PhD-FSTC-2016-36

Faculty of Life Sciences, Technology and Communication

DISSERTATION

Defense held on 01/09/ 2016 in Luxembourg

to obtain the degree of

DOCTEUR DE L 'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Maria Irene PIRES PACHECO

Born on 22^{th} of December 1979 in Luxembourg

FAST RECONSTRUCTION OF COMPACT CONTEXT-SPECIFIC METABOLIC NETWORK MODELS

Dissertation Defence Committee:

Dr. Thomas Sauter, dissertation supervisor

Professor, Université du Luxembourg

Dr.-Ing. Steffen Klamt

Group leader, Max Planck Institute-Magdeburg, Magdeburg

Dr. Jean-Luc Bueb, Chairman

Professor, Université du Luxembourg

Dr. Elmar Heinzle

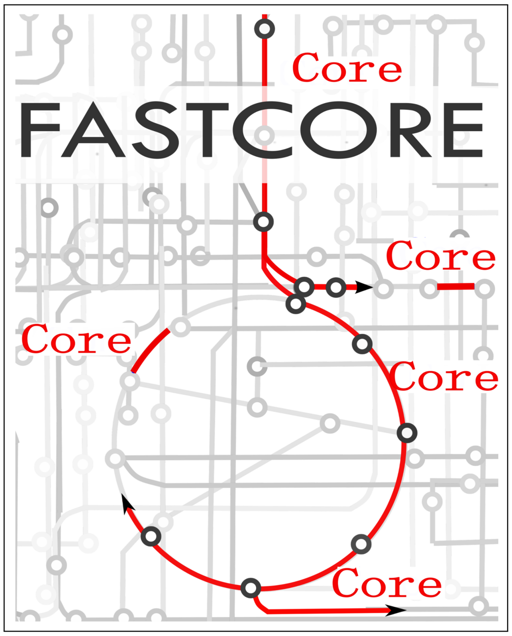
Professor, Université du Saarland

Dr. Francisco Azuaje, Vice Chairman

Principal Investigator, Luxembourg Institute of Health

UNIVERSITÉ DU LUXEMBOURG

Systems biology group of Professor Thomas Sauter at the University of Luxembourg



Dissertation Defence Committee:

Professor Jean-Luc Bueb, University of Luxemburg, Esch-Alzette, Luxembourg

Dr.-Ing. Steffen Klamt, Max Planck Institute-Magdeburg, Magdeburg, Germany

Dr. Francisco Azuaje, Luxembourg Institute of Health, Luxembourg, Luxembourg

Professor Elmar Heinzle, University of Saarland, Saarbrücken, Germany

Supervisor:

Professor Thomas Sauter, University of Luxemburg, Luxembourg

Affidavit

I hereby confirm that the PhD thesis entitled FAST RECONSTRUCTION OF COMPACT CONTEXT - SPECIFIC METABOLIC NETWORK MODELS has been written independently and without any other sources than cited.

Luxembourg,

Dedication

To my parents and to Paulo for their love, trust and support.

"Success consists of going from failure to failure without loss of enthusiasm."

Winston Churchill

Acknowledgments

The first person I would like to thank is Professor Thomas Sauter for his advises, his guidance, for taking the time to go over the scripts or concepts with me every time that one of the tools was not doing what it was supposed to do and for all his brilliant ideas. Beside of being a great researcher, Professor Sauter has the remarkable capacity to understand very complex subjects and to present them in a clear, well-structured and simple form which does greatly facilitate the communication between biologists and people with a more mathematical or computational background, which was greatly beneficial for my project. Every discussion with him made me re-evaluate all my assumptions and certitudes, which allowed me to spot and correct flaws. But mainly, I want to thank him for his infinite patience and calm, his positiveness, his kindness, his humbleness and his humanity.

The second person I would like to express my gratitude is Doctor Nikos Vlassis who had the fantastic ideas that allowed the implementation of FASTCORE. I admire him for his creativity, intelligence and his precise mathematical mind. But even more, I want to acknowledge him for his incredible enthusiasm (nothing seemed impossible), his kindness, his modesty and his enormous efforts to make himself understandable by a biologist.

The third person I would like also to acknowledge is Professor Bueb, for having been my personal encyclopaedia on everything related to pure biology. His comments were very precious to understand the subtleties of omics technology and therefore to understand how to integrate this type of data into metabolic models. I admire his open-mindedness towards the weird mathematical concepts, we were bombarding him with in every progress report, and for his great sense of humour and his kindness.

Further, I wish to thank Doctor Elisabeth John for having done all the wet lab experiments for the FASTCORMICS's paper and Doctor Lasse Sinkkonen, the epigenetic guru of the group, for his help in the integration of epigenetic data in metabolic models and all my colleagues and

friends that contributed to the papers included in this thesis. Also I wish to acknowledge all the members of the systems biology group and all the people working in the Life Science Research Unit (LSRU) for the friendly atmosphere and the interesting and stimulating discussions. Further, I want to thank the AFR - Fond National de la Recherche Luxembourg - for its financial support (AFR 6041230) and the jury for reviewing this thesis.

Finally I want to express my gratitude to my family, José Manuel, Albertina and Paulo for their love and unwavering support.

List of abbreviations

ACHR	Artificial Centering Hit-and-Run algorithm
ASL	Arginosuccinate lyase
ASS	Arginosuccinate synthethase
BCAT1	Branched chain amino-acid transaminase 1
BQ	Bio qualifiers
C13	Carbon 13 labelling
CBM	Constraint-based modelling
cDNA	complementary DNA
CH	Hard core reactions
ChIP-seq	Chromatin immunoprecipitation
CL	Light core reactions
CM	Medium core reactions
CS	Chondroitin sulphates
CTX	Cerebrotendinous xanthomatosis
DG	Dentale gyrus
EFM	Elementary flux mode
EMU	Elementary metabolite units
ER	Estrogen receptor
ExPA	Extreme Pathway
FASTCC	Fast Consistency check
FASTCORE	Fast reconstruction of compact context-specific algorithm
FASTCORMICS	Fast reconstruction of compact context-specific algorithm via the integration of omics data

FBA	Flux Balance Analysis
FDR	False discovery rate
fRMA	frozen Robust multi-array average
FVA	Flux Variability Analysis
FPKM	Fragments per kilobase of exon per million fragments mapped
GEM	Genome-scale models
GENRE	Genome-scale reconstructions
GEO	Gene Expression Omnibus
GIMME	Gene Inactivity Moderated by Metabolism and Expression
GO	Gene Ontology
GPR	Gene Protein Reaction Rules
GR	Glucocorticoid receptor
H3K27ac	Histone H3 K27 acetylation
HepatoNet	a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology
HMR	Human Metabolic Reconstruction
HPA	Human Protein Atlas
HS	Heparan Sulphates
HR	Hit-rejection approach
IEM	Inborn errors of metabolism
ILP	Integer Linear Programming
iMAT	Intiegrative metabolic analysis tool
INIT	Integrative Network Inference for Tissues
KEGG	Kyoto Encyclopedia of Genes and
KYNU	Kynurininase
KO	Knock-out
KS-test	Kolmogorov-Smirnov test
LXR	Liver X receptors
LP	Linear Programming
LPS	Lipopolysaccharide

MBA	Model Building Algorithm
mCADRE	metabolic Context-specificity Assessed by Deterministic Reaction Evaluation
MILP	Mixed Integer Linear Programming
MPA	Metabolic Phenotypic Analysis
mRNA	messenger RNA
MOMA	Minimization of metabolic adjustment
NAFLD	Non-alcoholic fatty liver disease
ODE	Ordinary differentially equation
OH	Olfactory Habitual test
ORF	Open Reading Frame
OTC	Ornithine transcarbanylase
PBS	Phosphate buffered saline buffer
PCA	Principal component analysis
PRIME	Personalized ReconstructIon of Metabolic models
Recon1	Reconstruction1
REcon2	Reconstruction2
ReconX	Recon1 and Recon2
RegrEx	Regularized Context-specific model Extraction method
RH	Reaction high expressed
RL	Reaction low expressed
RMA	Robust multi-array average
RMF	Required Metabolic Function
RNA-seq	RNA sequencing
RPKM	Reads per kilobase of exon per million fragments mapped
rRNA	ribosomial RNA
SNP	Single Nucleotide Polymorphism
TF	Transcription factor
UGCG	UDP glucose ceramide glucosyltransferase

Summary

Recent progress in high-throughput data acquisition has shifted the focus from data generation to the processing and understanding of now easily collected patient-specific information. Metabolic models, which have already proven to be very powerful for the integration and analysis of such data sets, might be successfully applied in precision medicine in the near future. Context-specific reconstructions extracted from generic genome-scale models like Reconstruction X (ReconX) (Duarte et al., 2007; Thiele et al., 2013) or Human Metabolic Reconstruction (HMR) (Agren et al., 2012; Mardinoglu et al., 2014a) thereby have the potential to become a diagnostic and treatment tool tailored to the analysis of specific groups of individuals. The use of computational algorithms as a tool for the routinely diagnosis and analysis of metabolic diseases requires a high level of predictive power, robustness and sensitivity. Although multiple context-specific reconstruction algorithms were published in the last ten years, only a fraction of them is suitable for model building based on human high-throughput data. Beside other reasons, this might be due to problems arising from the limitation to only one metabolic target function or arbitrary thresholding.

The aim of this thesis was to create a family of robust and fast algorithms for the building of context-specific models that could be used for the integration of different types of omics data and which should be sensitive enough to be used in the framework of precision medicine. FASTCORE (Vlassis et al., 2014) (Chapter 3), which was developed in the frame of this thesis is among the first context-specific building algorithms that do not optimize for a biological function and that has a computational time around seconds. Furthermore, FASTCORE is devoid of heuristic parameter settings. FASTCORE requires as input a set of reactions that are known to be active in the context of interest (core reactions) and a genome-scale reconstruction. FASTCORE uses an approximation of the cardinality function to force the core set of reactions to carry a flux above a threshold. Then an L1-minimization is applied to penalize the activation of

reactions with low confidence level while still constraining the set of core reactions to carry a flux. The rationale behind FASTCORE is to reconstruct a compact consistent (all the reactions of the model have the potential to carry non zero-flux) output model that contains all the core reactions and a small number of non-core reactions.

Then, in order to cope with the non-negligible amount of noise that impede direct comparison within genes, FASTCORE was extended to the FASTCORMICS workflow (Pires Pacheco and Sauter, 2014; Pires Pacheco et al., 2015a) for the building of models via the integration of microarray data (Chapter 4). FASTCORMICS was applied to reveal control points regulated by genes under high regulatory load in the metabolic network of monocyte derived macrophages (Pires Pacheco et al., 2015a) (Chapter 5) and to investigate the effect of the TRIM32 mutation on the metabolism of brain cells of mice (Hillje et al., 2013) (Appendix Chapter A).

The use of metabolic modelling in the frame of personalized medicine, high-throughput data analysis and integration of omics data calls for a significant improvement in quality of existing algorithms and generic metabolic reconstructions used as input for the former. To this aim and to initiate a discussion in the community on how to improve the quality of context-specific reconstruction, benchmarking procedures were proposed and applied to seven recent context-specific algorithms including FASTCORE and FASTCORMICS (Pires Pacheco et al., 2015a) (Chapter 6). Further, the problems arising from a lack of standardization of building and annotation pipelines and the use of non-specific identifiers was discussed in the frame of a review. In this review, we also advocated for a switch from gene-centred protein rules (GPR rules) to transcript-centred protein rules (Pfau et al., 2015) (Appendix Chapter B).

List of Publications

Parts of this thesis were previously published in the articles listed below. The contribution of the different authors is stated in the original articles or/and in the Summary and contributions sections.

- N Vlassis*, **MP Pacheco***, T Sauter (2014), Fast reconstruction of compact context-specific metabolic network models, PLoS Computational Biology 10(1): e1003424.
  doi:10.1371/journal.pcbi.1003424 (Chapter 3)
  (*equal contribution)

- **MP Pacheco***, E John*, T Kaoma, M Heinäniemi, N Nicot, L Vallar, JL Bueb, L Sinkkonen,T Sauter (2015, Integrated metabolic modelling reveals cell-type specific epigenetic control points of the macrophage metabolic network, BMC Genomics 16 (1), 809
  doi: 10.1186/s12864-015-1984-4 (Chapter 4 and 5)
  (*equal contribution)

- **MP Pacheco**, T Pfau, T Sauter, (2016), Benchmarking procedures for high-throughput context specific reconstruction algorithms, Frontiers in Physiology.
  doi: 10.3389/fphys.2015.00410 (Chapter 6)

- AL Hillje, E Beckmann, MAS Pavlou, C Jaeger, **MP Pacheco**, T Sauter, JC Schwamborn, L Lewejohann, (2015) The neural stem cell fate determinant TRIM32 regulates complex behavioral traits, Frontiers in Cellular Neuroscience. 2015;9:75.
  doi:10.3389/fncel.2015.00075. (Appendix Chapter A)

- T Pfau*, **MP Pacheco***, T Sauter, (2015) Towards improved genome-scale metabolic network reconstructions: unification, transcript specificity and beyond, Briefings in Bioinformatics
  doi: 10.1093/bib/bbv100 (Appendix Chapter B)
  (*equal contribution)

Contents

List of Figures

List of Tables

Introduction

Contents

1.1 Systems biology

The sequencing of the human genome in 2003 and the advent of high-throughput technology revolutionized biology by shifting the focus from a gene-centred to a system-wide approach. The suffix -omics, i.e. in genomics, proteomics, transcriptomics or metabolomics became very popular, as the promise of high-throughput technology was that once the complete information of a cell was available, every disease could be easily diagnosed by some rather simple analysis.

This infatuation for high-throughput data, was not shared by the Nobel laureate Professor Sydney Brenner, who qualified high-throughput technology as "factory science" and who dropped the quote: "So we now have a culture which is based on everything must be high-throughput," Professor Brenner continued: "I like to call it low-input, high-throughput, no-output biology" (Brenner, 2010).

For half a century, biologists tried to deduce key principles and underlying mechanisms governing biological processes instead of simply describing them. The advent of omics technology aroused the fear that this more theoretical trend could be reversed, that complex models or analysis were no longer necessary, explaining the aversion of Brenner for the high-throughput technologies (Brenner, 2010).

But the omics technologies did not yet hold all their promises as 13 years after the completion of the Human Genome Project, retrieving knowledge out of high-throughput data is still challenging.

Professor Brenner was proven right by the fact that data without proper integration and analysis has only a limited use. But high-throughput technologies did not mark the end of the more theoretical biology approaches. It allowed the rise of Systems biology, a branch of the systems science that appeared in the middle of the 20th century, but only really exploded after the advent of high-throughput and sequencing technologies, that allowed finally obtaining the amount of data necessary for the building of system-wide models (Kitano, 2002; Trewavas, 2006).

Systems biology, a science that Professor Brenner disapproved as much as omics data and whose fail Professor Brenner predicted, defends a holistic view of biology (probably the reason why the reductionist Brenner predicted its fail) in which cells like other complex systems exhibit emerging properties that cannot be foreseen from the study of its individual components, created in return the computational analysis, mathematical modelling and statistical tools that allow analysing system-wide data instead of focussing on a single gene or protein.

Despite the predictions of the Nobel laureate, the symbiosis between high-throughput technologies and systems biology was so efficient that it caused an explosion of the number of genome scale models ranging from unicellular organisms like *Escherichia coli* or *Saccharomyces cerevisiae* to multicellular organism models including humans or plants of interest (*Hordeum vulgare*, *Arabidopsis thaliana* or *Zea mays*). Genome-scale models have applications in agriculture, industry and medicine. Plant generic genome-scale models are commonly used for crop improvement and yield stability. Whereas unicellular genome scale models have applications that range from enhanced production of valuable bioproducts (Ranganathan et al., 2010) to the study of parasitism or pathogens i.e. *Mycobacterium tuberculosis* (Jamshidi and Palsson, 2007; Beste et al., 2007) *Staphylococcus aureus* (Becker and Palsson, 2005; Heinemann et al., 2005; Lee et al., 2009) *Helicobacter pylori* (Schilling et al., 2002; Thiele et al., 2005) and *Salmonella typhimurium* (Raghunathan et al., 2009; AbuOun et al., 2009; Thiele et al., 2011) and industrially relevant bacteria (e.g. *Escherichia coli* (Feist et al., 2007; Reed et al., 2003), *Geobacter metallireducens* (Mahadevan et al., 2006; Sun et al., 2009) and *Saccharomyces cerevisiae* (Mo et al., 2009)).

In the context of medical research, the systemic and multi-factorial nature of metabolic diseases turns them into ideal study objects for systems biologists, as they often result of minimal variations to the optimal conversion rate of enzymes due to small alterations of the genome like i.e. single nucleotide polymorphism (SNPs), epigenetic modifications in the regulation sites of metabolic genes or are the result of slightly higher concentrations of metabolites due to rich diet over decades that shift the equilibrium of reactions causing imbalances at the metabolic level. Taken alone, none of these imbalances or SNPs could cause the appearance of pathologies. The emergence of metabolic diseases can only be explained by several enzymes working slightly outside the optimal range, SNPs, and a rich diet over decades.

1.2 Constraint-based modelling (CBM)

The extreme complexity of the metabolism of a cell that involves thousands of metabolites implicated in a comparable number of reactions, can only be fully understood at a system-wide level using simplified mathematical representations called models. Traditionally metabolic models were quantitative, deterministic models, based on ordinary differential equations (ODE), which attempt to precisely describe the metabolic fluxes and transformations. Building quantitative models, in general, is rather difficult and is currently infeasible for cell-wide-models due to the lack of detailed information regarding enzyme kinetics and metabolites concentrations. Moreover, even if data on every enzyme implicated in the studied system was available, the integration of these data into the models is not straightforward. Indeed, enzyme kinetics are usually obtained through *in vitro* assays and might not always be transposed *in vivo* (Bailey, 2001). Furthermore these parameters are very sensitive to external and internal factors and vary from cell-to-cell. To cope with the requirements of cell-wide models, new strategies had to be adopted like constraint-based modelling (CBM).

Constraint-based modelling, unlike other modelling approaches, does not require precise knowledge of metabolites' concentrations or enzyme kinetics (Palsson et al., 2000). Instead CBM restrains the space of allowed solutions by eliminating flux distributions that are not permitted by the system due to some pre-defined constraints. In CBM, the chemical constraints imposed to the flux of metabolites are represented as stoichiometric matrix, where the rows and columns stand for the metabolites and the reactions, respectively, in which the metabolites are implicated. The entries in the matrix correspond to the stoichiometric coefficients. Negative and positive numbers symbolize respectively the consumption and production of a metabolite by a given reaction.

Constraints can be represented as balances. An example of equality constraints is the conservation of mass that is imposed to the system in constraint-based modelling at the assumed quasi-steady state. The principle of conservation of mass states that the total flux for every single metabolite entering the system is equal to the total flux exiting it.

This constraint can be written as:

$$S * v = 0,$$

where S represents the stoichiometric matrix and v the flux vector (Orth et al., 2010; Varma and Palsson, 1994). Others equality constraints can be imposed to the system to guarantee the conservation of energy, the redox potential or osmotic pressure (Price et al., 2004). The system can also be constrained by a set of inequality equations, called boundaries or bounds, which limit the range of values of the metabolic flux carried by a reaction due to e.g. the maximal capacity of enzymes, the thermodynamic laws (reversibility of reactions), spatial localization, metabolite sequestration, gene expression or protein level regulation (Figure 1.1).

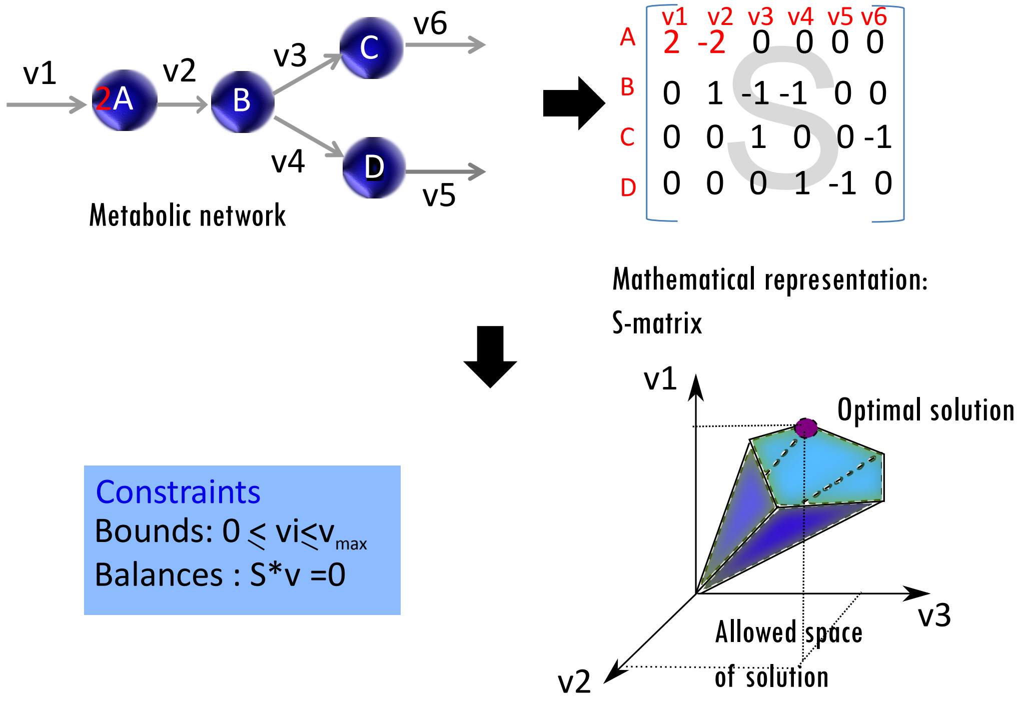


Figure 1.1. Constraint-based modelling uses constraints to define the space of solutions: Chemical reactions that compose a metabolic network can be represented as a stoichiometric matrix. Two types of constraints are imposed on the system to specify the space of solutions: 1) balances represented as equality equations e.g. the conservation of mass and 2) bounds, inequalities equations that limit the range of values the fluxes v can adopt. The system is under-determined therefore a unique solution can only be obtained when optimizing the solution space for a given task.

The use of constraints reduces the space of allowed flux distributions but does not permit to obtain a unique solution as the system is usually under-determined. A unique distribution for some fluxes can only be found when maximizing for an objective for example growth (Palsson et al., 2000).

CBM assumes quasi-steady state conditions, which impede capturing dynamic behaviours of the system (Orth et al., 2010; Varma and Palsson, 1994), but as metabolic diseases evolve over decades, the dynamics of glucose in blood after a rich diet is less relevant than the effects of high glucose levels after decades of unhealthy diet habits, at the quasi steady state.

1.2.1 Constraint-based modelling methods

Metabolic models are the inputs for a wide spectrum of CBMs that were e.g. reviewed and classified by (Price et al., 2004):

Pathway analysis The cell metabolism is a connected network. Pathways inside this network are collections of reactions that were grouped together based often on historical and arbitrary reasons that are not justified by the fulfilling of a metabolic function. Elementary flux modes (EFM) (Schuster and Hilgetag, 1994), which are the minimal sets of reactions able to carry a flux at steady state and extreme pathways (ExPa) (Schilling et al., 2000) that correspond to the edges of the solution cone, (Figure 1.1), allow defining pathways in an unbiased manner. ExPA and EFM use convex analysis to find basis vectors which carry one valid solutions to $S * v = 0$. ExPA and EFM have applications e.g. in the designing of *Escherichia coli* strains (Trinh et al., 2008) and in the analysis of global pathway regulation (Stelling et al., 2002). But, ExPA and EFM are computationally demanding and therefore are mostly applied to small subnetworks.

Random sampling Random sampling (Becker et al., 2007) uses a Monte-Carlo approach to explore the space of solutions by computing the calculations on a few random selected points. The first attempts were based on a hit-rejection approach (HR) (Wiback et al., 2004), where the flux cone was enclosed in a simpler regular shape such as a parallelepiped. Points were then randomly chosen inside this simple shape and only solutions allowed by the constraints were retained. But, the high-dimensionality of metabolic networks turns it difficult to find a shape that fits the flux cone, causing a large number of rejected solutions.

More advanced methods are based on the hit and run approach (Smith, 1984) where random points, direction and step size falling inside the flux cone are iteratively selected. But as the flux cone is elongated for reactions that are loosely constrained, the size of steps tends to remain rather small and the points are often trapped at the boundaries. Therefore, to circumvent these issues, the Artificial Centering Hit-and-Run algorithm (ACHR) (Kaufman and Smith, 1998) chooses the direction within the elongated direction, which allows larger steps but does not longer guarantee that the whole space is uniformly covered. Whereas, optGpDampler (Megchelenbrink et al., 2014) combines a warm-up points strategy, (already used by some of its predecessors), in which the flux is minimized and maximized with random sampling but the directionality is not systematically fixed at each iteration. Furthermore, the whole process is run in parallel on several cores to speed up the random sampling process. Unlike ExPA and EFM, random sampling can be applied to the whole network.

Flux Balance Analysis Flux Balance Analysis (FBA) (Edwards et al., 2002)) allows selecting among all the possible flux distributions, the optimal distribution for a given context or purpose like for example the maximization of growth. For some networks multiple flux distributions, called alternative optimal flux distributions can be found for the same optimization framework. Alternative optima can then be explored using Flux Variability Analysis (FVA) (Mahadevan and Schilling, 2003).

Knock-out experiment Mutations events can be simulated *in silico* by setting the bounds of the target reactions of the mutated genes to 0. The effect on growth is then evaluated by FBA while maximization for the growth. Knock-out experiments allow moreover identifying essential genes (Metabolic Metabolite Essentiality Analysis) or potential new drug targets.

Minimal adjustment After a mutation event, in order to survive, the flux distribution has to be rearranged to recover the lost metabolic function(s). Algorithms like MOMA (Segre et al., 2002) and ROOM (Shlomi et al., 2005) compute the optimal solution using FBA for the wild type phenotype then find the nearest solution after a perturbation of the network.

Methods based on thermodynamics constraints The definition of reversibility are based on *in vitro* assays that might not reflect the reversibility found *in vivo*. Moreover when the information is missing, reactions are considered by default to be reversible. The problem with this approach is that reactions wrongly defined as reversible might cause the formation of loops.

FBA cannot find a finite solution for these cases and therefore metabolites are infinitely cycling through these loops. Looping flux violates a law similar to Kirchhoff's second law for electrical circuits and therefore a loop should carry a net-zero flux. Approaches based on extreme pathways and network-based pathways analyses can be applied to identify loops.

1.3 Genome-scale models

Constraint-based modelling allows building Genome-scale models (GEMs) that are obtained after the mapping of annotated sequenced transcripts corresponding to the open reading frames (ORF) of an organism to known biochemical reactions retrieved from databases such as KEGG (Kanehisa et al., 2010), Reactome (Joshi-Tope et al., 2005), MetaCyc (Caspi et al., 2014) or BRENDA (Schomburg et al., 2013) (Figure 1.2).

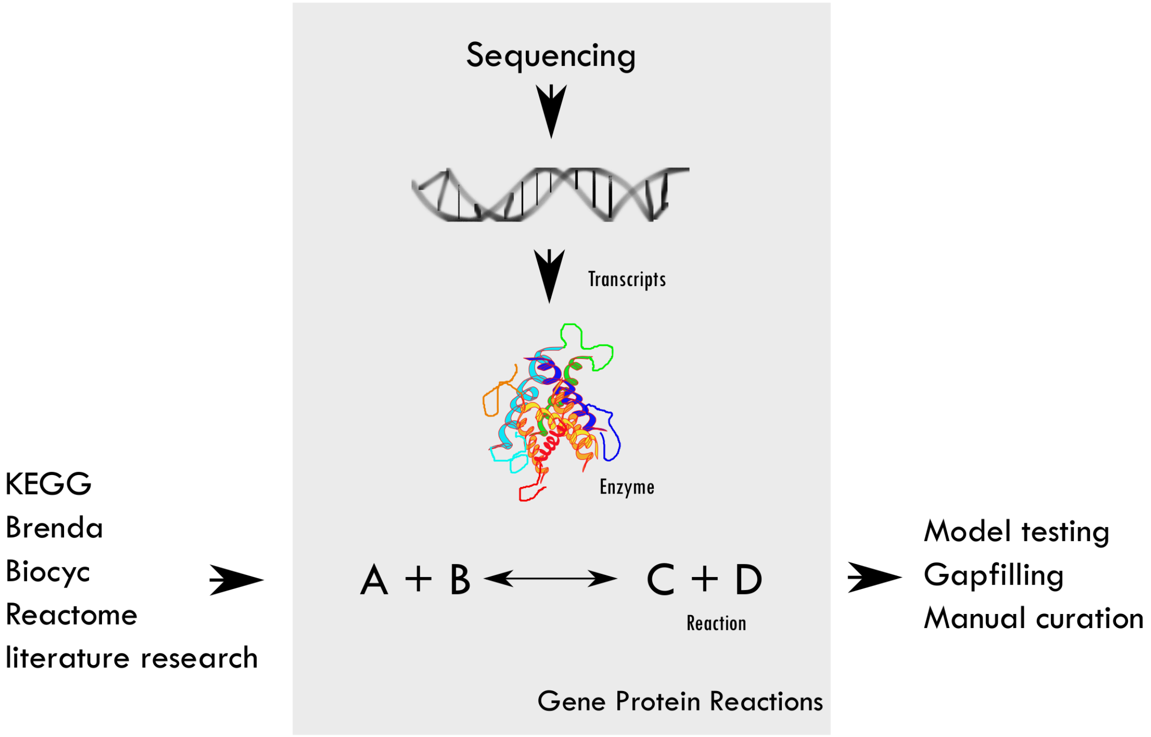


Figure 1.2. Generic metabolic model are obtained after the mapping of sequenced annotated transcripts to the reactome retrieved from diverse databases. The relationship between the genes, the proteins and the reactions are given by the gene-protein-reaction rules (GPRs).

The building of a reconstruction requires several rounds of manually curation based on bibliographic research and database information to set the reversibility of reactions, to establish the gene-protein-reaction association rules that link genetic information to the reactome and to identify the cofactors required for a chemical reaction to take place. Genome-wide reconstructions have dual and sometimes interfering functionalities: as mathematical model that can be used for simulations and as repository of all known reactions that take place in an organism. To be complete a repository will tend to add a reaction multiple times with different cofactors whereas multiple entries of a reaction controlled by the same genes might decrease the prediction power of a model especially if the reactions are reversible which might be responsible for the formation of loops. The currently available reconstructions cover multiple types of organisms, from micro-organisms like *Escherichia coli* (Reed et al., 2003; Orth et al., 2011; Keseler et al., 2013) and *Saccharomyces cerevisiae* (Förster et al., 2003; Aung et al., 2013) to pluricellular organisms like *Arabidopsis thaliana* (Poolman et al., 2009; de Oliveira Dal'Molin et al., 2010; Mintz-Oron et al., 2012). For human cells, several global or generic reconstructions were published in the last years: The Edinburgh Human Metabolic Network (Ma et al., 2007) Recon X (Duarte et al., 2007; Thiele et al., 2013) and the Human metabolic Reconstructions (HMRs) (Agren et al., 2012; Mardinoglu et al., 2014a). Note that for human and other pluricellular organisms, generic GEMs often do not model a real existing cell, as GEMs contain all the reactions known to take place in an organism, but allow extracting subnetworks that account for the different cell types. Beside generic GEMs, cell-specific models like HepatoNet1 (Gille et al., 2010) and *iAdipocytes1809*, (Mardinoglu et al., 2013) were partially manually reconstructed.

1.4 Context-specific models

GEMs, as they are built from genome annotations, do not take into account the variability of phenotypes such as that result from epigenetic modifications or external stimuli, which cause cells of the same organism with identical genetic information to adopt different cell fates, phenotypes, and metabolisms to fulfil very different functions. Following this idea, in 2007 (Shlomi et al., 2007) used transcriptomics data, to further constrain the space of solutions and to extract from the generic reconstruction subnetworks that contain exclusively reactions active in the studied cell type. As all the transcripts are not translated into proteins and the presence of a protein does not guarantee that the latter is an active conformation, and therefore flux rates do not correlate completely with expression levels, the expression level were merely used by iMAT, the method developed by (Shlomi et al., 2007), as clues of the likelihood that the associated reactions carry a flux. iMAT maximizes then the consistency between the expression data and the flux distribution. The reactions predicted to be active respectively inactive by iMAT correlated significantly with the cell-specific data retrieved from GeneNote and Human protein reference database. Nevertheless, 18% of the predicted genes activity did not match their expression levels. A fact that could be explained by post-transcriptional regulation or be attributed to limitations of the input model or of the method.

Extracting subnetworks that only contain active reactions or constraining the bounds of a generic model was proven to increase the predictability of metabolic model in different test such in *in silico* knock-out studies, in which generic models were shown to underestimate the number of essential genes as they tend to include alternative pathways that do not take place simultaneously in the same cell (Folger et al., 2011; Pires Pacheco et al., 2015a). The elimination of inactive pathways allows also obtaining more accurate flux distribution prediction as showed in the simulation of liver disorders such as hyperammonemia and hyperglutamenia (Jerby et al., 2010; Vlassis et al., 2014) and in the prediction of ATP and NO production rates in murine Raw 264.7 macrophages (Bordbar et al., 2012).

1.5 Context-specific reconstruction algorithms

1.5.1 The algorithms

In their review, (Estévez and Nikoloski, 2014) subdivide the context-specific algorithms in 3 families of algorithms in function of the optimization strategy (Figure 1.3, Table 1.1).

Gimme-like family The GIMME-like family of algorithms maximizes a required metabolic function such as the biomass function. By doing so, when using the GIMME-like family for the reconstruction of a metabolic model, the user assumes that the metabolism of the modelled cell can be reduced to the fulfilment of this task. This assumption might be valid for unicellular organisms like *Saccharomyces cerevisiae*, *Escherichia coli* or for cancer cells, but cannot be applied to pluricellular organisms that need to fulfil several tasks simultaneously. Further, pluricellular organisms are composed of specialized cells that allow a division of the tasks or functions among the different cell types. Therefore if a unique function is required, the latter should be considered at the whole organism level.

The GIMME-like family comprises GIMME (Becker and Palsson, 2008), GIMMEp (Bordbar et al., 2012) and GIM3E (Schmidt et al., 2013). GIMME optimizes for a Require Metabolic Function (RMF) while penalizing reactions with expression levels below a user-defined threshold. GIM3E and GIMMEp are variants of GIMME for the integration of metabolic and proteomic data, respectively. GIM3E forces a non-zero flux through reactions implicated in the conversion or the transit of metabolites experimentally shown to be present in the context of interest. Further, GIM3E penalizes all the genes associated reactions of the model. Reactions with lower expression values being more penalized than reactions associated with high expression levels. Another notable difference is the optimization strategy: Mixed Integer Linear Programming (MILP) instead of Linear Programming (LP) for GIMME and GIMMEp.

iMAT-like family The iMAT-like family of algorithms maximizes for the consistency between the flux distribution and the data. The iMAT-like family that comprises iMAT, INIT (Agren et al., 2012) and tINIT (Agren et al., 2014), do not require optimizing for a biological function. But, the drawback of these approaches is their high computational demand.

iMAT maximizes the number of reactions that carry a flux consistent with the expression data (high flux rates associated with high expression values or low flux rates with low expression). In practice, iMAT uses Flux Variance Analysis (FVA) (Mahadevan and Schilling, 2003) to force reactions to carry a flux then to be inactive and evaluate for each case the similarity of the flux distribution to the data in order to determine the activity state of each reaction. The main differences between iMAT and INIT is that INIT uses weights based on expression or arbitrary values to favour the inclusion of reactions that are most supported by the data instead using hard discretization thresholds. Further, INIT imposes a net positive flux through reactions responsible for the production of some metabolites by imposing non-zero bounds (b>0). tINIT uses an additional input, which is the function task that a model must be able to fulfil.

MBA-like family The MBA-like family of algorithms takes as input a core reaction set, defined as set of reactions that have a high confidence level to be expressed in the context of interest. This family searches for a compact consistent model that includes all or most of the core reactions that are complemented with a minimum number of non-core reactions, reactions that are not supported by the data. The concept of core reactions allows the integration of several data types and multiple datasets. The MBA-like family comprises MBA (Model building algorithm) (Jerby et al., 2010), mCADRE (Wang et al., 2012) and the FASTCORE family (Vlassis et al., 2014; Pires Pacheco et al., 2015a) (that are part of this thesis). MBA takes as input two core sets: a core high (CH) and core medium (CM) set in function of the confidence level of the reactions. MBA tests one by one the remaining reactions, called core light (CL). If the removal of the latter does not prevent the high core reactions and a fraction of the medium core reactions to carry a flux, the CL is pruned out with the CM and CL reactions that are no longer able to carry a flux. As the pruning order has a non-negligible impact on the output model, the process is repeated thousand times. In order to speed up this computationally demanding process (2 hours per pruning step), mCADRE uses connectivity and confidence level scores besides expression to not only form the core set but also to determine the pruning order. Reactions with an overall lower confidence level based on the before-mentioned scores are the first tested.

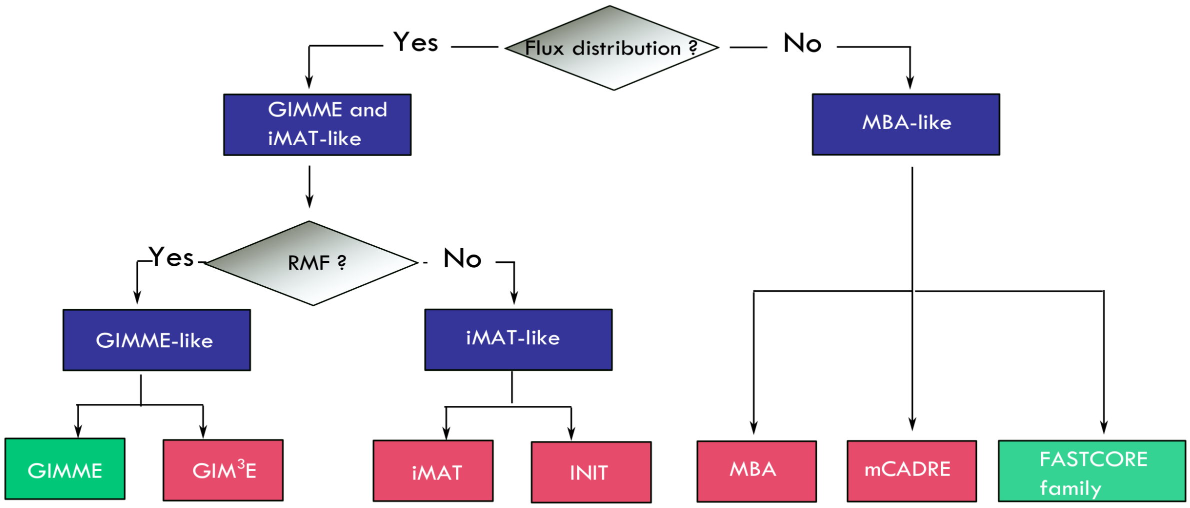


Figure 1.3. (Figure adapted from (Estévez and Nikoloski, 2014) **The 3 families of context-specific reconstruction algorithms use different strategies to build models that uniquely contain reactions that are active in the context of interest.** Context-specific reconstruction algorithms are clustered into three families: a) The GIMME-like family that maximizes a required metabolic function (RMF), b) the iMAT-like that maximizes for the consistency between the data and the flux distribution and c) MBA-like family that maximizes the consistency of the core set. In green and red, the algorithm with low and high computational demands, respectively.

Strategies	Drawbacks
Maximization of a required metabolic function	- not convenient for pluricellular organism - should be considered at the whole organism level - required function might differ strongly among cell types
Maximization of the consistency between the flux and the data	- higher computational demands model
Maximization of consistency of a core set	- the quality of the model is dependent of the core reactions set

Table 1.1. The strategies used by the different families of algorithms and the respective drawbacks.

1.5.2 LP versus ILP

Context-specific model building algorithms can be further classified in function of the optimization framework. Most algorithms use real linear programming (LP). LP, which is the simplest constraint-based optimization framework, is used to solve optimization (maximization or minimization) problems defined as a linear objective, described by polynomes of degree 0 or 1, with unknown decision variables. Note that in this framework, absolute values, square roots or multiplication of decision variables are not allowed. The decision variables are further subjected to some equalities or inequality constraints that limit the space of possible solutions. The optimization of an objective function allows finding the optimal solution(s) corresponding to a flux distribution in the polyhedron of possible solutions (Figure 1.4), in blue) delimited by inequality constraints. Convex or concave objective functions are of particular interest as they have a unique optimal solution when the objective function is minimized, respectively maximized (Smith).

Integer Linear programming, is a special case of LP, where all the decision variables are integers. And therefore the solution space is a Z-polyhedron (represented by black dots in the space of solutions, Figure 1.4) in opposition to (Real) LP where the decision variables can be any real number (blue space, Figure 1.4). Whereas, for MILP only some of the decision variables are constrained to be integers. (Real) LP is more efficient in terms of computational demands as solvable in polynomial time $(n^{constant})$, where n is the number of variables, whereas the computational time for ILP and MILP is more often non polynomial (2^n) (Smith). Among the above described algorithms, the FASTCORE family and GIMME are LP-based whereas the others algorithms use MILP.

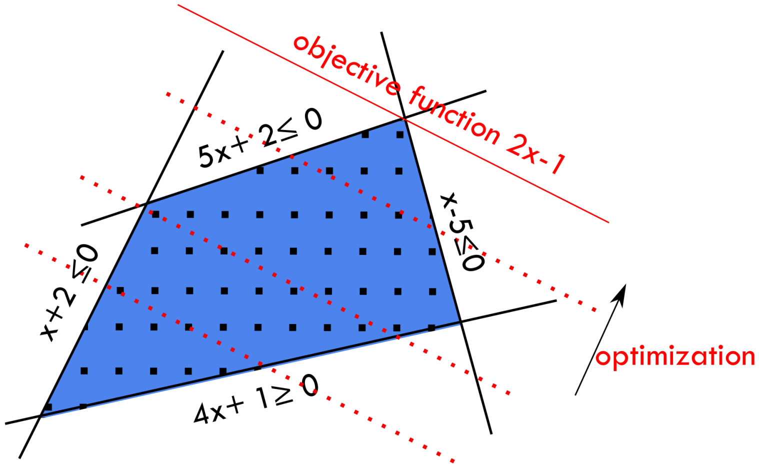


Figure 1.4. Linear Programming optimization framework is commonly used in Constraint-based modelling to find the optimal flux distribution(s) or to reconstruct models. The space of solution(s) is constrained by inequality equations that impose the solution (value of the decision variables) to be found in a given space (blue space). As the system is undetermined, a unique solution can only be found when optimizing for a given function (red line). The blue space represents the solution obtained by real linear programming whereas the black dots illustrate the values that decision values can take in ILP.

1.5.3 Integration of omics data

1.5.3.1 A trade-off between robustness and resolution power

Data obtained from high-throughput technologies like RNA-sequencing (RNA-seq) or protein arrays can be used to constraint generic metabolic models, to extract context-specific models, or to determine what part of the network is active in a given context. The type of data and the choice of the dataset(s) is crucial and should be coherent with the purpose of the model. If the objective is to build a generic tissue or cell-specific reconstruction that is able to simulate the cell or the tissue in all potential conditions and contexts, the inclusion of literature-based evidences from different studies or the inclusion of other types of omics data e.g. transcriptomics or proteomics data gathered from diverse datasets is recommended.

The obtained model can mimic the cell in all possible contexts by adapting the bounds of the model reactions. Further, the inclusion of information on the functionalities of the cell and the medium composition allows increasing the knowledge on the potentialities of the cell and therefore the predictive power.

The drawback of this approach is that the integration of arrays from different contexts reduces the resolution power, defined as the ability to capture small but significant variations between different contexts. Generic tissue-specific models might not capture fine differences between different context like healthy versus disease or different cells of a same tissue with a slightly different phenotype. Furthermore, inducible reactions, such as reactions of the nitric oxide synthesis pathway that are only active in stress conditions or in presence of pathogens might not be included in the model because they are only expressed in a small fraction of the input data.

If the aim is to obtain a snapshot of the metabolism of a cell for a given context, the inclusion of a single dataset is more appropriate as it allows obtaining a greater resolution than the generic counterparts (Figure 1.5). But the use of a unique dataset increases the risk of over-fitting the dataset. Noise might wrongly be interpreted as real meaningful biological variations and further the integration of a single array does not allow obtaining a complete model that is able to simulate the metabolism of a given cell or tissue for all potential contexts.

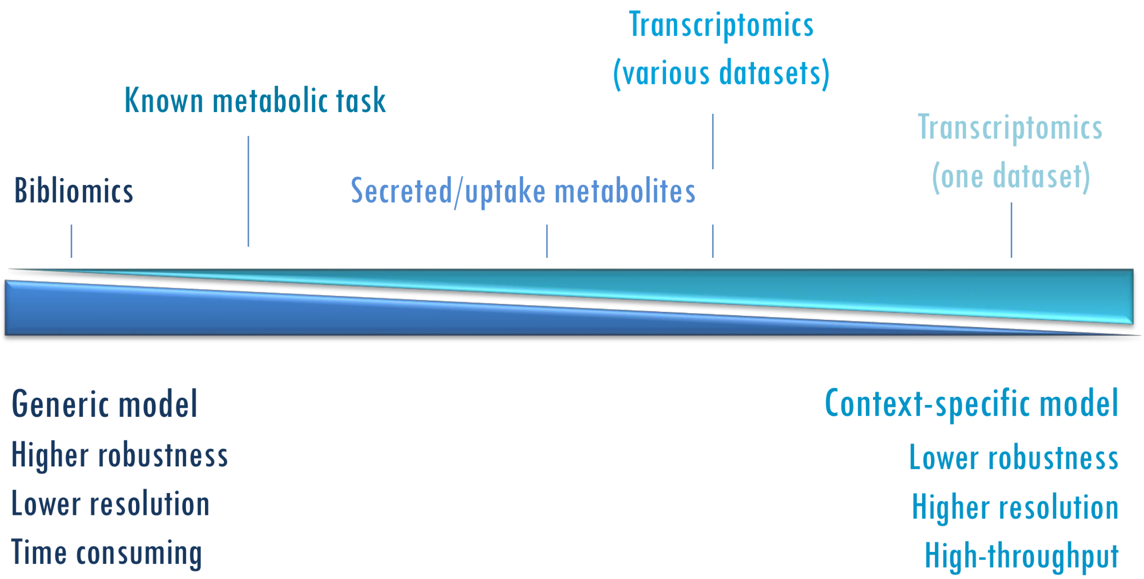


Figure 1.5. Building context-specific metabolic models requires a trade-off between robustness and resolution power: The integration of bibliomics data and data from different datasets allows getting more generic and robust data whereas the integration of one single dataset produce a context-specific, sensitive models with a high resolution power that allow distinguishing similar though different metabolic pattern. Further, the second approach allows also speeding up the building process.

1.5.3.2 Different types of omics data

Different types of omics data i.e. genomics, transcriptomics, proteomics and metabolomics can be considered as input for context-specific building algorithms. Transcriptomic data does not completely correlate with the activity of enzymes. The expression of a gene does not guarantee the presence of a protein or that the protein if present is in an active conformation. Further, a high concentration does not necessary imply a high flux rate. Nevertheless, to this date, transcriptomics is the most used omics data type in metabolic modelling as it is relatively cheap and allows high-throughput acquisition of data. Transcriptomic data is followed by proteomic data in the rank of the most used omics data type. Proteomics data was mainly used as input by the INIT algorithm (Agren et al., 2012) and the main source of high-throughput proteomic data are tissue arrays used within the Human Protein Atlas (HPA) project (Uhlen et al., 2010). In this database, the proteomic data has more a qualitative value and is categorized in high, medium or low confidence level, or as undetected.

