grabherr *et al.*, 2011] as MT assembler • MEGAHIT [Li *et al.*, 2016, 2015] as an additional option for co-assembly of MG and MT data • Update software and dependencies
1.3	16 June 2016	• Update of tools • Enhancement of Python wrapper script • Enhancements on Docker container
1.4	14 October 2016	• Implementation of Binning procedure (MaxBin 2.0 [Wu *et al.*, 2014]) • Enhancement of workflow for improved modularity • Enhancement of Python wrapper script • Enhancements on Docker container
*1.4.1	January 2017	• New binning tools [Kang *et al.*, 2015; Heintz-Buschart *et al.*, 2016]
*1.4.2	February 2017	• New assembler(s) [Nurk *et al.*, 2016] • New reference based workflow [Schaeffer *et al.*, 2015; Ye and Tang, 2016]
*2.0	July 2017	• Digital normalization for preprocessing [Brown *et al.*, 2012] • Metaproteomic analysis engine [Tang *et al.*, 2016]

*Foreseen future versions

4.2.1 Updates with state-of-the-art tools

In order to keep up with the ever-changing world of bioinformatic software, IMP will require updates through inclusion of additional state-of-the art tools. IMP is centred on extensive iterative co-assemblies of metagenomic and metatranscriptomic data to produce high-quality assemblies. Recent years have not only witnessed steady improvements to metagenomic *de novo* assemblers either in the form of assembly quality [Peng *et al.*, 2012] and/or efficiency (i.e. speed and memory usage) [Li *et al.*, 2015, 2016], but also various enhancements in various pre- and post-analysis steps of NGS datasets (e.g. digital normalization of data and pseudo alignment based methodologies).

Efficient preprocessing of NGS reads prior to assembly was shown to improve downstream *de novo* assemblies [Mende *et al.*, 2012]. While IMP already incorporates a rather stringent preprocessing procedure (**Section 2.4.1**), it could still be improved by the incorporation of digital normalization. This recently introduced concept is a *k*mer-based method to normalize the coverage of shogun metagenome data. The method was shown to: i) reduce sampling variation, ii) discard redundant reads possibly representing highly abundant populations, and iii) the removal of sequencing errors through removal of unique *k*mers [Brown *et al.*, 2012]. The use of digital normalization would then reduce the overall volume of assembly input, while reducing the complexity of the data, to result in higher quality assemblies. The application of this normalization method would be particularly useful for very large (deeply sequenced) datasets and/or datasets that combine multiple samples/extractions/sequencing runs, and may prove especially powerful when combined with the use of highly efficient *de novo* assembly programs such as MEGAHIT, which is already implemented as an option within IMP [Li *et al.*, 2015, 2016]. Such a method could also be applied to MT data as it has similar uneven depth characteristics when compared to MG data. To that end, it would be important to incorporate additional *de novo* assemblers to further enhance the quality of the assemblies generated by IMP, while also providing users with additional choice of assemblers. One notable assembler that could be integrated in the future versions of IMP (**Table 4.1**) would be MetaSPAdes, which promises high-quality, microdiversity-aware assemblies [Nurk *et al.*, 2016].

In addition, there are possibilities for improving the post-assembly steps within IMP, such as the estimation of sequence abundance (contigs and genes). The introduction of pseudo-alignments was shown to be as accurate as direct mapping of reads with the additional advantage of being rapid and memory efficient compared to available mapping algorithms [Teo and Neretti, 2016]. The incorporation of rapid methods such as these would reduce the presently extensive runtime of IMP (**Section 2.4.2**). Furthermore, the quality of IMP output could also be enhanced by updating the gene annotation databases [Seemann, 2014] or by the inclusion of highly customized databases, such as the manually curated viral gene databases of VIRSorter [Roux *et al.*, 2015a]. However, the latter would be highly dependent on the aims of the study.

4.2.2 Integration of reference-based analysis

Although the present work advocates the use of reference-independent methods for the analyses of microbiome NGS data, there are certain advantages associated with reference-based analysis methods (**Section 1.4.4**), especially when applied to well explored microbiomes, with comprehensive collections of sequenced isolates and/or gene catalogues such as the human GIT (**Section 2.4.2**). Therefore, providing the user with an option to perform analyses based on alignment to reference databases (i.e. isolate genomes or gene catalogues) would enhance the overall functionality and flexibility of the IMP (**Section 1.4.4**). More specifically, this would warrant the inclusion of a standard/traditional reference-based workflow within IMP, which would rely on standard read aligners such as bwa [Li and Durbin, 2009] (already implemented within IMP) or bowtie2 [Langmead *et al.*, 2009], whereas a more modern approach would incorporate the previously described pseudo-alignment based software (**Section 4.2.1**). In addition, information such as depth of coverage (for abundance estimation) and SNPs (for observing population-level heterogeneity) could be drawn from the aforementioned alignment information.

Moreover, the present work has highlighted the possibility of combining reference-dependent and reference-independent methods to further improve overall data usage and thereby increasing information gain (**Section 2.4.2**). Given the complementarity of reference-based and referenced-independent methods, a possible integrated solution could leverage on the advantages of both methods (described in **Sections 1.4.4** and **1.4.4**). More specifically, reference-based methods would be effective in recovering well known and lowly abundant bacterial taxa and/or genes from the data. On the opposite side, reference-independent methods would enable the recovery of abundant bacterial taxa, and more importantly genomes from previously uncharacterised bacterial taxa (and thereby novel genes). Furthermore, the mapping of reads to a reference prior to assembly may reduce the total number of reads in a subsequent *de novo* assembly, thus speeding up the overall process. In conclusion, this prospective integration of reference-based and reference-independent methods will further enhance data usage and increase information gain.

4.2.3 Extension to multi-sample analyses

Given the lowering cost of NGS data production, it is currently possible to perform coupled metagenomic and metatranscriptomic studies on a much larger scale than previously envisioned. These include studies that involve: i) a large number of environmental samples [Muller *et al.*, 2014b; Roume *et al.*, 2015; Satinsky *et al.*, 2015], ii) large cohorts [Franzosa *et al.*, 2014; Heintz-Buschart *et al.*, 2016] and iii) long term time-series studies (**Chapter 3**). Therefore, an important future expansion of IMP would be to incorporate multi-sample analysis. This would be further possible with the incorporation of rapid and memory efficient tools described in **Section 4.2.1**.

The facility to handle multiple samples would further allow the incorporation of multi-sample binning algorithms such as CONCOCT [Alneberg *et al.*, 2014], MetaBAT [Kang *et al.*, 2015] and canopy clustering-based binning [Nielsen *et al.*, 2014]. Finally, there is also an option to link bins defined within different samples (based on single sample analyses) using highly conserved marker genes [Herold *et al.*, unpublished].

4.2.4 Metaproteomic analyses engine

Proteins and metabolites can be measured using methods which couple chromatography to mass spectrometry and result in metaproteomic and (meta-)metabolomics data, respectively. Although this work does not specifically cover metaproteomic and metabolomic analyses, they are recognized as a crucial step for downstream data integration for enhanced understanding of the actual functional capacity of microbial communities. Metaproteomic data analysis is reference-based, and thereby require protein sequence databases to perform mass spectrometry based peptide sequence searches. Therefore, the quality of the reference database determines the effectiveness of these searches and thus impacts the final output volume and quality.

It has previously been highlighted that protein sequences predicted from concomitant MG and/or MT data (derived from the same sample) enhance the detection of peptide sequences (**Section 1.4.4**) [Ram *et al.*, 2005; Muller *et al.*, 2014a; Roume *et al.*, 2015; Heintz-Buschart *et al.*, 2016; Ye and Tang, 2016]. IMP generates a large number of predicted sequences, and thereby provides ideal databases for such protein searches (**Chapter 2**). Therefore, the possible incorporation of the current range of available proteomic or metaproteomic analysis engine, such as MetaProteomeAnalyzer [Muth *et al.*, 2015], MetaProSIP [Sachsenberg *et al.*, 2015], Pipasic [Penzlin *et al.*, 2014] and/or the recently introduced integrated de Bruijn graph-based approach for protein identification [Tang *et al.*, 2016]. In summary, the addition of a metaproteomic analysis engine will further extend the functionality of IMP to cover a larger spectrum of omic data sets.

4.2.5 Keeping up with technological advancements

There will be significant technological advancements in all high-throughput measurement techniques particularly in the area of long-read sequencing, chromatography as well as mass spectrometry. Naturally, these technological improvements will be complemented by equally sophisticated *in silico* data processing and analysis methods, which in turn will allow integrated omics to provide comprehensive multi-level snapshots of microbial population structures and functions *in situ* (**Figure 1.1**; step 3).

Long read sequencing platforms, such as PacBio and Oxford Nanopore, will further increase the quality of assemblies. However, different types of *de novo* assemblers will need to be used for these data types. The main issue with these “third generation” sequencing technologies is lower throughput (i.e. bases per run or sequencing depth) and lower accuracy compared to current NGS technologies (**Section 1.4.3**). Therefore, it is not foreseeable that third generation sequencing methods will take over in the near future. Rather, it is more realistic that a combination of NGS and long read sequencing, such that it would combine the high-throughput capabilities from NGS, while longer reads from the third-generation sequencing technologies would be used as a means to scaffold the shorter NGS reads, thus increasing the quality of the assemblies. On the other hand, metaproteomic and (meta-)metabolomic data types are currently limited in their profiling depth, compared to NGS platforms. While the situation for metaproteomics is rapidly improving [Hettich *et al.*, 2012], community-wide metabolomic studies are still limited in their scope due to the poor detection/sensitivity of high-throughput metabolomic instruments and high dependency on a limited knowledgebase reflected in current metabolite databases.

In summary, a long term outlook of IMP would have to include continual updates of the pipeline to handle data from these new technologies, which may either come from different formats and or require specialized programs for their analysis. It would be an important measure to ensure parallel development of technologies and reproducible (standardized) analyses workflows.

4.2.6 Standardized benchmarking for integrated omics

Standardized benchmarks are required for ground truth assessment of any bioinformatic analyses. This is especially relevant for *de novo* assembly (reference-independent) based methods. There are several efforts for standardizing benchmarking and analysis of *de novo* assemblies, such as GAGE (http://gage.cbcb.umd.edu/, [Salzberg *et al.*, 2012]) and Assemblathon (http://assemblathon.org/, [Bradnam *et al.*, 2013]) for isolate genome, while CAMI (http://www.cami-challenge.org/) was founded specifically for assessment of metagenomic data sets.

In addition to these standardized efforts, metagenomic data sets could be computationally simulated using the wide array of metagenomic simulation tools available. However, the opposite is true for metatranscriptomic data sets, whereby there is a sparsity of simulated mock metatranscriptome and even more so for metaproteomic data sets, for which there are no known simulated data sets. This work utilized a simulated mock community, for which the simulated MT data was obtained from previously published work and the MG data was simulated using available metagenome simulators. In my opinion, efforts to produce good software should in fact be centralized around a widely accepted benchmarking dataset (**Section 2.3.9**). This particular data set was also made available on a long term archiving platform, to hopefully standardize the efforts involving integrated omic analyses of MG and MT datasets. To that end, I believe that efforts such as CAMI, which simulated a wide range of metagenomes, should be replicated for other meta-omic datasets (MT and metaproteomic).

Computationally simulated mock communities are the most widely used benchmarking datasets because they can be generated rather easily, especially for MG data. However, computationally simulated mock communities of any omic data types will not be able to simulate data derived from real high-throughput omic measurements of real microbial communities. This is mainly due to the underlying complex characteristics of microbial communities (**Section 1.1**) and the technical biases stemming from the high-throughput measurements themselves. In order to mitigate this issue, wet-lab based mock communities are currently generated by mixing cultures of different bacterial strains in known amounts (HMP, [Shakya *et al.*, 2013]). MG data derived from mock communities are widely available, despite certain datasets, e.g. HMP mock community remain obsolete (i.e. single end sequencing and shallow sequencing) while more recent datasets provide a better option for benchmarking exercises [Shakya *et al.*, 2013]. However, concomitant MT and metaproteomic data derived from these same mock communities are yet to be made available to date. In my opinion, there should be a concerted and collaborative effort to generate all possible meta-omic data sets from a single mock community such that integrated omic analysis could be effectively evaluated from the metagenome up to the metaproteome. While this would be presently challenging, I believe that a concerted collaborative initiative would be worthwhile such that it would be able to standardize all efforts towards the improvement of integrated omic analyses, more effectively compared to the present standard of relying on computationally simulated data sets.

4.3 Bacterial-phage interactions

The study of bacteriophages could be improved through targeted analysis of viral/phage genetic material derived from microbial communities *in situ*. This would specifically involve the filtering and purification of viral-like particles from microbial community samples to perform targeted analyses of these viromes subsequently. NGS sequencing measurements on viromes bypass the limitations of NGS technologies which tend to be biased towards highly abundant organisms with larger genomes, such as the *M. parvicella* bacterial population within LAMPs, and thereby provide better access to the viral/phage genomic components within the community. With regards to the current work, it would be of interest to validate these finding using a combination of phage-specific culture and single-cell methodologies. Therefore, the extracted virus-like particles could also be used for co-culturing with bacterial species, which would be useful for the study of low abundant populations such as LCSB005. Finally, virome sequencing would also allow higher quality *de novo* reconstructions of phage genomes.

Laboratory validation is crucial in confirming observations within large-scale bioinformatic analyses. With regards to the phage host interactions, the first step could involve the quantification of phages (and associated) hosts using quantitative polymerase chain reaction (on DNA) and/or reverse-transcription quantitative polymerase chain reaction (on RNA). These protocols provide more robust quantitative estimations of population abundances and expression of genes compared to NGS-based estimations. This approaches would be particularly advantageous in the quantification of lowly abundant populations such as LCSB005. In addition one would be apply more advanced targeted methods such as digital PCR, viral tagging, PhageFISH and single-cell sequencing [Edwards *et al.*, 2015; Jover *et al.*, 2016]. Such targeted methods would be able to conclusively confirm phage-host relationships and interactions.

4.3.1 Moving beyond associations and hypotheses

This work represents the conversion of data into information and associations, allowing the formulation of hypotheses which may be tested for ascertaining causality. Hypotheses can be generated using a combination of appropriate statistical and mathematical modelling methods to enable the deconvolution of the information to uncover unprecedented insights into the structure and function of microbial communities (**Figure 1.2**; step 4) [Muller *et al.*, 2013; Segata *et al.*, 2013; Zarraonaindia *et al.*, 2013]. Data mining, machine learning and/or modelling approaches will be useful for extracting features of interest, e.g. known and unknown populations/genes, and also to derive associations (or links) between desired features utilizing measures such as correlation, co-occurrence, mutual information and hyper-geometric overlap [Muller *et al.*, 2013; Segata *et al.*, 2013]. Such associations may allow the prediction of gene functions using the concept of 'guilt by association' and interactions/dependencies between community members [Wolfe *et al.*, 2005; Muller *et al.*, 2013; Segata *et al.*, 2013; Solomon *et al.*, 2014]. Within the scope of the present work, the respective information and associations drawn through the longitudinal-based integrated omic analyses, will enable downstream hypothesis generation through the postulation of phage and host interaction networks and/or dynamic modelling thereof. The application of these methodologies will enable the elucidation of phage-host interaction as well as gaining new knowledge about phage biology and thereby the generation of novel hypothesis.

However, the derived associations will always be 'mere' hypotheses, which will require rigorous testing through targeted laboratory experiments (**Figure 1.2**; step 5) and/or *in situ* perturbation experiments followed by additional omic measurements [Muller *et al.*, 2013; Segata *et al.*, 2013]. Although integrated omics-based approaches are highly effective for large-scale analysis and formulation of hypotheses (including within the context of BWWT plant communities), these efforts are limited due to current high-throughput measurement methods (see previous section) and the reliance on *a priori* knowledge for both taxonomical and functional inferences [Röling *et al.*, 2010]. Hence, there is a need to validate newly generated hypotheses using full-scale plants, customized laboratory-based experiments, such as batch cultures, bioreactors or pilot plants (**Figure 1.2**; step 5) and/or single-cell methods. Hypotheses may be tested using additional integrated omic datasets generated from ancillary samples [Muller *et al.*, 2014b] by using molecular biology techniques such as heterologous gene expression [Wexler *et al.*, 2005; Maixner *et al.*, 2008] or single-cell approaches using microautoradiography-fluorescent *in situ* hybridisation (MAR-FISH), nano-scale secondary-ion mass spectrometry (nanoSIMS) and/or Raman spectroscopy [Huang *et al.*, 2007; Lechene *et al.*, 2007; Musat *et al.*, 2012; Sheik *et al.*, 2014, 2016]. Such a combination of technologies can be used to test hypotheses regarding: (i) community dynamics, (ii) gene expression patterns/interactions, (iii) metabolite abundances, (iv) effect of physico-chemical factors on distinct microbial species and functionalities, (v) gene function associations between any of these. With regards to the current work, it would be of interest to validate these finding using a combination of phage specific culture and single-cell methodologies, including, but not limited to, viral tagging, microfluidic-PCR, single-cell sequencing, Hi-C sequencing and/or phage-FISH [Edwards *et al.*, 2015]. Identified patterns may be subsequently formulated as cues and can be used as input to facilitate knowledge-driven control of different microbial community structures and/or functions using bacteriophages (**Figure 1.2**; step 6) [Withey *et al.*, 2005; Jassim *et al.*, 2016]. Thus, large-scale integrated omic analyses and modelling of *in situ* biological samples, coupled to carefully controlled laboratory experiments, will allow the effective elucidation of novel functions within LAMPs with potential biotechnological applications.

4.3.2 From Eco-Systems Biology to applications

Integrated-omics under the framework of Eco-Systems Biology will aid in the understanding of biotechnological and biomedical processes by dissecting interactions among its constituent populations, their genes and the biotope, with the ultimate aim of maximizing outcomes through various control strategies [Sheik *et al.*, 2014; Muller *et al.*, 2014b]. This work demonstrated that information of gene function, regulation and physiological potential derived from integrated omic data over different spatial and temporal scales holds great promise in harnessing the biotechnological potential of microbial consortia. In particular, advancements in integrated omics followed by hypothesis testing may generate new knowledge [Muller *et al.*, 2013], which may for example be exploited in new approaches for the optimized production of biotechnologically relevant compounds under varying environmental conditions (Chen and Nielsen, 2013). The derived knowledge-base may further be used to fine-tune metabolic pathways at the transcriptional, translational and post-translational levels using the ever-expanding synthetic biology toolbox [Peralta-Yahya *et al.*, 2012]. Examples of possible future applications may include, for instance the bioengineering of fatty acid utilization and production for the production of biodiesel from ‘dirty’ mixed substrates, the engineering of different gene combinations for the production of various alcohols from mixed substrates [Lee *et al.*, 2008] and the generation of hybrid processes by combining biological and chemical production steps resulting in new compounds that could serve as biofuels [Román-Leshkov *et al.*, 2007].

In the context of this work, the use of bacteriophages for manipulation of LAMPs could be foreseen as a viable control strategy. For instance, specific phages could be used to target selected bacterial species to alleviate competition for carbon sources and, thus, promoting the growth of a specific bacterial populations of interest, i.e. populations that are efficient in accumulating lipids such as *M. parvicella* and thus making BWWT plants as a viable source for biodiesel production [Sheik *et al.*, 2014; Muller *et al.*, 2014a]. However, this could only be achieved by thorough understanding of the different aspects that may influence LAMPs (e.g. which are the competing populations, effect of physico-chemical parameters and effect of substrate availability) and a generalized fundamental understanding on the influence of phages on community structure, dynamics and phenotype. I therefore believe that the combination of such knowledge would bring us closer towards the ultimate goal of controlling LAMPs and perhaps other biotechnologically interesting microbial communities *in situ.*

Integrated omics through facilitating direct linkages between genetic potential and final phenotype may become an essential tool in future bioprospecting. Therefore, in my opinion, integrated omics should become the standard means of analysing microbial consortia in the near future and will allow meta-omics to fulfil their promise for the comprehensive discovery of biotechnology- relevant microbial traits in natural consortia. Integrated-omics under the framework of Eco-Systems Biology will aid in the understanding of biotechnological and biomedical processes by dissecting interactions among its constituent populations, their genes and the biotope, with the ultimate aim of maximizing outcomes through various control strategies [Withey *et al.*, 2005; Muller *et al.*, 2013; Sheik *et al.*, 2014; Muller *et al.*, 2014a].

REFERENCES

O. O. Abudayyeh, J. S. Gootenberg, S. Konermann, J. Joung, I. M. Slaymaker, D. B. Cox, S. Shmakov, K. S. Makarova, E. Semenova, L. Minakhin, K. Severinov, A. Regev, E. S. Lander, E. V. Koonin, and F. Zhang. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science*, Epub, 2016.

M. Albertsen, P. Hugenholtz, A. Skarshewski, K. L. Nielsen, G. W. Tyson, and P. H. Nielsen. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature biotechnology*, 31:533–538, 2013a.

M. Albertsen, A. Stensballe, K. L. Nielsen, and P. H. Nielsen. Digging into the extracellular matrix of a complex microbial community using a combined metagenomic and metaproteomic approach. *Water science and technology*, 67:1650–1656, 2013b.

J. Alneberg, B. S. Bjarnason, I. de Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. Binning metagenomic contigs by coverage and composition. *Nature methods*, 11:1144–1146, 2014.

R. I. Amann, W. Ludwig, and K. H. Schleifer. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological reviews*, 59:143–169, 1995.

G. Amitai and R. Sorek. CRISPR-Cas adaptation: insights into the mechanism of action. *Nature reviews microbiology*, 14:67–76, 2016.

P. Amstutz, M. R. Crusoe, N. Tijanić, B. Chapman, J. Chilton, M. Heuer, A. Kartashov, D. Leehr, H. Ménager, M. Nedeljkovich, M. Scales, S. Soiland-Reyes, and L. Stojanovic. Common Workflow Language, v1.0. 2016.

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome biology*, 11:R106, 2010.

A. F. Andersson and J. F. Banfield. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science*, 320:1047–1050, 2008.

K. Andrade, J. Logemann, K. B. Heidelberg, J. B. Emerson, L. R. Comolli, L. A. Hug, A. J. Probst, A. Keillar, B. C. Thomas, C. S. Miller, E. E. Allen, J. W. Moreau, J. J. Brocks, and J. F. Banfield. Metagenomic and lipid analyses reveal a diel cycle in a hypersaline microbial ecosystem. *The ISME journal*, 9:2697–2711, 2015.

K. R. Arrigo. Marine microorganisms and global nutrient cycles. *Nature*, 437:349–355, 2005.

F. O. Aylward, K. E. Burnum, J. J. Scott, G. Suen, S. G. Tringe, S. M. Adams, K. W. Barry, C. D. Nicora, P. D. Piehowski, S. O. Purvine, G. J. Starrett, L. A. Goodwin, R. D. Smith, M. S. Lipton, and C. R. Currie. Metagenomic and metaproteomic insights into bacterial communities in leaf-cutter ant fungus gardens. *The ISME journal*, 6:1688–1701, 2012.

R. Bargiela, F.-A. Herbst, M. Martínez-Martínez, J. Seifert, D. Rojo, S. Cappello, M. Genovese, F. Crisafi, R. Denaro, T. N. Chernikova, C. Barbas, M. von Bergen, M. M. Yakimov, M. Ferrer, and P. N. Golyshin. Metaproteomics and metabolomics analyses of chronically petroleum-polluted sites reveal the importance of general anaerobic processes uncoupled with degradation. *Proteomics*, 15:3508–3520, 2015.

J. J. Barr, B. E. Dutilh, C. T. Skennerton, T. Fukushima, M. L. Hastie, J. J. Gorman, G. W. Tyson, and P. L. Bond. Metagenomic and metaproteomic analyses of Accumulibacter phosphatis-enriched floccular and granular biofilm. *Environmental microbiology*, 18:273–287, 2016.

R. Barrangou and J. van der Oost. *CRISPR-Cas system*. Springer, 2013.

R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, and P. Horvath. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 315:1709–1711, 2007.

P. Belmann, J. Dröge, A. Bremges, A. C. McHardy, A. Sczyrba, and M. D. Barton. Bioboxes: standardised containers for interchangeable bioinformatics software. *GigaScience*, 4:47, 2015.

D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E. Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D.

Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. Huw Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. VandeVondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–59, 2008.

P. N. Bertin, A. Heinrich-Salmeron, E. Pelletier, F. Goulhen-Chollet, F. Arsène-Ploetze, S. Gallien, B. Lauga, C. Casiot, A. Calteau, D. Vallenet, V. Bonnefoy, O. Bruneel, B. Chane-Woon-Ming, J. Cleiss-Arnold, R. Duran, F. Elbaz-Poulichet, N. Fonknechten, L. Giloteaux, D. Halter, S. Koechler, M. Marchal, D. Mornico, C. Schaeffer, A. A. T. Smith, A. Van Dorsselaer, J. Weissenbach, C. Médigue, and D. Le Paslier. Metabolic diversity among main microorganisms inside an arsenic-rich ecosystem revealed by meta- and proteo-genomics. *The ISME journal*, 5:1735–1747, 2011.

F. Beulig, T. Urich, M. Nowak, S. E. Trumbore, G. Gleixner, G. D. Gilfillan, K. E. Fjelland, and K. Küsel. Altered carbon turnover processes and microbiomes in soils under long-term extremely high CO2 exposure. *Nature microbiology*, 1:15025, 2016.

E. Biagi, D. Zama, C. Nastasi, C. Consolandi, J. Fiori, S. Rampelli, S. Turroni, M. Centanni, M. Severgnini, C. Peano, G. D. Bellis, G. Basaglia, R. Gotti, R. Masetti, A. Pession, P. Brigidi, and M. Candela. Gut microbiota trajectory in pediatric patients undergoing hematopoietic SCT. *Bone marrow transplantation*, 50:992–998, 2015.

A. Biswas, J. N. Gagnon, S. J. Brouns, P. C. Fineran, and C. M. Brown. CRISPRTarget. *RNA biology*, 10: 817–827, 2013.

L. L. Blackall, H. Stratton, D. Bradford, T. D. Dot, C. Sjörup, E. M. Seviour, and R. J. Seviour. "Candidatus Microthrix parvicella", a filamentous bacterium from activated sludge sewage treatment plants. *International journal of systematic bacteriology*, 46:344–346, 1996.

C. Bland, T. L. Ramsey, F. Sabree, M. Lowe, K. Brown, N. C. Kyrpides, and P. Hugenholtz. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC bioinformatics*, 8:209, 2007.

N. Blow. DNA sequencing: generation next-next. *Nature methods*, 5:267–274, 2008.

A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30:2114–2120, 2014.

K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W.-C. Chou, J. Corbeil, C. Del Fabbro, T. R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N. A. Fonseca, G. Ganapathy, R. A. Gibbs, S. Gnerre, É. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. B. Hiatt, I. Y. Ho, J. Howard, M. Hunt, S. D. Jackman, D. B. Jaffe, E. D. Jarvis, H. Jiang, S. Kazakov, P. J. Kersey, J. O. Kitzman, J. R. Knight, S. Koren, T.-W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. MacCallum, M. D. MacManes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T. D. Otto, B. Paten, O. S. Paulo, A. M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. J. Ribeiro, S. Richards, D. S. Rokhsar, J. G. Ruby, S. Scalabrin, M. C. Schatz, D. C. Schwartz, A. Sergushichev, T. Sharpe, T. I. Shaw, J. Shendure, Y. Shi, J. T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B. M. Vieira, J. Wang, K. C. Worley, S. Yin, S.-M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I. F. Korf. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2:10, 2013.

N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, 34:525–527, 2016.

M. Breitbart and F. Rohwer. Here a virus, there a virus, everywhere the same virus? *Trends in microbiology*, 13:278–284, 2005.

A. Bremges, I. Maus, P. Belmann, F. Eikmeyer, A. Winkler, A. Albersmeier, A. Pühler, A. Schlüter, and A. Sczyrba. Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant. *GigaScience*, 4:33, 2015.

C. T. Brown, A. Howe, Q. Zhang, A. B. Pyrkosz, and T. H. Brom. A reference-free algorithm for computational normalization of shotgun sequencing data. *bioRxiv*, 2012.

B. Buchfink, C. Xie, and D. H. Huson. Fast and sensitive protein alignment using DIAMOND. *Nature methods*, 12:59–60, 2015.

J. J. Bull, J. J. Gill, M. Rebecca, and J. Clokie. The habits of highly effective phages : population dynamics as a framework for identifying therapeutic phages. *Frontiers in microbiology*, 5:618, 2014.

C. J. Bult, O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J.-F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. M. Geoghagen, J. F. Weidman, J. L. Fuhrmann, D. Nguyen, T. R. Utterback, J. M. Kelley, J. D. Peterson, P. W. Sadow, M. C. Hanna, M. D. Cotton, K. M. Roberts, M. A. Hurst, B. P. Kaine, M. Borodovsky, H.-P. Klenk, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter. Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. *Science*, 273:1058–1073, 1996.

M. Burrows and D. J. Wheeler. A block-sorting lossless data compression algorithm. Technical report, 1994.

C. N. Butterfield, Z. Li, P. F. Andeer, S. Spaulding, B. C. Thomas, A. Singh, R. L. Hettich, K. B. Suttle, A. J. Probst, S. G. Tringe, T. Northen, C. Pan, and J. F. Banfield. Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. *PeerJ*, 4:e2687, 2016.

B. J. Cairns, A. R. Timms, V. A. A. Jansen, I. F. Connerton, and R. J. H. Payne. Quantitative models of in vitro bacteriophage-host dynamics and their application to phage therapy. *PLoS pathogens*, 5:e1000253, 2009.

A. Celaj, J. Markle, J. Danska, and J. Parkinson. Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation. *Microbiome*, 2:39, 2014.

R. Claeys. Study on a product with antibiotic action: antibiophagin. *Acta oto-rhino-laryngologica Belgica*, 16:32–8, 1962.

P. E. C. Compeau, P. A. Pevzner, and G. Tesler. How to apply de Bruijn graphs to genome assembly. *Nature biotechnology*, 29:987–991, 2011.

D. J. Conley, H. W. Paerl, R. W. Howarth, D. F. Boesch, S. P. Seitzinger, K. E. Havens, C. Lancelot, and G. E. Likens. Controlling eutrophication: nitrogen and phosphorus. *Science*, 323:1014–1015, 2009.

H. Daims, M. W. Taylor, and M. Wagner. Wastewater treatment: a model system for microbial ecology. *Trends in biotechnology*, 24:483–9, 2006.

K. a. Datsenko, K. Pougach, A. Tikhonov, B. L. Wanner, K. Severinov, and E. Semenova. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nature communications*, 3:945, 2012.

N. de Bruijn and W. van der Woude. A combinatorial problem. *Proceedings of the Nederlandse Akademie van Watenschappen*, 49:758–764, 1946.

M. de la Bastide, W. R. McCombie, M. Bastide, and W. R. McCombie. Assembling genomic DNA sequences with PHRAP. In *Current protocols in bioinformatics* 11.4.1–11.4.15. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2007.

N. Delmotte, C. Knief, S. Chaffron, G. Innerebner, B. Roschitzki, R. Schlapbach, C. von Mering, and J. A. Vorholt. Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 106:16428–33, 2009.

V. J. Denef, L. H. Kalnejais, R. S. Mueller, P. Wilmes, B. J. Baker, B. C. Thomas, N. C. VerBerkmoes, R. L. Hettich, and J. F. Banfield. Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proceedings of the National Academy of Sciences of the United States of America*, 107:2383–90, 2010.

X. Deng, S. N. Naccache, T. Ng, S. Federman, L. Li, Y. Chiu, and E. L. Delwart. An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. *Nucleic acids research*, 43:e46, 2015.

H. Deveau, R. Barrangou, J. E. Garneau, J. Labonté, C. Fremaux, P. Boyaval, D. A. Romero, P. Horvath, and S. Moineau. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *Journal of bacteriology*, 190:1390–400, 2008.

P. D'haeseleer, J. M. Gladden, M. Allgaier, P. S. G. Chain, S. G. Tringe, S. A. Malfatti, J. T. Aldrich, C. D. Nicora, E. W. Robinson, L. Paša-Tolić, P. Hugenholtz, B. A. Simmons, and S. W. Singer. Proteogenomic analysis of a thermophilic bacterial consortium adapted to deconstruct switchgrass. *PLoS ONE*, 8:e68465, 2013.

P. Di Tommaso, E. Palumbo, M. Chatzou, P. Prieto, M. L. Heuer, and C. Notredame. The impact of Docker containers on the performance of genomic pipelines. *PeerJ*, 3:e1273, 2015.

G. J. Dick, A. F. Andersson, B. J. Baker, S. L. Simmons, B. C. Thomas, a. P. Yelton, and J. F. Banfield. Community-wide analysis of microbial genome sequence signatures. *Genome biology*, 10:R85, 2009.

S. Dulal and T. O. Keku. Gut microbiome and colorectal adenomas. *Cancer journal*, 20:225–31, 2014.

S. R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–63, 1998.

S. R. Eddy. Hidden Markov models. *Current opinion in structural biology*, 6:361–365, 1996.

R. C. Edgar. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC bioinformatics*, 8:18, 2007.

R. A. Edwards and F. Rohwer. Opinion: viral metagenomics. *Nature reviews microbiology*, 3:504–510, 2005.

R. A. Edwards, K. McNair, K. Faust, J. Raes, B. E. Dutilh, A. Biswas, J. N. Gagnon, S. J. Brouns, P. C. Fineran, and C. M. Brown. Computational approaches to predict bacteriophage-host relationships. *FEMS microbiology reviews*, 40:fuv048, 2015.

J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. DeWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323:133–138, 2009.

J. B. Emerson, K. Andrade, B. C. Thomas, A. Norman, E. E. Allen, K. B. Heidelberg, and J. F. Banfield. Virus-host and CRISPR dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia. *Archaea*, 2013:370871, 2013a.

J. B. Emerson, B. C. Thomas, K. Andrade, K. B. Heidelberg, and J. F. Banfield. New approaches indicate constant viral diversity despite shifts in assemblage structure in an Australian hypersaline lake. *Applied and environmental microbiology*, 79:6755–6764, 2013b.

A. M. Eren, Ö. C. Esen, C. Quince, J. H. Vineis, H. G. Morrison, M. L. Sogin, and T. O. Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, 2015.

A. R. Erickson, B. L. Cantarel, R. Lamendella, Y. Darzi, E. F. Mongodin, C. Pan, M. Shah, J. Halfvarson, C. Tysk, B. Henrissat, J. Raes, N. C. Verberkmoes, C. M. Fraser, R. L. Hettich, and J. K. Jansson. Integrated metagenomics/metaproteomics reveals human host-microbiota signatures of Crohn's Disease. *PLoS ONE*, 7:e49138, 2012.

R. Feiner, T. Argov, L. Rabinovich, N. Sigal, I. Borovok, and A. A. Herskovits. A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nature reviews microbiology*, 13:641–650, 2015.

P. C. Fineran and E. Charpentier. Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information. *Virology*, 434:202–209, 2012.

R. D. Finn, J. Clements, and S. R. Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, 39:W29–37, 2011.

E. A. Franzosa, X. C. Morgan, N. Segata, L. Waldron, J. Reyes, A. M. Earl, G. Giannoukos, M. R. Boylan, D. Ciulla, D. Gevers, J. Izard, W. S. Garrett, A. T. Chan, and C. Huttenhower. Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences of the United States of America*, 111:E2329–E2338, 2014.

L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28:3150–3152, 2012.

W. M. Gelbart and C. M. Knobler. The physics of phages. *Physics today*, 61:42–47, 2008.

J. A. Gilbert, F. Meyer, L. Schriml, I. R. Joint, M. Mühling, and D. Field. Metagenomes and metatranscriptomes from the L4 long-term coastal monitoring station in the Western English Channel. *Standards in genomic sciences*, 3:183–193, 2010.

J. B. Glass, H. Yu, J. A. Steele, K. S. Dawson, S. Sun, K. Chourey, C. Pan, R. L. Hettich, and V. J. Orphan. Geochemical, metagenomic and metaproteomic insights into trace metal utilization by methane-oxidizing microbial consortia in sulphidic marine sediments. *Environmental microbiology*, 16:1592–1611, 2014.

J. S. Godde and A. Bickerton. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *Journal of molecular evolution*, 62:718–729, 2006.

D. S. A. Goltsman, V. J. Denef, S. W. Singer, N. C. VerBerkmoes, M. Lefsrud, R. S. Mueller, G. J. Dick, C. L. Sun, K. E. Wheeler, A. Zemla, B. J. Baker, L. Hauser, M. Land, M. B. Shah, M. P. Thelen, R. L. Hettich, and J. F. Banfield. Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing "Leptospirillum rubarum" (Group II) and "Leptospirillum ferrodiazotrophum" (Group III) bacteria in acid mine drainage biofilms. *Applied and environmental microbiology*, 75:4599–615, 2009.

M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. Palma, B. W. Birren, C. Nusbaum, K. Lindblad-toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29:644–652, 2011.

K. Greenhalgh, K. M. Meyer, K. M. Aagaard, and P. Wilmes. The human gut microbiome in health: establishment and resilience of microbiota over a lifetime. *Environmental microbiology*, 18:2103–2116, 2016.

C. Grob, M. Taubert, A. M. Howat, O. J. Burns, J. L. Dixon, H. H. Richnow, N. Jehmlich, M. von Bergen, Y. Chen, and J. C. Murrell. Combining metagenomics with metaproteomics and stable isotope probing reveals metabolic pathways used by a naturally occurring marine methylotroph. *Environmental microbiology*, 17:4007–4018, 2015.

D. H. Haft, J. Selengut, E. F. Mongodin, and K. E. Nelson. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS computational biology*, 1: e60, 2005.

A. Hanreich, U. Schimpf, M. Zakrzewski, A. Schlüter, D. Benndorf, R. Heyer, E. Rapp, A. Pühler, U. Reichl, and M. Klocke. Metagenome and metaproteome analyses of microbial communities in mesophilic biogas-producing anaerobic batch fermentations indicate concerted plant carbohydrate degradation. *Systematic and applied microbiology*, 36:330–338, 2013.

C. H. Hawkes, K. Del Tredici, and H. Braak. Parkinson's disease: a dual-hit hypothesis. *Neuropathology and applied neurobiology*, 33:599–614, 2007.

A. K. Hawley, H. M. Brewer, A. D. Norbeck, L. Paša-Tolić, and S. J. Hallam. Metaproteomics reveals differential modes of metabolic coupling among ubiquitous oxygen minimum zone microbes. *Proceedings of the National Academy of Sciences of the United States of America*, 111:11395–400, 2014.

J. He and M. W. Deem. Heterogeneous diversity of spacers within CRISPR (clustered regularly interspaced short palindromic repeats). *Physical review letters*, 105:128102, 2010.

S. He, N. Ivanova, E. Kirton, M. Allgaier, C. Bergin, R. H. Scheffrahn, N. C. Kyrpides, F. Warnecke, S. G. Tringe, and P. Hugenholtz. Comparative metagenomic and metatranscriptomic analysis of hindgut paunch microbiota in wood- and dung-feeding higher termites. *PLoS ONE*, 8:e61126, 2013.

A. Heintz-Buschart, P. May, C. C. Laczny, L. A. Lebrun, C. Bellora, A. Krishna, L. Wampach, J. G. Schneider, A. Hogan, C. de Beaufort, and P. Wilmes. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature microbiology*, 2:16180, 2016.

R. L. Hettich, R. Sharma, K. Chourey, and R. J. Giannone. Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. *Current opinion in microbiology*, 15:373–380, 2012.

T. Hitch and C. Creevey. Spherical: an iterative workflow for assembling metagenomic datasets. *bioRxiv*, 2016.

Z.-S. Hua, Y.-J. Han, L.-X. Chen, J. Liu, M. Hu, S.-J. Li, J.-L. Kuang, P. S. Chain, L.-N. Huang, and W.-S. Shu. Ecological roles of dominant and rare prokaryotes in acid mine drainage revealed by metagenomics and metatranscriptomics. *The ISME journal*, 9:1280–1294, 2015.

W. E. Huang, K. Stoecker, R. Griffiths, L. Newbold, H. Daims, A. S. Whiteley, and M. Wagner. Raman-FISH: combining stable-isotope Raman spectroscopy and fluorescence in situ hybridization for the single cell analysis of identity and function. *Environmental microbiology*, 9:1878–1889, 2007.

X. Huang and A. Madan. CAP3: A DNA sequence assembly program. *Genome research*, 9:868–877, 1999.

L. A. Hug, B. C. Thomas, I. Sharon, C. T. Brown, R. Sharma, R. L. Hettich, M. J. Wilkins, K. H. Williams, A. Singh, and J. F. Banfield. Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environmental microbiology*, 18:159–173, 2016.

L. W. Hugerth, J. Larsson, J. Alneberg, M. V. Lindh, C. Legrand, J. Pinhassi, and A. F. Andersson. Metagenome-assembled genomes uncover a global brackish microbiome. *Genome biology*, 16:279, 2015.

J. Hultman, M. P. Waldrop, R. Mackelprang, M. M. David, J. Mcfarland, S. J. Blazewicz, J. Harden, M. R. Turetsky, A. D. Mcguire, M. B. Shah, N. C. Verberkmoes, and L. H. Lee. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature*, 521:208–212, 2015.

D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11:119, 2010.

R. M. Idury and M. S. Waterman. A new algorithm for DNA sequence assembly. *Journal of computational biology*, 2:291–306, 1995.

M. Imelfort, D. Parks, B. J. Woodcroft, P. Dennis, P. Hugenholtz, and G. W. Tyson. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, 2:e603, 2014.

Y. Ishino, H. Shinagawa, K. Makino, M. Amemura, and a. Nakata. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *Journal of bacteriology*, 169:5429–5433, 1987.

R. Jansen, J. D. a. V. Embden, W. Gaastra, and L. M. Schouls. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular microbiology*, 43:1565–1575, 2002.

S. A. A. Jassim, R. G. Limoges, and H. El-Cheikh. Bacteriophage biocontrol in wastewater treatment. *World journal of microbiology and biotechnology*, 32:70, 2016.

R. R. Jenq, C. Ubeda, Y. Taur, C. C. Menezes, R. Khanin, J. A. Dudakov, C. Liu, M. L. West, N. V. Singer, M. J. Equinda, A. Gobourne, L. Lipuma, L. F. Young, O. M. Smith, A. Ghosh, A. M. Hanash, J. D. Goldberg, K. Aoyama, and B. R. Blazar. Regulation of intestinal inflammation by microbiota following allogeneic bone marrow transplantation. *The journal of experimental medicine*, 209:903–911, 2012.

B. Jia, L. Xuan, K. Cai, Z. Hu, L. Ma, and C. Wei. NeSSM: a next-generation sequencing simulator for metagenomics. *PloS ONE*, 8:e75448, 2013.

M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337:816–821, 2012.

M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden. NCBI BLAST: a better web interface. *Nucleic acids research*, 36:W5–9, 2008.

L. F. Jover, J. Romberg, and J. S. Weitz. Inferring phage-bacteria infection networks from time-series data. *Royal society open science*, 3:160654, 2016.

M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28: 27–30, 2000.

D. D. Kang, J. Froula, R. Egan, and Z. Wang. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3:e1165, 2015.

F. V. Karginov and G. J. Hannon. The CRISPR system: small RNA-guided defense in bacteria and archaea. *Molecular cell*, 37:7–19, 2010.

A. Kenall, S. Edmunds, L. Goodman, L. Bal, L. Flintoft, D. R. Shanahan, and T. Shipley. Better reporting for better research: a checklist for reproducibility. *BMC neuroscience*, 16:44, 2015.

A. Keshavarzian, S. J. Green, P. A. Engen, R. M. Voigt, A. Naqib, C. B. Forsyth, E. Mutlu, and K. M. Shannon. Colonic bacterial composition in Parkinson's disease. *Movement disorders*, 30:1351–1360, 2015.

N. E. Kimes, A. V. Callaghan, D. F. Aktas, W. L. Smith, J. Sunner, B. Golding, M. Drozdowska, T. C. Hazen, J. M. Suflita, and P. J. Morris. Metagenomic analysis and metabolite profiling of deep-sea sediments from the Gulf of Mexico following the Deepwater Horizon oil spill. *Frontiers in microbiology*, 4:50, 2013.

M. Kleiner, C. Wentrup, C. Lott, H. Teeling, S. Wetzel, J. Young, Y.-J. Chang, M. Shah, N. C. VerBerkmoes, J. Zarzycki, G. Fuchs, S. Markert, K. Hempel, B. Voigt, D. Becher, M. Liebeke, M. Lalk, D. Albrecht, M. Hecker, T. Schweder, and N. Dubilier. Metaproteomics of a gutless marine worm and its symbiotic microbial community reveal unusual pathways for carbon and energy use. *Proceedings of the National Academy of Sciences of the United States of America*, 109:E1173–E1182, 2012.

H.-P. Klenk, R. A. Clayton, J.-F. Tomb, O. White, K. E. Nelson, K. A. Ketchum, R. J. Dodson, M. Gwinn, E. K. Hickey, J. D. Peterson, D. L. Richardson, A. R. Kerlavage, D. E. Graham, N. C. Kyrpides, R. D. Fleischmann, J. Quackenbush, N. H. Lee, G. G. Sutton, S. Gill, E. F. Kirkness, B. A. Dougherty, K. McKenney, M. D. Adams, B. Loftus, S. Peterson, C. I. Reich, L. K. McNeil, J. H. Badger, A. Glodek, L. Zhou, R. Overbeek, J. D. Gocayne, J. F. Weidman, L. McDonald, T. Utterback, M. D. Cotton, T. Spriggs, P. Artiach, B. P. Kaine, S. M. Sykes, P. W. Sadow, K. P. D'Andrea, C. Bowman, C. Fujii, S. A. Garland, T. M. Mason, G. J. Olsen, C. M. Fraser, H. O. Smith, C. R. Woese, and J. C. Venter. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. *Nature*, 390:364–370, 1997.

C. Knief. Analysis of plant microbe interactions in the era of next generation sequencing technologies. *Frontiers in plant science*, 5:216, 2014.

C. Knief, N. Delmotte, S. Chaffron, M. Stark, G. Innerebner, R. Wassmann, C. von Mering, and J. A. Vorholt. Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. *The ISME journal*, 6:1378–1390, 2012.

E. V. Koonin and P. Starokadomskyy. Are viruses alive? The replicator paradigm sheds decisive light on an old but misguided question. *Studies in history and philosophy of biological and biomedical science*, 59: 125–134, 2016.

E. Kopylova, L. Noé, and H. Touzet. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, 28:3211–3217, 2012.

J. Köster. *Reproducibility in next-generation sequencing analysis*. PhD thesis, Technischen Universitat Dortmund, 2014.

J. Köster and S. Rahmann. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics*, 28: 2520–2522, 2012.

I. Kozarewa, Z. Ning, M. A. Quail, M. J. Sanders, M. Berriman, and D. J. Turner. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature methods*, 6:291–5, 2009.

J. R. Kultima, S. Sunagawa, J. Li, W. Chen, H. Chen, D. R. Mende, M. Arumugam, Q. Pan, B. Liu, J. Qin, J. Wang, and P. Bork. MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS ONE*, 7: e47656, 2012.

V. Kunin, R. Sorek, and P. Hugenholtz. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome biology*, 8:R61, 2007.

V. Kunin, S. He, F. Warnecke, S. B. Peterson, H. Garcia Martin, M. Haynes, N. Ivanova, L. L. Blackall, M. Breitbart, F. Rohwer, K. D. McMahon, and P. Hugenholtz. A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome research*, 18:293–297, 2008.

S. J. Labrie, J. E. Samson, and S. Moineau. Bacteriophage resistance mechanisms. *Nature reviews microbiology*, 8:317–327, 2010.

C. C. Laczny. *Visualization and binning of metagenomic data*. PhD thesis, University of Luxembourg, 2015.

C. C. Laczny, N. Pinel, N. Vlassis, and P. Wilmes. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Scientific reports*, 4:4516, 2014.

C. C. Laczny, T. Sternal, V. Plugaru, P. Gawron, A. Atashpendar, H. H. Margossian, S. Coronado, L. van der Maaten, N. Vlassis, and P. Wilmes. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*, 3:1, 2015.

C. C. Laczny, E. E. Muller, A. Heintz-Buschart, M. Herold, L. A. Lebrun, A. Hogan, P. May, C. De Beaufort, and P. Wilmes. Identification, recovery, and refinement of hitherto undescribed population-level genomes from the human gastrointestinal tract. *Frontiers in microbiology*, 7:884, 2016.

N. F. Lahens, I. H. Kavakli, R. Zhang, K. Hayer, M. B. Black, H. Dueck, A. Pizarro, J. Kim, R. Irizarry, R. S. Thomas, G. R. Grant, and J. B. Hogenesch. IVT-seq reveals extreme bias in RNA-sequencing. *Genome biology*, 15:R86, 2014.

B. Lai, F. Wang, X. Wang, L. Duan, and H. Zhu. InteMAP: integrated metagenomic assembly pipeline for NGS short reads. *BMC bioinformatics*, 16:244, 2015.

B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10:R25, 2009.

C. P. Lechene, Y. Luyten, G. McMahon, and D. L. Distel. Quantitative imaging of nitrogen fixation by individual bacteria within animal cells. *Science*, 317:1563–1566, 2007.

S. K. Lee, H. Chou, T. S. Ham, T. S. Lee, and J. D. Keasling. Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Current opinion in biotechnology*, 19: 556–563, 2008.

M. M. Leimena, J. Ramiro-Garcia, M. Davids, B. van den Bogert, H. Smidt, E. J. Smid, J. Boekhorst, E. G. Zoetendal, P. J. Schaap, and M. Kleerebezem. A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. *BMC genomics*, 14:530, 2013.

R. Leinonen, H. Sugawara, M. Shumway, and International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic acids research*, 39:D19–21, 2011.

J. Leipzig. A review of bioinformatic pipeline frameworks. In *Briefings in bioinformatics* bbw020. Oxford University Press, 2016.

H. C. M. Leung, S.-M. Yiu, J. Parkinson, and F. Y. L. Chin. IDBA-MT: de novo assembler for metatranscriptomic data generated from next-generation sequencing technology. *Journal of computational biology*, 20: 540–550, 2013.

H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin. IDBA-MTP : a hybrid metatranscriptomic assembler based on protein information. In *Research in computational molecular biology* 160–172, 2014.

D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, 31:1674–1676, 2015.

D. Li, R. Luo, C.-M. Liu, C.-M. Leung, H.-F. Ting, K. Sadakane, H. Yamashita, and T.-W. Lam. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*, 102:3–11, 2016.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25:589–595, 2009.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25:2078–2079, 2009.

J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J. R. Kultima, E. Prifti, T. Nielsen, A. S. Juncker, C. Manichanh, B. Chen, W. Zhang, F. Levenez, J. Wang, X. Xu, L. Xiao, S. Liang, D. Zhang, Z. Zhang, W. Chen, H. Zhao, J. Y. Al-Aama, S. Edris, H. Yang, J. Wang, T. Hansen, H. B. Nielsen,

S. Brunak, K. Kristiansen, F. Guarner, O. Pedersen, J. Doré, S. D. Ehrlich, P. Bork, and J. Wang. An integrated catalog of reference genes in the human gut microbiome. *Nature biotechnology*, 32:834–841, 2014.

Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, H. Zhang, J. Gan, N. Li, X. Hu, B. Liu, B. Yang, and W. Fan. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Briefings in functional genomics*, 11:25–37, 2012.

Y. W. Lim, R. Schmieder, M. Haynes, D. Willner, M. Furlan, M. Youle, K. Abbott, R. Edwards, J. Evangelista, D. Conrad, and F. Rohwer. Metagenomics and metatranscriptomics: windows on CF-associated viral and microbial communities. *Journal of cystic fibrosis*, 12:154–164, 2013.

M. Liu, L. Fan, L. Zhong, S. Kjelleberg, and T. Thomas. Metaproteogenomic analysis of a community of sponge symbionts. *The ISME journal*, 6:1515–1525, 2012.

R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S.-M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.-W. Lam, and J. Wang. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1:18, 2012.

F. Maixner, M. Wagner, S. Lücker, E. Pelletier, S. Schmitz-esser, K. Hace, E. Spieck, R. Konrat, D. L. Paslier, and H. Daims. Environmental genomics reveals a functional chlorite dismutase in the nitrite-oxidizing bacterium ‘Candidatus Nitrospira defluvii’. *Environmental microbiology*, 10:3043–3056, 2008.

K. S. Makarova, Y. I. Wolf, S. Snir, and E. V. Koonin. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *Journal of bacteriology*, 193:6039–6056, 2011.

E. A. Manrao, I. M. Derrington, A. H. Laszlo, K. W. Langford, M. K. Hopper, N. Gillgren, M. Pavlenok, M. Niederweis, and J. H. Gundlach. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature biotechnology*, 30:349–353, 2012.

D. D. Mara and N. J. Horan. *The handbook of water and wastewater microbiology*. Academic, 2003.

M. B. Marcó, S. Moineau, and A. Quiberoni. Bacteriophages and dairy fermentations. *Bacteriophage*, 2: 149–158, 2012.

M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, 2005.

D. C. B. Mariano, T. d. J. Sousa, F. L. Pereira, F. Aburjaile, D. Barh, F. Rocha, A. C. Pinto, S. S. Hassan, T. D. L. Saraiva, F. A. Dorella, A. F. de Carvalho, C. A. G. Leal, H. C. P. Figueiredo, A. Silva, R. T. J. Ramos, and V. A. C. Azevedo. Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002. *BMC genomics*, 17:315, 2016.

L. A. Marraffini and E. J. Sontheimer. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature reviews genetics*, 11:181–190, 2010.

A. Martínez, L.-A. Ventouras, S. T. Wilson, D. M. Karl, and E. F. DeLong. Metatranscriptomic and functional metagenomic analysis of methylphosphonate utilization by marine bacteria. *Frontiers in microbiology*, 4: 340, 2013.

X. Martinez, M. Pozuelo, V. Pascal, D. Campos, I. Gut, M. Gut, F. Azpiroz, F. Guarner, and C. Manichanh. MetaTrans: an open-source pipeline for metatranscriptomics. *Scientific reports*, 6:26447, 2016.

O. U. Mason, T. C. Hazen, S. Borglin, P. S. G. Chain, E. A. Dubinsky, J. L. Fortney, J. Han, H.-Y. N. Holman, J. Hultman, R. Lamendella, R. Mackelprang, S. Malfatti, L. M. Tom, S. G. Tringe, T. Woyke, J. Zhou, E. M. Rubin, and J. K. Jansson. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *The ISME journal*, 6:1715–1727, 2012.

S. J. McIlroy, R. Kristiansen, M. Albertsen, S. M. Karst, S. Rossetti, J. L. Nielsen, V. Tandoi, R. J. Seviour, and P. H. Nielsen. Metabolic model for the filamentous 'Candidatus Microthrix parvicella' based on genomic and metagenomic analyses. *The ISME journal*, 7:1161–72, 2013.

S. J. McIlroy, A. M. Saunders, M. Albertsen, M. Nierychlo, B. McIlroy, A. A. Hansen, S. M. Karst, J. L. Nielsen, and P. H. Nielsen. MiDAS: the field guide to the microbes of activated sludge. *Database*, 2015: bav062, 2015.

K. J. McKernan, H. E. Peckham, G. L. Costa, S. F. McLaughlin, Y. Fu, E. F. Tsung, C. R. Clouser, C. Duncan, J. K. Ichikawa, C. C. Lee, Z. Zhang, S. S. Ranade, E. T. Dimalanta, F. C. Hyland, T. D. Sokolsky, L. Zhang, A. Sheridan, H. Fu, C. L. Hendrickson, B. Li, L. Kotler, J. R. Stuart, J. A. Malek, J. M. Manning, A. A. Antipova, D. S. Perez, M. P. Moore, K. C. Hayashibara, M. R. Lyons, R. E. Beaudoin, B. E. Coleman, M. W. Laptewicz, A. E. Sannicandro, M. D. Rhodes, R. K. Gottimukkala, S. Yang, V. Bafna, A. Bashir, A. MacBride, C. Alkan, J. M. Kidd, E. E. Eichler, M. G. Reese, F. M. De La Vega, and A. P. Blanchard. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research*, 19:1527–1541, 2009.

D. R. Mende, A. S. Waller, S. Sunagawa, A. I. Järvelin, M. M. Chan, M. Arumugam, J. Raes, and P. Bork. Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE*, 7: e31386, 2012.

P. Menzel, K. L. Ng, and A. Krogh. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature communications*, 7:11257, 2016.

F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9:386, 2008.

A. Mikheenko, V. Saveliev, and A. Gurevich. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, 32:1088–1090, 2015.

S. Mocali and A. Benedetti. Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. *Research in microbiology*, 161:497–505, 2010.

F. J. M. Mojica, C. Diez-Villasenor, E. Soria, and G. Juez. MicroCorrespondence: biological significance of a family of regularly spaced repeats in genomes of Archaea, Bacteria and mitochondria. *Molecular microbiology*, 36:244–246, 2000.

E. E. L. Muller, N. Pinel, J. D. Gillece, J. M. Schupp, L. B. Price, D. M. Engelthaler, C. Levantesi, V. Tandoi, K. Luong, N. S. Baliga, J. Korlach, P. S. Keim, and P. Wilmes. Genome sequence of "Candidatus Microthrix parvicella" Bio17-1, a long-chain-fatty-acid-accumulating filamentous actinobacterium from a biological wastewater treatment plant. *Journal of bacteriology*, 194:6670–6671, 2012.

E. E. L. Muller, E. Glaab, P. May, N. Vlassis, and P. Wilmes. Condensing the omics fog of microbial communities. *Trends in microbiology*, 21:325–333, 2013.

E. E. L. Muller, A. R. Sheik, and P. Wilmes. Lipid-based biofuel production from wastewater. *Current opinion in biotechnology*, 30C:9–16, 2014a.

E. E. Muller, N. Pinel, C. C. Laczny, M. R. Hoopman, S. Narayanasamy, L. A. Lebrun, H. Roume, J. Lin, P. May, N. D. Hicks, A. Heintz-Buschart, L. Wampach, C. M. Liu, L. B. Price, J. D. Gillece, C. Guignard, J. M. Schupp, N. Vlassis, N. S. Baliga, R. L. Moritz, P. S. Keim, and P. Wilmes. Community integrated omics links the dominance of a microbial generalist to fine-tuned resource usage. *Nature communications*, 5:5603, 2014b.

N. Musat, R. Foster, T. Vagner, B. Adam, and M. M. M. Kuypers. Detecting metabolic activities in single cells, with emphasis on nanoSIMS. *FEMS microbiology reviews*, 36:486–511, 2012.

T. Muth, A. Behne, R. Heyer, F. Kohrs, D. Benndorf, M. Hoffmann, M. Lehtevä, U. Reichl, L. Martens, and E. Rapp. The MetaProteomeAnalyzer: a powerful open-source software suite for metaproteomics data analysis and interpretation. *Journal of proteome research*, 14:1557–1565, 2015.

E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. J. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H.-H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter. A whole-genome assembly of drosophila. *Science*, 287:2196–2204, 2000.

O. U. Nalbantoglu, S. F. Way, S. H. Hinrichs, and K. Sayood. RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC bioinformatics*, 12:41, 2011.

T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara. MetaVelvet: An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic acids research*, 40:e155, 2012.

S. Narayanasamy, E. E. L. Muller, A. R. Sheik, and P. Wilmes. Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. *Microbial biotechnology*, 8: 363–368, 2015.

H. B. Nielsen, M. Almeida, A. S. Juncker, S. Rasmussen, J. Li, S. Sunagawa, D. R. Plichta, L. Gautier, A. G. Pedersen, E. Le Chatelier, E. Pelletier, I. Bonde, T. Nielsen, C. Manichanh, M. Arumugam, J.-M. Batto, M. B. Quintanilha Dos Santos, N. Blom, N. Borruel, K. S. Burgdorf, F. Boumezbeur, F. Casellas, J. Doré, P. Dworzynski, F. Guarner, T. Hansen, F. Hildebrand, R. S. Kaas, S. Kennedy, K. Kristiansen, J. R. Kultima, P. Léonard, F. Levenez, O. Lund, B. Moumen, D. Le Paslier, N. Pons, O. Pedersen, E. Prifti, J. Qin, J. Raes, S. Sørensen, J. Tap, S. Tims, D. W. Ussery, T. Yamada, P. Renault, T. Sicheritz-Ponten, P. Bork, J. Wang, S. Brunak, and S. D. Ehrlich. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature biotechnology*, 32:822–828, 2014.

H. Nishimasu, F. A. Ran, P. D. Hsu, S. Konermann, S. I. Shehata, N. Dohmae, R. Ishitani, F. Zhang, and O. Nureki. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, 156:935–49, 2014.

V. Nolla-Ardèvol, M. Strous, and H. E. Tegetmeyer. Anaerobic digestion of the microalga Spirulina at extreme alkaline conditions: biogas production, metagenome, and metatranscriptome. *Frontiers in microbiology*, 6: 597, 2015.

J. K. Nuñez, P. J. Kranzusch, J. Noeske, A. V. Wright, C. W. Davies, and J. A. Doudna. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nature structural & molecular biology*, 21:528–534, 2014.

J. K. Nuñez, L. B. Harrington, P. J. Kranzusch, A. N. Engelman, and J. A. Doudna. Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature*, 527:535–538, 2015.

S. Nurk, D. Meleshko, A. Korobeynikov, and P. Pevzner. MetaSPAdes: a new versatile de novo metagenomics assembler. *bioRxiv*, 2016.

E. Odum and G. Barrett. *Fundamentals of ecology*. 1971.

B. D. Ondov, N. H. Bergman, and A. M. Phillippy. Interactive metagenomic visualization in a Web browser. *BMC bioinformatics*, 12:385, 2011.

V. Ortseifen, Y. Stolze, I. Maus, A. Sczyrba, A. Bremges, S. P. Albaum, S. Jaenicke, J. Fracowiak, A. Pühler, and A. Schlüter. An integrated metagenome and -proteome analysis of the microbial community residing in a biogas production plant. *Journal of biotechnology*, 231:268–279, 2016.

D. Paez-Espino, I. Sharon, W. Morovic, B. Stahl, B. C. Thomas, R. Barrangou, and J. F. Banfield. CRISPR immunity drives rapid phage genome evolution in Streptococcus thermophilus. *mBio*, 6:e00262–15, 2015.

D. Paez-Espino, E. A. Eloe-Fadrosh, G. A. Pavlopoulos, A. D. Thomas, M. Huntemann, N. Mikhailova, E. Rubin, N. N. Ivanova, and N. C. Kyrpides. Uncovering Earth's virome. *Nature*, 536:425–430, 2016.

R. J. Parsons, M. Breitbart, M. W. Lomas, and C. A. Carlson. Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea. *The ISME journal*, 6:273–284, 2012.

R. K. Patel and M. Jain. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE*, 7:e30619, 2012.

Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin. IDBA - a practical iterative de Bruijn graph de novo assembler. *Research in computational molecular biology*, 6044:426–440, 2010.

Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28:1420–1428, 2012.

Y. Peng, H. C. M. Leung, S.-M. M. Yiu, M.-J. J. Lv, X.-G. G. Zhu, and F. Y. L. Chin. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*, 29:i326–i334, 2013.

A. Penzlin, M. S. Lindner, J. Doellinger, P. W. Dabrowski, A. Nitsche, and B. Y. Renard. Pipasic: similarity and expression correction for strain-level identification and quantification in metaproteomics. *Bioinformatics*, 30:149–156, 2014.

P. P. Peralta-Yahya, F. Zhang, S. B. del Cardayre, and J. D. Keasling. Microbial engineering for the production of advanced biofuels. *Nature*, 488:320–328, 2012.

P. A. Pevzner, H. Tang, and M. S. Waterman. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, 98:9748–53, 2001.

C. Pourcel, G. Salvignol, and G. Vergnaud. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*, 151:653–663, 2005.

A. A. Price, T. R. Sampson, H. K. Ratner, A. Grakoui, and D. S. Weiss. Cas9-mediated targeting of viral RNA in eukaryotic cells. *Proceedings of the National Academy of Sciences of the United States of America*, 112:6164–6169, 2015.

D. T. Pride, J. Salzman, M. Haynes, F. Rohwer, C. Davis-Long, R. a. White, P. Loomer, G. C. Armitage, and D. a. Relman. Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome. *The ISME journal*, 6:915–926, 2012.

K. Pruitt, G. Brown, T. Tatusova, and D. Maglott. *The reference sequence (RefSeq) database*. 2002.

J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, M. Antolin, F. Artiguenave, H. Blottiere, N. Borruel, T. Bruls, F. Casellas, C. Chervaux, A. Cultrone, C. Delorme, G. Denariaz, R. Dervyn, M. Forte, C. Friss, M. van de Guchte, E. Guedon, F. Haimet, A. Jamet, C. Juste, G. Kaci, M. Kleerebezem, J. Knol, M. Kristensen, S. Layec, K. Le Roux, M. Leclerc, E. Maguin, R. Melo Minardi, R. Oozeer, M. Rescigno, N. Sanchez, S. Tims, T. Torrejon,

E. Varela, W. de Vos, Y. Winogradsky, E. Zoetendal, P. Bork, S. D. Ehrlich, and J. Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464:59–65, 2010.

M. a. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13:341, 2012.

A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26:841–842, 2010.

J. Raes and P. Bork. Molecular eco-systems biology: towards an understanding of community function. *Nature reviews microbiology*, 6:693–699, 2008.

R. J. Ram, N. C. VerBerkmoes, M. P. Thelen, G. W. Tyson, B. J. Baker, R. C. Blake, M. Shah, R. L. Hettich, and J. F. Banfield. Community proteomics of a natural microbial biofilm. *Science*, 308:1915–1919, 2005.

D. Raoult and P. Forterre. Redefining viruses: lessons from Mimivirus. *Nature reviews microbiology*, 6: 315–319, 2008.

D. Rath, L. Amlinger, A. Rath, and M. Lundgren. The CRISPR-Cas immune system: biology, mechanisms and applications. *Biochimie*, 117:119–128, 2015.

T. B. K. Reddy, A. D. Thomas, D. Stamatis, J. Bertsch, M. Isbandi, J. Jansson, J. Mallajosyula, I. Pagani, E. A. Lobos, and N. C. Kyrpides. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic acids research*, 43:D1099–106, 2015.

A. Reyes, N. P. Semenkovich, K. Whiteson, F. Rohwer, and J. I. Gordon. Going viral: next-generation sequencing applied to phage populations in the human gut. *Nature reviews microbiology*, 10:607–617, 2012.

A. Reyes, M. Wu, N. P. McNulty, F. L. Rohwer, and J. I. Gordon. Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. *Proceedings of the National Academy of Sciences of the United States of America*, 110:20236–41, 2013.

A. Reyes, L. V. Blanton, S. Cao, G. Zhao, M. Manary, I. Trehan, M. I. Smith, D. Wang, H. W. Virgin, F. Rohwer, and J. I. Gordon. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proceedings of the National Academy of Sciences of the United States of America*, 112:11941–11946, 2015.

A. Rimmer, H. Phan, I. Mathieson, Z. Iqbal, S. R. F. Twigg, A. O. M. Wilkie, G. McVean, and G. Lunter. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics*, 46:912–918, 2014.

S. Rodrigue, A. C. Materna, S. C. Timberlake, M. C. Blackburn, R. R. Malmstrom, E. J. Alm, and S. W. Chisholm. Unlocking short read sequencing for metagenomics. *PLoS ONE*, 5:e11840, 2010.

S. Rogers, Y. Shtarkman, Z. Koçer, R. Edgar, R. Veerapaneni, and T. D'Elia. Ecology of subglacial Lake Vostok (Antarctica), based on metagenomic/metatranscriptomic analyses of accretion ice. *Biology*, 2: 629–650, 2013.

W. F. M. Röling, M. Ferrer, and P. N. Golyshin. Systems approaches to microbial communities and their functioning. *Current opinion in biotechnology*, 21:532–538, 2010.

Y. Román-Leshkov, C. J. Barrett, Z. Y. Liu, and J. A. Dumesic. Production of dimethylfuran for liquid fuels from biomass-derived carbohydrates. *Nature*, 447:982–985, 2007.

H. Roume, A. Heintz-Buschart, E. E. L. Muller, and P. Wilmes. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. *Methods in enzymology*, 531:219–236, 2013a.

H. Roume, E. E. L. Muller, T. Cordes, J. Renaut, K. Hiller, and P. Wilmes. A biomolecular isolation framework for eco-systems biology. *The ISME journal*, 7:110–121, 2013b.

H. Roume, A. Heintz-Buschart, E. E. L. Muller, P. May, V. P. Satagopam, C. C. Laczny, S. Narayanasamy, L. A. Lebrun, M. R. Hoopmann, J. M. Schupp, J. D. Gillece, N. D. Hicks, D. M. Engelthaler, T. Sauter, P. S. Keim, R. L. Moritz, and P. Wilmes. Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. *NPJ biofilms and microbiomes*, 1:15007, 2015.

S. Roux, M. Faubladier, A. Mahul, N. Paulhe, A. Bernard, D. Debroas, and F. Enault. Metavir: a web server dedicated to virome analysis. *Bioinformatics*, 27:3074–3075, 2011.

S. Roux, F. Enault, B. L. Hurwitz, and M. B. Sullivan. VirSorter: mining viral signal from microbial genomic data. *PeerJ*, 3:e985, 2015a.

S. Roux, S. J. Hallam, T. Woyke, and M. B. Sullivan. Viral dark matter and virus-host interactions resolved from publicly available microbial genomes. *eLife*, 4:e08490, 2015b.

E. Rybicki. The classification of organisms at the edge of life or problems with virus systematics. *South African journal of science*, 86:182–186, 1990.

T. Sachsenberg, F.-A. Herbst, M. Taubert, R. Kermer, N. Jehmlich, M. von Bergen, J. Seifert, and O. Kohlbacher. MetaProSIP: automated inference of stable isotope incorporation rates in proteins for functional metaproteomics. *Journal of proteome research*, 14:619–627, 2015.

S. L. Salzberg and J. A. Yorke. Beware of mis-assembled genomes. *Bioinformatics*, 21:4320–4321, 2005.

S. L. Salzberg, A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marçais, M. Pop, and J. A. Yorke. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome research*, 22:557–567, 2012.

J. E. Samson, A. H. Magadán, M. Sabri, and S. Moineau. Revenge of the phages : defeating bacterial defences. *Nature reviews microbiology*, 11:675–687, 2013.

F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94:441–448, 1975.

F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74:5463–5467, 1977.

D. G. Sashital, B. Wiedenheft, and J. A. Doudna. Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Molecular cell*, 46:606–615, 2012.

B. B. M. Satinsky, C. S. Fortunato, M. Doherty, C. B. C. Smith, S. Sharma, N. D. N. N. D. Ward, A. A. V. Krusche, P. L. Yager, J. E. Richey, M. A. Moran, and B. B. C. Crump. Metagenomic and metatranscriptomic inventories of the lower Amazon River, May 2011. *Microbiome*, 3:39, 2015.

B. M. Satinsky, B. L. Zielinski, M. Doherty, C. B. Smith, S. Sharma, J. H. Paul, B. C. Crump, and M. Moran. The Amazon continuum dataset: quantitative metagenomic and metatranscriptomic inventories of the Amazon River plume, June 2010. *Microbiome*, 2:17, 2014.

L. Schaeffer, H. Pimentel, N. Bray, A. Melsted, and L. Pachter. Pseudoalignment for metagenomic read assignment. *bioRxiv* 1–13, 2015.

A. C. Schürch, D. Schipper, M. A. Bijl, J. Dau, K. B. Beckmen, C. M. E. Schapendonk, V. S. Raj, A. D. M. E. Osterhaus, B. L. Haagmans, M. Tryland, and S. L. Smits. Metagenomic survey for viruses in Western Arctic caribou, Alaska, through iterative assembly of taxonomic units. *PLoS ONE*, 9:e105227, 2014.

T. Seemann. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30:2068–2069, 2014.

N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, C. Huttenhower, D. Boernigen, T. L. Tickle, X. C. Morgan, W. S. Garrett, and C. Huttenhower. Computational meta'omics for microbial community studies. *Molecular systems biology*, 9:666, 2013.

K. Selle and R. Barrangou. Harnessing CRISPR-Cas systems for bacterial genome editing. *Trends in microbiology*, 23:225–232, 2015.

S. Shah, S. Erdmann, F. J. M. Mojica, and R. a. Garrett. Protospacer recognition motifs: mixed identities and functional diversity. *RNA biology*, 10:891–899, 2013.

K. Shahzad and J. J. Loor. Application of top-Down and bottom-up systems approaches in ruminant physiology and metabolism. *Current genomics*, 13:379–394, 2012.

M. Shakya, C. Quince, J. H. Campbell, Z. K. Yang, C. W. Schadt, and M. Podar. Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental microbiology*, 15:1882–1899, 2013.

Y. Shao and I.-N. Wang. Bacteriophage adsorption rate and optimal lysis time. *Genetics*, 180:471–482, 2008.

A. R. Sheik, E. E. L. Muller, and P. Wilmes. A hundred years of activated sludge: time for a rethink. *Frontiers in microbiology*, 5:47, 2014.

A. R. Sheik, E. E. Muller, J.-N. Audinot, L. A. Lebrun, P. Grysan, C. Guignard, and P. Wilmes. In situ phenotypic heterogeneity among single cells of the filamentous bacterium Candidatus Microthrix parvicella. *The ISME journal*, 10:1274–1279, 2016.

Y. Shi, G. W. Tyson, J. M. Eppley, and E. F. DeLong. Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *The ISME journal*, 5:999–1013, 2011.

B. Shirley, V.-L. Alejandra, C.-G. Fernanda, R. Karina, C.-Q. Samuel, S. Xavier, D. P.-Y. Luis, and O.-L. Adrián. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiomes. *Computational and structural biotechnology journal*, 13:390–401, 2015.

S. Silas, G. Mohr, D. J. Sidote, L. M. Markham, A. Sanchez-Amat, D. Bhaya, A. M. Lambowitz, and A. Z. Fire. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science*, 351:aad4234, 2016.

J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol. ABySS: a parallel assembler for short read sequence data. *Genome research*, 19:1117–1123, 2009.

C. T. Skennerton, M. Imelfort, and G. W. Tyson. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic acids research*, 41:1–10, 2013.

K. V. Solomon, C. H. Haitjema, D. A. Thompson, and M. A. O'Malley. Extracting data from the muck: deriving biological insight from complex microbial communities and non-model organisms with next generation sequencing. *Current opinion in biotechnology*, 28C:103–110, 2014.

R. Staden. A strategy of DNA sequencing employing computer programs. *Nucleic acids research*, 6:2601–10, 1979.

J. T. Staley and A. Konopka. Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual review of microbiology*, 39:321–346, 1985.

A. Stern, E. Mick, I. Tirosh, O. Sagy, and R. Sorek. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome research*, 22:1985–1994, 2012.

E. J. Stewart. Growing unculturable bacteria. *Journal of bacteriology*, 194:4151–4160, 2012.

R. Stokke, I. Roalkvam, A. Lanzen, H. Haflidason, and I. H. Steen. Integrated metagenomic and metaproteomic analyses of an ANME-1-dominated community in marine cold seep sediments. *Environmental microbiology*, 14:1333–1346, 2012.

S. Sunagawa, D. R. Mende, G. Zeller, F. Izquierdo-Carrasco, S. A. Berger, J. R. Kultima, L. P. Coelho, M. Arumugam, J. Tap, H. B. Nielsen, S. Rasmussen, S. Brunak, O. Pedersen, F. Guarner, W. M. de Vos, J. Wang, J. Li, J. Doré, S. D. Ehrlich, A. Stamatakis, and P. Bork. Metagenomic species profiling using universal phylogenetic marker genes. *Nature methods*, 10:1196–1199, 2013.

C. A. Suttle. Marine viruses - major players in the global ecosystem. *Nature reviews microbiology*, 5:801–812, 2007.

H. Tang, S. Li, and Y. Ye. A graph-centric approach for metagenome-guided peptide and protein identification in metaproteomics. *PLoS computational biology*, 12:e1005224, 2016.

Y. Taur, J. B. Xavier, L. Lipuma, C. Ubeda, J. Goldberg, A. Gobourne, Y. J. Lee, K. A. Dubin, N. D. Socci, A. Viale, M.-A. Perales, R. R. Jenq, M. R. M. van den Brink, and E. G. Pamer. Intestinal domination and the risk of bacteremia in patients undergoing allogeneic hematopoietic stem cell transplantation. *Clinical infectious diseases*, 55:905–914, 2012.

Y. V. Teo and N. Neretti. A comparative study of metagenomics analysis pipelines at the species level. *bioRxiv* 1–21, 2016.

T. J. Treangen, S. Koren, D. D. Sommer, B. Liu, I. Astrovskaya, B. Ondov, A. E. Darling, A. M. Phillippy, and M. Pop. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome biology*, 14:R2, 2013.

W. L. Trimble, K. P. Keegan, M. D'Souza, A. Wilke, J. Wilkening, J. Gilbert, and F. Meyer. Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. *BMC bioinformatics*, 13:183, 2012.

P. J. Turnbaugh, R. E. Ley, M. Hamady, C. Fraser-Liggett, R. Knight, and J. I. Gordon. The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature*, 449:804–810, 2007.

T. Urich, A. Lanzén, R. Stokke, R. B. Pedersen, C. Bayer, I. H. Thorseth, C. Schleper, I. H. Steen, and L. Ovreas. Microbial community structure and functioning in marine sediments associated with diffuse hydrothermal venting assessed by integrated meta-omics. *Environmental microbiology*, 16:2699–2710, 2014.

R. P. van Rij and R. Andino. The silent treatment: RNAi as a defense against virus infection in mammals. *Trends in biotechnology*, 24:186–193, 2006.

I. Vanwonterghem, P. D. Jensen, D. P. Ho, D. J. Batstone, and G. W. Tyson. Linking microbial community structure, interactions and function in anaerobic digesters using new molecular techniques. *Current opinion in biotechnology*, 27:55–64, 2014.

S. Varrette, P. Bouvry, H. Cartiaux, and F. Georgatos. Management of an academic HPC cluster : the UL experience. In *International conference on high performance computing & simulation* 959–967, 2014.

E. Vogtmann and J. J. Goedert. Epidemiologic studies of the human microbiome and cancer. *British journal of cancer*, 114:237–242, 2016.

M. Wagner and A. Loy. Bacterial community composition and function in sewage treatment systems. *Current opinion in biotechnology*, 13:218–227, 2002.

J. Wang, J. Li, H. Zhao, G. Sheng, M. Wang, M. Yin, and Y. Wang. Structural and mechanistic basis of PAM-dependent spacer acquisition in CRISPR-Cas systems. *Cell*, 163:840–853, 2015.

