PhD-FSTC-2017-19
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defense held on 29/03/2017 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Anne KAYSEN
Born on 21 April 1988 in Luxembourg (Luxembourg)

DYNAMIC CHANGE OF THE HUMAN GASTROINTESTINAL MICROBIOME IN RELATION TO MUCOSAL BARRIER EFFECTS DURING CHEMOTHERAPY AND IMMUNE ABLATIVE INTERVENTION

Dissertation defence committee

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

Dr. Stephanie Kreis, Chair of committee
Université du Luxembourg

Dr. Jochen Schneider, Vice-chair of committee
Université du Luxembourg

Dr. Christoph Reinhardt
Junior professor, University Medical Center of the Johannes Gutenberg University Mainz

Dr. Jörg Bittenbring
Saarland University Medical Center

Dynamic change of the human gastrointestinal microbiome in relation to mucosal barrier effects during chemotherapy and immune ablative intervention

A dissertation

by

Anne Kaysen

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

in collaboration with the
Saarland University Medical Center

To obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

This work was funded by the University of Luxembourg.

March, 2017

Affidavit

I hereby confirm that the PhD thesis entitled 'Dynamic change of the human gastrointestinal microbiome in relation to mucosal barrier effects during chemotherapy and immune ablative intervention' has been written independently and without any other sources than cited.

Luxembourg, _____

Name and Signature: _____

Acknowledgments

I would like to thank my supervisor Associate Professor Dr. Paul Wilmes, for giving me the opportunity to work on this project within a great team. Your enthusiasm, optimism and interest kept inspiring and pushing me when it was needed.
I want to thank Prof. Dr. Jochen Schneider for guidance and critical feedback throughout the project; Dr. Stephanie Kreis for valuable feedback and ideas during the CET meetings and agreeing to be chairperson for the defense.
My gratitude goes to Prof. Dr. Christoph Reinhardt and Dr. Jörg Bittenbring for agreeing to be on my dissertation committee. Thank you for taking the time to attend my defense.

I want to express my highest gratitude to all the collaborators involved in this project, either by recruiting patients, taking samples, organizing sample storage and transfer or documenting clinical data and blood counts: Prof. Dr. Norbert Graf, Prof. Dr. Arne Simon, Dr. Jörg Bittenbring, Dr. Eyad Torfah, Dr. Michael Ehrhardt, Dr. Holger Stenzhorn, Prof. Dr. Martina Sester, Dr. Tina Schmidt, Lisa Lieblang, Katharina Franke, Manuela Faust, Mariana Borbely, Yvonne Hemmer and Hannah Bender. Thank you for investing valuable time into this project, despite an already packed work schedule. Without this dedication, this project would not have been feasible.
I wish to thank all the patients and/or their legal guardian who have agreed to participate despite their difficult situation. Thank you for having enough trust in research to participate in this study.

To the whole Eco-Systems Biology group, former and current members - thank you for being an amazing team. In tough times, the great working atmosphere and moral support kept drawing me back to my desk. Dr. Anna Heintz-Buschart, this work would not have been possible without your guidance and support throughout these years. Thank you for untiringly answering my innumerable questions. Laura Lebrun, thank you for keeping the laboratory organized and running and especially for your technical assistance! Dr. Emilie Muller, thank you for stimulating discussions and valuable ideas. Linda Wampach, I especially enjoyed our time as desk neighbors. Thanks for always cheering me up. Dr. Shaman Narayanasamy, thank you for always having time to help me, especially with those very IMPortant things. Dr. Abdul Sheik, Dr. Cédric Laczny and Claire Battin, thank you for your support and help in different areas – in the lab, during data analysis and outside of the lab. I want to thank our extraction robot 'Eva' for carefully handling the precious samples without major incidents (at least most of the time).

I thank the University of Luxembourg for financing this project, the Luxembourg Centre for Systems Biomedicine and the Doctoral School for their support throughout my PhD project.

Many thanks go to my family, especially my parents, Liliane a René. You have taught me some of the most important things in life. Thanks for your unconditional love and support. My brother, Marc, for always believing in me. Especially in the last few months, your kind words have brought motivation and energy when they were much needed.
My partner, Raoul - thanks for being at my side, even and especially in difficult times. Thanks for your love, support and understanding. I can't wait for our next adventure together. No matter what and where it may be. Fiischen (†16.12.2016) – your calming purrs have been deeply missed in the last few months.

Summary

Numerous studies have demonstrated that the gastrointestinal tract (GIT) microbiota plays important roles for the human host. Since the GIT microbiota interfaces with the immune system and represents a first line of defense against infectious agents, interest has grown in whether the GIT microbiota may influence the outcome of different anticancer treatments. In this study, the GIT of pediatric patients with different cancer types as well as adult patients with hematologic malignancies undergoing an allogeneic hematopoietic stem cell transplantation were sampled throughout their treatment. In order to deeply profile not only the composition of the community, but also the functional capacity and expression, recently developed wet- and dry-lab methodologies for integrated multi-omic analyses were applied. The trajectories of the prokaryotic and microeukaryotic GIT communities of the patients were described in detail using 16S, 18S rRNA gene amplicon sequencing, as well as metagenomic and metatranscriptomic shotgun sequencing. Indeed, changes in the GIT microbiome in response to treatment were detected. Some changes that are generally thought to be detrimental for human health were detected during treatment, such as a decrease in alpha-diversity, a decrease in relative abundance of bacteria associated with health-promoting properties (such as *Blautia* spp., *Roseburia* spp. and *Faecalibacterium* spp.), as well as an increase in the relative abundance of antibiotic resistance genes. These changes were more pronounced in the adult hematology patients than in the pediatric patients, which is likely due to the more intensive treatment. Some observations need further investigation in order to explain their implication in human health. For example, in the pediatric patients, lower relative abundance of *Akkermansia muciniphila* was associated with mucositis and functional gene categories that are linked to bacteriophages or the bacterial defense mechanism against bacteriophages were associated with the overall status of the patient and mucositis development. Importantly, in both cohorts, high inter-individual but also high intra-individual variation in the prokaryotic communities were detected while the microeukaryotic community did not exhibit drastic changes. In conclusion, the employed integrated multi-omics analysis allowed detailed profiling of the GIT community including archaea, bacteria, eukaryotes and viruses as well as the functional potential including antibiotic resistance genes. In the future, analysis of the individual-specific processes within the GIT microbial community of patients throughout treatment might allow to adjust therapy regimens accordingly and improve the overall outcome of the therapy.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ALL	acute lymphoid leukemia
allo-HSCT	allogeneic hematopoietic stem cell transplantation
aGvHD	acute graft-versus-host disease
AML	acute myeloid leukemia
AMP	antimicrobial peptide
APC	antigen presenting cell
ARG	antibiotic resistance gene
ATG	antithymocyte globulin
bp	base pair
CDI	*Clostridium difficile* infection
DNA	DNA
contig	contiguous sequence
CpG-ODN	CpG-oligodeoxynucleotide
CRISPR	clustered regularly inter-spaced palindromic repeats
CRP	C-reactive protein
CSF	colony-stimulating factor
CTX	cyclophosphamide
DC	dendritic cell
DNA	deoxyribonucleic acid
FDR	false discovery rate
FMT	Fecal microbiota transplantation
G-CSF	granulocyte colony-stimulating factor
GALT	gut-associated lymphoid tissue
GF	germ-free
GIT	gastrointestinal tract
GVC	graft-versus-cancer
GvHD	graft-versus-host disease
GVT	graft-versus-tumor
HLA	human leukocyte antigen
HMM	hidden Markov model
ICI	immune checkpoint inhibitor
IgA	immunoglobulin A
IHF	integration host factor
IL	interleukin

IMP	Integrated Meta-Omic Pipeline
ITS	internal transcribed spacers
KO	KEGG orthologous group
MAMP	microbe-associated molecular pattern
MDR	multidrug resistant
MG	metagenomic
MHC	major histocompatibility complex
mOTU	metagenomic operational taxonomic unit
MT	metatranscriptomic
NHL	non-Hodgkin's lymphoma
NK	natural killer cell
NLR	nucleotide-binding oligomerisation domain-like receptors
nt	nucleotide
OTU	operational taxonomic unit
PAMP	pathogen-associated molecular pattern
PBSC	peripheral-blood stem cells
PCA	principal component analysis
PCR	polymerase chain reaction
PMMoV	pepper mild mottle virus
PRR	pattern recognition receptor
RHM	reference healthy microbiome
RNA	ribonucleic acid
ROS	reactive oxygen species
rRNA	ribosomal RNA
SCFA	short-chain fatty acid
SNAPP	structurally nanoengineered antimicrobial peptide polymer
SNV	single nucleotide variant
SPF	specific-pathogen-free
TG	treatment group
T_H	T helper cell
TLR	Toll-like receptor
TNF	tumor necrosis factor
TP	time point
T_{reg}	regulatory T cell
VRE	vancomycin resistant *Enterococcus*

1 Introduction

1.1 The human microbiome

The human body is colonized by a multitude of different microorganisms, commonly referred to as the 'microbiota', with their associated genomes being referred to as the 'microbiome' (Ursell, Metcalf, Parfrey, & Knight, 2013). Various sites such as the skin, the oral cavity and the gastrointestinal tract (GIT) are highly colonized by these organisms. This ecosystem is assumed to be composed of about 100 trillion microorganisms, including 500 – 1500 different species of bacteria, archaea, fungi, unicellular eukaryotes and viruses (Hooper & Gordon, 2001; Kinross, von Roon, Holmes, Darzi, & Nicholson, 2008; Schwiertz, 2016; Sekirov, Russell, Antunes, & Finlay, 2010). The human microbiome, especially the GIT microbiome has recently gained much research interest worldwide (Figure 1.1.1). For a long time, it was believed that the number of human cells that makes up the human body is outnumbered by the number of microorganisms living in and on it, by at least a factor of ten (Luckey, 1972). However, this was based on a rough estimate and has recently been revised and rectified. According to more recent studies, the number of microorganisms associated with the human body is approximately equal to the number of human cells (Sender, Fuchs, & Milo, 2016). Recent studies focussing on the microbiome of different body sites have highlighted that the community structure of a specific body site remains relatively constant within one person and that the inter-individual variation of the community structure is in general higher than the variation over time within one person (The Human Microbiome Project Consortium, 2012; Zhou et al., 2013).

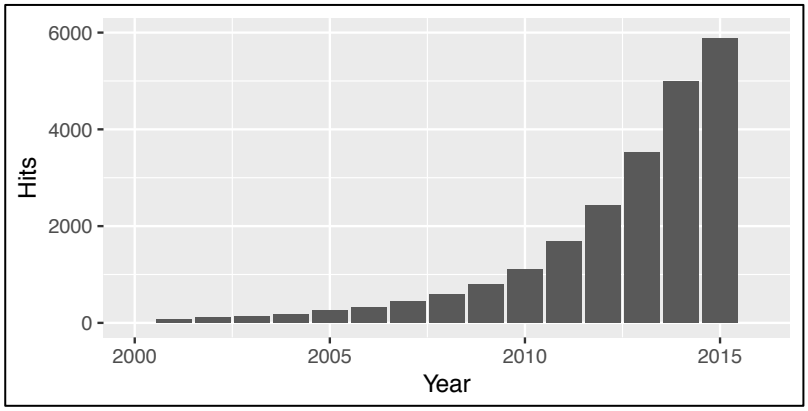


Figure 1.1.1: Number of published articles including the words 'human microbiome' per year. The plot indicates the number of published articles in PuBMed (Medline) including the words 'human microbiome' per year (Corlan, 2004).

1.1.1 The technology-driven revolution of human microbiome research

It is known that only a small part of the microorganisms present in the GIT are culturable in isolation under laboratory conditions. Therefore, for a long time, culture-dependent analyses have allowed to only characterize and inspect a part of the microbiome. Novel culture-independent methods have allowed a much deeper characterization of the human microbiome. Targeted amplification and sequencing of specific phylogenetic marker genes such as 16S (for prokaryotes) and 18S (for eukaryotes) ribosomal ribonucleic acid (rRNA) gene or internal transcribed spacer (ITS, mostly used for fungi) has led to a revolution in microbiome research. 16S and 18S rRNAs are part of the small ribosomal subunits, meaning that they are present in each prokaryotic and eukaryotic organism, respectively. Furthermore, they contain conserved regions, allowing the construction of universal primers used for polymerase chain reaction (PCR) amplification, as well as hypervariable regions (Figure 1.1.2), which can be utilized to identify different species. These traits make the 16S rRNA gene the 'gold standard' genetic marker for bacterial phylogeny (Case et al., 2007).

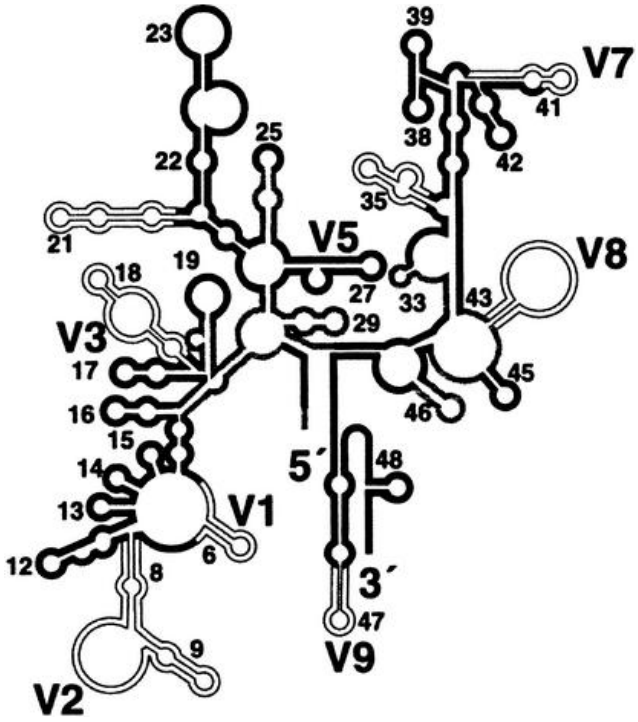


Figure 1.1.2: Secondary structure of the 16S ribosomal RNA. Double lines represent variable or hypervariable regions, single lines represent highly conserved regions. V1 to V9 represent hypervariable regions (Tortoli, 2003).

Organisms are classified into a hierarchical system, the taxonomic classification. Different levels or ranks have been defined. Ideally, taxonomy reflects evolutionary relationships

among organisms. In the past, bacteria were classified based on their morphologic and phenotypic characteristics (e.g. shape, Gram stain, motility) while more recently, gene sequences (including the 16S rRNA gene) are used to identify relationships. Table 1.1.1 includes the taxonomic classification of *Escherichia coli* as an example.

Table 1.1.1: Example of the taxonomic classification of *Escherichia coli.*

Taxonomic rank	**Taxon**
Domain	Bacteria
Phylum	Proteobacteria
Class	Gammaproteobacteria
Order	Enterobacteriales
Family	Enterobacteriaceae
Genus	*Escherichia*
Species	*Escherichia coli*

On one hand, amplicon sequencing has allowed to characterize the composition of the microbiome much more in detail than cultivation. Around 80 % of the sequenced taxa had not been cultivated before (Eckburg et al., 2005). On the other hand, this method allows to solely identify different taxa, whereas culturing of isolates is necessary to describe the biological and genetic nature of the organisms. However, the creation of isolate cultures is difficult and often even impossible. Reasons for this 'unculturability' include the need for specific growth conditions such as nutrients, pH, incubation temperatures, or dependence on the presence of other community members (Vartoukian, Palmer, & Wade, 2010).

The third approach, metagenomics, refers to culture-independent genomic analysis of an assemblage of microorganisms. The emergence of high-throughput random shotgun sequencing and the continuous decrease in sequencing costs (Figure 1.1.3) have made it possible to explore complex communities using metagenomic (MG) sequencing without *a priori* knowledge, without reference database.

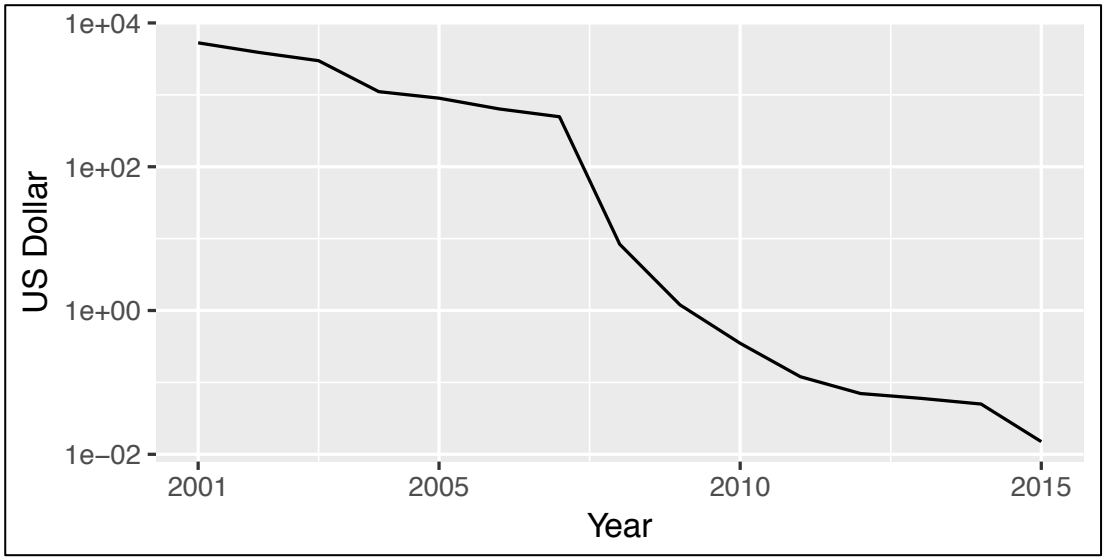


Figure 1.1.3: Development of DNA sequencing costs over time. The sequencing cost is represented in US Dollar per megabase. Data taken from https://www.genome.gov/sequencingcostsdata (Wetterstrand, 2016).

The collective genomes of a community within a sample is called the metagenome (Handelsman, Rondon, Brady, Clardy, & Goodman, 1998), while the associated RNA is called metatranscriptome. Metatranscriptomic (MT) sequencing allows community-wide gene expression to be resolved. At a sufficient sequencing depth, *de novo* MG assemblers are now able to assemble genomes of a complex community, such as a fecal microbiome (Segata et al., 2013). Large-scale international MG studies (Human Microbiome Project, MetaHIT) are concentrating on characterizing the human microbiome and the complex interplay between this microbial community and its human host (Peterson et al., 2009; The Human Microbiome Project Consortium, 2012). A reference gene catalogue of the GIT microbiome has been assembled, including 9.9 Mio non-redundant genes (Li et al., 2014). The functional potential of the microbiome is estimated to be two orders of magnitude greater than that encoded by the human genome (Bäckhed, Ley, Sonneburg, Peterson, & Gordon, 2005). Meta-omic studies, combining metagenomics and metatranscriptomics are currently arising and will change microbiome studies, allowing an even more precise characterization of the community, its functional potential and gene expression and thus, its relationship and importance for the human host.

1.1.2 Characteristics of the human gastrointestinal microbiome

While every body site has its own unique microbial community, the composition of each community varies between individuals and over time, due to external influences such as changes in diet, antibiotic administration and important lifestyle changes (Lozupone, Stomabaugh, Gordon, Jansson, & Knight, 2012).

The majority of the microorganisms associated with the human body live in the GIT. This organ system includes the most stable and diverse microbiome, the colon being the most densely colonized compartment, with densities of around 10^8 cells per ml in the cecum to up to 10^{12} cells per ml in stool (Dethlefsen, Eckburg, Bik, & Relman, 2006). Although this microbiome is considered to be quite stable in healthy individuals, it is difficult to determine features of a 'healthy' microbiome, as it might be different for people according to their age, geographical location and genetics (Greenhalgh, Meyer, Aagaard, & Wilmes, 2016). Upon birth, neonates are exposed to a high number of microorganisms which influence colonization if the neonate GIT. Gestational age and the mode of delivery affect the initial colonization and following succession of the infant GIT (Arboleya et al., 2015; Jakobsson et al., 2013). After birth, the first food (breast milk versus formula milk) as well as the subsequent diet (including solid food) influences the GIT microbiome (Thompson, Monteagudo-Mera, Cadenas, Lampl, & Azcarate-Peril, 2015). Increasingly improved hygiene and antibiotic usage in early childhood are believed to negatively affect the GIT microbiome (Shen & Wong, 2016). In addition, throughout lifetime, external environmental factors such as the living area (urban versus rural environment) (Nakayama et al., 2015), siblings (Penders et al., 2006), pets (Azad et al., 2013) and the general familial environment further influence the GIT microbiome (Figure 1.1.4).

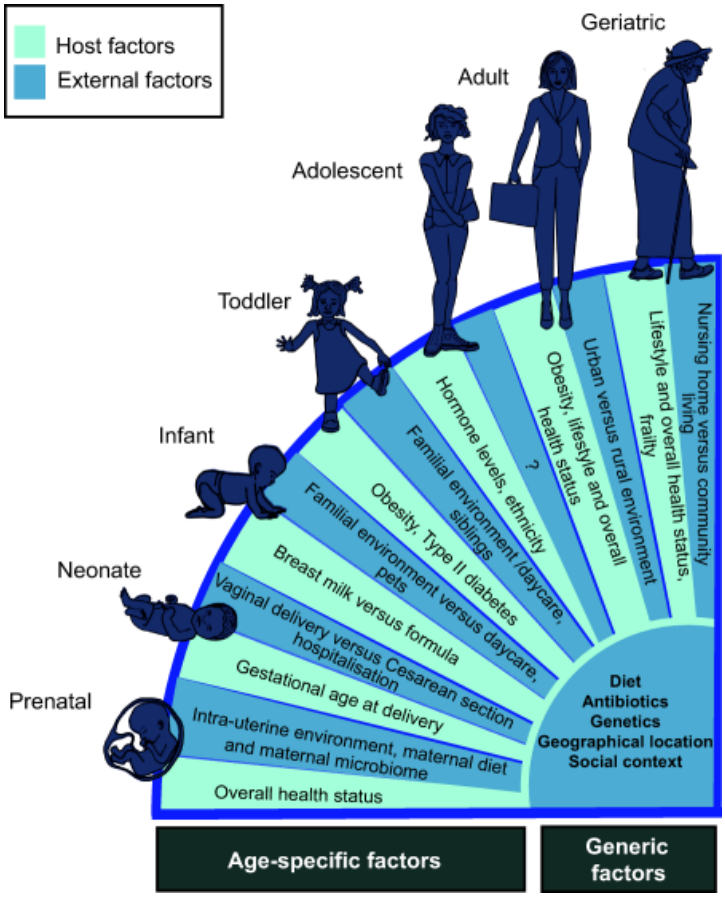


Figure 1.1.4: Generic and age-specific factors which influence the GIT microbiome (Greenhalgh et al., 2016).

1.1.3 The gastrointestinal microbiome in human health and disease

On the one hand, colonization by microorganisms can negatively affect humans, infectious diseases being one example. On the other hand, microorganisms perform essential functionalities such as carbohydrate metabolism, modulation of epithelial barrier function and nutrient absorption (Hollister, Gao, & Versalovic, 2014). In the GIT, they also play important roles for the host as for example in shaping the immune system (Mazmanian, Cui, Tzianabos, & Kasper, 2005), synthesis of vitamins (Qin et al., 2010), providing the host with short-chain fatty acids (SCFAs) (Qin et al., 2010), prevention of colonization by pathogens (Ivanov et al., 2009; Stecher & Hardt, 2011) and the metabolism of xenobiotics (Maurice, Haiser, & Turnbaugh, 2013) (Figure 1.1.5).

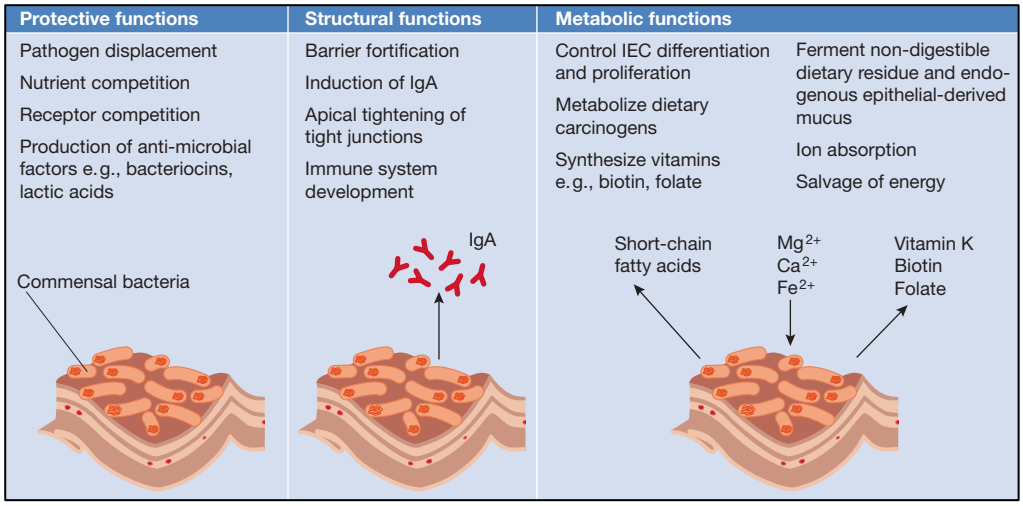


Figure 1.1.5: Roles of the GIT microbiome. The GIT microbiome exerts many protective, structural and metabolic functions within the host. IgA = Immunoglobulin A; IEC = intestinal epithelial cell (O'Hara & Shanahan, 2006).

Recent MG analysis has shown that the GIT microbial community is mainly comprised of bacteria, with 97.6 % of the reads belonging to this domain, while 2.2 % of the reads belonged to the archaea and less than 0.01 % to eukaryotes. An additional 0.2 % of the reads could be associated with viruses (Zhernakova et al., 2016). The intestinal microbiota is primarily composed of four different bacterial phyla, Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria, with Firmicutes and Bacteroidetes accounting for more than 90 % of the total bacterial community (Ley, Peterson, & Gordon, 2006). Proportions between these phyla can vary between individuals and also within one individual over time.

In a healthy person, the GIT microbiota usually includes a balanced composition of different organisms, which is essential for human health. These include different commensals, symbionts and pathobionts. Commensals are residents that do not provide any benefit nor harm to the host. Symbionts are organisms that benefit the host. Pathobionts are organisms that do not normally elicit an inflammatory response, but have the potential to cause inflammation and lead to disease (Round & Mazmanian, 2009). Several factors might lead to a shift and significant alterations in the composition of the microbiota with either a decrease in symbionts and commensals and/or an increase in pathobionts (Figure 1.1.6). This disruption of the balanced state between the intestinal microbes is called dysbiosis. The term 'dysbiosis' was already used in 1890 by Dr. E. E. Furney where he related to plant, animal and human resiliency (Furney, 1890). This state refers to an altered composition with associated functional changes. In this scenario,

pathobionts can become pathogenic and cause non-specific inflammation and possibly diseases (Cerf-Bensussan & Gaboriau-Routhiau, 2010; Round & Mazmanian, 2009).
Dysbiosis has recently been linked to different diseases or conditions such as inflammatory bowel disease, obesity, colorectal cancer, cardiovascular diseases, diabetes, allergies, infection and multiple sclerosis (Bollyky et al., 2009; Hill et al., 2012; Kamada, Seo, Chen, & Núñez, 2013; Ley, Turnbaugh, Klein, & Gordon, 2006; Manichanh et al., 2006; Moore & Moore, 1995; Qin et al., 2012; Z. Wang et al., 2011). Most of these conditions involve inflammation, which is a result of an altered immune response, in this case as consequence of an altered intestinal microbiome. Thus, immune dysregulation and GIT dysbiosis often concur, one as a result of the other.

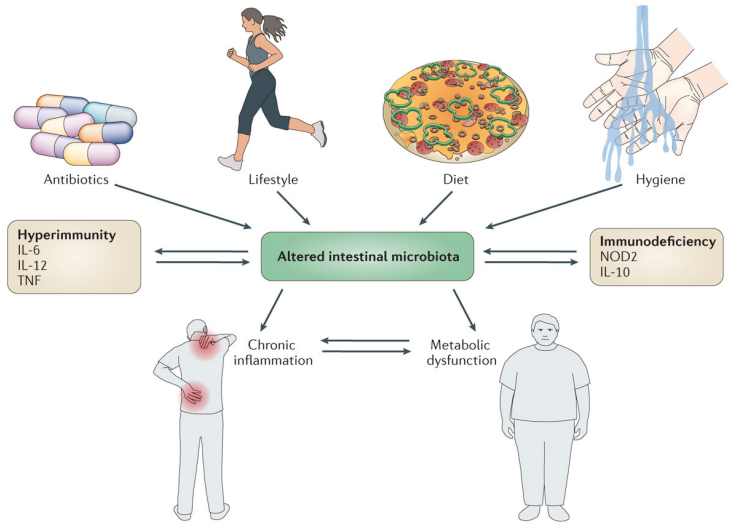


Figure 1.1.6: Factors influencing the GIT microbiome composition and subsequent effects on host health (F. Sommer & Bäckhed, 2013).

As mentioned before (section 1.1.2 and Figure 1.1.4), many factors such as diet, lifestyle, hygiene, antibiotics and other medication, shape and influence the GIT microbiome. The altered composition of the GIT microbiome can lead to modulation of production of pro-inflammatory mediators (such as (interleukin) IL-6, IL-12 and tumor necrosis factor (TNF)) or anti-inflammatory mediators (such as IL-10), while nucleotide-binding oligomerization domain-containing protein 2 (NOD2) polymorphisms have been linked to Crohn's disease and GvHD (Penack, Holler, & van den Brink, 2010). This regulation of the immune system can in turn influence the GIT microbiome and lead to dysbiosis (Figure 1.1.6).
Perturbations of the GIT microbiota may allow 'blooms' of harmful bacteria that are usually only lowly abundant, contributing to development of a disease. Especially blooms of Enterobacteriaceae (such as *E. coli*, *Proteus* spp. or *Klebsiella* spp.) are often observed in

a dysbiosis. Additionally, this may result in a vicious cycle of pathobiont blooms induced by inflammation and further inflammation induced by pathobionts.

In individuals with colorectal cancer, the abundance of *Fusobacterium nucleatum* is frequently increased in the fecal microbiome (Castellarin et al., 2011). In both Crohn's disease and ulcerative colitis, the two main forms of inflammatory bowel disease, a lower diversity in the GIT microbiome as well as changes in the microbiome composition (i.e. an increased abundance of mucosal-associated aerotolerant bacteria compared to healthy individuals) have been observed (Gevers et al., 2014; Wlodarska, Kostic, & Xavier, 2015).

Although specific patterns or shifts in the GIT microbial communities could be linked to diseases and disorders, often it is still unknown what is cause and consequence. Considering the multitude of different microbial populations living in and on the human body and the important roles they are playing in health but also in disease, it is clear that a precise characterization of the microbial community composition, its functional potential and actual expression of genes, but also specifically of changes in the community composition throughout time is needed, in order to be able to link these changes to the development of a disease or disorder.

1.1.4 Relationship between the gastrointestinal microbiome and the immune system

The GIT is densely populated with an important variety of microorganisms and they are in close contact with the host intestinal mucosa and its innate and adaptive immune systems. Here, the immune system helps in maintaining a balanced community of commensals, which in turn plays a part in protecting from pathogen invasion, for example by occupation of specific niches and by nutrient competition.

Physical barriers such as the mucus layer (produced by goblet cells), tight junctions and secretion of certain anti-microbial peptides (by Paneth cells) regulate the relationship between the microbiota and the host (Hooper, Littman, & Macpherson, 2012). Peyer's patches are collections of lymphoid follicles in the intestinal mucosa, which harbor antigen-presenting cells (such as macrophages and dendritic cells). Together with mesenteric lymph nodes and lamina propria lymphocytes, they are part of the gut-associated lymphoid tissue (GALT) (Cerf-Bensussan & Gaboriau-Routhiau, 2010; Schuijt, van der Poll, de Vos, & Wiersinga, 2013). Epithelial cells, dendritic host cells and macrophages express pattern recognition receptors (PRRs) such as Toll-like receptors (TLRs) and nucleotide-binding oligomerization domain-like receptors (NLRs) which recognize microbe-associated molecular patterns (MAMPs), evolutionary conserved molecular structures produced by various microorganisms (Clemente, Ursell, Parfrey, &

Knight, 2012). Activation of these receptors can induce a spectrum of signalling events from a pro-inflammatory cytokine response up to the presentation of antigens to regulatory T cells (T_{regs}). Activation of these T_{regs} conveys tolerance towards commensal bacteria from the initial colonization of the GIT during early life (Fukata, Vamadevan, & Abreu, 2009). Some GIT bacteria produce SCFAs such as butyrate, propionate and acetate. Butyrate is one of the most important energy sources for enterocytes and has important anti-inflammatory properties by inhibiting nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) signaling (Vinolo, Rodrigues, Nachbar, & Curi, 2011). It was also shown to upregulate expression of tight junction proteins, thereby strengthening the physical epithelial barrier (Peng, Li, Green, Holzman, & Lin, 2009).

While most bacteria occupy the GIT lumen, segmented filamentous bacteria (SFB) seem to be able to penetrate the mucus layer and interact closely with the epithelial cells, inducing signaling events that lead to differentiation of T helper (T_H) 17 cells (Ivanov et al., 2009). These are cytokine (mainly IL-17A, IL-17F, IL-21 and IL-22) producing CD4+ effector T cells which are specialized in responses to extracellular bacteria and fungi (Ouyang, Kolls, & Zheng, 2012). On the other hand, polysaccharide A (PSA) produced by *Bacteroides fragilis* prevents expansion of T_H17 cells. It has anti-inflammatory or regulatory characteristics, which include induction of IL-producing T_{regs} cells, which in turn suppresses the production of pro-inflammatory cytokines (Troy & Kasper, 2010).

Additionally, epithelial innate antimicrobial effector molecules, called antimicrobial peptides (AMPs) are important effectors of innate immunity and can shape and regulate the composition of the GIT community. The peptide RegIIIγ depends on microbiome induced TLR signalling. Besides killing Gram-positive bacteria, it also prevents overstimulation of the immune system by limiting penetration of bacteria to the epithelial surface (Ubeda, Djukovic, & Isaac, 2017). Probiotics such as *Bifidobacterium breve* induce RegIII expression (Natividad et al., 2013). α-defensins are antibacterial peptides which are secreted by Paneth cells in response to bacteria or their antigens (Ayabe, Ashida, Kohgo, & Kono, 2004). Secretory immunoglobulin A (sIgA) is the most abundant class of antibodies in the intestinal lumen. Besides protecting the intestinal epithelium from pathogenic bacteria, viruses and toxins, it is also capable of downregulating pro-inflammatory responses (Mantis, Rol, & Corthésy, 2011).

Table 1.1.2 includes some of the most common representatives of the GIT microbiome along with the generally associated immune response or role in the GIT. However, it should be kept in mind that this is only the generally assumed consensus within a healthy adult person. Additionally, also commensals can become pathogenic under specific

circumstances and specific strains of otherwise nonpathogenic bacteria can harbor virulence factors, turning them into pathogens.

Table 1.1.2: Common bacterial members of the GIT microbiome and the general response in the GIT

Taxon	Response in the GIT
Roseburia	anti-inflammatory (butyrate producer)
Blautia	anti-inflammatory (SCFA producer)
Faecalibacterium	anti-inflammatory (butyrate producer)
SFB	formation of T_H17 cells
Clostridium cluster IV and XIVa	induction of T_{reg} cells
Escherichia	pro-inflammatory, TLR4 mediated NF-κB activation (response to lipopolysaccharide)
Bifidobacterium	induction of RegIII expression
Bacteroides	induction of T_{reg} cells
Akkermansia muciniphila	mucin degrading, anti-inflammatory

Studies with germ-free (GF) mice show that microbial colonization has consequences on the development of the immune system, especially of lymphoid structures (Macpherson & Harris, 2004). GF mice develop smaller Peyer's patches than mice grown under specific-pathogen-free (SPF) conditions and are deficient in secretory immunoglobulin A (IgA) (Round & Mazmanian, 2009). A mixture of 17 SCFA-producing bacterial strains from the order Clostridiales has been shown to be important for induction of colonic T_{regs} (Atarashi et al., 2013). This indicates how diet, microbes, their products and the immune system are interconnected and complexly regulated.

In short, one role of the immune system is not to extinguish all microorganisms living in the GIT but to confer tolerance, establish homeostasis, which includes commensals, which exert essential functions on the host as for example in digestion. This homeostasis is maintained by an intricate balance of pro-inflammatory cells (like T_H1 cells and T_H17 cells) and anti-inflammatory T_{regs} (Hooper et al., 2012). This highly sensitively regulated relationship between the microbiota and the host can be disturbed by many factors (such as antibiotic intake, drastic dietary changes or the invasion by pathogens) and this can lead to dysbiosis, an imbalance in the intestinal microbial community. Resulting inflammation can lead to epithelial damage and a resulting decreased intestinal barrier function. In immunocompromised patients, for example in cancer patients during intensive treatment, this 'leaky gut' allows translocation of microorganisms and microbial products

from the GIT lumen to neighboring tissues and/or bloodstream (Yu et al., 2014), putting the host at risk for local and systemic infections and sepsis (Khosravi & Mazmanian, 2013; Stecher, Maier, & Hardt, 2013).

1.2 Cancer and anticancer treatment

1.2.1 Cancer

Cancer is one of the leading causes of mortality with 8.2 million deaths worldwide in 2012 (Stewart & Wild, 2014). In adults, 80 % of the cancer types affect the respiratory, gastrointestinal and reproductive organs, while in children, less than 5 % of the malignancies affect these organs (Imbach, Kühne, & Arceci, 2004). More than 100 types of cancer have been categorized, with around 85 % of cancers affecting epithelial cells (carcinomas). Cancers derived from mesodermal cells (e.g. bone, muscle cells) are called sarcomas and cancers arising of glandular tissue (e.g. breast) are called adenocarcinomas (Pecorino, 2012). Every year, around 1 out of 500 children under the age of 16 years are diagnosed with childhood cancer, with acute lymphoblastic leukemia and brain tumors being the most frequent kinds of childhood cancer, accounting for 47 % of all pediatric neoplasia (Imbach et al., 2004). Within the past 30 years, chances for longterm survival have significantly increased to over 70 % in pediatric oncology (Imbach et al., 2004). Overall, in the United States the 5-year survival rate has increased by 23 % within the last 30 years, to now 69 % (Siegel, Miller, & Jemal, 2016). Globally, the cancer mortality rate decreases by approximately 1 % per year (Hashim et al., 2016).

All cancers arise due to changes in the DNA sequence of the cancer cell genomes. Tumorigenesis is thought to be a multistep process which is often compared to Darwin's theory of evolution, where cells with mutations are selected to survive, due to increased replicative and survival abilities (Stratton, Campbell, & Futreal, 2009). In 2000, Hanahan and Weinberg defined six key 'hallmarks of cancer', which include: sustained proliferative signalling, evasion of growth suppressors, resistance against apoptosis, immortality, induction of angiogenesis and inflammation as well as activation of invasion and metastasis (Hanahan & Weinberg, 2000). In 2011, emerging hallmarks and enabling characteristics were added to this list, including inflammation as tumor-promoting characteristic (Hanahan & Weinberg, 2011).

Mutations leading to this abnormal behaviour can be acquired or inherited. Acquired mutations are the most common cause of cancer in adults and can be caused due to factors like chemicals in tobacco smoke, ultraviolet radiation and viruses (Hyndman,

2016). These mutations often affect tumor suppressor genes, proto-oncogenes or DNA repair genes. Tumors of different origins show a high complexity and heterogeneity in their patterns of mutations (Luo, Solimini, & Elledge, 2009).

The lifetime risk of cancer for people has increased over time. In Great Britain, 1 in 2 people born after 1960 will be diagnosed with some type of cancer during their lifetime (Ahmad, Ormiston-Smith, & Sasieni, 2015). Considering this, it is clear that there is an urgent need for effective anticancer treatments. However, the high number of different types of cancer and their heterogeneity make this very difficult.

1.2.2 Anticancer therapies

The main treatment options for malignancies are radiotherapy, chemotherapy and surgery. After discovery of X-rays in 1895 by Wilhelm Conrad Röntgen, they were used diagnostically and first successful treatments of different skin tissue malignancies were reported in 1899 (Tomlinson & Kline, 2005). The most common forms of ionizing radiation used for the treatment of cancer are high-energy photons, gamma-rays or X-rays (Stockham, Balagamwala, Macklis, Wilkinson, & Singh, 2014). Other types of radiation include proton therapy and electron beams (Stockham et al., 2014). While passing through cells, energy of the ionizing radiation can directly result in DNA damage or indirectly in production of free radicals which provokes DNA damage, and thereby, lead to cell death (Baskar, Ann-Lee, Yeo, & Yeoh, 2012). Radiation treatment alone can be curative for some kinds of cancer (for example cervix carcinomas, head and neck carcinomas) but for others such as for example pediatric tumors and breast carcinomas, it is used together with other treatments (Baskar et al., 2012). Often, it is used before surgery to shrink the tumor, or after surgery to destroy cancer cells that could not be removed surgically.

Single drugs or combinations of chemotherapy drugs are used for treatment of many kinds of cancer, often together with surgery and/or radiotherapy. Chemotherapeutic agents either arrest cell growth (cytostatic) or kill rapidly dividing cells (cytotoxic). The principle of most cytotoxic drugs is that they attack actively dividing cells, therefore cancer cells, which divide more rapidly, are more affected than normal cells (Newman, 2010). However, due to the lack of selectivity for cancerous cells, non-malignant cells, especially those that undergo rapid division (e.g. hematopoietic, mucosal and gastrointestinal cells), are also affected, culminating in some of the side effects like bone marrow suppression and mucositis (Tomlinson & Kline, 2005).

Chemotherapy agents can be classified into cell cycle phase-specific agents (antimetabolites and plant derivatives) and cell cycle phase non-specific agents (alkylating

agents, antitumor antibiotics, corticosteroids and others) (Tomlinson & Kline, 2005). In general, these agents interfere with DNA synthesis, culminating in cell death. There are however differences in their modes of action. Antimetabolites are folic acid, pyrimidine or purine analogues and interfere with DNA production by inhibiting enzymes needed for nucleic acid production. Alkylating agents such as busulfan and cyclophosphamide act by attaching an alkyl group to DNA. Angiogenesis inhibitors interfere with the binding of angiogenesis-signalling molecules to receptors on endothelial cells (Shewach & Kuchta, 2009).

Which drugs and combinations of treatments are used depends on the underlying disease, the stage and the status of the patient. In pediatric oncology, dosage is based on body size, usually considering surface area or body weight (Ratain, 1998). Trials have shown, which combination of drugs (called combination chemotherapy) is most effective against a specific malignancy. As an example, in the following I will briefly explain the mode of action of different therapeutic agents, which were frequently used in this project for treatment of the pediatric patients. One combination often used for treatment of Hodgkin's lymphoma in children is 'OEPA', composed of vincristine (oncovin), etoposide, prednisone and doxorubicin (adriamycin) (Imbach et al., 2004). Vincristine is a vinca alkaloid, which works by binding to the tubulin protein, preventing the polymerization of tubulins, stopping the cell from separating chromosomes during metaphase, thereby leading to apoptosis. In addition to this, vincristine can also damage DNA and interfere with DNA, RNA and protein synthesis (Mohammadgholi, Rabbani-Chadegani, & Fallah, 2013). Etoposide derives from a toxin found in the American mayapple (also known as mandrake) and was first approved for cancer therapy in 1983 (Montecucco, Zanetta, & Biamonti, 2015). This cytotoxic cancer drug belongs to the topoisomerase inhibitor drug class. Topoisomerases are involved in essential cellular functions such as DNA replication, repair and transcription. Etoposide interferes with the topoisomerase II and DNA complex (called cleavable complex) and prevents religation of the DNA strands, causing DNA strand breaks and thus, apoptosis (Montecucco et al., 2015). Prednisone is a glucocorticoid that reduces inflammation and has been shown to cause regression of lymphoid tumors (Walsh & Avashia, 1992). Doxorubicin is an anthracycline antitumor antibiotic that was first extracted from *Streptomyces peucetius* var. *caesius* in the 1970's. Two mechanisms of action are proposed: it can inhibit topoisomerase II-mediated DNA repair by intercalation into DNA and it can damage membranes, DNA and proteins via production of free radicals (Thorn, Caroline; Oshiro, Connie; Marsh, Sharon; Hernandez-Boussard, Tina; McLeod, Howard; Klein, Teri; Altman, 2012).

Another approach is the activation of the immune system in order to fight cancer cells, in the so-called immunotherapy. It can be subdivided into active immunotherapy and adoptive immunotherapy. Common targets in this area are immune checkpoint inhibitors (ICI), monoclonal antibodies, cancer vaccines and cytokines. Some proteins (like CTLA-4 and PD-1) on the surface of T cells act like checkpoints or brakes, preventing the cells from attacking cancer cells. Checkpoint inhibitors block these checkpoints, allowing the T cells to fight cancer (Ott, Hodi, & Robert, 2013). Monoclonal antibodies (a form of targeted therapy) identify abnormalities on cancer cell surfaces, bind to them and mark them for the immune system, while cancer vaccines help to recognize cancer cells and stimulate the immune system. Cytokines including interleukins, interferons and colony-stimulating factors all increase the immune system's reaction to cancer cells (S. Lee & Margolin, 2011).

1.2.3 Allogeneic stem cell transplantation

For many relapsed hematologic malignancies, allogeneic hematopoietic stem cell transplantation (allo-HSCT) is the only potentially curative treatment (Figure 1.2.1). It consists of a preparative conditioning treatment, which often comprises high-dose chemotherapy, sometimes combined with total body irradiation. This conditioning has three aims: It is used to eradicate the malignancy, to make space to allow engraftment of the donor stem cells and to prevent host-versus-graft reactions via immunosuppression (Vriesendorp, 2003). These conditioning treatments damage and kill the patient's immune cells and often lead to immunodeficiency and myeloablation. The conditioning treatment is followed by the stem cell transplantation whereby hematopoietic stem cells (harvested from the bone marrow, peripheral blood or umbilical cord blood) from a healthy donor are given to the patient intravenously. The stem cells travel to the bone marrow and restore hematopoiesis, differentiating into the different types of blood cells. Engraftment is usually defined as the first of three consecutive days with an absolute neutrophil count $> 0.5 * 10^9$/L. The time of engraftment depends on different factors such as the source of the graft. It usually occurs around three to four weeks after transplantation. However, complete restoration of the immune system can take up to two years.

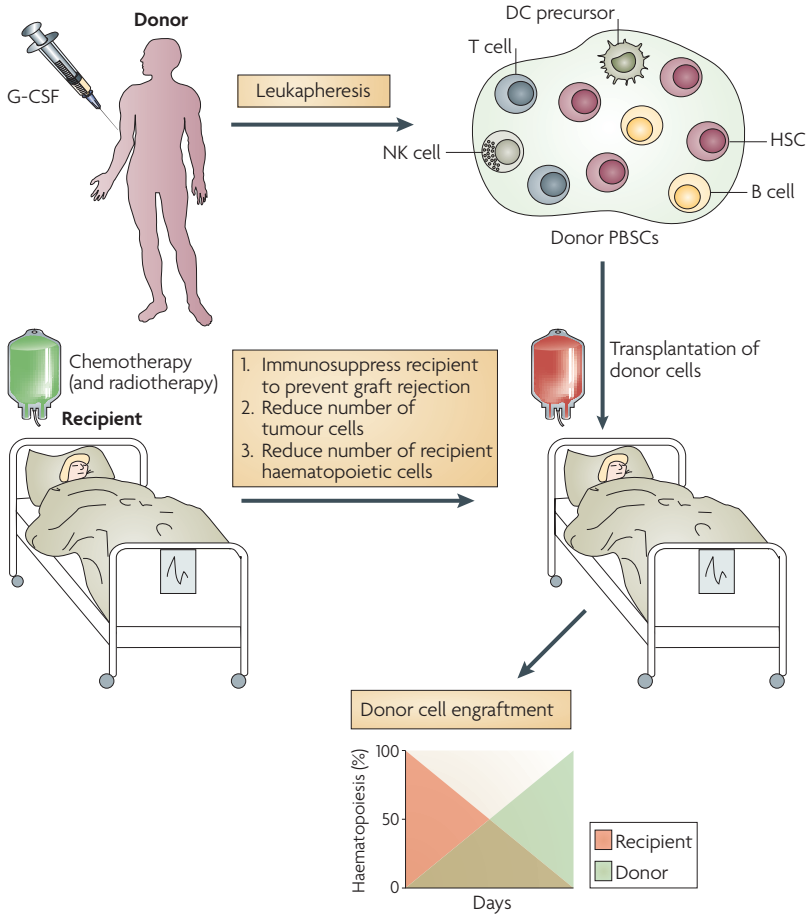


Figure 1.2.1: Allogeneic hematopoietic stem cell transplantation. Donor stem cells are collected after treatment with granulocyte colony-stimulating factor (G-CSF). Recipients receive a treatment, which has different goals (prevent graft rejection, reduce the number of tumor cells and create a niche for engraftment). The patient receives the peripheral-blood stem cells (PBSCs) and engraftment takes place. DC: dendritic cell; HSC: hematopoietic stem cell; NK: natural killer (Shlomchik, 2007).

While the patient's immune cells have failed at recognizing and fighting the cancer cells, immune cells from the donor might succeed at eradicating residual cancer cells, an effect called graft-versus-cancer (GVC) or graft-versus-tumor (GVT) (Blazar, Murphy, & Abedi, 2012). However, donor T cells can also respond to proteins on the surface of host cells, especially to human leukocyte antigens (HLA) which are encoded by the major histocompatibility complex (MHC), resulting in an immunological disorder typically referred to as graft-versus-host disease (GvHD) (Ferrara, Levine, Reddy, & Holler, 2009). Class I HLA (A, B and C) proteins are expressed on nearly all nucleated cells. Consequently, donor selection is mainly based on their HLA profile with preferentially choosing an identical donor, although allo-HSCT is also feasible with near-identical donors. However, these HLAs are highly polymorphic and in addition to GvHD development, HLA disparity is

also associated with graft failure, delayed immune reconstitution and mortality (Park & Seo, 2012). Even with a matched donor, many allo-HSCT receivers develop GvHD, which is caused by differences in the minor histocompatibility antigens. As GvHD prophylaxis, T-cell depletion, monoclonal antibodies and immunosuppression by mycophenolate mofetil, cyclosporine and/or tacrolimus is often applied (Ferrara et al., 2009). As the patients are immunocompromised due to the treatment, they are prone to severe infections, which is the most common cause of mortality after allo-HSCT (Sahin, Toprak, Atilla, Atilla, & Demirer, 2016). Therefore, additional supportive treatment includes antibiotic and antifungal prophylaxis.

1.2.4 Side effects

1.2.4.1 Mucositis

Mucositis is probably the most common complication of anticancer treatment. Each year, it affects over two million people worldwide and occurs in approximately 40 % of patients receiving standard chemotherapy and nearly all the patients undergoing high-dose chemotherapy and allo-HSCT or radiation for head and neck cancers (Elting et al., 2003; Elting, Cooksley, Chambers, & Garden, 2007; Legert, Remberger, Ringdén, Heimdahl, & Dahllöf, 2014). It is the painful inflammation and ulceration (disintegration of tissue) of the mucosal lining of the GIT. Symptoms evolve gradually, from atrophic lesions to deeper ulceration, causing severe pain and diarrhea (Van Sebille et al., 2015; Sonis, 2004). In particular, the lesions in the oral cavity can be very painful, requiring parenteral nutrition and treatment with narcotics and antibiotics, which may lead to a longer hospital stay. This condition often implies reducing the dosages of chemotherapeutics or postponing treatment. Ultimately, this can decrease the chance for remission or cure and compromise survival outcome in cancer patients.

In 2004, Sonis introduced a model which divides mucositis pathogenesis into five phases (Figure 1.2.2): initiation with formation of reactive oxygen species (ROS), primary damage response including induction of messenger molecules, signal amplification, disruption of the epithelial barrier with ulceration and finally healing lead by cell proliferation (Sonis, 2004).

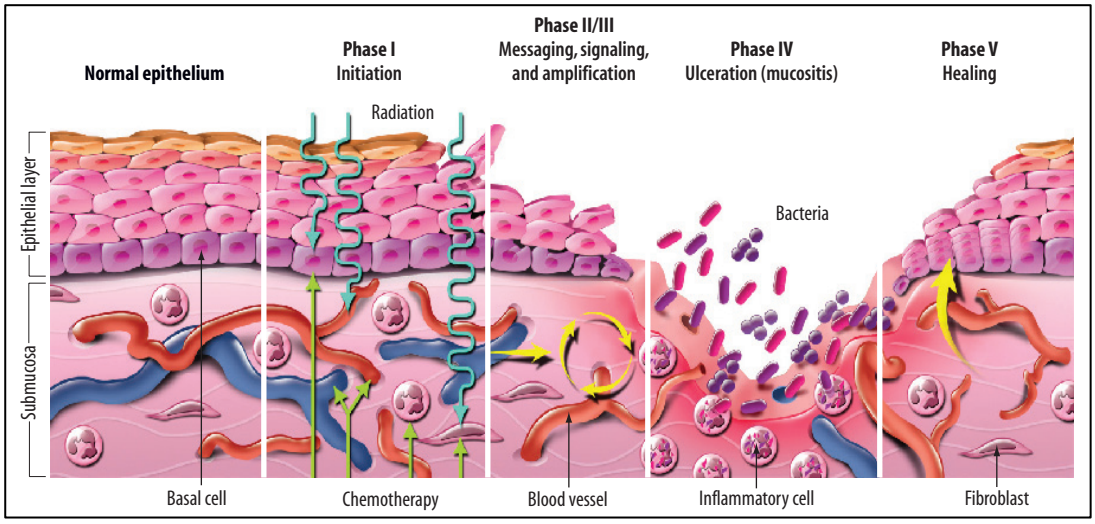


Figure 1.2.2: The five stages in the pathobiology of mucositis. (Sonis, 2004)

More precisely, cytotoxic agents and radiation lead to damage and apoptosis of quickly dividing cells, such as cells of the oral and GIT mucosa. Additionally, ROS are formed, damaging cells and tissues, while also stimulating macrophages and inflammatory pathways. Altogether, different control mechanisms including the transcription factor NF-κB are activated, which mediates gene expression and synthesis of different inflammatory molecules including pro-inflammatory cytokines (such as TNF-α, IL-6 and IL-1β) and tissue injury begins. These cytokines further activate NF-κB in other cells, marking the signal amplification phase followed by ulceration, the formation of painful lesions. At this stage, the loss of mucosal integrity increases the risk for infection, especially in neutropenic patients. In most cases, ulcers heal spontaneously once the cancer treatment has been halted (Al-Dasooqi et al., 2013; Sonis, 2004).

1.2.4.2 Graft-versus-host disease

Graft-versus-host disease is an immunological disorder that mostly affects the skin, GIT and liver and is the main complication after allo-HSCT (Ferrara et al., 2009). The severity of acute GvHD (aGvHD) is scaled according to the involvement of the three mainly affected organs with overall grades ranging from I (mild) to IV (very severe). In 1966, Billingham defined three requirements for development of GvHD: (1) the graft must contain immunologically competent cells; (2) the recipient must express antigens that are not present in the donor; and (3) the recipient must be incapable of immunologically eliminating the transplanted cells (Billingham, 1966).

In general, aGvHD is defined as GvHD occurring within 100 days after allo-HSCT while GvHD arising later is defined as chronic GvHD. Symptoms of aGvHD include skin rash,

nausea, watery or bloody diarrhea, vomiting and severe abdominal pain. Chronic GvHD comprises different symptoms as for example skin dispigmentation, nail dystrophy, vaginal sclerosis, weight loss and pericarditis (Ferrara et al., 2009).

Different factors influence the risk of developing GvHD, such as the source of the graft (Bensinger, 2013), recipient age (Hahn et al., 2008), sex disparity between donor and recipient (Flowers et al., 2011; Jagasia et al., 2012), related or unrelated donor (Flowers et al., 2011) and number and kind of HLA-mismatches (Flowers et al., 2011). GvHD occurs in 50-80 % of patients where there is an HLA mismatch, but even with HLA identity, 40 % of recipients develop aGvHD, due to differences in the minor histocompatibility antigens (Ferrara et al., 2009; Tabbara, Zimmerman, Morgan, & Nahleh, 2002).

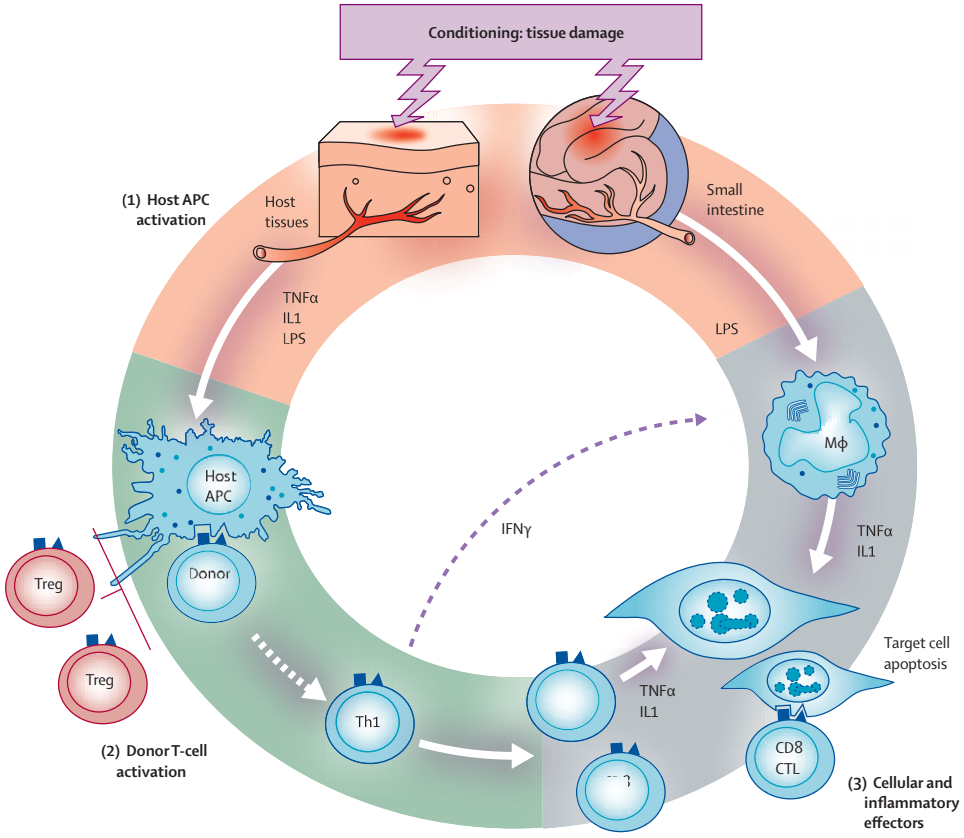


Figure 1.2.3: The pathophysiology of aGvHD. APC: antigen presenting cell; TNFα: tumor necrosis factor α; IL 1=interleukin 1; IFN γ=interferon γ; LPS=lipopolysaccharide; Treg=regulatory T cell; Th1=T helper 1 cell; CTL=cytotoxic T lymphocyte; Mϕ=macrophage (Ferrara et al., 2009).

GvHD occurs when donor T cells respond to proteins on the host cells such as the HLAs, which are highly polymorphic and individual-specific. Pathophysiology of aGvHD can be divided into three phases (Figure 1.2.3). Conditioning damages patient cells and causes

release of inflammatory cytokines (such as TNF-α, IL-1 and IL-6), which leads to activation of antigen-presenting cells (APCs, such as dendritic cells, macrophages, Langerhans cells and B cells) (Ferrara et al., 2009; Harris, Ferrara, & Levine, 2013). In the second step, host APCs activate mature donor cells, producing T helper cell type 1 (T_H1) cytokines (like interferon γ, IL-2 and TNF-α). In the third phase, T_H1 cells activate proliferation of activated cytotoxic T lymphocytes and natural killer cells. Local tissue injury is amplified by the interplay of these cellular mediators and inflammatory agents.

1.2.5 Influence of the microbiome on anticancer treatment side effects and on treatment outcome

A relatively new area of investigation is whether the GIT microbiome influences the occurrence of side effects during anticancer treatment and the outcome of therapy. Especially interesting in this context are development of mucositis and GvHD. The main affected organs in both cases are body parts, which are highly colonized by bacteria, such as the oral cavity and the GIT, which points towards a possible involvement of bacteria in development of those side effects.

According to the generally accepted model introduced by Sonis, the GIT microbiome plays no role in the development and pathophysiology of mucositis (Sonis, 2004). A review by van Vliet and co-authors however, suggests an important role of the commensal GIT microbiome in the development of inflammatory digestive tract diseases, including mucositis (van Vliet et al., 2010). Their review suggests different pathways in which the intestinal microbiome can influence development of mucositis, which are listed hereafter. Some bacteria or bacterial parts decrease NF-κB activation, herewith they decrease production of inflammatory cytokines and influence the inflammatory process. Commensal bacteria increase tight junction strength, influence the composition of the mucus layer and contribute to epithelial repair. Therefore, they improve the epithelial barrier function in different ways. Moreover, the resident microbiota in a healthy intestine regulate the expression of immune effector molecules such as IgA (van Vliet et al., 2010). On the other hand, anticancer treatment can damage the GIT epithelium, reduce intestinal integrity and lead to a loss of barrier function. Tissue damage and translocation of microbes and microbial products can elicit an inflammatory response, aggravating mucositis.

A systematic review by Touchefeu et al. indicates that cytotoxic and radiation therapy leads to important changes in the composition of the GIT microbiota, generally with a decrease in *Bifidobacterium*, *Clostridium* cluster XIVa, *Faecalibacterium prausnitzii* and an increase in Enterobacteriaceae and *Bacteroides* (Touchefeu et al., 2014). It is suggested that these changes might be linked to development of mucositis, diarrhea and bacteremia.