Some attempts were done using metabolomics or fluxometrics data as input data to constrain metabolic models. Fluoxometrics is the only layer that allows acquiring directly data on the flux distribution using e.g. carbon 13 labelled molecules (C13). Elementary metabolite units (EMU) that computes the minimal information required to simulate the flux distribution of the labelled atoms was introduced to cope with the large number of isotopomers due the labelled atoms on (Antoniewicz et al., 2007). Although fluxometrics allows measuring directly flux distribution, labelling experiments are mainly available for the central carbon metabolism, explaining why this source of information is not more intensively used. Only very few datasets in *Escherichia coli* and *Saccharomyces cerevisiae* cover other pathways. Therefore, at the moment, the usage of metabolic data is restricted to uptake and secretion rate of key metabolites that are used to fix the bounds of the metabolic models (Figure 1.6).

RNA-Sequencing (RNA-seq): messenger RNA-seq and microarray assays are the most popular sources of transcriptomic data. A messenger RNA-seq starts by enriching for messenger RNAs (mRNA) using poly-T tagged beads or by depleting the sample of ribosomal RNA (rRNA). Once purified the mRNAs are used as template to synthesise a complementary strand of DNA, called complementary DNA or cDNA. The cDNA is then sequenced using a deep-sequencing approach that produces short reads (fragments) of a few nucleotides. These short reads are finally mapped to a reference genome. The level of expression of a gene or abundance is given

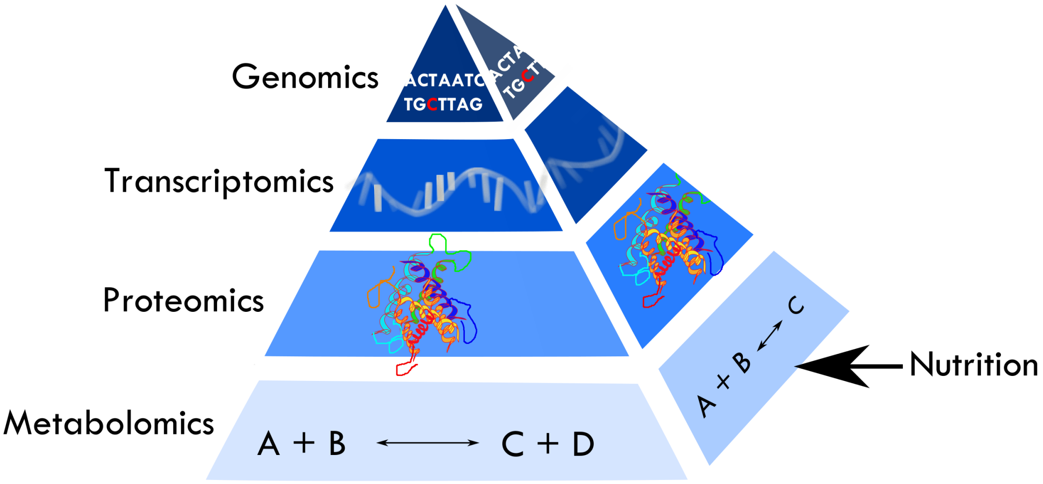


Figure 1.6. Metabolomics integrate the genome information background and nutrition. The enzymes and transporters that are the main actors of the metabolism are coded by genes and therefore mutations in the open reading frame or in the regulation regions have an impact on the metabolism. Another important component is the nutrition, which is a provider of metabolites. The excess of metabolites causes a shift of the equilibrium of metabolic reactions.

in fragments per kilobase of exon per million fragments mapped (FPKM) also called reads per kilobase of exon per million reads mapped (RPKM).

DNA microarray assay: In a microarray experiment, the transcripts are coupled with a dye and then hybridized to short complementary sequences, called probes, that are immobilized on a solid surface. The strength of the hybridization depends on the temperature, the number of perfect matches between the probes and the mRNAs and the base composition of the probes. Guanine-cytosine bounds are more solid that the ones formed by adenine and thymine. Non-specific hybridization, that are characterized by weaker bounds are mostly eliminated by the washing process. Nevertheless, for each probe set, a non-negligible number of probes will be engaged in a non-specific binding. As the strength of the binding depends on the base composition of the probes, the associated noise strongly varies within probe-sets. This non negligible background, so called probe-effect, affects the probe sets differently, and prevents the use of absolute intensity level to quantify gene expression in the frame of microarray experiments.

RNA-seq technology is getting more popular than microarrays as it provides an unbiased measurement of the whole transcriptome, including alternative splicing, non-coding RNAs and unknown transcripts. Whereas microarray assays are limited to the probes present on the solid surface. Moreover, unlike microarrays, RNA-seq is not prone to hybridization saturation and background noise that grants RNA-seq technology with a larger dynamic range and a higher sensitivity.

Moreover, RNA-seq requires less input material and correlates better to protein expression (Van Vliet, 2010; Wang et al., 2009). RNA-seq has also some drawbacks like the large amount of ribosomal RNA (rRNA) in the data, less base accuracy, and variation of read density along the length of the transcript, posing a challenge for this high-throughput method. Although microarray data is getting less popular than RNA-seq, the availability of repository of millions arrays turns microarray data into an extremely valuable resources for modelling purposes.

1.5.3.3 What does expression mean?

The microarray technology is subjected to an unnegligible noise amount that affects probe sets differently and does not allow direct comparison the intensity levels between probe-sets. Probe effects turns questionable the direct integration of transcriptomics into metabolic models using the absolute values of intensity levels. The use of the Barcode (McCall et al., 2011; Zilliox and Irizarry, 2007) a method for discretization of microarray data, allows overcoming this hurdle, by correcting for the noise affecting the different probes. Barcode uses previous knowledge over thousands of conditions to determine for each probe set the distribution of intensities for an unexpressed condition. To determine if a probe set is expressed for a given array, the intensity of the latter is compared to the intensity distribution established by the Barcode project for genes in unexpressed state across thousands of arrays and conditions. Genes are associated with intensity values that are 5 standard deviations apart from the unexpressed intensity distribution mode are considered as expressed. A very conservative threshold allows excluding false positives in the identification process of expressed genes. Nevertheless, if required to form a consistent model, reactions associated with expression levels below this threshold, are included.

Although RNA-seq data is not subjected to probe effects, the question at which level of intensity or number of reads a gene can be considered as expressed is not easy to answer. Is a single molecule of messenger RNA sufficient to consider a gene to be expressed or a simple erratic phenomena that has no impact on the phenotype or metabolism is not solved. And if the latter is true then at how many copy numbers a gene should be considered as expressed? Moreover, in how many cells a gene must be expressed to have an impact on the tissue or the population to which it belongs? The question of the threshold is crucial for the building of context-specific models because, if the former is too conservative active pathways might be excluded in the output models whereas if the threshold was set too low inactive pathways might be included in the context-specific models. Obviously the prediction power is radically affected by the expression threshold(s). A good example are knock-out assays: in the first scenario, the essential reactions will be over-estimated whereas in the second their number will be underestimated.

1.6 Metabolic modelling and personalized medicine

One of the future challenges for the next five to ten years, is to apply metabolic modelling to biomedical engineering and medicine. The metabolic layer, which is directly affected by the nutrition and which stands under the genetic control through the expression of enzymes and transporters, constitute an ideal layer for the study of metabolic diseases as this layer allows integrating and monitoring all aspects contributing to the appearance of a metabolic disease (Figure 1.6).

Metabolic modelling that was already applied with success for bioengineering e.g. for the designing of novel strains of micro-organisms that are able to synthesize chemicals of interest, is one of the most promising field in systems biology that could be applied for precision or personalized medicine. The applications of CBM in biotechnology were reviewed in (Fondi and Liò, 2015) and (Price et al., 2003).

In personalized medicine, metabolic medicine have numerous potential applications reaching from prediction of drugs targets (Folger et al., 2011; Frezza et al., 2011; Kim et al., 2010; Agren et al., 2014), identification of biomarkers (Shlomi et al., 2009) (or reporter metabolites that are often associated to cancer (Patil and Nielsen, 2005)), and anti-metabolites (Agren et al., 2014) (chemical compounds that bind to enzymes and prevent the usage of endogenous metabolites) to a routine integration and analysis of high-throughput patient omics data using metabolic models as scaffold for the understanding of the underlying process leading to the appearance of diseases (Mardinoglu et al., 2013, 2014a).

1.6.1 Metabolic modelling and cancer

The main focus in metabolic modelling of diseases is set on cancer as in humans, cancer cells are the unique cells that justify the optimization of growth, which allows using knock-out and minimal adjustment assays. In this frame, metabolic modelling was mainly used to understand how cancer cells modify their metabolism to increase their proliferation or resistance to hypoxia e.g. by the Warburg effect (Resendis-Antonio et al., 2010; Vazquez et al., 2010; Shlomi et al., 2011; Yizhak et al., 2014b).

Further, several algorithms dedicated specifically to cancer like MPA (Jerby et al., 2012) and PRIME (Yizhak et al., 2014a) were published. MPA uses an approach that is based on MOMA (Segre et al., 2002) to capture phenotype changes between healthy and cancer patients or between patients with different phenotypes using the respective transcriptome or proteome as input. MPA computes a score that translate the consistency between the flux distribution required to optimize for a metabolic task and the expression data. A high consistency score translates to a high likelihood that the function actually takes place in the studied cell. Notably, MPA predicted that several pathways were different between estrogen receptor positive (ER+) and negative (ER-) patients. In turn PRIME is a cancer-specific context-building algorithm that promises predicting the phenotype in an individual-specific manner. PRIME uses as input expression data and growth rates to change the bounds of the input model. Though PRIME was validated only with cancer cell lines and the requirement of growth rates and the difficulty to obtain them for patient limits the application of PRIME for personalized medicine (Yizhak et al., 2014a).

Other more generic context-specific metabolic reconstruction algorithms were also applied for the building of cancer models. The latter were used for the comparison to MBA (Jerby et al., 2010) that allowed determining cytostastics genes, mCADRE (Wang et al., 2012) and INIT (Agren et al., 2012) or pathways that differ between cancer patients and healthy donors. Further, models of cell lines that express different levels of p53 due to the silencing with shRNA constructions targeting p53 and the incubation with Nutlin-3a that activates p53, were reconstructed with iMAT (Zur et al., 2010). The iMAT algorithm predicted that p53 promotes gluconeogenesis and generally decreases the flux through the glycolysis and the pentose phosphate pathways by favouring the consumption of glucose by other metabolic processes explaining to some extent the tumour-suppressor effect of p53 (Goldstein et al., 2013). Another, very popular algorithm in the context of cancer is INIT, that was among others, used to build 46 patient-specific reconstruction suffering from hepatocellular carcinoma (Agren et al., 2014). The models were then employed to predict anti-metabolites. The same approach revealed that the transcript variant ACSS1 is associated with tumour growth and malignancy under hypoxic conditions in human HCC, whereas ACSS2 is linked to a less severe phenotype (Björnson et al., 2015). Beside the numerous draft reconstructions built by the above mentioned algorithms, a population based functional semi-manual genome reconstruction was built using RNA-seq and proteomic data from human colon carcinoma tumours (iHCC2578) (Björnson et al., 2015).

1.6.2 Metabolic modelling and metabolic diseases

The second promising application area are diseases that result from the metabolic syndrome, such as obesity, diabetes and cardiovascular diseases. The mapping of omics data on metabolic reconstruction is a key tool in the identification of disease-related pathways. Among others the catabolism of branched amino acids was predicted to be decreased in obese by *iAdipocyte1809* model (Mardinoglu et al., 2014b). The same model was used as scaffold for the integration of transcriptomic data and uptake rates of glucose and fatty acids of lean versus obese people. Disease-enriched pathways were then identified through random sampling (Mardinoglu et al., 2013). Further, a semi-manual liver reconstruction based on HMR 2.0 (iHepatocytes2322) (Mardinoglu et al., 2014a) and tINIT (Agren et al., 2014) allowed the identification of increased blood levels of chondroitin sulphates (CS) and blood decreased level of heparan sulphates (HS) as potential biomarkers for the non-alcoholic fatty liver disease (NAFLD). The same study demonstrated that branched chain amino-acid transaminase 1 (BCAT1) was up-regulated in tissue-samples of patients suffering of NAFLD.

Metabolic modelling was successfully used in numerous studies and existing algorithms contributed to unravel disease-specific subpathways. Nevertheless, the whole field has to increase in predictability and accuracy to allow the tools, strategies and approaches to be applied in a high-throughput and standardized manner outside the lab. Models still require manually curation, algorithms are too slow, lack accuracy or are often unable to differentiate between the metabolism of related but distinct cells or tissues (Pires Pacheco et al., 2015b).

1.7 Improving the quality of automated reconstructions

1.7.1 Better models call for better algorithms

An important aspect for the reconstruction of metabolic models is time. A manually curated generic model that was obtained after months or years of testing and corrections is expected to be of higher quality than automatic reconstructions. But the time necessary to build a curated model of quality turns their use impracticable for numerous applications like the integration or analysis of patient data that require a more high-throughput approach. Further, automatic reconstructions are getting more popular and are even used to speed up the manual reconstruction process. For example (Becker and Palsson, 2008) was used for the building of subnetworks that were then manually assembled in a macrophage model (Bordbar et al., 2010, 2012). Context-specific algorithms like FASTCORE (Vlassis et al., 2014) (fastgapfill (Thiele et al., 2014)) and MBA (Jerby et al., 2010) (Mirage (Vitkin and Shlomi, 2012)) were slightly modified and applied for gap-filling purposes in the frame of manually curated reconstructions.

Nevertheless, to be used in a high-throughput manner in the frame of diagnosis purposes, the quality of draft reconstructions has to be improved. The required improvements call obviously for better algorithms and inputs models. But also for unbiased and thorough validation methods that allows assessing the quality of algorithms and models and eventually to correct flaws. Two benchmarks evaluating flux prediction algorithms in *Saccharomyces cerevisiae* and *Escherichia coli* (Machado and Herrgård, 2014) and context-specific building algorithms in human cells respectively, (Pires Pacheco et al., 2015a) (the latter was published by ourselves).

1.7.2 Towards functional correct models

Concerns about the ability of draft metabolic models to simulate known metabolic tasks like gluconeogenesis or the conversion of valine into glucose were first raised by (Gille et al., 2010) that tested the functional capacity of a draft liver model built by the iMAT algorithm and realized that the draft model failed in a great number of cases. This negative result led the authors to build the first manually curated liver model, based on experimental evidences (proteomic and transcriptomic level), databases and bibliography. The obtained model was intensively tested for the functional capacity and modified accordingly. HepatoNet1 distinguishes itself from other manually curated model not only by the fact that HepatoNet1 was the first tissue-specific manually curated model but more importantly, HepatoNet1 is functions-and reactions-oriented. This means that the starting point of the models was not the annotated transcripts but the parts of the reactome supported by omics data and the metabolic task (functions) expected to take place in the liver. As a consequence the chosen reactions among the all possible reaction combinations that allow the fulfilment of a biological function might not be the ones expressed in liver cells.

In order to build of a functionally correct macrophage model, (Bordbar et al., 2010, 2012), Bordbar *et al.* created for each metabolic task a sub-network with GIMME. The sub-networks were then manually assembled. The process is faster than the building of manually curated models but could still not be applied in a high-throughput manner. Further, the optimization of several tasks is not so straightforward. The optimization of a function is often only possible at the expense of another. Some flux distributions may be mutual exclusive as requiring too much energy or resources.

Finally, tINIT (Agren et al., 2014), a variant of the INIT algorithm that takes as input metabolic tasks besides transcriptomic and proteomic data, was published recently, offering the possibility to build more correct models. tINIT has nevertheless a serious drawback as it is unable to capture metabolic variations between different tissues (Uhlen et al., 2010). Note also that the identification of the functions that take place in a given cell type is a fastidious process that requires an intensive bibliographic research and as the gene expression changes continuously to cope with external stimuli, the metabolic functions is also expected to vary.

1.8 The metabolic identity and epigenetic control

One possibility to assign the quality of models is to determine the percentage of the reactions that are shown to be active in the context of interest. A reaction that is under gene control is assumed to be active when the associated genes are expressed. But as discussed in the chapter "What does expression mean?", the set of expressed genes in a cell is far from being clearly defined and therefore it is impossible with the current knowledge to identify completely the set of reactions that actually take place in a given tissue.

The identification of the reactome of a cell is an aspect of a wider problematic about cell identity: What defines the different cell types? What are the key elements that make a progeny cell differentiate into a liver or a muscle cell? The first question can be answered in numerous ways. The different cell types distinguish among themselves by their morphologies, their functions and the set of expressed genes and therefore by their transcriptome, proteome and metabolome. Further a resident macrophage and a pathogens fighting macrophage have a different transcriptome and metabolome but both remain macrophages and will never become liver cells. Calling for the next question: Why do progenitors cells have the potential to give rise to every cell type whereas differentiated cells lose this ability?

1.8.1 Enhancers and cell identity

Epigenetics and the study of gene regulation might deliver some initial responses to these questions. One of the main characteristics of Metazoa is the subdivision in cell types with different expression patterns, functions and morphologies, which calls for a regulation of the gene expression. Since their unicellular ancestors, Metozoa have a complex network of genes responsible for transcriptional regulation of genes linked to specific functions (Sebé-Pedrós et al., 2016), advocating for a highly conserved gene expression regulation system. The different expression patterns are under the control of the condensation level of the DNA molecules and of one or a combination of multiple types of transcription factors (Adams and Workman, 1995) that bind to cis-and-trans regulating elements of target genes, called enhancers. Whereas unicellular organisms are characterized by small regulatory regions that are mainly proximal to the target gene advocating for the hypothesis that the combinatory binding of transcription factors and the chromatin loop formation is restricted to metazoa.

1.8.1.1 Regulation of the gene expression by the chromatin condensation level and the transcription network

The DNA strands are wrapped around basic proteins, called histones (H3, H4, H2A, H2B and H1), forming a nucleosome, the building block of eucaryotic chromatin fibres (Luger and Richmond, 1998; Van Holde, 2012). The chromatin can be in a different state of relaxation or condensation. Genes located for example in a highly condensed hetero-chromatin regions are not translated into messenger RNA because the transcription machinery cannot access the promoter of these genes. The level of condensation of the chromatin depends on the affinity of the basic histone tails to the negatively charged DNA strands. This charge can be masked by the addition of covalent modifications such as acetyl-residues to the lysines or neutralized by phosphorylation, causing a relaxation of the chromatin structure (Wolffe and Hayes, 1999).

The level of condensation of the chromatin is a dynamic process that depends on a large number of possible modifications (phosphorylation, methylation, ubiquitination, sumoylation, ADP ribosylation, glycosylation, biotinylation and carbonylation). The complexity of histone modifications and their action on the chromatin structure led to the formulation of the histone code (Strahl and Allis, 2000) Hypothesis that the combination of various modifications allows amplifying the read-out of upstream signalling pathways, causing a larger and more controlled impact on the chromatin condensation state (Strahl and Allis, 2000).

In summary, the layer of regulation of the cell identity depends on the binding sites (enhancers) that are accessible to transcription factors, that allow the binding of transcription factors. These binding sites, for most of them of few hundred base pairs in length (Carey, 1998), are characterized by a depletion of histones (Bernstein et al., 2002; Lee et al., 2004) due among others on a enrichment of histone acetylation (H3K9ac, H3K27ac) and methylation (H3K4me1) of the histones tails in this area. Enhancers harbour usually multiple binding sites that in some cases only allow a stable binding of a complex of transcription factors.

The binding of transcription factors permits the fixation of co-activators, that, unlike transcription factors, do not bind to a specific sequence. Co-activators cause covalent modifications such as methylation or acetylation state of the DNA and histones through the recruitment of p300, a covactor that is also a histone acetyltransferase, that increases the accessibility of the DNA binding sites to the RNA polymerase II (Lelli et al., 2012).

1.8.1.2 Cell-specific expression patterns depend on the activity of enhancers

Cell-specific expression patterns that define the cell identity are less related to the expression a cell-specific transcription factor, as most transcription factors were shown to be ubiquitously expressed (Uhlén et al., 2015) than to an unique combination of transcription factors that bind together in a stable manner to large regions of more than 50 kilobases containing several cluster of enhancers, called super-enhancer, that are active in a cell-specific manner (Bulger and Groudine, 2011).

Conformingly, the activity of super-enhancers and enhancers in general were shown to be largely cell-type specific (Consortium et al., 2012; Heintzman et al., 2009), as shown by large studies focusing on histones marks like H3K27ac and H3K4me1 (Creyghton et al., 2010; Rada-Iglesias et al., 2011; Heintzman et al., 2009), the binding of p300 (Visel et al., 2009), and DNAse hypersensibility (Consortium et al., 2012; Hnisz et al., 2013) that act as surrogate for enhancers.

Moreover, gene ontology analysis showed that super-enhancer are largely associated with process linked with cell identity (Hnisz et al., 2013), association studies showed that super-enhancers are linked to genes enriched for diseases and that single nucleotide polymorphism (SNP) are often localized within super-enhancer regions of disease-relevant cells types, (Hnisz et al., 2013; Lovén et al., 2013; Chapuy et al., 2013), suggesting that super-enhancers marks and regulate cell-specific genes like master transcription factor or genes implicated in the identity of the cell and that any small variation in the regulation of these central genes is likely to cause the appearance of diseases.

1.8.2 Enhancers and macrophages differentiation

Each cell type contains a different set of enhancers (Heintzman et al., 2009) to which transcription factors can bind, after being activated by an external or internal stimuli, and initiate the expression of the target genes. This predefined set of enhancers prevents the transformation of a cell type into a very different one as the required enhancers for the expression of another set of genes are simply missing. Nevertheless, several cells types demonstrate a certain level of plasticity like the cells of myeloid lineage can differentiate into a different subtype in function of the stimuli. Turning this lineage into an ideal model for the study of enhancers and cell identity.

Histones marks are often used as surrogate for large cluster of active enhancers. Enhancers are also found in a latent or inactive state that are not enriched in H3K27ac histones marks or to which transcription factors do not bind unless the cell are submitted to the right stimuli. In the similar way, when submitted to a given stimuli, active enhancer can lose the associated histones marks (Ostuni et al., 2013). This latent enhancers are characterised by only presence of H3K4me1 histone marks. Interestingly, during development, H3K4me1 was shown to be a mark of early stage of enhancer formation that precedes nucleosome depletion and H3K27ac histone marks (Creyghton et al., 2010; Rada-Iglesias et al., 2011), supporting the idea that the latent enhancers allow the cells to acquire new functionalities in order to cope with unexpected stimuli like i.e. (lipopolysaccharide) LPS (Ostuni et al., 2013).

Scope and aims

The first aim of this work is the implementation of a family of algorithms that allow the automated reconstruction of high quality metabolic models that could be used in a high-throughput manner. To achieve this aim, the algorithms must fulfil a certain number of criteria. The reactions included in the model must be supported by independent data with a high confidence level. The reconstruction algorithms must be robust in order to avoid modelling noise but nevertheless be able to capture significant metabolic variations between different cell types, tissues and contexts. Further, algorithms must also have reduced computational demands and be devoid of free parameters to allow their use in high-throughput framework. Finally, metabolic models must be able to predict, among others, the key flux distributions, essential genes, drug targets or secretions rates that could subsequently be validated *in vitro*.

The second aim is to establish benchmarking procedures to assign the quality of context-specific algorithms and validate metabolic models. Unbiased validation procedures allow identifying caveats of the respective approaches in order to permit a correction in the next versions. Furthermore, a benchmarking using the same input data allows selecting among the alternative strategies proposed by the existing algorithms the most efficient ones, so that the more efficient strategy could be incorporated in the next generation of algorithms and by such guarantee that the every newly published algorithm outperforms its predecessors.

The last aim of this thesis was to understand how gene regulatory network and epigenetics define the metabolic identity of monocyte-derived macrophages by enhancing the expression of a set of key target genes. Previous studied showed that the number of enhancers associated with a gene was related with the cell-specificity of its expression. Therefore, the integration of epigenetic data in context-specific models should allow identifying key control points in the metabolic network that regulate the establishment of the macrophage metabolic identity.

FASTCORE

The following chapter was published as the ***'Fast reconstruction of compact context-specific metabolic models networks'*** *article in Plos Computational biology in January 2014.*

Contents

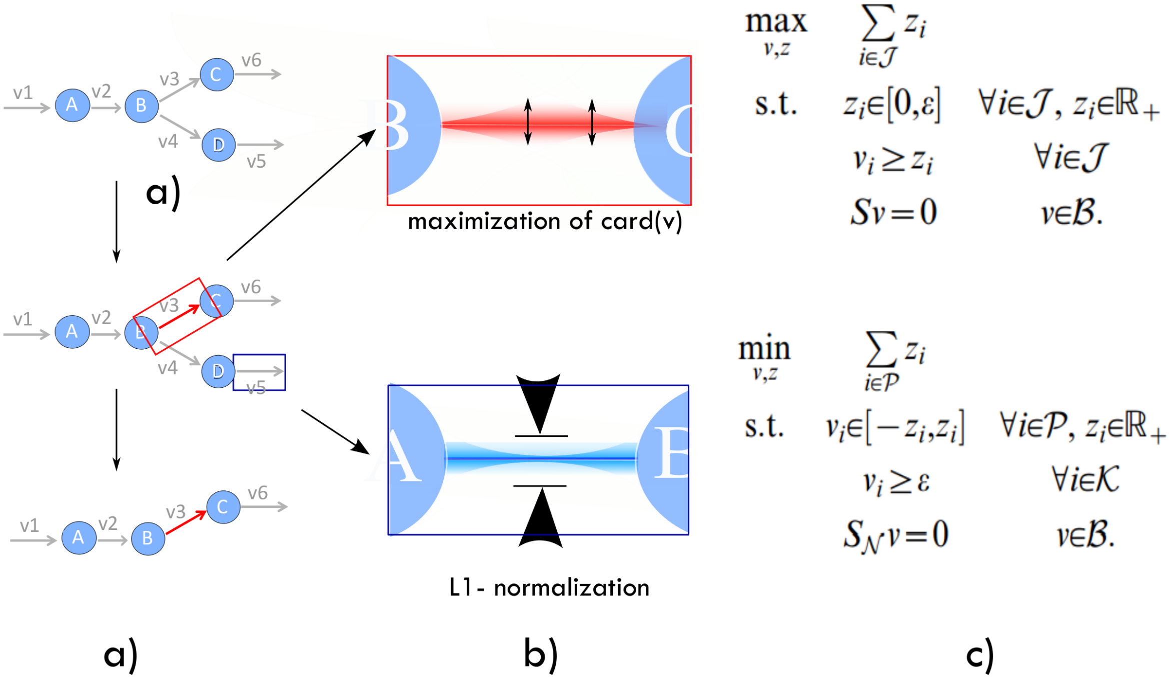
3.1 Summary and contributions

The idea of Prof. Thomas Sauter to create a fast context-specific reconstruction algorithm stems from the fact that although several algorithms were already published to fulfil this specific task, none of them could practically be used in high-throughput way for the building of human metabolic models. Indeed most algorithms have too high computational demands as using Mixed Integer linear Programming (MILP), require the setting of arbitrary thresholds or/and optimize for a biological function. Free parameters should be avoided when designing an algorithm as they call for computationally demanding parameter tuning that should be ideally repeated for each input model and dataset. Whereas the optimization of a biological function is a strategy developed for unicellular organism that assume that the metabolism of a cell can be reduced to the fulfilment of one single task (the optimization of more than one task is not straightforward due to the pareto effect). For humans, this category of algorithms can only be applied for the modelling of cancer cells, the other cell types usually do not divide.

The philosophy behind FASTCORE (Vlassis et al., 2014), for which I share the first co-authorship (shared contributions) with Dr. Nikos Vlassis and for which Prof. Thomas Sauter is the last author, was to create a context-specific reconstruction algorithm devoid of arbitrary thresholds and with building times in the scale of seconds. The low computational demands of FASTCORE open a whole range of new possibilities like cross-validation assays that are practically not doable with most algorithms. Leave-out cross-validation assays allow assigning a confidence score to each reaction of the output model, which facilitates manual curation and eventually the high-throughput building of models for e.g. diagnosis purposes.

FASTCORE requires as input a set of core reactions, reactions, known to be expressed in context-specific of interest (e.g. obtained from data mining), and a generic genome-wide reconstruction (such as Recon X (Duarte et al., 2007; Thiele et al., 2013), HMR (Agren et al., 2012), HMR 2.0 (Mardinoglu et al., 2014a), etc).

FASTCORE, is based on the idea of Dr. Vlassis, who developed the mathematical concepts of FASTCORE with the precious help of Prof. Thomas Sauter, to approximate the cardinality function by the minimum of two linear functions. As this approximation is concave for positive flux values, the function can be maximized using linear programming (LP). The maximization of an approximation of the cardinality of the flux v that flow through the core reactions forces the latter to carry a non-zero flux. Further, to obtain compact context-specific models, the inclusion of non-core reactions are penalized using an L1-minimization, therefore only a minimal set of non-core reactions are added to the core reactions in order to obtain a consistent model (Figure 3.1).



$$\max_{v,z} \sum_{i\in\mathcal{J}} z_i$$
$$\text{s.t.} \quad z_i\in[0,\varepsilon] \quad \forall i\in\mathcal{J},\ z_i\in\mathbb{R}_+$$
$$v_i \geq z_i \quad \forall i\in\mathcal{J}$$
$$Sv=0 \quad v\in\mathcal{B}.$$

$$\min_{v,z} \sum_{i\in\mathcal{P}} z_i$$
$$\text{s.t.} \quad v_i\in[-z_i,z_i] \quad \forall i\in\mathcal{P},\ z_i\in\mathbb{R}_+$$
$$v_i \geq \varepsilon \quad \forall i\in\mathcal{K}$$
$$S_{\mathcal{N}}v=0 \quad v\in\mathcal{B}.$$

Figure 3.1. Principle of FASTCORE: 1) Maximization of an approximation of the cardinality function and 2) minimization of the flux through non-core reactions. a) Toy network were the circles represent the metabolites and the arrows the reactions b) illustrations of the maximization of the cardinality of the flux carried by the core reaction and of penalization of non-core reactions through L1-normalization c) equation describing both processes.

The experiments performed to validate the FASTCORE namely the reconstruction of a liver and a model macrophage model, the comparison of the performance of FASTCORE to MBA (Jerby et al., 2010) and GIMME (Becker and Palsson, 2008), 2 competitors algorithms, the cross-validation assays, the functional analysis of the FASTCORE liver model and the simulation of the flux distribution in the context of liver disorders affecting three enzymes arginosuccinate synthetase (ASS), arginosuccinate lyase (ASL) and ornithine transcarbanylase (OTC) (Figure 3.6) through random sampling and the comparison to the *in vivo* fluxes as described in the paper were mainly performed by me under the guidance and help of Prof. Thomas Sauter and Dr. Nikos Vlassis. Figure 3.4 and Table 3.2 and 3.1 and the supplementary data were mainly created by me with the help of Prof. Sauter and Dr. Vlassis. Whereas Dr. Vlassis, developed the mathematical concepts with help of Prof. Thomas Sauter, wrote box 1 and 2, draw Figure 3.3 and 3.5. He also run the simulation for 3.3. The manuscript was written and the development of the tool was done by all three authors (see for contribution section of the original paper for more details).

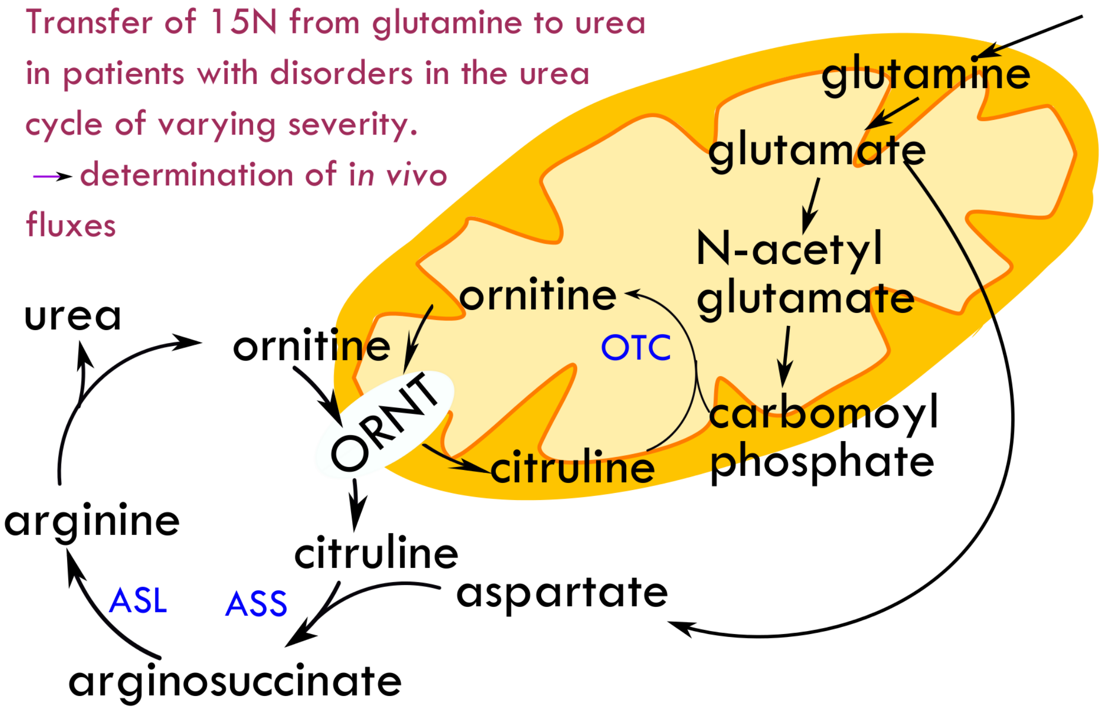


Figure 3.2. Simulation of the flux distribution in the context of diseases linked to mutations in the urea cycle: Several disorders are due to mutations in 3 genes: arginosuccinate synthetase (ASS), arginosuccinate lyase (ASL) and ornithine transcarbanylase (OTC). The flux distributions were simulated *in silico* to validate the FASTCORE liver model by random sampling and were then compared to the flux distributions determined by labeled 15-N glutamine in the study of Lee *et al*.

3.2 Abstract

Systemic approaches to the study of a biological cell or tissue rely increasingly on the use of context-specific metabolic network models. The reconstruction of such a model from high-throughput data can routinely involve large numbers of tests under different conditions and extensive parameter tuning, which calls for fast algorithms. We present , a generic algorithm for reconstructing context-specific metabolic network models from global genome-wide metabolic network models such as Recon X. FASTCORE takes as input a core set of reactions that are known to be active in the context of interest (e.g., cell or tissue), and it searches for a flux consistent subnetwork of the global network that contains all reactions from the core set and a minimal set of additional reactions. Our key observation is that a minimal consistent reconstruction can be defined via a set of sparse modes of the global network, and FASTCORE iteratively computes such a set via a series of linear programs. Experiments on liver data demonstrate speedups of several orders of magnitude, and significantly more compact reconstructions, over a rival method. Given its simplicity and its excellent performance, FASTCORE can form the backbone of many future metabolic network reconstruction algorithms.

3.3 Introduction

Cell metabolism is known to play a key role in the pathogenesis of various diseases (DeBerardinis and Thompson, 2012) such as Parkinson's disease (Pourfar et al., 2013) and cancer (Hiller and Metallo, 2013). The study of human metabolism has been greatly advanced by the development of computational models of metabolism, such as Recon 1 (Duarte et al., 2007), the Edinburgh human metabolic network (Hao et al., 2010), and Recon 2 (Thiele et al., 2013). These are genome-scale metabolic network models that have been reconstructed by combining various sources of 'omics' and literature data, and they involve a large set of biochemical reactions that can be active in different contexts, e.g., different cell types or tissues (Thiele and Palsson, 2010).

To maximize the predictive power of a metabolic model when conditioning on a specific context, for instance the energy metabolism of a neuron or the metabolism of liver, recent efforts go into the development of *context-specific* metabolic models (Becker and Palsson, 2008; Christian et al., 2009; Jerby et al., 2010; Chang et al., 2010; Lewis et al., 2010b; Agren et al., 2012). These are network models that are derived from global models like Recon 1, but they only contain a subset of reactions, namely, those reactions that are active in the given context. Such context-specific metabolic models are known to exhibit superior explanatory and predictive power than their global counterparts (Jerby et al., 2010; Folger et al., 2011; Bordbar et al., 2012).

Most algorithms for context-specific metabolic network reconstruction (see Section 3.4.5 for a short overview) first identify a relevant subset of reactions according to some 'omics' information (typically expression data and bibliomics), and then search for a subnetwork of the global network that satisfies some mathematical requirements and contains all (or most of) these reactions (Becker and Palsson, 2008; Shlomi et al., 2008; Jerby et al., 2010; Chandrasekaran and Price, 2010; Jensen and Papin, 2011; Agren et al., 2012). The mathematical requirements are typically imposed via flux balance analysis, which characterizes the steady-state distribution of fluxes in a metabolic network via linear constraints that are derived from the stoichiometry of the network and physical conservation laws (Schuster and Hilgetag, 1994; Stephanopoulos et al., 1998; Price et al., 2004; Gagneur and Klamt, 2004; Fleming et al., 2012). The search problem may target the optimization of a specific functionality of the model (e.g., biomass production) or some other objective (Blazier and Papin, 2012), and it may involve repeated tests under different conditions and parameter tuning (Becker and Palsson, 2008; Folger et al., 2011; Orth et al., 2011; Wang et al., 2012). The latter calls for fast algorithms.

We present FASTCORE, a generic algorithm for context-specific metabolic network reconstruction. FASTCORE takes as input a core set of reactions that are supported by strong evidence to be active in the context of interest. Then it searches for a *flux consistent* subnetwork of the global network that contains all reactions from the core set and a minimal set of additional reactions. Flux consistency implies that each reaction of the network is active (i.e., has nonzero flux) in at least one feasible flux distribution (Schuster and Hilgetag, 1994; Acuña et al., 2009). An attractive feature of FASTCORE is its generality: As it only relies on a preselected set of reactions and a simple mathematical objective (flux consistency), it can be applied in different contexts and it allows the integration of different pieces of evidence ('multi-omics') into a single model.

Computing a minimal consistent reconstruction from a subset of reactions of a global network is, however, an NP-hard problem (Acuña et al., 2009), and hence some approximation is in order. Our key observation is that a minimal consistent reconstruction can be defined via a set of *sparse* modes of the global network, and FASTCORE is designed to compute a minimal such set. Every iteration of the algorithm computes a new sparse mode via two linear programs that aim at maximizing the support of the mode inside the core set while minimizing that quantity outside the core set. FASTCORE's search strategy is in marked contrast to related approaches, in which the search for a minimal consistent reconstruction involves, for instance, incremental network pruning (Jerby et al., 2010). FASTCORE is simple, devoid of free parameters, and its performance is excellent in practice: As we demonstrate on experiments with liver data, FASTCORE is several orders of magnitude faster, and produces much more compact reconstructions, than the main competing algorithm MBA (Jerby et al., 2010).

3.4 Methods

3.4.1 Background

A metabolic network of m metabolites and n reactions is represented by an $m \times n$ *stoichiometric* matrix S, where each entry S_{ij} contains the stoichiometric coefficient of metabolite i in reaction j. A *flux* vector $v \in \mathbb{R}^n$ is a tuple of reaction rates, $v = (v_1, \ldots, v_n)$, where v_i is the rate of reaction i in the network. Reactions are grouped into *reversible* ones ($\mathcal{R}$) and *irreversible* ones ($\mathcal{I}$). For a reaction $i \in \mathcal{I}$ it holds that $v_i \geq 0$; this and other imposed flux bounds, e.g., lower and upper bounds per reaction, are collectively denoted by $\mathcal{B}$ (which defines a convex set). A flux vector is called *feasible* or a *mode* if it satisfies a set of steady-state mass-balance constraints that can be compactly expressed as:

$$Sv = 0, \quad v \in \mathcal{B}\,. \tag{3.1}$$

An *elementary* mode is a feasible flux vector $v \neq 0$ with minimal support, that is, there is no other feasible flux vector $w \neq 0$ with $supp(w) \subset supp(v)$, where $supp(v) = \{j \in \{1, 2, \ldots, n\} : v_j \neq 0\}$ is the support (i.e., the set of nonzero entries) of v (Schuster and Hilgetag, 1994; Gagneur and Klamt, 2004). A reaction i is called *blocked* if it cannot be active under any mode, that is, there exists no mode $v \in \mathbb{R}^n$ such that $v_i \neq 0$ (in practice $|v_i| \geq \varepsilon$, for some small positive threshold ε). A metabolic network model that contains no blocked reactions is called *(flux) consistent* (Schuster and Hilgetag, 1994; Acuña et al., 2009).

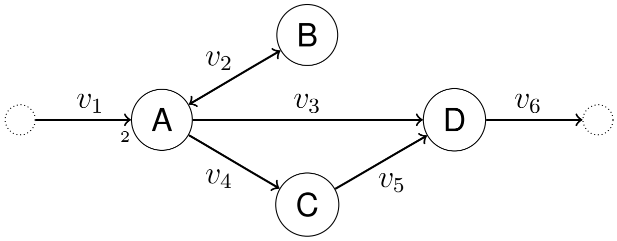


Figure 3.3. Consistency testing allows identifying and removing reactions that cannot carry a flux due to the existence of gaps or dead-ends: A metabolic network with one blocked reaction (A↔B). Note that A appears with stoichiometric coefficient 2 in the boundary reaction →2A.

3.4.2 Network consistency testing

Given a metabolic network model with stoichiometric matrix S, a problem of interest is to test whether the network is consistent or not. Additionally, if the network is inconsistent, it would be desirable to have a method that detects all blocked reactions.

It has been suggested that network consistency can be detected by a single linear program (LP) (Acuña et al., 2009). The idea is to first convert each reversible reaction into two irreversible reactions (and define a reversible flux as the difference of two irreversible fluxes), and then test if the minimum feasible flux on the new set $\mathcal{J}$ of irreversible-only reactions is strictly positive (in practice, at least ε). This is equivalent to testing if the following LP is feasible:

$$\begin{aligned} \max_{v,z} \quad & z & & \\ \text{s.t.} \quad & z \geq \varepsilon & & z \in \mathbb{R} \\ & v_i \geq z & & \forall i \in \mathcal{J} \\ & Sv = 0 & & v \in \mathcal{B}\,. \end{aligned} \tag{LP-3.2}$$

This test of consistency, however, can produce spurious solutions. In Figure 3.3 we show a toy metabolic network comprising four metabolites (A,B,C,D) and six reactions annotated with corresponding fluxes $v_1, \ldots, v_6$. Fluxes are bounded as $0 \leq v_i \leq 3$ for $i \neq 2$, and $|v_2| \leq 3$. All stoichiometric coefficients are equal to one, except for the reaction →2A. The only reversible reaction is A↔B, which is a dead-end reaction and therefore blocked, whereas all other reactions are irreversible and unblocked. After converting A↔B to a pair of irreversible reactions, LP-3.2 achieves optimal value $z^* = 1.5$, which implies (wrongly) that the network is consistent. The test here fails because the two irreversible copies of A↔B have equal flux at the solution, thereby nullifying the actual net flux of A↔B.

A straightforward solution to the problem would involve iterating through all reactions, computing the maximum and minimum feasible flux of each reaction via an LP that satisfies the constraints in (3.1). Reactions with minimum and maximum flux zero would then be blocked. This is the idea behind the FVA (Flux Variability Analysis) algorithm and the *reduceModel* function of the COBRA toolbox (Mahadevan and Schilling, 2003; Schellenberger et al., 2011a). However, iterating through all reactions can be inefficient. A faster variant is fastFVA (Gudmundsson and Thiele, 2010), which achieves acceleration over FVA via LP warm-starts. Another fast algorithm is CMC (CheckModelConsistency) (Jerby et al., 2010), which involves a series of LPs, where each LP maximizes the sum of fluxes over a subset $\mathcal{J}$ of reactions:

$$\begin{aligned} \max_{v} \quad & \sum_{j \in \mathcal{J}} v_j \\ \text{s.t.} \quad & Sv = 0 \qquad v \in \mathcal{B}\,. \end{aligned} \tag{LP-3.3}$$

The set $\mathcal{J}$ is initialized by $\mathcal{J} = \mathcal{R} \cup \mathcal{I}$ (all reactions in the network), and it is updated after each run of LP-3.3 so that it contains the reactions whose consistency has not been established yet. When $\mathcal{J}$ cannot be reduced any further, we can reverse the signs of the columns of S corresponding to the reversible reactions in $\mathcal{J}$ and resume the iterations. Eventually, all remaining reactions may have to be tested one by one for consistency, as in FVA. Such an iterative scheme is complete, in the sense that it will always report consistency if the network is consistent, and if not, it will reveal the set of blocked reactions. However, as we will clarify in the next section, LP-3.3 is not optimizing the 'correct' function, which may result in unnecessarily many iterations. For example, when applied to the network of Figure 3.3, LP-3.3 will pick up the elementary mode that corresponds to the pathway A→C→D (because this pathway achieves maximum sum of fluxes $v_1 + v_4 + v_5 + v_6 = 1.5 + 3 + 3 + 3$), and it will set $v_3 = 0$. To establish the consistency of the reaction A→D, an additional run of LP-3.3 would be needed, where the set $\mathcal{J}$ would only involve the reactions A↔B and A→D. Hence, an iterative algorithm like CMC that relies on LP-3.3 would need two iterations to detect the consistent part of this network. However, one LP suffices to detect the consistent subnetwork in this example, as we explain in the next section. In more general problems involving larger and more realistic networks, CMC may involve unnecessarily many iterations, as we demonstrate in the experiments.

3.4.3 Fast consistency testing

In most problems of interest there will be no single mode that renders the whole network consistent, and an iterative algorithm like the one described in the previous section must be used. For performance reasons it would therefore be desirable to be able to establish the consistency of as many reactions as possible in each iteration of the algorithm.

Since consistency implies nonzero fluxes, it is sufficient to optimize a function that just 'pushes' all fluxes away from zero. Formally, this amounts to searching for modes v whose *cardinality*—denoted by $card(v)$ and defined as $card(v) = \#supp(v)$, i.e., the number of nonzero entries of v—is as large as possible. Directly maximizing $card(v)$ is, however, not straightforward, for the following reasons: First, the $card$ function is quasiconcave only for $v \in \mathbb{R}^n_+$ (the nonnegative orthant), and it is nonconvex for general $v \in \mathbb{R}^n$ (Boyd and Vandenberghe, 2004). Second, even if we restrict attention to nonnegative fluxes in each iteration (which we can do without loss of generality by flipping the signs of the corresponding columns of S), it is not obvious how to efficiently maximize the quasiconcave $card(v)$. Third, in practice consistency implies fluxes that are ε-distant from zero, in which case some adaptation of the $card$ function is in order.