L. K. Wang and N. C. Pereira. *Biological Treatment Processes*. Humana Press, 1987.

B. Wemheuer, F. Wemheuer, J. Hollensteiner, F.-D. Meyer, S. Voget, and R. Daniel. The green impact: bacterioplankton response toward a phytoplankton spring bloom in the southern North Sea assessed by comparative metagenomic and metatranscriptomic approaches. *Frontiers in microbiology*, 6:805, 2015.

S. T. Westreich, I. Korf, D. A. Mills, D. G. Lemay, M. Moran, M. Leimena, M. Embree, K. McGrath, D. Dimitrov, I. Cho, M. Blaser, J. Round, S. Mazmanian, M. Gosalbes, G. Giannoukos, M. Reck, E. Hainzl, A. Bolger, M. Lohse, B. Usadel, T. Magoc, S. Salzberg, F. Meyer, T. Tatusova, A. Wilke, R. Overbeek, M. Love, W. Huber, S. Anders, V. Costa, C. Neut, F. Guillemot, J. Colombel, and S. Ohkawara. SAMSA: a comprehensive metatranscriptome analysis pipeline. *BMC bioinformatics*, 17:399, 2016.

M. Wexler, P. L. Bond, D. J. Richardson, and A. W. B. Johnston. A wide host-range metagenomic library from a waste water treatment plant yields a novel alcohol/aldehyde dehydrogenase. *Environmental microbiology*, 7:1917–1926, 2005.

B. Wiedenheft, E. van Duijn, J. B. Bultema, J. Bultema, S. P. Waghmare, S. Waghmare, K. Zhou, A. Barendregt, W. Westphal, A. J. R. Heck, A. Heck, E. J. Boekema, E. Boekema, M. J. Dickman, M. Dickman, and J. A. Doudna. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 108:10092–10097, 2011.

P. Wilmes, A. F. Andersson, M. G. Lefsrud, M. Wexler, M. Shah, B. Zhang, R. L. Hettich, P. L. Bond, N. C. VerBerkmoes, and J. F. Banfield. Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *The ISME journal*, 2:853–864, 2008.

P. Wilmes, S. L. Simmons, V. J. Denef, and J. F. Banfield. The dynamic genetic repertoire of microbial communities. *FEMS microbiology reviews*, 33:109–132, 2009.

P. Wilmes, B. P. Bowen, B. C. Thomas, R. S. Mueller, V. J. Denef, N. C. VerBerkmoes, R. L. Hettich, T. R. Northen, and J. F. Banfield. Metabolome-proteome differentiation coupled to microbial divergence. *mBio*, 1:e00246–10, 2010.

S. Withey, E. Cartmell, L. Avery, and T. Stephenson. Bacteriophages-potential for application in wastewater treatment processes. *Science of the total environment*, 339:1–18, 2005.

C. J. Wolfe, I. S. Kohane, and A. J. Butte. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. *BMC bioinformatics*, 6:227, 2005.

K. E. Wommack and R. R. Colwell. Virioplankton: viruses in aquatic ecosystems. *Microbiology and molecular biology reviews*, 64:69–114, 2000.

K. E. Wommack, J. Bhavsar, S. W. Polson, J. Chen, M. Dumas, S. Srinivasiah, M. Furman, S. Jamindar, and D. J. Nasko. VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in genomic sciences*, 6:427–439, 2012.

D. E. Wood and S. L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15:R46, 2014.

J. Wu, W. Gao, R. Johnson, W. Zhang, and D. Meldrum. Integrated metagenomic and metatranscriptomic analyses of microbial communities in the meso- and bathypelagic realm of North Pacific Ocean. *Marine drugs*, 11:3777–3801, 2013.

Y.-W. Wu, Y.-H. Tang, S. G. Tringe, B. A. Simmons, and S. W. Singer. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, 2:26, 2014.

Y. Xia, Y. Wang, H. H. P. Fang, T. Jin, H. Zhong, and T. Zhang. Thermophilic microbial cellulose decomposition and methanogenesis pathways recharacterized by metatranscriptomic and metagenomic analysis. *Scientific reports*, 4:6708, 2014.

T. Yasunori, M. Katsunori, Y. Masatoshi, M. Masatomo, H. Katsutoshi, and U. Hajime. Fate of coliphage in a wastewater treatment process. *Journal of bioscience and bioengineering*, 94:172–174, 2002.

Y. Ye and H. Tang. Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis. *Bioinformatics*, 32:1001–1008, 2016.

H. Yin, W. Xue, S. Chen, R. L. Bogorad, E. Benedetti, M. Grompe, V. Koteliansky, P. A. Sharp, T. Jacks, and D. G. Anderson. Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype. *Nature biotechnology*, 32:551–553, 2014.

K. Yu and T. Zhang. Metagenomic and metatranscriptomic analysis of microbial community structure and gene expression of activated sludge. *PLoS ONE*, 7:e38183, 2012.

I. Zarraonaindia, D. P. Smith, and J. a. Gilbert. Beyond the genome: community-level analysis of the microbial world. *Biology & philosophy*, 28:261–282, 2013.

M. Zeimes. *Functional genomics and population dynamics of lipid-accumulating bacterial strains in biological wastewater treatment plants*. Masters thesis, University of Luxembourg, 2015.

K. Zengler. Central role of the cell in microbial ecology. *Microbiology and molecular biology reviews*, 73: 712–729, 2009.

K. Zengler and B. O. Palsson. A road map for the development of community systems (CoSy) biology. *Nature reviews microbiology*, 10:366–372, 2012.

D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18:821–829, 2008.

Q. Zhang, M. Rho, H. Tang, T. G. Doak, and Y. Ye. CRISPR-Cas systems target a diverse collection of invasive mobile genetic elements in human microbiomes. *Genome biology*, 14:R40, 2013.

Q. Zhang, T. G. Doak, and Y. Ye. Expanding the catalog of cas genes with metagenomes. *Nucleic acids research*, 42:2448–2459, 2014.

Y. Zhang, F. Zhao, Y. Deng, Y. Zhao, and H. Ren. Metagenomic and metabolomic analysis of the toxic effects of trichloroacetamide-induced gut microbiome and urine metabolome perturbations in mice. *Journal of proteome research*, 14:1752–1761, 2015.

W. Zhu, A. Lomsadze, and M. Borodovsky. Ab initio gene identification in metagenomic sequences. *Nucleic acids research*, 38:e132, 2010.

GLOSSARY

k-mer Sequence of short length, e.g., between 4 and 6 symbols. Symbols are from a fixed alphabet, e.g., "A","C","G", or "T" in the case of DNA.

De novo Starting from the beginning.

In silico Conducted or produced by means of computational representations, modeling or simulation (of scientific measurements, experiments, research).

In situ In its original place.

Amplicon sequencing The process of sequencing a set of target sequences, e.g., genetic fragments from a specific genomic region. Typically, the respective sequences are amplified prior to sequencing, e.g., via polymerase chain reactions.

AWS Amazon Web Services.

BG biogas.

BH-SNE Barnes-Hut stochastic neighbour embedding.

Binning Process of grouping sequence fragments derived from closely related taxa.

bp base pair.

BWWT biological wastewater treatment.

CAMI Critical Assessment of Metagenome Interpretation.

cDNA complementary-DNA.

CDS coding DNA sequence.

CG composite genome.

Community Collection of populations of (micro)organisms.

Community structure Composition of a community with respect to individual, constituent populations.

Contig Contiguous sequence, generally a product of sequence assembly.

CRISPR clustered regularly inter-spaced palindromic repeats.

crRNA CRISPR RNA.

DBG de Bruijn graph.

DNA deoxyribonucleic acid.

Dysbiosis Microbial imbalance.

FISH fluorescence *in situ* hybridisation.

Genomic signature Signature of a genomic sequence defined using, e.g., %GC content or k-mer composition. Can be seen as a fingerprint of the respective genome.

GFF general feature format.

GIT gastrointestinal tract.

HF human fecal.

HMM hidden Markov model.

HMP Human Microbiome Project.

ICG Integrated Gene Catalogue.

IMP Integrated Meta-omic Pipeline.

INDELs insertions and deletions.

Kb kilo base.

KEGG Kyoto Encyclopaedia of Genes and Genomes.

LAMPs lipid accumulating microbial populations.

MetaHIT Metagenomics of Human Intestinal Tract.

MG metagenomic.

MT metatranscriptomic.

NCBI National Center for Biotechnology Information.

NCBI NR National Center for Biotechnology Information non-redundant.

NGS Next-Generation Sequencing.

OLC overlap concensus layout.

Omics Collective technologies used to explore the roles, relationships, and actions of the various types of biomolecules.

ONT Oxford Nanopore Technologies®.

PacBio Pacific Biosciences®.

PCR polymerase chain reaction.

Population A collection of microbial cells of the same species/subtype present in the same place and at the same time.

qPCR quantitative polymerase chain reaction.

Read Sequencing product.

RIGe(s) RNA-based invasive genetic elements.

RNA ribonecleic acid.

rRNA ribosomal RNA.

SM simulated mock.

SMRT single molecule real-time.

SNE stochastic neighbour embedding.

SNPs single nucleotide polymorphisms.

SRA Sequence read archive.

t-SNE t-distributed stochastic neighbour embedding.

Taxon Unit of classification, e.g., species, family, or phylum.

tRNA transfer RNA.

VCF variant call format.

WGS whole genome shotgun.

WW wastewater.

ZMW zero-mode waveguide.

Appendices

APPENDIX A

ARTICLE MANUSCRIPTS

This appendix contains all manuscripts authored as a first author or co-author. Journal formatted articles are provided for published manuscripts. Submitted manuscripts or manuscripts that are ready for submission are provided as the submitted versions.

A.1 Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities.

Shaman Narayanasamy, Emilie E.L. Muller, Abdul R. Sheik, Paul Wilmes
2015
Microbial Biotechnology **8**: 363-368
DOI: 10.1111/1751-7915.12255

This manuscript was written upon an invitation by journal editor to submit an opinions article about microbial communities within biological wastewater treatment plants. Contributions of author include:

- Coordination
- Figure creation
- Writing and revision of manuscript

microbial biotechnology

Open Access

Opinion

Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities

Shaman Narayanasamy, Emilie E. L. Muller, Abdul R. Sheik and Paul Wilmes*
Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7 avenue des Hauts-Fourneaux, Esch-Sur-Alzette L-4362, Luxembourg.

Summary

Biological wastewater treatment plants harbour diverse and complex microbial communities which prominently serve as models for microbial ecology and mixed culture biotechnological processes. Integrated omic analyses (combined metagenomics, metatranscriptomics, metaproteomics and metabolomics) are currently gaining momentum towards providing enhanced understanding of community structure, function and dynamics *in situ* as well as offering the potential to discover novel biological functionalities within the framework of Eco-Systems Biology. The integration of information from genome to metabolome allows the establishment of associations between genetic potential and final phenotype, a feature not realizable by only considering single 'omes'. Therefore, in our opinion, integrated omics will become the future standard for large-scale characterization of microbial consortia including those underpinning biological wastewater treatment processes. Systematically obtained time and space-resolved omic datasets will allow deconvolution of structure–function relationships by identifying key members and functions. Such knowledge will form the foundation for discovering novel genes on a much larger scale compared with previous efforts. In general, these insights will allow us to optimize microbial biotechnological processes either through better control of mixed culture processes or by use of more efficient enzymes in bioengineering applications.

Biological wastewater treatment as a model system for Eco-Systems Biology

Biological wastewater treatment (BWWT), including the standard activated sludge process and other ancillary processes, relies on microbial community-driven remediation of municipal and industrial wastewater. Biological wastewater treatment plants host diverse and dynamic microbial communities possessing varied metabolic capabilities over changing environmental conditions, e.g. microorganisms accumulating various storage compounds of biotechnological importance. Given their structural and functional diversity, BWWT processes hold great potential for future sustainable production of various commodities from wastewater as well as from other mixed substrates (Muller *et al.*, 2014; Sheik *et al.*, 2014). Eco-Systems Biology is an integrative framework that includes systematic measurements, data integration, analysis, modelling, prediction, experimental validation (e.g. through targeted perturbations) and ultimately control of microbial ecosystems (Muller *et al.*, 2013). This framework will aid in the understanding of BWWT processes by dissecting interactions among its constituent populations, their genes and the biotope, with the ultimate aim of maximizing biotechnological outcomes through various control strategies (Muller, Pinel *et al.*, 2014; Sheik *et al.*, 2014).

Biological wastewater treatment plants typically possess a relatively homogeneous environment (compared with most natural ecosystems) with well-defined physico-chemical boundaries and are widespread in developed and developing countries (Daims *et al.*, 2006; Muller, Pinel *et al.*, 2014; Sheik *et al.*, 2014). Furthermore, contrary to other microbial habitats, e.g. the marine environment, acid mine drainage biofilms, the human gastrointestinal tract, etc., BWWT plants represent a

Received 21 August, 2014; revised 11 November, 2014; accepted 13 November, 2014. *For correspondence. E-mail paul.wilmes@uni.lu; Tel. +352 46 66 44 6188; Fax +352 46 66 44 6949.
doi:10.1111/1751-7915.12255
Funding Information This work was supported by an ATTRACT programme grant (A09/03) and a European Union Joint Programming in Neurodegenerative Diseases grant (INTER/JPND/12/01) to PW and Aide à la Formation Recherche (AFR) grants to SN (PHD-2014-1/7934898), EELM (PDR-2011-1/SR) and ARS (PDR-2013-1/5748561) all funded by the Luxembourg National Research Fund (FNR).

© 2015 The Authors. *Microbial Biotechnology* published by John Wiley & Sons Ltd and Society for Applied Microbiology.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

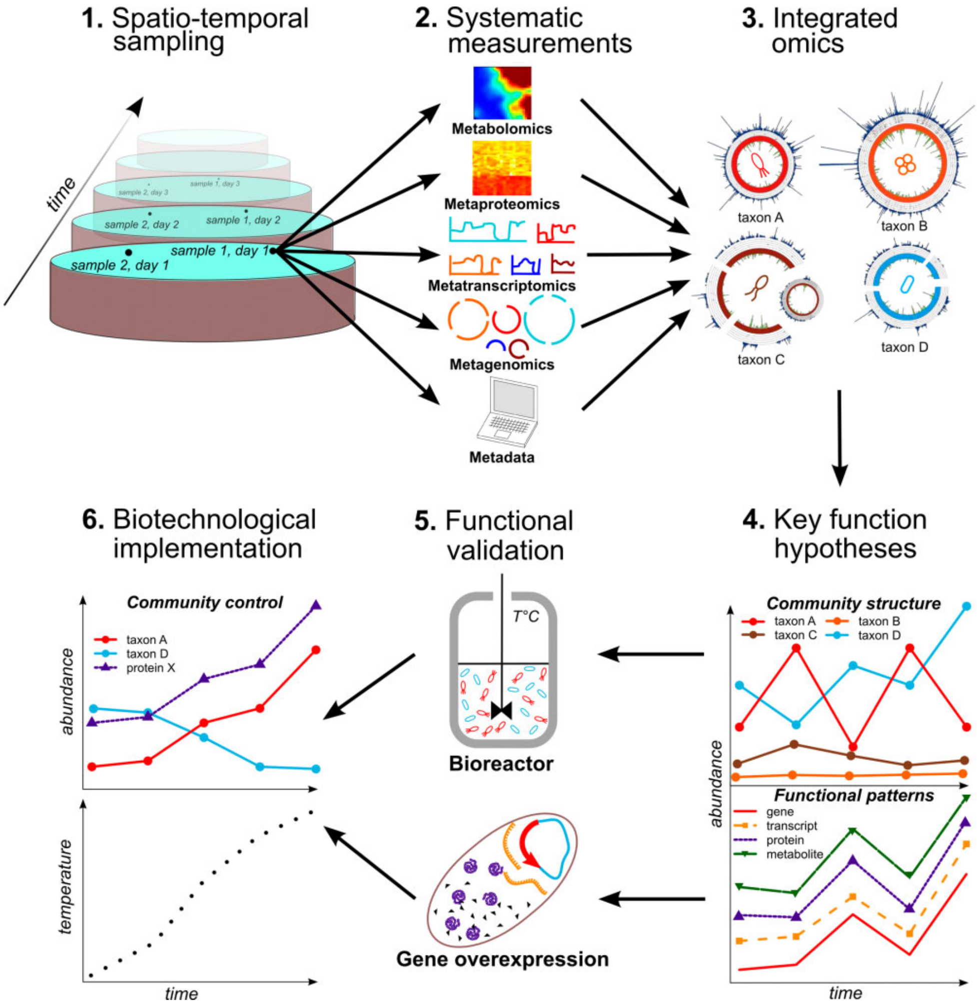


Fig. 1. The path from large-scale integrated omics to hypothesis testing and biotechnological application in the context of biological wastewater treatment.

convenient and virtually unlimited source of spatially and temporally resolved samples (Fig. 1; step 1). Physico-chemical parameters such as temperature, pH, oxygen and nutrient concentrations are routinely monitored and recorded, thereby facilitating hypothesis formulation and verification in rapid succession. This allows for example, the establishment of causal links between the influence of certain environmental parameters on microbial community structure and/or function derived from temporal sampling. Importantly, microbial consortia from BWWT plants are very amenable to experimental validation at differing scales, ranging from laboratory-scale bioreactors to full-scale plants (see section “From Eco-Systems Biology to biotechnology” below).

While being highly dynamic, microbial communities within BWWT plants maintain a medium to high range of diversity/complexity, thereby exhibiting a baseline stability over time such that there is temporal succession of repeatedly few quantitatively dominant populations (Albertsen *et al*., 2012; Zhang *et al*., 2012; Muller, Pinel *et al*., 2014; N. Pinel, pers. comm.). These characteristics reduce the complexity of downstream omic data analyses. In particular, given sufficient sequencing depth, current *de novo* metagenomic assemblers are highly effective for medium complexity communities, such as BWWT plant microbial communities (Segata *et al*., 2013; Muller, Pinel *et al*., 2014). Representative population-level genomic reconstructions can now be obtained for abundant community members (Albertsen *et al*., 2013; Muller, Pinel *et al*., 2014), and such genomic information is vital for the meaningful interpretation of additional functional omic data. Overall, BWWT plant microbial communities represent an important intermediary step/model between communities of lower diversity, e.g. acid mine drainage biofilms (Denef *et al*., 2010), and complex communities such as those from soil environments (Mocali and Benedetti, 2010), while retaining important hallmarks of both extremes including, for example, quantitative

© 2015 The Authors. *Microbial Biotechnology* published by John Wiley & Sons Ltd and Society for Applied Microbiology

dominance of specific taxa (a characteristic of acid mine drainage biofilm communities), rapid stochastic environmental fluctuations (a characteristic of soil environments). Therefore, BWWT plant microbial communities exhibit important properties rendering them an ideal model for microbial ecology (Daims *et al*., 2006), and more specifically eco-systematic omic studies in line with a discovery-driven planning approach (Muller *et al*., 2013).

Laboratory protocols, systematic measurements and *in silico* analyses

Mixed microbial communities, such as those present in BWWT plants, exhibit varying degrees of inter- and intra-sample heterogeneity, rendering standard (i.e. originally designed for pure isolate culture systems) biomolecular extractions protocols and computational analyses ineffective (Muller *et al*., 2013; Roume *et al*., 2013a). In our opinion, it is therefore absolutely essential to apply tailored and systematic approaches such as the biomolecular isolation protocol designed by Roume and colleagues (Roume *et al*., 2013a) to microbial communities. The protocol allows the sequential isolation of high-quality genomic deoxyribonucleic acid (DNA), ribonucleic acid (RNA), small RNA, proteins and metabolites from a single, undivided sample for subsequent systematic multi-omic measurements (Fig. 1, step 2). Importantly, this eliminates the need for subsampling the heterogeneous biomass and, therefore, reduces the noise arising from incongruous omics data in the subsequent downstream integration and analysis steps (Fig. 1, step 3; Muller *et al*., 2013; Roume *et al*., 2013a,b).

Following standardized and systematized biomolecular isolations, multi-omic datasets are generated in addition to the physico-chemical parameters recorded at the time of sampling (Fig. 1; step 2). The multi-omic data are then subjected to bioinformatic pre-processing and analyses. Preliminary characterization of microbial communities can be facilitated either by high-throughput ribosomal RNA gene amplicon sequencing to determine broad community composition from shotgun metagenomic analyses to resolve the overall structure as well as the functional potential of the communities (Vanwonterghem *et al*., 2014). More importantly, hybrid *de novo* assemblies of metagenomic and metatranscriptomic reads promises higher quality compared with conventional *de novo* metagenomic assemblies due to the ability to reconstruct and resolve genomic complements of low abundance (i.e. low metagenomic coverage) yet highly active populations (i.e. high metatranscriptomic coverage for expressed genes; Muller, Pinel *et al*., 2014). Hybrid assemblies allow high-quality population-level genomic reconstructions after the application of binning/classification methods, such as those developed for a single sample (Laczny *et al*., 2014) or for spatio-temporally resolved samples (Albertsen *et al*., 2013; Alneberg *et al*., 2014; Nielsen *et al*., 2014). Furthermore, hybrid metagenomic and metatranscriptomic data assemblies allow the resolution of genetic variations with higher confidence through replication and highlights their potential relative importance, thereby allowing more detailed short-term evolutionary inferences regarding specific populations and while increasing sensitivity for downstream metaproteomic analysis (Muller, Pinel *et al*., 2014). Thus, the generation of metatranscriptomic and metaproteomic data is crucial to fully understand the functional capacity of microbial communities. Therefore, we believe that the integrated omic approach as elucidated by Muller and colleagues (Muller, Pinel *et al*., 2014), from systematic measurements to *in silico* analysis, is highly effective in: (i) minimizing errors by cancelling out noise and biases stemming from single omic analyses and (ii) optimizing/maximizing overall data usage.

Although high-throughput metagenomics and metatranscriptomics allow deep profiling of microbial communities at relatively low cost, existing sequence-based approaches do have some important limitations. Given the availability of omic technologies and their non-prohibitive costs (in particular for metagenomics and metatranscriptomics), fully integrated omic analyses should be applied routinely in the study of microbial consortia for greater effectiveness. For instance, despite this wealth of information, current metagenomic assemblies and analysis schemes, metagenomic (and metatranscriptomic) data resulting from the use of current short-read sequencing and assembly approaches do not allow the comprehensive resolution of microdiversity, e.g. genetic heterogeneity of microbial populations (Wilmes *et al*., 2009). Furthermore, RNAseq technologies are subject to biases stemming from the extensive, yet compulsory pre-processing steps (Lahens *et al*., 2014), thereby affecting the resulting metatranscriptomic data. On the other hand, chromatography and mass spectrometry-based metaproteomics and metabolomics currently remain limited in their profiling depth. While the situation for metaproteomics is rapidly improving (Hettich *et al*., 2012), community-wide metabolomic studies are still limited in their scope due to the poor detection/sensitivity of high-throughput metabolomic instruments and high dependency on a limited knowledgebase reflected in current metabolite databases. Overall, we anticipate significant technological advancements in all high-throughput measurement techniques particularly in the area of long-read sequencing, chromatography as well as mass spectrometry. Naturally, these technological improvements will be complemented by equally sophisticated *in silico* data processing and analysis methods, which in turn will allow integrated omics to provide comprehensive multi-level snapshots of microbial population structures and functions *in situ* (Fig. 1; step 3).

In our opinion, the real power of the integrated omics approach within the Eco-Systems Biology framework will stem from applying the approach to temporally and spatially resolved samples (Fig. 1, steps 1 to 4; Muller *et al*., 2013; Zarraonaindia *et al*., 2013). In combination with appropriate statistical and mathematical modelling methods, the deconvolution of the data will unveil unprecedented insights into the structure and function of microbial communities (Fig. 1; step 4; Muller *et al*., 2013; Segata *et al*., 2013; Zarraonaindia *et al*., 2013). Data mining, machine learning and/or modelling approaches will be useful for extracting features of interest, e.g. known and unknown populations/genes, and also to derive associations (or links) between desired features utilizing measures such as correlation, co-occurrence, mutual information and hyper-geometric overlap (Muller *et al*., 2013; Segata *et al*., 2013). Such associations may allow the prediction of gene functions using the concept of ‘guilt by association’ and interactions/dependencies between community members (Wolfe *et al*., 2005; Segata *et al*., 2013; Solomon *et al*., 2014). Biological wastewater treatment plants offer particularly exciting opportunities to link responses in community structure and function to fluctuating environmental conditions because of the relative ease of sampling and routine recording of metadata (Muller *et al*., 2013; Segata *et al*., 2013; Vanwonterghem *et al*., 2014). Systematic omic analyses of BWWT microbial communities may therefore uncover (i) the effect of physico-chemical parameters on the expression of specific genes or phenotypes and (ii) the linkage of unknown genes to specific metabolites as well as to both known and unknown community members. However, the derived associations will always be ‘mere’ hypotheses, which will require rigorous testing through targeted laboratory experiments (Fig. 1; step 5) and/or *in situ* perturbation experiments followed by additional omic measurements (Muller *et al*., 2013; Segata *et al*., 2013).

Moving beyond associations and hypotheses

Although integrated omics-based approaches are highly effective for large-scale analysis and formulation of hypotheses (including within the context of BWWT plant communities), these efforts are limited due to current high-throughput measurement methods (see previous section) and the reliance on *a priori* knowledge for both taxonomical and functional inferences (Röling *et al*., 2010). Hence, there is a need to validate newly generated hypotheses using full-scale plants, customized laboratory-based experiments, such as batch cultures, bioreactors or pilot plants (Fig. 1; step 5) and/or single-cell methods. Hypotheses may be tested using additional integrated omic datasets generated from ancillary samples (e.g. Muller, Pinel *et al*, 2014) by using molecular biology techniques such as heterologous gene expression (e.g. Wexler *et al*., 2005; Maixner *et al*., 2008) or single-cell approaches using microautoradiography-fluorescent in situ hybridisation (MAR-FISH), nano-scale secondary-ion mass spectrometry (nanoSIMS) and/or Raman spectroscopy (e.g. Huang *et al*., 2007; Lechene *et al*., 2007; Musat *et al*., 2012). Such a combination of technologies can be used to test hypotheses regarding (i) community dynamics, (ii) gene expression patterns/interactions, (iii) metabolite abundances, (iv) effect of physico-chemical factors on distinct microbial species and functionalities, (v) gene function associations between any of these. Identified patterns may be subsequently formulated as cues and can be used as input to facilitate knowledge-driven control of different microbial community structures and/or functions (Fig. 1; step 6). Thus, large-scale integrated omic analyses of *in situ* biological samples (section “Laboratory protocols, systematic measurements and *in silico* analyses”), coupled to carefully controlled laboratory experiments, will allow the effective elucidation of novel functions within BWWT plant microbial communities with potential biotechnological applications.

From Eco-Systems Biology to biotechnology

Knowledge of gene function, regulation and physiological potential derived from integrated omic data over different spatial and temporal scales holds great promise in harnessing the biotechnological potential of microbial consortia. In particular, advancements in integrated omics followed by hypothesis testing may generate new knowledge (Muller *et al*., 2013), which may for example be exploited in new approaches for the optimized production of biotechnologically relevant compounds under varying environmental conditions (Chen and Nielsen, 2013). The derived knowledge-base may further be used to fine-tune metabolic pathways at the transcriptional, translational and post-translational levels using the ever-expanding synthetic biology toolbox (Peralta-Yahya *et al*., 2012). Examples of possible future applications may include, for instance the bioengineering of fatty acid utilization and production for the production of biodiesel from ‘dirty’ mixed substrates, the engineering of different gene combinations for the production of various alcohols from mixed substrates (Lee *et al*., 2008) and the generation of hybrid processes by combining biological and chemical production steps resulting in new compounds that could serve as biofuels (Román-Leshkov *et al*., 2007). Through exploration of BWWT plant microbial consortia using integrated omics, we are therefore poised to unravel key functionalities, which will find applications in a whole range of different biotechnologies. In this context, integrated omics through facilitating direct linkages between genetic potential and final phenotype may become an essential tool in future bioprospecting. Therefore, in our opinion, integrated omics will become the standard means of analysing microbial consortia in the near future and will allow meta-omics to fulfil their promise for the comprehensive discovery of biotechnology-relevant microbial traits in natural consortia.

Conflict of interest

None declared.

References

Albertsen, M., Hansen, L.B.S., Saunders, A.M., Nielsen, P.H., and Nielsen, K.L. (2012) A metagenome of a full-scale microbial community carrying out enhanced biological phosphorus removal. *ISME J* **6:** 1094–1106.

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K.L., Tyson, G.W., and Nielsen, P.H. (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat Biotechnol* **31:** 533–538.

Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., *et al.* (2014) Binning metagenomic contigs by coverage and composition. *Nat Methods* **11:** 1–7.

Chen, Y., and Nielsen, J. (2013) Advances in metabolic pathway and strain engineering paving the way for sustainable production of chemical building blocks. *Curr Opin Biotechnol* **24:** 965–972.

Daims, H., Taylor, M.W., and Wagner, M. (2006) Wastewater treatment: a model system for microbial ecology. *Trends Biotechnol* **24:** 483–489.

Denef, V.J., Mueller, R.S., and Banfield, J.F. (2010) AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *ISME J* **4:** 599–610.

Hettich, R.L., Sharma, R., Chourey, K., and Giannone, R.J. (2012) Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. *Curr Opin Microbiol* **15:** 373–380.

Huang, W.E., Stoecker, K., Griffiths, R., Newbold, L., Daims, H., Whiteley, A.S., and Wagner, M. (2007) Raman-FISH: combining stable-isotope Raman spectroscopy and fluorescence in situ hybridization for the single cell analysis of identity and function. *Environ Microbiol* **9:** 1878–1889.

Laczny, C.C., Pinel, N., Vlassis, N., and Wilmes, P. (2014) Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Sci Rep* **4:** 4516: 1–12.

Lahens, N.F., Kavakli, I.H., Zhang, R., Hayer, K., Black, M.B., Dueck, H., *et al.* (2014) IVT-seq reveals extreme bias in RNA-sequencing. *Genome Biol* **15:** R86.

Lechene, C.P., Luyten, Y., McMahon, G., and Distel, D.L. (2007) Quantitative imaging of nitrogen fixation by individual bacteria within animal cells. *Science* **317:** 1563–1566.

Lee, S.K., Chou, H., Ham, T.S., Lee, T.S., and Keasling, J.D. (2008) Metabolic engineering of microorganisms for biofuels production: from bugs to synthetic biology to fuels. *Curr Opin Biotechnol* **19:** 556–563.

Maixner, F., Wagner, M., Lücker, S., Pelletier, E., Schmitz-esser, S., Hace, K., *et al.* (2008) Environmental genomics reveals a functional chlorite dismutase in the nitrite-oxidizing bacterium 'Candidatus Nitrospira defluvii'. *Environ Microbiol* **10:** 3043–3056.

Mocali, S., and Benedetti, A. (2010) Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. *Res Microbiol* **161:** 497–505.

Muller, E.E.L., Glaab, E., May, P., Vlassis, N., and Wilmes, P. (2013) Condensing the omics fog of microbial communities. *Trends Microbiol* **21:** 325–333.

Muller, E.E.L., Sheik, A.R., and Wilmes, P. (2014) Lipid-based biofuel production from wastewater. *Curr Opin Biotechnol* **30C:** 9–16.

Muller, E.E.L., Pinel, N., Laczny, C.C., Hoopmann, M.R., Narayanasamy, S., Lebrun, L.A., *et al.* (2014) Community integrated omics links the dominance of a microbial generalist to fine-tuned resource usage. *Nat Commun.* **5:** 5603; 1–10.

Musat, N., Foster, R., Vagner, T., Adam, B., and Kuypers, M.M.M. (2012) Detecting metabolic activities in single cells, with emphasis on nanoSIMS. *FEMS Microbiol Rev* **36:** 486–511.

Nielsen, H.B., Almeida, M., Juncker, A.S., Rasmussen, S., Li, J., Sunagawa, S., *et al.* (2014) Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* **32:** 1–11.

Peralta-Yahya, P.P., Zhang, F., del Cardayre, S.B., and Keasling, J.D. (2012) Microbial engineering for the production of advanced biofuels. *Nature* **488:** 320–328.

Román-Leshkov, Y., Barrett, C.J., Liu, Z.Y., and Dumesic, J.A. (2007) Production of dimethylfuran for liquid fuels from biomass-derived carbohydrates. *Nature* **447:** 982–985.

Röling, W.F.M., Ferrer, M., and Golyshin, P.N. (2010) Systems approaches to microbial communities and their functioning. *Curr Opin Biotechnol* **21:** 532–538.

Roume, H., Heintz-Buschart, A., Muller, E.E.L., and Wilmes, P. (2013a) Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. *Methods Enzymol* **531:** 219–236.

Roume, H., Muller, E.E.L., Cordes, T., Renaut, J., Hiller, K., and Wilmes, P. (2013b) A biomolecular isolation framework for eco-systems biology. *ISME J* **7:** 110–121.

Segata, N., Boernigen, D., Tickle, T.L., Morgan, X.C., Garrett, W.S., and Huttenhower, C. (2013) Computational meta'omics for microbial community studies. *Mol Syst Biol* **9:** 666; 1–15.

Sheik, A.R., Muller, E.E.L., and Wilmes, P. (2014) A hundred years of activated sludge: time for a rethink. *Front Microbiol* **5:** 47; 1–7.

Solomon, K.V., Haitjema, C.H., Thompson, D.A., and O'Malley, M.A. (2014) Extracting data from the muck: deriving biological insight from complex microbial communities and non-model organisms with next generation sequencing. *Curr Opin Biotechnol* **28C:** 103–110.

Vanwonterghem, I., Jensen, P.D., Ho, D.P., Batstone, D.J., and Tyson, G.W. (2014) Linking microbial community structure, interactions and function in anaerobic digesters using new molecular techniques. *Curr Opin Biotechnol* **27:** 55–64.

Wexler, M., Bond, P.L., Richardson, D.J., and Johnston, A.W.B. (2005) A wide host-range metagenomic library from a waste water treatment plant yields a novel alcohol/aldehyde dehydrogenase. *Environ Microbiol* **7:** 1917–1926.

Wilmes, P., Simmons, S.L., Denef, V.J., and Banfield, J.F. (2009) The dynamic genetic repertoire of microbial communities. *FEMS Microbiol Rev* **33:** 109–132.

Wolfe, C.J., Kohane, I.S., and Butte, A.J. (2005) Systematic survey reveals general applicability of ‘guilt-by-association’ within gene coexpression networks. *BMC Bioinformatics* **6:** 227; 1–10.

Zarraonaindia, I., Smith, D.P., and Gilbert, J.A. (2013) Beyond the genome: community-level analysis of the microbial world. *Biol Philos* **28:** 261–282.

Zhang, T., Shao, M.-F., and Ye, L. (2012) 454 pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. *ISME J* **6:** 1137–1147.

© 2015 The Authors. *Microbial Biotechnology* published by John Wiley & Sons Ltd and Society for Applied Microbiology

A.2 IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses.

Shaman Narayanasamy†, Yohan Jarosz†, Emilie E.L. Muller, Anna Heintz-Buschart, Malte Herold, Anne Kaysen, Cédric C. Laczny, Nicolàs Pinel, Patrick May, Paul Wilmes
2016
Genome Biology **17**: 260
DOI: 10.1186/s13059-016-1116-8

Contributions of author include:

- Coordination
- Analytical research design
- Software development
- Figure creation
- Data analysis and visualization
- Writing and revision of manuscript

---

†Co-first author

Narayanasamy *et al. Genome Biology* (2016) 17:260
DOI 10.1186/s13059-016-1116-8

Genome Biology

SOFTWARE Open Access

IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses

Shaman Narayanasamy[1†], Yohan Jarosz[1†], Emilie E. L. Muller[1,2], Anna Heintz-Buschart[1], Malte Herold[1], Anne Kaysen[1], Cédric C. Laczny[1,3], Nicolás Pinel[4,5], Patrick May[1] and Paul Wilmes[1*]

Abstract

Existing workflows for the analysis of multi-omic microbiome datasets are lab-specific and often result in sub-optimal data usage. Here we present IMP, a reproducible and modular pipeline for the integrated and reference-independent analysis of coupled metagenomic and metatranscriptomic data. IMP incorporates robust read preprocessing, iterative co-assembly, analyses of microbial community structure and function, automated binning, as well as genomic signature-based visualizations. The IMP-based data integration strategy enhances data usage, output volume, and output quality as demonstrated using relevant use-cases. Finally, IMP is encapsulated within a user-friendly implementation using Python and Docker. IMP is available at http://r3lab.uni.lu/web/imp/ (MIT license).

Keywords: Multi-omics data integration, Metagenomics, Metatranscriptomics, Microbial ecology, Microbiome, Reproducibility

Background

Microbial communities are ubiquitous in nature and govern important processes related to human health and biotechnology [1, 2]. A significant fraction of naturally occurring microorganisms elude detection and investigation using classic microbiological methods due to their unculturability under standard laboratory conditions [3]. The issue of unculturability is largely circumvented through the direct application of high-resolution and high-throughput molecular measurements to samples collected in situ [4–6]. In particular, the application of high-throughput next-generation sequencing (NGS) of DNA extracted from microbial consortia yields metagenomic (MG) data which allow the study of microbial communities from the perspective of community structure and functional potential [4–6]. Beyond metagenomics, there is also a clear need to obtain functional readouts in the form of other omics data. The sequencing of reverse transcribed RNA (cDNA) yields metatranscriptomic (MT) data, which provides information about gene expression and therefore allows a more faithful assessment of community function [4–6]. Although both MG and MT data allow unprecedented insights into microbial consortia, the integration of such multi-omic data is necessary to more conclusively link genetic potential to actual phenotype in situ [4, 6]. Given the characteristics of microbial communities and the resulting omic data types, specialized workflows are required. For example, the common practice of subsampling collected samples prior to dedicated biomolecular extractions of DNA, RNA, etc. has been shown to inflate variation, thereby hampering the subsequent integration of the individual omic datasets [7, 8]. For this purpose, specialized wet-lab methods which allow the extraction of concomitant DNA, RNA, proteins, and metabolites from single, unique samples were developed to ensure that the generated data could be directly compared across the individual omic levels [7, 8]. Although standardized and reproducible wet-lab methods have been developed for integrated omics of microbial communities, corresponding bioinformatic analysis workflows have yet to be formalized.

---

* Correspondence: paul.wilmes@uni.lu
†Equal contributors
[1] Luxembourg Centre for Systems Biomedicine, 7, avenue des Hauts-Fourneaux, Esch-sur-Alzette L-4362, Luxembourg
Full list of author information is available at the end of the article

© The Author(s). 2016 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Bioinformatic analysis methods for MG and MT NGS data can be broadly classified into reference-dependent or reference-independent (de novo) methods [5]. Reference-dependent methods are based on the alignment/mapping of sequencing reads onto isolate genomes, gene catalogs, or existing MG data. A major drawback of such methods is the large number of sequencing reads from uncultured species and/or divergent strains which are discarded during data analysis, thereby resulting in the loss of potentially useful information. For example, based on analyses of MG data from the human gut microbiome (arguably the best characterized microbial community in terms of culture-derived isolate genomes), approximately 43% of the data are typically not mappable to the available isolate genomes [9]. Conversely, reference-independent methodologies, such as approaches based on de novo assemblies, enable the retrieval of the actual genomes and/or potentially novel genes present in samples, thereby allowing more of the data to be mapped and exploited for analysis [4, 5, 10]. Furthermore, it has been demonstrated that the assembly of sequencing reads into longer contiguous sequences (contigs) greatly improves the taxonomic assignments and prediction of genes as opposed to their direct identification from short sequencing reads [11, 12]. Finally, de novo MG assemblies may be further leveraged by binning the data to resolve and retrieve population-level genomes, including those from hitherto undescribed taxa [13–21].

Given the advantages of reference-independent methods, a wide array of MG-specific assemblers such as IDBA-UD [22] and MEGAHIT [23] have been developed. Most MT data analyses involve reference-based [24–26] or MG-dependent analysis workflows [27–29]. A comparative study by Celaj et al. [12] demonstrated that reference-independent approaches for MT data analyses are also applicable using either specialized MT assemblers (e.g., IDBA-MT [12, 30]), MG assemblers (e.g., IDBA-UD [22, 30, 31] and MetaVelvet [12, 32]) or single-species transcriptome assemblers (e.g., Trinity [12, 33]). In all cases, the available assemblers are able to handle the uneven sequencing depths of MG and MT data. Although dedicated assembly methods have been developed for MG and MT data, formalized pipelines allowing the integrated use of both data types are not available yet.

Automated bioinformatic pipelines have so far been mainly developed for MG data. These include MOCAT [34] and MetAMOS [10], which incorporate the entire process of MG data analysis, ranging from preprocessing of sequencing reads, de novo assembly, and post-assembly analysis (read alignment, taxonomic classification, gene annotation, etc.). MOCAT has been used in large-scale studies such as those within the MetaHIT Consortium [35, 36], while MetAMOS is a flexible pipeline which allows customizable workflows [10]. Both pipelines use SOAPdenovo [37] as the default de novo assembler, performing single-length *k*mer-based assemblies which usually result in fragmented (low contiguity) assemblies with low gene coverage values [38].

Multi-omic analyses have already provided new insights into microbial community structure and function in various ecosystems. These include studies of the human gut microbiome [28, 39], aquatic microbial communities from the Amazon river [27], soil microbial communities [40, 41], production-scale biogas plants [29], hydrothermal vents [42], and microbial communities from biological wastewater treatment plants [43, 44]. These studies employed differing ways for analyzing the data, including reference-based approaches [27, 28, 42], MG assembly-based approaches [29, 40], MT assembly-based approaches [42], and integrated analyses of the meta-omic data [39, 42–44]. Although these studies clearly demonstrate the power of multi-omic analyses by providing deep insights into community structure and function, standardized and reproducible computational workflows for integrating and analyzing the multi-omic data have so far been unavailable. Importantly, such approaches are, however, required to compare results between different studies and systems of study.

Due to the absence of established tools/workflows to handle multi-omic datasets, most of the aforementioned studies utilized non-standardized, ad hoc analyses, mostly consisting of custom workflows, thereby creating a challenge in reproducing the analyses [10, 45–47]. Given that the lack of reproducible bioinformatic workflows is not limited to those used for the multi-omic analysis of microbial consortia [10, 45–47], several approaches have recently been developed with the explicit aim of enhancing software reproducibility. These include a wide range of tools for constructing bioinformatic workflows [48–50] as well as containerizing bioinformatic tools/pipelines using Docker [29, 46–48].

Here, we present IMP, the Integrated Meta-omic Pipeline, the first open source de novo assembly-based pipeline which performs standardized, automated, flexible, and reproducible large-scale integrated analysis of combined multi-omic (MG and MT) datasets. IMP incorporates robust read preprocessing, iterative co-assembly of metagenomic and metatranscriptomic data, analyses of microbial community structure and function, automated binning, as well as genomic signature-based visualizations. We demonstrate the functionalities of IMP by presenting the results obtained on an exemplary data set. IMP was evaluated using datasets from ten different microbial communities derived from three distinct environments as well as a simulated mock microbial community dataset. We compare the assembly and data integration measures of IMP against standard MG analysis strategies (reference-based and reference-independent) to demonstrate that IMP vastly improves overall data usage. Additionally, we benchmark our assembly procedure against available MG analysis pipelines to show that IMP consistently produces high-quality assemblies across all the processed datasets. Finally, we describe a number of particular use cases which highlight biological applications of the IMP workflow.

Results

Overview of the IMP implementation and workflow

IMP leverages Docker for reproducibility and deployment. The interfacing with Docker is facilitated through a user-friendly Python wrapper script (see the "Details of the IMP implementation and workflow" section). As such, Python and Docker are the only prerequisites for the pipeline, enabling an easy installation and execution process. Workflow implementation and automation is achieved using Snakemake [49, 51]. The IMP workflow can be broadly divided into five major parts: i) preprocessing, ii) assembly, iii) automated binning, iv) analysis, and v) reporting (Fig. 1).

The preprocessing and filtering of sequencing reads is essential for the removal of low quality bases/reads, and potentially unwanted sequences, prior to assembly and analysis. The input to IMP consists of MG and MT (the latter preferably depleted of ribosomal RNA prior to sequencing) paired-end reads in FASTQ format (section "Input data"). MG and MT reads are preprocessed independently of each other. This involves an initial quality control step (Fig. 1 and section "Trimming and quality filtering") [52] followed by an optional screening for host/contaminant sequences, whereby the default screening is performed against the human genome while other host genome/contaminant sequences may also be used (Fig. 1 and section "Screening host or contaminant sequences"). In silico rRNA sequence depletion is exclusively applied to MT data (Fig. 1 and section "Ribosomal RNA filtering").

The customized assembly procedure of IMP starts with an initial assembly of preprocessed MT reads to generate an initial set of MT contigs (Additional file 1: Figure S1). MT reads unmappable to the initial set of MT contigs undergo a second round of assembly. The process of assembling unused reads, i.e., MG or MT reads unmappable to the previously assembled contigs, is henceforth referred to as "iterative assembly". The assembly of MT reads is performed, first as transcribed regions are covered much more deeply and evenly in MT data. The resulting MT-based contigs represent high-quality scaffolds for the subsequent co-assembly with MG data, overall leading to enhanced assemblies [43]. Therefore, the combined set of MT contigs from the initial and iterative MT assemblies are used to enhance the subsequent assembly with the MG data. MT data are assembled using the MEGAHIT de novo assembler using the appropriate option to prevent the merging of bubbles within the de Bruijn assembly graph [23, 36]. Subsequently, all preprocessed MT and MG reads, together with the generated MT contigs, are used as input to perform a first co-assembly, producing a first set of co-assembled contigs. The MG and MT reads unmappable to this first set of co-assembled contigs then undergo an additional iterative co-assembly step. IMP implements two assembler options for the de novo co-assembly step, namely IDBA-UD or MEGAHIT. The contigs resulting from the co-assembly procedure undergo a subsequent assembly refinement step by a contig-level assembly using the cap3 [53] de novo assembler. This aligns highly similar contigs against each other, thus reducing overall redundancy by collapsing shorter contigs into longer contigs and/or improving contiguity by extending contigs via overlapping contig ends (Additional file 1: Figure S1). This step produces the final set of contigs. Preprocessed MG and MT reads are then mapped back against the final contig set and the resulting alignment information is used in the various downstream analysis procedures (Fig. 1). In summary, IMP employs four measures for the de novo assembly of preprocessed MG and MT reads, including: i) iterative assemblies of unmappable reads, ii) use of MT contigs to scaffold the downstream assembly of MG data, iii) co-assembly of MG and MT data, and iv) assembly refinement by contig-level assembly. The entire de novo assembly procedure of IMP is henceforth referred to as the "IMP-based iterative co-assembly" (Additional file 1: Figure S1).

Contigs from the IMP-based iterative co-assembly undergo quality assessment as well as taxonomic annotation [54] followed by gene prediction and functional annotation [55] (Fig. 1 and section "Annotation and assembly quality assessment"). MaxBin 2.0 [20], an automated binning procedure (Fig. 1 and section "Automated binning") which performs automated binning on assemblies produced from single datasets, was chosen as the de facto binning procedure in IMP. Experimental designs involving single coupled MG and MT datasets are currently the norm. However, IMP's flexibility does not forego the implementation of multi-sample binning algorithms such as CONCOCT [16], MetaBAT [18], and canopy clustering [15] as experimental designs evolve in the future.

Non-linear dimensionality reduction of the contigs' genomic signatures (Fig. 1 and section "Non-linear dimensionality reduction of genomic signatures") is performed using the Barnes-Hut Stochastic Neighborhood Embedding (BH-SNE) algorithm allowing visualization of the data as two-dimensional scatter plots (henceforth referred to as VizBin maps [13, 56]). Further analysis steps include, but are not limited to, calculations of the contig- and gene-level depths of coverage (section

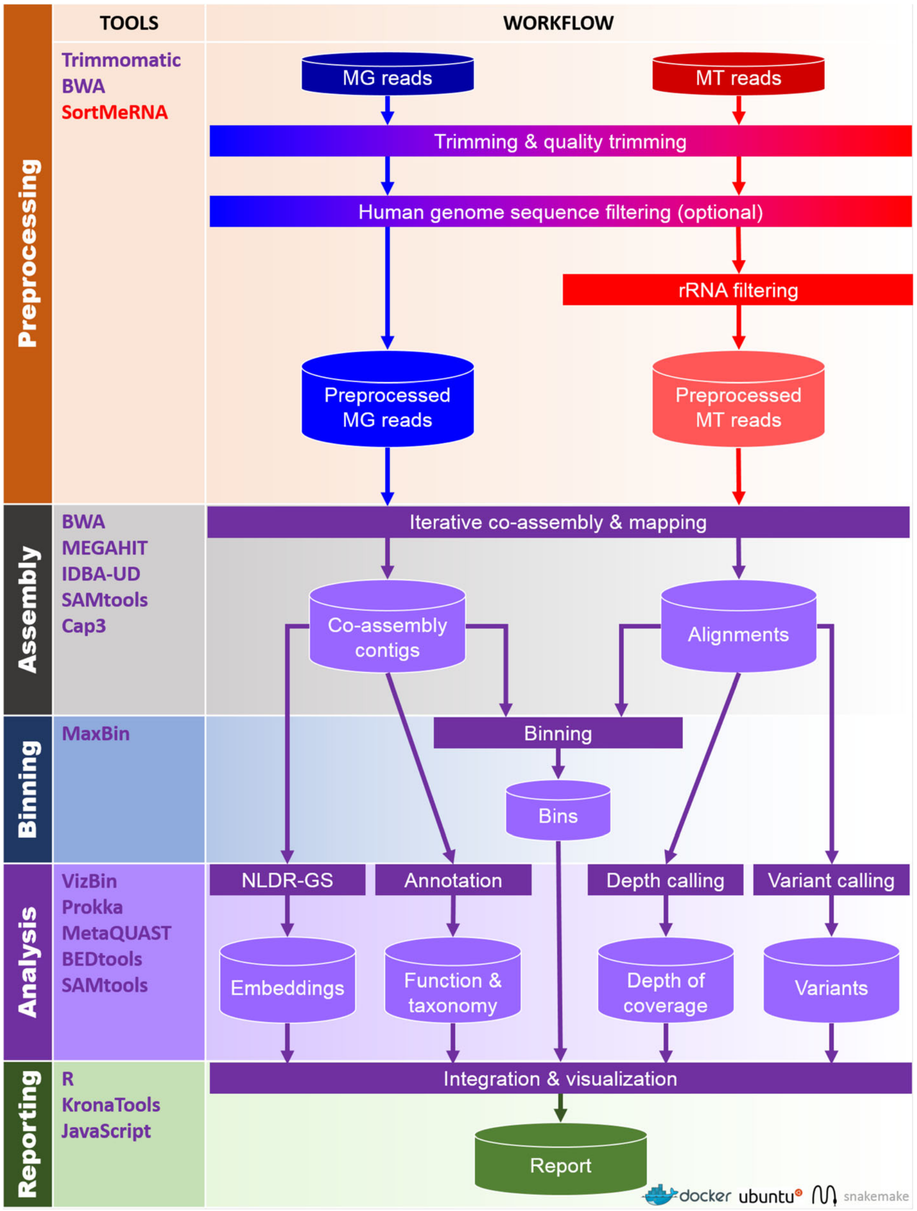


Fig. 1 Schematic overview of the IMP pipeline. *Cylinders* represent input and output while *rectangles* represent processes. *Arrows* indicate the flow between input, processes, and output. *MG* — Metagenomic data, *MT* — Metatranscriptomic data, *rRNA* — ribosomal RNA, *NLDR-GS* — genomic signature non-linear dimensionality reduction. Processes, input, and output specific to MG and MT data are labeled in *blue* and *red*, respectively. Processes and output that involve usage of both MG and MT data are represented in *purple*. A detailed illustration of the "iterative co-assembly" is available in Additional file 1: Figure S1

"Depth of coverage") as well as the calling of genomic variants (variant calling is performed using two distinct variant callers; section "Variant calling"). The information from these analyses are condensed and integrated into the generated VizBin maps to produce augmented visualizations (sections "Visualization and reporting"). These visualizations and various summaries of the output are compiled into a HTML report (examples of the HTML reports available via Zenodo [57]).

Exemplary output of IMP (using the default IDBA-UD assembler) based on a human fecal microbiome dataset is summarized in Fig. 2. The IMP output includes taxonomic (Fig. 2a) and functional (Fig. 2b, c) overviews. The representation of gene abundances at the MG and MT levels enables comparison of potential (Fig. 2b) and actual expression (Fig 2c) for specific functional gene categories (see Krona charts within HTML S1 [57]). IMP provides augmented VizBin maps [13, 56], including, for

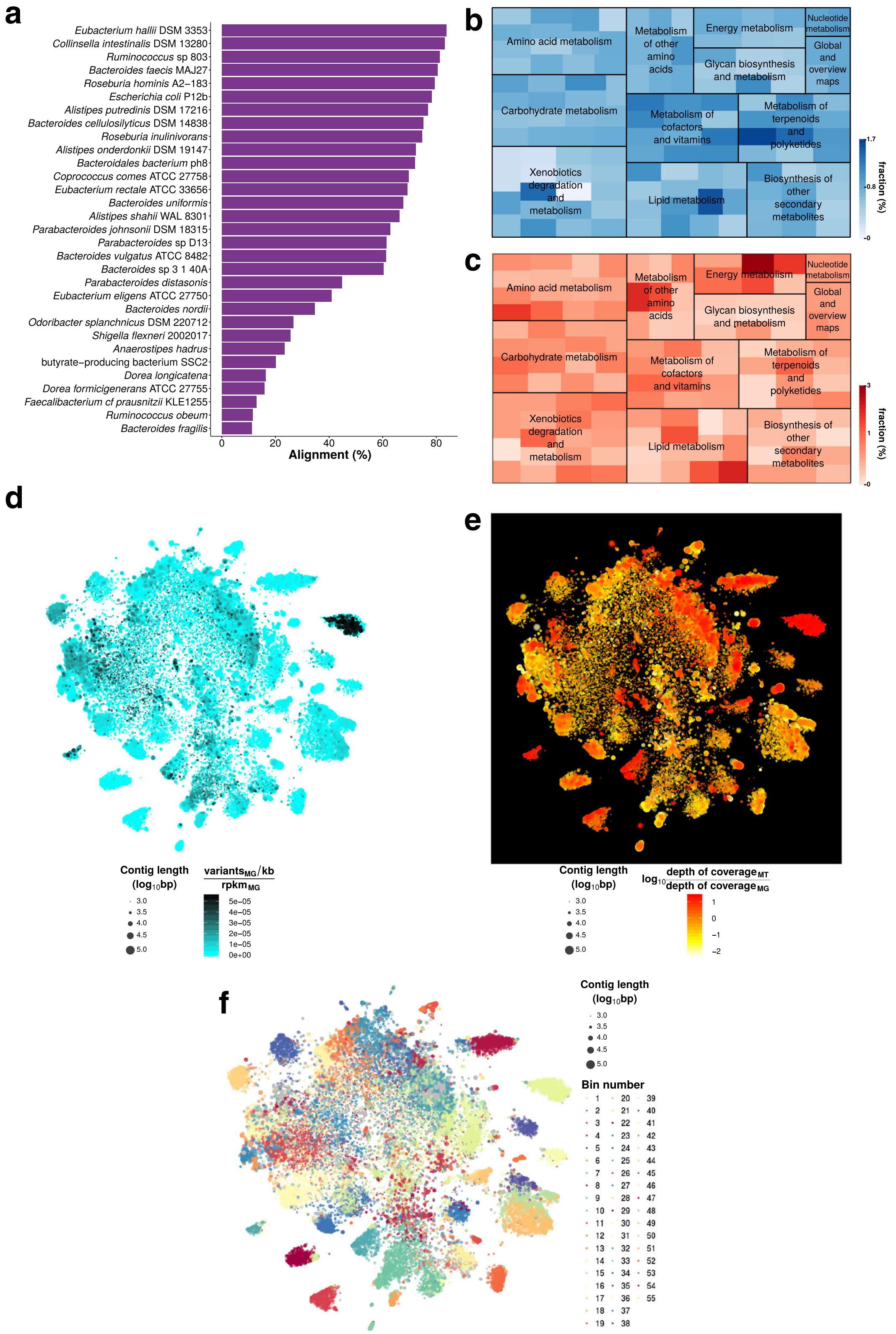


Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 Example output from the IMP analysis of a human microbiome dataset (HF1). **a** Taxonomic overview based on the alignment of contigs to the most closely related genomes present in the NCBI genome database (see also HTML report S1 [57]). **a**, **b** Abundances of predicted genes (based on average depths of coverage) of various KEGG Ontology categories represented both at the MG (**b**) and MT (**c**) levels (see also Krona charts within HTML report S1). **d**–**f** Augmented VizBin maps of contigs ≥1 kb, representing contig-level MG variant densities (**d**), contig-level ratios of MT to MG average depth of coverage (**e**), and bins generated by the automated binning procedure (**f**). Please refer to the HTML reports [57] for additional examples

example, variant densities (Fig. 2d) as well as MT to MG depth of coverage ratios (Fig. 2e). These visualizations may aid users in highlighting subsets of contigs based on certain characteristics of interest, i.e., population heterogeneity/homogeneity, low/high transcriptional activity, etc. Although an automated binning method [20] is incorporated within IMP (Fig. 2f), the output is also compatible with and may be exported to other manual/interactive binning tools such as VizBin [56] and Anvi'o [17] for additional manual curation. Please refer to the HTML reports for additional examples [57].

The modular design (section "Automation and modularity") and open source nature of IMP allow for customization of the pipeline to suit specific user-defined analysis requirements (section "Customization and further development"). As an additional feature, IMP also allows single-omic MG or MT analyses (section "Details of the IMP implementation and workflow"). Detailed parameters for the processes implemented in IMP are described in the section "Details of the IMP implementation and workflow" and examples of detailed workflow schematics are provided within the HTML reports [57].

Assessment and benchmarking

IMP was applied to ten published coupled MG and MT datasets, derived from three types of microbial systems, including five human fecal microbiome samples (HF1, HF2, HF3, HF4, HF5) [28], four wastewater sludge microbial communities (WW1, WW2, WW3, WW4) [43, 44], and one microbial community from a production-scale biogas (BG) plant [29]. In addition, a simulated mock (SM) community dataset based on 73 bacterial genomes [12], comprising both MG and MT data was generated to serve as a means for ground truth-based assessment of IMP (details in section "Coupled metagenomic and metatranscriptomic datasets"). The SM dataset was devised given the absence of a standardized benchmarking dataset for coupled MG and MT data (this does solely exist for MG data as part of the CAMI initiative (http://www.cami-challenge.org)).

Analysis with IMP was carried out with the two available de novo assembler options for the co-assembly step (Fig. 1; Additional file 1: Figure S1), namely the default IDBA-UD assembler [22] (hereafter referred to as IMP) and the optional MEGAHIT assembler [23] (henceforth referred to as IMP-megahit). IMP was quantitatively assessed based on resource requirement and analytical capabilities. The analytical capabilities of IMP were evaluated based on data usage, output volume, and output quality. Accordingly, we assessed the advantages of the iterative assembly procedure as well as the overall data integration strategy.

Resource requirement and runtimes

IMP is an extensive pipeline that utilizes both MG and MT data within a reference-independent (assembly-based) analysis framework which renders it resource- and time-intensive. Therefore, we aimed to assess the required computational resource and runtimes of IMP.

All IMP-based runs on all datasets were performed on eight compute cores with 32 GB RAM per core and 1024 GB of total memory (section "Computational platforms"). IMP runtimes ranged from approximately 23 h (HF1) to 234 h (BG) and the IMP-megahit runtimes ranged from approximately 21 h (HF1) up to 281 h (BG). IMP was also executed on the Amazon cloud computing (AWS) infrastructure, using the HF1 dataset on a machine with 16 cores (section "Computational platforms") whereby the run lasted approximately 13 h (refer to Additional file 1: Note S1 for more details). The analysis of IMP resulted in an increase in additional data of around 1.2–3.6 times the original input (Additional file 2: Table S1). Therefore, users should account for the disc space for both the final output and intermediate (temporary) files generated during an IMP run. Detailed runtimes and data generated for all the processed data sets are reported in Additional file 2: Table S1.

We further evaluated the effect of increasing resources using a small scale test dataset (section "Test dataset for runtime assessment"). The tests demonstrated that reduced runtimes are possible by allocating more threads to IMP-megahit (Additional file 2: Table S2). However, no apparent speed-up is achieved beyond allocation of eight threads, suggesting that this would be the optimal number of threads for this particular test dataset. Contrastingly, no speed-up was observed with additional memory allocation (Additional file 2: Table S3). Apart from the resources, runtime may also be affected by the input size, the underlying complexity of the dataset and/or behavior of individual tools within IMP.

Data usage: iterative assembly

De novo assemblies of MG data alone usually result in a large fraction of reads that are unmappable to the assembled contigs and therefore remain unused, thereby leading to suboptimal data usage [43, 58–60]. Previous studies have assembled sets of unmappable reads iteratively to successfully obtain additional contigs, leading to an overall increase in the number of predicted genes, which in turn results in improved data usage [43, 58–60]. Therefore, IMP uses an iterative assembly strategy to maximize NGS read usage. In order to evaluate the best iterative assembly approach for application within the IMP-based iterative co-assembly strategy, we attempted to determine the opportune number of assembly iterations in relation to assembly quality metrics and computational resources/runtimes.

The evaluation of the iterative assembly strategy was applied to MG and MT datasets. For both omic data types, it involved an "initial assembly" which is defined as the de novo assembly of all preprocessed reads. Additional iterations of assembly were then conducted using the reads that remained unmappable to the generated set of contigs (see section "Iterative single-omic assemblies" for details and parameters). The evaluation of the iterative assembly procedure was carried out based on the gain of additional contigs, cumulative contig length (bp), numbers of genes, and numbers of reads mappable to contigs. Table 1 shows the evaluation results of four representative data sets and Additional file 2: Table S4 shows the detailed results of the application of the approach to 11 datasets. In all the datasets evaluated, all iterations (1 to 3) after the initial assembly lead to an increase in total length of the assembly and numbers of mappable reads (Table 1; Additional file 2: Table S4). However, there was a notable decline in the number of additional contigs and predicted genes beyond the first iteration. Specifically, the first iteration of the MG assembly yielded up to 1.6% additional predicted genes while the equivalent on the MT data yielded up to 9% additional predicted genes (Additional file 2: Table S4). Considering the small increase (<1%) in the number of additional contigs and predicted genes beyond the first assembly iteration on one hand and the extended runtimes required to perform additional assembly iterations on the other hand, a generalized single iteration assembly approach was retained and implemented within the IMP-based iterative co-assembly (Fig. 1; Additional file 1: Figure S1). This approach aims to maximize data usage without drastically extending runtimes.

Despite being developed specifically for the analysis of coupled MG and MT datasets, the iterative assembly can also be used for single omic datasets. To assess IMP's performance on MG datasets, it was applied to the simulated MG datasets from the CAMI challenge (http://www.cami-challenge.org) and the results are shown in Additional file 1: Figure S2. IMP-based MG assembly using the MEGAHIT assembler on the CAMI dataset outperforms well-established MG pipelines such

Table 1 Statistics of iterative assemblies performed on MG and MT datasets

Dataset	Iteration	MG iterative assembly				MT iterative assembly			
		Number of contigs (≥1 kb)	Cumulative length of assembled contigs (bp)	Number of predicted genes	Number of mapped reads	Number of contigs (all)	Cumulative length of assembled contigs (bp)	Number of predicted genes	Number of mapped reads
SM	Initial assembly	29063	182673343	186939	18977716	13436	8994518	13946	822718
	1	16	483336	329	9515	1286	502535	1272	16038
	2	6	213094	126	3425	48	18460	49	656
	3	1	86711	47	1536	0	0	0	0
HF1	Initial assembly	27028	145938650	154760	20715368	40989	45300233	66249	17525586
	1	15	966872	274	39839	2471	969614	2238	329400
	2	−1	26822	5	1276	26	10315	24	45642
	3	0	4855	0	172	3	1640	6	54788
WW1	Initial assembly	14815	77059275	81060	6513708	45118	22525759	49859	8423603
	1	28	3146390	1136	73511	2115	723904	1589	529441
	2	2	175634	114	4031	250	82048	201	13335
	3	1	30032	16	572	31	10280	18	65866
BG	Initial assembly	105282	545494441	593688	109949931	47628	27493690	60566	3754432
	1	417	10998269	3902	456821	3956	1397409	3061	130131
	2	5	335313	219	21647	717	250223	754	12766
	3	7	79022	20	2511	24	9060	22	5827

Results for all datasets available in Additional file 2: Table S2 as MOCAT in all measures. In addition, IMP-based iterative assemblies also exhibit comparable performance to the gold standard assembly with regards to contigs ≥1 kb and number of predicted genes (http://www.cami-challenge.org). Detailed results of the CAMI assemblies are available in Additional file 2: Table S5. However, as no MT and/or coupled MG and MT datasets so far exist for the CAMI challenge, the full capabilities of IMP could not be assessed in relation to this initiative.

Data usage: multi-omic iterative co-assembly

In order to assess the advantages of integrated multi-omic co-assemblies of MG and MT data, IMP-based iterative co-assemblies (IMP and IMP-megahit) were compared against MG-only-based assemblies which include single-omic iterative MG assemblies generated using IMP (referred to as IMP_MG) and standard MG assemblies by MOCAT (hereafter referred to as MOCAT_MG) and MetAMOS (hereafter referred to as MetAMOS_MG). Furthermore, the available reads from the human fecal microbiome dataset (preprocessed with IMP) were mapped to the MetaHIT Integrated Gene Catalog (IGC) reference database [35] to compare the data usage of the different assembly procedures against a reference-dependent approach.

IMP-based iterative co-assemblies consistently recruited larger fractions of properly paired MG (Fig. 3a) and/or MT (Fig. 3b) reads compared to single-omic

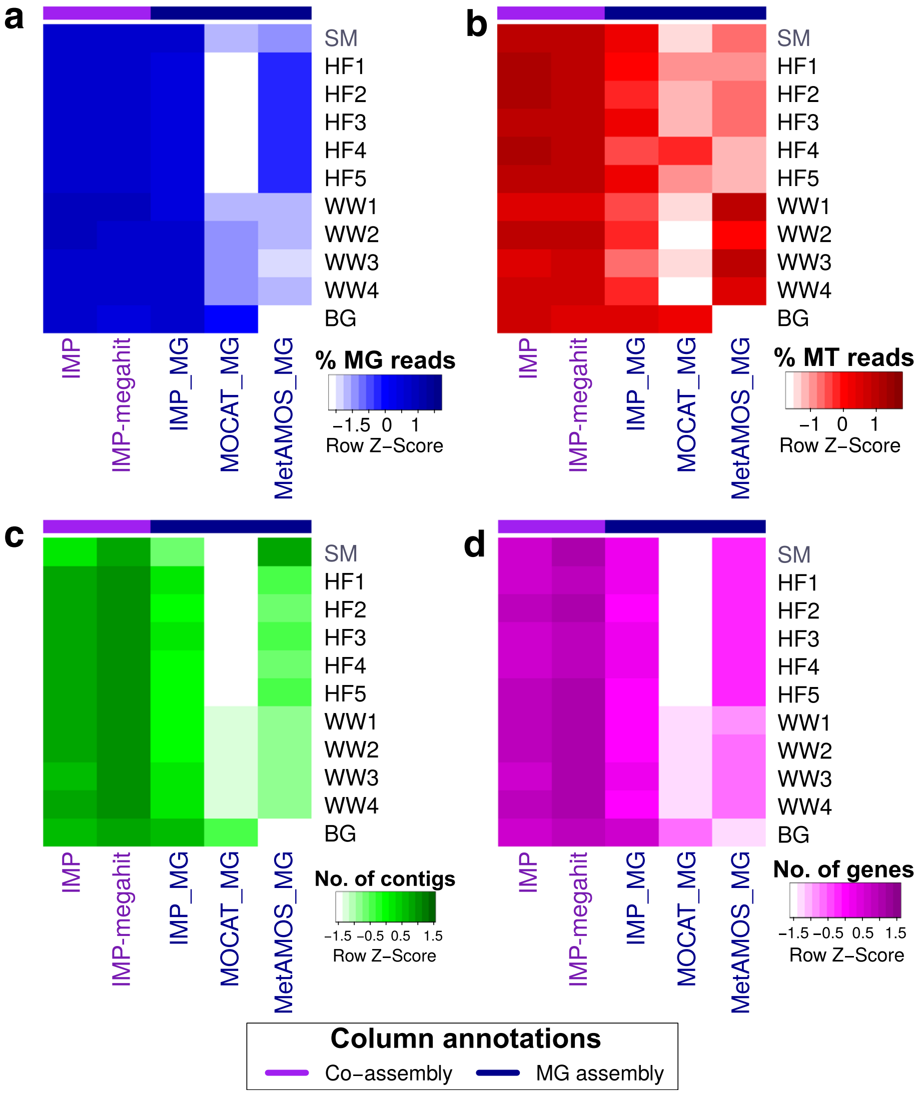


Fig. 3 Assessment of data usage and output generated from co-assemblies compared to single-omic assemblies. Heat maps show (**a**) fractions of properly mapped MG read pairs, (**b**) fractions of properly mapped MT read pairs, (**c**) numbers of contigs ≥1 kb, and (**d**) numbers of unique predicted genes. IMP and IMP-megahit represent integrated multi-omic MG and MT iterative co-assemblies while IMP_MG, MOCAT_MG, and MetAMOS_MG represent single-omic MG assemblies. All numbers were row Z-score normalized for visualization. Detailed results available in Additional file 2: Table S5

assemblies. The resulting assemblies also produced larger numbers of contigs ≥1 kb (Fig. 3c), predicted non-redundant unique genes (Fig. 3d), and, even more important, complete genes as predicted with start and stop codon by Prodigal [61] (Additional file 2: Table S5). Using the reference genomes from the SM data as ground truth, IMP-based iterative co-assemblies resulted in up to 25.7% additional recovery of the reference genomes compared to the single-omic MG assemblies (Additional file 2: Table S5).

IMP-based iterative co-assemblies of the human fecal microbiome datasets (HF1–5) allowed recruitment of comparable fractions of properly paired MG reads and an overall larger fraction of properly paired MT reads compared to those mapping to the IGC reference database (Table 2). The total fraction (union) of MG or MT reads mapping to either IMP-based iterative co-assemblies and/or the IGC reference database was higher than 90%, thus demonstrating that the IMP-based iterative co-assemblies allow at least 10% of additional data to be mapped when using these assemblies in addition to the IGC reference database. In summary, the complementary use of de novo co-assembly of MG and MT datasets in combination with iterative assemblies enhances overall MG and MT data usage and thereby significantly increases the yield of useable information, especially when combined with comprehensive reference catalogs such as the IGC reference database.

Assembly quality: multi-omic iterative co-assembly

In order to compare the quality of the IMP-based iterative co-assembly procedure to simple co-assemblies, we compared the IMP-based iterative co-assemblies against co-assemblies generated using MetAMOS [10] (henceforth referred to as MetAMOS_MGMT) and MOCAT [34] (henceforth referred to as MOCAT_MGMT). Although MetAMOS and MOCAT were developed for MG data analysis, we extended their use for obtaining MG and MT co-assemblies by including both MG and MT read libraries as input (section "Execution of pipelines"). The assemblies were assessed based on contiguity (N50 length), data usage (MG and MT reads mapped), and output volume (number of contigs above 1 kb and number of genes; Additional file 2: Table S5). Only the SM dataset allowed for ground truth-based assessment by means of aligning the generated de novo assembly contigs to the original 73 bacterial genomes used to simulate the data set (section "Simulated coupled metagenomic and metatranscriptomic dataset") [12, 54]. This allowed the comparison of two additional quality metrics, i.e., the recovered genome fraction and the composite performance metric (CPM) proposed by Deng et al. [62].

Table 2 Mapping statistics for human microbiome samples

Reference	Average MG pairs mapping (%)	Average MT pairs mapping (%)
IGC	70.91	53.57
IMP	70.25	86.21
IMP-megahit	70.62	83.33
IMP_MG	68.08	58.54
MetAMOS_MG	57.31	37.34
MOCAT_MG	36.73	36.68
IMP + IGC	92.66	95.77
IMP-megahit + IGC	92.80	93.24

Average fractions (%) of properly paired reads from the human microbiome datasets (HF1–5) mapping to various references, including IMP-based iterative co-assemblies (IMP and IMP-megahit) and single-omic co-assemblies (IMP_MG, MetAMOS_MG, and MOCAT_MG) as well as the IGC reference database. IMP + IGC and IMP-megahit + IGC reports the total number of properly paired reads mapping to IMP-based iterative co-assemblies and/or the IGC reference database. Refer to Additional file 2: Table S3 for detailed information

Assessments based on real datasets demonstrate comparable performance between IMP and IMP-megahit while both outperform MetAMOS_MGMT and MOCAT_MGMT in all measures (Fig. 4a–c). The ground truth assessment using the SM dataset shows that IMP-based iterative co-assemblies are effective in recovering the largest fraction of the original reference genomes while achieving a higher CPM score compared to co-assemblies from the other pipelines. Misassembled (chimeric) contigs are a legitimate concern within extensive de novo assembly procedures such as the IMP-based iterative co-assembly. It has been previously demonstrated that highly contiguous assemblies (represented by high N50 lengths) tend to contain higher absolute numbers of misassembled contigs compared to highly fragmented assemblies, thereby misrepresenting the actual quality of assemblies [38, 62, 63]. Therefore, the CPM score was devised as it represents a normalized measure reflecting both contiguity and accuracy for a given assembly [62]. Based on the CPM score, both IMP and IMP-megahit yield assemblies that balance high contiguity with accuracy and thereby outperform the other methods (Fig. 4c, d). In summary, cumulative measures of numbers of contigs ≥1 kb, N50 lengths, numbers of unique genes, recovered genome fractions (%), and CPM scores (the latter two were only calculated for the SM dataset), as well as the mean fractions (%) of mappable MG and MT reads, show that the IMP-based iterative co-assemblies (IMP and IMP-megahit) clearly outperform all other available methods (Fig. 4e; Additional file 2: Table S5).

Use-cases of integrated metagenomic and metatranscriptomic analyses in IMP

The integration of MG and MT data provides unique opportunities for uncovering community- or population-specific traits, which cannot be resolved from MG or MT data alone. Here we provide two examples of

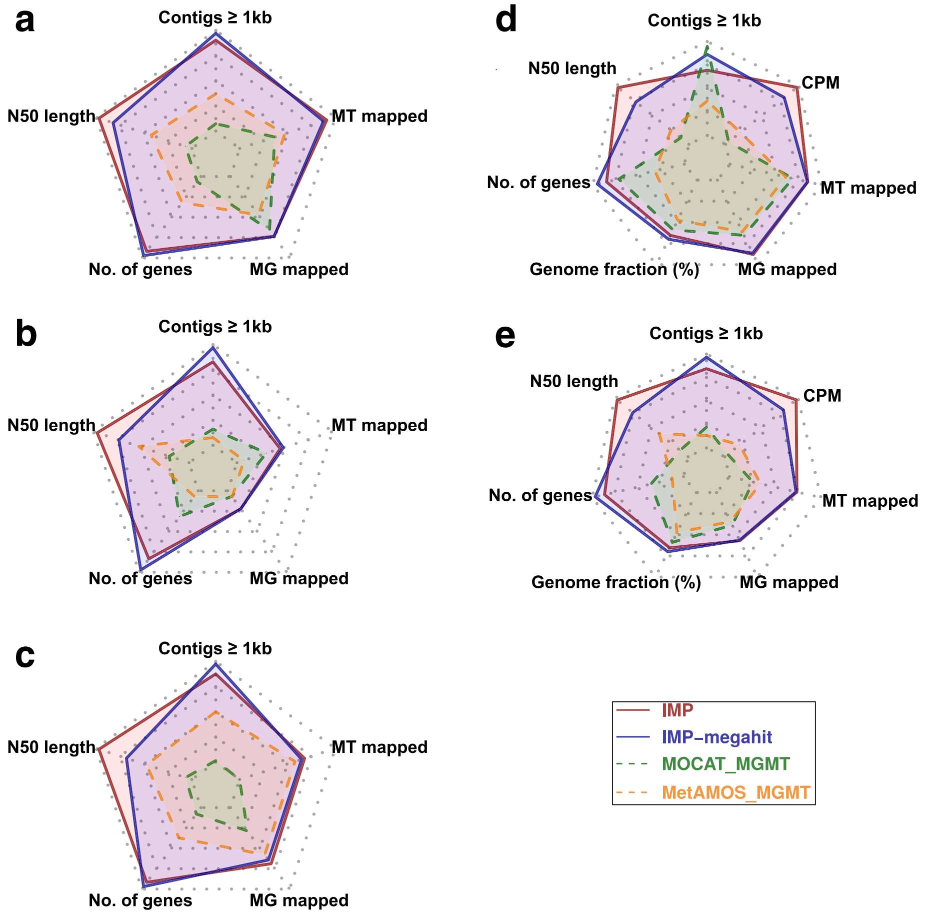


Fig. 4 Assessment of the IMP-based iterative co-assemblies in comparison to MOCAT- and MetAMOS-based co-assemblies. Radar charts summarizing the characteristics of the co-assemblies generated using IMP, MetAMOS, and MOCAT pipelines on: **a** human fecal microbiome, **b** wastewater sludge community, **c** biogas reactor, **d** simulated mock community. IMP co-assemblies were performed with two de novo assembler options, IDBA_UD and MEGAHIT, whereas MetAMOS and MOCAT were executed using default settings. Assessment metrics within the radar charts include number of contigs ≥1 kb, N50 length (contiguity, cutoff 500 bp), number of predicted genes (unique), and fraction of properly mapped MG and MT read pairs. N50 statistics are reported using a 500-bp cutoff. Additional ground truth assessments for simulated mock dataset included recovered genome fractions (%) and the composite performance metric (CPM) score with a cutoff of 500 bp [62]. **e** Summary radar chart reflecting the cumulative measures and mean fraction of properly mapped MG and MT read pairs from all analyzed 11 datasets while incorporating ground truth-based measures from the simulated mock dataset. Higher values within the radar charts (furthest from center) represent better performance. Detailed information on the assembly assessments is available in Additional file 2: Table S5

insights gained through the direct inspection of results provided by IMP.