As mentioned above, studies have established that NF-κB is implicated in regulation of mucositis and is being upregulated in the signal amplification phase (Logan et al., 2008; Stringer et al., 2013). Different bacterial molecules such as lipopolysaccharides (LPS) and flagellin activate TLRs, which, in turn upregulate the NF-κB pathway. Changes in the composition of the GIT microbiome and following activation of TLRs and NF-κB which is usually down-regulated by commensals, might link the microbiota to development of mucositis. Prophylactic fluconazole mouthwash has been shown to decrease the rate of severe mucositis (Rao et al., 2013), indicating that also fungi can amplify mucositis.

According to the generally accepted pathophysiology of GvHD, the microbiome could also be involved in its development or amplification. Both in the first and in the third phase (as described in the previous section), pathogen-associated molecular patterns (PAMPs) such as LPS could lead to further activation of immune cells and secretion of pro-inflammatory cytokines (Ferrara et al., 2009). Additionally, the damaged intestinal mucosa could allow translocation of PAMPs and thereby trigger additional inflammatory cytokine production causing apoptosis.

Intensive cancer treatment results in damage to the GIT and decreases its epithelial barrier function, which allows microorganisms or bacterial products to enter the blood circulation (Fuji, Kapp, & Einsele, 2014; van der Velden et al., 2013). Immune response in form of inflammatory cytokine production, followed by sepsis or systemic infections is a consequence. Therefore, it used to be a quite common approach to apply total or selective gut decontamination on patients receiving allo-HSCT for prevention of GvHD development and infection. This is based on fundamental murine studies by van Bekkum and colleagues in which they treated mice with irradiation and allogeneic bone marrow transplantation. Less severe GvHD respectively no GvHD development was observed in mice treated with bacterial decontamination respectively when working with germ-free (GF) mice (Van Bekkum & Knaan, 1977; Van Bekkum, Roodenburg, Heidt, & Van der Waaij, 1974). However, in clinical practice, incomplete rather than complete decontamination is often achieved (Holler et al., 2014). In addition, newer studies have shown that a diverse microbiome correlates with a lower occurrence of GvHD and a higher survival rate (Jenq et al., 2015; Y Taur et al., 2014).

Several recent studies, which have focused on the analysis of the microbiota composition before and after allo-HSCT observed a drastic loss in microbial diversity in the patient after receiving the treatment. This shift was more pronounced in patients who developed GIT GvHD (Holler et al., 2014; Jenq et al., 2012). Lower diversity has been linked to higher mortality, especially to death due to transplant related causes like infection and

GvHD (Y Taur et al., 2014) whereas higher diversity, likewise, has been associated with reduced GvHD-related mortality (Jenq et al., 2015). A retrospective study indicated that treatment of neutropenic fever with different antibiotics (imipenem-cilastatin and piperacillin-tazobactam) was linked to increased GvHD-related mortality at 5 years (Shono et al., 2016). Similarly, a recent retrospective study linked early use of antibiotics to higher transplant-related mortality (Weber et al., 2017). Although it is not exactly clear how diversity relates to the lower mortality, one aspect might be that the normal commensals present in a diverse community enhance resistance against infection by GIT pathogens by colonization resistance, filling specific ecological niches within the GIT. Likewise, a diverse microbial community might contain SCFA producers, which contribute to induction of colonic T_{regs}.

The onset of GvHD was often accompanied by an increase in the abundance of members of the genus *Enterococcus* (Biagi et al., 2015; Holler et al., 2014) as well as by a reduction in abundance of Clostridiales or specific members of this order, including *Faecalibacterium* spp. and *Ruminococcus* spp. (Biagi et al., 2015; Jenq et al., 2012; Simms-Waldrip et al., 2017). A higher abundance of another Clostridiales member, the genus *Blautia* has been linked to reduced GvHD-related mortality (Jenq et al., 2015). Many members of the order Clostridiales (including the genera mentioned beforehand *Faecalibacterium* spp. and *Ruminococcus* spp.) are known butyrate-producers (Morrison & Preston, 2016). Butyrate is known to reinforce the epithelial barrier and inhibit inflammatory response (Canani et al., 2011). In this way, depletion of these health-promoting bacteria might result in higher inflammation rate and thus, onset of GvHD. In a recent murine study, butyrate was the main component found to be linked to mitigation of GvHD, overall survival and improved intestinal epithelial cell junction integrity (Mathewson et al., 2016). The same effect could be observed when administering the mice 17 rationally selected strains of Clostridia that have been shown to increase amounts of butyrate and induction of T_{reg} cells (Atarashi et al., 2013; Mathewson et al., 2016).

Severe GIT GvHD has been seen to correlate with a loss in Paneth cells in biopsies (Levine et al., 2013). α-defensins, the principal AMPs secreted by Paneth cells, play an important part in keeping a balanced microbiome by having bactericidal activity against noncommensals. Eriguchi et al. observed that Paneth cells in mice were affected by GIT GvHD (Eriguchi et al., 2012). Therefore, GIT GvHD itself could lead to a loss in Paneth cells, which would in turn promote a dysbiotic state, allowing the overgrowth of pathogens and thus, result in an overall decrease in diversity.

Although the attention to commensal intestinal bacteria in human health and disease is growing, the link between the intestinal microbiome and anticancer treatment-induced mucositis and GvHD is still obscure. Understanding the clinical relevance of imbalance in the GIT microbiome and further the possibility to restore homeostasis could improve the treatment of cancer patients and therefore decrease the overall mortality from cancer.

1.2.6 Influence of the gastrointestinal microbiome on the efficacy of anticancer treatments

Recently, the GIT microbiome has been shown to influence efficacy of anticancer components (Iida et al., 2013; Sivan et al., 2015; Vétizou et al., 2015; Viaud et al., 2013). Cyclophosphamide (CTX) is an important cancer drug. This alkylating agent is used for treating different kinds of cancer but also autoimmune diseases. It has the ability to initiate antitumor immune responses by stimulating the generation of T_H1 and T_H17 cells which in turn control cancer outgrowth (Viaud et al., 2011). Viaud and colleagues showed that treatment with antibiotics, for example vancomycin and colistin, decreases the efficacy of CTX against MCA205 sarcoma. Also, size reduction of the tumor due to CTX treatment was greater in SPF mice than in GF mice (Viaud et al., 2013). In fact, CTX decreases intestinal barrier integrity and facilitates translocation of specific Gram-positive bacteria into secondary lymphoid organs. Here, these bacteria stimulate an anticancer immune response (Viaud et al., 2013) (Figure 1.2.4).

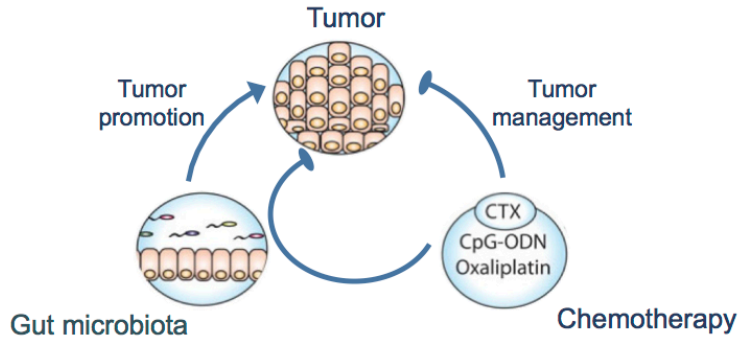


Figure 1.2.4: Influence of the GIT microbiome on tumor promotion and tumor management. GIT microbiome can drive tumor development. On the other hand, it can influence the efficacy of chemotherapeutic drugs, thus, influence tumor management, CTX = cyclophosphamide; CpG-ODN = CpG-oligodeoxynucleotides (Modified from Perez-Chanona & Jobin, 2014).

A similar effect was observed in a study by Iida and colleagues, where mice were pretreated with broad-spectrum antibiotics (vancomycin, imipenem and neomycin) before being inoculated with subcutaneous tumors and treated with immunotherapy or chemotherapy. Comparison of gene expression of three transplantable tumor models in

mice treated with these antibiotics showed a down-regulation of genes related to inflammation, phagocytosis, antigen presentation and adaptive immune response whereas genes related to tissue development, cancer and metabolism were upregulated (Iida et al., 2013). Mice treated with antibiotics and GF mice showed reduced tumor regression and survival when treated with oxaliplatin and cisplatin, two platinum salts, or with CpG-oligodeoxynucleotides (CpG-ODN) (Figure 1.2.4). Oxaliplatin damages tumor cells with ROS (Laurent et al., 2005). Mice treated with antibiotics and GF mice displayed immune cells producing less ROS-generating enzymes. The GIT microbiome seems to play a role in preparing the immune system and the antitumor immune response.

Recently, two studies have shown that the efficacy of ICIs is also influenced by specific bacteria. In both studies, efficacy of ICI was reduced in GF mice and mice treated with antibiotics. Polysaccharide A produced by different species of *Bacteroides* spp. was found to induce maturation of intratumoral dendritic cells and T_H1 cells. An increased antitumoral immune response could be observed when administering these bacteria to wild-type mice (Vétizou et al., 2015). In another study, different bifidobacteria were positively linked to the amount of antigen-presenting cells in tumors, leading to a better response to ICI treatment (Sivan et al., 2015).

In vitro studies and subsequent *in vivo* murine studies have shown that the efficacy of several chemotherapeutic agents is increased or decreased due to interaction with specific bacteria. The available data suggests that specific bacterial enzymes are affecting the efficacy of the drugs (Lehouritis et al., 2015). These recent results indicate that the microbial composition and products in the intestine influence the efficacy of anticancer components, often by influencing the immune system. The ability to increase the efficacy of the anticancer treatment regimens would allow to use a less intense treatment or fewer cycles, possibly resulting in less adverse side effects and ultimately in a better outcome for patients. Altogether, this proves the importance of the microbiome in cancer treatment and therapeutic outcome and the perspective of incorporating microbiome in cancer treatment, but also the need for further investigations in the field. The GIT microbiome could be seen as an additional factor influencing cancer treatment and its outcome and should be modulated in a way that it could improve treatment outcome. However, currently, this is not possible. Rather, prophylactic broad-spectrum antibiotics are used, which do not distinguish between commensals and pathobionts and therefore can culminate in dybiosis, potentially eliminating bacteria that could have a beneficial effect on treatment outcome.

1.3 Antibiotics, antibiotic resistance genes and alternative treatments

Patients undergoing anticancer treatment are especially prone to infections, as they are often neutropenic or completely immunodeficient. Thus, antibiotic treatment is needed as prophylaxis or for the treatment of infections. However, the use of antibiotics may trigger dysbiosis by selection for pathogens expressing antibiotic resistance genes (ARGs) and lead to the emergence of multi-drug resistant (MDR) bacteria.

The discovery of antibiotics has greatly impacted modern medicine and has helped to control infectious diseases that were major causes for morbidity and mortality. As many antibiotics and antimicrobial molecules are produced by microorganisms, resistance mechanisms have also been present in microbial communities, long before humans have started to use antibiotics in clinical settings. For example, ARGs have been identified in a cave in New Mexico that had been isolated for 4 million years (Bhullar et al., 2012), in an 11th century AD mummy (Santiago-Rodriguez et al., 2015), and in 30,000 year-old permafrost sediments (D'Costa et al., 2011). This shows that antibiotic resistance predates selective pressure of antibiotics usage in clinical settings. However, the recent prevalence in the use of antibiotics drives the emergence of higher rate of antibiotic resistance by microorganisms and the appearance of MDR organisms, a problem that we are facing today (Goossens, Ferech, Vander Stichele, & Elseviers, 2005; Sun, Klein, & Laxminarayan, 2012). This in turn has led to an increased risk of fatal infections in susceptible patients, as for example in immune compromised cancer patients during therapy, who are usually treated with antibiotics as prophylactic agents. Also, antibiotics given to livestock or sprayed on fruit trees eventually affect the environmental microbiome and favour propagation of resistance mechanisms (Angenent, Mau, George, Zahn, & Raskin, 2008; Levy, 2002).

Antibiotic resistance has recently been listed to one of the greatest threats to human health in the World Economic Forum Global Risks report (World Economic Forum, 2016), with 700,000 people dying of resistant infections each year (Neill, 2016). In Europe alone, around 25,000 patients die yearly as a result of MDR bacterial infections and this adds to € 1.5 billion hospital and treatment costs (ECDC/EMEA, 2009). It is estimated that by 2050, MDR pathogens will kill 10 million people a year (Neill, 2016). Therefore, there is an urgent need for discovery of new antibiotics and modification of existing antibiotics. Similarly, there is a high interest in new databases, which allow detection of ARGs in sequencing datasets. Updated databases as well as new tools, allowing detection of the genes in large MG datasets are currently being developed and published (Jia et al., 2016; Lakin et al., 2016).

Antibiotic resistance can be intrinsic or acquired. Intrinsic resistance is inherent to an organism and includes the absence of the antibiotic target, low cell permeability preventing access to the target, efflux or inactivation of the drug (Cox & Wright, 2013). Resistances can be acquired by mutation of pre-existing target genes or by horizontal gene transfer of ARGs via transformation, transduction or conjugation (Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O, Piddock, 2015; Huddleston, 2014). Although it is generally assumed that acquired ARGs represent a fitness cost, studies have shown that resistant populations can persist even four years after antibiotic treatment, showing that carrying these genes does not lead to a reduced fitness (Jakobsson et al., 2010; Jernberg, Löfmark, Edlund, & Jansson, 2007; Sjölund, Wreiber, Andersson, Blaser, & Engstrand, 2003). Studies have shown that the number of ARGs within an individual's microbiome increases over time as the person ages and is exposed to more antibiotics. But also in subjects that have never been treated with quinolones (one group of broad-spectrum antibiotics), 40 % of the bacteria in humans and animals carry corresponding resistance genes (Field & Hershberg, 2015). Consequently, the human GIT microbial community, even of healthy individuals without antibiotic treatment, comprises a diversity of different ARGs (Bartoloni et al., 2004; Gibson, Forsberg, & Dantas, 2014; M. O. A. Sommer, Church, & Dantas, 2010).

Alternatives to currently used traditional antibiotics could consist of bacteriocins, microcins, structurally nanoengineered antimicrobial peptide polymers (SNAPPs) and phage therapy. Bacteriocins and microcins largely target the same bacterial compartments and processes as conventional antibiotics (Cavera, Arthur, Kashtanov, & Chikindas, 2015). An advantage is that broad- but also narrow-spectrum bacteriocins exist, however as for traditional antibiotics, emergence of resistances is a threat (Cotter, Ross, & Hill, 2013). SNAPPs are antimicrobial agents, which interact with microbial membranes and therefore elicit less resistance development than traditional antibiotics, which act inside of bacterial cells. SNAPPs have recently been shown to be active against gram-negative bacteria, incuding different colistin-resistant and multidrug-resistant pathogens (Lam et al., 2016). One advantage of phages as treatment is that they usually only target a limited range of bacterial strains (Kuntz & Gilbert, 2017). Concerns regarding this treatment method however include the possibility of emergence of resistance mechanisms or a system able to degrade viral DNA (such as clustered regularly inter-spaced palindromic repeats (CRISPRs)), as well as the risk of phages carrying and transferring ARGs into different bacteria via transduction (Nobrega, Costa, Kluskens, & Azeredo, 2015). Also, phages could be recognized and rapidly removed by the immune

system. Thus, several questions and concerns have to be solved before phage therapy could be applied to humans.

Commensal bacteria prevent from overgrowth of pathogens in different ways, i.e. by activating the innate immune defences, via nutrient competition, occupation of specific niches or production of AMPs. Antibiotic treatment allows overgrowth of pathogens not only by selecting antibiotic-resistant bacteria, but also by killing bacteria that provide colonization resistance. Opening of ecological niches allows proliferation of pathogens. Loss of commensals also results in higher availability of nutrients, especially of sugars, which favors pathogen expansion. One important example is *Clostridium difficile* infection (CDI), which is a leading cause of hospital-associated diarrhea, causing an estimated cost of 5.4 billion US dollars annually (Desai et al., 2016). The two biggest risk factors for this infection are antibiotic intake and exposure to this organism (Slimings & Riley, 2014; Surawicz et al., 2013). CDI is also a major concern in patients who undergo an allo-HSCT (Alonso et al., 2012) with reported ratios of around 15 % or even 18 % (Chopra et al., 2011; Trifilio, Pi, & Mehta, 2013). It is generally accepted that a severe perturbation (dysbiosis, caused by intake of broad-spectrum antibiotics) allows the expansion of pathogenic strains of *C. difficile*. The first treatment of CDI includes different antibiotics (metronidazole or vancomycin), however, many patients develop recurrent chronic CDI. Fecal microbiota transplantation (FMT) is recently being applied in the treatment of patients with recurrent CDI, where cure rates of 90 % have been reported (Brandt et al., 2012). Intestinal domination of vancomycin-resistant *Enterococcus* (VRE) is often observed in patients undergoing allo-HSCT as result of nosocomial acquisition or extensive exposure to broad-spectrum antibiotics. VRE are one of the most frequently encountered bloodstream infections in patients undergoing allo-HSCT and this has been associated with high mortality in these patients (Kamboj et al., 2010; Y. Taur et al., 2012; Y Taur et al., 2014).

1.4 Aims of this work

The GIT microbiome has recently gained a lot of interest and there is growing evidence, that this complex community of organisms can affect the efficacy of anticancer treatment, development or severity of severe side effects and influence overall treatment outcome.

1.4.1 Identify changes in the GIT microbiome during treatment

The first objective of this work is to describe in detail how the GIT microbiome changes during different chemotherapy regimens and myeloablative treatment followed by allo-HSCT. As there is a tight interaction between the host intestinal mucosa and immune system and the commensal microbiome, changes in the status of the patient might be reflected in dynamics of the GIT microbiome community composition before, during and after treatment.

1.4.2 Discern how the GIT microbiome might be involved in the development of anti-cancer treatment side effects

Recent studies indicate that the GIT microbiome might be implicated in development or severity of important treatment side effects, such as mucositis and GvHD (Biagi et al., 2015; Holler et al., 2014; Touchefeu et al., 2014). Patterns within the microbial community might be linked to development of these side effects. Importantly, this knowledge could help in the formulation of measures to prevent mucositis and GvHD development.

1.4.3 Assess if and how metagenomic and metatranscriptomic sequencing could be used in personalized medicine

In addition to taxonomic profiling of the GIT microbiome by 16S and 18S rRNA gene amplicon sequencing, this study includes MG and MT sequencing, which allows the description of the functional capacity and expression within the community. Additional information gathered from these methods could help in expanding our understanding of the complex interactions within the GIT microbiome and with its host. Ultimately, this knowledge might help in adjusting therapy regimens, thereby improving the overall outcome.

To assess the structural and functional changes within the GIT microbiome throughout anticancer treatment, fecal samples from patients following different treatment regimens were collected at different time points throughout this treatment. An integrated-omics approach was used, including 16S, 18S rRNA gene amplicon sequencing, metagenomics and metatranscriptomics.

Phylogenetic marker gene sequencing (16S and 18S rRNA gene amplicon sequencing) allows taxonomic profiling of the microbial community. In addition to marker gene sequencing, this study also includes MG sequencing, as it provides information about the functional potential of the microbial community or individual taxa within the community. This gives further insights into the interaction between the host and the microbiome and analysis of for example antibiotic resistance genes, which might play an important role in this setting.

Although culture-independent methods such as rRNA gene amplicon sequencing and whole MG sequencing are not yet part of clinical routine, these applications could in the future play an important role in diagnostics and in tailoring treatments to the specific individual needs. Drastic reductions in sequencing costs (Wetterstrand, 2016) but also in time and higher throughput sequencing technologies (Neelapu & Surekha, 2016) now permit to apply next-generation sequencing in the clinical setting as diagnostic tool. Recent studies have demonstrated how MG sequencing could be applied for example to identify and quickly treat foodborne outbreaks of *Salmonella* (Quick et al., 2015) and bioinformatic tools which help to identify pathogens have been developed (Flygare et al., 2016; Greninger et al., 2015).

This work gives first insights into the detailed information and knowledge that these methods provide and shows the importance of understanding the microbiome dynamics for future personalized strategies for anticancer treatment.

2 Materials and methods

Parts of the materials and methods section are taken and modified from a manuscript that has been submitted to Translational Research. The respective manuscript is attached in the appendix:

Appendix A.1: Anne Kaysen, Anna Heintz-Buschart, Emilie E. L. Muller, Shaman Narayanasamy, Linda Wampach, Cédric C. Laczny, Norbert Graf, Arne Simon, Katharina Franke, Jörg Bittenbring, Paul Wilmes, Jochen G. Schneider. (2017) Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic stem cell transplantation. *Translational Research.* (in revision).

2.1 Study participants and collection of fecal samples

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). Fecal samples were collected by the pediatric oncology department and the hematology and oncology department of the Saarland University Medical Center. All patients, respectively their parent or guardian signed a written informed consent. Fecal samples were collected at different time points (TPs) throughout the anticancer treatment. The sampling plan is illustrated in Figure 2.1.1. The patients were grouped into different treatment groups (TGs) according to their treatment. Pediatric oncology patients with different types of cancer were included into TG1 – TG3. Patients within TG1 received a low intensity treatment, TG2 received a standard intensity treatment and TG3 underwent an intensive treatment followed by an autologous stem cell transplantation (auto-HSCT). In general, the first sample was taken at admission, before treatment was started. The second sample was collected after the first cycle of treatment at the nadir, the point where the leukocyte count was at its lowest. The third sample was taken when the blood counts recovered before the next cycle of treatment was started and the last sample was taken after the end of the therapy. For patients within TG3 or patients who were switched to TG3, samples were collected before and after the auto-HSCT.

Patients recruited in the hematology and oncology department were grouped into TG4, whose treatment includes an intensive immune ablative treatment followed by an

allogeneic stem cell transplantation (allo-HSCT). Due to several reasons such as the overall health status of the patient or poor nutritional intake (due to loss of appetite, mucositis or GvHD), it was especially difficult to collect samples within this TG. Whenever possible, the first sample was collected before allo-HSCT (generally during the conditioning treatment). The next sample was taken directly after allo-HSCT, the third sample was taken around the engraftment period (around one month after allo-HSCT) and follow-up samples were taken at later time points.

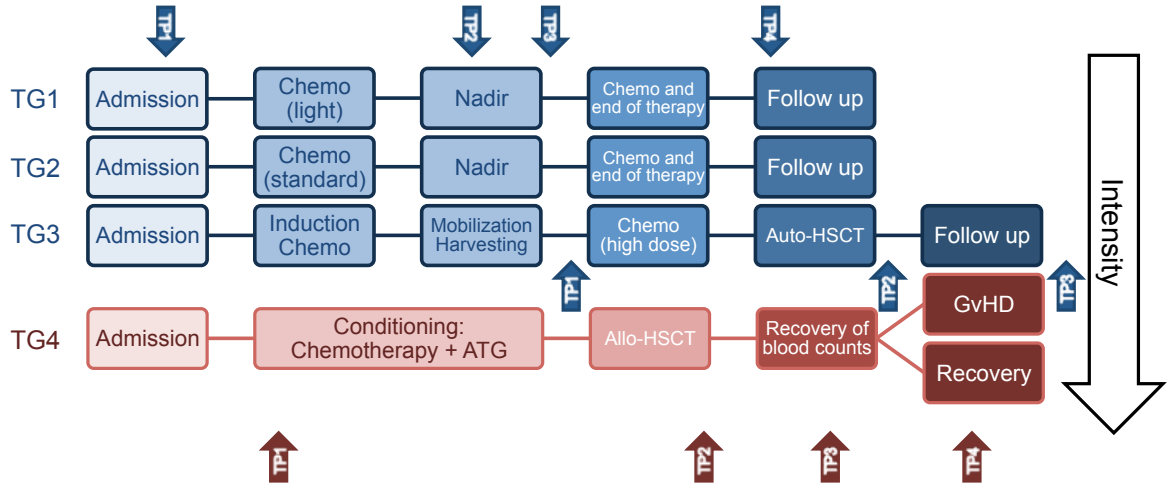


Figure 2.1.1: Sampling plan. Vertical arrows indicate sampling time points throughout treatment. Blue boxes and arrows represent treatment phases for pediatric patients grouped into three different treatment groups (TG1 – TG3). Red boxes and arrows indicate the treatment phases for adult patients with hematologic malignancies. Abbreviations: TG= treatment group; TP= time point; chemo=chemotherapy; ATG= antithymocyte globulin; Auto-HSCT: autologous stem cell transplantation; Allo-HSCT: allogeneic stem cell transplantation; GvHD: graft-versus-host disease.

18 pediatric patients were recruited with the majority belonging to TG2. Anthropometric and clinical information on the study participants including age, sex, underlying disease, antimicrobial treatment, occurrence of severe mucositis as well as overall treatment outcome were collected (Table 2.1.1). As antimicrobial prophylaxis, trimethoprim/sulfamethoxazole (cotrimoxazole) was given twice a week, to prevent *Pneumocystis jirovecii* pneumonia.

27 adult hematology patients were recruited. Anthropometric and clinical information on the study participants including age, sex, underlying disease, stem cell donor type, conditioning regimen, antimicrobial treatment, aGvHD development and grade as well as overall treatment outcome were collected (Table 2.1.2).

Table 2.1.1: Anthropometric and clinical information of the pediatric study cohort

Patient	Sex	Age	Underlying disease[a]	Treatment group	Mucositis (min. grade 3)	Antimicrobials[b]	Outcome 1.5 years after start of treatment
P01	f	17	Hodgkin's lymphoma	2	no	T/S, P-T	alive
P02	f	13	Ovarian germ cell tumor	2	no	T/S	alive
P03	f	16	Nephroblastoma	2	yes	T/S	alive
P04	f	3	Nephroblastoma	1	yes	T/S, P-T	alive
P05	f	8	Nephroblastoma	2 & 3	no	T/S, C	deceased day 521, tumor progression
P07	f	4	Nephroblastoma	1	no	T/S, C	alive
P08	m	4	Medulloblastoma	2 & 3	no	T/S	deceased day 364, respiratory failure
P09	m	3	Neuroblastoma	3	yes	T/S, P	alive
P10	f	14	Ewing sarcoma	2	no	T/S, P-T, AF	deceased day 504, tumor progression
P11	f	12	Ewing sarcoma	2	no	T/S, P-T	deceased day 383, tumor progression
P12	m	19	Relapsed ALL	2	no	T/S	alive
P13	m	4	ALL	2	no	T/S	alive
P14	f	8	Germ cell tumor of the brain	2 & 3	no	T/S	alive
P15	m	8	ALL	2	no	T/S	alive
P16	m	14	NHL	2	no	T/S	alive
P17	m	14	Large cell NHL	2	yes	T/S	alive
P18	f	11	Hodgkin's lymphoma	2	no	T/S	alive
P19	m	12	Relapsed ALL	2	yes	T/S	alive

[a]: ALL: acute lymphoblastic leukemia, NHL: non-Hodgkin's lypmphoma
[b]: T/S: trimethoprim/sulfamethoxazole, P-T: piperacillin-tazobactam, C: cefuroxime, P: phosphomycin, AF: antifungal

Table 2.1.2: Anthropometric and clinical information of the study cohort recruited in the hematology department

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
A02	m	46	lymphoma	MRD	FluBuCy	AF, M, P-T, V, other	-	deceased day 17, relapse
A03	m	56	lymphoma	MRD	FluBuCy	AF, F, M, P-T, other	-	deceased day 66, 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 day 275, 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 day 268, 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 day 212, pneumonia
A10	m	54	AML	MRD	BuCy	F, M, P-T	Skin I°, GIT II°	alive
A11	m	58	AML	MMUD	BuCy	AF, M, P-T, V, other	Skin 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
A14	f	22	lymphoma	MUD	FluBuCy	P-T, V, other	-	alive
A16	m	67	AML	MRD	BuCy	AF, M, P-T, V, other	Skin 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 day 184, GvHD
A19	f	58	myeloma	MUD	Treo/Flu	F, M, P-T	-	deceased day 39, 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
A25	m	52	lymphoma	MUD	FluBuCy	AF, F, M, P-T, V, other	Skin III°	alive
A27	m	56	CML	MRD	BuCy	AF, F, M, V, other	**Skin III°, GIT IV°, liver III**	deceased day 53, GvHD
A29	f	60	ALL	MRD	TBI/Cy	M, P-T, V	-	alive
A34	m	52	AML	MMUD	Treo/Flu	F, M, V, other	-	alive
A35	m	45	lymphoma	MMUD	FluBuCy	AF, M, P-T, V, other	-	alive
A41	f	54	AML	MMUD	Treo/Flu	F	-	alive
A44	f	24	lymphoma	MUD	FluBuCy	AF, F, M, V, other	-	deceased day 4, sepsis

[a]: AML: acute myeloid leukemia, ALL: acute lymphoblastic leukemia, CML: chronic myeloid leukemia

[b]: MRD: matched related, MUD: matched unrelated, MMUD: mismatched unrelated

[c]: Bu: busulfan, Cy: cyclophosphamide, Flu: fludarabine, FLAMSA: fludarabine, cytarabine, amsacrine, Treo: treosulfan

[d]: AF: antifungal, F: fluoroquinolone, M: meropenem; P-T: piperacillin-tazobactam, V: vancomycin

[e]: Organ involvement, stages according to Glucksberg et al. (1974)

[f]: Bold: GvHD with summed stages ≥ 4 considered as severe GvHD

2.2 Extraction of biomolecules from fecal samples

Fecal samples were immediately flash-frozen in liquid nitrogen on-site and preserved at -80 °C to ensure integrity of the biomolecules of interest. The general workflow from sample collection to data integration is illustrated in Figure 2.2.1.

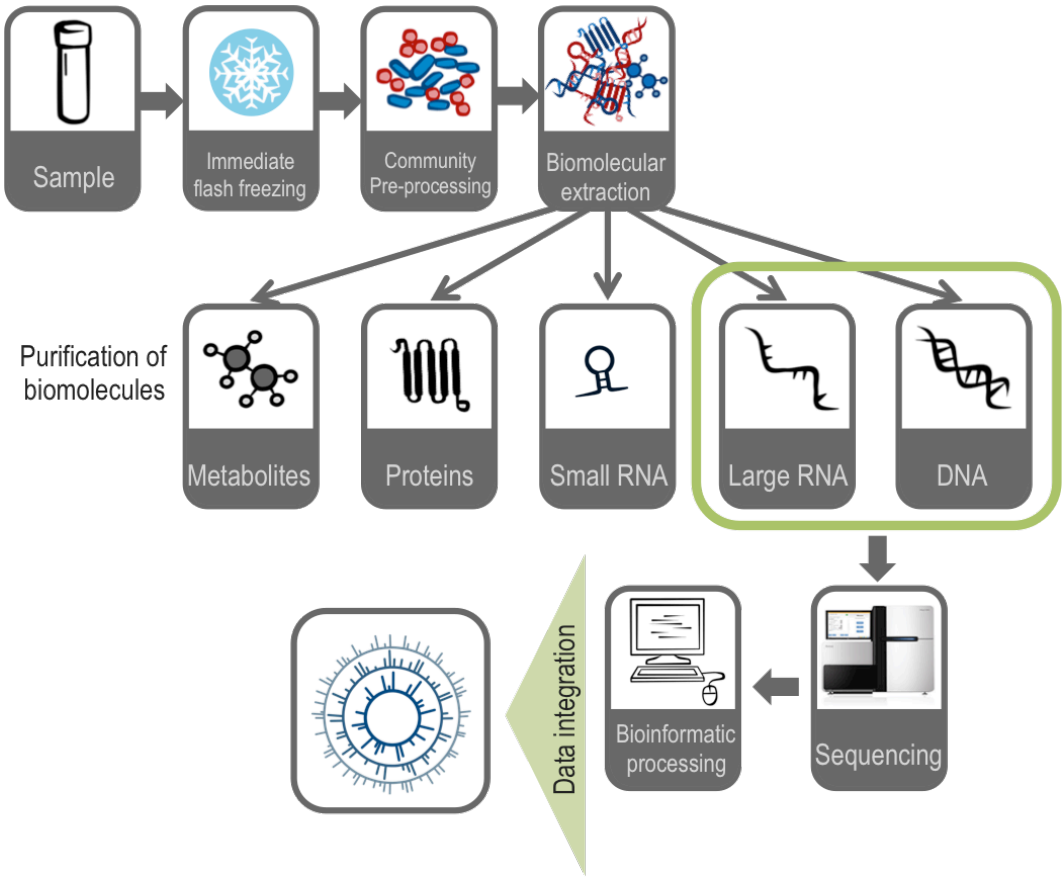


Figure 2.2.1: General workflow from sample collection to data integration. Original graphics by Linda Wampach and from http://support.illumina.com/sequencing/sequencing_instruments/hiseq_2500.html (picture of the sequencer).

Biomolecules 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 in Roume, Muller, et al., 2013; Roume, Heintz-Buschart, Muller, & Wilmes, 2013. Briefly, the mixture was homogenized and biomolecules were extracted using the AllPrep DNA/RNA/Protein kit (Qiagen) as described in Roume, Heintz-Buschart, et al., 2013. To increase the overall yield, DNA fractions were supplemented with DNA extracted from 200 mg subsamples using the PowerSoil DNA isolation kit (MO BIO).

To samples with a very low yield, an adjusted protocol of the PowerSoil DNA isolation kit (MO BIO) was applied, which includes a few additional steps in comparison to the standard protocol. It starts with a phenol-chloroform-isoamyl treatment (200 µl) and an incubation at 65 °C for 10 minutes. Additional wash steps with ethanol were included. As this protocol elutes DNA and RNA, it was always followed by an RNA digestion. 1 µl of a 100 µg/ml RNase A solution was added per 20 µl of DNA and incubated at 37 °C for 30 min. Quality and quantity of the DNA were verified using 1 % agarose gel electrophoresis with Tris Acetate-EDTA running buffer (Sigma-Aldrich) and the MassRuler DNA Ladder Mix (Thermo Fisher Scientific), NanoDrop 2000c spectrophotometer (Thermo Fisher Scientific) or Qubit fluorometer (Life Technologies, Carlsbad, CA, USA), following the manufacturers' recommendations. Quality and quantity of RNA extracts were verified using the Agilent 2100 Bioanalyzer (Agilent Technologies) or LabChip GXII Touch HT (PerkinElmer) following the manufacturers' recommendations. Only fractions with RNA integrity number (RIN, Agilent Technologies) > 7 or RNA Quality Score (PerkinElmer) > 6 were sequenced. Extracted biomolecules were stored at -80 °C until shipment to the sequencing center on dry ice.

2.3 16S and 18S rRNA gene amplicon sequencing

Amplification and paired-end sequencing of extracted and purified DNA was performed at the Groupe Interdisciplinaire de Génoprotéomique Appliquée (GIGA, Belgium). Sequencing with 2 * 300 nt was performed using the V2 MiSeq kit on a MiSeq platform (Illumina). The V4 region of the 16S rRNA gene, which allows resolution of bacteria and archaea, was amplified using the primers listed in Table 2.3.1. Furthermore, the V4 region of the 18S rRNA gene, which allows resolution of eukaryotes, was amplified and sequenced using the primers listed in Table 2.3.1.

Table 2.3.1: Primers used for 16S and 18S rRNA gene amplicon sequencing.

Amplicon	Primer name	Sequence	Reference
16S rRNA	515F	GTGBCAGCMGCCGCGGTAA	L. W. Hugerth, Wefer, et al., 2014
16S rRNA	805R	GACTACHVGGGTATCTAATCC	Herlemann et al., 2011
18S rRNA	574*f	CGGTAAYTCCAGCTCYV	L. W. Hugerth, Muller, et al., 2014
18S rRNA	1132R	CCGTCAATTHCTTYAART	L. W. Hugerth, Muller, et al., 2014

The number of samples that have been sequenced per department are indicated in Table 2.3.2.

Table 2.3.2: Number of patients and sequenced samples per department.

Department	Patients	Samples
Pediatrics	18	60
Hematology	27	78

2.4 16S and 18S rRNA gene amplicon sequencing data analysis

16S rRNA gene sequencing reads were processed using the less operational taxonomic units scripts (LotuS) pipeline (version 1.34) (Hildebrand, Tadeo, Voigt, Bork, & Raes, 2014) with default parameters. Within this pipeline, sequences are filtered with a C++ program, simple demultiplexer (sdm) and grouped into 'high-' and 'mid-' quality sequences. The high quality sequences are used by the software UPARSE (Edgar, 2013) to generate operational taxonomic unit (OTU) clusters, while the medium quality sequences are mapped to the established list of OTUs. Default sdm options for MiSeq sequences were used, consisting of a minimum sequence length (after primer removal and trimming) of 170 and a minimal average quality of 27 respectively 20 for high and medium quality sequences. Processed reads were clustered into OTUs, taxa with similar amplicon sequences at 97 % identity level. For taxonomic assignment, the Ribosomal Database Project (RDP) classifier (Q. Wang, Garrity, Tiedje, & Cole, 2007) was used. OTUs with a confidence level below 0.8 at the domain level were filtered out, as well as OTUs that were not represented by more than 10 reads in any given sample. OTUs that were unclassified on phylum level were selected and aligned to the NCBI nucleotide collection (nr/nt) database using the BLAST webservice (using program blastn). OTUs aligning to human, fungal or other eukaryotic genomes were removed. The final hand-curated OTU table included 2,789 unique OTUs. Samples with a low number of reads (overall < 5,000 reads) were removed, concluding with an average (± standard deviation) of 208,000 ± 73,000 sequencing reads in each of the 134 samples.

To process the 18S rRNA gene sequencing reads, a workflow specifically designed to process reads that are not overlapping was used (L. Hugerth, 2015). For classification, the PR2 database (Chevenet, Brun, Bañuls, Jacq, & Christen, 2006; Chevenet, Croce, Hebrard, Christen, & Berry, 2010; Guillou et al., 2013) was employed. After processing, OTUs represented by less than 10 reads in all samples were removed. OTUs belonging to the taxa Craniata and Streptophyta were removed, since they were most likely derived from human sequences or ingested food. Unclassified OTUs were aligned to the NCBI

nucleotide collection (nr/nt) database using the BLAST webservice (using program blastn). The results were hand-curated, removing OTUs that were related to human sequences, food, plants, viruses or phages. After removing samples from the 18S rRNA gene sequencing data due to a low number of reads (overall < 1,000 reads), 1381 unique OTUs were represented with an average (± standard deviation) of 69,000 ± 77,000 reads in each of the 103 samples.

2.4.1 Diversity and statistical analyses

Alpha-diversity describes within-sample diversity of a community. Different indices such as richness, evenness and the Shannon diversity index can be calculated to determine the diversity of a community within a sample. Richness quantifies how many different species or OTUs are present within a community, without taking into account different relative abundances. Evenness measures how species or OTUs are distributed in a community. The Shannon (or Shannon-Weaver) diversity index takes into account both richness and evenness. It is calculated as follows:

$$H = -\sum_{i=1}^{S} p_i \, ln \, p_i$$

where p_i is the relative abundance of species *i* and *S* is the total number of species in the community (Shannon & Weaver, 1948).

Statistical analyses and plots were generated in R (version 3.2.1) (R Development Core Team, 2008). Alpha-diversity was determined at the OTU level, by calculating the Shannon diversity index and the Chao1 richness estimator after rarefaction, using the vegan package (Oksanen et al., 2015). The 16S and 18S rRNA gene sequencing data were rarefied to the lowest number of respective reads for any sample. Plots were generated using the R base graphics or the ggplot2 package (Wickham, 2009).

Comparison of diversity and richness was carried out using the non-parametric Wilcoxon rank sum test, or, when applicable, Wilcoxon signed-rank test. When *p* values < 0.05 were observed, groups were considered as statistically significantly different. Differential analysis of taxa based on 16S rRNA gene sequencing data was performed using the DESeq2 package (Love, Huber, & Anders, 2014) and significant differences on taxonomic levels were determined using the Wald test, after multiple-testing adjustment with the false discovery rate (FDR) method after Benjamini and Hochberg. For further statistics, the Fisher's exact test, Spearman's rank correlation coefficient test and the Kolmogorov-Smirnov test of the stats package were used (R Development Core Team, 2008).

2.5 Metagenomic and metatranscriptomic sequencing

Metagenomic (MG) and metatranscriptomic (MT) sequencing of the extracted DNA and RNA fractions was conducted by GATC Biotech AG, European Genome and Diagnostics Centre, 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 or 125 bp paired-end approach on an Illumina HiSeq 2500 using HiSeq V3 reagents. A total of 97 samples from 17 pediatric and 21 hematology patients were sequenced. For 41 of those samples, MT and MG sequencing was possible, while for 56 samples, only the metagenome could be sequenced (Table 2.5.1).

Table 2.5.1: Number of patients, samples, MGMT and MG only datasets per department.

Department	Patients	Samples	MGMT	MG
Pediatrics	17	54	32	22
Hematology	21	43	9	34

Detailed timelines indicating for each patient the day at which a sample was taken, as well as the sequencing method that could be applied to this sample are illustrated in Figure 2.5.1 (pediatric department) and Figure 2.5.2 (hematology department). Vertical color shadings in Figure 2.5.2 indicate the time periods defined as specific TPs. 16S and 18S rRNA gene sequencing was performed for each sample, while it is indicated in the figures only if MG and MT sequencing were not possible.

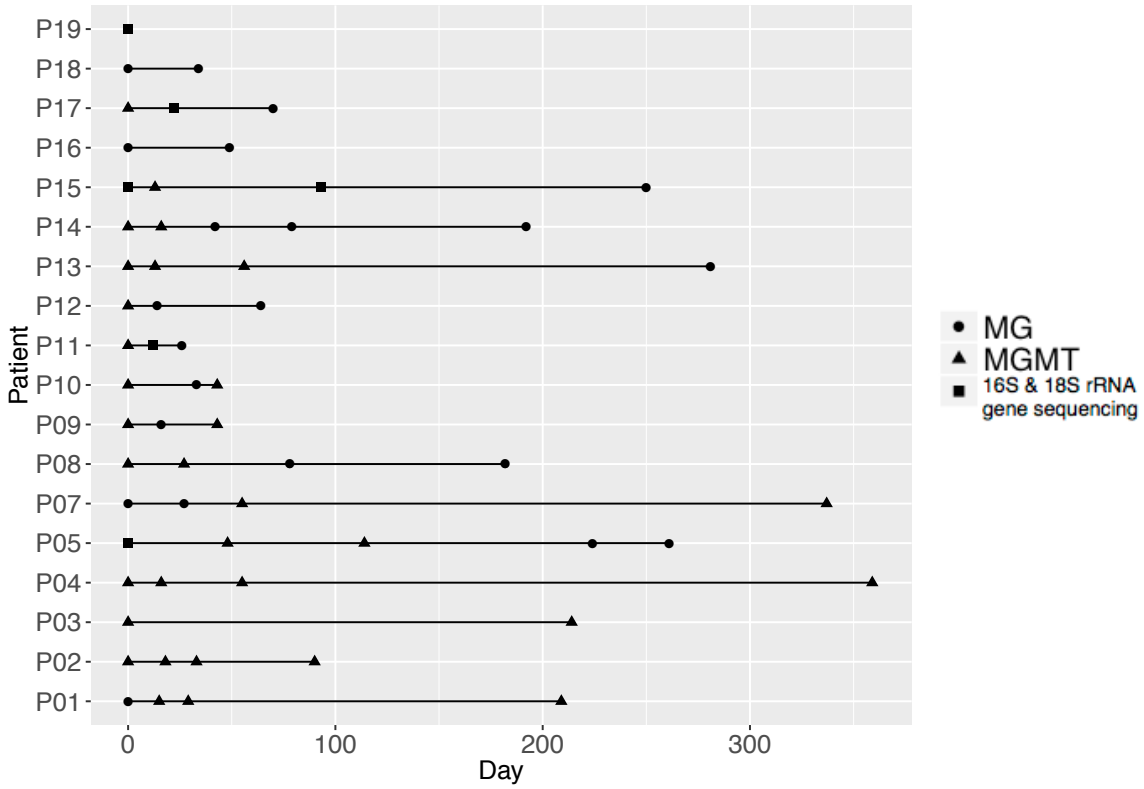


Figure 2.5.1: Sampling timeline for pediatric patients. Timelines indicate the days samples were taken (in relation to the first sample). 16S and 18S rRNA gene amplicon sequencing was performed on all samples and additionally, applied metagenomic (MG) and metatranscriptomic (MT) sequencing is indicated for each sample.

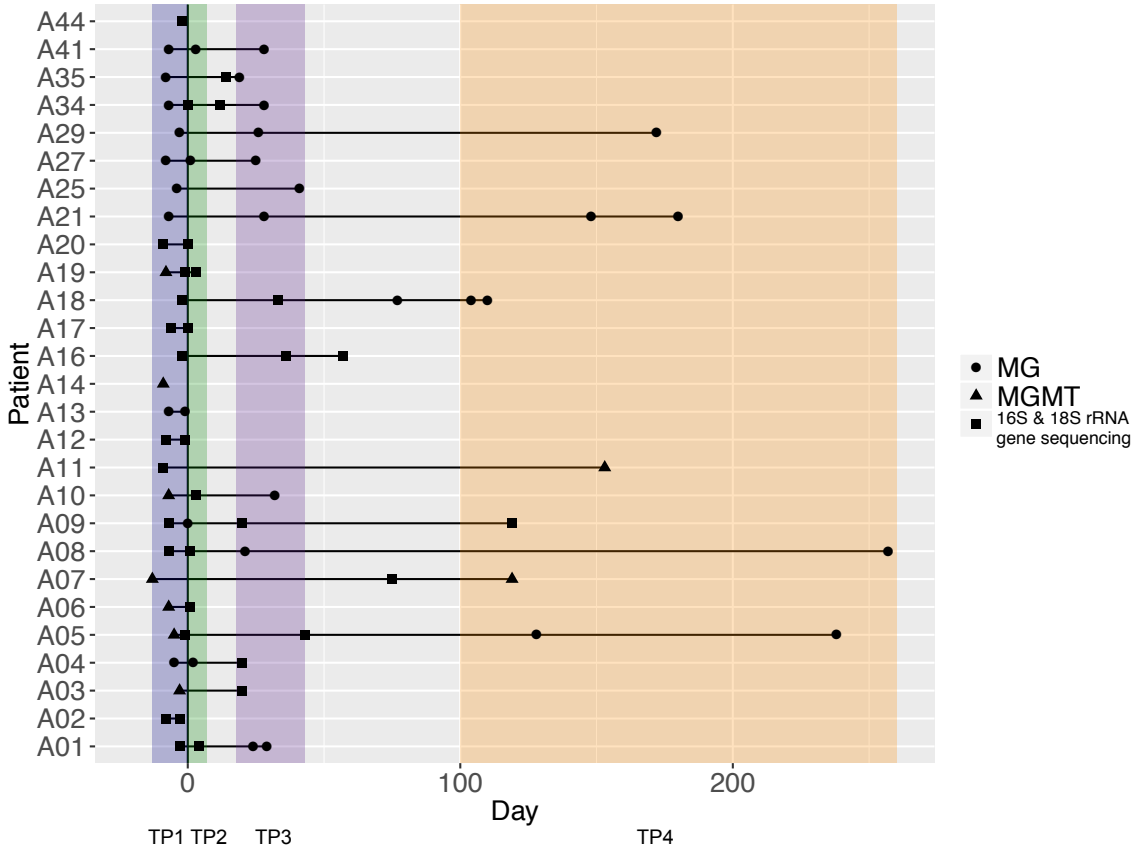


Figure 2.5.2: Sampling timeline for patients recruited at the hematology department. The timeline indicates the day the samples were taken (in relation to allo-HSCT at day 0). 16S and 18S rRNA gene amplicon sequencing was performed on all samples and additionally, applied metagenomic (MG) and metatranscriptomic (MT) sequencing is indicated for each sample.

2.6 Processing and assembly of metagenomic and metatranscriptomic datasets

MG and MT datasets were processed using the Integrated Meta-omic Pipeline (IMP) (Narayanasamy et al., 2016a). For datasets from patient A07 (section 3.3), version 1.1 was used. All other patient datasets were processed using version 1.2.2. Published human GIT microbiome MG and MT read data from four healthy individuals was obtained from the NCBI Sequence Read Archive [MG: SRX247379, SRX247391, SRX247401, SRX247405; MT: SRX247335, SRX247345, SRX247349, SRX247340] (Franzosa et al., 2014). These sequencing reads were processed using IMP version 1.2.1 (Narayanasamy et al., 2016a). Data from the individuals 'X310763260', 'X316192082', 'X317690558' and 'X316701492' are in the following referred to as the 'reference healthy microbiome', averaged as 'RHMs' or individually referred to as 'RHM1', 'RHM2', 'RHM3' and 'RHM4'.

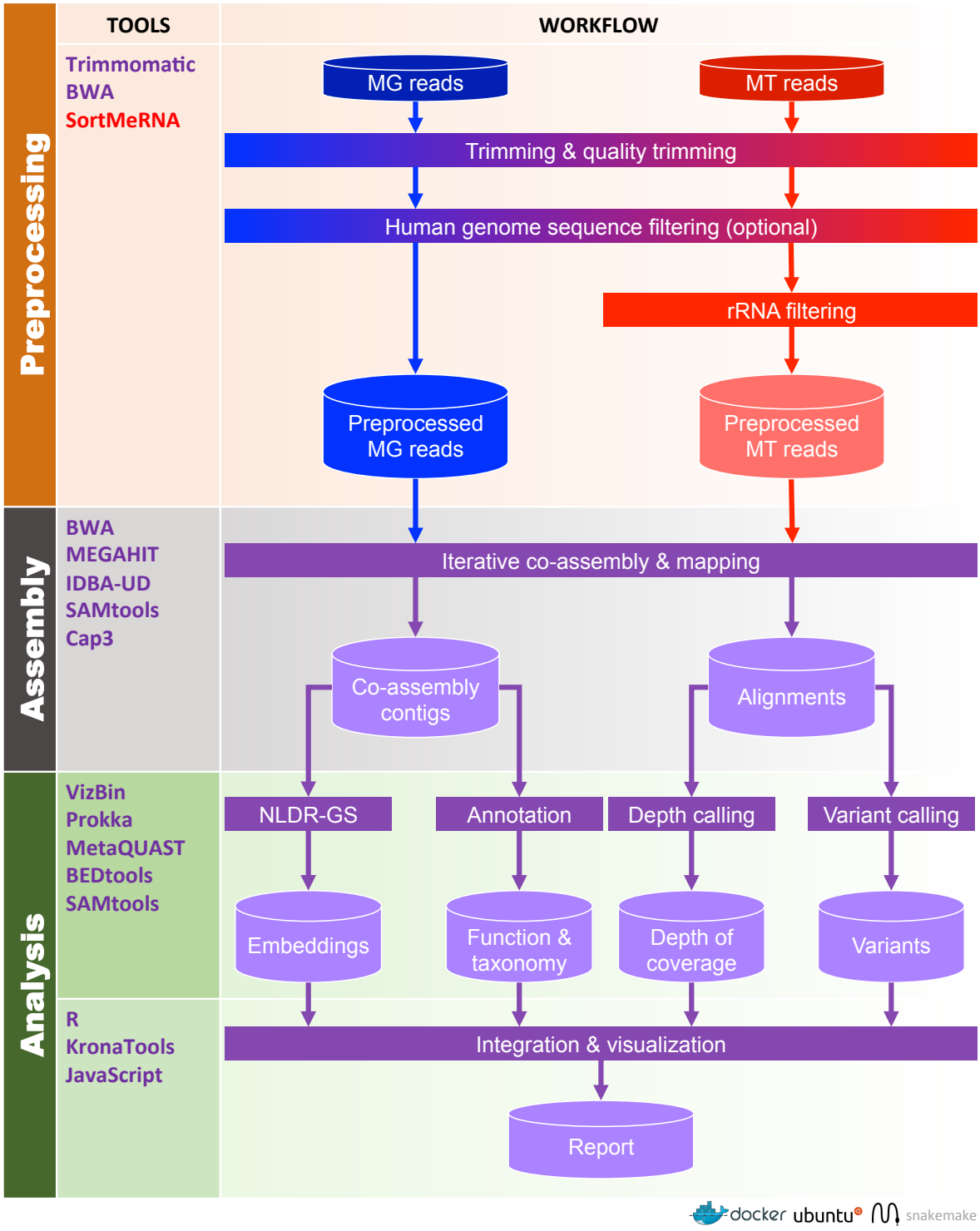


Figure 2.6.1: Overview of the IMP workflow. Cylinders represent input and ouput. Rectangles represent processes. MG: metagenomic, MT: metatranscriptomic, NLDR-GS: genomic signature non-linear dimensionality reduction (Narayanasamy et al., 2016b).

This fully automated pipeline comprises preprocessing of the reads, a customized iterative co-assembly of MG and MT reads, augmented visualizations of the resulting contigs

(Laczny et al., 2015; Laczny, Pinel, Vlassis, & Wilmes, 2014), variant calling, functional annotation of predicted genes, and provides information on the depth of coverage of genes based on the MG and MT datasets (Figure 2.6.1).

In a first step, raw reads are being preprocessed, including removal of adapter sequences, bad quality reads, ribosomal RNA (for MT datasets) and reads mapping to the human genome. This is followed by the assembly of preprocessed MG and MT reads. This 'iterative assembly' includes several assembly rounds, each time including reads that were unmappable in the previous step, thereby increasing the amount of information used in the final set of contigs, the final assembly. The initial MG and MT reads are mapped back onto the final contigs, resulting alignment information is used in different analysis procedures such as variant calling and determination of depth of coverage. The average depth of coverage D_x of a gene or contig x is determined both for the metagenome and the metatranscriptome by calculating the average number of reads mapping to each nucleotide within a gene, respectively in a contig.

$$D_x = \frac{\sum r_x}{length_x}$$

where r_x is the number of reads mapping to a gene or contig x at each nucleotide.

Here, gene expression of a gene x is calculated as the ratio of average metatranscriptomic depth of coverage to the average metagenomic depth of coverage for individual genes x.

$$E_x = \frac{D_x(MT)}{D_x(MG)}$$

An additional output of the pipeline are VizBin maps, which will be explained in the following paragraph.

2.7 Population-level binning of contigs from the co-assembly and inference of population size

To analyze and compare the population-level structure of the microbial communities based on the assembled genomic information, contigs were binned into (partial) population-level genomes. Using VizBin (Laczny et al., 2015, 2014), 2D embeddings based on BH-SNE (Barnes-Hut Stochastic Neighborhood Embedding) of the pentamer frequency profiles of all contigs of at least 1,000 nt were produced, as part of IMP. In these embeddings, contigs with similar genomic signatures (pentamer frequencies) are closer together, hence, individual clusters of contigs represent individual populations (Muller et al., 2014). Population-level clusters were selected following the method

described in Heintz-Buschart et al., 2016. In short, the automatic workflow is based on DBSCAN (Ester, Kriegel, Sander, & Xu, 1996). Clusters were selected based on neighborhood points in a first step. Using a collection of 107 single-copy marker genes or 'essential genes' (Albertsen et al., 2013), completeness and homogeneity of clusters were determined. By analyzing the MG depth of coverage of these essential genes, clusters with multiple copies of the same genes were further divided. Resulting bins are referred to as 'population-level genomes' in the following.

Within a community, the relative population size of a cluster (*i*) was determined by dividing the number of MG reads mapping to the contigs forming this cluster (c_i), by the total number of MG reads mapping to all the contigs used in the assembly (*C*) according to the following formula:

$$N_i = \frac{c_i * 100}{C}$$

2.8 Taxonomic affiliation of reconstructed population-level genomes

Taxonomic affiliations of population-level genomes were determined using complementary methods. Contigs forming the population-level genomes were first aligned to NCBI nucleotide collection (nr/nt) database using the BLAST webservice (Madden, 2002). Parameters were left at default (using program megablast), and the output was analyzed using the MEtaGenome ANalyzer (MEGAN) (D. Huson, Mitra, & Ruscheweyh, 2011). Whenever the *rpoB* gene could be recovered within a population-level genome, the closest neighbor was determined in the nucleotide collection (nr/nt) database using the MOLE-BLAST webservice (Boratyn et al., 2014). Additionally, AMPHORA2 (Wu & Scott, 2012) was used to identify the taxonomic affiliation of up to 31 bacterial or 104 archaeal phylogenetic marker genes.

2.9 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 (Bankevich et al., 2012) (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 variant density.

2.10 Sequence comparison of population-level genomes

The average nucleotide identity (ANI) calculator ('ANI Average Nucleotide Identity', at http://enve-omics.ce.gatech.edu/ani/; Goris et al., 2007) 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 (Aziz et al., 2008) using standard parameters.

2.11 Detection of antibiotic resistance genes

Antibiotic resistance genes (ARGs) within a community or population were searched against Resfams version 1.2 (Gibson et al., 2014) using HMMer version 3.1b2 (Eddy, 2011). 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.

2.12 Variant identification and density

Variants were identified in population-level reassembled genomes using SAMtools mpileup (H. Li et al., 2009) with default settings, which include the calling of single nucleotide variants (SNVs) as well as the identification of small insertions/deletions (indels). The output was filtered using a conservative heuristic established in Eren et al., 2015, 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.

Variant density (*V*) in a population *i* was calculated by dividing the number of single nucleotide variant and indel positions (*P*) by the relative population size, more precisely the ratio between reads mapping to the population-level genomes (c_i) and total number of reads (*C*).

$$V_i = \frac{P_i}{\frac{c_i}{C}}$$

2.13 Extraction, sequencing and analysis of bacterial DNA from a blood culture

A bacterium was isolated from a patient's blood using the BD BACTEC FX system (Becton Dickinson) following the manufacturer's recommendations, by a consultant in medical microbiology from the institute for medical microbiology and hygiene of the Saarland University Medical Center. It was identified as a multidrug-resistant *E. coli*. Subcultures were grown on TSA blood agar plates and on MacConkey agar plates. DNA was isolated from a culture grown in standard growth medium (TSB, BHI or LB) using the Maxwell 16 Tissue LEV Total RNA Purification Kit on a Maxwell 16 MDx Instrument (Promega) according to manufacturer's instructions.
DNA was sequenced on an Illumina MiSeq, 300 bp paired-end at GIGA. The genome was assembled using SPAdes (Bankevich et al., 2012) and ARGs were identified using Resfams, as described previously. Using ANI, nucleotide identity between *E. coli* genomes from GIT samples and the *E. coli* genome from the isolate were assessed. Using PanPhlAn (Scholz et al., 2016) and the provided database including 118 *E. coli* reference strains, their relation was assessed based on their gene set. While the PanPhlAn database includes 31,734 genes, only genes present in 10 or more genomes were included, resulting in 7,845 genes for comparison.

2.14 Virus profiling

The analysis tool ViromeScan (Rampelli et al., 2016) was separately applied to trimmed and filtered MG and MT reads to identify reads that map to eukaryotic viral genomes. Default parameters were used along with the eukaryotic DNA/RNA reference database which includes genomes of viruses that have the human as natural host, in addition to viruses of other vertebrates, invertebrates, fungi algae and plants. The database excluded bacteriophages. The relative abundance of the viral community in each sample was calculated as percentage of reads mapping to viral genomes of the total number of reads.

2.15 Read-based taxonomic analyses

Metagenomic operational taxonomic units (mOTUs) analysis for taxonomic profiling was performed on trimmed and filtered paired MG and MT reads individually using the MOCAT pipeline (Kultima et al., 2012; Sunagawa et al., 2013). This was used with standard parameters and the mOTU.v1.padded reference database including 10 single-copy marker genes (Sunagawa et al., 2013).

2.16 Functional analyses

Within IMP, Prokka was used for functional annotation (Seemann, 2014), with Prodigal for gene prediction (Hyatt et al., 2010). Additionally, KEGG (Kyoto Encyclopedia of Genes and Genomes) orthologous groups (KOs) were annotated as described in Heintz-Buschart et al., 2016 (Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016). The featureCounts tool of the Subread package was used to count (MG or MT) reads mapping to the predicted genes (Liao, Smyth, & Shi, 2014). Differential analysis of functional gene categories and enzymes was performed using the DESeq2 package (Love et al., 2014) and significant differences on functional gene abundances were determined using the Wald test, after multiple-testing adjustment with FDR method after Benjamini and Hochberg. Additionally, KOs were grouped according to their KEGG pathway affiliation. Heatmaps were plotted using the heatmap.2 function from the gplots package.

3 Results and discussion

This chapter is divided into three sections, based on different patient cohorts that are focussed on in the individual sections. The first section focuses on patients from the pediatric department who underwent different anticancer treatments. The second section focuses on the adult hematology patients who underwent an allo-HSCT and the third section focuses on one specific adult patient. Each section first describes the general changes observed in the GIT microbiome community throughout the treatment, including assessment of alpha-diversity and changes on different taxonomic levels. The following parts describe the viral community and functional properties of the GIT microbiome, especially trends observed in the abundance of ARGs. The last part focuses on the possible link between the microbiome and development of mucositis or GvHD.

3.1 Meta-omic analyses of the gastrointestinal tract microbiome in pediatric patients undergoing different anticancer treatments

The first section focuses on the changes within the GIT microbiome of pediatric oncology patients undergoing anticancer treatment with different intensities and is based on 60 fecal samples collected from 18 different patients. After processing and filtering of the 16S rRNA gene amplicon sequences (as described in section 2.4), 58 datasets were left for the following analyses, with 243,000 ± 75,000 (mean ± standard deviation) sequences per sample. Similarly, 67,500 ± 76,000 sequences per sample of the 40 sets of 18S rRNA gene sequencing data were retained. Of 32 samples, MG and MT combined datasets could be produced. Additionally, 22 MG-only datasets were produced. After processing with IMP (as described in section 2.6) which included filtering out low quality reads and reads mapping to the human genome, per dataset, 56,000,000 ± 21,000,000 MG sequences and 50,000,000 ± 15,000,000 MT sequences were kept for the following analyses. Co-assemblies of the preprocessed MG and MT reads were constructed which comprised longer contiguous sequences (contigs). For MG-only datasets, MG-only assemblies were contructed. Within the co-assemblies and MG-only assemblies, 158,500 ± 75,000 genes were predicted.