Here we propose an approach to approximately maximize $card(v)$ over a nonnegative flux subspace indexed by a set of reactions $\mathcal{J}$. First note that the cardinality function can be expressed as

$$card(v) = \sum_{i \in \mathcal{J}} \theta(v_i)\,, \tag{3.4}$$

where $\theta : \mathbb{R} \to \{0, 1\}$ is a step function:

$$\theta(v_i) = \begin{cases} 0 & \text{if } v_i = 0 \\ 1 & \text{if } v_i > 0\,. \end{cases} \tag{3.5}$$

The key idea is to approximate the function θ by a concave function that is the minimum of a linear function and a constant function:

$$\theta(v_i) \approx \min\{\frac{v_i}{\varepsilon}, 1\}\,, \tag{3.6}$$

where ε is the flux threshold. The problem of approximately maximizing $card(v)$ can then be cast as an LP: We introduce an auxiliary variable $z_i \in \mathbb{R}_+$ for each flux variable v_i, for $i \in \mathcal{J}$, and take epigraphs (Boyd and Vandenberghe, 2004), in which case maximizing $card(v) = \sum_{i \in \mathcal{J}} \theta(v_i)$ can be expressed as

$$\begin{aligned}
\max_{v,z} \quad & \sum_{i\in\mathcal{J}} z_i \\
\text{s.t.} \quad & z_i \le \theta(v_i) && \forall i \in \mathcal{J},\ z_i \in \mathbb{R}_+ \\
& v_i \ge 0 && \forall i \in \mathcal{J} \\
& Sv = 0 && v \in \mathcal{B}\,.
\end{aligned}$$

Using (3.6) and assuming constant ε, this simplifies to

$$\begin{aligned}
\max_{v,z} \quad & \sum_{i\in\mathcal{J}} z_i \\
\text{s.t.} \quad & z_i \in [0,\varepsilon] && \forall i \in \mathcal{J},\ z_i \in \mathbb{R}_+ \\
& v_i \ge z_i && \forall i \in \mathcal{J} \\
& Sv = 0 && v \in \mathcal{B}\,.
\end{aligned} \tag{LP-3.7}$$

Note that LP-3.7 tries to maximize the number of feasible fluxes in $\mathcal{J}$ whose value is at least ε (contrast this with LP-3.2).

Returning to the network of Figure 3.3, if $\mathcal{J}$ comprises all network reactions, then note that the flux vector $[v_1, v_2, v_3, v_4, v_5, v_6] = [\varepsilon, 0, \varepsilon, \varepsilon, \varepsilon, 2\varepsilon]$ is an optimal solution of LP-3.7. Hence, a single run of the latter can detect all unblocked reactions of that network. More generally, a single run of LP-3.7 on an arbitrary subset $\mathcal{J}$ of a given network will typically detect all unblocked *irreversible* reactions of $\mathcal{J}$. The intuition is that LP-3.7 prefers flux 'splitting' over flux 'concentrating' in order to maximize the number of participating reactions in the solution, which, in the case of irreversible reactions, corresponds to flux cardinality maximization.

By construction, the above approximation of the cardinality function applies only to nonnegative fluxes. In order to deal with reversible reactions that can also take negative fluxes, we can embed LP-3.7 in an iterative algorithm (as in the previous section), in which reversible reactions are first considered for positive flux via LP-3.7, and then they are considered for negative flux. The latter is possible by flipping the signs of the columns of the stoichiometric matrix that correspond to the reversible reactions under testing, in which case the fluxes of the transformed model are again all nonnegative, and the above approximation of the cardinality function can be used. This gives rise to an algorithm for detecting the consistent part of a network that we call FASTCC (for fast consistency check). Since FASTCC is just a variant of FASTCORE, we defer its detailed description until the next section.

Independently to this work, a similar approach to network consistency testing was recently proposed, called OnePrune (Dreyfuss, 2014). OnePrune first converts each reversible reaction into two irreversible reactions, forming an augmented set $\mathcal{J}$ of irreversible-only reactions (as in LP-3.2 above), and then it employs an LP that coincides with LP-3.7 for the above choice of $\mathcal{J}$ and $\varepsilon = 1$. However, such an approach is prone to the same drawback as LP-3.2, namely, that the two irreversible copies of a blocked reaction can carry equal positive flux at the solution of LP-3.7 due to the presence of cycles introduced by the transformation. The authors acknowledge this problem but they do not fully resolve it. In our case, we avoid this problem by working with the original reactions and a series of LPs with appropriate sign flips of the stoichiometric matrix, thereby guaranteeing the completeness of the algorithm.

3.4.4 Context-specific network reconstruction

The reconstruction problem involves computing a minimal consistent network from a global network and a 'core' set of reactions that are known to be active in a given context. Formally, given (i) a *consistent* global network $\{\mathcal{N}, S_{\mathcal{N}}\}$ with reaction set $\mathcal{N} = \{1, 2, \ldots, n\}$ and stoichiometric matrix $S_{\mathcal{N}}$, and (ii) a set $\mathcal{C} \subset \mathcal{N}$, the problem is to find the smallest set $\mathcal{A} \subseteq \mathcal{N}$ such that $\mathcal{C} \subseteq \mathcal{A}$ and the subnetwork $\{\mathcal{A}, S_{\mathcal{A}}\}$ induced by the reaction set $\mathcal{A}$ is consistent. (By $S_{\mathcal{A}}$ we denote the submatrix of $S_{\mathcal{N}}$ that contains only the columns indexed by $\mathcal{A}$.) This problem is known to be NP-complete (Acuña et al., 2009), suggesting that a practical solution should entail some approximation. (We note that Acuña et al. (Acuña et al., 2009) prove NP-completeness of this problem by noting that a special case involves $\mathcal{C}$ being the empty set, in which case the problem comes down to finding the smallest elementary mode of the global network, which, as the authors show, is NP-complete. However, this leaves open the case of a nonempty core set $\mathcal{C}$, since a solution to the minimal reconstruction problem need not constitute an elementary mode. We conjecture that the problem remains NP-hard when $\mathcal{C}$ is nonempty, but we are not pursuing this question here.)

Our approach hinges on the observation that a consistent induced subnetwork of the global network can be defined via a set of modes of the latter:

Theorem 1. *Let $\mathcal{V}$ be a set of modes of the global network $\{\mathcal{N}, S_{\mathcal{N}}\}$, and let $\mathcal{A} = \cup_{v \in \mathcal{V}}\, supp(v)$ be the union of the supports of these modes. The induced subnetwork $\{\mathcal{A}, S_{\mathcal{A}}\}$ is consistent.*

Proof. For each $v \in \mathcal{V}$, let $v_{\mathcal{A}}$ be the 'truncated' v after dropping all dimensions not indexed by $\mathcal{A}$.

Clearly, $S_{\mathcal{A}} v_{\mathcal{A}} = 0$, therefore each $v_{\mathcal{A}}$ is a mode in the reduced model $\{\mathcal{A}, S_{\mathcal{A}}\}$. By construction of $\mathcal{A}$, each reaction in $\mathcal{A}$ is in the support of some $v \in \mathcal{V}$, and hence also in the support of some mode $v_{\mathcal{A}}$ of the reduced model. □

This simple result allows one to cast the reconstruction problem as a search problem over sets of modes of the global network:

$$\begin{aligned} \min_{\mathcal{V}} \quad & card(\mathcal{A}) \\ \text{s.t.} \quad & \mathcal{A} = \bigcup_{v \in \mathcal{V}} supp(v) \\ & \mathcal{C} \subseteq \mathcal{A} \\ & \forall v \in \mathcal{V}: \quad S_{\mathcal{N}} v = 0, \ v \in \mathcal{B}\,. \end{aligned} \tag{NLP-3.8}$$

Note that this optimization problem involves searching for a set $\mathcal{V}$ of modes of $\{\mathcal{N}, S_{\mathcal{N}}\}$, such that the union of the support of these modes (the set $\mathcal{A}$) is a minimal-cardinality set that contains the core set $\mathcal{C}$. In order to practically make use of this theorem, one has to define a search strategy over modes. Next we discuss two possibilities. The first gives rise to an exact algorithm, but this algorithm does not scale to large networks. The second is a scalable greedy approach that gives rise to FASTCORE.

Exact reconstruction via mixed integer linear programming

Note that, without loss of generality, in NLP-3.8 we can restrict the search for $\mathcal{V}$ over all *elementary modes* of the global network, since the union of their supports covers the whole set $\mathcal{N}$. As we show next, if all elementary modes are available, NLP-3.8 can be cast as a mixed integer linear program (MILP) and solved exactly. This MILP is defined as follows. Let r be the number of elementary modes, and $\{m_1, \dots, m_r\}$ be a set of length-n binary vectors, where each vector m_j captures the support of elementary mode j (so, its ith entry is 1 if reaction i is included in elementary mode j, and 0 otherwise). Also, let $c = (c_1, \dots, c_n)$ be a length-n binary vector with $c_i = 1$ if reaction i is included in the core set $\mathcal{C}$, and $c_i = 0$ otherwise. The decision variables of the MILP are a length-n binary vector $x = (x_1, \dots, x_n)$ and a length-r real vector $y = (y_1, \dots, y_r)$. At an optimal solution of the MILP, the set $\mathcal{A}$ is defined as $\mathcal{A} = \{i \in \mathcal{N}: \ x_i^* = 1\}$.

Theorem 2. *When all elementary modes are available, the following MILP-3.9 solves NLP-3.8* *exactly.*

$$\begin{aligned}
\min_{x,y} \quad & \sum_i x_i \\
\textit{s.t.} \quad & x \geq \frac{1}{r} \sum_j m_j y_j \\
& c \leq \sum_j m_j y_j \\
& y \in [0,1] \\
& x \in \{0,1\} \,.
\end{aligned} \qquad \text{(MILP-3.9)}$$

Proof. By definition, $x_i^* = 1$ implies that reaction i will be included in the reconstruction $\mathcal{A}$, hence the objective minimizes the cardinality of $\mathcal{A}$. The sum $\sum_j m_j y_j^*$ is a vector whose support is the union of the supports of all selected elementary modes at the solution, where an elementary mode j is selected when $y_j^* > 0$. The first constraint $x \geq \frac{1}{r} \sum_j m_j y_j$ therefore imposes that the set $\mathcal{A}$ must contain the union of the supports of the selected elementary modes at the solution. (The factor $\frac{1}{r}$ ensures that $\frac{1}{r} \sum_j m_j y_j \leq 1$). Since superfluous reactions are removed by the minimization of $\sum_i x_i$ in the objective, the above implies that $\mathcal{A}$ is precisely the union of the supports of the selected elementary modes at the solution. The second constraint $c \leq \sum_j m_j y_j$ imposes that the core set must be included in the union of the supports of the selected elementary modes at the solution, and hence the core set must be included in $\mathcal{A}$. Therefore, all constraints of NLP-3.8 are satisfied at the optimal solution of MILP-3.9, and since the two programs minimize the same objective, an optimal solution of MILP-3.9 must be an optimal solution of NLP-3.8. □

Note, however, that MILP-3.9 does not scale to large networks, for the following reasons: First, it requires computing all elementary modes of the global network, which can be a very large number (Gagneur and Klamt, 2004). Second, the binary decision variables x_i index all reactions of the global network, and therefore MILP-3.9 needs to search over a binary hypercube of dimension n, which can be prohibitive for large n. Nonetheless, it is reassuring to know that an exact solution to the context-specific network reconstruction problem is possible, albeit with high complexity. Next we describe FASTCORE, an approximate greedy algorithm that scales much better to large networks, and we compare it to MILP-3.9 in the Results section.

Algorithm 1 The FASTCORE algorithm

Input: A consistent metabolic network model $\{\mathcal{N}, S_{\mathcal{N}}\}$ and a reaction set $\mathcal{C} \subset \mathcal{N}$.
Output: A consistent induced subnetwork $\{\mathcal{A}, S_{\mathcal{A}}\}$ of $\{\mathcal{N}, S_{\mathcal{N}}\}$ such that $\mathcal{C} \subseteq \mathcal{A}$.

```
function FASTCORE($\mathcal{N}$,$\mathcal{C}$)
    $\mathcal{J} \leftarrow \mathcal{C} \cap \mathcal{I}, \ \mathcal{P} \leftarrow \mathcal{N} \setminus \mathcal{C}$
    $flipped \leftarrow False, singleton \leftarrow False$
    $\mathcal{A} \leftarrow$ FINDSPARSEMODE($\mathcal{J}, \mathcal{P}, singleton$)
    $\mathcal{J} \leftarrow \mathcal{C} \setminus \mathcal{A}$
    while $\mathcal{J} \neq \emptyset$ do
        $\mathcal{P} \leftarrow \mathcal{P} \setminus \mathcal{A}$
        $\mathcal{A} \leftarrow \mathcal{A} \cup$ FINDSPARSEMODE($\mathcal{J}, \mathcal{P}, singleton$)
        if $\mathcal{J} \cap \mathcal{A} \neq \emptyset$ then
            $\mathcal{J} \leftarrow \mathcal{J} \setminus \mathcal{A}, \ flipped \leftarrow False$
        else
            if $flipped$ then
                $flipped \leftarrow False, \ singleton \leftarrow True$
            else
                $flipped \leftarrow True$
                if $singleton$ then
                    $\tilde{\mathcal{J}} \leftarrow \mathcal{J}(1)$    (the first element of $\mathcal{J}$)
                else
                    $\tilde{\mathcal{J}} \leftarrow \mathcal{J}$
                end if
                for each $i \in \tilde{\mathcal{J}} \setminus \mathcal{I}$ do
                    flip the sign of the $i$'th column of $S_{\mathcal{N}}$ and
                    swap the upper and lower bounds of $v_i$
                end for
            end if
        end if
    end while
    return $\mathcal{A}$
end function
```

Greedy approximation and the FASTCORE algorithm

An alternative search strategy for computing $\mathcal{V}$ in NLP-3.8 is a greedy approach, reminiscent of greedy heuristics for the related *set covering problem* (Chvátal, 1979). This is the idea behind the proposed FASTCORE algorithm: We build up the set $\mathcal{V}$ in a greedy fashion, by computing in each iteration a new mode of the global network. Further, as a means to approximately minimize $card(\mathcal{A})$, each added mode is constrained to have *sparse* support outside $\mathcal{C}$. This is implemented via L_1-norm minimization, which is a standard approach to computing sparse solutions to (convex) optimization problems (Boyd and Vandenberghe, 2004; Julius et al., 2008).

The overall FASTCORE algorithm is shown in Algorithm 1. The algorithm maintains a set $\mathcal{J} \subseteq \mathcal{C}$ that is initialized with the irreversible reactions in $\mathcal{C}$, and a 'penalty' set $\mathcal{P} = (\mathcal{N} \setminus \mathcal{C}) \setminus \mathcal{A}$ that contains all reactions outside $\mathcal{C}$ that have not been added yet to the set $\mathcal{A}$. Each iteration

Algorithm 2 The FINDSPARSEMODE function

Input: A set $\mathcal{J} \subseteq \mathcal{C}$, a penalty set $\mathcal{P} \subseteq \mathcal{N} \setminus \mathcal{C}$, and the $singleton$ flag.
Output: The support of a mode that is dense in $\mathcal{J}$ and sparse in $\mathcal{P}$.

```
function FINDSPARSEMODE($\mathcal{J}, \mathcal{P}, singleton$)
    if $\mathcal{J} = \emptyset$ then
        return $\emptyset$
    end if
    if $singleton$ then
        $v^* \leftarrow$ LP-3.7 on set $\mathcal{J}(1)$
    else
        $v^* \leftarrow$ LP-3.7 on set $\mathcal{J}$
    end if
    $\mathcal{K} \leftarrow \{i \in \mathcal{J} : v_i^* \geq \varepsilon\}$
    if $\mathcal{K} = \emptyset$ then
        return $\emptyset$
    end if
    $v^* \leftarrow$ LP-3.10 on sets $\mathcal{K}, \mathcal{P}$
    return $\{i \in \mathcal{N} : |v_i^*| \geq \varepsilon\}$
end function
```

adds to the set $\mathcal{A}$ the support of a mode that is dense in $\mathcal{J}$ (i.e., contains as many nonzero fluxes in $\mathcal{J}$ as possible) and sparse in $\mathcal{P}$ (i.e., contains as many zero fluxes in $\mathcal{P}$ as possible), computed by the function FINDSPARSEMODE (Algorithm 2). This function first applies an LP-3.7 to compute an active subset $\mathcal{K}$ of $\mathcal{J}$, and then it applies the following L_1-norm minimization LP constrained by the set $\mathcal{K}$:

$$\begin{aligned} \min_{v,z} \quad & \sum_{i \in \mathcal{P}} z_i \\ \text{s.t.} \quad & v_i \in [-z_i, z_i] && \forall i \in \mathcal{P},\ z_i \in \mathbb{R}_+ \\ & v_i \geq \varepsilon && \forall i \in \mathcal{K} \\ & S_{\mathcal{N}}\, v = 0 && v \in \mathcal{B}\,. \end{aligned} \tag{LP-3.10}$$

The LP-3.10 minimizes $\sum_{i \in \mathcal{P}} |v_i|$, the L_1 norm of fluxes in the penalty set $\mathcal{P}$ (expressed via epigraphs), subject to a minimum flux constraint on the set $\mathcal{K}$. However, some care is needed to preempt false negative solutions arising from the minimization of L_1 norm in LP-3.10. For example, suppose in the network of Figure 3.3 that the global network comprises all reactions except A↔B, and $\mathcal{C} = \mathcal{J} = \mathcal{K} = \{6\}$ and $\mathcal{P} = \{1, 3, 4, 5\}$. In this case, LP-3.10 could settle to a solution $[v_1, v_3, v_4, v_5, v_6] = [\frac{\varepsilon}{2}, \varepsilon, 0, 0, \varepsilon]$. The flux v_1, being below ε, would be treated as zero by FINDSPARSEMODE, in which case the reaction →2A would be erroneously excluded from the reconstruction. A simple way to avoid this is to use a scaled version of ε (we used $10^5\varepsilon$) in the second constraint of LP-3.10, with an equal scaling of all flux bounds in $\mathcal{B}$.

The FASTCORE algorithm first goes through the $\mathcal{I} \cap \mathcal{C}$ reactions (step 2), and then through the $\mathcal{R} \cap \mathcal{C}$ ones (and eventually through each individual reversible reaction in the core set; when $singleton = True$). The $flipped$ variable ensures that a reversible reaction is tested in both the forward and negative direction. The algorithm terminates when all reactions in $\mathcal{C}$ have been added to $\mathcal{A}$, which is guaranteed since in the main loop the set $\mathcal{J}$ never expands (step 10) and the global network is consistent. Note that FASTCORE has no free parameters besides the flux threshold ε.

The FASTCC algorithm for detecting the consistent part of an input network (see previous section) can be viewed as a variant of FASTCORE$(\mathcal{N}, \mathcal{N})$ in which the steps 10–14 of FindSparseMode are omitted (and there is no $\mathcal{P}$ set). It is easy to verify that FASTCC is complete, in the sense that it will always report consistency if the network is consistent, and if not, it will reveal the set of blocked reactions.

3.4.5 Related work

Table 3.1. Summary of the main characteristics of GIMME (Becker and Palsson, 2008), MBA (Jerby et al., 2010), iMAT (Zur et al., 2010), mCADRE (Wang et al., 2012), INIT (Agren et al., 2012), and FASTCORE (this paper) reconstruction algorithms.

	GIMME	MBA	iMAT	mCADRE	INIT	FASTCORE
Optimization	LP	MILP	MILP	MILP	MILP	LP
Computational cost	low	high	high	high	high	low
Function required	yes	no	no	yes	yes	no
Omics required	yes	optional	yes	yes	yes	no
Code available	yes	yes	yes	yes	no	yes

Several algorithms have been published in the last years for extracting condition-specific models from generic genome-wide models like Recon 1. Among them, mCADRE (Wang et al., 2012), INIT (Agren et al., 2012), iMAT (Zur et al., 2010), MBA (Jerby et al., 2010) and GIMME (Becker and Palsson, 2008) are the most commonly used (see Table 3.1 for an overview). Here we provide a short outline of the different algorithms, and refer to (Blazier and Papin, 2012) for a more extensive overview. For GIMME, iMAT, and MBA, we briefly discuss some notable differences to FASTCORE.

GIMME (Becker and Palsson, 2008) takes as input microarray data and a biological function to optimize for, such as biomass production. GIMME starts by removing reactions with associated expression levels below a user-defined threshold, and then it optimizes for the specified biological function using linear programming. In case the pruning steps compromise the input biological function, GIMME reintroduces some previously removed reactions that are in minimal disagreement with the expression data. Since GIMME has not been designed to include all core reactions in the solution (as FASTCORE does), the reconstructions obtained by GIMME and FASTCORE can differ significantly: Running the *createTissueSpecific* function of the COBRA toolbox on a set of liver core reactions (see Section 3.5) treating them as expressed reactions (and adding a biomass reaction (Wang et al., 2012) and a sink reaction for glycogen to be used as optimization function), only about 50% of the core reactions of the GIMME model were consistent at the solution. A fairer comparison would require adapting FASTCORE to explicitly deal with omics data, which is outside the scope of the current work.

iMAT (Zur et al., 2010) was originally designed for the integration of transcriptomic data. iMAT optimizes for the consistency between the experimental data and the activity state of the model reactions. iMAT tries to include modes composed of reactions associated to genes with high expression value, and therefore a threshold needs to be chosen to segregate between low, medium, and highly expressed genes. The computational demands of iMAT are high due to the repeated use of mixed integer linear programming. As with GIMME, direct comparison of iMAT to FASTCORE is problematic. Nevertheless, we applied iMAT (own implementation) on the liver problem (see Section 3.5), by setting the liver core reactions to RH (reaction high) and all non-core reactions to RL (reaction low). iMAT determined 549 core reactions as active, while 182 and 338 reactions were classified as undetermined and inactive, respectively. This means that about 50% of the core reactions were lost during iMAT model building. As with GIMME, this demonstrates the difficulty of directly comparing FASTCORE to algorithms that optimize different objectives.

mCADRE (Wang et al., 2012) is similar to MBA, except that the pruning order is not random, but it depends on the tissue-specific expression evidence and weighted connectivity to other reactions of the network. Reactions that are associated to genes that are never tagged as expressed and which are not connected to reactions associated to highly expressed genes are first evaluated in the pruning step. Reactions are effectively removed if the removal does not impair core reactions and metabolic functions to carry a flux (mCADRE removes core reactions if the core/non-core reaction ratio is below a user-given threshold). mCADRE uses mixed integer linear programming and therefore it does not scale up to large networks (but it is in general faster than MBA).

INIT (Agren et al., 2012) uses data retrieved from public databases in order to assess the presence of a certain reaction-respective metabolites in the cell type of interest. INIT uses mixed integer linear programming to build a model in which all reactions can carry a flux. Contrary to other algorithms, INIT does not rely on the assumption of a steady state, but it allows small net accumulation of all metabolites of the model.

The closest algorithm to FASTCORE is the MBA algorithm of Jerby et al. (Jerby et al., 2010). MBA takes as input two core sets of reactions, and it searches for a consistent network that contains all reactions from the first set, a maximum number of reactions from the second set (for a given tradeoff), and a minimal number of reactions from the global network. (FASTCORE can be easily adapted to work with multiple core sets, by introducing a set of weights that reflect the confidence of each reaction to be active in the given context, and adding appropriate regularization terms in the objective functions of LP-3.7 and LP-3.10 that capture the given tradeoff. We will address this variant in future work.) Both FASTCORE and MBA involve a search for a minimal consistent subnetwork, however the search strategy of FASTCORE is very different to MBA: Whereas FASTCORE iteratively expands the active set $\mathcal{A}$ starting with $\mathcal{A} = \emptyset$, MBA starts with $\mathcal{A} = \mathcal{N}$ and iteratively prunes the set $\mathcal{A}$ by checking whether the removal of each individual reaction (selected in random order) compromises network consistency. As the pruning order affects the output model, this step of MBA is repeated multiple times. MBA builds a final model by adding one by one non-core reactions with the highest presence rate over all pruning runs, and it stops when a consistent final model is obtained. Due to the multiple pruning runs, MBA has very high computational demands. Consistency testing in MBA is carried out with the CMC algorithm that is based on LP-3.3, as explained earlier. Hence, FASTCORE's search strategy differs to MBA in two key aspects: First, consistency testing in FASTCORE involves the maximization of flux cardinality (LP-3.7) instead of sum of fluxes (LP-3.3), which results in fewer LP iterations. Second, the search for compact solutions in FASTCORE involves L_1-norm minimization instead of pruning. The advantage of the former is that it can be encoded by a single LP, resulting in significant overall speedups (see Section 3.5).

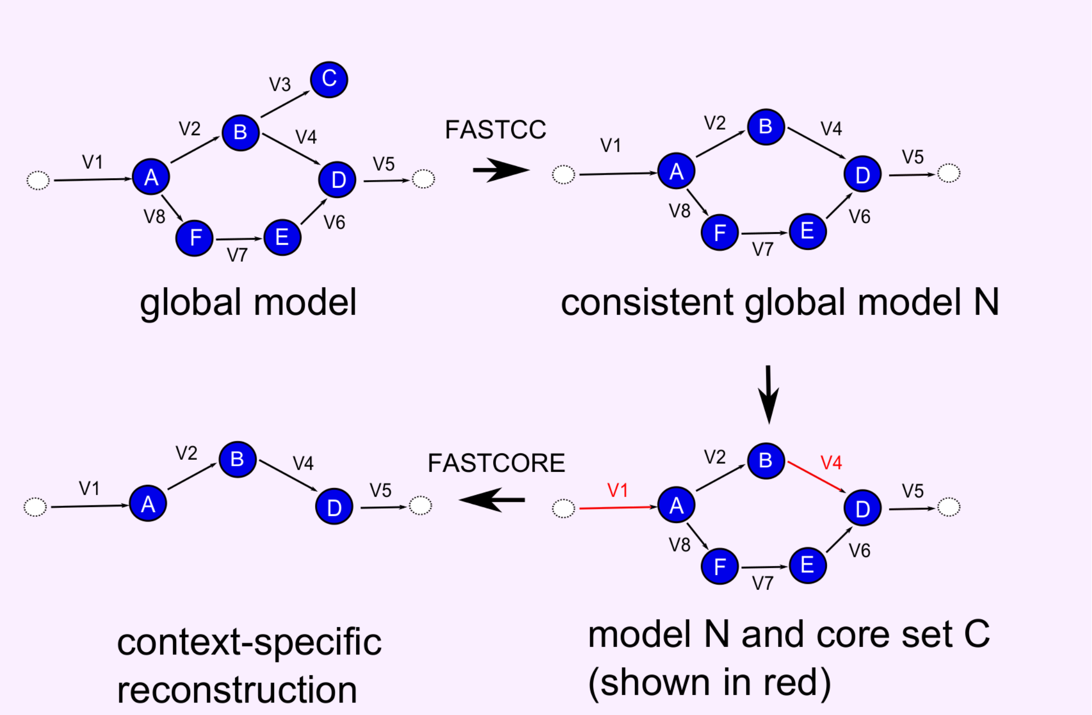

V1
V2
V3
V4
V5
V6
V7
V8
A
B
C
D
E
F
global model
FASTCC
consistent global model N
FASTCORE
context-specific reconstruction
model N and core set C (shown in red)

Figure 3.4. Fastcore pipeline: Flowchart of the overall pipeline for generating consistent context-specific models.

3.5 Results

Generic metabolic reconstructions like Recon 2 are inconsistent models as they contain reactions that are not able to carry nonzero flux due to gaps in the network (see next section). The first step towards obtaining a consistent context-specific reconstruction is therefore to extract the consistent part of a global generic model. This can be achieved by FASTCC or other similar methods (see Section 3.4.2). The consistent global model serves then as input for the context-specific reconstruction with FASTCORE. In Figure 3.4 we show a flowchart of the overall pipeline.

We report results on two sets of problems, the first involving consistency verification of an input model, and the second involving the reconstruction of a context-specific model from an input model and a core set of reactions. The FASTCORE algorithm was implemented in the COBRA toolbox (Schellenberger et al., 2011a), using Matlab 2013a and the IBM CPLEX solver (version 12.5.0.0). Test runs were performed on a standard 1.8 GHz Intel Core i7 laptop with 4 GB RAM running Mac OS X 10.7.5. In all experiments we used flux threshold ε =1e-4. The software is available from `bio.uni.lu/systems_biology/software/`

Table 3.2. Comparing FASTCC to fastFVA (Gudmundsson and Thiele, 2010) and CMC (Jerby et al., 2010) on four input models.

	c-Yeast		c-Ecoli		c-Recon1		c-Recon2	
	# LPs	time*	# LPs	time	# LPs	time	# LPs	time
fastFVA	2408	3	3436	3	4938	9	11668	207
CMC	18	0.5	25	1	49	2	42	11
FASTCC	7	0.1	2	0.2	9	0.4	19	5

*in seconds

3.5.1 Consistency testing

In the first set of experiments we applied FASTCC, the consistency testing variant of FASTCORE, for consistency verification of four input models, and compared it against the FastFVA algorithm of Gudmundsson and Thiele (Gudmundsson and Thiele, 2010), and an own implementation (based on FASTCC but with LP-3.3 replacing LP-3.7) of the CMC algorithm of Jerby et al. (Jerby et al., 2010). We also tested the FVA algorithm of the *reduceModel* function of the COBRA toolbox (Schellenberger et al., 2011a), and the MIRAGE algorithm of Vitkin and Shlomi (Vitkin and Shlomi, 2012), but we do not include them in the results as they performed worse than the reported ones. The input models were the following:

- c-Yeast ($\#\mathcal{N} = 1204$), the consistent part of a yeast model (Zomorrodi and Maranas, 2010).

- c-Ecoli ($\#\mathcal{N} = 1718$), the consistent part of an *E. coli* model (Orth et al., 2011). (Here we set to 1000 the upper bounds of all fluxes that were fixed to zero, and we multiplied all bounds by 1000 to avoid numerical issues.)

- c-Recon1 ($\#\mathcal{N} = 2469$), the consistent part of Recon 1 (Duarte et al., 2007). (Recon 1 was found to contain 1273 blocked reactions.)

- c-Recon2 ($\#\mathcal{N} = 5834$), the consistent part of Recon 2 (Thiele et al., 2013). (Recon 2 was found to contain 1606 blocked reactions.)

The results are shown in Table 3.2. FASTCC is faster and it uses much fewer LPs than the other two algorithms. We note that fastFVA is based on an optimized Matlab/C++ implementation with LP warm-starts, while FASTCC is based on standard Matlab. These results confirm

Table 3.3. Comparing FASTCORE to MBA (Jerby et al., 2010) on liver model reconstruction from c-Recon1.

	liver core set (C=1069)				strict liver core set (C=1083)			
	A	**IR**	**LPs**	**time**	**A**	**IR**	**LPs**	**time**
MBA	1826	1573	72279	7383	1888	1630	71546	6730
FASTCORE	1746	1546	20	1	1818	1627	20	1

the appropriateness of flux cardinality (LP-3.7) as a metric for network consistency testing, in agreement with the theoretical analysis and the discussions above.

3.5.2 Reconstruction of a liver model

In the second set of experiments, we used the FASTCORE algorithm to reconstruct a liver specific metabolic network model from the consistent part of Recon 1 (c-Recon1, $\#\mathcal{N} = 2469$), and we compared against an own implementation of the MBA algorithm of Jerby et al. (Jerby et al., 2010). We applied the two algorithms in two settings. The first setting involves the liver specific input reaction set of Jerby et al. (Jerby et al., 2010), which is based on 779 ‘high’ core and 290 ‘medium’ core reactions (the latter set is supported by weaker biological evidence than the former). To allow a comparison with FASTCORE, we defined a single core set as the union of the high and medium core reaction sets, and we applied the two algorithms on this core set. The second setting uses the ’strict’ liver model of Jerby et al. (Jerby et al., 2010), which contains 1083 high core reactions and no medium core reactions, and therefore allows a direct comparison with FASTCORE.

The results for the two settings are shown in Table 3.3. We note that for MBA, the reported number of LPs and the runtime refer to a single pruning iteration of the algorithm, whereas the size of each reconstruction refers to the final model after 1000 pruning iterations. In both settings, FASTCORE is several orders of magnitude faster than MBA, achieving a full reconstruction of a liver specific model in about one second, using a much smaller number of LPs. As MBA employs a greedy pruning strategy for optimization, the number of LPs that it uses and its total runtime can be very high, as also indicated by Wang et al. (Wang et al., 2012) who reported runtime of a single pruning pass of MBA in the order of 10 hours on a 2.34 GHz CPU computer.

The reconstructed models by FASTCORE are also more compact than those obtained by

MBA, with a difference of 70-80 non-core reactions. For the standard liver model, 1687 out of the 1746 reactions (96%) of the FASTCORE reconstruction appear also in the MBA reconstruction, whereas for the strict liver model the common reactions are 1739 out of 1818 (95%). The two algorithms turned out to use alternative transporters to connect the core reactions: In the standard liver model, 46 out of 59 reactions that are present exclusively in the FASTCORE reconstruction are transporter reactions or other reactions which are not associated to a specific gene and thus are not sufficiently supported in the core set, whereas in MBA the corresponding numbers are 116 out of 139 reactions. Note that both MBA and FASTCORE try to minimize the number of added non-core reactions in order to obtain a compact consistent model. The above difference in the number of added non-core reactions between MBA and FASTCORE is the result of the different optimization approaches taken by the two algorithms, and no biological relevance should be attributed to each reconstruction other than the one implied by the makeup of the core set. From this point of view, FASTCORE performs in general better than MBA, as it tends to add fewer unnecessary reactions.

We also compared the solutions of FASTCORE to those of MILP-3.9, using core sets that are randomly generated from a consistent subset of *E. coli core* (Orth et al., 2010). This is a small model with $\#\mathcal{N} = 53$ and 414 elementary modes (unfortunately, the dependence of the MILP-3.9 solver to the number of elementary modes did not allow testing larger models). In Figure 3.5 we show the size of the reconstructed models (mean values) obtained with the exact MILP solver vs. FASTCORE, as a function of the size of the core set. FASTCORE is capable of obtaining very good approximations to the optimal solutions, which improve with the size of the core set.

To evaluate FASTCORE's performance in correctly identifying liver reactions, we performed repeated random sub-sampling validation in which FASTCORE was used to reconstruct the liver metabolism based on a reduced, randomly selected 'subcore' set of 80% of the original core reactions. As in (Jerby et al., 2010), we wanted to test whether FASTCORE is able to recover a significant number of the 20% left-out core reactions. To test for the enrichment of the left-out core reactions in the reconstructed model, we used a hypergeometric test, in which the total population is defined by all non-subcore reactions in the global network, the number of draws is defined as the number of non-subcore reactions included in the reconstruction, and the left-out core reactions are the 'successes'. Under the null-hypothesis that there is no enrichment for the left-out core reactions when reconstructing the liver model based on the subcore set, we

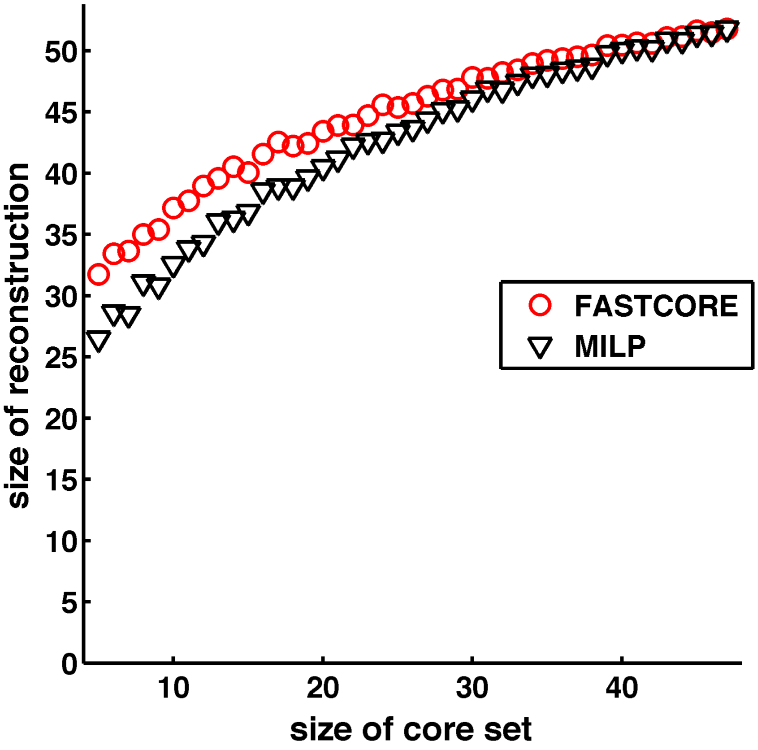


Figure 3.5. Comparing FASTCORE to an exact MILP solver on a small *E. coli* model (Orth et al., 2010). Shown are mean values of sizes of reconstructed models (over 50 repetitions for each core set; standard deviations were small and are omitted to avoid clutter) as a function of the size of the core set. FASTCORE computes near-optimal reconstructions, which improve with the size of the core set.

can compute a p-value for including at least the number of observed left-out core reactions in the reconstruction. We repeated this random sub-sampling procedure 500 times and computed the corresponding p-values. The median of these p-values was 0.0025, indicating the ability of FASTCORE to capture liver-specific reactions that were included in the original core set.

As argued above, the reconstructions obtained by FASTCORE need not optimize for cellular functions other than the ones implied by the composition of the input core set, and it is an interesting research question how to modify FASTCORE so that it can explicitly capture functional requirements in its reconstructions. Nevertheless, it is of interest to test whether the current version of FASTCORE can produce reconstructions that *are* functionally relevant, perhaps for slight variations of the core set. To this end, as in (Jerby et al., 2010), we checked whether the (standard) liver model reconstructed by FASTCORE can perform gluconeogenesis from glucogenic amino acids, glycerol, and lactate (altogether 21 metabolites). If not yet included, transporters from the extracellular medium to the cytosol were added to the model (glycerol, glutamate, glycine, glutamine, and serine). This was necessary as the transport reactions were not sufficiently supported in the core set. This ‘extended’ liver model was able to convert 17/21 metabolites (vs 12/21 metabolites of the non-extended model). The extended liver model was then used to simulate the liver disorders hyperammonemia and hyperglutamenia, which affect the capacity to metabolize dietary amino acids into urea (Jerby et al., 2010). Loss of function mutations of three enzyme-coding genes, argininosuccinate synthetase (ASS), argininosuccinate lyase (ASL), and ornithine transcarbamylase (OTC) were identified in patients suffering from these disorders. The rates of the reactions controlled by the three genes were fixed to 500, 250, or zero, to mimic the healthy homozygote (no mutation), heterozygote (loss of one allele), and the complete loss of function, respectively. To allow for a comparison with the experimental study of Lee et al. (Lee et al., 2000) where labeled 15N-glutamine was administrated to patients suffering from inborn errors affecting the three genes, we explicitly shut down the influx of other potential nitrogen sources in the liver model, thereby simulating only the uptake and metabolism of glutamine. By allowing the influx of only one nitrogen source, the fate of the latter could be determined exactly in the model. The ratio of urea secretion level over glutamine absorption was computed by sampling over the feasible space (Price et al., 2004). In accordance with the wet lab observations (Lee et al., 2000), the severity of the disorders, characterized by the mean urea over glutamine ratio, increased with the level of loss of function of the three genes ASS, ASL, and OTC (see Figure 3.6). Null patients showed no native production of urea. Overall, the

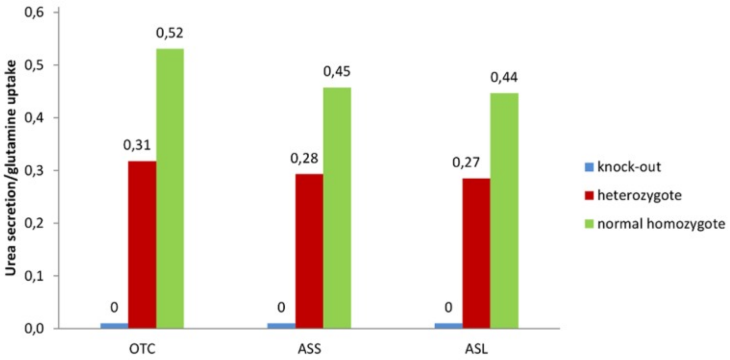


Figure 3.6. Mean urea/glutamine ratio in the extended liver model obtained by FASTCORE. Healthy (normal homozygote), partial (heterozygote) and full knock-out cases. See text for details.

ratios predicted by the FASTCORE model faithfully match the experimentally observed ones (Lee et al., 2000). (The corresponding ratios reported by Jerby et al. when using the MBA algorithm (Jerby et al., 2010) matched less well the experimental observations, probably because of the cross-feeding of nitrogen to urea from multiple nitrogen sources. By running the above procedure on the MBA model, we noticed that both models attained comparable urea / glutamine flux ratios.) To summarize, the above experiments demonstrate that, by an informed choice of the core set and influx bounds, FASTCORE can indeed give rise to functionally relevant models.

3.5.3 Reconstruction of a murine macrophage model

We also used the FASTCORE algorithm to build a cell-type specific murine macrophage model from the consistent part of Recon1bio (comprising $\#\mathcal{N} = 2474$ reactions). Recon1bio ($\#\mathcal{N} = 3745$) is a modified Recon 1 model that contains three extra reactions (biomass, NADPOX, and a sink reaction to balance the glycogenin self-glucosylation reaction) (Bordbar et al., 2012). We used a core set comprising 300 (out of 382) proteomics derived Raw264.7 macrophage reactions, as described by Bordbar et al. (Bordbar et al., 2012). (The remaining 82 reactions could not be added to the core set as they are situated in an inconsistent region of Recon 1 and therefore carry a permanent zero net flux.) For their macrophage reconstruction, Bordbar et al. used, among other methods, GIMMEp—a variant of the GIMME algorithm (Becker and Palsson, 2008) that is similar to the MBA algorithm—and they obtained a network model containing 1026 intracellular reactions. Our main interest was to investigate whether FASTCORE can obtain a functional network that is at least as compact as the one obtained with GIMMEp. FASTCORE generated (in about one second and using 11 LPs) a consistent network model of 953 reactions, 831 of which are intracellular reactions. This is a much more compact model than the one obtained with GIMMEp.

3.6 Discussion

FASTCORE is a generic algorithm for context-specific metabolic network reconstruction from genome-wide metabolic models, and it was motivated by requirements of fast computation and compactness of the output model.

The key advantage of having a fast reconstruction algorithm is that it permits the execution of multiple runs in order to optimize for extra parameters or test different core sets extracted from the input data (Folger et al., 2011; Wang et al., 2012). For example, when working with gene expression data, the definition of the core set may depend on the threshold used to segregate between high expression genes (core reactions) and low expression genes (non-core reactions) (Becker and Palsson, 2008). As the choice of threshold is rather arbitrary, a practical approach could involve evaluating the robustness of the output model as a function of the chosen threshold. FASTCORE can perform this analysis in a few minutes, whereas for the same problem other algorithms would need hours or days. (Algorithms like GIMME or GIMMEp that require manual curation and assembly of subnetworks, would also fail in this kind of task.) Another example where fast computation is imperative is cross-validation. In the current study (see Section 3.5) we ran a random sub-sampling validation procedure 500 times, an operation that took a few minutes with FASTCORE but that would barely be manageable with other reconstruction algorithms. Other examples where fast computation is important are time-course experiments or experiments involving different patients or conditions (Jerby and Ruppin, 2012). There FASTCORE could more easily identify differential models over time and/or input conditions.

Compactness is a key concept in various research areas of biology, such as the minimal genome (Morowitz, 1984; Maniloff, 1996). Notwithstanding, the requirement of model compactness seems to be in disagreement with the observation that biological systems are fairly redundant and this redundancy serves a specific purpose, namely, the fast adaptation to changes in the environment. Alternative pathways that perform similar functions are known to be expressed in different environmental conditions, allowing for instance to metabolize another type of sugar when glucose is not available (Suckow et al., 1996). At any rate, the pursuit of compactness in metabolic network reconstruction need not be in conflict with the notion of redundancy. Alternative pathways will be included in a reconstructed model as long as ’redundant’ reactions that are supported by biological evidence are included in the core set.

FASTCORMICS

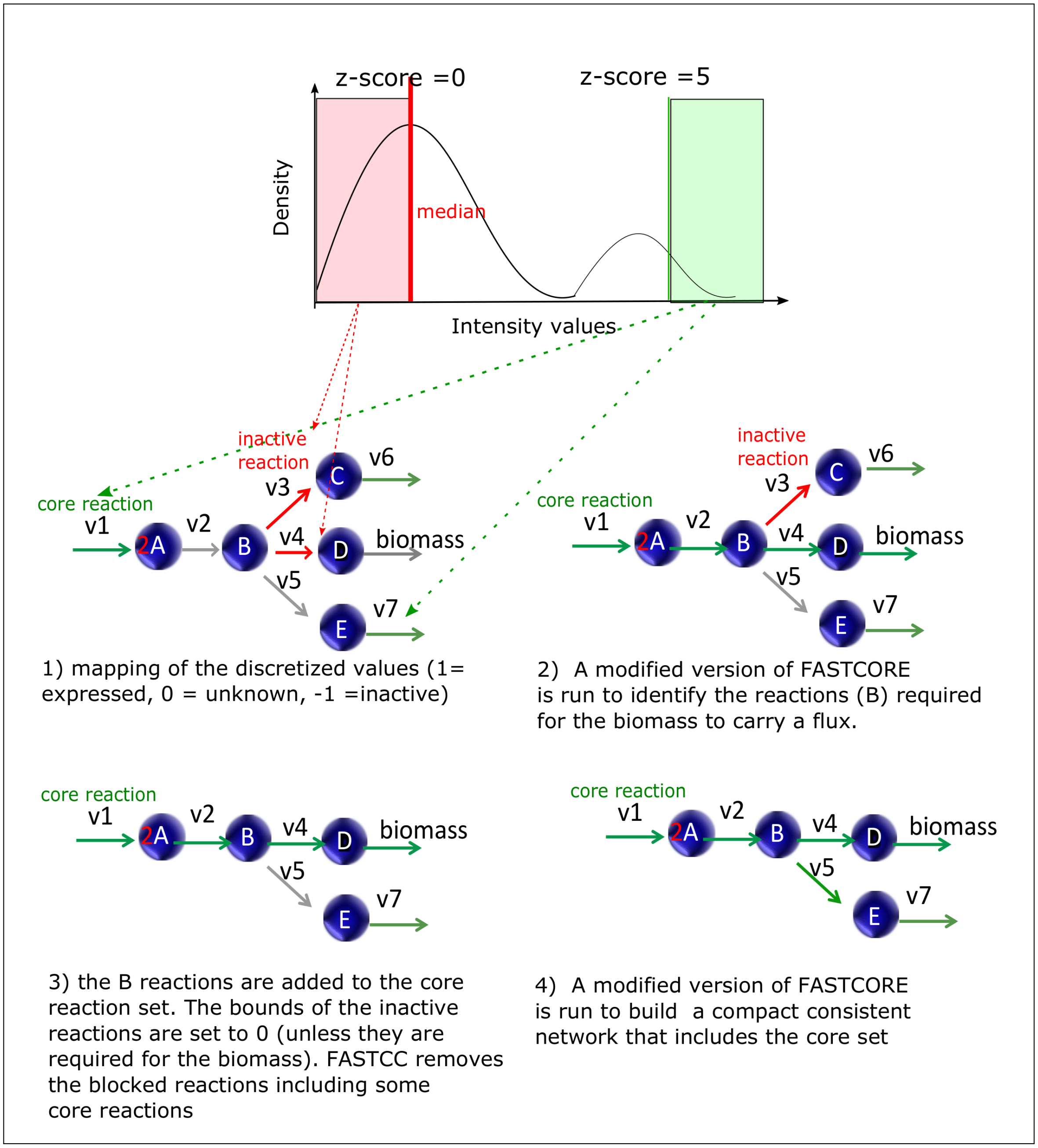
The following chapter has been originally published as supplementary file of the ***Integrated metabolic modelling reveals cell-type specific epigenetic control points of the macrophage metabolic network*** *paper. The text was slightly modified and some tables and figures were removed to form a coherent chapter.*

Contents

4.1 Summary and contributions

FASTCORMICS is a version of FASTCORE for context-specific model building via the integration of microarray data. The FASTCORMICS workflow (Pires Pacheco and Sauter, 2014; Pires Pacheco et al., 2015a) was developed by me with the help of Professor Sauter to address the problem of probe effects, non-negligible amounts of noise that affect in different proportions each probe set and that do not allow an direct comparison of the absolute intensity values within probe sets in a microarray experiment. FASTCORMICS shares the specifications of FASTCORE, with the exception that FASTCORMICS does not require a set of core reactions as input. The establishment of a core set is automatically performed by the FASTCOMICS workflow

(Figure **??**) based on previous knowledge on the intensity distribution for each probe set across thousands of different conditions obtained fron the BARCODE discretization workflow (Zilliox and Irizarry, 2007; McCall et al., 2011). Genes in the context of interest with intensities of expression 5 standard deviations $(Z - score > 5)$ apart from the lowest intensity mode observed for the same genes (corresponding to a non-expressed state), are considered as expressed. Whereas genes with intensities values below the median intensity level are considered as not expressed. The z-scores were mapped to the reactions according to the Gene-Protein-Reaction (GPR) rules. Reactions controlled by expressed genes form the core set whereas the bounds of reactions controlled by unexpressed genes are set to zero. Transporter reactions are removed from the core set as the genes controlling this types of reactions tend to be more promiscuous but are not penalized during the reconstruction. The designing of the FASTCORMICS workflow and the redaction of the supplementary files were performed by me and Prof. Sauter. The implementation of the workflow, the experiments performed to validate the workflow and the figures 4.2 and 4.3, the tables 4.1, 4.2, 4.3 and 4.4 were made by me under the supervision of Prof. Sauter.