Tailored preprocessing and filtering of MG and MT data

The preprocessing of the datasets HF1–5 included filtering of human-derived sequences, while the same step was not necessary for the non-human-derived datasets, WW1–4 and BG. MT data analyzed within this article included RNA extracts which were not subjected to wet-lab rRNA depletion, i.e., BG [29], and samples which were treated with wet-lab rRNA removal kits (namely HF1–5 [28] and WW1–4 [43]). Overall, the removal of rRNA pairs from the MT data showed a large variation, ranging from as low as 0.51% (HF5) to 60.91% (BG), demonstrating that wet-lab methods vary in terms of effectiveness and highlighting the need for such MT-specific filtering procedures (Additional file 1: Note S2; Additional file 2: Table S6).

Identification of RNA viruses

To identify differences in the information content of MG and MT complements, the contigs generated using IMP were inspected with respect to coverage by MG and MT reads (Additional file 2: Table S7). In two exemplary datasets HF1 and WW1, a small fraction of the contigs resulted exclusively from MT data (Additional file 2: Table S7). Longer contigs (≥1 kb) composed exclusively of MT reads and annotated with known viral/bacteriophage genes were retained for further inspection (Table 3; complete list contigs in Additional file 2: Table S8

Table 3 Contigs with a likely viral/bacteriophage origin/function reconstructed from the metatranscriptomic data

Sample	Contig ID*	Contig length	Average contig depth of coverage	Gene product	Average gene depth of coverage
HF1	Contig_34	6468	20927	Virus coat protein (TMV like)	30668
				Viral movement protein (MP)	26043
				RNA-dependent RNA polymerase	22578
				Viral methyltransferase	18817
	Contig_13948	2074	46	RNA-dependent RNA polymerase	41
				Viral movement protein (MP)	56
WW2	Contig_6405	4062	46	Tombusvirus p33	43
				Viral RNA-dependent RNA polymerase	42
				Viral coat protein (S domain)	36
	Contig_7409	3217	21	Viral RNA-dependent RNA polymerase	18
				Viral coat protein (S domain)	21
	Contig_7872	2955	77	Hypothetical protein	112
				Phage maturation protein	103

*Contigs of ≥1 kb and average depth of coverage ≥20 were selected

and S9). A subsequent sequence similarity search against the NCBI NR nucleotide database [64] of these candidate contigs revealed that the longer contigs represent almost complete genomes of RNA viruses (Additional file 2: Table S10 and S11). This demonstrates that the incorporation of MT data and their contrasting to the MG data allow the identification and recovery of nearly complete RNA viral genomes, thereby allowing their detailed future study in a range of microbial ecosystems.

Identification of populations with apparent high transcriptional activity

To further demonstrate the unique analytical capabilities of IMP, we aimed to identify microbial populations with a high transcriptional activity in the HF1 human fecal microbiome sample. Average depth of coverage at the contig- and gene-level is a common measure used to evaluate the abundance of microbial populations within communities [14, 16, 43]. The IMP-based integrative analysis of MG and MT data further extends this measure by calculation of average MT to MG depth of coverage ratios, which provide information on transcriptional activity and which can be visualized using augmented VizBin maps [56].

In our example, one particular cluster of contigs within the augmented VizBin maps exhibited high MT to MG depth of coverage ratios (Additional file 1: Figure S3). The subset of contigs within this cluster aligned to the genome of the *Escherichia coli* P12B strain (henceforth referred to as *E. coli*). For comparison, we also identified a subset, which was highly abundant at the MG level (lower MT to MG ratio), which aligned to the genome of *Collinsella intestinalis* DSM 13280 strain (henceforth referred to as *C. intestinalis*). Based on these observations, we highlighted the subsets of these contigs in an augmented VizBin map (Fig. 5a). The *C. intestinalis* and *E. coli* subsets are mainly represented by clear peripheral clusters which exhibit consistent intra-cluster MT to MG depth of coverage ratios (Fig. 5a). The subsets were manually inspected in terms of their distribution of average MG and MT depths of coverage and were compared against the corresponding distributions for all contigs. The MG-based average depths of coverage of the contigs from the entire community exhibited a bell-shape like distribution, with a clear peak (Fig. 5b). In contrast, MT depths of coverage exhibited more spread, with a relatively low mean (compared to MG distribution) and no clear peak (Fig. 5b). The *C. intestinalis* subset displays similar distributions to that of the entire community, whereas the *E. coli* subset clearly exhibits unusually high MT-based and low MG-based depths of coverage (Fig. 5b). Further inspection of the individual omic datasets revealed that the *E. coli* subset was not covered by the MG contigs, while approximately 80% of the *E. coli* genome was recoverable from a single-omic MT assembly (Fig. 5c). In contrast, the *C. intestinalis* subset demonstrated genomic recovery in all co-assemblies (IMP, IMP-megahit, MOCAT_MGMT, MetAMOS_MGMT) and the single-omic MG assemblies (IMP_MG, MOCAT_MG, MetAMOS_MG; Fig. 5c).

As noted by the authors of the original study by Franzosa et al. [28], the cDNA conversion protocol used to produce the MT data is known to introduce approximately 1–2% of *E. coli* genomic DNA into the cDNA as contamination which is then reflected in the MT data. According to our analyses, 0.12% of MG reads and

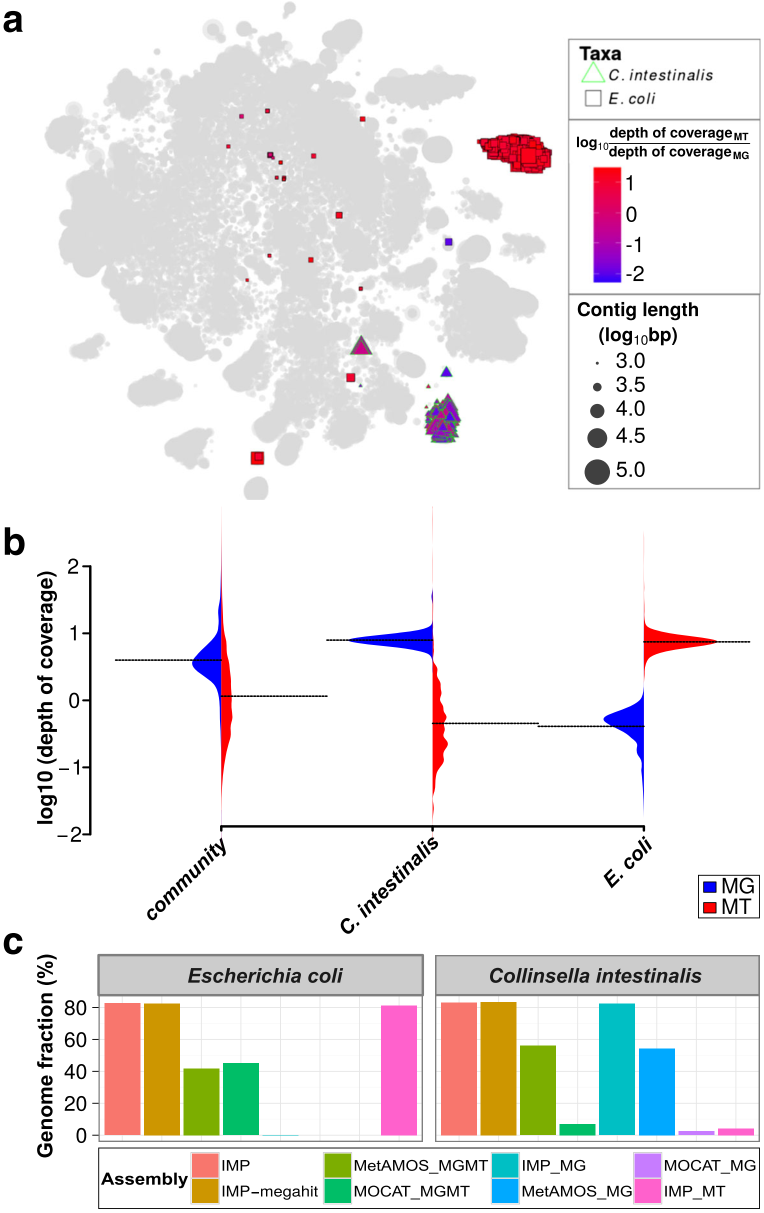


Fig. 5 Metagenomic and metatranscriptomic data integration of a human fecal microbiome. **a** Augmented VizBin map highlighting contig subsets with sequences that are most similar to *Escherichia coli* P12b and *Collinsella intestinalis* DSM 13280 genomes. **b** Beanplots representing the densities of metagenomic (*MG*) and metatranscriptomic (*MT*) average contig-level depth of coverage for the entire microbial community and two subsets (population-level genomes) of interest. The *dotted lines* represent the mean. **c** Recovered portion of genomes of the aforementioned taxa based on different single-omic assemblies and multi-omic co-assemblies (Additional file 2: Table S5)

1.95% of MT reads derived from this sample could be mapped onto the *E. coli* contigs, which is consistent with the numbers quoted by Franzosa et al. [28].

Consistent recovery of the *E. coli* genome was also observed across all other assemblies of the human fecal microbiome datasets (HF2–5) which included their respective MT data (Additional file 1: Figure S4; Additional file 2: Table S12). The integrative analyses of MG and MT data within IMP enables users to efficiently highlight notable cases such as this and to further investigate inconsistencies and/or interesting characteristics within these multi-omic datasets.

Discussion

The microbiome analysis workflow of IMP is unique in that it allows the integrated analysis of MG and MT data. To the best of our knowledge, IMP represents the only pipeline that spans the preprocessing of NGS reads to the binning of the assembled contigs, in addition to being the first automated pipeline for reproducible reference-independent metagenomic and metatranscriptomic data analysis. Although existing pipelines such as MetAMOS or MOCAT may be applied to perform co-assemblies of MG and MT data [44], these tools do not include specific steps for the two data types in their pre- and post-assembly procedures, which is important given the disparate nature of these datasets. The use of Docker promotes reproducibility and sharing, thereby allowing researchers to precisely replicate the IMP workflow with relative ease and with minimal impact on overall performance of the employed bioinformatic tools [29, 46–48]. Furthermore, static websites will be created and associated with every new version of IMP (Docker image), such that users will be able to download and launch specific versions of the pipeline to reproduce the work of others. Thereby, IMP enables standardized comparative studies between datasets from different labs, studies, and environments. The open source nature of IMP encourages a community-driven effort to contribute to and further improve the pipeline. Snakemake allows the seamless integration of Python code and shell (bash) commands and the use of *make* scripting style, which are arguably some of the most widely used bioinformatic scripting languages. Snakemake also supports parallel processing and the ability to interoperate with various tools and/or web services [49, 51]. Thus, users will be able to customize and enhance the features of the IMP according to their analysis requirements with minimal training/learning.

Quality control of NGS data prior to de novo assemblies has been shown to increase the quality of downstream assembly and analyses (predicted genes) [63]. In addition to standard preprocessing procedures (i.e., removal of low quality reads, trimming of adapter sequences and removal), IMP incorporates additional tailored and customizable filtering procedures which account for the different sample and/or omic data types. For instance, the removal of host-derived sequences in the context of human microbiomes is required for protecting the privacy of study subjects. The MT-specific in silico rRNA removal procedure yielded varying fractions of rRNA reads between the different MT datasets despite the previous depletion of rRNA (section “Tailored preprocessing and filtering of MG and MT data”), indicating that improvements in wet-lab protocols are necessary. Given that rRNA sequences are known to be highly similar, they are removed in IMP in order to mitigate any possible misassemblies resulting from such reads and/or regions [65, 66]. In summary, IMP is designed to perform stringent and standardized preprocessing of MG and MT data in a data-specific way, thereby enabling efficient data usage and resulting in high-quality output.

It is common practice that MG and MT reads are mapped against a reference (e.g., genes, genomes, and/or MG assemblies) [28, 29, 40] prior to subsequent data interpretation. However, these standard practices lead to suboptimal usage of the original data. IMP enhances overall data usage through its specifically tailored iterative co-assembly procedure, which involves four measures to achieve better data usage and yield overall larger volumes of output (i.e., a larger number of contigs ≥1 kb and predicted unique and complete genes).

First, the iterative assembly procedure leads to increases in data usage and output volume in each additional iterative assembly step (section “Data usage: iterative assembly”). The exclusion of mappable reads in each iteration of the assembly serves as a means of partitioning the data, thereby reducing the complexity of the data and overall, resulting in a higher cumulative volume of output [60, 63, 67].

Second, the initial assembly of MT-based contigs enhances the overall assembly, as transcribed regions are covered much more deeply and evenly in MT data, resulting in better assemblies for these regions [43]. The MT-based contigs represent high-quality scaffolds for the subsequent co-assembly with MG data.

Third, the co-assembly of MG and MT data allows the integration of these two data types while resulting in a larger number of contigs and predicted complete genes against which, in turn, a substantially higher fraction of reads can be mapped (section “Data usage: multi-omic iterative co-assembly”). Furthermore, the analyses of the human fecal microbiome datasets (HF1–5) demonstrate that the numbers of MG reads mapping to the IMP-based iterative co-assemblies for each sample are comparable to the numbers of reads mapping to the comprehensive IGC reference database (Table 2). Previously, only fractions of 74–81% of metagenomic reads mapping to the IGC have been reported [35]. However, such numbers have yet to be reported for MT data, in which case we observe lower mapping rates to the IGC reference database (35.5–70.5%) compared to IMP-based assemblies (Additional file 2: Table S3). This may be attributed to the fact that the IGC reference database was generated from MG-based assemblies only, thus creating a bias [35]. Moreover, an excess of 90% of MG and MT reads from the human fecal datasets (HF1–5) are mappable to either the IGC reference database and/or IMP-based iterative co-assemblies, emphasizing that a combined reference-based and IMP-based integrated-omics approach vastly improves data usage (Table 2). Although large fractions of MG and/or MT reads can be mapped to the IGC, a significant advantage of using a de novo reference-independent approach lies within the fact that reads can be linked to genes within their respective genomic context and microbial populations of origin.

Exploiting the maximal amount of information is especially relevant for microbial communities with small sample sizes and which lack comprehensive references such as the IGC reference database.

Fourth, the assembly refinement step via a contig-level assembly with cap3 improves the quality of the assemblies by reducing redundancy and increasing contiguity by collapsing and merging contigs (section "Assembly quality: multi-omic iterative co-assembly"). Consequently, our results support the described notion that the sequential use of multi-*k*mer-based de Bruijn graph assemblers, such as IDBA-UD and MEGAHIT, with overlap-layout-consensus assemblers, such as cap3, result in improved MG assemblies [38, 62] but importantly also extend this to MG and MT co-assemblies.

When compared to commonly used assembly strategies, the IMP-based iterative co-assemblies consisted of a larger output volume while maintaining a relatively high quality of the generated contigs. High-quality assemblies yield higher quality taxonomic information and gene annotations while longer contigs (≥1 kb) are a prerequisite for unsupervised population-level genome reconstruction [14, 19, 56] and subsequent multi-omics data integration [39, 43, 44]. Throughout all the different comparative analyses which we performed, IMP performed more consistently across all the different datasets when compared to existing methods, thereby emphasizing the overall stability and broad range of applicability of the method (section "Assembly quality: multi-omic iterative co-assembly").

Integrated analyses of MG and MT data with IMP provide the opportunity for analyses that are not possible based on MG data alone, such as the detection of RNA viruses (section "Identification of RNA viruses") and the identification of transcriptionally active populations (section "Identification of populations with apparent high transcriptional activity"). The predicted/annotated genes may be used for further analyses and integration of additional omic datasets, most notably metaproteomic data [39, 43, 44]. Furthermore, the higher number of complete genes improves the downstream functional analysis, because the read counts per gene will be much more accurate when having full length transcript sequences and will increase the probability to identify peptides. More specifically, the large number of predicted genes may enhance the usage of generated metaproteomic data, allowing more peptides, and thus proteins, to be identified.

Conclusions

IMP represents the first self-contained and standardized pipeline developed to leverage the advantages associated with integrating MG and MT data for large-scale analyses of microbial community structure and function in situ [4, 6]. IMP performs all the necessary large-scale bioinformatic analyses, including preprocessing, assembly, binning (automated), and analyses within an automated, reproducible, and user-friendly pipeline. In addition, we demonstrate that IMP vastly enhances data usage to produce high-volume and high-quality output. Finally, the combination of open development and reproducibility should promote the general paradigm of reproducible research within the microbiome research community.

Methods

The details of the IMP workflow, implementation, and customizability are described in further detail. We also describe the additional analyses carried out for assessment and benchmarking of IMP.

Details of the IMP implementation and workflow

A Python (v3) wrapper script was implemented for user-friendly execution of IMP via the command line. The full list of dependencies, parameters (see below), and documentation is available on the IMP website (http://r3lab.uni.lu/web/imp/doc.html). Although IMP was designed specifically for integrated analysis of MG and MT data, it can also be used for single MG or MT analyses as an additional functionality.

Reproducibility

IMP is implemented around a Docker container that runs the Ubuntu 14.04 operating system, with all relevant dependencies. Five mounting points are defined for the Docker container with the -v option: i) input directory, ii) output directory, iii) database directory, iv) code directory, and v) configuration file directory. Environment variables are defined using the -e parameter, including: i) paired MG data, ii) paired MT data, and iii) configuration file. The latest IMP Docker image will be downloaded and installed automatically upon launching the command, but users may also launch specific versions based on tags or use modified/customized versions of their local code base (documentation at http://r3lab.uni.lu/web/imp/doc.html).

Automation and modularity

Automation of the workflow is achieved using Snakemake 3.4.2 [49, 51], a Python-based make language implemented specifically for building reproducible bioinformatic workflows and pipelines. Snakemake is inherently modular and thus allows various features to be implemented within IMP, including the options of i) executing specific/selected steps within the pipeline, ii) check-pointing, i.e., resuming analysis from a point of possible interruption/termination, iii) analysis of single-omic datasets (MG or MT). For more details regarding the functionalities of IMP, please refer to the documentation of IMP (http://r3lab.uni.lu/web/imp/doc.html).

Input data

The input to IMP includes MG and/or MT FASTQ paired files, i.e., pairs-1 and pairs-2 are in individual files. The required arguments for the IMP wrapper script are metagenomic paired-end reads ("-m" options) and/or metatranscriptomic paired-end reads ("-t" option) with the specified output folder ("-o" option). Users may customize the command with the options and flags described in the documentation (http://r3lab.uni.lu/web/imp/doc.html) and in the "Customization and further development" section.

Trimming and quality filtering

Trimmomatic 0.32 [52] is used to perform trimming and quality filtering of MG and MT Illumina paired-end reads, using the following parameters: ILLUMINACLIP: TruSeq3-PE.fa:2:30:10; LEADING:20; TRAILING:20; SLIDINGWINDOW:1:3; MAXINFO:40:0.5; MINLEN:40. The parameters may be tuned via the command line or within the IMP config file. The output from this step includes retained paired-end and single-end reads (mate discarded), which are all used for downstream processes. These parameters are configurable in the IMP config file (section "Customization and further development")

Ribosomal RNA filtering

SortMeRNA 2.0 [68] is used for filtering rRNA from the MT data. The process is applied on FASTQ files for both paired- and single-end reads generated from the trimming and quality filtering step. Paired-end FASTQ files are interleaved prior to running SortMeRNA. If one of the mates within the paired-end read is classified as an rRNA sequence, then the entire pair is filtered out. After running SortMeRNA, the interleaved paired-end output is split into two separate paired-end FASTQ files. The filtered sequences (without rRNA reads) are used for the downstream processes. All available databases provided within SortMeRNA are used for filtering and the maximum memory usage parameter is set to 4 GB (option: "-m 4000"), which can be adjusted in the IMP config file (section "Customization and further development").

Read mapping

The read mapping procedure is performed using the bwa mem aligner [69] with settings: " -v 1" (verbose output level), "-M" (Picard compatibility) introducing an automated samtools header using the "-R" option [69]. Paired- and single-end reads are mapped separately and the resulting alignments are merged (using samtools merge [70]). The output is written as a binary aligment map (BAM) file. Read mapping is performed at various steps in the workflow, including: i) screening for host or contaminant sequences (section "Screening host or contaminant sequences"), ii) recruitment of unmapped reads within the IMP-based iterative co-assembly (section "Extracting unmapped reads"), and iii) mapping of preprocessed MG and MT reads to the final contigs. The memory usage is configurable in the IMP config file (section "Customization and further development").

Extracting unmapped reads

The extraction of unmapped reads (paired- and single-end) begins by mapping reads to a given reference sequence (section "Read mapping"). The resulting BAM file is used as input for the extraction of unmapped reads. A set of paired-end reads are considered unmappable if both or either one of the mates do not map to the given reference. The unmapped reads are converted from BAM to FASTQ format using samtools [70] and BEDtools 2.17.0—bamToFastq utility [71]. Similarly, unmapped single-end reads are also extracted from the alignment information.

Screening host or contaminant sequences

By default, the host/contaminant sequence screening is performed by mapping both paired- and single-end reads (section "Read mapping") onto the human genome version 38 (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/), followed by extraction of unmapped reads (section "Extracting unmapped reads"). Within the IMP command line, users are provided with the option of i) excluding this procedure with the "--no-filtering" flag, ii) using other sequence(s) for screening by providing the FASTA file (or URL) using "--screen" option, or iii) specifying it in the configuration file (section "Customization and further development").

Parameters of the IMP-based iterative co-assembly

The IMP-based iterative co-assembly implements MEGAHIT 1.0.3 [23] as the MT assembler while IDBA-UD 1.1.1 [22] is used as the default co-assembler (MG and MT), with MEGAHIT [23] as an alternative option for the co-assembler (specified by the "-a" option of the IMP command line). All de novo assemblies are performed on *k*mers ranging from 25-mers to 99-mers, with an incremental step of four. Accordingly, the command line parameters for IDBA-UD are "--mink 25 --maxk 99 --step 4 --similar 0.98 --pre-correction" [22]. Similarly, the command line parameters for MEGAHIT are "--k-min 25 --k-max 99 --k-step 4", except for the MT assemblies which are performed with an additional "--no-bubble" option to prevent merging of bubbles within the assembly graph [23]. Furthermore, contigs generated from the MT assembly are used as "long read" input within the "-l" flag of IDBA-UD or "-r" flag of MEGAHIT [22, 23]. *K*mer ranges for the IDBA-UD and MEGAHIT can be adjusted/specified in the configuration file (section "Customization and further development"). Cap3 is used to reduce the redundancy and improve contiguity of the assemblies using a minimum alignment identity of 98% (“-p 0.98”) with a minimum overlap of 100 bases (“-o 100”), which are adjustable in the configuration file (section “Customization and further development”). Finally, the extraction of reads that are unmappable to the initial MT assembly and initial co-assembly is described in the “Extracting unmapped reads” section.

Annotation and assembly quality assessment

Prokka 1.11 [55] with the “- -metagenome” setting is used to perform functional annotation. The default BLAST and HMM databases of Prokka are used for the functional annotation. Custom databases may be provided by the user (refer to the “Databases” and “Customization and further development” sections for details).

MetaQUAST 3.1 [54] is used to perform taxonomic annotation of contigs with the maximum number of downloadable reference genomes set to 20 (“- -max-ref-number 20”). In addition, MetaQUAST provides various assembly statistics. The maximum number of downloadable reference genomes can be changed in the IMP config file (see “Customization and further development” for details).

Depth of coverage

Contig- and gene-wise depth of coverage values are calculated (per base) using BEDtools 2.17.0 [71] and aggregated (by average) using awk, adapted from the CONCOCT code [16] (script: map-bowtie2-markduplicates.sh; https://github.com/BinPro/CONCOCT) and is non-configurable.

Variant calling

The variant calling procedure is performed using Samtools 0.1.19 [70] (mpileup tool) and Platypus 0.8.1 [72], each using their respective default settings and which are non-configurable. The input is the merged paired- and single-end read alignment (BAM) against the final assembly FASTA file (section “Read mapping”). The output files from both the methods are indexed using tabix and compressed using gzip. No filtering is applied to the variant calls, so that users may access all the information and filter it according to their requirements. The output from samtools mpileup is used for the augmented VizBin visualization.

Non-linear dimensionality reduction of genomic signatures

VizBin [56] performs non-linear dimensionality reduction of genomic signatures onto contigs ≥1 kb, using default settings, to obtain two-dimensional embeddings. Parameters can be modified in the IMP config file (section “Customization and further development”).

Automated binning

Automated binning of the assembled contigs is performed using MaxBin 2.0. Default setting are applied and paired-end reads are provided as input for abundance estimation [20]. The sequence length cutoff is set to be same as VizBin (section “Non-linear dimensionality reduction of genomic signatures”) and is customizable using the config file (section “Customization and further development”).

Visualization and reporting

IMP compiles the multiple summaries and visualizations into a HTML report [57]. FASTQC [73] is used to visualize the quality and quantity of reads before and after preprocessing. MetaQUAST [54] is used to report assembly quality and taxonomic associations of contigs. A custom script is used to generate KEGG-based [74] functional Krona plots by running KronaTools [75] (script: genes.to.kronaTable.py, GitHub URL: https://github.com/EnvGen/metagenomics-workshop). Additionally, VizBin output (two-dimensional embeddings) is integrated with the information derived from the IMP analyses, using a custom R script for analysis and visualization of the augmented maps. The R workspace image is saved such that users are able to access it for further analyses. All the steps executed within an IMP run, including parameters and runtimes, are summarized in the form of a workflow diagram and a log-file. The visualization script is not configurable.

Output

The output generated by IMP includes a multitude of large files. Paired- and single-end FASTQ files of preprocessed MG and MT reads are provided such that the user may employ them for additional downstream analyses. The output of the IMP-based iterative co-assembly consists of a FASTA file, while the alignments/mapping of MG and MT preprocessed reads to the final co-assembly are also provided as BAM files, such that users may use these for further processing. Predicted genes and their respective annotations are provided in the various formats produced by Prokka [55]. Assembly quality statistics and taxonomic annotations of contigs are provided as per the output of MetaQUAST [54]. Two-dimensional embeddings from the NLDR-GS are provided such that they can be exported to and further curated using VizBin [56]. Additionally, abundance and expression information is represented by contig- and gene-level average depth of coverage values. MG and MT genomic variant information (VCF format), including both SNPs and INDELs (insertions and deletions), is also provided. The results of the automated binning using MaxBin 2.0 [20] are provided in a folder which contains the default output from the program (i.e., fasta files of bins and summary files).

The HTML reports [57], e.g., HTML S1 and S2, compile various summaries and visualizations, including, i) augmented VizBin maps, ii) MG- and MT-level functional Krona charts [75], iii) detailed schematics of the steps carried out within the IMP run, iv) list of parameters and commands, and v) additional reports (FASTQC report [73], MetaQUAST report [54]). Please refer to the documentation of IMP for a detailed list and description of the output (http://r3lab.uni.lu/web/imp/doc.html).

Databases

The IMP database folder (db) contains required databases required for IMP analysis. The folder contains the following subfolders and files with their specific content:

i. adapters folder — sequencing adapter sequences. Default version contains all sequences provided by Trimmomatic version 0.32 [52]
ii. cm, genus, hmm, and kingdom folders — contains databases provided by Prokka 1.11 [55]. Additional databases may be added into the corresponding folders as per the instructions in the Prokka documentation (https://github.com/tseemann/prokka#databases)
iii. sortmerna folder — contains all the databases provided in SortMeRNA 2.0 [68]. Additional databases may be added into the corresponding folders as per the instructions in the SortMeRNA documentation (http://bioinfo.lifl.fr/RNA/sortmerna/code/SortMeRNA-user-manual-v2.0.pdf)
iv. ec2pathways.txt — enzyme commission (EC) number mapping of amino acid sequences to pathways
v. pathways2hierarchy.txt — pathway hierarchies used to generated for KEGG-based functional Krona plot (section "Visualization and reporting")

Customization and further development

Additional advanced parameters can be specified via the IMP command line, including specifying a custom configuration file ("-c" option) and/or specifying a custom database folders ("-d" option). Threads ("- -threads") and memory allocation ("- -memcore" and "- -memtotal") can be adjusted via the command line and the configuration file. The IMP launcher script provides a flag ("- -enter") to launch the Docker container interactively and the option to specify the path to the customized source code folder ("-s" option). These commands are provided for development and testing purposes (described on the IMP website and documentation: http://r3lab.uni.lu/web/imp/doc.html). Further customization is possible using a custom configuration file (JSON format). The customizable options within the JSON file are specified in individual subsections within the "Details of the IMP implementation and workflow" section. Finally, the open source implementation of IMP allows users to customize the Docker image and source code of IMP according to their requirements.

Iterative single-omic assemblies

In order to determine the opportune number of iterations within the IMP-based iterative co-assembly strategy an initial assembly was performed using IMP preprocessed MG reads with IDBA-UD [22]. Cap3 [53] was used to further collapse the contigs and reduce the redundancy of the assembly. This initial assembly was followed by a total of three assembly iterations, whereby each iteration was made up of four separate steps: i) extraction of reads unmappable to the previous assembly (using the procedure described in the "Extracting unmapped reads" section), ii) assembly of unmapped reads using IDBA-UD [22], iii) merging/collapsing the contigs from the previous assembly using cap3 [53], and iv) evaluation of the merged assembly using MetaQUAST [54]. The assembly was evaluated in terms of the per-iteration increase in mappable reads, assembly length, numbers of contigs ≥1 kb, and numbers of unique genes.

Similar iterative assemblies were also performed for MT data using MEGAHIT [23], except CD-HIT-EST [76] was used to collapse the contigs at ≥95% identity ("-c 0.95") while MetaGeneMark [77] was used to predict genes. The parameters and settings of the other programs were the same as those defined in the "Details of the IMP implementation and workflow" section.

The aforementioned procedures were applied to all the datasets analyzed within this article. The merged contig sets (non-redundant) from the first iteration of both the MG and MT iterative assemblies were selected to represent the IMP single-omics assemblies (IMP_MG and IMP_MT) and were compared against co-assemblies.

Execution of pipelines

MetAMOS v1.5rc3 was executed using default settings. MG data were provided as input for single-omic assemblies (MetAMOS_MG) while MG and MT data were provided as input for multi-omic co-assemblies (MetAMOS_MGMT). All computations using MetAMOS were set to use eight computing cores ("-p 8").

MOCAT v1.3 (MOCAT.pl) was executed using default settings. Paired-end MG data were provided as input for single-omic assemblies (MOCAT_MG) while paired-end MG and MT data were provided as input for multi-omic co-assemblies (MOCAT_MGMT). All computations using MOCAT were set to use eight computing cores ("-cpus 8"). Paired-end reads were first preprocessed using the read_trim_filter step of MOCAT ("-rtf"). For the human fecal microbiome datasets (HF1–5), the preprocessed paired- and single-end reads were additionally screened for human genome-derived sequences ("-s hg19"). The resulting reads were afterwards assembled with default parameters ("-gp assembly -r hg19") using SOAPdenovo.

IMP v1.4 was executed for each dataset using different assemblers for the co-assembly step: i) default setting using IDBA-UD, and ii) MEGAHIT ("-a megahit"). Additionally, the analysis of human fecal microbiome datasets (HF1–5) included the preprocessing step of filtering human genome sequences, which was omitted for the wastewater sludge datasets (WW1–4) and the biogas (BG) reactor dataset. Illumina TruSeq2 adapter trimming was used for wastewater dataset preprocessing since the information was available. Computation was performed using eight computing cores ("--threads 8"), 32 GB memory per core ("--memcore 32") and total memory of 256 GB ("--memtotal 256 GB"). The customized parameters were specified in the IMP configuration file (exact configurations listed in the HTML reports [57]). The analysis of the CAMI datasets were carried using the MEGAHIT assembler option ("-a megahit"), while the other options remained as default settings.

In addition, IMP was also used on a small scale dataset to evaluate performance of increasing the number of threads from 1 to 32 and recording the runtime ("time" command). IMP was launched on the AWS cloud computing platform running the MEGAHIT as the assembler ("-a megahit") with 16 threads ("--threads 16") and 122 GB of memory ("--memtotal 122").

Data usage assessment

Preprocessed paired-end and single-end MG and MT reads from IMP were mapped (section Read mapping) onto the IMP-based iterative co-assemblies and IMP_MG assembly. Similarly, preprocessed paired-end and single-end MG and MT reads from MOCAT were mapped onto the MOCAT co-assembly (MOCAT_MGMT) and the MOCAT single-omic MG assembly (MOCAT_MG). MetAMOS does not retain single-end reads; therefore, preprocessed MG and MT paired-end reads from MetAMOS were mapped onto the MetAMOS co-assembly (MetAMOS_MGMT) and MetAMOS single-omic MG assembly (MetAMOS_MG).

Preprocessed MG and MT reads from the human fecal datasets (HF1–5) were mapped using the same parameters described in the "Read mapping" section to the IGC reference database [35] for evaluation of a reference-based approach. Alignment files of MG and MT reads mapping to the IMP-based iterative co-assemblies and the aforementioned alignments to the IGC reference database were used to report the fractions of properly paired reads mapping in either IMP-based iterative co-assembly, IGC reference database, or both. These fractions were then averaged across all the human fecal datasets (HF1–5).

Assembly assessment and comparison

Assemblies were assessed and compared using MetaQUAST by providing contigs (FASTA format) from all different (single- and multi-omic) assemblies of the same dataset as input [54]. The gene calling function ("-f") was utilized to obtain the number of genes which were predicted from the various assemblies. An additional parameter within MetaQUAST was used for ground truth assessment of the simulated mock (SM) community assemblies by providing the list of 73 FASTA format reference genomes ("-R"). The CPM measure was computed based on the information derived from the results of MetaQUAST [54]. In order to be consistent with the reported values (i.e., N50 length), the CPM measures reported within this article are based on alignments of 500 bp and above, unlike the 1-kb cutoff used in the original work [62]. Prodigal was also used for gene prediction to obtain the number of complete and incomplete genes [61].

Analysis of contigs assembled from MT data

A list of contigs with no MG depth of coverage together with additional information on these contigs (contig length, annotation, MT depth of coverage) was retrieved using the R workspace image, which is provided as part IMP output (sections "Visualization and reporting" and "Output"). The sequences of these contigs were extracted and subjected to a BLAST search on NCBI to determine their potential origin. Furthermore, contigs with length ≥1 kb, average depth of coverage ≥20 bases, and containing genes encoding known virus/bacteriophage functions were extracted.

Analysis of subsets of contigs

Subsets of contigs within the HF1 dataset were identified by visual inspection of augmented VizBin maps generated by IMP. Specifically, detailed inspection of contig-level MT to MG depth of coverage ratios was carried out using the R workspace provided as part of IMP output (sections "Visualization and reporting" and "Output"). The alignment information of contigs to isolate genomes provided by MetaQUAST [54] was used to highlight subsets of contigs aligning to genomes of the *Escherichia coli* P12B strain (*E. coli*) and *Collinsella intestinalis* DSM 13280 (*C. intestinalis*).

An additional reference-based analysis of MetaQUAST [54] was carried out for all the human fecal microbiome assemblies (HF1–5) by providing the genomes of *E. coli* P12B and *C. intestinalis* DSM 13280 as reference (flag: "-R") to assess the recovery fraction of the aforementioned genomes within the different assemblies.

Computational platforms

IMP and MetAMOS were executed on a Dell R820 machine with 32 Intel(R) Xeon(R) CPU E5-4640 @ 2.40GHz physical computing cores (64 virtual), 1024 TB of DDR3 RAM (32 GB per core) with Debian 7 Wheezy as the operating system. MOCAT, IMP single-omic assemblies, and additional analyses were performed on the Gaia cluster of the University of Luxembourg HPC platform [78].

IMP was executed on the Amazon Web Services (AWS) cloud computing platform using EC2 R3 type (memory optimized) model r3.4xlarge instance with 16 compute cores, 122 GB memory, and 320 GB of storage space running a virtual Amazon Machine Image (AMI) Ubuntu v16.04 operating system.

Additional files

Additional file 1: Supplementary figures and notes. **Figures S1–S3** and **Notes S1–S2**. Detailed figure legends available within file. (PDF 1047 kb)

Additional file 2: Supplementary tables. **Tables S1–S12.** Detailed table legends available within file. (XLSX 4350 kb)

Abbreviations

AWS: Amazon Web Services; BAM: Binary Alignment Maps; BG: Biogas; bp: Base pair; CAMI: Critical Assessment of Metagenome Interpretation; cDNA: Complementary DNA; Contigs: Contiguous sequence(s); HF: Human fecal; IGC: Integrated Gene Catalog; IMP: Integrated Meta-omic Pipeline; INDELs: Insertions and deletions; kb: Kilo base; KEGG: Kyoto Encyclopedia of Genes and Genomes; MetaHIT: Metagenomics of the Human Intestinal Tract; MG: Metagenomic; MT: Metatranscriptomic; NCBI: National Center for Biotechnology Information; NGS: Next-generation sequencing; rRNA: Ribosomal RNA; SM: Simulated mock; SNPs: Single nucleotide polymorphisms; SRA: Sequence read archive; VCF: Variant call format; WW: Wastewater

Acknowledgements

We would like to acknowledge John Larsson from SciLifeLab (Sweden) for kindly providing the KEGG-based functional Krona plot scripts. Albi Celaj from the University of Toronto is thanked for supplying the in silico simulated metatranscriptomic data and the corresponding reference genomes. The University of Luxembourg High Performance Computing (HPC) facility is duly thanked for providing and maintaining the computing platform. The Reproducible Research Results (R3) team of the Luxembourg Centre for Systems Biomedicine is acknowledged for support of the project and for promoting reproducible research. Finally, we acknowledge the Luxembourg National Research Fund (FNR) for funding the project via the ATTRACT, AFR, CORE, INTER and PoC grants.

Funding

This work was supported by an ATTRACT program grant (A09/03), a European Union Joint Programming in Neurodegenerative Diseases grant (INTER/JPND/12/01), a proof-of-concept grant (PoC/13/02), an ERASysAPP grant (INTER/8888488), and CORE programme grant (CORE/15/BM/10404093) to PW, an Aide à la Formation Recherche (AFR) grant to SN (PHD-2014-1/7934898), and a CORE junior (C15/SR/10404839) to EELM, all funded by the Luxembourg National Research Fund (FNR).

Availability and requirements

All the data, software, and source code related to this manuscript are publicly available.

Coupled metagenomic and metatranscriptomic datasets

The published human fecal microbiome datasets (MG and MT) were obtained from NCBI Bioproject PRJNA188481 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA188481). They include samples from individuals X310763260, X311245214, X316192082, X316701492, and X317690558 [28], designated within this article as HF1–5, respectively. Only samples labeled as "Whole" (samples preserved by flash-freezing) were selected for analysis [28]. The published wastewater sludge microbial community datasets (MG and MT) were obtained from NCBI Bioproject with the accession code PRJNA230567 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA230567). These include samples A02, D32, D36, and D49, designated within this article as WW1–4, respectively [43].

The published biogas reactor microbial community data set (MG and MT) was obtained from the European Nucleotide Archive (ENA) project PRJEB8813 (http://www.ebi.ac.uk/ena/data/view/PRJEB8813) and is designated within this article as BG [29].

Simulated coupled metagenomic and metatranscriptomic dataset

The simulated MT data were obtained upon request from the original authors [12]. A complementary metagenome was simulated using the same set of 73 bacterial genomes used for the aforementioned simulated MT [12]. Simulated reads were obtained using the NeSSM MG simulator (default settings) [79]. The simulated mock community is designated as SM within this article [79]. The simulated data along with the corresponding reference genomes used to generate the MG data are made available via LCSB WebDav (https://webdav-r3lab.uni.lu/public/R3lab/IMP/datasets/) and is archived on Zenodo [80].

CAMI simulated community metagenomic datasets

The medium complexity CAMI simulated MG data and the corresponding gold standard assembly were obtained from the CAMI website (http://www.cami-challenge.org).

Test dataset for runtime assessment

A subset of ~5% of reads from both the WW1 MG and MT datasets (section "Coupled metagenomic and metatranscriptomic datasets") was selected and used as the data to perform runtime assessments. This dataset could be used to test IMP on regular platforms such as laptops and desktops. It is made available via the LCSB R3 WebDav (https://webdav-r3lab.uni.lu/public/R3lab/IMP/datasets/) and is archived on Zenodo [81].

Software and source code

IMP is available under the MIT license on the LCSB R3 website (http://r3lab.uni.lu/web/imp/), which contains necessary information related to IMP. These include links to the Docker images on the LCSB R3 WebDav (https://webdav-r3lab.uni.lu/public/R3lab/IMP/dist/) and is archived on Zenodo [82]. Source code is available on LCSB R3 GitLab (https://git-r3lab.uni.lu/IMP/IMP), GitHub (https://github.com/shaman-narayanasamy/IMP), and is archived on Zenodo [83]. Scripts and commands for additional analyses performed specifically within this manuscript are available on LCSB R3 GitLab (https://git-r3lab.uni.lu/IMP/IMP_manuscript_analysis) and on GitHub (https://github.com/shaman-narayanasamy/IMP_manuscript_analysis). Frozen pages containing all necessary material related to this article are available at http://r3lab.uni.lu/frozen/imp/.

Authors' contributions

SN, NP, EELM, PM, and PW conceived the analysis and designed the workflow. SN, YJ, MH, and CCL developed the software, wrote the documentation and tested the software. YJ ensured reproducibility of the software. SN, PM, and MH performed data analyses. EELM, PM, AHB, AK, NP, and PW participated in discussions and tested the software. SN, EELM, AHB, PM, NP, AK, MH, and PW wrote and edited the manuscript. PW designed and supported the project. All authors read and agreed on the final version of the manuscript.

Authors' information

Current affiliations: CCL—Saarland University, Building E2 1, 66123 Saarbrücken, Germany; NP—Universidad EAFIT, Carrera 49 No 7 sur 50, Medellín, Colombia; EELM—Department of Microbiology, Genomics and the Environment, UMR 7156 UNISTRA—CNRS, Université de Strasbourg, Strasbourg, France.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Luxembourg Centre for Systems Biomedicine, 7, avenue des Hauts-Fourneaux, Esch-sur-Alzette L-4362, Luxembourg. [2]Present address: Department of Microbiology, Genomics and the Environment, UMR 7156 UNISTRA—CNRS, Université de Strasbourg, Strasbourg, France. [3]Present address: Saarland University, Building E2 1, Saarbrücken 66123, Germany. [4]Institute of Systems Biology, 401 Terry Avenue North, Seattle, WA 98109, USA. [5]Present address: Universidad EAFIT, Carrera 49 No 7 sur 50, Medellín, Colombia.

Received: 18 October 2016 Accepted: 22 November 2016
Published online: 16 December 2016

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett C, Knight R, Gordon JI. The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature. 2007;449:804–10.
2. Rittmann BE. Microbial ecology to manage processes in environmental biotechnology. Trends Biotechnol. 2006;24:261–6.
3. Stewart EJ. Growing unculturable bacteria. J Bacteriol. 2012;194:4151–60.
4. Narayanasamy S, Muller EEL, Sheik AR, Wilmes P. Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. Microb Biotechnol. 2015;8:363–8.
5. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C, Boernigen D, Tickle TL, Morgan XC, Garrett WS, Huttenhower C. Computational meta'omics for microbial community studies. Mol Syst Biol. 2013;9:666.
6. Muller EEL, Glaab E, May P, Vlassis N, Wilmes P. Condensing the omics fog of microbial communities. Trends Microbiol. 2013;21:325–33.
7. Roume H, Muller EEL, Cordes T, Renaut J, Hiller K, Wilmes P. A biomolecular isolation framework for eco-systems biology. ISME J. 2013;7:110–21.
8. Roume H, Heintz-Buschart A, Muller EEL, Wilmes P. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. Methods Enzymol. 2013;531:219–36.
9. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, Rasmussen S, Brunak S, Pedersen O, Guarner F, de Vos WM, Wang J, Li J, Doré J, Ehrlich SD, Stamatakis A, Bork P. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013;10:1196–9.
10. Treangen TJ, Koren S, Sommer DD, Liu B, Astrovskaya I, Ondov B, Darling AE, Phillippy AM, Pop M. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biol. 2013;14:R2.
11. Nalbantoglu OU, Way SF, Hinrichs SH, Sayood K. RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. BMC Bioinformatics. 2011;12:41.
12. Celaj A, Markle J, Danska J, Parkinson J. Comparison of assembly algorithms for improving rate of metatranscriptomic functional annotation. Microbiome. 2014;2:39.
13. Laczny CC, Pinel N, Vlassis N, Wilmes P. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. Sci Rep. 2014;4:4516.
14. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol. 2013;31:533–8.
15. Nielsen HB, Almeida M, Juncker AS, Rasmussen S, Li J, Sunagawa S, Plichta DR, Gautier L, Pedersen AG, Le Chatelier E, Pelletier E, Bonde I, Nielsen T, Manichanh C, Arumugam M, Batto J-M, Quintanilha Dos Santos MB, Blom N, Borruel N, Burgdorf KS, Boumezbeur F, Casellas F, Doré J, Dworzynski P, Guarner F, Hansen T, Hildebrand F, Kaas RS, Kennedy S, Kristiansen K, et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. Nat Biotechnol. 2014;32:822–8.
16. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11:1144–6.
17. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3:e1319.
18. Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ. 2015;3:e1165.
19. Laczny CC, Muller EEL, Heintz-Buschart A, Herold M, Lebrun LA, Hogan A, May P, De Beaufort C, Wilmes P. Identification, recovery, and refinement of hitherto undescribed population-level genomes from the human gastrointestinal tract. Front Microbiol. 2016;7:884.
20. Wu Y-W, Tang Y-H, Tringe SG, Simmons BA, Singer SW, Metzker M, Dick G, Andersson A, Baker B, Simmons S, Thomas B, Yelton A, Banfield J, Tyson G, Chapman J, Hugenholtz P, Allen E, Ram R, Richardson P, Solovyev V, Rubin E, Rokhsar D, Banfield J, Mackelprang R, Waldrop M, DeAngelis K, David M, Chavarria K, Blazewicz S, Rubin E, et al. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome. 2014;2:26.
21. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ. 2014;2:e603.
22. Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics. 2012;28:1420–8.
23. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674–6.
24. Westreich ST, Korf I, Mills DA, Lemay DG, Moran M, Leimena M, Embree M, McGrath K, Dimitrov D, Cho I, Blaser M, Round J, Mazmanian S, Gosalbes M, Giannoukos G, Reck M, Hainzl E, Bolger A, Lohse M, Usadel B, Magoc T, Salzberg S, Meyer F, Tatusova T, Wilke A, Overbeek R, Love M, Huber W, Anders S, Costa V, et al. SAMSA: a comprehensive metatranscriptome analysis pipeline. BMC Bioinformatics. 2016;17:399.
25. Martinez X, Pozuelo M, Pascal V, Campos D, Gut I, Gut M, Azpiroz F, Guarner F, Manichanh C, Li J, Gosalbes MJ, Helbling DE, Ackermann M, Fenner K, Kohler HP, Johnson DR, Tulin S, Aguiar D, Istrail S, Smith J, Leimena MM, He S, Murakami S, Fujishima K, Tomita M, Kanai A, Manichanh C, Li R, McDonald D, Wilke A, et al. MetaTrans: an open-source pipeline for metatranscriptomics. Sci Rep. 2016;6:26447.
26. Leimena MM, Ramiro-Garcia J, Davids M, van den Bogert B, Smidt H, Smid EJ, Boekhorst J, Zoetendal EG, Schaap PJ, Kleerebezem M. A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. BMC Genomics. 2013;14:530.
27. Satinsky BMM, Fortunato CS, Doherty M, Smith CBC, Sharma S, Ward NDNND, Krusche AAV, Yager PL, Richey JE, Moran MA, Crump BBC, Richey JE, Devol A, Wofsy S, Victoria R, Riberio M, Nebel G, Dragsted J, Vega A, Hedges J, Clark W, Quay P, Richey JE, Devol A, Santos U, Spencer R, Hernes P, Aufdenkampe A, Baker A, Gulliver P, et al. Metagenomic and metatranscriptomic inventories of the lower Amazon River, May 2011. Microbiome. 2015;3:39.
28. Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G, Boylan MR, Ciulla D, Gevers D, Izard J, Garrett WS, Chan AT, Huttenhower C. Relating the metatranscriptome and metagenome of the human gut. Proc Natl Acad Sci U S A. 2014;111:E2329–38.
29. Bremges A, Maus I, Belmann P, Eikmeyer F, Winkler A, Albersmeier A, Pühler A, Schlüter A, Sczyrba A. Deeply sequenced metagenome and metatranscriptome of a biogas-producing microbial community from an agricultural production-scale biogas plant. Gigascience. 2015;4:33.
30. Leung HCM, Yiu S-M, Parkinson J, Chin FYL. IDBA-MT: de novo assembler for metatranscriptomic data generated from next-generation sequencing technology. J Comput Biol. 2013;20:540–50.
31. Leung HCM, Yiu SM, Chin FYL. IDBA-MTP: a hybrid metatranscriptomic assembler based on protein information. Res Comput Mol Biol. 2014; 160–172.
32. Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res. 2012;40;e155.
33. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Palma F, Birren BW, Nusbaum C, Lindblad-toh K, Friedman N, Regev A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
34. Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, Arumugam M, Pan Q, Liu B, Qin J, Wang J, Bork P. MOCAT: a metagenomics assembly and gene prediction toolkit. PLoS One. 2012;7:e47656.
35. Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, Arumugam M, Kultima JR, Prifti E, Nielsen T, Juncker AS, Manichanh C, Chen B, Zhang W, Levenez F, Wang J, Xu X, Xiao L, Liang S, Zhang D, Zhang Z, Chen W, Zhao H, Al-Aama JY, Edris S, Yang H, Wang J, Hansen T, Nielsen HB, Brunak S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol. 2014;32:834–41.
36. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010;464:59–65.
37. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu

S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 2012;1:18.
38. Lai B, Wang F, Wang X, Duan L, Zhu H. InteMAP: Integrated metagenomic assembly pipeline for NGS short reads. BMC Bioinformatics. 2015;16:244.
39. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, Wampach L, Schneider JG, Hogan A, de Beaufort C, Wilmes P. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2016;2:16180.
40. Hultman J, Waldrop MP, Mackelprang R, David MM, Mcfarland J, Blazewicz SJ, Harden J, Turetsky MR, Mcguire AD, Shah MB, Verberkmoes NC, Lee LH. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. Nature. 2015;521:208–12.
41. Beulig F, Urich T, Nowak M, Trumbore SE, Gleixner G, Gilfillan GD, Fjelland KE, Küsel K. Altered carbon turnover processes and microbiomes in soils under long-term extremely high CO2 exposure. Nat Microbiol. 2016;1:15025.
42. Urich T, Lanzén A, Stokke R, Pedersen RB, Bayer C, Thorseth IH, Schleper C, Steen IH, Ovreas L. Microbial community structure and functioning in marine sediments associated with diffuse hydrothermal venting assessed by integrated meta-omics. Environ Microbiol. 2014;16:2699–710.
43. Muller EEL, Pinel N, Laczny CC, Hoopman MR, Narayanasamy S, Lebrun LA, Roume H, Lin J, May P, Hicks ND, Heintz-Buschart A, Wampach L, Liu CM, Price LB, Gillece JD, Guignard C, Schupp JM, Vlassis N, Baliga NS, Moritz RL, Keim PS, Wilmes P. Community integrated omics links the dominance of a microbial generalist to fine-tuned resource usage. Nat Commun. 2014;5:5603.
44. Roume H, Heintz-Buschart A, Muller EEL, May P, Satagopam VP, Laczny CC, Narayanasamy S, Lebrun LA, Hoopmann MR, Schupp JM, Gillece JD, Hicks ND, Engelthaler DM, Sauter T, Keim PS, Moritz RL, Wilmes P. Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. npj Biofilms Microbiomes. 2015;1:15007.
45. Kenall A, Edmunds S, Goodman L, Bal L, Flintoft L, Shanahan DR, Shipley T. Better reporting for better research: a checklist for reproducibility. BMC Neurosci. 2015;16:44.
46. Belmann P, Dröge J, Bremges A, McHardy AC, Sczyrba A, Barton MD. Bioboxes: standardised containers for interchangeable bioinformatics software. Gigascience. 2015;4:47.
47. Di Tommaso P, Palumbo E, Chatzou M, Prieto P, Heuer ML, Notredame C. The impact of Docker containers on the performance of genomic pipelines. PeerJ. 2015;3:e1273.
48. Leipzig J. A review of bioinformatic pipeline frameworks. Brief Bioinform. 2016. http://bib.oxfordjournals.org/content/early/2016/03/23/bib.bbw020.full.
49. Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics. 2012;28:2520–2.
50. Amstutz P, Crusoe MR, Tijanić N, Chapman B, Chilton J, Heuer M, Kartashov A, Leehr D, Ménager H, Nedeljkovich M, Scales M, Soiland-Reyes S, Stojanovic L. Common Workflow Language, v1.0. 2016. https://figshare.com/articles/Common_Workflow_Language_draft_3/3115156.
51. Koster J. Reproducibility in next-generation sequencing analysis. 2014.
52. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
53. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999;9:868–77.
54. Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics. 2015;32:1088–90.
55. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
56. Laczny CC, Sternal T, Plugaru V, Gawron P, Atashpendar A, Margossian HH, Coronado S, der Maaten L, Vlassis N, Wilmes P. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome. 2015;3:1.
57. IMP HTML reports. October 17, 2016. http://dx.doi.org/10.5281/zenodo.161321.
58. Schürch AC, Schipper D, Bijl MA, Dau J, Beckmen KB, Schapendonk CME, Raj VS, Osterhaus ADME, Haagmans BL, Tryland M, Smits SL. Metagenomic survey for viruses in Western Arctic caribou, Alaska, through iterative assembly of taxonomic units. PLoS One. 2014;9:e105227.
59. Reyes A, Blanton LV, Cao S, Zhao G, Manary M, Trehan I, Smith MI, Wang D, Virgin HW, Rohwer F, Gordon JI. Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. Proc Natl Acad Sci U S A. 2015;112:11941–6.
60. Hitch T, Creevey C. Spherical: an iterative workflow for assembling metagenomic datasets. bioRxiv. 2016. http://biorxiv.org/content/early/2016/08/02/067256.
61. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ, Delcher A, Bratke K, Powers E, Salzberg S, Lukashin A, Borodovsky M, Benson D, Karsch-Mizrachi I, Lipman D, Ostell J, Sayers E, Larsen T, Krogh A, Zhu H, Hu G, Yang Y, Wang J, She Z, Ou H, Guo F, Zhang C, Tech M, Pfeifer N, Morgenstern B, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
62. Deng X, Naccache SN, Ng T, Federman S, Li L, Chiu Y, Delwart EL. An ensemble strategy that significantly improves de novo assembly of microbial genomes from metagenomic next-generation sequencing data. Nucleic Acids Res. 2015;43:e46.
63. Mende DR, Waller AS, Sunagawa S, Järvelin AI, Chan MM, Arumugam M, Raes J, Bork P. Assessment of metagenomic assembly using simulated next generation sequencing data. PLoS One. 2012;7:e31386.
64. Pruitt K, Brown G, Tatusova T, Maglott D. The Reference Sequence (RefSeq) Database. In: NCBI Handbook. 2002. p. 1–24.
65. Salzberg SL, Yorke JA. Beware of mis-assembled genomes. Bioinformatics. 2005;21:4320–1.
66. Mariano DCB, Sousa Tde J, Pereira FL, Aburjaile F, Barh D, Rocha F, Pinto AC, Hassan SS, Saraiva TDL, Dorella FA, de Carvalho AF, Leal CAG, Figueiredo HCP, Silva A, Ramos RTJ, Azevedo VAC, Dorella F, Pacheco LC, Oliveira S, Miyoshi A, Azevedo V, Aleman M, Spier S, Wilson W, Doherr M, Soares S, Silva A, Trost E, Blom J, Ramos R, et al. Whole-genome optical mapping reveals a mis-assembly between two rRNA operons of Corynebacterium pseudotuberculosis strain 1002. BMC Genomics. 2016;17:315.
67. Hug LA, Thomas BC, Sharon I, Brown CT, Sharma R, Hettich RL, Wilkins MJ, Williams KH, Singh A, Banfield JF. Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. Environ Microbiol. 2016;18:159–73.
68. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28:3211–7.
69. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:589–95.
70. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
71. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
72. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, McVean G, Lunter G. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014;46:912–8.
73. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One. 2012;7:e30619.
74. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28:27–30.
75. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011;12:385.
76. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
77. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010;38:e132.
78. Varrette S, Bouvry P, Cartiaux H, Georgatos F. Management of an Academic HPC Cluster: the UL Experience. In: Proceedings of the 2014 International Conference on High Performance Computing Simulation. 2014. p. 959–67.
79. Jia B, Xuan L, Cai K, Hu Z, Ma L, Wei C. NeSSM: a next-generation sequencing simulator for metagenomics. PLoS One. 2013;8:e75448.
80. IMP simulated mock community data set. October 12, 2016. http://doi.org/10.5281/zenodo.160261.
81. IMP small scale test dataset. October 14, 2016. http://doi.org/10.5281/zenodo.160708.
82. IMP v1.4 docker image. October 12, 2016. http://doi.org/10.5281/zenodo.160263.
83. IMP v1.4 source code. October 14, 2016. http://doi.org/10.5281/zenodo.160703.

A.3 Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage.

Emilie E.L. Muller, Nicolás Pinel, Cédric C. Laczny, Michael R. Hoopmann, **Shaman Narayanasamy**,
Laura A. Lebrun, Hugo Roume, Jake Lin, Patrick May, Nathan D. Hicks, Anna Heintz-Buschart, Linda
Wampach, Cindy M. Liu, Lance B. Price, John D. Gillece, Cédric Guignard, Jim M. Schupp, Nikos Vlassis,
Nitin S. Baliga, Robert L. Moritz, Paul S. Keim, and Paul Wilmes
2014
Nature Communications **5**: 5603
DOI: 10.1038/ncomms6603

The authors involvement in this work resulted in the conception of the methods and strategies applied within the eventual development of the integrated omics pipeline.

Contributions of author include:

- Data analysis
- Research design
- Data visualization
- Revision of manuscript

nature COMMUNICATIONS

ARTICLE

Received 15 Aug 2014 | Accepted 20 Oct 2014 | Published 26 Nov 2014

DOI: 10.1038/ncomms6603

OPEN

Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage

Emilie E. L. Muller[1,*], Nicolás Pinel[1,2,*], Cédric C. Laczny[1], Michael R. Hoopmann[2], Shaman Narayanasamy[1], Laura A. Lebrun[1], Hugo Roume[1,†], Jake Lin[1], Patrick May[1], Nathan D. Hicks[3], Anna Heintz-Buschart[1], Linda Wampach[1], Cindy M. Liu[3], Lance B. Price[3], John D. Gillece[3], Cédric Guignard[4], James M. Schupp[3], Nikos Vlassis[1,†], Nitin S. Baliga[2], Robert L. Moritz[2], Paul S. Keim[3] & Paul Wilmes[1]

Microbial communities are complex and dynamic systems that are primarily structured according to their members' ecological niches. To investigate how niche breadth (generalist versus specialist lifestyle strategies) relates to ecological success, we develop and apply an integrative workflow for the multi-omic analysis of oleaginous mixed microbial communities from a biological wastewater treatment plant. Time- and space-resolved coupled metabolomic and taxonomic analyses demonstrate that the community-wide lipid accumulation phenotype is associated with the dominance of the generalist bacterium *Candidatus* Microthrix spp. By integrating population-level genomic reconstructions (reflecting fundamental niches) with transcriptomic and proteomic data (realised niches), we identify finely tuned gene expression governing resource usage by *Candidatus* Microthrix parvicella over time. Moreover, our results indicate that the fluctuating environmental conditions constrain the accumulation of genetic variation in *Candidatus* Microthrix parvicella likely due to fitness trade-offs. Based on our observations, niche breadth has to be considered as an important factor for understanding the evolutionary processes governing (microbial) population sizes and structures *in situ*.

---

[1] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7 Avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg. [2] Institute for Systems Biology, 401 Terry Avenue North, Seattle, Washington 98109, USA. [3] TGen North, 3051 West Shamrell Boulevard, Flagstaff, Arizona 86001, USA. [4] Centre de Recherche Public-Gabriel Lippmann, 41 rue du Brill, L-4422 Belvaux, Luxembourg. * These authors contributed equally to this work. † Present address: Laboratory of Microbial Ecology and Technology, Ghent University, Coupure Links 653, B-9000 Ghent, Belgium (H.R.) or Adobe Research, 345 Park Avenue, San Jose, California 95110, USA (N.V.). Correspondence and requests for materials should be addressed to P.W. (email: paul.wilmes@uni.lu).

© 2014 Macmillan Publishers Limited. All rights reserved.

Microorganisms are ubiquitous and form complex, heterogeneous and dynamic assemblages[1]. They represent essential components of the Earth's biogeochemical cycles[2], human metabolism[3] and biotechnological processes[4]. Microbial population sizes and structures are governed by resource availability and usage[5–7], and mainly develop in response to the breadths of the fundamental and realized niches of constituent populations[8,9] (narrow for specialists; wide for generalists) albeit being influenced by stochastic neutral processes[8,9]. Microbial niche breadths remain poorly described *in situ* (for an earlier study, see ref. 10). The application of high-fidelity, high-resolution and high-throughput molecular analyses to microbial consortia holds great promise for resolving population-level phenotypes and defining their corresponding niche breadths *in situ*[11]. However, to obtain community-wide multi-omic data that can be meaningfully integrated and analysed, systematic measurements are essential[1].

We have recently developed the required laboratory methods that enable us to isolate representative biomolecular fractions (DNA, RNA, proteins and small molecules) from single microbial community samples[12,13]. Here, we expand this concept by performing integrated omic analyses of purified biomolecular fractions from oleaginous mixed microbial communities (OMMCs) located on the surface of an anoxic biological wastewater treatment tank to study how microbial lifestyle strategies relate to ecological success and the associated community-level phenotypes in this fluctuating but well-characterized environment. In addition, OMMCs are typically enriched in lipid-accumulating filamentous bacteria and often associated with operational difficulties, such as phase separation and bulking in biological wastewater treatment plants[14]. However, the phenotypic traits of OMMCs may allow for the recovery of lipids from wastewater streams for subsequent chemical energy recovery through biodiesel synthesis[15]. As for other microbial consortia, a detailed understanding of OMMC ecology is essential for the formulation of strategies to shape microbial community structure and function (in this case, enriching for lipid-accumulating bacteria) in the future. Building on the recently developed methodologies for the systematic molecular characterization of microbial consortia[12,13], and for resolving and reconstructing population-level genomic complements from community-wide sequence data[16], here, we integrate multi-omic data sets to resolve microbial lifestyle strategies *in situ*, identify finely tuned gene expression governing resource usage by a dominant bacterial generalist population and reveal that genetic variation within this population is constrained likely due to fitness trade-offs.

Results

Coupled metabolomic—taxonomic analyses over space and time. To obtain a detailed view of OMMC (Supplementary Fig. 1a) lipid accumulation and bacterial composition, we first applied coupled metabolomics and 16S rRNA gene sequencing to samples taken over space and time (see the Methods section). The initial sample set included four distinct biological replicates (Supplementary Fig. 1b) from four representative time points (two in autumn, two in winter; Methods and Supplementary Fig. 2a,b). Using gas chromatography coupled to tandem mass spectrometry (GC-MS/MS), absolute quantifications of the 14 major long-chain fatty acids were obtained for OMMC biomass and wastewater, respectively (Methods). The V3 and V6 hypervariable regions of the 16S rRNA gene were amplified from the DNA fraction of the samples (Methods). The barcoded amplicons were pyrosequenced on a 454 GS FLX platform, yielding a total of 265,592 reads ($n = 10{,}574 \pm 3{,}451$ (mean ± s.d.) per OMMC sample after quality control and chimera filtering). Direct taxonomic classification of the obtained sequencing reads demonstrates that, at the phylum level, the OMMCs of the studied treatment plant were dominated across the studied seasons by Proteobacteria and Actinobacteria, which constituted 43% ± 14% and 21% ± 5% (mean ± s.d.) of the community, respectively (Fig. 1a). Similar results were obtained when operational taxonomic unit clustering was applied before classification (Supplementary Fig. 3a). Among the two most dominating taxa over time (Fig. 1b; Supplementary Fig. 3b), *Candidatus* Microthrix spp., a well-known lipid-accumulating genus[17,18], correlated with a more pronounced community-wide lipid accumulation phenotype (Spearman correlation coefficient $\rho \geq 0.8$ for *Candidatus* Microthrix spp. and palmitoleic and oleic acids, respectively; Fig. 1c). This trend, despite the metabolic versatility

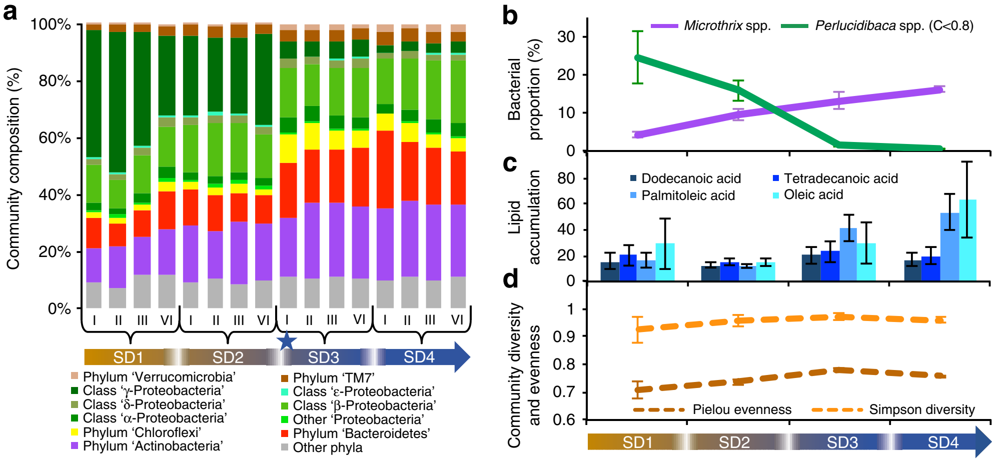


Figure 1 | Microbial community dynamics and lipid accumulation from wastewater. (**a**) Fractions of taxa identified across the communities sampled on four distinct dates (SD1–SD4). Roman numerals refer to the four biological replicates sampled per time point. The blue star indicates the representative sample from SD3 subjected to the integrated omic analysis. (**b**) Average genus-level abundances of the two dominant populations. The most abundant microbial population in winter was identified as *Candidatus* Microthrix spp., whereas a population tentatively identified (confidence level <0.8) as *Perlucidibaca* spp. was dominant in autumn. (**c**) Long-chain fatty acid intracellular accumulation per sampling date expressed as ratios between quantified intracellular and extracellular long-chain fatty acid abundances. (**d**) Genus-level alpha diversity and evenness. (**b**–**d**), error bars represent s.d. ($n = 4$).

© 2014 Macmillan Publishers Limited. All rights reserved.

of *Candidatus* Microthrix spp.[17,18] and the statistically significant lower levels of lipids available compared with other carbon and energy sources particularly in winter (Wilcoxon rank-sum test, $P<0.001$, $n=15$; Supplementary Fig. 4), suggests optimal foraging behaviour[19,20] by *Candidatus* Microthrix spp.

High-throughput multi-omic analyses. To obtain an initial view of the population-level characteristics that determine the ecological success of *Candidatus* Microthrix spp. in winter in comparison with other co-occurring microbial populations, we first conducted a detailed integrated omic analysis of a single representative sample obtained on sampling date 3 (SD3; Methods). This sample was selected on the basis of the relatively large *Candidatus* Microthrix spp. population size (Fig. 1b), the desired community-wide lipid accumulation phenotype (Fig. 1c) and its even but diverse community composition (Fig. 1d). Concomitantly isolated DNA, RNA and protein fractions were processed and subjected to high-throughput metagenomic, metatranscriptomic and metaproteomic analyses. Massive parallel sequencing of DNA and cDNA resulted in the generation of 1.47×10^7 and 1.65×10^7 metagenomic and metatranscriptomic paired-end reads, respectively. Shotgun proteomic analyses based on liquid chromatography followed by tandem mass spectrometry (LC-MS/MS) resulted in the generation of 271,915 mass spectra (Methods).

Assembly-free community profiling. To assess the composition of the winter OMMC, we first carried out an assembly-free community analysis. For this, the results obtained using two shotgun sequence data profiling tools, namely MetaPhlAn[21] and MG-RAST[22], were compared with the profiles obtained using the previous 16S rRNA gene sequencing. Given the poor taxonomic classification of *Candidatus* Microthrix spp. in the databases used by these profiling tools, all analyses were limited to phylum-level classification. At this level, similar community structures were apparent for the representative sample from SD3 using the three distinct approaches (Supplementary Fig. 3a and Methods).

Second, to resolve the functions encoded and expressed by the OMMC from SD3, the proportions of genes and transcripts belonging to different cluster of orthologous group (COG) functional categories were compared for both the metagenomic and metatranscriptomic data sets (Supplementary Fig. 5). Similar proportions were observed for most of the different functional categories in both data sets. Nevertheless, major differences were observed for the categories 'J—translation, ribosomal structure and biogenesis', 'O—posttranslational modification, protein turnover, chaperones', and 'C—energy production and conversion'. Differences in gene copy numbers and transcript abundances may be expected for these functional genes because of their typical high levels of constitutive expression. The proportion of gene copies and transcript numbers were similar for the COG category 'I—lipid transport and metabolism' although these genes are expected to have essential roles in OMMCs and, therefore, overall high levels of expression may be expected. The previous findings suggest that key members in the OMMC, that is, *Candidatus* Microthrix spp., are involved in lipid transformations. Consequently, key processes related to lipid transport and metabolism, that is, resource usage, have to be resolved at the population level. Therefore, to deconvolute the activities of the constituent OMMC members, a detailed population-resolved analysis was subsequently performed.

Population-resolved integrated omic analyses. To resolve the traits of the dominant populations within the community obtained on SD3, composite genomes (CGs) were reconstructed using a newly developed iterative binning and *de novo* assembly procedure for the combined metagenomic and metatranscriptomic sequence data (Methods). Detailed profiling and grouping of the assembled contiguous sequence fragments (contigs) were performed on the basis of centred log-ratio transformed pentanucleotide signatures and visualization using the Barnes–Hut Stochastic Neighbour Embedding (BH-SNE) algorithm[16], followed by human-augmented clustering (Methods). Using this approach, we identified nine CG groups (Fig. 2a) that displayed homogeneous G + C percentage (Fig. 2b). The assembled contigs from the nine CG groups were subjected to gene calling and annotation (Methods), which led to the identification of 23,317 coding sequences, with 16,841 and 1,533 of these being represented at the transcript and protein levels, respectively.

The refinement of the CGs by depth of read coverage resulted in the splitting of CG8 into low (CG8a) and high (CG8b) coverage populations (Supplementary Fig. 6). The average amino-

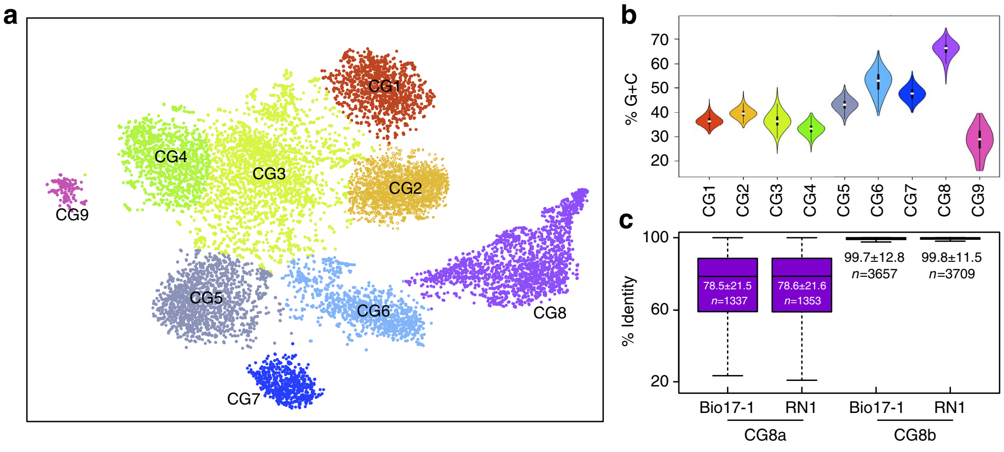


Figure 2 | Identification of genomic fragments derived from distinct populations. (**a**) Binning of assembled contigs (≥1,000 bp in length) on the basis of pentanucleotide signatures and visualization using the BH-SNE algorithm followed by human-augmented clustering of composite genome (CG) groups. (**b**) Violin plots of the G + C percentage for contigs within each of the CG groups. (**c**) Percentage amino-acid identity of the two subpopulations in CG8 (CG8a and CG8b) compared with the two sequenced *Candidatus* Microthrix parvicella (Bio17-1 (ref. 16) and RN1 (ref. 17)) genomes. The values are median ± s.d. and *n* is the number of putative orthologues identified as best BLAST hits. Boxplots represent the lower quartile, median and upper quartile. Whiskers are placed at ×1.5 interquartile range beyond the lower and upper quartiles.

© 2014 Macmillan Publishers Limited. All rights reserved.

acid sequence identities of CG8b with recently obtained genome sequences of *Candidatus* Microthrix parvicella strains Bio17-1 (ref. 17) and RN1 (ref. 18) were $>99\%$ (Fig. 2c), an identity level usually observed among strains of the same bacterial species[23]. In contrast, the CG8a identity levels compared with the same reference sequences were around 78% (Fig. 2c). Consequently, CG8b represents a *Candidatus* Microthrix parvicella population in this OMMC sample from SD3.

The identities of the other CGs were determined using 31 phylogenetic marker genes[24] resulting, for example, in the tentative identification of CG5 as a population belonging to the Moraxellaceae family of the γ-Proteobacteria (Supplementary Data 1).

Eight of the 10 reconstructed CGs were estimated to be $>60\%$ complete with CG8b, CG4 and CG5 being 97.5, 90 and 85% complete, respectively (Supplementary Data 1). On the basis of population sizes inferred from the mapping of the metagenomic read data onto the CGs (Methods), these three community members represent the first, the seventh and the fourth most abundant OMMC populations, respectively. Because CG8b and CG5 represent the most deeply covered nearly complete genomic reconstructions, a detailed analysis of the ecophysiology of these populations based on the generated functional omic data was performed.

By mapping the metatranscriptomic and metaproteomic data onto the reconstructed CGs (Methods), all 10 populations were found to be active albeit exhibiting differing levels of gene expression (Fig. 3a, Supplementary Figs 7–9, Supplementary Data 2). Observed patterns of gene expression were not necessarily consistent at the transcript and protein levels for the different CGs. Discrepancies between the levels of gene expression inferred from transcriptomic and proteomic data have been well described in eukaryotes[25] and these have also recently been observed for microbial communities[26]. The lack of correlation may have different reasons including differing molecular half-lives[26] and/or possible posttranscriptional or posttranslational modifications, which are not detectable using the transcriptomic and proteomic methodologies used in this study.

Despite its large population size, population CG8b expressed only a comparatively small fraction (45.8% of possible transcripts detected) of its genetic complement *in situ* (Fig. 3b, Supplementary Data 1, Supplementary Fig. 8) and this at a moderate level of expression both at the transcript and protein levels (Supplementary Fig. 9a), suggesting the fine-tuning of gene expression by CG8b. On the contrary, the other CGs exhibited expression of the majority of their genetic repertoire (Fig. 3, Supplementary Data 1, Supplementary Fig. 8). In particular, 92.7% of CG5 genes were detected at the transcript level.

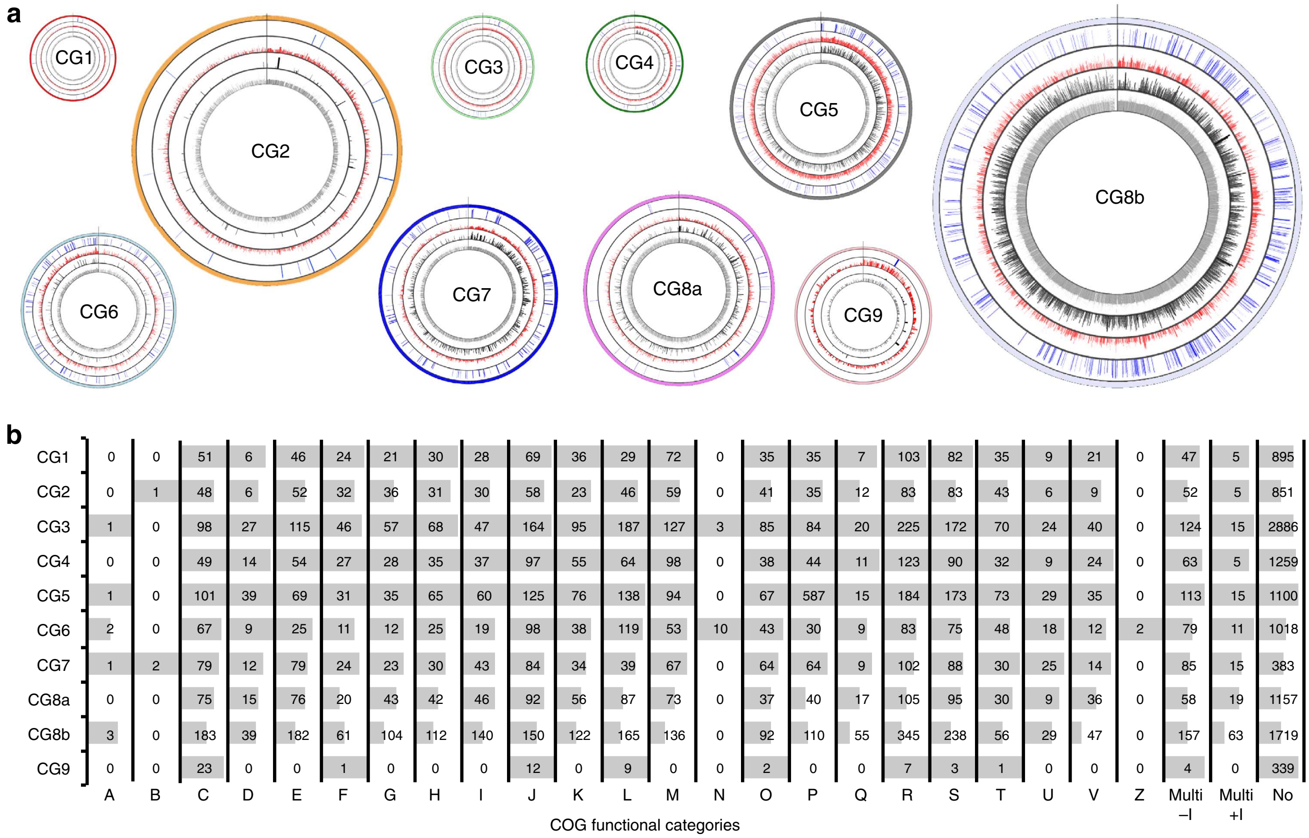
On the basis of its genetic repertoire, *Candidatus* Microthrix parvicella appears to be physiologically versatile[17,18] which, combined with its enrichment under fluctuating environmental conditions, indicates a generalist lifestyle strategy[27]. The fine-tuning of gene expression is particularly apparent for genes involved in lipid transport and metabolism, which show a clear genomic enrichment within the CG8b population although only a limited subset, that is, 46%, are expressed (Fig. 3b, Supplementary Fig. 8). Among these genes, long-chain fatty acid-CoA ligases are essential for the assimilation and activation of extracellular fatty acids into their acyl-CoA conjugates[28] and are therefore required for resource usage by *Candidatus* Microthrix parvicella. CG8b encodes 29 genes annotated as homologues of this enzyme class, indicating that a broad spectrum of free fatty acids may be assimilated by this population. Only 11 and 14 of these genes were found to be expressed at the RNA and protein levels, respectively (Fig. 4a). In contrast, the five genes annotated as long-chain fatty acid-CoA ligase homologues in CG5 were all expressed (Supplementary Data 2). This observation suggests the fine-tuned expression of these genes by *Candidatus* Microthrix parvicella, likely through the tight regulation of gene expression, to facilitate preferential resource usage in accordance with optimal foraging behaviour[19,20].