3.1.1 Patient characteristics and treatment

Anthropometric and clinical information of the ten female and eight male patients included in the following analyses are provided in Table 2.1.1, and a summarized overview is presented in Table 3.1.1. At the start of the treatment, the patients aged between 3 and 19 years (median 11.5). They were grouped into different treatment groups (TGs) according to the intensity of their treatment. Patients within TG1 received a treatment with low intensity, TG2 received a standard intensive treatment and TG3 underwent an intensive treatment followed by an autologous stem cell transplantation (auto-HSCT). Four patients were treated for nephroblastoma, four for acute lymphoblastic leukemia, two for Hodgkin's lymphoma, two for Ewing sarcoma, two for non-Hodgkin's lymphoma, one for ovarian germ cell tumor, one for germ cell tumor of the brain, one for neuroblastoma and one for medulloblastoma. Of the eighteen patients enrolled in the study, two were grouped into TG1, fifteen were grouped into TG2 and one into TG3. Three patients were changed from TG2 to TG3 during the course of their treatment. They first received a standard treatment and later an auto-HSCT. Five patients developed mucositis with minimum grade 3, meaning occurrence of severe, painful ulcers needing pain medication, with impaired intake of solid food. 1.5 years after start of the treatment, fourteen patients were still alive while four had deceased due to tumor progression or respiratory failure.

Table 3.1.1: Summarized anthropometric and clinical information of the pediatric study cohort

Underlying disease	Number of patients	Treatment group	Age
Nephroblastoma	4	2, 1, 2 & 3, 1	16, 3, 8, 4
Acute lymphoblastic leukemia	4	2, 2, 2, 2	19, 4, 8, 12
Hodgkin's lymphoma	2	2, 2	17, 11
Ewing sarcoma	2	2, 2	14, 12
Non-Hodgkin's lymphoma	2	2, 2	14, 14
Ovarian germ cell tumor	1	2	13
Germ cell tumor of the brain	1	2 & 3	8
Neuroblastoma	1	3	3
Medulloblastoma	1	2 & 3	4

3.1.2 Changes in the prokaryotic GIT microbiome in pediatric patients throughout cancer treatment

Based on 16S rRNA gene amplicon sequencing of DNA extracted from 60 fecal samples of the patients, the prokaryotic (bacterial and archaeal) community was assessed. After filtering and removal of samples with a low number of reads (as described in section 2.4), 58 of the sequenced samples were kept for the following analyses. The overall 14 most abundant genera within all 58 samples were identified to get an overview of the

composition of the GIT microbiome of the patients (Figure 3.1.1). The samples were usually dominated by few different genera, the most abundant ones being *Bacteroides*, *Faecalibacterium* and *Parabacteroides*. For some patients, the different samples taken at different time points (TPs) did not reveal drastic differences, but rather an individual specific composition, such as for example P02 revealing a high relative abundance of the genus *Prevotella*, at all four TPs, which expanded over a time period of over three months. For other patients, such as P15, strong differences in the composition between different TPs could be observed.

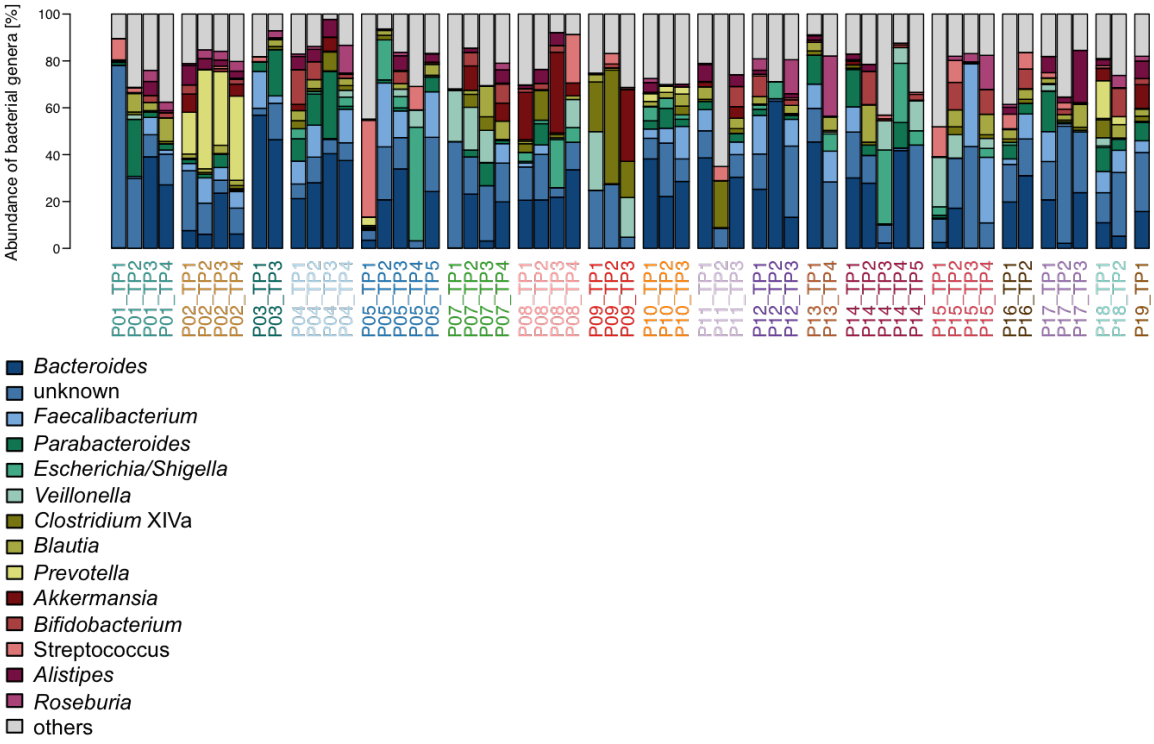


Figure 3.1.1: Relative abundance of the 14 most abundant bacterial genera in fecal samples from pediatric cancer patients, grouped according to patient. Genera which were not comprised in the 14 most abundant genera are combined as 'others'. Operational taxonomic units (OTUs) which could not be classified at the genus level are grouped as 'unknown'. Patient ID and sampling time point (TP) are indicated below each respective bar and are colored according to the patient.

In the following, the samples were grouped according to the TG and according to the TP within the TG (Figure 3.1.2). In TG1 and TG2, the first sample was taken at admission, before treatment was started. The second sample was collected after the first cycle of treatment at the nadir, the point where the leukocyte count is at its lowest. The third sample was taken when the blood counts recovered before the next cycle of treatment was started and the last sample was taken after the end of the therapy. For patients within TG3 or patients who were switched to TG3, samples were collected before and after the

auto-HSCT. No clear specificity of the relative abundance of bacteria distinct to individual TGs or TPs was apparent.

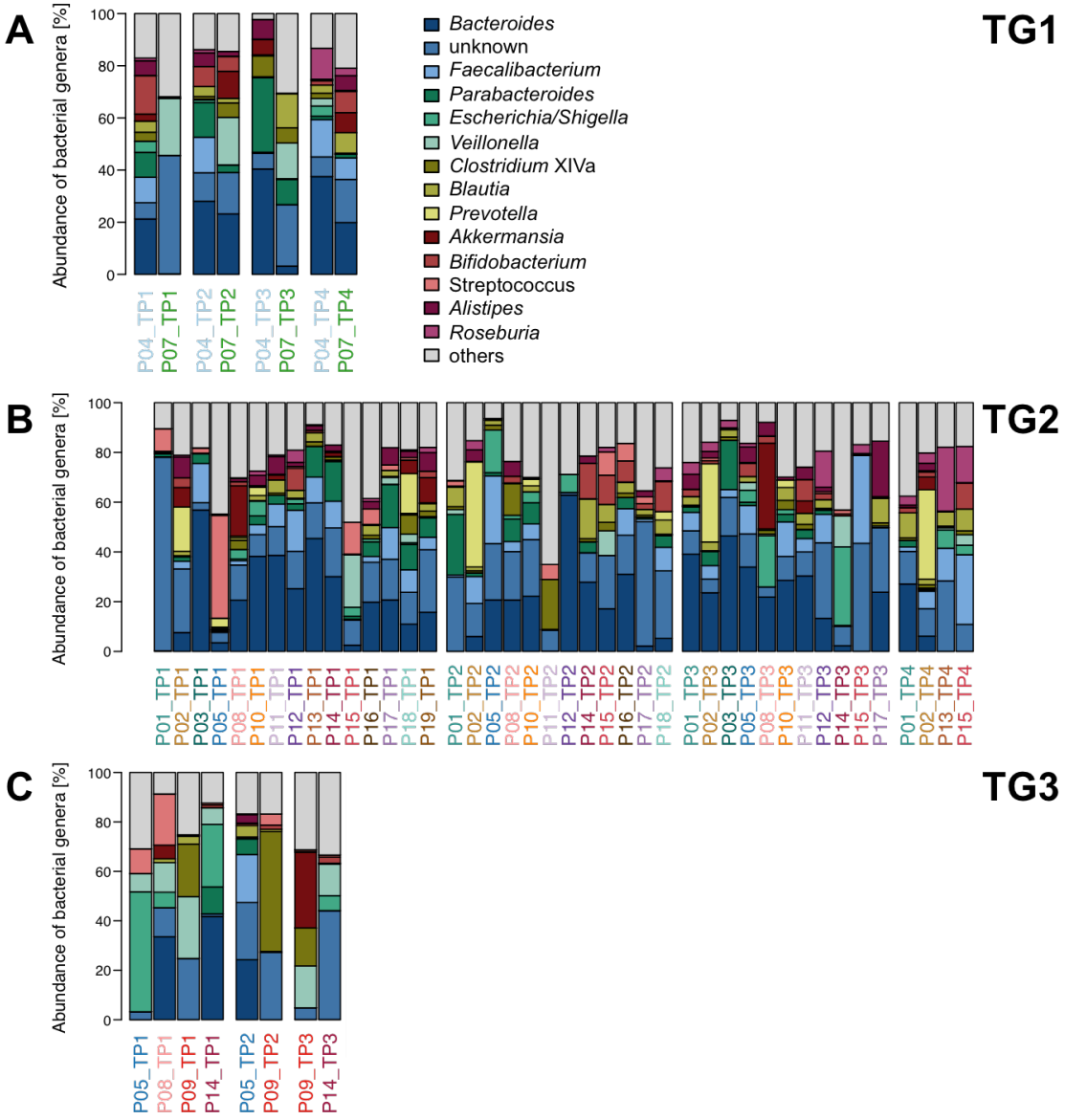


Figure 3.1.2: Relative abundance of the 14 most abundant bacterial genera in fecal samples from pediatric cancer patients. Samples are grouped according to TPs within TG: (A) TG1, (B) TG2 and (C) TG3. Genera which were not comprised in the 14 most abundant genera are combined as 'others'. OTUs which could not be classified at the genus level are grouped as 'unknown'. Patient ID and sampling TP are indicated below each respective bar and are colored according to the patient.

Differentially abundant OTUs between TP1 (before treatment) and TP2 (lowest leukocyte count after first cycle of treatment) were assessed. For this analysis, patients from TG2 were included, as this TG comprised the highest number of patients. Between those TPs, 27 differentially abundant bacterial OTUs were identified (absolute $\log_2$ fold change ≥ 1, FDR-adjusted *p* value < 0.05). Table 3.1.2 displays the 13 OTUs with the lowest adjusted

p value (< 0.02, Wald test, FDR-adjusted). A negative fold change indicates a lower relative abundance in samples collected after treatment (TP2).

Table 3.1.2: Differentially abundant bacterial OTUs in samples from collection TP1 and TP2

OTU	Taxon	log_2 fold change	adjusted *p* value
OTU_17	*Prevotella* sp.	-3.12	0.001
OTU_135	*Prevotella* sp.	-2.87	0.013
OTU_46	Ruminococcaceae (family)	-2.77	0.002
OTU_76	*Parasutterella* sp.	-2.62	0.002
OTU_11	*Akkermansia* sp.	-2.45	0.002
OTU_194	Lachnospiraceae incertae sedis	-2.37	0.002
OTU_90	*Clostridium* cluster XIVa	-2.22	0.002
OTU_30	*Flavonifractor* sp.	1.49	0.013
OTU_33	*Clostridium* cluster XVIII	1.49	0.013
OTU_384	Clostridiales (order)	1.88	0.013
OTU_31	Lachnospiraceae (family)	2.06	0.002
OTU_52	*Clostridium* cluster XI	2.20	0.003
OTU_35	*Lactobacillus* sp.	4.13	5.15E-07

While the observed decrease in *Clostridium* cluster XIVa is in agreement with observations in other studies (Touchefeu et al., 2014), a different study observed an increase in the relative abundance of *Akkermansia* spp. after chemotherapy (Zwielehner et al., 2011), which is contrary to the current observation. Some members of the families Ruminococcaceae and Lachnospiraceae have beneficial, health-promoting properties (Ying Taur, 2016) and loss of these bacteria might have negative effects on the host. However, here, both decreases and increases in the relative abundance of different OTUs belonging to these families were observed.

No correlations between the relative abundances of bacterial taxa (on any taxonomic level) and any clinical markers (including data such as calprotectin level, number or relative abundance of leukocytes, thrombocytes, CD3+ cells or C-reactive protein) were detected (Spearman's rank correlation test with multiple-testing adjustment). Similarly, no links between prokaryotic diversity and the clinical data were found.

Bacterial Shannon diversity and Chao1 richness for each TP within each TG at OTU level were assessed after rarefaction (Figure 3.1.3). Quite drastic differences in diversity in TG1 (Figure 3.1.3A) were observed with a median diversity of 2.6 at TP1, increasing to 3.3 at TP4. A similar pattern was observed for bacterial richness (Figure 3.1.3B). No statistically significant differences were found between different TPs in this TG. However, this TG comprised only two patients. With a higher number of patients and samples, the observed

pattern might look different, possibly more similar to the pattern observed in TG2. In TG2, a slight decrease in diversity throughout treatment was observed (Figure 3.1.3C). Bacterial richness displayed more pronounced changes with a significant decrease from TP1 to TP2 (samples of the same individuals: *p* value 0.017, Wilcoxon signed-rank test and over all patients: *p* value 0.018, Wilcoxon rank sum test, Figure 3.1.3D). In TG3, no drastic changes in diversity or richness could be observed (Figure 3.1.3E and Figure 3.1.3F). Both diversity and richness were in general lower in this TG than in the other TGs. This might be due to the higher intensity of the treatment and due to previous treatments of these patients.

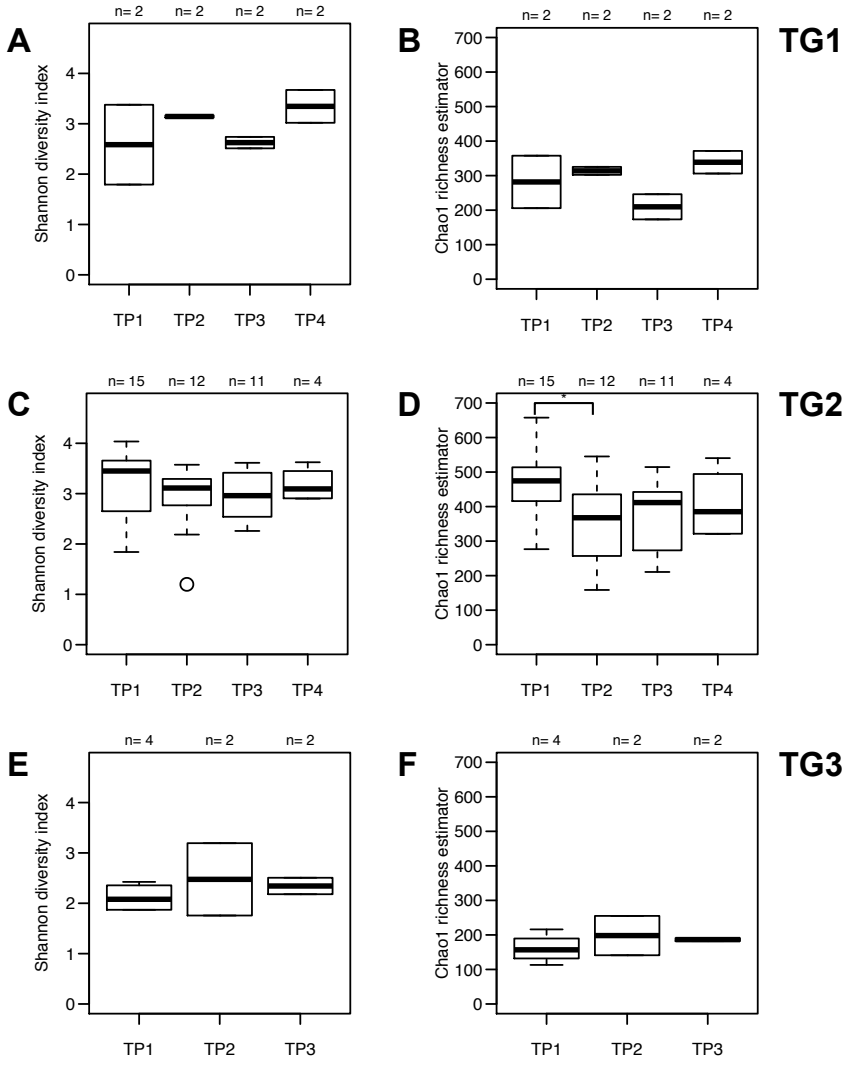


Figure 3.1.3: Changes within gastrointestinal bacterial community structure in patients receiving different anticancer treatments. Boxplots depicting (A, C and E) diversity (Shannon diversity index) and (B, D and F) richness (Chao1 richness estimator) per collection time point (TP), for prokaryotes in (A, B) TG1, (C, D) TG2 and (E, F) TG3 (determined by 16S rRNA gene amplicon sequencing) The number of samples per collection TP is indicated above each box. Diversity and richness were determined after rarefaction of the dataset (* when *p* value < 0.05, Wilcoxon rank sum test).

As the GIT microbiome of infants and young children is unstable and reaches a state similar to that of adults at around 3-4 years (Greenhalgh et al., 2016; Voreades, Kozil, & Weir, 2014), I was wondering whether a difference in age of the patients was reflected in the microbiome. TG1 and TG3 displayed an overall lower richness than TG2 and included a higher ratio of younger patients. Both patients included in TG1, and one out of four patients in TG3, while only two out of fifteen patients in TG2, were 3 or 4 years old at beginning of treatment. In the following, patients were grouped according to their age with five patients belonging to the group with younger patients (3 or 4 years old). All remaining thirteen patients belonged to the group with older patients (8 years or older). A significant difference in richness was observed when grouping the samples independently of the TP (*p* value 0.0025, Wilcoxon rank sum test), with a lower richness for the samples from younger patients (Figure 3.1.4A). Richness was also plotted per TP as boxplots (Figure 3.1.4B and Figure 3.1.4C) or as connected points (Figure 3.1.4D and Figure 3.1.4E) illustrating the development of richness per patient throughout the treatment.

In general, a lower bacterial richness was observed in younger patients, and the richness was already lower at TP1, before treatment start. In the younger patients, a striking decrease in richness was observed at TP3. In the older patients, bacterial richness was generally higher and showed no specific pattern. In some of the older patients, a low level of richness was observed at some point during treatment, similar to the richness observed in samples from younger patients. These results indicate that richness was indeed related to age. In our study however, richness did change throughout the treatments indicating that the treatment also had a large effect on bacterial richness.

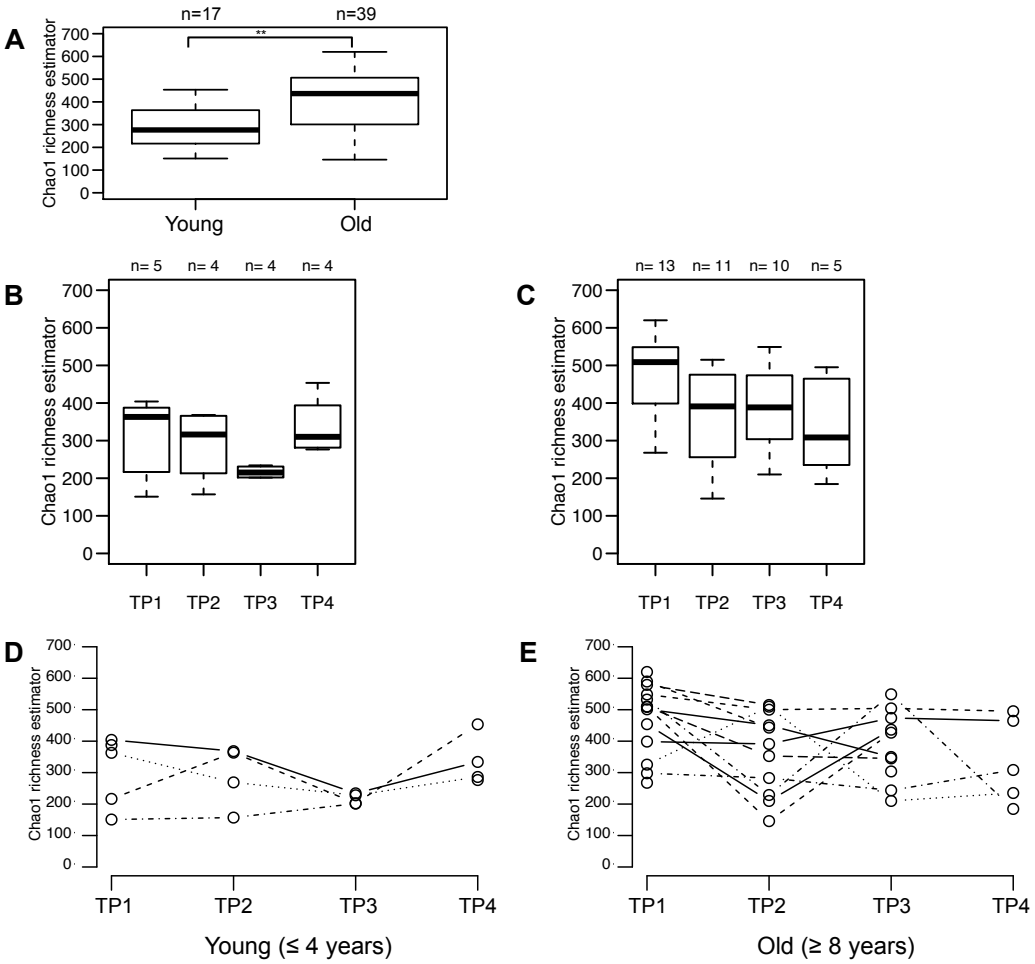


Figure 3.1.4: Bacterial richness in young and older children. Boxplots depicting richness (Chao1 richness estimator) in samples from (A) young (3-4 years, left) and older (8-19 years, right) patients. (B, C) Boxplots depicting richness of the same samples and patients per TP. (D, E) Richness per TP and patient indicated as point connected with lines.

The GIT microbiome is usually temporally stable and inter-individual differences significantly higher than intra-individual differences (Brandt et al., 2012; Heintz-Buschart et al., 2016). Here, to compare the intra-individual to inter-individual distance between microbial profiles, Bray-Curtis dissimilarity was determined based on the OTUs obtained from 16S rRNA gene sequencing data. The intra-individual distance describes the variation between the microbial profiles of different TPs from each patient while the inter-individual distance indicates the variation between the microbial profiles from different patients. A slightly higher inter-individual dissimilarity was observed (*p* value 0.024, Wilcoxon rank sum test, Figure 3.1.5).

Earlier studies with healthy individuals have shown that without the influence of external factors, the inter-individual dissimilarity was significantly higher than intra-individual

differences, because of the stability of the GIT microbiome. Here, the intra-individual dissimilarity was high, indicating that the treatment did have a strong effect on the GIT microbial community. As the TGs were exposed to different intensities of treatment, a higher intra-individual distance in TGs with the more intense treatment might be expected. However, no statistically significant differences between the intra-individual distances in the three TGs were detected. Similarly, a higher diversity or richness at TP1 might be a predictor of a more stable bacterial community, which would be less affected by the following treatment. However, no correlation between the diversity or richness at TP1 and the following changes in terms of intra-individual dissimilarity was observed. Diversity or richness at TP1 is therefore not predictive of the level of changes within the GIT microbial community throughout treatment.

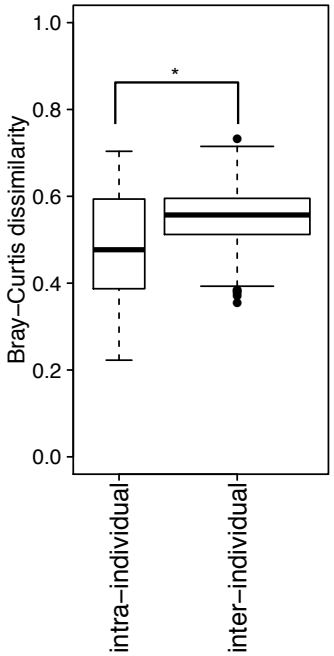


Figure 3.1.5: Comparison of intra-individual to inter-individual distances between bacterial profiles

3.1.3 Changes in the microeukaryotic populations of pediatric patients throughout cancer treatment

While other studies have focussed on the prokaryotic community only, here, also the microeukaryotic community composition was assessed using 18S rRNA gene amplicon sequencing of DNA extracted from 60 fecal samples of the patients. During the filtering steps, food related reads, such as reads mapping to plants (e.g the genera *Fragaria* (strawberry) or *Solanum* (potato and tomato)) as well as reads mapping to the human genome were removed. For some samples, the majority of the reads were human or food related, hence, these datasets were removed due to a low number of reads after filtering.

38 samples were kept for further analyses. The 14 most abundant taxa were identified to get an overview of the composition of the GIT microeukaryotic community of the patients (Figure 3.1.6). Most samples were dominated by the taxon Ascomycota, which includes the most common microeukaryotic representatives within the human GIT, such as *Candida* spp. and *Saccharomyces* spp. Some samples included other taxa, for example P01_TP3 showed relative high levels of *Blastocystis* spp., a unicellular, nonflagellated member of the Stramenopiles belonging to the taxon Opalinata (Scanlan et al., 2014), which is a common member of the GIT microbiome. In P18_TP1, a common flagellate protozoan parasite, *Dientamoeba fragilis* belonging to the taxon Parabasalia was detected. BLAST analysis revealed that some of the abundant OTUs included in the group 'unknown' also represented *Dientamoeba fragilis*. Infections with *Dientamoeba fragilis* can remain asymptomatic but can also cause abdominal symptoms such as pain, nausea and diarrhea (Elbakri, Al-qahtani, & Samie, 2015).

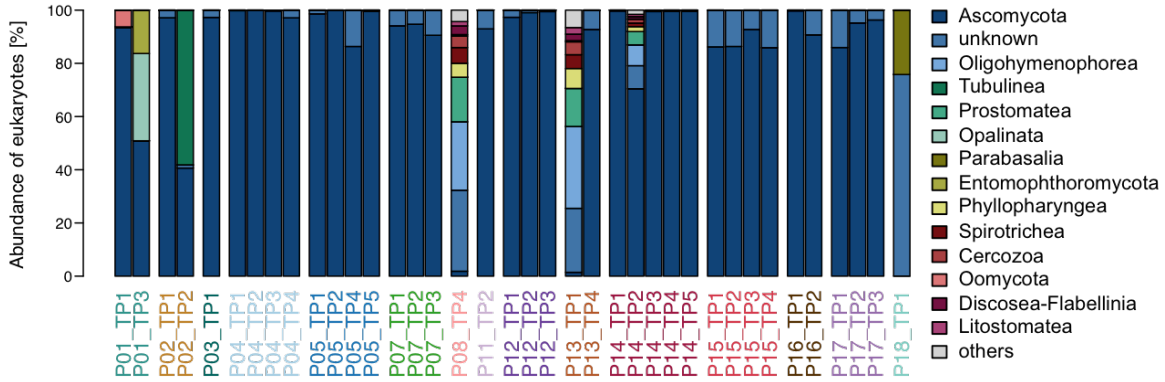


Figure 3.1.6: Relative abundance of the 14 most abundant microeukaryotic taxa in fecal samples from pediatric cancer patients grouped according to patient. Taxa which are not comprised in the 14 most abundant taxa are combined as 'others'. OTUs which could not be classified at any taxonomic level are grouped as 'unknown'. Patient ID and sampling TP are indicated below each respective bar and are colored according to the patient.

In the following, the samples were grouped according to the TG and according to the TP within the TG (Figure 3.1.7). In some samples, a large proportion of reads was grouped as 'unclassified', meaning that those OTUs could not be classified at the genus level. They could however be classified at higher levels and the largest part of these OTUs belonged to the taxon Ascomycota. No clear differences in the relative abundance of microeukaryotes specific to individual TGs or TPs were apparent. Some patients (such as P07, in Figure 3.1.7A and P15, in Figure 3.1.7B) showed similar microeukaryotic community compositions at each TP while others (such as P01 in Figure 3.1.7B) revealed a more unique community composition at different TPs.

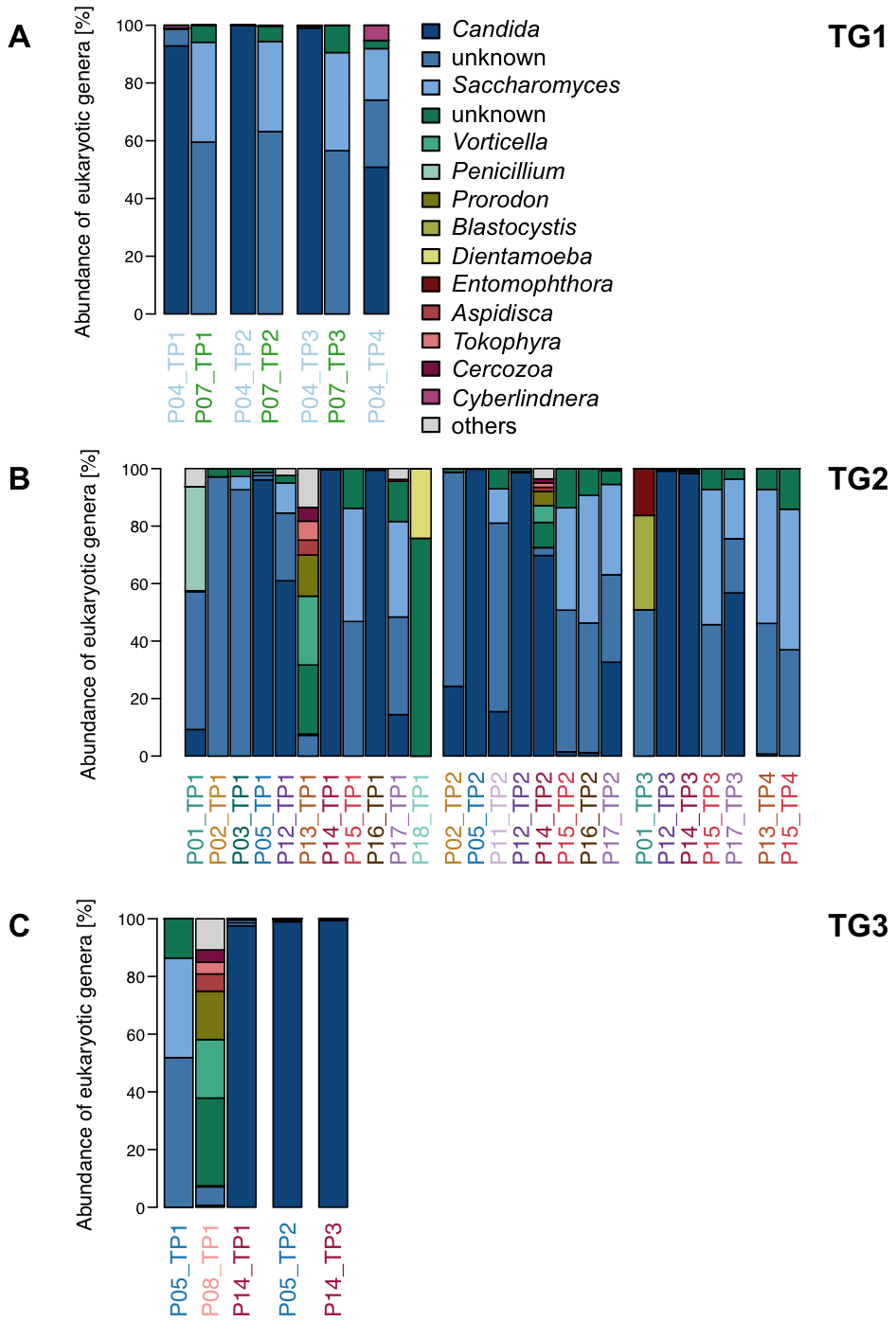


Figure 3.1.7: Relative abundance of the 14 most abundant microeukaryotic genera in fecal samples from pediatric cancer patients. Samples are grouped according to TPs within TG: (A) TG1, (B) TG2 and (C) TG3. Taxa which were not comprised in the 14 most abundant genera are combined as 'others'. OTUs which could not be classified at the genus level are grouped as 'unclassified'. OTUs which could not be classified at any taxonomic level are grouped as 'unknown'. Patient ID and sampling TP are indicated below each respective bar and are colored according to the patient.

In order to determine whether the microeukaryotic community evolved in the same way as the prokaryotic community, diversity and richness of the microeukarotic community were determined on the OTU level after rarefaction (Figure 3.1.8). No statistically significant differences in diversity or richness between different TPs within the TGs were determined. Overall, the microeukaryotic community exhibited a lower diversity and richness than the prokaryotic community. No correlation between prokaryotic and eukaryotic diversity or richness was detected (Spearman's rank correlation test). Development over time of both diversity and richness of the microeukaryotic community was different from the development of the prokaryotic community (Figure 3.1.3), which indicates that both communities were differently affected by the treatment.

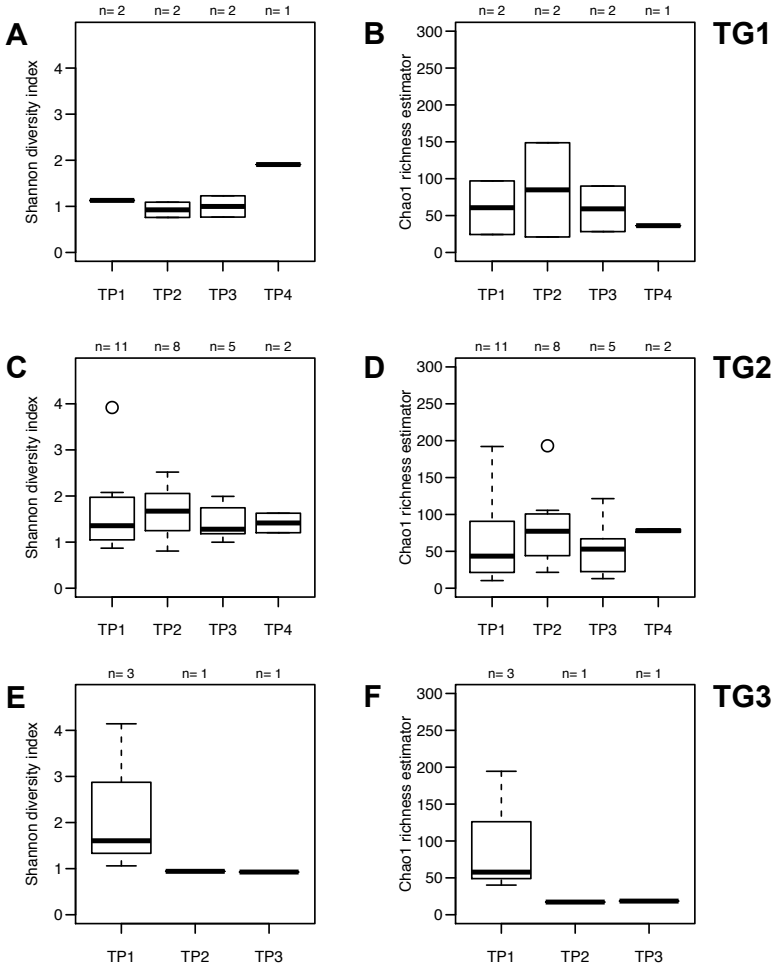


Figure 3.1.8: Changes in the gastrointestinal microeukaryotic community structure in patients receiving different anticancer treatments. Boxplots depicting (A, C and E) diversity (Shannon diversity index) and (B, D and F) richness (Chao1 richness estimator) per collection time point (TP), for microeukaryotes in (A, B) TG1, (C, D) TG2 and (E, F) TG3 (determined by 18S rRNA gene amplicon sequencing). The number of samples per collection TP is indicated above each box. Diversity and richness were determined after rarefaction of the dataset.

3.1.4 Variability of GIT microbiome trajectories throughout treatment

As apparent from the overview of the prokaryotic community composition in each sample (Figure 3.1.1), the GIT microbiome from some patients underwent pronounced changes throughout treatment while the microbiome of other patients stayed relatively stable. A principal component analysis (PCA) on the genus level revealed similar trends (Figure 3.1.9). One patient whose samples clustered relatively closely together (P02, marked in light brown in Figure 3.1.9) and one patient whose samples were quite dispersed on the PCA plot, especially in regards to PC1 (P05, marked in blue in Figure 3.1.9) were selected for a closer evaluation. Their respective samples are encircled in Figure 3.1.9. Overall, the samples also did not group according to the collection TP, or according to parameters characterizing the patient's status, such as the occurrence of mucositis, a low number of leukocytes, a reduced overall status or overall outcome.

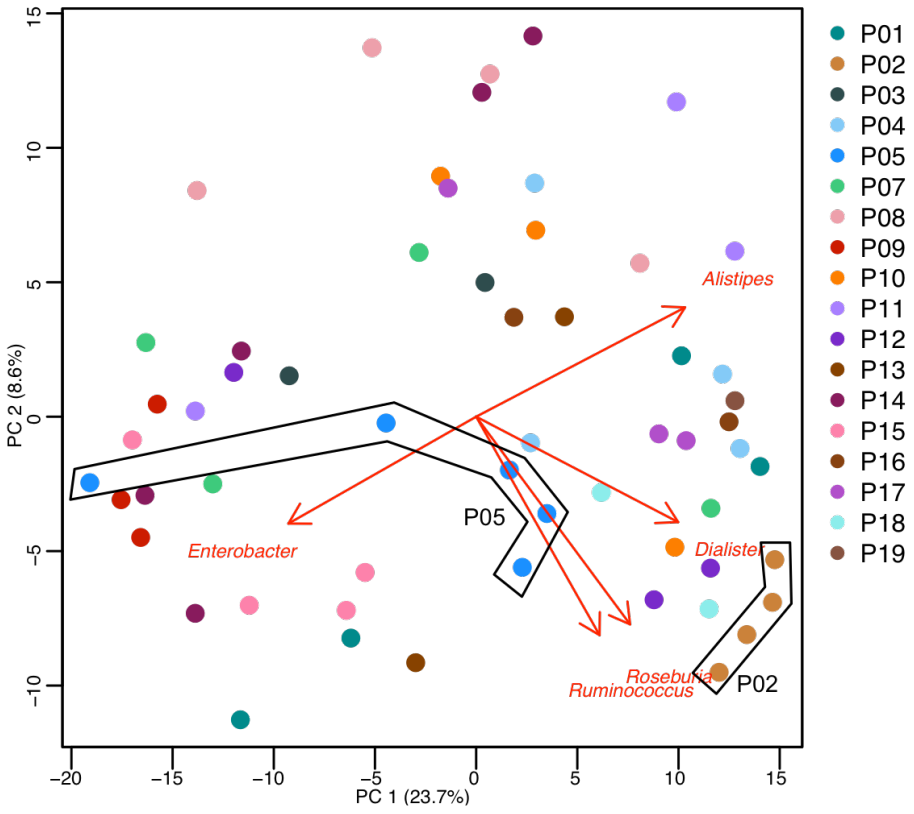


Figure 3.1.9: Principal component analysis (PCA) for GIT prokaryotic community composition. Each dot represents a sample colored according to which patient it derived from. PCA was performed on genus level, based on 16S rRNA gene amplicon sequencing.

In P02, the most abundant OTUs (such as *Prevotella* sp., *Dialister* sp. and *Faecalibacterium* sp.) were the same at each TP (Figure 3.1.10A), diversity stayed quite high (Figure 3.1.10C), as did the richness (ranging between 445 and 610). Within the

cohort, patient P02 displayed the lowest mean Bray-Curtis dissimilarity between the TPs (0.22). In P05, the most abundant OTUs differed at each TP (Figure 3.1.10B) and there were important changes in diversity between the TPs (Figure 3.1.10D) as well as in richness (ranging from 188 to 526). Mean intra-individual dissimilarity for this patient was 0.54.

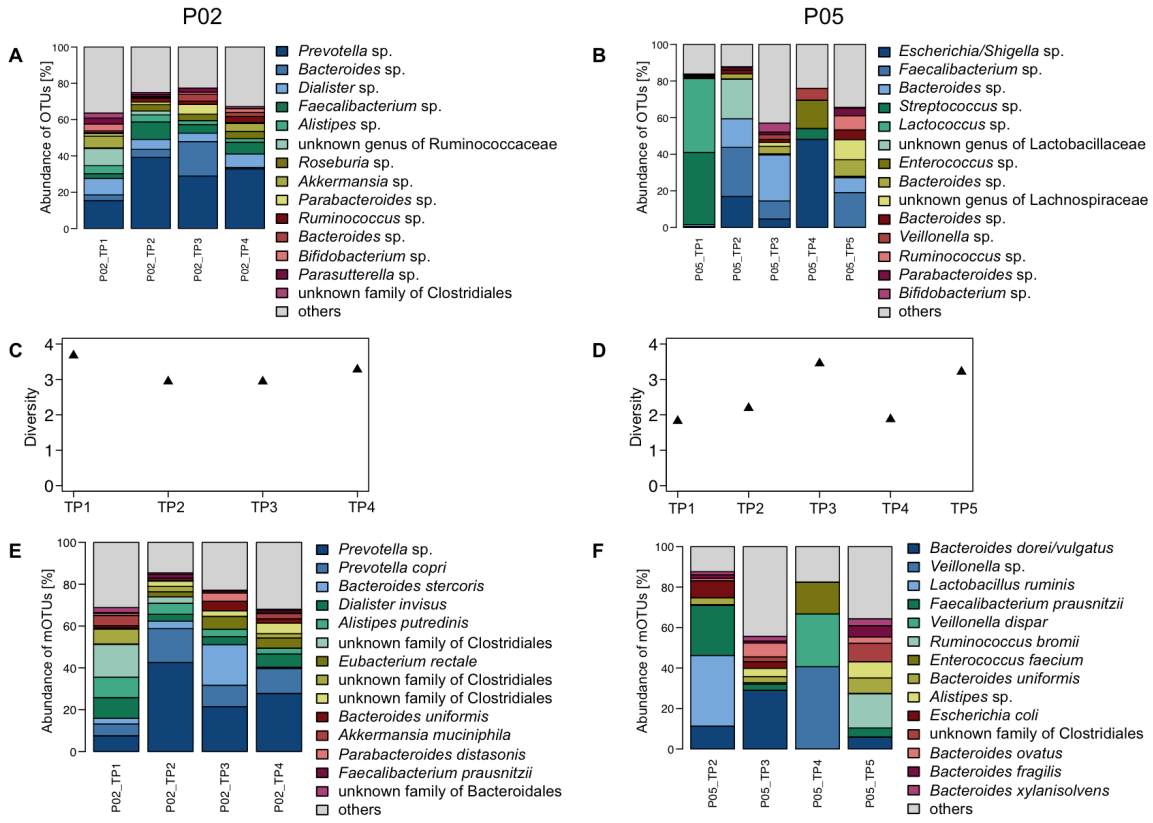


Figure 3.1.10: Variation of the microbial community structure over the course of the treatment in pediatric patients. (A) and (B) Relative proportions of the 14 most abundant operational taxonomic units (OTUs) based on 16S rRNA gene sequencing. The remaining OTUs are summarized as 'others'. (C) and (D) Prokaryotic diversity represented by Shannon diversity index at sampling TPs throughout the treatment. (E) and (F) Relative proportions of the 14 most abundant metagenomic operational taxonomic units (mOTUs) based on MG sequencing. The remaining mOTUs are summarized as 'others'. Plots represent the corresponding results for patient P02 (panels A, C and E) and patient P05 (panels B, D and F).

For the 54 samples where MG sequencing data was available, metagenomic operational taxonomic units (mOTUs) were identified for taxonomic profiling. For most of the abundant mOTUs, classification at species level was possible. While the bacterial abundance profile based on 16S rRNA gene sequencing and MG sequencing largely coincide for P02 (Figure 3.1.10A and Figure 3.1.10E), more differences are observed for P05 (Figure 3.1.10B and Figure 3.1.10F). For example, the most abundant OTU at TP4 based on 16S

rRNA gene sequencing was classified as *Escherichia/Shigella* sp., while this seemed to be only lowly abundant at TP4 in the MG sequencing dataset.
MG sequencing does not only allow taxonomic profiling but also functional profiling, which will be described in the next parts.

3.1.5 Detection of antibiotic resistance genes

Many cancer patients are immunocompromised, either due to the underlying disease, or due to the anticancer treatment. Therefore, they are often prone to infection, hence, prophylactic antibiotics or an antibiotic treatment during neutropenia or at occurrence of fever are often given. However, antibiotic resistance and especially the emergence of multi-drug resistant bacteria represent a threat. For the 54 samples with available MG datasets, antibiotic resistance genes (ARGs) were detected within the predicted genes using the Resfams database (section 2.11) and their relative abundance (percentage of ARGs relative to the total number of genes) was calculated (Figure 3.1.11A). The mean relative abundance and standard deviation within these 54 samples was 0.26 ± 0.06 %.
While most patients received only trimethoprim/sulfamethoxazole (cotrimoxazole) regularly to prevent *Pneumocystis jirovecii* pneumonia, some patients received also other antibiotics. For example, patient P01 was treated with piperacillin/tazobactam before TP1, patient P07 before TP3, patient P04 before TP2 and TP3 and patient P05 was treated with cefuroxime before TP1 and TP2. For the majority of patients, a slight increase in relative abundance of ARGs was observed (Figure 3.1.11A and Figure 3.1.11C), however, also other patterns were observed as for example a decrease for P01, and a decrease followed by an increase for P07 and P14 (Figure 3.1.11A). Within TG1 (Figure 3.1.11B) the relative abundance of ARGs stayed relatively constant while a decrease followed by an increase was observed in TG3 (Figure 3.1.11D). However, due to the low number of samples within these TGs, these observations cannot be fully attributed to the respective treatments. In TG2, a statistically significant increase from TP1 to TP2 (p value 0.034, Wilcoxon rank sum test) and from TP1 to TP3 (p value 0.045, Wilcoxon rank sum test) was observed (Figure 3.1.11C). To summarize, especially in TG2 (which included the highest number of patients), an increase in the relative abundance of ARGs throughout treatment was observed.

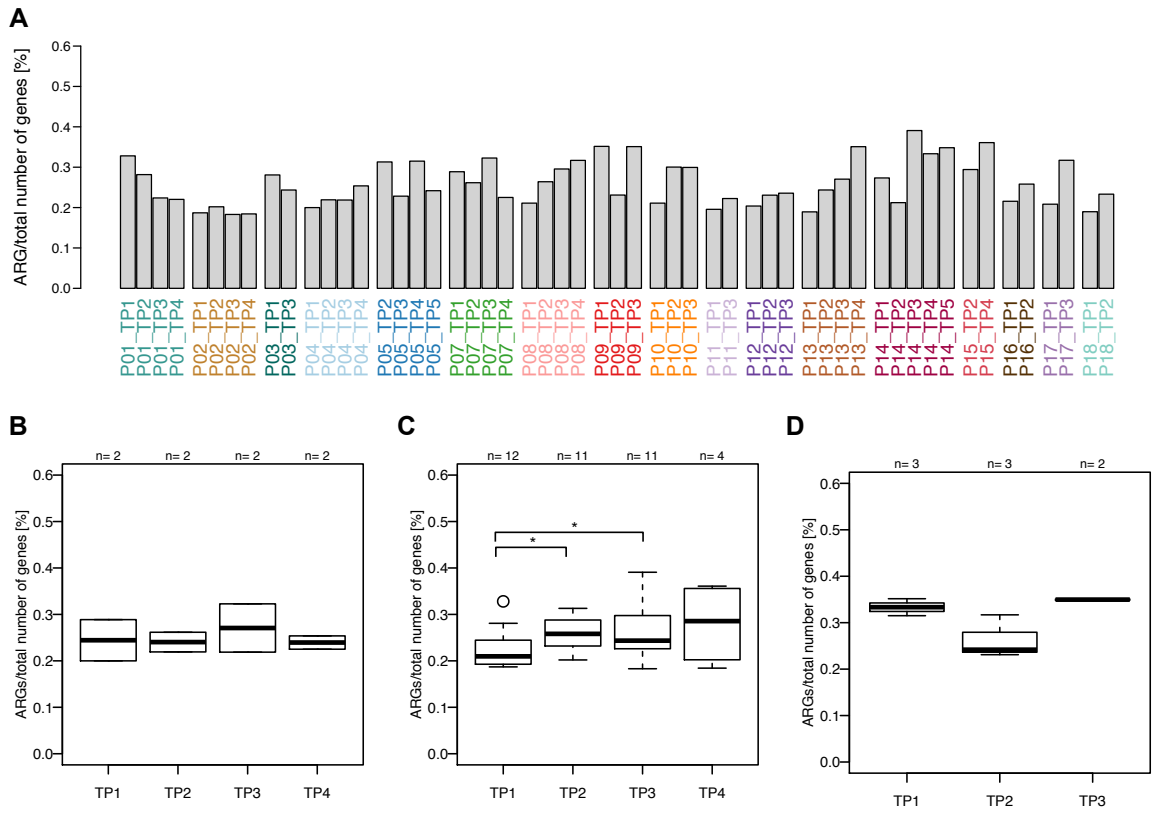


Figure 3.1.11: Relative abundance of antibiotic resistance genes in fecal samples from pediatric cancer patients. (A) Samples grouped according to patient. IDs are colored according to the patient. (B – D) Samples combined according to TP within (B) TG1, (C) TG2 and (D) TG3. (B-D) The number of samples per collection TP is indicated above each plot (* when *p* value < 0.05, Wilcoxon rank sum test).

3.1.6 Virome profiling within the GIT microbiome of pediatric cancer patients

Apart from prokaryotes and eukaryotes, the GIT community also harbors viruses which play a part in human health and disease (Popgeorgiev, Temmam, Raoult, & Desnues, 2013). Even viruses which are usually not harmful for their host can pose a serious threat in immunocompromised patients (Sahin et al., 2016). Shotgun sequencing may allow their detection and treatment, before they become harmful. On average, 0.014 ± 0.012 % MG and 0.033 ± 0.058 % MT reads were mapped onto viral genomes within the 54 (for MG) and 32 (for MT) samples.

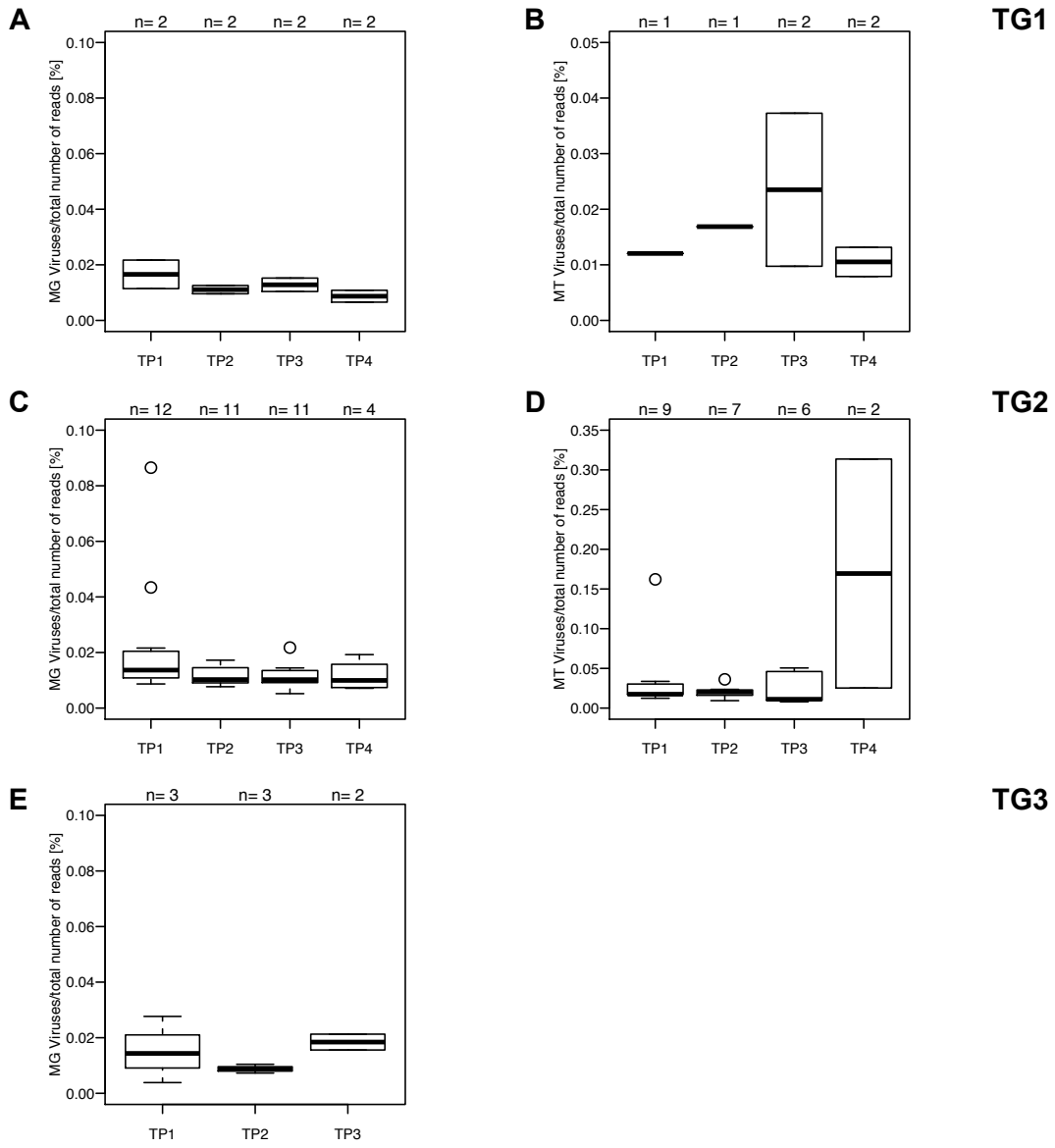


Figure 3.1.12: Relative abundance of reads mapping to viral genomes. Boxplots depict MG (A, C and E) or MT (B and D) reads mapping to viral genomes grouped per TP within TG1 (A and B), TG2 (C and D) and TG3 (E). The number of samples per collection TP is indicated above each box. For TG3, only two MT datasets from one patient were available, hence it was not possible to construct boxplots.

Compared to the average, one sample (P01_TP1, Figure 3.1.12C) contained many reads mapping to viral DNA genomes (0.087 %). The most abundant viruses found in this sample included viruses with large genomes belonging to the clade Megaviridae, such as the *Phaeocystis globosa* virus (Santini et al., 2013), which infects the unicellular algae *Phaeocystis* (however this was not detected in the 18S rRNA gene sequences), or the *Acanthamoeba polyphage* moumouvirus infecting the amoeba *Acanthamoeba polyphage*

(which was also not detected in the 18S rRNA gene sequences) (Yoosuf et al., 2012). Different Entomopoxviriniae were detected, which infect insects (Afonso et al., 1999).

Sample P01_TP4 (Figure 3.1.12D) contained many reads (0.31 %) mapping to RNA viruses, compared to the average. Tobamovirus, the pepper mild mottle virus (PMMoV) was found to account for 96 % of the RNA reads mapping to viral genomes within this sample. This is among the most important pathogens of peppers (*Capsicum* spp.) and is commonly detected in high abundance in fecal samples (Colson et al., 2010; Zhang et al., 2006). Although it is generally assumed that plant-associated viruses do not become resident in the GIT and are not pathogenic for humans, it cannot be completely excluded (Balique, Lecoq, Raoult, & Colson, 2015). In one study, the presence of PMMoV was linked to clinical symptoms such as abdominal pains and fever (Colson et al., 2010). However, abdominal pain could simply be linked to the consumption of pepper.

As for some samples, the majority of the reads mapping to viral genomes contained plant-associated viruses which were taken up via nutrition, or viruses, which infect insects or amoeba, the same analyses were repeated but mapping against human-associated viruses only (Figure 3.1.13). On average, 0.0004 ± 0.0003 % MG and 0.0009 ± 0.0006 % MT reads were mapped onto viral genomes within the 54 (for MG), respectively 32 (for MT) samples.

Sample P14_TP5 contained the highest ratio of MG reads (0.002 %) mapping to human-associated viral genomes, compared to the other samples. 66 % of those reads mapped to *Betapapillomavirus*. This virus has previously been detected in fecal samples (Di Bonito et al., 2015; Ma et al., 2014) and does not necessarily provoke any symptoms in the host. Patient P14 developed herpes zoster, a painful skin rash, directly after collection of the third sample. This is caused by the human herpesvirus 3, or herpes zoster virus. This virus was detected in P14_TP3 ($2.24 * 10^{-5}$ % of the total number of reads), P14_TP4 ($8.22 * 10^{-6}$ %) and P14_TP5 ($1.79 * 10^{-4}$ %).

One sample with a high RNA viral load was P15_TP2, where the human Coronavirus was detected. This virus commonly infects the human respiratory tract or GIT, has been isolated from fecal samples and has been observed to be shed for many months (Clarke, Caul, & Egglestone, 1979; Zhang et al., 2006). In P13_TP1 (69 %), P13_TP2 (2 %) but not in P13_TP3, Enterovirus of the order Picornaviridae was detected. This virus is thought to be swallowed, followed by replication in the intestinal mucosa, crossing the intestinal barrier to reach the blood. Most infections are asymptomatic, but symptoms can present as symptoms of a common summer cold to a threatening encephalitis (Tapparel, Siegrist, Petty, & Kaiser, 2013).

To summarize, shotgun sequencing allows deep profiling of the GIT microbiome, including the virome and might enable their detection and treatment, before they become harmful.

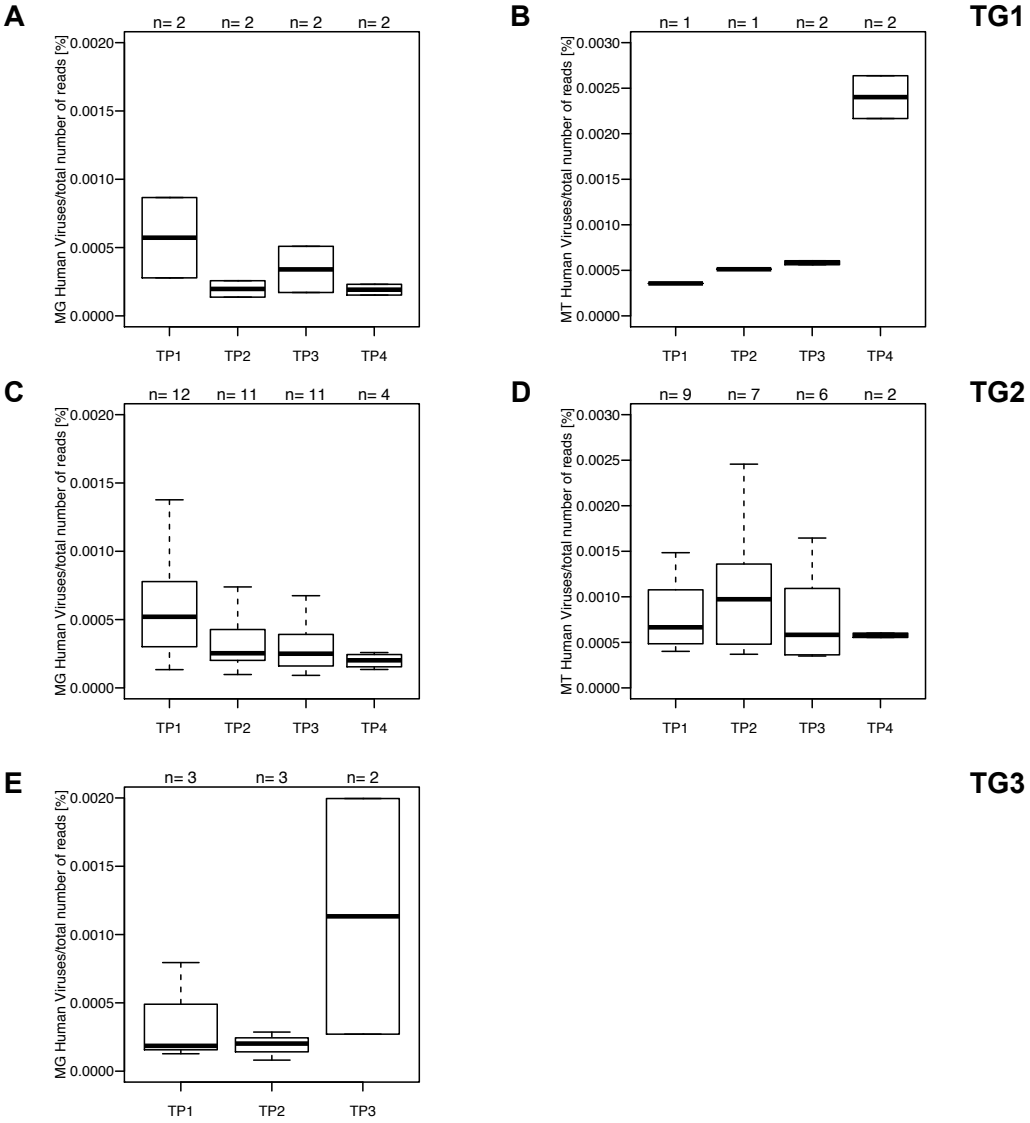


Figure 3.1.13: Relative abundance of reads mapping to human-associated viral genomes. Boxplots depict MG (A, C and E) or MT (B and D) reads mapping to human-associated viral genomes grouped per TP within TG1 (A and B), TG2 (C and D) and TG3 (E). The number of samples per collection TP are indicated above each box. For TG3, only two MT datasets from one patient were available, hence it was not possible to construct boxplots.

3.1.7 Does the microbiome influence development of mucositis?

3.1.7.1 Differences on taxonomic level

It has been hypothesized that the GIT microbiome might be involved in development of mucositis, a severe side effect of anticancer treatment. In the next part, I will focus on this subject. Mucositis is generally associated with the treatment, with intensive chemotherapy and/or radiation leading to higher incidence of mucositis. Five out of eighteen patients developed severe mucositis, three of these belonged to TG2, one to TG1 and one belonged to TG3. No link between the TG and occurrence of severe mucositis could be made. Similarly, no link between prokaryotic diversity or richness and development of mucositis was found.