1) mapping of the discretized values (1= expressed, 0 = unknown, -1 =inactive)

2) A modified version of FASTCORE is run to identify the reactions (B) required for the biomass to carry a flux.

3) the B reactions are added to the core reaction set. The bounds of the inactive reactions are set to 0 (unless they are required for the biomass). FASTCC removes the blocked reactions including some core reactions

4) A modified version of FASTCORE is run to build a compact consistent network that includes the core set

Figure 4.1. FASTCORMICS workflow. The FASTCORMICS workflow uses Barcode as preprocessing step. Barcode considers the intensity distribution across thousands of arrays to identify the mode associated to the lowest expression values. The median (red line) of this mode (corresponding to a z-score of 0) determines the unexpression threshold. Expression values below this threshold are assigned a value of -1. Expression values that are 5 z-scores higher than the unexpression threshold are associated with a z-score of 1. Expression in between of these two thresholds are assigned a value of 0. These discretized values are then mapped to the network. The user can defined a limited number of reactions such as the biomass reaction that have to be included in the output model. A modified version of FASTCORE that include beside the core reactions set, the non-core reactions set, and a third set of reactions corresponding to the unpenalized set is run. The biomass reaction (or required exchange reactions) is set as core whereas the reactions that are associated with a score of 1 (z-score >5) are set to the unpenalized set. The rational is to choose among alternative pathways the one that contains more core reactions, and by such the one that is the best supported by the data. The output of the modified version of FASTCORE contains the indices of the reactions that are required for the biomass reaction to carry a flux. These reactions are added to the core set composed of reaction with discretization score of 1. If among these reactions, reactions were previously associated with a score of -1, this score is changed to 1. So that they are no longer considered as inactive.
In the next step, the bounds of reactions with a score of -1 are set to zero. FASTCC is run to remove these reactions along with reactions that are no longer able to carry a flux due to the change of the bounds. A new consistent model is build. The modified version of FASTCORE is run with as core set, the reactions associated with a score 1. Transporter reactions that are associated to a score of 1 are moved to the unpenalized set as controlled to promiscuous genes. The red and green squares represent expression intensity levels that are considered as active and inactive, respectively. The red arrows symbolize inactive and the green ones stand for active reactions.

4.2 Introduction

The prospect of studying cell-type specific metabolism under numerous conditions or for example patient-specific metabolism in a diagnostic setting requires the capacity for fast creation of high-quality and robust metabolic models based on available data such as gene expression data. Recently we proposed an algorithm for the fast reconstruction of compact context-specific metabolic networks (FASTCORE) that reduces the reconstruction time of context-specific networks to the order of seconds (Vlassis et al., 2014). In order to adapt FASTCORE for the integration of transcriptomics data from microarrays, we have developed a new workflow named FASTCORMICS.

4.3 Methods and Results

As inputs FASTCORMICS (Figure 4.2) requires microarray data and a GENRE of the organism of interest. Like FASTCORE, FASTCORMICS is devoid of arbitrary parameter settings and has a low computational demand with overall building times in the order of a few minutes. FASTCORMICS pre-processes microarray data with the discretization tool Barcode (Zilliox and Irizarry, 2007). Barcode uses prior knowledge on the intensity distribution of each probe set for a given microarray platform to segregate between expressed and non-expressed genes. The preprocessing step with Barcode allows circumventing setting an arbitrary expression threshold to segregate between expressed and non-expressed genes, which is still commonly done (Becker and Palsson, 2008; Zur et al., 2010; Folger et al., 2011). As such a threshold is arbitrary and critical for the output metabolic models since due to this threshold complete branches, alternative pathways, or subsystems might be included or excluded, thereby significantly changing the functionalities of the model. Furthermore, Barcode shows a better correlation between predicted expression and protein expression than competing discretization methods and decreases batch and lab-effects that affect measurements (Zilliox and Irizarry, 2007).

4.3.1 Validation 1: *In silico* Knock-out experiments

To validate FASTCORMICS we performed an essentiality assay on two generic cancer models that are based on Recon 1 and Recon 2 (cancer1 and cancer2, respectively) and generated by the FASTCORMICS workflow using existing microarray expression data from 59 cancer cell lines (Shankavaram et al., 2009; Pfister et al., 2009). The first model (cancer1) is composed of 810 reactions and is therefore bigger than the cancer model previously derived by Folger et al. (772 reactions) (Folger et al., 2011). The second model (cancer2) is composed of 1322 reactions. All reconstructed models are available in SBML format (Additional File S6 of the original paper). The assays performed on cancer1 and cancer2 predict 183 and 78 genes essential for cell growth, respectively (Table 4.1). The predicted essential genes were compared to a list of 8000 genes ranked for essentiality by Luo et al. using a shRNA knock down screen in several different cancer cell lines (Luo et al., 2008) to assess the predictive power of the FASTCORMICS models. In general, metabolic genes are slightly overrepresented in the top of the list as shown by Folger et al (Folger et al., 2011; Luo et al., 2008), suggesting that metabolic genes are more essential than non metabolic genes on average. As expected, the Recon 1 and Recon

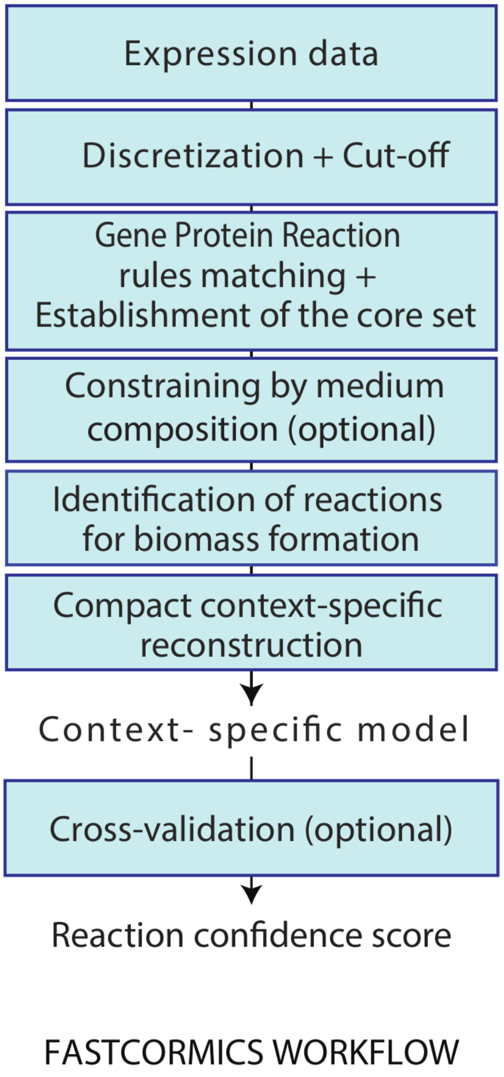

Expression data
Discretization + Cut-off
Gene Protein Reaction rules matching + Establishment of the core set
Constraining by medium composition (optional)
Identification of reactions for biomass formation
Compact context-specific reconstruction
Context- specific model
Cross-validation (optional)
Reaction confidence score
FASTCORMICS WORKFLOW

Figure 4.2. FASTCORMICS workflow. Microarray data are discretized with Barcode in expressed (z-score > 5) and unexpressed genes (z-score < 0) that are mapped to the input model according to the Gene-Protein-Reactions rules. The FASTCORE core set is composed of reactions under the control of Barcode supported genes. Optionally, the model can be constrained in function of the medium composition and a biomass function or the requirement to produce given metabolites can be added to the model. A modified version of FASTCORE, that allows the definition of a set of non-penalized reactions (in this study: Barcode-supported core reactions) is run. The modified version of FASTCORE forces the biomass function to carry a non-zero flux while penalizing the inclusion of non-core reactions. The output of the modified FASTCORE is then added to the core set and the modified FASTCORE is run again, now forcing all core reactions to carry a flux while penalizing non-core reactions. Transporters are removed from the core set, but are not penalized as explained in the main text. Finally a left-out cross-validation experiment can optionally be run to assign a confidence score to each reaction of the context specific output model.

Output model	Generic model	Contextualization method	Time	Size	Essential genes	KS test p-value	Permutation p-value
Recon 1	Recon 1	None		2471	14	0.7623	0.7212
Recon2	Recon 2	None		5317	4	0.0231	0.0210
Medium constrained Recon1	Recon 1	Medium constrained		1922	78	0.1908	0.1444
Medium constrained Recon 2	Recon 2	Medium constrained		4246	32	0.8260	0.7919
GIMME cancer model	Recon 1	GIMME	2497 sec	1749	69	0.0814	0.0465
PRIME cancer models	Recon 1	PRIME		3788	112	0.0286	0.0152
mCADRE cancer model	Recon1	mCADRE	26356 sec	1037	169	0.1248	0.0228
MBA cancer model	Recon 1	MBA	2000 hours	772	178*	0.0284	0.0060
cancer1	Recon 1	FASTCORMICS	184	810	183	0.0314	0.0063
cancer2	Recon 2	FASTCORMICS	184	1322	78	0.0502	0.0351

Table 4.1. Essentiality testing of different cancer models. Comparison of the number of essential genes found by an in silico essentiality assay to a ranked gene list established by Luo et al. based on the effect of shRNA knock-downs on the proliferation of cancer cells (Luo et al., 2008). In Folger et al. (Folger et al., 2011) a gene is considered as essential if its knock-down results in a decrease of the growth rate of at least 1%. To allow for a comparison of the different methods the 1% criteria was applied here as well. *The number of essential genes was taken from Additional Table 3 Cancer Cytostatic Genes column KO Growth Rate (relative to WT) of (Luo et al., 2008).

2 models, even when further constrained by the medium composition (Additional Table S2 of the original paper, medium composition sheet), allowed identification of only a smaller set of essential genes and their distribution along the ranked list of essential genes was not significantly different from the distribution of all metabolic genes (Table 4.1). Therefore, the predictive power of the reconstructed context-specific models is much better than either of the original GENREs. In contrast, the distribution of essential genes in the FASTCORMICS cancer models is different from the remaining metabolic genes and shifted towards the top of the ranked list as shown by a one-side KS-test (p-value=0.0314 for cancer1 and p-value=0.0502 for cancer2), demonstrating that FASTCORMICS predictions are much more coherent with the experimental data.

Moreover, comparison of the p-values to those obtained previously using the MBA algorithm (p-value=0.0284) (Folger et al., 2011; Jerby et al., 2010) suggests that FASTCORMICS performs with similar accuracy but with significantly lower running time (Table 4.1) Consistently, a permutation test showed that the likelihood of finding a gene set of the same size with a better KS-score by chance is low (p-value=0.0063 for cancer1 and pvalue= 0.0351 for cancer2). In order to benchmark our workflow we also built cancer models using GIMME (Becker and Palsson, 2008), iMAT (Zur et al., 2010) and mCADRE (Wang et al., 2012). For GIMME and iMAT, the implementation of the Cobra toolbox (Schellenberger et al., 2011a) was run using as thresholds respectively the 75 and the 25 percentile for high and low expressed genes. For mCADRE the data was first discretized using Barcode (McCall et al., 2011) and then the implementation provided in the supplementary files of (Wang et al., 2012) was run. We also compared our workflow to PRIME (Yizhak et al., 2014a). PRIME uses microarray data and respective growth rate information to adapt the bounds of the input generic reconstruction. Thus it does not extract a context-specific sub-network from a general reconstruction and thereby differs from FASTCORMICS and the others algorithms discussed in this paper. Building a generic cancer model using PRIME was not possible as there is no generic growth rate. Instead the 32 models built by (Yizhak et al., 2014b), were used to perform KO assays. 112 genes were essential in at least 90% of the 32 models (in fact these 112 genes were essential in all models). Out of the 112 genes, 81 were found in the ranked list of essential genes by (Luo et al., 2008) and used for p-value calculation.

We also tested iMat (Zur et al., 2010), but the algorithm does not guarantee that the biomass function is included in the model and therefore the knockout experiment could not be performed here. In general, (Table 4.1), more compact models, i.e. mCADRE cancer model, MBA cancer model, and cancer1 generated with FASTCORMICS, tend to predict a higher number of essential genes, respectively 169, 178 and 183, compared to models with a larger number of reaction, i.e. the GIMME cancer model that includes twice as many reactions as cancer1 and only 69 predicted essential genes. The aforementioned models also tend to perform better in the KO assay with the exception of mCADRE that identifies essential with a lower rank in the ranked essentiality list of Luo et al (Luo et al., 2008).

4.3.2 Validation 2: Prediction of the secretion rate of lactate

Context-specific models were built for the 59 cell lines integrating Recon1 and the cell line specific expression data with the FASTCORMICS workflow. The medium composition was used to constrain the inputs of the models (only input reactions for metabolites present in the medium were allowed to carry a flux). To obtain lactate secretion rates predictions in fmol/cell/h, the biomass coefficients were multiplied by 550 as described in (Gatto et al., 2015). Further, the bounds of the obtained models were multiplied with 1.5 to obtain a flux range consistent with the measured lactate rate. In order to guarantee lactate, glucose, oxygen and glutamine exchanges, the respective exchange reactions were added to the core set. To allow quantitative predictions for each context-specific models, the bounds of the inputs reactions of glucose and glutamine were fixed to match the experimental data. Additionally, the maximal uptake respectively production rate of alanine, serine, leucine, lysine, isoleucine, valine, arginine, threonine, tyrosine, phenylalanine, methionine, asparagine, choline, glycine, and tryptophan were constrained according to the experimental data. The uptake rates of cysteine, histidine, and myo-Inositol, which were not reported in the table, were set to zero. Random sampling was performed while optimizing for biomass production. A solution could not be found for 7 cancer models, with these settings. For the other models a R^2 value of 0.7 was obtained, indicating a good correlation of context specific predicted and measured lactate secretion rates.

As a second quality control step, a hypergeometric test showed that the neoplasia associated genes retrieved from the DisGeNet database (Queralt-Rosinach and Furlong, 2013) are over-represented in the essential genes of both FASTCORMICS models (Table 4.2). This indicates that FASTCORMICS can help to identify medically relevant genes. Further, among essential genes predicted in cancer1 and cancer 2 130 (71%) and 46 (59%) were known to be associated to cancer, respectively (DisGeNET (Queralt-Rosinach and Furlong, 2013), CCGD database (Starr et al.)) or to be already predicted as essential by the generic model from which they were extracted. Taken together, FASTCORMICS outperforms competing algorithms in speed and therefore allows generating robust high-quality models in a high-throughput manner. This will enable the use of metabolic modelling as a routine process for the analysis of large microarray data sets across different cell types and contexts.

4.3.3 Confidence levels of the reactions of the macrophage model

We compared the reactions of the macrophage model built with the FASTCORMICS workflow to a table (Table 4.3) established by (Bordbar et al., 2010) that assigned confidence levels to the reactions of Recon1 in function of the evidence of expression in macrophage. 759 reactions of our model were found in the supplementary data 7 of (Bordbar et al., 2010), with 595 having

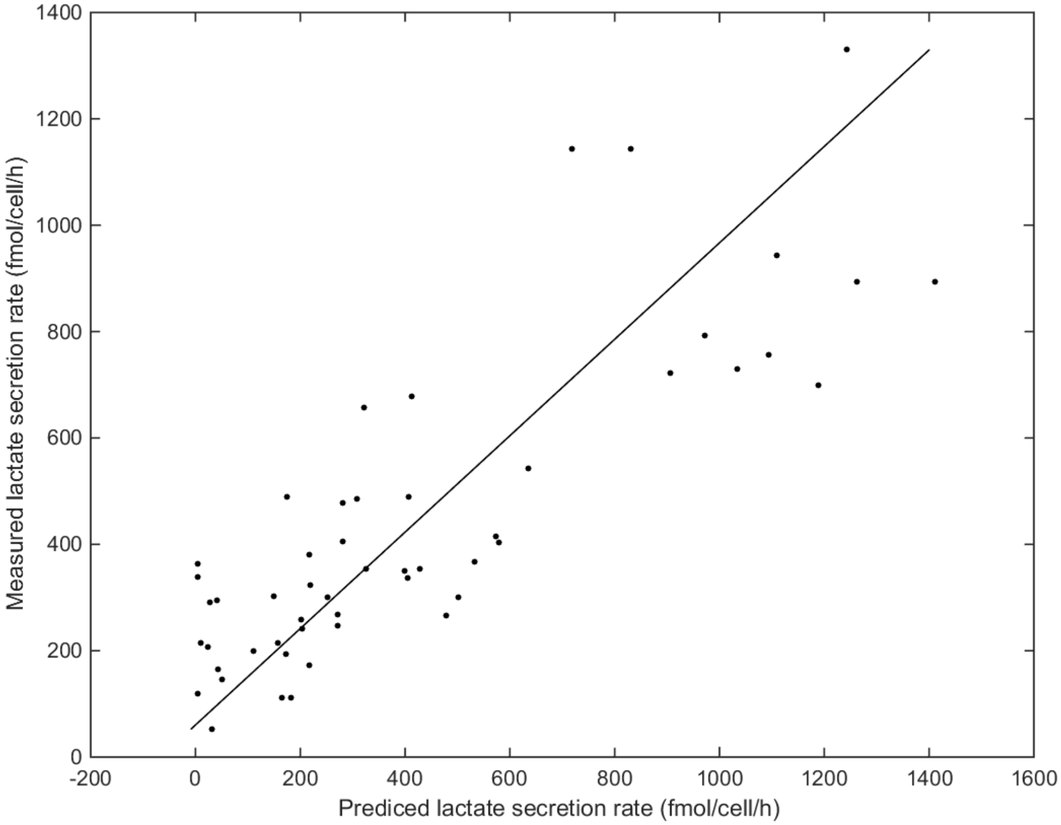


Figure 4.3. Prediction power of FASTCORMICS: Correlation plot of the predicted lactate secretion rates by context-specific cancer cell models and the lactate secretion rates measured by (Jain et al., 2012).

Output model	Essential genes (EG)	EG in DisGeNet	Genes in the generic models (GG)	GG in DisGeNet	p-value
Recon1 (unconstrained)	14	6	1168	377	0.2792
Recon 2 (unconstrained)	4	1	1599	433	0.7176
Medium constrained Recon 1 + biomass	78	33	1168	377	0.0350
Medium constrained Recon 2 + biomass	32	14	1599	433	0.0299
GIMME cancer model	69	32	1168	377	0.0083
PRIME cancer models	124	50	1496	449	4.69 e-4
mCADRE cancer model	169	73	1168	2635	2.474e-4
MBA cancer model	178	84	1168	449	4.63e-6
cancer1	183	86	1168	377	4.28e-6
cancer2	106	45	1599	433	0.0295

Table 4.2. Hypergeometric test quantifying the enrichment of neoplasia-related genes retrieved from DisGeNet (Queralt-Rosinach and Furlong, 2013), a database of disease-gene associations, in the set of essential genes of the different cancer models. In (Folger et al., 2011), a gene is considered essential if its knock-downs resulted in a decrease of the growth rate of at least 1%. To allow, a comparison with (Folger et al., 2011), the 1% criteria was applied as well.

Model	reactions	metabolites	genes	Core reactions	Inactive reactions
Day 2 model	978	858	614	462	806
Day 4 model	1055	918	594	605	759
Day 7 model	1202	1034	706	671	646
Day 11 model	1149	993	689	623	656

Table 4.3. Summary of the monocyte-macrophage models

a confidence level assigned. The remaining 410 reactions of our model not being listed in the Bordbar table (Bordbar et al., 2010) are due to a different annotation of Recon1 and Recon2 that was taken as input for our macrophage model. Of the 595 reactions with confidence information, 485 (82%) were assigned a high or medium confidence level by (Bordbar et al., 2010), 16 had a low and 94 are Exchanges / Transports added for modelling purposes, disassociations or spontaneous reactions to which no specific confidence level was assigned. No reactions were added that were shown not to be expressed in macrophages. Overall, this indicates a high confidence level for our reconstructed macrophage model (Table 4.4).

Confidence levels

	high	moderate	weak	High	Weak
Day 2 model	306	387	285	721	85
Day 4 model	391	491	173	667	96
Day 7 model	490	540	172	563	83
Day 11 model	441	507	201	593	63

Table 4.4. Confidence level of the included and excluded reactions of the monocytes macrophage models determined through the cross-validation step. Reactions with a high level of confidence are supported by at least two core reactions. Reactions with moderate confidence level are reactions only supported by barcode. Reactions with a weak confidence level are not supported by barcode, but needed to generate a consistent network model. Excluded reactions with a high confidence score were never included in any simulations suggesting the presence of other excluded reactions in the branch. Excluded reactions with a low confidence level were removed only due to their low expression level.

FASTCORMICS: Application

*This chapter was originally published as **Integrated metabolic modelling reveals cell-type specific epigenetic control points of the macrophage metabolic network** in October 2015 in BMC Genomics.*

Contents

5.1 Summary and contributions

FASTCORMICS was used in the paper **"Integrated metabolic modelling reveals cell-type specific epigenetic control points of the macrophage metabolic network"** Pires Pacheco et al. (2015a) to create models over hundred samples of different cell types. The wet lab experiments were performed by Doctor Elisabeth John (under the supervision of Dr. Lasse Sinkkonen), with whom I share the first co-authorship. I performed the *in silico* experiments and the integration of the data under the guidance of Prof. Thomas Sauter and Dr. Lasse Sinkkonen, except for the ones clearly stated in the contribution section of the paper to have been performed by another co-author. Figure 5.2, 5.4, 5.5 panel D and E, 5.6, 5.7 and 5.8 were produced by myself under the guidance and the help of Prof. Thomas Sauter and Dr. Lasse Sinkkonen.

The FASTCORMICS workflow was used to capture the variations in metabolism across 156 models corresponding to 63 primary cell types (Figure 5.2). The models were clustered in function of the Jaccard Similarity Index showing that a) the workflow is able to capture small but significant metabolic variations and b) that each model shares at least 30 % of the reactions with every other model. Further, the fraction of activation of the pathways across the models was computed (Figure 5.2) to determine the level of variance across the models and by such illustrate the cell-specificity of the pathway. The FASTCORMICS workflow was further used to reveal epigenetic control points in the metabolic network of macrophage. Therefore the expression levels and the specificity of expression in macrophages of high-regulatory load genes defined as the set of genes with the top 10% highest number of enhancers was compared to remaining metabolic gene set. Finally, strategic positions in the network controlled by high-regulatory genes, namely transporters and entry points, were identified (Figure 5.1).

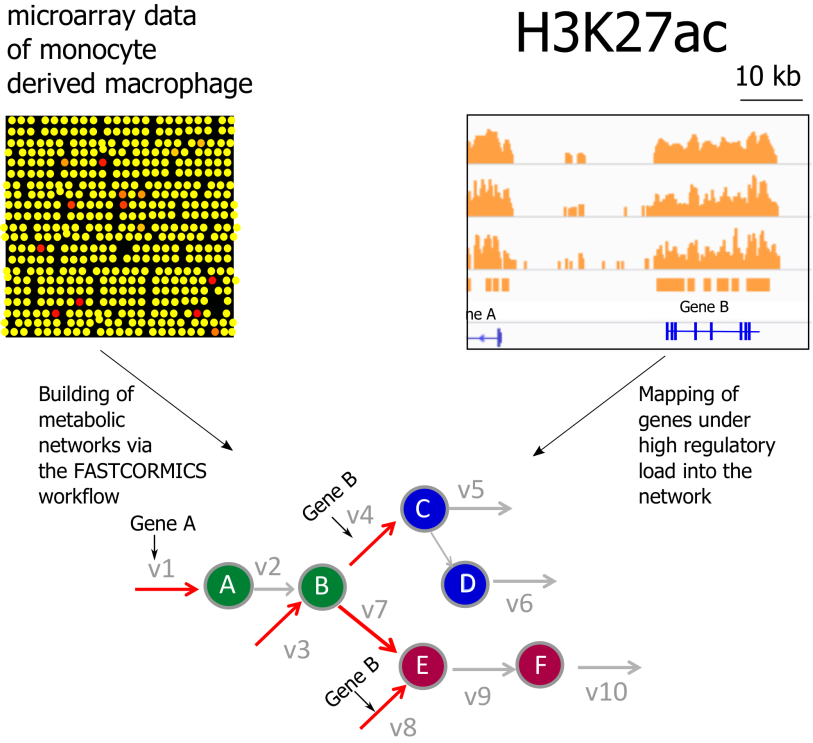


Figure 5.1. Entry points and transporters are under tight regulation: Entry points are defined as the first gene-associated reaction of a system or the first reaction after a pathway change. The different pathways of the network are represented in green, blue and red. Entry points are marked in red.

5.2 Abstract

5.2.1 Introduction

The reconstruction of context-specific metabolic models from easily and reliably measurable features such as transcriptomics data will be increasingly important in research and medicine. Current reconstruction methods suffer from high computational effort and arbitrary threshold setting. Moreover, understanding the underlying epigenetic regulation might allow the identification of putative intervention points within metabolic networks. Genes under high regulatory load from multiple enhancers or super-enhancers are known key genes for disease and cell identity. However, their role in regulation of metabolism and their placement within the metabolic networks has not been studied.

5.2.2 Methods

Here we present FASTCORMICS, a fast and robust workflow for the creation of high-quality metabolic models from transcriptomics data. FASTCORMICS is devoid of arbitrary parameter settings and due to its low computational demand allows cross-validation assays. Applying FASTCORMICS, we have generated models for 63 primary human cell types from microarray data, revealing significant differences in their metabolic networks.

5.2.3 Results

To understand the cell type-specific regulation of the alternative metabolic pathways we built multiple models during differentiation of primary human monocytes to macrophages and performed ChIP-Seq experiments for histone H3 K27 acetylation (H3K27ac) to map the active enhancers in macrophages. Focusing on the metabolic genes under high regulatory load from multiple enhancers or super-enhancers, we found these genes to show the most cell type-restricted and abundant expression profiles within their respective pathways. Importantly, the high regulatory load genes are associated to reactions enriched for transport reactions and other pathway entry points, suggesting that they are critical regulatory control points for cell type-specific metabolism.

Conclusions

By integrating metabolic modelling and epigenomic analysis we have identified high regulatory load as a common feature of metabolic genes at pathway entry points such as transporters within the macrophage metabolic network. Analysis of these control points through further integration of metabolic and gene regulatory networks in various contexts could be beneficial in multiple fields from identification of disease intervention strategies to cellular reprogramming.

5.3 Introduction

Metabolism is a highly regulated dynamic process that involves transport and chemical reactions of thousands of metabolites to fulfill hundreds of metabolic functions. Metabolic dysfunction is a major contributor to many diseases which have become prevalent in human population in the last decades, e.g. cardiovascular diseases (Gluckman et al., 2009), neurodegenerative diseases (Lin and Beal, 2006) and cancer (Cairns et al., 2011) amongst many others. Alternative pathways and branches are continuously activated or shut down to maximize metabolic efficiency in a specific context (Lewis et al., 2010a), resulting in disease and patient-specific alterations.

Metabolism is regulated at multiple-levels with abundance and expression of the metabolic enzymes being one of the most decisive mechanisms. Gene expression control has to integrate multiple signals both at transcriptional and post-transcriptional levels. At the epigenetic level the availability of various transcription factor (TF) binding sites through chromatin decondensation at context-specific enhancers is regulated by the interplay of TFs and post-translational histone modifications deposited by the recruited co-activators (Maston et al., 2012). Enhancers adhere to unique chromatin states defined by features such as deposition of histone variants, presence of coactivators and monomethylation of histone H3 at lysine 4 (H3K4me1) (Calo and Wysocka, 2013). More recently, acetylation of histone H3 at lysine 27 (H3K27ac) was described to specifically mark active enhancers engaged in regulation of RNA polymerase activity through chromatin looping (Rada-Iglesias et al., 2011; Creyghton et al., 2010). Recent work on genome-wide analysis of active enhancers has revealed that important genes determining cellular identity, such as TFs, are often controlled by large and strong clusters of multiple enhancers called super-enhancers or stretch-enhancers that are active in a cell type-specific manner (Whyte et al., 2013; Parker et al., 2013; Hnisz et al., 2013). Moreover, these enhancer clusters usually reside in insulated chromatin loops or domains and often overlap with so called TF hotspots, suggesting that their target genes are under high regulatory load from multiple TFs and enhancers, integrating numerous different signals to promote proper cellular phenotype, including the appropriate metabolic network (Dowen et al., 2014; Siersbæk et al., 2014). However, the role of high regulatory load genes in the metabolic networks has not been studied previously.

Metabolic networks are highly complex and can hardly be understood without using mathematical representations. The most comprehensive descriptions of metabolism are genome-scale reconstructions (GENREs). There are several human reconstructions available, like Recon 1 and Recon 2 (Duarte et al., 2007; Thiele et al., 2013) or the Edinburgh Human Metabolic Network (Ma et al., 2007). Alongside with these reconstructions extensive reaction databases were developed, like HMR (Agren et al., 2012, 2014) or HumanCyc (Caspi et al., 2010; Romero et al., 2005), which collect additional information to refine the available models. Mathematical models derived from GENREs were successfully used to understand how perturbations in the metabolism lead to severe pathologies (Agren et al., 2014; Folger et al., 2011; Mardinoglu et al., 2014a).

GENREs are usually generic representations of a cell or organism comprising all reactions that can potentially become active regardless of the specific environment and cell type. Therefore they do not cover the fact that the set of expressed genes and thereby the set of active reactions vary significantly in function of the cellular context. The generation of context-specific models that include only pathways predicted to be active in the given context is highly desirable and has lead to the development of various algorithms like GIMME (Becker and Palsson, 2008), IMAT (Zur et al., 2010), MADE (Jensen and Papin, 2011), mCADRE (Wang et al., 2012), INIT (Agren et al., 2012) or MBA (Jerby et al., 2010) that use omics data for building of context-specific model. While allowing the generation of models with higher predictive power than the GENREs from which they were derived from (Becker and Palsson, 2008; Jerby et al., 2010), these algorithms suffer from high computational demands due to the application of mixed integer linear programming, and/or the required setting of one or several expression thresholds.

Recently we proposed an LP-based algorithm for the fast reconstruction of compact context specific metabolic networks (FASTCORE) that allowed decreasing the reconstruction time of context-specific networks to the order of seconds, using as input a GENRE and a set of core reactions being active in the context of interest (Vlassis et al., 2014). FASTCORE identifies a close to minimal set of non-core reactions from the input model, to be added to the core set in order to obtain a consistent model.

To adapt FASTCORE for the direct integration of microarray data, we propose here a new workflow: FASTCORMICS pre-processes microarray data with the discretization tool Barcode (McCall et al., 2011; Zilliox and Irizarry, 2007), is devoid of arbitrary parameter settings and has a low computational demand with overall context-specific model building times in the order of a few minutes. We use FASTCORMICS to generate multiple metabolic models across tens of primary cell types and analyze the cell-type-specific usage of the alternative branches in metabolic networks. To address the question of epigenetic regulation of metabolism in different cell types we performed genomewide mapping of active enhancers in primary human macrophages and integrated these data with metabolic models of monocyte-to-macrophage differentiation to expose the metabolic genes under high regulatory load by multiple enhancers. We show that high regulatory load genes have a cell type-selective expression profile within any metabolic pathway and a specific positioning of many of these genes at transport or entry point reactions of pathways.

5.4 Results

5.4.1 Cell type-specific metabolic networks of primary cells

In order to adapt FASTCORE for the integration of transcriptomics data from microarrays, we developed a new workflow named FASTCORMICS (Additional file 1: Figure 4.2), requiring as inputs microarray data, which are first pre-processed with the discretization tool Barcode (Zilliox and Irizarry, 2007), and a GENRE of the organism of interest. Like FASTCORE, FASTCORMICS is devoid of arbitrary parameter settings and has a low computational demand with overall building times in the order of a few minutes.

To validate FASTCORMICS, we first performed an essentiality assay on two generic cancer models based on Recon 1 and Recon 2 and existing microarray expression data from 59 cancer cell lines (Shankavaram et al., 2009; Pfister et al., 2009) (for full description, please see Additional file 1). Comparison to a ranked gene list based on an shRNA essentiality screen in several different cancer cell lines (Luo et al., 2008) shows the significant predictive power of the FASTCORMICS models (Additional file 1: Table S1). Benchmarking against similar algorithms shows that FASTCORMICS clearly outperforms competitors in speed, while predicting the highest number of essential genes and achieving best significance levels among other algorithms (for results and medium composition, see Additional file 1: Table S1 and Additional file 2: Table S2, respectively). A hypergeometric test also showed that the neoplasia-associated genes retrieved from the DisGeNet database (Queralt-Rosinach and Furlong, 2013) are overrepresented in the essential genes of both FASTCORMICS models (Additional file 1: Table S3). Finally, predicted lactate secretion rates based on cancer cell line specific reconstructions showed a good correlation with measured rates indicating the capability of FASTCORMICS to also generate context specific reconstructions (Additional file 1: Figure 4.3).

In order to identify cell type-specific differences in the usage of the human metabolic network and to further validate the FASTCORMICS workflow, we generated context-specific metabolic models based on Recon 2 for different cell types across most human lineages. From an existing collection of 745 microarrays (Mabbott et al., 2013), we selected a subset of 156 microarrays (Additional file 3: Table S4), corresponding to 63 primary human cell types at their resting states, and took advantage of the low computational demands of FASTCORMICS to generate a model for each microarray. All reconstructed models are available in SBML format (Additional file 4). Interestingly, clustering the different models according to their active reactions allowed clear separation between the cell types largely along their developmental origin or cellular function, suggesting significant differences in the metabolism across cell types (Figure 5.2). The most unique metabolism was predicted for the gametocytes, oocytes and spermatocytes, which at lowest showed only around 30% similarity to other cell types. Some of the largest clusters were formed by the different blood cells that clustered together with their progenitors as well as CD34+ hematopoietic stem cells, suggesting many shared features in their metabolism.

To investigate how much the different pathways contribute to the differential metabolism between the cell types, and what are the most unique pathways in different cell types, we looked into the activity state of all reactions according to the pathways they belong to. Figure 5.2b lists all the Recon 2 pathways consisting of more than one reaction, ordered by their combined median activity in all analyzed cell types with the first pathways (from left to right) showing no activity in almost none of the analyzed cell types and the last pathways being fully active in almost all cell types. The distribution of these values indicates the variation in the number of active reactions for each pathway between cell types and, for example, the usage of additional or alternative branches of the pathways. By focusing on the most deviant values of any pathway one can identify the cell types that show very high or very low number of active reactions for that pathway compared to other cell types, and can thereby identify the cell type-specific branches of those pathways.

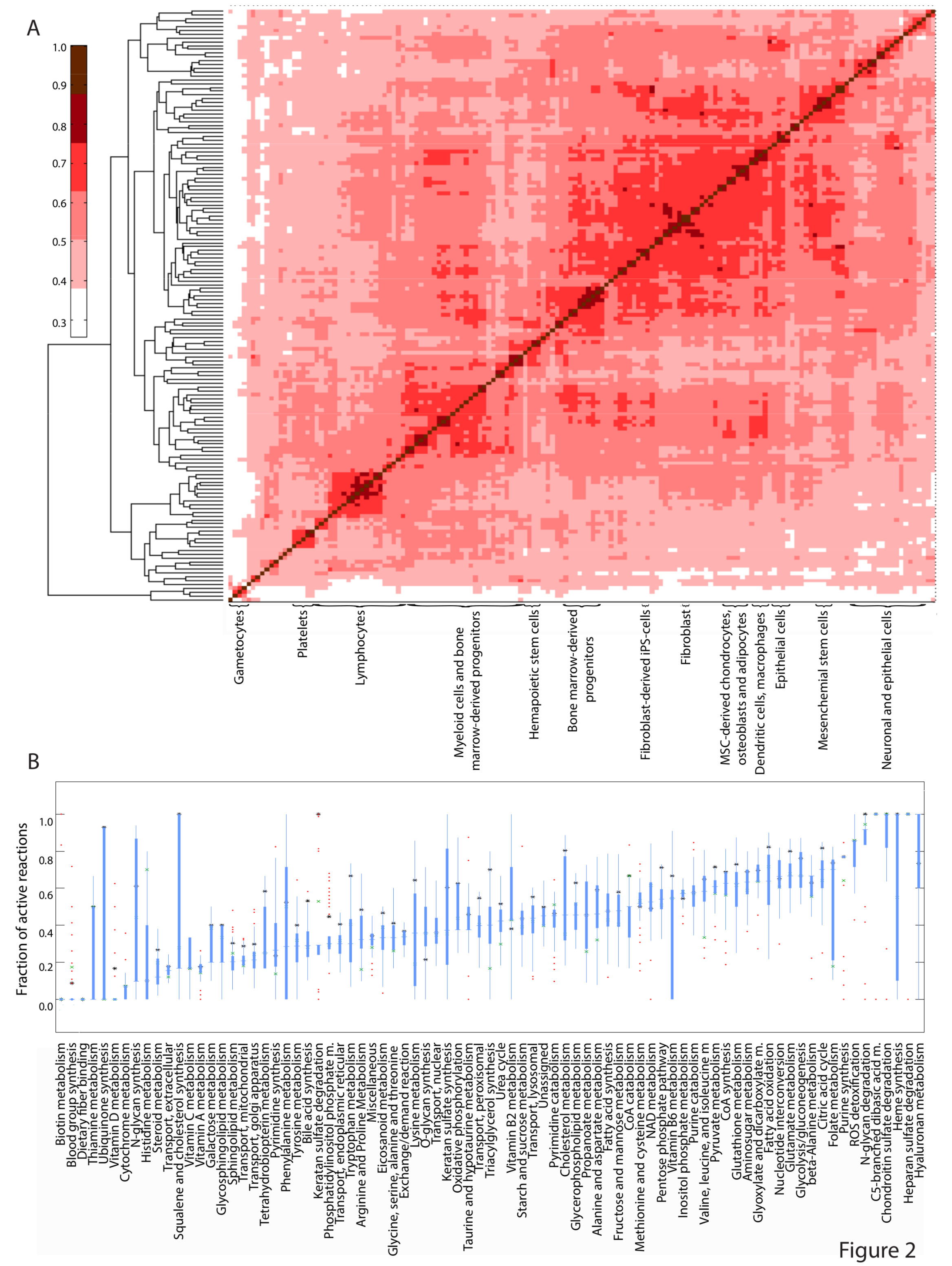


Figure 2

Figure 5.2. Cell type-specific metabolic pathways in primary human cells. a 156 metabolic models based on an equal number of microarrays and corresponding to 63 primary human cell types were built using the FASTCORMICS workflow and microarray collection from Primary Cell Atlas (GSE49910). The level of similarity between the different model pairs was determined via the Jaccard index. The Jaccard index matrix was then clustered in function of the similarity level. b For each pathway, the level of activity, given as percentage of reactions of the consistent version of Recon 2 (5317 reactions) that are present in each context-specific model was computed. The distribution of the activity levels of each pathway across the 156 models are shown as box plots and sorted according to the median value across the pathways. Pathways that contain less than 4 reactions were not included. Mean percentage of active reactions across macrophage and monocyte samples are depicted by black asterisk (*) and green cross (x), respectively

Altogether, as expected, the different cell types exhibit differential usage of their metabolic pathways, ranging from ubiquitous to cell-type specific. This variation can be captured by FASTCORMICS and allows clustering of the cell types according to their functions and developmental origins.

5.4.2 Metabolic modelling of primary monocyte-to macrophage differentiation

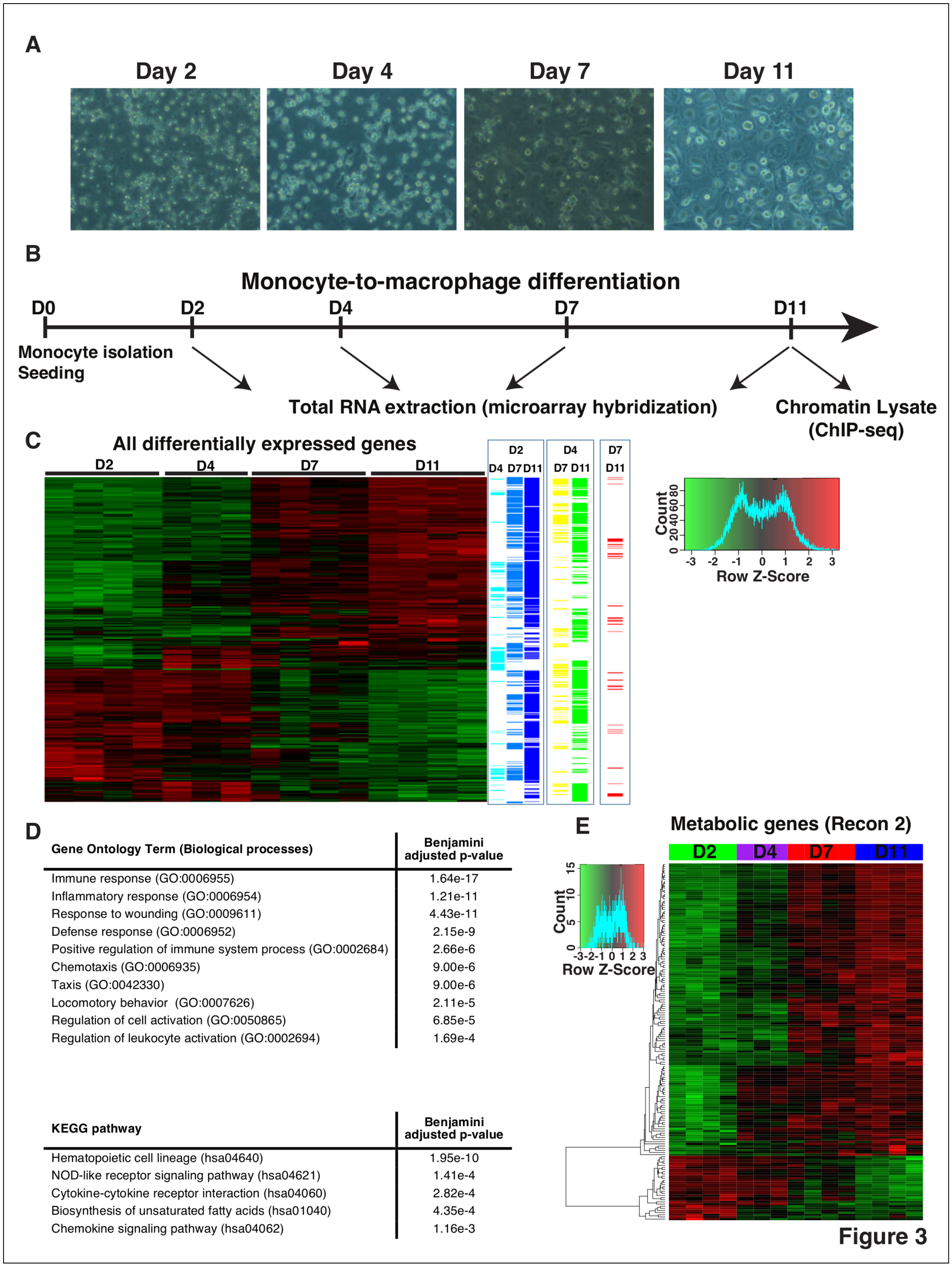
One of the cell types with particularly high proportion of active reactions compared to other cell types across many metabolic pathways are macrophages. This is true when comparing to the median of all cell types as well as when comparing to the immediate precursor cells, the monocytes (Figure 5.2b, Additional file 1: Figure S3). To gain more detailed understanding of the differential usage and regulation of the metabolic pathways in macrophages, we chose to generate our own expression data with sampling at multiple time points during differentiation of primary human monocytes to macrophages as well as regulatory data from macrophages by mapping active enhancer regions (Figure 5.3). This was done in multiple biological replicates to stringently focus on regulatory regions that are active in most healthy individuals (please see next chapter for details).

For the expression profiling we chose to isolate primary human monocytes from blood samples of four healthy donors and to differentiate those to adherent mature macrophages over a time course of 11 days (Figure 5.3a). Total RNA was collected at four time points, 2, 4, 7 and 11 days after isolation, and used for gene expression profiling by microarrays (Figure 5.3b). Time points before 2 days were not considered as the cells at these early stages are affected by the stress from the collection and isolation. During the differentiation (comparing day 11 to day 2), a total of 882 genes were significantly upregulated (FDR < 0.05, log2 fold change

1) while 519 were down-regulated (Figure 5.3c, Additional file 5: Table S5). Most expression changes occurred already early in the differentiation and were not too dynamic, as most genes that changed significantly during the differentiation (day 4 or day 7), also remained differentially expressed in day 11 macrophages (Figure 5.3c). Gene Ontology (GO) and KEGG Pathway analysis of the differentially expressed genes revealed enrichment for many categories and pathways related to macrophage function, suggesting the differentiation had been successful (Figure 5.3d). The differentially expressed genes included also 57 TFs. Among the highest expressed TFs in macrophages we found CEBP-family factors (CEBPB, CEBPA, CEBPD and CEBPG), EGR2, SPI1 (also known as PU.1), SREBF2, and FLI1, most of which are known regulators of macrophage differentiation and phenotype (Huber et al., 2014; Pham et al., 2012; Najafi-Shoushtari et al., 2010; Suzuki et al., 2013). RREB1 was the only factor among the 20 highest expressed TFs for which we did not find any previously described role in macrophages. Finally, 164 metabolic genes became differentially expressed with a log2 fold change 1 (FDR < 0.05) during the differentiation, most of which were up-regulated (Figure 5.3e).

The microarray data was used as an input for FASTCORMICS to generate four metabolic models that correspond to each tested time point of macrophage differentiation (Figure 5.4). All reconstructed models are available in SBML format (Additional file 4). Out of 5317 reactions in the consistent Recon 2 (version 3), 660 reactions were predicted to be active in each time point of macrophage differentiation (Additional file 1: Table S6). The complete size of the day 2 monocyte model was 978 active reactions (corresponding to 64 pathways), which increased to 1149 active reactions (67 pathways) in day 11 macrophages, suggesting that many inactive alternative branches become active during differentiation. Many of the newly activated reactions were turned on already early on day 4 of differentiation with most of the remaining reactions becoming active by day 7. The number of reactions that became inactive in macrophages is smaller with only one pathway decreasing its overall number of active reactions.