Population-level genetic diversity. To assess the overall frequencies of population-level genetic variation and determine how these variations may reflect the lifestyle strategy of CG8b, the number of single-nucleotide polymorphisms (SNPs) identified in individual CGs were normalized according to reconstructed CG length and population sizes inferred from the proportion of metagenomic reads mapped to the reconstructed CGs (Supplementary Data 1). CG8b displayed a relatively limited level of genetic variation. For example, it exhibited one order of magnitude fewer SNPs compared with CG5, the other almost complete reconstructed CG with enough coverage to confidently infer SNP densities (Supplementary Data 1). Given the generalist lifestyle strategy of CG8b, the relatively high within-population genetic homogeneity may be explained by fitness trade-offs resulting, for example, from antagonistic pleiotropy[29,30]. In a fluctuating environment, most of the beneficial or neutral mutations under one condition may be deleterious under other conditions, thereby restricting the evolutionary rate of generalists[29,30]. An alternative hypothesis positing that this low population-level variation may be due to a recent genetic bottleneck (selective sweep, colonization, and so on.) followed by population expansion[31] may be rejected on the basis of the high genetic similarity between the reconstructed CG and the two available *Candidatus* Microthrix parvicella genome sequences from strains isolated from distant wastewater treatment plants 7 and 16 years before the present study (Fig. 2c).

Fine-tuned gene expression and limited genetic diversity over time. To validate the previous snapshot views of the ecophysiology and structure of populations CG8b and CG5, identical multi-omic analyses were carried out on three additional, rationally selected OMMC samples from the same wastewater treatment tank (Methods). A first additional sample was collected on SD7 approximately a year after SD3 when the measured physico-chemical parameters were very similar to those measured on SD3 (Supplementary Fig. 2b,c). In addition, samples were selected from SD5 and SD6 because the physico-chemical parameters on these dates were at variance with SD3 and SD7 (Supplementary Fig. 2c). Importantly, the additional samples also contain populations CG8b and CG5 at sufficient quantities to obtain the necessary coverages at the genomic and transcriptomic levels for the subsequent analyses of genetic diversity and gene expression over time (Table 1). Massive parallel sequencing of DNA and cDNA resulted in the generation of an additional 5×10^7 and 4.45×10^7 metagenomic and metatranscriptomic paired-end reads, respectively. In addition, a total of 326,630 additional mass spectra were generated using LC-MS/MS (Methods).

To corroborate the fine-tuning of gene expression of the generalist population CG8b (*Candidatus* Microthrix parvicella) deduced from the analysis of the sample from SD3, patterns of gene expression were assessed for SD5–SD7. Although the CG5 population consistently expressed the vast majority of its genetic repertoire, only a comparatively small fraction of the genetic complement of CG8b was expressed at each additional time point despite its relatively consistent population size (Table 1, Supplementary Fig. 9b, Supplementary Data 3). These observations support the previous results obtained on the OMMC from SD3. In addition, analogous to the patterns observed for SD3, the

© 2014 Macmillan Publishers Limited. All rights reserved.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	Z	Multi −I	Multi +I	No
CG1	0	0	51	6	46	24	21	30	28	69	36	29	72	0	35	35	7	103	82	35	9	21	0	47	5	895
CG2	0	1	48	6	52	32	36	31	30	58	23	46	59	0	41	35	12	83	83	43	6	9	0	52	5	851
CG3	1	0	98	27	115	46	57	68	47	164	95	187	127	3	85	84	20	225	172	70	24	40	0	124	15	2886
CG4	0	0	49	14	54	27	28	35	37	97	55	64	98	0	38	44	11	123	90	32	9	24	0	63	5	1259
CG5	1	0	101	39	69	31	35	65	60	125	76	138	94	0	67	587	15	184	173	73	29	35	0	113	15	1100
CG6	2	0	67	9	25	11	12	25	19	98	38	119	53	10	43	30	9	83	75	48	18	12	2	79	11	1018
CG7	1	2	79	12	79	24	23	30	43	84	34	39	67	0	64	64	9	102	88	30	25	14	0	85	15	383
CG8a	0	0	75	15	76	20	43	42	46	92	56	87	73	0	37	40	17	105	95	30	9	36	0	58	19	1157
CG8b	3	0	183	39	182	61	104	112	140	150	122	165	136	0	92	110	55	345	238	56	29	47	0	157	63	1719
CG9	0	0	23	0	0	1	0	0	0	12	0	9	0	0	2	0	0	7	3	1	0	0	0	4	0	339

Figure 3 | Population-resolved integrated omics. (**a**) Circos plots[63] of genome-wide gene expression levels for the 10 reconstructed composite genomes (CGs). Tracks (from the innermost concentric track to the outermost): $\log_{10}$ of metagenomic fragments per 1 kb of sequence per 10^6 mapped reads (FPKM[46]; dark grey), $\log_{10}$ of the numbers of detected SNPs per gene (black), $\log_{10}$ of metatranscriptomic FPKM (red), $\log_2$ of the protein expression levels as Normalized Spectral Indices (NSI; blue), and reconstructed contigs ordered by size. The track scales are identical across the plots. The sizes of the individual Circos plots are weighted according to the $\log_{10}$ of inferred population size (Methods). (**b**) The number of genes for each COG category encoded by the different CGs and their corresponding relative transcript levels (grey bars). COG categories are: A, RNA processing and modification; B, chromatin structure and dynamics; C, energy production and conversion; D, cell cycle control, cell division, chromosome partitioning; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure and biogenesis; K, transcription; L, replication, recombination and repair; M, cell wall/membrane/envelope biogenesis; N, cell motility; O, posttranslational modification, protein turnover, chaperones; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport and catabolism; R, general function prediction only; S, function unknown; T, signal transduction mechanism; U, intracellular trafficking, secretion and vesicular transport; V, defence mechanisms; Z, cytoskeleton; Multi − I, multiple COG category excluding I; Multi + I, multiple COG category including I; No, no COG category assigned.

expression of genes involved in lipid transport and processing encoded by CG8b was highly variable over time (Fig. 4b). In contrast, the CG5 population consistently expressed the vast majority of this functional category (Fig. 4b, Supplementary Fig. 10). These comprehensive additional data reinforce the notion of finely tuned gene expression for resource usage by the *Candidatus* Microthrix parvicella generalist population.

The patterns of low SNP density in the generalist population CG8b were also consistent over time, with one order of magnitude fewer SNPs generally apparent in CG8b compared with CG5 (Table 1, Supplementary Data 4). In contrast to CG5, the variant counts of CG8b remain relatively constant over time (Table 1). The observations from the three additional time points reinforce the previous notion that a generalist lifestyle under fluctuating environmental conditions constrains the accumulation of population-level genetic variation.

Discussion

Here, the application of systematic integrated multi-omic measurements to mixed microbial communities has allowed us to obtain fundamental insights into the ecology of the constituent dominant populations. On the basis of its genetic repertoire and enrichment under temporally changing environmental conditions, the dominant population within the winter OMMCs, that is, *Candidatus* Microthrix parvicella, can be classified as a generalist species. The low proportion of genes expressed over time indicates that its ecological success most likely results from finely tuned gene expression facilitating optimal foraging behaviour. In addition, the *Candidatus* Microthrix parvicella population exhibits low levels of genetic variation that may be explained by evolutionary constraints resulting from fitness trade-offs. Elucidating the exact mechanisms driving these trade-offs in *Candidatus* Microthrix parvicella, for example, antagonistic pleiotropy or others, will require additional integrated omic data sets to be generated from many more samples taken over space and time, as well as surveys of other wastewater treatment plants. Overall, our results call for similar studies on other microbial communities to determine whether fine-tuning of gene expression is a general feature of generalists and whether lifestyle strategies provide an explanation for the varying degrees of within-population genetic heterogeneity so far observed in metagenomic data sets[32].

© 2014 Macmillan Publishers Limited. All rights reserved.

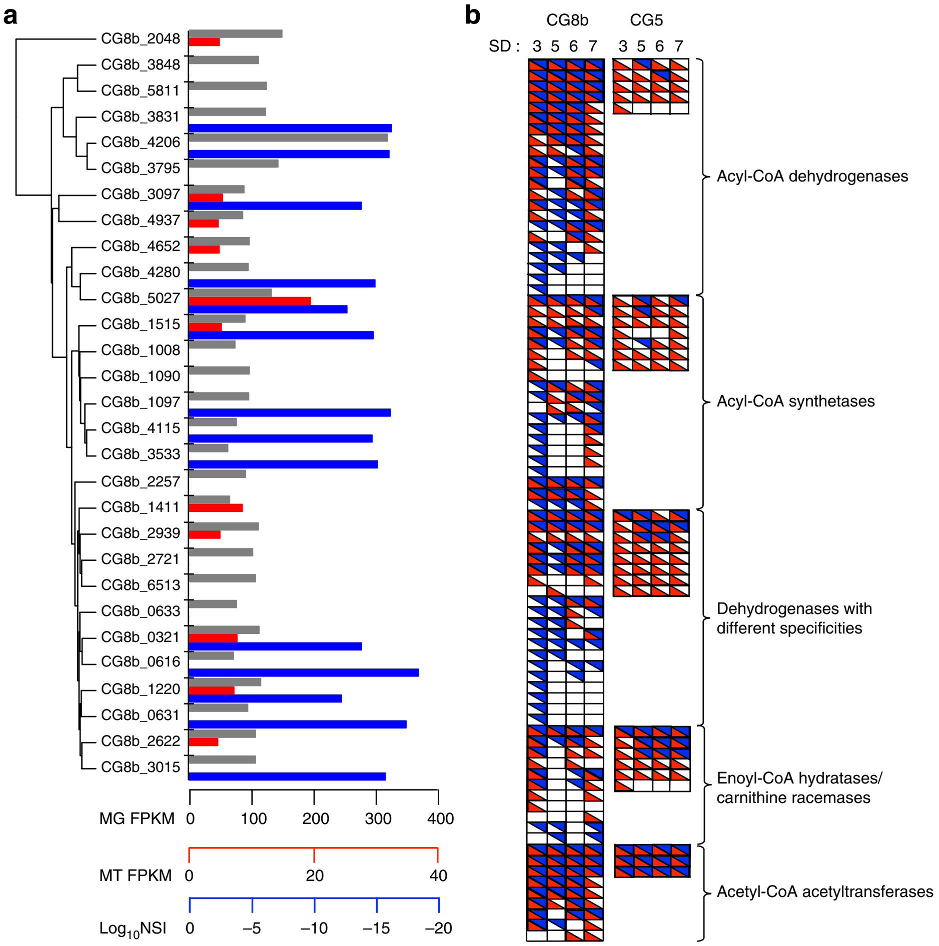


Figure 4 | *In situ* fine-tuning of gene expression by *Candidatus* Microthrix parvicella. (**a**) Dendrogram based on the amino acid similarities of 29 predicted long-chain fatty acid-CoA ligases encoded by the *Candidatus* Microthrix parvicella population (CG8b). Metagenomic (grey) and metatranscriptomic (red) data represented as fragments per 1 kb of sequence per 10^6 mapped reads (FPKM), the protein abundance data (blue) is represented as the $\log_{10}$ of the Normalized Spectral Indices (NSI). (**b**) Qualitative gene expression patterns of the five most prevalent homologous gene sets belonging to the COG category 'I—lipid transport and metabolism' encoded by *Candidatus* Microthrix parvicella (CG8b) and the CG5 population at the metatranscriptomic (red) and metaproteomic (blue) levels across four sampling time points.

Table 1 | The characteristics of composite genomes CG5 and CG8b at different sampling time points.

	SD3		SD5		SD6		SD7	
	CG5	**C8b**	**CG5**	**CG8b**	**CG5**	**CG8b**	**CG5**	**CG8b**
Average composite genome coverage (×)	9.65	23.36	30.02	45.74	20.54	65.57	35.06	81.81
Proportion of total metagenomic reads mapped per composite genome (%)	8.10	36.50	16.81	37.48	11.27	51.38	13.71	47.85
Percentage of ORFs expressed at the RNA level	92.7	45.8	78.9	25.1	85.3	32.0	87.0	36.8
Number of detected variants (based on the metagenomic data)	5,428	11,702	42,250	11,588	29,431	12,596	37,699	11,517
Number of detected variants (based on the metatranscriptomic data)	11,481	777	24,353	1,366	26,227	2,923	28,247	3,504
Variant density per CG population	2.34E – 04	7.43E – 05	8.78E – 04	7.17E – 05	9.13E – 04	5.68E – 05	9.61E – 04	5.58E – 05

Methods

Sample processing. Oleaginous mixed microbial communities (OMMCs) were sampled at four representative time points from the surface of the anoxic treatment phase of a biological wastewater treatment plant treating residential effluents (Schifflange, Esch-sur-Alzette, Luxembourg; 49°30′48.29″N; 6°1′4.53″E).
For each sampling date (SD), four distinct 'islets' were collected (herein referred to as biological replicates; Supplementary Fig. 1), transferred into a sterile tube, snap frozen on site and maintained at −80 °C until processing. Initial samples were taken on 4 October 2010 (SD1; anoxic tank wastewater temperature of 20.7 °C), 25 October 2010 (SD2; 18.9 °C), 25 January 2011 (SD3; 14.5 °C) and 23 February 2011 (SD4; 13.9 °C). Since the prevalence of OMMCs is dependent on wastewater temperature[14], these samples were chosen to be representative of the range of wastewater temperatures at which OMMCs are highly abundant within the system. Due to heavy precipitation, which leads to dispersion of the OMMC islets, and due to excess nitrate concentrations (Supplementary Fig. 2a), no samples from December/early January were included in the present study. However, given the range of water temperatures encountered in this biological wastewater treatment plant (Supplementary Fig. 2a), the October as well as January and February samples are representative of autumn and winter OMMCs, respectively.

© 2014 Macmillan Publishers Limited. All rights reserved.

For the integrated omic analyses of a representative sample, a single biological replicate (SD3-I; Fig. 1a) from the 25 January 2011 samples was selected for subsequent high-resolution omic analyses. This sample was chosen on the basis of its pronounced community-wide lipid accumulation phenotype, dominance of *Candidatus* Microthrix spp., and because it exhibited the highest bacterial diversity and evenness. This in turn should allow a comprehensive community-wide overview and reconstruction of composite genomes (CGs) from the most abundant populations within the OMMC.

To validate findings from the integrated omic analysis of the sample from SD3, three additional OMMC samples were rationally selected. On the basis of hierarchical clustering analysis of physico-chemical parameters (Supplementary Fig. 2c), a first additional sample was collected on SD7 (11 January 2012) approximately a year after SD3 when the measured physico-chemical parameters were very similar to those for SD3 (Supplementary Fig. 2b,c). In addition, samples were selected from SD5 (5 October 2011) and SD6 (12 October 2011) because the physico-chemical parameters measured on these dates (especially wastewater temperature) were at variance with those of SD3 and SD7 (Supplementary Fig. 2b,c).

Biomolecular isolation. All biomolecular fractions were obtained using a recently developed methodological framework, which allows recovery of high-quality biomolecular fractions (DNA, RNA, proteins, polar and non-polar metabolites) from unique undivided samples[12,13]. For biomacromolecular purification we used the AllPrep DNA/RNA/Protein Mini kit (Qiagen). Resulting biomolecular fractions comprising genomic DNA, RNA, proteins and small molecules were further processed and analysed as detailed below.

Quantification of biomolecular resources. Intracellular and extracellular non-polar metabolite fractions of the four biological replicates collected on SD1 to SD4 were obtained using the biomolecular extraction procedure described earlier[12,13] (only on three biological replicates for SD3). The non-polar phases were aliquoted in four vials (analytical replicates) of 100 µl each, dried overnight and the resulting pellets were then preserved at −80 °C. The intracellular and extracellular dried extracts were redissolved in 100 and 40 µl of dichloromethane, respectively. Derivatisation was carried out on 40 µl of solubilized extract with 40 µl of N,O-bis(trimethylsilyl)trifluoroacetamide:trimethylchlorosilane 99:1 (Sylon BFT, Supelco) for 1 h at 70 °C. Samples were analysed by gas chromatography coupled to tandem mass spectrometry (GC-MS/MS) on a Thermo Trace GC and a Thermo TSQ Quantum XLS triple-quadrupole MS (Thermo Fisher). Samples were injected in PTV splitless mode onto a Rxi-5Sil MS column (20 m × 0.18 mm × 0.18 µm, Restek). Helium was used as the carrier gas at a constant flow rate of 1.0 $\mathrm{ml\,min^{-1}}$. Metabolite detection was performed in Multiple Reaction Monitoring mode, with two Multiple Reaction Monitoring transitions per target compound. Quantification was carried out by external calibration using standard mixtures of pure hexanoic acid, octanoic acid, decanoic acid, dodecanoic acid, tetradecanoic acid, palmitoleic acid, hexadecanoic acid, linoleic acid, oleic acid, linolenic acid, octadecanoic acid, eicosanoic acid, docosanoic acid and tetracosanoic acid, respectively (Sigma-Aldrich).

Total carbohydrate and protein quantities were determined on supernatant of the same samples comprising 200 mg of OMMC biomass for each sampling date using a Total Carbohydrate Assay Kit and a Total Protein Assay Kit (Micro Lowry, Peterson's Modification; Sigma-Aldrich) according to the manufacturer's instructions.

16S rRNA amplicon sequencing and analysis. *Amplification*. The wet-laboratory and bioinformatic procedures for analysing the bacterial community composition based on the V3–V6 region of the bacterial 16S rRNA gene are described in detail elsewhere[33]. Briefly, we generated barcoded V3–V6 amplicons using broad-coverage fusion PCR primers (forward primer: 5′-CCATCTCATCCCTGCGTGTCTCCGACTCAG*nnnnnnnn***CCTACGGGDGGCWGCA**-3′ and reverse primer 5′- CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**CTGACGACRRCCRTGCA**-3′ with the underlined portion denoting FLX Lib-L adaptor sequences, italicized portion denoting the sample-specific 8-nt barcode sequences and bold portion denoting 16S rRNA gene primer sequences) on 10 µl of DNA extracts in 50 µl PCR reactions. The barcoded amplicons were pooled and sequenced on a Roche/454 Genome Sequencer FLX platform (Roche Applied Sciences). Resulting pyrosequencing data underwent processing and stringent filtering, which included chimera-checking, demultiplexing and quality-based trimming.

Direct classification. The processed pyrosequences were classified at each taxonomic level using a bootstrap confidence level of ≥80 and using a re-trained version of the Ribosomal Database Project (RDP) Naive Bayesian Classifier 2.4 (refs 34,35), which includes the genus *Candidatus* Microthrix as a separate taxon. The training set consisted of the RDP 16S rRNA training set #9, with sequences S001942289, S000724117, S000724133, S001448117, S001448118, S001942070, S001942206, S002416756, S002416776, S000014283, S000588187, S000588192, S000010408, S000011228, S000021841, S000267158, S000588182, S000588183, S000588184, S000588185, S000588186, S000588188, S000588189, S000588190, S000588191, S000588193, S000724113, S000724122, S000832952, S000935760, S001294363, S001942173 reclassified as bacteria > Actinobacteria > Actinobacteria > unclassified > unclassified *Candidatus* Microthrix, by placing the full 16S rRNA gene sequence of the recently sequenced strain Bio17-1 (ref. 17) into the same taxon. Classification results from each sample were used to produce an abundance matrix for data analysis.

The 16S rRNA gene-based data of the four different biological replicates (islets) per sampling date were used for calculating Simpson diversity and Pielou evenness indices with 10 replications of subsampling of 6,359 reads per sample using the R Vegan package.

Operational taxonomic unit-based classification. In addition to the direct classification, processed pyrosequences were also analysed by clustering the reads into operational taxonomic units using Mothur[36] v.1.32.1. To allow appropriate sample-specific classifications, the *Candidatus* Microthrix parvicella Bio17-1 (ref. 17) 16S rRNA gene sequence was added to the Mothur-formatted version of the RDP training set v9 and the related taxonomy file. Operational taxonomic units clustering was performed at a cut-off level of 0.03 before the assignment of taxonomy. Scripts are available from the authors upon request.

Metagenome and metatranscriptome sequencing and assembly. *DNA library preparation*. The purified DNA fractions[12,13] from the unique biological replicates I of SD3 and from SD5 to SD7 suspended in elution buffer (pH 8.0) were used to prepare a paired-end library with the AMPure XP/Size Select Buffer Protocol as previously described by Kozarewa *et al.*[37], modified to allow for size selection of fragments using the double solid phase reversible immobilization procedure described earlier[38]. Size selection yielded metagenomic library fragments with a mean size of 450 bp. All enzymatic steps in the protocol were performed using the Kapa Library Preparation Kit (Kapa Biosystems) with the addition of 1 M PCR-grade betaine in the PCR reaction to aid in the amplification of high G + C percentage content templates.

RNA library preparation. Following RNA purification[12,13] from the unique biological replicates I of SD3 and from SD5 to SD7, RNA fractions were ethanol precipitated, overlayed with RNAlater solution (Ambion) and stored at −80 °C. Before sequencing library preparation, the RNA pellet was rinsed twice in 80% ethanol and twice in 100% ethanol to remove any excess RNAlater solution. The pellet was then left on ice to dry. After ethanol evaporation, the RNA pellets were resuspended in 1 mM sodium citrate buffer at pH 6.4. Ribosomal RNAs were depleted using the Ribo-Zero Meta-Bacteria rRNA Removal Kit (Epicentre) according to the manufacturer's instructions. Transcriptome libraries were subsequently prepared using the ScriptSeq v2 RNA-Seq Library Preparation Kit (Epicentre) according to the manufacturer's instructions. The resulting cDNA was subjected to Illumina sequencing.

Nucleic acid sequencing. Nucleic acid fractions were sequenced on an Illumina Genome Analyser (GA) IIx sequencer. Massive parallel sequencing of DNA and cDNA resulted in the generation of 1.47×10^7 and 1.65×10^7 metagenomic and metatranscriptomic paired-end reads for SD3, respectively. Similarly, the sequencing of SD5-derived DNA and cDNA generated 1.57×10^7 and 1.47×10^7 metagenomic and metatranscriptomic paired-end reads, SD6-derived DNA and cDNA sequencing generated 1.47×10^7 and 1.48×10^7 metagenomic and metatranscriptomic paired-end reads and SD7-derived DNA and cDNA sequencing generated 1.96×10^7 and 1.80×10^7 metagenomic and metatranscriptomic paired-end reads.

Nucleic acid sequence data analysis. MetaPhlAn[21] (using default parameters) was used on 5' seven base pairs hard-clipped raw paired-end reads, collapsed, filtered at or above a mean QV of 30 and a minimum length of 60 bp.

Raw metagenomic paired-end reads were submitted to MG-RAST[22] using the 'join fastq-formatted paired-reads' option retaining the non-overlapping reads, dynamic trimming and dereplication options. Raw metatranscriptomic reads were submitted to MG-RAST as described for metagenomic data, except that the dereplication option was not selected. As MG-RAST also supports the analysis of eukaryotic sequences, to allow comparison to MetaPhlAn and the 16S rRNA gene sequencing results, the MG-RAST output was filtered to only include bacterial and archaeal taxa. MG-RAST complete functional annotations of both the metagenomic and metatranscriptomic data were used for the assembly-free analysis of the community function.

Apart from these assembly-free community analyses, any overlapping paired-end reads from SD3 were joined with PANDASeq[39] (with threshold parameter $t = 0.9$) before the removal of potential PCR duplicates using custom scripts (available upon request). Read clipping, quality trimming and filtering of sequence reads was performed with the trim-fastq.pl script from the PoPoolation suite[40]. Four base pairs were hard-clipped from the 5' of all raw reads, and reads were filtered at or above a mean QV of 30 , and a minimum length of 40 bp. The quality of the resulting reads and the presence of remaining adaptor sequence contamination were assessed using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

To reduce the sample complexity and to improve the efficiency of the assembly process, quality-filtered, combined metagenomic and metatranscriptomic reads were initially mapped against the draft genome sequence of *Candidatus* Microthrix parvicella Bio17-1 (ref. 17). Mapped reads were extracted from the pool of reads and assembled separately with IDBA-UD[41] (v.1.1.10), with the following parameters: --pre_correction --mink 35 --maxk 75 --step 2 --num_threads 12 --similar 0.97 --min_count 3. The remaining unmapped reads were binned as pairs

© 2014 Macmillan Publishers Limited. All rights reserved.

according to low and high G + C percentage content, with an inclusive cut-off value of 50% G + C, and assembled as above. This strategy resulted in the generation of 1,739,837 additional base pairs of assembled sequence data compared with a direct assembly. Assemblies were merged using minimus and scaffolded using Bambus2 (AMOS tool suite[42]).

Assembled contigs longer than 500 bp, representing the first set, were grouped by a reference-free binning algorithm that has recently been developed by some of the authors[16]. The algorithm first computes the pentanucleotide frequency of each contig, which then allows representation of each contig as a point in a 512-dimensional Euclidean space (512 is the number of unique pentanucleotides after taking reverse complements into account). After a centred log-ratio transformation on each point[43], the sets of points were used as input for the Barnes–Hut Stochastic Neighbourhood Embedding (BH-SNE) algorithm[44], which produced a two-dimensional map (embedding) of the original signatures (Fig. 2a). Binning of points was then carried out using the Expectation–Maximisation (EM) algorithm on a postulated two-dimensional Gaussian Mixture model[45], where the means of the Gaussian components of interest were initialized by the user and the covariance matrices were initialized by diagonal matrices with small positive entries. For the results reported in this work, we initialized EM with one Gaussian component per expected cluster following visual inspection. Contigs from the resulting clusters were extracted as contig groups and used as reference sequences to recruit sequence reads from the original, quality-filtered data set. Non-mapped paired-end reads were extracted and merged. A second iterative round of assembly was performed on each set of recruited reads separately and BH-SNE profiling was conducted as described above (except that this time a minimal contig size of 1,000 bp was used). Contig groups resulting from the second BH-SNE iteration were used once more for read recruitment. G + C percentage was calculated per base using in-house scripts (available upon request).

For coverage and gene expression analyses (SD3, SD5, SD6 and SD7), metagenomic and metatranscriptomic reads were mapped onto reconstructed genomic fragments from SD3 using Bowtie2 (ref. 46) (using 'very sensitive-local' parameters) and BWA[47] (using default parameters except for the –M option). Gene expression levels were determined using Cufflinks[48] on the basis of the BWA read mappings.

Metagenomic FPKM[48] (fragments per 1 kb of sequence per 10^6 mapped reads) and coverage values corresponding to each predicted gene in each of the CGs were obtained for the different time points (Supplementary Data 2 and Supplementary Data 3).

To estimate relative population sizes within the community, we devised a measure analogous to the RPKM[49] (reads per 1 kb of sequence per 10^6 mapped reads) measure, widely used for reporting the normalized abundance of, for example, transcripts and we defined this as follows:

$$N_i = \frac{c_i \times 10^6}{C \times l_i}$$

where N_i is the relative size of the population corresponding to CGi; c_i is the number of reads mapped to CGi; C is the total number of metagenomic reads mappable to all of the CGs; and l_i is the length of CGi in bp.

To account for the differences in observed expression levels linked to differing population sizes and to allow comparative analyses between different CGs as well as the different time points, genes were only considered to be expressed per CG per time point, if their metatranscriptomic FPKM values were $\geq 50 \times N_i$ (Supplementary Data 2 and Supplementary Data 3).

Metaproteome processing and analysis. Five microlitres of the protein extract obtained from SD3, SD5, SD6 and SD7 as previously described[12,13] were mixed with 1.25 µl of XT sample buffer and 0.25 µl of XT reducing agent (Bio-Rad). After 10 min of denaturation at 70 °C, 5 µl the sample was subsequently separated by 1D SDS–PAGE (Criterion precast 1D gel, Bio-Rad). The gel was then stained with Imperial stain (Coomassie-Blue R250, Thermo Fisher Scientific) and cut into uniform 2 mm bands[50]. After in-gel reduction and alkylation, tryptic digestion was performed. Resulting peptides were separated by liquid chromatography (LC) using an Easy-nLC column (Proxeon, Thermo Fisher Scientific). Separation was performed using a 75 µm ID fused silica column packed with 20 cm of ReproSil Pur C18-AQ 3 µm beads (Dr Maisch). Before column separation, the samples were loaded onto a fritted 100 µm ID fused silica trap packed with 2 cm of the same material. The peptide mixture was separated using a binary solvent gradient to elute the peptides. Solvent A was 0.1% formic acid in water. Solvent B was 0.1% formic acid in acetonitrile. The peptide fractions were pooled in consecutive pairs, concentrated and resuspended up to 20 µl in solvent A. Eight microlitres of each pooled sample was injected per LC analysis. The three-step elution programme was operated at a flow rate of 0.3 ml min^{-1} consisting of (1) a gradient from 2 to 35% solvent B over 60 min, (2) a 10-min wash at 80% solvent B and (3) a 20-min column re-equilibration step at 2% solvent B.

Mass spectra were acquired on an LTQ-Orbitrap Elite (Thermo Fisher Scientific). The instrument was operated on an 11-scan cycle consisting of a single Fourier transformed (FT) precursor scan at 30,000 resolution followed by 10 data-dependent MS/MS scan events using higher-energy collisional dissociation at 15,000 resolution in the FT Orbitrap. The precursor scans had a mass range of 300–2,000 m/z, and an automatic gain control setting of 10^6 ions. The MS/MS scans were performed using a normalized collision energy of 35 and an isolation width of 2 m/z. The data-dependent settings included monoisotopic precursor selection and charge state filtering that excluded unassigned and single charge states. Dynamic exclusion was enabled with a repeat count of 1, a repeat duration of 10 s, an exclusion list size of 500 and an exclusion duration of 180 s. Exclusion mass width was ± 5 p.p.m. relative to mass.

LC-MS/MS analysis resulted in the generation of 271,915 mass spectra for SD3, 118,386 mass spectra for SD5, 102,916 mass spectra for SD6 and 105,328 mass spectra for SD7.

Composite genome and expression analyses. *Gene calling and annotation*. The assembled CGs were submitted for gene calling and annotation to RAST[51] with default parameters except for Domain (Bacteria), Genetic code (11), Sequencing method (other), FIGfam version (release 63) and with the Build metabolic model option selected.

Taxonomic affiliation. The taxonomies of the reconstructed CGs were determined using the AmphoraNet[24] webserver. A taxon name was assigned when at least 75% of the identified marker genes resulted in a concordant taxonomy, and a putative taxon name was assigned when at least 50% of the identified markers resulted in a concordant taxonomy.

Completeness and composition of composite genomes. Genome completeness of the reconstructed CGs was estimated on the basis of 40 universal single copy genes[52]. For this, the functional annotation of the predicted proteins in each CG was obtained using the WebMGA server[53] using the 'cog' analysis option. Functional compositions of the CGs and of their expressed genes were obtained from COG category counts, which were normalized by the total number of predicted features per CG.

Comparative analysis of Candidatus Microthrix parvicella-like sequences. Draft genome sequences for *Candidatus* Microthrix parvicella strains Bio17-1 (ref. 17) and RN1 (ref. 18) were obtained from the GenBank database (Assembly ID GCA_000299415.1 and GCA_000455525.1, respectively). Sets of orthologous genes were built using RAST's 'sequence based comparison' tool.

Variant identification. SNPs were identified by separately mapping metagenomic and metatranscriptomic reads against the reconstructed CG assemblies using Bowtie2 and BWA (as described above). SNPs were identified from each of the mappings using mpileup (SAMtools[54]), the UnifiedGenotyper (Genome Analysis Tool Kit[55]) and Freebayes[56]. The intersection of identified SNPs from all the aforementioned methods was obtained using the vcf-isec utility from the VCFtools suite and was considered for subsequent analyses. SD3 variant amino-acid sequences were included in the amino-acid sequence databases generated on the basis of called and annotated genes. This database was used for subsequent protein identification on the basis of the generated metaproteomic data (see below). Variant frequencies were separately estimated from mapped metagenomic and metatranscriptomic reads. Only variants in regions with a minimum read depth (coverage) of 10 for both metagenomic and metatranscriptomic data were considered. Variant density per CG population was calculated by normalizing the SNP density (number of SNPs per kb) by the relative population size, which in turn was inferred from the fraction of metagenomic sequencing reads mapped onto the individual genomic reconstructions.

Protein identification. MS/MS spectra were searched against the generated amino-acid sequence database (containing the predicted proteins including all variants of the 10 reconstructed CGs and common contaminants) using the X!Tandem algorithm[57]. The resulting peptide identifications were validated using the Trans-Proteomic Pipeline[58]. The X!Tandem parameters included precursor and fragment ion mass tolerances of 15 p.p.m., a static modification of 57.021464 Da on cysteine residues and a potential modification mass of 15.994915 Da on methionine residues. The search allowed for semi-tryptic cleavages up to two missed cleavages. The database search results were validated and proteins were inferred at ~1% false discovery rate using the PeptideProphet, ProteinProphet and iProphet tools from the Trans-Proteomic Pipeline software suite[58–60].

Protein quantification. Relative protein quantification was performed using the normalized spectral index (NSI) measure using an in-house software tool called NSICalc (details available upon request). The tool was adapted from the method by Griffin *et al.*[61] Briefly, the NSI combines peptide count, spectral count and MS/MS fragment ion intensity for quantification and normalizes these values by the length of each protein. This strategy incorporates measurable peptide intensities while removing some of the biases of using spectral counts when comparing large and small proteins. NSI values were $\log_2$ normalized before comparison across proteins to obtain relative quantification ratios.

Metaproteomic analyses led to peptide matching against the amino-acid database of 43,214 spectra, which in turn provided abundance data on a total of 1,815 proteins for SD3.

Analysis of the long-chain fatty acid-CoA ligases of CG8b. Amino-acid sequences of genes annotated as long-chain-fatty-acid-CoA ligases by RAST from CG8b were aligned using Expresso[62] using default parameters. Sequence similarities were determined using the SIAS server (http://imed.med.ucm.es/Tools/sias.html). A dendrogram based on pairwise comparisons of amino-acid sequence similarities was obtained using the hierarchical clustering function in R. Abundance values were extracted from the mapped metagenomic, metatranscriptomic and metaproteomic data.

© 2014 Macmillan Publishers Limited. All rights reserved.

References

1. Muller, E. E. L., Glaab, E., May, P., Vlassis, N. & Wilmes, P. Condensing the omics fog of microbial communities. *Trends Microbiol.* **21,** 325–333 (2013).
2. Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's biogeochemical cycles. *Science* **320,** 1034–1039 (2008).
3. Nicholson, J. K., Holmes, E. & Wilson, I. D. Gut microorganisms, mammalian metabolism and personalized health care. *Nat. Rev. Microbiol.* **3,** 431–438 (2005).
4. Rittmann, B. E. Microbial ecology to manage processes in environmental biotechnology. *Trends Biotechnol.* **24,** 261–266 (2006).
5. Waldrop, M. P., Zak, D. R., Blackwood, C. B., Curtis, C. D. & Tilman, D. Resource availability controls fungal diversity across a plant diversity gradient. *Ecol. Lett.* **9,** 1127–1135 (2006).
6. Langenheder, S. & Prosser, J. I. Resource availability influences the diversity of a functional group of heterotrophic soil bacteria. *Environ. Microbiol.* **10,** 2245–2256 (2008).
7. Rasche, F. *et al.* Seasonality and resource availability control bacterial and archaeal communities in soils of a temperate beech forest. *ISME J.* **5,** 389–402 (2011).
8. Dumbrell, A. J., Nelson, M., Helgason, T., Dytham, C. & Fitter, A. H. Relative roles of niche and neutral processes in structuring a soil microbial community. *ISME J.* **4,** 1078–1078 (2010).
9. Jeraldo, P. *et al.* Quantification of the relative roles of niche and neutral processes in structuring gastrointestinal microbiomes. *Proc. Natl Acad. Sci. USA* **109,** 9692–9698 (2012).
10. Mou, X., Sun, S., Edwards, R. A., Hodson, R. E. & Moran, M. A. Bacterial carbon processing by generalist species in the coastal ocean. *Nature* **451,** 708–711 (2008).
11. Wilmes, P. *et al.* Metabolome-proteome differentiation coupled to microbial divergence. *MBio* **1,** e00246-10 (2010).
12. Roume, H. *et al.* A biomolecular isolation framework for eco-systems biology. *ISME J.* **7,** 110–121 (2013).
13. Roume, H., Heintz-Buschart, A., Muller, E. E. L. & Wilmes, P. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. *Methods Enzymol.* **531,** 219–236 (2013).
14. Rossetti, S., Tomei, M. C., Nielsen, P. H. & Tandoi, V. '*Microthrix parvicella*', a filamentous bacterium causing bulking and foaming in activated sludge systems: a review of current knowledge. *FEMS Microbiol. Rev.* **29,** 49–64 (2005).
15. Muller, E. E. L., Sheik, A. R. & Wilmes, P. Lipid-based biofuel production from wastewater. *Curr. Opin. Biotechnol.* **30,** 9–16 (2014).
16. Laczny, C. C., Pinel, N., Vlassis, N. & Wilmes, P. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Sci. Rep.* **4,** 4516 (2014).
17. Muller, E. E. L. *et al.* Genome sequence of '*Candidatus* Microthrix parvicella' Bio17-1, a long-chain-fatty-acid-accumulating filamentous actinobacterium from a biological wastewater treatment plant. *J. Bacteriol.* **194,** 6670–6671 (2012).
18. Jon McIlroy, S. *et al.* Metabolic model for the filamentous '*Candidatus Microthrix parvicella*' based on genomic and metagenomic analyses. *ISME J.* **7,** 1161–1172 (2013).
19. Emlen, J. M. The role of time and energy in food preference. *Am. Nat.* **100,** 611 (1966).
20. MacArthur, R. H. & Pianka, E. R. On optimal use of a patchy environment. *Am. Nat.* **100,** 603–609 (1966).
21. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9,** 811–814 (2012).
22. Meyer, F. *et al.* The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9,** 386 (2008).
23. Konstantinidis, K. T. & Tiedje, J. M. Towards a genome-based taxonomy for prokaryotes. *J. Bacteriol.* **187,** 6258–6264 (2005).
24. Kerepesi, C., Bánky, D. & Grolmusz, V. AmphoraNet: the webserver implementation of the AMPHORA2 metagenomic workflow suite. *Gene* **533,** 538–540 (2014).
25. Pradet-Balade, B., Boulmé, F., Beug, H., Müllner, E. W. & Garcia-Sanz, J. A. Translation control: bridging the gap between genomics and proteomics? *Trends Biochem. Sci.* **26,** 225–229 (2001).
26. Urich, T. *et al.* Microbial community structure and functioning in marine sediments associated with diffuse hydrothermal venting assessed by integrated meta-omics. *Environ. Microbiol.* **16,** 2699–2710 (2013).
27. Kassen, R. The experimental evolution of specialists, generalists, and the maintenance of diversity. *J. Evol. Biol.* **15,** 173–190 (2002).
28. Black, P. N. & DiRusso, C. C. Transmembrane movement of exogenous long-chain fatty acids: proteins, enzymes, and vectorial esterification. *Microbiol. Mol. Biol. Rev.* **67,** 454–472 (2003).
29. Whitlock, M. C. The red queen beats the jack-of-all-trades: the limitations on the evolution of phenotypic plasticity and niche breadth. *Am. Nat.* **148,** S65–S77 (1996).
30. Elena, S. F. & Lenski, R. E. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat. Rev. Genet.* **4,** 457–469 (2003).
31. Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G. & Hanage, W. P. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323,** 741–746 (2009).
32. Wilmes, P., Simmons, S. L., Denef, V. J. & Banfield, J. F. The dynamic genetic repertoire of microbial communities. *FEMS Microbiol. Rev.* **33,** 109–132 (2009).
33. Liu, C. M. *et al.* Male circumcision significantly reduces prevalence and load of genital anaerobic bacteria. *MBio* **4,** e00076 (2013).
34. Cole, J. R. *et al.* The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37,** D141–D145 (2009).
35. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73,** 5261–5267 (2007).
36. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75,** 7537–7541 (2009).
37. Kozarewa, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of GC-biased genomes. *Nat. Methods* **6,** 291–295 (2009).
38. Rodrigue, S. *et al.* Unlocking short read sequencing for metagenomics. *PLoS ONE* **5,** e11840 (2010).
39. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13,** 31 (2012).
40. Kofler, R. *et al.* PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS ONE* **6,** e15925 (2011).
41. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28,** 1420–1428 (2012).
42. Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S. & Pop, M. Next generation sequence assembly with AMOS. *Curr. Protoc. Bioinformatics* Chapter 11, Unit 11.8 (2011).
43. Aitchison, J. *The Statistical Analysis of Compositional Data* (Blackburn Press, 2003).
44. Van der Maaten, L. Barnes-Hut-SNE. in *Proceedings of the International Conference on Learning Representations* arXiv:1301.3342 (2013).
45. Redner, R. A. & Walker, H. F. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26,** 195–239 (1984).
46. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
47. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
48. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
49. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5,** 621–628 (2008).
50. Simpson, R. J. *et al.* Proteomic analysis of the human colon carcinoma cell line (LIM 1215): development of a membrane protein database. *Electrophoresis* **21,** 1707–1732 (2000).
51. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9,** 75 (2008).
52. Stark, M., Berger, S. A., Stamatakis, A. & von Mering, C. MLTreeMap—accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics* **11,** 461 (2010).
53. Wu, S., Zhu, Z., Fu, L., Niu, B. & Li, W. WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* **12,** 444 (2011).
54. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
55. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
56. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *ArXiv e-print* 1207:3907 (2012).
57. Craig, R. & Beavis, R. C. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* **17,** 2310–2316 (2003).
58. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75,** 4646–4658 (2003).
59. Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **74,** 5383–5392 (2002).

© 2014 Macmillan Publishers Limited. All rights reserved.

60. Shteynberg, D. *et al.* iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics* **10,** M111.007690 (2011).
61. Griffin, N. M. *et al.* Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat. Biotechnol.* **28,** 83–89 (2010).
62. Armougom, F. *et al.* Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Res.* **34,** W604–W608 (2006).
63. Lin, J. *et al.* POMO—plotting omics analysis results for multiple organisms. *BMC Genomics* **14,** 918 (2013).

Acknowledgements

We thank the scientists and technical staff of the Luxembourg Centre for Systems Biomedicine and TGen North, particularly Dr René Hussong and Dr Owain Moss for their assistance and support; Mr Bissen and Mr Di Pentima from the Syndicat Intercommunal à Vocation Ecologique (SIVEC), for their permission to collect samples and access to the monitoring platform of the Schifflange wastewater treatment plant; Joëlle Fritz, Sheri Simmons and Rudi Balling for their comments on the manuscript. Bioinformatic analyses presented in this paper were carried out using the high-performance computing facilities of the University of the Luxembourg. This work was supported by an ATTRACT programme grant (ATTRACT/A09/03) and a European Union Joint Programming in Neurodegenerative Diseases grant (INTER/JPND/12/01) both to P.W., and Aide à la Formation Recherche (AFR) grants to H.R. (PHD/MARP-04), C.C.L. (PHD/4964712), S.N. (PHD-2014-1/7934898), L.W. (PHD-2013-1/BM) and E.E.L.M. (PRD-2011-1/SR), all funded by the Luxembourg National Research Fund (FNR). The project received financial support from the Integrated Biobank of Luxembourg with funds from the Luxembourg Ministry of Higher Education and Research. The work of N.P., J.L. and P.M. was funded by the 'Plan Technologies de la Santé du Gouvernement du Grand-Duché de Luxembourg' through the Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg. The work was also, in part, supported by US federal funds from the National Science Foundation (MRI grant No. 0923536), by funds from the National Institutes of Health and by funds from the US National Institute of General Medical Sciences (grant No. 2P50 GM076547/Center for Systems Biology and S10RR027584) to R.L.M. and M.R.H.

Author contributions

E.E.L.M., H.R., L.A.L. and P.W. sampled the treatment plant and extracted biomolecules; E.E.L.M., N.P., C.C.L., S.N., J.L., P.M., N.D.H., A.H.-B., L.W., C.M.L., N.V. and P.W. analysed the (meta)genomic and/or metatranscriptomic data; E.E.L.M., N.P., M.R.H. and R.L.M. analysed the metaproteomic data; E.E.L.M., N.P. and C.G. analysed the metabolomic data; E.E.L.M, N.P. and P.W. designed the study and wrote the manuscript. All the authors discussed the results and commented on the manuscript.

Additional information

Accession codes: All raw sequence data have been deposited in the NCBI Bioproject with the accession code PRJNA230567. MG-RAST-based analysis results of the metagenomic and metatranscriptomic data can be found at http://metagenomics.anl.gov/linkin.cgi?-metagenome=4566023.3 and http://metagenomics.anl.gov/linkin.cgi?metagenome=4566620.3, respectively. RAST IDs from 6666666.37195 to 6666666.37203 correspond to CG1 to CG9 assemblies and annotations are available through the RAST guest account at http://rast.nmpdr.org, using 'guest' as login as well as password. All proteomic data has been deposited in the PeptideAtlas mass spectrometry raw file repository at http://www.peptideatlas.org/ under the 'datasetIdentifier' PASS00359, PASS00576, PASS00577 and PASS00578 for SD3, SD5, SD6 and SD7, respectively.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Muller, E. E. L. *et al.* Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nat. Commun.* 5:5603 doi: 10.1038/ncomms6603 (2014).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© 2014 Macmillan Publishers Limited. All rights reserved.

A.4 Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks.

Hugo Roume, Anna Heintz-Buschart, Emilie E.L. Muller, Patrick May, Venkata P. Satagopam, Cédric C. Laczny, **Shaman Narayanasamy**, Laura A. Lebrun, Michael R. Hoopmann, Jim M. Schupp, John D. Gillece, Nathan D. Hicks, David M. Engelthaler, Thomas Sauter, Paul S. Keim, Robert L. Moritz, and Paul Wilmes
2015
NPJ Biofilms and Microbiomes **1**: 15007
DOI: 10.1038/npjbio

Contributions of author include:

- Data analysis
- Revision of manuscript

npj Biofilms and Microbiomes

www.nature.com/npjbiofilms
All rights reserved 2055-5008/15

ARTICLE OPEN

Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks

Hugo Roume[1,5,6], Anna Heintz-Buschart[1,6], Emilie EL Muller[1], Patrick May[1], Venkata P Satagopam[1], Cédric C Laczny[1], Shaman Narayanasamy[1], Laura A Lebrun[1], Michael R Hoopmann[2], James M Schupp[3], John D Gillece[3], Nathan D Hicks[3], David M Engelthaler[3], Thomas Sauter[4], Paul S Keim[3], Robert L Moritz[2] and Paul Wilmes[1]

BACKGROUND: Mixed microbial communities underpin important biotechnological processes such as biological wastewater treatment (BWWT). A detailed knowledge of community structure and function relationships is essential for ultimately driving these systems towards desired outcomes, e.g., the enrichment in organisms capable of accumulating valuable resources during BWWT.
METHODS: A comparative integrated omic analysis including metagenomics, metatranscriptomics and metaproteomics was carried out to elucidate functional differences between seasonally distinct oleaginous mixed microbial communities (OMMCs) sampled from an anoxic BWWT tank. A computational framework for the reconstruction of community-wide metabolic networks from multi-omic data was developed. These provide an overview of the functional capabilities by incorporating gene copy, transcript and protein abundances. To identify functional genes, which have a disproportionately important role in community function, we define a high relative gene expression and a high betweenness centrality relative to node degree as gene-centric and network topological features, respectively.
RESULTS: Genes exhibiting high expression relative to gene copy abundance include genes involved in glycerolipid metabolism, particularly triacylglycerol lipase, encoded by known lipid accumulating populations, e.g., *Candidatus Microthrix parvicella*. Genes with a high relative gene expression and topologically important positions in the network include genes involved in nitrogen metabolism and fatty acid biosynthesis, encoded by *Nitrosomonas* spp. and *Rhodococcus* spp. Such genes may be regarded as 'keystone genes' as they are likely to be encoded by keystone species.
CONCLUSION: The linking of key functionalities to community members through integrated omics opens up exciting possibilities for devising prediction and control strategies for microbial communities in the future.

npj Biofilms and Microbiomes (2015) **1,** 15007; doi:10.1038/npjbiofilms.2015.7; published online 17 June 2015

INTRODUCTION

Our ability to study microbial communities in natural settings as well as in engineered systems, e.g., biological wastewater treatment (BWWT) plants, has dramatically improved in recent years owing to rapid advances in high-throughput DNA sequencing technologies and other 'meta-omic' analyses which are driving molecular microbial ecology into the era of Eco-Systems Biology.[1] Although metagenomic data provide gene inventories, without any proof of their functionality, the analysis of community-wide transcripts facilitates an assessment of community-wide functions,[2] and community proteomics provide representation of the actual phenotypic traits of individual community members.[3] Metabolomics, through resolving the final and intermediate products of cellular metabolism, should theoretically be the most sensitive indicator of community-wide phenotypes and allow inference of key metabolic processes.[4] However, current metabolomic methodologies are limited in the number of metabolites that can be measured as well as their limited identifiability.[5]

The reconstruction of metabolic networks based on genomic data presents a compelling alternative to metabolomics for resolving the metabolic capabilities of organisms.[6] So far, the conventional approach used to progress from single to multi-species metabolic network reconstructions has involved treating the metabolic networks of individual species as an input–output system to build network-based[7] or constraint-based[8] models of metabolic interactions. However, these multi-species models, which are usually limited to only a few species, fail to explain how variations in gene or species composition affect the overall metabolic state of ecosystems.[9] Given the complexity of microbial communities, as well as the inability to isolate and sequence representative single cultures of all organisms within a community, such bottom-up approaches may be limited by the inherent impossibility to extrapolate community-wide networks and behaviour from individual isolate omic data sets.[1] Recently developed alternative approaches involve the determination of community-wide metabolic potential[10] and the reconstruction of community-wide metabolic networks based directly on

---

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-sur-Alzette, Luxembourg; [2]Institute for Systems Biology, Seattle, WA, USA; [3]The Translational Genomic Research Institute-North, Flagstaff, AZ, USA and [4]Life Science Research Unit, University of Luxembourg, Luxembourg, Luxembourg.
Correspondence: P Wilmes (paul.wilmes@uni.lu)
[5]Current address: Laboratory of Microbial Ecology and Technology, Ghent University, Ghent, Belgium
[6]These authors contributed equally to this work.

Received 16 February 2015; revised 24 April 2015; accepted 6 May 2015

© 2015 Nanyang Technological University/Macmillan Publishers Limited

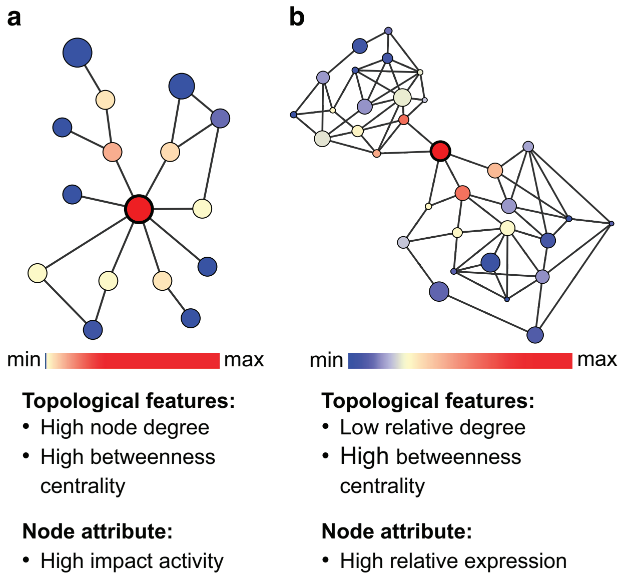


Figure 1. Criteria for defining keystone nodes in microbial species interaction and community-wide metabolic networks. (**a**) Criteria for identifying keystone species in reconstructed species interaction networks. Nodes represent taxa and edges represent associations between them. Node sizes reflect activity. (**b**) Criteria for identifying genes encoding key functionalities in reconstructed community-wide metabolic networks. Nodes represent enzyme-coding genes and edges correspond to shared metabolites (either reactants, products or educts). Node sizes reflect relative expression.

metagenomic data,[11] thereby ignoring the contribution of individual species.[12] Through this population-independent approach, Greenblum *et al.*[12] identified enzyme-coding genes, either enriched or depleted, in stool samples of human individuals with obesity or inflammatory bowel disease, highlighting the potential of such approaches for the identification of key metabolic traits within microbial consortia. Ideally, top-down and bottom-up approaches should be combined to identify links between microbial community structure and function, thereby bridging the gap between population-level metabolic networks and the larger community-wide networks to ultimately build a systems-level model of interactions between species.[13]

Here, we discuss a framework for comparative integrated omic analyses, which allows integration of systematically generated multi-omic data within reconstructed community-level metabolic networks. The resulting networks allow assessment of gene expression and protein abundances in combination with network topological features. We propose the use of these networks as an alternative to identifying keystone species through co-occurrence networks[14] (Figure 1a). Reconstruction of co-occurrence networks requires large numbers of highly resolved samples and spurious correlations can affect interpretability of the resulting networks.[15] Here, we identify genes encoding key functionalities in reconstructed community-wide metabolic networks and trace these back to the community members which encode them. Through their activity, keystone species are expected to have a disproportionately large effect on their environment, relative to their abundance.[16] Their removal would greatly impact community structure and function.[17] For example, in the human colon, specialist primary degraders such as *Ruminococcus bromii* are considered keystone species because of their ability to initiate the degradation of recalcitrant substrates.[18] Herein, we define key functionalities as specific functions which have an overall pronounced effect on ecosystem functioning, because they exhibit a high relative gene expression and are represented by a node with a prominent topological position within a community-wide metabolic network (Figure 1b). The loss of such nodes would result in a lack of connectivity and this would greatly impact the overall topology of the community-wide metabolic network. In addition, the expression of these genes will likely be rate-limiting, similar to the effect of 'load points' on reconstructed single-organism metabolic networks,[19] and thereby will govern the metabolic outcomes of the entire community. Therefore, by altering the expression of such genes, the community-wide phenotype could be influenced. By extension, members of the microbial community carrying out these functions would likely also be keystone species.

We apply the developed methodological framework to oleaginous mixed microbial communities (OMMCs) sampled from the surface of an anoxic BWWT tank in autumn and winter, respectively (Figure 2a,b). BWWT plants exhibit well-defined physical boundaries and represent a convenient and virtually unlimited source of spatially and temporally resolved samples. The microbial communities found in BWWT plants represent an ideal model system for microbial ecology[20] because these communities are comparatively well described and lie between communities of low diversity, e.g., acid mine drainage biofilms,[21] and complex communities such as those found in the human gastrointestinal tract[22] or soil environments[23] while retaining important hallmarks of both ends of the spectrum. These characteristics include (i) levels of dominance of individual taxa typically associated with low diversity communities (up to 30% of the community), most notably either *Candidatus* Microthrix parvicella (henceforth referred to as *Microthrix parvicella*) or *Perlucidibaca* spp. depending on the time of year;[24] and (ii) the functional potential to adapt to rapid environmental changes typically observed in more diverse communities. Compared with BWWT microbial communities that are more typically studied, e.g., bulk activated sludge, OMMCs have additional important attributes which render them ideally suited as a model for the development and implementation of eco-systematic approaches. These include (i) limited species richness, i.e., operational taxonomic unit (OTU) richness of approximately 600 (Chao[25] estimate from previous data[24]) compared with more than 1,000 (ref. 26) for activated sludge; (ii) high reproducibility between samples taken at the same time point.[4,27] Apart from these characteristics, the targeted enrichment of OMMCs is of biotechnological interest as this would allow the reclamation of a significant fraction of the chemical energy contained within wastewater through lipid recovery and subsequent biodiesel synthesis.[28,29] However, for such enrichment strategies to be successful, a detailed understanding of community function is necessary.[30] For example, identified key functionalities may ultimately serve as driver nodes[31] for controlling these communities.

MATERIALS AND METHODS

Sampling

OMMCs were sampled from the anoxic tank of the Schifflange (Esch-sur-Alzette, Luxembourg; 49°30′48.29″N; 6°1′4.53″E) BWWT plant as described previously.[4] Samples were taken on 4 October 2010 (referred to herein as the autumn OMMC) and 25 January 2011 (referred to herein as the winter OMMC; physico-chemical characteristics of the wastewater on the sampling dates are provided in Supplementary Table 1). These dates were chosen because they are representative of both extremes of OMMC-wide phenotypes, whereby, during the autumn sampling date, the tank exhibited only sparse amounts of OMMC biomass (Figure 2a) and, on the winter sampling date, ample amounts of OMMC biomass were present (Figure 2b).

Biomolecular extractions

A previously developed biomolecular isolation framework for community-integrated omics[4,27] was used to sequentially extract total RNA, genomic

© 2015 Nanyang Technological University/Macmillan Publishers Limited

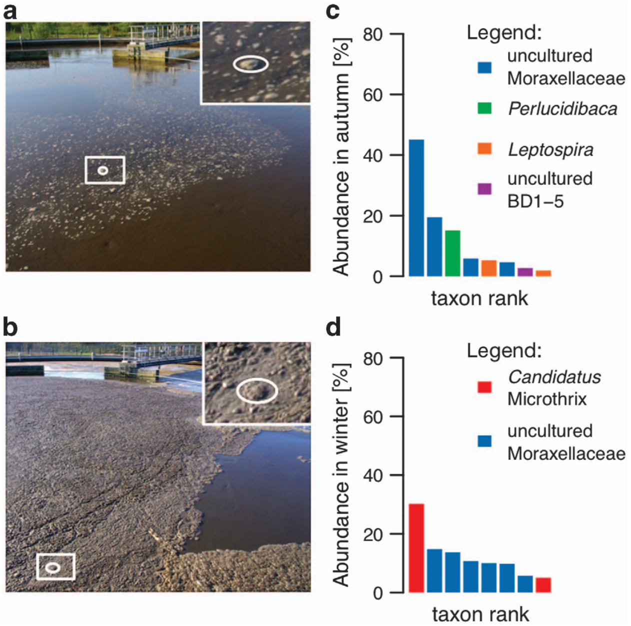


Figure 2. OMMC composition in autumn and winter seasons. Photographs of the OMMCs located at the water surface of the anoxic tank at the Schifflange BWWT plant in (**a**) autumn and (**b**) winter sampling dates. Abundance of genera of dominant community members based on reconstructed 16S rRNA gene sequences from metagenomic data in (**c**) autumn and (**d**) winter. OMMC, oleaginous mixed microbial community; rRNA, ribosomal RNA.

DNA and proteins from single OMMCs based on the Qiagen AllPrep DNA/RNA/Protein Mini kit (QA, Qiagen, Venlo, The Netherlands). The quality and quantity of isolated biomacromolecules were assessed as described previously[4] (Supplementary Table 2, Supplementary Materials and methods).

High-throughput sequencing

Total genomic DNA and ribosomal RNA-depleted retrotranscribed cDNA from both samples were sequenced on an Illumina Genome Analyzer IIx (Supplementary Materials and methods). Raw metagenomic and metatranscriptomic sequence data files are accessible in nucleic acid databases under BioProject PRJNA230567, sample LAO-A01 (SRX612782 and SRX612783) and LAO-A02 (SRX389533 and SRX389534).

Metagenomic and metatranscriptomic sequence assembly, gene annotation and determination of gene abundances

Raw 100 nt paired-end sequencing reads from the metagenome and metatranscriptome libraries from each of the two sampling dates were first trimmed and quality filtered using the *trim-fastq.pl* script from the *PoPoolation* package[32] and overlapping read pairs were assembled using the PAired-eND Assembler[33] (*PANDAseq*). Non-redundant assembled *PANDAseq* read pairs and non-assembled reads from metagenomic and metatranscriptomic data sets of both sampling dates were then used as a single input for the *MOCAT* assembly pipeline.[34] The resulting non-redundant contigs and *PANDAseq*-assembled read pairs that had not been used were then combined and filtered with a minimum length threshold of 150 bp. Protein-coding genes were predicted using the *Prodigal* gene finder[35] (v2.60, contigs above 500 bp) or *FragGeneScan*[36] (contigs between 150 and 500 bp). The resulting amino acid sequences from both contig sets were merged and made non-redundant using *CD-HIT*.[37] All predicted gene sequences are accessible through *MG-RAST*[38] as ID MGM4550606.3. The Kyoto Encyclopedia of Genes and Genome[39] database version 64.0 was used to functionally annotate genes with Kyoto Encyclopedia of Genes and Genome orthologous groups (KOs) for ensuing metabolic network reconstruction (Supplementary Materials and methods, Supplementary Figure 1).

To allow meaningful comparisons between gene copy and transcript numbers from the two seasons, identical numbers of reads were sampled from the metagenomic and the metatranscriptomic libraries of both seasons (Supplementary Materials and methods) using an in-house developed Perl-script. The resulting reads were then mapped to the annotated gene sets. Cross-mapping reads were equally weighted according to the number of genes they mapped to and mapped reads were counted per gene. Finally, metagenomic and metatranscriptomic counts were normalised by the effective length of the gene sequence,[40] yielding normalised gene copy abundances and normalised transcript abundances, respectively. KO abundances were inferred from the sums of normalised gene copy or transcript abundances of all genes belonging to a given KO (Supplementary Materials and methods). Relative gene expression values were determined per KO by calculating the ratio of normalised transcript abundances to normalised gene copy abundances (Supplementary Materials and methods, Supplementary Dataset 3).

Metaproteome processing and analysis

Isolated and purified protein fractions were separated using one-dimensional SDS polyacrylamide gel electrophoresis. The proteins were reduced, alkylated, and digested with trypsin. The resulting peptides were then analysed by liquid chromatography coupled to tandem mass spectrometry. Peptide identification was carried out by database searching using the *X!Tandem* software[41] with the amino acid sequence database generated from the genes predicted from the combined metagenomic and metatranscriptomic assembly. Protein identification was carried out using peptide-spectrum matches using the Trans-Proteomic Pipeline,[42] with a probability of being correctly assigned to each protein determined by *PeptideProphet*.[43] The protein inferences from each fraction were determined using *ProteinProphet* and then combined with *iProphet*[44] to obtain a master set of identified proteins at a 1% false discovery rate. All proteomic data have been deposited in the PeptideAtlas mass spectrometry raw file repository at http://www.peptideatlas.org/PASS/PASS00512. Identified proteins were assigned KO numbers using BLAT-based[45] alignment against the Kyoto Encyclopedia of Genes and Genome database v64.0 (Supplementary Materials and methods). Relative protein abundances were obtained using the normalised spectral index, as described previously[24] (Supplementary Materials and methods, Supplementary Figure 4).

Community-wide metabolic network reconstructions

Community-wide metabolic networks were reconstructed from the KOs with metabolic functions identified in the predicted gene sets from the combined metagenomic and metatranscriptomic assembly. The network reconstructions were rendered season-specific by using only KOs with mapped metatranscriptomic reads from each of the two sampling dates. The reconstructed networks reflect a connectivity-centred view of metabolism whereby enzymes grouped by KOs are represented by nodes and metabolites are represented by undirected edges, which represent either substrate or products of reactions catalysed by the respective KOs.[12] Each KO was assigned a pair-set of substrate and product metabolites according to the RPAIR[46] annotation in Kyoto Encyclopedia of Genes and Genome database version 67.1 (Supplementary Materials and methods).

Topological network analysis and selection criteria for genes encoding key functionalities

To carry out a topological analysis of the reconstructed metabolic network, nodes and edges were rendered non-redundant, by representing multiple KOs with identical substrate and product metabolites as a single node. A comparison between the non-redundant network and a redundant version was also carried out (Supplementary Materials and methods). As most of the nodes that regroup several KOs represent subunits of the same enzyme, the small changes incurred on betweenness centrality and load by making the nodes non-redundant enhance the ability of these topological measures to identify key enzymes in the reconstructed community-wide metabolic networks (see also Supplementary Results and Discussion). Key functionalities were identified on the basis of topological criteria and relative gene expression. The topological selection criterion was defined in analogy to 'load points' as defined by Rahman and Schomburg[19] in the context of reconstructed single-cell metabolic networks. Load points have the highest ratio of betweenness centrality (the number of valid shortest paths passing through them) relative to node degree (the number of neighbouring nodes; referred to as 'neighbourhood

© 2015 Nanyang Technological University/Macmillan Publishers Limited

connectivity' by Rahman and Schomburg[19]). Node degree and betweenness centrality, among other topological measures, of each node were computed using the *Cytoscape Network-Analyzer* plug-in,[47] considering the reconstructed network as undirected. These parameters were used to calculate load scores as defined in Equation (1).

$$\mathit{load\ score}_n = \frac{\sum_{s \neq n \neq t} (\sigma_{st}(n)/\sigma_{st})}{\frac{k_n}{\sum e}} \tag{1}$$

where s and t are nodes in the network different from n, σ_{st} is the number of shortest paths from s to t, and $\sigma_{st}(n)$ is the number of shortest paths from s to t that n lies on, k_n denotes the node degree of n, and Σe denotes the total number of edges in the network. Thus, load score describes the number of reaction paths or conversions between metabolites that utilise a given enzyme, relative to its connectivity. It therefore serves as a proxy for an enzyme's contribution to the metabolic fluxes of the overall community.

We prioritised the nodes with the top 10 per cent of load scores. In addition to this topological criterion, the relative gene expression of a node (either from a single KO or nodes regrouping several KOs) was also taken into account, such that only KOs with a high relative expression (top 10 per cent) were regarded as genes encoding key functionalities (Supplementary Materials and methods). Key functionalities were analysed for their involvement in the metabolism of uniquely occurring metabolites, i.e., to assess whether they represent 'choke points' as defined by Rahman and Schomburg.[19] For the calculation of an alternative load score weighted according to the occurrence of the metabolites which should restrict 'load points' to nodes within pathways[46] and a detailed analysis of sensitivity to the chosen cut-offs, see Supplementary Materials and methods.

Linking genes encoding key functionalities to specific organisms

The presence of the identified genes in genomes of bacterial isolates was determined by aligning contigs bearing these genes to the contigs from genome assemblies of these strains using BLAST (Supplementary Materials and methods).

Isolate strain culture and whole-genome sequencing

OMMC biomass sampled on 12 October 2011 was cultured on different growth media recommended for the culture of bacteria from water and wastewater and isolation procedures followed (Supplementary Materials and methods). In all, 140 pure bacterial cultures were obtained and screened for lipid inclusions using the Nile Red fluorescent dye.[48] Following DNA extraction using the Power Soil DNA isolation kit (MO BIO, Carlsbad, CA, USA), the genomes of 85 Nile Red-positive isolates were sequenced on an Illumina HiSeq Genome Analyzer IIx using the same sequencing approach as described for the metagenomic samples. The resulting sequencing reads were assembled using either the *Abyss*[49] or the *SPAdes*[50] assemblers (Supplementary Materials and methods). Based on the presence of a gene encoding a key functionality, one isolate (Isolate LCSB065) was selected for refinement of genome assembly as well as phylogenetic and genomic analysis (Supplementary Materials and methods).

Code availability and computational resources

All in-house developed scripts are available from the authors upon request. *In silico* analysis results were obtained using the high performance computing facilities of the University of Luxembourg.[51]

RESULTS AND DISCUSSION

Identification of functions encoded and expressed in OMMCs in autumn and winter

High-resolution coupled metagenomic, metatranscriptomic and metaproteomic data were generated from the OMMCs sampled in autumn and winter. A total of 16.2 gigabases (Gb) of shotgun metagenomic paired-end 100 nt read sequence data as well as 38.6 Gb of metatranscriptomic sequence data were obtained. 6.5 million genes were predicted from a 6.7 million contigs of a combined assembly (1.6 Gb total length) of all metagenomic and metatranscriptomic reads (Supplementary Table 3). Based on reconstructed 16S ribosomal RNA gene sequences from the metagenomic data (Supplementary Materials and methods), the autumn and winter communities are dominated by *Perlucidibaca* spp. and *Microthrix* spp., respectively (Figure 2c,d, Supplementary Dataset 1). A total 830,679 predicted genes were annotated with KOs and regrouped (Materials and methods), yielding a total of 7,270 unique KOs. In the autumn sample, 10,074 protein groups (identified proteins grouped together because they share detected peptides) were identified using 19,248 non-redundant peptides out of a total of 727,155 mass spectra. In the winter sample, 7,106 protein groups were identified from 15,966 non-redundant peptides out of a total of 620,488 tandem mass spectra. A total 4,906 and 5,007 proteins were unambiguously identified in the autumn and winter samples, respectively.

The congruency between the metagenomic and metatranscriptomic data was high, as 92% of KOs represented in the metagenomic data are also present in the metatranscriptomic data for both autumn and winter data sets (Supplementary Dataset 2). The coverage of KOs was lower in the proteomic data, as 1,357 KOs (26% of KOs annotated in the metagenomic data set) and 1,236 KOs (23%) were identified in autumn and winter OMMCs, respectively. These proportions were mirrored by KOs within metabolic pathways (Figure 3a,b). This comparatively low metaproteomic coverage is due to current limitations in proteomic technologies for metaproteomic analyses.[52]

Analysis of highly expressed genes in winter and autumn communities

Given the limited depth of coverage in the proteomic data, we mainly focused our subsequent comparative analyses on the metagenomic and metatranscriptomic data. Metaproteomic results were, however, used to corroborate and validate interpretations based on the analysis of the metatranscriptomic data. The comparison of KOs present in the metagenomic and metatranscriptomic data sets highlighted 757 (12%) and 210 (4%) unique KOs in autumn and winter OMMCs, respectively. Similar results were found in the comparison of KOs from metabolic pathways (Figure 3c). This analysis highlights a relatively limited difference in terms of genetic potential and gene expression between the two seasonally distinct OMMCs despite stark differences in community structure (Figure 2c,d).

For each identified KO, we calculated relative gene expression, which is considered to be more informative than simple transcript abundance because expression levels are normalised to metagenomic gene copy numbers.[53] Furthermore, it allows quantitative insights into the contribution of low abundance members (such populations may be potential keystone species) to overall community activity to be obtained.[54] KOs with high relative expression in both seasons (Figure 3d,e, Supplementary Dataset 3) were further analysed, as these are good candidates for genes which likely affect the overall community phenotype. Among these, enrichments were found in KOs linked to nitrogen metabolism, as well as oxidative phosphorylation and non-ribosomal peptide synthesis in both seasons (Supplementary Dataset 3). The highly expressed KOs involved in nitrogen metabolism represent enzymes for ammonium assimilation and oxidation, denitrification and nitrification. In particular, they include genes encoding likely subunits of ammonia monooxygenase (AMO; K10944, K10945 and K10946). AMO has a key role in the first step of nitrification carried out by aerobic ammonia-oxidising bacteria, mainly belonging to *Nitrosomonas* spp. and *Nitrosospira* spp.[54] AMO was previously found to be highly expressed in BWWT biomass.[55] In addition to the nitrogen metabolism enzymes expressed at a high level in both seasons, a nitrite reductase gene (K00363) was highly expressed in the autumn sample.

In the winter sample, the glycerolipid metabolism was enriched within highly expressed KOs. In particular, triacylglycerol lipase

© 2015 Nanyang Technological University/Macmillan Publishers Limited

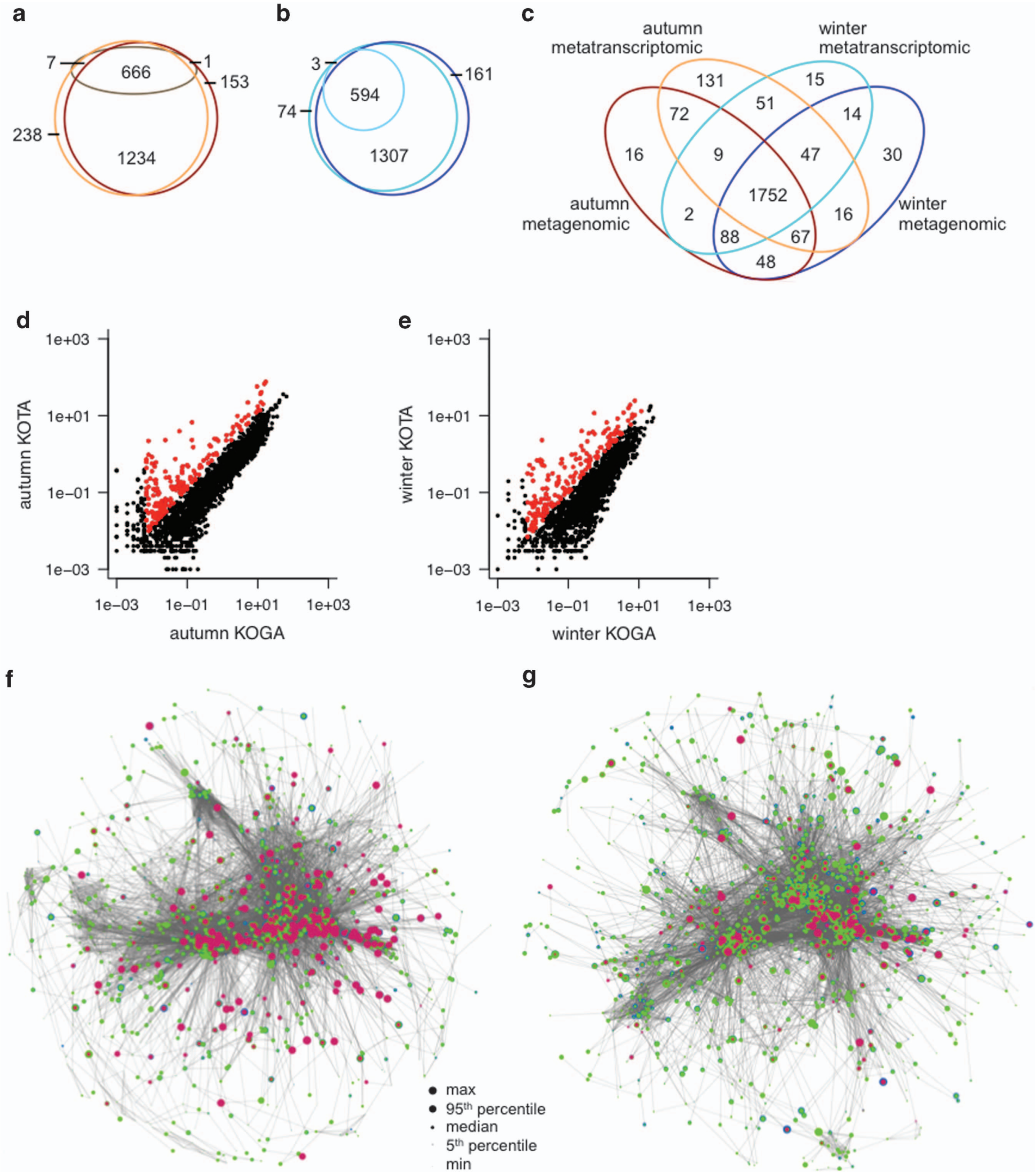


Figure 3. Integration of metagenomic, metatranscriptomic and metaproteomic data. (**a**) Venn diagram highlighting subsets of KEGG orthologous groups (KOs) in metabolic pathways present in the metagenomic (dark brown), metatranscriptomic (orange) and metaproteomic (pale brown) data from the autumn sample. (**b**) Subsets of KOs in metabolic pathways present in the metagenomic (dark blue), metatranscriptomic (cyan) and metaproteomic (pale blue) data from the winter sample. (**c**) Comparison of occurrence of KOs in metabolic pathways in metagenomic and metatranscriptomic data sets from autumn and winter. (**d**) Comparison of KO gene copy abundance (KOGA) and transcript abundance (KOTA) of KOs in metabolic pathways in the autumn data set. (**e**) Comparison of KO gene copy abundance (KOGA) and transcript abundance (KOTA) in metabolic pathways in the winter data set. In **d** and **e**, highly expressed KOs are highlighted in red. (**f**) Simplified autumn-specific metabolic network reconstruction. (**g**) Simplified winter-specific metabolic network reconstruction. In **f** and **g**, size of nodes represents KO abundance at metagenomic (blue), metatranscriptomic (green) and metaproteomic (magenta) levels, respectively. KEGG, Kyoto Encyclopedia of Genes and Genome.

(K01046) exhibited pronounced transcript levels and its expression was also confirmed at the protein level (Supplementary Dataset 2). The most highly expressed genes of the 6,222 genes belonging to this KO could be matched to *Acinetobacter* spp., which are known to occur in BWWT plants and accumulate triacylglycerols.[56] Furthermore, out of the genes with detectable expression, the two gene sequences with the highest gene copy numbers (i.e., abundance in the metagenomic data) were matched to the genome sequence of *Microthrix parvicella* BIO17-1 (ref. 57), which is enriched in KOs involved in lipid metabolism[57] (11.3% of its annotated genes). The presence of these enzymes was recently suggested to be essential for lipid accumulation in a metabolic

© 2015 Nanyang Technological University/Macmillan Publishers Limited

model reconstruction of *Microthrix parvicella,*[58] but not until now were they found to be expressed in biological wastewater treatment communities. The pronounced expression of the aforementioned KOs involved in ammonium oxidation and the hydrolysis of triacylglycerols during both seasons emphasises the capability of the OMMCs to remove two of the main compounds present in wastewater, i.e., ammonia[59] and lipids.[60]

In the winter sample, KOs from the TCA cycle were also strongly expressed and the majority could be detected at the proteome level. Rather surprisingly, in the autumn sample, photosynthesis KOs were enriched. Expression of photosystem I in autumn was also confirmed by proteomics suggesting that phototrophic organisms are part of the floating OMMC during this season.