Differentially abundant prokaryotic taxa between patients who developed mucositis and those who did not were identified based on 16S rRNA gene datasets. Since a specific microbial community profile might possibly play a part in the development of mucositis, and the community could also be altered after the active phase of mucositis, not only samples from TPs with active mucositis (n=3), but all the samples from patients who developed mucositis (n=13), were included for the following analyses and compared to samples from patients who did not develop severe mucositis (n=45). 68 different OTUs were observed to be differentially abundant. Table 3.1.3 includes 34 OTUs with the lowest adjusted *p* value (< 0.02, Wald test, FDR-adjusted). A negative fold change indicates a lower relative abundance in samples from patients with mucositis.

Table 3.1.3: Differentially abundant OTUs in relation to mucositis

OTU	Taxon	log_2 fold change	adjusted *p* value
OTU_54	*Phascolarctobacterium* sp.	-4.91	5.88E-11
OTU_94	Lachnospiracea incertae sedis	-3.35	4.03E-05
OTU_278	*Alistipes* sp.	-3.18	0.001
OTU_28	*Bacteroides* sp.	-2.92	1.75E-04
OTU_366	*Anaerorhabdus* sp.	-2.81	0.003
OTU_135	*Prevotella* sp.	-2.8	0.017
OTU_466	*Phascolarctobacterium* sp.	-2.8	0.02
OTU_488	*Oxalobacter* sp.	-2.79	0.006
OTU_82	*Prevotella* sp.	-2.74	0.004
OTU_72	*Ruminococcus* sp.	-2.61	0.003
OTU_52	*Clostridium* cluster XI	-2.47	1.02E-04
OTU_78	Lachnospiracea incertae sedis	-2.39	0.003
OTU_139	*Sutterella* sp.	-2.32	0.013
OTU_17	*Prevotella* sp.	-2.3	0.016
OTU_55	*Clostridium* sensu stricto	-2.23	9.37E-05
OTU_191	*Roseburia* sp.	-2.23	0.002
OTU_262	*Clostridium* cluster XIVa	-2.2	0.02
OTU_203	*Escherichia/Shigella* sp.	-2.16	0.002
OTU_92	*Clostridium* sensu stricto	-1.98	0.006
OTU_33	*Clostridium* cluster XVIII	-1.85	0.001
OTU_359	*Clostridium* cluster XVIII	-1.81	0.006
OTU_32	*Clostridium* cluster XI	-1.73	0.003
OTU_12	Erysipelotrichaceae incertae sedis	-1.58	0.003
OTU_45	*Blautia* sp.	-1.58	0.02
OTU_3	*Escherichia/Shigella* sp.	-1.41	0.004
OTU_286	Erysipelotrichaceae incertae sedis	-1.41	0.016
OTU_242	*Clostridium* cluster XIVa	1.23	0.006
OTU_532	*Eggerthella* sp.	1.42	0.004
OTU_75	*Eggerthella* sp.	1.55	0.003
OTU_160	*Clostridium* cluster IV	1.59	0.016
OTU_10	*Clostridium* cluster XIVa	1.64	2.41E-04
OTU_115	*Clostridium* sensu stricto	2.81	0.005
OTU_127	Erysipelotrichaceae incertae sedis	3.12	8.22E-05
OTU_376	*Alistipes* sp.	3.36	0.004

While OTU_278, belonging to the genus *Alistipes* was observed to be less abundant in samples from patients with mucositis, OTU_376, also classified as the genus *Alistipes*, showed increased levels in these patients. Similarly, different OTUs classified into the genus *Clostridium* or *Clostridium* cluster XIVa showed differing directions of change. Some of the OTUs that were less abundant in the samples from patients with mucositis

belonged to genera which are among the most common members of the GIT microbiome, such as the genera *Bacteroides*, *Prevotella*, *Ruminococcus* and *Escherichia* (Arumugam et al., 2011). Other OTUs with a lower relative abundance in patients with mucositis were bacteria, which are usually considered to have health-promoting properties, such as the acetate- or butyrate-producing genera *Blautia* and *Roseburia*.

When comparing only the samples that were collected when patients had active severe mucositis (n=3) against all the other samples (n=55), a significant decrease in only one genus, *Akkermansia*, was observed (log_2 fold change -3.35, FDR-adjusted *p* value 0.04, Figure 3.1.14). *Akkermansia muciniphila* is a mucin-degrading bacterium which resides in the mucus layer and uses the proteins of the epithelial mucus layer as its main source of carbon and nitrogen (Reunanen et al., 2015). Lower levels of *A. muciniphila* have been observed in patients with ulcerative colitis and Crohn's disease, indicating that this organism is important for human GIT health and inversely correlated with inflammation (Derrien, Belzer, & de Vos, 2016; Png et al., 2010; Schneeberger et al., 2015; Wu & Scott, 2012). Also, it was observed that *A. muciniphila* strengthens the epithelial barrier function (Reunanen et al., 2015). On the one hand, loss of this bacterium could add to the impaired integrity due to treatment-provoked damage and facilitate translocation of bacteria and bacterial products such as LPS. On the other hand, chemotherapy induced damage of the intestinal epithelial wall and a concomitant decrease of mucin, could lead to a decrease in relative abundance of *A. muciniphila*, as the mucus layer is its most important nutrient source (Yamamoto, Ishihara, Takeda, Koizumi, & Ichikawa, 2013). In this scenario, the decrease in relative abundance of *A. muciniphila* would only represent a consequence of the treatment and the damage, rather than being a cause of mucositis development or its aggravation.

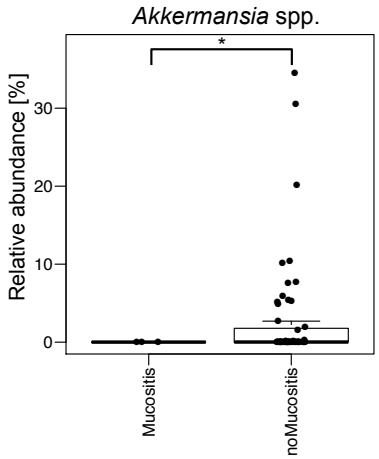


Figure 3.1.14: Relative abundance of the genus *Akkermansia* in samples from TPs with active mucositis compared to TPs without mucositis (* FDR-adjusted *p* value < 0.05, Wald test).

3.1.7.2 Differences on the functional level

Recently, several studies have stressed the importance of SCFAs, especially of butyrate, as energy source for colonocytes, possibly reinforcing epithelial barrier function (Mathewson et al., 2016; Peng et al., 2009). The list of differentially abundant OTUs (Table 3.1.3) includes many known important SCFA producers. Therefore, I was interested in whether a difference in the potential for production of butyrate by the microbiome was apparent in the samples of patients who developed severe mucositis and those who did not. Presence and expression of the genes coding for the three enzymes catalyzing the final step in butyrate production (*buk, but* and *ato*) were compared (Vital, Howe, & Tiedje, 2014). Median copy number of the three genes on MG level was higher in the samples from patients with mucositis (0.029 % relative abundance) than in the samples from patients without severe mucositis (0.016 %). On MT level, transcript levels were slightly higher in samples from patients without mucositis (0.027 % compared to 0.025 %). However, both on MG and on MT level, no statistically significant differences were observed between both groups.

In the 54 MG and 32 MT datasets, differentially expressed functions between samples from patients with mucositis and those without were detected. Genes from different bacterial genomes with the same functions were grouped together. These will in the following be referred to as 'functional gene categories'. Differential analysis of functional gene categories in the MG datasets between both groups resulted in 15 differentially abundant functional gene categories including an apparent increase in the peptidoglycan-associated lipoprotein precursor in samples from patients without mucositis ($\log_2$ fold change ≥ 1, FDR-adjusted p value < 0.05, Figure 3.1.15). This precursor contains a characteristic LVAC motif, which is cleaved during translocation across the cytoplasmic membrane. One role of this lipoprotein is to link the outer membrane to the peptidoglycan layer. The peptidoglycan-associated lipoprotein is a TLR2 agonist and is highly conserved in different genera (Liang et al., 2005). Lipoproteins are PAMPS, which are usually associated with bacterial pathogenicity and the possibility to cause inflammatory responses and even sepsis. It seems counter-intuitive that a gene associated with this should be found to be increased in samples from patients who did not develop severe mucositis. Possibly, the intestinal barrier in these patients was more resistant, less damaged by the treatment and therefore these patients did not develop severe mucositis despite the risk posed by these PAMPs.

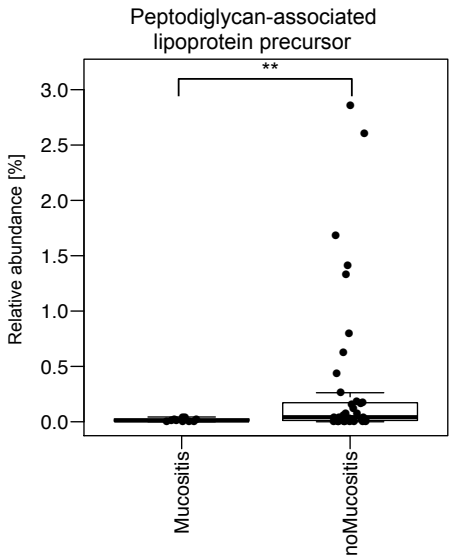


Figure 3.1.15: One differentially abundant functional gene category on MG level when grouping according to development of severe mucositis (* FDR-adjusted *p* value < 0.05, Wald test).

In the MT dataset, 59 functional gene categories were differentially expressed when grouping according to occurrence of severe mucositis. Some of these genes, which could possibly be implicated in the development of the side effect are illustrated in Figure 3.1.16. One functional gene category with higher expression in samples from patients with mucositis was the integration host factor (IHF) subunit alpha (Figure 3.1.16A). IHF is a DNA-binding protein and plays a role in several cellular functions in gram-negative bacteria. It was shown to positively regulate virulence gene expression in different bacteria (Porter & Dorman, 1997; Stonehouse, Kovacikova, Taylor, & Skorupski, 2008). The alpha subunit of this gene showed a higher expression level in samples from patients with mucositis. Similarly, the type II secretion system protein F expression was higher in samples from patients with mucositis (Figure 3.1.16B) and this secretion system has also been associated with bacterial virulence (Sandkvist, 2001). Both functions being implicated in bacterial virulence could initiate a cascade of inflammatory processes and therefore be involved in the development of mucositis.

Other functional gene categories which were more highly expressed in samples from patients without severe mucositis were associated with bacteriophages (such as the phage P2 GpU, the phage tail tube protein FII and the baseplate J-like protein, Figure 3.1.16C, D and E) or with the prokaryotic defense mechanism against bacteriophages (such as the CRISPR-associated protein Cas6, Figure 3.1.16F), indicating that bacteriophages might play a role in the prevention of mucositis. Studies have shown that bacteriophages can decrease the level of ROS produced by phagocytes (Przerwa et al.,

2006). This process involves not only phages and phagocytes, but an intermediate step with phage-LPS interaction, thus, implicating the GIT microbiome as well as the host immune system (Kaur et al., 2012). ROS are believed to play a role in the first and second stages of mucositis development, initiation and the primary damage response (Sonis, 2004). Presence of bacteriophages as well as the bacteria in the patients' GIT might have resulted in lower ROS production and therefore contributed towards the prevention of severe mucositis in these patients.

As phages cannot actively move toward their host, the targeted bacterial community has to be above a so called 'replication threshold', allowing the phage to encounter its host by chance. This scenario is also called 'kill the winner', where more abundant bacteria are targeted. In this scenario, bacteriophages could possibly act as regulator of the microbial ecosystem, maintaining bacterial diversity (De Paepe, Leclerc, Tinsley, & Petit, 2014). Functional gene categories detected in these samples (phage tail tube protein FII, phage P2 GpU and baseplate J-like protein, Figure 3.1.16B, C and D) are associated with the bacteriophage P2. This is a temperate phage, meaning that it can insert its genome into the host genome and be maintained as prophage (lysogenic cycle) or use the host cell to produce phage progenies and finally lyze the host cell (lytic cycle). This phage has been shown to infect *E. coli*, as well as *Shigella*, *Klebsiella*, *Yersinia* and *Serratia* (Kahn et al., 1991). Phage therapy has been suggested as alternative to antibiotics to fight bacterial infections. Here, patients who did not develop severe mucositis might have had higher titers of these bacteriophages, helping to maintain a diverse bacterial community. As this was detected in the MT dataset, the presence of viral RNA indicates active production of phage progenies, thus a lytic cycle with lysis of the bacterial host cell.

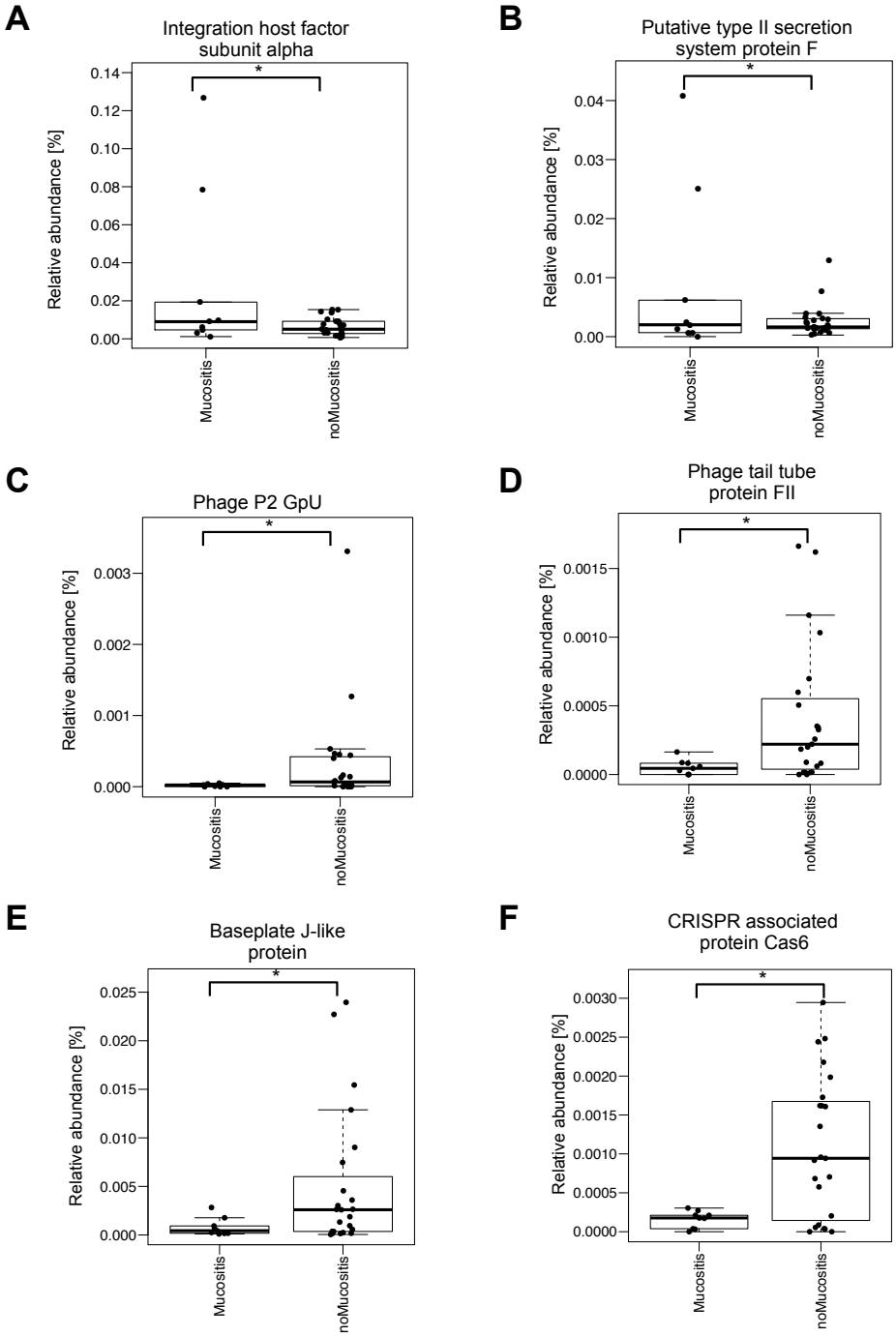


Figure 3.1.16: Selection of differentially abundant functional gene categories on MT level when grouping according to development of severe mucositis. Relative abundances of (A) integration host factor subunit alpha (B) putative type II secretion system protein F (C) phage P2 GpU (D) phage tail tube protein FII (E) baseplate J-like protein and (F) CRISPR associated protein Cas6 (* FDR-adjusted *p* value < 0.05, Wald test).

3.1.8 Functional changes in the GIT microbiome in relation to the overall health status

At each TP, the status of the patient was defined either as 'stable' or as 'reduced'. The reduced status could include different clinical symptoms such as fever, neurological symptoms (for example agitation), efflorescence of the skin or other symptoms. In the MG dataset, 136 functional gene categories were differentially abundant when accounting for the status of the patient, with *p* value < 0.01. These included many functional gene categories with normal bacterial cellular functions, where a link to the status of the patient could not be made. Examples of functional gene categories which could possibly be implicated in the patient's overall condition are illustrated in Figure 3.1.17. These include two genes associated with bacterial pathogenicity and one gene playing a part in the prokaryotic defense mechanism against phages. Type IV secretion systems can for example mediate injection of virulence proteins into mammalian cells or contribute to the spread of ARGs among pathogenic bacteria (Fronzes, Christie, & Waksmas, 2009; Wallden, Rivera-Calzada, & Waksman, 2010). Type IV pili are involved in adherence to different surfaces and in pathogenicity (Bieber et al., 1998). Both genes were more abundant in patients whose status was reduced which indicates the possible activity of pathogens and their implication in the overall health status of the patients (Figure 3.1.17A and Figure 3.1.17B). The CRISPR-associated gene Csn2 (Figure 3.1.17C) is involved in spacer acquisition in the type II CRISPR system, so in the first stage of the immune response against bacteriophages (Nam, Kurinov, & Ke, 2011; Sapranauskas et al., 2011). So far, this gene has not been detected in a wide range of bacterial species. A higher bacterial diversity in the GIT microbiome of patients with a stable health status could possibly be linked to a higher detection of this functional gene category. Incidentally, the median diversity of the bacterial community was higher in samples from patients with a stable health status than in samples from patients with reduced health status (2.77 in stable, 2.30 in reduced, *p* value 0.024, Wilcoxon rank sum test). However, there was no statistically significant correlation between the relative abundance of this functional gene category and the bacterial diversity.

On MT level, grouping according to the status of the patient did not reveal differentially abundant functional gene categories.

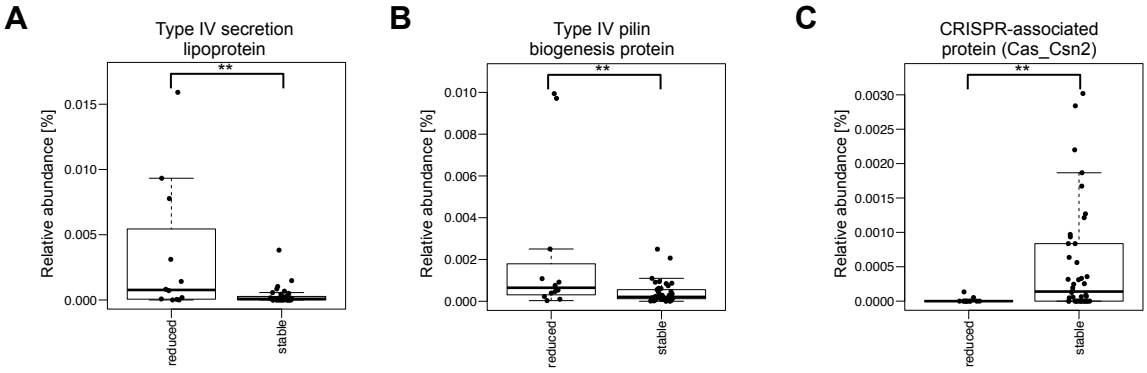


Figure 3.1.17: Selection of differentially abundant functional gene categories on MG level when grouping according to the status of the patient. Relative abundances of (A) type IV secretion lipoprotein (B) type IV pilin biogenesis protein and (C) CRISPR-associated protein Cas_Csn2 (** FDR-adjusted *p* value < 0.01, Wald test).

3.1.8.1 Case study – a patient with severe mucositis

In the following, I will focus on a patient who developed severe mucositis, patient P04. This patient, 3 years old at diagnosis, was treated with a light chemotherapy (TG1) for nephroblastoma. Fecal samples were collected at day 0 (TP1, beginning of treatment), 16 (TP2, after first cycle of treatment, lowest leukocyte count), 55 (TP3, before next cycle of treatment) and 359 (TP4, after end of therapy). The patient received trimethoprim/sulfamethoxazole to prevent *Pneumocystis jirovecii* pneumonia and piperacillin-tazobactam around TP2 and TP3 (for 2, respectively 3 days) because the patient had developed a fever (reason unknown). At TP2, when the leukocyte count was low (500/µl), the patient developed severe mucositis, which included impaired intake of solid food and needing pain medication. From this patient, MG and MT data from four TPs were produced. Although this patient received antibiotics, including a broad spectrum antibiotic (piperacillin-tazobactam), no drastic changes were observed in the bacterial community composition (Figure 3.1.18A). The biggest differences in the relative abundances of the most common mOTUs in P04 between TP1 and TP4 were an increase in relative abundance of *Bacteroides dorei/vulgatus* and *Bacteroides fragilis* (from 8.2 % to 29.7 %, respectively from 1.4 % to 14.8 %), an increase in relative abundance of *Roseburia intestinalis* (from 1.0.3 % to 8.0 %) and a decrease in relative abundance of *Bifidobacterium longum* (from 26.7 % to 0.5 %). There was an interval of almost one year between TP1 and TP4, which is the largest time span within this cohort (Figure 2.5.1). However, the Bray-Curtis dissimilarity index between TP1 and TP4 was 0.36, which is lower than the median intra-individual variability (Figure 3.1.5), indicating that the GIT microbial community within this patient was nonetheless quite stable. The genus

Akkermansia was detected in each sample and accounted for 2.70 %, 0.01%, 5.89 % and 0.01 % relative abundance, respectively. The observed changes in the microbial community including the evolution of high abundance of the genus *Bifidobacterium* to higher abundance of the genus *Bacteroides* might have been age related (Ottman, Smidt, de Vos, & Belzer, 2012). Shannon diversity of the bacterial population, both on MG and on MT level stayed relatively high throughout the treatment with values ranging between 2.8 and 3.4 (Figure 3.1.18B).

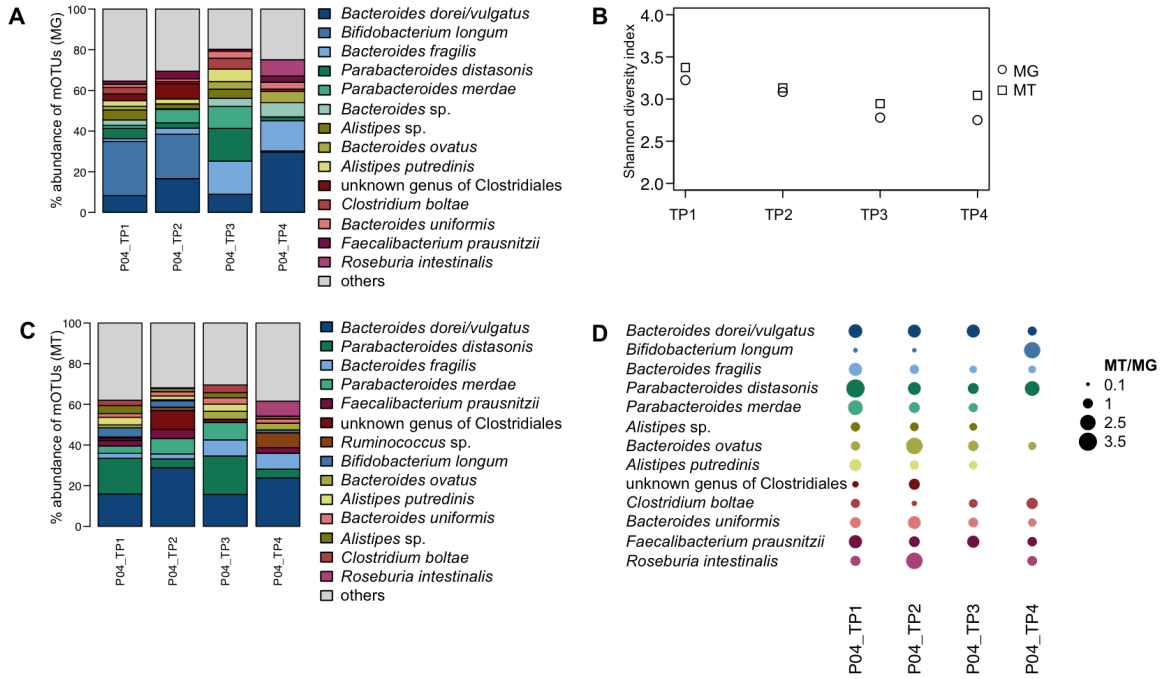


Figure 3.1.18: Variation of the microbial community structure over the course of the treatment in a patient who developed severe mucositis. (A) Relative proportions of the 14 most abundant metagenomic OTUs (mOTUs), based on MG and (C) MT reads. The remaining OTUs are summarized as 'others'. (B) Bacterial diversity represented by Shannon diversity index at sampling TPs throughout the treatment, based on MG (circle) and MT (square) reads. (D) Ratio of the MT to MG relative proportion of the 13 most abundant mOTUs (in both datasets). The size of the dot indicates the ratio. Absence of a dot indicates that the organism was not detected in the MT dataset.

Compared to the median intra-individual variability as indicated by the Bray-Curtis dissimilarity index (Figure 3.1.5), patient P04 showed relatively little variation between different TPs with a Bray-Curtis dissimilarity index of 0.23 between TP1 and TP2, 0.33 between TP2 and TP3 and 0.43 between TP3 and TP4. Although this patient developed severe mucositis, the core microbiome appears to have stayed relatively unaffected and stable, indicating that the colonic microbiome might not be strongly affected by the treatment and mucositis. Treatment and mucositis might have stronger effects on bacteria

such as the mucin-degrading *Akkermansia muciniphila*, which adheres to the intestinal epithelium (Reunanen et al., 2015).

Comparing the 14 most abundant mOTUs on MG and on MT levels (Figure 3.1.18A and Figure 3.1.18C), 13 mOTUs were found to agree. Some mOTUs (such as *Bacteroides dorei/vulgatus* and *Bacteroides fragilis*) were even positioned at the same rank on MG and MT level. For most mOTUs however, their abundance rank was not exactly the same on MG and MT level. In Figure 3.1.18D, the ratio of the MT to MG reads that were mapped onto a specific mOTU, is represented. This representation indicates that the relative activity of the mOTUs varied over time and that the organism displaying the highest activity differed at each TP. This also demonstrates that the organism showing the highest relative abundance in a sample cannot be expected to also display the highest activity. Thus, the most abundant organisms might not be the most important ones within an ecosystem. This new layer of information (MT sequencing) might be important in order to reveal the microbial activity and its possible implications in human health.

3.2 Meta-omic analyses of the gastrointestinal tract microbiome in adult patients undergoing allogeneic stem cell transplantation

The second section focuses on the changes within the GIT microbiome of hematology oncology patients undergoing an allogeneic stem cell transplantation (allo-HSCT). The results within this chapter are based on 78 fecal samples collected from 27 patients. After processing and filtering of the 16S rRNA gene amplicon sequences, 76 datasets were left, with 183,000 ± 55,000 (mean ± standard deviation) reads per sample. 73,000 ± 79,000 18S rRNA gene amplicon sequences per sample were retained for 61 datasets.

In order to compare the changes within the GIT during and following treatment, specific time frames relative to the day of transplantation were defined as time points (TP). These time frames are visualized in (Figure 2.5.2 in section 2.5). The first TP (TP1) includes samples that were taken up to two weeks before the transplantation. TP2 includes samples that were collected within one week after transplantation, TP3 includes samples that were collected around engraftment (3-6 weeks after transplantation). TP4 includes samples that were taken later, between 100 and 260 days after transplantation. Together, these TPs included 62 samples from 24 different patients.

Of 9 samples, MG and MT combined datasets could be produced. Additionally, 34 MG datasets were produced. After processing with IMP which included filtering out low quality reads and reads mapping to the human genome, per dataset 47,000,000 ± 18,000,000 MG sequences and 61,000,000 ± 23,000,000 MT sequences were kept for the following analyses and assembly. From the assemblies, 92,000 ± 71,500 genes per sample were predicted.

3.2.1 Patient characteristics and treatment

Anthropometric and clinical information of the seventeen male and ten female patients included in the following analyses are provided in Table 2.1.2. The patients were between 22 and 67 years old (median 54). Fourteen patients were treated for acute myeloid leukemia (AML), ten were treated for lymphoma, one for acute lymphoblastic leukemia (ALL), one for chronic myleloid leukemia (CML) and one for myeloma. Eleven patients received stem cells from a matched unrelated donor, nine donors were matched and related and seven were mismatched and unrelated. Ten patients were conditioned with fludarabine (Flu), busulfan (Bu) and cyclophosphamide (Cy). Nine patients were treated with BuCy, one with FluBu, three with treosulfan (Treo) and Flu, one with total body irradiation (TBI) and Flu, one with TBI and Cy and two with FLAMSA-Bu (Flu, cytarabine,

amsacrine, Bu). Of the 27 patients, sixteen developed GvHD. Of those, six developed GvHD with implication of two or more organs and of those, four patients developed severe GvHD (with summed stages ≥ 4). 1.5 years after allo-HSCT, nine patients had deceased; three due to relapse, three due to GvHD, two due to pneumonia and one due to sepsis. The sex of the patient was not determined as risk factor for higher GvHD grade (Kolmogorov-Smirnov test). No correlation between the sex or the age of the patient and the overall outcome was determined (Fisher's exact test). In this cohort, the ratio of patients who deceased was higher in lymphoma patients than in leukemia patients compared to a random distribution. The grade of GvHD was positively correlated with the age of the patient (*p* value 0.042, Spearman's rho 0.393).

3.2.2 Changes in the prokaryotic GIT microbiome of patients undergoing allo-HSCT

The prokaryotic (bacterial and archaeal) community composition was assessed based on 16S rRNA gene amplicon sequencing of DNA extracted from 78 fecal samples from 27 patients. After filtering and removal of samples with a low number of reads (as described in section 2.4), 76 of the sequenced samples were kept for the following analyses. The overall 14 most abundant orders within all 76 samples were identified to get an overview of the composition of the GIT microbiome of the patients (Figure 3.2.1). Overall, the majority of the samples displayed a high relative abundance of Bacteroidales, often in combination with a high abundance of Clostridiales. Some samples were almost completely dominated by Lactobacillales (such as A05_4, A08_2, A13_1, A18_4, A20_2, A27_3, A29_3 and A35_3), while others displayed high relative abundance of Enterobacteriales (as A05_5, A07_2, A08_3 and A11_1). Even on this taxonomic level, drastic changes in the GIT microbiome composition in different samples from one patient were observed. For example, A03 displayed high relative abundance of Burkholderiales in the second sample (90 %), which made up only 0.01 % in the first sample, or A41 which showed a high relative abundance of Verrucomicrobiales in the first sample (40 %) which decreased to 0.03 % and 0.23 % in the second and third sample, respectively. For other patients such as A21, no drastic changes from one sample to the next were observed.

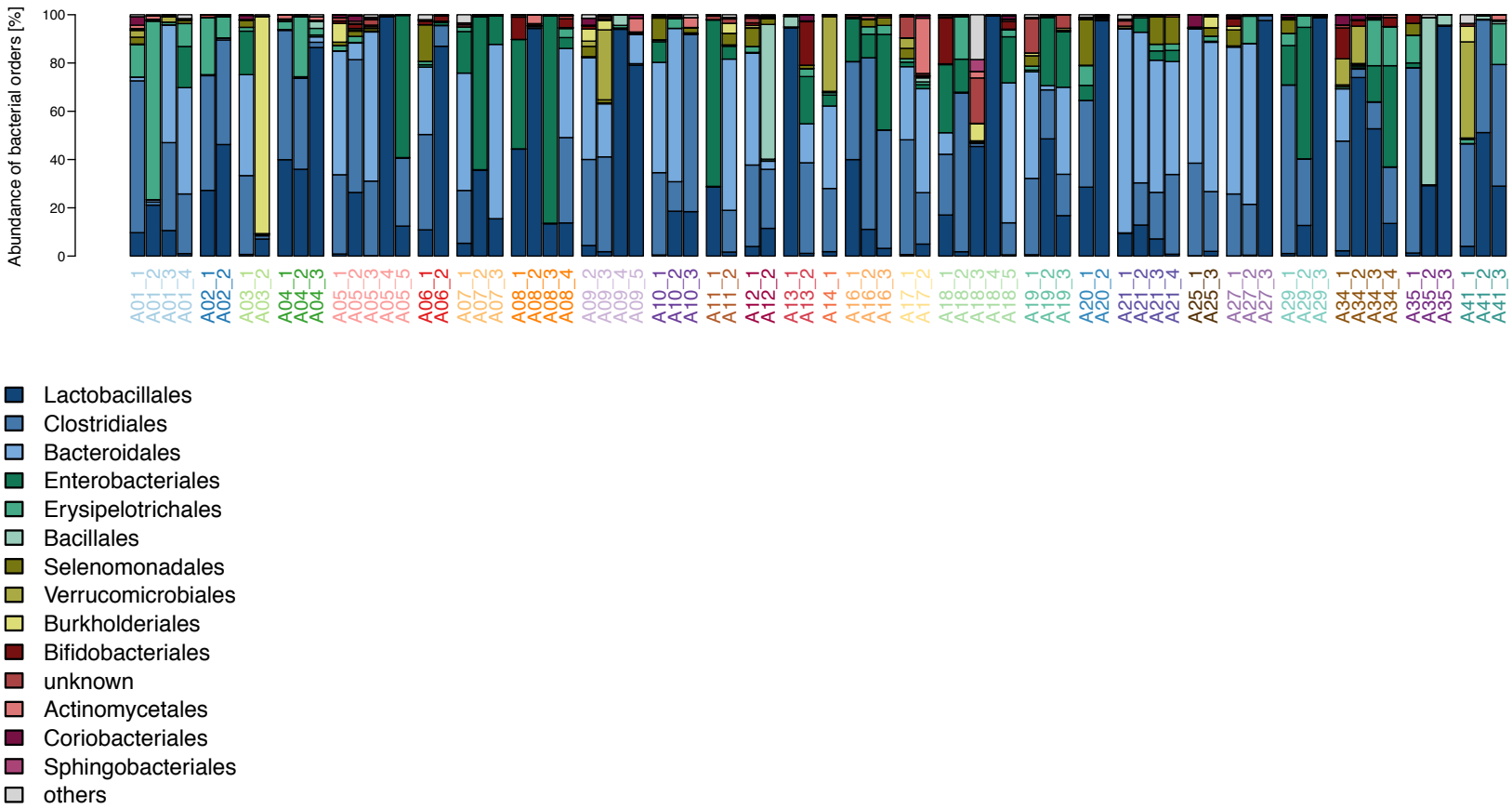


Figure 3.2.1: Relative abundance of the 14 most abundant bacterial orders in fecal samples from patients undergoing an allogeneic stem cell transplantation (allo-HSCT), grouped according to patient. Orders which were not comprised in the 14 most abundant orders are combined as 'others'. OTUs which could not be classified at the order level are grouped as 'unknown'. Patient ID and number of the sample are indicated below each respective bar and colored according to the patient.

In the following, the samples were grouped according to the TP (Figure 3.2.2). Only samples for which the collection fell into one of the defined time periods (as displayed in Figure 2.5.2) were included and designated TP1-4. Analyses based on these TPs included 62 samples from 24 different patients. In these barplots, the 14 most abundant genera are displayed. For some samples, dominance of one genus was observed, often by *Enterococcus* spp.(e.g. A08_TP2 and A29_TP4), *Escherichia/Shigella* spp. (e.g. A08_TP3 and A29_TP3) and sometimes by *Lactobacillus* spp. (A27_TP3). On this taxonomic level, the drastic changes between TPs from one patient are even more apparent than on the level of orders. Similarly, large changes between samples from different patients were observed, indicating that both the initial composition of the GIT microbiome but also the changes throughout treatment are individual-specific. No clear specificity of the relative abundance of bacteria specific to individual TGs or TPs was apparent.

Overall, the majority of the reads were assigned to the domain bacteria, with only an average of 0.06 % ± 0.42 % reads over the 62 samples being assigned to the domain archaea. Archaea were detected in 21 out of 62 samples and comprised three different genera, the most abundant genus being *Methanobrevibacter*, followed by *Methanosarcina* and *Methanosaeta*.

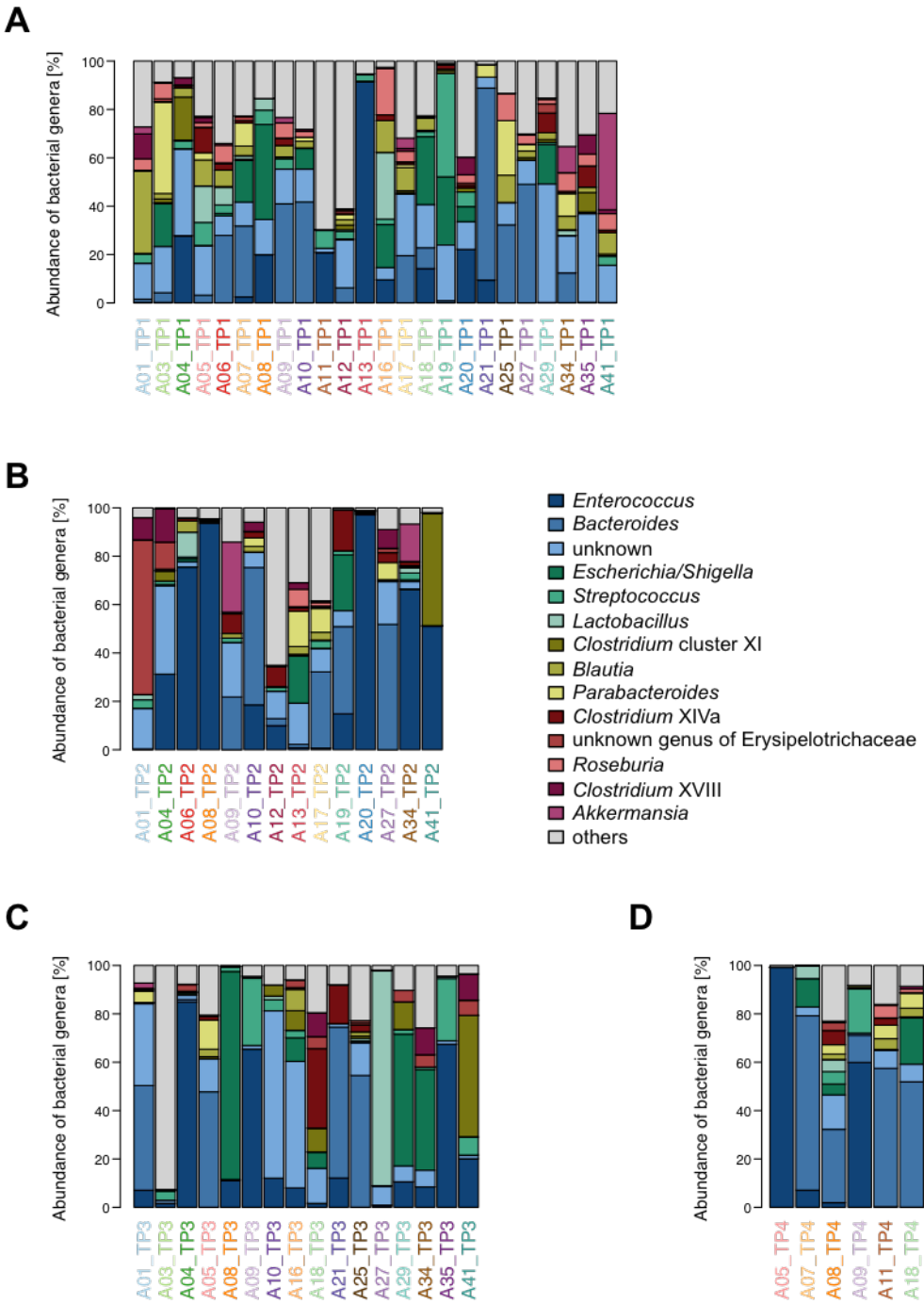


Figure 3.2.2: Relative abundance of the 14 most abundant bacterial genera in fecal samples from patients undergoing an allo-HSCT. Samples are grouped according to TPs: (A) TP1, (B) TP2, (C) TP3 and (D) TP4. Genera which were not comprised in the 14 most abundant genera are combined as 'others'. OTUs which could not be classified at the genus level are grouped as 'unknown'. Patient ID and sampling TP are indicated below each respective bar and are colored according to the patient.

As the taxonomic overview revealed drastic changes within the community, I wondered whether this was also reflected in diversity and richness. Figure 3.2.3 represents bacterial Shannon diversity and Chao1 richness for each TP, which were assessed after rarefaction. Drastic changes both in diversity (Figure 3.2.3A) and richness (Figure 3.2.3B) were observed. A decrease in median diversity from 2.9 at TP1 to 1.7 at TP2 (samples of all patients: *p* value 0.007, Wilcoxon rank sum test and samples from the same individuals: p value 0.02, Wilcoxon signed-rank test) and a further decrease to 1.6 at TP3 (TP1 to TP3: *p* value 0.0006, Wilcoxon rank sum test, *p* value 0.001 Wilcoxon signed-rank test) were observed, followed by an increase to 2.2 at TP4. The variation of the diversity within each TP was high, ranging for example between 0.43 and 3.89 at TP1 and from 0.09 to 3.59 at TP4.

The same trends were observed for the bacterial richness, decreasing from median richness of 280 to 229 (*p* value 0.04, Wilcoxon rank sum test, 0.03 Wilcoxon signed-rank test), to 155 (TP1 to TP3: *p* value 0.001, Wilcoxon rank sum test, 0.002 Wilcoxon signed-rank test, TP2 to TP3: *p* value 0.03, Wilcoxon rank sum test and Wilcoxon signed rank-test) and increasing to 204 at TP4.

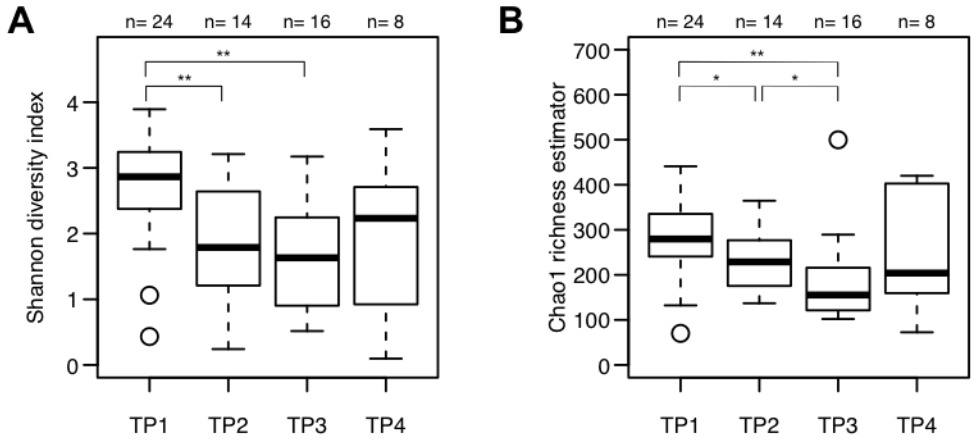


Figure 3.2.3: Changes within gastrointestinal bacterial community structure in patients undergoing allo-HSCT. Boxplots depicting (A) diversity (Shannon diversity index) and (B) richness (Chao1 richness estimator) per collection time point (TP), for prokaryotes (determined by 16S rRNA gene amplicon sequencing). The number of samples per collection TP is indicated above each box. Diversity and richness were determined after rarefaction of the dataset. (* when *p* value < 0.05, ** when *p* value < 0.01, Wilcoxon rank sum test)

In the following, the same patients and corresponding samples were grouped according to the overall outcome and to their collection TP (as visualized in Figure 2.5.2 in section 2.5). This included 24 patients of whom, 17 patients had survived (1.5 years after allo-HSCT)

while 7 patients had deceased due to different reasons, such as relapse, GvHD or pneumonia.

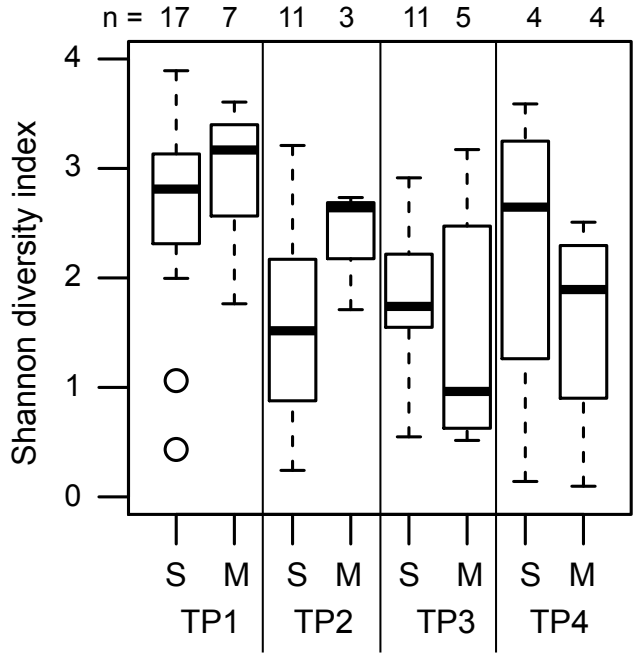


Figure 3.2.4: Shannon diversity indices of samples from patients who survived 1.5 years after allo-HSCT (S) compared to those who deceased (M). The samples are grouped per collection TP. The number of samples per collection TP is indicated above each box. Diversity was determined after rarefaction of the (16S rRNA gene amplicon sequencing) dataset.

For both groups, a statistically significant decrease in diversity from TP1 to TP3 was observed (Figure 3.2.4). No statistically significant difference between diversity in both groups at any TP was detected. However, at TP3 (around engraftment), a trend towards lower diversity in patients who later deceased compared to those who survived, was observed. This link between lower diversity at engraftment and higher mortality rate was also seen in a study including 80 patients (Y Taur et al., 2014). Possibly, antibiotic usage led to a lower diversity and its pressure selected pathogens, which could result in sepsis or infection. It is however not known whether this link between microbial diversity and the outcome is directly causal. An intermediate link in the form of infection and antibiotics might exist, meaning that possibly a severe infection required intensive antibiotic treatment, resulting in a lower bacterial diversity. Hence, the link between infection and outcome would only be indirectly reflected in the diversity index. Also, microbial diversity is more likely to be linked to mortality due to for example infection, pneumonia, sepsis or GvHD, than to mortality due to relapse. Due to the low number of patients in this group (7) however, I did not further discern the reasons for mortality in this analysis.

Similarly, the patients were in the following grouped according to development of severe GvHD (with summed stages ≥ 4) or no GvHD at all (Figure 3.2.5).

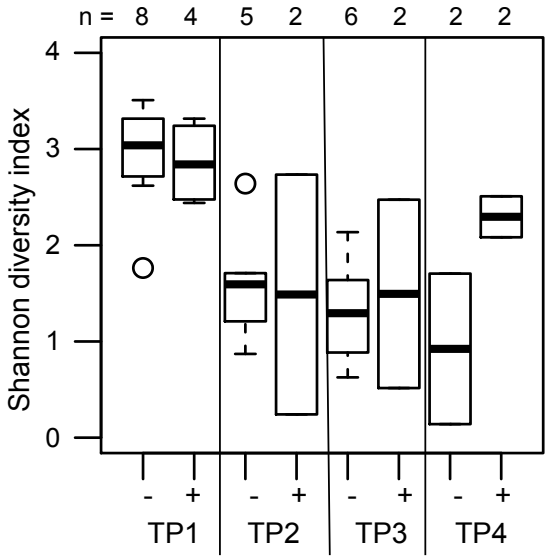


Figure 3.2.5: Shannon diversity indices of samples from patients who did not develop GvHD (-) compared to those who developed severe GvHD (+). The samples are grouped per collection TP. The number of samples per collection TP is indicated above each box. Diversity was determined after rarefaction of the (16S rRNA gene amplicon sequencing) dataset.

A statistically significant decrease in diversity from TP1 to TP2 in patients without GvHD (*p* value 0.006, Wilcoxon rank sum test) was observed. At TPs 1, 2 and 3, median diversity in both groups was similar, while at TP4, patients who had developed severe GvHD displayed a higher bacterial diversity. However, at TP4, both groups included only 2 samples, each. No link between bacterial diversity at TP1 and subsequent development of severe GvHD was identified. Hypothesizing that a GIT microbial dysbiosis (possibly caused by usage of several broad-spectrum antibiotics) with selection of pathogens would be implicated in initiation of GvHD (Khoruts et al., 2016), a lower Shannon diversity index in patients who developed severe GvHD would be expected. This was however not the case, median diversities in both groups stayed similar. GIT microbial diversity was therefore not predictive of development of severe GvHD. As the risk of developing GvHD is influenced by different factors such as the number of HLA-mismatches and sex disparity between donor and recipient, one should also account for these factors. Further grouping according to this was however not done due to the low number of patients with severe GvHD.

As drastic changes on different taxonomic levels and in bacterial diversity and richness had been observed, I wondered whether this was also reflected in the intra-individual and

inter-individual distance between microbial profiles. Bray-Curtis dissimilarity was determined based on the OTUs obtained from 16S rRNA gene sequencing data. High intra-individual dissimilarities were observed, reflecting the marked changes in individual microbial profiles. Still, a statistically significantly higher inter-individual dissimilarity (*p* value 0.003, Wilcoxon rank sum test) was observed (Figure 3.2.6). Both the intra- and the inter- individual dissimilarity in the adult hematology patients were higher than the ones observed in patients from the pediatric department (Figure 3.1.5) indicating even stronger perturbations in the GIT microbial community compositions in this patient cohort as well as strong effects of the intensive treatments on the GIT microbial community. Prokaryotic diversity at TP1 was negatively correlated with the dissimilarity index between TP1 and the following TP (*p* value 0.028, Spearman's rho -0.452). This indicates that a diverse microbiome was more stable and less affected by the treatment, while a GIT community with a low diversity underwent more drastic changes due to the treatment.

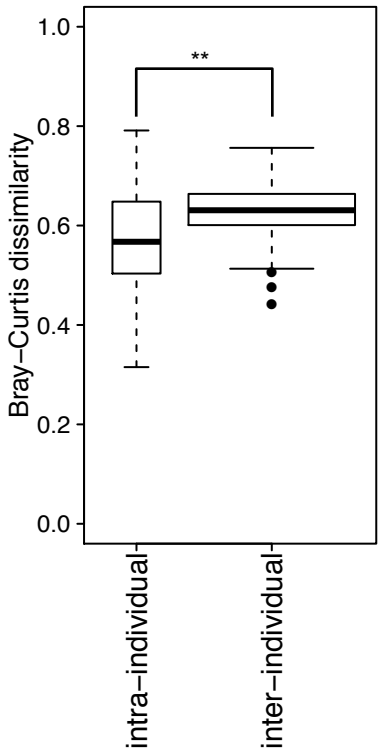


Figure 3.2.6: Comparison of intra-individual to inter-individual distances between bacterial profiles.

As the biggest changes in diversity and richness were observed from TP1 to TP3, I wondered whether this was reflected by general changes in relative abundance of specific taxa over the whole cohort. Between those TPs, 30 differentially abundant bacterial genera were identified (absolute $\log_2$ fold change ≥ 1, FDR-adjusted *p* value < 0.05, Table 3.2.1). A negative fold change indicates a decrease in abundance in samples from TP3

(n=16) compared to TP1 (n=24). Of those 30 genera, 14 genera decreased and 16 genera increased in relative abundance from TP1 to TP3. Figure 3.2.7 illustrates examples of bacterial genera, which decreased in relative abundance from TP1 to TP3.

Table 3.2.1: Differentially abundant bacterial genera in samples from collection TP1 and TP3

Genus	log_2 fold change	adjusted *p* value
Roseburia	-2.32	1.90E-10
Dorea	-2.20	1.20E-04
Barnesiella	-2.15	0.002
Butyricicoccus	-1.89	1.20E-04
Blautia	-1.58	4.11E-05
Dialister	-1.55	0.012
Gemella	-1.55	0.012
Collinsella	-1.42	0.016
Lachnospiracea incertae sedis	-1.38	0.004
Parabacteroides	-1.26	0.004
Ruminococcus	-1.25	0.002
Bifidobacterium	-1.20	0.005
Gemmiger	-1.06	0.022
Faecalibacterium	-1.03	0.022
Flavonifractor	1.02	0.020
Prevotella	1.13	0.045
Enterococcus	1.15	0.045
Lactobacillus	1.16	0.008
Streptophyta	1.40	0.020
Corynebacterium	1.45	0.008
Clostridium sensu stricto	1.53	0.002
Acinetobacter	1.69	0.008
Lactococcus	1.87	3.25E-04
Rothia	2.07	1.58E-04
Elizabethkingia	2.29	0.015
Clostridium cluster XI	2.29	1.39E-06
Brevundimonas	2.30	0.025
Erysipelotrichaceae incertae sedis	2.44	1.06E-09
Flavobacterium	2.45	1.20E-04
Staphylococcus	2.51	5.87E-06

After treatment, a decrease in the abundance of several bacterial genera that are considered to have health-promoting properties, such as *Blautia* spp., and the butyrate producers *Roseburia* spp. and *Faecalibacterium* spp., was observed. These, as well as *Dorea* spp. have been shown to diminish inflammation by modulation of the NF-κB

pathway (Lakhdari et al., 2011). Another decreased genus, *Barnesiella*, was shown to confer resistance to domination by vancomycin-resistant *Enterococcus* (Ubeda et al., 2013). On the other hand, an increase in for example the genera *Enterococcus*, *Clostridium* cluster XI, *Lactobacillus* and *Staphylococcus* was observed. Overall, the GIT microbiome composition after treatment might harbor a higher risk for infection and GvHD.

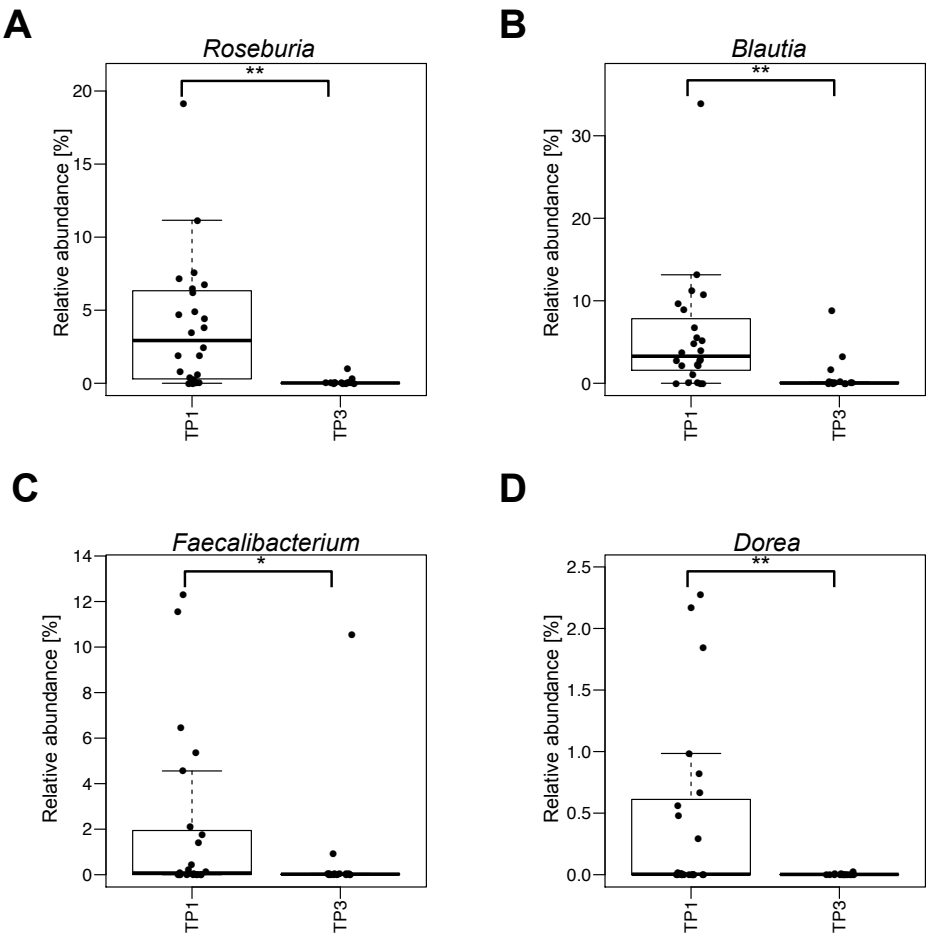


Figure 3.2.7: Examples of differentially abundant genera in samples from TP1 (n=24) and TP3 (n=16). Relative abundances of (A) *Roseburia*, (B) *Blautia*, (C) *Faecalibacterium* and (D) *Dorea* (* FDR-adjusted *p* value < 0.05, ** < 0.01, Wald test).

3.2.3 Changes in the microeukaryotic GIT microbiome of patients undergoing allo-HSCT

The microeukaryotic community composition was assessed using 18S rRNA gene amplicon sequencing of DNA extracted from 78 fecal samples of the patients. During the filtering steps, food-related reads as well as reads matching to the human genome were removed. Samples with a low number of reads after this filtering step were removed. 41 datasets were kept for further analyses. The 14 most abundant taxa were identified to get

an overview of the composition of the GIT microeukaryotic community of the patients (Figure 3.2.8). Overall, the most abundant taxon was Ascomycota. Sample A12_1 contained mostly Dinophyceae. Some samples (A05_3, A11_2, A18_3, A18_5, A19_1 and A21_1) displayed a similar combination of different taxa (including Oligohymenophorea, Prostomatea, Phylopharyngea and others).

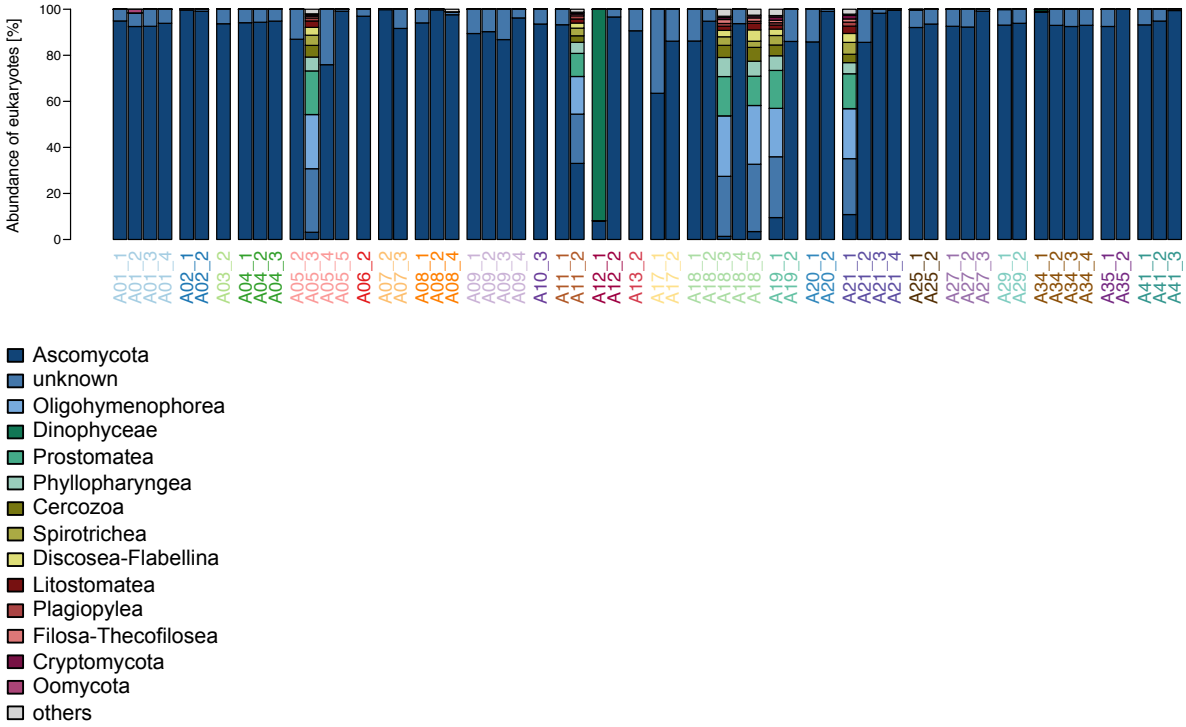


Figure 3.2.8: Relative abundance of the 14 most abundant microeukaryotic taxa in fecal samples from patients undergoing allo-HSCT, grouped according to patient. Taxa which were not comprised in the 14 most abundant taxa are combined as 'others'. OTUs which could not be classified at any taxonomic level are grouped as 'unknown'. Patient ID and number of the sample are indicated below each respective bar and are colored according to the patient.

In the following, the samples were grouped according to the TPs and different sequencing control data were included in this representation (Figure 3.2.9). The first four controls included extraction controls from different extraction batches. The last control (NTC) included a sequencing control (no template control, Figure 3.2.9E). The first extraction control (NC) contained only few human associated reads, which were removed during the filtering steps. The other controls included many reads that could not be classified at the genus level, as well as genera that were not or only lowly abundant in the patients' samples. Most samples from the patients included mainly the genus *Saccharomyces*, next to organisms, which could not be classified at this taxonomic level. On higher taxonomic levels, they were mostly assigned to Ascomycota.