Among the pathways with highest relative number of active reactions in macrophages were several fairly ubiquitously active pathways such as hyaluronan metabolism, chondroitin sulfate degradation, and N-glycan degradation (Figs. 5.2 and 5.4). However, most of these, as well as many other pathways with steady overall number of active reactions (such as triacylglycerol synthesis and cholesterol metabolism) still showed a significant increase in the expression of



Gene Ontology Term (Biological processes)	Benjamini adjusted p-value
Immune response (GO:0006955)	1.64e-17
Inflammatory response (GO:0006954)	1.21e-11
Response to wounding (GO:0009611)	4.43e-11
Defense response (GO:0006952)	2.15e-9
Positive regulation of immune system process (GO:0002684)	2.66e-6
Chemotaxis (GO:0006935)	9.00e-6
Taxis (GO:0042330)	9.00e-6
Locomotory behavior (GO:0007626)	2.11e-5
Regulation of cell activation (GO:0050865)	6.85e-5
Regulation of leukocyte activation (GO:0002694)	1.69e-4

KEGG pathway	Benjamini adjusted p-value
Hematopoietic cell lineage (hsa04640)	1.95e-10
NOD-like receptor signaling pathway (hsa04621)	1.41e-4
Cytokine-cytokine receptor interaction (hsa04060)	2.82e-4
Biosynthesis of unsaturated fatty acids (hsa01040)	4.35e-4
Chemokine signaling pathway (hsa04062)	1.16e-3

Figure 3

Figure 5.3. Transcriptomic profiling of primary human monocyte-to-macrophage differentiation. a-b Primary human monocytes isolated from donated blood samples were differentiated into monocyte-derived macrophages in vitro, and microarrays were performed with total RNA extracted on time points day 2, day 4, day 7 and day 11. In addition chromatin was isolated for Chip-seq experiments from day 11 macrophages. c Relative expression levels of differentially expressed genes during monocyte-to-macrophage differentiation selected with a FDR cut-off of 0.05 and absolute fold change greater or equal to 2 were clustered and represented as a heatmap. Genes with a positive Z-score are represented in red and negative in green. On the right of the heatmap, the time points where the differentially expressed genes show significant changes are indicated for a comparison between D2 and the remaining time points in different shades of the blue, between D4 and D7 or D11 in yellow or green and between D7 and D11 in red, indicating that most significant expression changes occur already at early time points. d. A Gene ontology analysis for enriched biological processes and KEGG pathways was performed on the differentially expressed genes using DAVID (Huang et al., 2009). The top ten gene ontology terms for the biological processes and the top five KEGG pathways are listed. e Relative expression levels of Recon 2 genes with differential expression (absolute fold change greater or equal to 2 and FDR < 0.05) during monocyte-to-macrophage differentiation are represented as a heatmap as in panel c

the genes corresponding to their active reactions, suggesting a further increased flux for these pathways in macrophages (Figure 5.4). The total of 42 subsystems that showed an increase in the expression of genes controlling them, are listed in Figure 5.4, together with the 2 subsystems showing decreased activity. Cross-validation based determination of confidence levels of the included model reactions (Additional file 1: Table S7) show high or moderate confidence for approx. 80% of the reactions, indicating that only approx. 20% of the reactions did not have expression based support, i.e. were added by FASTCORMICS to generate a consistent network model. And approx. 88% of the excluded reactions had multiple evidences (reactions with low expression) for not being included.

Next we aimed to find out which subsystems are particularly active in macrophages when compared to other cell types, including monocytes, and therefore possibly under macrophage-specific regulation. Since our own data were generated with more recent Affymetrix arrays (Human Gene 1.0 ST platform) where limited possibilities for comparisons to public data exist, we focused here also on the macrophage samples from the Primary Cell Atlas (Mabbott et al., 2013). Results are depicted in Figure 5.2b and Additional file 1: Figure S3. Among the interesting subsystems, for example, more than 60% of reactions in tryptophan metabolism are predicted active in macrophages while the median value across cell types is 30%. This is consistent with the models of monocyte-to-macrophage differentiation from our own data, which suggest over 3-fold increase in active tryptophan metabolism reactions over the time course (Figure 5.4). Similarly, approximately 80% of reactions in cholesterol metabolism are predicted active in macrophages, compared to a median of 45%. Also here there is a comparable 2-fold increase in active reactions from day 2 monocytes to day 11 macrophages in the models based on our own microarrays. Consistently with increased cholesterol metabolism, also bile acid synthesis, a major cholesterol catabolism pathway, is predicted to have more active reaction in macrophages (>50%) than the median across other cell types (29%), Other interesting pathways with particularly high numbers of reactions in macrophages include triacylglycerol synthesis and valine, leucine and isoleucine metabolism, both of which show further increase in expression during differentiation from monocytes to macrophages. Overall these results suggest that some alternative branches of the above-mentioned pathways could be under cell type-specific regulation in macrophages.

Taken together, time course analysis of metabolic models during macrophage differentiation predicts changed activities for hundreds of reactions, many of which occur already at early time points and, in contrast to what could be assumed from transcriptome-wide expression level changes, consist largely of increased reaction activities, especially in alternative branches of already active pathways.

5.4.3 Metabolic genes under high regulatory load in macrophages

Recent work has shown that active enhancers directly involved in transcriptional activation via chromatin looping are marked by specific chromatin modifications such as acetylation of lysine 27 of histone H3 (H3K27ac) (Rada-Iglesias et al., 2011; Creyghton et al., 2010). Moreover, we and others have shown that genes under high regulatory load from multiple TFs are often disease-associated and acting as cell type-specific key regulators of cellular identity (Hnisz et al., 2013; Galhardo et al., 2014; Pasquali et al., 2014). Importantly, these genes are marked by a high number of strong enhancers, collectively also called super-enhancers or stretch-enhancers (Whyte et al., 2013; Parker et al., 2013), allowing their identification using epigenomic mapping of active enhancers.

In order to identify metabolic genes under high regulatory load in macrophages, we performed chromatin immunoprecipitation coupled to high throughput sequencing (ChIP-Seq) with an antibody against H3K27ac in primary human macrophages derived from additional three donors different on top of the donors used for the microarray analysis. Analysis of the obtained sequencing data identified approximately 27,000-28,000 active enhancer regions in macrophages, depending on the sample, with 16,290 regions detected in all three samples (Figure 5.5a).

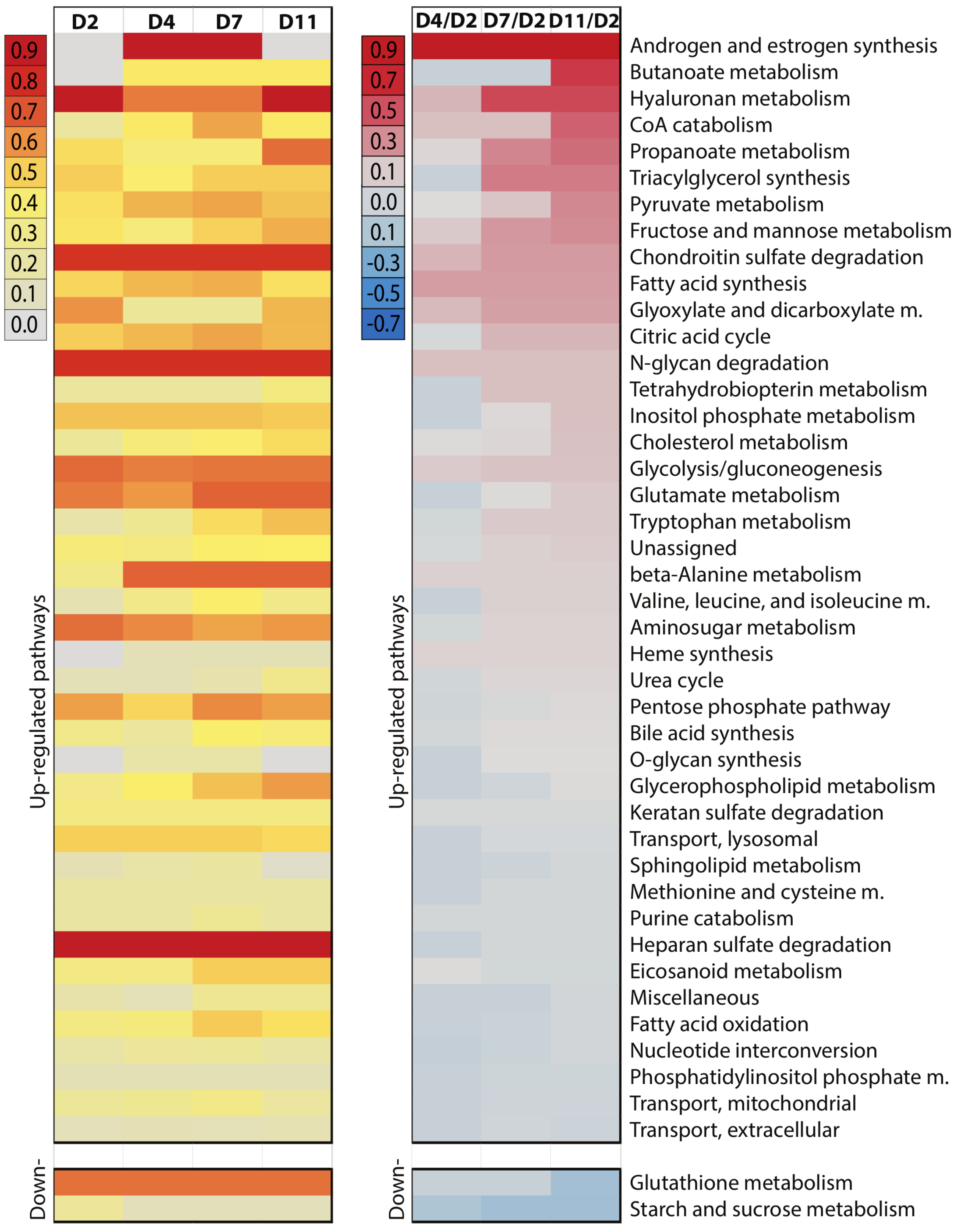


Figure 4

Figure 5.4. Monocyte-to-macrophage differentiation is accompanied by activation of alternative metabolic branches and increased activity of already active pathways. For each pathway, the level of activity (percentage of reactions in the input model that are present in the context-specific model) was computed for each time point (left panel). Each column represents the model built by the FASTCORMICS workflow for the given time point whereas each line stands for a different pathway. The fraction of active reactions ranges from 0 to 1 and is represented in shades of gray for low, yellow for intermediate and red for high number of active reactions per pathway. Additionally, (right panel) the significantly differentially expressed genes (FDR <0.05 and absolute log2 fold change > 1) were mapped to the models via the GPR rules. The percentage of up-regulated reactions in a pathway was computed after summing up the significantly up-regulated reactions. The number of significantly down-regulated reactions was then removed from this sum and the total was then normalized by the number of reactions in the pathway. The fraction of reactions associated with differentially expressed genes ranges between 0.7 for down-regulated pathways in blue and 0.9 for unregulated pathways in red. Only pathways that show a differential expression over time are represented.

The reproducibly identified enhancers in proximity of induced genes correspond to binding sites of known macrophage TFs such as SREBF2, FLI1, CEBP-family and SPI1, as suggested by the de novo motif analysis of the underlying sequences for enriched motifs (Figure 5.5b, see Additional file 1: Figure S4 for the complete list).

When assigning the enhancer regions to their putative target genes (see Materials and Methods; Generation of enhancer-to-gene associations), we observed that almost 8000 genes were associated with at least one active enhancer in macrophages, despite our stringent selection (Figure 5.5c). Ranking the genes according to their regulatory load (number of associated enhancers) revealed that the number of enhancers per gene ranged from 1 up to 59 with only the top 10% of the associated genes having 7 or more enhancers. Among these top genes were numerous TFs, many of which were already identified as highly expressed and enriched for their binding site motifs, including CEBP-family members, SPI1, and FLI1. As an example of a high regulatory load gene, the genomic locus of SPI1 the well-known pioneering factor and key regulator of macrophage differentiation with two large clusters of multiple enhancers, is depicted in Figure 5.5c. In contrast another abundantly expressed macrophage gene, CD4, is using only one intragenic enhancer region. Interestingly, RREB1, which we had previously noticed among highly expressed TFs in our microarray data, but for which no role in macrophages has been described, was the gene with third highest enhancer load of all genes in our experiments, suggesting that RREB1 might play an important role in macrophages or their differentiation. Finally, analysis of the expression levels of the top genes with 7 associated enhancers confirmed them to be on average significantly higher expressed than the genes with fewer enhancers (KS-test, p-value = 4.63e-38; Figure 5.5d).

Next we focused on the identification of the metabolic genes under high regulatory load. In total there are 689 metabolic genes expressed in the macrophages that are consistent with our metabolic model and 55 of them belong to genes under high regulatory load of 7 or more enhancers in our data set (based on manual curation of the enhancer to gene association, see Materials and Methods). Importantly, the expression of the metabolic genes under high regulatory load is even more shifted towards high expression levels when compared with other expressed metabolic genes (KS-test, p-value = 1.8537e-11; Figure 5.5e).

In summary, we reproducibly identified over 16,000 active enhancers in primary human macrophages, a large proportion of which could be associated to the top 10% of genes with high regulatory load. These genes are expressed at high levels and include many of the known key regulators of macrophage phenotype as well as 55 metabolic genes.

5.4.4 Genes under high regulatory load control monocyte-derived macrophages specific control points of metabolic pathways

Given that genes with high regulatory load are important for the cell identity and often expressed in a cell type specific manner, we decided to analyze the expression levels of the macrophage metabolic model genes across numerous different cell types. To this end, we again used the microarray data collection from Mabbott et al. (Mabbott et al., 2013), this time taking advantage of all 756 arrays corresponding to a total of 188 different cell types and conditions, and analyzed the expression level of each metabolic gene across the 188 conditions and ranked it according to its average level in the monocyte-derived-macrophage samples contained in the data set. Figure 5.6 depicts these ranks for all genes of the macrophage-specific metabolic model that belong to a subsystem containing at least one high regulatory load gene. Analysis of the distribution of the expression ranks along the cell types and subsystems reveals that; 1) the genes under high regulatory load (marked in orange) show an overall shift towards the upper ranks of macrophage metabolic genes, arguing they are generally expressed in a macrophage-specific manner, and 2) they are the more selectively expressed genes within each metabolic subsystem (Figure 5.6). At the same time the other genes contained in the macrophage model show an even distribution across the ranks, suggesting a more ubiquitous expression between cell types.

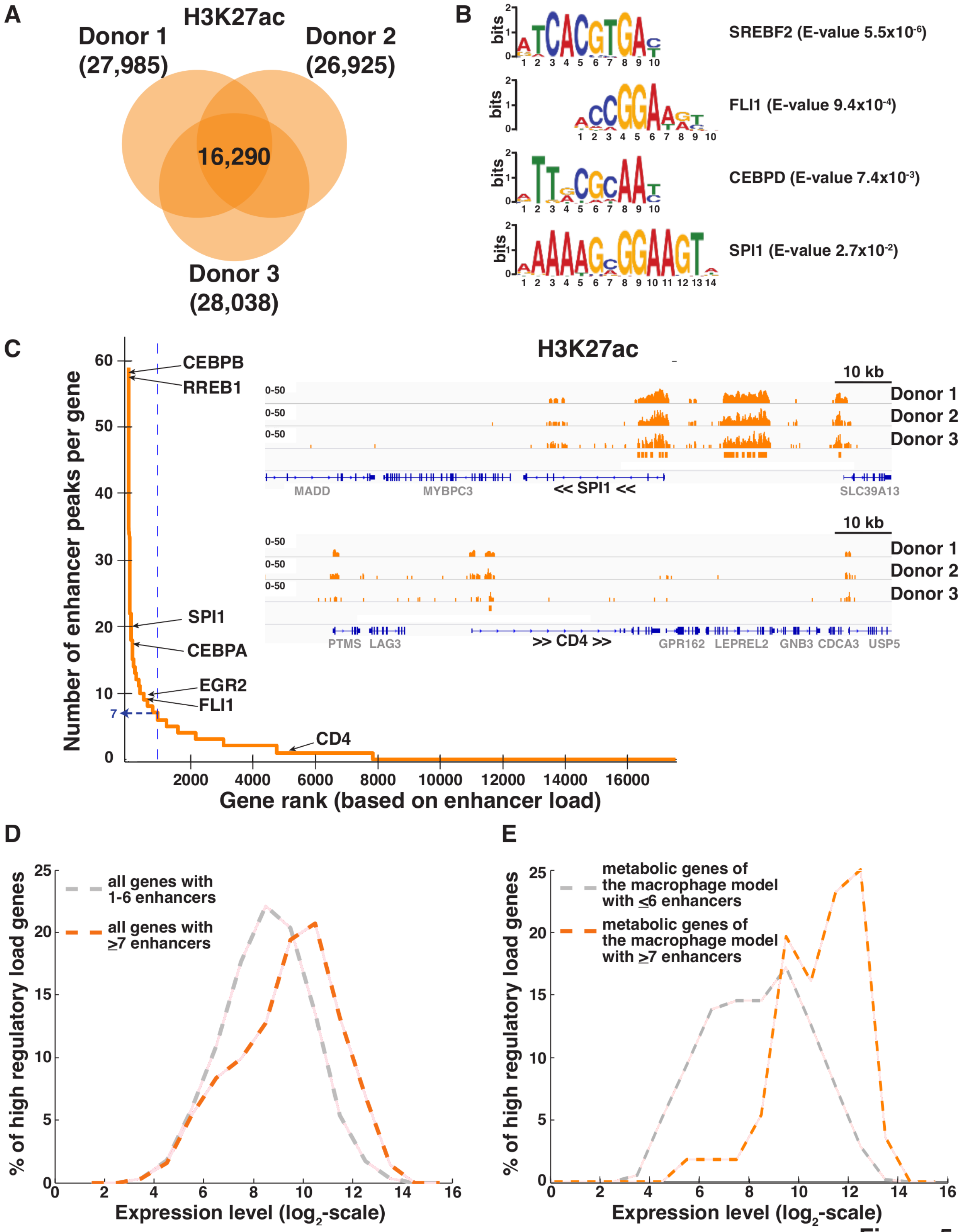

A
H3K27ac
Donor 1 (27,985)
Donor 2 (26,925)
16,290
Donor 3 (28,038)
B
bits
SREBF2 (E-value 5.5x10^-6)
FLI1 (E-value 9.4x10^-4)
CEBPD (E-value 7.4x10^-3)
SPI1 (E-value 2.7x10^-2)
C
Number of enhancer peaks per gene
Gene rank (based on enhancer load)
CEBPB
RREB1
SPI1
CEBPA
EGR2
FLI1
CD4
7
H3K27ac
10 kb
0-50
Donor 1
Donor 2
Donor 3
MADD
MYBPC3
<< SPI1 <<
SLC39A13
PTMS
LAG3
>> CD4 >>
GPR162
LEPREL2
GNB3
CDCA3
USP5
D
% of high regulatory load genes
all genes with 1-6 enhancers
all genes with ≥7 enhancers
Expression level (log2-scale)
E
% of high regulatory load genes
metabolic genes of the macrophage model with ≤6 enhancers
metabolic genes of the macrophage model with ≥7 enhancers
Expression level (log2-scale)

Figure 5

Figure 5.5. Identification of high-regulatory load genes in human macrophages. a Active enhancer regions were identified via chromatin immunoprecipitation coupled to high throughput sequencing (ChIP-Seq) with an antibody against H3K27ac using chromatin from monocyte-derived day 11 macrophages from 3 anonymous donors. Enhancer regions were considered reproducibly detectable when their genomic coordinates overlapped by at least one nucleotide in all biological replicates. b Selected enriched sequence motifs located within the identified active enhancer regions associated to upregulated genes in macrophages and corresponding to known transcription factor binding sites are shown. See full list in Additional file 1: Figure S4. c Genes associated with at least one active enhancer region were ranked in function of the number of active enhancer regions. A threshold (blue line) corresponding to the top 10% and at least 7 active enhancer regions was set to segregate between high regulatory load genes and the remaining expressed genes (please see Discussion for details on the threshold selection). 105 kb genomic regions surrounding SPI1 and CD4 loci, mapped reads indicating H3K27ac enrichment from the three donor samples, and called reproducible peaks are shown as examples of high regulatory load and low regulatory load genes, respectively. d The distribution of the expression levels of the high regulatory load genes was compared to genes that have a number of enhancers below the threshold of seven enhancers but that are associated to at least one enhancer (KS-test, p-value = 4.63e-38). e The enhancer load of the metabolic genes present in the consistent version of Recon2 was determined and then manually curated to minimize false peaks-to-gene assignments allowing identification of 74 high-regulatory load genes (7 enhancers), 55 of which mapped to the macrophage model. The distribution of expression levels of these metabolic high regulatory load genes was compared to the distribution of expression of the remaining metabolic genes of the macrophage model (KS-test, p-value = 1.8537e-11)

Since most of the metabolic genes with high regulatory load in macrophages are preferentially expressed in macrophages, and are usually the most abundantly expressed genes within their respective pathway, we asked in addition whether the positioning of the reactions they control within the macrophage metabolic network is also different from other reactions. Indeed, we could observe clear differences when focusing on the genes associated to transporters or entry points of the pathways predicted active in the macrophage model (Figure 5.7).

While 53.1% of all gene-associated reactions in our macrophage metabolic model are transport or entry point reactions, this fraction increases significantly to 67.1% when focusing on reactions associated to high regulatory load genes (KS-test, p-value = 9.0e-5). Furthermore, when looking only on transport reactions that constitute 17.4% of all macrophage reactions, we observe an even more significant enrichment (KS-test, p-value = 1.8e-7) to 32.9% of the reactions associated with high regulatory load. Finally, when excluding the transport reactions and focusing on the reactions corresponding to the remaining entry points of the different pathways (44.7% of all macrophage reactions) we also see an enrichment for the high regulatory load genes (51.6% of high regulatory load reactions), although with clearly higher p-value (KS-test, p-value = 0.0839).

Importantly, similar results could not be obtained using a generic metabolic

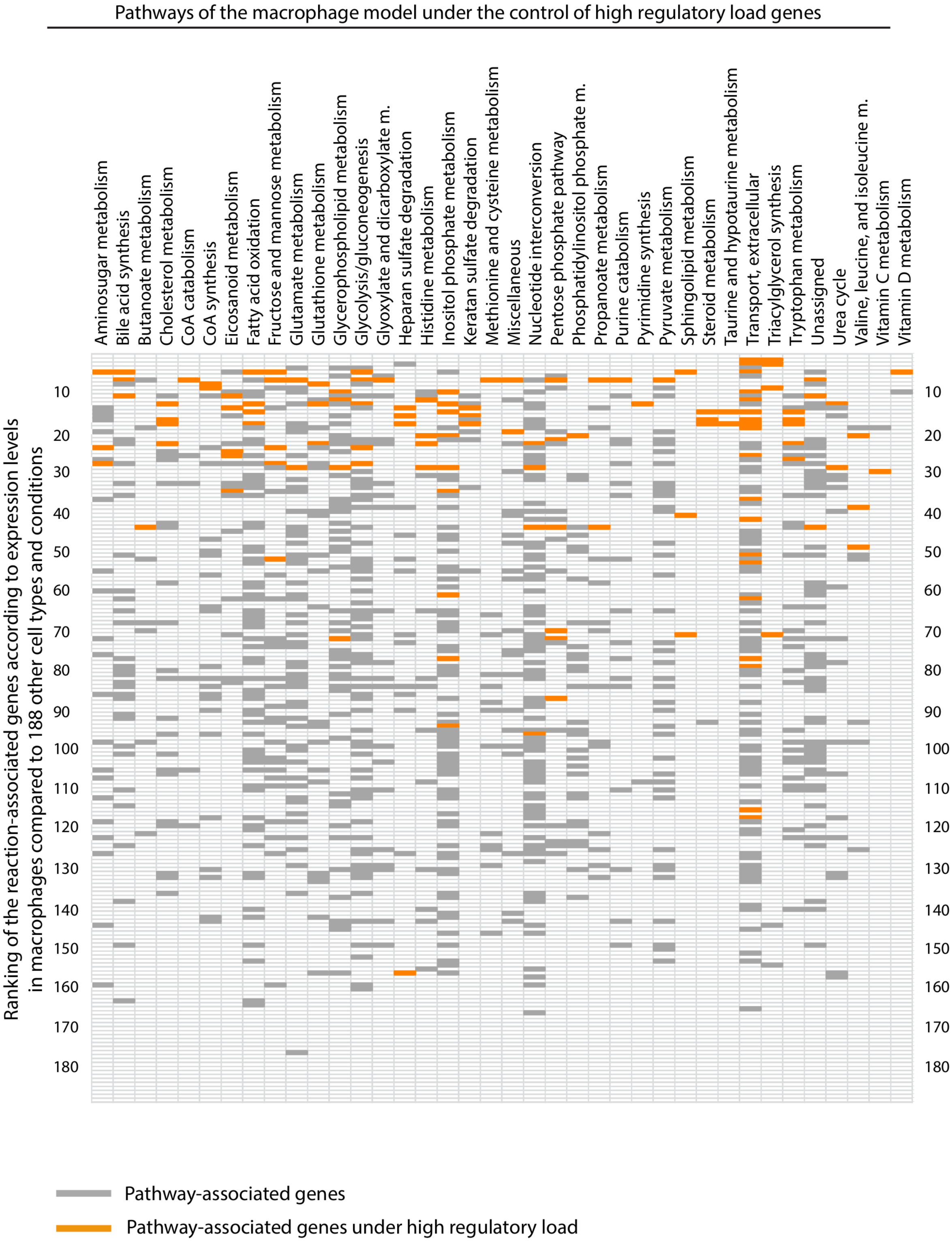


Figure 6

Figure 5.6. High-regulatory load genes show macrophage specific expression and are the highest expressed genes in their respective pathways. The normalized expression values of the 745 arrays of Primary Cells Atlas were downloaded from the Gene Expression Omnibus repository (GSE49910). The 745 arrays are subdivided in 188 separate cellular contexts. For each reactions-related gene of a pathway, the normalized expression value was retrieved and for each gene, the 188 conditions were ranked from the highest expressed to the lowest expression level with the ImpAvRank function from (Baumuratova et al., 2013). For each pathway, the genes are plotted in function of the rank of the monocyte-derived macrophages among the 188 conditions. Each rank position is represented as a box along the y-axis. High-regulatory load genes are mapped along this axis in function of their rank and depicted in orange, whereas the remaining genes in the pathway are depicted in dark gray

reconstruction such as Recon2 (data not shown), further highlighting the importance of using context-specific models and cell type-specific epigenomic data.

Taken together, genes associated to reactions at important control points of the macrophage metabolic network such as transporters or other pathway entry points are particularly enriched for high regulatory load, and exhibit abundant and cell type-specific expression patterns, possible enabling cell type-specific control of the downstream pathways.

5.4.5 Entry to alternative bile acid synthesis pathway via CYP27A1 is under high regulatory load and depends on multiple transcription factors

An interesting example among pathways with differential activity in macrophages is the bile acid synthesis pathway, which also serves as the major cholesterol catabolism pathway. Consequently, it also produces intermediates like oxysterols that serve as regulators of gene expression through their role as endogenous ligands for transcription factors like liver X receptors (LXRs). The bile acid synthesis pathway has two genes with high regulatory load in macrophages, CYP27A1 and ACP2, which are also the highest expressed genes of the pathway throughout the differentiation from monocytes to macrophages (Fig. 5.8a). Both genes are the most macrophage-specifically expressed genes of the pathway (Fig. 5.8b) and CYP27A1 shows the most abundant expression in different macrophage cells and selected dendritic cells (Fig. 5.8c). CYP27A1 is known to be involved in catalyzing the mitochondrial reactions of the classic, or neutral, bile acid synthesis pathway in the liver (Pasquali et al., 2014; Björkhem, 1992). In addition, CYP27A1 is also responsible for the first reaction of the alternative, or acidic, pathway to hydroxylate cholesterol directly in the mitochondria to 27-hydroxycholesterol in extrahepatic tissues, in particular in macrophages (Fig. 5.8d) (Björkhem et al., 1994). Therefore CYP27A1 is a prime example of a high regulatory load gene potentially integrating multiple signals to control

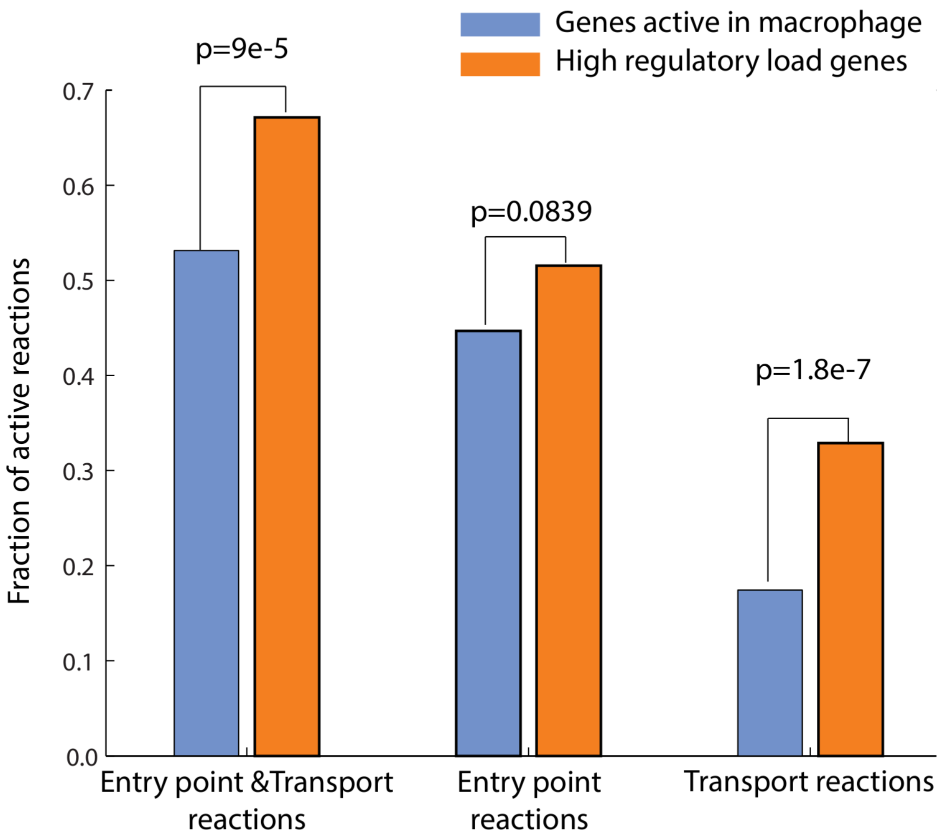


Figure 5.7. High regulatory load genes control transport and entry point reactions in macrophages. The enrichment of transport reactions and other entry point reactions under high-regulatory load among the gene-regulated reactions of the macrophage model was computed using hypergeometric test. An entry point is defined as the first reaction of a pathway change when considering the flux direction. In addition, the transport reactions and entry point reactions were tested separately to estimate their contributions to the observed enrichment

an entry point reaction of an alternative pathway.

Finally, to test which transcription factors could be responsible for the high regulatory load of CYP27A1, we analyzed microarray data from the FANTOM consortium for knock-down experiments of 53 transcriptional regulators in THP1 monocytes (Fig. 5.8e) (Suzuki et al., 2009). Interestingly, almost half of the tested knock-downs affected CYP27A1 expression directly or indirectly with 18 TFs showing significant downregulation after transfection and additional 4 regulators causing a significant upregulation (Fig. 5.8e). Among the TFs causing significant change in CYP27A1 expression upon knock-down were many known myeloid regulators that were also predicted as key TFs based on our de novo motif analysis (Additional file 1: Figure S4), including CEBPfamily members, Forkhead-family members, and FLI1. Moreover, CEBPB and SREBF1 knock-downs both led to decreased expression levels just above the significance cut-off with p-values of 0.055 and 0.054, respectively, altogether indicating that CYP27A1 expression is controlled by multiple transcription factors in monocyte-derived macrophages.

5.5 Discussion

Here we present a novel workflow, FASTCORMICS, for the fast, robust and accurate generation of metabolic models based on transcriptomics data generated by microarrays and use FASTCORMICS to generate multiple metabolic models across tens of primary cell types. This analysis reveals a cell type-specific usage of the alternative branches in metabolic networks and raises the question about the epigenetic regulation of metabolism in different cell types. To address this question we performed genome-wide mapping of active enhancers in primary human macrophages and integrated these data with metabolic models of monocyte-to-macrophage differentiation to expose the metabolic genes under high regulatory load in macrophages and general features of these genes within metabolic networks. Interestingly, the high regulatory load genes show the most abundant and cell type-selective expression profiles of the genes within any metabolic pathway and control in particular the different transport and entry point reactions of the pathways.

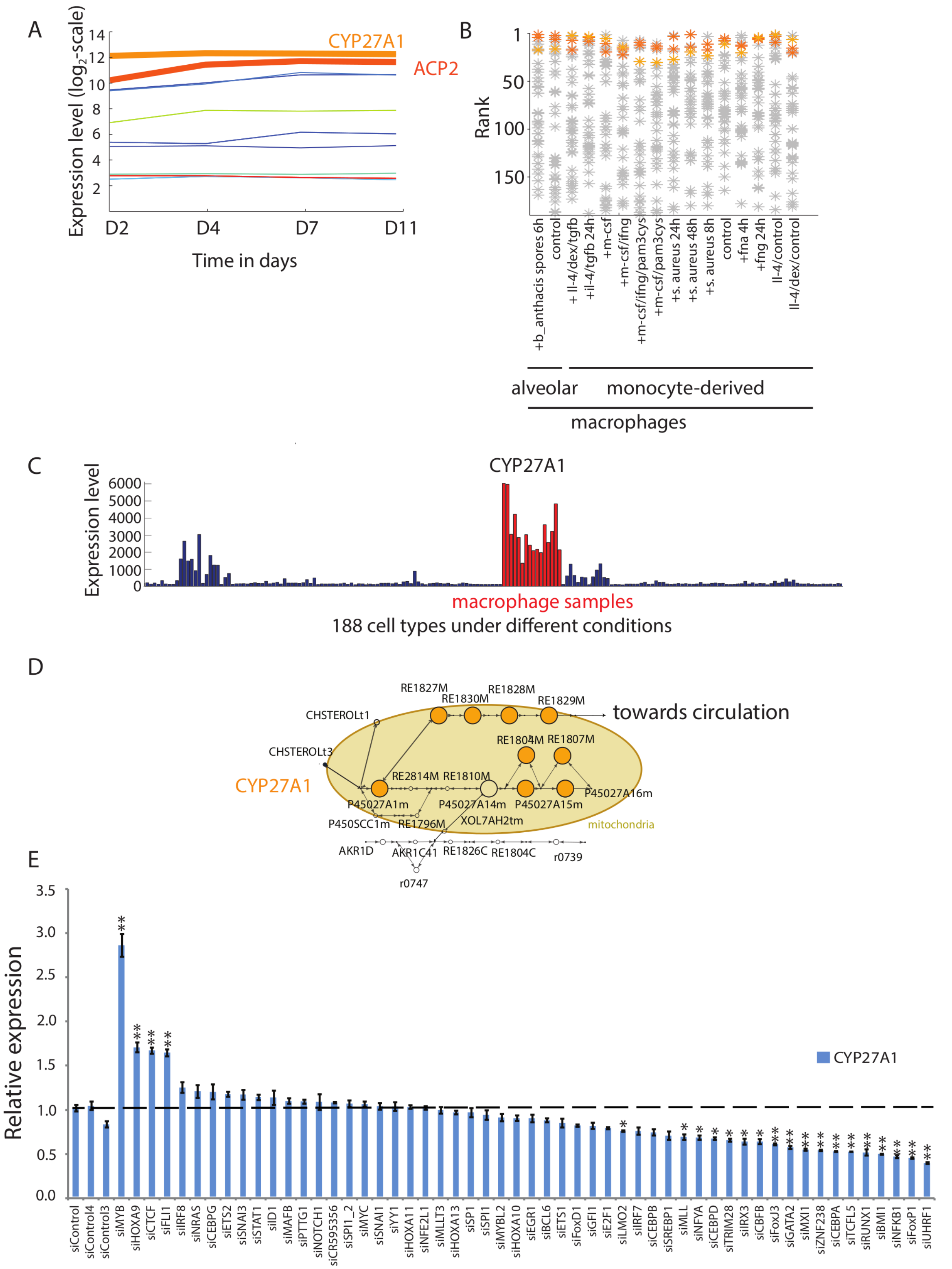

A
Expression level (log_2-scale)
14
12
10
8
6
4
2
CYP27A1
ACP2
D2
D4
D7
D11
Time in days
B
1
50
100
150
Rank
+b_anthacis spores 6h
control
+ Il-4/dex/tgfb
+il-4/tgfb 24h
+m-csf
+m-csf/ifng
+m-csf/ifng/pam3cys
+m-csf/pam3cys
+s. aureus 24h
+s. aureus 48h
+s. aureus 8h
control
+fna 4h
+fng 24h
Il-4/control
Il-4/dex/control
alveolar
monocyte-derived
macrophages
C
CYP27A1
Expression level
6000
5000
4000
3000
2000
1000
0
macrophage samples
188 cell types under different conditions
D
RE1827M
RE1830M
RE1828M
RE1829M
towards circulation
CHSTEROLt1
CHSTEROLt3
RE1804M
RE1807M
CYP27A1
RE2814M
RE1810M
P45027A16m
P45027A1m
P45027A14m
P45027A15m
P450SCC1m
RE1796M
XOL7AH2tm
mitochondria
AKR1D
AKR1C41
RE1826C
RE1804C
r0739
r0747
E
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Relative expression
CYP27A1
siControl
siControl4
siControl3
siMYB
siHOXA9
siCTCF
siFLI1
siIRF8
siNRAS
siCEBPG
siETS2
siSNAI3
siSTAT1
siID1
siMAFB
siPTTG1
siNOTCH1
siCR595356
siSPI1_2
siMYC
siSNAI1
siYY1
siHOXA11
siNFE2L1
siMLLT3
siHOXA13
siSP1
siSPI1
siMYBL2
siHOXA10
siEGR1
siBCL6
siETS1
siFoxD1
siGFI1
siE2F1
siLMO2
siIRF7
siCEBPB
siSREBP1
siMLL
siNFYA
siCEBPD
siTRIM28
siIRX3
siCBFB
siFoxJ3
siGATA2
siMXI1
siZNF238
siCEBPA
siTCFL5
siRUNX1
siBMI1
siNFKB1
siFoxP1
siUHRF1
**
*

Figure 8

Figure 5.8. The alternative pathway of bile acid synthesis is controlled by high regulatory load on CYP27A1 gene. a The mean normalized expression values of the genes implicated in the bile acid synthesis pathway based on the microarray data across the four differentiation time points are depicted. High regulatory load genes (CYP27A1 and ACP2) are presented in different shades of orange and with a thicker line than other genes of the pathway. b For each gene of the bile acid synthesis pathway, the rank of the expression level in the macrophage samples among the 188 conditions and cell types of the Primary Cell Atlas are shown by an orange or gray star (*) for high regulatory load genes and genes that are not under high regulatory load, respectively. Genes in the top ranks are situated in the top of the figure. c The expression profile for CYP27A1 across all 188 conditions and cell types from Primary Cell Atlas as arbitrary expression units. Macrophage samples are depicted in red. d The alternative pathway of bile acid synthesis was visualized in Cytoscape. To allow the alternative pathway to carry a flux, an exchange reaction was added, enabling the export of the last metabolite from the cell. Reactions predicted active in this modified macrophage model (day 11) are depicted as filled black circles or filled orange circles for reactions under control of high-regulatory load genes. The size of the nodes correlates with the number of associated enhancers. The reaction names correspond to reaction-identifiers of Recon 2. e The normalized expression levels of CYP27A1 in microarray analysis of THP-1 monocytes in a series of knock-down experiments for 53 different transcription factors or regulators and three unspecific control siRNAs retrieved from the FANTOM consortium database. Expression values were normalized to the first control siRNA (siNC) and represent the mean expression values SD (n 3). $Student's$ t-test determined the significance of changes in response to siRNA transfection (*, p <0.05; **, p <0.01)

An interesting example of a metabolic enzyme controlling an entry point of an alternative pathway is CYP27A1, which is encoded by one of the 55 metabolic genes under high regulatory load in macrophages. The alternative bile acid synthesis, which is initiated by CYP27A1 in mitochondria, is also the major cholesterol catabolism pathway in macrophages. Therefore the regulation of CYP27A1 can be used to control cholesterol homeostasis in macrophages, and other extra-hepatic cell types, on one hand through initiating cholesterol catabolism, and on the other hand due to production of intermediate oxysterols that indirectly influence cholesterol efflux and biosynthesis (Escher et al., 2003). CYP27A1 has therefore many implications to the development of atherosclerosis and cardiovascular disease. Moreover, a mutation of CYP27A1 in humans causes a disease called cerebrotendinous xanthomatosis (CTX), which leads to accumulation of cholesterol in brain and tendons and is accompanied by neurological dysfunctions, including parkinsonism, as well as increased rate of atherosclerosis (Cali et al., 1991; Shanahan et al., 2001).

The disease-association of CYP27A1 is consistent with previous findings from us and others that genes under high regulatory load, or controlled by so called superenhancers, are often associated with disease (Galhardo et al., 2014; Hnisz et al., 2013). Indeed, our current findings suggest that within any cell type the top 10th percentile of highest regulated genes are significantly enriched for disease-association (which is also the reasoning behind the applied cut-off for high regulatory load in this study) (Galhardo et al.). This is possibly due to their central roles as network hubs within gene regulatory networks, forming integration points for multiple signals. While this combinatorial regulation can be robust, it might also increase the likelihood of being affected by alterations such as single nucleotide polymorphisms (SNPs) in the regulatory regions. This would be consistent with the experiments of Siersbaek et al. who showed that omission of one TFs activity, that of glucocorticoid receptor (GR), in early adipocyte differentiation had more potent effect on super-enhancer activity than on activity of more isolated GR binding sites (Siersbæk et al., 2014).

Integrating gene regulatory networks with metabolic networks is an important and necessary step for truly global understanding of metabolism and its regulation. However, the role of high regulatory load genes in control of metabolism has not been previously specifically addressed. We find that high regulatory load genes, the central hubs of the gene regulatory networks, are significantly enriched for controlling transport reactions or other entry points of pathways, like in the case of CYP27A1, with almost 70% of such reactions located at transporters/entry points (Figure 5.7). They are the most abundantly expressed genes within the pathways and show most variation between cell types, suggesting they are used as the control points for cell type-specific metabolism. This is consistent with the findings in metabolic control analysis that for linear pathways with similar individual kinetics assigned to the different enzymes the flux control exerted at the upper part of the pathway and especially at the first step is much higher than in the lower part (Klipp et al., 2008).

While most high regulatory load genes do control entry point reactions, there remains a large proportion of them that do not. An interesting question is what other network positions are controlled by high regulatory load and to which end. Among the non-entry point reactions associated to high regulatory load genes in macrophages many are situated immediately downstream of branch points where a metabolite can follow two different fates within the pathway. For example, kynurininase (KYNU) is a high regulatory load gene catalyzing branch point reactions in tryptophan metabolism pathway to decide the faith of tryptophan metabolite kynurenine into downstream metabolites with inflammatory and neuroactive functions (Schwarcz et al., 2012). Similarly, UDPglucose ceramide glucosyltransferase (UGCG) is a macrophage high regulatory load gene controlling the commitment of sphingolipids to glycosphingolipid branch (Ishibashi et al., 2013). Interestingly, the enzyme is also required for capture of HIV-1 viral particles into dendritic cells and useful for the virus upon infection (Puryear et al., 2012). In addition to branch point reactions, many high regulatory genes also control reactions along the metabolic pathways. Regulation at such positions might be important for example to control accumulation of harmful or beneficial metabolic intermediates. However, it should also be pointed out that the consistency of the current human GENREs like Recon 2 is only approximately 75% and many branch or entry points might still remain unannotated.

In general the context-specific reconstruction of metabolic network models with FASTCORMICS as presented here might be severely influenced by the quality of the used GENRE, especially when applying automated annotation pipelines. As the overall runtime of FASTCORMICS is very low, it allows performing cross-validation studies as described earlier and thereby detecting high-confidence reactions with multiple evidence for their presence in the context-specific model of interest. In general, with run-times in the order of seconds FASTCORMICS clearly outperforms competing algorithms and might serve as an important corner stone of many future applications.

We've used FASTCORMICS to generate metabolic models of hundreds of human cell types, including a time-course of monocyte-to-macrophage differentiation. As discussed above, the cholesterol metabolism was predicted to be increased between day 2 and day 11 of the differentiation (Figure 5.4), consistent with the ability of healthy resident macrophages to uptake and release lipids, as part of their generic cleaning role or in a targeted way through low density lipoproteins (LDLs) (Brown and Goldstein, 1985; Ross, 1999). This may also be correlated to the observed increase in the active reactions in phospholipid (more precisely glycerophospholipids in Figure 5.4) metabolism or overall increase in expression of genes associated to reactions in triacylglycerol synthesis, the two other main lipid families that constitute LDLs. Also, the differentiation process between day 2 and 11 predicts an increase in the metabolism of the essential amino acid tryptophan, in particular with respect to its kynurenin metabolite (Halaris, 2013). In addition, also the metabolism of other relevant metabolites like the eicosanoids, another important signaling family (Norris and Dennis, 2014), or glutamate (Gras et al., 2006), were increased, as well as pathways with fewer specific implications for macrophage biology like inositol phosphate, pyruvate and propanoate metabolisms. FASTCORMICS is therefore able to contextualize a qualitative and quantitative difference between monocytes and macrophages.

More detailed analysis of pathophysiologic states of monocyte-to-macrophage differentiation in inflammatory conditions could be another informative application of the predictive efficacy of FASTCORMICS. Indeed, inflammation of the vascular wall is for example disturbing the uptake and release equilibrium of lipids by macrophages, making them become lipid-loaded foam cells by mechanisms involving oxidized LDL, and thus participate to the development of atherosclerosis (Ross, 1999). Also, inflamed microglia (the resident brain macrophages) have been shown to produce enhanced quantities of quinolinic acid, a metabolite of the tryptophan-derived kynurenin, which can become toxic to the brain and could participate to the development of various neurodegenerative processes among which Alzheimer's and Parkinson's diseases (Schwarcz et al., 2012; Tan and Yu, 2012).

FASTCORMICS allows in a modular fashion to use medium information and/or a biomass function for improved contextualization. This would allow generating more accurate context specific network models. However, it might be challenging to obtain specific medium and biomass information for reconstructing a cell's metabolism residing within a multi-cellular context. In the presented work a general biomass function was used. Future progress in the respective analytical methods will therefore help to further improve the contextualization via FASTCORMICS by providing more accurate specific medium and biomass information.

FASTCORMICS is based on the discretization of the expression data with Barcode, which to our knowledge currently is the most robust and reliable discretization method. The pre-processing step with Barcode allows circumventing the need of setting an arbitrary expression threshold that segregates between expressed and non-expressed genes as e.g. in (Folger et al., 2011; Becker and Palsson, 2008; Zur et al., 2010). As such a threshold is arbitrary and critical for the output metabolic models as in response to this threshold complete branches, alternative pathways, or subsystems might be included or excluded, thereby heavily changing the functionalities of the model. Further, Barcode shows a better correlation between predicted expression and protein expression than competing discretization methods for the segregation of gene expression and allows reducing batch and lab-effects that affect measurements (Zilliox and Irizarry, 2007).

An interesting future research question is if better context-specific reconstruction could be obtained by applying continuous weights instead of discrete core assignments or by a combination of the two approaches. While in general continuous weights might be able to better capture the continuous distribution of expression values, this would require the setting of arbitrary parameters to convert expression values into optimization weights, thus rendering this approach biased to arbitrary settings as also stated by Machado et al. (Machado and Herrgård, 2014). Thus the overall performance of such approach needs to be investigated in more detail in future work. FASTCORE can form a valuable building block here as well. Such continuous approach might also be suitable to treat genes with reactions associated in multiple pathways (like the discussed CYP27A1 example) more efficiently, where a stringent including of core reactions without integration of the expression context of the remaining reactions in the pathway might not be the best approach.

Furthermore FASTCORMICS can easily be adapted for the integration of other omics types, like data from next generation sequencing methods such as RNA-seq, while special attention has to be paid to the data type specific discretization step.

5.6 Conclusion

FASTCORMICS allows obtaining high-quality, robust models in a high-throughput manner. This allows the use of metabolic modelling as routine process for the analysis of expression data. Further integration with gene regulatory network data opens possibilities for better understanding of the upstream events and identification of novel drug targets such as the genes under high regulatory load which we here find to control entry points of pathways in the macrophage metabolic network.