Reconstruction of a generalised and season-specific OMMC-wide metabolic networks

A community-wide metabolic network was reconstructed using the KOs expressed in the autumn and winter samples (Materials and methods, Supplementary Figure 5, Supplementary Dataset 4). The reconstructed network comprised 1,432 KO nodes with 29,988 edges representing non-unique metabolites.

Season-specific networks were reconstructed analogous to the generalised OMMC-wide network, but by only using the 1,885 KOs or 1,775 KOs expressed in autumn or winter, respectively (Figure 3f,g, Supplementary Datasets 5 and 6). This yielded networks comprising 1,298 nodes with 25,842 edges and 1,375 nodes with 27,370 edges forming a connected network for winter and autumn, respectively.

Among the KOs specific to the autumn network, functions in the metabolic pathways for porphyrin and chlorophyll metabolism, sesquiterpenoid, triterpenoid and carotenoid biosynthesis pathways (ko00860, ko00909 and ko00906) were found to be enriched. This reinforces the notion that photosynthesis occurs in the OMMC sampled in autumn, while photosynthetic gene appear to be below the detection limit in the winter sample.

Identification of season-specific metabolic traits

The autumn- and winter-specific community-wide metabolic network reconstructions exhibit similar structures (Figure 3f,g) and represent 1,605 common KOs (i.e., 88 or 94% of the KOs included in the autumn or winter network reconstructions, respectively). Based on the reconstructed networks, a detailed network topological analysis was carried out (Supplementary Dataset 7).

Load scores (Equation 1) were determined in the reconstructed season-specific community-wide metabolic networks (Materials and methods). Most of the nodes in both the autumn- and winter-specific networks, which feature a high degree, represent KOs involved in amino acid synthesis. The relative small average shortest path lengths of 3.21 and 3.29 in the autumn and winter network reconstructions demonstrate that these represent ‘small world’ networks.[61] Among the nodes with the highest betweenness centrality, i.e., the highest number of shortest paths passing through a node,[62] in both metabolic reconstructions, KOs with functions in pyruvate metabolism, glycolysis or gluconeogenesis and glycerolipid metabolism were enriched (false discovery rate-adjusted P value <0.05). In contrast, relatively higher betweenness centrality of the nodes representing KOs in fatty acid metabolism pathway (ko01212) was observed in the network reconstruction from the winter data set (median fold change of 4; Wilcoxon signed rank test P value <0.001; enriched with false discovery rate-adjusted P value <0.00001; Supplementary Figure 6, Supplementary Dataset 7) suggesting distinct substrate usage in both seasons. Other functions, in which this subset of KOs was enriched, included porphyrin and chlorophyll metabolism, biotin metabolism, polyketide sugar unit biosynthesis, lipoic acid metabolism and fluorobenzoate degradation (ko00860, ko00780, ko00523, ko00785 and ko00364), while only phosphoinositol metabolism (ko00562) was significantly enriched among the functions of the nodes with a higher betweenness centrality in the autumn network.

Identification of genes encoding key functionalities

Keystone species occupy topologically important positions in species interaction networks[63] and are characterized by a high relative activity.[17] Within a community-wide metabolic network reconstruction, key functionalities contributed by keystone populations should be encoded by genes which exhibit a high relative gene expression and these genes should also occupy important topological positions in relation to the community-wide metabolic network, i.e., they should represent ‘load points’[19] (Figure 1b). Herein, we therefore identify genes having a high load score (Equation 1) within the season-specific metabolic networks as well as high relative expression in the respective data sets (Figure 2b, Figure 4, Materials and methods). Selected genes are reported and potential ‘choke points’ are indicated in Supplementary Dataset 7. According to Rahman and Schomburg, choke points are special cases of load points, which consume and/or produce unique metabolites. Given that uniqueness of a metabolite is a strong claim in the context of the reconstructed community-wide metabolic networks as much of community metabolism remains unknown (only 13% of the predicted genes could be confidently annotated with a function), the identification of key functionalities by using load points was chosen as a more robust and appropriate measure in the present case. The positions of the key functionalities within the networks as per our criteria (Figure 1b) are indicated in Figure 4 and Supplementary Figure 7. KOs involved in porphyrin and chlorophyll metabolic pathways are enriched among the selected genes in the autumn community, as are KOs with a function in degradation of aromatic compounds. Among the genes encoding key functionalities in the winter OMMCs, no significant enrichment among KOs from a particular pathway could be observed. However, one of these genes is K03921, coding for an acyl-[acyl-carrier-protein] desaturase, which is part of the biosynthesis pathway for polyunsaturated fatty acids.

In both the autumn and winter sets of season-specific key genes, the subunits of ammonia or methane monooxygenase (AMO or MMO) stand out. As discussed above and given the sampling from a nitrifying–denitrifying wastewater treatment plant, this is likely an AMO which catalyses the first essential step of nitrification by converting ammonia to hydroxylamine.[64] In contrast, MMO is involved in methane oxidation, which is less likely to be expressed in the sampled environment.

Linking genes encoding key functionalities to community members

Having selected genes encoding key functionalities within the sampled OMMCs using the reconstructed community-wide metabolic networks (Supplementary Dataset 7), we were interested in revealing which organisms expressed these genes within the community. As these genes contribute essential functionalities to the community and are characterized by relatively high expression, they are likely to be encoded by keystone species. Contigs containing genes annotated with one of the genes encoding key functionalities were selected from the combined metagenomic and metatranscriptomic data sets. These contigs were aligned to the NCBInr nucleotide database (Supplementary Dataset 7) to identify organisms encoding genes with similarity to the expressed genes of interest.

For five such genes (K03921, K01186, K01576, K01709 and K03335), no significant matches could be identified. On the other

© 2015 Nanyang Technological University/Macmillan Publishers Limited

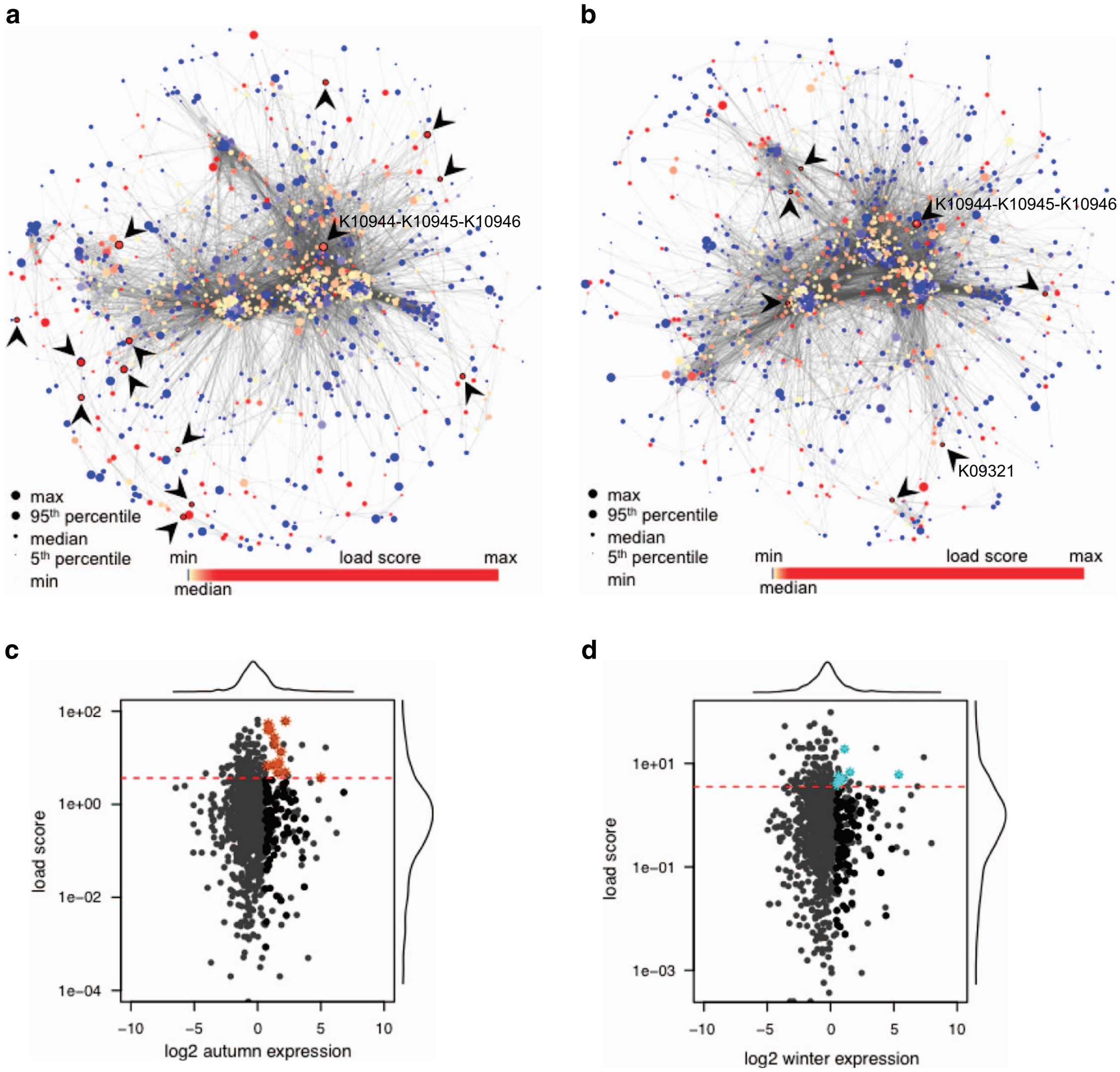


Figure 4. Topological analysis of the reconstructed season-specific community-wide metabolic networks and assessment of relative gene expression. (**a**) Autumn- and (**b**) winter-specific networks. In (**a**) and (**b**) node colours refer to *load score* and node sizes represent relative gene expression. KOs encoding key functionalities are encircled and highlighted by arrow heads. (**c** and **d**) Results of the topological analysis of KOs in simplified season-specific networks for (**c**) autumn and (**d**) winter. Highly expressed genes are indicated as black dots and KOs encoding key functionalities are indicated by brown (autumn) or cyan (winter) asterisks. Dotted red lines indicate minimal *load score* of KOs deemed to encode key functionalities.

hand, three of these key genes from the winter-specific network (K01251, K00789 and K03527) were expressed from a multitude of contigs, which could be aligned well to over 50 different species. Half of the matched contigs encoding the five autumn key genes from the chlorophyll- and porphyrin-synthesis pathway (K03403, K03404, K03405, K04034, K04035) were most similar to sequences encoded by the genome of the cyanobacterium *Oscillatoria nigro-viridis* PCC 712. The relative expression of these genes accounted for 85% of the expression of these genes in autumn (Supplementary Dataset 7). Some *Oscillatoria* spp. are found in wastewater, where they have been found to participate in nitrate removal.[65]

From the list of genes encoding key functionalities, we further selected the acyl-[acyl-carrier protein] desaturase (K03921) and the three subunits of AMO or MMO (K10944, K10945 and K10946) for further analysis. In all, 922 out of 1,067 contigs belonging to the AMO or MMO complex matched best to sequences of *Nitrosomonas* spp. a well-known genus of nitrifiers. The other contigs matched sequences from uncultured organisms or, in two cases, to a MMO from *Methylovulum miyakonense*. These two cases only represented 0.1% of the total contig length of the KOs K10944–K10946. Furthermore, less than 1% of the metatranscriptomic reads mapped to these two contigs, suggesting that the major function of these KOs is in ammonia rather than methane oxidation. In addition, a refined assembly of contigs belonging to K10944–K10946 using additional metagenomic data from a third sampling date (Supplementary Materials and methods) yielded a new contig containing complete sequences for *amoA* (an established phylogenetic marker for nitrifying microorganisms[66]), and *amoB*, both also matching best to *Nitrosomonas* spp. A phylogenetic tree was reconstructed using the predicted amino acid sequence of *AmoA* from this contig and the tree clearly places it closest to sequences of *Nitrosomonas* spp. (Figure 5a, Supplementary Table 4). To estimate the abundance of *Nitrosomonas* spp. in the sampled OMMCs, metagenomic and metatranscriptomic reads were mapped against the genome

© 2015 Nanyang Technological University/Macmillan Publishers Limited

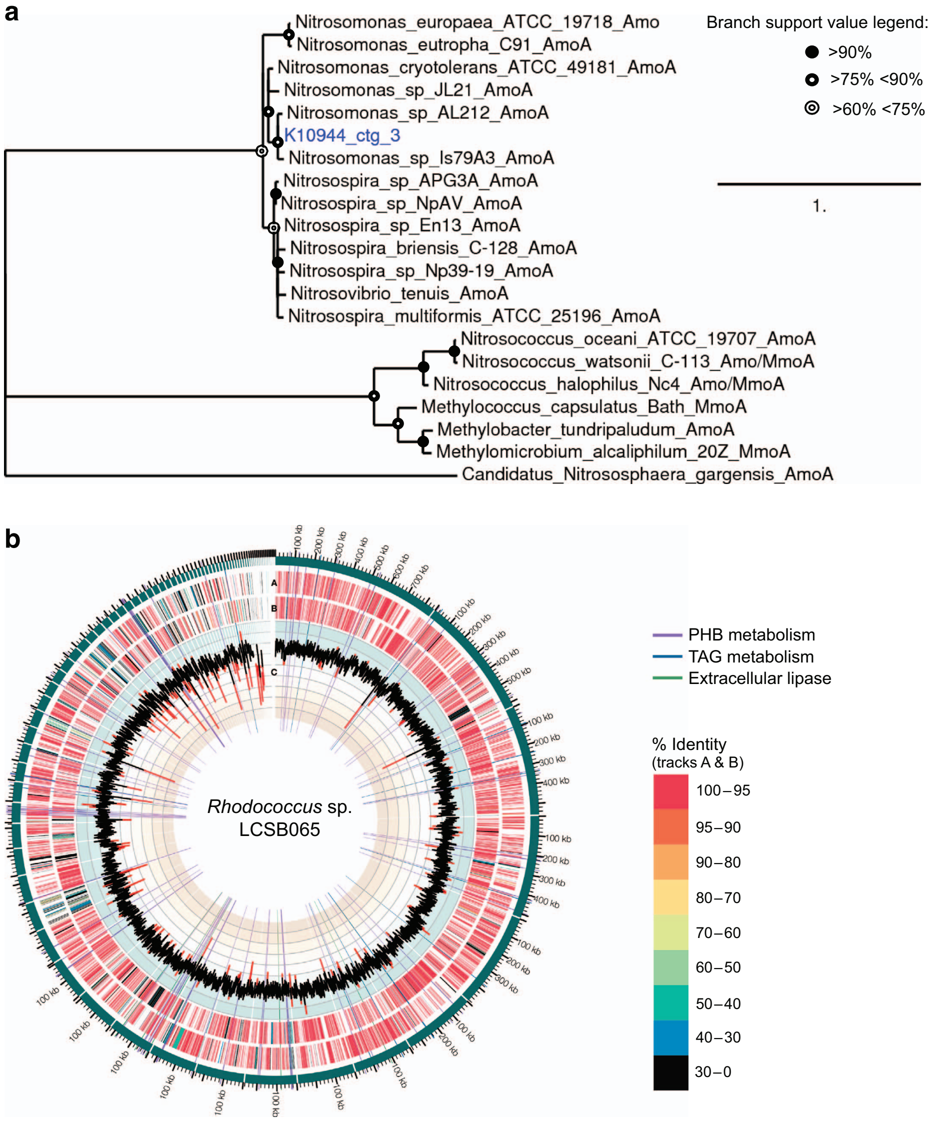


Figure 5. Linking key functionalities to important community members. (**a**) Phylogenetic tree based on the AmoA amino acid sequence derived from a contig extended using combined metagenomic and metatranscriptomic data (K10944_ctg_3). (**b**) Circos plot of the genome of Isolate LCSB065, highlighting amino acid similarity of encoded proteins to the *Rhodococcus erythropolis* PR4 genome and genes involved in poly-hydroxybutyrate (PHB) and TAG accumulation as well as encoded extracellular lipases. From the outside to the inside track: contigs (green) arranged by size; A: open reading frames in forward direction; B: open reading frames in reverse direction; colours in tracks A and B indicate %-similarity to the *Rhodococcus erythropolis* PR4 genome; C: %G+C in 1,000 bp sliding windows. Highlighted rays indicate the location of genes involved in PHB metabolism (violet), genes involved in TAG metabolism (blue) and extracellular lipase genes (green). TAG, triacylglycerol.

sequence of *Nitrosomonas* sp. Is79 (ref. 67), yielding approximately twice as many metagenomic reads in winter compared with autumn (Supplementary Table 5). The ratio of metatranscriptomic to metagenomic coverage was four times higher in winter than in autumn, indicating a higher general level of activity of *Nitrosomonas* spp. in the winter OMMC, although AMO activity was high in both seasons.

In contrast to the compelling link between the putative AMO genes and *Nitrosomonas* spp., linking the acyl-[acyl-carrier protein] desaturase unambiguously to an organismal group could not be achieved by simple alignment to reference genomes in public databases. Of the 14 contigs which harboured genes annotated with K03921 expressed in the winter sample, 9 did not yield any hits with a percentage identity >80% and query coverage >50%. The remaining five contigs yielded hits with 82 to 86% identity to sequences from *Rhodococcus erythropolis, Amycolatopsis mediterranei* and *Nocardia cyriacigeorgica*. As none of these alignments were of high confidence, we aligned the contigs encoding acyl-[acyl-carrier protein] desaturases to genomes of an in-house bacterial isolate collection from the same BWWT plant. Three of the contigs containing expressed genes matched to the same gene of the genome of Isolate LCSB065 with 88 to 100% identity over a total of 459 nt of the combined metagenomic contig length of 678 nt. Isolate LCSB065's 81 contigs contain an almost complete 7.67 Mbp genome with a GC-content of 62.4% (Figure 5b, Supplementary Dataset 8). Based on the use of 31 bacterial protein coding marker genes, this isolate was identified as a *Rhodococcus* sp.[68] (Supplementary Dataset 8). A detailed genomic analysis revealed a high number of genes involved in lipid metabolism encoded by this organism (Supplementary Results and Discussion) and non-polar

© 2015 Nanyang Technological University/Macmillan Publishers Limited

storage granules were also observed microscopically (Supplementary Figure 8). As *Rhodococcus* spp. are known to exhibit lipid accumulation phenotypes,[69] it is likely that this organism is a keystone species within the OMMC. Recruitment of metagenomic and metatranscriptomic reads to the isolate's genome (Supplementary Dataset 8) revealed a low abundance of this organismal group in both autumn and winter, with a relative high transcriptional activity only in winter (Figure 5b, Supplementary Table 5) potentially directly linking its activity to the high community-wide lipid accumulation phenotype observed in winter.[24] Low abundance combined with an activity with a great impact on their environment are hallmarks of keystone species and the *Rhodococcus* population fulfils these criteria in the context of the sampled OMMC.

CONCLUSION

Despite stark differences in the appearance and structure of the sampled autumn and winter OMMCs, the comparative analysis of integrated metagenomic, metatranscriptomic and metaproteomic data contextualised in reconstructed community-wide metabolic networks uncovered surprisingly few global differences in terms of functional genetic potential and gene expression between the two communities. This result confirms previous observations that taxonomic profiles can be very variable whereas global functional profiles are typically more conserved.[70,71] Nonetheless, our approach highlighted genes coding for essential enzymes involved in nitrogen metabolism (genes involved in nitrification, denitrification and ammonium oxidation) as being relatively highly expressed in both seasons despite exhibiting only low gene copy numbers. Identified differences between the two seasons include a marked expression of enzymes involved in glycerolipid metabolism in winter when OMMC biomass is most pronounced (Figure 2a,b) and lipid accumulation is higher.[24] In particular, our analyses highlight the importance of triacylglycerol lipases, which are essential for hydrolysis of lipids into long-chain fatty acids and their subsequent assimilation and intracellular storage. The pronounced expression of this particular enzyme group suggests the possibility to enrich for lipid accumulating organisms (LAOs) in BWWT plants through lipase supplementation and environmental biocatalysis. Enhancing the growth of LAOs through such a strategy would result in the availability of excess amounts of lipid-rich biomass at the air–water interface of anoxic tanks and this could, for example, be transesterified to biodiesel, thereby allowing recovery of a significant fraction of the chemical energy contained within wastewater.[28,29]

The topological analysis of the OMMC-wide metabolic networks confirms the metabolic similarity of both autumn and winter communities, with a high centrality of central carbon metabolism. The measure of betweenness centrality demonstrates seasonal variability in fatty acid metabolism, which is more enriched in the sampled winter OMMC. The identification of genes encoding key functionalities involved the detailed analysis of topological features within the reconstructed community-wide metabolic networks as well as an assessment of relative gene expression by enzyme-coding genes. This analysis highlighted genes such as AMO, expressed by *Nitrosomonas* spp., and an acyl-[acyl-carrier protein] desaturase, expressed by *Rhodococcus* spp., as fulfilling key functions in OMMCs.

The developed framework allows the integration of structural and functional measurements through contextualisation in reconstructed community-wide metabolic networks to result in the identification of genes encoding key functionalities, which can in turn be linked to functionally important community members. These potential 'keystone genes' could ultimately serve as driver nodes[31] for controlling such complex microbial ecosystems. Therefore, the application of our methodological framework to other microbial communities for the identification of keystone species may allow community-wide control strategies to be formulated where other community-wide phenotypic outcomes may be desirable, e.g., in the human gastrointestinal tract. *In silico* analysis results presented in this paper were obtained using the high performance computing facilities of the University of Luxembourg[51].

ACKNOWLEDGEMENTS

The authors thank Mr Bissen and Mr Di Pentima from the Syndicat Intercommunal à Vocation Ecologique (SIVEC) for their permission to collect samples and gain access to the monitoring platform of the Schifflange wastewater treatment plant; Dr Nicolás Pinel for his refinement of genome assembly and visualisation of *Rhodococcus* sp. LCSB065 and helpful discussions; and Dr Aidos Baumuratov for his support for microscopic analyses of the OMMC isolates. This work was supported by an ATTRACT programme grant (ATTRACT/A09/03) and a European Union Joint Programming in Neurodegenerative Diseases grant (INTER/JPND/12/01) both to PW, and Aide à la Formation Recherche (AFR) grants to HR (PHD-MARP-04), EELM (PRD-2011-1/SR), CCL (PHD/4964712) and SN (PHD-2014-1/7934898), all funded by the Luxembourg National Research Fund (FNR). The project received financial support from the Integrated Biobank of Luxembourg with funds from the Luxembourg Ministry of Higher Education and Research. The work of PM was funded by the 'Plan Technologies de la Santé du Gouvernement du Grand-Duché de Luxembourg' through the Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg. This work was also supported in part with US federal funds from the National Science Foundation MRI grant No. 0923536, by funds through National Institutes of Health; from the National Institute of General Medical Sciences under grant No. 2P50 GM076547/Centre for Systems Biology and S10RR027584.

CONTRIBUTIONS

HR, AH-B and PW designed the study; HR, EELM, LAL and PW sampled the treatment plant and extracted biomolecules; AH-B, HR, PM, VPS, CCL, SN, JMS, JDG, NDH, DME and PSK analysed the metagenomic and metatranscriptomic data; MRH, AH-B, HR, PM, RLM and VPS analysed the metaproteomic data; HR, AH-B, TS and VPS reconstructed and analysed the metabolic networks; AH-B, HR and PW wrote the manuscript. All the authors discussed the results and commented on the manuscript.

COMPETING INTERESTS

The authors declare no conflict of interest.

REFERENCES

1 Muller EE, Glaab E, May P, Vlassis N, Wilmes P. Condensing the omics fog of microbial communities. *Trends Microbiol* 2013; **7**: 325–333.
2 Helbling DE, Ackermann M, Fenner K, Kohler H-PE, Johnson DR. The activity level of a microbial community function can be predicted from its metatranscriptome. *ISME J* 2012; **6**: 902–904.
3 Wilmes P, Bond PL. Microbial community proteomics: elucidating the catalysts and metabolic mechanisms that drive the Earth's biogeochemical cycles. *Curr Opin Microbiol* 2009; **12**: 310–317.
4 Roume H, Muller EE, Cordes T, Renaut J, Hiller K, Wilmes P. A biomolecular isolation framework for eco-systems biology. *ISME J* 2013; **7**: 110–121.
5 Tang J. Microbial metabolomics. *Curr Genomics* 2011; **12**: 391–403.
6 Oberhardt MA, Palsson BO, Papin JA. Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* 2009; **5**: 320.
7 Cottret L, Milreu PV, Acuña V, Marchetti-Spaccamela A, Stougie L, Charles H *et al.* Graph-based analysis of the metabolic exchanges between two co-resident intracellular symbionts, *Baumannia cicadellinicola* and *Sulcia muelleri*, with their insect host, *Homalodisca coagulata*. *PLoS Comput Biol* 2010; **6**: e1000904.
8 Wintermute EH, Silver PA. Emergent cooperation in microbial metabolism. *Mol Syst Biol* 2010; **6**: 407.
9 Greenblum S, Chiu H-C, Levy R, Carr R, Borenstein E. Towards a predictive systems-level model of the human microbiome: progress, challenges, and opportunities. *Curr Opin Biotech* 2013; **24**: 810–820.
10 Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL *et al.* Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* 2012; **8**: e1002358.
11 Konwar KM, Hanson NW, Pagé AP, Hallam SJ. MetaPathways: a modular pipeline for constructing pathway/genome databases from environmental sequence information. *BMC Bioinformatics* 2013; **14**: 202.

© 2015 Nanyang Technological University/Macmillan Publishers Limited

12 Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci USA* 2012; **109**: 594–599.
13 Borenstein E. Computational systems biology and *in silico* modeling of the human microbiome. *Brief Bioinform* 2012; **13**: 769–780.
14 Steele JA, Countway PD, Xia L, Vigil PD, Beman JM, Kim DY *et al.* Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *ISME J* 2011; **5**: 1414–1425.
15 Berry D, Widder S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Front Microbiol* 2014; **5**: 219.
16 Paine RT. A conversation on refining the concept of keystone species. *Conserv Biol* 1995; **9**: 962–964.
17 de Visser S, Thébault E, de Ruiter PC. Ecosystem Engineers, Keystone Species. In: Leemans R (ed). *Ecological Systems*. Springer: New York, NY, USA, 2013; 59–68.
18 Ze X, Duncan SH, Louis P, Flint HJ. *Ruminococcus bromii* is a keystone species for the degradation of resistant starch in the human colon. *ISME J* 2012; **6**: 1535–1543.
19 Rahman SA, Schomburg D. Observing local and global properties of metabolic pathways:'load points' and 'choke points' in the metabolic networks. *Bioinformatics* 2006; **22**: 1767–1774.
20 Daims H, Taylor MW, Wagner M. Wastewater treatment: a model system for microbial ecology. *Trends Biotechnol* 2006; **24**: 483–489.
21 Denef VJ, Mueller RS, Banfield JF. AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *ISME J* 2010; **4**: 599–610.
22 Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M *et al.* Diversity of the human intestinal microbial flora. *Science* 2005; **308**: 1635–1638.
23 Mocali S, Benedetti A. Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. *Res Microbiol* 2010; **161**: 497–505.
24 Muller EE, Pinel N, Laczny CC, Hoopmann MR, Narayanasamy S, Lebrun LA *et al.* Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nat Commun* 2014; **5**: 1–10.
25 Chao A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 1987; **43**: 783–791.
26 Zhang T, Shao M-F, Ye L. 454 Pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. *ISME J* 2012; **6**: 1137–1147.
27 Roume H, Heintz-Buschart A, Muller EE, Wilmes P. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. Microbial Metagenomics, Metatranscriptomics, and Metaproteomics. *Method Enzymol* 2013; **531**: 219–236.
28 Sheik AR, Muller E, Wilmes P. A hundred years of activated sludge: time for a rethink. *Front Microbiol* 2014; **5**: 47.
29 Muller EE, Sheik AR, Wilmes P. Lipid-based biofuel production from wastewater. *Curr Opin Biotechnol* 2014; **30**: 9–16.
30 Narayanasamy S, Muller EEL, Sheik AR, Wilmes P. Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. *Microb Biotechnol* 2015; **8**: 363–368.
31 Liu Y-Y, Slotine J-J, Barabási A-L. Controllability of complex networks. *Nature* 2011; **473**: 167–173.
32 Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A *et al.* PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* 2011; **6**: e15925.
33 Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* 2012; **13**: 31.
34 Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR *et al.* MOCAT: a metagenomics assembly and gene prediction toolkit. *PLoS ONE* 2012; **7**: e47656.
35 Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; **11**: 119.
36 Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* 2010; **38**: e191.
37 Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012; **28**: 3150–3152.
38 Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al.* The metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 2008; **9**: 386.
39 Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2014; **42**: D199–D205.
40 Lee S, Seo CH, Lim B, Yang JO, Oh J, Kim M *et al.* Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res* 2011; **39**: e9.
41 Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res* 2004; **3**: 1234–1242.
42 Deutsch EW, Mendoza L, Shteynberg D, Farrah T, Lam H, Tasman N *et al.* A guided tour of the trans-proteomic pipeline. *Proteomics* 2010; **10**: 1150–1159.
43 Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* 2002; **74**: 5383–5392.
44 Shteynberg D, Deutsch EW, Lam H, Eng JK, Sun Z, Tasman N *et al.* iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol Cell Proteomics* 2011; **10**: M111,007690.
45 Kent WJ. BLAT - the BLAST-like alignment tool. *Genome Res* 2002; **12**: 656–664.
46 Faust K, Croes D, van Helden J. Metabolic pathfinding using RPAIR annotation. *J Mol Biol* 2009; **388**: 390–414.
47 Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 2011; **27**: 431–432.
48 Fowler SD, Greenspan P. Application of Nile red, a fluorescent hydrophobic probe, for the detection of neutral lipid deposits in tissue sections: comparison with oil red O. *J Histochem Cytochem* 1985; **33**: 833–836.
49 Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol İ. ABySS: a parallel assembler for short read sequence data. *Genome Res* 2009; **19**: 1117–1123.
50 Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012; **19**: 455–477.
51 Varrette S, Bouvry P, Cartiaux H, Georgatos F. Management of an Academic HPC Cluster: The UL Experience. *Proceedings of the 2014 International Conference on High Performance Computing Simulation (HPCS 2014)*. IEEE: Bologna, Italy, 2014.
52 Hettich RL, Sharma R, Chourey K, Giannone RJ. Microbial metaproteomics: identifying the repertoire of proteins that microorganisms use to compete and cooperate in complex environmental communities. *Curr Opin Microbiol* 2012; **15**: 373–380.
53 Carvalhais LC, Dennis PG, Tyson GW, Schenk PM, de Bruijn F. Rhizosphere metatranscriptomics: challenges and opportunities. de Bruijn FJ (ed). *Molecular Microbiology of the Rhizosphere*. Wiley-Blackwell: New Jersey, NJ, USA, 2013; 1137–1144.
54 Tsementzi D, Poretsky R, Rodriguez-R LM, Luo C, Konstantinidis KT. Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. *Environ Microbiol Rep* 2014; **6**: 640–655.
55 Zhu G, Peng Y, Li B, Guo J, Yang Q, Wang S. Biological removal of nitrogen from wastewater. In: *Reviews of Environmental Contamination and Toxicology*. Springer: New York, NY, USA, 2008; 159–195.
56 Yu K, Zhang T. Metagenomic and metatranscriptomic analysis of microbial community structure and gene expression of activated sludge. *PLoS ONE* 2012; **7**: e38183.
57 Kalscheuer R. Genetics of wax ester and triacylglycerol biosynthesis in bacteria. In: Timmis KN (ed). *Handbook of Hydrocarbon and Lipid Microbiology*. Springer: Berlin Heidelberg, Germany, 2010; 527–535.
58 Muller EEL, Pinel N, Gillece JD, Schupp JM, Price LB, Engelthaler DM *et al.* Genome Sequence of '*Candidatus* Microthrix parvicella' Bio17-1, a long-chain-fatty-acid-accumulating filamentous actinobacterium from a biological wastewater treatment plant. *J Bacteriol* 2012; **194**: 6670–6671.
59 McIlroy SJ, Kristiansen R, Albertsen M, Karst SM, Rossetti S, Nielsen JL *et al.* Metabolic model for the filamentous '*Candidatus* Microthrix parvicella' based on genomic and metagenomic analyses. *ISME J* 2013; **7**: 1161–1172.
60 De Clippeleir H, Vlaeminck SE, De Wilde F, Daeninck K, Mosquera M, Boeckx P *et al.* One-stage partial nitritation/anammox at 15 °C on pretreated sewage: feasibility demonstration at lab-scale. *App Microbiol Biotechnol* 2013; **97**: 10199–10210.
61 Raunkjær K, Hvitved-Jacobsen T, Nielsen PH. Measurement of pools of protein, carbohydrate and lipid in domestic wastewater. *Water Res* 1994; **28**: 251–262.
62 Watts DJ, Strogatz SH. Collective dynamics of 'small-world'networks. *Nature* 1998; **393**: 440–442.
63 Brandes U. On variants of shortest-path betweenness centrality and their generic computation. *Soc Networks* 2008; **30**: 136–145.
64 Faust K, Raes J. Microbial interactions: from networks to models. *Nat Rev Microbiol* 2012; **10**: 538–550.
65 Martens-Habbena W, Berube PM, Urakawa H, José R, Stahl DA. Ammonia oxidation kinetics determine niche separation of nitrifying Archaea and Bacteria. *Nature* 2009; **461**: 976–979.
66 Attasat S, Wanichpongpan P, Ruenglertpanyakul W. Cultivation of microalgae (*Oscillatoria okeni* and *Chlorella vulgaris*) using tilapia-pond effluent and a comparison of their biomass removal efficiency. *Water Sci Technol* 2013; **67**: 271–277.
67 Liu W, Li L, Khan MA, Zhu F. Popular molecular markers in bacteria. *Mol Genet Microbiol Virol* 2012; **27**: 103–107.

© 2015 Nanyang Technological University/Macmillan Publishers Limited

68 Bollmann A, Sedlacek CJ, Norton J, Laanbroek HJ, Suwa Y, Stein LY *et al.* Complete genome sequence of *Nitrosomonas* sp. Is79, an ammonia oxidizing bacterium adapted to low ammonium concentrations. *Stand Genomic Sci* 2013; **7**: 469.
69 Kerepesi C, Bánky D, Grolmusz V. AmphoraNet: the webserver implementation of the AMPHORA2 metagenomic workflow suite. *Gene* 2014; **533**: 538–540.
70 Alvarez HM, Mayer F, Fabritius D, Steinbüchel A. Formation of intracytoplasmic lipid inclusions by *Rhodococcus opacus* strain PD630. *Arch Microbiol* 1996; **165**: 377–386.
71 Ju F, Guo F, Ye L, Xia Y, Zhang T. Metagenomic analysis on seasonal microbial variations of activated sludge from a full-scale wastewater treatment plant over 4 years. *Environ Microbiol Rep* 2014; **6**: 80–89.
72 Xu Z, Malmer D, Langille MGI, Way SF, Knight R. Which is more important for classifying microbial communities: who's there or what they can do? *ISME J* 2014; **8**: 2357–2359.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

Supplementary Information accompanies the paper on the *npj Biofilms and Microbiomes* website (http://www.nature.com/npjbiofilms)

© 2015 Nanyang Technological University/Macmillan Publishers Limited

A.5 Colonization and succession within the human gut microbiome by archaea, bacteria and microeukaryotes during the first year of life

Linda Wampach, Anna Heintz-Buschart, Angela Hogan, Emilie E.L. Muller, **Shaman Narayanasamy**, Cédric C. Laczny, Luisa W. Hugerth, Lutz Bindl, Jean Bottu, Anders F. Andersson, Carine de Beaufort, Paul Wilmes
Submitted
Frontiers in Microbiology

Contributions of author include:

- Data analysis
- Revision of manuscript

Colonization and succession within the human gut microbiome by archaea, bacteria and microeukaryotes during the first year of life

Linda Wampach[1], Anna Heintz-Buschart[1], Angela Hogan[2], Emilie E. Muller[1, 3], Shaman Narayanasamy[1], Cédric C. Laczny[1, 4], Luisa W. Hugerth[5], Lutz Bindl[6], Jean Bottu[6], Anders F. Andersson[5], Carine de Beaufort[1, 6], Paul Wilmes[1*]

[1]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg, [2]Integrated BioBank of Luxembourg, Luxembourg, [3]Department of Microbiology, Genomics and the Environment, Université de Strasbourg, France, [4]Clinical Bioinformatics, Saarland University, Germany, [5]Science for Life Laboratory, KTH Royal Institute of Technology, Sweden, [6]Centre Hospitalier de Luxembourg, Luxembourg

In review

Submitted to Journal:
Frontiers in Microbiology

Specialty Section:
Food Microbiology

Article type:
Original Research Article

Manuscript ID:
241593

Received on:
08 Nov 2016

Frontiers website link:
www.frontiersin.org

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Author contribution statement

LW carried out the qPCR assays, data processing, comparative analyses of the 16S and 18S rRNA gene amplicon sequencing data and data interpretation, participated in the biomolecular extractions and wrote the manuscript. AHB, AH, CCL, CdB and PW were involved in data analysis and interpretation. AHB contributed towards biomolecular extractions and writing the manuscript. AH participated in the design of the study and in the sample and data collection. EELM, SN and CCL participated in data analysis and critical revision of the manuscript. LWH and AFA participated in the data processing of the 18S rRNA gene amplicon sequencing data and provided important advice. LB and JB participated in the sample collection and processing. CdB and PW conceived and coordinated the study and participated in its design and in writing the manuscript. All authors read and approved the final manuscript.

Keywords

infant gut microbiome, Microbial colonization, Fungi, succession, delivery mode, Amplicon sequencing, quantitative real-time PCR

In review

Abstract

Word count: 330

Perturbations to the colonization process of the human gastrointestinal tract have been suggested to result in adverse health effects later in life. Although much research has been performed on bacterial colonization and succession, much less is known about the other two domains of life, archaea and eukaryotes. Here we describe the colonization and succession by bacteria, archaea and microeukaryotes during the first year of life (samples collected around days 1, 3, 5, 28, 150 and 365) within the gastrointestinal tract of infants delivered either vaginally or by caesarean section and using a combination of quantitative real-time PCR as well as 16S and 18S rRNA gene sequencing. Sequences from organisms belonging to all three domains of life were detectable in all of the collected meconium samples. The microeukaryotic community composition fluctuated strongly over time and early diversification was delayed in infants receiving formula milk. Caesarean section-delivered (CSD) infants experienced a delay in colonization and succession, which was observed for all three domains of life. Shifts in prokaryotic succession in CSD infants compared to vaginally delivered (VD) infants were apparent as early as on days 3 and 5, which were characterized by increased relative abundances of the genera Streptococcus and Staphylococcus, and a decrease in relative abundance for the genera Bifidobacterium and Bacteroides. Generally, a depletion in Bacteroidetes was detected as early as day 5 postpartum in CSD infants, causing a significantly increased Firmicutes/Bacteroidetes ratio between days 5 and 150 when compared to VD infants. Although the delivery mode appeared to have the strongest influence on differences between the infants, other factors such as a younger gestational age or maternal antibiotics intake likely contributed to the observed patterns as well. Our findings complement previous observations of a delay in colonization and succession of CSD infants, which likely affects not only bacteria but also archaea and microeukaryotes. This further highlights the need for resolving bacterial, archaeal and microeukaryotic dynamics in future longitudinal studies of microbial colonization and succession within the neonatal gastrointestinal tract.

Funding statement

This work was partially funded by the IBBL and the continuation of the study has been ensured with the generous financial support of the Fondation André et Henriette Losch.

Ethics statements

(Authors are required to state the ethical considerations of their study in the manuscript, including for cases where the study was exempt from ethical approval procedures)

Does the study presented in the manuscript involve human or animal subjects: Yes

Please provide the complete ethics statement for your manuscript. Note that the statement will be directly added to the manuscript file for peer-review, and should include the following information:

- Full name of the ethics committee that approved the study

- Consent procedure used for human participants or for animal owners
- Any additional considerations of the study in cases where vulnerable populations were involved, for example minors, persons with disabilities or endangered animal species

As per the Frontiers authors guidelines, you are required to use the following format for statements involving human subjects:
This study was carried out in accordance with the recommendations of 'name of guidelines, name of committee' with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the 'name of committee'.
For statements involving animal subjects, please use:
This study was carried out in accordance with the recommendations of 'name of guidelines, name of committee'. The protocol was approved by the 'name of committee'.

If the study was exempt from one or more of the above requirements, please provide a statement with the reason for the exemption(s).
Ensure that your statement is phrased in a complete way, with clear and concise sentences.

This study was carried out in accordance with the recommendations of good clinical practices established by the ‘International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use’ with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Luxembourgish ‘Comité National d'Ethique de Recherche’ in 2011 (reference number 201110/06).

In review

Colonization and succession within the human gut microbiome by archaea, bacteria and microeukaryotes during the first year of life

Linda Wampach[1], Anna Heintz-Buschart[1], Angela Hogan[2], Emilie E.L. Muller[1+], Shaman Narayanasamy[1], Cédric C. Laczny[1°], Luisa W. Hugerth[3], Lutz Bindl[4], Jean Bottu[4], Anders F. Andersson[3], Carine de Beaufort[1,4] and Paul Wilmes[1]*

Author affiliations:

1. University of Luxembourg, Luxembourg Centre for Systems Biomedicine, Esch-sur-Alzette, Luxembourg
linda.wampach@uni.lu,
anna.buschart@uni.lu
emilie.muller@unistra.fr
shaman.narayanasamy@uni.lu
cedric.laczny@ccb.uni-saarland.de
carine.debeaufort@uni.lu
paul.wilmes@uni.lu

[+] Current affiliation: Department of Microbiology, Genomics and the Environment, UMR 7156 UNISTRA – CNRS, Université de Strasbourg, Strasbourg, France

° Current affiliation: Chair for Clinical Bioinformatics, Saarland University, Building E2.1, 66123 Saarbrücken, Germany

2. Integrated BioBank of Luxembourg, Luxembourg, Luxembourg
Angela.Hogan@ibbl.lu

3. KTH Royal Institute of Technology, Science for Life Laboratory, School of Biotechnology, Division of Gene Technology, Stockholm, Sweden
luisa.hugerth@scilifelab.se
anders.andersson@scilifelab.se

4. Centre Hospitalier de Luxembourg, Luxembourg, Luxembourg
Bindl.Lutz@chl.lu
Bottu.Jean@chl.lu
debeaufort.carine@chl.lu

* for correspondence: paul.wilmes@uni.lu
Tel. +352 46 66 44 6188
Fax +352 46 66 44 6949

Number of words: 9,810
Number of figures: 7
Number of tables: 2

In review

Abstract

Perturbations to the colonization process of the human gastrointestinal tract have been suggested to result in adverse health effects later in life. Although much research has been performed on bacterial colonization and succession, much less is known about the other two domains of life, archaea and eukaryotes. Here we describe the colonization and succession by bacteria, archaea and microeukaryotes during the first year of life (samples collected around days 1, 3, 5, 28, 150 and 365) within the gastrointestinal tract of infants delivered either vaginally or by caesarean section and using a combination of quantitative real-time PCR as well as 16S and 18S rRNA gene sequencing. Sequences from organisms belonging to all three domains of life were detectable in all of the collected meconium samples. The microeukaryotic community composition fluctuated strongly over time and early diversification was delayed in infants receiving formula milk. Caesarean section-delivered (CSD) infants experienced a delay in colonization and succession, which was observed for all three domains of life. Shifts in prokaryotic succession in CSD infants compared to vaginally delivered (VD) infants were apparent as early as on days 3 and 5, which were characterized by increased relative abundances of the genera *Streptococcus* and *Staphylococcus*, and a decrease in relative abundance for the genera *Bifidobacterium* and *Bacteroides.* Generally, a depletion in Bacteroidetes was detected as early as day 5 *postpartum* in CSD infants, causing a significantly increased Firmicutes/Bacteroidetes ratio between days 5 and 150 when compared to VD infants. Although the delivery mode appeared to have the strongest influence on differences between the infants, other factors such as a younger gestational age or maternal antibiotics intake likely contributed to the observed patterns as well. Our findings complement previous observations of a delay in colonization and succession of CSD infants, which likely affects not only bacteria but also archaea and microeukaryotes. This further highlights the need for resolving bacterial, archaeal and microeukaryotic dynamics in future longitudinal studies of microbial colonization and succession within the neonatal gastrointestinal tract.

Keywords

fungi
succession
delivery mode
infant gut microbiome
amplicon sequencing
microbial colonization
quantitative real-time PCR

In review

1. Introduction

The human microbiome contributes essential functionalities to human physiology and is thought to play a crucial role in governing human health and disease (Greenhalgh et al., 2016). A growing body of evidence suggests that chronic diseases such as allergies (Abrahamsson et al., 2012; Abrahamsson et al., 2014), type 2 diabetes (Delzenne et al., 2015), obesity (Turnbaugh et al., 2006) and metabolic syndrome (Vrieze et al., 2012) are associated with a disequilibrium in the microbiome of the human gastrointestinal tract (GIT).

The initial microbiome colonization process is crucial for the development and maturation of the GIT as well as the immune system of the developing infant (Houghteling et al., 2015; Björkstén, 2004; Caicedo et al., 2005; Rautava and Walker, 2007; Eberl and Lochner, 2009). During vaginal delivery, a subset of the maternal bacterial community is supposedly transferred to the infant; in contrast, early-stage microbiome profiles from infants delivered by caesarean section (C-section) are typically not as reflective of the mothers' vaginal or gastrointestinal environment (Dominguez-Bello et al., 2010; Bäckhed et al., 2015; Nayfach et al., 2016). Based on spatio-temporal studies in humans (Abrahamsson et al., 2014), it has been suggested that various disturbances in the initial microbiome colonization process as early as one month after birth may increase chronic disease susceptibilities over the course of human life (Houghteling et al., 2015; Cox et al., 2014; Arrieta et al., 2014). It has been previously observed that the delivery mode is the most important factor in determining the early colonization pattern(s) (Dominguez-Bello et al., 2010; Jakobsson et al., 2014), although other factors, such as diet (breast milk versus formula milk; Le Huërou-Luron et al., 2010), gestational age (term delivery versus preterm delivery; Barrett et al., 2013) or the maternal intake of antibiotics (Sekirov et al., 2008) have also been observed to have effects on this process.

Even though the colonization and succession within the GIT have been studied extensively, the focus has mostly been directed to the bacterial domain. However, such a constrained view may lead to an underestimation of the contribution of the archaeal and eukaryotic domains, in particular microeukaryotes, such as unicellular parasites or yeasts, and could ultimately lead to incomplete conclusions (Horz, 2015).

Within the archaeal domain, methanogenic archaea (mainly those belonging to the order Methanobacteriales) have been estimated to comprise between 10^8 and 10^{10} cells per gram dry weight of stool (Miller and Wolin, 1986) and are considered almost ubiquitous inhabitants of the intestinal microbiome with a presence in up to 95.7% of all adult humans (Dridi et al., 2009). Methanogenic archaea are functionally important due to their ability to consume molecular hydrogen, which is both an end product and a concentration-dependent inhibitor of bacterial fermentation (Thauer at el., 2008). Consequently, methanogens drive the effective degradation of organic substances and play an important role in interspecies hydrogen transfer through maintaining syntrophic relationships with bacterial populations (Hansen et al., 2011). Additionally, gut methanogens have been linked to energy metabolism and adipose tissue deposition of the human host (Samuel et al., 2007), and the ability of certain archaea to produce methane may play a role in the pathogenesis of several intestinal disorders (Roccarina et al., 2010). Despite these observations, the simultaneous presence of archaea and bacteria has been ignored in the majority of studies on the gastrointestinal microbiome to date and details about neonatal colonization by archaea

In review

remain limited. Previous studies have detected archaea transiently and almost exclusively in the first few weeks of life, and considerably less in samples collected after the fifth week of life (Palmer et al., 2007). Archaea have been sporadically detected in the vaginal environment before, although exclusively in women with bacterial vaginosis (Belay et al., 1990). As archaea are mainly inhabitants of the human GIT, but also colonize the skin surface (Probst et al., 2013) as well as the oral cavity (Nguyen-Hieu et al., 2013), a transfer from mother to infant by faecal-oral or oral-oral route seems thereby most probable.

Eukaryotes and microeukaryotes, which form part of the human microbiota, have been shown to exert immunomodulatory effects on the host (Weinstock, 2012; Rizzetto et al., 2014). Furthermore, infections by parasitic eukaryotes have been linked to decreased allergic and autoimmune disease prevalence (Weinstock, 2012) and have been used for therapeutic interventions in that context (McFarland and Bernasconi, 1993; Williamson et al., 2016). However, the role of microeukaryotes within the human GIT microbiome and the resulting impact on the human host remain so far unresolved (Andersen et al., 2013). It has been previously reported that the overall microeukaryotic diversity of the adult human GIT is low but largely temporally stable (Scanlan and Marchesi, 2008), whereas other research suggested that the adult GIT microbiome harbors a complex microeukaryotic community with the most abundant taxa by far being fungi (Hamad et al., 2012). To date, a single study followed the initial colonization of the GIT by microeukaryotes using 18S rRNA gene amplicon sequencing in four newborn infants (Pandey et al., 2012), but failed to detect any microeukaryotes at the timepoints analyzed. However, this study might have been substantially limited by its sample collection as well as the applied sequencing technique.

In our present work, a longitudinal study was conducted to describe the colonization and succession of the three domains of life within the GIT of newborns. More specifically, we investigated the microbiome changes during the first year of life among eight vaginally delivered (VD) infants and seven infants delivered by C-section (CSD). The latter are statistically at a higher risk of developing metabolic disease such as obesity (Mueller et al., 2015) and/or related diseases like type 2 diabetes (Nguyen and El-Serag, 2010), as well as allergic diseases such as atopic eczema (Abrahamsson et al., 2012) and asthma (Abrahamsson et al., 2014) in childhood and/or adulthood. Fecal samples were collected from all infants (VD and CSD) at six time points between day 1 and 1 year *postpartum* and, using quantitative real-time PCR (qPCR), we determined the sizes of prokaryotic (bacteria and archaea) and fungal populations, the relative quantities of archaea and validated the amounts of four selected bacterial genera and two phyla in the collected samples. Additionally, targeted high-throughput 16S and 18S rRNA gene amplicon sequencing was conducted on the isolated DNA. After processing and filtering of the resulting data, we compared the prokaryotic and microeukaryotic community structures in relation to the delivery mode and a multitude of other recorded maternal/neonatal characteristics. The resulting data provides a detailed overview of the neonatal colonization and succession patterns of members of all three domains of life.

In review

2. Material and methods

2.1 Sample collection, processing and biomolecular extraction

2.1.1 Study context

In the context of the national COSMIC study, pregnant women were recruited in Luxembourg starting in 2012. The 15 pregnant women included in the presented study were aged between 24 and 42 years and gave birth in the maternity department of the Centre Hospitalier de Luxembourg (CHL). This study was carried out in accordance with the recommendations of good clinical practices established by the 'International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use' with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Luxembourgish 'Comité National d'Ethique de Recherche' in 2011 (reference number 201110/06).

2.1.2 Sample and data collection

To mitigate pre-analytical confounders, fecal samples were immediately snap-frozen in liquid nitrogen or placed on dry ice following collection and were stored at -80 °C until further processing. Fecal samples were scheduled to be collected at day 1, day 3, day 5, day 28, day 150 and day 365. The medical histories of both parents and medication intake of the mother were recorded, as well as weight, date of birth, gender, mode of delivery and gestational age of the infant. Additional data, which was collected subsequently for all infants included weight, type of milk fed, medication intake including antibiotics and time point at which solid food was introduced. If an infant received formula at a specific point in time, it was considered as receiving combined feeding for the entire remainder of the study, as even short-term formula-feeding has been shown to cause profound and long lasting shifts to the gastrointestinal microbiome composition (Guaraldi and Salvatori, 2012). Hospitalization in the neonatal care unit and administration of antibiotics to infants immediately *postpartum* as well as birth prior to 34 weeks of gestation were exclusion criteria. Additionally, a control fecal sample from a single healthy adult individual was collected and preserved under the same conditions as described previously. Samples and associated data were collected and stored at the Integrated BioBank of Luxembourg (IBBL) following ISO17025:2005 standards and the International Society for Biological and Environmental Repositories (ISBER) best practices.

2.1.3. DNA extraction from fecal samples

Pre-processing of all fecal samples (150-200 mg of weighed material) was carried out according to Shah et al. (2016, in press; subsection 3.2, steps 1-4). After high-speed centrifugation, DNA was extracted from the resulting interphase pellet using the PowerSoil® DNA isolation kit (MOBIO Laboratories, Belgium). The method was optimized for mechanical disruption with bead-beating to ensure a realistic representation of microbial communities (Walker et al., 2015). DNA quality and quantity were determined on 1 % agarose gels, by NanoDrop 2000c spectrophotometer (Thermo Fisher Scientific, USA) and Qubit 2.0 fluorometer (Thermo Fisher Scientific, USA). The extracted DNA was stored at -80 °C until qPCR validation and sequencing library construction.

2.2 DNA analyses and sequencing

2.2.1 Quantitative real-time PCR

Extracted DNA was diluted, when applicable, to a concentration of 5 ng/µl and

In review

amplified in duplicates, using previously published primers targeting prokaryotes, archaea or specific fungi as well as specific bacterial genera and phyla (Table 1), which were ordered and received from Eurogentec (Belgium). The reaction mixture contained 1 µl template DNA, 5 µl of Mastermix (iQ SYBR Green Supermix; Bio-Rad Laboratories, USA), and 500 nMol of each primer, in a final reaction volume of 10 µl. Genomic DNA isolated from *Salmonella* Typhimurium LT2 and *Saccharomyces cerevisiae* BY4743 was used to prepare standard curves for the universal prokaryotic and fungal primers, respectively. A sample pool, comprised of 1 µl of undiluted DNA from each of the 65 samples, was used to prepare standard curves for all assays. All standard curves were prepared with a total of at least five successive 10-fold dilutions. qPCR was performed on a LightCycler 480 (Roche Diagnostics, Germany) with an initial denaturation step of 1 min at 95 °C followed by primer-specific cycling times (Table 1), a single fluorescence acquisition step at the end of the extension step and a final melting curve. Crossing point (Cp) values were calculated using the second derivative method within the Roche LightCycler 480 software version 1.5. Absolute copy numbers of prokaryotic 16S and fungal 18S rRNA genes were calculated using the Cp values and the reaction efficiencies based on the standard curves obtained from defined DNA samples and extractions yields were estimated from these numbers. Relative concentrations of specific taxa compared to all 16S rRNA genes were calculated using Cp values and the standard curves obtained for the sample pool. Only samples where the target was positively detected in both duplicate reactions were considered for further analyses.

2.2.2 16S/18S rRNA gene amplicon sequencing

Specific sets of primers targeting 16S and 18S rRNA genes were chosen for the amplification and subsequent sequencing to broadly cover bacterial, archaeal and eukaryotic diversity. The bacterial and archaeal community structures of the 65 samples were resolved by amplifying the V4 region of the 16S rRNA gene using the universal primers 515F and 805R (515F_GTGBCAGCMGCCGCGGTAA; 805R_GACTACHVGGGTATCTAATCC) (Hugerth et al., 2014; Herlemann et al., 2011). This primer pair covers the bacterial domain, including the phylum Actinobacteria and additionally resolves the archaeal domain.

The eukaryotic community structures for 63 samples were analyzed by amplifying the V4 region of the 18S rRNA gene using primers 574*F and 1132R (574*F_ CGGTAAYTCCAGCTCYV; 1084r_CCGTCAATTHCTTYAART) (Hugerth et al., 2014). Two samples did not yield sufficient amplicons (CSD infant 7 collected on days 1 and 3).

The KAPA HiFi HotStart ReadyMix (Kapa Biosystems, Wilmington, MA, USA) was used for amplification with 25 cycles and according to the service provider's standards. Paired-end sequencing with 2 x 300 nt was performed on an Illumina MiSeq platform with the V3 MiSeq kit at the Center of Analytical Research and Technology – Groupe Interdisciplinaire de Génoprotéomique Appliquée (CART-GIGA; Liège, Belgium).

2.2.3 16S rRNA and 18S rRNA gene sequencing data processing

The raw 16S rRNA gene amplicon sequencing data were processed using the LotuS software (version 1.35) with default parameters (Hildebrand et al., 2014) and using the SILVA database (Quast et al., 2013; Yilmaz et al., 2014). After clustering the

In review

reads into operational taxonomic units (OTUs) at 97% identity level, they were classified using the Ribosomal Database Project (RDP) classifier (Wang et al. 2007). OTUs with a confidence level below 0.8 at the domain level were discarded. The amplicon sequences belonging to the 100 most abundant OTUs were additionally manually curated for unspecific amplification. As only few archaeal reads were detected, the overall quality of the archaeal reads were manually assessed using the FASTQC results[1]. As the paired-end 18S rRNA gene amplicon reads obtained in this study did not overlap, a specifically tailored workflow was used to process the raw 18S rRNA gene amplicon sequencing data[2]. For the classification step and the taxonomic assignment, the PR2 database (Guillou et al., 2013) was used according to Hu et al. (2016).

2.2.4 16S rRNA and 18S rRNA gene sequencing data analysis

For both prokaryotic and eukaryotic datasets, we removed OTUs that were represented by less than 10 reads in all of the sequenced samples. Samples yielding less than 5,000 16S rRNA gene amplicon reads necessary for assessing bacterial diversity (Lundin et al., 2012) were excluded. As the complexity of the microeukaryotic community structure is largely undetermined and no previous recommendations exist, no cutoff for the number of 18S rRNA gene amplicon reads was applied. All statistical analyses and visualisations were performed using the R statistical software package (version 3.2.0) (R Development Core Team, 2008). Per-sample normalization, calculations of richness, diversity (Shannon's diversity index), evenness (Pielou's evenness index), dissimilarity index (distance to the most mature sample, calculated using Soerensen's similarity index of presence/absence of taxa at each individual time point compared to samples collected at the last individual time point) and non-parametric estimation of minimum community richness according to Chao et al. (1984) were performed using the 'vegan' package[3]. For the calculations of diversity and evenness indices for microeukaryotes, only samples with a total of more than 10 reads were considered. Differential analysis of relative OTU abundances based on read count data for the 16S rRNA gene amplicon sequencing dataset was done using the 'DESeq2' package (Love et al., 2014), which allows testing for differential abundance using negative binomial generalized linear models and multiple-testing adjustment by controlling the false discovery rate (Benjamini and Hochberg, 1995). Adobe Illustrator (version 19.1.0) was used for labeling axes and creating multi-plot graphs.

Various neonatal characteristics that were previously shown to have an impact on the microbiome (e.g. delivery mode, fed milk type, gestational age, maternal antibiotic and probiotic intake, positive screening for Group B *Streptococcus* (*Streptococcus agalactiae*) colonization of the mother) were compared between samples using the Wilcoxon rank sum test or Kruskal-Wallis test where applicable and comparisons with *P*-value <0.05 were considered statistically significant. Principal coordinate analysis (PCoA) graphs were generated using the Jensen-Shannon distance as implemented in the R package 'phyloseq' (McMurdie and Holmes, 2013) and clusters were defined using the partitioning around medoids (pam) function contained in the R package 'cluster' (Maechler et al., 2015).

In review

3. Results

3.1 Cohort characteristics

65 fecal samples were collected between September 2012 and April 2014 at the CHL from eight healthy VD and seven healthy CSD infants at six time points (samples collected around days 1, 3, 5, 28, 150 and 365). The birth weights as well as the gestational ages of the infants were similar, while the ratios of genders, the maternal age and the maternal postnatal BMI differed between both groups, with the CSD group comprising more male infants as well as mothers with a higher average age and postnatal BMI (Table 2). Three mothers who gave birth vaginally screened positively for Group B *Streptococcus*, whereas all mothers giving birth by C-section were screened negatively. Clinical healthcare guidelines in Luxembourg recommend that mothers who were screened positively for Group B *Streptococcus* should be treated intravenously with antibiotics prior to birth. Although mothers undergoing C-section were preferentially treated with antibiotics prior to birth, the majorities of both cohorts received antibiotic treatment (Table 2). Two of the three mothers who did not receive any antibiotics prior to birth chose to take probiotics during their pregnancies, whereas none of the other mothers recorded any probiotic supplementation. Out of eight VD infants, four were fed purely with maternal breast milk, while two others received formula milk and the remaining two were fed a combination of formula and breast milk. Out of the seven CSD infants, five were purely fed breast milk and the remaining two received a combination of breast milk and formula (Supplementary File 1, Table S1). According to the self-assessment of mothers that were purely breastfeeding, both the frequency and duration of feeding were not significantly different between VD and CSD infants. Introduction of solid food occurred in average around day 150 for all infants.

In review

3.2 Assessment of bacterial, fungal and archaeal load using real-time PCR

Specific qPCR assays using previously published primers were used to obtain quantitative information on the individual taxonomic groups of interest (Table1). Absolute yields of extracted DNA were quantified and prokaryotic and fungal DNA, as well as the relative quantities of archaea were calculated based on the ratio between the relative concentrations obtained for the universal prokaryotic primer pair and the relative concentrations obtained for archaea (Fig 1). As negative controls for the qPCR quantifications, sample-free 'DNA mock extracts' were prepared and subjected to qPCR analyses. The detection of organisms in the mock extracts reflecting the three domains of life was negative for the archaea- and fungi-specific primer sets whereas the universal prokaryotic primer set resulted in the detection of a minimal amount of DNA close to the qPCR detection limit, (average concentration of 0.002 ng/µl measured for the 'DNA mock extracts' as opposed to 0.3 ng/µl measured for meconium samples, i.e. the earliest fecal material excreted by infants, which had the lowest observed concentrations amongst all study samples). Therefore, the mock extracts and subsequent analyses did not indicate the presence of reagent-derived contaminants.

The qPCR-based quantification of prokaryotic DNA was successful for 64 out of 65 samples, with yields ranging from 0.2 ± 0.4 ng of DNA per mg of stool (average ± standard deviation) in the meconium samples of day 1, to 16.6 ± 6.4 on day 365. Generally, the prokaryotic load of both cohorts increased considerably after the introduction of food. The DNA yields were dependent on the collection time point, and the greatest differences were observed between day 1 and all other collection time points (Fig.1A; for all significant differences between collection time points, see Supplementary File 1, Table S2). Moreover, at day 5 significantly lower extraction yields (*P*-value = 0.03; Wilcoxon rank sum test) were observed for samples derived from infants whose mothers received antibiotics prior to birth (Supplementary File 1, Fig. S1).

The presence of archaea was detected in 91% of all samples (59 out of 65 samples) and the relative concentration of archaeal DNA in relation to the mean of all samples ranged from 5.5 ± 7.8 on day 1 to 0.5 ± 0.4 on day 365. Generally, more samples were found to be positive in VD (97 % of VD infant samples) than in CSD infants (86 % of CSD infant samples) and archaeal presence was as well detected in the samples from the very first time points (Fig. 1B).

Presence of fungal organisms was detected in 37 % (24 out of 65 samples) of all samples, ranging from 0.0007 ± 0.0005 ng of fungal DNA per mg of stool on day 3 to 0.002 ± 0.002 ng of fungal DNA per mg of stool on day 365, with generally more samples being positive for fungi in VD (43 % of VD infant samples) compared to CSD infants (31 % CSD infant samples). Fungi were detected earliest at day 3 in VD and at day 5 in CSD infants. The fungal DNA yield tended to increase over time, even though the magnitude of the increase was smaller compared to prokaryotes (Fig. 1C).

3.3 Validation of GIT microbiome profiles in low-yield samples

The absolute quantification of prokaryotic 16S rRNA gene copy numbers in all samples showed that the earliest samples contained significantly less microbial DNA compared to all other visits (Fig. 1, Supplementary File 1, Table S2). In order to exclude any biases by low-yield samples (Salter et al., 2014; Jervis-Bardy et al., 2015), we extracted an additional adult stool sample using the same protocol and created a dilution series ranging from 2 to 0.002 ng/µl. The four DNA dilution samples were 16S rRNA gene amplicon sequenced using the same primer pair as for the collected study samples (see the reported results below).

The undiluted sample, reflecting the concentration of most samples in the study (Fig. 1), and all three dilutions, simulating low-yield samples, showed highly comparable diversity and evenness indices (Supplementary File 1, Fig. S2A). For richness, the undiluted sample and both 10-fold and 100-fold diluted samples had highly comparable results, while the 1,000-fold dilution caused a slight decrease. This loss of observed richness is also reflected in a slightly increased dissimilarity index for the 1,000-fold diluted sample compared to the undiluted sample. Considering the observed taxonomic composition with decreasing DNA concentration, all three dilutions showed high resemblance to the undiluted sample, while the 100-fold and 1,000-fold dilutions showed slightly over-estimated relative abundances for *Roseburia* spp. and *Collinsella* spp. and a slight under-estimation for *Bacteroides* spp. (Supplementary File 1, Fig. S2B). However in each case, a similar taxonomic profile to the one in the undiluted sample was observed and potential reagent contaminants or sequencing artifacts did not have a significant effect on the taxonomic composition in the low-yield samples. These data indicated that the chosen approach allowed the comparison of samples with low extraction yields to those with higher yields.

In review

3.4 Generated amplicon sequencing data

After the 16S rRNA gene sequencing and following the primary data processing and filtering, a total of 13,136,451 reads were retained and used for the subsequent analyses. With 205,000 ± 90,000 reads per sample (average ± standard deviation), a total of 1,053 unique OTUs were identified. One out of the 65 samples was excluded from further 16S rRNA gene sequencing analysis due to poor coverage (sample collected at day 3 for VD infant 8).

For the processed 18S rRNA gene amplicon sequence data, only OTUs reflecting the microeukaryotic members of the microbiome were considered. To achieve this, we manually curated the dataset of initially 3,376,004 reads by removing classified OTUs that belonged to the following clades containing multicellular organisms: Metazoa (total of 3,302,231 reads), Chlorophyta (total of 4,611 reads), Streptophyta (total of 7,414 reads) and Agaricomycetes (7,038 reads). After filtering out OTUs that were represented by less than 10 reads, a total of 60,476 reads (average of 960 ± 1,540 reads per sample) and 152 microeukaryotic OTUs were retained for the subsequent analyses.

3.5 Prominent bacterial, archaeal and microeukaryotic taxa

In order to resolve which specific taxa were present during neonatal GIT colonization, we first identified the most common and abundant OTUs in the 16S rRNA gene amplicon sequencing data, which belonged to the phyla Proteobacteria, Actinobacteria, Firmicutes, Bacteroidetes and Verrucomicrobia (Fig. 2A). Bacterial genera present in all samples ('core populations') included *Bifidobacterium* spp., *Escherichia/Shigella* spp., *Bacteroides* spp., *Streptococcus* spp. and *Enterococcus* spp., with the first three genera also being the bacterial taxa represented by the most reads out of the total of sequencing reads in all samples (Supplementary File 2).

Within the 16S rRNA gene sequencing data, two OTUs belonging to the domain archaea were identified. OTU 1128 was assigned to the genus *Methanosphaera* and comprised a total of 25 reads in a single sample (day 1 for a VD5, 0.02 % of reads; Supplementary File 2). Despite being low in abundance, reads of OTU 1128 (*Methanosphaera* sp.) were of good quality and allowed us to confidently ascertain the presence of this organism in this sample (Supplementary File 1, Fig. S3A-B). Meanwhile, OTU 693, assigned to the genus *Methanobrevibacter*, was found in four samples represented by one to 11 reads but showed insufficient sequence quality for a confident classification (Supplementary File 1, Fig. S3C-D).

Overall, microeukaryotic taxa were less frequent in the individual samples compared to bacterial taxa, with fewer OTUs and without specific 'core' OTUs, which were detected in all samples. The most represented fungal phyla in all samples belonged to the phyla Basidiomycota and Ascomycota (Fig. 2B), with the genus *Saccharomyces* and the class Exobasidiomycetes having been detected in more than 40 % of the samples (Supplementary File 3).

Interestingly, meconium samples already presented a relatively large diversity of different prokaryotic and microeukaryotic populations. For prokaryotes, a total of 674 OTUs were detected in the 10 collected meconium samples (miniumum of 109 OTUs, maximum of 347; Supplementary File 4). OTUs that were detected in all meconium samples included *Escherichia/Shigella* spp. and *Bifidobacterium* spp., which were

In review

also two of the taxa with the highest read counts over all samples. *Enterobacter* spp., *Staphylococcus* spp., *Streptococcus* spp., *Veillonella* spp., *Bacteroides* spp., *Prevotella* spp., *Clostridium sensu stricto* spp., *Delftia* spp. and *Blautia* spp. were also detected across all meconium samples. For the microeukaryotic community, a total of 45 OTUs were detected in the 9 sequenced meconium samples (Supplementary File 5). The most frequently detected OTU (in 77.8 % of meconium samples) belonged to *Exobasidiomycetes* spp., while *Saccharomyces* spp., represented by the two most dominant OTUs with the highest relative abundances, were detected in more than half of the meconium samples.

3.6 Colonization and succession

As the amount of microbial DNA in the infants' stool increased with time, we analyzed whether the increase in microbial biomass was accompanied by a change in community characteristics such as richness or diversity. Based on the 16S and 18S rRNA gene amplicon data, we calculated overall richness, diversity, evenness and dissimilarity indices for the prokaryotic (bacterial and archaeal) (Fig. 3A-D) and microeukaryotic (Fig. 3E-H) datasets over the entire cohort. Non-parametric estimation of community richness for the individual time points according to Chao et al. (1984) for prokaryotes and microeukaryotes showed comparable trends to the estimation of richness based on the numbers of different OTUs (Supplementary File 1, Fig S4). Given the sparseness and low abundance of archaeal OTUs detected by 16S rRNA gene amplicon sequencing, the observed patterns regarding prokaryotic diversity were mostly driven by bacterial taxa.

In review

A significantly higher bacterial richness (number of different OTUs) was observed for the meconium samples compared to all other collection time points (Fig. 3A, Supplementary File 1, Table S3). In general, the inter-individual variability in richness was high on the first two sampling dates. The lowest richness of any sample was observed on day 3 *postpartum* and the overall median richness was lowest on day 5. The median richness increased subsequently and stabilized between day 28 and 150 (Fig. 3A). The observed microeukaryotic richness tended towards a lower median richness at the end of the first year and showed a high level of variability throughout the first year of life (Fig. 3E; Supplementary File 1, Table S4).

Shannon diversity and evenness metrics (Fig. 3B and Fig. 3C respectively) showed comparable trends for prokaryotic OTUs, i.e. a decrease in diversity and evenness with a concomitant decrease in variation in both diversity and evenness between individuals until day 5 *postpartum*. This was followed by a gradual increase for the subsequent collection time points. The observed microeukaryotic diversity and evenness (Fig. 3F and Fig. 3G respectively) followed no discernible trends compared to the bacterial data and exhibited constantly high levels of inter-individual variation. When linking samples according to the type of milk the infants received per time point, it became apparent that at day 5 and 28, infants that received combined feeding and formula-fed infants had a significantly lower microeukaryotic diversity compared to breast milk-fed infants (*P*-value = 0.01 at day 5 and *P*-value = 0.03 at day 28; Kruskal-Wallis test).

We calculated the Soerensen distance between the community structure at each time point and the community structure of the same individual in the most mature sample, i.e. usually the sample collected at 1 year, and compared the distances as a measure for maturity. For the prokaryotic dataset, the distances to the most mature sample exhibited a decreasing trend over time (Fig. 3D). The observed patterns suggested a gradual development towards the 1 year samples, with day 150 exhibiting significantly more similarities to the most mature samples compared to the samples collected at day 1 (*P*-value = 0.009; Wilcoxon rank sum test). The same trend was observed for the Spearman correlation between the different time points (Supplementary File 1, Fig. S5A), with samples of day 150 being significantly more correlated to the most mature microbiome than samples of day 1 (*P*-value = 0.004; Wilcoxon rank sum test). In contrast, the distances to the most mature microbial composition for the microeukaryotic microbiota (Fig. 3H) as well as the Spearman correlation (Supplementary File 1, Fig. S5B) displayed high variability among infants and between time points, and remained variable over time without reaching a certain level of maturity in regard to the 1 year samples.

3.7 Comparison of microbiome community profiles of VD and CSD infants

Absolute quantification of 16S rRNA gene counts by qPCR showed that CSD infants carried significantly lower bacterial loads and thereby a decreased colonisation density at day 3 and day 150 (*P*-value = 0.03 and *P*-value = 0.04 respectively; Fig. 1A; Wilcoxon rank sum test). At the same time, CSD infants had microbial community structures with a significantly higher richness compared to VD infants at day 3 (*P*-value = 0.02; Wilcoxon rank sum test; Fig. 3A).

To provide an overview of the development of the microbiome of the eight VD (34 samples) and the seven CSD infants (30 samples), the 16S and 18S rRNA gene amplicon data were represented by an ordination of their respective Jensen–Shannon distances (Fig. 4), a method that is commonly used for human microbial community structure analyses (Koren et al., 2013). Clusters on the PCoA plots were defined by partitioning around medoids (Maechler et al., 2015). For the prokaryotic community structure, samples collected at one year clustered together independently of delivery mode (Cluster I in Fig. 4A and B), whereas most samples collected for CSD infants around days 3 and 5 *postpartum* were located in Cluster II (Fig. 4B). In order to identify cluster-specific taxa, we compared the taxa in both clusters using DESeq2, resulting in 52 OTUs that were significantly different in their DESeq2-normalized read numbers between both clusters (Supplementary File 6). Among the top 10 OTUs with the smallest adjusted *P*-values ranging from $1.41*10^{-18}$ to $3.06*10^{-04}$, 6 OTUs belonged to the genus *Streptococcus* and always one OTU belonged to the genera *Proteus*, *Haemophilus* and *Rothia*, which all exhibited increased abundances in Cluster II; and one OTU classified as *Bifidobacterium* spp. which was more abundant in Cluster I.

Similar to the 16S rRNA gene sequence data, the 18S rRNA data exhibited two clusters (Fig. 4C and 4D). One cluster (Cluster III) comprised all samples except for the samples belonging to three VD infants (Cluster IV), while the microeukaryotic community composition of one VD infant transitioned between both clusters (Fig. 4C). When comparing the taxonomic compositions in samples between both clusters (III and IV) using the Wilcoxon rank sum test and adjusting for multiple testing, eight OTUs, with six unclassified OTUs and two OTUs classified as *Candida* spp., were detected to be differentially abundant in both clusters with *P*-values ranging between $5.94*10^{-10}$ to $2.63*10^{-02}$ (Supplementary File 7). These OTUs were increased in

In review

abundance in samples belonging to Cluster IV, but were most often missing or decreased in abundance in samples from Cluster III.

3.8 Depletion of Bacteroidetes in CSD infants

The most profound difference between CSD and VD infants was observed for the Firmicutes/Bacteroidetes ratio. While both phyla were approximately equally abundant in the VD infants (Fig. 5), the corresponding ratio was significantly higher for CSD infants at days 5 (*P*-value = 0.006), 28 (*P*-value = 0.005) and 150 (*P*-value = 0.01; Wilcoxon rank sum test) while the proportional abundance for the phylum Bacteroidetes was significantly decreased in samples from CSD infants over most of the sampling time points (day 5: *P*-value = 0.006, day 28: *P*-value = 0.003, day 150: *P*-value = 0.01, day 365: *P*-value = 0.04; Wilcoxon rank sum test; Supplementary File 1, Fig. S6A). At the same time, there was a concomitant increase in Firmicutes at day 5 in CSD infants (*P*-value = 0.01; Wilcoxon rank sum test). Preceding the drastic decrease in Bacteroidetes at day 5, there was already a significant difference at day 3 between infants born at different gestational ages, whereby full term (≥39 weeks) infants showed a higher relative abundance of Bacteroidetes when compared to late preterm (34-36 weeks) and early term (37-38 weeks) born infants (*P*-value = 0.05; Kruskal Wallis test; Supplementary File 1, Fig. S7).

In addition, we also more specifically analyzed richness, evenness and diversity within the Bacteroidetes phylum (Fig. 6). We observed a significant decrease in the Bacteroidetes richness in CSD infants at day 28 compared to VD infants (*P*-value = 0.01; Wilcoxon rank sum test; Fig. 6A). The relative abundance of the genus *Bacteroides*, which made up more than 10 % of the reads in most VD infants at days 28 and 150, exhibited a significant decrease in abundance associated with a delayed colonization in CSD infants (*P*-value = 0.04 at day 28 and 0.01 at day 150; Wilcoxon rank sum test; Supplementary File 1, Fig. S6B). Due to this significant decrease in relative abundance of *Bacteroides* spp. compared to earlier and later time points in CSD infants and the subsequent shift in dominance inside the Bacteroidetes phylum, the diversity and evenness inside this phylum at day 28 were significantly increased (P-value = 0.005 for both; Wilcoxon rank sum test; Fig. 6B and Fig. 6C). The different measures of diversity and evenness within the Firmicutes phylum did not show any significant differences between both delivery modes.

3.9 Additional differences in prokaryotic community structure in CSD infants

We further aimed to determine whether other bacterial taxa also showed different changes in CSD infants compared to VD infants during the first year of life. We identified taxa that were differentially abundant according to delivery mode and at each collection time point. After filtering the resulting 88 differentially abundant OTUs according to a cumulative read count above 10,000, we retrieved 29 OTUs with a positive fold change in CSD infants compared to VD infants and four OTUs that exhibited a negative fold change (Supplementary File 8). The same analysis was performed at the genus level and resulted in three genera with a negative fold change and 20 with a positive fold change in CSD compared to VD infants (Supplementary File 9).

The fecal microbiome of CSD infants was associated with increased proportional abundances of, amongst others, OTUs assigned to the genera *Haemophilus* spp., *Streptococcus* spp., *Enterobacter* spp., *Propionibacterium* spp., *Staphylococcus* spp.

In review

and the genus *Lactobacillus* over the first year of life. Furthermore, the microbiome of CSD infants contained lower proportions of *Bacteroides* spp. and *Parabacteroides* spp.

In order to validate that CSD infants harboured substantially different relative abundances of certain prokaryotic populations compared to VD infants at certain time points, we amplified specific target regions of the genera *Staphylococcus* spp. and *Streptococcus* spp. (at days 3 and 5), *Haemophilus* spp. and *Lactobacillus* spp. (at days 3 and 28) and the two phyla Firmicutes and Bacteroidetes (at days 5 and 28), to calculate their relative abundances. Validation by qPCR was done on samples that were collected on days on which the differences in relative abundances between both delivery modes were most pronounced. All targeted differences between CSD and VD children obtained in the previous differential analysis could be confirmed by qPCR analysis for the specific collection time points (Fig. 7).