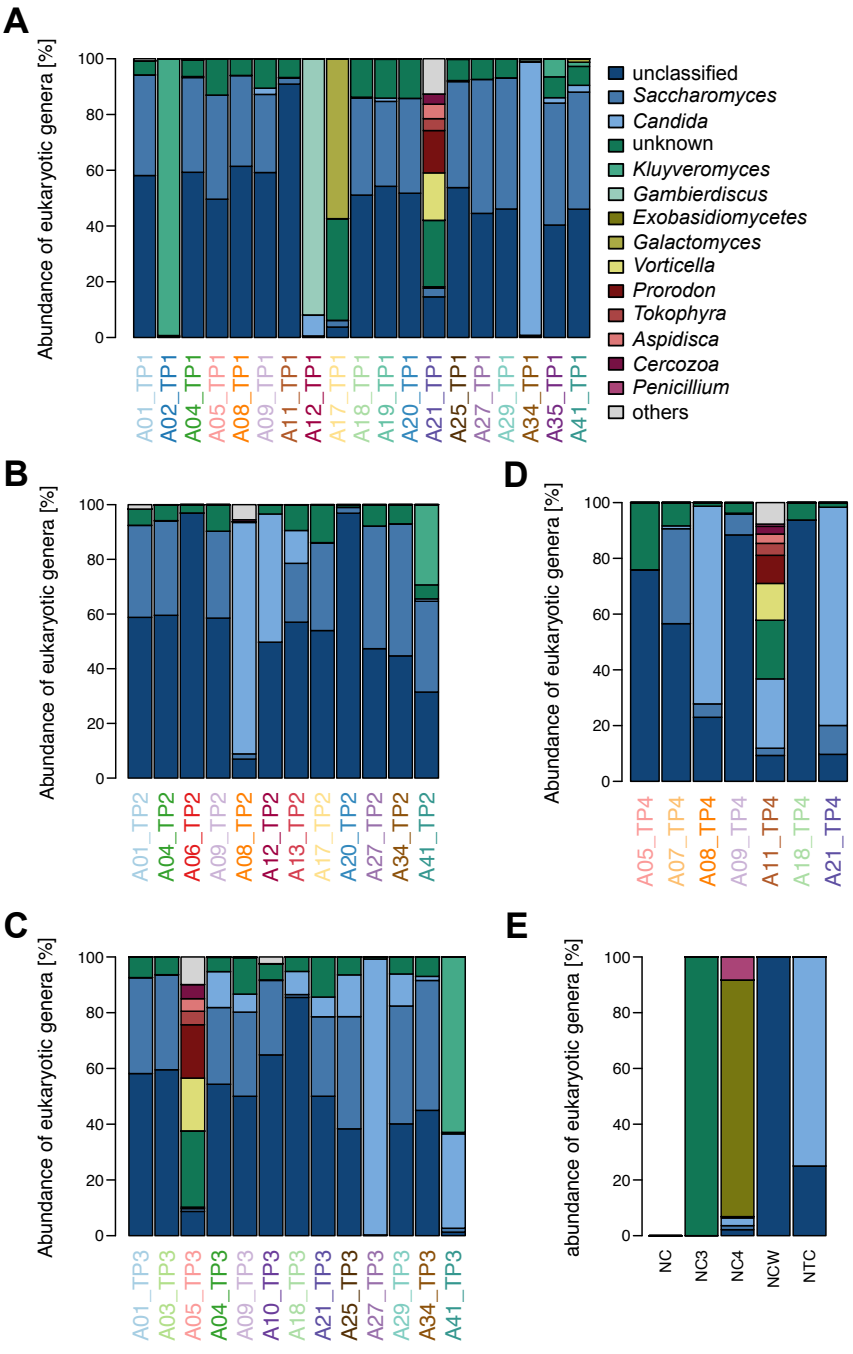


Figure 3.2.9: Relative abundance of the 14 most abundant microeukaryotic genera in fecal samples from patients undergoing allo-HSCT. Samples are grouped according to TPs: (A) TP1, (B) TP2, (C) TP3 and (D) TP4. Taxa which were not comprised in the 14 most abundant genera are combined as 'others'. OTUs which could not be classified at the genus level are grouped as 'unclassified'. OTUs which could not be classified at any taxonomic level are grouped as 'unknown'. Patient ID and sampling TP are indicated below each respective bar and are colored according to the patient. (E) includes the relative abundance of microeukaryotes sequenced in different extraction and sequencing controls.

Some samples, such as A34_TP1, A08_TP2, A27_TP3 and A21_TP4 mostly contained reads assigned to the genus *Candida*. No clear difference in the relative abundance of microeukaryotes specific to individual TPs was apparent. As in the previous representation, several samples displayed a similar taxonomic composition including a high ratio of unclassified microeukaryotic genera, as well as reads assigned to the genera *Vorticella*, *Prorodon* and others. However, the different extraction and sequencing controls included in Figure 3.2.9E did not show this specific composition of genera, indicating that it was not artifactual. No explanation for this peculiar composition of microeukaryotic taxa in different samples was found.

As the prokaryotic community displayed drastic changes in terms of diversity and richness from one TP to another, I wondered whether the microeukaryotic community showed similar trends. Diversity and richness of the microeukarotic community were determined on OTU level after rarefaction (Figure 3.2.10). No statistically significant differences in diversity or richness between different TPs were observed. Overall, in the microeukaryotic community, a lower median diversity (ranging between 1.3 and 1.7) and lower median richness (ranging between 65 and 90) than in the prokaryotic community (Figure 3.2.3) were observed. No drastic changes between different TPs were observed, indicating that the microeukaryotic community was not strongly affected by the treatment.

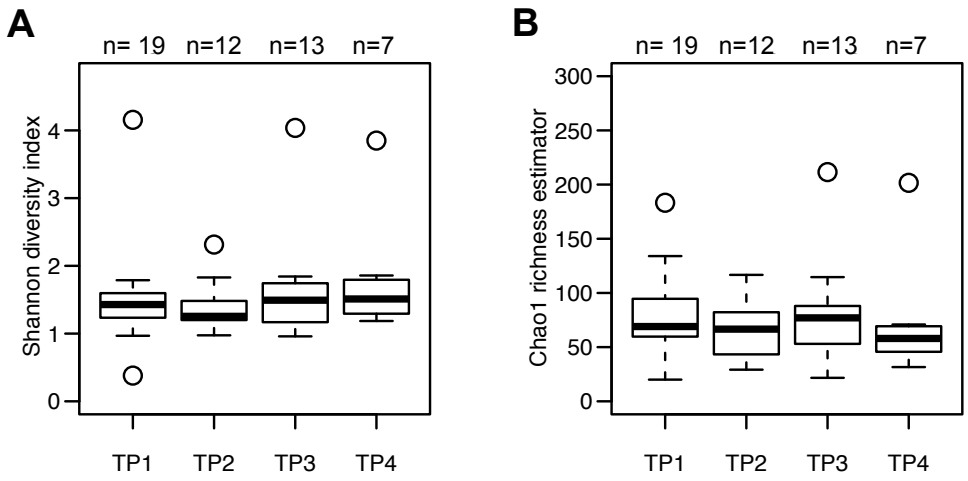


Figure 3.2.10: Changes in the gastrointestinal microeukaryotic community structure in patients undergoing allo-HSCT. Boxplots depicting (A) diversity (Shannon diversity index) and (B) richness (Chao1 richness estimator) per collection time point (TP), for microeukaryotes (determined by 18S rRNA gene amplicon sequencing). The number of samples per collection TP is indicated above each box. Diversity and richness were determined after rarefaction of the dataset.

When grouping the patients according to antifungal treatment, a slightly lower median diversity at TP2 (1.24 against 1.41) in patients who were treated with antifungals,

compared to no antifungal treatment was observed. However, this difference was not statistically significant.

3.2.4 Virome profiling within the GIT microbiome of hematology cancer patients

Besides 16S and 18S rRNA gene amplicon sequencing, which allow identification of prokaryotes and eukaryotes, this project also included MG and MT shotgun sequencing of the samples. This allows identification of viruses, which are also numerous in the GIT community. While many of them do not pose problems in healthy individuals, they might represent a serious threat in immunocompromised and immunosuppressed patients (Sahin et al., 2016).

On average (± standard deviation), 0.017 % ± 0.014 % MG and 0.089 % ± 0.171 % MT reads were mapped onto viral genomes within the 43 MG, respectively 9 MT samples (Figure 3.2.11A and Figure 3.2.11B). One of the highest ratios of MG reads mapping onto viral genomes was found in A18_3 (0.055 %), with 25 % of those reads mapping to the torque teno virus, which is very common in the human population and has been suspected to cause several pathologies. However, due to its ubiquitous state in the population, it is difficult to discern its pathogenic potential (Kincaid, Burke, Cox, de Villiers, & Sullivan, 2013; Okamoto, 2009). The highest ratio of MT reads mapping to viral genomes was detected in A11_2 (TP4), adding up to 0.54 %. 98 % of these reads mapped to Tobamovirus genomes, including the tomato mosaic virus and the pepper mild mottle virus.

As many reads mapped to plant-associated viruses, the same analyses were repeated, but with human-associated viral genomes only (Figure 3.2.11C and Figure 3.2.11D). On average, 0.0012 % ± 0.0036 % MG and 0.005 % ± 0.012 % MT reads were mapped onto human-associated viral genomes within the 43 MG, respectively 9 MT samples. The highest ratio of MG reads mapping to human-associated viruses was detected in A18_4 (0.018 %), with 86 % of those reads mapping to the BK polyomavirus. This is a widepread virus and infections are usually harmless. However, infection in immunocompromised individuals can have severe consequences, such as hemorrhagic cystitis (Bennett, Broekema, & Imperiale, 2012; Dropulic & Jones, 2008). In sample A34_4, 42 % of the viral reads mapped to cytomegalovirus, which can cause severe colitis in immunosuppressed patients with symptoms similar to those of GIT GvHD such as diarrhea and abdominal pain (Jacobsohn & Vogelsang, 2007; Sahin et al., 2016).

Within the MT datasets, one sample showed a drastically higher ratio of reads mapping to viral genomes than the average: A07_3 (0.039 %). These reads mainly mapped to the Torque teno virus (89 %), which has a ssDNA genome which should not be detected in

the MT datasets. However, during an active phase in which the virus replicates, viral RNA of this virus might also be detected. Incidentally, this virus was also detected within the MG dataset of this sample and accounted for 78 % of the viral reads. Other human RNA viruses detected include for example the human enteric coronavirus (in A06_1).

To summarize, shotgun sequencing allows deep profiling of the GIT microbiome, including the virome and might enable detection of for example the cytomegalovirus, which can cause serious infections after allo-HSCT which are difficult to treat (Kharfan-Dabaja et al., 2012).

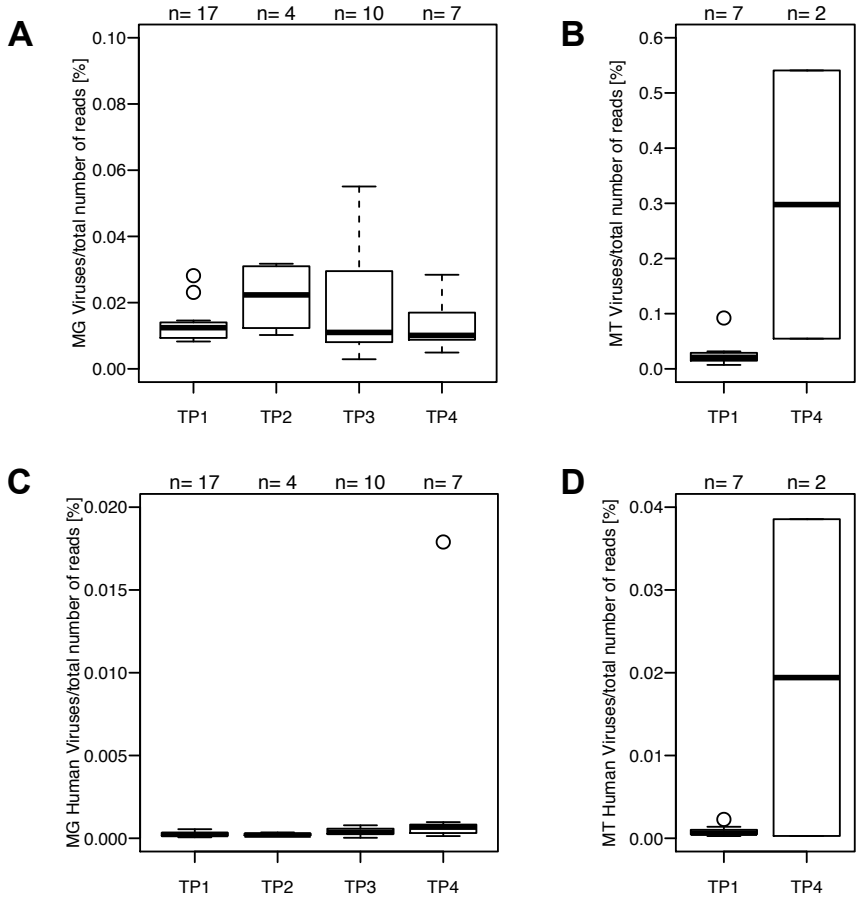


Figure 3.2.11: Relative abundance of reads mapping to viral genomes. Boxplots depict MG (A, and C) or MT (B and D) reads mapping to viral genomes grouped per TP. (A and B) include reads mapping to all eukaryote-associated viruses. (C And D) include reads mapping to human-associated viruses. The number of samples per collection TP is indicated above each box.

3.2.5 Variability of GIT microbiome trajectories in patients throughout treatment

In the following, I will focus on the 43 MG datasets obtained from 21 patients. Within those 43 MG datasets, after filtering out reads mapping to the human genome, on average, 82.88 % (median 94.24 %) ± 24.90 % reads were retained, which were of microbial origin.

Only for few samples (A05_4, A18_3, A27_3, A29_3 and A35_3), less than 50 % of the reads were retained as the rest mapped to the human genome. This high amount of human DNA within the sequenced sample indicates the presence of human cells in the corresponding fecal samples, which could be due to the presence of intestinal epithelial cells and might indicate a bad health status of the patient. The detection of high amounts of human DNA did not always coincide with the occurrence of GIT GvHD and did not correlate with the overall outcome.

The overview of the most abundant bacteria in different samples (Figure 3.2.1) as well as the intra-individual dissimilarity index (Figure 3.2.6), both based on 16S rRNA gene sequencing, indicated big changes from one collection TP to the next. A principal component analysis (PCA) based on metagenomic OTUs (mOTUs) identified in MG sequencing data revealed similar trends (Figure 3.2.12) as samples from the individual patients do not cluster together.

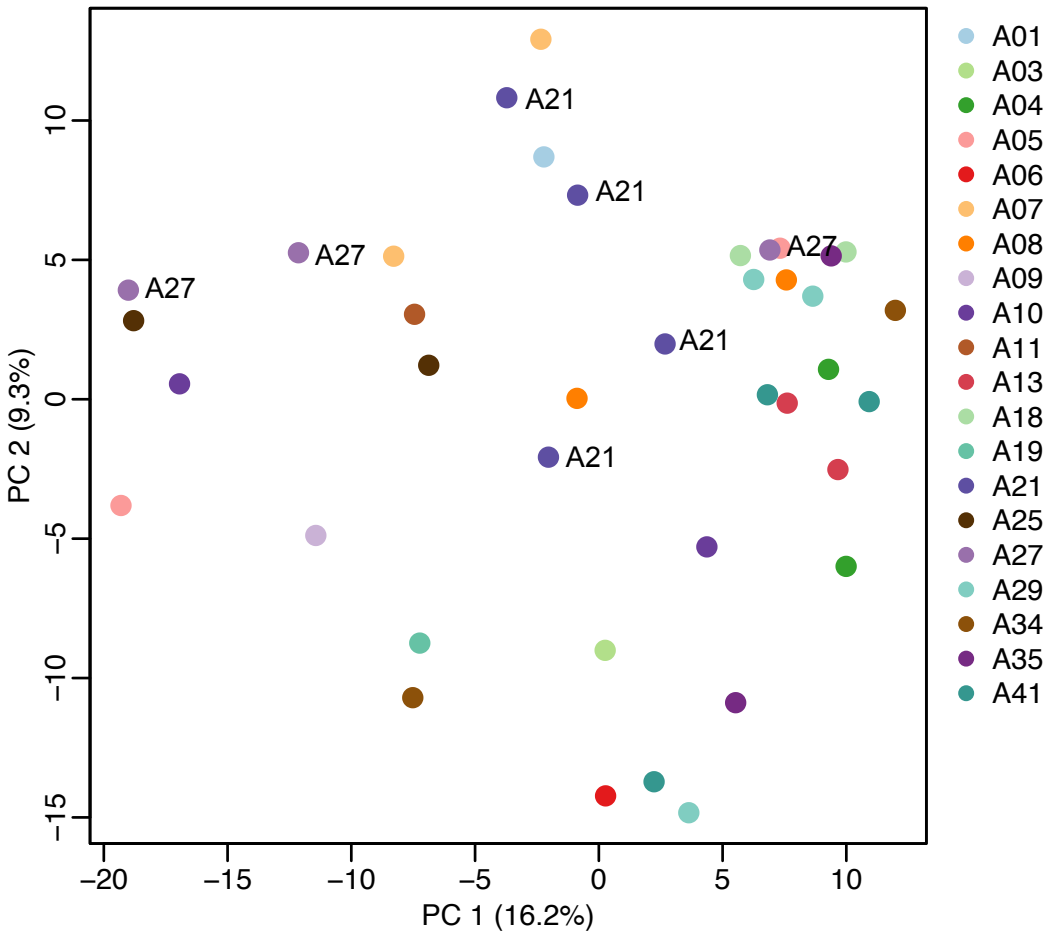


Figure 3.2.12: Principal component analysis (PCA) for GIT prokaryotic community composition. Each dot represents a sample colored according to the corresponding patient. PCA was performed based on the metagenomic operational taxonomic units (mOTUs) identified in MG sequencing data. Samples from patient A21 and from patient A27 are highlighted.

In the following, I will focus on two patients (A21 and A27) who displayed strongly diverging evolutions of their health status. The corresponding samples are highlighted in Figure 3.2.12 and Figure 3.2.13.

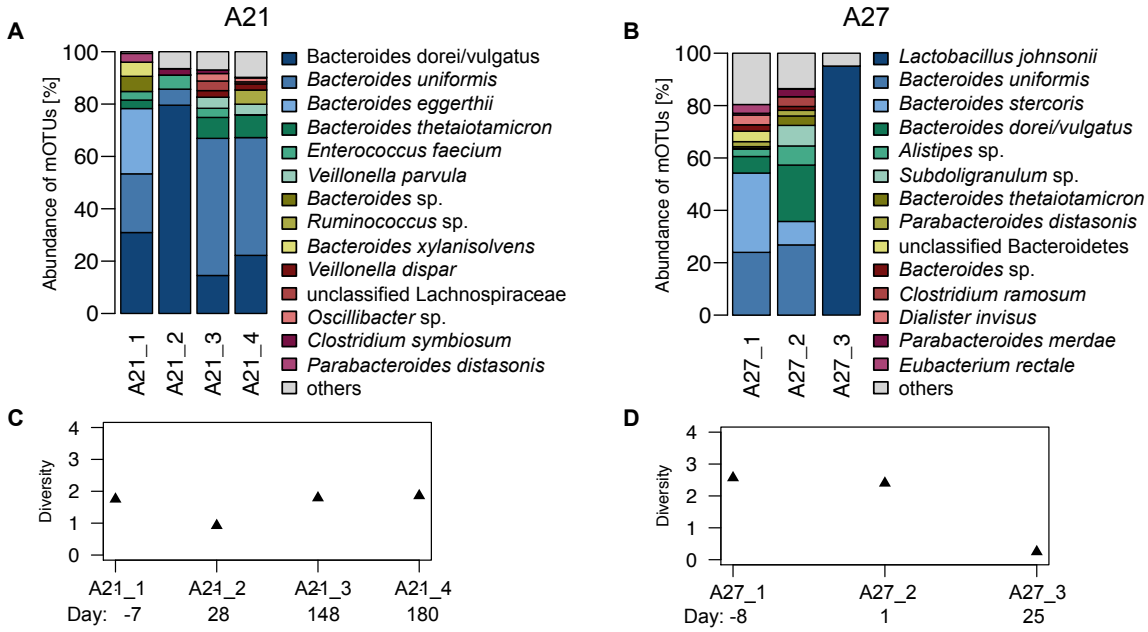


Figure 3.2.13: Variation of the microbial community structure over the course of the treatment in two hematology patients. (A) and (B) Relative proportions of the 14 most abundant metagenomic operational taxonomic units (mOTUs) based on MG sequencing. The remaining mOTUs are summarized as 'others'. (C) and (D) Bacterial diversity represented by Shannon diversity index in different samples taken throughout the treatment. Days of the collection TP relative to allo-HSCT are indicated below the plot. Plots represent the corresponding results for patient A21 (panels A and C) and patient A27 (panels B and D).

Patient A21 was 64 years old at the beginning of treatment and was treated with a treosulfan/fludarabine conditioning regimen for acute myeloid leukemia. The graft was from a matched unrelated donor. The patient developed light GvHD with implication only of the skin (stage II°) and was still alive two years after allo-HSCT. The most abundant mOTUs in all four samples included different species of the genera *Bacteroides*, as well as *Veillonella* and *Ruminococcus* (Figure 3.2.13A). *Enteroccocus faecium* was detected in each sample, however never in high abundance. This patient was already treated with many different antibiotics (piperacillin/tazobactam, meropenem, vancomycin, linezolid) before the first sample was collected and the intensive antibiotic treatment was continued for some time, which might have lead to the overall relatively low bacterial diversity (Shannon diversity index), ranging between 0.9 (A21_2) and 1.9 (A21_4) (Figure 3.2.13C). Although one species was dominating in sample A21_2, the majority of the other bacteria did not disappear completely but were still detectable. The number of

different mOTUs detected from sample 1 to sample 4 evolved from 31, over 48, to 41 and finally 34.

Patient A27 was 56 years old at beginning of the treatment and was treated with a busulfan/cyclophosphamide conditioning regimen for chronic myeloid leukemia. The graft was harvested from a matched related donor. 22 days after allo-HSCT, the patient developed severe GvHD (skin stage III°, GIT IV°, liver III°). The patient deceased 53 days after allo-HSCT due to steroid-resistant GvHD. Some of the most abundant taxa in the first two samples included different species of the genus *Bacteroides*, *Alistipes* sp. and *Parabacteroides distasonis*. Bacterial diversity was relatively high at the beginning of the treatment with a Shannon diversity index of 2.6 in the first sample and 2.4 in the second sample. The third sample (collected three days after onset of GvHD) had an extremely low diversity (0.3), in which only eight different taxa could be detected (Figure 3.2.13D). Of the eight different taxa identified in the third sample, only the lowest abundant (*Clostridium butyricum*, 0.04 %) was a strictly anaerobic bacterium. All the other bacteria were facultative anaerobic bacteria (such as *Lactobacillus johnsonii*, *L. rhamnosus*, *Enterococcus faecium* and *Streptococcus thermophilus*).

After quality filtering, 84 % of the MG reads from the third sample were removed, as they mapped to the human genome. The collection of this sample coincided with severe aGvHD implicating the GIT. Thus, the sample could have contained human cells, possibly intestinal epithelial cells. This sample was dominated by *L. jonhsonii*, which was also detected in the first samples but in low relative abundance. Considering that the third sample contained a large proportion of human DNA and the microbial community composition is represented in relative abundance, in relation to the total bacterial DNA within the sample, this probably means that there were only very little bacteria left in the GIT and almost only facultative aerobic bacteria were able to survive, which can possibly be linked to aGvHD (Jenq et al., 2012).

This detailed look at the taxonomic GIT community profiles of two patients again makes clear how drastic but also how individual-specific the changes within the microbial community can be, especially in such a heterogeneous cohort undergoing intensive treatments.

As in the PCA plot (Figure 3.2.12), the samples did not cluster according to the patients, I wondered whether the samples clustered according to their collection TP. Figure 3.2.14 represents the same analysis, a PCA plot based on the mOTUs identified in MG sequencing data. Here, the samples are colored according to their collection TP. No clear clustering is observed in this plot. However, the TP1 samples (dark blue) tend to be

separate from samples from other TPs. The main drivers are indicated and show that the main driver for TP1 samples was *Ruminococcus* sp. 5_1_39BFAA. One sample from TP1 is closer to the samples from later TPs. This is A13_TP1, which has a very high relative abundance of *Enterococcus* spp. (Figure 3.2.2), which is the main driver guiding the samples from later TPs in this opposite direction of the plot. Different species of the genus *Bacteroides* act as additional drivers. Loss of health-promoting bacteria such as *Ruminococcus* spp. as well as expansion of *Enterococcus* spp. after allo-HSCT has been observed in several studies (Biagi et al., 2015; Holler et al., 2014; Ubeda et al., 2010).

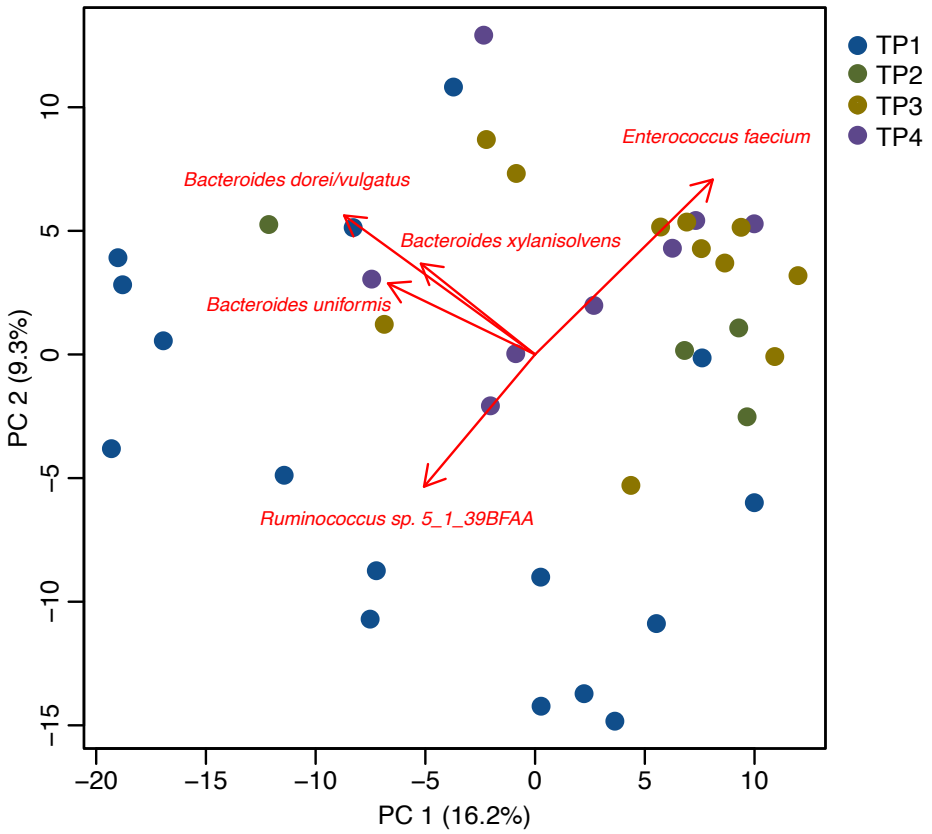


Figure 3.2.14: Principal component analysis (PCA) for GIT prokaryotic community composition. Each dot represents a sample colored according to the collection TP. PCA was performed based on the metagenomic operational taxonomic units (mOTUs) identified in MG sequencing data. The main drivers are indicated.

Besides taxonomic profiling, MG sequencing also allows functional profiling, which will be described in the next parts.

3.2.6 Detection of antibiotic resistance genes

Due to the underlying disease and/or the intensive conditioning treatment, hematology patients undergoing an allo-HSCT are immunocompromised, at least from the day the conditioning treatment takes effect, until engraftment takes place. As they are prone to infection, they are usually treated with prophylactic antibiotics, and treatment is continued during neutropenia or at occurrence of fever. Intensive antibiotic treatment can favor emergence of multi-drug resistant (MDR) bacteria. For the 43 samples where MG data was available, ARGs were detected (as described in section 2.11) and their relative abundance (percentage of ARGs relative to the total number of genes) was calculated (Figure 3.2.15). The mean relative abundance and standard deviation within these 43 samples was 0.33 % ± 0.11 %. As indicated in Table 2.1.2, the patients were treated with several antibiotics including broad-spectrum antibiotics such as fluoroquinolones and meropenem. Treatment with these antibiotics sometimes continued for several weeks. For almost all of the patients, an increase in the ratio of ARGs from one TP to the next is observed (Figure 3.2.15). Only in patient A08, a strong decrease was observed between A08_3 and A08_4. These samples were collected more than 7 months apart, during which the patient was only heavily and regularly treated with antibiotics for around a month. Except for four samples (A05_1, A27_1, A27_3 and A18_3), the ratio of ARGs within the samples from patients was higher than the ratio within four reference microbiomes from healthy individuals (RHM1-4). All of the ARGs identified within the sample with the highest ARG ratio, A29_3, belonged to one population-level genome. This was the only population-level genome that could be reconstructed within this sample. It was classified as *Enterococcus faecium*. During processing of this MG dataset, only 5.22 % of the (quality filtered) reads were retained, as the rest mapped to the human genome.

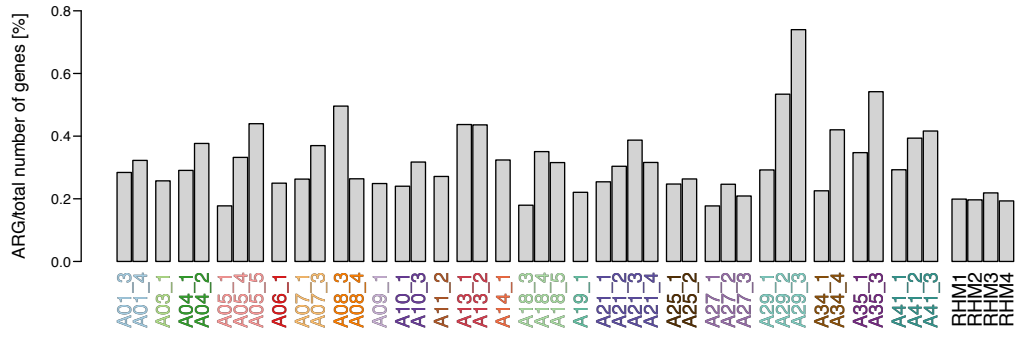


Figure 3.2.15: Relative abundance of antibiotic resistance genes in fecal samples from patients undergoing allo-HSCT. Samples are grouped according to patient. The labels are colored according to the patient. RHM1-4 represent the relative abundance of ARGs in 4 reference microbiomes from healthy individuals.

A drastic increase in the relative abundance of ARGs from TP1 to TP2 was observed (Figure 3.2.16), although the difference was not statistically significant, which might be due to the low number of samples at TP2 (n=4). Significant increases were observed from TP1 to TP3 (samples from all the patients: *p* value 0.013, Wilcoxon rank sum test and samples from the same individuals: *p* value 0.014, Wilcoxon signed-rank test) and from TP1 to TP4 (*p* value 0.008, Wilcoxon rank sum test). The biggest increase was observed from TP1 to TP2, probably because the intensive antibiotic treatment had begun in the time span between collection of both samples. The median ARG ratio at TP3 was lower than at TP2, (but still higher than at TP1). Here, the range of ARG ratio between different samples was high, going from 0.20 % to 0.54 %. This might be due to different intensities and duration of antibiotic treatment in the patients. Also at TP4, median ARG ratio was higher than at TP1 but lower than at TP2.

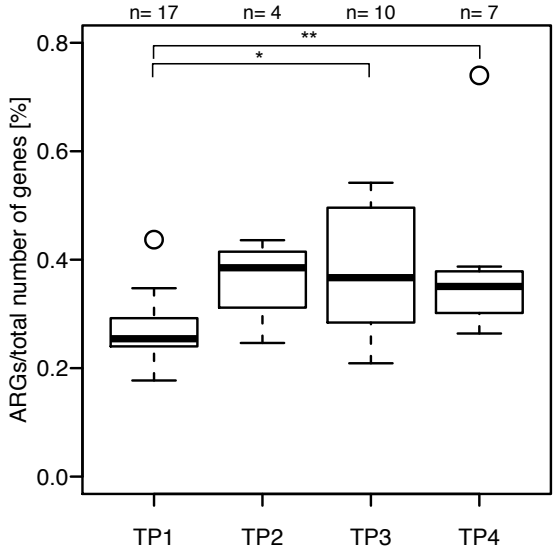


Figure 3.2.16: Relative abundance of antibiotic resistance genes in fecal samples from patients undergoing allo-HSCT. Samples are combined according to TP. The number of samples per collection TP is indicated above each plot.

3.2.7 Changes in the functional potential of the GIT microbiome during treatment

As the biggest changes in bacterial diversity and richness were observed from TP1 to TP3 (Figure 3.2.3), I wondered whether this was reflected in the functional potential of the GIT microbiome. Within the MG datasets, functions were detected. Genes with similar functions from different bacterial genomes were grouped together. These will in the following be referred to as 'functional gene categories'.

A total of 12,917 different functional gene categories were detected. Large differences in the number of functional gene categories between different TPs were detected. The

number of different functional gene categories (measured by the Chao1 richness estimator) at each TP indicated a decrease from a median value of around 7,000 at TP1 to 5,400 at TP2 (TP1 to TP2: *p* value 0.014, Wilcoxon rank sum test) and further to around 5,000 at TP3 (TP1 to TP3: *p* value 0.0005, Wilcoxon rank sum test).

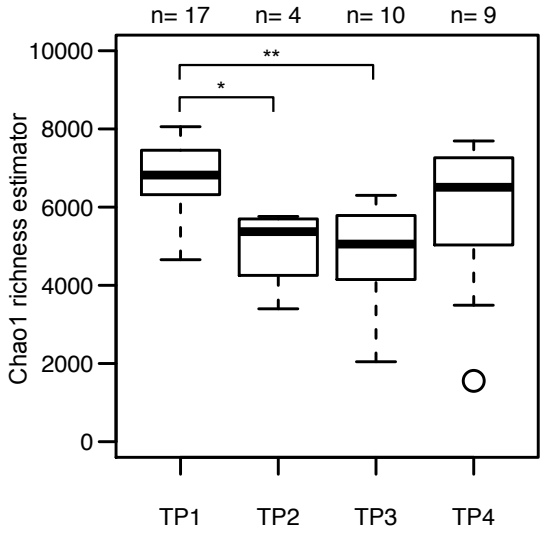


Figure 3.2.17: Richness of functional gene categories at different TPs. Boxplot depicting richness (Chao1 richness estimator) per collection time point (TP) of different functional gene categories. The number of samples per collection TP is indicated above each box. Richness was determined after rarefaction of the dataset. (* when *p* value < 0.05, ** when *p* value < 0.01, Wilcoxon rank sum test)

730 functional gene categories were found to be differentially abundant in the MG datasets between TP1 (n=17) and TP3 (n=10), with FDR-adjusted *p* values < 0.05 and absolute $\log_2$ fold change ≥ 1. Of those, 530 were decreased at TP3. As seen in Figure 3.2.17, in samples from TP3, there was a loss of around 2000 functional gene categories. Results of the differential analysis indicate, that the functional gene categories that were lost, were not always the same in every sample, which again shows that there is a large inter-individual variation and can probably be linked to the high taxonomic inter-individual variation within the GIT microbiome.

Upon further filtering, 64 functional gene categories were found to be differentially abundant with FDR-adjusted *p* value < 0.01 and absolute $\log_2$ fold change ≥ 3 (Figure 3.2.18). Of those, only 1 functional gene category (the putative carnobacteriocin-B2 immunity protein) were found to be increased, all the other functional gene categories displayed a decreased abundance at TP3. In healthy humans, even if there are changes within the composition of the GIT microbiome, the functional profile usually stays relatively stable (The Human Microbiome Project Consortium, 2012). Contrary to this, the functional capacity of the GIT microbiome of the adult hematology patients underwent drastic

changes throughout treatment. The decrease in abundance of a high number of functional gene categories at TP3 can probably be linked to a loss in diversity and richness of the GIT microbiome (Figure 3.2.3). In the less complex ecosystems in samples collected around one month after allo-HSCT, also the functional potential of this community seems to be decreased. The functional gene categories that displayed a statistically significantly lower abundance at TP3 included various functions such as genes needed for flagella construction, chaperones, spore formation, metabolic pathways (such as cellobiose degradation, sulfur metabolism) and many others, which is probably linked to loss of several bacterial populations.

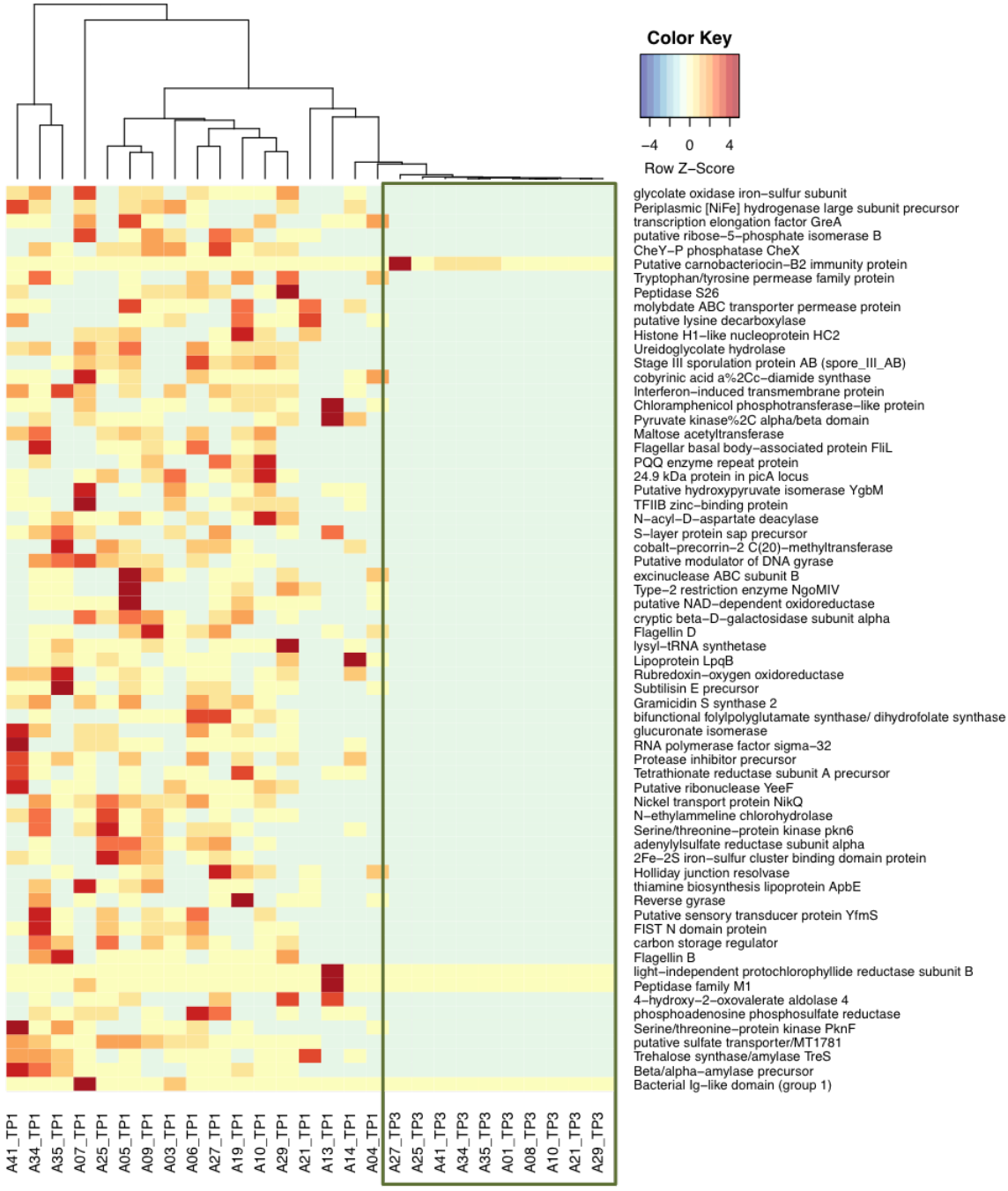


Figure 3.2.18: Heatmap of differentially abundant functional gene categories between collection TP1 and TP3 with FDR-adjusted *p* value < 0.01 and absolute $\log_2$ fold change ≥ 3. Samples from collection TP3 are highlighted in green. Heatmap of normalized functional gene category abundances is scaled as indicated in the color key. The dendrogram represents a hierarchical clustering with Ward's minimum variance method of the euclidian distance between the abundances displayed in the heatmap.

3.2.8 Does the microbiome influence development of GvHD?

One aim of this project was to assess whether the GIT microbiome is involved in development or aggravation of GvHD. Of the 27 patients, sixteen developed GvHD. Of those, four patients developed severe GvHD (with summed stages ≥ 4), considered as severe GvHD.

Previously detected differences of relative genus abundances between TP1 and TP3 (Table 3.2.1) encompassed a shift towards a GIT microbiome composition that might lead to a more inflammatory environment, thereby potentially harbouring a higher risk for GvHD development. As a specific composition of the GIT microbiome could potentially play part in development of GvHD, I compared the early samples (TP1 and TP2) from patients who later developed severe GvHD (n=12), to those who never developed GvHD (n=14). Ten genera were found to be differentially abundant with five being decreased in samples from patients who later developed severe GvHD (Table 3.2.2). The highest difference in relative abundance was seen for the genus *Akkermansia*, with a 10-fold reduction in samples from patients who did develop GvHD. As previously described in relation to mucositis development (section 3.1.7.1), this bacterium strengthens the epithelial barrier function and is inversely correlated with onset of inflammation (Derrien et al., 2016; Png et al., 2010; Schneeberger et al., 2015; Wu & Scott, 2012). Thus, loss of this bacterium could have led to impaired integrity and higher translocation of microbial products, possibly adding to initiation of GvHD. On the other hand, damage of the intestinal epithelial wall might have led to a decrease in the availability of the main nutrient source for *Akkermansia* (mucin), and thereby caused this decrease in relative abundance of *Akkermansia*. In this scenario, possibly a higher damage of the intestinal epithelial wall in these patients would represent a factor leading to GvHD development. In general, the differences observed in the GIT microbiome between both groups do not point towards a higher microbiome-induced inflammatory environment in the GIT of patients who later developed severe GvHD.

Table 3.2.2: Differentially abundant bacterial genera in (TP1 and TP2) samples from patients with severe GvHD compared to those who never developed GvHD

Genus	log_2 fold change	adjusted *p* value
Akkermansia	-3.32	1.29E-06
Bilophila	-2.67	0.006
Lachnoanaerobaculum	-1.88	0.040
Streptococcus	-1.61	0.005
Lactobacillus	-1.56	0.030
Parabacteroides	1.45	0.034
Coprococcus	1.70	0.040
Mogibacterium	1.83	0.050
Alistipes	2.18	0.001
Barnesiella	2.31	0.034

In the following, I compared samples that were collected while or after the patient had developed severe GvHD (n=3) against samples from patients who did not develop GvHD (n=21), including all TPs. Only two genera were differentially abundant, the genus *Lactobacillus* being more abundant in samples from patients with GvHD (2.71 log_2 fold change, FDR-adjusted *p* value 0.013) and the genus *Streptococcus* less abundant (-2.36 log_2 fold change, FDR-adjusted *p* value 0.013). In patients with GvHD, an increase in facultative anaerobic (aerotolerant) Lactobacillales has been observed and it was suggested, that they might be able to occupy this opening niche, in contrast to obligate anaerobes (Jenq et al., 2012). Here however, contradicting trends were observed, with increase in one facultative anaerobe and decrease in another facultative anaerobe. In the samples from early collection TPs from patients who later developed severe GvHD (Table 3.2.2), both genera were decreased. This indicates that, rather than being involved in development of GvHD, physiological changes within the GIT during onset of GvHD might have enabled an expansion of *Lactobacillus* spp.

When comparing samples collected from patients during or after onset of severe GvHD (n=3) to samples from patients who did not develop GvHD (n=15), 797 functional gene categories (with FDR-adjusted *p* value < 0.01 and absolute log_2 fold change ≥ 2) were found to be differentially abundant. Of those, only 5 functional gene categories were more abundant in samples from patients with GvHD. Similar to the changes in the functional profile throughout treatment in general, a drastic loss in the functional potential was observed in samples from patients during or after onset of severe GvHD.

To get a better overview of what pathways those functions belong to, the functional gene categories were annotated for corresponding KEGG orthologous groups (KOs). Applying

the same analysis on these KOs as previously, 646 were found to be differentially abundant (FDR-adjusted *p* value < 0.05 and absolute $\log_2$ fold change ≥ 1), with only 41 being higher abundant in samples from patients with GvHD and 605 lower abundant. The KOs with lower abundance are illustrated in Figure 3.2.19. The corresponding pathways which were enriched (meaning that a significantly higher number of KOs belonging to this pathway were observed in the differentially abundant set than expected by chance, adjusted *p* value < 0.05, hypergeometric test) are indicated with colored lines.

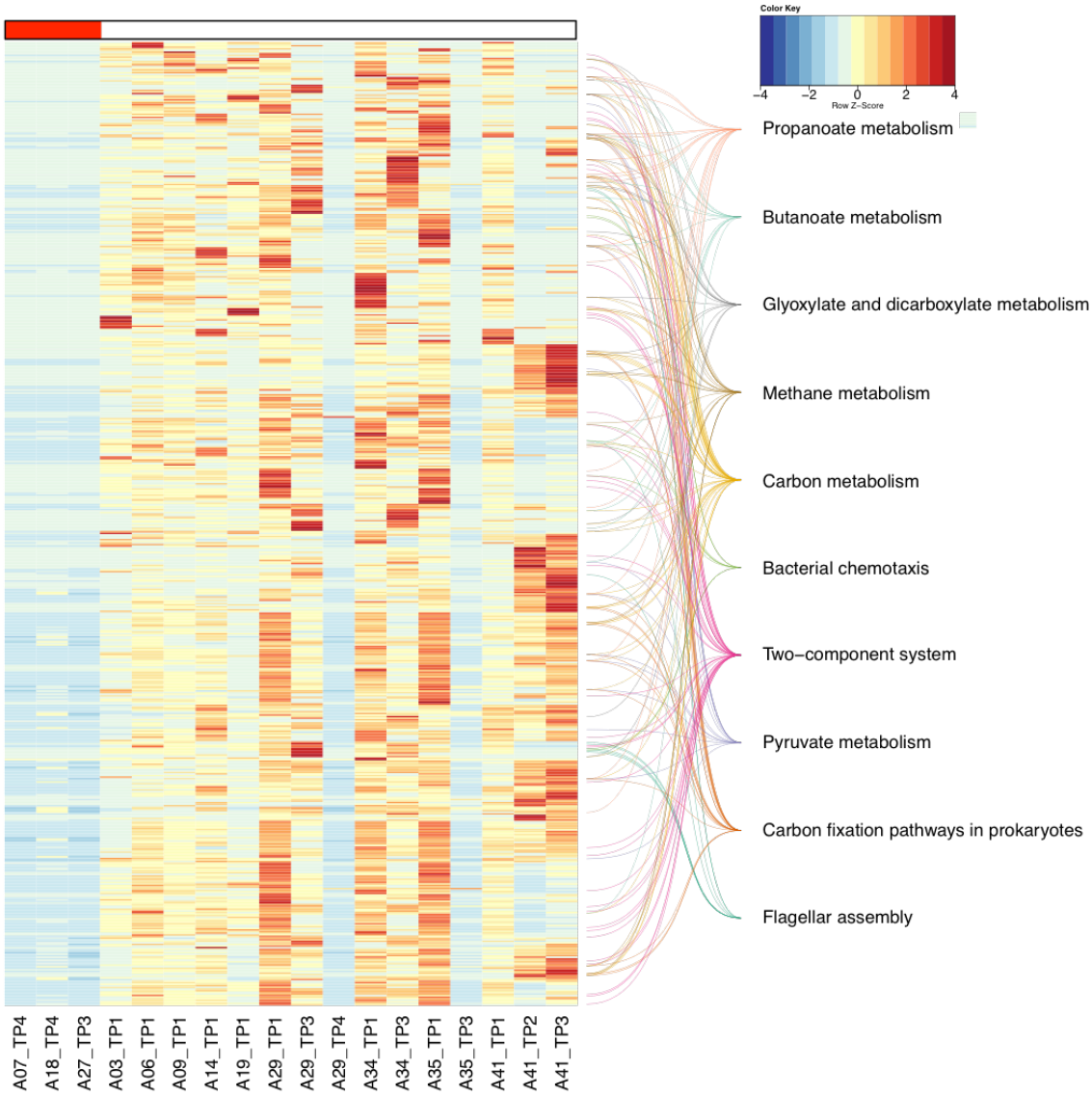


Figure 3.2.19: Heatmap of differentially abundant KOs between samples from patients with severe active GvHD and samples from patients who never developed GvHD (with FDR-adjusted *p* value < 0.05 and negative $\log_2$ fold change ≥ 1). Heatmap of normalized gene abundances is scaled as indicated in the color key. Pathway correspondence of the KOs is indicated next to the heatmap. At the top of the heatmap, samples from patients with severe GvHD are marked in red, samples from patients without GvHD are marked in white. Samples from patients without GvHD are ordered according to the patient ID.

Figure 3.2.20A represents the ratio of KOs with significantly decreased abundance in samples from patients with GvHD compared to the total number of KOs per pathway. Figure 3.2.20B represents the number of significantly differentially abundant KOs in samples from patients with GvHD compared to the total number of KOs detected in the 18 samples included in this analysis. A higher ratio of differentially abundant KOs belonging to these mentioned pathways (Figure 3.2.20A and Figure 3.2.19) than the overall ratio of differentially abundant KOs (Figure 3.2.20B), is observed.

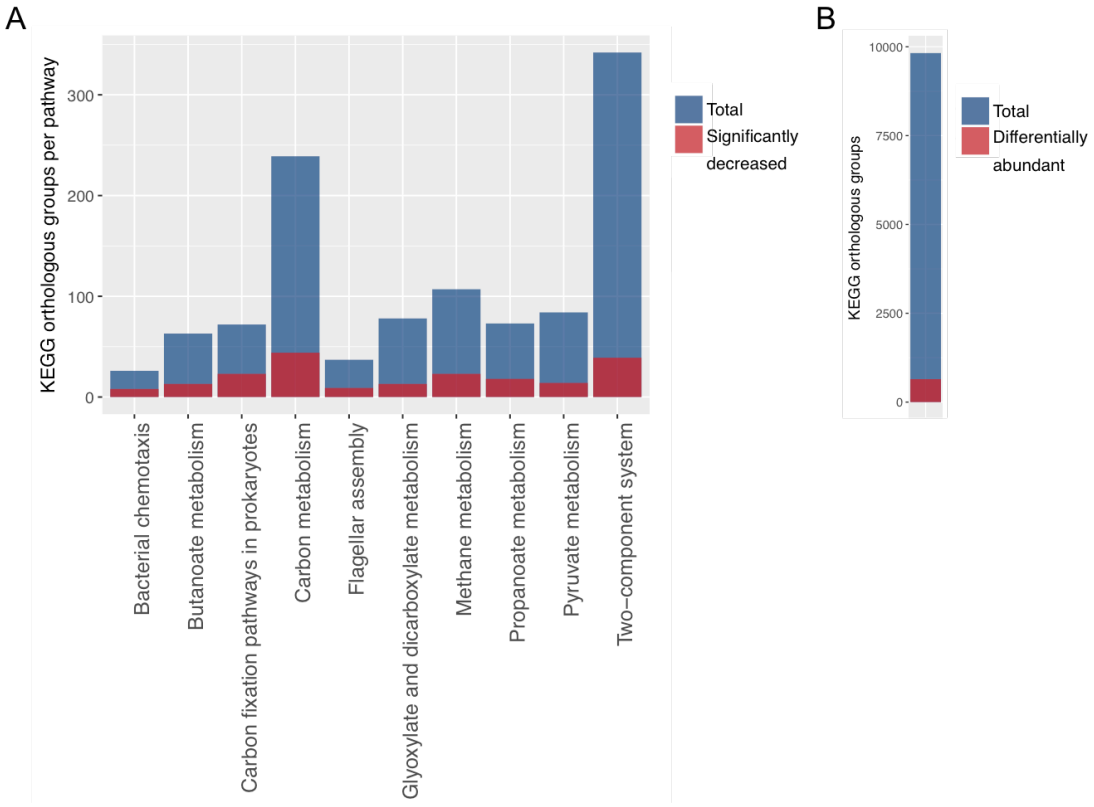


Figure 3.2.20: Barplots indicating number of differentially abundant KOs in samples from patients with GvHD, compared to the total number of KOs per pathway. (A) Ratios of KOs with decreased abundance are represented in red, total numbers of KOs per pathway are represented in blue. Corrresponding pathways are indicated below each bar. (B) Ratio of differentially abundant KOs (red) compared to the total number of KOs (blue).

These pathways included some that are generally needed by bacterial cells, such as two-component systems or the carbon metabolism. However, also genes belonging to the propanoate and the butanoate metabolism pathway were overrepresented within the functions with lower abundances in samples from patients with GvHD. This might indicate lower production of butyrate and propionate, two SCFAs, which are associated with important health-related functions, for example promoting absorption of electrolytes and

fluid (Scheppach, 1994). As mentioned above, a higher relative abundance of the genus *Lactobacillus* was detected in samples from patients with GvHD, which mainly produce lactate. This is usually converted by other bacteria into acetate, propionate and butyrate. If these bacteria and the corresponding pathways are missing or down-regulated however, lactate might accumulate in the GIT. In feces from individuals with ulcerative colitis for example, high concentrations of lactate have been detected and linked to diarrhea (Hashizume, Tsukahara, Yamada, Koyama, & Ushida, 2003; Vernia et al., 1988). Thus, changes within the GIT microbiome community and functional capacity during GvHD could further worsen the status of the patient.

3.3 Case study: Patient A07 – severe GvHD and dysbiosis

This section focuses on a specific patient who underwent an allo-HSCT, developed severe GvHD in parallel with a marked GIT microbial dysbiosis. Parts of the results and discussion section focussing on this are taken and slightly modified from a manuscript that has been submitted to Translational Research. The respective manuscript is attached in the appendix:

Appendix A.1: Anne Kaysen, Anna Heintz-Buschart, Emilie E. L. Muller, Shaman Narayanasamy, Linda Wampach, Cédric C. Laczny, Norbert Graf, Arne Simon, Katharina Franke, Jörg Bittenbring, Paul Wilmes, Jochen G. Schneider. (2017) Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic stem cell transplantation. *Translational Research.* (in revision).

3.3.1 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 (Figure 3.3.1A and Figure 3.3.1B) 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 (Schmid, Schleuning, Ledderose, Tischer, & Kolb, 2005), a modified sequential conditioning regimen 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 (Figure 3.3.1C). 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 (Figure 3.3.1C).

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 (Figure 3.3.1D).

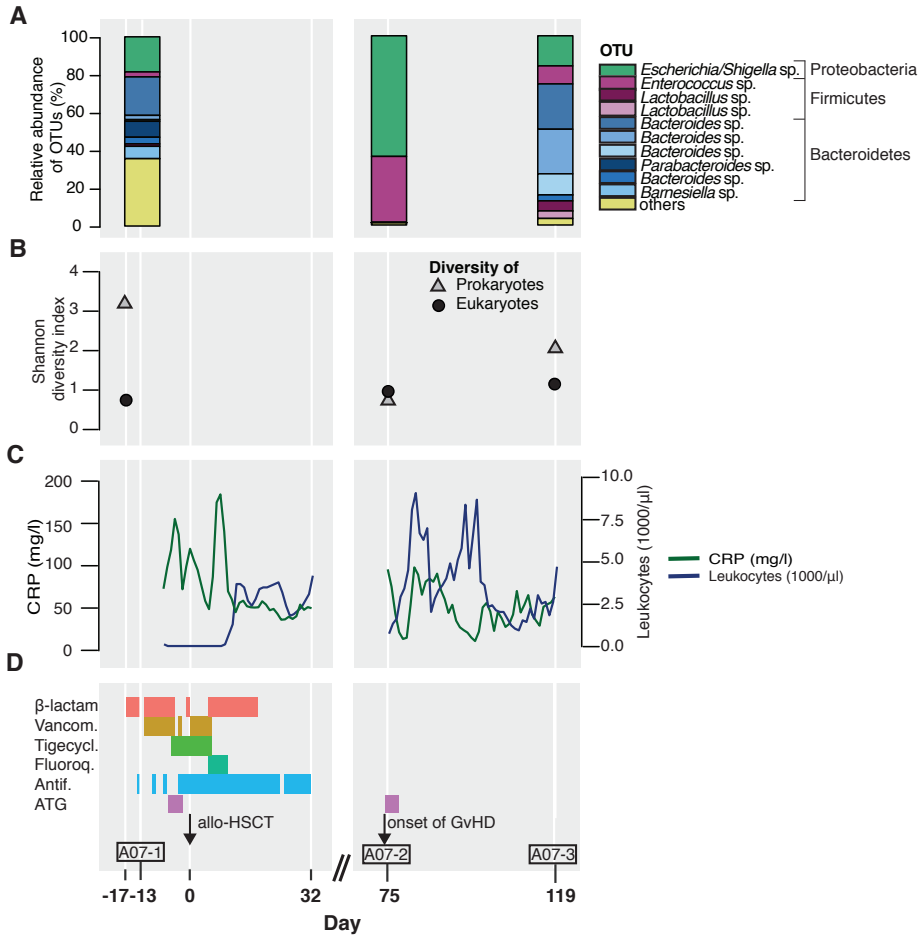


Figure 3.3.1: Variation of the microbial community structure over the course of the allo-HSCT treatment in patient A07. (A) Relative proportions of the 10 most abundant operational taxonomic units (OTUs) based on 16S rRNA gene sequencing. The remaining OTUs are summarized as 'others'. Similar shades 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.

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 MDR *Escherichia coli* was isolated both from the abscess and from a blood culture, and was analyzed further. After surgery the patient's health status improved, she 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.

3.3.2 Patient A07 – changes in the microbial community structure during the treatment

Fecal samples were taken, as indicated in Figure 3.1.1D, 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 (Figure 3.3.1B). 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. (Figure 3.3.1A). However, many of the less abundant OTUs, belonging to 25 different genera, disappeared entirely, including for example *Anaerostipes* and *Clostridium* cluster IV. 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 (Figure 3.3.1A). 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 in the second sample 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 before, the eukaryotic microbial community did not exhibit pronounced changes over time (Figure 3.2.10 and Figure 3.3.1B). Taken together, a drastic decrease in prokaryotic diversity, with relative expansion of few bacteria, including potential pathogens, was observed.

Only one study so far has followed the GIT microbiome trajectories up to three months after allo-HSCT (Biagi et al., 2015). Contrary to this study, which observed that the richness and metabolic capacity of the microbial community recovered after two months (Biagi et al., 2015), 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) (Abreu & Peek, 2014; Jiang et al., 2015; Perez-Chanona & Jobin, 2014). The genus *Blautia* for instance, has been linked to reduced aGvHD-associated death and improved overall survival (Jenq et al., 2015) and the genus *Barnesiella* with resistance to intestinal domination with vancomycin-resistant enterococci in allo-HSCT patients (Ubeda et al., 2013). 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 (Holler et al., 2014; Y. Taur et al., 2012). 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 (Shono et al., 2016), 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 (Brandl et al., 2008; Ubeda & Pamer, 2012). 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 (Figure 3.3.1A). However, they can also cause infections with associated mortality (Wexler, 2007). *Bacteroides* spp. might be able to penetrate the colonic mucus and persevere within crypt channels. These reservoirs might persist even during antibiotic treatment (S. M. Lee et al., 2013). Different species of the genus *Bacteroides* produce bacteriocins (Avelar et al., 1999; Booth, Johnson, & Wilkins, 1977; Nakano, Ignacio, Fernandes, Fukugaiti, & Avila-campos, 2006), 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).

3.3.3 Population-level structure of the pre- and post-treatment microbial community

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.

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 (Figure 3.3.2). 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 (Figure 3.3.2B). 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 (Figure 3.3.1A).

Representation of both samples within a single plot allows visual discrimination of clusters that are specific to one sample, or present in both samples (Figure 3.3.2C). In accordance with the results from 16S rRNA gene sequencing (Figure 3.3.1A), 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 (Figure 3.3.2C).

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 (Jenq et al., 2012). *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 (Borriello et al., 2003; Servin, 2004; Spinler et al., 2008). Bacteria found in probiotic formulations, especially *Lactobacillus* species have occasionally also caused bloodstream infections (Cohen et al., 2016). 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.

Given the potential role of opportunistic pathogens in aGvHD (Penack et al., 2010), 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 ARGs encoded by them and detected in each sample.