5.7 Methods

5.7.1 The FASTCORMICS workflow

The general workflow of FASTCORMICS (Additional file 1: Figure 4.2) contains a discretization step with Barcode to obtain for each gene a z-score which indicates the number of standard deviations of the gene of the considered array above the mean expression value of the same probe set in an unexpressed context measured across thousands of arrays. Genes with a z-score equal or below zero, corresponding to the mean of the distribution of the non-expressed genes, are considered as inactive and are associated with a discretization score of âĹŠ1. Genes with z-score above 5, corresponding to the threshold value benchmarked by Zilliox et al. (Zilliox and Irizarry, 2007), are considered as expressed and get a discretization score equal to

1. Genes with z-score larger than 0 but smaller than 5 form the undetermined gene set and get a discretization score of zero. The discretization score 1 is then mapped to the consistent generic model via the model's Gene-Protein-Reactions Rules (GPR) to obtain a list of active reactions (core reactions). For reactions that are under the control of one gene only, the discretized gene score is directly mapped to the reaction. If more genes are associated to a reaction, the relationship between the genes and the reaction is given by Boolean Rules. A Boolean AND means that all the genes have to be expressed to activate the reaction, which is typically the case when a reaction is controlled by a complex of proteins. Therefore the minimum of the discretization score is mapped to the reaction. A Boolean OR signifies that only one gene has to be expressed. The maximal discretization score value is then mapped to the reaction. Boolean ANDs and ORs can be combined inside the same rule, e.g. ((A AND B) OR C), in this example the minimal value D is computed of A and B, and then the maximum between D and C is matched to the reaction. Reactions associated to a discretization score of -1, are considered as inactive and removed from the model by setting their bounds to zero. Reactions with a discretization score of 1, form the set of core reactions that are fed into a modified version of FASTCORE (mFC) that allows leaving a set of reactions non-penalized besides defining core and non-core reactions. The inclusion of non-penalized reactions is, unlike core reactions, not forced, but only preferred over the inclusion of penalized non-core reactions. Barcode-supported transporters are put to the set of non-penalized reactions. Transport reactions are generally under the control of promiscuous genes (in the consistent version of Recon 2 e.g. the gene SLC7A6 controls 294 reactions) and therefore transporters are not included into the core set as otherwise whole subsystems would be included in the output model due to one gene. Nevertheless, the inclusion of Barcode-supported genes should be preferred over non-core reactions (which are not supported) and therefore Barcode-supported transporters are not penalized. For more details on FASTCORE see the original paper (Vlassis et al., 2014). A MATLAB implementation of the FASTCORE and FASTCORMICS algorithms will be available for download from $bio.uni.lu/systems_biology/software$. Three optional steps can be included in the workflow. The first one allows further constraining the model with respect to the medium composition, if this information is available. Uptake reactions for metabolites not being present in the medium are shut down and FASTCC (Vlassis et al., 2014) is run to remove reactions that cannot carry a flux due to these additional constraints. The second optional step allows adding a biomass function or of production reactions of specific metabolites to the model. FAST-

CORMICS forces the biomass function or/and the corresponding exchange reactions to carry a flux while penalizing the inclusion of non-core reactions (Additional file 1: Figure 4.2). Core reactions, including core transporters, are not penalized in order to find, within the different alternatives sets of reactions that allow the production of biomass or required metabolites, the one that contain the highest number of core reactions. The output reactions of the modified FASTCORE are then added to the core set and the modified FASTCORE is run a second time to now force all the core reactions to carry a flux while penalizing the non-core reactions. Transport reactions are removed from the core set, but are not penalized during the reconstruction to favor Barcode-supported transporters over non-core reactions that are not supported. If no biomass function is added, FASTCORMICS is only run once. Finally a cross-validation step can be performed to assign a confidence score to the reactions included in the model. For the latter, the building process is repeated multiple times, leaving at each run one core reaction out. Reactions (core and non-core reactions) present in all the runs are supported by at least 2 core reactions and therefore are assigned a high confidence score, whereas core reactions that were not recovered during their left-out run are supported by the expression value of their own gene(s) only. The remaining non-core reactions have a low confidence score as they themselves are not supported by Barcode and their inclusion in the model depends on a single core reaction only. The same process can also be repeated with the non-expressed reactions set in order to estimate if a sub-branch of a pathway was removed from the model due to the presence of a single unexpressed reaction or to multiple inactive reactions that interrupts the flux.

5.7.2 Reconstruction of generic cancer models

The NCI dataset composed of 174 Hgu133plus2 arrays corresponding to 59 cancer cell lines was downloaded from the Cell miner web page (Shankavaram et al., 2009) and read in R version 2.15.1 using the affy package (1.36.1). The arrays were normalized with the frozen Robust Multi-array Average package (fRMA version 1.14.0) (McCall et al., 2010) using the core target and the median polish option. The normalized values were then processed with Barcode using the hgu133plus2frmavrecs vector (version 1.1.12) into a list of probe sets IDs with the respective z-score (Additional file 1: Figure 4.2). The list of probe sets was then converted in Entrez IDs via the hgu133plus2.db package (Carlson M. R package version 3.0.0). The z-scores are converted into discretization scores (1, 0, -1) using the above mentioned expression threshold of 5 and non-expression threshold of 0. The ubiquity of expression (sum of the discretization score for a gene over all arrays) was computed for each gene and a list of genes Entrez IDs with their respective score was then loaded in Matlab (version 2013a) and mapped via the Gene Protein Reactions Rules (GPR) to the consistent version of Recon1 (consistRecon1, 2469 reactions) and Recon2 (consistRecon2, 5317 reactions, the lower bound of the AATAI reaction was set to zero to be consistent with the reversibility information of the model) obtained with FASTCC. To be consistent with the experimental setup of Folger et al. (Folger et al., 2011) reactions tagged as active in 90 174 arrays were included in the core set with the exception of Barcode-supported transport reactions. Reactions with ubiquity of expression score equal below zero in 90% were removed from the model as explained previously. To be comparable to the results of Folger et al. and Luo et al. (Folger et al., 2011; Luo et al., 2008) the growth of the cancer cells was simulated on RPMI medium, the uptake reactions of the consistent versions of Recon 1 and Recon 2 were constrained with respect to the medium composition (Additional file 2: Table S2, medium composition sheet). Uptake reactions for the metabolites present in the medium were automatically added within FASTCORMICS if required by the biomass function taken from Wang et al. or for the inclusion of a barcode-supported pathway. Beside a biomass function, a sink reaction was added to Recon 1 to balance the glycogenin self-glucosylation reaction (Folger et al., 2011; Luo et al., 2008). The exchange reaction of triacyglycerides in Recon 2 was left unconstrained. FASTCC was run to remove reactions that are not able to carry a flux due to these additional medium constraints (Additional file 1: Figure 4.2).

The modified FASTCORE was then run on the medium-constrained models forcing the biomass function to carry a flux while penalizing the inclusion of non-core reactions. The reactions required to allow a biomass production were then added to the core set and the modified FASTCORE was run again now forcing the inclusion of all core reaction while penalizing the noncore reactions with the exception of core transporters.

The pre-processing step with Barcode for large data sets was performed due to memory issues on a Linux compute server with 3.0 GHz Intel Xeon CPU and 16 GB RAM and took 3 min. The model reconstructions were performed on a standard 3.40 GHz Intel Core i5 computer with 4 GB RAM in 38 and 288 s for cancer 1 and cancer 2 respectively, so that the overall computational time of the FASTCORMICS workflow is below 5 min.

5.7.3 Validation of the cancer models by comparison to an shRNA screen on cancer cell lines

A in silico knock-out experiment was performed on the obtained cancer models as previously described by Folger et al. applying Flux Balance Analysis (FBA) (Wang et al., 2012; Bordbar et al., 2010). In Folger et al. a gene is considered essential if its knock-downs results in a decrease of the growth rate of more below 1% of the maximum. To allow, a comparison with Folger et al. the 1% criteria was kept. The lists of essential genes were compared to the ranked list of 8000 genes established by Luo et al. based on an shRNA knockdown screen on cancer cell lines. The rank of the essential metabolic genes were compared to the rank of the remaining metabolic genes (set of genes associated to Recon2 minus the essential genes) with a Kolmogorov-Smirnov test (KS-test). In addition 1,000,000 random sets of genes of the same size were created and the respective KS-test was computed for evaluating the likelihood to obtain the same or better KS-score by chance (Additional file 1: Table S1).

To further validate the predicted essential genes, a list of neoplasia-related genes was retrieved from DisGeNET, a database for gene-disease associations (Folger et al., 2011; Queralt-Rosinach and Furlong, 2013). A hypergeometric test was performed to evaluate the enrichment of neoplasia-related genes in the predicted essential genes (Additional file 1: Table S3).

5.7.4 Reconstruction of 156 context-specific models of selected primary cells

The Primary Cells Atlas (GSE49910) gathering 745 arrays of the HG-U133_Plus_2 platform taken from 100 separate studies, corresponding to >180 different experimental conditions in tens of primary cell types, was downloaded from the Gene Expression Omnibus repository (Mabbott et al., 2013). 156 arrays corresponding to 63 cell types were selected favoring control samples in order to derive undisturbed cell-specific metabolic pathways in resting cells (see Additional file 3: Table S4 for the list of selected arrays). The arrays were normalized with fRMA using the median polish and core target option and then discretized with the Barcode package (as in Reconstruction of generic cancer models). The probe set IDs were converted to Entrez IDs with the (hgu133plus2.db) package as above, which were mapped to the consistent version of Recon2. 156 models (one model per array) were built using the previously described FASTCORMICS workflow. The high efficiency of FASTCORMICS allowed to perform this task within 4.5 h (5 min for the pre-processing with Barcode on 3.0 GHz Intel Xeon CPU and 4.5 h for the model reconstructions on a standard 3.40 GHz Intel Core i5 computer with 4 GB RAM). The primary context-specific models were represented as a matrix of 5317 rows corresponding to the reactions of the consistent Recon2 version and 156 columns for the number of models. The presence of the reactions in the different models was indicated by ones and the absence by zeros. The level of similarity between the different models pairs was determined via the Jaccard index. The resulting Jaccard index matrix of size 156 times 156 was then clustered with the MATLAB clustergram function (Figure 5.2a).

5.7.5 Isolation of primary human monocytes from blood

Primary human monocytes were extracted from the blood samples of anonymous healthy male donors, donated by the blood transfer centre of the Luxembourgish Red Cross and were used for diverse experiments in agreement with the convention between the Luxembourgish Red Cross and the University of Luxembourg from 16.05.2011 and following the principles of Helsinki Declaration.

The blood was diluted 1:1 with phosphate buffered saline (PBS) (Invitrogen, Life Technologies). Afterwards the peripheral blood mononuclear cells (PBMC), were isolated by Ficoll density gradient separation. Therefore the blood-PBS suspension was transferred to leucosep tubes (Greiner Bio One,) containing 15 ml of ficoll (VWR). After a 10 min centrifugation (1000 x g, room temperature, without break), the mixture separated into an upper phase of plasma, followed by the white peripheral blood mononuclear cell (PBMC) layer, the separation gel ficoll and erythrocytes in the bottom of a 50 ml tube. The PBMC layer was collected and washed twice with ice-cold MACS buffer (PBS, pH 7.2; 0.5% bovine serum albumin (BSA) (Sigma-Aldrich, Seelze, Germany) and 2 mM ethylenediaminetetraacetic acid (EDTA) (Sigma-Aldrich, Seelze, Germany) at 4 ℃ for 10 min at 300g. From this step on cells were kept on ice. Following the separation of the PBMCs the CD14+ cells (monocytes) were isolated from the total PBMC fraction by using the MACS technology from Miltenyi Biotec. In this method, anti-CD14+-antibodies are conjugated with superparamagnetic particles CD14 MicroBeads (Miltenyi Biotec) and bind to the CD14 antigen on the cell surface of CD14+ cells. By using a magnet MACS separator (Miltenyi Biotec) and LS Columns (Miltenyi Biotec) the CD14+ cells can be separated from the rest of the PBMCs. Before the CD14+ cells were separated, the PBMCs of one blood preservation were mixed with 200μl of CD14 MicroBeads and incubated for 30 min at 4 ℃ on a rotating wheel. Afterwards the cells were washed with MACS buffer and centrifuged at 300 g for 10 min at 4 ℃. The cells were again suspended in MACS buffer and loaded on a pre-washed LS-column which was put on a MACS separator, and contained a preseparation filter (Miltenyi Biotec) on top, in order to avoid a blocking of the column. Subsequently the column with the CD14+ cells was washed and the CD14+ cells were eluted from the column with MACS buffer, after taking away the MACS separator.

5.7.6 Differentiation of primary human monocytes into macrophages

After the successful isolation of the CD14+ monocytes, the cells were counted and seeded in a density of 2 x 10e6 cells/ml, either in a 10 cm2 plates (of about 20 x 10e6 cells) (Thermo scientific) in order to perform ChIP experiments or in 6-well plates (of about 4 x 10e6 cells/well) (Thermo scientific) to extract RNA. For culturing and differentiation of monocytes to macrophages RPMI 1640 medium (VWR) was supplemented with 10% human serum off the clot, type AB (A&E Scientific, PAA, Pasching, Austria, lot number: C02108-1021), 0.1 mg/ml streptomycin (Invitrogen, Life Technologies), 100 U/ml penicillin (Invitrogen, Life Technologies) and 0.1 mM Lglutamine (Invitrogen, Life Technologies). The cells were kept at 37 ℃ under a 5% CO2 atm. The medium was changed, during the differentiation process of monocytes to macrophages, 4 and 7 days after seeding. For the RNA extraction and the subsequent array analysis, the cells were extracted 2 days, 4 days, 7 days and 11 days after seeding (see Figure 5.3). In order to perform ChIP experiments the chromatin of day 11 cells was cross-linked (see Figure 5.3).

5.7.7 Morphology of primary monocytes and macrophages by microscopy

The morphology of the monocytes and macrophages was visualized by using the microscope Axiovert 40C (Zeiss) with a magnification between 10x and 20x, the camera AxioCAM MRC (Zeiss) and the software Zen blue (Zeiss). Unstained cells were used to generate pictures of the monocytes, macrophages and intermediate states (see Figure 5.3).

5.7.8 Total RNA extraction

The RNA was extracted by using TRI Reagent (Sigma- Aldrich). The cells in the 6-well plate were lysed with 500 μl of TRI Reagent per well. Following complete lysis, 100 μl of chloroform (Sigma-Aldrich) were added to the lysate, vortexed for 20 s and incubated at room temperature for 3 min. These steps were followed by 15 min centrifugation at 4 ℃ with full speed, during which the mixture separated into different phases, with the upper phase containing the

RNA. This RNA containing phase was mixed with equal volume of ice-cold isopropanol (Sigma-Aldrich) in order to precipitate the RNA overnight at 20 ℃ to recover also all small RNAs. The pelleting of the RNA was done at full speed for 20 min at 4 ℃. Then the RNA was washed with 70% ice-cold ethanol (VWR) and centrifuged for 5 min at full speed and 4 ℃ . Finally, the RNA pellet was dried and solved in RNase-free water. The concentration and the purity of the RNA were measured with the NanoDrop 2000c (Thermo scientific). The quality of the RNA was measured with the 2100 Bioanalyzer from Agilent Technologies and all the RNA samples had a RIN number greater or equal to 8.

5.7.9 Data analysis of mRNA microarrays

One hundred ng of total RNA was used to process Affymetrix Human Gene 1.0st microarrays. The Ambion WT Expression Kit was used to reverse transcribe the RNA into cDNA and to purify it according to manufacturer's instructions (The Ambion WT Expression Kit Protocol For Affymetrix GeneChip Whole Transcript (WT) Expression Arrays Part Number 4425209 Rev.B 05/2009). Then the cDNA was fragmented, labeled and hybridized on the arrays according to The GeneChip Whole Transcript (WT) Sense target Labeling Assay Manual Version 4 from Affymetrix (P/N 701880 Rev.4). The arrays were washed and scanned after 16 h of hybridization.

Microarray data were analyzed using Partek Genomics Suite, R Software (http://www.R-project.org/). First, 15 CEL files containing raw probe intensities were imported into Partek and data were preprocessed using the robust multi-array average (RMA) algorithm (Irizarry et al., 2003). Preprocessing aims at estimating transcript cluster (gene) expression values from probe signal intensities. Boxplot and relative log expression calculated on resulting gene expression values were then used to assess the quality of data; no outlier was found. Principal component analysis (PCA) was then performed for data reduction and factor analysis. PCA was able to separate data according to the time. According to this observation, the Linear Models for Microarray (Limma) (Smyth) package was used to identify genes for which expression changed throughout the time. Gene expression values were imported into R, Limma was applied and all times were compared to the gene expression values generated from D2 cells. Resulting p-value was adjusted for multiple testing errors using false discovery rate (FDR) (Benjamini and Hochberg, 1995). The microarray expression data can be found at ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) with accession number E-MTAB-3089.

5.7.10 Reconstruction of the monocyte-macrophage models

The 15 microarrays of the Hugene.1.0.st.v1 platform were read into R version 2.15.2, with the oligo package (1.22.0) and normalized with the fRMA package (1.14.0)and the hugene.1.0.st.v1frmavecs (1.0.0) vector and then discretized with Barcode. The probe sets were converted in Entrez ID via the hugene10sttranscriptcluster.db package (MacDonald JW. R package version 8.2.0). The discretized values were then mapped to the consistent version of Recon 2 (version 3, the lower bound of the AATAI reaction was set to zero to be consistent with the reversibility information of the model). In order to minimize the effect of patient-specific variation on the models, reactions tagged as active in the cells of 3 out of 4 donors for each time point, respectively 2 out of 3 for time point D4 were included in the core set, with the exception of the core transporters that were removed from the core set, but not penalized during the building process. Similarly, reactions tagged as inactive in 3 out of 4 or 2 out 3 donors were removed from the models as explained previously. Cross-validation was used to determine the confidence levels of the included and excluded reactions. Reactions with a high level of confidence are supported by at least two core reactions. Reactions with moderate confidence level are reactions only supported by barcode. Reactions with a weak confidence level are not supported by expression, but needed to generate a consistent network model. Excluded reactions with a high confidence score were never included in any simulations suggesting the presence of other excluded reactions in the branch. Whereas, inactive reactions with a low confidence level were excluded only due to their low expression level.

5.7.11 Chromatin immunoprecipitation (ChIP)

The primary human macrophages (15.5-21 x 106 cells/ 10 cm2 dish) were fixed for 8 min with 1% formaldehyde in PBS (Sigma-Aldrich) and were washed before with PBS. Then the formaldehyde was quenched for 5 min with a final concentration of 125 mM of glycine (Sigma-Aldrich). The fixed cells were washed twice with PBS, the PBS of the second washing step contained protease inhibitor (PI, Roche Applied Sciences). This step was followed by scraping the primary human macrophages in the PBS-PI solution and spinning them down at 4 °C for 5 min at 1300 rpm. The pellet was resuspended in 1500 μl of ice-cold lysis buffer (5 mM 1,4-piperazinediethanesulfonic acid (PIPES) pH 8.0 (Sigma-Aldrich); 85 mM potassium chloride (KCl) (Sigma-Aldrich); 0.5% NP-40 (VWR)) containing PI and incubated for 30 min on ice.

Afterwards, the cell lysate was centrifuged at 5000 rpm for 10 min at 4 ℃. The pellet was resuspended in 750 μl of ice-cold shearing buffer 50 mM Tris Base pH 8.1 (Sigma-Aldrich); 10 mM EDTA, disodium salt (Sigma-Aldrich); 0.1 sodium dodecyl sulfate (SDS) (Sigma-Aldrich); 0.5 sodium deoxycholate (Sigma-Aldrich,) into which fresh PI was added. After 30 min incubation on ice, the chromatin was sheared with a sonicator (BioruptorTM Next Gene, Diagenode) during 30 cycles at high intensity (30 s off and 30 s on). The sheared chromatin samples were then centrifuged at 15.000 rpm for 10 min at 4 ℃ in order to pellet the remaining cell debris. The supernatant, which contains the chromatin, was transferred to a new tube.

Twenty-five μl of the sheared chromatin was purified to check the size of the sheared DNA on an agarose gel. The concentration of the DNA was determined by the Qubit dsDNA HS Assay Kit (Invitrogen) and the Qubit 2.0 Fluorometer (Invitrogen) according to the manufacturer's instructions.

For each immunoprecipitation 5 μg of sheared chromatin and 0.5 μg as input were used. In order to preclean the chromatin, the sheared chromatin was diluted with modified RIPA Buffer (140 mM NaCl; 10 mM Tris pH 7.5; 1 mM EDTA; 0.5 mM ethylene glycol-bis(2-aminoethylether)−N,N,N',N'-tetraacetic acid (EGTA) (Sigma-Aldrich); 1% Triton X-100 (Sigma-Aldrich); 0.01% SDS; 0.1% sodium deoxycholate (Sigma-Aldrich) containing PI, up to 1200 μl, and incubated for 30 min with 25 μl of protein A magnetic (PAM) beads (Millipore) at 4 ℃ on a rotating wheel. Afterwards, the PAM beads were captured with a magnet and the supernatant containing the pre-cleared chromatin was transferred to a new tube. This pre-cleared chromatin was then incubated overnight with 5 μg of an antibody against the active enhancer mark H3K27ac (Abcam, product No.: ab4729). On the next day the antibodies were captured with 25 μl of PAM beads during 2 h on a rotating wheel at 4 ℃. This step was followed by pelleting the magnetic beads on the tube side by using a magnetic stand. The supernatant was discarded and the PMA beads, linked with the antibodies and, due to this, to chromatin, were washed twice with 800 μl of wash buffer 1 (20 mM Tris pH 8.1; 50 mM NaCl; 2 mM EDTA; 1% TX-100 (Sigma-Aldrich); 0.1% SDS), once with 800 μl Wash Buffer 2 (10 mM Tris, pH 8.1; 150 mM NaCl; 1 mM EDTA; 1% NP40; 1% sodium deoxycholate (Sigma-Aldrich); 250 mM lithium chloride (LiCl) (Sigma-Aldrich) and twice with 800 μl TE buffer (10 mM Tris pH 8.1; 1 mM EDTA pH 8). All the washing steps were performed for 2 min on a rotating wheel at room temperature, followed by pelleting the beads on a magnetic stand. In order to detach the chromatin from the PMA beads and to get rid of the proteins, the washed beads as well as the input were incubated with 100 μl elution buffer (0.1 M sodium bicarbonate (NaHCO3) (Sigma-Aldrich); 1% SDS) and 10 μg RNase at 65 ℃ overnight on a shaking platform. 5 μg of proteinase K were added after the overnight step for 90 min at 42 ℃. Afterwards, the DNA was purified with a QIAquick PCR clean-up kit. Again, the DNA concentration was measured by using the Qubit dsDNA HS Assay Kit and the Qubit 2.0 Fluorometer according to the manufacturer's instructions.

5.7.12 ChIP-Seq

ChIP-Seq was performed with chromatin from three different donors. For each donor one ChIP sample using an antibody against H3K27ac and one input sample were sequenced. The sequencing of the ChIP samples was done at the Genomics Core Facility in EMBL Heidelberg. For sequencing, single-end-reads were used and the samples were processed in an Illumina CBot and sequenced in an Illumina HiSeq 2000 machine. The sequencing data can be found at Gene Expression Omnibus GEO (http://www.ncbi.nlm.nih.gov/geo/) with accession number GSE68798.

5.7.13 Quality control and identification of enriched genomic regions

After sequencing the quality of the raw reads was controlled by applying the software FastQC v.0.10.1 (http:// www.bioinformatics.babraham.ac.uk/projects/fastqc/). The reads that had a low quality base pair calling or the ones, which were detected as read artefacts were removed from the dataset (minimum quality score of phred 10 across the read length was required). Furthermore, these reads were read stacks collapsed using the FASTX software v.0.0.13 (http://hannonlab.cshl.edu/ fastx_toolkit/index.html). The reads, which were not rejected by the quality control, were aligned to the human genome version 19 (hg19). This was done by applying the software Bowtie v0.1.25 (Langmead et al., 2009) (one mismatch allowed, maximum three locations in the genome from which the highest quality match was reported).

The software QuEST v.2.4 (Valouev et al., 2008) was used in order to identify enriched regions. The 44-mers were aligned to the hg19 by using the mappability parameter 0.88. The ChIP enrichment was set to 15 and the ChIP to background enrichment to 3. BigWig files were generated, which were used to visualize the data with the software Integrated Genome Viewer (IGV) v.2.3 (http:// www.broadinstitute.org/software/igv/home) 69] (Figure 5.5).

5.7.14 Generation of enhancer-to-gene associations

The identified enriched regions were extended to both sides to create sequences of 450 bp in length that would include the sequences immediately flanking the modified histone and thereby capture the potential TF binding sites. Afterwards the lists of enriched regions were overlapped, generating a list of 16290 common enriched regions. For this the software Galaxy (http:// galaxyproject.org) (07.07.2015, version 15.05) was used. Afterwards the common enriched regions were analyzed for their association to all genes by using the software GREAT v.2.0.2 70] with the setting 'single nearest gene within 500 kb'. For the metabolic genes these associations were further manually curated to make sure that the identified loci do not contain alternative highly expressed genes or previously unannotated transcripts such as noncoding genes.

For de novo motif analysis only enhancers associated to genes upregulated $2-fold$ in macrophages (D11 cells compared to D2 cells) with a FDR < 0.05 were considered. These were derived independently of the GREAT analysis by taking the TSSs of the up-regulated genes and extending ± 200 kb and overlapped with the common enriched enhancer regions by using Galaxy (http:// galaxyproject.org).

5.7.15 De novo motif identification

In order to identify the common sequences of the putative enhancers associated to genes upregulated in macrophages, de novo motif identification was performed by using the software MEME-ChIP (http://meme-suite.org/ tools/meme-chip) (08.07.2015) version 4.10.01 (Machanick and Bailey, 2011). Database Jolma 2013 (Jolma et al., 2013) for known TF binding motifs was used to identify the TFs that might have bound to the putative enhancers. For the analysis of the data default settings were applied. For the MEME options the expected motif site distribution was set to zero or one occurrence per sequence. The count of motifs was set to 10. The minimum width of the motifs was set to 6 bp and the maximum width to 25 bp. For the CentriMo analysis, the software was asked to find uncentered regions and include sequence IDs.

These analyses generated a list of enriched motifs, which were linked to TFs that potentially bind these motifs. The enriched motifs were identified by analysing the meme-chip.html data generated by the MEME-ChIP software. Therefore, the e-value generated for the motifs found by the different programs were considered. The motifs with a known TF binding site are listed in the Additional file 1: Figure S4. The motifs with an e-value < 0.05 were considered. In addition, if a TF was associated to a motif that occurred several times, the motif with a lower e-value was considered.

5.7.16 The control points regulated by high-regulatory load genes

The enrichment of transport reactions under high-regulatory load among gene-regulated reactions was computed via a hypergeometric test. For this test, the population size N, the number of successes states in the population K, the number of draws n and the number of successes k are respectively equal to the number of reactions in the macrophage model under gene control, the number of transporter under gene regulation, the number of genes under high-regulatory load and the number of transporter under high-regulatory load. In Recon 2, most transporters reactions were assigned to so-called transport subsystems (i.e. transport nuclear) but nevertheless some transporters are part of other pathways. For this study, we defined transporters as reactions that carry metabolites between compartments and therefore both types mentioned above were included.

To investigate if others reactions, beside transporters, were under high regulatory load, an enrichment test for reactions under the control of high-regulatory load genes situated at entry points of pathways was performed. An entry point is defined as the first reaction after a pathway change as annotated in the input models. Thereby the flux direction is taken into account. To identify entry points of pathways, for each reactions of the macrophage model under gene control flux variability analysis was performed to determine in which direction the reaction can carry a flux. It is important to note that the reversibility of the reactions given by the bounds only partially addresses this question, as the reversibility of adjacent reactions constraint the overall flux direction as well. Consumed metabolites in each reaction were identified in order to determine if one or more reactions producing these metabolites were part of a different pathway. Inorganic metabolites, CO2, known cofactor combinations (see Additional file 6: Table S8) and metabolite couples that do not change during the chemical reaction and therefore have the same chemical formula, were not considered. Reactions only composed of metabolites defined previously as cofactor are not taken into consideration as this would lead to a high numbers of false positives. Further Acetyl-coA is considered as a co-factor if it acts as a coA donor in the reaction. For each of the remaining metabolites, the producing reactions and the pathways to which they belong were determined. If the producing reaction is not a transporter and belongs to a different pathway than the considered reaction, the latter is an entry point of the pathway. In case that the producing reaction is a transporter, the transported metabolites and initial compartment were determined. If the pathway is the same as the considered reaction, the latter is an entry point after a compartment change, otherwise it is an entry point after a pathway change.

To avoid a bias of the enrichment test due to the transporters reactions, the latter were not considered for the test and therefore the population size N was defined as the number gene regulated reactions in macrophage being not part of the transporter set. The success in population K was defined as the set of reactions being entry points, the number of draw n are the reactions being entry points while excluding transporters and the successes k are the entry points under high regulatory load. The test was repeated without excluding transporters and successes were then defined as transporters or entry point reactions under high regulatory load. Finally, pathways were visualized in Cytoscape via the outputNetworkCytoscape function of the Cobra toolbox in MATLAB.

Benchmarking procedures

This chapter was previously published as ***Benchmarking Procedures for High-Throughput Context Specific Reconstruction Algorithms*** *in Frontiers in Physiology (2016).*

Contents

6.1 Summary and contributions

One of the aims of metabolic modelling is to design tools that could be used in a standard and high-throughput manner in clinics or in industry. To this purpose, the accuracy and predictability of existing tools and workflows have to be improved to allow their use in a high-throughput and automated way. The first step in order to realize these aims is to assess the tool in a standardized and objective manner through the establishment of a benchmark that could be applied to existing but more importantly be used to validate new context-specific algorithms. The need of a benchmark is shown by the fact that although having been fed the same input data, the seven tested algorithms produced very different models that share only 30% of reactions for the most different ones. This difference can be explained by algorithm related bias, aims and philosophy of the algorithm that must be known and taken in account by the users. Though the main aim of this work is to initiate a discussion in the community to establish new validation procedures, to increase the knowledge of the users on the specificity of the existing algorithms and finally to ultimately increase the prediction power of the existing and new algorithms.

The benchmark assesses three important criteria:

- Confidence level of the reactions included in the model
- Robustness
- Resolution power

Note that a trade-off must be found between robustness and resolution power. A high robustness can only be obtained in the detriment of the resolution power. An algorithm that build very similar algorithms regardless of the fraction of missing data will in same way build very similar models when the inputs varies due to real biological variances.

The experiment with experimental data were performed by myself whereas the experiments with artificial data were implemented by Dr. Thomas Pfau with Prof. Thomas Sauter and me.

6.2 Abstract

Recent progress in high-throughput data acquisition has shifted the focus from data generation to processing and understanding of how to integrate collected information. Context specific reconstruction based on generic genome scale models like ReconX or HMR has the potential to become a diagnostic and treatment tool tailored to the analysis of specific individuals. The respective computational algorithms require a high level of predictive power, robustness and sensitivity. Although multiple context specific reconstruction algorithms were published in the last ten years, only a fraction of them is suitable for model building based on human high-throughput data. Beside other reasons, this might be due to problems arising from the limitation to only one metabolic target function or arbitrary thresholding.

This review describes and analyses common validation methods used for testing model building algorithms. Two major methods can be distinguished: consistency testing and comparison based testing. The first is concerned with robustness against noise, e.g. missing data due to the impossibility to distinguish between the signal and the background of non-specific binding of probes in a microarray experiment, and whether distinct sets of input expressed genes corresponding to i.e. different tissues yield distinct models.

The latter covers methods comparing sets of functionalities, comparison with existing networks or additional databases. We test those methods on several available algorithms and deduce properties of these algorithms that can be compared with future developments. The set of tests performed, can therefore serve as a benchmarking procedure for future algorithms.

6.3 Introduction

Metabolic network reconstructions become ever more complicated and complete with reconstructions like Recon2 (Thiele et al., 2013) or HMR (Mardinoglu et al., 2014a) containing more than 7000 reactions. While these reconstructions are a great tool for the analysis of the potential capabilities of an organism, one challenge faced by many researchers is that different cell types in multicellular organisms exhibit diverse functionality and the global generic network is too flexible. This issue has been addressed in two ways, by manually generating tissue specific models (Gille et al., 2010; Quek et al., 2014) or by creating algorithms for automatic reconstructions (Becker and Palsson, 2008; Zur et al., 2010; Jerby et al., 2010; Agren et al., 2012;

Wang et al., 2012; Vlassis et al., 2014; Yizhak et al., 2014a; Robaina Estévez and Nikoloski, 2015). (Ryu et al., 2015) and (Robaina Estévez and Nikoloski, 2014) recently reviewed this field and give a good overview of the available reconstructions and point to many algorithms used in this context. While (Ryu et al., 2015) are more concerned with the state of the reconstructions, (Robaina Estévez and Nikoloski, 2014) focused on the applicability and properties of the available algorithms. With that many methods available, the method selection is difficult, and it is an enormous effort to try and distinguish which network, of a set of generated networks is best. Quality assessment is therefore essential but the methods used to evaluate the currently available algorithms are very diverse and it is difficult to compare them with each other. There are several approaches for validation which can essentially be split into two different categories: Consistency testing and Comparison based testing. The first is concerned with robustness against noise, e.g. missing data, and whether distinct sets of input data yield distinct models. The second commonly aims at validating the resulting model against other models or against additional data. Comparison tends to be the more common approach so far, while consistency is often ignored. This leads to the problem that algorithms are often prone to be over-specific to the comparison dataset (e.g. parameters like expression thresholds or weights working well for only one specific tissue). While comparison methods validate the reconstructed model, they are however not validating the consistency. Thus, it is possible that small differences in the input dataset can lead to vastly different networks, or even very diverse datasets yield the same models. The latter is particularly true if e.g. a biomass function is set as objective function, since it will lead to the inclusion of a multitude of reactions, which might not be necessary if a specific tissue is supplied with some metabolites by other tissues. To investigate the quality of automatically reconstructed networks it is therefore necessary to rigorously test them. In the following paragraphs, we describe multiple methods that were used in the past. Table 6.1 also gives an overview of these approaches, and details which concept was used for validation of which algorithm.

6.3.1 Methods for testing algorithmic consistency

The idea of consistency testing covers two major aspects: Robustness of the method and its capacity to distinguish slightly different contexts.

Left-out cross-validation allows identifying the reactions that if left-out from the input set would nevertheless be included (or excluded for inactive reactions) in the output model as their in-

Method	Used by
Consistency testing	
Cross validation	PRIME, FASTCORE, MBA, FASTCORMICS, iMAT
Diversity of generated models	GIMME, mCADRE, tINIT, FASTCORMICS
Comparison based testing	
Comparison with manually curated network	INIT, MBA
Comparison with additional databases	mCADRE, RegrEx, iMAT
Comparison with shRNA knockdown screens	MBA, FASTCORMICS
Comparison with literature mining	iMAT
Comparison with metabolic exchange rates	PRIME
Comparison with known metabolic functions	MBA, mCADRE, FASTCORE

Table 6.1. Overview of methods used for validation of automated tissue specific reconstruction algorithms.

clusion is supported by other reactions of the input set (Pires Pacheco et al., 2015a). The robustness of algorithms against noise can also be assessed by adding noise to the expression data i.e. by using a weighted combination of real and random data (Machado and Herrgård, 2014). The main issue using random and left-out cross validation with most of the current algorithms is that running times of several hours makes decent cross-validation with hundreds of test and validation sets infeasible. While small cross validation runs (e.g. when multiple sources of input data are available and only some sets are considered (Jerby et al., 2010)) can give an indication of robustness, they cannot replace random sampling runs, which reflect noisy data much better.

To test the diversity of generated networks, many algorithms are employed to generate multiple networks and those networks are then investigated for dissimilarity (Becker and Palsson, 2008; Wang et al., 2012; Uhlén et al., 2015; Agren et al., 2014; Pires Pacheco et al., 2015a). If networks of similar cell types group together in a clustering and networks of divergent cell types are further apart, this indicates that the method does indeed generate specific networks. While it is desirable to obtain distinct networks for distinct tissues, the optimal method should not be too sensitive to small changes in the input data. Otherwise the resulting networks are prone to overfitting to the provided input data.

6.3.2 Methods for comparison based testing

Comparison based testing is commonly employed to show the advantages of the presented algorithm compared over previous algorithms or to show the quality of the reconstructed network based on additional, formerly unknown, data. While the former has been employed for the validation of some algorithms (Wang et al., 2012; Vlassis et al., 2014; Robaina Estévez and Nikoloski, 2015), and becomes more important with an increasing number of available methods, it has also recently been used to compare multiple methods systematically (Machado and Herrgård, 2014; Robaina Estévez and Nikoloski, 2014). In the review by (Machado and Herrgård, 2014) 8 different methodologies (including GIMME (Becker and Palsson, 2008), iMAT (Zur et al., 2010) and a Method by (Lee et al., 2012)) where tested on an independent dataset. However, their focus was on comparing the quality of flux value predictions, i.e. flux bounds specific to a condition in *Escherichia coli* and yeast, and not the reconstruction of tissue specific networks, i.e. the extraction of an active sub-network.

6.3.2.1 Comparison against manually curated networks

Comparison to a manually curated tissue was employed by (Agren et al., 2012) for the INIT algorithm, when they compared their automatically generated liver reconstruction to HepatoNet. However, they were restricted to a comparison on the gene level, since the source network used by INIT was the HMR database (Mardinoglu et al., 2013), while HepatoNet used its own identifiers. As they mention the difference between the reconstructed and manually curated models was partially due to absence of genes from HMR that were present in HepatoNet. Simultaneously, it is likely that the curators of HepatoNet lacked information on some of the genes present in HMR. Thus to validate a methodology it is necessary for both the “reference” network and the source network to be compatible.

6.3.2.2 Comparison against additional datasets and databases

Similarly, many methods compare the resulting reconstructions to additional databases that contain tissue localisation data (like BRENDA (Schomburg et al., 2013), HPA (Uhlén et al., 2015) or the Gene Expression Omnibus (Barrett et al., 2013)), which was performed for multiple reconstruction methods (Shlomi et al., 2008; Wang et al., 2012; Robaina Estévez and Nikoloski, 2015). The common approach is to check for matches of either genes or proteins that the algorithm assigned to the tissue. This validation (and the results) are however highly dependent on whether the reconstruction method aims at creating a consistent network, or whether it allows inconsistent reactions to be part of the reconstruction. The latter will very likely increase the amount of correctly assigned genes, as enzymatic activities that cannot carry flux in the source reconstruction, would otherwise be excluded. In addition, when extracting reactions from a source network, the associated gene-protein reaction relations are commonly not altered. Thus genes, which are inactive in a specific tissue show up as assigned to the tissue. Removing them however, could potentially be problematic if the tissue does express the removed gene under a specific condition. In this instance the tissue reconstruction would no longer contain information about this fact, and would indicate wrong potentials of the tissue. Another method that could be used as an assessment for predictive quality of an algorithm was performed by (Folger et al., 2011) and subsequently by (Pires Pacheco et al., 2015a). They used gene silencing data from a shRNA screen and compared it with gene essentiality predictions from a flux balance analysis (FBA) analysis screen. The cancer network generated in this work showed an enrichment of essential genes in the genes indicated in the shRNA screen. In (Pires Pacheco et al., 2015a), the list of essential genes predicted by FASTCORMICS was further compared to essential genes predicted by PRIME, MBA, mCADRE and GIMME. Likewise bibliographic approaches have been employed to determine the agreement of reactions belonging to a certain subsystem in the reconstructed network and those subsystems being mentioned in connection with the reconstructed tissue in the literature (Shlomi et al., 2008).

To assess the predictive capability of the Model Building Algorithm (MBA) (Jerby et al., 2010) used flux data from a study performed in primary rat hepatocytes and compared the ability of the source reconstruction and the generated reconstruction to predict internal fluxes given the exchange fluxes (and vice versa). This allowed them to assess whether the tissue specific network was indeed performing better in estimating the internal fluxes than the generic reconstruction (in this instance Recon1). They could show that indeed the tissue specific network had a better capability to capture the actual fluxes than the generic reconstruction. This concept was also used by (Machado and Herrgård, 2014) in their assessment of multiple methods for network contextualisation. However, while contextualisation commonly aims at altering flux bounds, which leads to a good comparability of flux measurements with predictions, tissue specific reconstruction is aiming at determining the network available in a given tissue. This means that bounds from the underlying source reconstruction are used and these are often unsuitable for the tissue of interest. But as shown by (Jerby et al., 2010), even the pure network structure alteration can already improve the agreement between network fluxes and measured data, at least on a qualitative level.

A method developed by (Shlomi et al., 2009) to compare the resulting network for the effects of inborn errors of metabolism (IEM) is also often used in model quality assessment. The concept is, briefly, to analyse flux ranges of the exchange reactions of the created network and compare them with clinical indications of increased or decreased metabolite levels. This concept has also been used for assessment of Recon2 (Thiele et al., 2013) who investigated a diverse set of IEMs and could show their effect even on the level of a generic reconstruction. Similarly, the authors of PRIME (Yizhak et al., 2014a) used experimentally measured uptake and excretion rates and compared them to the secretion rates determined by the models their algorithm generated. While the former approach is commonly used to provide a qualitative assessment of increase or decrease in production potential, the latter results in a quantitative comparison. However, it requires the availability of uptake and secretion rates, which are commonly only available for cell lines and could be largely different in real tissues.

Another common approach to investigate the quality of reconstructions is the comparison with lists of metabolic functions. This approach is both used to validate automated reconstructions (Jerby et al., 2010; Wang et al., 2012) as well as manual reconstructions (Gille et al., 2010). The aim is to establish whether the reconstruction supports the current knowledge of the target tissue (e.g. a liver reconstruction should support the conversion of ammonia to urea), and to show that there are no structural issues in the reconstructed network (e.g. free regeneration of ATP or reductants).

6.3.3 A benchmark for testing tissue specific reconstruction algorithms

In this paper we present a potential benchmark that is using several of the mentioned methodologies to assess the consistency and quality of reconstructed networks and tested it with several of the available algorithms.

There are however multiple obstacles, when defining a benchmark for contextualisation algorithms. There is no such thing as a “perfect” measurement, which will always leave us with noisy data to incorporate. Furthermore, we do not yet have a contextualised model that perfectly reflects a given context which could be used as a target model. In addition, the global reconstructions are not yet complete, and will likely never be and finally, there is a wide variety of data that can be used to contextualise models. Thus, to define a benchmark we will address these questions by generating networks which we define as reference networks for out testing. The actual benchmark is preceded by a characterization of the algorithms, in which the similarity level of the context-specific reconstructions obtained with real and artificial input data is assessed. In the latter test, artificial models of different sizes were built and 50%, 60%, 70%, 80% and 90% of the reactions of these networks were used as input for the tested algorithms. The capacity of the algorithm to distinguish between different models was compared for the different percentages of input data.

In the actual benchmark, the confidence level of the reactions included in the context-specific reconstructions using real data was assessed by matching z-scores obtained by the Barcode (McCall et al., 2011) method that basically indicate the difference in intensity between the measured intensity and the intensity distribution observed in an unexpressed state and through a comparison against the confidence score at the proteomic level of the Human Protein Atlas (Uhlén et al., 2015). In a second comparison, artificial models were built and 50%, 60%, 70%, 80% and 90% of the reactions of these networks were used as input for the tested algorithms and the output models were then compared to the complete input model. The context-specific networks obtained with the real data were also tested for the functionalities established by (Gille et al., 2010).

6.4 Material and Methods

6.4.1 Models used for Benchmarking

There are currently two competing global reconstructions for humans available: Recon2 (Thiele et al., 2013) and HMR2 (Mardinoglu et al., 2013). To be able to test multiple validation techniques, we needed to select one of those reconstructions as the source network used by the tested algorithms. We decided to employ Recon2, as we used functionalities originating from HepatoNet (Gille et al., 2010), a model based on Recon1 (Duarte et al., 2007) and largely incorporated into Recon2. However we still had to modify Recon2 to allow the algorithms to fully reconstruct HepatoNet (the procedure can be found in Supplementary File 1). HepatoNet was also adapted to match reactions and metabolites with Recon2. This modified Recon2 was used as source model for all runs.

In addition to HepatoNet as a comparison model for real data, we constructed ten artificial sub-

Algorithm	Input	Publication
Akesson04	Set of inactive genes	(Åkesson et al., 2004)
FASTCORE	Set of active reactions	(Vlassis et al., 2014)
FASTCORMICS	Gene expression data	(Pires Pacheco et al., 2015a)
GIMME	Gene expression data, objective function	(Becker and Palsson, 2008)
GIM^{3}E	Gene expression data, metabolomics data, objective function	(Becker and Palsson, 2008)
iMAT	Gene expression data	(Zur et al., 2010)
INIT	Gene expression data and metabolite presence data	(Agren et al., 2012)
MBA	High, medium and low reaction sets	(Jerby et al., 2010)
mCADRE	Gene expression data	(Wang et al., 2012)
PRIME	Growth rates, gene expression data	(Yizhak et al., 2014a)
RegrEx	Gene expression data	(Robaina Estévez and Nikoloski, 2015)
tINIT	Gene expression data, functions, metabolite presence	(Agren et al., 2014)

Table 6.2. Algorithms available for tissue specific metabolic network reconstruction. Most methods can use expression data as input but there are some that need additional inputs.

networks from Recon2. Those networks were generated to be approximately equally spaced in a range between 1000 and 3500 reactions. They were generated by randomly removing up to 4500 reactions from our Recon2 version and determining the consistent part of the remaining model. The first model within ±50 reactions of equally spaced points in the interval [1000 - 3500] was selected as representative for this point. The models and model sizes can be found Supplementary File 5.

6.4.2 Characterization of the algorithms

There are many algorithms available for tissue-specific metabolic network reconstructions (see Table 6.2). In this section we will detail the algorithms used in our study and give reasons, why others were excluded.

In order to test the algorithms with real data, liver models were built by the tested algorithms using as input 22 arrays from different datasets downloaded from the Gene Expression Omnibus (GEO) (Edgar et al., 2002) database (Supplementary File 2). The same data was also used for the cross-validation assays.

6.4.2.1 GIMME (Becker and Palsson, 2008) and iMAT (Zur et al., 2010)

For the benchmarking of the GIMME (Becker and Palsson, 2008) and the iMAT (Zur et al., 2010) algorithms, the implementation provided by the COBRA toolbox (Schellenberger et al., 2011a) was used with an expression threshold corresponding to the 75th percentile. The proceedExp option was set to 1 as the data was preprocessed. For GIMME, the biomass objective coefficient was set to 10^{-4}.

6.4.2.2 INIT (Agren et al., 2012)

In the original paper, INIT (Agren et al., 2012) assigns weights to the genes associated to the input model that were computed by dividing the gene expression in the tissue of interest by the average expression across all tissues. As for the first experiment, only liver arrays were available, z-scores obtained by the Barcode (Zilliox and Irizarry, 2007; McCall et al., 2011) discretization method, were used as weights (see below).

6.4.2.3 RegrEx (Robaina Estévez and Nikoloski, 2015)

The RegrEx implementation in the supplementary files of (Robaina Estévez and Nikoloski, 2015) was used. This algorithm has previously only been used with RNA-seq data and therefore no established discretization method exist for microarray data. In order to allow a comparison with the others methods, the intensity values after frma normalization and the standard variation were directly mapped to the reactions of the model using the Gene-Protein-Reaction rules (GPR). For reactions that are not associated to any gene, the expression and the standard deviation were set to 0 and 1000 respectively.