4. Discussion

4.1 Detection of prokaryotic and microeukaryotic communities in meconium

A number of recent studies indicate that meconium samples are not sterile but contain complex bacterial communities (Jiménez et al., 2008; Gosalbes et al., 2013; Ardissone et al., 2014). In this context, the previously accepted dogma of intrauterine sterility has been questioned. According to our results based on qPCR analyses as well as 16S and 18S rRNA gene amplicon sequencing, representatives of all three domains of life were present in meconium samples. Given that DNA yield out of meconium samples was limited (Fig. 1), it could be possible that this microbial DNA might not be derived from the samples but may in fact represent contaminants of the reagents used for DNA extraction (Salter et al., 2014; Jervis-Bardy et al., 2015). However, according to simultaneously conducted analyses, even a 1,000-fold dilution of DNA extracted from an adult stool sample did not considerably change the taxonomic composition compared to the undiluted and 10- to 100-fold diluted samples (Supplementary File 1 Fig. S2B). From these results, we deduced that potential reagent contaminants did not have any significant impact on the overall composition observed in our study. Moreover, the fact that we observed a significantly increased prokaryotic richness and diversity in meconium samples (Fig.3A-B) stood in stark contrast to the results from the dilution series, which revealed a decreased richness along with a stable diversity in the low-yield samples due to several taxa being diluted out of the adult stool sample during the 1,000-fold DNA dilution process (Supplementary File 1 Fig. S2A). Additionally, the sequencing of all 'DNA mock extracts' yielded very low coverage, while the detection of representatives of all three domains of life by qPCR could be considered negative as well. Taking these results into account, we suggest that the detection of taxa inside the meconium samples was not an artifact but had to be considered genuine. Whether the neonatal GIT was colonized prenatally or whether detected microbial populations were acquired perinatally could not be assessed in the context of our study.

The bacterial richness was significantly higher in meconium samples than at later time points. Samples from the first day were also highly diverse and the taxa were evenly distributed compared to subsequent collection time points, which suggests that these samples captured the potential early pioneering microbiota, most of which did not stably colonize the GIT thereafter. Subsequently, the richness decreased during the following days as the initial colonizers took hold in the GIT. Some of the taxa detected in the meconium samples may have been present in later samples but were

In review

not captured due to the masking by the dominant taxa. At day 1, the most abundant bacterial taxa in all infants were *Escherichia/Shigella* spp., *Bifidobacterium* spp., *Enterobacter* spp., *Staphylococcus* spp., *Streptococcus* spp., *Prevotella* spp. and *Veillonella* spp., which have all been previously described in meconium samples as being pioneering genera of the human GIT (Gosalbes et al., 2013; Ardissone et al., 2014; Hansen et al., 2015). The latter four are either present predominantly on skin (Dominguez-Bello et al., 2010), in colostrum or are typical inhabitants of the oral cavity (Cabrera-Rubio et al., 2012). Pioneering bacterial colonizers of the microbiome are usually facultative anaerobes, such as *Escherichia* spp. (Jiménez et al., 2008), as also observed in our study. These pioneers shape the gastrointestinal microbiome environment, promoting the subsequent colonization by strict anaerobes such as *Bacteroides* spp., *Clostridium* spp., and *Bifidobacterium* spp., which were already detected in samples collected on day 1 in our study. Overall, the earliest bacterial colonizers detected in all meconium samples included both facultative and strict anaerobic taxa suggesting that the GIT rapidly transitions towards an anaerobic environment after birth. *Bifidobacterium* spp., which was the taxon with the highest read counts across all samples, are important for neonatal health and are known to have beneficial effects for the host through their breakdown of dietary carbohydrates, the products of which directly feed into host metabolism (Davis et al., 2011). *Bifidobacterium* spp. are colonizers of the vaginal microbiome and are supposedly transferred to the infant during vaginal delivery (Dominguez-Bello et al., 2010). However, while in line with previous findings (Jakobsson et al., 2014), no significant difference in *Bifidobacterium* spp. abundances between VD and CSD infants could be detected for meconium samples, suggesting that other routes of transmission are also very likely during neonatal colonization. Additionally, the growth of this specific taxon is promoted selectively by prebiotic oligosaccharides present in the maternal colostrum and breast milk (Zivkovic et al., 2011; Yu et al., 2013).

Results from the quantitative real-time PCR assay suggested that archaea, even if low in abundance, were amongst the earliest colonizers of the neonatal GIT microbiome. The only methanogenic archaeon that was identified using the 16S rRNA gene amplicon sequencing was *Methanosphaera* spp., which was exclusively detected in VD infant 5 at day 1. This human archaeal commensal has a highly restricted energy metabolism (Fricke et al., 2006), which makes it a specialized member of the gastrointestinal microbiome. Archaea have been shown to be ubiquitous members of the adult GIT microbiome (Dridi et al., 2009), were sporadically detected in the vaginal environment (Belay et al., 1990), and were shown to colonize the skin surface (Probst et al., 2013) and the oral cavity (Nguyen-Hieu et al., 2013). As the presence of archaea was also apparent in CSD infants and also in samples collected at day 1 in our study, we can postulate that transmission paths besides vaginal transmission, such as fecal-oral, oral-oral or by skin contact most probably occur perinatally.

The earliest microeukaryotic colonizers included *Exobasidiomycetes* spp. and 2 OTUs classified as *Saccharomyces* spp., which were detected in meconium from CSD infants, whereas *Dothideomycetes* spp. and *Pezizomycotina* were detected mostly in VD infants. A recent study found *Saccharomyces* spp. and *Dothideomycetes* spp. to be present in more than half of the analyzed adult stool samples (Mar Rodríguez et al., 2015), which make them common taxa of the human GIT microbiome. As the vaginal tract is largely colonized by yeasts such as *Saccharomyces* spp., vaginal delivery is supposedly linked to neonatal colonization by yeasts through vertical transmission

In review

from the mother's vaginal microbiome or through horizontal transmission from the environment and hands of family members as well as health care workers (Bliss et al., 2008; Lupetti et al., 2002).

If pioneering microbiota, including representatives from all three domains of life, have the potential to colonize the GIT microbiome prenatally (Greenhalgh et al., 2016), according to our results, birth still marked the time point of extensive microbial colonization, which further defined microbial succession. Clearly, more work needs to be undertaken on meconium and the crucial first hours of life to ascertain the different sources of the pioneering microbiota.

4.2 Colonization and succession within the neonatal GIT microbiome by prokaryotes and microeukaryotes during the first year of life

The progressive nature of neonatal GIT colonization and succession by prokaryotes was apparent through an increase in absolute prokaryotic DNA load (Fig. 1), overall alterations to community compositions (Fig. 2) as well as changes in richness, diversity and evenness (Fig. 3). A general trend regarding the prokaryotic community members is that their structure matures over the course of the first year of life. This maturation was reflected by increases in diversity and evenness over time, which has already been reported in previous studies (Jakobsson et al., 2014; Yatsunenko et al., 2012). However, in our study, significant differences in diversity and evenness between subsequently sampled time points were observed as early as between days 5 and 28 (Fig. 3B-C). The prokaryotic richness stabilized between days 28 and 150 (Fig. 3A). Similarly, the dissimilarity index, reflecting the distance of the taxonomic composition of each sample to the last collected sample per child, showed a decreasing trend (Fig. 3D), highlighting that the microbiome composition gradually changed from a neonatal profile towards the most mature composition available by 1 year of age.

A previous study, focusing on neonatal colonization, has found archaea to be transiently and almost exclusively present in the first few weeks of life during their sample collection, which was conducted until around 17 months (Palmer et al., 2007), whereas archaea are considered core members of the adult GIT microbiome (Dridi et al., 2009). While archaea could not be identified confidently by amplicon sequencing in our study after the first day, the more sensitive qPCR assays suggested that they were indeed present in 90% of all samples, opposing previous results and highlighting their potential importance in the maintenance of inter-species community networks (Hansen et al., 2011). Although the 16S rRNA gene amplifying primer used for sequencing covered both domains bacteria and archaea, the nature of GIT microbiome profiles, with bacteria making up the large majority of the composition, likely caused a lack of primer availability for archaea, potentially explaining why this domain was more extensively detected with qPCR using the archaea-specific primers rather than using the more generic 16S rRNA gene primers used for the amplicon sequencing. In the future, dedicated archaeal and bacterial primer sets may be used to allow better resolution of the archaea.

When considering the microeukaryotic community, no clear successional patterns were discernible. In line with previous studies involving culture-independent analyses of the GIT microbiome, most detected fungal taxa belonged to the phyla Ascomycota and Basidiomycota (Scanlan and Marchesi, 2008; Ott et al., 2008). In contrast to

In review

previous reports on adult GIT microbiota (Scanlan and Marchesi, 2008), identities and abundances of detected microeukaryotic taxa fluctuated strongly throughout the first year of life. Similarly, richness, diversity and evenness indices did not follow discernible trends over time (Fig. 3E-G). However, we found a more rapid microeukaryotic diversification in infants who were fed exclusively breast milk between days 5 and 28. This suggests a possible link between the infants' feeding regimes and early changes to microeukaryotic community development in the human GIT. When considering the intra-individual dissimilarity index in addition to the apparent large inter-individual variation, our findings indicated that the microeukaryotic community members were more dynamic compared to their prokaryotic counterparts (Fig. 3H). A previous study in the mouse GIT observed similar results with fungal populations varying substantially, while bacterial populations remained relatively stable over time (Dollive et al., 2013). Typically, only a small number of common genera, such as the genus *Saccharomyces*, and a large number of spurious taxa that have been barely reported previously have been described to form part of the human GIT microbiome (Suhr et al., 2015). The specific characteristics of these rare taxa suggest that they do not persist inside the GIT microbiome but are likely more transient in nature when compared to bacteria (Suhr et al., 2015). Also, fewer microeukaryotic species and individual microeukaryotes are found in the human GIT than bacteria, potentially explaining why the microeukaryotic community may be less robust in comparison to bacteria (Underhill and Iliev, 2014). Furthermore, according to our results, the general lack in successional patterns with regards to the microeukaryotes suggested that either the neonatal GIT would not allow any durable colonization by microeukaryotes, including known common microbiome members such as *Blastocystis* spp. or *Dientamoeba fragilis* (Scanlan et al., 2014), that the required ecological niches did not exist in the GIT during the first year of life or that those microeukaryotes never actually stably colonize the GIT as suggested before by Suhr et al. (2015).

In review

4.3 Prokaryotic differences in colonization and succession between CSD and VD infants

Diversity and evenness measures were not significantly different between CSD and VD infants (Fig. 3B-C), in contrast to the results from another recent study (Jakobsson et al., 2014). However, a difference between VD and CSD infants was observed early on in terms of the prokaryotic richness, which was significantly increased in CSD infants (Fig. 3A). This finding could reflect the different pioneering taxa between both delivery groups. Furthermore, we found that generally lower amounts of DNA were extracted from stool of CSD infants compared to VD infants using the same extraction protocol, suggesting a delay in the acquisition of prokaryotic biomass in the GIT of CSD infants. While the DNA yields quickly increased over time for VD infants, CSD infants showed a slower acquisition of a similar colonization density, which could be explained by either a delay in exposure to bacteria or the inoculation by fundamentally different microbial taxa, which could be less adapted to the human GIT and therefore exhibited lower growth rates.

In addition to differences in microbial loads during the first days after birth (Fig. 1A), we identified apparent differences in early prokaryotic succession. For instance, several samples taken from CSD infants during days 3 and 5 were found to share similarities in community structure (Cluster II) that were not typically observed in samples from VD children (Fig. 4A-B). These similarities included increased relative abundances of *Streptococcus* spp. and *Staphylococcus* spp (Supplementary File 6). These taxa are typically found in the oral cavity and on the skin surface and are supposedly transferred from mother to infant through skin contact in CSD infants (Dominguez-Bello et al., 2010). Furthermore, these samples showed significantly decreased relative abundances of *Bacteroides* spp. and *Bifidobacterium* spp., whose colonization has been shown to be delayed in CSD infants (Adlerberth et al., 2006; Penders et al., 2006; Sufang et al., 2007; Dominguez-Bello et al., 2010). Interestingly, allergic diseases have been previously associated with a low prevalence of *Bacteroides* spp. and *Bifidobacterium* spp. (Björkstén et al., 1999; Watanabe et al., 2003), and low levels of *Bifidobacterium* spp. together with significantly increased levels of *Staphylococcus* spp. have been associated with childhood obesity (Kalliomäki et al., 2008). Generally, the genus *Bifidobacterium* is associated with an enhanced epithelial barrier function (Cani et al., 2009). These findings are in line with the statistically higher risks of CSD infants of developing obesity (Mueller et al., 2015) or allergic diseases (Abrahamsson et al., 2012; Abrahamsson et al., 2014). Although the differences observed in our study were compelling, whether the observed microbiome signatures in CSD infants are directly causally linked to disease development later in life has yet to be established in larger infant cohorts with longer-term follow-up. After day 150, the observed differences between CSD and VD infants became less pronounced. This observed trend could have been driven by weaning the infants from an exclusive milk diet and/or the introduction of solid food around the same time. Previous studies showed that through the introduction of new and diverse nutrients, the microbiome quickly changes towards a more adult-like profile, thereby decreasing early differences in profiles caused by delivery mode or other maternal and neonatal characteristics (Koenig et al., 2011; Fallani et al., 2011).

Although the delivery mode appeared to have the strongest influence on differences between the infants, other factors may also contribute to the observed patterns. Most notably, reduced gestational age, higher maternal age, a higher maternal BMI and specific maternal antibiotic treatments are commonly observed in the context of CSD (van Schalkwyk et al., 2010; Al-Kubaisy et al., 2014; Euro-Peristat Preterm Group et al., 2014; Klemetti et al., 2016). For example, gestational age may have been an additional factor driving the early Bacteroidetes depletion. Already at day 3, Bacteroidetes were significantly decreased in five infants that were born late preterm (34-36 weeks) or early term (37-38 weeks) compared to four full term infants (≥39 weeks) (Supplementary File 1 Fig. S7). Known effects of preterm delivery on neonatal microbiome colonization include reduced levels of strict anaerobes such as *Bifidobacterium* spp. and *Bacteroides* spp. (Arboleya et al., 2012; Arboleya et al., 2016) and a slower microbial succession (La Rosa et al., 2014), all of which were observed in our study for samples collected from CSD infants. Another factor may have been maternal perinatal antibiotics intake which was associated with significantly lower amount of prokayotic DNA at day 5 (and a similar trend at days 1 and 28; Supplementary File 1 Fig. S1). Importantly, the antibiotic intake of the mother may have effects on the GIT microbiome of the infant, either directly, e.g. transfer from maternal blood via the blood-placental barrier prior to birth (Pacifici, 2006), or indirectly, e.g. transfer of antibiotics via breast milk *postpartum* (Zhang et al., 1997). As antibiotic administration is recommended in case of delivery by C-section, this could be yet another factor that had a negative influence on the observed delay in colonization and succession in CSD infants while even potentially inhibiting the succession rate in VD infants to a certain extent.

In review

Besides shifts in the early successional patterns and factors that could enhance the observed delay in colonization, we also observed fundamental differences in the taxonomic composition of CSD infants compared to VD infants and over all time points, such as a significantly decreased relative abundance of Bacteroidetes (Supplementary File 1, Fig. S6A), which remained prominent even at 1 year. The most drastic difference in microbiome composition was an elevated Firmicutes/Bacteroidetes ratio observed in CSD infants between days 5 and 150 (Fig. 5). An elevated Firmicutes/Bacteroidetes ratio has been previously linked to an increased energy harvesting capacity by the host and its potential contribution to the development of metabolic disorders such as diabetes, obesity or metabolic syndrome in adulthood (Turnbaugh et al., 2006), although more recent findings seem to suggest that evidence for the implication of the Firmicutes/Bacteroidetes ratio in human health may be weaker than previously assumed (Sze and Schloss, 2016). The differential analysis detected statistically significant alterations of additional bacterial taxa in CSD infants over all time points, of which several were also validated by qPCR (Fig. 7, Supplementary Files 8 and 9). As already highlighted previously, CSD infants harbored lower proportions of *Bacteroides* spp. and *Parabacteroides* spp., which again point out that CSD infants were subject to a delayed rate of colonization for the phylum Bacteroidetes and more specifically the genera *Bacteroides* and *Parabacteroides*. Taxa commonly derived from skin, the oral cavity and the environment exhibited an enrichment in CSD infants. These taxa included *Haemophilus* spp., *Streptococcus* spp., *Enterobacter* spp., *Propionibacterium* spp. and *Staphylococcus* spp., which have been previously found to be enriched in CSD infants, supposedly through skin microbiome transfer from mother to the newborn after birth (Dominguez-Bello et al., 2010; Bäckhed et al., 2015). Interestingly, CSD infants in our study were also enriched in the genus *Lactobacillus*. As *Lactobacillus* spp. are usually dominant in the vaginal microbiome, they are supposedly transferred from mother to infant during vaginal delivery, thereby being deficient and delayed in CSD infants (Grönlund et al., 1999; Adlerberth et al., 2006; Dominguez-Bello et al., 2010). Other routes of colonization however also include the administration of breast milk (Bäckhed et al., 2015).

4.4 General delay in colonization rates in CSD infants

Overall, archaea and fungi were more often detected by qPCR in VD infants compared to CSD infants, and the yield of fungal DNA was lower in CSD infants compared to VD infants, except at 1 year and after introduction of solid food in all infants. These findings indicate that the previously described delay in colonization and succession observed for bacteria in CSD infants may actually affect all three domains of life, adding valuable information to our current knowledge regarding neonatal colonization of the GIT microbiome.

The initial microbiome colonization process is especially crucial for the early stimulation and maturation of the immune system (Rizzetto et al., 2014; Houghteling and Walker, 2015; Mueller et al., 2015), such that the observed delay of all three domains of life in CSD infants may result in an altered immunostimulatory effect, which in turn may potentially have long-lasting effects in relation to human health. Whether the early disturbance and delay of the colonization and succession processes in CSD infants could potentially exacerbate or contribute to the higher risk of CSD infants to develop certain diseases, therefore requires additional immunological data. However, what has been observed so far is that due to the close contact between the

In review

developing GIT, the underlying immune system and the colonizing bacteria, the early microbiome acts as an important interface in the neonatal development of the immune system (Björkstén, 2004; Caicedo et al., 2005; Rautava et al. 2007, Eberl and Lochner, 2009). Substantial shifts of neonatal taxonomic compositions or disruptions of natural colonization and succession processes may thereby lead to changes in the long-term developmental processes and subsequent altering of immune development. Additionally, the timing of colonization plays an important role in neonatal immune programming. Previous studies on mouse models observed that a delayed microbial GIT colonization of germ-free mice caused long-term changes in the immune system (Sudo et al., 1997; Rautava et al. 2007, Eberl and Lochner, 2009; Hansen et al., 2012; Olszak et al., 2012).

These findings demonstrate that the composition and timing of early neonatal colonization in CSD infants are important factors influencing the microbial education of the developing immune system, which could result in long-term persistent alterations in systemic gene expression and increased disease predispositions.

5. Conclusions

Here, we describe for the first time the colonization of the neonatal human GIT resolved to all three domains of life. We demonstrate that bacteria but also archaea and microeukaryotes, predominantly fungi, were detectable in meconium samples and are thereby among the earliest colonizers of the neonatal GIT microbiome.

In review

In contrast to the patterns observed for prokaryotes, microeukaryotic abundances fluctuated strongly over time, suggesting that the microeukaryotic community did not reach a stable colonization state during the first year of life. Based on our results, the milk-feeding regime appeared to impact the early microeukaryotic colonization and diversification process. An important question in this context is whether a diverse microeukaryotic microbiome is more resilient to disturbances and beneficial for the host as it has been proposed in respect for bacterial constituents of the GIT microbiome.

As for the differences in colonization and succession between VD and CSD infants during the first year of life, our findings highlight that CSD infants experience a delay in colonization and succession affecting all three domains of life, generally complementing and further extending previous observations. Substantial shifts in the community compositions started as early as day 5 and were potentially caused by differences in time of incidental exposure of bacteria from the environment in CSD infants. We further suggest a potential link to earlier gestational age and maternal antibiotics intake. Given that the early microbiome supposedly shapes the immune system, our observations that CSD infants exhibited a different succession pattern early on raises the hypothesis that disturbances to the microbiome in the early stages of neonatal development might have long-lasting health effects. Although major differences between VD and CSD infants were less apparent at 1 year of age, the question whether differences in the early stimulation of the immune system by either the VD or the CSD microbiomes may change the infants' response to later perturbations, such as during the introduction of solid food, will require further in-depth studies. In order to answer these questions, high-frequency sampling of GIT microbiota along with resolving crucial immune characteristics over longer periods of time should be undertaken.

Further additional work is required to determine at which stage of infant development the GIT microbiome acquires a mature archaeal community structure, as well as when the transition between the highly dynamic early microeukaryotic microbiota and the stable adult microeukaryotic community is occurring. Open questions in this context revolve around which role these two domains play with respect to neonatal host metabolism, how they influence the host's immune system and how they influence the GIT microbiome through providing specific metabolic functions.

Our findings provide an important account of the neonatal colonization and succession within the human GIT microbiome by bacteria, archaea and microeukayotes. In particular, our findings highlight the need for studying all three domains of life in future longitudinal studies of microbial colonization and succession within the human GIT to finally understand how the individual taxa affect host physiology and how differences in colonization and succession of all three domains may contribute to the development of diseases later in life.

6. Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

In review

7. Authors' contributions

LW carried out the qPCR assays, data processing, comparative analyses of the 16S and 18S rRNA gene amplicon sequencing data and data interpretation, participated in the biomolecular extractions and wrote the manuscript. AHB, AH, CCL, CdB and PW were involved in data analysis and interpretation. AHB contributed towards biomolecular extractions and writing the manuscript. AH participated in the design of the study and in the sample and data collection. EELM, SN and CCL participated in data analysis and critical revision of the manuscript. LWH and AFA participated in the data processing of the 18S rRNA gene amplicon sequencing data and provided important advice. LB and JB participated in the sample collection and processing. CdB and PW conceived and coordinated the study and participated in its design and in writing the manuscript. All authors read and approved the final manuscript.

8. Funding

This work was partially funded by the IBBL as well as by an ATTRACT program grant (ATTRACT/A09/03) and a European Union Joint Programming in Neurodegenerative Diseases grant (INTER/JPND/12/01) awarded to PW, as well as Aide à la Formation Recherche grants to LW (AFR PHD-2013-5824125), CCL (AFR PHD/4964712) and SN (AFR PHD-2014-1/7934898) and a CORE junior to EELM (C15/SR/10404839), all funded by the Luxembourg National Research Fund (FNR). The continuation of the study has been ensured with the generous financial support of the Fondation André et Henriette Losch. AFA was funded by a grant from the Swedish Research Council VR (grant 2011-5689). The funding bodies had no impact on the design of the study, data collection, analysis, interpretation of data, nor in writing the manuscript.

9. Abbreviations

BMI: body mass index; **Cp:** Crossing point; **CSD:** caesarean section delivery; **C-section:** caesarean section; **FNR:** Luxembourg National Research Fund; **CART-GIGA:** Center of Analytical Research and Technology - Groupe Interdisciplinaire de Génoprotéomique Appliquée; **GIT:** gastrointestinal tract; **IBBL:** Integrated BioBank of Luxembourg; **ISBER:** International Society for Biological and Environmental Repositories; **OTU:** operational taxonomic units; **pam:** partitioning around medoids; **PCoA:** Principal Coordinate Analysis; **qPCR:** quantitative real-time PCR; **RDP:** Ribosomal Database Project; **VD:** vaginally delivered.

10. Acknowledgment

In silico analyses presented in this paper were carried out using the HPC facilities of University of Luxembourg (Varrette et al., 2014). We are deeply grateful for all the parents and infants that have and will participate in the study. We also thank the dedicated clinical staff for participant recruitment, neonatologists of the Clinique pédiatrique and gynecologists, especially Isabelle Cimino and Roxane Batutu and all others involved in sample and data collection at the CHL as well as for sample storage at the IBBL. Patrick May, Enrico Glaab, Jean-Pierre Trezzi and Anne Kaysen from the Luxembourg Centre for Systems Biomedicine, are thanked for regular and valuable discussions. Abdul Sheik (LCSB) is thanked for additional editing of the manuscript. Wouter Coppieters, Latifa Karim and Benoît Hennuy from CART-GIGA are thanked for their work and discussion on the sequencing. Furthermore, we are thankful for the assistance of Patricia Martins Conde, Sandy Collignon and Laura A. Lebrun (LCSB) with fecal sample extractions and the efforts of Lola Kourouma on metadata entry and cross-checking of the medical data as well as Falk Hildebrand for being readily helpful with questions regarding the LotuS software.

In review

11. Supplementary Material

All processed and filtered 16S and 18S rRNA gene amplicon sequencing datasets as well as additional analyses are given as supplementary files.

Supplementary File 1

Fig. S1.	Impact of maternal antibiotic intake prior to birth on the yield of prokaryotic DNA from neonatal stool samples.
Fig. S2.	Analysis of 16S rRNA gene amplicon data from a DNA dilution series.
Fig. S3.	Quality of the archaeal sequencing reads.
Fig. S4.	Microbial richness during colonization and succession of the infant GIT.
Fig. S5.	Spearman correlations etween samples from each time point compared to the individual most mature microbial community profiles represented by samples collected at the final time point per infant.
Fig. S6.	Differences between delivery modes in relation to relative abundances of Bacteroidetes and *Bacteroides* spp..
Fig. S7.	Relative abundances of Bacteroidetes in children born at different gestational ages.
Table S1.	Milk feeding regime of the infants prior to the different sample collection time points.
Table S2.	Results of Wilcoxon rank sum test comparing the yields measured for the prokaryotic and fungal DNA at different collection time points.
Table S3.	Results of Wilcoxon rank sum tests for prokaryotic diversity, evenness, richness and dissimilarity indices at different collection time points.
Table S4.	Results of Wilcoxon rank sum tests for microeukaryotic diversity, evenness, richness and dissimilarity indices at different collection time points.

Supplementary File 2

Number of reads obtained from the 16S rRNA gene amplicon sequencing data at OTU level and in all samples.

Supplementary File 3

Number of reads obtained from the 18S rRNA gene amplicon sequencing data at OTU level and in all samples.

Supplementary File 4

Sequencing reads obtained from the 16S rRNA gene amplicon sequencing data at OTU level in meconium samples.

Supplementary File 5

Sequencing reads obtained from the 18S rRNA gene amplicon sequencing data at OTU level in meconium samples.

Supplementary File 6

List of OTUs with significantly different relative abundances between Clusters I and II.

Supplementary File 7

List of OTUs with significantly different relative abundances between Clusters III and IV.

Supplementary File 8

List of OTUs resulting from the differential analysis (DESeq2) that were significantly differentially abundant across all collection time points and comparing VD and CSD infants.

Supplementary File 9

List of genera resulting from the differential analysis (DESeq2) that were significantly differentially abundant across all collection time points and comparing VD and CSD infants.

In review

12. References

Abrahamsson, T. R., Jakobsson, H. E., Andersson, A. F., Björkstén, B., Engstrand, L., and Jenmalm, M. C. (2012). Low diversity of the gut microbiota in infants with atopic eczema. *J. Allergy Clin. Immunol.* 129, 434–440.e2. doi:10.1016/j.jaci.2011.10.025.

Abrahamsson, T. R., Jakobsson, H. E., Andersson, A. F., Björkstén, B., Engstrand, L., and Jenmalm, M. C. (2014). Low gut microbiota diversity in early infancy precedes asthma at school age. *Clin. Exp. Allergy.* 44, 842–850. doi:10.1111/cea.12253.

Adlerberth, I., Lindberg, E., Åberg, N., Hesselmar, B., Saalman, R., Strannegård, I.-L., et al. (2006). Reduced enterobacterial and increased staphylococcal colonization of the infantile bowel: an effect of hygienic lifestyle? *Pediatr. Res.* 59, 96–101. doi:10.1203/01.pdr.0000191137.12774.b2.

Al-Kubaisy, W., Al-Rubaey, M., Al-Naggar, R. A., Karim, B., and Mohd Noor, N. A. (2014). Maternal obesity and its relation with the cesarean section: a hospital based cross sectional study in Iraq. *BMC Pregnancy Childbirth.* 14. doi:10.1186/1471-2393-14-235.

Andersen, L. O., Nielsen, H. V., and Stensvold, C. R. (2013). Waiting for the human intestinal Eukaryotome. *ISME J.* 7, 1253–1255.

Arboleya, S., Binetti, A., Salazar, N., Fernández, N., Solís, G., Hernández-Barranco, A., et al. (2012). Establishment and development of intestinal microbiota in preterm neonates. *FEMS Microbiol. Ecol.* 79, 763–772. doi:10.1111/j.1574-6941.2011.01261.x.

Arboleya, S., Sánchez, B., Solís, G., Fernández, N., Suárez, M., Hernández-Barranco, A., et al. (2016). Impact of prematurity and perinatal antibiotics on the developing intestinal microbiota: a functional inference study. *Int. J. Mol. Sci.* 17, 649. doi:10.3390/ijms17050649.

Ardissone, A. N., de la Cruz, D. M., Davis-Richardson, A. G., Rechcigl, K. T., Li, N., Drew, J. C., et al. (2014). Meconium microbiome analysis identifies bacteria correlated with premature birth. *PLoS ONE.* 9, e90784. doi:10.1371/journal.pone.0090784.

Arrieta, M.-C., Stiemsma, L. T., Amenyogbe, N., Brown, E. M., and Finlay, B. (2014). The intestinal microbiome in early life: health and disease. *Front. Immunol.* 5. doi:10.3389/fimmu.2014.00427.

Bacchetti De Gregoris, T., Aldred, N., Clare, A. S., and Burgess, J. G. (2011). Improvement of phylum- and class-specific primers for real-time PCR quantification of bacterial taxa. *J. Microbiol. Methods.* 86, 351–356. doi:10.1016/j.mimet.2011.06.010.

Bäckhed, F., Roswall, J., Peng, Y., Feng, Q., Jia, H., Kovatcheva-Datchary, P., et al. (2015). Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* 17, 690–703. doi:10.1016/j.chom.2015.04.004.

Barrett, E., Kerr, C., Murphy, K., O'Sullivan, O., Ryan, C. A., Dempsey, E. M., et al. (2013). The individual-specific and diverse nature of the preterm infant microbiota. *Arch. Dis. Child-Fetal.* 98, F334–340. doi:10.1136/archdischild-2012-303035.

Belay, N., Mukhopadhyay, B., Conway de Macario, E., Galask, R., and Daniels, L. (1990). Methanogenic bacteria in human vaginal samples. *J. Clin. Microbiol.* 28, 1666–1668.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. St. B.* 57, 289–300.

Björkstén, B., Naaber, P., Sepp, E., and Mikelsaar, M. (1999). The intestinal microflora in allergic Estonian and Swedish 2-year-old children. *Clin. Exp. Allergy.* 29, 342–346.

Björkstén, B. (2004). Effects of intestinal microflora and the environment on the development of asthma and allergy. *Springer Semin. Immunopathol.* 25, 257–270. doi:10.1007/s00281-003-0142-2.

Bliss, J. M., Basavegowda, K. P., Watson, W. J., Sheikh, A. U., and Ryan, R. M. (2008). Vertical and horizontal transmission of *Candida albicans* in very low birth weight infants using DNA fingerprinting techniques. *Pediatr. Infect. Dis. J.* 27, 231–235. doi:10.1097/INF.0b013e31815bb69d.

Cabrera-Rubio, R., Collado, M. C., Laitinen, K., Salminen, S., Isolauri, E., and Mira, A. (2012). The human milk microbiome changes over lactation and is shaped by maternal weight and mode of delivery. *Am. J. Clin. Nutr.* 96, 544–551. doi:10.3945/ajcn.112.037382.

Caicedo, R. A., Schanler, R. J., Li, N., and Neu, J. (2005). The developing intestinal ecosystem: implications for the neonate. *Pediatr. Res.* 58, 625–628. doi:10.1203/01.PDR.0000180533.09295.84.

Cani, P. D., Possemiers, S., Van de Wiele, T., Guiot, Y., Everard, A., Rottier, O., et al. (2009). Changes in gut microbiota control inflammation in obese mice through a mechanism involving GLP-2-driven improvement of gut permeability. *Gut* 58, 1091–1103. doi:10.1136/gut.2008.165886.

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scand. J. Stat.* 11, 265–270.

In review

Cox, L. M., Yamanishi, S., Sohn, J., Alekseyenko, A. V., Leung, J. M., Cho, I., et al. (2014). Altering the intestinal microbiota during a critical developmental window has lasting metabolic consequences. *Cell.* 158, 705–721. doi:10.1016/j.cell.2014.05.052.

Davis, L. M. G., Martínez, I., Walter, J., Goin, C., and Hutkins, R. W. (2011). Barcoded pyrosequencing reveals that consumption of galactooligosaccharides results in a highly specific bifidogenic response in humans. *PLoS ONE.* 6, e25200. doi:10.1371/journal.pone.0025200.

Euro-Peristat Preterm Group, Delnord, M., Blondel, B., Drewniak, N., Klungsøyr, K., Bolumar, F., et al. (2014). Varying gestational age patterns in cesarean delivery: an international comparison. *BMC Pregnancy Childbirth.* 14. doi:10.1186/1471-2393-14-321.

Delzenne, N. M., Cani, P. D., Everard, A., Neyrinck, A. M., and Bindels, L. B. (2015). Gut microorganisms as promising targets for the management of type 2 diabetes. *Diabetologia*. 1–12. doi:10.1007/s00125-015-3712-7.

Dollive, S., Chen, Y.-Y., Grunberg, S., Bittinger, K., Hoffmann, C., Vandivier, L., et al. (2013). Fungi of the murine gut: episodic variation and proliferation during antibiotic treatment. *PLOS ONE.* 8, e71806. doi:10.1371/journal.pone.0071806.

Dominguez-Bello, M. G., Costello, E. K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., et al. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proc. Natl. Acad. Sci. U.S.A.* 107, 11971–11975. doi:10.1073/pnas.1002601107.

Dridi, B., Henry, M., El Khéchine, A., Raoult, D., and Drancourt, M. (2009). High Prevalence of *Methanobrevibacter smithii* and *Methanosphaera stadtmanae* detected in the human gut using an improved DNA detection protocol. *PLoS ONE.* 4, e7063. doi:10.1371/journal.pone.0007063.

Eberl, G., and Lochner, M. (2009). The development of intestinal lymphoid tissues at the interface of self and microbiota. *Mucosal. Immunol.* 2, 478–485. doi:10.1038/mi.2009.114.

Einsele, H., Hebart, H., Roller, G., Löffler, J., Rothenhofer, I., Müller, C. A., et al. (1997). Detection and identification of fungal pathogens in blood by using molecular probes. *J. Clin. Microbiol.* 35, 1353–1360.

Fallani, M., Amarri, S., Uusijarvi, A., Adam, R., Khanna, S., Aguilera, M., et al. (2011). Determinants of the human infant intestinal microbiota after the introduction of first complementary foods in infant samples from five European centres. *Microbiology+.* 157, 1385–1392. doi:10.1099/mic.0.042143-0.

Fierer, N., Jackson, J. A., Vilgalys, R., and Jackson, R. B. (2005). Assessment of soil microbial community structure by use of taxon-specific quantitative PCR assays. *Appl. Environ. Microbiol.* 71, 4117–4120. doi:10.1128/AEM.71.7.4117-4120.2005.

Fricke, W. F., Seedorf, H., Henne, A., Krüer, M., Liesegang, H., Hedderich, R., et al. (2006). The genome sequence of *Methanosphaera stadtmanae* reveals why this human intestinal archaeon is restricted to methanol and H2 for methane formation and ATP synthesis. *J. Bacteriol.* 188, 642–658. doi:10.1128/JB.188.2.642-658.2006.

Gosalbes, M. J., Llop, S., Vallès, Y., Moya, A., Ballester, F., and Francino, M. P. (2013). Meconium microbiota types dominated by lactic acid or enteric bacteria are differentially associated with maternal eczema and respiratory problems in infants. *Clin. Exp. Allergy.* 43, 198–211. doi:10.1111/cea.12063.

In review

Greenhalgh, K., Meyer, K. M., Aagaard, K. M., and Wilmes, P. (2016). The human gut microbiome in health: establishment and resilience of microbiota over a lifetime: the human gut microbiome in health. *Environ. Microbiol.*. doi:10.1111/1462-2920.13318.

Grönlund, M. M., Lehtonen, O. P., Eerola, E., and Kero, P. (1999). Fecal microflora in healthy infants born by different methods of delivery: permanent changes in intestinal flora after cesarean delivery. *J. Pediatr. Gastroenterol. Nutr.* 28, 19–25.

Guaraldi, F., and Salvatori, G. (2012). Effect of breast and formula feeding on gut microbiota shaping in newborns. *Front. Cell. Infect. Microbiol*. 94. doi:10.3389/fcimb.2012.00094.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., et al. (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucl. Acids Res.* 41, D597–D604. doi:10.1093/nar/gks1160.

Hamad, I., Sokhna, C., Raoult, D., and Bittar, F. (2012). Molecular detection of eukaryotes in a single human stool sample from Senegal. *PLoS ONE.* 7, e40888. doi:10.1371/journal.pone.0040888.

Hansen, C. H. F., Nielsen, D. S., Kverka, M., Zakostelska, Z., Klimesova, K., Hudcovic, T., et al. (2012). Patterns of early gut colonization shape future immune responses of the host. *PLoS ONE.* 7, e34043. doi:10.1371/journal.pone.0034043.

Hansen, E. E., Lozupone, C. A., Rey, F. E., Wu, M., Guruge, J. L., Narra, A., et al. (2011). Pan-genome of the dominant human gut-associated archaeon, *Methanobrevibacter smithii*, studied in twins. *Proc. Natl. Acad. Sci. U.S.A*. 108, 4599–4606. doi:10.1073/pnas.1000071108.

Hansen, R., Scott, K. P., Khan, S., Martin, J. C., Berry, S. H., Stevenson, M., et al. (2015). First-pass meconium samples from healthy term vaginally-delivered neonates: an analysis of the microbiota. *PLOS ONE.* 10, e0133320. doi:10.1371/journal.pone.0133320.

Herlemann, D. P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J. J., and Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J.* 5, 1571–1579. doi:10.1038/ismej.2011.41.

Hermann-Bank, M. L., Skovgaard, K., Stockmarr, A., Larsen, N., and Mølbak, L. (2013). The Gut Microbiotassay: a high-throughput qPCR approach combinable with next generation sequencing to study gut microbial diversity. *BMC Genomics.* 14, 788. doi:10.1186/1471-2164-14-788.

Hildebrand, F., Tadeo, R., Voigt, A. Y., Bork, P., and Raes, J. (2014). LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome.* 2, 30. doi:10.1186/2049-2618-2-30.

Horz, H.-P. (2015). Archaeal lineages within the human microbiome: absent, rare or elusive? *Life.* 5, 1333–1345. doi:10.3390/life5021333.

Houghteling, P. D., and Walker, W. A. (2015). Why is initial bacterial colonization of the intestine important to infants' and children's health? *J. Pediatr. Gastroenterol. Nutr.* 60, 294–307. doi:10.1097/MPG.0000000000000597.

Hu, Y. O. O., Karlson, B., Charvet, S., and Andersson, A. F. (2016). Diversity of pico- to mesoplankton along the 2000 km salinity gradient of the baltic Sea. *Front. Microbiol.*, 679. doi:10.3389/fmicb.2016.00679.

Hugerth, L. W., Muller, E. E. L., Hu, Y. O. O., Lebrun, L. A. M., Roume, H., Lundin, D., et al. (2014). Systematic design of 18S rRNA gene primers for determining

In review

eukaryotic diversity in microbial consortia. *PLoS ONE.* 9, e95567. doi:10.1371/journal.pone.0095567.

Hugerth, L. W., Wefer, H. A., Lundin, S., Jakobsson, H. E., Lindberg, M., Rodin, S., et al. (2014). DegePrime, a program for degenerate primer design for broad-taxonomic-range PCR in microbial ecology studies. *Appl. Environ. Microbiol.* 80, 5116–5123. doi:10.1128/AEM.01403-14.

Jakobsson, H. E., Abrahamsson, T. R., Jenmalm, M. C., Harris, K., Quince, C., Jernberg, C., et al. (2014). Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by caesarean section. *Gut.* 63, 559–566. doi:10.1136/gutjnl-2012-303249.

Jervis-Bardy, J., Leong, L. E. X., Marri, S., Smith, R. J., Choo, J. M., Smith-Vaughan, H. C., et al. (2015). Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome.* 3. doi:10.1186/s40168-015-0083-8.

Jiménez, E., Marín, M. L., Martín, R., Odriozola, J. M., Olivares, M., Xaus, J., et al. (2008). Is meconium from healthy newborns actually sterile? *Res. Microbiol.* 159, 187–193. doi:10.1016/j.resmic.2007.12.007.

Kalliomäki, M., Collado, M. C., Salminen, S., and Isolauri, E. (2008). Early differences in fecal microbiota composition in children may predict overweight. *Am. J. Clin. Nutr.* 87, 534–538.

Klemetti, R., Gissler, M., Sainio, S., and Hemminki, E. (2016). At what age does the risk for adverse maternal and infant outcomes increase - nationwide register-based study on first births in Finland in 2005-2014? *Acta. Obstet. Gyn. Scan.* doi:10.1111/aogs.13020.

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., et al. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4578–4585. doi:10.1073/pnas.1000081107.

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., et al. (2013). A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS. Comput. Biol.* 9, e1002863. doi:10.1371/journal.pcbi.1002863.

La Rosa, P. S., Warner, B. B., Zhou, Y., Weinstock, G. M., Sodergren, E., Hall-Moore, C. M., et al. (2014). Patterned progression of bacterial populations in the premature infant gut. *Proc. Natl. Acad. Sci. U.S.A.* 111, 12522–12527. doi:10.1073/pnas.1409497111.

Le Huërou-Luron, I., Blat, S., and Boudry, G. (2010). Breast- v. formula-feeding: impacts on the digestive tract and immediate and long-term health effects. *Nutr. Res. Rev.* 23, 23–36. doi:10.1017/S0954422410000065.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 550. doi:10.1186/s13059-014-0550-8.

Lundin, D., Severin, I., Logue, J. B., Östman, Ö., Andersson, A. F., and Lindström, E. S. (2012). Which sequencing depth is sufficient to describe patterns in bacterial α- and β-diversity? *Environ. Microbiol. Rep.* 4, 367–372. doi:10.1111/j.1758-2229.2012.00345.x.

Lupetti, A., Tavanti, A., Davini, P., Ghelardi, E., Corsini, V., Merusi, I., et al. (2002). Horizontal transmission of *Candida parapsilosis* Candidemia in a neonatal intensive care unit. *J. Clin. Microbiol.* 40, 2363–2369. doi:10.1128/JCM.40.7.2363-2369.2002.

In review

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2016). cluster: cluster analysis basics and extensions. R package version 2.0.5.

Mar Rodríguez, M., Pérez, D., Javier Chaves, F., Esteve, E., Marin-Garcia, P., Xifra, G., et al. (2015). Obesity changes the human gut mycobiome. *Sci. Rep.* 5. doi:10.1038/srep14600.

Martineau, F., Picard, F. J., Ke, D., Paradis, S., Roy, P. H., Ouellette, M., et al. (2001). Development of a PCR assay for identification of Staphylococci at genus and species levels. *J. Clin. Microbiol.* 39, 2541–2547. doi:10.1128/JCM.39.7.2541-2547.2001.

McFarland, L. V., and Bernasconi, P. (1993). *Saccharomyces boulardii'*. A review of an innovative biotherapeutic agent. *Microb. Ecol. Health. Dis.* 6, 157–171. doi:10.3109/08910609309141323.

McMurdie, P. J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE.* 8, e61217. doi:10.1371/journal.pone.0061217.

Miller, T. L., and Wolin, M. J. (1986). Methanogens in human and animal intestinal Tracts. *Syst. Appl. Microbiol.* 7, 223–229. doi:10.1016/S0723-2020(86)80010-8.

Mueller, N., Whyatt, R., Hoepner, L., Oberfield, S., Dominguez-Bello, M., Widen, E., et al. (2015). Prenatal exposure to antibiotics, cesarean section and risk of childhood obesity. *Int. J. Obes. (Lond)* 39, 665–670. doi:10.1038/ijo.2014.180.

Nayfach, S., Rodriguez-Mueller, B., Garud, N., and Pollard, K. S. (2016). An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Research* 26, 1612–1625. doi:10.1101/gr.201863.115.

Nguyen, D. M., and El-Serag, H. B. (2010). The Epidemiology of Obesity. *Gastroenterol. Clin. North Am.* 39, 1–7. doi:10.1016/j.gtc.2009.12.014.

Nguyen-Hieu, T., Khelaifia, S., Aboudharam, G., and Drancourt, M. (2013). Methanogenic archaea in subgingival sites: a review. *APMIS* 121, 467–477. doi:10.1111/apm.12015.

Olszak, T., An, D., Zeissig, S., Vera, M. P., Richter, J., Franke, A., et al. (2012). Microbial exposure during early life has persistent effects on natural killer T cell function. *Science.* 336, 489–493. doi:10.1126/science.1219328.

Ott, S. J., Kühbacher, T., Musfeldt, M., Rosenstiel, P., Hellmig, S., Rehman, A., et al. (2008). Fungi and inflammatory bowel diseases: alterations of composition and diversity. *Scan. J. Gastroentero..* 43, 831–841. doi:10.1080/00365520801935434.

Pacifici, G. M. (2006). Placental transfer of antibiotics administered to the mother: a review. *Int. J. Clin. Pharmacol. Ther.* 44, 57–63.

Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., and Brown, P. O. (2007). Development of the human infant intestinal microbiota. *PLoS Biol.* 5, e177. doi:10.1371/journal.pbio.0050177.

Pandey, P. K., Siddharth, J., Verma, P., Bavdekar, A., Patole, M. S., and Shouche, Y. S. (2012). Molecular typing of fecal eukaryotic microbiota of human infants and their respective mothers. *J. Bioscience.* 37, 221–226. doi:10.1007/s12038-012-9197-3.

Penders, J., Thijs, C., Vink, C., Stelma, F. F., Snijders, B., Kummeling, I., et al. (2006). Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics.* 118, 511–521. doi:10.1542/peds.2005-2824.

Probst, A. J., Auerbach, A. K., and Moissl-Eichinger, C. (2013). Archaea on human skin. *PLoS ONE.* 8, e65388. doi:10.1371/journal.pone.0065388.

In review

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596. doi:10.1093/nar/gks1219.

R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2008) ISBN 3-900051-07-0.

Rautava, S., and Walker, W. A. (2007). Commensal bacteria and epithelial cross talk in the developing intestine. *Curr. Gastroenterol. Rep.* 9, 385–392.

Rizzetto, L., De Filippo, C., and Cavalieri, D. (2014). Richness and diversity of mammalian fungal communities shape innate and adaptive immunity in health and disease. *Eur. J. Immunol.* 44, 3166–3181. doi:10.1002/eji.201344403.

Roccarina, D., Lauritano, E. C., Gabrielli, M., Franceschi, F., Ojetti, V., and Gasbarrini, A. (2010). The role of methane in intestinal diseases. *Am. J. Gastroenterol.* 105, 1250–1256. doi:10.1038/ajg.2009.744.

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., et al. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology.* 12, 87. doi:10.1186/s12915-014-0087-z.

Samuel, B. S., Hansen, E. E., Manchester, J. K., Coutinho, P. M., Henrissat, B., Fulton, R., et al. (2007). Genomic and metabolic adaptations of *Methanobrevibacter smithii* to the human gut. *Proc. Natl. Acad. Sci. U.S.A.* 104, 10643–10648. doi:10.1073/pnas.0704189104.

Scanlan, P. D., and Marchesi, J. R. (2008). Micro-eukaryotic diversity of the human distal gut microbiota: qualitative assessment using culture-dependent and -independent analysis of faeces. *ISME J.* 2, 1183–1193. doi:10.1038/ismej.2008.76.

Scanlan, P. D., Stensvold, C. R., Rajilić-Stojanović, M., Heilig, H. G. H. J., Vos, W. M. D., O'Toole, P. W., et al. (2014). The microbial eukaryote Blastocystis is a prevalent and diverse member of the healthy human gut microbiota. *FEMS Microbiol. Ecol.* 90, 326–330. doi:10.1111/1574-6941.12396.

Sekirov, I., Tam, N. M., Jogova, M., Robertson, M. L., Li, Y., Lupp, C., et al. (2008). Antibiotic-induced perturbations of the intestinal microbiota alter host susceptibility to enteric infection. *Infect. Immun.* 76, 4726–4736. doi:10.1128/IAI.00319-08.

Shah, P., Muller, E.E.L., Lebrun, L.A., Wampach, L., Wilmes. P. (2016) Sequential isolation of DNA, RNA, protein and metabolite fractions from murine organs and intestinal contents for integrated omics of host-microbiota interactions. *Methods Mol. Biol*. In press.

Sudo, N., Sawamura, S., Tanaka, K., Aiba, Y., Kubo, C., and Koga, Y. (1997). The requirement of intestinal bacterial flora for the development of an IgE production system fully susceptible to oral tolerance induction. *J. Immunol.* 159, 1739–1745.

Sufang, G., Padmadas, S. S., Fengmin, Z., Brown, J. J., and Stones, R. W. (2007). Delivery settings and caesarean section rates in China. *Bull. World Health Organ.* 85, 755–762.

Suhr, M. J., and Hallen-Adams, H. E. (2015). The human gut mycobiome: pitfalls and potentials - a mycologist's perspective. *Mycologia.* 107, 1057–1073. doi:10.3852/15-147.

Sze, M. A., and Schloss, P. D. (2016). Looking for a signal in the noise: revisiting obesity and the microbiome. *mBio.* 7, e01018–16. doi:10.1128/mBio.01018-16.

In review

Thauer, R. K., Kaster, A.-K., Seedorf, H., Buckel, W., and Hedderich, R. (2008). Methanogenic archaea: ecologically relevant differences in energy conservation. *Nat. Rev. Micro.* 6, 579–591. doi:10.1038/nrmicro1931.

Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature.* 444, 1027–131. doi:10.1038/nature05414.

Underhill, D. M., and Iliev, I. D. (2014). The mycobiota: interactions between commensal fungi and the host immune system. *Nat. Rev. Immunol.* 14, 405–416. doi:10.1038/nri3684.

van Ketel, R. J., de Wever, B., and van Alphen, L. (1990). Detection of *Haemophilus influenzae* in cerebrospinal fluids by polymerase chain reaction DNA amplification. *J. Med. Microbiol.* 33, 271–276. doi:10.1099/00222615-33-4-271.

van Schalkwyk, J., Van Eyk, N., Yudin, M. H., Boucher, M., Cormier, B., Gruslin, A., et al. (2010). Antibiotic prophylaxis in obstetric procedures. *J. Obstet. Gynaecol. (Canada)* 32, 878–884. doi:10.1016/S1701-2163(16)34662-X.

Varrette, S., Bouvry, P., Cartiaux, H., and Georgatos, F. (2014). Management of an academic HPC cluster: the UL experience. in (IEEE), 959–967. doi:10.1109/HPCSim.2014.6903792.

Vrieze, A., Van Nood, E., Holleman, F., Salojärvi, J., Kootte, R. S., Bartelsman, J. F. W. M., et al. (2012). Transfer of intestinal microbiota from lean donors increases insulin sensitivity in individuals with metabolic syndrome. *Gastroenterology.* 143, 913–916.e7. doi:10.1053/j.gastro.2012.06.031.

Walker, A. W., Martin, J. C., Scott, P., Parkhill, J., Flint, H. J., and Scott, K. P. (2015). 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome.* 3, 26. doi:10.1186/s40168-015-0087-4.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naïve bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* 73, 5261–5267. doi:10.1128/AEM.00062-07.

Watanabe, S., Narisawa, Y., Arase, S., Okamatsu, H., Ikenaga, T., Tajiri, Y., et al. (2003). Differences in fecal microflora between patients with atopic dermatitis and healthy control subjects. *J. Allergy. Clin. Immunol.* 111, 587–591.

Weinstock, J. V. (2012). The worm returns. *Nature.* 491, 183–185. doi:10.1038/491183a.

Williamson, L. L., McKenney, E. A., Holzknecht, Z. E., Belliveau, C., Rawls, J. F., Poulton, S., et al. (2016). Got worms? Perinatal exposure to helminths prevents persistent immune sensitization and cognitive dysfunction induced by early-life infection. *Brain Behav. Immun.* 51, 14–28. doi:10.1016/j.bbi.2015.07.006.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature*. doi:10.1038/nature11053.

Yilmaz, P., Parfrey, L. W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., et al. (2014). The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* 42, D643–D648. doi:10.1093/nar/gkt1209.

Yu, Y., Lee, C., Kim, J., and Hwang, S. (2005). Group-specific primer and probe sets to detect methanogenic communities using quantitative real-time polymerase chain reaction. *Biotechnol. Bioeng.* 89, 670–679. doi:10.1002/bit.20347.

Yu, Z.-T., Chen, C., Kling, D. E., Liu, B., McCoy, J. M., Merighi, M., et al. (2013). The principal fucosylated oligosaccharides of human milk exhibit prebiotic

In review

properties on cultured infant microbiota. *Glycobiology.* 23, 169–177. doi:10.1093/glycob/cws138.

Zhang, Y., Zhang, Q., and Xu, Z. (1997). Tissue and body fluid distribution of antibacterial agents in pregnant and lactating women. *Zhonghua. Fu. Chan. Ke. Za. Zhi.* 32, 288–292.

Zivkovic, A. M., German, J. B., Lebrilla, C. B., and Mills, D. A. (2011). Human milk glycobiome and its impact on the infant gastrointestinal microbiota. *Proc. Natl. Acad. Sci. U.S.A.* 108, 4653–4658. doi:10.1073/pnas.1000083107.

13. Footnotes

1. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
2. https://github.com/EnvGen/Tutorials/blob/master/amplicons-no_overlap.rst.
3. https://cran.r-project.org/web/packages/vegan/index.html

14. Figure legends

Figure 1. Detection of prokaryotes, fungi and archaea in infant stool during the first year of life. (**A**) Absolute quantification of for prokaryotic and (**B**) fungal DNA (ng DNA per mg of stool) and (**C**) relative quantification of archaeal read counts by quantitative real-time PCR and over the course of the first year of life. The numbers of samples per collection time point are provided at the top of the graph. For the purpose of clarity, only significant differences between subsequent time points are shown in the figure; for all significant differences between collection time points, see Supplementary File 1, Tables S3 and S4. Significant differences obtained by Wilcoxon rank sum test between consecutive time points are represented by asterisks (* when < 0.05; ** when <0.01). CSD: C-section delivery, VD: vaginal delivery. Fecal samples originating from VD infants are represented on the left side of each barplot and by green points, samples from CSD infants are represented on the right side of each barplot and by blue points.

Figure 2. **Prokaryotic and microeukaryotic microbiome compositions in infants over the first year of life.** Barplots of relative abundances of the 49 most abundant taxa per sample for (**A**) prokaryotes and (**B**) microeukaryotes for both delivery modes. All OTUs with the same taxonomy were regroupedd into the same taxa, whereas taxa that did not belong to the 49 most abundant were regrouped under 'Others'. Sequences were classified to the highest taxonomic level that could be confidently assigned. Aggregated OTUs are color-coded according to the phylum they belong to. Numbers below the barplots are representative of the different infants in the study. CSD: C-section delivery, VD: vaginal delivery. * Twins.

Figure 3. Colonization of prokaryotes and microeukaryotes. Depiction of (**A**,**E**) richness (number of OTUs), (**B**,**F**) diversity (Shannon's diversity index), (**C**,**G**) evenness (Pielou's evenness index) and (**D**,**H**) dissimilarity index reflecting the distance to the most mature sample (Soerensen's similarity index of presence/absence of taxa at each individual time point compared to the most mature microbial community structures represented by samples collected at the last individual time point) for prokaryotes and microeukaryotes, respectively. The numbers of samples per collection time point are provided at the top of the graph. For the purpose of clarity, only significant differences as assessed by Wilcoxon rank sum test between subsequent time points are shown in the figure; for all significant differences between collection time points, see Supplementary File 1, Table S3 for the prokaryotic and

In review

Table S4 for the microeukaryotic datasets. Significant differences between consecutive time points are represented by asterisks (* when *P*-value < 0.05; ** when *P*-value < 0.01). CSD: C-section delivery, VD: vaginal delivery. Fecal samples originating from VD infants are represented on the left side of each barplot and by green points, samples from CSD infants are represented on the right side of each barplot and by blue points.

Figure 4. Principal coordinate analyses of Jensen–Shannon distances for prokaryotic and microeukaryotic rRNA gene amplicon sequencing data. Depiction of (**A,C**) data from VD infants in green and (**B,D**) CSD infants in blue for prokaryotes and microeukaryotes, respectively. Sampling time points are represented by shadings, with lighter colors depicting an earlier sampling time point. Lines connect samples which originated from the same infant according the order of sampling. Samples that are the focus of the corresponding other sub-panel are shaded in grey. Cluster delineations were added manually after computing the cluster membership of each sample using the partitioning around medoids (pam) function contained in the R package 'cluster' (Maechler et al., 2015).

Figure 5. Firmicutes/Bacteroidetes ratio over time. The numbers of samples per collection time point are given at the top of the graph. Significant differences obtained by Wilcoxon rank sum test and according to delivery mode are represented by asterisks (* when *P*-value < 0.05; ** when *P*-value < 0.01). CSD: C-section delivery, VD: vaginal delivery. Fecal samples originating from VD infants are represented on the left side of each barplot and by green points, samples from CSD infants are represented on the right side of each barplot and by blue points.

Figure 6. C**olonization by Bacteroidetes phylum.** Per collection time point depiction of (**A**) richness (number of OTUs), (**B**) diversity (Shannon's diversity index) and (**C**) evenness (Pielou's evenness index) of the phylum Bacteroidetes. The numbers of samples per collection time point are provided at the top of the graph. Significant differences as assessed by Wilcoxon rank sum test and according to delivery mode are represented by asterisks (* when *P*-value < 0.05; ** when *P*-value < 0.01). CSD: C-section delivery, VD: vaginal delivery. Fecal samples originating from VD infants are represented on the left side of each barplot and by green points, samples from CSD infants are represented on the right side of each barplot and by blue points.

Figure 7. qPCR validation of 16S rRNA gene sequencing data based differences according to delivery mode. Comparison of the DESeq2-normalized 16S rRNA read numbers and relative abundances (given on log scale) measured by qPCR for two phyla and four genera that were found to be significantly different between birth modes. For each comparison the Spearman correlation coefficient (ρ) was calculated and figures next to the taxa. The numbers of samples per collection time point are given at the top of each barplot. Significant differences according to a Wilcoxon rank sum test for delivery mode are represented by asterisks (* when *P*-value < 0.05; ** when *P*-value < 0.01). CSD: C-section delivery, VD: vaginal delivery. Fecal samples originating from CSD infants are represented on the left side of each barplot and by blue points, samples from VD infants are on the right side of each barplot and by green points.

In review

15. Tables

Table 1. Primer pairs and conditions of quantitative real-time PCR.

In review

Main target (target gene)	Designation	Oligonucleotide sequence (5' -> 3')	Annealing temperature (°C)	Cycling	Reference
Fungi (18S rRNA)	Fungi2F Fungi2R	F: ATT GGA GGG CAA GTC TGG TG R: CCG ATC CCT AGT CGG CAT AG	55	60 cycles: 15 sec at 95°C, 10 sec at 55°C, 25 sec at 72°C	Einsele et al., 1997
Staphylococcus (tuf)	TStaG422-F TStag765-R	F: GGC-CGT-GTT-GAA-CGT-GGT-CAA-ATC-A R: TAT-HAC-CAT-TTC-AGT-ACC-TTC-TGG-TAA	55	45 cycles: 20 sec at 95°C, 30 sec at 55°C, 1 min at 72°C	Martineau et al., 2001
Haemophilus (P6)	HI-IV HI-V	F: ACT-TTT-GGC-GGT-TAC-TCT-GT R: TGT-GCC-TAA-TTT-ACC-AGC-AT	55		van Ketel et al., 1990
Universal archaea (16S rRNA)	ARC787F ARC1059R	F: ATT-AGA-TAC-CCS-BGT-AGT-CC R: GCC-ATG-CAC-CWC-CTC-T	60	45 cycles: 15 sec at 95°C, 30 sec at 60°C, 1 min at 72°C	Yu et al., 2005
Lactobacillus (16S rRNA)	Lac774F Lac989R	F: GCG-GTG-AAA-TTC-CAA-ACG R: GGG-ACC-TTA-ACT-GGT-GAT	60		Hermann-Bank et al., 2013
Streptococcus (16S rRNA)	Strep488F Strep824R	F: CTW-ACC-AGA-AAG-GGA-CGG-CT R: AAG-GRY-CYA-ACA-CCT-AGC	60		Hermann-Bank et al., 2013
Firmicutes (16S rRNA)	Lgc353 Eub518	F: GCA-GTA-GGG-AAT-CTT-CCG R: ATT-ACC-GCG-GCT-GCT-GG	60		Fierer et al, 2005
Bacteroidetes (16S rRNA)	798cfbF cfb967R	F: CRA-ACA-GGA-TTA-GAT-ACC-CT R: GGT-AAG-GTT-CCT-CGC-GTA-T	61	45 cycles: 15 sec at 95°C, 20 sec at 61°C, 30 sec at 72°C	Bacchetti De Gregoris et al., 2011
Universal prokaryotes (16S rRNA)	926F 1062R	F: AAA-CTC-AAA-KGA-ATT-GAC-GG R: CTC-ACR-RCA-CGA-GCT-GAC	61		Bacchetti De Gregoris et al., 2011

Table 2. Neonatal and maternal characteristics (n=15). Study groups are defined according to delivery mode (VD: n=8; CSD: n=7). CSD: C-section delivery, VD: vaginal delivery.

	Total cohort *n*=15	VD *n*=8	CSD *n*=7[1]
Infant characteristics			
Female gender	7 (46.7%)	5 (62.5%)	2 (28.6%)
Gestational age at delivery (weeks)	38.7 ± 1.8	39 ± 1.5	38.3 ± 2.1
Birth weight (g)	3273 ± 416	3311 ± 543	3230 ± 236
Maternal characteristics			
Positive group B *Streptococcus* screening	3 (21.4%)	3 (37.5%)	0
Age	33.6 ± 4.6	32.5 ± 4.4	35 ± 4.8
Postnatal body mass index	24 ± 4.3	21.8 ± 2.7	26.8 ± 4.6
Ethnicity			
Caucasian	12 (85.7%)	7 (87.5%)	5 (83.3 %)
African	2 (14.3%)	1 (12.5%)	1 (16.7%)
Perinatal antibiotic intake[2]	11 (78.6%)	6 (75%)	5 (83.3 %)
Penicillin[3]	6 (42.9%)	6 (75%)	0
Cephalosporin	4 (28.6%)	0	4 (66.7%)
Clindamycin	1 (7.1%)	0	1 (16.7%)
Probiotic use during pregnancy	2 (14.3%)	1 (12.5%)	1 (16.7%)

1. 2 C-section infants are twins.
2. Considering all antibiotics administered to the mother 12 hours prior and after the delivery.
3. As ampicillin belongs to the penicillin group, ampicillin and penicillin intake were both categorized as ‘penicillin’.

In review

Figure 1.JPEG

In review

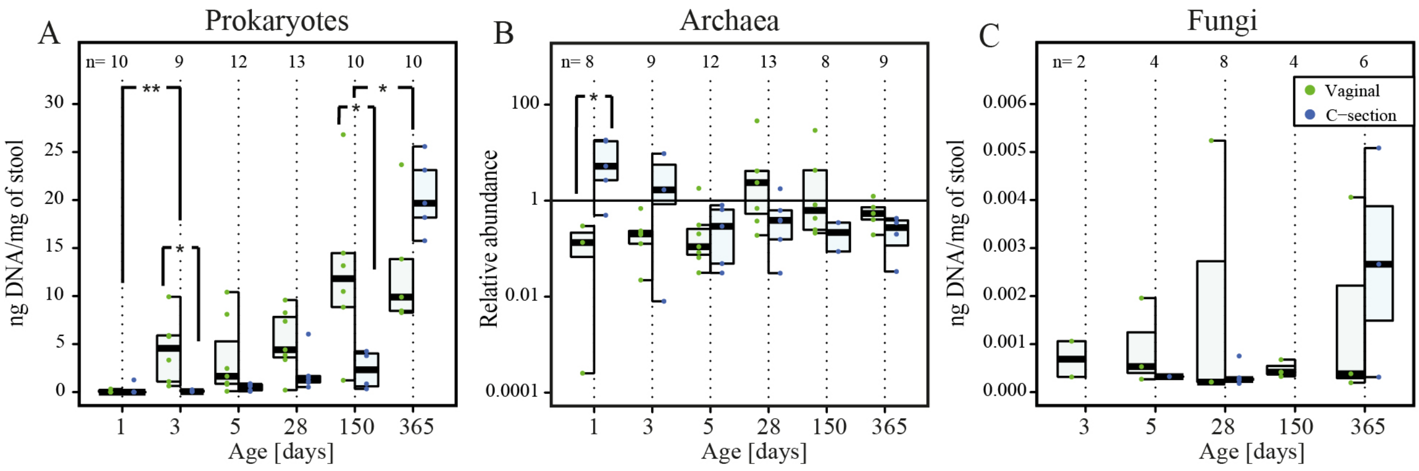
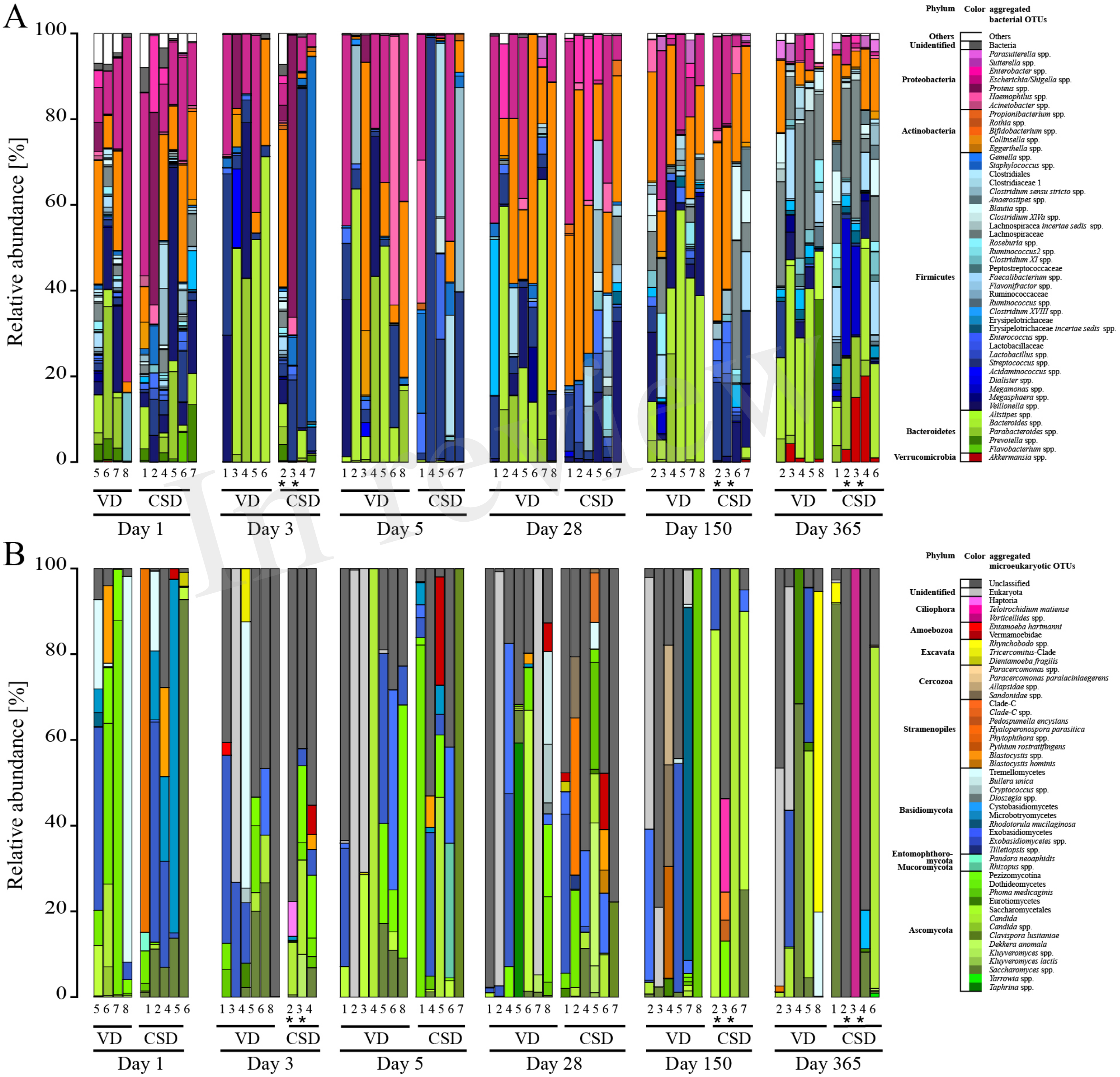
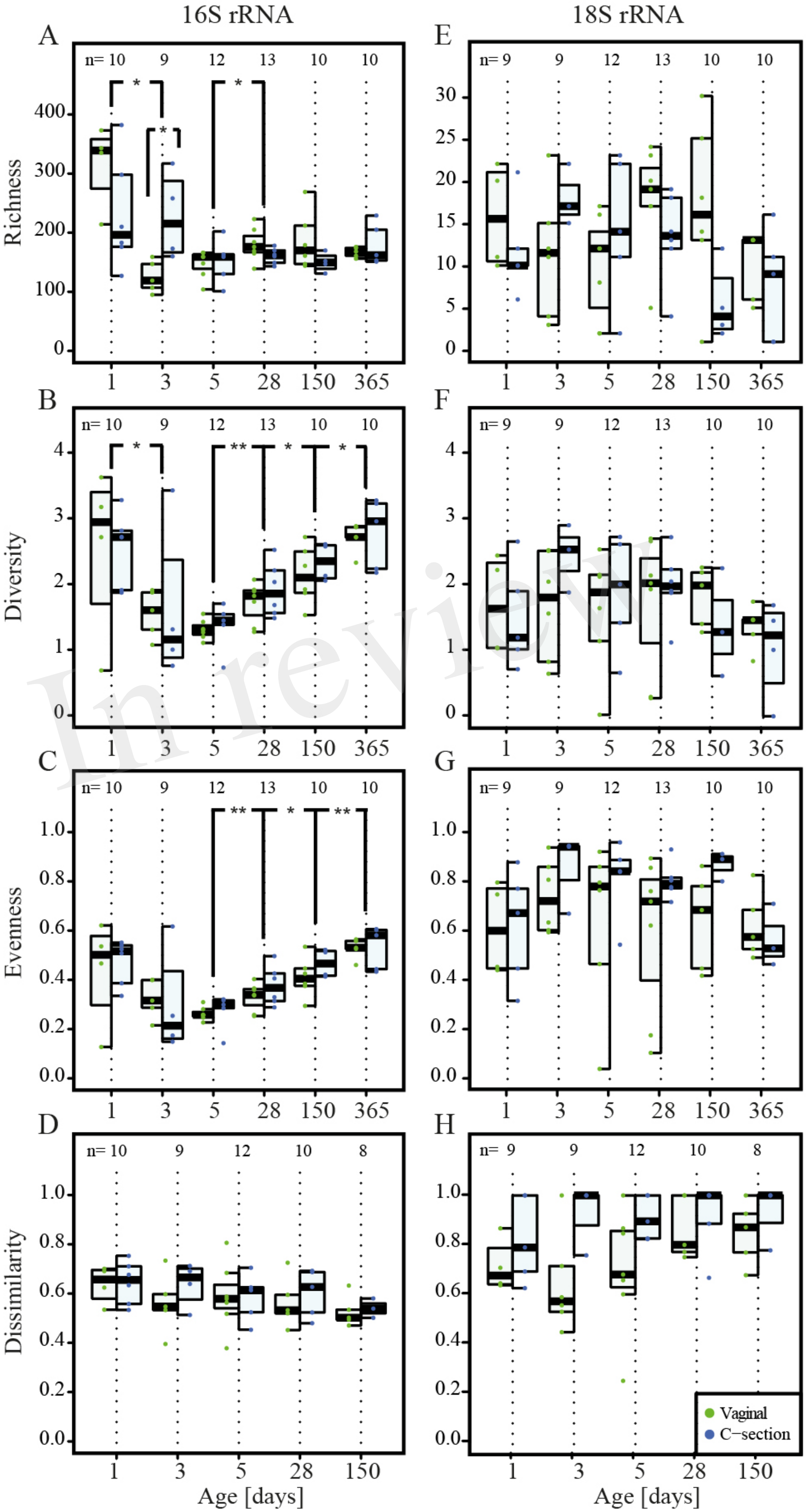


Figure 2.JPEG



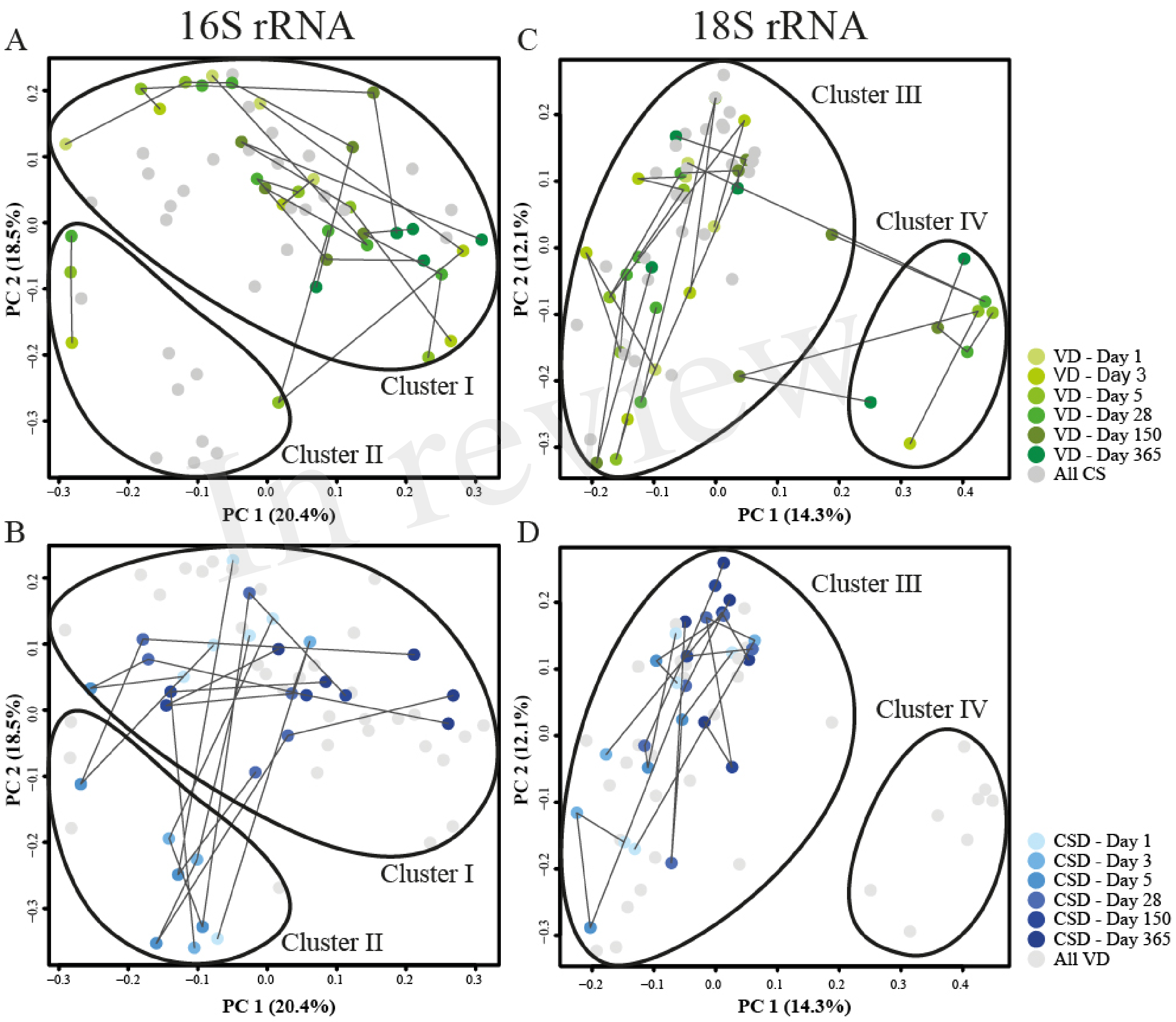
In review

Figure 3.JPEG



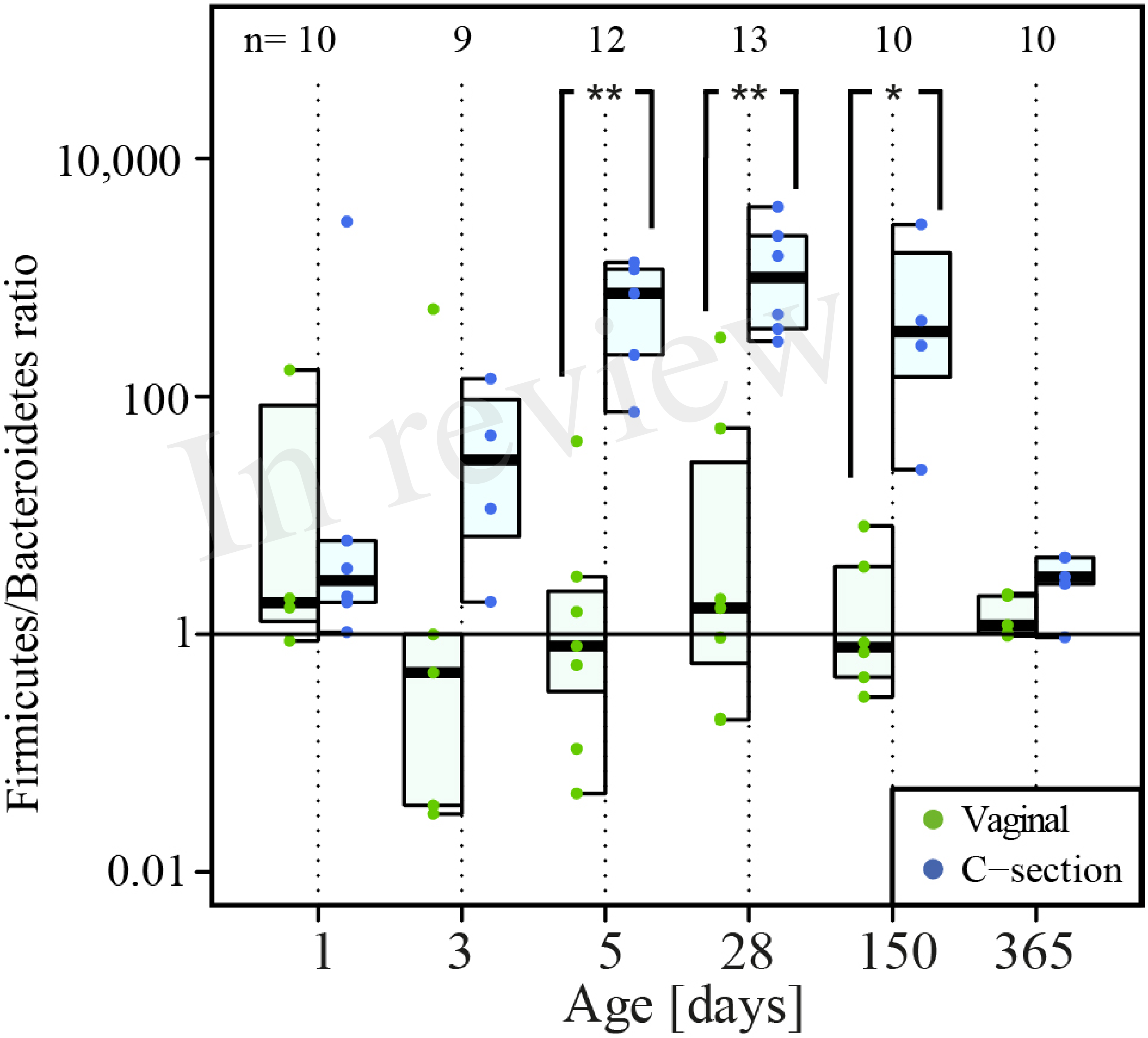
In review

Figure 4.JPEG



In review

Figure 5.JPEG



In review

Figure 6.JPEG

In review

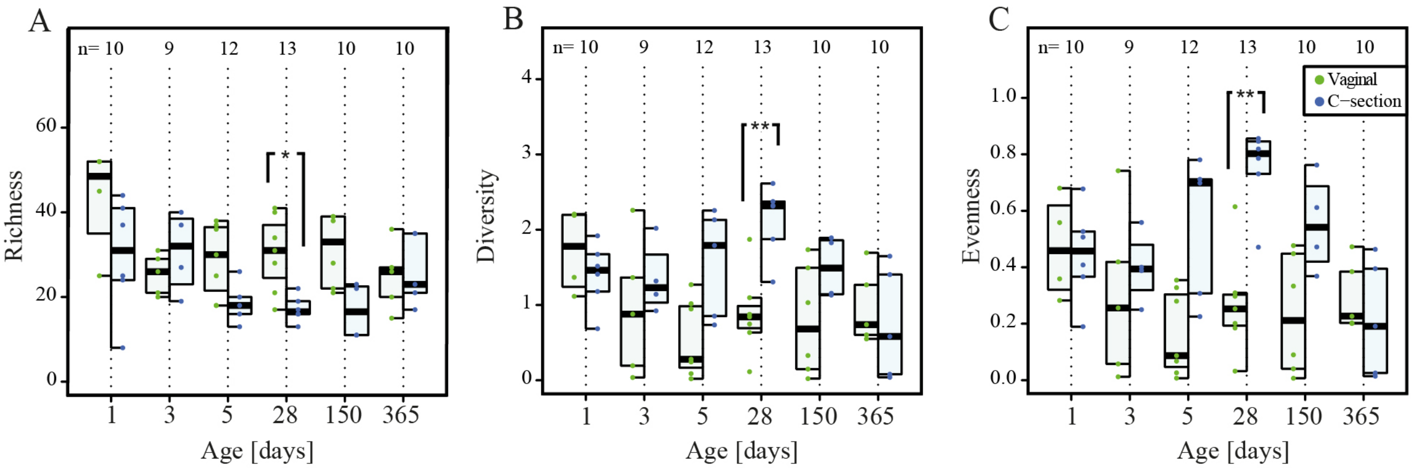
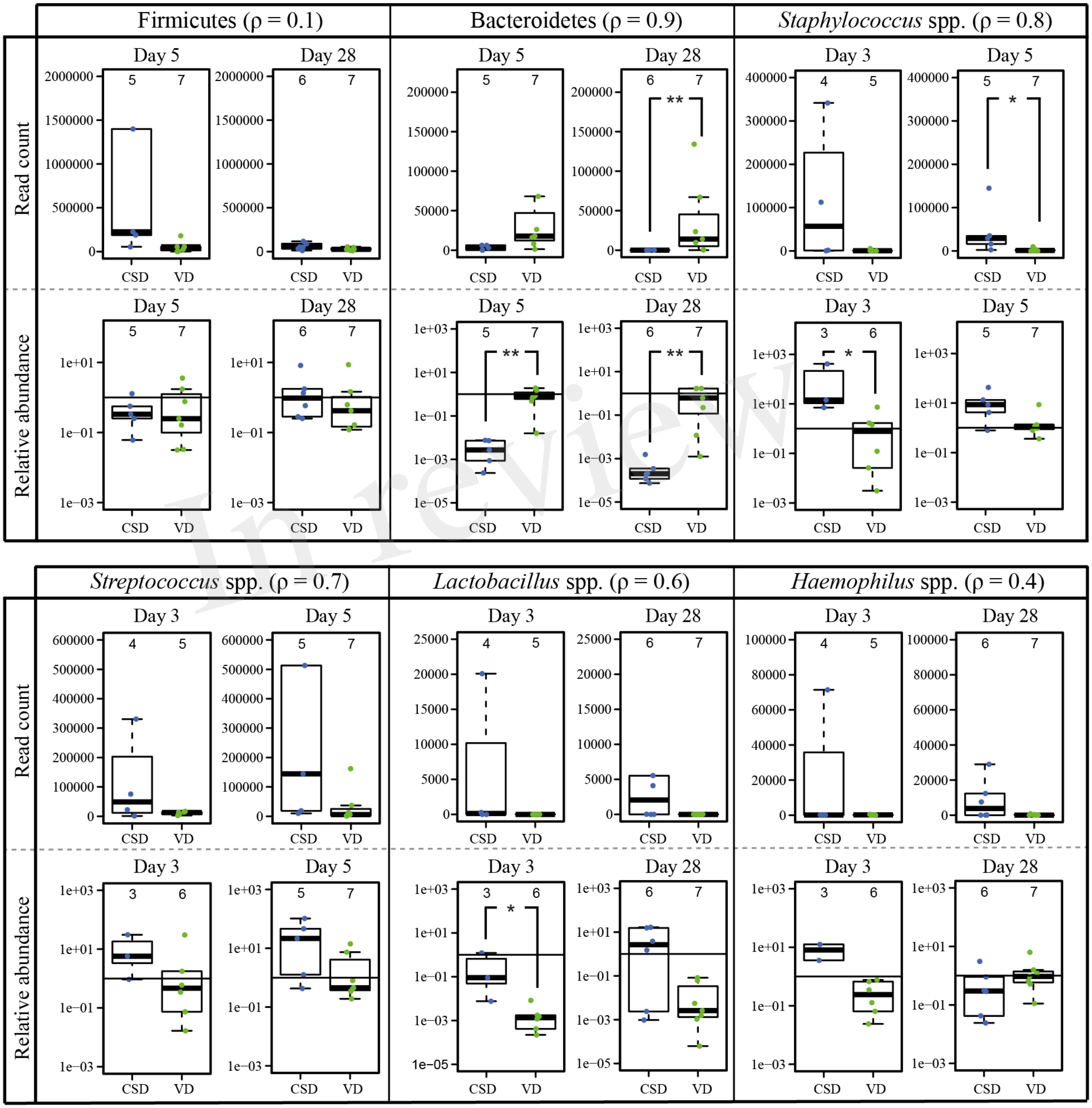


Figure 7.JPEG



In review

A.6 Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic stem cell transplantation.