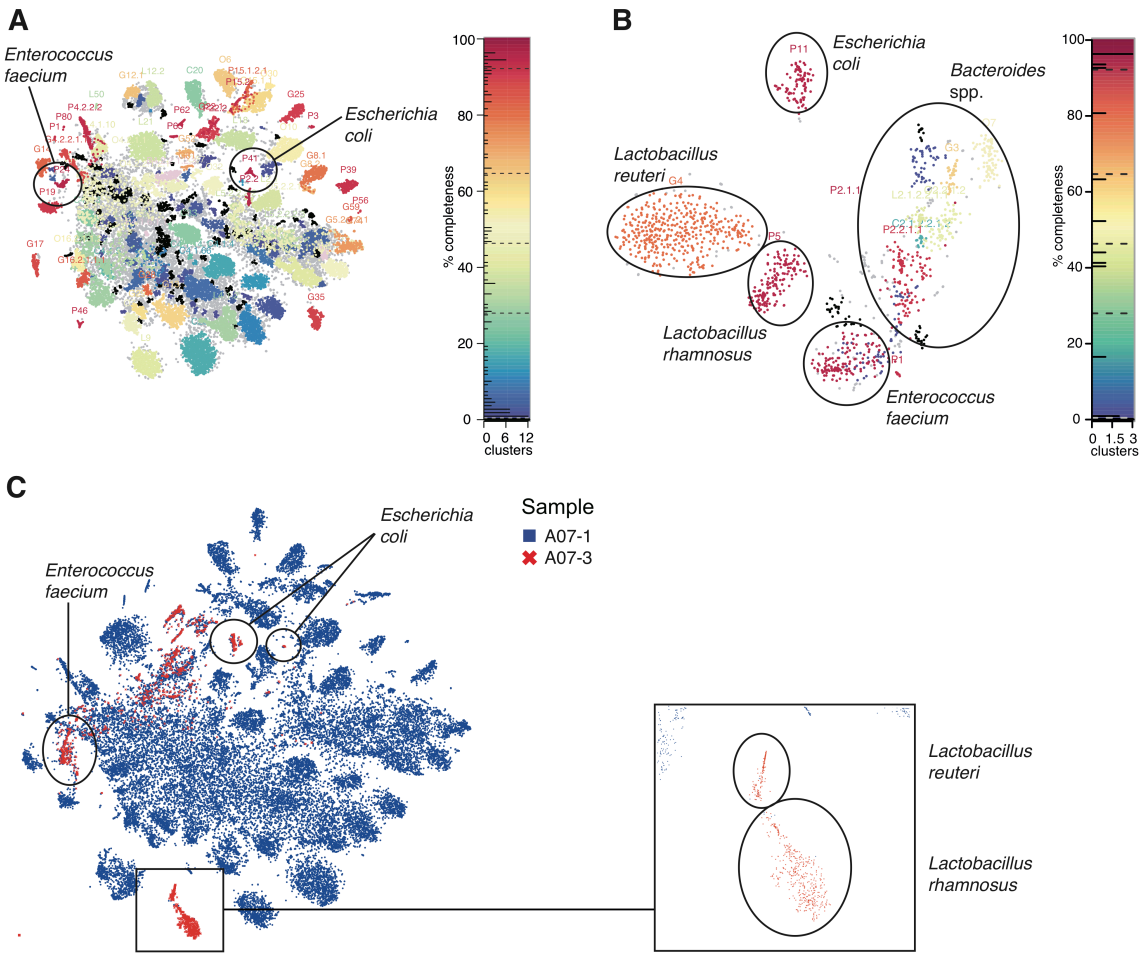


Figure 3.3.2: BH-SNE-based visualization of genomic fragment signatures of microbial communities present in samples of patient A07. Points represent contigs ≥ 1000 nt. Clusters are formed by contigs with similar genomic signatures. (A) Visualization of pre-treatment sample contigs. (B) Visualization of post-treatment sample contigs. (A and B) Points within clusters are colored according to the cluster 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.

3.3.4 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 indicated that the population size of *E. coli* was smaller after allo-HSCT (Figure 3.3.3A), while the population size of *E. faecium* remained rather constant (Figure 3.3.3C). 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 (Figure 3.3.3C), whereas only a few variants were present in *E. faecium* of both samples (Figure 3.3.3D). A similar pattern of variant distributions in both samples was observed for *E. coli* (Figure 3.3.3E and Figure 3.3.3G), while the variant pattern in *E. faecium* (Figure 3.3.3F and Figure 3.3.3H) 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.

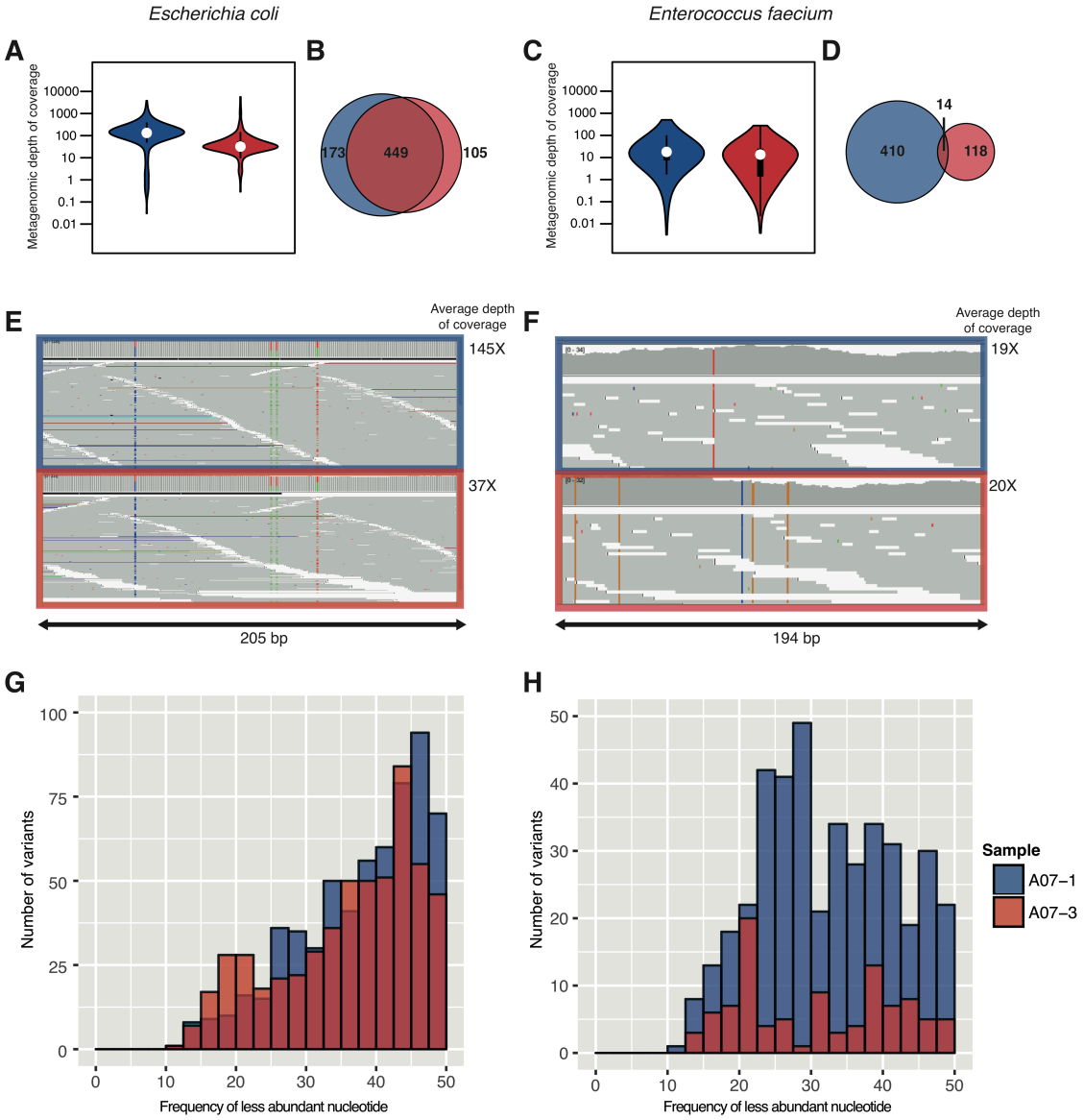


Figure 3.3.3: Number and distribution of variants in *Escherichia coli* and *Enterococcus faecium*. (A and C) Violin plots representing distribution of depth of coverage of the contigs contained in each population-level genome. (B and D) Venn diagrams indicating the number of variant positions exclusive to each sample respectively the number of variant positions found in both samples. (E and F) 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. (G and H) 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).

3.3.5 Coupled metagenomic and metatranscriptomic analysis of antibiotic resistance genes in the pre- and post-treatment samples from patient A07

The relative abundance of detected ARGs (percentage of ARGs relative to the total number of genes, Figure 3.3.4A) 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}). Moreover, the expression of ARGs was higher in both samples from patient A07 when compared to the RHMs (Figure 3.3.4B).

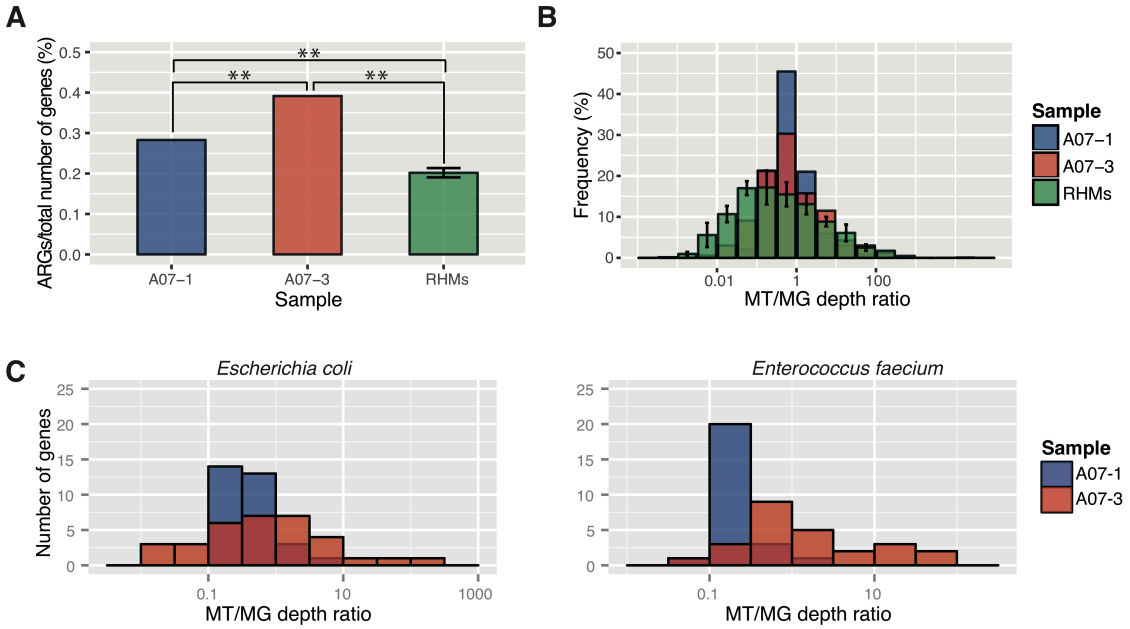


Figure 3.3.4: Expression levels and relative abundances of antibiotic resistance genes (ARGs). (A) Percentage of identified ARGs (in relation to total number of genes) in the pre-treatment (A07-1) and post-treatment (A07-3) sample and in the GIT microbiomes of four healthy untreated individuals (RHMs; ** *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. For the RHMs, the average of four datasets is represented with standard deviation as error bar.

3.3.6 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 (summaries of the ARGs identified in each population-level genome are listed in Table 3.3.1 and Table 3.3.2). 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 (Figure 3.3.4C). Although patient A07 was only treated with antibiotics until day 18 (Figure 3.3.1D), expression of the ARGs was in general higher in the post-treatment sample, both in the whole sample (Figure 3.3.4B), as well as in the specific populations (Figure 3.3.4C).

Table 3.3.1: ARGs 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 3.3.2: ARGs 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	Erm23SRibosomalRNAMethyltransferase	rRNA 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

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.

3.3.7 Genomic characterization of a blood culture *E. 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* (Figure 3.3.5). 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*.

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.

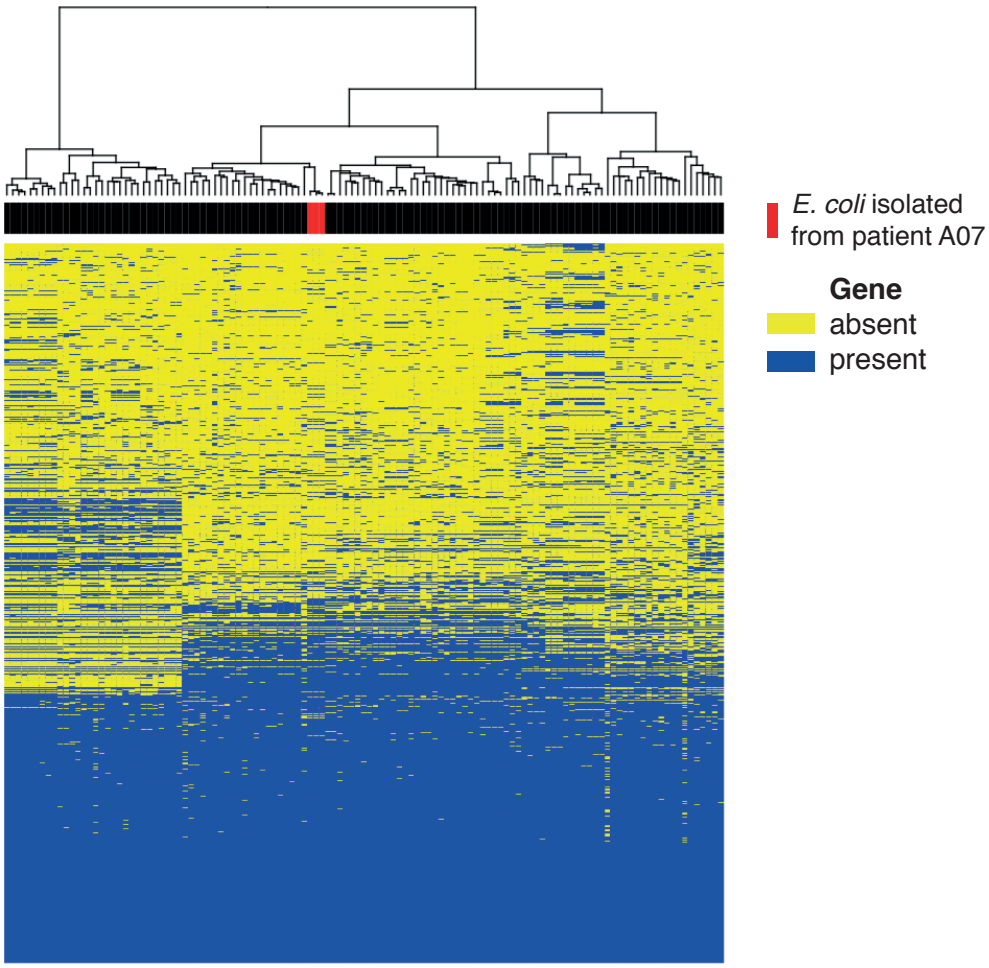


Figure 3.3.5: Gene set profiles of the 118 reference strains and 3 *E. coli* isolated from patient A07 (marked in red). Each row represents a gene (blue: present, yellow: absent), each column represents a strain.

4 Conclusion and perspectives

4.1 Is there a general response of the GIT microbiome to anticancer treatment and is the GIT microbiome implicated in development of treatment side effects?

Different studies analyzing the GIT microbiome using 16S rRNA gene amplicon sequencing, in relation to changes in response to different anticancer treatments have been published (Biagi et al., 2015; Montassier et al., 2014). However, these studies have included different cohorts with several underlying diseases, various treatments and also varying analysis methods. Therefore, it is not surprising that the results are not necessarily in complete agreement. A systematic review (Touchefeu et al., 2014) lists a decrease in *Bifidobacterium* spp., *Clostridium* cluster XIVa, *Faecalibacterium prausnitzii* and an increase in Enterobacteriaceae and *Bacteroides* spp. as the most commonly observed changes after different anticancer treatments, including different combinations of cytotoxic and radiation therapies. In the current study, the changes within the GIT microbiome of the pediatric patients did not agree with these findings, except for a slight decrease in one member of the *Clostridium* cluster XIVa (Table 3.1.2).

In samples from patients who developed mucositis, a decrease in one OTU classified in the *Clostridium* cluster XIVa and a decrease in *Bacteroides* sp. was observed, among other changes (Table 3.1.3). This was observed over all TPs including the TP before the treatment, thus, these differences were not likely due to the treatment. However, these GIT microbiome profiles might have contributed to mucositis development. In samples from patients who had severe mucositis*, Akkermansia muciniphila* was decreased. This bacterium was shown to strengthen the epithelial barrier function and is assumed to have anti-inflammatory properties (Derrien et al., 2016; Reunanen et al., 2015; Schneeberger et al., 2015). Thus, loss of this bacterium could lead to increased translocation of microbial molecules and favor an inflammatory response, thereby contributing to mucositis development. On the other hand, the decrease could represent a consequence of the damaged epithelium and a degradation of the mucus layer, which is its most important nutrient source (Yamamoto et al., 2013). In this case, the observed decrease in relative abundance of this bacterium would only represent a consequence of mucositis.

To this date, this is the first study of MG and MT datasets of fecal samples from cancer patients. On the MT level, differentially abundant functional gene categories in samples from patients with mucositis were detected, which included functions associated with

bacteriophages, suggesting a potential role of these entities in the development of mucositis. Further investigations are needed in order to validate this finding and elucidate a mechanism.

In the adult patient cohort, a decrease in many bacterial genera that are generally associated with health-promoting properties, such as *Bifidobacterium* spp., *Faecalibacterium* spp., *Dorea* spp., *Roseburia* spp., *Blautia* spp. and *Barnesiella* spp. was observed, along with many other changes (Table 3.2.1), which is partly in agreement with results commonly found in previous studies (Touchefeu et al., 2014).

Similar to the observations in the pediatric cohort, the relative abundance of *Akkermansia muciniphila* was decreased in samples from patients who later developed severe GvHD. As mentioned before, lower levels of this bacterium might have contributed inflammation or to a weaker intestinal epithelial barrier and thus enabled translocation of bacterial components and activation of the pro-inflammatory cascade. In other studies, onset of GvHD was accompanied by an increase in the abundance of *Enterococcus* spp. (Biagi et al., 2015; Holler et al., 2014) and a reduction in *Faecalibacterium* spp. and *Ruminococcus* spp. (Biagi et al., 2015; Jenq et al., 2012). In this study, trends in the abundance of these taxa were not observed with regards to GvHD but more generally over time. In the current study, *Ruminococcus* spp. was shown to be characteristic of samples collected at TP1 (Figure 3.2.12), while *Enterococcus* spp. was more abundant in later TPs.

One general response to treatment that was observed in both cohorts was a decrease in alpha-diversity of the prokaryotic GIT community (measured by the Shannon diversity index and the Chao1 richness estimator in Figure 3.1.3 and Figure 3.2.3). This decrease was substantially more pronounced in the adult patients undergoing allo-HSCT, which could be due to the intensive conditioning treatment, but also due to the antibiotic treatment. Furthermore, a general decrease in the functional potential was observed including a decrase in relative abundance or loss of various metabolic and cellular functions, which is probably linked to the decrease in microbial diversity and associated loss in bacterial populations (Figure 3.2.17 and Figure 3.2.18). This is the first study to show this decrease in functional capacity in these patients' GIT microbiomes. Overall, a detailed look at changes in samples from patients who have been sampled more frequently might help to discern and detect mechanisms that are involved in development of treatment side effects. Therefore, it might be useful to add this layer of information in future studies.

A decrease in alpha-diversity after allo-HSCT (Figure 3.2.3) was also observed in other studies and lower diversity at engraftment was linked to higher transplant-related mortality

(Y Taur et al., 2014). This was also observed in this study (Figure 3.2.4). This raises questions and concerns regarding intensive antibiotic treatment after allo-HSCT, which of course has a drastic effect on microbial diversity. Based on the observation, one strategy to avoid a treatment-induced intestinal domination by pathogens could consist in the tailored administration of several, rather than single probiotic strains, composed based on the individual GIT microbiome changes during therapy. A different approach to restore specific beneficial microbes could consist in fecal microbiome transplantation, either as 'autologous' (transplanting the pre-transplant microbiome) or 'allogeneic' graft (from the donor of the stem cells or a relative of the patient). A randomized trial of autologous fecal microbiome transplantation (FMT) for prevention of *Clostridium difficile* infection after allo-HSCT is currently recruiting participants (https://clinicaltrials.gov/ct2/show/NCT02269150). This auto-FMT could possibly also prevent other treatment-related side effects. FMT with the patient's spouse or relative as donor was successfully performed on four allo-HSCT patients with steroid-resistant aGvHD of the GIT. In three cases, a complete response was observed along with one partial response (Kakihana et al., 2016). Preservation or re-establishment of a diverse microbiome able to inhibit expansion of potential pathogens might be a new approach to avoid treatment related side effects. Bacteria with anti-inflammatory properties, such as several members of the class Clostridia, might help in regulating the immune response by induction of T_{reg} and IL-10 and lower the incidence or severity of GvHD and mucositis. Specifically *Blautia* spp. (a member of the class Clostridia with anti-inflammatory properties) was linked to lower GvHD-related mortality and improved overall survival after allo-HSCT (Jenq et al., 2015).

In both cohorts, the microeukaryotic and the archaeal community did not display drastic changes in response to the treatment (section 3.1.3, section 3.2.2 and section 3.2.3). No correlation between bacterial diversity and microeukaryotic diversity was found which indicates that the individual microbial communities were differently affected by the treatment.

In both cohorts, a high intra-individual dissimilarity but also inter-individual dissimilarity between different GIT bacterial profiles were observed (Figure 3.1.5 and Figure 3.2.6). Here, individual-specific changes rather than general trends within the GIT microbiome following treatment were observed. Additionally, as there is a high complexity within this ecosystem including numerous interactions between the GIT microbiome, the host immune system, treatment and dietary changes, it is not possible to relate development of mucositis or GvHD to one factor (the microbiome) alone. In this study, it was not possible

to detect changes in the GIT microbiome leading to development or aggravation of the side effects.

4.2 How important are SCFAs?

Numerous studies have focused on and discovered new effects of short-chain fatty acids (SCFAs), especially of butyrate, on the human body and implications in human health. For example, butyrate is important for maintenance of the intestinal barrier and it is a histone deacetylase inhibitor (Davie, 2003). Thereby, butyrate acts as an anti-inflammatory agent, inhibiting NF-κB activation (Inan et al., 2000). In patients with inflammatory bowel diseases, similar GIT microbial profiles were found as in patients with GvHD; marked by reduced richness, lower abundance of butyrate producers (such as members of the *Clostridium* clusters IV and XIVa) (Kolho et al., 2015; Zama et al., 2016). In a recent murine study, butyrate or a bacterial community including butyrate-producing Clostridia, were shown to mitigate GvHD (Mathewson et al., 2016), suggesting an important role of butyrate and butyrate-producers in relation to GvHD.

In this project, in general, a decrease in the relative abundance of many SCFA producers was observed, in response to the allo-HSCT (Table 3.2.1). However, this was not observed when grouping according to development of severe GvHD (section 3.2.8). On this level, no link between SCFA-producers and GvHD was found. In both cohorts within the metagenome and metatranscriptome, no statistically significant differences in the copy number of the genes coding for the three enzymes catalyzing the final step in butyrate production were detected when grouping according to the occurrence of severe mucositis or GvHD (section 3.1.7.2). In samples from patients with active GvHD, a statistically significant decrease in the relative abundances of some genes belonging to different SCFA biosynthetic pathways was observed. These pathways included the propionate and butyrate biosynthesis pathway (Figure 3.2.19). Still, with the current study results, no direct link between the SCFA-producers in the GIT microbiome and occurrence of treatment side effects can be made. As it seems, there are more factors playing a role in development of these side effects and the interconnection between the microbiome, the health status of the patient and especially the immune system is more complicated.

4.3 Could shotgun sequencing of the GIT microbiome revolutionize personalized medicine?

Metagenomic sequencing allows description of the composition of the GIT microbiome including identification of potential pathogens. One advantage in comparison to 16S

and/or 18S rRNA gene amplicon sequencing is that with MG sequencing, genomic fragments from all domains of life and even of DNA viruses are contained within one dataset. 16S rRNA gene sequencing, in contrast, does not allow to draw conclusions on the actual composition of the GIT community as eukaryotes and viruses are not resolved. In addition, the total community size or the proportion of human DNA content are not measured. All of this information is retained in MG sequencing datasets. In this project, for example the presence of high amounts of human DNA in several samples was noticed (sample A27_3, Figure 3.2.13 and section 3.2.5), which in turn indicates that the overall microbial content was low.

MG sequencing allows identification of potential pathogens without culturing. Culturing can lead to false negatives, as some bacteria are difficult to culture (Vartoukian et al., 2010). Also, culturing does not reflect relative abundance. In addition, MG sequencing allows identification of all the ARGs comprised within the microbial community, i.e. the resistome. As observed in this study, there was an important increase in the ratio of ARGs within the community throughout treatment (Figure 3.1.11 and Figure 3.2.15), which was especially pronounced in the adult hematology patients and this can probably be linked to the intensive antibiotic treatment. A specifically tailored treatment could then be chosen, in response to the detected pathogens and corresponding ARGs. Although this may sound good in theory, we are still far away from being able to apply this workflow in the clinical setting. Especially immunocompromised patients are in a critical state, where timely actions and direct treatments are a necessity. At the onset of fever, indicating an infection, immediate action is crucial. At this point, it is not possible to wait for results of processed sequencing data, identifying the infecting pathogen and corresponding ARGs. However, if the first choice of treatment (empiric treatment) fails, these results may help to find a suitable medication (informed treatment). Furthermore, if next-generation sequencing became cheaper and faster, the changes in the microbiome of patients could be monitored and dangerous bacteria could be recognized and kept at bay before a fever or infection were to develop. Implementation of screening of ARGs and potential pathogens in baseline samples (at admission) as well as at sign of infection could possibly help to find the appropriate treatment and infection control measures, thereby improve patient treatment. Such a workflow could potentially result in faster and more reliable diagnostics (Figure 4.3.1).

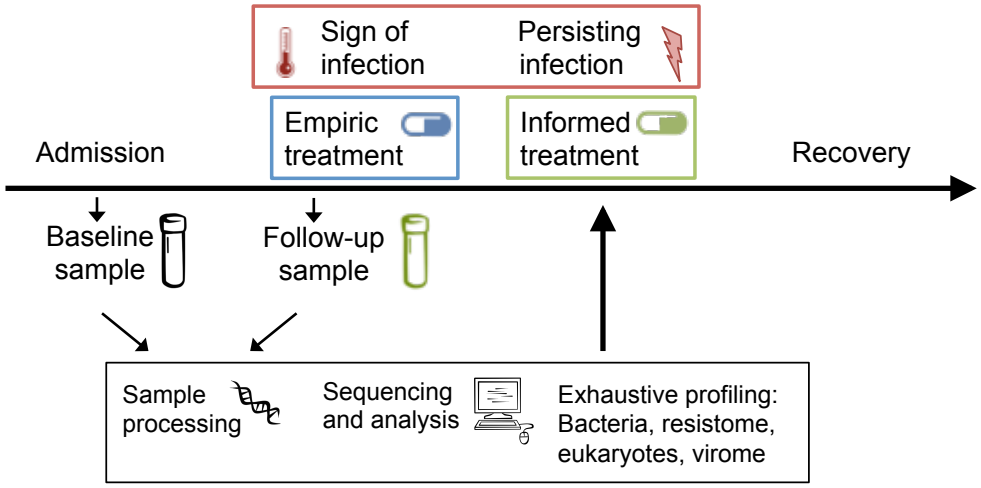


Figure 4.3.1: Workflow suggesting possible usage of shotgun sequencing in personalized medicine to compile individually tailored treatments.

Furthermore, MG sequencing allows detection and identification of viruses before they pose a threat to the patient. Often, symptoms of viral infections can vary and be quite unspecific, making it difficult to discern the virus causing the symptoms. For example, the symptoms of cytomegalovirus colitis, diarrhea and abdominal pain, are very similar of those of aGvHD. Therefore, a biopsy is advised at suspicion of infection (Jacobsohn & Vogelsang, 2007). Another method to detect viral infection is an antibody assay, however, in immunocompromised patients this test might result in a false negative result due to a low antibody response (Woods, 2013). Similarly, culturing of viruses (in cell culture) can be difficult and lead to false negatives if only a low number of viruses is present, which would then be growing slowly and take up to weeks to eventuate in a positive result. Shotgun sequencing allows early detection of viruses in a non-invasive and exhaustive way (section 3.1.6 and section 3.2.4).

As seen in this study and mentioned previously, shotgun sequencing allows identification of ARGs (section 3.1.5 and section 3.2.6). A suggested response following this would be a specifically tailored antibiotic treatment. The majority of the currently used antibiotics however, are not specific to particular pathogens. Modes of action for example are the inhibition of cell wall synthesis (e.g. meropenem) or the inhibition of nucleic acid synthesis (e.g. fluoroquinolones). Generally, the fact that antibiotics have broad-spectrum activity is seen as an advantage. However, treatment with prophylactic broad-spectrum antibiotics will not only kill pathogenic and opportunistic bacteria, but also important beneficial and commensal bacteria. Resulting antibiotic pressure can lead to microbial dysbiosis and emergence of MDR bacteria, a threat that is well known. Therefore, development of other narrow-spectrum treatments is indispensable. Alternatives to currently used traditional

antibiotics have been presented in section 1.3. Another option for changing the composition of the microbiome may be the ingestion of specific beneficial microbes, i.e. probiotics. Generally, probiotics are defined as living microbes, which confer benefits to the host. Mainly lactobacilli and bifidobacteria are used as probiotics. In this study, the relative abundance of *Lactobacillus* spp. was in general higher in patients with severe aGvHD than in those without GvHD (section 3.2.8), as for example in A27_3 (Figure 3.2.13) and A07_3 (Figure 3.3.1). The results in this project suggest, that the high relative abundance of this bacterial genus was a consequence of GvHD, rather than a cause. Bloodstream infection with *Lactobacillus* spp. has been observed in patients, including patients after allo-HSCT (Cohen et al., 2016; Salminen et al., 2004). Although their administration is generally considered safe (Ladas et al., 2016), these organisms and their interaction in particular with the immunocompromised host should be further investigated. Similarly, FMT is recently being applied in the treatment of patients with multiple recurrent *Clostridium difficile* infections (CDI), where cure rates of 90 % have been reported (Brandt et al., 2012). However, FMT has yet to be further investigated and improved before becoming a standard therapy option. In order to apply this method to patients undergoing cancer treatment, mechanisms of action need to be better understood. This is especially important in immunocompromised patients such as patients undergoing an allo-HSCT. After allo-HSCT, immunologic recovery occurs gradually, over time. On one hand, it might be beneficial to have an 'intact' GIT microbiome from a healthy donor, which could positively influence development of the hematopoietic cells and immune recovery. On the other hand, it is not known how the patient's newly acquired immune system might react to a completely foreign microbiome, even though the donor might be in good health.
In short, if cost and turnaround time of shotgun sequencing could be addressed, it could potentially allow exhaustive profiling of the GIT microbiome of patients including members of all domains of life, as well as viruses and ARGs and thereby enable individually tailored treatments.

4.4 General challenges for GIT microbiome studies in the clinical setting

Numerous studies applying high-throughput sequencing such as rRNA gene amplicon sequencing on fecal samples have characterized the human GIT microbiome in detail and suggested that the GIT microbiome performs many important functions for the host, and that alterations and disorders in the microbial community can be associated to different diseases. Eventually, this deep insight into the human-microbiome interaction might enable the application of high-throughput sequencing in the clinical setting. However, a

standardized procedure and methodology should be put in place before it can be used in personalized medicine. Several recent studies have compared how different steps in microbiome studies influence the results (Blekhman et al., 2016; Gerasimidis et al., 2016; Walker et al., 2015). These factors include: storage (flash-freezing samples versus keeping them at room temperate, addition of preservation buffers, storage at -20 °C or -80°C), the period of storage, the extraction method, sequencing method (including the choice of primers for example for amplicon sequencing) and the pipeline used for processing of the reads (for example mothur (Schloss et al., 2009), QIIME (Caporaso et al., 2010) or LotuS (Hildebrand et al., 2014)). All of those steps have been shown to introduce bias and influence the final output. Within individual studies therefore, one is usually mindful to use the same strategy for all of the samples, which allows comparison of the results within the study. However, it is a general aim in research to share the gained knowledge and build upon it. Usage of different strategies makes it difficult or impossible to truly compare results from different studies. However, it is difficult to establish a general workflow that accounts for everybody's needs. While the majority of the studies only focus on DNA, it would be crucial to add and integrate different omic data (such as RNA, proteins or metabolites). In this case, a framework that allows extraction of all of those biomolecules from one undivided subsample would be ideal (Roume, Muller, et al., 2013; Roume, Heintz-Buschart, et al., 2013).

To reduce some of the introduced bias, the same storage method, and a generally applied lysis protocol, allowing disruption of the cell walls of the different organisms present in one sample could be set in place. After lysis, subsequent extraction of DNA only or of different biomolecules (DNA, RNA, proteins and metabolites), according to the needs of the individual project would be possible. Additionally, if raw sequencing data of individual projects were publicly available, they could be processed using different pipelines, making the results from different studies more comparable.

In the GIT microbiome field, fecal samples are generally used as proxy for the content of the large intestine, as they are easier to collect than for example biopsies. However, they might not accurately reflect the microbial community in the colon as for example organisms living in the mucus layer might be underrepresented in the sample. Additionally, transit time and stool consistency influence microbial richness (Vandeputte et al., 2016). Thus, it has to be kept in mind that fecal samples might not represent the exact reality. A solution to get a more representative overview of the actual microbial community within the colon is the IntelliCap (MediMetrics), an electronic capsule, which allows

sampling within the GIT in a non-invasive way. The included pH temperature sensors enable position tracking and thereby sampling from different locations within the GIT.
Also, different subsamples from one fecal sample result in high variability of the different microbial taxa (Gorzelak et al., 2015; D. H. Huson et al., 2016). In order to circumvent some of these difficulties in this study, we applied an extraction method that allows extraction of DNA and RNA from one undivided sample was applied (Roume, Muller, et al., 2013; Roume, Heintz-Buschart, et al., 2013). Also, the same sequencing methodologies were applied for individual samples and each dataset type was processed using the same pipeline, in order to allow comparison of the different results. Different omic layers (rRNA gene amplicon sequencing, metagenomic and metatranscriptomic datatsets) were produced and analyzed, allowing higher confidence. In addition, metagenomic sequencing is less bias prone than rRNA gene amplicon sequencing, as this method does not include PCR amplification (Shakya et al., 2013).

4.5 Challenges in this study

In this project, two different departments were included to be able to include a higher number of patients, resulting in two different patient cohorts, each cohort displaying high levels of heterogeneity in itself. The patients in these cohorts had different underlying diseases and received different antibiotic and anticancer treatments. Thus, many confounding factors could have influenced the GIT microbiome, which makes it difficult to link the observed changes to one factor alone. For example, the GIT microbiome has been shown to evolve throughout lifetime and to display different patterns according to the age of the host. Even within the pediatric department, where the age of the patients ranged only from 3 to 19 years, this effect could be observed. Younger children (3-4 years) displayed a microbial community with a lower diversity and richness (Figure 3.1.4 (Ringel-Kulka et al., 2013)). It was noticed that the lower richness was not related to the treatment group, but mainly to the age of the patient. The treatment however did also have an influence on richness.
It has been observed that the GIT microbiome of healthy individuals is strikingly stable over time and also capable of reverting roughly to its initial state, even after external influences have induced changes in the complex ecosystem (David et al., 2014; Heintz-Buschart et al., 2016). However, in this study, a high inter-individual but also a high intra-individual variation and dissimilarity between different GIT microbial profiles were observed (Figure 3.1.5 and Figure 3.2.6). Currently, large-scale studies including a higher number of patients usually conduct analyses based solely on the mean or median over

the whole cohort or individual study groups. Considering the huge variation that was observed between different patients and the very individual-specific patterns that were observed in this study (for example illustrated in Figure 3.1.10), averages were often not representative of any member of the cohort. Therefore, detailed personalized profiles were presented and discussed. The compelling individual-specific results suggest, that one way to resolve the dynamics within the microbiome following treatments should include studies focussing on patient dynamics individually. Similary, in the future, treatments should be more individually tailored (as suggested in Figure 4.3.1).

In this study, high fluctuations in different blood counts, as well as in the GIT, especially in patients undergoing an allo-HSCT, were observed (for example in patient A07, Figure 3.3.1). In order to relate changes in the composition of the GIT microbiome to the status of the immune system, a more frequent sampling scheme (ideally daily), would have to be adopted. For future studies, it would be interesting to include only a low number of patients, but to sample them frequently, over a longer time period. This has been done in other studies, focusing for example on the GIT microbiome of healthy individuals or of newborn babies (David et al., 2013, 2014; Palmer, Bik, DiGiulio, Relman, & Brown, 2007).

4.6 Perspectives

As observed in this study, many factors influence the complex microbial ecosystem in the GIT. Already in a simple bacterial population with a single taxon, there is a complicated interconnection and signalling network between the microorganisms resulting in responses of the population to simple parameters such as nutrient availability, temperature and pH. In general in the clinical setting, as it was the case in this project, a complex ecosystem of microorganisms, interacting with each other, but being also influenced by factors such as dietary changes of the host, drastic changes of the status of the immune system and treatment with chemotherapeutics and antibiotics, is studied. In this system, new layers and levels of interconnections are added, resulting in a highly intricate network.

To get an overview of the mechanisms within the microbiome, in this project, in addition to taxonomic profiling, assessment of the functional potential within the microbial community was included, adding a whole new level of information. However, the current study design does not allow identification of causality. It was not possible to identify whether changes in the GIT microbiome composition led to the development or aggravation of adverse effects, a worse outcome, or changes in the status of the immune system, or, if the latter resulted in further changes in the microbiome. In order to find out specific causalities and

mechanisms, a much more simplified and controlled system might be more suitable. Mouse studies allow a more controlled environment with less confounding factors, frequent sampling and also the analysis of the GIT content instead of the stool samples. However, as for example mice are different from humans in anatomy, genetics and physiology, murine studies can never mimic human systems (Nguyen TL, Vieira-Silva S, Liston A, 2015). Another possibility would be the usage of a microfluidics-based co-culture device such as HuMiX (Shah et al., 2016), which allows growth of bacteria, intestinal epithelial cells and human immune cells in different chambers, separated by membranes, which allow molecular cross-talk. Thus, interactions between those layers can be examined in a controlled environment with specific user-defined properties. Distinct bacterial communities (mimicking either a healthy GIT microbiome or a community in a dysbiotic state including potential pathogens) can be introduced or a non confluent layer of epithelial cells and membranes with larger pores, mimicking a 'leaky gut'. Observational studies in the clinical setting are still important and are needed. Validation of specific observations in different clinical studies is difficult, as heterogeneous patient cohorts in different studies might not reflect the same results. Ideally, studies in different settings, using different methodologies, would complement each other and finally allow integration of the knowledge.

In order to tackle the complexity of this highly interconnected network, in a future study, only a small number of patients, ideally with similar age, disease and treatment should be included and followed throughout their treatment with a high sampling frequency. In order to get from observation to causality, specific, interesting findings could then be further analyzed using a meta-omic approach and validated using different methods. For example, concerning the results from this study, lactobacilli present in high abundance in patients after allo-HSCT could be isolated and cultured together with human (epithelial and immune) cells, to see if they were able to elicit an immune response, using cytokine detection assays and transcriptomics. The amount of SCFA (especially butyrate) in fecal samples as well as in the serum could be determined in order to see if the observed decrease in SCFA-producing bacteria actually resulted in lower amounts of SCFA.

Despite extensive research efforts in the field of GIT microbiome, also in the context of cancer development and treatment, we are still only at the beginning of understanding this complex ecosystem and its interconnection with human health and disease. Clearly, the GIT microbiome plays an important role in human health and disease and carries the potential to be used as therapeutic agent (in the form of probiotics or FMT), emphasizing the importance and value of future research in this field. The results from this project

indicate that integrated omics, including different layers of information are needed to understand the complex mechanisms and interconnection of the GIT microbiome and its host. Ultimately, exhaustive personalized analyses could possibly enable an accordingly tailored treatment in the future.

References

Abreu, M. T., & Peek, R. M. (2014). Gastrointestinal malignancy and the microbiome. *Gastroenterology*, *146*(6), 1534–46.

Afonso, C. L., Tulman, E. R., Lu, Z., Oma, E., Kutish, G. F., & Rock, D. L. (1999). The genome of *Melanoplus sanguinipes* entomopoxvirus. *Journal of virology*, *73*(1), 533–52.

Ahmad, A. S., Ormiston-Smith, N., & Sasieni, P. D. (2015). Trends in the lifetime risk of developing cancer in Great Britain: comparison of risk for those born from 1930 to 1960. *British journal of cancer*, *112*(5), 943–7.

Al-Dasooqi, N., Sonis, S. T., Bowen, J. M., Bateman, E., Blijlevens, N., Gibson, R. J., Logan, R. M., et al. (2013). Emerging evidence on the pathobiology of mucositis. *Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer*, *21*(7), 2075–83.

Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., & Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, *31*(6), 533–8.

Alonso, C. D., Treadway, S. B., Hanna, D. B., Huff, C. A., Neofytos, D., Carroll, K. C., & Marr, K. a. (2012). Epidemiology and outcomes of *Clostridium difficile* infections in hematopoietic stem cell transplant recipients. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, *54*(8), 1053–63.

Angenent, L. T., Mau, M., George, U., Zahn, J. A., & Raskin, L. (2008). Effect of the presence of the antimicrobial tylosin in swine waste on anaerobic treatment. *Water Research*, *42*(10–11), 2377–84.

ANI Average Nucleotide Identity. Retrieved April 19, 2016, from http://enve-omics.ce.gatech.edu/ani/

Arboleya, S., Sánchez, B., Milani, C., Duranti, S., Solís, G., Fernández, N., De Los Reyes-Gavilán, C. G., et al. (2015). Intestinal microbiota development in preterm neonates and effect of perinatal antibiotics. *Journal of Pediatrics*, *166*(3), 538–44.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., et al. (2011). Enterotypes of the human gut microbiome. *Nature*, *473*(7346), 174–80.

Atarashi, K., Tanoue, T., Oshima, K., Suda, W., Nagano, Y., Nishikawa, H., Fukuda, S., et al. (2013). Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota. *Nature*, *500*(7461), 232–6.

Avelar, K. E. S., Pinto, L. J. F., Antunes, L. C. M., Lobo, L. A., Bastos, M. C. F., Domingues, R. M. C. P., & De Souza Ferreira, M. C. (1999). Production of bacteriocin by *Bacteroides fragilis* and partial characterization. *Letters in Applied Microbiology*, *29*(4), 264–8.

Ayabe, T., Ashida, T., Kohgo, Y., & Kono, T. (2004). The role of Paneth cells and their antimicrobial peptides in innate host defense. *Trends in Microbiology*, *12*(8), 394–8.

Azad, M. B., Konya, T., Maughan, H., Guttman, D. S., Field, C. J., Sears, M. R., Becker, A. B., et al. (2013). Infant gut microbiota and the hygiene hypothesis of allergic disease: impact of household pets and siblings on microbiota composition and diversity. *Allergy, asthma, and clinical immunology : official journal of the Canadian Society of Allergy and Clinical Immunology*, *9*(1), 15.

Aziz, R. K., Bartels, D., Best, A. a, DeJongh, M., Disz, T., Edwards, R. a, Formsma, K., et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC genomics*, *9*, 75.

Bäckhed, F., Ley, R. E., Sonneburg, J. L., Peterson, D. A., & Gordon, J. I. (2005). Host-bacterial mutualism in the human intestine. *Science*, *307*(5717), 1915–20.

Balique, F., Lecoq, H., Raoult, D., & Colson, P. (2015). Can plant viruses cross the kingdom border and be pathogenic to humans? *Viruses*, *7*(4), 2074–98.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. a., Dvorkin, M., Kulikov, A. S., Lesin, V. M., et al. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–77.

Bartoloni, A., Bartalesi, F., Mantella, A., Dell'Amico, E., Roselli, M., Strohmeyer, M., Barahona, H. G., et al. (2004). High prevalence of acquired antimicrobial resistance unrelated to heavy antimicrobial consumption. *Journal of Infectious Diseases*, *189*(7), 1291–4.

Baskar, R., Ann-Lee, K., Yeo, R., & Yeoh, K.-W. (2012). Cancer and Radiation Therapy: Current Advances and Future Directions. *International Journal of Medical Science*, *9*(3), 193–9.

Van Bekkum, D. W., & Knaan, S. (1977). Role of bacterial microflora in development of intestinal lesions from graft-versus-host reaction. *Journal of the National Cancer Institute*, *58*(3), 787–90.

Van Bekkum, D. W., Roodenburg, J., Heidt, P. J., & Van der Waaij, D. (1974). Mitigation of secondary disease of allogeneic mouse radiation chimeras by modification of the intestinal microflora. *Journal of the National Cancer Institute*, *52*(2), 401–4.

Bennett, S. M., Broekema, N. M., & Imperiale, M. J. (2012). BK polyomavirus: Emerging pathogen. *Microbes and Infection*, *14*(9), 672–83.

Bensinger, W. I. (2013). Allogeneic Transplantation: Peripheral Blood versus Bone Marrow. *Curr Opin Oncol*, *24*(2), 191–6.

Bhullar, K., Waglechner, N., Pawlowski, A., Koteva, K., Banks, E. D., Johnston, M. D., Barton, H. A., et al. (2012). Antibiotic resistance is prevalent in an isolated cave microbiome. *PLoS ONE*, *7*(4), 1–11.

Biagi, E., Zama, D., Nastasi, C., Consolandi, C., Fiori, J., Rampelli, S., Turroni, S., et al. (2015). Gut microbiota trajectory in pediatric patients undergoing hematopoietic SCT. *Bone marrow transplantation*, *50*(7), 992–8.

Bieber, D., Ramer, S. W., Wu, C. Y., Murray, W. J., Tobe, T., Fernandez, R., & Schoolnik, G. K. (1998). Type IV pili, transient bacterial aggregates, and virulence of enteropathogenic *Escherichia coli*. *Science*, *280*(5372), 2114–8.

Billingham, R. E. (1966). The biology of graft-versus-host reactions. *Harvey Lect*, *62*, 21–78.

Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O, Piddock, L. V. J. (2015). Molecular mechanisms of antibiotic resistance. *Nature Reviews Microbiology*, *42*(13), 42–51.

Blazar, B. R., Murphy, W. J., & Abedi, M. (2012). Advances in graft-versus-host disease biology and therapy. *Nature Reviews Immunology*, *12*(6), 443–58.

Blekhman, R., Tang, K., Archie, E., Barreiro, L., Johnson, Z., Wilson, M., Kohn, J., et al. (2016). Common methods for fecal sample storage in field studies yield consistent signatures of individual identity in microbiome sequencing data. *bioRxiv*, (August), 1–5.

Bollyky, P. L., Bice, J. B., Sweet, I. R., Falk, B. a, Gebe, J. a, Clark, A. E., Gersuk, V. H., et al. (2009). The toll-like receptor signaling molecule *Myd88* contributes to pancreatic beta-cell homeostasis in response to injury. *PloS one*, *4*(4), e5063.

Di Bonito, P., Della Libera, S., Petricca, S., Iaconelli, M., Sanguinetti, M., Graffeo, R., Accardi, L., et al. (2015). A large spectrum of alpha and beta papillomaviruses are detected in human stool samples. *Journal of General Virology*, *96*(3), 607–13.

Booth, S. J., Johnson, J. L., & Wilkins, T. D. (1977). Bacteriocin production by strains of *Bacteroides* isolated from human feces and the role of these strains in the bacterial ecology of the colon. *Antimicrobial Agents and Chemotherapy*, *11*(4), 718–24.

Boratyn, G., Camacho, C., Federhen, S., Merezhuk, Y., Madden, T., Schoch, C., & Zaretskaya, I. (2014). MOLE - BLAST a new tool to search and classify multiple

sequences (Vol. 4, p. 2004).
Borriello, S. P., Ammes, W. P., Holzapfel, W., Marteau, P., Schrezenmeir, J., Vaara, M., & Valtonen, V. (2003). Safety of probiotics that contain lactobacilli or bifidobacteria, *36*(6), 775–80.
Brandl, K., Plitas, G., Mihu, C. N., Ubeda, C., Jia, T., Schnabl, B., Dematteo, R. P., et al. (2008). Vancomycin-resistant enterococci exploit antibiotic-induced innate immune deficits. *Nature*, *455*(7214), 804–7.
Brandt, L. J., Aroniadis, O. C., Mellow, M., Kanatzar, A., Kelly, C., Park, T., Stollman, N., et al. (2012). Long-term follow-up of colonoscopic fecal microbiota transplant for recurrent *Clostridium difficile* infection. *The American Journal of Gastroenterology*, *107*(7), 1079–87.
Canani, R. B., Costanzo, M. Di, Leone, L., Pedata, M., Meli, R., & Calignano, A. (2011). Potential beneficial effects of butyrate in intestinal and extraintestinal diseases. *World Journal of Gastroenterology*, *17*(12), 1519–28.
Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., et al. (2010). QIIME allows analysis of high- throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing. *Nature Publishing Group*, *7*(5), 335–6.
Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., & Kjelleberg, S. (2007). Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology*, *73*(1), 278–88.
Castellarin, M., Warren, R. L., Freeman, J. D., Warren, L., Freeman, J. D., Dreolini, L., Castellarin, M., et al. (2011). *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma, 299–306.
Cavera, V. L., Arthur, T. D., Kashtanov, D., & Chikindas, M. L. (2015). Bacteriocins and their position in the next wave of conventional antibiotics. *International Journal of Antimicrobial Agents*, *46*(5), 494–501.
Cerf-Bensussan, N., & Gaboriau-Routhiau, V. (2010). The immune system and the gut microbiota: friends or foes? *Nature reviews. Immunology*, *10*(10), 735–44.
Chevenet, F., Brun, C., Bañuls, A.-L., Jacq, B., & Christen, R. (2006). TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC bioinformatics*, *7*, 439.
Chevenet, F., Croce, O., Hebrard, M., Christen, R., & Berry, V. (2010). ScripTree: Scripting phylogenetic graphics. *Bioinformatics*, *26*(8), 1125–6.
Chopra, T., Chandrasekar, P., Salimnia, H., Heilbrun, L. K., Smith, D., & Alangaden, G. J. (2011). Recent epidemiology of *Clostridium difficile* infection during hematopoietic stem cell transplantation. *Clin Transplant*, *25*(1), e1-8.
Clarke, S. K., Caul, E. O., & Egglestone, S. I. (1979). The human enteric coronaviruses. *Postgraduate medical journal*, *55*(640), 135–42.
Clemente, J. C., Ursell, L. K., Parfrey, L. W., & Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. *Cell*, *148*(6), 1258–70.
Cohen, S. A., Woodfield, M. C., Boyle, N., Stednick, Z., Boeckh, M., & Pergam, S. A. (2016). Incidence and outcomes of bloodstream infections among hematopoietic cell transplant recipients from species commonly reported to be in over-the-counter probiotic formulations. *Transplant infectious disease : an official journal of the Transplantation Society*, (1), 699–705.
Colson, P., Richet, H., Desnues, C., Balique, F., Moal, V., Grob, J. J., Berbis, P., et al. (2010). Pepper mild mottle virus, a plant virus associated with specific immune responses, fever, abdominal pains, and pruritus in humans. *PLoS ONE*, *5*(4).
Corlan, A. D. (2004). Medline trend: automated yearly statistics of PubMed results for any query, 2004. Retrieved from http://dan.corlan.net/medline-trend.html
Cotter, P. D., Ross, R. P., & Hill, C. (2013). Bacteriocins - a viable alternative to antibiotics? *Nature reviews. Microbiology*, *11*(2), 95–105.
Cox, G., & Wright, G. D. (2013). Intrinsic antibiotic resistance: Mechanisms, origins,

challenges and solutions. *International Journal of Medical Microbiology*, *303*(6–7), 287–92.
D'Costa, V. M., King, C. E., Kalan, L., Morar, M., Sung, W. W. L., Schwarz, C., Froese, D., et al. (2011). Antibiotic resistance is ancient. *Nature*, *477*(7365), 457–61.
David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., Erdman, S. E., et al. (2014). Host lifestyle affects human microbiota on daily timescales. *Genome Biology*, *15*(7), R89.
David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V, et al. (2013). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, *505*(7484), 559–63.
Davie, J. R. (2003). Inhibition of Histone Deacetylase Activity. *J. Nutr*, *133*, 2485–93.
Derrien, M., Belzer, C., & de Vos, W. M. (2016). Akkermansia muciniphila and its role in regulating host functions. *Microbial Pathogenesis*, *410*(15), 30178–9.
Desai, K., Gupta, S. B., Dubberke, E. R., Prabhu, V. S., Browne, C., Mast, T. C., Gupta, S., et al. (2016). Epidemiological and economic burden of *Clostridium difficile* in the United States: estimates from a modeling approach. *BMC Infectious Diseases*, *16*(1), 303.
Dethlefsen, L., Eckburg, P. B., Bik, E. M., & Relman, D. A. (2006). Assembly of the human intestinal microbiota. *Trends in Ecology and Evolution*, *21*(9), 517–23.
Dropulic, L., & Jones, R. (2008). Polyomavirus BK infection in blood and marrow transplant recipients. *Bone Marrow Transplantation*, *41*(1), 11–8.
ECDC/EMEA. (2009). *The bacterial challenge : time to react*. *Technical Report*.
Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Sargent, M., Gill, S. R., Nelson, K. E., et al. (2005). Diversity of the human intestinal microbial flora. *Science (New York, N.Y.)*, *308*, 1635–8.
Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Computational Biology*, *7*(10), e1002195.
Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, *10*(10), 996–8.
Elbakri, A., Al-qahtani, A., & Samie, A. (2015). Advances on *Dientamoeba fragilis* infections. *An Overview of Tropical Diseases*.
Elting, L. S., Cooksley, C., Chambers, M., Cantor, S. B., Manzullo, E., & Rubenstein, E. B. (2003). The burdens of cancer therapy: Clinical and economic outcomes of chemotherapy-induced mucositis. *Cancer*, *98*(7), 1531–9.
Elting, L. S., Cooksley, C. D., Chambers, M. S., & Garden, A. S. (2007). Risk, outcomes, and costs of radiation-induced oral mucositis among patients with head-and-neck malignancies. *International Journal of Radiation Oncology Biology Physics*, *68*(4), 1110–20.
Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., & Delmont, T. O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, *3*, e1319.
Eriguchi, Y., Takashima, S., Oka, H., Shimoji, S., Nakamura, K., Uryu, H., Shimoda, S., et al. (2012). Graft-versus-host disease disrupts intestinal microbial ecology by inhibiting Paneth cell production of α-defensins. *Blood*, *120*(1), 223–31.
Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Published in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining* (Vol. KDD-96, pp. 1–6).
Ferrara, J. L., Levine, J. E., Reddy, P., & Holler, E. (2009). Graft-versus-host disease. *The Lancet*, *373*(9674), 1550–61.
Field, W., & Hershberg, R. (2015). Alarmingly high segregation frequencies of quinolone resistance alleles within human and animal microbiomes are not explained by direct clinical antibiotic exposure. *Genome biology and evolution*, *7*(6), 1743–57.

Flowers, M. E. D., Inamoto, Y., Carpenter, P. a, Lee, S. J., Kiem, H., Petersdorf, E. W., Pereira, S. E., et al. (2011). Comparative analysis of risk factors for acute graft-versus-host disease and for chronic graft-versus-host disease according to National Institutes of Health consensus criteria. *Blood*, *117*(11), 3214–9.

Flygare, S., Simmon, K., Miller, C., Qiao, Y., Kennedy, B., Di Sera, T., Graf, E. H., et al. (2016). Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biology*, *17*(1), 111.

Franzosa, E. A., Morgan, X. C., Segata, N., Waldron, L., Reyes, J., Earl, A. M., Giannoukos, G., et al. (2014). Relating the metatranscriptome and metagenome of the human gut. *Proceedings of the National Academy of Sciences*, *111*(22), E2329-38.

Fronzes, R., Christie, P. J., & Waksmas, G. (2009). The structural biology of type IV secretion systems. *Nature Review Microbiology*, *7*(10), 1–25.

Fuji, S., Kapp, M., & Einsele, H. (2014). Possible implication of bacterial infection in acute graft-versus-host disease after allogeneic hematopoietic stem cell transplantation. *Frontiers in oncology*, *4*(April), 89.

Fukata, M., Vamadevan, A. S., & Abreu, M. T. (2009). Toll-like receptors (TLRs) and Nod-like receptors (NLRs) in inflammatory disorders. *Seminars in immunology*, *21*(4), 242–53.

Furney, E. E. (1890). *Culture: A Modern Method*.

Gerasimidis, K., Bertz, M., Quince, C., Brunner, K., Bruce, A., Combet, E., Calus, S., et al. (2016). The effect of DNA extraction methodology on gut microbiota research applications. *BMC Research Notes*, *9*(1), 365.

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., et al. (2014). The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host and Microbe*, *15*(3), 382–92.

Gibson, M. K., Forsberg, K. J., & Dantas, G. (2014). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME journal*, *9*, 207–16.

Glucksberg, H., Storb, R., Fefer, A., Buckner, C. D., Neiman, P. E., Clift, R. A., Lerner, K. G., et al. (1974). Clinical manifestations of graft-versus-host disease in human recipients of marrow from HL-A-matched sibling donors. *Transplantation*, *18*(4), 295–304.

Goossens, H., Ferech, M., Vander Stichele, R., & Elseviers, M. (2005). Outpatient antibiotic use in Europe and association with resistance: a cross-national database study. *The Lancet*, *365*(9459), 579–87.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, *57*(1), 81–91.

Gorzelak, M. A., Gill, S. K., Tasnim, N., Ahmadi-Vand, Z., Jay, M., & Gibson, D. L. (2015). Methods for improving human gut microbiome data by reducing variability through sample processing and storage of stool. *PLoS ONE*, *10*(8), 1–14.

Greenhalgh, K., Meyer, K. M., Aagaard, K. M., & Wilmes, P. (2016). The human gut microbiome in health: establishment and resilience of microbiota over a lifetime. *Environmental Microbiology*, *18*, 2103–16.

Greninger, A. L., Naccache, S. N., Federman, S., Yu, G., Mbala, P., Bres, V., Stryke, D., et al. (2015). Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. *Genome medicine*, *7*(1), 99.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., et al. (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids*

Research, *41*(D1), D597–D604.
Hahn, T., McCarthy Jr., P. L., Zhang, M. J., Wang, D., Arora, M., Frangoul, H., Gale, R. P., et al. (2008). Risk factors for acute graft-versus-host disease after human leukocyte antigen-identical sibling transplants for adults with leukemia. *J Clin Oncol*, *26*(35), 5728–34.
Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell*, *100*, 57–70.
Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: The next generation. *Cell*, *144*(5), 646–74.
Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, *5*(10), R245–R249.
Harris, A. C., Ferrara, J. L. M., & Levine, J. E. (2013). Advances in predicting acute GVHD. *British journal of haematology*, *160*(3), 288–302.
Hashim, D., Boffetta, P., La Vecchia, C., Rota, M., Bertuccio, P., Malvezzi, M., & Negri, E. (2016). The global decrease in cancer mortality: Trends and disparities. *Annals of Oncology*, *27*(5), 926–33.
Hashizume, K., Tsukahara, T., Yamada, K., Koyama, H., & Ushida, K. (2003). *Megasphaera elsdenii* JCM1772T normalizes hyperlactate production in the large intestine of fructooligosaccharide-fed rats by stimulating butyrate production. *The Journal of Nutrition*, 3187–90.
Heintz-Buschart, A., May, P., Laczny, C. C., Lebrun, L. A., Bellora, C., Krishna, A., Wampach, L., et al. (2016). Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nature Microbiology*, *2*, 16180.
Herlemann, D. P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J. J., & Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *The ISME journal*, *5*(10), 1571–9.
Hildebrand, F., Tadeo, R., Voigt, A., Bork, P., & Raes, J. (2014). LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome*, *2*(1), 30.
Hill, D. a, Siracusa, M. C., Abt, M. C., Kim, B. S., Kobuley, D., Kubo, M., Kambayashi, T., et al. (2012). Commensal bacteria–derived signals regulate basophil hematopoiesis and allergic inflammation. *Nature Medicine*, *18*(4), 538–46.
Holler, E., Butzhammer, P., Schmid, K., Hundsrucker, C., Koestler, J., Peter, K., Zhu, W., et al. (2014). 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. *Biology of blood and marrow transplantation : journal of the American Society for Blood and Marrow Transplantation*, *20*(5), 640–5.
Hollister, E. B., Gao, C., & Versalovic, J. (2014). Compositional and functional features of the gastrointestinal microbiome and their effects on human health. *Gastroenterology*, *146*(6), 1449–58.
Hooper, L. V., & Gordon, J. (2001). Commensal host-bacterial relationships in the gut. *Science (New York, N.Y.)*, *292*(5519), 1115–8.
Hooper, L. V, Littman, D. R., & Macpherson, A. J. (2012). Interactions between the microbiota and the immune system. *Science (New York, N.Y.)*, *336*(6086), 1268–73.
Huddleston, J. R. (2014). Horizontal gene transfer in the human gastrointestinal tract: Potential spread of antibiotic resistance genes. *Infection and Drug Resistance*, *7*, 167–76.
Hugerth, L. (2015). Processing amplicons with non-overlapping reads. Retrieved April 19, 2016, from https://github.com/EnvGen/Tutorials/blob/master/amplicons-no_overlap.rst
Hugerth, L. W., Muller, E. E. L., Hu, Y. O. O., Lebrun, L. A. M., Roume, H., Lundin, D., Wilmes, P., et al. (2014). Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS ONE*, *9*(4), e95567.

Hugerth, L. W., Wefer, H. A., Lundin, S., Jakobsson, H. E., Lindberg, M., Rodin, S., Engstrand, L., et al. (2014). DegePrime, a program for degenerate primer design for broad-taxonomic-range PCR in microbial ecology studies. *Applied and Environmental Microbiology*, *80*(16), 5116–23.

Huson, D. H., Steel, M., El-Hadidi, M., Mitra, S., Peter, S., & Willmann, M. (2016). A simple statistical test of taxonomic or functional homogeneity using replicated microbiome sequencing samples. *Journal of Biotechnology*, http://dx.doi.org/doi:10.1016/j.jbiotec.2016.10.02.

Huson, D., Mitra, S., & Ruscheweyh, H. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, *21*(9), 1552–60.

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

Hyndman, I. J. (2016). Review: the Contribution of both Nature and Nurture to Carcinogenesis and Progression in Solid Tumours. *Cancer Microenvironment*, *9*(1), 63–9.

Iida, N., Dzutsev, A., Stewart, C. A., Smith, L., Bouladoux, N., Weingarten, R. a, Molina, D. a, et al. (2013). Commensal bacteria control cancer response to therapy by modulating the tumor microenvironment. *Science (New York, N.Y.)*, *342*(6161), 967–70.

Imbach, P., Kühne, T. R., & Arceci, R. (2004). *Pediatric Oncology: A Comprehensive Guide*.

Inan, M. S., Rasoulpour, R. J., Yin, L., Hubbard, A. K., Rosenberg, D. W., & Giardina, C. (2000). The luminal short-chain fatty acid butyrate modulates NF-κB activity in a human colonic epithelial cell line. *Gastroenterology*, *118*(4), 724–734.

Ivanov, I. I., Atarashi, K., Manel, N., Brodie, E. L., Shima, T., Karaoz, U., Wei, D., et al. (2009). Induction of Intestinal Th17 Cells by Segmented Filamentous Bacteria. *Cell*, *139*(3), 485–498.