6.4.2.4 Akesson (Åkesson et al., 2004)

For this algorithm, the data was normalized with the frma normalization method and then discretized with Barcode. Genes with z-scores below 0 in 90% of the arrays, were considered inactive and the bounds of the associated reactions, taking into account the Gene-Protein-Reaction rules (GPR), were set to 0. FASTCC (Vlassis et al., 2014) was then run to remove reactions that are unable to carry a flux.

6.4.2.5 FASTCORE z-score

For FASTCORE z-score, the expression data was normalized with frma method and discretized using Barcode. Barcode uses previous knowledge on the intensity distribution across thousands of arrays to calculate for each probe set of the analyzed array the number of standard deviations to the median of the intensity distribution for the same probe set in an unexpressed state. Genes with a z-score above 5 in 90% of arrays are considered as expressed and mapped to the reactions according to the Gene-Protein-Reaction rules (GPR) to obtain a core set that is fed into FASTCORE (Vlassis et al., 2014).

6.4.2.6 FASTCORMICS (Pires Pacheco et al., 2015a)

The expression values were first normalized with frma, converted into z-scores using Barcode (McCall et al., 2011) and further discretized using an expression threshold of 5 z-scores and an unexpression threshold of 0 z-score. Genes with 90% of the arrays above the expression threshold are assigned a score of 1 while those below the unexpression threshold are assigned a score of -1. All other genes are associated with a discretization score of 0. These scores are then mapped onto the model using the Gene-Protein-Reactions rules to obtain lists of core and unexpressed reactions. Unexpressed reactions are excluded from the model.

The FASTCORMICS workflow allows the inclusion of a medium composition, which was not used in the tests, as the aim was to provide the same information to all algorithms. A modified version of FASTCORE is then run that maximizes the inclusion of core reactions while penalizing the entry of non-core reactions. Note that transporter reactions are excluded from the core set but are not penalized.

6.4.2.7 Context-specific reconstruction algorithm that were not tested

PRIME and tINIT were not included in the tests as they require, in addition to expression data, growth rates for PRIME and information on tissue functionalities for tINIT. Determination of growth rates in multicellular organisms is restricted to cell lines or cancerous cells, as most other cell types are finally differentiated and therefore no longer divide. Since growth rates are an essential part of PRIME it was excluded from the tests. While functionalities are available for some metabolically very active tissues (like kidney and liver), they are often not available for others. Since we wanted to test a wide range of potential tissues, we decided not to employ functionalities in our input set. Therefore tINIT would be reduced to INIT as the remaining functionality is the same. Since we wanted to focus on gene expression data, which is currently the most readily available type of data, we did not add metabolomic information into our screens. GIM^{3}E would need this type of information and was therefore not tested. Finally, MBA, Lee and mCADRE took more than 5 days for a single run on 2 cores of our cluster and where therefore not included.

6.4.2.8 Similarity of the context-specific models and algorithm-related bias

The similarity level between the context-specific models built by the tested algorithms was assessed by computing the Jaccard index between each pair of models ($(A \cap B/A \cup B)$). The matrix containing the Jaccard indices was then clustered using Euclidian distance. Further, for each context-specific model, the number of reactions found by only 1, 2 up to all of the methods was computed and represented as a stacked boxplot. The colored areas represent the different models built by the tested algorithms and for each bin the colored area is proportional to the number of shared reactions.

6.4.2.9 Sensitivity and Robustness testing using artifical data

While there are methods that take continuous expression measurements into account (Colijn et al., 2009; Lee et al., 2012) (and reviewed in (Machado and Herrgård, 2014)), other methods require the user to define sets of reactions that are present (FASTCORE, MBA) or perform some form of discretization to determine the presence or absence of a gene or a reaction (Akesson, GIMME, iMAT, FASTCORMICS). The latter types of methods, using some form of presence/absence calls can be more rigorously tested for robustness, as a target model can be used to provide the present and absent genes/reactions.

We also tested these algorithms using the artificially created networks. The test was performed as follows:

The potential available information was defined as the sets of reactions present in each submodel and absent from each submodel. Based on this data different percentages of input information (50%, 60%, 70%, 80%, 90%) were provided to the algorithms. The same random samples were provided to the tested algorithms to allow a further comparison between the algorithms (generating a total of 5000 models for each algorithm). To be able to use reaction data, we modified the implementation of the GIMME algorithm to allow the direct provision of the *ExpressedRxns* and *UnExpressedRxns* fields. The model similarities were assessed by calculating the Jaccard index between each pair of models generated for input sets from different target models. In addition, the internal distances of all models generated for one target model were calculated (a total of 50000 comparisons per algorithm). Furthermore, the corresponding models for each algorithm and each tested input percentage were compared, to obtain the inter-algorithm distance.

6.4.2.10 Robustness testing using real data

For the cross-validation, 20% of the reactions were removed from the core set and transferred to the validation set. The number of these reactions that were included in the output model was determined and a hypergeometic test was computed. The process was repeated 100 times randomizing at each iteration the core set to form different validation sets. For algorithms that take continuous data as input, the cross-validation assay was adapted as follows: 20% of the gene-associated reactions were removed from the input set by setting the expression to 0 and the standard deviation to 1000 for RegrEX and the rxnsScores to 0 for INIT. But only reactions considered to be expressed with a high confidence level formed the validation set i.e. for INIT only reaction with z-scores above 5 and with expression value above 10 for RegrEX. For Akesson the validation set was composed of inactive reactions. The results for Akesson have to be taken with care as the validation set is only composed of 4 reactions. This is due to Barcode only indicating very few genes as absent, which led to only about 40 reactions being removed from Recon2.

6.4.3 Benchmarking with real data

6.4.3.1 Confidence level of the reactions

The z-scores computed by Barcode translate the number of standard deviations to the intensity distribution of the same genes in an unexpressed state. The z-scores of the genes were mapped to the reactions of Recon2 (Thiele et al., 2013), HepatoNet (Gille et al., 2010) and to the context-specific models built by the different workflows using the Gene Protein Rules (GPR). In the same way, the confidence levels assigned by the Human Protein Atlas (HPA) to the proteins of the database were mapped to the reactions of the different context-specific models.

6.4.3.2 Comparison between different tissue models

The aptitude of the algorithm to capture metabolic variations among tissues was tested using the GSE2361 dataset (Ge et al., 2005) downloaded from Gene Expression Omnibus (GEO) that contains 36 types of normal human tissues. 21 of the 36 tissues matched tissues in the Human Protein Atlas. The confidence levels of the proteins in the different tissues were first matched to the modified version of Recon2 to determine if proteins with high and medium confidence level are ubiquitously expressed or expressed in a more tissue specific manner. Then the confidence levels were matched to the corresponding context-specific models to verify if the variation observed among the tissue context-specific models matched the one observed in the Human Protein Database.

To further access the quality of the reconstructed models, the fraction of reactions of the Recon2 pathways that are active in the output models were computed. The obtained matrix was then clustered in function of the Euclidean distance (see Supplementary Figure 6.)

6.4.4 Benchmarking with artificial data

The runs using artificial data, performed for sensitivity and robustness analysis, were also used to provide an additional benchmarking measurement for the algorithms. Sensitivity and specificity and false discovery rate were calculated by comparison of the reconstructed networks with the respective target network. The artificial nature of these networks allowed us a complete knowledge of the actual target thus making these calculations possible.

6.4.5 Network functionality testing

Function testing is commonly achieved, by defining a set of metabolites that are available and can be excreted and requiring other metabolites to be produced/consumed or a reaction to be able to carry flux. The input and output can either be cast into a linear problem by adding importers and exporters or by relaxing the steady state requirement for the imported and exported metabolites. (Gille et al., 2010) used the latter definition and we adapted this approach using the following modification of the standard FBA approach:

$$
\begin{aligned}
min \quad & \sum v_i^+ + v_i^- \\
s.t \quad & b_l \leq S' * v' \leq b_u \\
& 0 \leq v_i^+ \leq ub_i \qquad \forall i \in internal\ reactions \\
& 0 \leq v_i^- \leq -lb_i \qquad \forall i \in internal\ reactions \\
& v_i^+ - v_i^- = 0 \qquad \forall i \in exchange\ reactions
\end{aligned}
$$

$$
with\ S' = [S, -S] and\ v' = \begin{bmatrix} v^+ \\ v^- \end{bmatrix}
$$

$$b_{l,i} = \begin{cases} -10000 & \forall i \in imported\ metabolites(-/=) \\ -1 & \forall i \in produced\ objectives(+) \\ 1 & \forall i \in consumed\ objectives(-) \\ 0 & else \end{cases}$$

$$and\ b_{u,i} = \begin{cases} 10000 & \forall i \in exported\ metabolites(+/=) \\ -1 & \forall i \in produced\ objectives(+) \\ 1 & \forall i \in consumed\ objectives(-) \\ 0 & else \end{cases}$$

The test is considered to be successful if there is a non zero value for all evaluators when calculating $S' \cdot v'$.

6.4.6 Computational resources

Except for RegrEx, all runs using the liver data were performed on two cores of a 2.26Ghz Xeon L5640 processor on the HPC system of the University of Luxembourg (Varrette et al., 2014) to achieve comparable running times. Tissue comparison runs and artificial simulation runs were performed on the same cluster but not limited to specific node types.

6.5 Results

6.5.1 Characterization of the algorithms

6.5.1.1 Similarity of the context-specific models and algorithm-related bias

The aim of this characterization step is to categorize the algorithms based on the similarity of their output models in order to gain insight into algorithm-related bias, requirements of the algorithms i.e. thresholds and more importantly when to use which algorithms. In an ideal case, one would expect that when fed with the same input data, the different algorithms would produce similar networks. But when comparing the context-specific liver models generated with the different algorithms and HepatoNet, only 530 reactions were found in all networks and 77 reactions of our version of Recon2 were inactive in all context-specific models and HepatoNet. The 530 reactions were found among 54 different subsystems, including reactions belonging to pathways

Model	Size	Input reactions	Gene-associated reactions	Time in seconds
GIMME	3513	2441	2087	4458
iMAT	3649	2441	2440	2098
INIT	3913	2020	2787	36002
RegrEx*	3239	1626	2576	64
Akesson	5740	1594	3715	54
FASTCORE z-score	2882	1595	2084	17
FASTCORMICS	2663	1595	1906	112

Table 6.3. Model numerics: Size, number of input reactions with high expression respectively z-score levels, fractions of input reactions set included in the output models, number of genes-associated reactions in the model and running time. *Note that RegrEx was run on a different computer with an Intel(R)Xeon(R)CPU E3 1241-v3 @ 3.50 GHz processor

expected in all tissues like i.e. the Krebs cycle, glycolysis/gluconeogenesis, but also pathways that were described to take place mainly in the liver, like i.e. bile acid synthesis ((Wang et al., 2012; Rosenthal and Glew, 2009)) or some reactions of the vitamin B6 pathway (pyridoxamine kinase, pyridoxamine 5'-phosphate oxidase and pyridoxamine 5'-phosphate oxidase) ((Merrill Jr et al., 1984)). This huge variability is due to workflow-related bias and to different strategies and aims of the algorithms. FASTCORE (Vlassis et al., 2014), expects as input a set of reactions with a high confidence level which are assumed to be active in the context of interest and therefore all core reactions are included in the output model (Table 6.3). In contrast, FASTCORMICS (Pires Pacheco et al., 2015a) only includes a core reaction if it does not require the activation of reactions with low z-scores. The main objective of GIMME (Becker and Palsson, 2008) is to build a model by maximizing a biological function. The input expression data is used to identify, which reactions are not required for the objective and can function therefore be removed from the model due to low expression values (Table 6.3). iMAT (Zur et al., 2010), (Lee et al., 2012) and RegrEx (Robaina Estévez and Nikoloski, 2015) maximize the consistency between the flux and the expression discarding reactions that have high expression values if necessary, which might be problematic if reactions have to be included in the model like i.e. the biomass function. INIT (Agren et al., 2012) uses weighted activity indicators as objective, with those having stronger evidence being weighted higher. Whereas the Akesson's (Åkesson et al., 2004) algorithm aims to eliminate non expressed reactions.

The models, when clustered in function of the Jaccard Similarity Index (Figure 6.1), form 2 branches and an outlier: HepatoNet. The first cluster is composed of algorithms that take as input continuous data and attempt to maximize the consistency between the data and the

Akesson algorithm that eliminates inactive reactions. The second cluster is composed of algorithms that discretize the data in expressed and non-expressed genes. Among this cluster, a second subdivision is observed between the algorithms that used z-score converted data (i.e. FASTCORE z-score and FASTCORMICS) and the ones that use normalized data without further transformation.

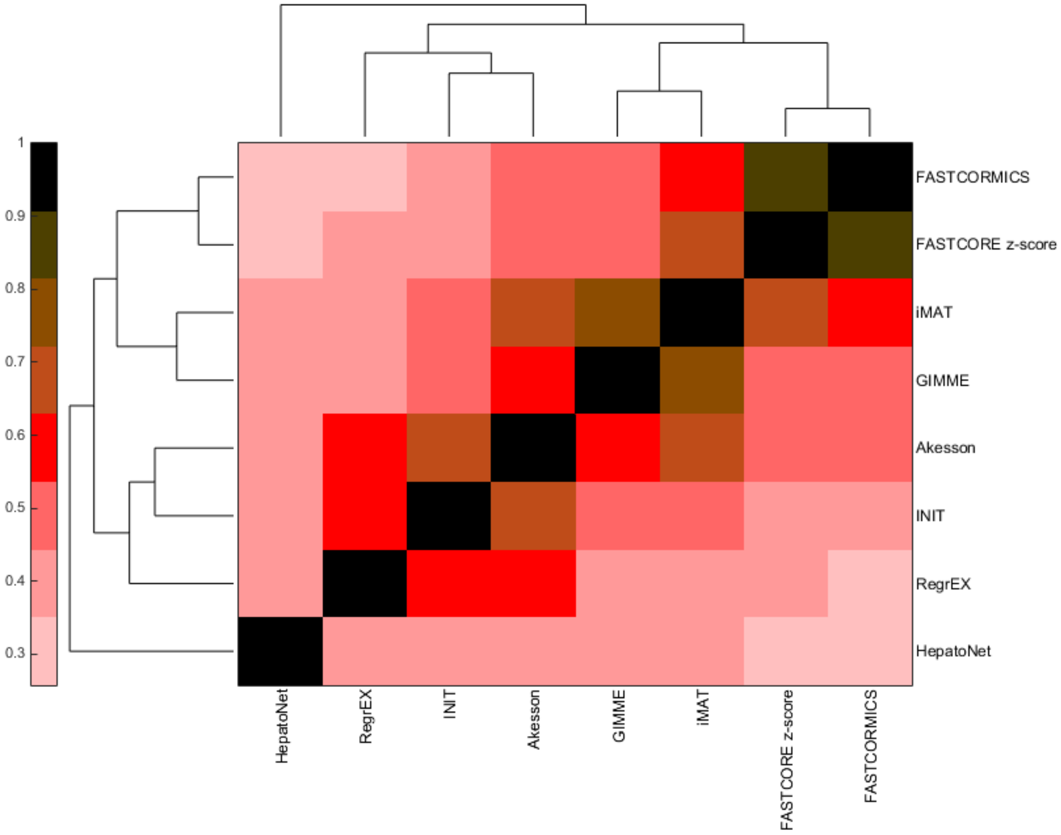


Figure 6.1. Similarity index of the models built by the different algorithms. The Jaccard index was computed for each pair of models, the rows and column were then clustered in function of the euclidean distance.Contrary to what was expected, the output models of the tested algorithms, despite having been fed with the same input show a huge variability.The descritization-based algorithms (GIMME, iMAT, Akesson, FASTCORE and FASTCORMICS) show the highest similarity levels.

Overall the highest similarity level are found between FASTCORE z-score and FASTCORMICS with a score of 85% of similarity followed by iMAT and GIMME with 77% of similarity. The lowest similarity level is found between FASTCORMICS and HepatoNet with only 26% of overlap. The largest overlap between HepatoNet and context-specific reconstructions is found for INIT with 43% of similarity. Note that the INIT model although having as input Barcode discretized data does not cluster with FASTCORE z-score or with the FASTCORMICS models but with RegrEx, suggesting that the choice to consider continuous data rather than defined core set has a larger impact on the output models.

As the algorithms were fed with the same input data, reactions that are predicted by one or only few algorithms are more likely to be algorithm-related bias (Figure 6.2).

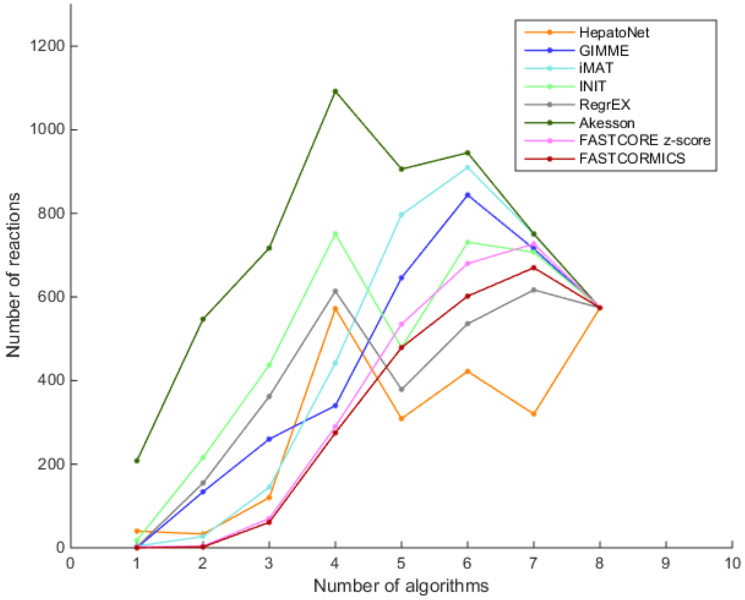


Figure 6.2. Reactions overlap: The number of reactions that are shared by the models built by the tested algorithms.. Each line represents HepatoNet or a model built by one of the tested algorithm. The plot illustrates the number of reactions that are common to 1, 2, 3 up to all of the models.

The Akesson model that contains 98.56% of the input model includes the largest number of reactions (201) that are absent in the others models.

The reactions included in the FASTCORE, FASTCORMICS, iMAT and GIMME models are for 97%, 98%, 96% respectively 89% supported by at least 3 other algorithms and display a similar profile shifted to the right. HepatoNet, INIT and the Akesson's model share 92%, 83% respectively 91% with 3 other algorithms and have different profiles from the algorithms of the first group composed of algorithm that include a discretization step.

In summary, discretization-based algorithms show the highest similarity level and therefore the lowest number of reactions due to potential algorithm-related bias.

6.5.1.2 Sensitivity and Robustness testing using artifical data

Since we noticed that there are two sets of algorithms among the discretizing algorithms, we decided to further test their properties with artificial networks by comparing resulting models from multiple runs for different models and levels of completeness of input data.

Figure 6.3 provides the average similarities for all models reconstructed for each target model at different available information percentages. (A full set of mean similarities for each percentage and each artificial model along with the data for the plots is provided in Supplementary File 1). Each square represents the mean Jaccard index of the all combinations of networks generated for different input networks (e.g. (1,2) is the average similarity of all networks generated for models 1 to all networks generated for model 2). The diagonal represents the internal similarity of all networks generated for one model. When 90% of the data is available, all the algorithms are able to distinguish variation between the different models. But with a less complete data set, inclusive algorithms lose in specificity and therefore also progressively lose the capacity to distinguish between different models. Further with 30% and 50% reactions missing, it would be expected that the algorithms get less robust, but Akesson and GIMME only show a modest decrease of robustness (as shown in the diagonal). A similar behavior for the GIMME algorithm was also described by (Machado and Herrgård, 2014) in an experiment where noise was progressively added to the input data to finally obtain a random input dataset. GIMME showed the same average error in prediction for the random and original data (Machado and Herrgård, 2014), suggesting that due to the optimization of the biomass function, the expression data has a reduced impact on the model building.

Comparing the models resulting from runs with different completeness of input data illustrates that the methods tend to converge on more complete data sets, with the Akesson approach and GIMME being more inclusive and the FASTCORE family being more exclusive (see Figure 6.4). While initially, with incomplete data, the methods are distinguishable by the networks generated, this difference becomes smaller with additional knowledge.

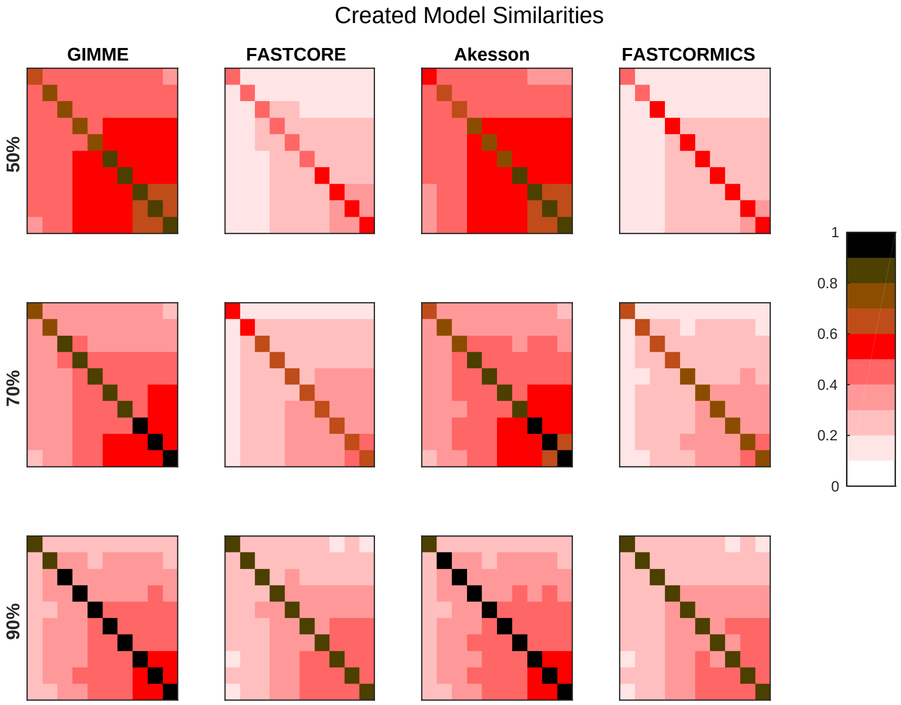


Figure 6.3. Resolution power: The plot shows Jaccard distances for the networks generated by the algorithms, when trying to create the artificial networks. For each of the ten artificial models 100 runs were performed and each square represents the mean Jaccard distance between these networks. E.g. For each percentage and algorithm, the tenth square in the first row is the mean of all pairwise Jaccard distances between the 100 models generated for artificial model 1 (the smallest) and the 100 models generated for artificial model 10 (the largest) generated for the respective algorithm and percentage. The diagonal is the mean of the pairwise Jaccard distances between 100 runs performed. The diagonal can therefore be an indicator for robustness (the brighter, the more similar the models) while the off diagonal indicates similarities between the generated models and is therefore an indicator for specificity to the input (the darker, the more distinct the generated models). When 90% of the data is available, all the algorithms are able to distinguish variations between the different models. But with a less complete data set, inclusive algorithms (here GIMME and Akesson) lose in specificity. It would also be expected that when only 50% of the data is available, the robustness decreases.

6.5.1.3 Robustness testing using real data

In order to further evaluate the confidence level of the reactions included in the different context-specific models a 5 fold cross-validation was performed. The experiment was repeated 100 times with a different validation set. GIMME, iMAT, and FASTCORMICS show the highest robustness, followed by FASTCORE and FASTCORE z-score (See Table 6.4).

Algorithms that maximize the consistency between the data and the flux, e.g. INIT and RegrEx,

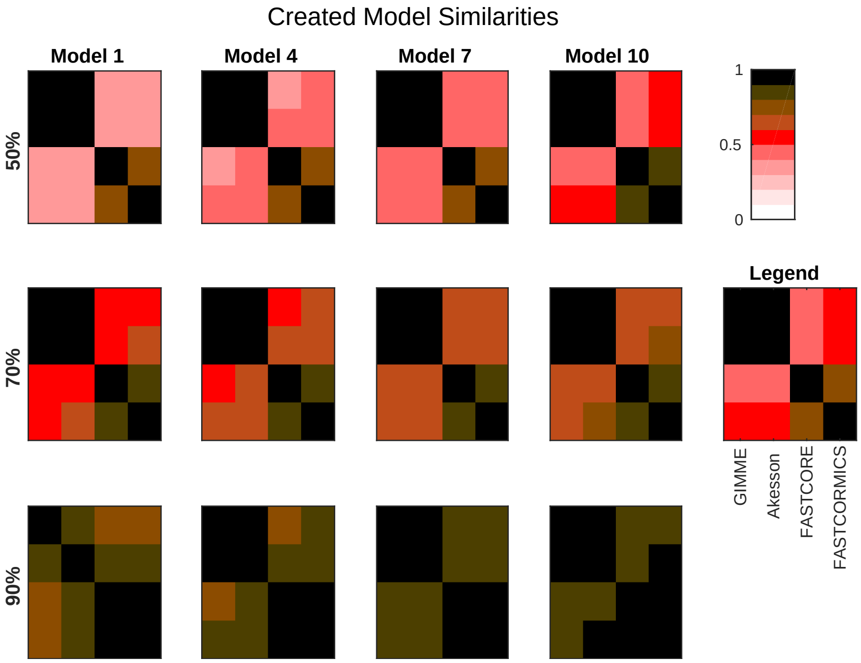


Figure 6.4. The plots show the mean Jaccard distance between the networks generated by the different algorithms for several artificial models and input percentages. For each algorithm, the corresponding networks (using the same input data) are compared. The models are provided in Supplementary File 5. Sizes are: Model 1: 961; Model 4: 1876; Model 7: 2629; Model 10: 3455. Smaller models (e.g. Model 1) tend to yield more distinguishable results, while larger models (due to a larger fraction of common reactions), tend to yield more similar networks. Overall, the difference between inclusive (GIMME/Akesson) and exclusive (Fastcore/FASTCORMICS) algorithms is clearly visible.

	Validation Set	Recovered reactions	% of Recovered reactions	Sample size	Input	hypergeometric p-value
GIMME	488	408 (6.42)	83.57%	1878 (6.42)	3871	$< 1e - 100$
iMAT	488	335 (10.85)	68.68%	1631 (29.85)	3871	$< 1e - 100$
INIT	345 (7.16)	83.7	24.26%	1931 (113.63)	4469 (7.16)	1
RegrEX	326 (12.79)	160 (19.25)	48.9%	2528 (201)	4524 (12.79)	0.96
Akesson	4	0.98 (1.41)	24.5%	5343 (6.54)	5828 (24.5)	ND
FASTCORE z-score	319	121.6 (8.26)	38.12%	1332 (27.33)	4548	0.0051
FASTORMICS	335(0.4)	192(7.79)	57.14%	1516 (27.13)	4782 (7.57)	1e-18

Table 6.4. Number and percentage of reactions recovered from the validation set, average model size over 100 reconstruction processes

are less robust with insignificant p-value. For Akesson no hyper-geometric test was performed as the validation set was too small to obtain a reliable p-value. Note that for context-specific reconstruction algorithms a trade-off has to be found between robustness and the capacity to capture differences between similar contexts. For this reason, a too high robustness might not be desirable as it would imply that the algorithm might loose in resolution power, i.e. the ability to distinguish between different sets of input data. Therefore it is also advisable to not test for robustness without testing the resolution power.

6.5.2 Benchmarking with real data

6.5.2.1 Confidence level of the reactions included in the different models

As shown by the previous similarity test, there are several alternative approaches to build context-specific models. To assess the confidence level of a reconstruction, one can quantify the confidence level of the reactions included by each algorithm. Context-specific algorithms assume that the higher the reactions associated expression levels, the more likely the reactions to be active. Following this logic, context-specific reconstructions should be enriched for higher expression levels. As the background level is non negligible and highly dependent on the probes, we corrected for probe effect using the Barcode method. The z-scores computed by Barcode translate the number of standard deviations to the intensity distribution of the same genes in an

Model 1	Model	KS p-value
FASTCORE z-score	FASTCORMICS	1e-10
GIMME	FASTCORE z-score	3e-111
iMAT	GIMME	2e-24
INIT	iMAT	$< 1e - 100$
HepatoNet	INIT	9 e-18
Akesson	Hepatonet	6e-20
consistRecon	Akesson	0.04
RegRexp	consistRecon	3e-14

Table 6.5. Comparison between the z-score distribution associated to the models build by the different methods. The p-values indicate the likelihood that the z-score associated with the model on the left side is larger than the one on the right side of the table.

unexpressed state. The z-scores of the genes mapped to the reactions of Recon2 (Thiele et al., 2013), HepatoNet (Gille et al., 2010) and to the context-specific models built by the different algorithms show that the distribution of the z-scores are for most models shifted, as expected, toward higher z-scores values with a significant p-value for all context-specific models except RegrEX (Robaina Estévez and Nikoloski, 2015). Algorithms that use a discretization method show a larger shift to the right than algorithms that maximize the consistency between the flux and the data. Within this group the FASTCORMICS (Pires Pacheco et al., 2015a) shows the most significant shift towards the highest z-score values followed by FASTCORE z-score, GIMME (Becker and Palsson, 2008) and iMAT (Zur et al., 2010). (Figure 6.5 and Table 6.5). Surprisingly, the consistent version of HepatoNet (Gille et al., 2010) is associated to slightly higher z-scores than Recon2 (Thiele et al., 2013) but significantly lower than most discretization based automated context-specific reconstructions.

Further, unlike their competitors, all the discretization-based context-specific reconstructions show an enrichment of genes with a high and medium confidence scores to be expressed at the protein level (Uhlén et al., 2015). A stronger enrichment is observed for FASTCORE z-score and FASTCORMICS with 46% and 50% of the gene associated reactions having a high or medium confidence level Table 6.6, respectively. GIMME and iMAT include 28% and 30% reaction with high or medium confidence levels, respectively. Again suprisingly, HepatoNet does not show an enrichment for high and medium confidence levels.

In summary, dicretization-based algorithms include reactions with a higher confidence level at the transcriptomic and proteomic level than their competitors.

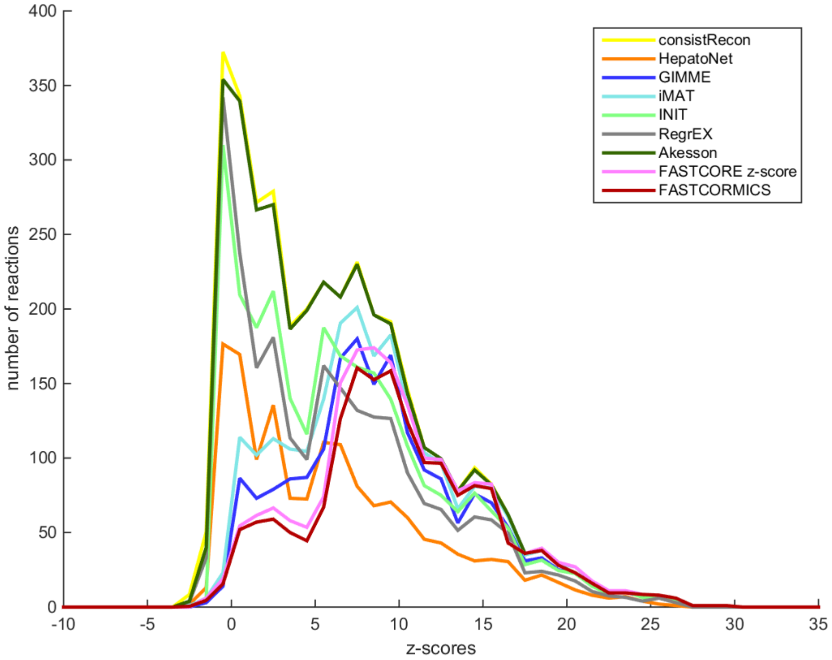


Figure 6.5. Confidence score at the transcriptomic level: Median z-score of the intensity measured in the liver samples to the median intensity distribution for the genes in an unexpressed context mapped the genes-associated reactions of Recon2 (yellow), HepatoNet (orange) the GIMME (dark blue), iMAT (light blue), INIT (green), RegrEx (gray), Akesson (dark green), FASTCORE z-score (pink) and FASTCORMICS (brown)

Discretization-based algorithms (GIMME, iMAT, FASTCORE and FASTCORMICS) are enriched for higher z-score values.

6.5.2.2 Comparison between different tissue models

The aim of a context-specific algorithm, as indicated by the name, is to build models that capture the metabolism of a cell for a given context and therefore these algorithms have to be able to capture variations in the metabolism of different tissues. To pass the following test, context-specific algorithms not only have to be sensitive (or to have a high resolution power) in order capture metabolic difference between tissues, but the reconstructions for different tissues have to be enriched for high or medium confidence levels based on HPA. The last criteria allows to identify algorithms that build different models based on noise or algorithm-related bias. In order to assess the variation among tissues in HPA, the genes with high, medium and low confidence levels for 48 different tissues were mapped to the input model Recon2, showing that very few reactions have a high or medium confidence level in all tissues. In summary, most reactions

algorithms	description	high	medium	low	not detected
Recon	number of reactions	628	641	65	265
	% of the reactions of the model	11%	11 %	1 %	5 %
	% of the gene-associated reactions	17 %	17 %	2 %	7 %
HepatoNet	number of reactions	213	266	47	108
	% of the reactions of the model	9 %	11 %	2 %	5 %
	% of the gene-associated reactions	12 %	15 %	3 %	6 %
GIMME	number of reactions	518	444	47	126
	% of the reactions of the model	15 %	13 %	1 %	4 %
	% of the gene-associated reactions	25 %	21 %	2 %	6 %
iMAT	number of reactions	574	525	55	153
	% of the reactions of the model	16 %	14 %	2 %	4 %
	% of the gene-associated reactions	24 %	22 %	2 %	6 %
iNIT	number of reactions	453	499	55	155
	% of the reactions of the model	12 %	13 %	1 %	4 %
	% of the gene-associated reactions	16 %	18 %	2 %	6 %
RegrEX	number of reactions	376	418	41	186
	% of the reactions of the model	12 %	13 %	1 %	6 %
	% of the gene-associated reactions	15 %	16 %	2 %	7 %
Akesson08	number of reactions	624	637	64	260
	% of the reactions of the model	11 %	11 %	1 %	5 %
	% of the gene-associated reactions	17 %	17 %	2 %	7 %
FASTCORE z-score	number of reactions	584	413	21	123
	% of the reactions of the model	20 %	14 %	1 %	4 %
	% of the gene-associated reactions	28 %	20 %	1 %	6 %
FASTCORMICS	number of reactions	570	391	15	73
	% of the reactions of the model	21 %	15 %	1 %	3 %
	% of the gene-associated reactions	30 %	21 %	1 %	4 %

Table 6.6. Number, percentage of gene-associated reactions and percentage of reactions of each context-specific reconstruction that have a high, medium and low confidence score to be expressed at the protein level. An enrichment in high and medium confidence level is observed for discretization-based algorithms (GIMME, iMAT, FASTCORE z-score and FASTCORMICS.

with high and medium confidence scores have a more tissue-specific expression (Figure 6.6). A similar experiment was performed with context-specific reconstructions built by the tested algorithms, in which the number of algorithms that shared a reactions was assessed (see Figure 6.7).

For RegrEX, INIT and Akesson models, the majority of reactions are found in all tissues. For GIMME, most reactions are either tissue-specific or present in all the tissues. In contrast, the models built by the members of the FASTCORE family show a distribution similar to the that obtained in Figure 6.6 for HPA. For iMAT only 8 models could obtained as the computational demands for the reconstructions of the others tissues surpasses the number of core available and the maximal running of 5 days. When looking at the confidence levels associated with the

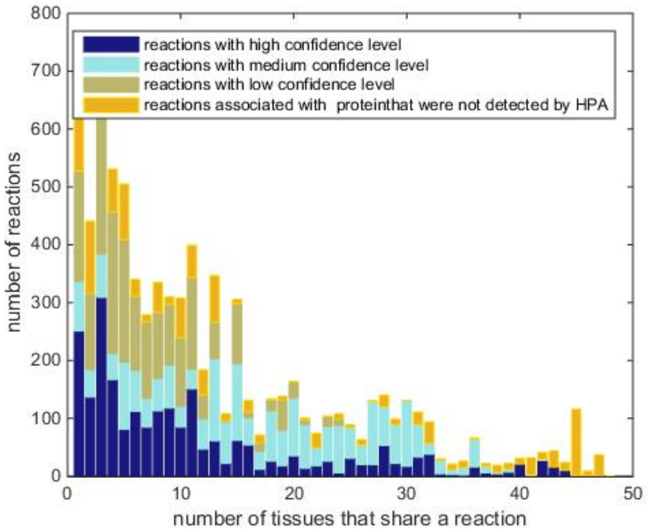


Figure 6.6. Ubiquity of expression: Number of reactions of Recon2 with a high or medium confidence level that are shared between 1, 2, 3 up to 48 tissues of the Human Protein Atlas. Reactions with a high confidence level tend to have a tissue-specific expression.

21 different tissue-specific models, FASTCORE z-score and FASTCORMICS show in 20 out 21 the highest percentage of reaction with a high or medium confidence level (see Figure 6.8). The size of the different tissue metabolic models built by the tested algorithm can be found in the Supplementary File 6).

The quality of the tissue-specific models built by the different algorithm were accessed by focusing on selected pathways known to have a more tissue-specific expression, namely bile acid synthesis and heme synthesis. The bile acid synthesis occurs in liver, although one or the other enzyme of the pathways might occasionally be expressed by other tissues ((Wang et al., 2012; Rosenthal and Glew, 2009)). As expected the FASTCORE family, GIMME and iMAT predicted that the highest fraction of active reactions are found in the liver followed by the foetal liver for the FASTCORE family members and iMAT and by placenta and foetal liver for GIMME. Whereas, the INIT models of skin, bone marrow, corpus, thalamus, pituitary gland and foetal liver had a higher fraction of active reactions than the liver model. 13 out of 36 of the tested Akesson models predicted 90% and more reactions of the bile acid pathway as active. RegrEX predicted a slightly higher fraction in the thalamus than in the liver and a comparable fraction in the ovary, the foetal brain and the corpus (Supplementary File 6, Supplementary File 1).

The heme synthesis that occurs mainly in the developing erythrocytes and in the liver ((Ajioka

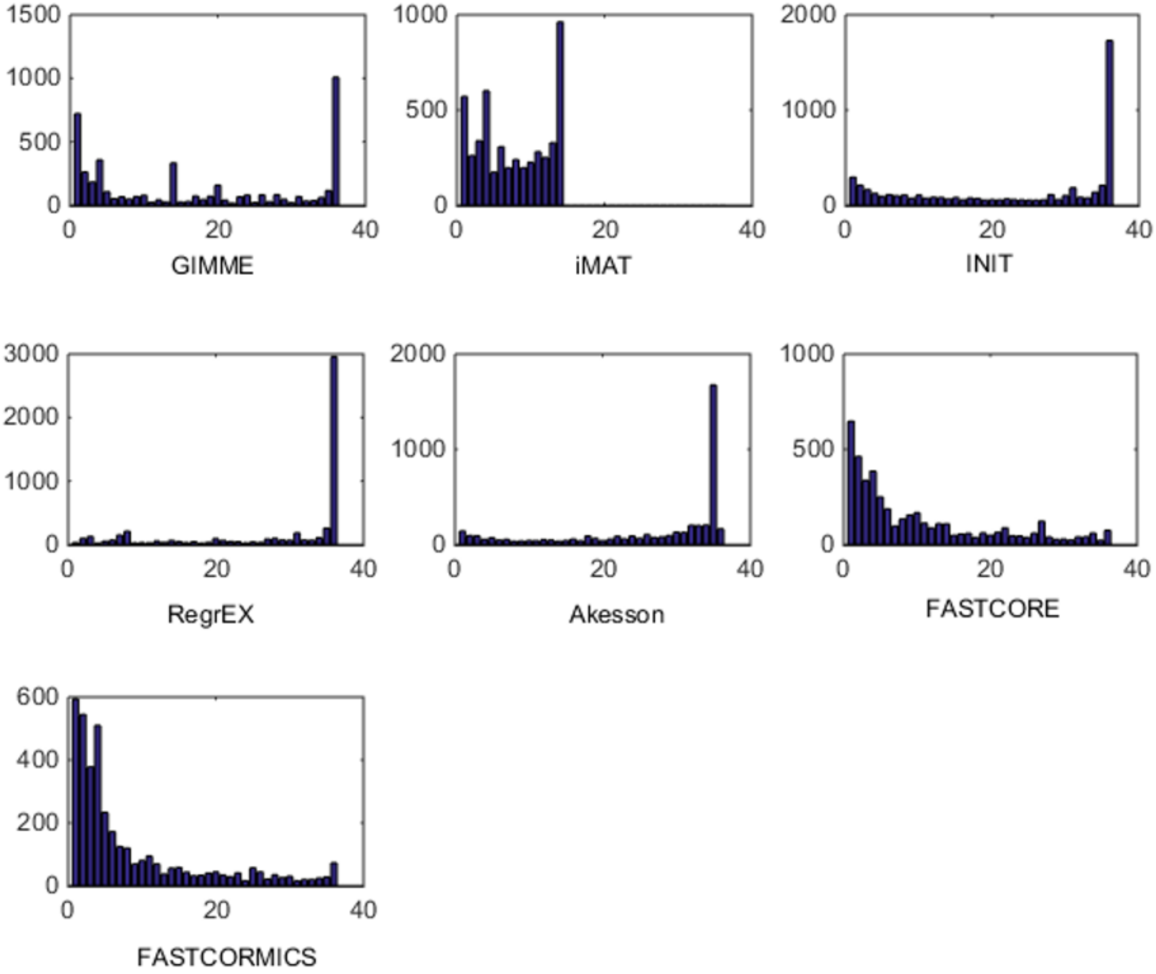


Figure 6.7. Tissue specificity of reconstructed models. Number of reactions that are present in 1, 2, 3 up to 36 tissues models. For INIT and RegrEX, more than 1500 and 3000 reactions are present in all tissues models, while a similar number is present in all but one model created by the Akesson method. Due to computational complexity of iMAT it was only possible to generate 14 out of 36 tissue models

.

et al., 2006)), was given as 100% active by the FASTCORE family and completely inactive by GIMME and iMAT in the liver. But these two algorithms predicted the pathway to be active in other tissues. As a matter fact, all the algorithms predicted the pathway to be active in others tissues than the liver. INIT, RegrEX and Akesson included this pathway in 20, 22 and all tested 36 tissues, respectively. Fewer models of the FASTCORE family contained reactions of this pathway: uterus and tyroid for FASTCORMICS and spleen, placenta, uterus, thyroid, skin, bone marrow, amygdala, lung and foetal liver for FASTCORE.

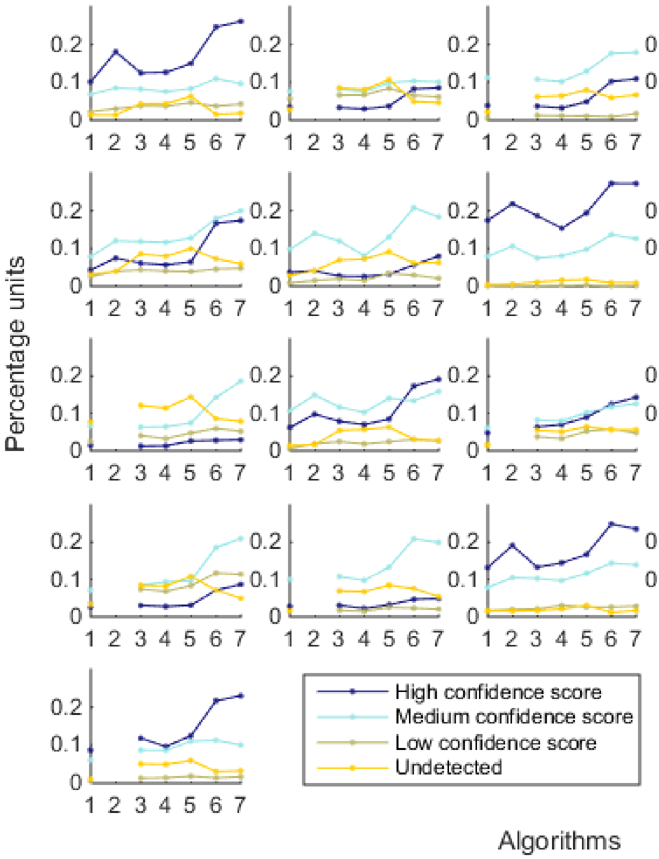


Figure 6.8. Percentage of reactions that are associated with high confidence (dark blue), medium confidence level (light blue), low confidence level (khaki) and not detected (yellow). Each subplot represent a different tissue. The x-axis represent the different algorithms: 1-GIMME, 2-iMAT, 3-INIT, 4-RegrEX, 5-Akesson, 6-FASTCORE z-score and 7-FASTCORMICS and the y-axis the percentage of reactions.

6.5.3 Benchmarking with artificial data

To further evaluate the quality of the algorithms, we also used the artificial data (see Section 6.5.1.2) to benchmark the algorithms. Comparing the resulting models with the target models, we again see that for more complete input sets, the model quality tends to become more similar (see Figure 6.9).

It is interesting to note that the false discovery rate (FDR) of FASTCORE for higher percentages is similar to those of the inclusive models, while FASTCORMICS achieves a better FDR. This indicates alternative routes to activate reactions. In general, there is again the tradeoff

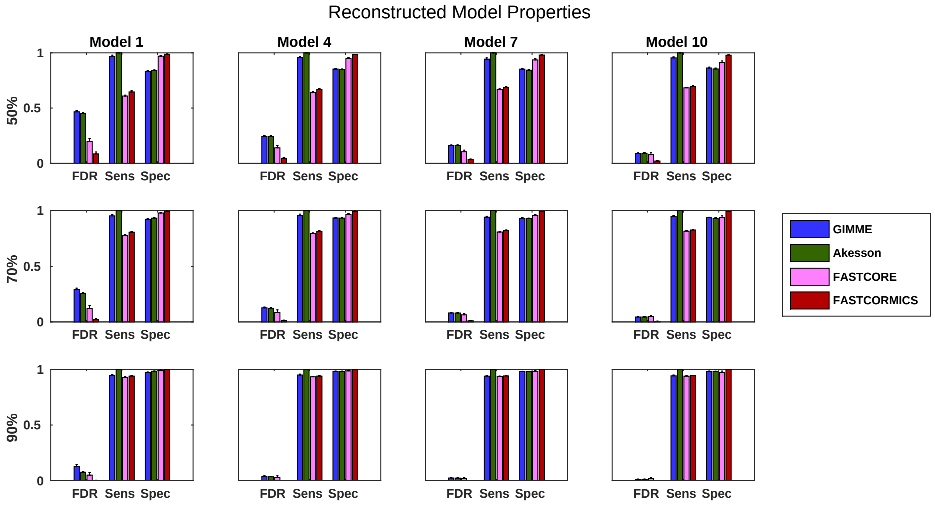


Figure 6.9. Quality measurements of the algorithms. FDR - False discovery rate, Spec - Specificity, Sens - Sensitivity. Data shown is a the mean of 100 runs for each model/input data. The model sizes are: Model 1: 961, Model 4: 1876, Model 7:2629, Model 10: 3455
While the quality of the FASTCORE models is independent of the target model size, the inclusive approaches tend to largely overestimate smaller models, when insufficient data is available. A plot with all Models can be found in Supplementary File 1.

between adding too much or too little. It is however interesting that the exclusive algorithms tend to miss targets and their sensitivity is independent on the size of the target model while this is different on inclusive algorithms. Exclusive algorithms show a better FDR than inclusive algorithms. Further, for smaller target models, the loss in precision of inclusive algorithms (1-FDR) is more pronounced for 50% and 70% of the input data, as the inclusive algorithms tend to overestimate the actual model. Similar to the previous experiment, it would be expected that the sensitivity (robustness) would decrease with an increased percentage of missing data. But the inclusive algorithms show an invariant sensitivity in function of the available data suggesting that the expression data has reduced impact on the model building. The specificity for the exclusive algorithms is independent of the target model size and are less affected by the increased missing data than the inclusive algorithms. The sizes of the different reconstructed models also indicates the trend for convergence, and a figure showing the converging sizes is provided in Supplementary File 1.