Anne Kaysen, Anna Heintz-Buschart, Emilie E. L. Muller, **Shaman Narayanasamy**, Linda Wampach,
Cédric C. Laczny, Katharina Franke, Jörg Bittenbring, Jochen G. Schneider, Paul Wilmes
Submitted
Journal of Experimental & Clinical Cancer Research

Contributions of author include:

- Data analysis
- Revision of manuscript

BMC Medicine

Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic stem cell transplantation

--Manuscript Draft--

Manuscript Number:	BMED-D-16-01531	
Full Title:	Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic stem cell transplantation	
Article Type:	Research article	
Section/Category:	Oncology	
Funding Information:	Université du Luxembourg (ImMicroDyn1)	Prof. Jochen G. Schneider
	Fonds National de la Recherche Luxembourg (A09/03)	Prof. Paul Wilmes
	Fonds National de la Recherche Luxembourg (INTER/JPND/12/01)	Prof. Paul Wilmes
	Fonds National de la Recherche Luxembourg (PoC/13/02)	Prof. Paul Wilmes
	Fonds National de la Recherche Luxembourg (PHD/4964712)	Dr. Cédric C. Laczny
	Fonds National de la Recherche Luxembourg (AFR PHD-2014-1/7934898)	Mr. Shaman Narayanasamy
	Fonds National de la Recherche Luxembourg (C15/SR/10404839)	Dr. Emilie E. L. Muller
	Fonds National de la Recherche Luxembourg (PHD-2013-5824125)	Ms. Linda Wampach

Abstract:

Background
In patients undergoing allogeneic hematopoietic stem cell transplantation (allo-HSCT), treatment-induced changes to the gastrointestinal tract (GIT) microbiome have been linked to adverse treatment outcomes, most notably graft-versus-host disease (GvHD). However, it is not known whether this relationship is directly causal. Here, we performed an integrated meta-omic analysis to gain deeper insight into GIT microbiome changes during allo-HSCT and accompanying treatments.
Methods
We used 16S and 18S rRNA gene amplicon sequencing to resolve archaea, bacteria and eukaryotes in the GIT microbiomes of 16 patients undergoing allo-HSCT for treatment of hematologic malignancies. To obtain a more detailed assessment of microbiome changes and their potential relation to acute GvHD (aGvHD), an integrated analysis of metagenomic and metatranscriptomic data was performed on samples collected from one patient before and after treatment for acute myeloid leukemia. This patient developed severe aGvHD, which led to death nine months after the transplantation.
Results
This study reveals a major shift in the GIT microbiome after allo-HSCT, including a marked reduction in bacterial diversity but limited changes among eukaryotes and archaea following the appropriate treatment protocols and accompanying interventions. Data from pre- and post-treatment samples of the patient who developed severe aGvHD revealed a drastically decreased bacterial diversity. Furthermore, the post-treatment sample showed a higher overall number and higher expression levels for antibiotic resistance genes (ARGs), discovered as a long-term effect of the treatment on the microbial community. An organism causing a paravertebral abscess was shown to be linked to the GIT dysbiosis, suggesting loss of intestinal barrier integrity.
Conclusions
The apparent selection for bacteria expressing ARGs suggests that prophylactic

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

	antibiotic administration may adversely affect overall treatment outcome. Detailed analyses including information about the selection of pathogenic bacteria expressing ARGs may help to support clinicians in tailoring the procedural therapy protocols in a personalized fashion to improve overall outcome in the future.
Corresponding Author:	Jochen Schneider Universite du Luxembourg LUXEMBOURG
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universite du Luxembourg
Corresponding Author's Secondary Institution:	
First Author:	Anne Kaysen
First Author Secondary Information:	
Order of Authors:	Anne Kaysen
	Anna Heintz-Buschart
	Emilie E. L. Muller
	Shaman Narayanasamy
	Linda Wampach
	Cédric C. Laczny
	Norbert Graf
	Arne Simon
	Katharina Franke
	Joerg Thomas Bittenbring
	Paul Wilmes
	Jochen G. Schneider
Order of Authors Secondary Information:	
Suggested Reviewers:	Christof von Kalle christof.kalle@nct-heidelberg.de
	Jesús Bañales jesus.banales@biodonostia.org
Opposed Reviewers:	

Manuscript

Click here to view linked References

Click here to download Manuscript Manuscript_ImMicroDyn1_20161118.docx

Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic stem cell transplantation

Anne Kaysen[1], Anna Heintz-Buschart[1], Emilie E. L. Muller[1,°], Shaman Narayanasamy[1], Linda Wampach[1], Cédric C. Laczny[1,°], Norbert Graf[2], Arne Simon[2], Katharina Franke[3], Jörg Bittenbring[3], Paul Wilmes[1], Jochen G. Schneider[1,4]*

[1] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 6 avenue du Swing, L-4367 Belvaux, Luxembourg

° Current affiliations: EELM: Department of Microbiology, Genomics and the Environment, UMR 7156 UNISTRA – CNRS, Université de Strasbourg, Strasbourg, France. CCL: Chair for Clinical Bioinformatics, Saarland University, Building E2.1, 66123 Saarbrücken, Germany

[2] Saarland University Medical Center, Klinik für Pädiatrische Onkologie und Hämatologie, Geb. 9, Kirrberger Straße, 66421 Homburg, Germany

[3] Saarland University Medical Center, Klinik für Innere Medizin I, Geb. 41.1, Kirrberger Straße, 66421 Homburg, Germany

[4] Saarland University Medical Center, Klinik für Innere Medizin II, Geb. 77, Kirrberger Straße, 66421 Homburg, Germany

AK: anne.kaysen@uni.lu, AHB: anna.buschart@uni.lu, EELM: emilie.muller@unistra.fr, SN: shaman.narayanasamy@uni.lu, LW: linda.wampach@uni.lu, CCL: cedric.laczny@ccb.uni-saarland.de, NG: norbert.graf@uniklinikum-saarland.de, AS: arne.simon@uniklinikum-saarland.de, FK: franke-katharina@gmx.de, JB: joerg.thomas.bittenbring@uniklinikum-saarland.de, PW: paul.wilmes@uni.lu, JS: josc012@outlook.com

* Correspondence: josc012@outlook.com, paul.wilmes@uni.lu

Abstract

Background

In patients undergoing allogeneic hematopoietic stem cell transplantation (allo-HSCT), treatment-induced changes to the gastrointestinal tract (GIT) microbiome have been linked to adverse treatment outcomes, most notably graft-versus-host disease (GvHD). However, it is not known whether this relationship is directly causal. Here, we performed an integrated meta-omic analysis to gain deeper insight into GIT microbiome changes during allo-HSCT and accompanying treatments.

Methods

We used 16S and 18S rRNA gene amplicon sequencing to resolve archaea, bacteria and eukaryotes in the GIT microbiomes of 16 patients undergoing allo-HSCT for treatment of hematologic malignancies. To obtain a more detailed assessment of microbiome changes and their potential relation to acute GvHD (aGvHD), an integrated analysis of metagenomic and metatranscriptomic data was performed on samples collected from one patient before and after treatment for acute myeloid leukemia. This patient developed severe aGvHD, which led to death nine months after the transplantation.

Results

This study reveals a major shift in the GIT microbiome after allo-HSCT, including a marked reduction in bacterial diversity but limited changes among eukaryotes and archaea following the appropriate treatment protocols and accompanying interventions. Data from pre- and post-treatment samples of the patient who developed severe aGvHD revealed a drastically decreased bacterial diversity. Furthermore, the post-treatment sample showed a higher overall number and higher expression levels for antibiotic resistance genes (ARGs), discovered as a long-term effect of the treatment on the microbial community. An organism causing a paravertebral abscess was shown to be linked to the GIT dysbiosis, suggesting loss of intestinal barrier integrity.

Conclusions

The apparent selection for bacteria expressing ARGs suggests that prophylactic antibiotic administration may adversely affect overall treatment outcome. Detailed analyses including information about the selection of pathogenic bacteria expressing ARGs may help to support clinicians in tailoring the procedural therapy protocols in a personalized fashion to improve overall outcome in the future.

Keywords

Graft-versus-host disease, stem cell transplantation, dysbiosis, antibiotic pressure, antibiotic resistance genes, metagenomics, metatranscriptomics, amplicon sequencing

Background

Humans live in a close relationship with microorganisms that are referred to as the “microbiome”, comprising bacteria, archaea and eukaryotes. The most densely populated human body habitat is the gastrointestinal tract (GIT), which is estimated to contain 500 – 1000 different microbial species [1]. The GIT microbiome plays a myriad of important roles in human physiology, including for example in the digestion of food, the synthesis of vitamins, the production of short-chain fatty acids and the prevention of colonization by pathogens through exclusion [2]. It is generally accepted that, within a healthy human GIT, a homeostatic state exists among the different microorganisms which is tightly regulated by the host's immune system [3–5]. However, perturbations, such as the intake of antibiotics, infections or immunosuppression, can lead to a disruption of this balanced state, typically referred to as "dysbiosis" [3, 6]. In a dysbiotic state, pathogens can overgrow the community [6]. Furthermore, reduced intestinal barrier function can facilitate translocation of microorganisms and microbial products from the GIT lumen to mesenteric lymph nodes and/or the bloodstream [7], putting the host at risk for local infections and sepsis [6, 8].

Allogeneic hematopoietic stem cell transplantation (allo-HSCT) represents an effective treatment for several hematologic malignancies. It is preceded by an intense conditioning regime, consisting of either total body immune ablative irradiation or high doses of chemotherapy, to facilitate engraftment of transplanted stem cells. Allo-HSCT is known to greatly impact stability and integrity of the GIT microbiome [9]. A substantial loss in bacterial diversity and the dominance of single bacterial taxa have been observed in patients undergoing allo-HSCT [9].

Supportive care of patients receiving allo-HSCT includes prophylactic broad-spectrum antibiotic treatment [10], an intervention that also influences the GIT microbiome by selection for potential pathogens carrying antibiotic resistance genes (ARGs) [11] as well as driving transfer of ARGs among commensal bacteria, including many opportunistic pathogens [12]. In addition, loss of the normal bacterial GIT community following antibiotic treatment can facilitate expansion of yeasts including invasive *Candida albicans* infections with potentially fatal consequences [13, 14].

The intensive conditioning treatment for allo-HSCT may lead to mucositis along the GIT, which culminates in the formation of painful ulcers, dysphagia and diarrhea [15]. The most significant complication of allo-HSCT is acute graft-versus-host disease (aGvHD) which affects 35 % - 50 % of patients and is a major cause of mortality [16]. GvHD, a systemic, inflammatory disease, is provoked by a complex anti-allogeneic immune response, which primarily affects the skin, liver and GIT [17]. Glucksberg et al. [18] divided each organ involvement into four stages from mild to severe. These are integrated into an overall grade of GvHD, where I-II are considered as mild and III-IV are considered as severe. Usually, intestinal GvHD dominates the clinical picture in severe aGvHD, which typically occurs within 100 days after allo-HSCT and is initiated by alloreactive donor T cells that recognize antigens on host cells [19].

It has been suggested that the GIT microbiome might be implicated in the development or exaggeration of aGvHD, as the damaged GIT epithelial barrier in patients undergoing allo-HSCT allows translocation of microorganisms or pathogen-associated-molecular patterns (PAMPs) [20]. These PAMPs can activate antigen-presenting cells and thereby lead to alloactivation and proliferation of donor T cells which trigger aGvHD [20]. Antibiotic treatment has been shown to have ambiguous effects on treatment outcome. On the one hand, a low bacterial diversity at engraftment, possibly caused by a preceding combination of chemotherapy, total body irradiation and broad spectrum antibiotics has been linked to a worse outcome [21]. On the other hand, GIT decontamination using antimicrobials has been observed to lower the rate of aGvHD [22, 23].

Previous studies have investigated changes in the bacterial community structures of the GIT microbiome directly after allo-HSCT or conditioning treatment [21, 24–26]. However, it is not yet known how GIT microbial communities including archaea and eukaryotes evolve over longer periods of time and what effects the disruption of the microbiome, for example through the administration of antibiotic regimens, has on the human host with respect to aGvHD and overall treatment outcome.

Recent advances in high-throughput next-generation sequencing allow for a detailed analysis of the GIT microbiome in the context of allo-HSCT and treatment outcome. Here, a meta-omic approach was used to provide an exhaustive view of the changes which occur in the GIT microbial community of patients with hematologic malignancies undergoing allo-HSCT treatment. We expand upon previous studies by analyzing changes not only in the bacterial populations, but also among archaea and eukaryotes, thereby covering all three domains of life. Additionally, we present a detailed analysis of metagenomic (MG) and metatranscriptomic (MT) data from one patient with a fatal treatment outcome, including identification of ARGs, corresponding expression levels and genetic variation in dominant bacterial populations. This study serves as a proof of concept for future meta-omic studies of the GIT microbiome in the context of allo-HSCT treatment and other intensive medical treatments.

Methods

Study participants and fecal sample collection

After provision of written informed consent, 16 patients undergoing allo-HSCT were enrolled in the study.

For microbial diversity and richness analyses, patients were included only if fecal samples were obtained from at least two of the following time points: i) up to eight days before allo-HSCT (designated time point 1 (TP) 1), ii) directly after allo-HSCT (up to four days after allo-HSCT, designated TP2) and/or iii) around the time of engraftment between day 20 and day 33 after allo-HSCT (designated TP3). One additional patient was selected for a detailed analysis of the effects of the treatment over an extended period of time. From this patient, samples were collected 13 days before allo-HSCT, as well as 75 and 119 days after allo-HSCT. Fecal samples were immediately flash-frozen on-site and preserved at -80 °C to ensure integrity of the biomolecules of interest.

Extraction of biomolecules from fecal samples

DNA and RNA were extracted from unthawed subsamples of 150 mg, after pre-treatment of the weighed subsamples with 1.5 ml RNAlater-ICE (LifeTechnologies) overnight at -20 °C. The biomolecules were extracted from the mixture as described previously, using the AllPrep DNA/RNA/Protein kit (Qiagen) [27, 28]. To increase the overall yield, DNA fractions were supplemented with DNA extracted from 200 mg subsamples using the PowerSoil DNA isolation kit (MO BIO).

The quality and quantity of the DNA extracts were verified using 1 % agarose gel electrophoresis and NanoDrop 2000c spectrophotometer (Thermo Fisher Scientific), while RNA extracts were verified using Agilent 2100 Bioanalyzer

(Agilent Technologies). Only fractions with RNA integrity number (RIN, Agilent Technologies) > 7 were sequenced. Extracted biomolecules were stored at -80 °C until sequencing.

16S and 18S rRNA gene amplicon sequencing

Amplification and paired-end sequencing of extracted and purified DNA was performed on an Illumina MiSeq platform at the Groupe Interdisciplinaire de Génoprotéomique Appliquée (GIGA, Belgium). The V4 region of the 16S rRNA gene, which allows resolution of bacteria and archaea, was amplified and sequenced using the primers 515F_GTGBCAGCMGCCGCGGTAA and 805R_GACTACHVGGGTATCTAATCC [29, 30] with paired-end reads of 300 nt each. The V4 region of the 18S rRNA gene was amplified and sequenced using the primers 574*F and 1132R (574*F_CGGTAAYTCCAGCTCYV 1132r_CCGTCAATTHCTTYAART; [31]) to resolve the eukaryotic community structure.

16S and 18S rRNA gene amplicon sequencing and data analysis

16S rRNA gene sequencing reads were processed using the LotuS pipeline (version 1.34) [32] with default parameters. Processed reads were clustered into operational taxonomic units (OTUs), designating taxa with similar amplicon sequences at 97 % identity level. To process the 18S rRNA gene sequencing reads, a workflow specifically designed to process reads that are not overlapping was used [33].

Statistical analyses and plots were generated in R (version 3.2.1) [34]. Microbial alpha-diversity and richness were determined at the OTU level, by calculating the Shannon diversity index and the Chao1 index after rarefaction, using the vegan package [35]. Plots were generated using the R base graphics or the ggplot2 package [36].

Differential analysis of taxa based on 16S rRNA gene sequencing data was performed using the DESeq2 package [37] and significant differences on taxonomic levels were determined using the Wald test, after multiple-testing adjustment.

Further information on processing and analyses of rRNA gene amplicon sequencing data can be found in Additional file 1.

Metagenomic and metatranscriptomic sequencing, processing and assembly

MG and MT sequencing of the extracted DNA and RNA fractions was conducted by GATC Biotech AG, Konstanz, Germany. Ribosomal RNA (rRNA) was depleted from the RNA fractions using the Ribo-Zero Gold rRNA Removal kit (Epidemiology, Illumina) and a strand-specific cDNA library was prepared according to standard protocols, optimized by GATC. Libraries representing both nucleic acid fractions were sequenced using a 100 bp paired-end approach on an Illumina HiSeq 2500 using HiSeq V3 reagents.

MG and MT datasets were processed using a newly in-house developed workflow, the Integrated Meta-omics Pipeline (IMP) version 1.1 [38]. This pipeline includes a co-assembly of MG and MT reads. Further information on this pipeline and on calculations used in this work can be found in Additional file 1.

Population-level binning of contigs from the co-assembly

To analyze and compare the population-level structure of the microbial communities based on the assembled genomic information, contiguous sequences (contigs) were binned into (partial) population-level genomes. Using VizBin [39, 40], 2D embeddings based on BH-SNE of the contigs of at least 1,000 nt were produced, as part of IMP. In these embeddings, contigs with similar genomic signatures are closer together, hence, individual clusters of contigs represent individual microbial populations [41]. Population-level clusters were selected following the method described in [42]. Resulting bins are referred to as “population-level genomes” in the following. Details on the inference of population sizes can be found in Additional file 1.

Taxonomic affiliation of reconstructed population-level genomes

Taxonomic affiliation of population-level genomes was determined using complementary methods. Contigs forming the population-level genomes were first aligned to the NCBI nucleotide collection (nr/nt) database using the BLAST webservice [43]. Parameters were left at default (using program megablast), and the output was analyzed using the MEtaGenome ANalyzer (MEGAN version 5.10.5) [44]. Whenever the *rpoB* gene could be recovered within a population-level genome, the closest neighbour (in terms of sequence identity) was determined in the nucleotide collection (nr/nt) database using the MOLE-BLAST webservice [45]. Additionally, AMPHORA2 [46] was used to identify the taxonomic affiliation of up to 31 bacterial or 104 archaeal phylogenetic marker genes.

Reassembly

Population-level genomes were reassembled using all MG and MT reads mapping to the contigs of the population-level genomes with the same taxonomic assignment. Reassembly of all recruited reads was carried out using SPAdes [47] (version 3.5.0) using standard parameters. MG and MT reads were subsequently mapped to the contigs forming this reassembly to determine expression levels and.

Sequence comparison of population-level genomes

The average nucleotide identity (ANI) calculator [48] was used with standard settings to compare the reassembly from population-level genomes to publicly available reference genomes. A gene-wise protein sequence comparison of different population-level genomes was performed using the RAST server [49] using standard parameters.

Detection of antibiotic resistance genes

Antibiotic resistance genes (ARGs) within a community or population were searched against Resfams version 1.2 [50] using HMMer version 3.1b2 [51]. We used the core version of the Resfams database, which includes 119 protein families. In accordance with the HMMer user manual, only identified genes with a bitscore higher than the binary logarithm of the total number of genes (of the community or population) were retained.

Variant identification

Variants were identified in population-level reassembled genomes using

SAMtools mpileup [52] with default settings, which include the calling of single nucleotide variants (SNVs) as well as the identification of small insertions/deletions (indels). The output of SAMtools mpileup was filtered using a conservative heuristic established in [53], which takes into account the ratio of the frequencies of both bases and the depth of coverage at the corresponding nucleotide position, in order to reduce the effect of sequencing errors.

Extraction, sequencing and analysis of bacterial DNA from a blood culture

DNA was extracted from a blood culture of an organism identified as a multidrug-resistant *E. coli* and sequenced on an Illumina MiSeq, 300 bp paired-end at GIGA. The genome was assembled with SPAdes [47]. Using PanPhlAn [54] and the provided database including 118 *E. coli* reference strains, their relation was assessed based on their gene set. While the PanPhlAn database includes 31734 genes, only genes present in 10 or more genomes were considered, resulting in 7845 genes for comparison.

Results

Patient characteristics and treatment

Anthropometric and clinical information of the ten female and six male patients included in the study are provided in Table 1. They were between 30 and 67 years old (median 55). Five patients with relapsed or refractory lymphoma received FluBuCy (fludarabine, busulfan, cyclophosphamide) as conditioning treatment, six acute myeloid leukemia (AML) patients received

BuCy (busulfan, cyclophosphamide), one myeloma and one comorbid AML patient received Treo/Flu (treosulfan, fludarabine), one comorbid AML patient received FluBu (fludarabine, busulfan) and two refractory AML patients received FLAMSA-Bu (fludarabine, amsacrine, busulfan) conditioning treatment. Grafts from eight full match unrelated, three mismatch unrelated and five sibling donors were used. 1.5 years after allo-HSCT, ten patients were still alive, while six patients had deceased. Twelve patients developed aGvHD and were treated with steroids (0.5 – 2 mg/kg/day). Three of them progressed to at least grade III aGvHD.

As a prophylactic treatment, patients received a fluoroquinolone antibiotic during leukopenia. At occurrence of fever, patients were treated with piperacillin-tazobactam, followed by meropenem and subsequently vancomycin, if necessary. In case of suspected fungal infection, patients also received antifungal treatment with liposomal amphotericin B or caspofungin (Table 1).

Table 1 Anthropometric and clinical information of the study cohort

Patient	Sex	Age	Underlying disease[a]	Donor relationship and HLA[b]	Conditioning regimen[c]	Antimicrobials[d]	GvHD[e, f]	Outcome 1.5 years after allo-HSCT
A01	m	43	lymphoma	MRD	FluBuCy	F, M, P-T, V	Skin I°	alive
A03	m	56	lymphoma	MRD	FluBuCy	AF, F, M, P-T, other	-	deceased d66, relapse
A04	f	43	AML	MUD	BuCy	AF, F, M, V	Skin I°	alive
A05	m	49	lymphoma	MMUD	FluBuCy	AF, F, M, P-T, V	Skin II°	deceased d275, pneumonia
A06	m	52	AML	MRD	BuCy	AF, F, M, P-T, V, other	-	alive
A07	f	63	AML	MMUD	FLAMSA-Bu	AF, F, M, P-T, V, other	**Skin II°, GIT III°**	deceased d268, GvHD
A08	f	50	AML	MUD	BuCy	AF, F, M, P-T, V	Skin I°	alive
A09	m	30	lymphoma	MUD	FluBuCy	F, M, P-T	-	deceased d212, pneumonia
A10	m	54	AML	MRD	BuCy	F, M, P-T	Skin I°, GIT II°	alive
A12	m	57	lymphoma	MUD	FluBuCy	F, M, P-T, V, other	Skin III°	alive
A13	m	57	AML	MRD	BuCy	AF, F, M, V	Skin I°, lung II°	alive
A17	m	66	AML	MUD	BuCy	F, M, V	Skin II°	alive
A18	f	67	AML	MUD	FluBu	F, M, P-T, V, other	**Skin III°, GIT III°**	deceased d184, GvHD
A19	f	58	myeloma	MUD	Treo/Flu	F, M, P-T	-	deceased d39, relapse
A20	m	51	AML	MMUD	FLAMSA-Bu	AF, F, M, P-T, V, other	**Skin II°, GIT II°**	alive
A21	f	64	AML	MUD	Treo/Flu	AF, M, P-T, V, other	Skin II°	alive

[a]: AML: acute myeloid leukemia

[b]: MRD: matched related, MUD: matched unrelated, MMUD: mismatched unrelated

[c]: Bu: busulfan, Cy: cyclophosphamide, Flu: fludarabine, FLAMSA: fludarabine, amsacrine, Treo: treosulfan

[d]: AF: antifungal, F: fluoroquinolone, M: meropenem; P-T: piperacillin-tazobactam, V: vancomycin

[e]: Organ involvement, stages according to [18]

[f]: Bold: aGvHD with summed stages ≥ 4 considered as severe aGvHD

Changes within the GIT microbiome of patients undergoing allo-HSCT

We assessed the diversity and richness in the microbial community separately for the prokaryotic (bacteria and archaea; 16S rRNA gene sequencing) and eukaryotic (18S rRNA gene sequencing) community structures. The prokaryotic communities showed a drastic and statistically significant decrease in diversity from TP1 to TP3 (Fig. 1A). Similar to the observed changes in terms of diversity, prokaryotic richness (Fig. 1B) decreased over the course of the study, with a significant decrease between TP1 and TP3 over all samples. Differences in average relative abundance on different taxonomic levels were tested. On the genus level, average decreases of 119-, 47- and 44-fold in the relative abundances of the genera *Roseburia*, *Bifidobacterium* and *Blautia* (Fig. 1C) were observed from TP1 to TP3. On the order level, a decrease in Bacteroidales relative abundance was observed in parallel with an increase in Bacillales (Fig. 1D). Only one OTU belonging to the domain archaea could be identified, the methanogen *Methanobrevibacter smithii* [55]. It was detected in 13 out of the total 35 samples (and 10 out of 15 patients) with a total of 914 reads. A complete list of prokaryotic OTUs and the number of reads in each sample are listed in Additional file 2: Table S1.

The analysis of the eukaryotic community did not reveal statistically significant differences for Shannon diversity (Fig. 1E) or Chao1 richness (Fig. 1F) between the different TPs. Both indices stayed relatively constant from TP1 to TP2 and even increased slightly at TP3 with no apparent statistically significant difference being observed for the 8 patients who underwent antifungal treatments. Overall, per sample, around 99 % of classified eukaryotic OTUs belonged to the fungal domain with the majority representing the genera *Saccharomyces, Candida* and *Kluyveromyces*. Only few different and lowly abundant protists could be identified, including a *Vorticella* sp., *Prorodon teres*, and a *Phytophthora* sp. A complete list of eukaryotic OTUs and the number of reads in each sample are listed in Additional file 2: Table S2. We observed a lower prokaryotic diversity at TP of engraftment in patients who deceased (within 1.5 years after allo-HSCT), than in those who survived (Fig. 1G).

In summary, we found a general decrease in bacterial diversity after allo-HSCT while the eukaryotic community stayed relatively stable throughout the treatment.

To further explore the effects of treatment on the structure and function of the GIT microbiome, we applied a detailed meta-omic approach on one patient.

Patient A07 - description of treatment and status of the patient

We chose to focus on patient A07, a patient who displayed a marked reduction in bacterial diversity with high relative abundances of opportunistic pathogens (Fig. 2A and 2B) and a fatal treatment outcome. This 63 year old patient had acute myeloid leukemia with deletion 7q. The patient was refractory to conventional induction (3+7) and salvage chemotherapy with high-dose cytarabine and mitoxantrone and therefore needed further treatment. FLAMSA-Bu [56], a modified sequential conditioning regime for refractory acute myeloid leukemia was used (Fludarabine 30 mg/m² day -11 to -8, Cytarabine 2000 mg/m² day -11 to -8, Amsacrine 100 mg/m² day -11 to -8 and Busulfan 3,2 mg/kg day -7 to -4) for remission induction and transplantation. She received peripheral blood stem cells from a single HLA-C antigen mismatched unrelated donor. After engraftment on day 26, bone marrow was hypocellular, but free of leukemia. Planned immunosuppression consisted of antithymocyte globulin (ATG) on day -4 to -2, mycophenolate mofetil until day 28 and cyclosporine until day 100.

A high level of C-reactive protein (CRP) before and around allo-HSCT was observed which decreased slightly but stayed considerably high throughout the entire observation period (Fig. 2C and Additional file 2: Table S3). After leukocyte depletion around allo-HSCT, the count increased to around 3600/µl 20 days after allo-HSCT and further increased to a normal value around 80 days after-HSCT. However, high fluctuations and later a decrease in the leukocyte count were observed (Fig. 2C and Additional file 2: Table S3). Further information such as different blood counts and creatinine levels over the course of treatment are provided in Additional file 2: Table S3.

As the patient had prolonged neutropenia due to refractory leukemia and intensive chemotherapy, various antibiotics and antifungals were used to treat infectious complications before and during transplantation. More specifically beginning from day -17 she received piperacillin/tazobactam for neutropenic fever and this was changed to meropenem on day -14 for refractory fever. On day -11, vancomycin was added and on day -4, meropenem was exchanged for tigecycline. Additionally, the patient was treated with a fluoroquinolone (levofloxacin), ceftazidime and liposomal amphotericin B (Fig. 2D).

74 days after allo-HSCT, the patient developed aGvHD overall grade III, skin stage II and GIT stage III. As the patient did not respond to 2 mg/kg prednisolone and deteriorated rapidly, ATG (5 mg/kg body weight) was administered for four days as second line GvHD treatment. A partial remission

of intestinal GvHD was noted with reduction of diarrhea from > 20 stools per day to 4-5 per day. She was bedridden with general fatigue and malaise. With continuous signs of infection and lower back pain an MRI scan of the spine showed a paravertebral abscess which was removed surgically on day 126.

A multidrug-resistant *Escherichia coli* was isolated both from the abscess and from a blood culture, and was analysed further. After surgery the patient's health status improved, was able to walk again and could be discharged from hospital at day 209. She was readmitted on day 260 with suspected sepsis. The patient deceased at day 268 due to GvHD and systemic inflammatory response syndrome suspected to be bacterial sepsis. However, no pathogen could be recovered from blood cultures.

In order to explore the treatment-induced effects on the GIT microbiome in more detail and relate them to the detrimental treatment outcome, we used a meta-omic approach including MG and MT analyses in addition to rRNA gene amplicon sequencing. For this patient, samples at later time points were available, i.e. four months after allo-HSCT, which allowed investigation of the GIT microbiome over an extended period of time.

Patient A07 - changes in the microbial community structure during the treatment

Fecal samples were taken, as indicated in Fig. 2D, at days -13 (sample A07-1), day 75 (sample A07-2) and day 119 (sample A07-3). The prokaryotic diversity decreased markedly after allo-HSCT (Fig. 2B). Similarly, in sample A07-1 177 different OTUs were detected, while A07-2 and A07-3 only contained 62 and 79 OTUs, respectively.

Dominant OTUs of sample A07-1 reappeared in A07-3, more precisely several OTUs representing *Bacteroides* spp., *Escherichia/Shigella* sp. and *Enterococcus* sp. (Fig. 2A). However, many of the less abundant OTUs, belonging to 25 different genera, disappeared entirely, including for example *Anaerostipes* and *Clostridium IV* cluster (complete list of OTUs and their relative number of reads in each sample in Additional file 2: Table S4). OTUs with decreased abundance in sample A07-3 (compared to sample A07-1) represented 50 genera, for example *Alistipes*, *Barnesiella*, *Blautia*, *Clostridium* cluster XIVa and cluster XI, *Prevotella*, *Roseburia* and *Ruminococcus*. In addition, OTUs belonging to the genus *Lactobacillus* exhibited a 10-fold increase in relative abundance. Furthermore, different OTUs belonging to the genus *Bacteroides* increased in relative abundance resulting in a total relative abundance of *Bacteroides* spp. in A07-3 of 63 % compared to a total relative abundance of 27 % in A07-1 (Fig. 2A). This difference was mainly due to the increase in relative abundance of two *Bacteroides* OTUs, with an increase from 2.2 % to 23.5 % and from 0.9 % to 11.1 %, respectively. In total, 19 different OTUs belonging to the genus *Bacteroides* were detected in the first sample, 23 different OTUs in the last sample, and only 5 different *Bacteroides* OTUs were identified at TP2 which accounted for 0.07 % overall. One OTU belonging to the domain archaea could be identified, *Methanobrevibacter smithii*, which accounted for 3.4 % total relative abundance in A07-1. Similar to the short-term developments observed in the whole cohort and described above, the eukaryotic microbial community did not exhibit pronounced changes over time (Fig. 2B). Taken together, a drastic decrease in prokaryotic diversity, with relative expansion of few bacteria, including potential pathogens, was observed.

Metagenomic and metatranscriptomic data generation

Coupled MG and MT datasets of samples A07-1 (pre-treatment) and A07-3 (post-treatment) were generated and analyzed in order to inspect the changes in the GIT microbiome and the effects of allo-HSCT and concurrent antibiotics use after an extended period of time. As a comparison, samples from four healthy individuals (referred to as "reference healthy microbiomes" or "RHMs") were analyzed in the same way.

Statistics such as the number of genes, the number of raw read pairs, number of reads after preprocessing, number of contigs, maximum length, average length and total length of the contigs are provided in Additional file 2: Table S5 and Table S6.

Population-level structure of the pre- and post-treatment microbial communities

To gain a comprehensive overview of the populations present in either sample, a method for automated binning of the contigs based on the BH-SNE embedding was employed. This binning method allowed the identification of 134 and 14 individual population-level genomic complements, representing individual populations, in the pre-treatment and post-treatment samples, respectively (Fig. 3A and 3B). The visual impressions of the two embeddings reflect the drastic change in the GIT microbiome, in particular the decrease in diversity with the representation of the post-treatment sample A07-3 being exceptionally sparse (Fig. 3B). The most abundant populations were identified as *Escherichia coli*, *Enterococcus faecium*, *Lactobacillus reuteri*, *Lactobacillus rhamnosus* and several species assigned to the genus *Bacteroides*, which is in agreement with the 16S rRNA gene sequencing-based results (Fig. 2A). Representation of both samples within a single plot allows visual discrimination of clusters that are specific to one sample, or present in both samples (Fig. 3C). In accordance with the results from 16S rRNA gene sequencing (Fig. 2A), the majority of the clusters were only found in the pre-treatment sample, while other clusters comprised contigs from both samples and two clusters in the post-treatment sample were identified as *Lactobacillus reuteri* and *Lactobacillus rhamnosus*, which were either not present, or lowly abundant in sample A07-1 (Fig. 3C).

Given the potential role of opportunistic pathogens in aGvHD [20], we were specifically interested in two opportunistic pathogens that were found in both samples and whose genomes could be recovered with high completeness. We identified populations of *Escherichia coli* and *Enterococcus faecium*, which were inspected further. The population-level genomes from both samples were reassembled to allow direct comparison of identified variants as well as of the complement of antibiotic resistance genes (ARGs) encoded by them and detected in each sample.

Evidence for selective pressure at the strain-level

To uncover evidence of possible selective sweeps in the populations of interest (the opportunistic pathogens *Escherichia coli* and *Enterococcus faecium*), caused by administration of antibiotics, we performed a gene-wise protein sequence comparison of the different population-level genomes. This analysis revealed that 97.4 % of the genes found in the different population-level genomes of *E. coli*, reconstructed from samples A07-1 and A07-3, were 100 % identical and only 1.1 % of the genes were less than 95 % identical. In *E. faecium*, only 76 % of the genes were completely identical and 13.2 % of the genes showed less than 95 % identity.

The average MG depths of coverage (Additional file 3: Fig. S1A and S1B) indicated that the population size of *E. coli* was smaller after allo-HSCT (Fig. 4A), while the population size of *E. faecium* remained rather constant (Fig. 4B). In *E. coli*, a similarly high number of variants was identified in both the pre- and post-treatment samples, with an important overlap of variants identified in both populations (Fig. 4C), whereas only a few variants were present in *E. faecium* of both samples (Fig. 4D). A similar pattern of variant distributions in both samples was observed for *E. coli* (Additional file 3: Fig. S1A and S1C), while the variant pattern in *E. faecium* (Additional file 3: Fig. S1B and S1D) changed between both samples. Observed nucleotide variant frequencies and patterns of variant distributions indicated that the *E. coli* populations were composed of different strains in both samples, which persisted over the course of the treatment. In contrast, *E. faecium* was mainly represented by a single strain in each sample, and the strain of the first sample was replaced by a different strain in the second sample.

Coupled metagenomic and metatranscriptomic analysis of antibiotic resistance genes in pre- and post-treatment samples from patient A07

The relative abundance of detected ARGs (percentage of ARGs relative to the total number of genes, Fig. 5A) in the post-treatment sample (0.39 %) was significantly higher than the relative abundance of ARGs in the pre-treatment sample (0.28 % ARGs, *P* value 6.9 $*10^{-4}$, Fisher's exact test) while the relative abundances of ARGs of both the pre- and post-treatment sample were higher than the average relative abundance in the RHMs (0.20 % ± 0.01 %, *P* value 5.601 $* 10^{-7}$ and 3.278 $* 10^{-10}$, Additional file 2: Table S5). Moreover, the expression of ARGs was higher in both samples from patient A07 when compared to the RHMs (Fig. 5B).

Identification of antibiotic resistance genes in population-level genomes of opportunistic pathogens

Given the higher number and expression of ARGs in the post-treatment sample of patient A07, we were interested whether this could also be detected in the specific populations *E. coli* and *E. faecium*. Within the population-level genome of *E. coli*, 31 ARGs were identified in both samples and 2 additional genes were detected in the post-treatment sample only. In *E. faecium*, 25 ARGs were identified in both samples of which 21 genes were identical in both samples (a complete list of identified genes is provided in Additional file 2: Table S7, summaries of the ARGs identified in each population-level genome are listed in Table 2 and Table 3). In *E. coli*, 20 of the 31 ARGs that were found in both samples, exhibited higher levels of expression in the post-treatment sample while in *E. faecium*, 18 out of 21 ARGs showed higher expression post-HSCT (Fig. 5C, Additional file 2: Table S7). Although patient A07 was only treated with antibiotics until day 18 (Fig. 2D), expression of the ARGs was in general higher in the post-treatment sample, both in the whole sample (Fig. 5B), as well as in the specific populations (Fig. 5C).

Table 2: Antibiotic resistance genes identified in population-level genomes of GIT *E. coli* from patient A07

Resfams_ID	**Number of Genes**	**Resfam Family Name**	**Mechanism**
RF0005	1	AAC6-Ib	Aminoglycoside Modifying Enzyme
RF0007	3	ABCAntibioticEffluxPump	ABC Transporter
RF0027	1	ANT3	Aminoglycoside Modifying Enzyme
RF0035	1	baeR	Gene Modulating Resistance
RF0053	1	ClassA	Beta-Lactamase
RF0055	1	ClassC-AmpC	Beta-Lactamase
RF0056	1	ClassD	Beta-Lactamase
RF0065	1	emrB	MFS Transporter
RF0088	1	macA	ABC Transporter
RF0089	1	macB	ABC Transporter
RF0091	1	marA	Gene Modulating Resistance
RF0098	1	MexE	RND Antibiotic Efflux
RF0101	1	MexX	RND Antibiotic Efflux
RF0112	1	phoQ	Gene Modulating Resistance
RF0115	6	RNDAntibioticEffluxPump	RND Antibiotic Efflux
RF0121	1	soxR	Gene Modulating Resistance
RF0147	1	tolC	ABC Transporter
RF0168	6	TE_Inactivator	Antibiotic Inactivation
RF0172	1	APH3''	Phosphotransferase
RF0173	1	APH3'	Phosphotransferase
RF0174	1	ArmA_Rmt	rRNA Methyltransferase

Table 3: Antibiotic resistance genes identified in population-level genomes of GIT *E. faecium* from patient A07

Resfams_ID	**Number of Genes**	**Resfam Family Name**	**Mechanism**
RF0004	1	AAC6-I	Aminoglycoside Modifying Enzyme
RF0007	9	ABCAntibioticEffluxPump	ABC Transporter
RF0033	1	APH3	Aminoglycoside Modifying Enzyme
RF0066	1	emrE	Other Efflux
RF0067	1	Erm23SRibosomalRNAM	rRNA

		ethyltransferase	Methyltransferase
RF0104	1	MFSAntibioticEffluxPump	MFS Transporter
RF0134	1	Tetracycline_Resistance_MFS_Efflux_Pump	Tetracycline MFS Efflux
RF0154	1	vanR	Gylcopeptide Resistance
RF0155	2	vanS	Gylcopeptide Resistance
RF0168	1	TE_Inactivator	Antibiotic Inactivation
RF0172	2	APH3''	Aminoglycoside Modifying Enzyme
RF0173	2	APH3'	Aminoglycoside Modifying Enzyme
RF0174	6	ArmA_Rmt	Aminoglycoside Resistance

Genomic characterization of a blood culture *Escherichia coli* isolate and comparison to GIT populations

The genomes of a blood culture isolate and GIT population-level genomes of *E.* coli from patient A07 exhibited an average nucleotide identity of 100 %. A heatmap and corresponding dendrogram based on the *E. coli* pangenomes indicated that the genomes of the *E. coli* isolated from patient A07 and genomes from the GIT MG data were closer related to each other than to any other reference *E. coli* (Additional file 4: Fig. S2). In the genome of the *E. coli* isolate, the same ARGs as in the pre- and post-treatment GIT *E. coli* could be identified, with 4 additional ARGs compared to the post-treatment GIT *E. coli*.

Discussion

Short-term structural changes in the gastrointestinal microbiome following an allogeneic stem cell transplantation

We observed a strong impact of allo-HSCT and accompanying treatment including antibiotic use on the GIT microbiome, with a marked decrease in bacterial diversity. The observed diversity indices are in agreement with values found in an earlier study [9]. The observed trend of a reduced bacterial diversity at engraftment in patients who did not survive (Fig. 1G), is in accordance with a study focussing on this link [21]. A significant decrease in important short-chain-fatty-acid (SCFA) producers [57–59] (the three bacterial genera *Roseburia*, *Bifidobacterium* and *Blautia*, Fig. 1E) was observed. SCFAs, especially the histone deacetylase inhibitor butyrate, are the main energy source for colonocytes [57], as well as anti-inflammatory agents which regulate NF-κB activation in colonic epithelial cells [57]. Additionally, butyrate enhances the intestinal barrier function by regulating assembly of epithelial tight junctions [60] and a recent study showed that local administration of exogenous butyrate mitigated GvHD in mice [61]. Depletion of these important bacteria has been highlighted to pose an additional risk for developing GvHD or infections after allo-HSCT [26, 62]. Therefore, in addition to damage in epithelial cells due to chemotherapy, loss in SCFA-producing bacteria could further compromise intestinal barrier integrity and facilitate translocation of bacteria and PAMPs.

We found that fungi were the most prominent eukaryotes and that the eukaryotic diversity was stable during the treatment and thus not affected by antibiotic treatment and ensuing changes in bacterial community structure. However, antibiotic treatment might indirectly increase the risk for invasive fungal infections, by opening niches to these organisms, which were previously occupied by commensal bacteria. Although we did not observe any clear treatment-induced effects on the eukaryotic communities in the patient samples analyzed, it is nonetheless important to also account for the eukaryotes in future studies as overgrowth thereof has previously been linked to adverse treatment outcomes [14].

Long-term effect of allogeneic stem cell transplantation on the gastrointestinal microbiome

Employing detailed integrated meta-omic analyses of the samples from one patient, we demonstrate the effects of allo-HSCT and accompanying treatment on the GIT microbiome and consequently on the patient over an extended period of time. Only one study so far has followed the GIT microbiome trajectory up to three months after allo-HSCT [63]. Contrary to this study, which observed that the richness and metabolic capacity of the microbial community recovered after two months [63], our study found that the GIT microbial community in patient A07 did not regain its initial composition even four months after allo-HSCT, which is likely linked to the detrimental treatment outcome. Diversity was still decreased and many bacterial taxa remained absent or at drastically decreased relative levels. Taxa with decreased relative abundance were mainly bacteria whose presence in the human GIT is associated with health-promoting properties (such as butyrate production) and whose absence has been linked to negative consequences (such as inflammation) [64–66]. The genus *Blautia* for instance, has been linked to reduced aGvHD-associated death and improved overall survival [26] and *Barnesiella* with resistance to intestinal domination with vancomycin-resistant enterococci in allo-HSCT patients [67]. On the other hand, potential pathogens like *Fusobacterium* sp. and *Proteus* sp. appeared in the post-treatment sample, which were not detected in the first sample. Consecutive loss in intestinal barrier integrity could have allowed a GIT-borne *E. coli* to cause a paravertebral abscess.

Coinciding with the development of severe aGvHD (expressed by severe diarrhea) 75 days after allo-HSCT, 16S rRNA gene amplicon sequencing revealed a GIT microbiome in a notably dysbiotic state with a low diversity and dominance of two opportunistic pathogens, *E. coli* and *E. faecium*. The dominance of *E. faecium* has been observed to be quite common in allo-HSCT recipients and has been linked to higher occurrence of bacteremia and/or GIT GvHD [9, 24]. A high relative abundance of *E. faecium* is also observed in sample A07-2. Broad-spectrum antibiotic therapy, which has been associated with higher GvHD-related mortality [68], can reduce mucosal innate immune defences through elimination of commensal microbes, thereby allowing the expansion of specific bacterial taxa, such as *E. faecium*, which carry multiple antibiotic resistance mechanisms [69, 70]. Our findings suggest that this specific population expanded in response to antibiotic treatment.

Bacteroides spp. are normal commensals of the human GIT microbiome, they usually make up around 25 % of the community, as it is the case in sample A07-1 (Fig. 2A). However, they can also cause infections with associated mortality [71]. *Bacteroides* spp. might be able to penetrate the colonic mucus and persevere within crypt channels. These reservoirs might persist even during antibiotic treatment [72]. Different species of the genus *Bacteroides* produce bacteriocins [73–75], a trait that might have made it possible for these bacteria to repopulate the GIT and expand after the dysbiosis in A07-2, occupying specific niches, resulting in a relative abundance of 63 % in A07-3 (day 119).

Facultative anaerobes such as members of the orders Lactobacillales and Enterobacteriales are often observed to increase in relative abundance after treatment while obligate anaerobes such as members of the order Clostridiales often decrease in abundance [76]. *Lactobacillus rhamnosus* and *Lactobacillus reuteri* (which were detected in sample A07-3) are both often combined in probiotic formulations and are commonly considered safe and even beneficial through inhibition of potential pathogen (such as *E. coli* and *E. faecium*) expansion [77–79]. Even in patients undergoing allo-HSCT, *Lactobacillus plantarum* administration has not been found to result in higher incidence of bacteremia or aGvHD [80]. However, bacteria found in probiotic formulations, especially *Lactobacillus* species have occasionally also caused bloodstream infections [81]. Our data suggest that probiotics should be administered with great caution and should be subject to further investigations to clearly ensure safety of their usage.

Identification of antibiotic resistance genes in population-level genomes of opportunistic pathogens and evidence for selective pressure at the strain-level

A higher ratio of ARGs within the microbial community was observed post-treatment, even a few months after the antibiotic treatment was concluded (Fig. 5A). Importantly, the observed expression of ARGs was higher in the post-treatment sample (Fig. 5B) when compared to the pre-treatment sample. Strains that carry mutations which lead to higher expression of ARGs might have been selected for by the antibiotic treatment [82].

In *E. coli*, three different genes conferring resistance against β-lactams were identified, one of which was only detected in the post-treatment sample, which might have been acquired due to selective pressure given the administration of three different β-lactam antibiotics during the treatment.

Observed nucleotide variant frequencies and patterns of variant distributions indicated that the treatment may have constituted a genetic bottleneck for *E. faecium*, culminating in the observed lower genetic diversity. This also suggests that two different mechanisms influenced the respective compositions of *E. coli* and *E. faecium* populations. While the *E. coli* population remained relatively unaffected, the *E. faecium* population underwent a selective sweep in response to the antibiotic treatment with selection of a specific genotype expressing ARGs. Overall, our observations indicate that antibiotic pressure and associated selection of bacteria encoding ARGs are likely essential factors in governing the observed expansion in opportunistic pathogens.

Interestingly, the multidrug-resistant *E. coli* that was isolated from a blood culture, was closely related to the GIT-borne *E. coli* population. The overlap of ARGs identified in each genome further indicates their association. These findings are a proof for the potential fatal effects of dysbiosis associated pathogen dominance in the GIT and subsequent systemic infections on patient survival.

Based on our observation, one strategy to avoid a treatment-induced intestinal domination by pathogens could consist in the tailored administration of several, not single probiotic strains, composed in dependence of the individual GIT microbiome changes during therapy. A different approach could consist in fecal microbiome transplantation, either as "autologous" (transplanting the pre-transplant microbiome) or "allogeneic" graft (from the donor of the stem cells). Preservation of a diverse microbiome, able to inhibit expansion of potential pathogens, might be a new approach to avoid treatment related side effects.

Conclusions

We observed drastic changes in the prokaryotic composition of the gastrointestinal microbiome of patients after allo-HSCT and supportive care, with a decrease in bacterial diversity. Pronounced changes in community structure can persevere and present as extensive dysbiosis with potential fatal effects on patient outcome. We observed increases in the number and expression levels of ARGs and the different ways in which bacterial populations respond to antibiotic stress

An individually tailored treatment, either by i) limiting the usage of broad-spectrum antibiotics, ii) via fecal microbiome transplantation and/or iii) via administration of specific probiotics could help to modulate the microbiome to increase tolerance or improve the overall efficacy of the therapy.

Abbreviations

aGvHD: acute graft-versus-host disease **allo-HSCT:** allogeneic hematopoietic stem cell transplantation **ARG:** antibiotic resistance genes **ATG:** antithymocyte globulin **bp:** base pair **cDNA:** complementary DNA **RHM:** reference healthy microbiome **Contig(s):** contiguous sequence(s) **GIT:** gastrointestinal tract **GvHD:** graft-versus-host disease **IMP:** Integrated Meta-omic Pipeline **MG:** metagenomic **MT:** metatranscriptomic **NCBI:** National

Center for Biotechnology Information **nt:** nucleotide **OTU:** operational taxonomic unit **PAMP:** pathogen-associated molecular pattern **rRNA:** ribosomal RNA **SNV:** single nucleotide variant **TP:** time point

Declarations

Ethics approval and consent to participate

The study was approved by the Ethics review board of the Saarland amendment 1 and 2 (reference number 37/13), and by the Ethics Review Panel of the University of Luxembourg (reference number ERP-15-029). Written informed consent was obtained from study participants prior to enrolment.

Consent for publication

Not applicable.

Availability of data and materials

Reassembled population-level genomes of *Escherichia coli* (ID 6666666.166711) and *Enterococcus faecium* (ID 6666666.166708) are accessible via the RAST guest account (http://rast.nmpdr.org, login: guest; password: guest).

For samples A07-1 and A07-3, preprocessed MG and MT reads (after adapter trimming, quality filtering, rRNA removal and removal of reads mapping to the human genome) were submitted to the NCBI Sequence Read Archive (SRA) repository under the BioProject ID PRJNA317435 (http://www.ncbi.nlm.nih.gov/bioproject/317435).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was partially funded by the University of Luxembourg (ImMicroDyn1) as well as by an ATTRACT program grant (ATTRACT/A09/03) and a European Union Joint Programming in Neurodegenerative Disease grant (INTER/JPND/12/01) and a proof of concept grant (PoC/13/02) awarded to PW as well as Aide à la Formation Recherche grants to CCL (AFR PHD/4964712), LW (AFR PHD-2013-5824125) and to SN (AFR PHD-2014-1/7934898) and a CORE junior to EELM (C15/SR/10404839), all funded by the Luxembourg National Research Fund (FNR).

Authors' contributions

JGS, PW, NG, AS and JB initiated and designed the study. JB, NG and AS recruited patients and collected samples. KF collected clinical data. AK performed experiments. AHB, SN and CCL developed the bioinformatic analysis methods. AK, AHB, EELM and PW contributed to analysis and interpretation of the data. AK, AHB, EELM, JB, JGS, LW and PW wrote and revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank the medical staff of the Hematology Department and of the Pediatric Department of the Saarland University Medical Center in Homburg, Germany, especially Manuela Faust, Eyad Torfah and Michael Ehrhardt for sample and data collection. We are deeply grateful for all the patients that have participated in the study. Laura A. Lebrun and Claire Battin are thanked for their assistance with fecal sample extractions. Stephanie Kreis and Giulia Cesi are thanked for valuable discussions in the establishment of the project. *In silico* analyses presented in this paper were carried out using the HPC facilities of the University of Luxembourg [83].

Additional files

Additional file 1: Document S1. Additional information on methods used in this study. (DOCX 126 kb)

Additional file 2: Table S1. Number of reads per prokaryotic operational taxonomic unit (OTU) and sample from the cohort. **Table S2.** Number of reads per eukaryotic operational taxonomic unit (OTU) and sample from the cohort. **Table S3.** Blood cell counts and clinical data from patient A07. **Table S4.** Number of reads per operational taxonomic unit (OTU) and sample from patient A07. **Table S5.** Numbers of identified antibiotic resistance genes and total number of genes. Numbers of identified antibiotic resistance genes in relation to total numbers of genes in samples from patient A07 before and after allo-HSCT and from four healthy individuals. **Table S6.** Statistics of the metagenomic and metatranscriptomic datasets and the co-assembled contigs. **Table S7.** Antibiotic resistance genes in population-level genomes and their expression in the pre- and post-treatment sample. (XLSX 273 kb)

Additional file 3: Figure S1. Variant distribution in *E. coli* and *E. faecium* (A and B) Representation of exemplary sections of the reassembled population-level genomes with aligned reads of both samples highlighting occurrences of variants in each population, visualized with the Integrative Genomics Viewer. Length of the represented section is indicated as well as the average MG depth of coverage of each reconstructed population-level genome. (C and D) Histogram of the variant frequencies of the minor nucleotide at all variant positions. Panels on the left represent results for *E. coli*, panels on the right represent results for *E. faecium*. Blue figure elements refer to the pre-treatment sample (A07-1), red figure elements refer to the post-treatment sample (A07-3). (PDF 570 kb)

Additional file 4: Figure S2. Gene set profiles of the 118 reference strains and 3 *E.* coli isolated from patient A07 (highlighted in red and marked with a light blue box). Each row represents a gene (blue: present, yellow: absent), each column represents a strain. (PDF 347 kb)

References

1. Hooper L V., Gordon J. Commensal host-bacterial relationships in the gut. Science. 2001;292:1115–8.

2. Sekirov I, Russell SL, Antunes LCM, Finlay BB. Gut microbiota in health and disease. Physiol Rev. 2010;90:859–904.

3. Sommer F, Bäckhed F. The gut microbiota-masters of host development and physiology. Nat Rev Microbiol. 2013;11:227–38.

4. Round JL, Mazmanian SK. The gut microbiota shapes intestinal immune responses during health and disease. Nat Rev Immunol. 2009;9:313–23.

5. Atarashi K, Tanoue T, Oshima K, Suda W, Nagano Y, Nishikawa H, et al. Treg induction by a rationally selected mixture of Clostridia strains from the

human microbiota. Nature. 2013;500:232–6.

6. Stecher B, Maier L, Hardt W-D. “Blooming” in the gut: how dysbiosis might contribute to pathogen evolution. Nat Rev Microbiol. 2013;11:277–84.

7. Yu LC-H, Shih Y-A, Wu L-L, Lin Y-D, Kuo W-T, Peng W-H, et al. Enteric dysbiosis promotes antibiotic-resistant bacterial infection: systemic dissemination of resistant and commensal bacteria through epithelial transcytosis. Am J Physiol Gastrointest Liver Physiol. 2014;307:824–35.

8. Khosravi A, Mazmanian SK. Disruption of the gut microbiome as a risk factor for microbial infections. Curr Opin Microbiol. 2013;16:221–7.

9. Taur Y, Xavier JB, Lipuma L, Ubeda C, Goldberg J, Gobourne A, et al. Intestinal domination and the risk of bacteremia in patients undergoing allogeneic hematopoietic stem cell transplantation. Clin Infect Dis. 2012;55:905–14.

10. Einsele H, Bertz H, Beyer J, Kiehl MG, Runde V, Kolb HJ, et al. Infectious complications after allogeneic stem cell transplantation: Epidemiology and interventional therapy strategies - Guidelines of the Infectious Diseases Working Party (AGIHO) of the German Society of Hematology and Oncology (DGHO). Ann Hematol. 2003;82(Suppl 2):175–85.

11. Aminov RI, Mackie RI. Evolution and ecology of antibiotic resistance genes. FEMS Microbiol Lett. 2007;271:147–61.

12. Salyers A, Gupta A, Wang Y. Human intestinal bacteria as reservoirs for antibiotic resistance genes. Trends Microbiol. 2004;12:412–6.

13. Samonis G, P. KD, Maraki S, Alegakis D, Mantadakis E, Papadakis JA, et al. Levofloxacin and moxifloxacin increase human gut colonization by *Candida* species. Antimicrob Agents Chemother. 2005;49:5189.

14. Zollner-Schwetz I, Auner HW, Paulitsch A, Buzina W, Staber PB, Ofner-Kopeinig P, et al. Oral and intestinal *Candida* colonization in patients undergoing hematopoietic stem-cell transplantation. J Infect Dis. 2008;198:150–3.

15. Tuncer HH, Rana N, Milani C, Darko A, Al-Homsi SA. Gastrointestinal and hepatic complications of hematopoietic stem cell transplantation. World J Gastroenterol. 2012;18:1851–60.

16. Couriel D, Caldera H, Champlin R, Komanduri K. Acute graft-versus-host disease: pathophysiology, clinical manifestations, and management. Cancer. 2004;101:1936–46.

17. Jacobsohn DA, Vogelsang GB. Acute graft versus host disease. Orphanet J Rare Dis. 2007;2:35.

18. Glucksberg H, Storb R, Fefer A, Buckner CD, Neiman PE, Clift RA, et al. Clinical manifestations of graft-versus-host disease in human recipients of marrow from HL-A-matched sibling donors. Transplantation. 1974;18:295–304.

19. Ferrara JL, Levine JE, Reddy P, Holler E. Graft-versus-host disease. Lancet. 2009;373:1550–61.

20. Penack O, Holler E, van den Brink MRM. Graft-versus-host disease: regulation by microbe-associated molecules and innate immune receptors. Blood. 2010;115:1865–72.

21. Taur Y, Jenq RR, Perales M, Littmann ER, Morjaria S, Ling L, et al. The effects of intestinal tract bacterial diversity on mortality following allogeneic hematopoietic stem cell transplantation. Blood. 2014;124:1174–82.

22. Van Bekkum DW, Roodenburg J, Heidt PJ, Van der Waaij D. Mitigation of

secondary disease of allogeneic mouse radiation chimeras by modification of the intestinal microflora. J Natl Cancer Inst. 1974;52:401–4.

23. Vossen JM, Guiot HFL, Lankester AC, Vossen ACTM, Bredius RGM, Wolterbeek R, et al. Complete suppression of the gut microbiome prevents acute graft-versus-host disease following allogeneic bone marrow transplantation. PLoS One. 2014;9:e105706.

24. Holler E, Butzhammer P, Schmid K, Hundsrucker C, Koestler J, Peter K, et al. Metagenomic analysis of the stool microbiome in patients receiving allogeneic stem cell transplantation: loss of diversity is associated with use of systemic antibiotics and more pronounced in gastrointestinal graft-versus-host disease. Biol Blood Marrow Transplant. 2014;20:640–5.

25. Montassier E, Batard E, Massart S, Gastinne T, Carton T, Caillon J, et al. 16S rRNA gene pyrosequencing reveals shift in patient faecal microbiota during high-dose chemotherapy as conditioning regimen for bone marrow transplantation. Microb Ecol. 2014;67:690–9.

26. Jenq RR, Taur Y, Devlin SM, Ponce DM, Goldberg JD, Ahr KF, et al. Intestinal *Blautia* is associated with reduced death from graft-versus-host disease. Biol Blood Marrow Transplant. 2015;21:1373–83.

27. Roume H, Heintz-Buschart A, Muller EEL, Wilmes P. Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. Methods Enzymol. 2013;531:219–36.

28. Roume H, Muller EEL, Cordes T, Renaut J, Hiller K, Wilmes P. A biomolecular isolation framework for eco-systems biology. ISME J. 2013;7:110–21.

29. Hugerth LW, Wefer HA, Lundin S, Jakobsson HE, Lindberg M, Rodin S, et

al. DegePrime, a program for degenerate primer design for broad-taxonomic-range PCR in microbial ecology studies. Appl Environ Microbiol. 2014;80:5116–23.

30. Herlemann DP, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ, Andersson AF. Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. ISME J. 2011;5:1571–9.

31. Hugerth LW, Muller EEL, Hu YOO, Lebrun LAM, Roume H, Lundin D, et al. Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. PLoS One. 2014;9:e95567.

32. Hildebrand F, Tadeo R, Voigt A, Bork P, Raes J. LotuS: an efficient and user-friendly OTU processing pipeline. Microbiome. 2014;2:30.

33. Processing amplicons with non-overlapping reads https://github.com/EnvGen/Tutorials/blob/master/amplicons-no_overlap.rst. Accessed 17 Nov 2016

34. R Development Core Team. R: A language and environment for statistical computing. 2008;5.

35. Oksanen AJ, Blanchet FG, Kindt R, Minchin PR, Hara RBO, Simpson GL, et al. Package “ vegan .” 2015.

36. Wickham H. *ggplot2 Elegant Graphics for Data Analysis*. 1st edition. Springer-Verlag New York; 2009.

37. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

38. Narayanasamy S, Jarosz Y, Muller EEL, Laczny CC, Herold M, Kaysen A, et al. IMP: a pipeline for reproducible metagenomic and metatranscriptomic analyses. bioRxiv. 2016.

39. Laczny CC, Pinel N, Vlassis N, Wilmes P. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. Sci Rep. 2014;4:4516.

40. Laczny CC, Sternal T, Plugaru V, Gawron P, Atashpendar A, Margossian H, et al. VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome. 2015;3:7.

41. Muller EEL, Pinel N, Laczny CC, Hoopmann MR, Narayanasamy S, Lebrun LA, et al. Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. Nat Commun. 2014;5:5603.

42. Heintz-Buschart A, May P, Laczny CC, Lebrun LA, Bellora C, Krishna A, et al. Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. Nat Microbiol. 2016.

43. Madden T. Chapter 16 : The BLAST Sequence Analysis Tool. In *The NCBI Handbook[internet]*; 2002:1–15.

44. Huson D, Mitra S, Ruscheweyh H. Integrative analysis of environmental sequences using MEGAN4. Genome Res. 2011;21:1552–60.

45. MOLE-BLAST webservice https://blast.ncbi.nlm.nih.gov/moleblast/moleblast.cgi. Accessed 17 Nov 2016

46. Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics. 2012;28:1033–4.

47. Bankevich A, Nurk S, Antipov D, Gurevich A a., Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77.

48. ANI Average Nucleotide Identity http://enve-omics.ce.gatech.edu/ani/. Accessed 17 Nov 2016

49. Aziz RK, Bartels D, Best A a, DeJongh M, Disz T, Edwards R a, et al. The

RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008;9:75.

50. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J. 2014;9:207–16.

51. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7:e1002195.

52. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.

53. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ. 2015;3:e1319.

54. Scholz M, Ward D V, Pasolli E, Tolio T, Zolfo M, Asnicar F, et al. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. Nat Methods. 2016;13:435–8.

55. Scanlan PD, Shanahan F, Marchesi JR. Human methanogen diversity and incidence in healthy and diseased colonic groups using *mcrA* gene analysis. BMC Microbiol. 2008;8:79.

56. Schmid C, Schleuning M, Ledderose G, Tischer J, Kolb HJ. Sequential regimen of chemotherapy, reduced-intensity conditioning for allogeneic stem-cell transplantation, and prophylactic donor lymphocyte transfusion in high-risk acute myeloid leukemia and myelodysplastic syndrome. J Clin Oncol. 2005;23:5675–87.

57. Canani RB, Costanzo M Di, Leone L, Pedata M, Meli R, Calignano A.

Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. World J Gastroenterol. 2011;17:1519–28.

58. Zwielehner J, Lassl C, Hippe B, Pointner A, Switzeny OJ, Remely M, et al. Changes in human fecal microbiota due to chemotherapy analyzed by TaqMan-PCR, 454 sequencing and PCR-DGGE fingerprinting. PLoS One. 2011;6:e28654.

59. Montassier E, Gastinne T, Vangay P, Al-Ghalith G a., Bruley des Varannes S, Massart S, et al. Chemotherapy-driven dysbiosis in the intestinal microbiome. Aliment Pharmacol Ther. 2015;42:515–28.

60. Peng L, Li Z, Green RS, Holzman IR, Lin J. Butyrate enhances the intestinal barrier by facilitating tight junction assembly via activation of AMP-activated protein kinase. J Nutr. 2009;139:1619–25.

61. Mathewson ND, Jenq R, Mathew A V, Koenigsknecht M, Hanash A, Toubai T, et al. Gut microbiome–derived metabolites modulate intestinal epithelial cell damage and mitigate graft-versus-host disease. Nat Immunol. 2016;17:505–13.

62. Docampo MD, Auletta JJ, Jenq RR. The emerging influence of the intestinal microbiota during allogeneic hematopoietic cell transplantation: Control the gut and the body will follow. Biol Blood Marrow Transplant. 2015;21:1360–6.

63. Biagi E, Zama D, Nastasi C, Consolandi C, Fiori J, Rampelli S, et al. Gut microbiota trajectory in pediatric patients undergoing hematopoietic SCT. Bone Marrow Transplant. 2015;50:992–8.

64. Abreu MT, Peek RM. Gastrointestinal malignancy and the microbiome. Gastroenterology. 2014;146:1534–1546.e3.

65. Perez-Chanona E, Jobin C. From promotion to management: the wide impact of bacteria on cancer and its treatment. Bioessays. 2014;36:658–64.

66. Jiang W, Wu N, Wang X, Chi Y, Zhang Y, Qiu X, et al. Dysbiosis gut microbiota associated with inflammation and impaired mucosal immune function in intestine of humans with non-alcoholic fatty liver disease. Sci Rep. 2015;5:8096.

67. Ubeda C, Bucci V, Caballero S, Djukovic A, Toussaint NC, Equinda M, et al. Intestinal microbiota containing *Barnesiella* species cures vancomycin-resistant *Enterococcus faecium* colonization. Infect Immun. 2013;81:965–73.

68. Shono Y, Docampo MD, Peled JU, Perobelli SM, Velardi E, Tsai JJ, et al. Increased GVHD-related mortality with broad-spectrum antibiotic use after allogeneic hematopoietic stem cell transplantation in human patients and mice. Sci Transl Med. 2016;8:339ra71.

69. Brandl K, Plitas G, Mihu CN, Ubeda C, Jia T, Schnabl B, et al. Vancomycin-resistant enterococci exploit antibiotic-induced innate immune deficits. Nature. 2008;455:804–7.

70. Ubeda C, Pamer EG. Antibiotics, microbiota, and immune defense. Trends Immunol. 2012;33:459–66.

71. Wexler HM. *Bacteroides*: The good, the bad, and the nitty-gritty. Clin Microbiol Rev. 2007;20:593–621.

72. Lee SM, Donaldson GP, Mikulski Z, Boyajian S, Ley K, Mazmanian SK. Bacterial colonization factors control specificity and stability of the gut microbiota. Nature. 2013;501:426–9.

73. Nakano V, Ignacio A, Fernandes MR, Fukugaiti MH, Avila-campos MJ. Intestinal *Bacteroides* and *Parabacteroides* species producing antagonistic

substances. Curr Trends Microbiol. 2006;1.

74. Avelar KES, Pinto LJF, Antunes LCM, Lobo LA, Bastos MCF, Domingues RMCP, et al. Production of bacteriocin by *Bacteroides fragilis* and partial characterization. Lett Appl Microbiol. 1999;29:264–8.

75. Booth SJ, Johnson JL, Wilkins TD. Bacteriocin production by strains of *Bacteroides* isolated from human feces and the role of these strains in the bacterial ecology of the colon. Antimicrob Agents Chemother. 1977;11:718–24.

76. Jenq RR, Ubeda C, Taur Y, Menezes CC, Khanin R, Dudakov J a, et al. Regulation of intestinal inflammation by microbiota following allogeneic bone marrow transplantation. J Exp Med. 2012;209:903–11.

77. Borriello SP, Ammes WP, Holzapfel W, Marteau P, Schrezenmeir J, Vaara M, et al. Safety of probiotics that contain lactobacilli or bifidobacteria. 2003;36:775–80.

78. Servin AL. Antagonistic activities of lactobacilli and bifidobacteria against microbial pathogens. FEMS Microbiol Rev. 2004;28:405–40.

79. Spinler JK, Taweechotipatr M, Rognerud CL, Ou CN, Tumwasorn S, Versalovic J. Human-derived probiotic Lactobacillus reuteri demonstrate antimicrobial activities targeting diverse enteric bacterial pathogens. Anaerobe. 2008;14:166–71.

80. Ladas EJ, Bhatia M, Chen L, Sandler E, Petrovic A, Berman DM, et al. The safety and feasibility of probiotics in children and adolescents undergoing hematopoietic cell transplantation. Bone Marrow Transplant. 2016;51:262–6.

81. Cohen SA, Woodfield MC, Boyle N, Stednick Z, Boeckh M, Pergam SA. Incidence and outcomes of bloodstream infections among hematopoietic cell

transplant recipients from species commonly reported to be in over-the-counter probiotic formulations. Transpl Infect Dis. 2016:699–705.

82. Webber MA, Piddock LJ V. The importance of efflux pumps in bacterial antibiotic resistance. J Antimicrob Chemother. 2003;51:9–11.

83. Varrette S, Bouvry P, Cartiaux H, Georgatos F. Management of an academic HPC cluster: The UL experience. Proc 2014 Int Conf High Perform Comput Simulation, HPCS 2014. 2014:959–67.

Figures

Figure 1 Changes of gastrointestinal microbial community structure in patients receiving allo-HSCT. Boxplots depicting (A, E) diversity (Shannon diversity index) and (B, F) richness (Chao1 richness estimator) per collection time point (TP), for (A, B) prokaryotes (determined by 16S rRNA gene amplicon sequencing) and (E, F) eukaryotes (determined by 18S rRNA gene amplicon sequencing), respectively. The number of samples per collection TP is indicated above each box. Diversity and richness were determined after rarefaction of the dataset. Statistically significant decrease in prokaryotic diversity between TP1 and TP3 (*P* value 0.014 in Kruskal-Wallis rank sum test) and in prokaryotic richness between TP1 and TP3 (*P* value 0.026, Wilcoxon rank sum test) was observed. (C) Changes in relative abundance of three bacterial genera between TP1 and TP3. Genera with at least 1.5-fold decrease, adjusted *P* value < 0.05 and a relative abundance of at least 5 % in one sample are included (adjusted *P* value 0.0025, 0.026 and $3.68 * 10^{-5}$, Wald test). (D) Changes in relative abundance of two bacterial orders between TP1 and TP3 (adjusted *P* value 0.0054 and 0.009, Wald test). (G)

Prokaryotic diversity at TP1 and TP3 in relation to outcome 1.5 years after allo-HSCT. Samples from five patients who survived (S) and three patients who deceased (M) are represented. (C, D and G) Data from all eight patients who had samples collected at TP1 and TP3 are displayed. Collection TP1 includes samples that were taken (up to eight days) before allo-HSCT. TP2 includes samples that were taken up to four days after the transplantation. TP3 includes samples that were taken between day 20 and day 33 after the transplantation. Significant differences between TPs are indicated by asterisks (* *P* value < 0.05, ** *P* value < 0.01).