Jacobsohn, D. A., & Vogelsang, G. B. (2007). Acute graft versus host disease. *Orphanet Journal of Rare Diseases*, *2*(1), 35.

Jagasia, M., Arora, M., Flowers, M. E. D., Chao, N. J., McCarthy, P. L., Cutler, C. S., Urbano-Ispizua, A., et al. (2012). Risk factors for acute GVHD and survival after hematopoietic cell transplantation. *Blood*, *119*(1), 296–307.

Jakobsson, H. E., Abrahamsson, T. R., Jenmalm, M. C., Harris, K., Jernberg, C., Björkstén, B., Engstrand, L., et al. (2013). Decreased gut microbiota diversity, delayed Bacteroidetes colonisation and reduced Th1 responses in infants delivered by Caesarean section. *Gut*, http://dx.doi.org/10.1136/gutjnl-2012-303249.

Jakobsson, H. E., Jernberg, C., Andersson, A. F., Sjölund-Karlsson, M., Jansson, J. K., & Engstrand, L. (2010). Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLoS ONE*, *5*(3).

Jenq, R. R., Taur, Y., Devlin, S. M., Ponce, D. M., Goldberg, J. D., Ahr, K. F., Littmann, E. R., et al. (2015). Intestinal *Blautia* is associated with reduced death from graft-versus-host disease. *Biology of Blood and Marrow Transplantation*, *21*(8), 1373–83.

Jenq, R. R., Ubeda, C., Taur, Y., Menezes, C. C., Khanin, R., Dudakov, J. a, Liu, C., et al. (2012). Regulation of intestinal inflammation by microbiota following allogeneic bone marrow transplantation. *The Journal of experimental medicine*, *209*(5), 903–11.

Jernberg, C., Löfmark, S., Edlund, C., & Jansson, J. K. (2007). Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *The ISME journal*, *1*(1), 56–66.

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., Lago, B. A., et al. (2016). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Research*, *45*(2017), D566–D573.

Jiang, W., Wu, N., Wang, X., Chi, Y., Zhang, Y., Qiu, X., Hu, Y., et al. (2015). Dysbiosis gut microbiota associated with inflammation and impaired mucosal immune function in intestine of humans with non-alcoholic fatty liver disease. *Scientific Reports*, *5*(8096).

Kahn, M. L., Ziermann, R., Dehó, G., Ow, D. W., Sunshine, M. G., & Calendar, R. (1991). [11] Bacteriophage P2 and P4. *Methods in Enzymology*, *204*, 264–280.

Kakihana, K., Fujioka, Y., Suda, W., Najima, Y., Kuwata, G., Sasajima, S., Mimura, I., et al. (2016). Fecal microbiota transplantation for patients with steroid-resistant/dependent acute graft-versus-host disease of the gut. *Blood*, *128*(16), 2083–9.

Kamada, N., Seo, S.-U., Chen, G. Y., & Núñez, G. (2013). Role of the gut microbiota in immunity and inflammatory disease. *Nature Publishing Group*, *13*(5), 321–35.

Kamboj, M., Chung, D., Seo, S. K., Pamer, E. G., Sepkowitz, K. A., Jakubowski, A. A., & Papanicolaou, G. (2010). The changing epidemiology of Vancomycin-Resistant *Enterococcus* (VRE) bacteremia in allogeneic hematopoietic stem cell transplant (HSCT) recipients. *Biology of Blood and Marrow Transplantation*, *16*(11), 1576–81.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–D462.

Kaur, T., Nafissi, N., Wasfi, O., Sheldon, K., Wettig, S., & Slavcev, R. (2012). Immunocompatibility of bacteriophages as nanomedicines. *Journal of Nanotechnology*, *2012*, 1–13.

Kharfan-Dabaja, M. A., Boeckh, M., Wilck, M. B., Langston, A. A., Chu, A. H., Wloch, M. K., Guterwill, D. F., et al. (2012). A novel therapeutic cytomegalovirus DNA vaccine in allogeneic haemopoietic stem-cell transplantation: A randomised, double-blind, placebo-controlled, phase 2 trial. *The Lancet Infectious Diseases*, *12*(4), 290–299.

Khoruts, A., Hippen, K. L., Lemire, A. M., Holtan, S. G., Knights, D., & Young, J.-A. H. (2016). Toward revision of antimicrobial therapies in hematopoietic stem cell transplantation: target the pathogens, but protect the indigenous microbiota. *Translational research : the journal of laboratory and clinical medicine*, *179*, 116–25.

Khosravi, A., & Mazmanian, S. K. (2013). Disruption of the gut microbiome as a risk factor for microbial infections. *Current opinion in microbiology*, *16*(2), 221–7.

Kincaid, R. P., Burke, J. M., Cox, J. C., de Villiers, E. M., & Sullivan, C. S. (2013). A human Torque Teno Virus encodes a microRNA that inhibits interferon signaling. *PLoS Pathogens*, *9*(12), 1–14.

Kinross, J. M., von Roon, A. C., Holmes, E., Darzi, A., & Nicholson, J. K. (2008). The human gut microbiome: implications for future health care. *Current Gastroenterology Reports*, *10*(4), 396–403.

Kolho, K., Korpela, K., Jaakkola, T., Pichai, M. V. A., Zoetendal, E. G., Salonen, A., & de Vos, W. M. (2015). Fecal microbiota in pediatric inflammatory bowel disease and its relation to inflammation. *The American Journal of Gastroenterology*, *110*, 921–30.

Kultima, J. R., Sunagawa, S., Li, J., Chen, W., Chen, H., Mende, D. R., Arumugam, M., et al. (2012). MOCAT: A metagenomics assembly and gene prediction toolkit. *PLoS ONE*, *7*(10), 1–6.

Kuntz, T. M., & Gilbert, J. A. (2017). Introducing the microbiome into precision medicine. *Trends in Pharmacological Sciences*, *38*(1), 81–91.

Laczny, C. C., Pinel, N., Vlassis, N., & Wilmes, P. (2014). Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Scientific Reports*, *4*, 4516.

Laczny, C. C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H., Coronado, S., et al. (2015). VizBin - an application for reference-independent visualization and human-augmented binning of metagenomic data. *Microbiome*, *3*(1), 7.

Ladas, E. J., Bhatia, M., Chen, L., Sandler, E., Petrovic, A., Berman, D. M., Hamblin, F., et

al. (2016). The safety and feasibility of probiotics in children and adolescents undergoing hematopoietic cell transplantation. *Bone marrow transplantation*, *51*, 262–6.

Lakhdari, O., Tap, J., Béguet-Crespel, F., Le Roux, K., De Wouters, T., Cultrone, A., Nepelska, M., et al. (2011). Identification of NF-κB modulation capabilities within human intestinal commensal bacteria. *Journal of Biomedicine and Biotechnology*, *2011*, 1–9.

Lakin, S. M., Dean, C., Noyes, N. R., Dettenwanger, A., Ross, A. S., Doster, E., Rovira, P., et al. (2016). MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic acids research*, *45*(2016), D574–D580.

Lam, S. J., O'Brien-Simpson, N. M., Pantarat, N., Sulistio, A., Wong, E. H. H., Chen, Y.-Y., Lenzo, J. C., et al. (2016). Combating multidrug-resistant Gram-negative bacteria with structurally nanoengineered antimicrobial peptide polymers. *Nature microbiology*, *10.1038/nm*.

Laurent, A., Nicco, C., Chéreau, C., Che, C., Goldwasser, F., Panis, Y., Soubrane, O., et al. (2005). Controlling tumor growth by modulating endogenous production of reactive oxygen species controlling tumor growth by modulating endogenous production of reactive oxygen species. *Cancer research*, *65*(3), 948–56.

Lee, S. M., Donaldson, G. P., Mikulski, Z., Boyajian, S., Ley, K., & Mazmanian, S. K. (2013). Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature*, *501*(7467), 426–9.

Lee, S., & Margolin, K. (2011). Cytokines in cancer immunotherapy. *Cancers*, *3*(4), 3856–93.

Legert, K. G., Remberger, M., Ringdén, O., Heimdahl, A., & Dahllöf, G. (2014). Reduced intensity conditioning and oral care measures prevent oral mucositis and reduces days of hospitalization in allogeneic stem cell transplantation recipients. *Supportive Care in Cancer*, *22*(8), 2133–40.

Lehouritis, P., Cummins, J., Stanton, M., Murphy, C. T., McCarthy, F. O., Reid, G., Urbaniak, C., et al. (2015). Local bacteria affect the efficacy of chemotherapeutic drugs. *Scientific Reports*, *5*, 14554.

Levine, J. E., Huber, E., Hammer, S. T. G., Harris, A. C., Greenson, J. K., Braun, T. M., Ferrara, J. L. M., et al. (2013). Low Paneth cell numbers at onset of gastrointestinal graft-versus-host disease identify patients at high risk for nonrelapse mortality. *Blood*, *122*(8), 1505–9.

Levy, S. B. (2002). Factors impacting on the problem of antibiotic resistance. *The Journal of antimicrobial chemotherapy*, *49*(1), 25–30.

Ley, R. E., Peterson, D. A., & Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, *124*(4), 837–48.

Ley, R. E., Turnbaugh, P. J., Klein, S., & Gordon, J. I. (2006). Microbial ecology: human gut microbes associated with obesity. *Nature*, *444*(7122), 1022–3.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–9.

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., et al. (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat Biotech*, *advance on*(8), 834–41.

Liang, M. D., Bagchi, A., Warren, H. S., Tehan, M. M., Trigilio, J. A., Beasley-topliffe, L. K., Tesini, B. L., et al. (2005). Bacterial peptidoglycan-associated lipoprotein : A naturally occurring Toll-like receptor 2 agonist that is shed into serum and has synergy with lipopolysaccharide. *Journal of Infectious Diseases*, *191*(g), 939–48.

Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–30.

Logan, R. M., Gibson, R. J., Bowen, J. M., Stringer, A. M., Sonis, S. T., & Keefe, D. M. K. (2008). Characterisation of mucosal changes in the alimentary tract following administration of irinotecan: Implications for the pathobiology of mucositis. *Cancer Chemotherapy and Pharmacology*, *62*(1), 33–41.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Lozupone, C., Stomabaugh, J., Gordon, J., Jansson, J., & Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, *489*(7415), 220–30.

Luckey, T. (1972). Introduction to intestinal microecology. *American journal of clinical nutrition*, *25*(12), 1292–4.

Luo, J., Solimini, N. L., & Elledge, S. J. (2009). Principles of cancer therapy: Oncogene and non-oncogene addiction. *Cell*, *136*(5), 823–37.

Ma, Y., Madupu, R., Karaoz, U., Nossa, C. W., Yang, L., Yooseph, S., Yachimski, P. S., et al. (2014). Human papillomavirus community in healthy persons, defined by metagenomics analysis of human microbiome project shotgun sequencing data sets. *Journal of virology*, *88*(9), 4786–97.

Macpherson, A. J., & Harris, N. L. (2004). Interactions between commensal intestinal bacteria and the immune system. *Nature reviews. Immunology*, *4*(6), 478–85.

Madden, T. (2002). Chapter 16 : The BLAST Sequence Analysis Tool. *The NCBI Handbook[internet]* (pp. 1–15).

Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., Nalin, R., et al. (2006). Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut*, *55*(2), 205–11.

Mantis, N. J., Rol, N., & Corthésy, B. (2011). Secretory IgA's complex roles in immunity and mucosal homeostasis in the gut. *Mucosal immunology*, *4*(6), 603–11.

Mathewson, N. D., Jenq, R., Mathew, A. V, Koenigsknecht, M., Hanash, A., Toubai, T., Oravecz-Wilson, K., et al. (2016). Gut microbiome–derived metabolites modulate intestinal epithelial cell damage and mitigate graft-versus-host disease. *Nature Immunology*, *17*(5), 505–13.

Maurice, C. F., Haiser, H. J., & Turnbaugh, P. J. (2013). Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell*, *152*(1–2), 39–50.

Mazmanian, S. K., Cui, H. L., Tzianabos, A. O., & Kasper, D. L. (2005). An immunomodulatory molecule of symbiotic bacteria directs maturation of the host immune system. *Cell*, *122*(1), 107–18.

Mohammadgholi, A., Rabbani-Chadegani, A., & Fallah, S. (2013). Mechanism of the interaction of plant alkaloid vincristine with DNA and chromatin: spectroscopic study. *DNA and cell biology*, *32*(5), 228–35.

Montassier, E., Batard, E., Massart, S., Gastinne, T., Carton, T., Caillon, J., Le Fresne, S., et al. (2014). 16S rRNA gene pyrosequencing reveals shift in patient faecal microbiota during high-dose chemotherapy as conditioning regimen for bone marrow transplantation. *Microbial ecology*, *67*(3), 690–9.

Montecucco, A., Zanetta, F., & Biamonti, G. (2015). Molecular mechanisms of etoposide. *EXCLI Journal*, *14*, 95–108.

Moore, W. E., & Moore, L. H. (1995). Intestinal floras of populations that have a high risk of colon cancer. *Applied and environmental microbiology*, *61*(9), 3202–7.

Morrison, D. J., & Preston, T. (2016). Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes*, *7*(3), 189–200.

Muller, E. E. L., Pinel, N., Laczny, C. C., Hoopmann, M. R., Narayanasamy, S., Lebrun, L. A., Roume, H., et al. (2014). Community-integrated omics links dominance of a microbial generalist to fine-tuned resource usage. *Nature communications*, *5*, 5603.

Nakano, V., Ignacio, A., Fernandes, M. R., Fukugaiti, M. H., & Avila-campos, M. J. (2006). Intestinal Bacteroides and Parabacteroides species producing antagonistic substances. *Current Trends in Microbiology*, *1*, 61–4.

Nakayama, J., Watanabe, K., Jiang, J., Matsuda, K., Chao, S.-H., Haryono, P., La-Ongkham, O., et al. (2015). Diversity in gut bacterial community of school-age children in Asia. *Scientific Reports*, *5*, 8397.

Nam, K. H., Kurinov, I., & Ke, A. (2011). Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca2+-dependent double-stranded DNA binding activity. *Journal of Biological Chemistry*, *286*(35), 30759–68.

Narayanasamy, S., Jarosz, Y., Muller, E. E. L., Laczny, C. C., Herold, M., Kaysen, A., Heintz-Buschart, A., et al. (2016a). IMP: a pipeline for reproducible metagenomic and metatranscriptomic analyses. *Genome Biology*, *17*(260).

Narayanasamy, S., Jarosz, Y., Muller, E. E. L., Laczny, C. C., Herold, M., Kaysen, A., Heintz-Buschart, A., et al. (2016b). IMP: a pipeline for reproducible metagenomic and metatranscriptomic analyses. *bioRxiv*.

Natividad, J. M. M., Hayes, C. L., Motta, J. P., Jury, J., Galipeau, H. J., Philip, V., Garcia-Rodenas, C. L., et al. (2013). Differential induction of antimicrobial REGIII by the intestinal microbiota and Bifidobacterium breve NCC2950. *Applied and Environmental Microbiology*, *79*(24), 7745–54.

Neelapu, N., & Surekha, C. (2016). Next-Generation Sequencing and Metagenomics. In K.-C. Wong (Ed.), *Computational Biology and Bioinformatics* (pp. 331–51).

Neill, J. I. M. O. (2016). Tackling drug-resistant infections globally: final report and recommendations. The review on antimicrobial resistance.

Newman, W. G. (2010). *Pharmacogenetics: Making cancer treatment safer and more effective*.

Nguyen TL, Vieira-Silva S, Liston A, R. J. (2015). How informative is the mouse for human gut microbiota research? *Disease Models & Mechanisms*, *8*, 1–16.

Nielsen, H. B., Almeida, M., Juncker, A. S., Rasmussen, S., Li, J., Sunagawa, S., Plichta, D. R., et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology*, *32*(8), 822–8.

Nobrega, F. L., Costa, A. R., Kluskens, L. D., & Azeredo, J. (2015). Revisiting phage therapy: New applications for old resources. *Trends in Microbiology*, *23*(4), 185–91.

O'Hara, A. M., & Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Reports*, *7*(7), 688–93.

Okamoto, H. (2009). History of discoveries and pathogenicity of TT viruses. *Current Topics in Microbiology and Immunology* (Vol. 331, pp. 1–20).

Oksanen, A. J., Blanchet, F. G., Kindt, R., Minchin, P. R., Hara, R. B. O., Simpson, G. L., Soly-, P., et al. (2015). Package "vegan," 1–285.

Ott, P. A., Hodi, F. S., & Robert, C. (2013). CTLA-4 and PD-1/PD-L1 blockade: New immunotherapeutic modalities with durable clinical benefit in melanoma patients. *Clinical Cancer Research*, *19*(19), 5300–9.

Ottman, N., Smidt, H., de Vos, W. M., & Belzer, C. (2012). The function of our microbiota: who is out there and what do they do? *Front Cell Infect Microbiol*, *9*(2), 104.

Ouyang, W., Kolls, J., & Zheng, Y. (2012). The biological functions of Th17 cell effector cytokines in inflammation. *Immunity*, *28*(4), 454–67.

De Paepe, M., Leclerc, M., Tinsley, C. R., & Petit, M.-A. (2014). Bacteriophages: an underestimated role in human and animal health? *Frontiers in cellular and infection microbiology*, *4*, 39.

Palmer, C., Bik, E. M., DiGiulio, D. B., Relman, D. A., & Brown, P. O. (2007). Development of the human infant intestinal microbiota. *PLoS Biology*, *5*(7), 1556–73.

Park, M., & Seo, J. J. (2012). Role of HLA in hematopoietic stem cell transplantation. *Bone Marrow Research*, *2012*, 1–7.

Pecorino, L. (2012). *Molecular biology of cancer: mechanisms, tagets, and therapeutics*.

Penack, O., Holler, E., & van den Brink, M. R. M. (2010). Graft-versus-host disease:

regulation by microbe-associated molecules and innate immune receptors. *Blood*, *115*(10), 1865–72.

Penders, J., Thijs, C., Vink, C., Stelma, F. F., Snijders, B., Kummeling, I., van den Brandt, P. A., et al. (2006). Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics*, *118*(2), 511–21.

Peng, L., Li, Z., Green, R. S., Holzman, I. R., & Lin, J. (2009). Butyrate enhances the intestinal barrier by facilitating tight junction assembly via activation of AMP-activated protein kinase. *Journal of Nutrition*, *139*(9), 1619–25.

Perez-Chanona, E., & Jobin, C. (2014). From promotion to management: the wide impact of bacteria on cancer and its treatment. *BioEssays : news and reviews in molecular, cellular and developmental biology*, *36*(7), 658–64.

Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. a, Bonazzi, V., et al. (2009). The NIH Human Microbiome Project. *Genome research*, *19*(12), 2317–23.

Png, C. W., Linden, S. K., Gilshenan, K. S., Zoetendal, E. G., McSweeney, C. S., Sly, L. I., McGuckin, M. A., et al. (2010). Mucolytic bacteria with increased prevalence in IBD mucosa augment in vitro utilization of mucin by other bacteria. *American Journal of Gastroenterology*, *105*(11), 2420–8.

Popgeorgiev, N., Temmam, S., Raoult, D., & Desnues, C. (2013). Describing the silent human virome with an emphasis on giant viruses. *Intervirology*, *56*(6), 395–412.

Porter, M. E., & Dorman, C. J. (1997). Positive regulation of Shigella flexneri virulence genes by integration host factor. *Journal of bacteriology*, *179*(21), 6537–50.

Przerwa, A., Zimecki, M., Świtała-Jeleń, K., Dkabrowska, K., Krawczyk, E., Łuczak, M., Weber-Dkabrowska, B., et al. (2006). Effects of bacteriophages on free radical production and phagocytic functions. *Medical Microbiology and Immunology*, *195*(3), 143–50.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, *464*(7285), 59–65.

Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, *490*(7418), 55–60.

Quick, J., Ashton, P., Calus, S., Chatt, C., Gossain, S., Hawker, J., Nair, S., et al. (2015). Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella. *Genome Biology*, *16*(1), 114.

R Development Core Team. (2008). R: A language and environment for statistical computing, *5*.

Rampelli, S., Soverini, M., Turroni, S., Quercia, S., Biagi, E., Brigidi, P., & Candela, M. (2016). ViromeScan: a new tool for metagenomic viral community profiling. *BMC Genomics*, *17*(1), 165.

Rao, N. G., Han, G., Greene, J. N., Tanvetyanon, T., Kish, J. A., De Conti, R. C., Chuong, M. D., et al. (2013). Effect of prophylactic fluconazole on oral mucositis and candidiasis during radiation therapy for head-and-neck cancer. *Practical Radiation Oncology*, *3*(3), 229–33.

Ratain, M. (1998). Body-surface area as a basis for dosing of anti-cancer agents. *Journal of Clinical Oncology*, *16*, 2297–8.

Reunanen, J., Kainulainen, V., Huuskonen, L., Ottman, N., Belzer, C., Huhtinen, H., de Vos, W. M., et al. (2015). *Akkermansia muciniphila* adheres to enterocytes and strengthens the integrity of the epithelial cell layer. *Applied and Environmental Microbiology*, *81*(11), 3655–62.

Ringel-Kulka, T., Cheng, J., Ringel, Y., Salojärvi, J., Carroll, I., Palva, A., de Vos, W. M., et al. (2013). Intestinal microbiota in healthy U.S. young children and adults - A high throughput microarray analysis. *PLoS ONE*, *8*(5), e64315.

Roume, H., Heintz-Buschart, A., Muller, E. E. L., & Wilmes, P. (2013). Sequential isolation of metabolites, RNA, DNA, and proteins from the same unique sample. *Methods in Enzymology*, *531*, 219–36.

Roume, H., Muller, E. E. L., Cordes, T., Renaut, J., Hiller, K., & Wilmes, P. (2013). A biomolecular isolation framework for eco-systems biology. *The ISME Journal*, *7*(1), 110–21.

Round, J. L., & Mazmanian, S. K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nature reviews. Immunology*, *9*(5), 313–23.

Sahin, U., Toprak, S. K., Atilla, P. A., Atilla, E., & Demirer, T. (2016). An overview of infectious complications after allogeneic hematopoietic stem cell transplantation. *Journal of Infection and Chemotherapy*, *22*(8), 505–14.

Salminen, M. K., Rautelin, H., Tynkkynen, S., Poussa, T., Saxelin, M., Valtonen, V., & Järvinen, A. (2004). *Lactobacillus* bacteremia, clinical significance, and patient outcome, with special focus on probiotic *L. rhamnosus* GG. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, *38*(1), 62–9.

Sandkvist, M. (2001). Type II secretion and pathogenesis. *Infection and Immunity*, *69*(6), 3523–35.

Santiago-Rodriguez, T. M., Fornaciari, G., Luciani, S., Dowd, S. E., Toranzos, G. A., Marota, I., Cano, R. J., et al. (2015). Gut microbiome of an 11th century A.D. Pre-Columbian andean mummy. *PLoS ONE*, *10*(9), 1–23.

Santini, S., Jeudy, S., Bartoli, J., Poirot, O., Lescot, M., Abergel, C., Barbe, V., et al. (2013). Genome of *Phaeocystis globosa* virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(26), 10800–5.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., & Siksnys, V. (2011). The Streptococcus thermophilus CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Research*, *39*(21), 9275–82.

Scanlan, P. D., Stensvold, C. R., Rajilić-Stojanović, M., Heilig, H. G. H. J., De Vos, W. M., O'Toole, P. W., & Cotter, P. D. (2014). The microbial eukaryote *Blastocystis* is a prevalent and diverse member of the healthy human gut microbiota. *FEMS Microbiology Ecology*, *90*(1), 326–30.

Scheppach, W. (1994). Effects of short chain fatty acids on gut morphology and function. *Gut*, *35*(1 Suppl), S35-8.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., et al. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–41.

Schmid, C., Schleuning, M., Ledderose, G., Tischer, J., & Kolb, H. J. (2005). 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. *Journal of Clinical Oncology*, *23*(24), 5675–87.

Schneeberger, M., Everard, A., Gómez-Valadés, A. G., Matamoros, S., Ramírez, S., Delzenne, N. M., Gomis, R., et al. (2015). *Akkermansia muciniphila* inversely correlates with the onset of inflammation, altered adipose tissue metabolism and metabolic disorders during obesity in mice. *Scientific reports*, *5*(October), 16643.

Scholz, M., Ward, D. V, Pasolli, E., Tolio, T., Zolfo, M., Asnicar, F., Truong, D. T., et al. (2016). Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nature Methods*, *13*(5), 435–8.

Schuijt, T. J., van der Poll, T., de Vos, W. M., & Wiersinga, W. J. (2013). The intestinal microbiota and host immune interactions in the critically ill. *Trends in microbiology*, *21*(5), 221–9.

Schwiertz, A. (2016). The Human Gut Microbiota. In A. Schwiertz (Ed.), *Microbiota of the Human Body: Implications in Health and Disease* (pp. 95–108).

Van Sebille, Y. Z. A., Stansborough, R., Wardill, H. R., Bateman, E., Gibson, R. J., & Keefe, D. M. (2015). Management of mucositis during chemotherapy: From pathophysiology to pragmatic therapeutics. *Current Oncology Reports*, *17*(11), 50.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–9.

Segata, N., Boernigen, D., Tickle, T. L., Morgan, X. C., Garrett, W. S., & Huttenhower, C. (2013). Computational meta'omics for microbial community studies. *Molecular Systems Biology*, *9*(1), 666.

Sekirov, I., Russell, S. L., Antunes, L. C. M., & Finlay, B. B. (2010). Gut microbiota in health and disease. *Physiological reviews*, *90*(3), 859–904.

Sender, R., Fuchs, S., & Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biology*, *14*(8), 1–14.

Servin, A. L. (2004). Antagonistic activities of lactobacilli and bifidobacteria against microbial pathogens. *FEMS Microbiology Reviews*, *28*(4), 405–40.

Shah, P., Fritz, J. V., Glaab, E., Desai, M. S., Greenhalgh, K., Frachet, A., Niegowska, M., et al. (2016). A microfluidics-based in vitro model of the gastrointestinal human-microbe interface. *Nature Communications*, *11*(7), 11535.

Shakya, M., Quince, C., Campbell, J., Yang, Z., Schadt, C., & Podar, M. (2013). Comparative metagenomic and rRNA microbial diversity characterization using Archaeal and Bacterial synthetic communities. *Environmental Microbiology*, *15*(6), 1882–99.

Shannon, C. E., & Weaver, W. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shen, S., & Wong, C. H. (2016). Bugging inflammation: role of the gut microbiota. *Clinical & Translational Immunology*, *5*(4), e72.

Shewach, D. S., & Kuchta, R. D. (2009). Introduction to cancer chemotherapeutics. *Chemical Reviews*, *109*(7), 2859–61.

Shlomchik, W. D. (2007). Graft-versus-host disease. *Nature Reviews Immunology*, *7*(5), 340–52.

Shono, Y., Docampo, M. D., Peled, J. U., Perobelli, S. M., Velardi, E., Tsai, J. J., Slingerland, A. E., et al. (2016). Increased GVHD-related mortality with broad-spectrum antibiotic use after allogeneic hematopoietic stem cell transplantation in human patients and mice. *Science Translational Medicine*, *8*(339), 339ra71.

Siegel, R. L., Miller, K. D., & Jemal, A. (2016). Cancer statistics. *CA: A Cancer Journal for Clinicians*, *66*(1), 7–30.

Simms-Waldrip, T. R., Sunkersett, G., Coughlin, L. A., Savani, M. R., Arana, C., Kim, J., Kim, M., et al. (2017). Antibiotic-induced depletion of Anti-Inflammatory Clostridia is associated with the development of GVHD in pediatric stem cell transplant patients. *Biology of Blood and Marrow Transplantation*, *0*(0).

Sivan, A., Corrales, L., Hubert, N., Williams, J. B., Aquino-Michaels, K., Earley, Z. M., Benyamin, F. W., et al. (2015). Commensal *Bifidobacterium* promotes antitumor immunity and facilitates anti-PD-L1 efficacy. *Science*, *350*(6264), 1084–9.

Sjölund, M., Wreiber, K., Andersson, D. I., Blaser, M. J., & Engstrand, L. (2003). Long-term persistence of resistant *Enterococcus* species after antibiotics to eradicate *Helicobacter pylori*. *Annals of Internal Medicine*, *139*(6), 483–7.

Slimings, C., & Riley, T. V. (2014). Antibiotics and hospital-acquired *Clostridium difficile* infection: Update of systematic review and meta-analysis. *Journal of Antimicrobial Chemotherapy*, *69*(4), 881–91.

Sommer, F., & Bäckhed, F. (2013). The gut microbiota-masters of host development and physiology. *Nature reviews. Microbiology*, *11*(4), 227–38.

Sommer, M. O. A., Church, G. M., & Dantas, G. (2010). The human microbiome harbors a

diverse reservoir of antibiotic resistance genes. *Virulence*, *1*(4), 299–303.
Sonis, S. (2004). A biological approach to mucositis. *J Support Oncol*, 21–36.
Spinler, J. K., Taweechotipatr, M., Rognerud, C. L., Ou, C. N., Tumwasorn, S., & Versalovic, J. (2008). Human-derived probiotic *Lactobacillus reuteri* demonstrate antimicrobial activities targeting diverse enteric bacterial pathogens. *Anaerobe*, *14*(3), 166–71.
Stecher, B., & Hardt, W. D. (2011). Mechanisms controlling pathogen colonization of the gut. *Current Opinion in Microbiology*, *14*(1), 82–91.
Stecher, B., Maier, L., & Hardt, W.-D. (2013). "Blooming" in the gut: how dysbiosis might contribute to pathogen evolution. *Nature reviews. Microbiology*, *11*(4), 277–84.
Stewart, B., & Wild, C. (2014). *World Cancer Report 2014*.
Stockham, A. L., Balagamwala, E. H., Macklis, R., Wilkinson, A., & Singh, A. D. (2014). Principles of radiation therapy. *Clinical Ophthalmic Oncology: Basic Principles and Diagnostic Techniques, Second Edition* (pp. 89–98).
Stonehouse, E., Kovacikova, G., Taylor, R. K., & Skorupski, K. (2008). Integration host factor positively regulates virulence gene expression in Vibrio cholerae. *Journal of Bacteriology*, *190*(13), 4736–48.
Stratton, M., Campbell, P., & Futreal, A. (2009). The cancer genome. *Nature*, *458*(7239), 719–24.
Stringer, A. M., Al-Dasooqi, N., Bowen, J. M., Tan, T. H., Radzuan, M., Logan, R. M., Mayo, B., et al. (2013). Biomarkers of chemotherapy-induced diarrhoea: A clinical study of intestinal microbiome alterations, inflammation and circulating matrix metalloproteinases. *Supportive Care in Cancer*, *21*(7), 1843–52.
Sun, L., Klein, E. Y., & Laxminarayan, R. (2012). Seasonality and temporal correlation between community antibiotic use and resistance in the United States. *Clinical Infectious Diseases*, *55*(5), 687–94.
Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. a, Kultima, J. R., Coelho, L. P., et al. (2013). Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods*, *10*(12), 1196–9.
Surawicz, C. M., Brandt, L. J., Binion, D. G., Ananthakrishnan, A. N., Curry, S. R., Gilligan, P. H., McFarland, L. V, et al. (2013). Guidelines for diagnosis, treatment, and prevention of *Clostridium difficile* infections. *The American Journal of Gastroenterology*, *108*(4), 478–98.
Tabbara, I. A., Zimmerman, K., Morgan, C., & Nahleh, Z. (2002). Allogeneic hematopoietic stem cell transplantation. *Archives of Internal Medicine*, *162*(14), 1558–66.
Tapparel, C., Siegrist, F., Petty, T. J., & Kaiser, L. (2013). Picornavirus and enterovirus diversity with associated human diseases. *Infection, Genetics and Evolution*, *14*(1), 282–93.
Taur, Y. (2016). Intestinal microbiome changes and stem cell transplantation: Lessons learned. *Virulence*, *7*(8), 930–8.
Taur, Y., Jenq, R. R., Perales, M., Littmann, E. R., Morjaria, S., Ling, L., No, D., et al. (2014). The effects of intestinal tract bacterial diversity on mortality following allogeneic hematopoietic stem cell transplantation. *Blood*, *124*(7), 1174–82.
Taur, Y., Xavier, J. B., Lipuma, L., Ubeda, C., Goldberg, J., Gobourne, A., Lee, Y. J., et al. (2012). Intestinal domination and the risk of bacteremia in patients undergoing allogeneic hematopoietic stem cell transplantation. *Clinical Infectious Diseases*, *55*(7), 905–14.
The Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–14.
Thompson, A. L., Monteagudo-Mera, A., Cadenas, M. B., Lampl, M. L., & Azcarate-Peril, M. A. (2015). Milk- and solid-feeding practices and daycare attendance are associated with differences in bacterial diversity, predominant communities, and metabolic and immune function of the infant gut microbiome. *Frontiers in Cellular and*

Infection Microbiology, *5*(3), 1–15.
Thorn, Caroline; Oshiro, Connie; Marsh, Sharon; Hernandez-Boussard, Tina; McLeod, Howard; Klein, Teri; Altman, R. (2012). Doxorubicin pathways:pharmacodynamics and adverse effects. *Pharmacogenet Genomics*, *21*(7), 440–6.
Tomlinson, D., & Kline, N. E. (2005). *Pediatric oncology nursing advanced clinical handbook*.
Tortoli, E. (2003). Impact of genotypic studies on mycobacterial taxonomy: The new mycobacteria of the 1990s. *Clinical Microbiology Reviews*, *16*(2), 319–54.
Touchefeu, Y., Montassier, E., Nieman, K., Gastinne, T., Potel, G., Bruley des Varannes, S., Le Vacon, F., et al. (2014). Systematic review: the role of the gut microbiota in chemotherapy- or radiation-induced gastrointestinal mucositis - current evidence and potential clinical applications. *Alimentary Pharmacology & Therapeutics*, *40*(5), 409–21.
Trifilio, S. M., Pi, J., & Mehta, J. (2013). Changing epidemiology of *Clostridium difficile*-associated disease during stem cell transplantation. *Biology of Blood and Marrow Transplantation*, *19*(3), 405–9.
Troy, E. B., & Kasper, D. L. (2010). Beneficial effects of *Bacteroides fragilis* polysaccharides on the immune system. *Frontiers in Bioscience*, *15*(8), 25–34.
Ubeda, C., Bucci, V., Caballero, S., Djukovic, A., Toussaint, N. C., Equinda, M., Lipuma, L., et al. (2013). Intestinal microbiota containing *Barnesiella* species cures vancomycin-resistant *Enterococcus faecium* colonization. *Infection and Immunity*, *81*(3), 965–73.
Ubeda, C., Djukovic, A., & Isaac, S. (2017). Roles of the intestinal microbiota in pathogen protection. *Clinical & Translational Immunology*, *6*(2), e128.
Ubeda, C., & Pamer, E. G. (2012). Antibiotics, microbiota, and immune defense. *Trends in Immunology*, *33*(9), 459–66.
Ubeda, C., Taur, Y., Jenq, R. R., Equinda, M. J., Son, T., Samstein, M., Viale, A., et al. (2010). Vancomycin-resistant *Enterococcus* domination of intestinal microbiota is enabled by antibiotic treatment in mice and precedes bloodstream invasion in humans. *The Journal of Clinical Investigation*, *120*(12), 4332–41.
Ursell, K. L., Metcalf, L. J., Parfrey, L. W., & Knight, R. (2013). Definig the Human Microbiome. *Nutrition Reviews*, *70*(Suppl 1), S38–S44.
Vandeputte, D., Falony, G., Vieira-Silva, S., Tito, R. Y., Joossens, M., & Raes, J. (2016). Stool consistency is strongly associated with gut microbiota richness and composition, enterotypes and bacterial growth rates. *Gut*, *65*(1), 57–62.
Vartoukian, S. R., Palmer, R. M., & Wade, W. G. (2010). Strategies for culture of “unculturable” bacteria. *FEMS Microbiology Letters*, *309*(1), 1–7.
van der Velden, W. J. F. M., Netea, M. G., de Haan, A. F. J., Huls, G. A., Donnelly, J. P., & Blijlevens, N. (2013). Role of the mycobiome in human acute graft-versus-host disease. *Biology of Blood and Marrow Transplantation*, *19*(2), 329–32.
Vernia, P., Caprilli, R., Latella, G., Barbetti, F., Magliocca, F. M., & Cittadini, M. (1988). Fecal lactate and ulcerative colitis. *Gastroenterology*, *95*(6), 1564–8.
Vétizou, M., Pitt, J. M., Daillère, R., Lepage, P., Waldschmitt, N., Flament, C., Rusakiewicz, S., et al. (2015). Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science*, *350*(6264), 1079–84.
Viaud, S., Flament, C., Zoubir, M., Pautier, P., LeCesne, A., Ribrag, V., Soria, J.-C., et al. (2011). Cyclophosphamide induces differentiation of Th17 cells in cancer patients. *Cancer Research*, *71*(3), 661–5.
Viaud, S., Saccheri, F., Mignot, G., Yamazaki, T., Daillère, R., Hannani, D., Enot, D. P., et al. (2013). The intestinal microbiota modulates the anticancer immune effects of cyclophosphamide. *Science*, *342*(6161), 971–6.
Vinolo, M. A. R., Rodrigues, H. G., Nachbar, R. T., & Curi, R. (2011). Regulation of inflammation by short chain fatty acids. *Nutrients*, *3*(10), 858–76.

Vital, M., Howe, C., & Tiedje, M. (2014). Revealing the bacterial butyrate synthesis pathways by analyzing (meta) genomic data. *mBio*, *5*(2), 1–11.

van Vliet, M. J., Harmsen, H. H. J. M., de Bont, E. S. J. M., Tissing, W. J. E., Vliet, M. van, & Harmsen, H. H. J. M. (2010). The role of intestinal microbiota in the development and severity of chemotherapy-induced mucositis. *PLoS pathogens*, *6*(5), e1000879.

Voreades, N., Kozil, A., & Weir, T. L. (2014). Diet and the development of the human intestinal microbiome. *Frontiers in Microbiology*, *5*(494), 1–9.

Vriesendorp, H. M. (2003). Aims of conditioning. *Experimental Hematology*, *31*(10), 844–54.

Walker, A. W., Martin, J. C., Scott, P., Parkhill, J., Flint, H. J., & 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*(1), 10.1186/s40168-015-0087-4.

Wallden, K., Rivera-Calzada, A., & Waksman, G. (2010). Type IV secretion systems: Versatility and diversity in function. *Cellular Microbiology*, *12*(9), 1203–12.

Walsh, D., & Avashia, J. (1992). Glucocorticoids in clinical oncology. *Cleveland Clinic Journal of Medicine*.

Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–7.

Wang, Z., Klipfell, E., Bennett, B. J., Koeth, R., Levison, B. S., Dugar, B., Feldstein, A. E., et al. (2011). Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*, *472*(7341), 57–63.

Weber, D., Jenq, R. R., Peled, J. U., Taur, Y., Hiergeist, A., Koestler, J., Dettmer, K., et al. (2017). Microbiota disruption induced by early use of broad spectrum antibiotics is an independent risk factor of outcome after allogeneic stem cell transplantation. *Biology of Blood and Marrow Transplantation*, http://dx.doi.org/doi: 10.1016/j.bbmt.2017.02.006.

Wetterstrand, K. (2016). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Retrieved from www.genome.gov/sequencingcostsdata

Wexler, H. M. (2007). *Bacteroides*: The good, the bad, and the nitty-gritty. *Clinical Microbiology Reviews*, *20*(4), 593–621.

Wickham, H. (2009). *ggplot2 Elegant Graphics for Data Analysis* (1st ed.).

Wlodarska, M., Kostic, A. D., & Xavier, R. J. (2015). An integrative view of microbiome-host interactions in inflammatory bowel diseases. *Cell Host and Microbe*, *17*(5), 577–91.

Woods, C. R. (2013). False-positive results for immunoglobulin M serologic results: Explanations and examples. *Journal of the Pediatric Infectious Diseases Society*, *2*(1), 87–90.

World Economic Forum. (2016). The Global Risks Report 2016 11th Edition. *www.weforum.org*, 103.

Wu, M., & Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics*, *28*(7), 1033–4.

Yamamoto, H., Ishihara, K., Takeda, Y., Koizumi, W., & Ichikawa, T. (2013). Changes in the mucus barrier during cisplatin-induced intestinal mucositis in rats. *BioMed Research International*, *2013*, 1–8.

Yoosuf, N., Yutin, N., Colson, P., Shabalina, S. A., Pagnier, I., Robert, C., Azza, S., et al. (2012). Related giant viruses in distant locations and different habitats: *Acanthamoeba polyphaga moumouvirus* represents a third lineage of the *Mimiviridae* that is close to the *Megavirus* lineage. *Genome Biology and Evolution*, *4*(12), 1324–30.

Yu, L. C.-H., Shih, Y.-A., Wu, L.-L., Lin, Y.-D., Kuo, W.-T., Peng, W.-H., Lu, K.-S., et al. (2014). Enteric dysbiosis promotes antibiotic-resistant bacterial infection: systemic

dissemination of resistant and commensal bacteria through epithelial transcytosis. *AJP: Gastrointestinal and Liver Physiology*, *307*(8), 824–35.

Zama, D., Biagi, E., Masetti, R., Gasperini, P., Prete, A., Candela, M., Brigidi, P., et al. (2016). Gut microbiota and hematopoietic stem cell transplantation: where do we stand? *Bone marrow transplantation*, *52*(1), 7–14.

Zhang, T., Breitbart, M., Lee, W. H., Run, J. Q., Wei, C. L., Soh, S. W. L., Hibberd, M. L., et al. (2006). RNA viral community in human feces: Prevalence of plant pathogenic viruses. *PLoS Biology*, *4*(1), 108–18.

Zhernakova, A., Kurilshikov, A., Bonder, M. J., Tigchelaar, E. F., Schirmer, M., Vatanen, T., Mujagic, Z., et al. (2016). Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. *Science*, *352*(6285), 565–9.

Zhou, Y., Gao, H., Mihindukulasuriya, K. A., La Rosa, P. S., Wylie, K. M., Vishnivetskaya, T., Podar, M., et al. (2013). Biogeography of the ecosystems of the healthy human body. *Genome Biology*, *14*, R1.

Zwielehner, J., Lassl, C., Hippe, B., Pointner, A., Switzeny, O. J., Remely, M., Kitzweger, E., et al. (2011). Changes in human fecal microbiota due to chemotherapy analyzed by TaqMan-PCR, 454 sequencing and PCR-DGGE fingerprinting. *PloS one*, *6*(12), e28654.

Scientific Output

Submissions in peer-reviewed journals

- **Anne Kaysen**, Anna Heintz-Buschart, Emilie E. L. Muller, Shaman Narayanasamy, Linda Wampach, Cédric C. Laczny, Norbert Graf, Arne Simon, Katharina Franke, Jörg Bittenbring, Paul Wilmes, Jochen G. Schneider (2017). Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic stem cell transplantation. *Translational Research*. (in revision)
  Appendix A.1

Publications in peer-reviewed journals

- 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*. **Appendix A.2**

Oral presentations at scientific conferences, symposia and workshops

- 'Dynamic change of host gastrointestinal microbiome and immune status in relation to mucosal barrier effects during chemotherapy and immune ablative intervention in humans' (2015). *Life Sciences PhD days*. Belval, Luxembourg.

Poster presentations at scientific conferences, symposia and workshops

- 'Dynamic change of host gastrointestinal microbiome and immune status in relation to mucosal barrier effects during chemotherapy and immune ablative intervention in humans' (2013). *Life Sciences PhD days*. Luxembourg, Luxembourg.
- 'Dynamic change of host gastrointestinal microbiome and immune status in relation to mucosal barrier effects during chemotherapy and immune ablative intervention in humans' (2014). *Exploring Human Host-Microbiome Interactions in Health and Disease. Wellcome Trust.* Hinxton, UK
- 'Changes in the human gastrointestinal microbiome during cancer treatments' (2015). *International Human Microbiome Congress*. Kirchberg, Luxembourg.
- 'Changes in the human gastrointestinal microbiome during cancer treatments' (2015). *The Human Microbiome. EMBL*. Heidelberg, Germany.

Appendix A.1

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, Norbert Graf, Arne Simon, Katharina Franke, Jörg Bittenbring, Paul Wilmes, Jochen G. Schneider

Translational Research (in revision)

Contributions of author include:

- Sample processing
- Data processing
- Data analysis & interpretation
- Figure creation
- Writing and revision of manuscript

Manuscript Details

Manuscript number	TRANSRES_2017_50
Title	Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic stem cell transplantation
Article type	Full Length Article

Abstract

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. 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. 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. 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 allo-HSCT. In addition to a drastically decreased bacterial diversity, the post-treatment sample showed a higher overall number and higher expression levels for antibiotic resistance genes (ARGs). An organism causing a paravertebral abscess was shown to be linked to the GIT dysbiosis, suggesting loss of intestinal barrier integrity. 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
Manuscript region of origin	Europe

Corresponding Author	Jochen G. Schneider
Corresponding Author's Institution	University of Luxembourg
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, Jörg Bittenbring, Paul Wilmes, Jochen G. Schneider
Suggested reviewers	Christoph Reinhardt, Jesus Bañales, Christof von Kalle

Submission Files Included in this PDF

File Name [File Type]

Cover Letter.docx [Cover Letter]

Manuscript_ImMicroDyn1_20170206.docx [Manuscript File]

Table 1.docx [Table]

Table 2.docx [Table]

Table 3.docx [Table]

Submission Files Not Included in this PDF

File Name [File Type]

Figure 1.ai [Figure]

Figure 2.ai [Figure]

Figure 3.ai [Figure]

Figure 4.ai [Figure]

Figure 5.ai [Figure]

Supplementary Figure S1.ai [Figure]

To view all the submission files, including those not included in the PDF, click on the manuscript title on your EVISE Homepage, then click 'Download zip file'.

Prof. Dr. med. Jochen G. Schneider
UNIVERSITY OF LUXEMBOURG

Luxembourg Centre for Systems Biomedicine
Translational & Experimental Medicine (TEM)
CAMPUS BELVAL
House of Biomedicine II
6, avenue du Swing
L-4367 Esch-Belvaux
T +352 46 66 44 6154
F +352 46 66 44 36154
email: jg.schneider@outlook.com

Esch/Belvaux, 06.02.2017

Translational Research
Editorial Office
Elsevier Inc.
1600 John F. Kennedy Blvd., Suite 1800
Philadelphia, PA 19103
USA

Dear Dr. Laurence,

Please find enclosed our manuscript entitled *"Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic stem cell transplantation"* which we would like to submit as Research Article for publication in *Translational Research*. Our manuscript touches upon the scientifically challenging role of the gut microbiome in health and disease in a clinically extremely vulnerable patient entity, i.e. people undergoing an allogeneic stem cell transplantation for the treatment of malignant diseases. We think that this manuscript will be of interest to the readership of *Translational Research* because of several important reasons:

1. The high clinical relevance: the observed changes in the gastrointestinal microbiome, especially the apparent enrichment in facultative pathogens expressing antibiotic resistance genes suggests that prophylactic antibiotic administration may adversely affect overall treatment outcome. We even present novel data suggesting that bacteria commonly considered as safe microorganisms (lactobacilli) may produce adverse effects during a stem cell transplantation. Our analyses may initiate follow up studies that might enable clinicians to adjust the therapy regimens according to individuals.
2. The main findings: while we observed a drastic decrease in bacterial diversity, the eukaryotic community seemed to be more resilient against the intensive treatment. Furthermore, no specific pattern in response to treatment was found in the archaeal populations. Additionally, detailed metagenomic and metatranscriptomic analyses of samples from one patient highlight the long-term effects that this intensive treatment has on the gastrointestinal microbiome. We further observed a clear enrichment in the copy numbers and expression of antibiotic resistance genes.
3. The novelty of the analysis approach and the dataset: According to our knowledge, the manuscript describes the first detailed analysis of the changes in the gastrointestinal microbiome of patients undergoing an allogeneic stem cell transplantation. The study includes information on all three domains of life in addition to metagenomic and metatranscriptomic data of samples from one specific patient who displayed an extensive

dysbiosis with enrichment in facultative pathogens and who deceased due to treatment side-effects.

The reported material reflects our own original research and has not been submitted for publication elsewhere. We declare that we do not have any conflicts of interest in relation to the presented research. The manuscript has been read and approved by all of the authors.

Thank you very much in advance for considering our manuscript and we look forward to hearing from you soon.

Yours sincerely,

Jochen G. Schneider

Integrated meta-omic analyses of the gastrointestinal tract microbiome in patients undergoing allogeneic stem cell transplantation

Anne Kaysen[a], Anna Heintz-Buschart[a], Emilie E. L. Muller[a,b], Shaman Narayanasamy[a], Linda Wampach[a], Cédric C. Laczny[a,b], Norbert Graf[c], Arne Simon[c], Katharina Franke[d], Jörg Bittenbring[d], Paul Wilmes[a], Jochen G. Schneider[a,e]*

[a] Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 6 avenue du Swing, L-4367 Belvaux, Luxembourg

[b] 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

[c] Saarland University Medical Center, Klinik für Pädiatrische Onkologie und Hämatologie, Geb. 9, Kirrberger Straße, 66421 Homburg, Germany

[d] Saarland University Medical Center, Klinik für Innere Medizin I, Geb. 41.1, Kirrberger Straße 100, 66421 Homburg, Germany

[e] 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, JGS: jg.schneider@outlook.com

* Correspondence: jg.schneider@outlook.com, +352 4666 44 6154

Running head: Meta-omics of the GIT microbiome in allo-HSCT

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

Abstract

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. 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. 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. 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 GvHD, which led to death nine months after allo-HSCT. In addition to a drastically decreased bacterial diversity, the post-treatment sample showed a higher overall number and higher expression levels for antibiotic resistance genes (ARGs). An organism causing a paravertebral abscess was shown to be linked to the GIT dysbiosis, suggesting loss of intestinal barrier integrity. 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.

Introduction

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 regimen, 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.

Material and methods

Study participants and fecal sample collection

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). 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 (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 and data analysis

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

196 1132r_CCGTCAATTHCTTYAART[31]) to resolve the eukaryotic community
197 structure.
198 16S rRNA gene sequencing reads were processed using the LotuS pipeline[32]
199 (version 1.34) with default parameters. Processed reads were clustered into
200 operational taxonomic units (OTUs), designating taxa with similar amplicon
201 sequences at 97 % identity level. For taxonomic assignment of 16S rRNA
202 gene amplicon sequencing data, the Ribosomal Database Project (RDP)
203 classifier[33] was used. OTUs with a confidence level below 0.8 at the domain
204 level were filtered out, as well as OTUs that were not represented by more
205 than 10 reads in any given sample.
206 To process the 18S rRNA gene sequencing reads, a workflow specifically
207 designed to process reads that are not overlapping was used.[34] For
208 classification of 18S rRNA gene amplicon sequencing data, the PR2
209 database[35–37] was employed. After processing, OTUs represented by less
210 than 10 reads in all samples were removed as well as unclassified OTUs and
211 OTUs belonging to the taxon Craniata, since they were most likely derived
212 from human sequences. For following analyses, the 16S and 18S rRNA gene
213 sequencing data were rarefied to the lowest number of respective reads for
214 any sample (16S to 71,051 reads and 18S to 1,020 reads).
215 Statistical analyses and plots were generated in R (version 3.2.1).[38] Microbial
216 alpha-diversity and richness were determined at the OTU level, by calculating
217 the Shannon diversity index and the Chao1 index after rarefaction, using the
218 vegan package.[39] Statistical comparison of diversity and richness was carried
219 out using the Kruskal-Wallis test, the non-parametric Wilcoxon rank sum test,
220 or, when applicable, Wilcoxon signed-rank test. When P values < 0.05 were

observed, groups were considered as statistically significantly different. Differential analysis of taxa based on 16S rRNA gene sequencing data was performed using the DESeq2 package[40] and significant differences on taxonomic levels were determined using the Wald test, after multiple-testing adjustment.

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.[41]

Within IMP, the average depth of coverage D_x of a gene or contig x was determined both for the metagenome and the metatranscriptome by calculating the average number of reads mapping to each nucleotide within a gene, respectively in a contig.

$$D_x = \frac{\sum r_x}{length_x}$$

where r_x is the number of reads mapping to a gene or contig x at each nucleotide.

Here, gene expression of a gene *x* was calculated as the ratio of average metatranscriptomic depth of coverage to the average metagenomic depth of coverage for individual genes *x*.

$$E_x = \frac{D_x(MT)}{D_x(MG)}$$

Published human GIT microbiome MG and MT read data from four healthy individuals was obtained from the NCBI Sequence Read Archive [MG: SRX247379, SRX247391, SRX247401, SRX247405; MT: SRX247335, SRX247345, SRX247349, SRX247340].[42] The sequencing reads were processed using IMP version 1.2.1.[41] Data from the individuals "X310763260", "X316192082", "X317690558" and "X316701492" are in the following referred to as the "reference healthy microbiome", averaged as "RHMs" or individually referred to as "RHM1", "RHM2", "RHM3" and "RHM4".

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,[43,44] 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.[45] Population-level clusters were selected following the method described in Heintz-Buschart et al.[46] Resulting bins are referred to as “population-level genomes” in the following.

Within a community, the relative population size of a cluster (*i*) was determined by dividing the number of MG reads mapping to the contigs

forming this cluster (c_i), by the total number of MG reads mapping to all the contigs used in the assembly (C) according to the following formula:

$$N_i = \frac{c_i * 100}{C}$$

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.[47] Parameters were left at default (using program megablast), and the output was analyzed using the MEtaGenome ANalyzer (MEGAN version 5.10.5).[48] 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.[49] Additionally, AMPHORA2[50] 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[51] (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 variant density.

295 **Sequence comparison of population-level genomes**

296 The average nucleotide identity (ANI) calculator[52] was used with standard
297 settings to compare the reassembly from population-level genomes to publicly
298 available reference genomes. A gene-wise protein sequence comparison of
299 different population-level genomes was performed using the RAST server[53]
300 using standard parameters.

301

302 **Detection of antibiotic resistance genes**

303 Antibiotic resistance genes (ARGs) within a community or population were
304 searched against Resfams version 1.2[54] using HMMer version 3.1b2[55]. We
305 used the core version of the Resfams database, which includes 119 protein
306 families. In accordance with the HMMer user manual, only identified genes
307 with a bitscore higher than the binary logarithm of the total number of genes
308 (of the community or population) were retained.

309

310 **Variant identification**

311 Variants were identified in population-level reassembled genomes using
312 SAMtools mpileup[56] with default settings, which include the calling of single
313 nucleotide variants (SNVs) as well as the identification of small
314 insertions/deletions (indels). The output of SAMtools mpileup was filtered
315 using a conservative heuristic established in Eren et al.[57] which takes into
316 account the ratio of the frequencies of both bases and the depth of coverage
317 at the corresponding nucleotide position, in order to reduce the effect of
318 sequencing errors.

319

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.[51] Using PanPhlAn[58] and the provided database including 118 *E. coli* reference strains, their relation was assessed based on their gene set. While the PanPhlAn database includes 31,734 genes, only genes present in 10 or more genomes were considered, resulting in 7,845 genes for comparison.

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). Supplementary tables are archived on Zenodo (https://doi.org/10.5281/zenodo.268914).

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).

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*.[59] It was detected in 13 out of the total 35 samples (and 10 out of 15 patients) with a total of 914 reads.[60]

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.[60] 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

395 treatment. To further explore the effects of treatment on the structure and
396 function of the GIT microbiome, we applied a detailed meta-omic approach on
397 one patient.
398
399 **Patient A07 - description of treatment and status of the patient**
400 We chose to focus on patient A07, a patient who displayed a marked
401 reduction in bacterial diversity with high relative abundances of opportunistic
402 pathogens (Fig. 2A and 2B) and a fatal treatment outcome. This 63 year old
403 patient had acute myeloid leukemia with deletion 7q. The patient was
404 refractory to conventional induction (3+7) and salvage chemotherapy with
405 high-dose cytarabine and mitoxantrone and therefore needed further
406 treatment. FLAMSA-Bu,[61] a modified sequential conditioning regimen for
407 refractory acute myeloid leukemia was used (fludarabine 30 mg/m² day -11 to
408 -8, cytarabine 2000 mg/m² day -11 to -8, amsacrine 100 mg/m² day -11 to -8
409 and busulfan 3,2 mg/kg day -7 to -4) for remission induction and
410 transplantation. She received peripheral blood stem cells from a single HLA-C
411 antigen mismatched unrelated donor. After engraftment on day 26, bone
412 marrow was hypocellular, but free of leukemia. Planned immunosuppression
413 consisted of antithymocyte globulin (ATG) on day -4 to -2, mycophenolate
414 mofetil until day 28 and cyclosporine until day 100.
415 A high level of C-reactive protein (CRP) before and around allo-HSCT was
416 observed which decreased slightly but stayed considerably high throughout
417 the entire observation period[60] (Fig. 2C). After leukocyte depletion around
418 allo-HSCT, the count increased to around 3600/µl 20 days after allo-HSCT
419 and further increased to a normal value around 80 days after allo-HSCT.

However, high fluctuations and later a decrease in the leukocyte count were observed[60] (Fig. 2C).

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 2 and GIT stage 3. 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 analyzed further. After surgery, the patient's health status improved, she 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 contained only 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* cluster IV.[60] 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.[60]

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 MG depth of coverage and number of variants in each sample are displayed in Fig. 4A and 4B for *E. coli* and in 4C and 4D for *E. faecium*. The average MG depths of coverage (Fig. 4E and 4F) indicated that the population size of *E. coli* was smaller after allo-HSCT (in sample A07-3), while the population size of *E. faecium* remained rather constant. 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. 4B), whereas only a few variants were present in *E. faecium* of both samples (Fig. 4D). A similar pattern of variant distributions (Fig. 4E and 4G) in both samples was observed for *E. coli*, while the variant pattern in *E. faecium* (Fig. 4F and 4H) changed between both samples. Observed nucleotide

545 variant frequencies and patterns of variant distributions indicated that the *E.*
546 *coli* populations were composed of different strains in both samples, which
547 persisted over the course of the treatment. In contrast, *E. faecium* was mainly
548 represented by a single strain in each sample, and the strain of the first
549 sample was replaced by a different strain in the second sample.
550
551 **Coupled metagenomic and metatranscriptomic analysis of antibiotic**
552 **resistance genes in pre- and post-treatment samples from patient A07**
553 The relative abundance of detected ARGs (percentage of ARGs relative to the
554 total number of genes, Fig. 5A) in the post-treatment sample (0.39 %) was
555 significantly higher than the relative abundance of ARGs in the pre-treatment
556 sample (0.28 % ARGs, *P* value $6.9 * 10^{-4}$, Fisher's exact test) while the relative
557 abundances of ARGs of both the pre- and post-treatment sample were higher
558 than the average relative abundance in the RHMs[60] (0.20 % ± 0.01 %, *P*
559 value $5.601 * 10^{-7}$ and $3.278 * 10^{-10}$). Moreover, the expression of ARGs was
560 higher in both samples from patient A07 when compared to the RHMs (Fig.
561 5B).
562
563 **Identification of antibiotic resistance genes in population-level genomes**
564 **of opportunistic pathogens**
565 Given the higher number and expression of ARGs in the post-treatment
566 sample of patient A07, we were interested in whether this could also be
567 detected in the specific populations *E. coli* and *E. faecium*. Within the
568 population-level genome of *E. coli*, 31 ARGs were identified in both samples
569 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 (summaries of the ARGs identified in each population-level genome are listed in Table 2 and Table 3).[60] 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.[60] 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).

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* (Fig. S1). 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[62–64] (the three bacterial genera *Roseburia*, *Bifidobacterium* and *Blautia*, Fig. 1C) was observed. SCFAs, especially the histone deacetylase inhibitor butyrate, are the main energy source for colonocytes,[62] as well as anti-inflammatory agents which regulate NF-κB activation in colonic epithelial cells.[62] Additionally, butyrate enhances the intestinal barrier function by regulating assembly of epithelial tight junctions[65] and a recent study showed that local administration of exogenous butyrate mitigated GvHD in mice.[66] Depletion of these important bacteria has been highlighted to pose an additional risk for developing GvHD or infections after allo-HSCT.[26,67] 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.[68] Contrary to this study, which observed that the richness and metabolic capacity of the microbial community recovered after two months,[68] 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).[69–71] 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.[72] 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,[73] 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.[74–76] 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.[77] *Bacteroides* spp. might be able to penetrate the colonic mucus and persevere within crypt channels. These reservoirs might persist even during antibiotic treatment.[78] Different species of the genus *Bacteroides* produce bacteriocins,[79–81] a trait that might have made it possible for these bacteria to repopulate the GIT and expand after the dysbiosis in A07-2,

669 occupying specific niches, resulting in a relative abundance of 63 % in A07-3
670 (day 119).
671 Facultative anaerobes such as members of the orders Lactobacillales and
672 Enterobacteriales are often observed to increase in relative abundance after
673 treatment while obligate anaerobes such as members of the order
674 Clostridiales often decrease in abundance.[82] *Lactobacillus rhamnosus* and
675 *Lactobacillus reuteri* (which were detected in sample A07-3) are both often
676 combined in probiotic formulations and are commonly considered safe and
677 even beneficial through inhibition of potential pathogen (such as *E. coli* and *E.*
678 *faecium*) expansion.[83–85] Even in patients undergoing allo-HSCT,
679 *Lactobacillus plantarum* administration has not been found to result in higher
680 incidence of bacteremia or aGvHD.[86] However, bacteria found in probiotic
681 formulations, especially *Lactobacillus* species have occasionally also caused
682 bloodstream infections.[87] Our data suggest that probiotics should be
683 administered with great caution and should be subject to further investigations
684 to clearly ensure safety of their usage.
685
686 **Identification of antibiotic resistance genes in population-level genomes**
687 **of opportunistic pathogens and evidence for selective pressure at the**
688 **strain-level**
689 A higher ratio of ARGs within the microbial community was observed post-
690 treatment, even a few months after the antibiotic treatment was concluded
691 (Fig. 5A). Importantly, the observed expression of ARGs was higher in the
692 post-treatment sample (Fig. 5B) when compared to the pre-treatment sample.
693 Strains that carry mutations which lead to higher expression of ARGs might

694 have been selected for by the antibiotic treatment.[88] In *E. coli*, three different
695 genes conferring resistance against β-lactams were identified, one of which
696 was only detected in the post-treatment sample, which might have been
697 acquired due to selective pressure given the administration of three different
698 β-lactam antibiotics during the treatment.
699 Observed nucleotide variant frequencies and patterns of variant distributions
700 indicated that the treatment may have constituted a genetic bottleneck for *E.*
701 *faecium*, culminating in the observed lower genetic diversity. This also
702 suggests that two different mechanisms influenced the respective
703 compositions of *E. coli* and *E. faecium* populations. While the *E. coli*
704 population remained relatively unaffected, the *E. faecium* population
705 underwent a selective sweep in response to the antibiotic treatment with
706 selection of a specific genotype expressing ARGs. Overall, our observations
707 indicate that antibiotic pressure and associated selection of bacteria encoding
708 ARGs are likely essential factors in governing the observed expansion in
709 opportunistic pathogens.
710 Interestingly, the multidrug-resistant *E. coli* that was isolated from a blood
711 culture, was closely related to the GIT-borne *E. coli* population. The overlap of
712 ARGs identified in each genome further indicates their association. These
713 findings are a proof for the potential fatal effects of dysbiosis associated
714 pathogen dominance in the GIT and subsequent systemic infections on
715 patient survival.
716 Based on our observation, one strategy to avoid a treatment-induced
717 intestinal domination by pathogens could consist in the tailored administration
718 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 tolerance or improve the overall efficacy of the therapy.