Figure 2 Variation of the microbial community structure over the course of the allo-HSCT treatment in patient A07. (A) Relative proportions of the 10 most abundant operational taxonomic units (OTUs) based on 16S rRNA gene sequencing. The remaining OTUs are summarised as "others". Similar shades of the colors represent genera belonging to the same phylum. (B) Prokaryotic (triangle) and eukaryotic (circle) diversity represented by Shannon diversity index at sampling TPs throughout the treatment. (C) C-reactive protein (CRP) blood levels (green line) and leukocyte blood count (blue line). (D) Drugs (antibiotics, antifungals and antithymocyte globulin) administered throughout the treatment. Along the x-axis, days relative to the day of transplantation are indicated. Abbreviations: Vancom=vancomycin; Tigecycl=tigecycline; Fluoroq=fluoroquinolone; Antif=antifungal; ATG=antithymocyte-globulin.

Figure 3 BH-SNE-based visualization of genomic fragment signatures of microbial communities present in samples of patient A07.
Points represent contigs ≥ 1000 nt. Clusters are formed by contigs with similar genomic signatures. (A) Visualization of pre-treatment sample contigs. (B) Visualization of post-treatment sample contigs. (A and B) Points within clusters are colored according to the reconstructed genomes' completeness, based on the number of unique essential genes. Lines within the colored bar indicate the number of clusters at each percentage of completeness. (C) Combined visualization of contigs derived from pre-treatment sample (A07-1, blue squares) and post-treatment (A07-3, red crosses) samples. The inset displays a magnification of a section of the plot representing two populations (*Lactobacillus reuteri* and *Lactobacillus rhamnosus*), which are only present in the post-treatment sample. In each representation, clusters representing *Escherichia coli* and *Enterococcus faecium* are indicated.

Figure 4 Number and distribution of variants in *Escherichia coli* and *Enterococcus faecium*. (A and B) Violin plots representing distribution of depth of coverage of the contigs contained in each population-level genome. (C and D) Venn diagrams indicating the number of variant positions exclusive to each sample respectively the number of variant positions found in both samples. Panels on the left represent results for *E. coli*, panels on the right represent results for *E. faecium*. Blue figure elements refer to the pre-treatment sample (A07-1), red figure elements refer to the post-treatment sample (A07-3).

Figure 5 Expression levels and relative abundances of antibiotic resistance genes (ARGs). (A) Percentage of identified ARGs (in relation to total number of genes) in the pre-treatment (A07-1) and post-treatment (A07-3) sample and in the GIT microbiome of four healthy untreated individuals (RHMs; ** *P* value < 0.01, Fisher's exact test). (B) Histogram of the ratios of metatranscriptomic (MT) to metagenomic (MG) depths of coverage of ARGs in the pre-treatment and post-treatment sample and in the RHMs. (C) Histograms of the ratios of MT to MG depths of coverage of ARGs in population-level genomes of *Escherichia coli* and of *Enterococcus faecium* in the pre- and post-treatment samples. Bars representing the number of ARGs at a specific expression rate in the pre-treatment sample are blue, bars representing the genes in the post-treatment sample are red and bars representing the genes in the RHMs are green. For the RHMs, the average of four datasets is represented with standard deviation as error bar.

Figure 1

Click here to download Figure Figure 1.pdf

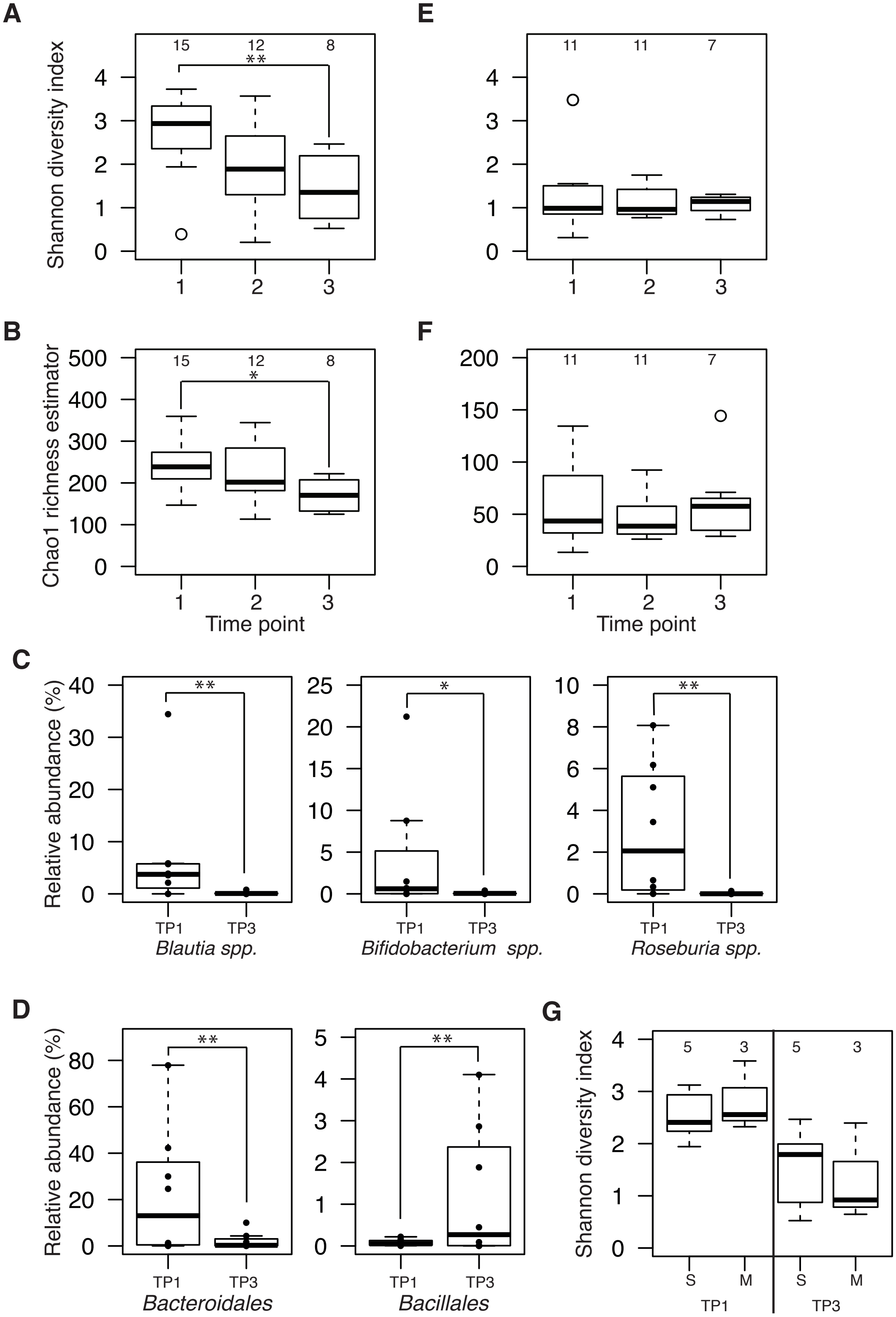


Figure 2

Click here to download Figure Figure 2.pdf

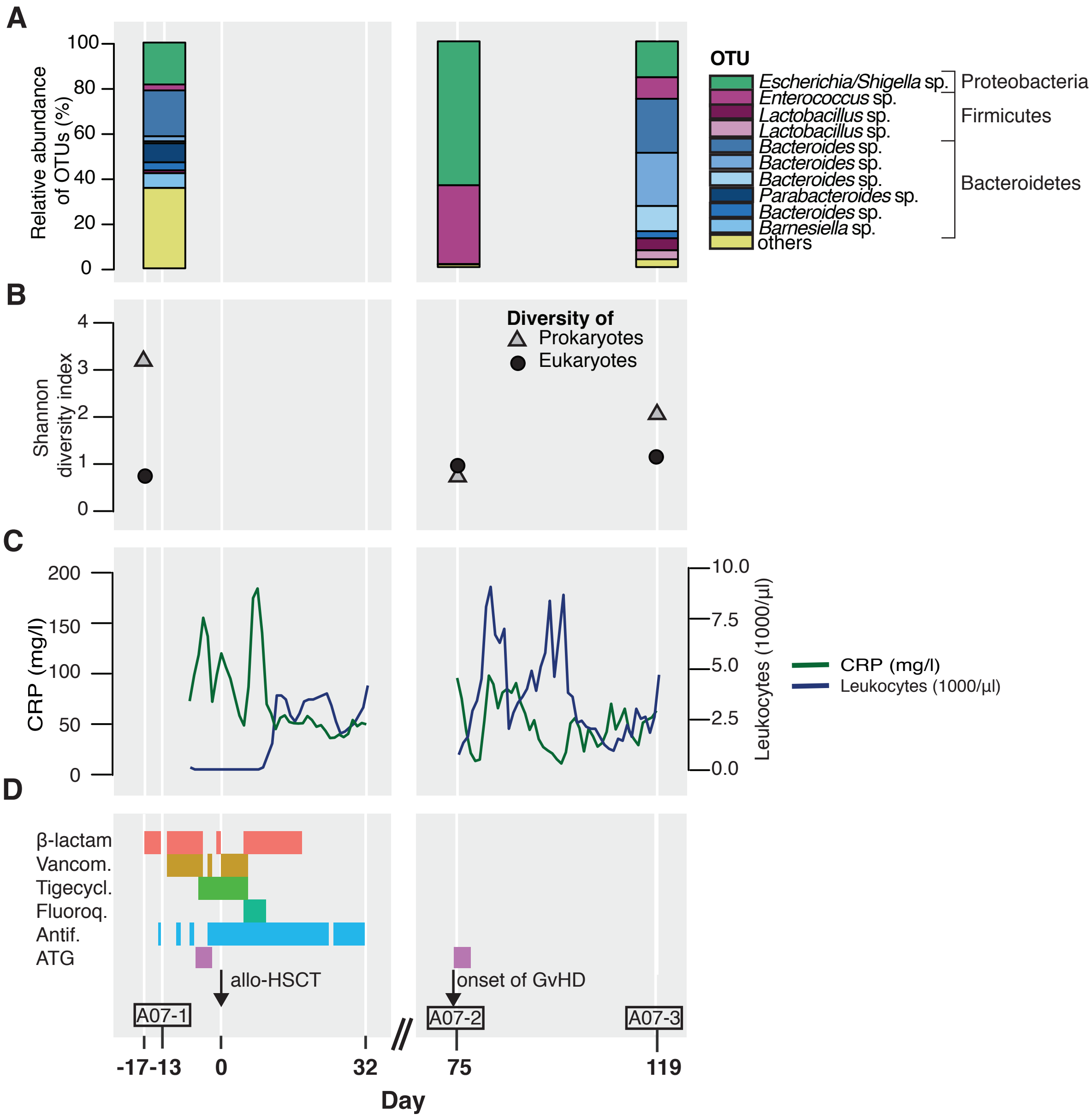


Figure 3

Click here to download Figure Figure 3.pdf

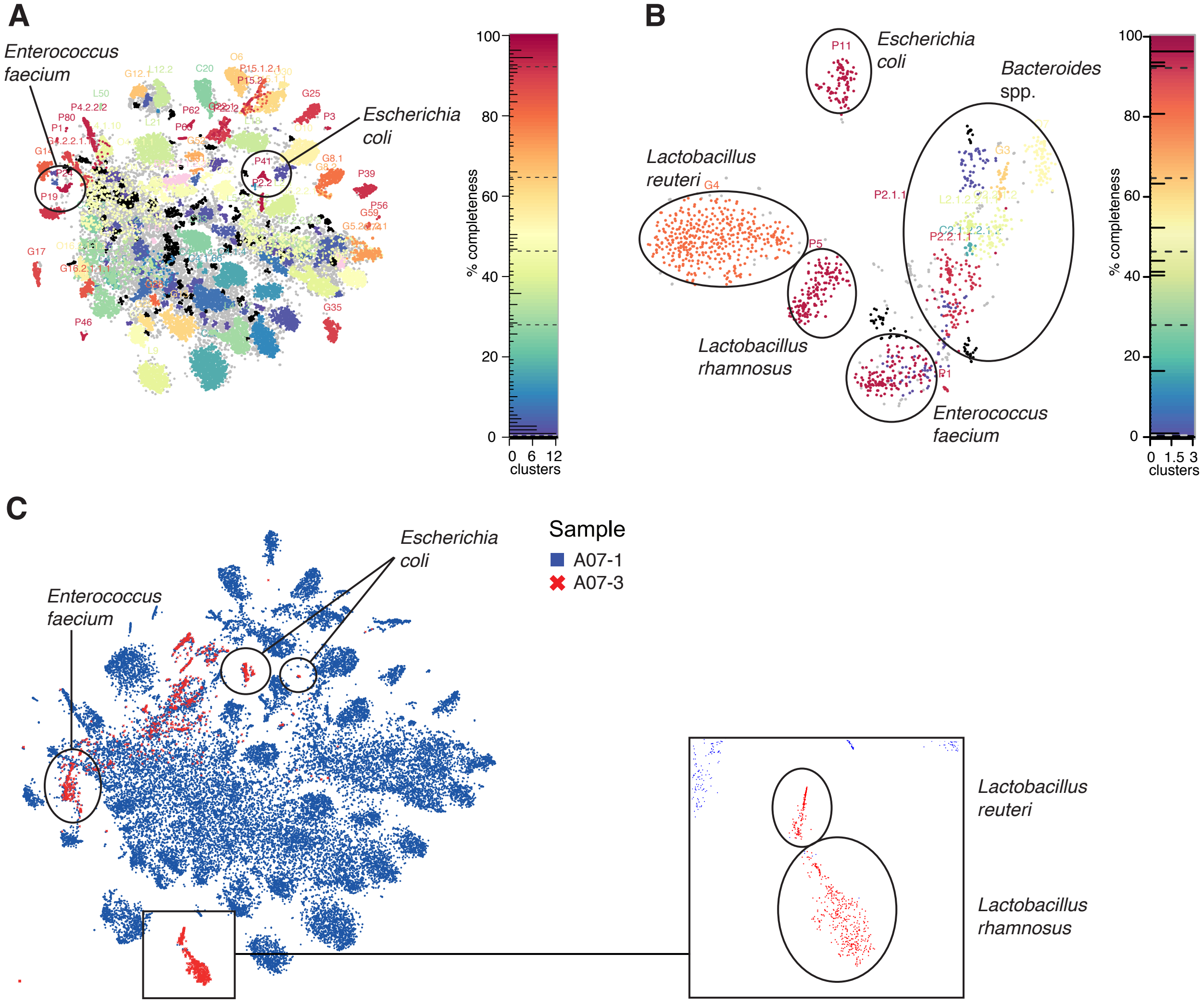


Figure 4

Click here to download Figure Figure 4.pdf

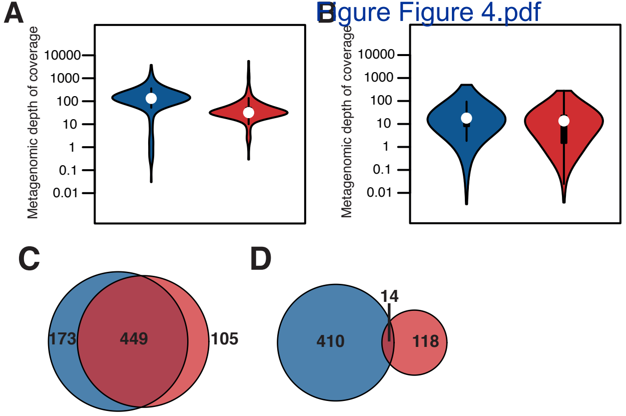
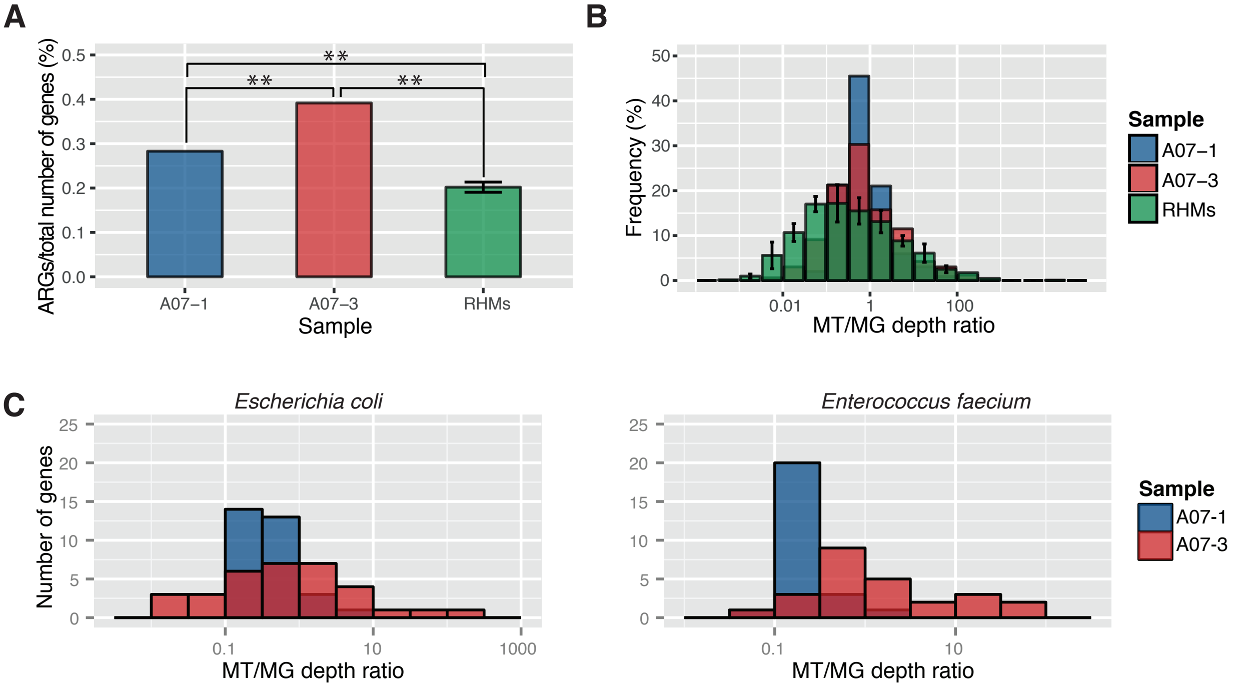


Figure 5

Click here to download Figure Figure 5.pdf



A.7 First draft genome sequence of a strain belonging to the *Zoogloea* genus and its gene expression *in situ*.

Emilie E.L. Muller[†], **Shaman Narayanasamy**[†], Myriam Zeimes, Laura A. Lebrun, Nathan D. Hicks, John D. Gillece, James M. Schupp, Paul Keim, Paul Wilmes

In preparation

Contributions of author include:

- Coordination
- Research design
- Data analysis and visualization
- Writing and revision of manuscript

[†] Co-first author

First draft genome sequence of a strain belonging to *Zoogloea* genus and its gene expression *in situ*

Authors

Emilie E.L. Muller[1*°], Shaman Narayanasamy[1*], Myriam Zeimes[1], Laura A. Lebrun[1], Nathan D. Hicks[2], John D. Gillece[2], James M. Schupp[2], Paul Keim[2], Paul Wilmes[1]

Institutional Affiliation

1. Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7 Avenue des Hauts-Fourneaux, L-4362 Esch-sur-Alzette, Luxembourg

2. TGen North, 3051 West Shamrell Boulevard, Flagstaff, Arizona 86001, USA

* authors contributed equally to this work

° current affiliation: Department of Microbiology, Genomics and the Environment, UMR 7156 UNISTRA – CNRS, Université de Strasbourg, Strasbourg, France

Corresponding author

Assistant-Professor Paul Wilmes: paul.wilmes@uni.lu

Dr Emilie Muller: emilie.muller@unistra.lfr

Abstract (Heading 1)

The Gram-negative beta-proteobacteria *Zoogloea schifflangensis* LCSB751 was originally isolated from foaming activated sludge. Here, we describe its draft genome and annotation together with a general physiological and genomic analysis, as the first sequenced representative of the *Zoogloea* genus. Moreover, *Zoogloea* gene expression in its environment is described using metatranscriptomic data obtained from the same treatment plant. The presented genomics and transcriptomics information demonstrate the pronounced capacity of this genus to synthesis PHA in wastewater.

Keywords: (Heading 1)

Genome assembly ; Genomic features ; Metatranscriptomics ; Poly-hydroxyalkanoate ; Wastewater treatement plant ; *Zoogloea schifflangensis*.

Abbreviations: (optional) (Heading 1)

Poly-β-hydroxyalkanoate (PHA)

Introduction (Heading 1)

Zoogloea spp. are chemoorganotrophic bacteria often found in organically enriched aquatic environments and known to be able to accumulate intracellular granules of poly-β-hydroxyalkanoate (PHA) [1]. The combination of these two characteristics render this genus particulary interesting from the perspective of leveraging chemical energy present in wastewater in order to produce high-value resources [2,3]. In particular, PHA may be used to synthesize biodegradable bioplastics or be chemically transformed into the biofuel hydroxybutyrate methyl ester [2].

The genus name *Zoogloea* is derived from the Greek meaning ‘animal glue’, refering to a phenotypic trait that was previously used to differentiate between *Zoogloea* species and other metabolically similar bacteria [1]. The polysaccharides composing this zoogloeal matrix was also proposed to be used as a polymer for heavy metals adsorption [4].

To date, no genome were sequenced for any of the representative strains of one of the five recognised species of *Zoogloea* and thus, very few information about *Zoogloea* genomic potential is available. In the context of a larger study designed to identify major bacterial populations naturally occuring in wastewater activated sludge that possess interesting features for biodiesel and bioplastic production, we sequenced the genome of a newly isolated representative of the *Zoogloea* genus, representing the first strain of a new species: *Zoogloea schifflangensis*.

Organism Information (Heading 1)

Classification and features (Heading 2)

Zoogloea schifflangensis LCSB751 was isolated from an activated sludge sample collected at the surface of the anoxic tank at Schifflange, Esch-sur-Alzette, Luxembourg (49°30′48.29″N; 6°1′4.53″E) on 12 October 2011. After dilution by a factor 10^4 with sterile physiological water of the activated sludge sample, the biomass was first cultivated on solid MSV peptone medium [5] at 20°C and in a glovebox with anoxic condition (less than 100 ppm oxygen). Single colonies were re-plated iteratively until obtaining a pure culture that was cryopreserved (aerobically) in 10% glycerol at -80°C.

As a facultative aerobe, *Zoogloea schifflangensis* LCSB751 can grow aerobically at 25°C in the liquid growth media R2A [6], MSV A+B [5] or Slijkhuis A [7] with 70 rpm agitation in which it forms cell clumps. More cell clumps were observed in MSV A+B than in R2A medium. When grown on R2A agar or on MSV peptone agar at 25°C under aerobe conditions, *Zoogloea schifflangensis* LCSB751 colonies were initially punctiform and after three days, they were white, circular, raised and with entire edges. The morphology of cells derived from these growth conditions are short rod-shaped bacteria (Figure 1A), Gram-negative in accordance with already described isolates species of *Zoogloea* spp. [8,9] (Table 1). Phylogenetic analysis based on 16S rRNA gene sequences confirms that strain LCB751 belong to the *Zoogloea* genus of the beta-proteobacterial class (Table 1). However, this strain formed a distinct phyletic linage from the five recognized species of *Zoogloea*, that are represented by the type strains *Z. caeni* EMB43^T [10], *Z. oleivorans* Buc^T [8], *Z. oryzea* A-7^T [11], *Z. ramigera* ATCC 19544^T [12] and *Z. resiniphila* DhA-35^T [13,14] (Figure 2).

Extended feature descriptions (optional Heading 3)

The capacity of *Zoogloea schifflangensis* LCSB751 to accumulate intracellular granules of lipids was tested using the dye Nile Red as described by Roume, Heintz-Buschart and collaborators [15]. Figure 1B shows the Nile Red positive phenotype of the described strain.

Additionally, the growth characteristics of the strain *Zoogloea schifflangensis* LCSB751 were determined aerobically and at 25°C with 700 rpm in 3 different liquid media. Its generation time was the longest in Slijkhuis A with the highest biomass production. MSV A+B allowed a generation time of 4 hours 30 minutes but lead to a poor biomass production as demonstrated by the low maximal optical density at 600 nm (OD_{600}) of 0.21. The tested liquid medium which allowed the most rapid growth for *Zoogloea schifflangensis* LCSB751 was R2A while the biomass production was close to the one observed for Slijkhuis A (Table 2).

Genome sequencing information (Heading 1)

Genome project history (Heading 2)

Overall, 140 pure bacterial cultures were obtained and screened for lipid inclusions using the Nile Red fluorescent dye and the genomes of 85 Nile Red-positive isolates were sequenced, of which isolate LCSB065 has already been published [15]. In particular, the genome of *Zoogloea schifflangensis* LCSB751 was sequenced to obtain information about the functional potential of this genus, which has no sequenced representative genome publically available, but also based on its particular phylogenetic position and to acquire knowledge on the genes related to lipid accumulation. The permanent draft genome sequence of this strain is available on NCBI with the accession number XXX. Table 3 summarizes the project information according to the MIGS compliance [16].

Growth conditions and genomic DNA preparation (Heading 2)

Zoogloea schifflangensis LCSB751 was grown on MSV peptone agar medium [5] at 20°C in anoxic condition. Half of the biomass was scrapped in order to cryopreserve the strain, while the second half was used for DNA extraction using the Power Soil DNA isolation kit (MO BIO, Carlsbad, CA, USA). This cryostock was used to distribute the strain to microorganism collection center (LMG 29444).

Genome sequencing and assembly (Heading 2)

The purified DNA was sequenced on an Illumina HiSeq Genome Analyzer IIx as described before for an other isolate of the same project [15]. Briefly, a paired-end sequencing library with a theoretical insert size of 300 bp was prepared with the AMPure XP/Size Select Buffer Protocol as previously described by Kozarewa & Turner [17], modified to allow for size-selection of fragments using the double solid phase reversible immobilisation procedure described earlier [18] and sequenced on an Illumina HiSeq with a read length of 100 bp at TGen North (AZ, USA). The leading end sequence of the resulting raw 2,638,115 paired-end reads was trimmed of N bases and filtered for a minimal quality score of 3, retaining 2,508,729 (~95%) of paired reads, 129,378 of forward read singletons and 8 (0.00%) of reverse read singletons. All reads retained after the pre-processing underwent a *de novo* assembly using SPAdes [19].

The total number of contigs (776), the mean contig length (7,497 bp) and the N50 value (180,423 bp) of the draft assembly of *Zoogloea schifflangensis* LCSB751 (Table 3) indicate quite a fragmented assembly despite the reasonable sequencing depth estimated at 150x fold coverage, 100x based on the counting 21-mer sequences, using KMC2 [20] and evaluated at approximately 120x average depth of coverage upon mapping reads back onto the contigs generated from the assembly [21–23].

Genome annotation (Heading 2)

As a part of RAST pipeline [24], ORFs encoding protein, tRNA and rRNA were predicted using GLIMMER2 [25], tRNAscan-SE [26] and "search_for_rnas" respectively, and were then annotated using the subsystems approach. The proteins predicted from RAST were submitted to WebMGA server [27], to SignalP server v.4.1 [28] and to TMHMM server v.2.0 [29] for COG functional annotation, signal peptides prediction and transmembrane helices prediction, respectively. 4202 of the predicted amino acid sequences were annotated with 13,030 Pfam IDs. Additionally, CRISPR loci were detected using metaCRT [30,31].

Genome Properties (Heading 1)

The draft genome assembly of *Zoogloea schifflangensis* LCSB751 consists of 5,817,831 bp with a G+C content of 64.2%, distributed among 776 contigs with an N50 value of 180,423 bp. Detailed statistics are provided in Table 4. The raw reads are available via GenBank nucleotide database under the accession number XXXX, while the assembly and the annotation (IDs 6666666.102999) can be access through the RAST guest account at http://rast.nmpdr.org.

Of the 5202 predicted genes, 77 are annotated as RNAs. The rRNA operon region is predicted to be occurring in multiple copies, because all reads from this region were assembled into a single contig with a higher depth of coverage (~1200x) compared to the rest of the genome. tRNAs anticodons covered all the 20 regular amino-acids. Less than 30% (1464) of the CDS were annotated as encoding hypothetical proteins or proteins of unknown function. The COG functional categories distribution is in Table 5. Additional functional classification based on subsystem is available via RAST.

Insights from the genome sequence (optional) (Heading l)

Central metabolism inferences from genomic information (Heading 2)

Regarding carbon central cycles, *Zoogloea schifflangensis* LCSB751 is predicted to have gene for a complete TCA cycle, but is missing some or the complete set of genes for the EMP pathway, the pentose phosphate pathway and Entner-Doudoroff pathway.

A periplasmic nitrate reductase as well as a nitrite reductase are found, suggesting the possible complete reduction of nitrate in ammonia by *Zoogloea schifflangensis* LCSB751. Furthermore, a complete set of *nif* genes involved in nitrogen fixation is present in this genome.

Genes for a complete electron transport chain are found as well as for the alternative RNF complex [32].

The genome of *Zoogloea schifflangensis* LCSB751 also encodes numerous genes for flagella synthesis and assembly, suggesting the motility of this strain. The strain is also predicted to be prototroph for all amino acids, nucleotides and vitamins B_2, B_6, B_9, H, and is missing a single gene for the synthesis of B_{12}.

Additionally, the catechol 2,3-dioxygenase that has been studied in *Z. oleivorans*, is also found in *Zoogloea schifflangensis* LCSB751.

Finally, three CRISPR loci were detected, accompanied by eight CRISPR-associated proteins. Two of these direct repeats are 37bp in length (sequence: GTTTCAATCCACGTCCGTTATTGCTAACGGACGAATC; GTGGCACTCGCTCCGAAGGGAGCGACTTCGTTGAAGC) while one of them is 32bp (sequence: CACTCGCTCCGGAGGGAGCGACTTCGTTGAAG). These CRISPRs contain 175, 51 and 11 spacers, respectively ranging from lengths of 33 to 46 bp. A total of 77 matches were found when blasting the spacers against the ACLAME phage/viral/plasmid gene database, NCBI phage and NCBI virus databases.

51 of the spacers match to phages, 6 to viruses, 11 to genes within plasmids and six to genes within phages/prophages.

Lipid metabolism (Heading 2)

In order to better understand the lipid accumulation phenotype of *Zoogloea* spp., the genome of *Zoogloea schifflangensis* LCSB751 was further analysed with special focus on genes related to lipid metabolism. With 202 genes annotated with a COG functional category I “Lipid transport and metabolism “, more that 3.8% of the genome of *Zoogloea schifflangensis* LCSB751 is devoted to lipid metabolism (Table 5). Using the SEED subsystem feature, similar results were obtained with 194 genes (3.8%) classified in the “Fatty acids, lipids and Isoprenoids” subsystem (details in Table 6).

In details, a complete set of genes necessary for the synthesis, polymerisation and depolymerisation of PHA [2] was found as well as the gene of the MEP/DOXP pathway for terpenoid synthesis. The gene necessary to convert diacylglycerol in triacylglycerol or fatty alcohol in wax ester has not been found, suggesting that the only lipid bodies that are accumulated in *Zoogloea schifflangensis* LCSB751 are PHA granules.

In situ gene expression (Heading 2)

While genomics data provides information about the genetic potential of *Zoogloea schifflangensis* LCSB751, it is possible to study expressed functions of the natural population of *Zoogloea* in the biological treatment plant it has been isolated from using metatranscriptomic data. Here, such analyses has been done on four temporally resolved samples, collected on the 25 January 2011, 11 January 2012, 5 October 2011, and 12 October 2011, as studied by Muller and collaborators [33]. Genes with an average depth of coverage equal or higher than 0.3, were considered as expressed by mapping the rRNA-depleted transcripts on the genome of *Zoogloea schifflangensis* LCSB751. 259, 312, 269 and 330 genes, respectively, were detected as expressed, with 160 of them being always expressed. For the vast majority, (4732 genes), no transcripts were detected, which can be explained by *Zoogloea* sp. low population size *in situ*. Indeed, by phylogenetic marker gene (16S rRNA) amplicon sequencing on the sample collected the 25 January 2011 (data from [33]), *Zoogloea* sp. population size was estimated at 0.1%.

Interestingly, at least one copy of the acetoacetyl-CoA reductase and of the polyhydroxyalkanoic acid synthase are found expressed at each time point, suggesting an important PHA accumulation in the population of *Zoogloea* sp. in this environment.

Conclusions (Heading 1)

We provide here the first draft genome of a strain belonging to the genus *Zoogloea*. The genetic inventory of *Zoogloea schifflangensis* LCSB751 makes it of particular interest for future wastewater treatment strategies based around the comprehensive reclamation of nutrients and chemical energy-rich biomolecules around the concept of a "wastewater biorefinery column" [3] as well as for industrial biotechnological application. Future comparative genomics studies would allow the scientific community to identify if this genomic repertoire is typical of this genus. Using metatranscriptomic data, we show that *Zoogloea* sp. population is active in the studied wastewater treatment plant despite its small size, and in particular that PHA accumulation occurs *in situ*.

Taxonomic and nomenclatural proposals (optional) (Heading 1).

Based on the phylogenetic analysis, we formally suggest the creation of *Zoogloea schifflangensis* sp. nov. with strain LCSB751 being the type strain.

Zoogloea Schifflangensis (Schif.flan.gen.sis gen. nov. Schifflangensis, based on the name of the city the strain has been isolated from). This type strain LCSB751 (=LMG 29444) has been obtained from foaming activated sludge of the municipal treatment plant of the city Schifflange. Growth was observed both aerobically and anaerobically in rich medium. Colonies are white, circular, raised and with entire edges. Cells are gram-negative, rod-shaped and accumulated lipid granules.

The G+C content of the genome is 64.2%. The genome sequence was deposited in GenBank under accession number XXX, and its annotation is available via RAST (IDs 6666666.102999) through the RAST guest account at http://rast.nmpdr.org, using ‘guest’ as login as well as password.

Authors' contributions

EELM and LAL isolated the strain, LAL prepared the DNA, NDH prepared the library and sequenced it, SN, MZ and EELM performed the bioinformatics analyses and EELM and PW designed and coordinated the project. All authors read and approved the final manuscript.

Acknowledgements (heading 1 style)

The authors thank Mr Bissen and Mr Di Pentima from the Syndicat Intercommunal à Vocation Ecologique (SIVEC) for their permission to collect samples and gain access to the monitoring platform of the Schifflange wastewater treatment plant; Dr Aidos Baumuratov for his support for microscopic analyses of the isolates; and the high performance computing facilities of the University of Luxembourg. This work was supported by an ATTRACT programme grant (ATTRACT/A09/03) and a European Union Joint Programming in Neurodegenerative Diseases grant (INTER/JPND/12/01) both to PW, and Aide à la Formation Recherche (AFR) grants to EELM (PRD-2011-1/SR) and SN (PHD-2014-1/7934898), all funded by the Luxembourg National Research Fund (FNR).

References (heading 1 style)

1. Dugan PR, Stoner DL, Pickrum HM. The Genus *Zoogloea*. Prokaryotes Vol 7 Proteobacteria Delta Epsil. Subclasses Deep. Rooting Bact. Springer Science & Business Media; 2006. p. 1105.

2. Muller EEL, Sheik AR, Wilmes P. Lipid-based biofuel production from wastewater. Curr. Opin. Biotechnol. 2014;30:9–16.

3. Sheik AR, Muller EEL, Wilmes P. A hundred years of activated sludge: time for a rethink. Front. Microbiol. 2014;5:47.

4. Sağ Y, Kutsal T. Biosorption of heavy metals by *Zoogloea ramigera*: use of adsorption isotherms and a comparison of biosorption characteristics. Chem. Eng. J. Biochem. Eng. J. 1995;60:181–8.

5. Levantesi C, Rossetti S, Thelen K, Kragelund C, Krooneman J, Eikelboom D, et al. Phylogeny, physiology and distribution of "*Candidatus* Microthrix calida", a new *Microthrix* species isolated from industrial activated sludge wastewater treatment plants. Environ. Microbiol. 2006;8:1552–63.

6. Reasoner DJ, Geldreich EE. A new medium for the enumeration and subculture of bacteria from potable water. Appl. Environ. Microbiol. 1985;49:1–7.

7. Slijkhuis H. *Microthrix parvicella*, a filamentous bacterium isolated from activated sludge: cultivation in a chemically defined medium. Appl. Environ. Microbiol. 1983;46:832–9.

8. Farkas M, Táncsics A, Kriszt B, Benedek T, Tóth EM, Kéki Z, et al. *Zoogloea oleivorans* sp. nov., a floc-forming, petroleum hydrocarbon-degrading bacterium isolated from biofilm. Int. J. Syst. Evol. Microbiol. 2015;65:274–9.

9. Huang T-L, Zhou S-L, Zhang H-H, Bai S-Y, He X-X, Yang X. Nitrogen removal characteristics of a newly isolated indigenous aerobic denitrifier from oligotrophic drinking water reservoir, *Zoogloea* sp. N299. Int. J. Mol. Sci. 2015;16:10038–60.

10. Shao Y, Chung BS, Lee SS, Park W, Lee S-S, Jeon CO. *Zoogloea caeni* sp. nov., a floc-forming bacterium isolated from activated sludge. Int. J. Syst. Evol. Microbiol. 2009;59:526–30.

11. Xie C-H, Yokota A. *Zoogloea oryzae* sp. nov., a nitrogen-fixing bacterium isolated from rice paddy soil, and reclassification of the strain ATCC 19623 as *Crabtreella saccharophila* gen. nov., sp. nov. Int. J. Syst. Evol. Microbiol. 2006;56:619–24.

12. Unz R. Neotype strain of *Zoogloea ramigera* Itzigsohn. Int J Syst Bacteriol. 1971;21:91–9.

13. Mohn WW, Wilson AE, Bicho P, Moore ER. Physiological and phylogenetic diversity of bacteria growing on resin acids. Syst. Appl. Microbiol. 1999;22:68–78.

14. unknown. Validation of the publication of new names and new combinations previously effectively published outside the IJSB. List No. 70. Int J Syst Bacteriol. 1999;49:935–6.

15. Roume H, Heintz-Buschart A, Muller EEL, May P, Satagopam VP, Laczny CC, et al. Comparative integrated omics: identification of key functionalities in microbial community-wide metabolic networks. Npj Biofilms Microbiomes. 2015;1:15007.

16. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, et al. The minimum information about a genome sequence (MIGS) specification. Nat. Biotechnol. 2008;26:541–7.

17. Kozarewa I, Turner DJ. 96-plex molecular barcoding for the Illumina Genome Analyzer. Methods Mol. Biol. Clifton NJ. 2011;733:279–98.

18. Rodrigue S, Materna AC, Timberlake SC, Blackburn MC, Malmstrom RR, Alm EJ, et al. Unlocking short read sequencing for metagenomics. PloS One. 2010;5:e11840.

19. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. J. Comput. Mol. Cell Biol. 2012;19:455–77.

20. Deorowicz S, Kokot M, Grabowski S, Debudaj-Grabysz A. KMC 2: fast and resource-frugal k-mer counting. Bioinformatics. 2015;31:1569–76.

21. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. 2009;25:2078–9.

22. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics. 2010;26:589–95.

23. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

24. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008;9:75.

25. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 1999;27:4636–41.

26. Lowe TM, Eddy SR. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25:0955–64.

27. Wu S, Zhu Z, Fu L, Niu B, Li W. WebMGA: a customizable web server for fast metagenomic sequence analysis. BMC Genomics. 2011;12:444.

28. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods. 2011;8:785–6.

29. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 2001;305:567–80.

30. CRISPR-related software [Internet]. [cited 2016 Aug 10]. Available from: http://omics.informatics.indiana.edu/CRISPR/

31. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics. 2007;8:209.

32. Biegel E, Schmidt S, González JM, Müller V. Biochemistry, evolution and physiological function of the Rnf complex, a novel ion-motive electron transport complex in prokaryotes. Cell. Mol. Life Sci. CMLS. 2011;68:613–34.

33. Muller EEL, Pinel N, Laczny CC, Hoopmann MR, Narayanasamy S, Lebrun LA, et al. Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. Nat. Commun. 2014;5:5603.

34. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. U. S. A. 1990;87:4576–9.

35. Garrity GM, Bell JA, Lilburn T. Phylum XIV. Proteobacteria phyl. nov. In: Brenner DJ, Krieg NR, Staley JT, editors. Bergey's Manual® Syst. Bacteriol. Springer US; 2005.

36. Garrity GM, Bell JA, Lilburn T. Class II. Betaproteobacteria class. nov. In: Brenner DJ, Krieg NR, Staley JT, editors. Bergey's Manual® Syst. Bacteriol. [Internet]. Springer US; 2005 [cited 2015 Oct 13]. p. 575–922. Available from: http://link.springer.com/chapter/10.1007/978-0-387-29298-4_2

37. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 2000;25:25–9.

38. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008;36:W465-469.

Tables and figures

Table 1. Classification and general features of *Zoogloea schifflangensis* strain LCSB751 according to the MIGS recommendation [16]

MIGS ID	**Property**	**Term**	**Evidence code**[a]
	Classification	Domain *Bacteria*	TAS [34]
		Phylum *Proteobacterium*	TAS [35]
		Class *Betaproteobacterium*	TAS [36]
		Order *Rhodocyclales*	TAS [10]
		Family *Rhodocyclaceae*	TAS [10]
		Genus *Zoogloea*	IDA
		Species *schifflengensis*	IDA
		Strain: LCSB0751	
	Gram stain	Negative	TAS [1]
	Cell shape	Rod	TAS [1]
	Motility	Motile	TAS [1]
	Sporulation	Not reported	NAS
	Temperature range	5-40°C	TAS [8,10,11]
	Optimum temperature	25-30°C	TAS [8,10]
	pH range; Optimum	6.0–9.0; 6.5-7.5	TAS [8,10]
MIGS-6	Habitat	Activated sludge	IDA
MIGS-6.3	Salinity	Inhibited at 0.5% NaCl (w/v)	TAS [11]
MIGS-22	Oxygen requirement	facultative anaerobe	IDA
MIGS-15	Biotic relationship	free-living	IDA
MIGS-14	Pathogenicity	non-pathogen	NAS
MIGS-4	Geographic location	Luxembourg	IDA
MIGS-5	Sample collection	2011	IDA
MIGS-4.1	Latitude	49°30′48.29″N;	IDA
MIGS-4.2	Longitude	6°1′4.53″E	IDA
MIGS-4.4	Altitude	275 m	IDA

[a] Evidence codes - IDA: Inferred from Direct Assay; TAS: Traceable Author Statement (i.e., a direct report exists in the literature); NAS: Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from the Gene Ontology project [37].

Table 2. Generation time, growth rate and maximal growth of *Zoogloea schifflangensis* LCSB751 under different aerobic culture conditions.

Medium	**Generation time ± standard deviation**[a]	**Growth rate (min^{-1})**	**Maximal OD_{600}**[b]
R2A	1h54min ± 3min	0.0058	0.46
MSV A+B	4h30min ± 53min	0.0026	0.21
Slijkhuis A	10h42min ± 1h51min	0.0011	0.73

[a] the values are an average of independent triplicate experiments

[b] OD_{600} stands for optical density measured at 600 nm with the spectrometer “Biochrom WPA CO 8000 Cell Density Meter” using BRAND disposable semi-micro UV cuvettes of 12.5 x 12.5 x 45 mm.

Table 3. Project information.

MIGS ID	**Property**	**Term**
MIGS 31	Finishing quality	Draft
MIGS-28	Libraries used	Illumina paired-end reads (insert size 30 bp)
MIGS 29	Sequencing platforms	Illumina HiSeq
MIGS 31.2	Fold coverage	150x
MIGS 30	Assemblers	SPAdes (version 3.1.1)
MIGS 32	Gene calling method	RAST
	Locus Tag	fig|6666666.102999
	Genbank ID	
	GenBank Date of Release	
	GOLD ID	
	BIOPROJECT	
MIGS 13	Source Material Identifier	LMG 29444
	Project relevance	Environmental, biodiversity, biotechnological

Table 4. Genome statistics of *Zoogloea schifflengensis* LCSB751.

Attribute	**Value**	**% of Total**[a]
Genome size (bp)	5,817,831	100.00
DNA coding (bp)[b]	4,966,077	85.36
DNA G+C (bp)	3,733,728	64.18
DNA scaffolds	773	100.00
Total genes	5,202	100.00
Protein coding genes	5,125	98.52
RNA genes	77	1.48
Pseudo genes	unknown	unknown
Genes in internal clusters	unknown	unknown
Genes with function prediction	3,661	70.38
Genes assigned to COGs	4,191	80.56
Genes with Pfam domains	4,202	80.78
Genes with signal peptides	505	9.71
Genes with transmembrane helices	1157	22.24
CRISPR repeats	2	2.85

[a] The total is based on either the size of the genome in base pairs, total number of scaffolds or the total number of genes in the annotated genome.

[b] The cumulative length of genes, without considering overlaps.

Table 5. Number of genes associated with general COG functional categories.

Code	**Value**	**%age**	**Description**
J	182	3.50	Translation, ribosomal structure and biogenesis
A	3	0.06	RNA processing and modification
K	342	6.57	Transcription
L	204	3.92	Replication, recombination and repair
B	3	0.06	Chromatin structure and dynamics
D	52	1.00	Cell cycle control, Cell division, chromosome partitioning
V	69	1.33	Defense mechanisms
T	564	10.84	Signal transduction mechanisms
M	252	4.84	Cell wall/membrane biogenesis
N	177	3.40	Cell motility
U	142	2.73	Intracellular trafficking and secretion
O	189	3.63	Posttranslational modification, protein turnover, chaperones
C	362	6.96	Energy production and conversion
G	130	2.50	Carbohydrate transport and metabolism
E	305	5.86	Amino acid transport and metabolism
F	85	1.63	Nucleotide transport and metabolism
H	185	3.56	Coenzyme transport and metabolism
I	202	3.88	Lipid transport and metabolism
P	283	5.44	Inorganic ion transport and metabolism
Q	126	2.42	Secondary metabolites biosynthesis, transport and catabolism
R	520	10.00	General function prediction only
S	351	6.75	Function unknown
-	1,011	19.43	Not in COGs

The total is based on the total number of protein coding genes in the genome.

Table 6. Gene abundance and frequency related to the lipid metabolism of *Zoogloea schifflengensis* LCSB751. The different categories (in **bold**) and subcategories of the subsystem “Fatty acids, lipids and isoprenoid” are represented

Subsystem	**Subsystem feature count**	**Subsystem feature (%)**
Fatty acids, lipids and isoprenoids	**194**	**100**
Phospholipids	**30**	**15.46**
Cardiolipin synthesis	2	6.67
Glycerolipid and glycerophospholipid metabolism in bacteria	28	93.33
Triacylglycerols	**3**	**1.55**
Triacylglycerol metabolism	3	100
Fatty acids	**71**	**36.60**
Fatty acid biosynthesis FASII	30	42.25
Fatty acid metabolism cluster	41	57.75
Fatty acids, lipids and isoprenoids - no subcategory	**56**	**28.87**
Polyhydroxybutyrate metabolism	56	100
Isoprenoids	**34**	**17.53**
Isoprenoids for quinones	5	14.71
Isoprenoid biosynthesis	18	52.94
Polyprenyl diphosphate biosynthesis	4	11.76
Nonmevalonate branch of isoprenoid Biosynthesis	7	20.59

Figure legends

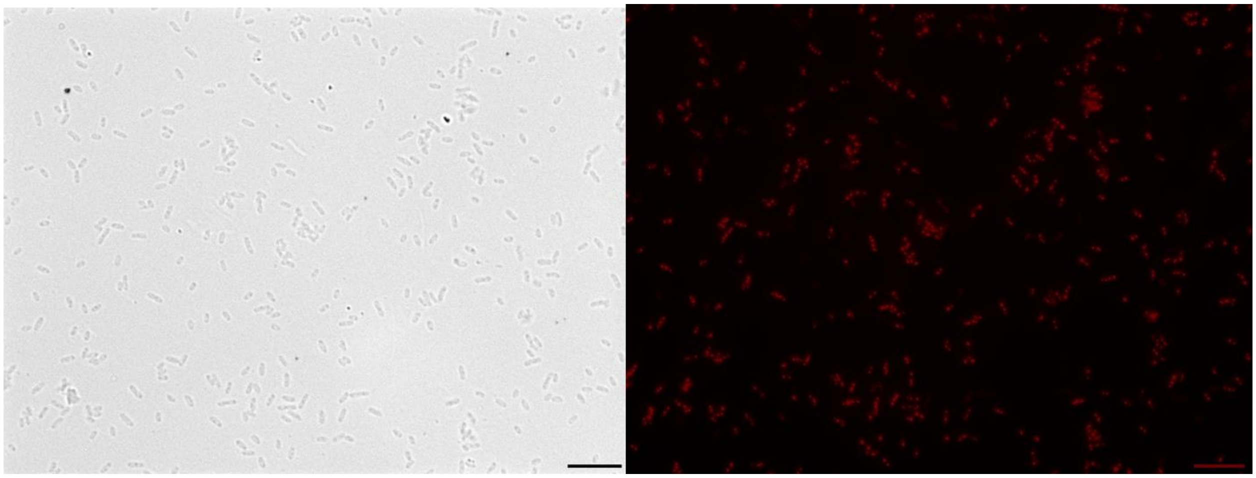


Figure 1: Photomicrograph of *Zoogloea schifflangensis* strain LCSB751. The cells were grown anaerobically at 20°C on plate with MSV Peptone medium and Nile Red stained after heat fixation. The image was taken using an inverted microscope (Nikon Ti) equipped with a 60× oil immersion Nikon Apo-Plan lambda objective (1.4N.A) and an intermediate magnification of 1.5x. The scale represents 10 µm. All imaging data were collected and analysed using the OptoMorph (Cairn Research, Kent, UK) and ImageJ. A: breigh field; B: same field observed in epifluorescence using an excitation light from a Xenon arc lamp. The beam was passed through an Optoscan monochromator (Cairn Research, Kent, UK) with 550/20nm selected band pass. Emitted light was reflected through a 620/60nm bandpass filter with a 565 dichroic connected to a cooled CCD camera (QImaging, Exi Blue).

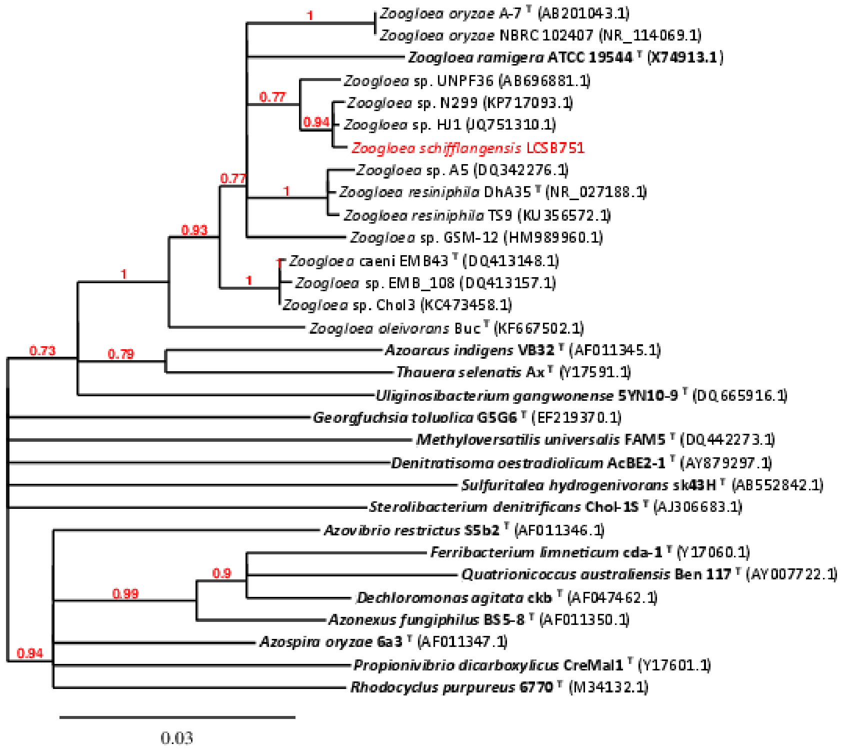

1
Zoogloea oryzae A-7 ^{T} (AB201043.1)
Zoogloea oryzae NBRC 102407 (NR_114069.1)
Zoogloea ramigera ATCC 19544 ^{T} (X74913.1)
0.77
Zoogloea sp. UNPF36 (AB696881.1)
Zoogloea sp. N299 (KP717093.1)
0.94
Zoogloea sp. HJ1 (JQ751310.1)
Zoogloea schifflangensis LCSB751
0.77
Zoogloea sp. A5 (DQ342276.1)
1
Zoogloea resiniphila DhA35 ^{T} (NR_027188.1)
Zoogloea resiniphila TS9 (KU356572.1)
0.93
Zoogloea sp. GSM-12 (HM989960.1)
1
Zoogloea caeni EMB43 ^{T} (DQ413148.1)
1
Zoogloea sp. EMB_108 (DQ413157.1)
Zoogloea sp. Chol3 (KC473458.1)
Zoogloea oleivorans Buc ^{T} (KF667502.1)
0.73
0.79
Azoarcus indigens VB32 ^{T} (AF011345.1)
Thauera selenatis Ax ^{T} (Y17591.1)
Uliginosibacterium gangwonense 5YN10-9 ^{T} (DQ665916.1)
Georgfuchsia toluolica G5G6 ^{T} (EF219370.1)
Methyloversatilis universalis FAM5 ^{T} (DQ442273.1)
Denitratisoma oestradiolicum AcBE2-1 ^{T} (AY879297.1)
Sulfuritalea hydrogenivorans sk43H ^{T} (AB552842.1)
Sterolibacterium denitrificans Chol-1S ^{T} (AJ306683.1)
Azovibrio restrictus S5b2 ^{T} (AF011346.1)
Ferribacterium limneticum cda-1 ^{T} (Y17060.1)
0.9
Quatrionicoccus australiensis Ben 117 ^{T} (AY007722.1)
0.99
Dechloromonas agitata ckb ^{T} (AF047462.1)
Azonexus fungiphilus BS5-8 ^{T} (AF011350.1)
Azospira oryzae 6a3 ^{T} (AF011347.1)
0.94
Propionivibrio dicarboxylicus CreMal1 ^{T} (Y17601.1)
Rhodocyclus purpureus 6770 ^{T} (M34132.1)
0.03

Figure 2: Phylogenetic tree based on 16S rRNA gene sequences. The type species strains of every species of the Rhodocyclaceae family were used as well as all the type strains of the genus *Zoogloea*, according to the List of prokaryotic names with sanding in nomenclature (LPSN; http://www.bacterio.net). The 16S sequences were aligned using CustalW, the alignment was curated using Gblocks conserving 81% of the initial positions and the phylogeny was computed with BioNJ using 100 bootstraps and the default (K2P) substitution model, using the pipeline Phylogeny.fr [38]. GenBank IDs of used whole genome sequences in order from top to bottom: xxx, xxx, xxx

A.8 IMP: a pipeline for reproducible metagenomic and metatranscriptomic analyses.

Shaman Narayanasamy*, Yohan Jarosz*, Emilie E.L. Muller, Cédric C. Laczny, Malte Herold, Anne Kaysen, Anna Heintz-Buschart, Nicolàs Pinel, Patrick May, Paul Wilmes

2016
BioRxiv
DOI: http://dx.doi.org/10.1101/039263

This article was a non-peer reviewed early release prior to submission within a scientific journal. Please refer to **Appendix A.2** or **Chapter 2** for updated versions.

APPENDIX B

ADDITIONAL FIGURES

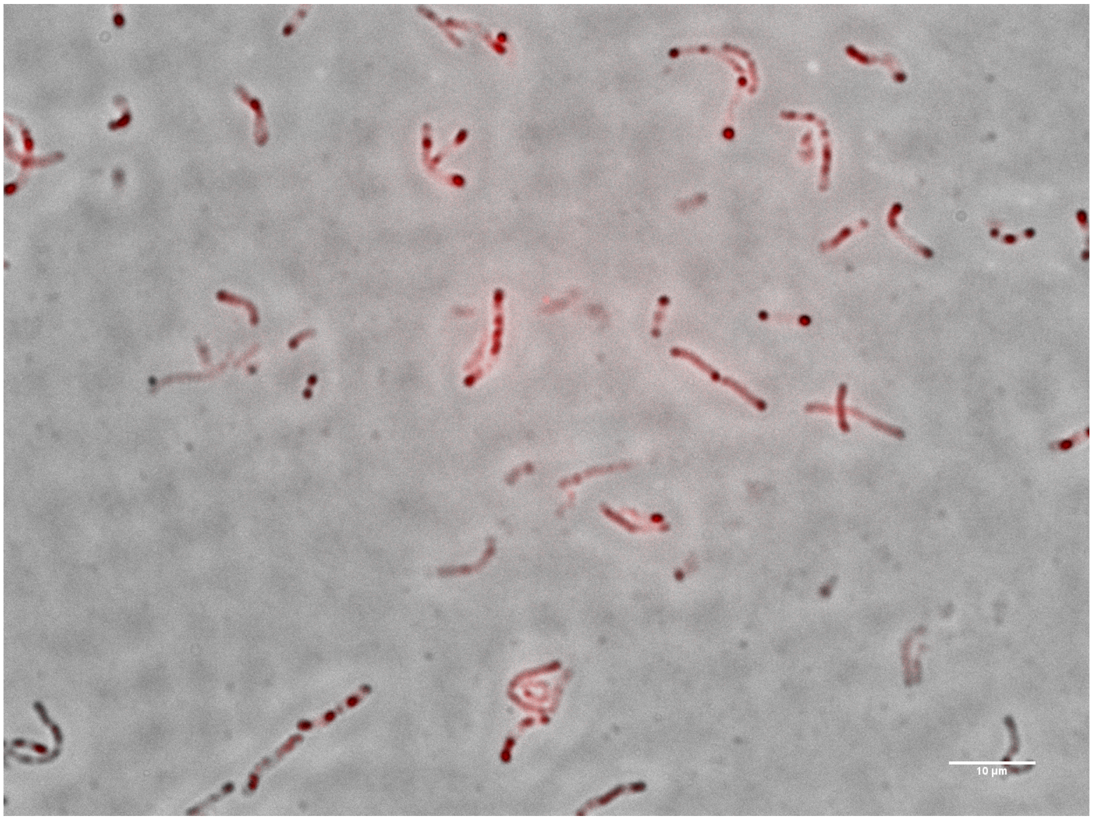


Figure B.1: Microscopy photo of bacterial strain LCSB005. The red dye stains lipids

APPENDIX C

ADDITIONAL TABLES

Table C.1: IMP time series analyses summary

Date	Sample ID	MG unprocessed	MG retained pairs	MG retained singles	MT unprocessed	MT retained pairs	MT retained singles	MT rRNA pairs	MT rRNA singles	MT contigs	Co-assembly contigs	Interval from previous sample (days)
2010-10-04	A01	17,531,630	16,368,148	1,078,317	56,764,165	46,430,755	1,833,300	8,162,922	211,837	57,491	126,390	1
2011-01-25	A02	14,699,402	13,108,258	1,199,766	16,493,393	9,119,940	1,951,440	4,110,417	775,696	47,726	208,345	113
2011-03-21	D37	35,999,747	32,911,158	2,199,347	35,975,301	13,400,496	2,315,369	17,642,665	2,323,133	61,242	431,078	55
2011-03-29	D02	34,135,175	27,721,744	2,837,683	33,912,215	13,612,603	1,826,984	16,504,703	1,719,392	81,321	443,307	8
2011-04-05	D39	28,908,243	26,605,828	1,673,538	38,592,454	17,367,979	2,504,144	16,717,068	1,685,117	118,913	479,608	7
2011-04-14	D24	29,187,268	27,401,321	1,354,513	36,395,674	18,174,395	2,162,663	14,460,847	1,445,338	130,962	517,506	9
2011-04-21	D25	25,320,222	20,487,745	4,083,256	33,954,127	13,847,740	1,736,581	16,687,126	1,534,912	91,152	331,870	7
2011-04-29	D35	30,371,838	27,587,583	1,797,893	29,681,622	12,018,746	1,441,649	14,775,098	1,342,298	73,557	377,432	8
2011-05-06	D46	28,122,705	25,044,633	2,528,946	34,116,436	13,596,582	2,102,990	15,704,281	1,559,060	86,116	389,063	7
2011-05-13	D13	31,173,943	29,336,922	1,467,456	35,894,482	12,307,140	1,424,327	20,133,094	1,889,977	70,840	478,799	7
2011-05-20	D40	25,651,909	24,085,668	1,018,577	27,707,752	9,331,432	1,515,442	14,726,480	1,669,177	44,671	326,463	7
2011-05-27	D33	30,235,782	28,307,886	1,506,824	27,369,514	10,155,882	1,139,082	14,561,598	1,374,840	72,686	401,437	7
2011-06-03	D06	30,216,637	26,954,409	1,576,187	32,430,876	12,235,924	1,998,252	16,026,593	1,959,118	89,444	396,383	7
2011-06-09	D27	27,497,422	25,777,061	1,306,125	30,331,588	11,628,944	1,328,863	15,895,274	1,332,740	94,475	413,711	6
2011-06-17	D42	27,303,250	25,503,950	1,313,167	34,544,359	14,363,643	1,328,939	17,320,121	1,449,468	117,872	468,428	8
2011-06-24	D03	32,347,395	29,220,419	1,797,971	32,699,565	18,755,346	2,503,880	9,988,214	944,102	145,997	553,152	7
2011-07-01	D47	30,109,054	27,730,630	1,936,737	36,766,788	19,790,420	2,959,780	12,327,661	1,279,728	132,644	494,971	7
2011-07-08	D18	23,391,679	22,164,878	1,066,058	31,924,022	13,802,753	1,538,828	15,093,605	1,388,116	107,211	404,870	7
2011-08-05	D45	29,930,665	27,829,830	1,644,918	35,378,306	15,997,138	2,372,196	14,915,695	1,347,833	103,844	517,913	28
2011-08-11	D51	29,852,842	26,901,508	2,115,968	44,174,409	18,564,123	2,671,740	20,249,314	2,173,170	108,969	490,163	6
2011-08-19	D43	26,075,473	24,281,057	1,444,335	27,171,859	14,640,598	1,251,988	10,516,375	682,408	119,962	438,208	8
2011-08-29	D34	30,403,109	28,058,664	2,030,863	28,421,236	14,285,422	1,622,021	11,327,996	1,029,180	115,348	425,668	10
2011-09-05	D30	30,532,662	28,718,715	1,490,886	29,515,165	15,646,084	2,024,810	10,561,939	1,116,334	128,807	482,611	7
2011-09-12	D04	29,014,224	25,778,552	1,730,152	35,007,981	19,311,266	2,775,618	11,158,731	1,232,688	154,729	500,533	7
2011-09-19	D31	25,456,037	23,744,583	1,302,766	30,499,924	18,342,790	1,783,389	9,426,018	755,439	144,450	448,666	7
2011-09-28	D29	29,839,175	27,199,081	2,226,114	33,708,983	17,931,937	2,782,354	10,936,159	1,380,297	133,424	504,283	9
2011-10-05	D49	29,042,305	26,641,829	1,400,099	29,639,609	14,248,134	1,675,031	12,356,024	1,169,930	104,818	443,476	7
2011-10-12	D32	28,460,125	26,905,139	1,280,844	31,894,222	17,448,520	1,520,300	11,887,316	902,949	149,581	538,414	7
2011-11-02	D23	23,448,618	18,851,501	3,923,449	32,127,592	18,036,232	2,631,682	9,851,426	1,340,869	134,971	353,499	21
2011-11-07	D44	27,963,482	24,897,807	2,376,671	23,803,138	16,386,146	1,843,356	4,949,382	514,343	113,204	415,829	5
2011-11-16	D20	32,139,021	30,421,843	1,283,367	33,469,570	28,141,065	1,936,802	3,111,109	208,282	137,445	466,269	9
2011-11-23	D15	26,973,299	24,536,053	1,365,590	35,418,302	26,585,849	4,005,009	4,108,201	572,155	114,025	338,261	7
2011-11-29	D05	25,621,580	22,949,103	773,926	32,499,996	25,199,023	3,264,871	3,342,329	446,283	117,807	334,496	6
2011-12-21	D28	29,329,592	27,574,163	1,289,080	34,336,596	20,937,720	2,097,177	10,285,053	900,638	111,373	442,661	22
2011-12-28	D09	31,138,147	25,898,148	3,088,153	33,883,001	17,654,965	2,767,161	11,810,528	1,504,504	112,721	399,510	7
2012-01-03	D38	24,493,806	22,681,193	991,272	23,571,029	12,051,161	1,421,055	9,055,794	858,754	84,128	339,691	6
2012-01-11	D36	34,928,806	32,014,419	2,069,892	37,853,692	16,360,403	3,039,172	15,527,661	2,517,756	114,618	482,520	8
2012-01-19	D12	26,256,782	24,765,672	1,183,841	32,078,032	11,415,158	1,142,644	17,950,311	1,459,721	87,563	417,909	8
2012-01-25	D10	32,576,299	28,822,117	2,112,865	35,356,100	12,792,706	2,064,782	18,080,293	2,224,365	98,151	403,369	6

Continued on next page...

Date	**Sample ID**	**MG unpro-cessed**	**MG retained pairs**	**MG retained singles**	**MT unpro-cessed**	**MT retained pairs**	**MT retained singles**	**MT rRNA pairs**	**MT rRNA singles**	**MT contigs**	**Co-assembly contigs**	**Interval from previous sample (days)**
2012-02-01	D14	26,995,145	22,067,691	2,719,767	33,583,347	12,404,420	1,860,511	17,209,617	1,986,249	91,357	354,881	7
2012-02-08	D08	33,462,713	29,590,993	2,191,955	32,414,295	11,974,617	1,443,747	17,301,018	1,546,768	94,461	501,333	7
2012-02-14	D26	29,242,844	27,118,676	1,565,421	31,735,326	9,954,805	1,106,915	19,045,314	1,542,360	57,897	423,528	6
2012-02-23	D21	24,350,474	22,754,752	1,033,309	34,301,246	11,506,763	1,627,570	19,117,272	1,858,032	86,515	375,114	9
2012-02-29	D19	32,574,065	28,288,700	3,165,796	36,394,504	12,496,491	1,734,078	19,915,026	2,069,786	97,348	449,815	6
2012-03-08	D17	23,876,138	22,226,625	1,033,784	30,362,735	10,236,878	1,340,616	17,044,898	1,653,723	79,024	333,653	8
2012-03-14	D41	30,626,165	28,490,276	1,515,664	27,396,429	8,434,111	1,040,702	16,251,236	1,553,860	61,024	395,166	6
2012-03-22	D11	29,495,784	27,691,057	1,348,071	35,677,196	8,436,575	1,195,654	23,290,747	2,631,821	50,236	401,414	8
2012-03-28	D16	28,186,198	26,162,520	1,081,186	36,820,589	10,406,024	1,424,590	22,364,722	2,538,008	69,141	370,833	6
2012-04-04	D07	21,699,617	15,940,878	3,465,313	34,514,784	13,143,860	1,791,203	17,520,794	1,800,547	93,079	264,999	7
2012-04-10	D22	33,671,265	30,670,757	2,094,658	34,097,677	17,817,055	3,145,665	11,029,695	1,793,094	116,509	432,627	6
2012-04-17	D01	32,692,858	27,975,133	2,329,138	33,418,152	15,582,015	2,155,159	13,975,373	1,477,695	113,332	465,002	7
2012-04-25	D50	24,359,112	22,277,277	1,450,478	32,882,193	16,384,859	1,888,715	12,992,641	1,329,066	108,632	380,672	8
2012-05-03	D48	27,267,336	25,359,794	1,619,284	32,450,531	15,443,434	2,107,587	13,248,562	1,405,761	106,065	412,436	8
Total		1,504,179,064	1,362,404,347	95,481,236	1,751,412,079	826,139,067	104,168,383	732,552,336	75,408,717	5,338,878	21,988,235	578
Average		28,380,737	25,705,742	1,801,533	33,045,511	15,587,530	1,965,441	13,821,742	1,422,806	100,734	414,872	8.18
Standard deviation		4,025,168	3,847,016	745,508	5,388,215	6,065,706	637,401	4,743,139	560,733	27,534	79,092	16.31703022

Table C.2: Summary statistics of LAMP community CRISPR elements based different time points

Date	Sample ID	Repeats	Spacers	Flanks	Protospacers	Protospacer containing contigs
04 October 2010	A01	825	8,846	328	63,650	14,903
25 January 2011	A02	578	3,287	159	62,425	14,130
21 March 2011	D37	1,528	7,379	247	167,096	28,208
29 March 2011	D02	1,517	7,060	282	62,192	23,734
05 April 2011	D39	1,792	8,500	297	97,101	29,725
14 April 2011	D24	1,892	10,078	371	122,766	31,312
21 April 2011	D25	1,664	5,818	186	59,043	18,270
29 April 2011	D35	1,620	6,738	221	64,890	22,072
06 May 2011	D46	1,779	7,237	232	86,692	24,698
13 May 2011	D13	1,641	7,621	282	65,239	25,653
20 May 2011	D40	1,101	5,193	211	54,957	19,215
27 May 2011	D33	1,327	7,236	256	67,537	22,515
03 June 2011	D06	1,240	7,249	248	64,487	21,428
09 June 2011	D27	1,467	7,502	258	63,327	22,142
17 June 2011	D42	1,586	7,954	325	89,132	26,333
24 June 2011	D03	1,734	10,889	372	83,559	29,843
01 July 2011	D47	1,856	12,119	370	110,021	31,075
08 July 2011	D18	1,598	8,667	261	88,639	23,323
05 August 2011	D45	2,039	11,501	403	110,526	31,695
11 August 2011	D51	2,049	10,338	384	144,888	35,844
19 August 2011	D43	1,620	7,664	379	84,975	25,036
29 August 2011	D34	1,539	8,902	315	93,107	25,778
05 September 2011	D30	1,844	10,409	445	359,352	35,261
12 September 2011	D04	1,882	11,479	421	81,870	28,812
19 September 2011	D31	1,650	10,015	382	112,779	28,742
28 September 2011	D29	2,031	15,387	470	508,357	39,549
05 October 2011	D49	1,733	10,015	344	102,073	29,044
12 October 2011	D32	2,144	11,556	346	102,564	31,175
02 November 2011	D23	1,661	9,149	262	183,658	27,171
07 November 2011	D44	1,914	13,443	373	124,125	26,507
16 November 2011	D20	2,153	21,430	563	102,499	27,069
23 November 2011	D15	1,730	18,745	401	90,195	21,686
29 November 2011	D05	1,632	15,662	375	94,697	22,039
21 December 2011	D28	1,832	15,446	378	110,959	29,162
28 December 2011	D09	1,829	13,765	316	99,623	26,758
03 January 2012	D38	1,359	10,852	267	95,808	24,657
11 January 2012	D36	2,128	13,083	377	384,385	40,682
19 January 2012	D12	1,951	11,138	296	83,323	25,643
25 January 2012	D10	1,931	11,098	328	77,970	24,866
01 February 2012	D14	1,983	9,340	249	67,645	21,188
08 February 2012	D08	2,684	12,742	390	85,554	28,888
14 February 2012	D26	1,531	8,470	310	76,139	24,191
23 February 2012	D21	1,542	7,305	268	61,864	21,261
29 February 2012	D19	1,729	9,003	300	73,013	25,415
08 March 2012	D17	1,490	7,180	233	55,877	18,707
14 March 2012	D41	1,546	7,801	270	73,315	23,718
22 March 2012	D11	1,627	7,921	274	87,678	23,207
28 March 2012	D16	1,512	7,965	276	68,028	21,155
04 April 2012	D07	1,219	5,047	182	51,559	15,112
10 April 2012	D22	1,696	9,806	323	250,766	32,752
17 April 2012	D01	1,768	11,037	352	73,249	25,511
25 April 2012	D50	1,568	10,082	281	117,110	26,872
03 May 2012	D48	1,516	9,727	291	97,198	25,569
Total		88,807	523,876	16,730	5,859,481	1,369,301
Average		1,676	9,884	316	110,556	25,836
Standard deviation		333.140842	3332.121838	76.968261	84833.45653	5483.53806

Table C.3: Summary statistics of *Candidatus* Microthrix parivicella Bio17-1 population CRISPR elements based different time points

Date	Sample ID	Repeats	Spacers	Flanks	Protospacers	Protospacer containing contigs
04 October 2010	A01	0	0	0	100	16
25 January 2011	A02	6	31	1	496	115
21 March 2011	D37	68	724	29	1634	519
29 March 2011	D02	54	542	29	1312	453
05 April 2011	D39	72	859	24	1382	450
14 April 2011	D24	61	726	30	1486	471
21 April 2011	D25	43	426	21	626	206
29 April 2011	D35	61	589	26	690	289
06 May 2011	D46	48	701	29	699	252
13 May 2011	D13	59	783	27	1257	428
20 May 2011	D40	56	582	12	632	230
27 May 2011	D33	53	760	20	718	297
03 June 2011	D06	62	730	20	767	274
09 June 2011	D27	52	594	14	804	265
17 June 2011	D42	61	600	14	795	339
24 June 2011	D03	51	728	20	1288	413
01 July 2011	D47	39	419	14	955	391
08 July 2011	D18	46	483	16	1001	323
05 August 2011	D45	34	379	16	931	429
11 August 2011	D51	50	487	16	977	420
19 August 2011	D43	55	645	23	975	424
29 August 2011	D34	73	731	23	880	361
05 September 2011	D30	54	615	19	1112	470
12 September 2011	D04	53	659	22	1513	505
19 September 2011	D31	70	762	25	1145	447
28 September 2011	D29	52	732	24	1497	616
05 October 2011	D49	38	431	13	1062	457
12 October 2011	D32	53	623	16	1298	519
02 November 2011	D23	48	516	26	703	217
07 November 2011	D44	48	455	20	946	323
16 November 2011	D20	32	273	7	877	292
23 November 2011	D15	8	54	3	459	174
29 November 2011	D05	3	57	2	466	183
21 December 2011	D28	47	614	19	1235	536
28 December 2011	D09	50	514	31	1157	394
03 January 2012	D38	54	584	15	1127	393
11 January 2012	D36	75	695	25	2089	614
19 January 2012	D12	57	657	22	2135	534
25 January 2012	D10	78	944	38	1372	483
01 February 2012	D14	82	776	33	1347	410
08 February 2012	D08	66	847	38	1478	552
14 February 2012	D26	70	626	13	1202	516
23 February 2012	D21	68	783	26	1131	457
29 February 2012	D19	80	949	45	1199	480
08 March 2012	D17	61	884	27	1417	434
14 March 2012	D41	82	891	32	1047	477
22 March 2012	D11	60	955	29	1481	491
28 March 2012	D16	52	850	22	1119	410
04 April 2012	D07	51	385	10	660	160
10 April 2012	D22	61	692	23	1447	457
17 April 2012	D01	59	821	36	1319	484
25 April 2012	D50	59	753	20	922	375
03 May 2012	D48	66	804	25	1034	407
Total		2,841	32,720	1,130	57,401	20,632
Average		54	617	21	1,083	389
Standard deviation		18.2254	229.6875	9.458	386.03	128.423934

Table C.4: Summary statistics of LCSB005 population CRISPR elements based different time points

Date	Sample ID	Repeats	Spacers	Flanks	Protospacers	Protospacer containing contigs
04 October 2010	A01	0	0	0	0	0
25 January 2011	A02	0	0	0	3	3
17 April 2012	D01	0	0	0	9	9
29 March 2011	D02	0	0	0	10	10
24 June 2011	D03	0	0	0	15	11
12 September 2011	D04	0	0	0	10	10
29 November 2011	D05	1	37	0	7	7
03 June 2011	D06	0	0	0	9	9
04 April 2012	D07	0	0	0	5	5
08 February 2012	D08	1	7	0	11	11
28 December 2011	D09	0	0	0	5	5
25 January 2012	D10	0	0	0	6	6
22 March 2012	D11	0	0	0	8	8
19 January 2012	D12	0	0	0	8	8
13 May 2011	D13	0	0	0	11	10
01 February 2012	D14	0	0	0	5	5
23 November 2011	D15	2	34	5	8	7
28 March 2012	D16	0	0	0	16	10
08 March 2012	D17	0	0	0	5	5
08 July 2011	D18	0	0	0	11	6
29 February 2012	D19	0	0	0	13	12
16 November 2011	D20	1	25	6	10	9
23 February 2012	D21	0	0	0	14	9
10 April 2012	D22	0	0	0	9	7
02 November 2011	D23	0	0	0	9	9
14 April 2011	D24	0	0	0	19	15
21 April 2011	D25	0	0	0	12	10
14 February 2012	D26	1	8	0	17	13
09 June 2011	D27	0	0	0	12	8
21 December 2011	D28	0	0	0	8	8
28 September 2011	D29	0	0	0	9	9
05 September 2011	D30	1	11	0	11	11
19 September 2011	D31	0	0	0	7	4
12 October 2011	D32	0	0	0	15	11
27 May 2011	D33	0	0	0	11	11
29 August 2011	D34	0	0	0	16	11
29 April 2011	D35	0	0	0	18	12
11 January 2012	D36	0	0	0	7	7
21 March 2011	D37	0	0	0	11	11
03 January 2012	D38	0	0	0	9	9
05 April 2011	D39	0	0	0	10	10
20 May 2011	D40	0	0	0	9	9
14 March 2012	D41	0	0	0	16	11
17 June 2011	D42	0	0	0	16	11
19 August 2011	D43	0	0	0	12	9
07 November 2011	D44	0	0	0	6	6
05 August 2011	D45	0	0	0	16	11
06 May 2011	D46	0	0	0	3	3
01 July 2011	D47	0	0	0	7	7
03 May 2012	D48	0	0	0	16	11
05 October 2011	D49	0	0	0	1	1
25 April 2012	D50	0	0	0	17	13
11 August 2011	D51	0	0	0	22	17
Total		7	122	11	550	460
Average		0	2	0	10	9
Standard deviation		0.39408	7.725	1.063	4.719951169	3.256789269

APPENDIX D

ADDITIONAL FILES

D.1 Additional file 2.1: Supplementary IMP HTML reports

HTML S1 and S2 are reports produced by IMP for the analysis of the human fecal microbial community and wastewater sludge microbial community datasets. HTML reports for the analyses of other datasets are also included. The file is available via the original publication [**Appendix A.2**] and **Zenodo**.

D.2 Additional file 2.2: Supplementary figures and notes

Supplementary figures and notes. Figures S1-S3 and Notes S1-S2. Detailed figure legends available within file. The file is available via the original publication [**Appendix A.2**].

D.3 Additional file 2.3: Supplementary tables

Tables S1 to S12 and their detailed table legends are available within file. The file is available via the original publication [**Appendix A.2**].