Acknowledgments

Conflicts of Interest: All authors have read the journal's authorship agreement. All authors have read the journal's policy on disclosure of potential conflicts of interest and have none to declare.

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).

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

744 their assistance with fecal sample extractions. Stephanie Kreis and Giulia
745 Cesi are thanked for valuable discussions in the establishment of the project.
746 In silico analyses presented in this paper were carried out using the HPC
747 facilities of the University of Luxembourg.[89]

748 Authors' contributions: JGS, PW, NG, AS and JB initiated and designed the
749 study. JB, NG and AS recruited patients and collected samples. KF collected
750 clinical data. AK performed experiments. AHB, SN and CCL developed the
751 bioinformatic analysis methods. AK, AHB, EELM and PW contributed to
752 analysis and interpretation of the data. AK, AHB, EELM, JB, JGS, LW and PW
753 wrote and revised the manuscript. All authors read and approved the final
754 manuscript.

755

756 References

757 1. Hooper L V., Gordon J. Commensal host-bacterial relationships in the
758 gut. Science. 2001;292:1115–8.
759 2. Sekirov I, Russell SL, Antunes LCM, Finlay BB. Gut microbiota in health
760 and disease. Physiol Rev. 2010;90:859–904.
761 3. Sommer F, Bäckhed F. The gut microbiota-masters of host
762 development and physiology. Nat Rev Microbiol. 2013;11:227–38.
763 4. Round JL, Mazmanian SK. The gut microbiota shapes intestinal
764 immune responses during health and disease. Nat Rev Immunol.
765 2009;9:313–23.
766 5. Atarashi K, Tanoue T, Oshima K, et al. Treg induction by a rationally
767 selected mixture of Clostridia strains from the human microbiota.
768 Nature. 2013;500:232–6.

769 6. Stecher B, Maier L, Hardt W-D. “Blooming” in the gut: how dysbiosis
770 might contribute to pathogen evolution. Nat Rev Microbiol.
771 2013;11:277–84.
772 7. Yu LC-H, Shih Y-A, Wu L-L, et al. Enteric dysbiosis promotes antibiotic-
773 resistant bacterial infection: systemic dissemination of resistant and
774 commensal bacteria through epithelial transcytosis. Am J Physiol
775 Gastrointest Liver Physiol. 2014;307:824–35.
776 8. Khosravi A, Mazmanian SK. Disruption of the gut microbiome as a risk
777 factor for microbial infections. Curr Opin Microbiol. 2013;16:221–7.
778 9. Taur Y, Xavier JB, Lipuma L, et al. Intestinal domination and the risk of
779 bacteremia in patients undergoing allogeneic hematopoietic stem cell
780 transplantation. Clin Infect Dis. 2012;55:905–14.
781 10. Einsele H, Bertz H, Beyer J, et al. Infectious complications after
782 allogeneic stem cell transplantation: Epidemiology and interventional
783 therapy strategies - Guidelines of the Infectious Diseases Working Party
784 (AGIHO) of the German Society of Hematology and Oncology (DGHO).
785 Ann Hematol. 2003;82:175–85.
786 11. Aminov RI, Mackie RI. Evolution and ecology of antibiotic resistance
787 genes. FEMS Microbiol Lett. 2007;271:147–61.
788 12. Salyers A, Gupta A, Wang Y. Human intestinal bacteria as reservoirs for
789 antibiotic resistance genes. Trends Microbiol. 2004;12:412–6.
790 13. Samonis G, P. KD, Maraki S, et al. Levofloxacin and moxifloxacin
791 increase human gut colonization by *Candida* species. Antimicrob
792 Agents Chemother. 2005;49:5189.
793 14. Zollner-Schwetz I, Auner HW, Paulitsch A, 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, 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, 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.

819 23. Vossen JM, Guiot HFL, Lankester AC, et al. Complete suppression of
820 the gut microbiome prevents acute graft-versus-host disease following
821 allogeneic bone marrow transplantation. PLoS One. 2014;9:e105706.
822 24. Holler E, Butzhammer P, Schmid K, et al. Metagenomic analysis of the
823 stool microbiome in patients receiving allogeneic stem cell
824 transplantation: loss of diversity is associated with use of systemic
825 antibiotics and more pronounced in gastrointestinal graft-versus-host
826 disease. Biol Blood Marrow Transplant. 2014;20:640–5.
827 25. Montassier E, Batard E, Massart S, et al. 16S rRNA gene
828 pyrosequencing reveals shift in patient faecal microbiota during high-
829 dose chemotherapy as conditioning regimen for bone marrow
830 transplantation. Microb Ecol. 2014;67:690–9.
831 26. Jenq RR, Taur Y, Devlin SM, et al. Intestinal *Blautia* is associated with
832 reduced death from graft-versus-host disease. Biol Blood Marrow
833 Transplant. 2015;21:1373–83.
834 27. Roume H, Heintz-Buschart A, Muller EEL, Wilmes P. Sequential
835 isolation of metabolites, RNA, DNA, and proteins from the same unique
836 sample. Methods Enzymol. 1st ed. 2013;531:219–36.
837 28. Roume H, Muller EEL, Cordes T, Renaut J, Hiller K, Wilmes P. A
838 biomolecular isolation framework for eco-systems biology. ISME J.
839 2013;7:110–21.
840 29. Hugerth LW, Wefer HA, Lundin S, et al. DegePrime, a program for
841 degenerate primer design for broad-taxonomic-range PCR in microbial
842 ecology studies. Appl Environ Microbiol. 2014;80:5116–23.
843 30. Herlemann DP, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ,

844 Andersson AF. Transitions in bacterial communities along the 2000 km
845 salinity gradient of the Baltic Sea. ISME J. 2011;5:1571–9.
846 31. Hugerth LW, Muller EEL, Hu YOO, et al. Systematic design of 18S
847 rRNA gene primers for determining eukaryotic diversity in microbial
848 consortia. PLoS One. 2014;9:e95567.
849 32. Hildebrand F, Tadeo R, Voigt A, Bork P, Raes J. LotuS: an efficient and
850 user-friendly OTU processing pipeline. Microbiome. 2014;2:30.
851 33. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive bayesian classifier for
852 rapid assignment of rRNA sequences into the new bacterial taxonomy.
853 Appl Environ Microbiol. 2007;73:5261–7.
854 34. Hugerth L. Processing amplicons with non-overlapping reads [Internet].
855 2015 [cited 2016 Apr 19]. Available from:
856 https://github.com/EnvGen/Tutorials/blob/master/amplicons-
857 no_overlap.rst
858 35. Guillou L, Bachar D, Audic S, et al. The Protist Ribosomal Reference
859 database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA
860 sequences with curated taxonomy. Nucleic Acids Res. 2013;41:D597–
861 604.
862 36. Chevenet F, Brun C, Bañuls A-L, Jacq B, Christen R. TreeDyn: towards
863 dynamic graphics and annotations for analyses of trees. BMC
864 Bioinformatics. 2006;7:439.
865 37. Chevenet F, Croce O, Hebrard M, Christen R, Berry V. ScripTree:
866 Scripting phylogenetic graphics. Bioinformatics. 2010;26:1125–6.
867 38. R Development Core Team. R: A language and environment for
868 statistical computing. 2008;5.

869 39. Oksanen AJ, Blanchet FG, Kindt R, et al. Package “vegan.” 2015;1–
870 285.
871 40. Love MI, Huber W, Anders S. Moderated estimation of fold change and
872 dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
873 41. Narayanasamy S, Jarosz Y, Muller EEL, et al. IMP: a pipeline for
874 reproducible metagenomic and metatranscriptomic analyses. Genome
875 Biol. 2016;17.
876 42. Franzosa EA, Morgan XC, Segata N, et al. Relating the
877 metatranscriptome and metagenome of the human gut. Proc Natl Acad
878 Sci. 2014;111:E2329-38.
879 43. Laczny CC, Pinel N, Vlassis N, Wilmes P. Alignment-free visualization
880 of metagenomic data by nonlinear dimension reduction. Sci Rep.
881 2014;4:4516.
882 44. Laczny CC, Sternal T, Plugaru V, et al. VizBin - an application for
883 reference-independent visualization and human-augmented binning of
884 metagenomic data. Microbiome. 2015;3:7.
885 45. Muller EEL, Pinel N, Laczny CC, et al. Community-integrated omics
886 links dominance of a microbial generalist to fine-tuned resource usage.
887 Nat Commun. 2014;5:5603.
888 46. Heintz-Buschart A, May P, Laczny CC, et al. Integrated multi-omics of
889 the human gut microbiome in a case study of familial type 1 diabetes.
890 Nat Microbiol. 2016;2:16180.
891 47. Madden T. Chapter 16 : The BLAST Sequence Analysis Tool. In: The
892 NCBI Handbook[internet]. 2002. p. 1–15.
893 48. Huson D, Mitra S, Ruscheweyh H. Integrative analysis of environmental

894 sequences using MEGAN4. Genome Res. 2011;21:1552–60.
895 49. MOLE-BLAST webservice [Internet]. [cited 2016 Apr 19]. Available
896 from: https://blast.ncbi.nlm.nih.gov/moleblast/moleblast.cgi
897 50. Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal
898 sequences with AMPHORA2. Bioinformatics. 2012;28:1033–4.
899 51. Bankevich A, Nurk S, Antipov D, et al. SPAdes: A new genome
900 assembly algorithm and its applications to single-cell sequencing. J
901 Comput Biol. 2012;19:455–77.
902 52. ANI Average Nucleotide Identity [Internet]. [cited 2016 Apr 19].
903 Available from: http://enve-omics.ce.gatech.edu/ani/
904 53. Aziz RK, Bartels D, Best A a, et al. The RAST Server: rapid annotations
905 using subsystems technology. BMC Genomics. 2008;9:75.
906 54. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic
907 resistance determinants reveals microbial resistomes cluster by
908 ecology. ISME J. 2014;9:207–16.
909 55. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol.
910 2011;7:e1002195.
911 56. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map
912 format and SAMtools. Bioinformatics. 2009;25:2078–9.
913 57. Eren AM, Esen ÖC, Quince C, et al. Anvi'o: an advanced analysis and
914 visualization platform for 'omics data. PeerJ. 2015;3:e1319.
915 58. Scholz M, Ward D V, Pasolli E, et al. Strain-level microbial epidemiology
916 and population genomics from shotgun metagenomics. Nat Methods.
917 2016;13:435–8.
918 59. Scanlan PD, Shanahan F, Marchesi JR. Human methanogen diversity

919 and incidence in healthy and diseased colonic groups using *mcrA* gene
920 analysis. BMC Microbiol. 2008;8:79.
921 60. Table S1 - Table S7: Integrated meta-omic analyses of the
922 gastrointestinal tract microbiome in patients undergoing allogeneic stem
923 cell transplantation [Internet]. [cited 2017 Feb 6]. Available from:
924 https://zenodo.org/record/268914
925 61. Schmid C, Schleuning M, Ledderose G, Tischer J, Kolb HJ. Sequential
926 regimen of chemotherapy, reduced-intensity conditioning for allogeneic
927 stem-cell transplantation, and prophylactic donor lymphocyte
928 transfusion in high-risk acute myeloid leukemia and myelodysplastic
929 syndrome. J Clin Oncol. 2005;23:5675–87.
930 62. Canani RB, Costanzo M Di, Leone L, Pedata M, Meli R, Calignano A.
931 Potential beneficial effects of butyrate in intestinal and extraintestinal
932 diseases. World J Gastroenterol. 2011;17:1519–28.
933 63. Zwielehner J, Lassl C, Hippe B, et al. Changes in human fecal
934 microbiota due to chemotherapy analyzed by TaqMan-PCR, 454
935 sequencing and PCR-DGGE fingerprinting. PLoS One. 2011;6:e28654.
936 64. Montassier E, Gastinne T, Vangay P, et al. Chemotherapy-driven
937 dysbiosis in the intestinal microbiome. Aliment Pharmacol Ther.
938 2015;42:515–28.
939 65. Peng L, Li Z, Green RS, Holzman IR, Lin J. Butyrate enhances the
940 intestinal barrier by facilitating tight junction assembly via activation of
941 AMP-activated protein kinase. J Nutr. 2009;139:1619–25.
942 66. Mathewson ND, Jenq R, Mathew A V, et al. Gut microbiome–derived
943 metabolites modulate intestinal epithelial cell damage and mitigate

944 graft-versus-host disease. Nat Immunol. 2016;17:505–13.
945 67. Docampo MD, Auletta JJ, Jenq RR. The emerging influence of the
946 intestinal microbiota during allogeneic hematopoietic cell
947 transplantation: Control the gut and the body will follow. Biol Blood
948 Marrow Transplant. 2015;21:1360–6.
949 68. Biagi E, Zama D, Nastasi C, et al. Gut microbiota trajectory in pediatric
950 patients undergoing hematopoietic SCT. Bone Marrow Transplant.
951 2015;50:992–8.
952 69. Abreu MT, Peek RM. Gastrointestinal malignancy and the microbiome.
953 Gastroenterology. 2014;146:1534–1546.e3.
954 70. Perez-Chanona E, Jobin C. From promotion to management: the wide
955 impact of bacteria on cancer and its treatment. Bioessays.
956 2014;36:658–64.
957 71. Jiang W, Wu N, Wang X, et al. Dysbiosis gut microbiota associated with
958 inflammation and impaired mucosal immune function in intestine of
959 humans with non-alcoholic fatty liver disease. Sci Rep. 2015;5:8096.
960 72. Ubeda C, Bucci V, Caballero S, et al. Intestinal microbiota containing
961 *Barnesiella* species cures vancomycin-resistant *Enterococcus faecium*
962 colonization. Infect Immun. 2013;81:965–73.
963 73. Shono Y, Docampo MD, Peled JU, et al. Increased GVHD-related
964 mortality with broad-spectrum antibiotic use after allogeneic
965 hematopoietic stem cell transplantation in human patients and mice. Sci
966 Transl Med. 2016;8:339ra71.
967 74. Brandl K, Plitas G, Mihu CN, et al. Vancomycin-resistant enterococci
968 exploit antibiotic-induced innate immune deficits. Nature.

2008;455:804–7.

75. Ubeda C, Pamer EG. Antibiotics, microbiota, and immune defense. Trends Immunol. 2012;33:459–66.

76. Khoruts A, Hippen KL, Lemire AM, Holtan SG, Knights D, Young J-AH. Toward revision of antimicrobial therapies in hematopoietic stem cell transplantation: target the pathogens, but protect the indigenous microbiota. Transl Res. 2016;1–10.

77. Wexler HM. *Bacteroides*: The good, the bad, and the nitty-gritty. Clin Microbiol Rev. 2007;20:593–621.

78. 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.

79. Nakano V, Ignacio A, Fernandes MR, Fukugaiti MH, Avila-campos MJ. Intestinal *Bacteroides* and *Parabacteroides* species producing antagonistic substances. Curr Trends Microbiol. 2006;1.

80. Avelar KES, Pinto LJF, Antunes LCM, et al. Production of bacteriocin by *Bacteroides fragilis* and partial characterization. Lett Appl Microbiol. 1999;29:264–8.

81. 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.

82. Jenq RR, Ubeda C, Taur Y, et al. Regulation of intestinal inflammation by microbiota following allogeneic bone marrow transplantation. J Exp Med. 2012;209:903–11.

994 83. Borriello SP, Ammes WP, Holzapfel W, et al. Safety of probiotics that
995 contain lactobacilli or bifidobacteria. 2003;36:775–80.
996 84. Servin AL. Antagonistic activities of lactobacilli and bifidobacteria
997 against microbial pathogens. FEMS Microbiol Rev. 2004;28:405–40.
998 85. Spinler JK, Taweechotipatr M, Rognerud CL, Ou CN, Tumwasorn S,
999 Versalovic J. Human-derived probiotic *Lactobacillus reuteri* demonstrate
1000 antimicrobial activities targeting diverse enteric bacterial pathogens.
1001 Anaerobe. 2008;14:166–71.
1002 86. Ladas EJ, Bhatia M, Chen L, et al. The safety and feasibility of
1003 probiotics in children and adolescents undergoing hematopoietic cell
1004 transplantation. Bone Marrow Transplant. 2016;51:262–6.
1005 87. Cohen SA, Woodfield MC, Boyle N, Stednick Z, Boeckh M, Pergam SA.
1006 Incidence and outcomes of bloodstream infections among
1007 hematopoietic cell transplant recipients from species commonly
1008 reported to be in over-the-counter probiotic formulations. Transpl Infect
1009 Dis. 2016;699–705.
1010 88. Webber MA, Piddock LJ V. The importance of efflux pumps in bacterial
1011 antibiotic resistance. J Antimicrob Chemother. 2003;51:9–11.
1012 89. Varrette S, Bouvry P, Cartiaux H, Georgatos F. Management of an
1013 academic HPC cluster: The UL experience. Proc 2014 Int Conf High
1014 Perform Comput Simulation, HPCS 2014. 2014;959–67.
1015

Figure legends

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 summarized 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 C) Violin plots representing distribution of depth of coverage of the contigs contained in each population-level genome. (B and D) Venn diagrams indicating the number of variant positions exclusive to each sample respectively the number of variant positions found in both samples. (E and F) 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. (G and H) 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).

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.

Appendix

Figure S1. 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.

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 day 66, 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 day 275, 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 day 268, 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 day 212, 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 day 184, GvHD
A19	f	58	myeloma	MUD	Treo/Flu	F, M, P-T	-	deceased day 39, 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 Glucksberg et al.[18]

[f]: Bold: aGvHD with summed stages ≥ 4 considered as severe aGvHD

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	Erm23SRibosomalRNAMethyltransferase	rRNA 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

Figure 1

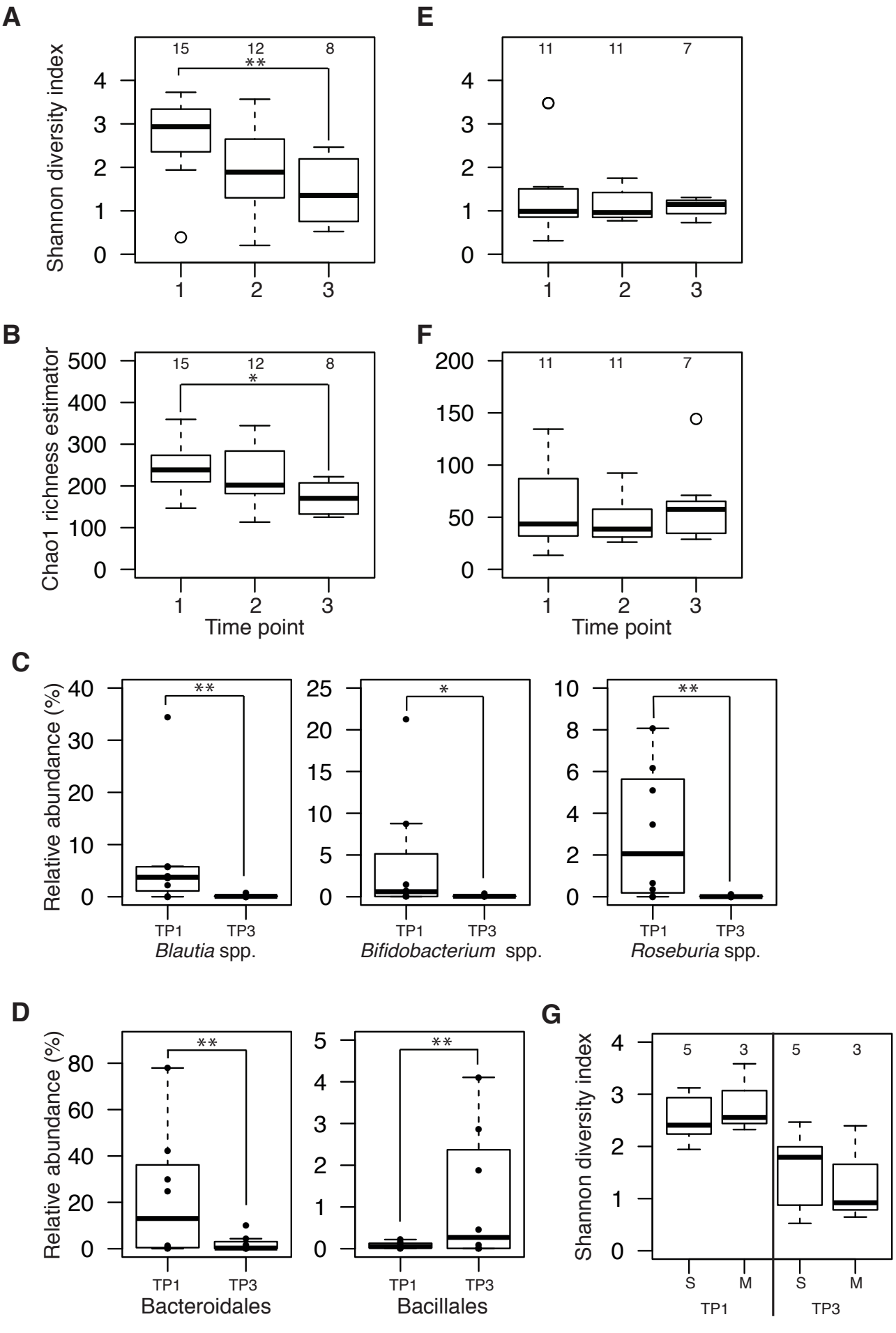

A
Shannon diversity index
15
12
8
**
0
1
2
3
4
1
2
3
E
11
11
7
0
1
2
3
4
1
2
3
B
Chao1 richness estimator
15
12
8
*
0
100
200
300
400
500
1
2
3
Time point
F
11
11
7
0
50
100
150
200
1
2
3
Time point
C
Relative abundance (%)
**
0
10
20
30
40
TP1
TP3
Blautia spp.
*
0
5
10
15
20
25
TP1
TP3
Bifidobacterium spp.
**
0
2
4
6
8
10
TP1
TP3
Roseburia spp.
D
Relative abundance (%)
**
0
20
40
60
80
TP1
TP3
Bacteroidales
**
0
1
2
3
4
5
TP1
TP3
Bacillales
G
Shannon diversity index
5
3
5
3
0
1
2
3
4
S
M
S
M
TP1
TP3

Figure 2

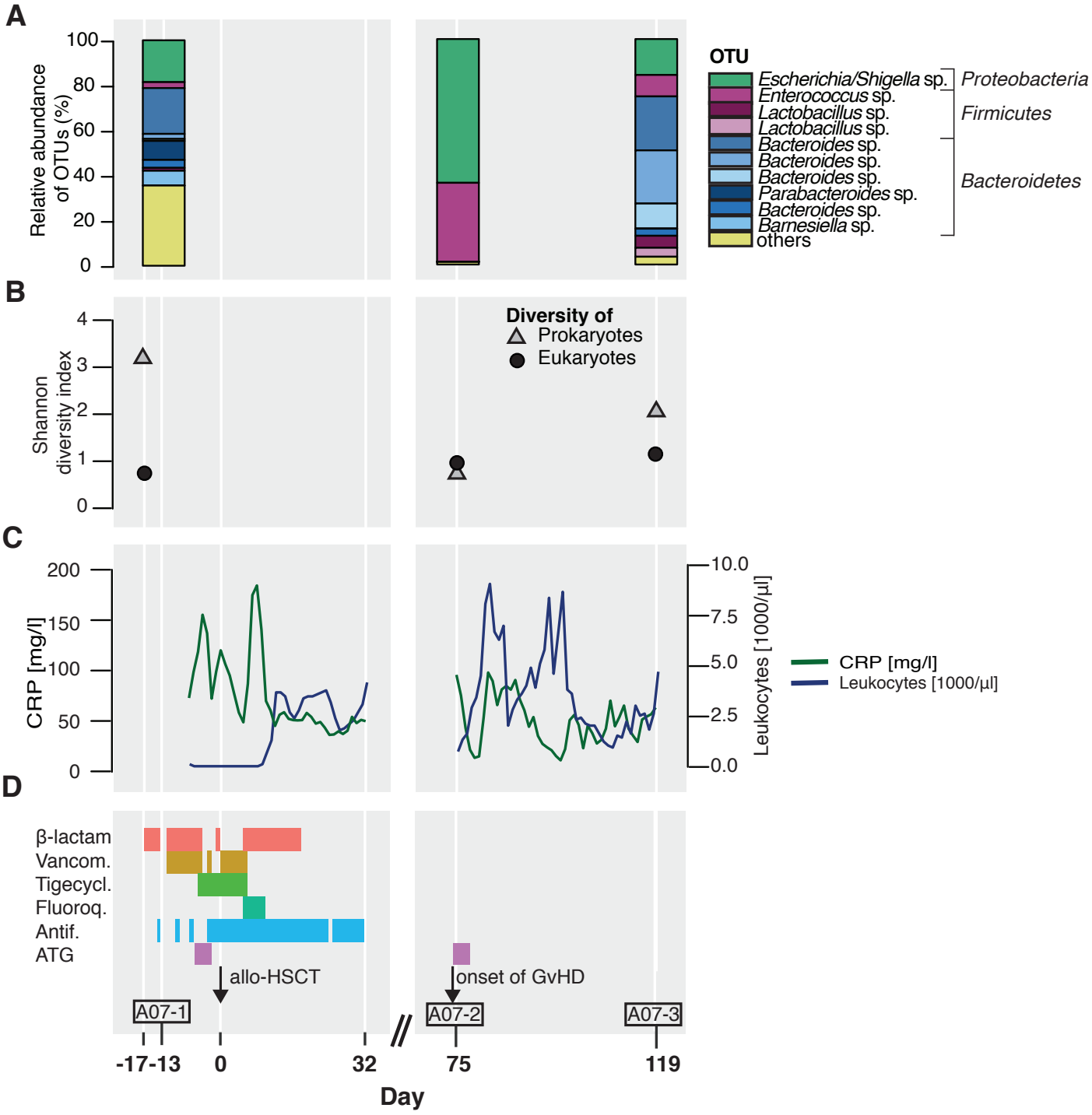
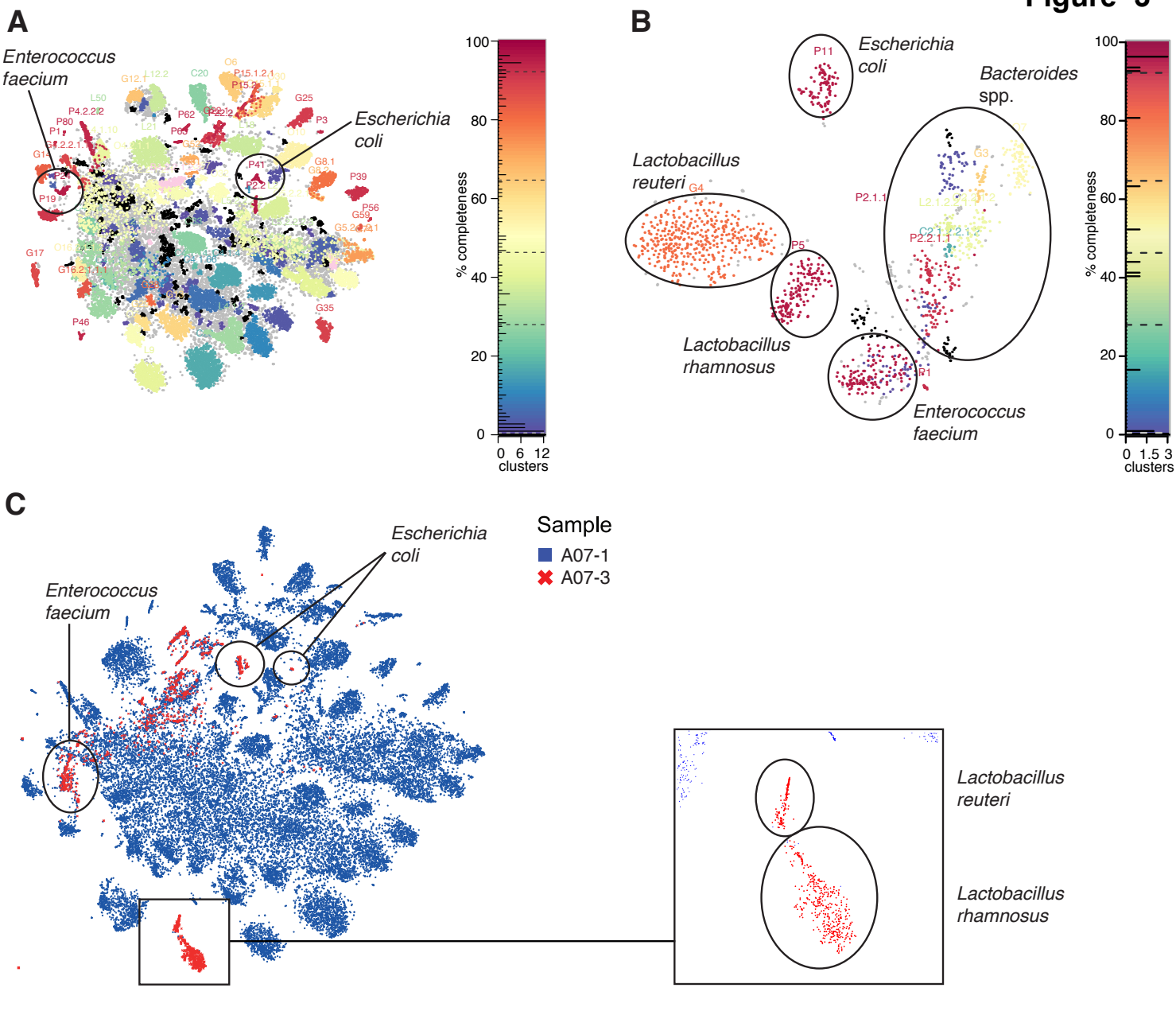

A
Relative abundance of OTUs (%)
100
80
60
40
20
0
OTU
Escherichia/Shigella sp.
Enterococcus sp.
Lactobacillus sp.
Lactobacillus sp.
Bacteroides sp.
Bacteroides sp.
Bacteroides sp.
Parabacteroides sp.
Bacteroides sp.
Barnesiella sp.
others
Proteobacteria
Firmicutes
Bacteroidetes
B
Shannon diversity index
4
3
2
1
0
Diversity of
Prokaryotes
Eukaryotes
C
CRP [mg/l]
200
150
100
50
0
Leukocytes [1000/µl]
10.0
7.5
5.0
2.5
0.0
CRP [mg/l]
Leukocytes [1000/µl]
D
β-lactam
Vancom.
Tigecycl.
Fluoroq.
Antif.
ATG
allo-HSCT
onset of GvHD
A07-1
A07-2
A07-3
-17
-13
0
32
75
119
Day

Figure 3


A
Enterococcus faecium
Escherichia coli
% completeness
clusters
B
Escherichia coli
Bacteroides spp.
Lactobacillus reuteri
Lactobacillus rhamnosus
Enterococcus faecium
% completeness
clusters
C
Escherichia coli
Enterococcus faecium
Sample
A07-1
A07-3
Lactobacillus reuteri
Lactobacillus rhamnosus

Escherichia coli

Enterococcus faecium

Figure 4

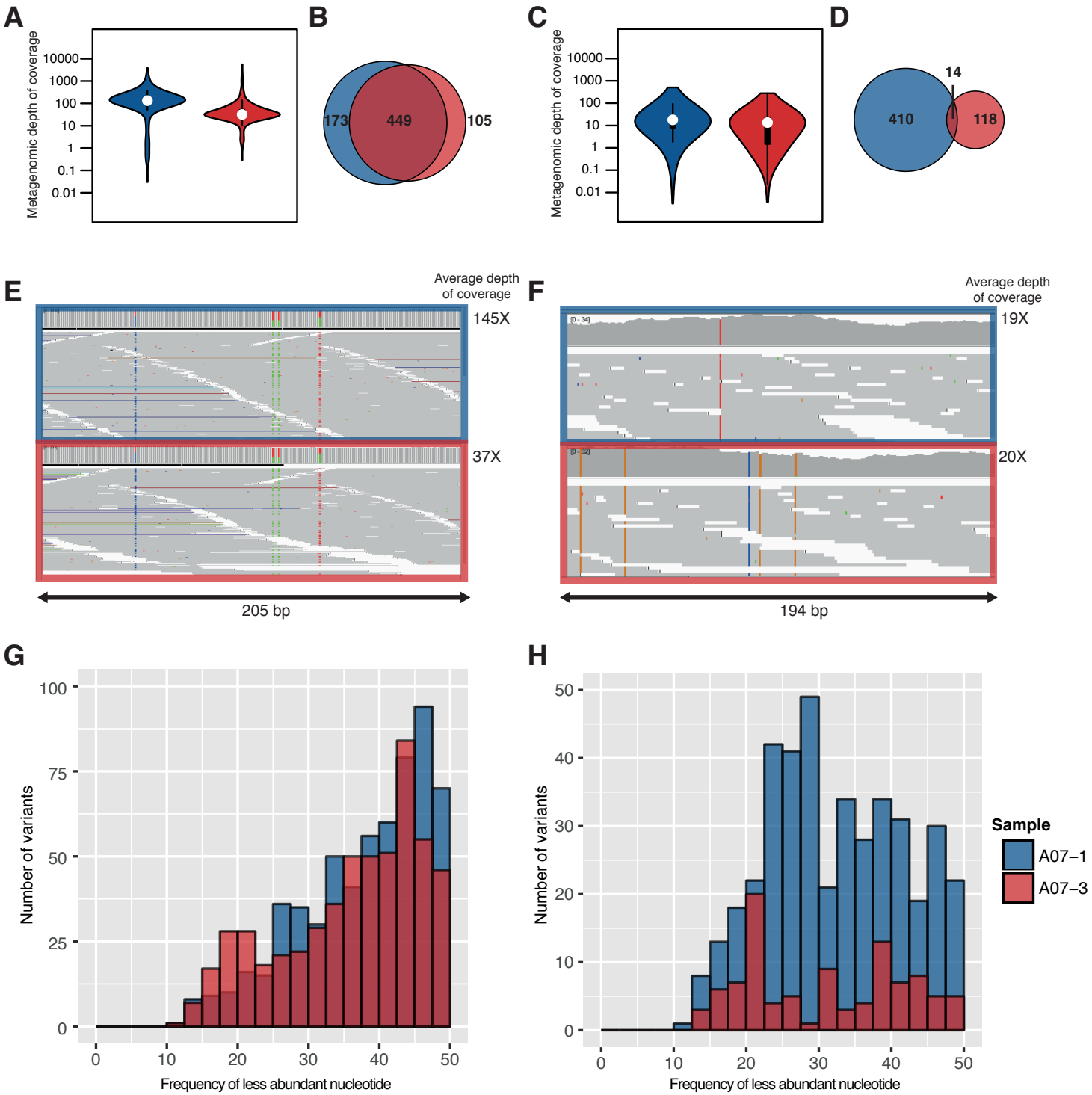


Figure 5

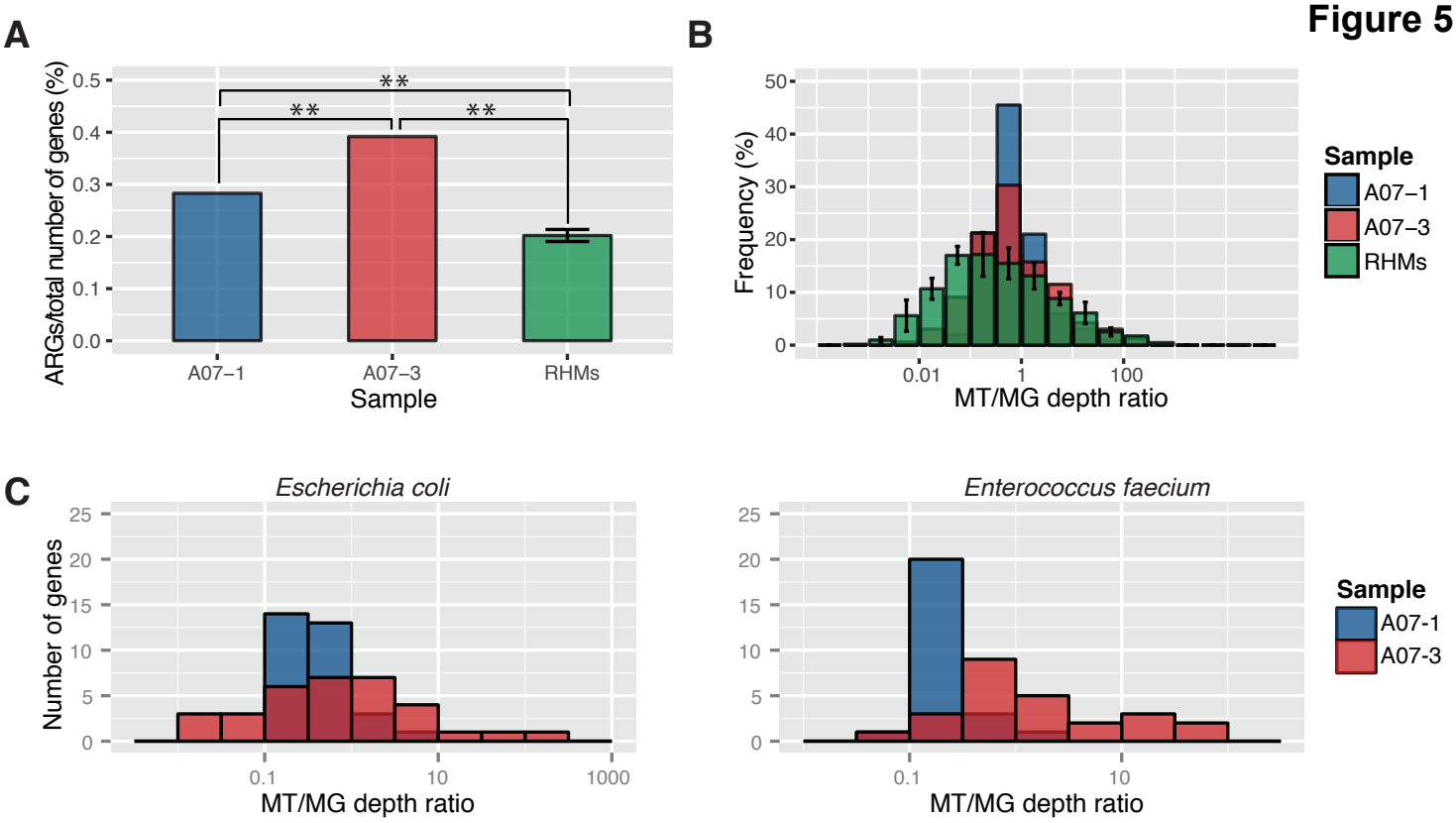
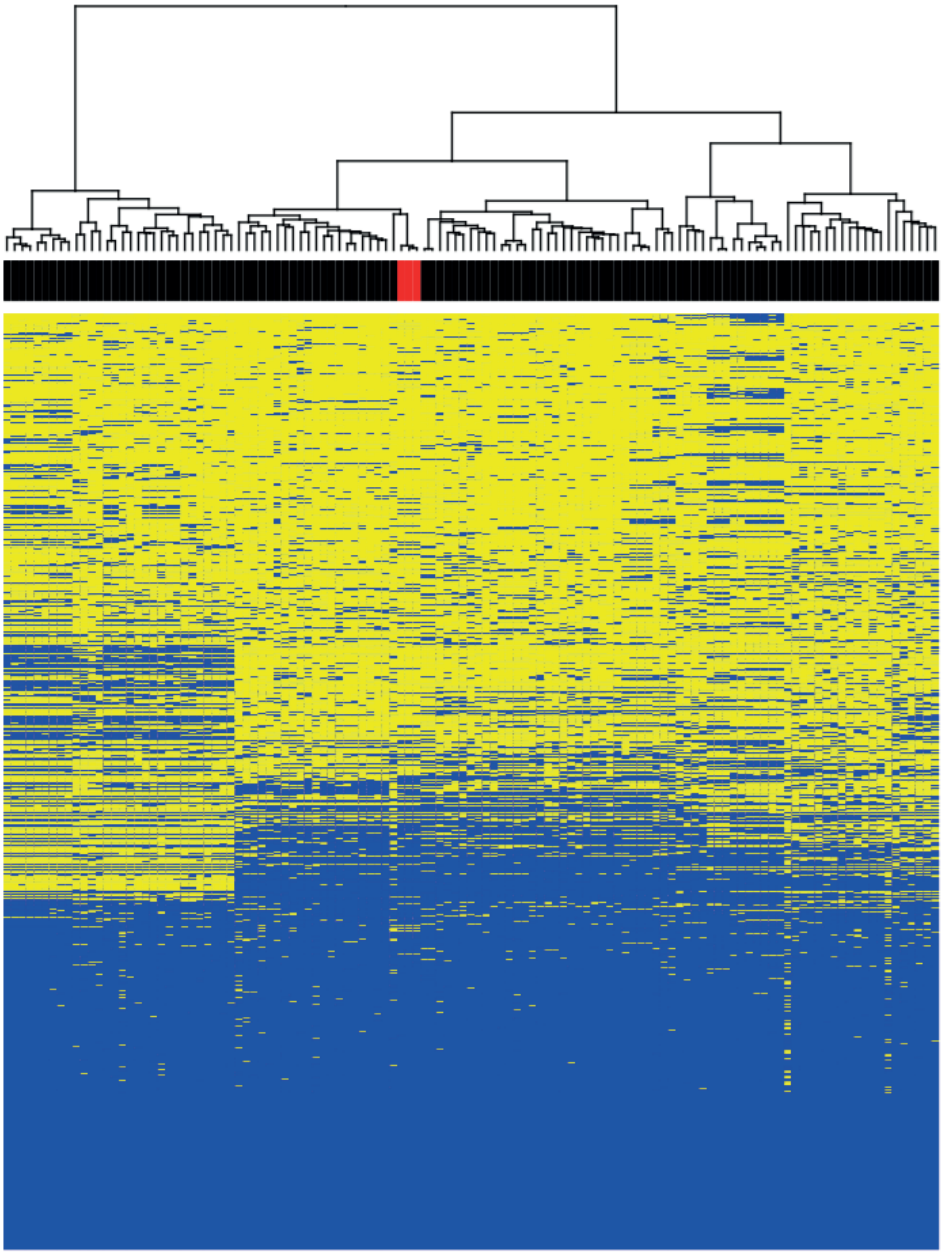

A
ARGs/total number of genes (%)
0.5
0.4
0.3
0.2
0.1
0.0
**
**
**
A07−1
A07−3
RHMs
Sample
B
Frequency (%)
50
40
30
20
10
0
0.01
1
100
MT/MG depth ratio
Sample
A07−1
A07−3
RHMs
C
Escherichia coli
Number of genes
25
20
15
10
5
0
0.1
10
1000
MT/MG depth ratio
Enterococcus faecium
25
20
15
10
5
0
0.1
10
MT/MG depth ratio
Sample
A07-1
A07-3

Figure S1



E. coli isolated from patient A07

Gene

absent

present

Appendix 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(1)

Contributions of author include:

- Software testing
- Participation in discussions
- Writing and revision of manuscript

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

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

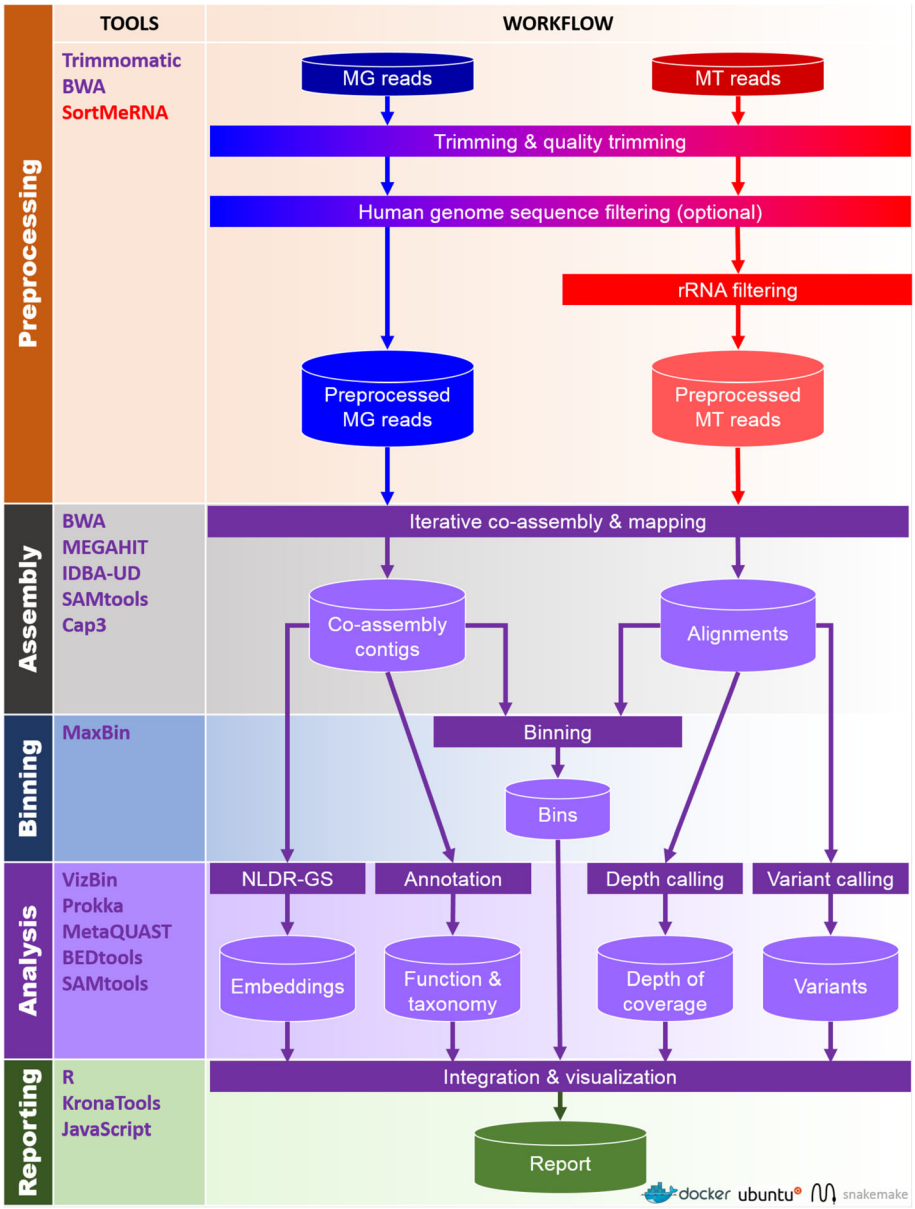


Fig. 1 Schematic overview of the IMP pipeline. *Cylinders* represent input and output while *rectangles* represent processes. *Arrows* indicate the flow between input, processes, and output. *MG* — Metagenomic data, *MT* — Metatranscriptomic data, *rRNA* — ribosomal RNA, *NLDR-GS* — genomic signature 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

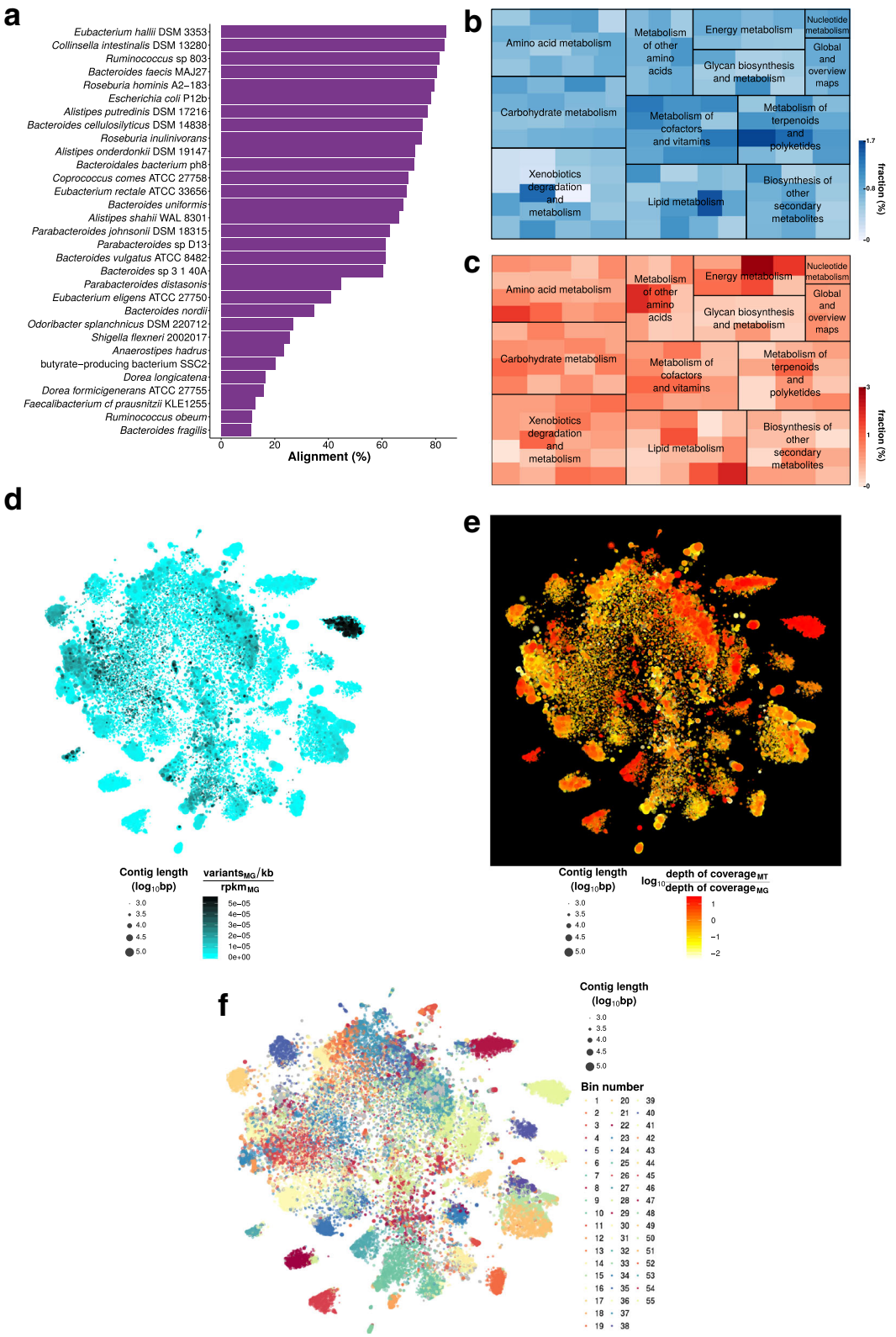


Fig. 2 (See legend on next page.)

(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

		MG iterative assembly				MT iterative assembly			
Dataset	Iteration	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

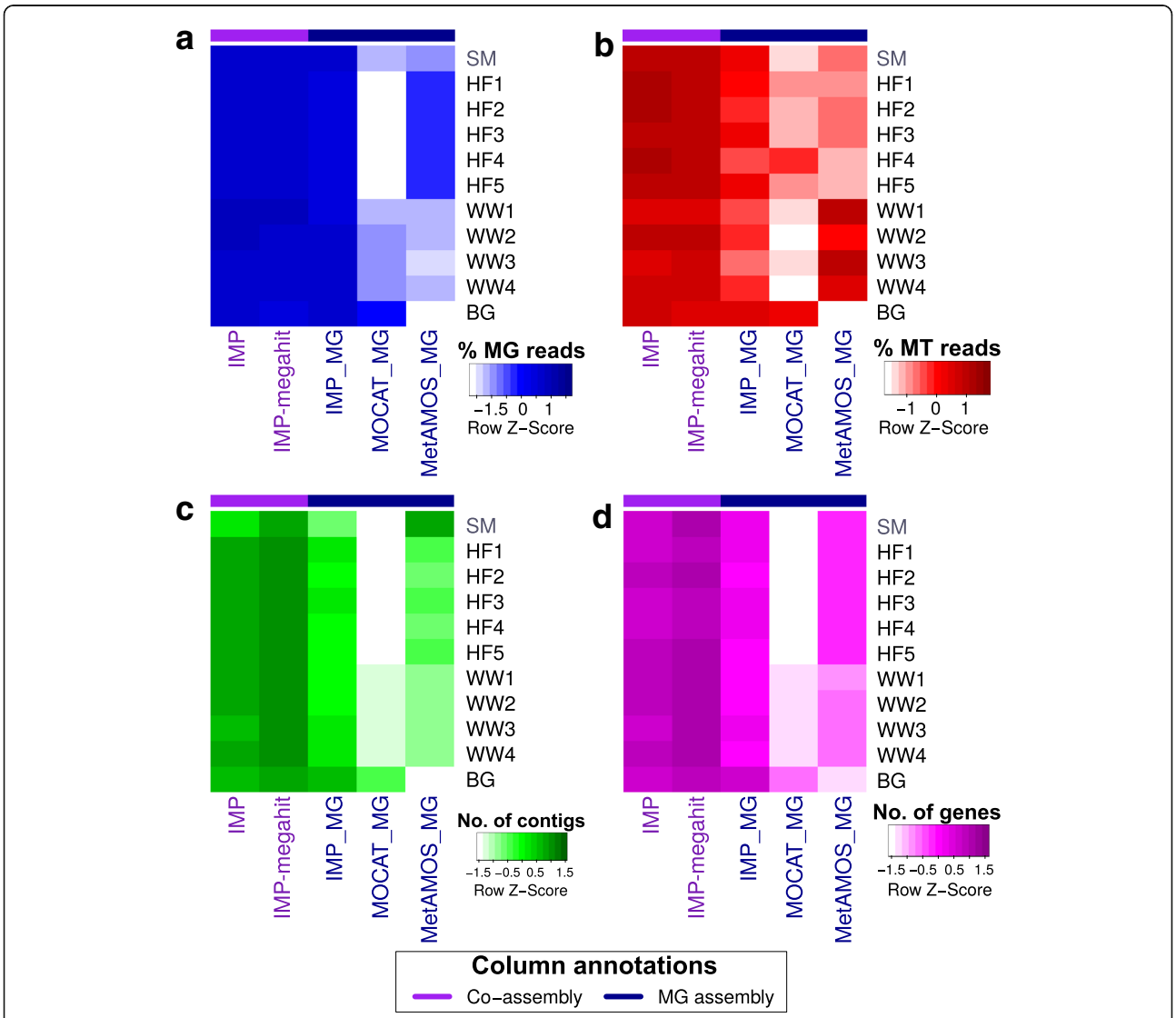


Fig. 3 Assessment of data usage and output generated from co-assemblies compared to single-omic assemblies. Heat maps show (**a**) fractions of properly mapped MG read pairs, (**b**) fractions of properly mapped MT read pairs, (**c**) numbers of contigs ≥1 kb, and (**d**) numbers of unique predicted 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).

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

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].

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

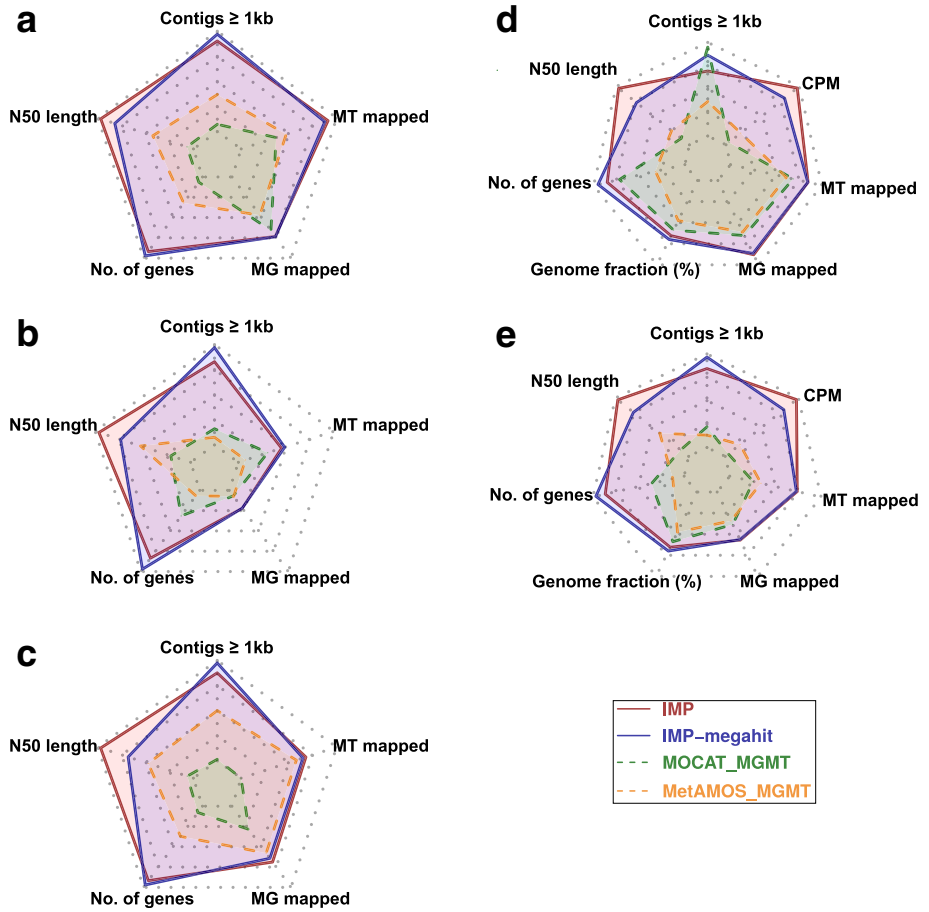


Fig. 4 Assessment of the IMP-based iterative co-assemblies in comparison to MOCAT- and MetAMOS-based co-assemblies. Radar charts summarizing the characteristics of the co-assemblies generated using IMP, MetAMOS, and MOCAT pipelines on: **a** human fecal microbiome, **b** wastewater sludge community, **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

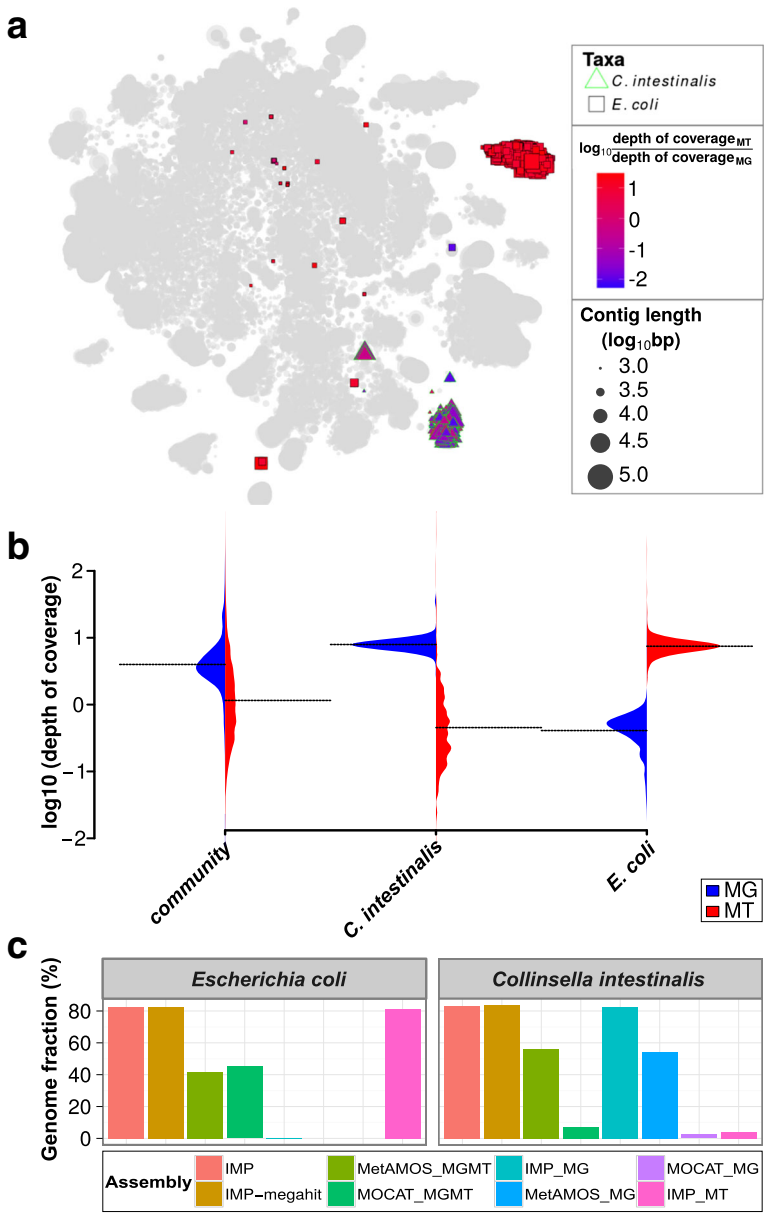


Fig. 5 Metagenomic and metatranscriptomic data integration of a human fecal microbiome. **a** Augmented VizBin map highlighting contig subsets with sequences that are most similar to *Escherichia coli* P12b and *Collinsella intestinalis* DSM 13280 genomes. **b** Beanplots representing the densities of 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. RAlphy: 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 BBM, 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.