6.5.4 Functionality testing

Functional testing allows us to assess which known functions of a specific tissue are captured by a reconstruction. We used the set of functions defined in HepatoNet and formalized in Section 6.4.5 for the liver and tested them on all reconstructed networks. We noticed that the success rate of HepatoNet and the generic reconstruction Recon2 are comparable with 244 vs 247 of 310 network tasks and 109 vs 98 of 123 physiological tasks for Recon2 and HepatoNet, respectively. The discrepancy with the original publication is likely due to alternative solutions and we noticed that HepatoNet allows free production of NADH and thereby ATP (see Table 2 in Supplementary File 1). The discrepancy between the consistent and inconsistent HepatoNet is due to the formulation of the functionalities, which do not require exchange reactions but modify the b vector, thus generating implicit importers and exporters and allowing inconsistent parts of the network to carry flux.

We also noticed an important issue with functional testing: For random models, the larger the models, the higher the functionality score (with R^2 = 0.869 and 0.915 for network and physiological functions, respectively). To illustrate this issue, we generated 400 random networks by removing a random number of up to 2000 reactions from the consistent part of Recon2 and subsequently removing all reactions which could no longer carry any flux. We then tested all network and physiological functions on these networks. The results can be seen in Figure 6.10, for both the network and physiological tests.

Blue circles represent the random networks; the consistent HepatoNet and the original HepatoNet are displayed in orange, and show a strong enrichment in functionalities. The higher number of functionalities covered in HepatoNet stems from several reactions which are inconsistent, but can be used in a functional testing as described above. We also marked the models generated using the GEO dataset for liver, which score similar to equally sized random models. One of the main reasons for the strong correlation between model size and successful tests is the amount of "positive" testing. Many tests are concerned with some type of biosynthesis or degradation and a larger model is more likely to be able to fulfil these requirements than a smaller model. But even using e.g. the biomass function (like GIMME) as part of the input, the models do not get significantly better than a random model on expression data for liver. None of the algorithms tested achieves high scores in the functionality test and several algorithms are on the lower end of the random network reference. A plot showing the tests passed by the different

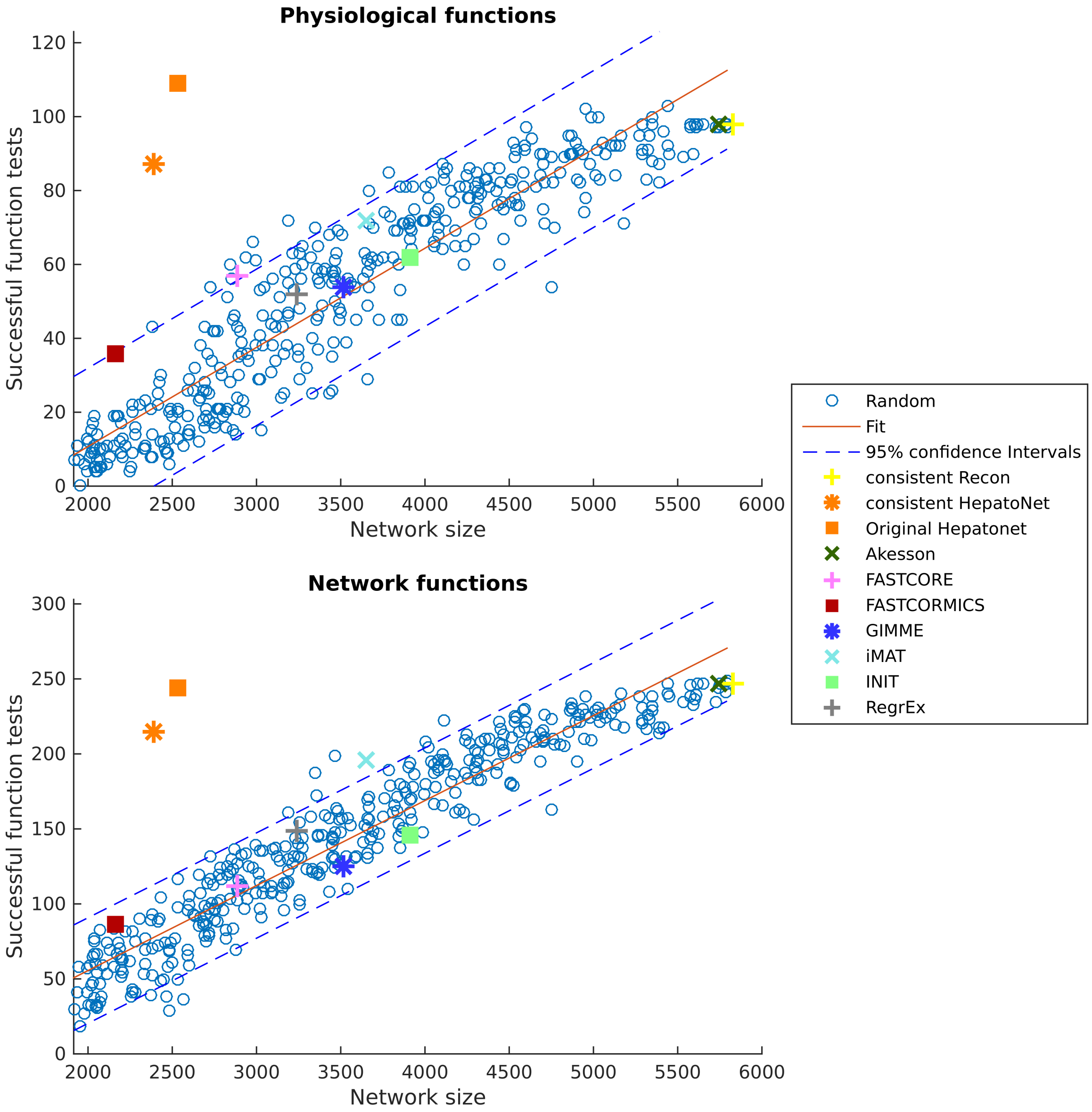


Figure 6.10. Scores in the physiological tests correlate with the size of the network. 260 Random Networks are shown with blue circles.

algorithms is supplied in Supplementary File 7. tINIT could potentially surpass most other algorithms on this test, as it includes functionality information in its reconstruction routine. However, the formulation of tINIT functions is again slightly different from the formulation in HepatoNet and thus not directly compatible.

6.6 Discussion

The primary aim of this work was to review and discuss the existing validation methods and to propose a unified benchmark for the assessment of context-specific reconstruction algorithms. This benchmark will help to identify potential deficiencies of existing and new algorithms and by such increase the quality of context-specific reconstruction algorithms and the models they generate. Although the tested algorithms were validated by their authors in order to be published, the validation methods applied are often incomplete, e.g. a particular aspect of the output model fitting the context of the paper is tested like the ability to produce lactate from glucose in cancer models, leaving other pathways unconsidered. Further, discretization thresholds and other free parameters of the algorithms are likely to be set to optimally fit a particular dataset. Thus, when used in another context the algorithm might perform worse than expected from the original publication. The need of a unified benchmark is nicely illustrated by Figure 6.1 which shows that despite being fed with the same inputs, the output models vary considerably from each other e.g. the output models of RegExp and FASTCORE that share only around 30% of the reactions. Part of the variance between the output models is due to different aims and philosophies of the tested algorithms but also due to algorithm-related bias. The second aim of this work was to demonstrate to the users that the context-specific reconstruction algorithms are not equivalent and that the choice of the algorithm and selection of parameter settings for the algorithms have to be performed with care respecting the philosophy of the tested algorithm. For example, GIMME maximizes a chosen biological function and when using GIMME the user assumes that the metabolism of a cell is aimed at the fulfilment of this function. While this biological function can be assumed to be growth for many microorganisms or cancer cells, it is likely to be more complex for multicellular organisms, where multiple “objectives” have to be balanced. In the same way, FASTCORE takes as input core reactions that are always included in the output model and therefore a higher threshold corresponding to a higher confidence level should be set when using FASTCORE.

Although the parameters were set according to the original papers, we are aware that some of the tested algorithms might perform better with a different parameter setting. We decided nevertheless when possible not to change the original parameter settings of the algorithm. First, because the main objective of this paper is not to assess existing algorithms but to propose a benchmark to validate context-specific algorithms. Second the finding of the optimal parameter setting is a computational demanding processes that would require i.e. cross-validations or other criteria that are not always available. Finding the optimal parameter setting is beyond the scope of a benchmark and rises other questions like overfitting to the data. Third, algorithms should be sufficiently robust to be applied to other datasets with the optimal settings as defined by the authors. As a general principle, in order to avoid overfitting, the parameter estimation should not be performed on the same data than the one used for model generation. We therefore encourage the authors and the users of these algorithms to test them with others parameter settings that might be more appropriate.

The benchmark that we suggest and for which we provide the scripts (http://systemsbiology.uni.lu/software) is based on several criteras:

First of all the algorithms have to produce models of high quality that include genes or reactions that are supported by some evidence to be expressed in the context of interest. This aspect was assessed in the workflow by mapping Barcode z-scored gene information and confidence levels established by the Human Protein Atlas to the models. Context-specific reconstruction that extract sub- networks composed only of active reactions in the context of interest from a general reconstruction tend to produce output models that are enriched for genes with high z-scores and a high confidence level to be expressed at the protein level. Indeed although the activity does not correlate perfectly with expression intensities, it was shown that algorithms that exclude reactions with low expression values show a better predictive power than the generic models from which they were extracted. Both tests show that algorithms that perform a discretization of the input data perform better in these tests than algorithms that maximize the consistency between flux values and the data.

We noticed that within the discretizing algorithms, there are two conceptually distinct approaches when considering unsupported reactions. An inclusive concept which considers unknown data as present and an exclusive concept that considers unknown data as absent. Inclusive concepts tend to produce larger networks and score lower, when comparing the networks to additional data, while exclusive concepts tend to produce smaller networks and score higher. This can be considered as algorithm related bias and it is likely that when multiple algorithms are supplied with the same inputs, reactions that are found by only one or only few algorithms are more likely to be due to algorithm-related bias. Algorithm related bias is not negligible as shown by the huge variability of liver reconstructions with e.g. up to 30% of the reactions being different between the FASTCORE and RegrExp algorithm (Figure 6.1).

Further, algorithms have to be robust to noise but nevertheless be precise enough to capture the variations in the metabolism of a cell in different contexts i.e. different cell types, different states e.g. healthy versus disease and eventually between different patients. These two criteria were tested using both experimental and artificial data. Algorithms like GIMME are performing extremely well in the cross-validation assay but score low in the tissue comparison test, as GIMME produces quite similar reconstructions for the different tissues tested. The algorithms using an inclusive concept tend to be more robust to noisy data but have a reduced resolution power. In contrast, exclusive algorithm are less robust as they tend to only recover reactions that are supported by the input data or reactions that are needed to obtain a consistent model, which allow a greater resolution power. Therefore among the tested algorithms, the FASTCORE family capture best the variation between the different tissues. Further, the confidence level of the reactions included in the 21 tissue models showed that the variability captured by the FASTCORE family models, was not due to noise or algorithm related bias. In the same aspect, the artificial model test gave some interesting insight into the quality of the reconstruction algorithms. While both groups of algorithms, including and excluding, generated about the same model when perfect information was available, they start to diverge at lower amounts of available data. In particular, with less information available the exclusive algorithms underestimate the target network and the including ones overestimate it. While this is to be expected it indicates that the use of two algorithms can give a good approximation of the quality of the available input data and completeness of the reconstruction. If both types of algorithms (inclusive and exclusive) do diverge substantially, it is likely that a relevant amount of input information is missing and that the “true” model is somewhere in between. Similarly, if the models are almost identical, it is likely that the input information and the reconstruction quality is high. GIMME will always include the objective function and all reactions necessary for this function to carry flux. Therefore, those reactions might influence the network size considerably. One advantage of an exclusive concept in this respect, is that it’s variability is less target model dependent than an inclusive approach. For smaller models, the FDR for inclusive models tends to rise much more rapidly with a more incomplete input data set than for larger models. As we commonly are unaware of the actual size of the target network, this might cause problems when using inclusive approaches.

Another important aspect is the computational demand. To determine the processing time we decided when possible not to change the solver used in the original paper as we noticed that algorithms like e.g. RegrEX are sensitive to the used solver, with gurobi finding an initial solution guess faster than e.g. cplex and thus the result returned by cplex being unusable for the algorithm. The range of computational times is however substantial, with fast algorithms running in seconds to minutes and others taking hours or even days. One of the greatest advantages of faster algorithms, is their capability to be more thoroughly evaluated using cross-validation techniques, which is infeasible for an algorithm running several days. We also observed an issue when running the INIT algorithm. For unknown reasons, the algorithm consistently stopped after 10 hours of computation. In particular, the resulting models were odd at best, as they should be close to the models generated by FASTCORE, and in the artificial test, should be optimal on optimal inputs. However, the artificial test was far from optimal, and we assume that the solver does terminate computation at some point.

Finally, we also assessed the capacity of the context-specific reconstruction to pass the functional test as established in (Gille et al., 2010). We found that no algorithm outperforms random models, but that a fitted model can indeed show higher scores without adding more reactions, as seen in Figure 6.10. Unfortunately, obtaining functional data is a very time consuming process and necessitates intensive literature research every time a new tissue model is created. The failure of the tested algorithms in the functional test is mainly due to the high number of non-gene associated reactions in the generic input model (one third of Recon2) and due to the reactions associated to genes with low expression levels. The tested algorithms extract a sub-network from the input model that includes all or most reactions associated with high expressions levels (core) and few reactions with low expression levels (non-core) in order to obtain a consistent model. In function of the chosen non-core reactions, the core reactions will be connected in a different way and the model will display different functionalities. As the choice of the non-core reactions is to a large extent not guided by the data, the obtained functions are random as shown by the functionality test. Interestingly, the reactions found in HepatoNet do have weak evidence when compared to HPA or z-scores, which partially provides another explanation for the inability of the tested algorithms to recover these activities. This however indicates that the general reconstruction currently used lacks either the correct gene-protein-reaction associations for several reactions necessary for the functionalities in liver, that there are alternative pathways missing in the reconstruction and the reactions used in HepatoNet are not the ”true” reactions, that the functions are incorrectly assumed to be available in liver or that the functionality lacks information about the consumed cofactors. Indeed, as all the exchange reactions are closed, some reactions might not carry a flux as the associated cofactor cannot be regenerated. This would also explain why bigger models accumulate more functions. The larger the models, the higher the likelihood of internal loops that could allow a regeneration of cofactors. Further it might also indicate that transcriptomics alone might not be sufficient to build functionally correct models. Information on the uptake and excreted metabolite added to the input reactions set would probably increase the score of most algorithms. We did nevertheless not include this type of information in the input data as the latter is not available for *in vivo* tissues. While presence of importers and exporters does not influence the functional tests, they are however highly influenced by the availability of internal transporters.

Assuming that the defined functions are indeed present in liver, this would indicate the importance of algorithms like tINIT which do take these functionalities into account and which could, given the right reference network, indicate potential missing links in the current reconstructions. tINIT is nevertheless not able to capture metabolic differences between different tissue as shown in (Uhlén et al., 2015), calling for a new generation of algorithms that capture metabolic variation and that are able to take as input functionalities. Note here that algorithms like PRIME that do not extract a subnetwork to obtain a context-specific model, but modifies the bounds of the reactions of the input model, will have regardless of the modelled cell-type or context the same functionalities as the input model. Therefore PRIME would score as high as the generic Recon2 in a qualitative test. Nevertheless, the approach used by PRIME is extremely dependant on the accuracy of the growth measurement and biomass formulation, leading to a very variable quality of the flux prediction (Yikzah et al, 2014). In a quantitative test aiming to predict the production rate of lactate by cancer cells, PRIME showed a lower correlation to the experimental data than FASTCORMICS (Pires Pacheco et al., 2015a). This suggests that building context-specific algorithms with the discretization-based algorithms and then constraining the uptakes rates of several key amino-acids and glucose as performed in (Pires Pacheco et al., 2015a) seems to be favourable. Further, as discussed in the main text, there is no unique function to which the metabolism of a non-cancerous pluricellular cell could be reduced and sofar is limited to handle one metabolic function.

In general, we would recommend to assess the quality of an algorithm based on a combination of functional tests for a reconstructed tissue always in comparison to random networks, confirmation using an independent source of information (e.g. proteomics data, when only using expression data for the reconstruction), and an assessment of algorithmic properties, like dependence on target or input model size and dependence on input data quality. For the latter we would suggest using artificial networks to provide a complete knowledge on the expected outcome.

Discussion

The advent of high-throughput technologies like RNA-seq, microarrays or protein-arrays allowed the generation of relatively cheap, systems-wide data for different cell types, diseases or patients. The huge amount of noisy data, that are several magnitudes larger than everything generated before, called for new concepts and integration methods that allow identifying key underlying principles, molecular signatures and biomarkers. Systems biology that focusses on the creation of computational analysis and modelling tools to extract knowledge from this large datasets, became suddenly very popular. Consequentially, the two decades following the advent of high-throughput technologies were marked by an explosion of developed integration and modelling approaches. One of the most promising fields, was and still is, constraint-based modelling (CBM) because it requires only relatively little knowledge about the system, namely the structure of the network and the stoichiometric coefficients, and therefore metabolic models based on CBM, unlike other modelling approaches such as kinetics models can scale to the size of a cell.

CBM was originally developed for the modelling of unicellular organism that have a simple metabolism that aims mostly to grant growth and homoeostasis. Human and in general multicellular organisms are characterized by hundreds of cell types with very different morphologies, metabolisms and objectives to fulfil. For example, hepatocytes are responsible for detoxification of exo- and endogenous substances, gluconeogenesis, bile salt, cholesterol and phospholipids synthesis whereas neurons are specialised in the transmission and processing of information in the form of electric signals. In order to account for the metabolic variability between different contexts and cell types, omics data was to extract from a generic reconstruction a subnetwork, composed only of active reactions in the given context (Becker and Palsson, 2008; Zur et al., 2010; Jensen and Papin, 2011; Colijn et al., 2009; Lee et al., 2012; Kim et al., 2012a; Navid and Almaas, 2012; Jerby et al., 2010; Åkesson et al., 2004). But as no clear standards existed for the integration of omics data in metabolic models, heuristic parameters in the implementation of algorithms were often used like such as the setting of the expression and unexpression thresholds to the top 25 upper and lower percentile of the intensity distribution (Zur et al., 2010). The setting of heuristic parameters allowed obtaining good predictions for a specific study as they were tailored for this particular dataset but nothing guarantees that this same parameter setting was not completely off for others. Further, the high computational demands of some of the context-specific algorithms prevented their use in a more high-throughput manner or for cross-validations assays. Moreover, as the building of a metabolic modelling is an iterative process consisting of building of a draft network, testing model predictions against some reference dataset and modifying the inputs or parameters to improve the accuracy of the predictions, high computational demands can become a serious handicap for manual curation purposes. To be published, the tools were of course submitted to a validation process but the latter was often designed to test the predictions made on a given dataset in the frame of a specific study, rather than testing the algorithm itself. Therefore the validation process did not allow concluding on the ability of context-specific algorithms to build accurate models for other datasets. And even when algorithms were benchmarked against their predecessors, usually the most ancient algorithms were selected (Wang et al., 2012; Yizhak et al., 2014a; Schultz and Qutub, 2016).

Now that we passed the frenetic early times of high-throughput technologies, more and more groups are getting concerned about the consolidation of the acquired knowledge and the validation of the tools. Among others, the fact that due to the lack of a unified representation such as different metabolite names, the building of a reconstruction from scratch was often easier than the extension of existing models, led to different initiatives like BIGG (Schellenberger et al., 2010), MetRxns (Kumar et al., 2012), MetaNetX (Ganter et al., 2013) or Model Seed (Henry et al., 2010). These initiatives serve among others as repositories for metabolic models and for some of them facilitate the comparison between metabolic models and databases like KEGG, Brenda or Reactome. The aim is to eventually impose a unified nomenclature for a same reaction or metabolite across different models. The need for these initiatives was demonstrated by MetRxns that showed that only 3 reactions could be found in common across 34 models, comprising 21 bacterial, 10 eukaryotic and three archaeal organisms, although most metabolic pathways are believed to be conserved (Kumar et al., 2012). The use of a unified nomenclature would allow to pool down the information from multiple sources, diminish the number of duplicated reactions that prevent the identification of essential genes, facilitate gap filling (Kumar et al., 2012) and more generally allow maximizing the efforts for the building of more accurate models instead of multiplying the models that can only to some extend be compared, updated or corrected.

Concerning context-specific algorithms, the lack of standardized validation procedures has led to a multiplication of algorithms. During the last ten years, dozens of algorithms were published without comparing their prediction in a standardize workflow against others algorithms. Therefore no unbiased prove of the superiority of most algorithms over their predecessors exists. Further, the lack of validation methods did not allow identifying serious limitations of most context-specific algorithms that partially are due to widespread wrong assumptions such as the idea that the use of omics data alone allows the building of functional models. Currently, due to a lack of an unbiased validation framework, the experience that could be collected from previous attempts is only partially taken into account, a new algorithm or model is built from scratch and published, addressing at most only partially the issues of its predecessors and sometimes by creating new ones.

This self-evaluation trend is further justified by the fact that systems biology and more specifically constraint-based modelling is facing a new challenge, namely its application to medicine. The use of constraint-based modelling for systems medicine calls for highly accurate tools with a very good prediction power. False negative and even false positives in medicine can have a very dramatic impact and therefore metabolic models, if used in this context, should be highly reliable. Another important factor is the affordability of the tools, if a tool requires an expert usage, fine tuning or has too high computational demands the latter might be too expensive to be commonly applied for most patients. Further an algorithm that is too difficult to use, increases the risk of a bad parameter setting and by extension a wrong diagnostic.

7.1 The required features of good context-specific reconstruction algorithms

One of the most promising metabolic modelling application for the next decade is the use of the tools of this field for personalized medicine. But, in order to be used by clinicians for diagnostics purposes, algorithms should share at least the following properties:

- low computational demand
- limited number of free parameters
- robustness
- high resolution power
- accuracy of the predictions

These properties will be discussed in more details in the following paragraphs.

7.1.1 Low computational demand

Low computational demands is a very useful feature for context-specific algorithms as it opens a wide range of new possibilities that allow increasing the quality and the predictability of metabolic models by leave-out cross-validations, parameter tuning and gap-filling. In regard of computational demands, (Real) Linear Programming that is solvable in polynomial time ($n^constant$, where n is the number of variables), has typical running times of seconds to few minutes and therefore should be favoured over Mixed Integer Programming (MILP), Quadratic Programming (QP) or Integer programming (IP) that are more often non-polynomial (2^n) (Smith).

7.1.1.1 Cross-validation

Cross-validations allow, among others, assigning a confidence level to the reactions of the model. For example in the case of FASTCORE, if a core reaction that was left-out from the core set, is nevertheless included in the model then this reaction is required to allow at least another core reaction to carry a non-zero flux. In this case, the left-out core reaction (hard core) has a higher confidence score than core reactions that are only supported by the expression level of its own associated genes. The same is true for non-expressed reactions. If the leaving out of a reaction from the non-expressed set does not allow its inclusion in the model, then either this reaction is not required as other alternative pathways ensure the consistency of the core reactions set or the existence of at least another inactive reaction causes the tested reaction to remain inactive (hard non-expressed reactions). In the same way, non-core reactions that are always included in the model, regardless of the left-out core reaction are supported by at least two core reactions and therefore this non-core reactions should be included in the hard core reactions set. Finally, non-core reactions that are never included in the model should be part of the hard non-expressed set. Obviously manual curation should then focus on the reactions that are neither included in the hard core nor in the hard non-expressed set.

7.1.1.2 Parameter tuning

Fast algorithms allow parameter tuning, which should be performed for each sensitive free parameter of an algorithm. And as the sensitivity of a parameter might depend on the input model and the used dataset, it is advisable to run a sensitivity analysis before performing the reconstruction. Sensitivity analysis allows identifying the parameters that when varied affect strongly the output model from the ones that have only a minimal impact. As the prediction power of metabolic models are dependent on the sensitive parameters, curation and identification efforts should be focussed on the sensitive parameters.

7.1.1.3 Gap-filling

Most human generic reconstructions contain around 30% of reactions that cannot carry a flux due to the presence of gaps and dead ends. Context-specific algorithms can be used to identify candidate missing reactions, fastening the gap-filling process and by such context-specific algorithms can improve the quality of manually curated models as shown by (Vitkin and Shlomi, 2012; Thiele et al., 2014) that used a slightly modified version of the MBA (Jerby et al., 2010) (MIRGAGE workflow) and the FASTCORE algorithms (Vlassis et al., 2014) (fastgapfill workflow), respectively, as gap-filling algorithms. Algorithms that were traditionally used for gap-filling purpose like SMILEY (Reed et al., 2006), Gapfill (Kumar et al., 2007), BNICE (Hatzimanikatis et al., 2005) have too high computational demand to be scalable to large models, especially if the models are compartmentalized. Consistent reconstruction algorithms do not only allow detecting but also filling gaps using reactions from a pool of potential candidate reactions retrieved from various databases. Further, the use of omics data allows reducing the number of potential candidates to the ones expressed in the context of interest.

7.1.2 Low number of free parameters

Context-specific algorithms should have a reduced number of free parameters. Having a fast algorithm is futile if the latter is obtained by multiplying the number of free parameters that call for a time-consuming parameter tuning process, which can become rapidly practically impossible for an increased number of free parameters. Arbitrary parameter setting should be avoided whereas heuristic parameter setting should be based on a solid statistical justification or on several bibliographical evidences. As shown previously, the obtained model and its associated proprieties like the number of essential genes are extremely dependant on the setting of the expression threshold. For microarrays, approaches based on previous knowledge on the intensity profile of non-expressed genes like Barcode (Zilliox and Irizarry, 2007; McCall et al., 2011) should be preferred over other arbitrary settings like the choice of the upper 25 percentile (Machado and Herrgård, 2014; Estévez and Nikoloski, 2015) or arbitrary chosen intensity values (Folger et al., 2011).

7.1.3 Robustness and resolution power

The third and fourth criteria are not dissociable as being the two facets of the same problem related to the differentiation between noise and signal. An algorithm with a perfect robustness is as undesired as an algorithm that captures and models each single variation. The first because it always builds the same model regardless of the input data and the latter because it over-fits the data and models noise. The perfect equilibrium between robustness and resolution depends on the purpose of the reconstruction. For the building of a patient-specific model, the equilibrium should be shifted towards a higher resolution whereas for the building of more generic models, a higher robustness might be preferable.

7.1.4 Accuracy of the model predictions

The accuracy of the model predictions is the most important criteria but also the most difficult to assess. As shown by the benchmarking procedures paper (Pires Pacheco et al., 2015b) presented in this thesis, although most algorithms were tested against other competing algorithms in order to get published, the validation methods used in these studies were often biased or focused on a given feature interesting for the particular object of the study. Two recently published papers (Machado and Herrgård, 2014) and (Pires Pacheco et al., 2015b) (ourselves) proposed standardized benchmarking procedures approaches. The first study tested the flux prediction of GIMME (Becker and Palsson, 2008), iMAT (Zur et al., 2010), MADE (Jensen and Papin, 2011), E-Flux (Colijn et al., 2009), Lee-12 (Lee et al., 2012), RELATCH (Kim et al., 2012a), pFBA and GX-FBA (Mahadevan and Schilling, 2003) in *Saccharomyces cerevisiae* and *Escherichia coli* against the experimentally determined fluxes through 13C labelling experiments. Whereas, we assessed the quality of model by determining the confidence of the reactions included in the context-specific model at the proteomic and transcriptomic level and by cross-validation assays using artificial and real data. Further, we tested the ability of liver models built by different algorithms to fulfil a set of biological functions established by (Gille et al., 2010), often defined as the capacity of the model to convert a metabolite A into a metabolite B after a chain of reactions. Although, the main aim of the human benchmark (Pires Pacheco et al., 2015b) was less to establish a ultimate validation procedure for context-specific reconstruction algorithms than to attract the attention of the community on the lack of valid and unbiased validation tests, the failure of most algorithms to capture metabolic variations between different tissues and the incapacity to fulfil known biological functions is evident. The publishing of the two benchmarking procedures should initiate a discussion in the community about validation methods commonly used to benchmark algorithms and their respective caveats.

7.1.4.1 Testing of biological functions

An example of the caveats of existing validations methods is the testing of the so-called biological functions. In order to validate metabolic models the capacity of producing a metabolite B from a metabolite A or from a well-defined medium is tested *in silico*. The benchmarking study (Pires Pacheco et al., 2015b), that we proposed, clearly showed that none of the tested algorithms performed better than random models of the same size. In other words the number and the functions that a model can fulfil is random and depends only on the size of the output model. The reason of this failure is partially due to the high fraction of reactions that are not associated to genes (30% of the reactions of ReconX (Thiele et al., 2013) models) and in general to non-core reactions (reactions associated to genes with low expression levels). To connect core reactions every algorithm has to include reactions with no (as not associated to any genes) or with weak evidence (non-core reactions). The issue is that several equivalent possibilities exist to connect core reactions via non-core reactions, the inclusion of one or the other non-core reaction being often more or less equivalent. The choice is more or less random. In function of the non-core reaction included, the output model can fulfil different functions. Further, a qualitative testing of function (is the model able to fulfil a specific model) without testing if the predicted fluxes are in the same range than the measured one might not be sufficient as it does not assigning if the predicted flux rate corresponds to the ones observed experimentally. Further a qualitative function testing favour large models and algorithms like PRIME (Yizhak et al., 2014a) that do not extract a sub-network but modifies the bounds of the input model.

PRIME models would by definition pass the same number of tests than the generic model from which they are derived. Whereas in a quantitative function test where the predicted flux rates are compared to measures ones the accuracy of predicted flux rates by PRIME cancer models are very versatile and correlated less well ($R^2 = 0.3$) with the experimentally observed secretion rates than the ones predicted by FASTCORMICS cancer models for which inputs of several key metabolites were constrained in function of the experimental data (Pires Pacheco et al., 2015a). The good correlation score of $R^2 = 0.7$ for the FASTCORMICS models suggests that when the correct non-core reactions necessary for biological functions are present in the model, the prediction of the flux rate is quite reasonable. A possible explanation for the bad scoring of PRIME is that as the bounds of non expressed reactions are only constrained and the reactions not excluded from the output model, the flux carried by reactions associated to unexpressed genes might contribute to the predicted lactate secretion rate or alternatively deviate a fraction of flux to inactive pathways.

Further, the set of functions that a model of cell-specific cell has to fulfil is at the moment unclear. Recon1 (Duarte et al., 2007), Recon2 (Thiele et al., 2013) models and HMR 1.0 and HMR 2.0 established around 300 biological tasks or functions, of which 56 are believed to be ubiquitously required across tissues (Uhlén et al., 2015). But beyond these 56 tasks, others more cell type-specific tasks are probably required. The identification of these tasks for a specific cell type would require a very time consuming literature search. And this might still be insufficient as the expression of genes in a given cell type varies in function of the external stimuli. The number of biological function that need to be fulfilled by a given cell might vary as well.

7.2 Critical assessment of the FASTCORE family

7.2.1 FASTCORE: The vanilla version

FASTCORE (Vlassis et al., 2014) is the vanilla version of a family of context-specific building algorithms, from which workflows like FASTCORMICS for the building of context-specific models via the integration of microarray data and fastgapfill (Thiele et al., 2014), as indicated by the name, for gap-filling purposes,were derived to cope with the specificity of different type of inputs data or different tasks. FASTCORE is devoid of free parameters (with the exception of epsilon, the flux activation threshold) and has very low computational demand with running times around one second, which turns FASTCORE into a perfect candidate for being used in a high-throughput framework. FASTCORE takes as input, core reactions and an input generic model like Recon2 (Thiele et al., 2013) models or HMR 2.0, then extracts a subnetwork that contains only reactions that are active in the context of interest. Further, FASTCORE uses as input binary data. The reactions are shown to be or not to be expressed in the context of interest. Therefore, FASTCORE is also suitable for the integration of proteomics or bibliomic data. But, as FASTCORE forces every core reaction to be included in the output model, FASTCORE is very sensitive to the input core set. A reaction wrongly tagged as a core reaction might cause the inclusion of entire inactive pathways. This is problematic as some genes control hundreds to thousands of reactions such as genes controlling transport reactions. One or few of the potential target reactions is probably really active in the context of interest whereas the others are inactive as the genes controlling the remaining reactions in the pathways are unexpressed. FASTCORE, in this case will wrongly, considers all target reactions and the required non-core as active.

For similar reasons, when used for input data subjected to large noise effects that prevent a clear segregation in expressed or unexpressed genes, a stringent threshold must be set or an additional discretization must be run before feeding the data to FASTCORE.

Moreover, the concept of compactness is a central assumption of the FASTCORE philosophy. The search of compact models is justified by the fact that bigger models are associated by a increased consumption of resources associated to the synthesis of the required enzymes and transporters. But the shortest path is not always the correct one as showed by the functional test. Metabolic Functions define which metabolites can be produced from some precursors and by such how core reactions are interconnected. The failing of FASTCORE and the other tested algorithms in the functional testing shows that the way core reactions were connected, (the shortest path), was not correct for these examples.

Furthermore, FASTCORE uses an approximation of cardinality function to allow a LP formulation of optimization problem. This approximation prevents FASTCORE to always construct the most compact model containing all core reactions. FASTCORE, nevertheless finds a good approximation of the optimal solution and the accuracy of FASTCORE was shown to increase with the size of the core set (Vlassis et al., 2014).

Additionally, the L1-normalization minimizes the flux through non-core reactions therefore a solution with a reaction R that carries a flux F, is equivalent to a solution with two reactions that carry each half of the flux F, further explaining why FASTCORE in some examples fails to build the most compact models.

Moreover, although FASTCORE adds all unblocked irreversible reactions and most irreversible reactions after the two first LPs (LP7 and LP10), branches only composed of reversible reactions are not included after this step. To address this problem, the sign of the remaining reversible reactions in the matrix is first flipped and then all remaining individual reversible are tested. One could argue that the order of the testing in the singleton might affect the output model. For the consistent Recon2, a maximum of 158 reactions are not included in the two first LPs. 52 of these reactions are exchange reactions and 46 are transporter reactions (mainly extracellular transport reactions). The remaining reactions are alternative branches composed mostly of one single reaction. So in general as these reversible branches are only composed of one or two reactions (for the case of exchange reaction and extracellular transporters), the order has no or a negligible impact on the output network, as 158 reactions are mostly not connected.

Nevertheless, the solution of FASTCORE and competing algorithms might not be unique due to alternative optimas. By default most solvers only display one solution and the displayed solution might vary in function on the version of cplex or the computer used.

7.2.2 FASTCORMICS: A variant of FASTCORE for the building of metabolic models via the integration of transcriptomic data.

FASTCORMICS uses BARCODE (Zilliox and Irizarry, 2007; McCall et al., 2011), a preprocessing tool that takes into consideration previous knowledge about the intensity distribution of genes in an unexpressed state across thousands of arrays and conditions to set an expression threshold corresponding to a z-score of 5. A threshold of 5 standard deviations is very stringent, which suits the specificity of the FASTCORE family that requires core reactions with a high confidence level. The use of thresholds is critical for output models as in function of the threshold setting whole pathways might be included or excluded of the model, which alters the functionalities of the model. Nevertheless, the Benchmarking paper showed that discretization based algorithms performed better than those that used continuous data.

Beside of the core set, FASTCORMICS sets an unexpressed set of reactions that allow excluding inactive reactions from the model and a set of reaction that is not penalized. These additional sets reduce the impact of reaction wrongly tagged as core reactions on the output model; a core reaction that requires the activation of reactions associated to unexpressed genes, is not included in the model. Further we recommend to remove reactions under the control of promiscuous genes from the core set and to place them in the not penalized set so that their inclusion is not forced but preferred over non-core reactions. Further, the MBA (Jerby et al., 2010) and CORDA algorithms (Schultz and Qutub, 2016) allow additionally the setting of core set with medium or low confidence, we decided not to multiply the categories of input reactions as it implies every time the heuristic or arbitrary threshold that have a non negligible impact on the output model.

Beside the above-mentioned modifications, FASTCORMICS (Pires Pacheco et al., 2015a) shares most of specifications of FASTCORE, like low computations times (few minutes due to the preprocessing step) and although FASTCORMICS requires the setting of 2 expression thresholds (expression and unexpression threshold). This thresholds were set based on solid statistical evidences. The FASTCORMICS and FASTCORE were shown to being able to capture variations between different cell types (Pires Pacheco et al., 2015b). Further, the FASTCORMICS cancer generic model allowed identifying a set of essential genes that were coherent with a published wet lab experiment performed on 12 cancer cell lines using shRNA screens. Moreover, the predicted lactate secretion rate by the NCI-60 cancer cell models showed a good correlation with the experimental data after constraining the inputs reactions of several key amino acids and glucose according to the experimental data. Taken together these two studies showed that the FASTCORMICS workflow allowed making prediction that are in conformity with the experimental data. To our best knowledge and to this date, Barcode is the best discretization tool for microarray data. But at the moment, Barcode is available uniquely for human and mouse microarrays and only for a restrained number of platforms. Therefore, we are working on others discretization procedures that could be also used for the integration of RNA-seq data. Further, a less conservative discretization method or simply a more relax thresholds setting might be necessary to differentiate the metabolism of different individuals.

7.3 Lessons from the benchmarking procedures

We initiated this study because we realised that the validation methods commonly used in publications were biased or had serious caveats like on the manner on how biological tasks are tested. We also wanted to illustrate the factthat unlike what was/is commonly believed, context-specific algorithms are far from being equivalent and that every algorithm is based on some assumptions that affect the output model and that users should choose the dataset and set the parameters according to the requirements of the selected algorithm (Pires Pacheco et al., 2015b). The benchmarking procedure paper (Pires Pacheco et al., 2015b) showed that the different algorithms although having been fed with the same inputs, reconstructed very different output models that only share for the more distinct ones around 30% of the reactions. Further, it demonstrated that discretization-based methods were performing better than methods that used continuous data as input. Unlike the discretization-based algorithms, the method using continuous did not favour the inclusion of highly expressed reactions over reactions associated with lower expressed genes. Using continuous data might a *priori* seem reasonable as it allows avoiding the setting of heuristic hard thresholds that would affect differentially reactions just below or above these thresholds. But using continuous data, requires a function that defines how the data is integrated, such as least squares error minimization. The problem is that there is an infinite number of possibilities how to tackle this problem and in regard of the function chosen, the output model will be different. Further some approaches as the least-square approach, might cause an over-fitting of the data, which is problematic if the algorithms are used on microarray data, a technique particularly prone to noise. In order to reduce the complexity a regularization approach can be applied like in (Estévez and Nikoloski, 2015). But regularization requires very time consuming cross-validations approaches.

Further, the Benchmark paper showed that the tested algorithms with the exception of the FASTCORE family could not capture metabolic variability between different tissues. Although when proteomic data was mapped on Recon2 only very few genes were shown to be ubiquitously expressed in all tissues. It also showed that the integration of omics data is not sufficient to guarantee that a model is able to fulfil a given metabolic task, defined as a metabolite Aconverted after a chain of reactions into a metabolite B.

The same was demonstrated by the benchmark procedures proposed by (Machado and Herrgård, 2014) that compared the flux distribution in *Escherichia coli* and *Saccharomyces cerevisiae* made by flux prediction algorithms (such as GIMME (Becker and Palsson, 2008), iMAT (Zur et al., 2010), E-flux(Colijn et al., 2009), MADE (Jensen and Papin, 2011), Lee-12 (Lee et al., 2012), RELATCH (Kim et al., 2012b), GX-FBA (Mahadevan and Schilling, 2003) and pFBA of the Cobra toolbox that are for some of them context-specific reconstruction) to carbon 13 (C13) labelling experiments.Unexpectedly, this study showed that none of the tested algorithms performed better than Flux Balance Analysis (FBA) and that in general omics data did not allow increasing prediction accuracy of the tested algorithms in *Escherichia coli* and *Saccharomyces* *cerevisiae.*

Another lesson from these studies is the lack of published flux rates using such as labelling experiments that cover other pathways than the central metabolism. The number of published data sets for *Escherichia coli* and *Saccharomyces cerevisiae* is around 5 (Machado and Hergård, 2014). For pluricellular organism this type of data seems even rarer, which is the reason why quantitative function tests of the predicted flux rate, are not often performed in mulitcellular organisms. The consumption and release (CORE) (Jain et al., 2012) dataset used in the (Pires Pacheco et al., 2015a), contains flux rate mainly of the central carbon systems of cancer cells. In other words it did not cover the entire metabolism. Furthermore, this type of experiments is very time consuming as besides of requiring the building of more than twenty cancer models for each algorithm, random sampling, which has high computational demand, is usually performed to estimate the range of predicted fluxes.

7.4 Genes under high regulatory load and the metabolic network

The FASTCORMICS workflow captures metabolic variation between different cell-types with a cell-specific activation of numerous branches of the metabolism pathways. Independently of the FASTCORMICS study, genes under high regulatory load were shown to have a gene specific expression and to play role in the cellular identity (Hnisz et al., 2013; Whyte et al., 2013), raising the question about the epigenetic regulation of metabolism in different cell types. Although genes under high regulatory load and more largely super-enhancers (large cluster of enhancers) are at the momentextensively studied (Creyghton et al., 2010; Rada-Iglesias et al., 2011; Heintzman et al., 2009), the regulation of the metabolism by high-regulatory load was not previously addressed. The mapping genes under high regulatory load on a macrophage metabolic network showed that high-regulatory load genes tend to rank among the highest expression within each pathway (Pires Pacheco et al., 2015a). Furthermore, the expression of metabolic genes under high regulatory load in macrophages tend to have a higher expression in macrophages than in other tissues, validating the fact that genes under high-regulatory load have cell type-specific expression profiles. Transporter and entry points reactions are enriched for genes under high-regulatory load suggesting that the latter control the activation of alternative pathways by enhancing the expression of the genes that control the first step of a pathways or a subpathway. These findings are consistent metabolic control analysis for linear pathways that showed that the control applied in the first steps (especially the first one) is stronger than in the last steps (Klipp et al., 2008).

For these study like in most study on enhancers and super-enhancers were associated with the nearest genes, which is problematic as chromosome conformation capture carbon copy (5C2)experiments and capture HI-C showed that only around 60% of the enhancer actually regulate the nearest gene. But at the moment no bioinformatic tool exist that allow identifying-targets of enhancers, requiring manual checking of candidate genes. We also defined as genes under high-regulated, thisheuristic threshold was set based on a previous study that the 10th percentile was significantly enriched for disease (Galhardo et al., 2014).

7.5 Conclusion and Outlook

The FASTCORE family outperforms its competitors in speed and in the capacity two distinguish between different tissues. The accuracy of FASTCORE family models can be further improved by leave cross-validations assays followed by manual curation. The FASTCORE family like all competitors algorithms with exception of tINIT cannot guarantee to build functional models. Whereas tINIT has high computational demands and fail in capturing variation between tissues. The outlook for the next years will be to create a version of FASTCORE that guarantees the building of functional models and the release of a version of FASTCORMICS with different discretization steps, so that the workflow could handle any type of omics data regardless of the platform used. A less stringent discrization step will also allow to capture variations between different individuals, which at the moment no algorithms manages to capture.

Article manuscripts

This section contains the remaining articles published in peer-reviewed manuscripts authored by the PhD candidate.

FASTCORMICS: Application 2

*The following chapter was integrally published as **The neural stem cell fate determinant TRIM32 regulates complex behavioural traits** in january 2015 in Frontiers cellular neuroscience.*

A.1 Summary and contributions

Contents

The FASTCORMICS workflow was used in the frame of the article: **"The neural stem cell fate determinant TRIM32 regulates complex behavioural traits"** published in 2015 in Frontiers in cellular neuroscience. Dr. Anna-Lena Hillje and Dr.Elisabeth Beckmann share the first authorship of this paper. For this project, we built together with Prof. Thomas Sauter, two metabolic models of murine brain cells, a wild type and a TRIM32 mutant model, using the FASTCORMICS workflow using expression data previously published by the team of Prof. Jens Schwamborn and the iSS1393 generic mouse model Heinken et al. (2013) as input. The murine brain wild type and the TRIM32 mutant model contain respectively 735 and 635 reactions, with 516 reactions being shared between the two models. The metabolic models were used to explain the metabolomics data obtained by GC/MS and analysed by Dr. Christian Jaeger. Of the two hundred and ten detected metabolites, five metabolites had concentration levels significantly different between the two genotypes $(p < 0.05)$. Random sampling, that uses a Monte Carlo based approach, was performed in order to explore the space of possible solutions and by such to obtain a qualitative estimation of the flux distribution allowed by the network topology and the constraints imposed to the model. As we were focusing on metabolites related to the glycolysis pathway, the flux through the latter was maximized. The Random sampling suggests that in the mutant model the flux through the branch leading from the glycolysis pathway through serine biosynthesis pathway is decreased, causing a drop in the consumption of 3-phosphoglyceric acid by the phosphoglycerate dehydrogenase, the first enzyme in the serine biosynthesis pathway. This decrease could explain accumulation the 3-phosphoglyceric acid observed in the mutant metabolite measurements. For this paper, my contribution and the one of Prof Thomas Sauter is related to the production and the analysis of Figure 6 of the paper and more precisely the building of the metabolic models and the random sampling experiments.

A.2 Abstract

In mammals, new neurons are generated throughout the entire lifespan in two restricted areas of the brain, the dentate gyrus (DG) of the hippocampus and the subventricular zone (SVZ)-olfactory bulb (OB) system. In both regions newborn neurons display unique properties that clearly distinguish them from mature neurons. Enhanced excitability and increased synaptic plasticity enables them to add specific properties to information processing by modulating the existing local circuitry of already established mature neurons. Hippocampal neurogenesis has been suggested to play a role in spatial-navigation learning, spatial memory, and spatial pattern separation. Cumulative evidences implicate that adult-born OB neurons contribute to learning processes and odor memory. We recently demonstrated that the cell fate determinant TRIM32 is upregulated in differentiating neuroblasts of the SVZ-OB system in the adult mouse brain. The absence of TRIM32 leads to increased progenitor cell proliferation and less cell death. Both effects accumulate in an overproduction of adult-generated OB neurons. Here, we present novel data from behavioral studies showing that such an enhancement of OB neurogenesis not necessarily leads to increased olfactory performance but in contrast even results in impaired olfactory capabilities. In addition, we show at the cellular level that TRIM32 protein levels increase during differentiation of neural stem cells (NSCs). At the molecular level, several metabolic intermediates that are connected to glycolysis, glycine, or cysteine metabolism are deregulated in TRIM32 knockout mice brain tissue. These metabolomics pathways are directly or indirectly linked to anxiety or depression like behavior. In summary, our study provides comprehensive data on how the impairment of neurogenesis caused by the loss of the cell fate determinant TRIM32 causes a decrease of olfactory performance as well as a deregulation of metabolomic pathways that are linked to mood disorders.

A.3 Introduction

Adult neurogenesis has been reported in the mammalian brain in two regions, the subventricular zone (SVZ) located in the wall of the lateral ventricles and the dentate gyrus (DG) of the hippocampus (Gage, 2000).

In the SVZ, neural stem cells (NSCs), also called type B cells, are astrocytes that are able to self-renew and at the same time give rise to transit amplifying cells (type C cells) (Doetsch et al., 1999). These transient amplifying cells differentiate into neuroblasts (type A cells) that migrate along the rostral migratory stream (RMS) to the olfactory bulb (OB) via chain migration (Doetsch et al., 1997). In the OB they finally become mature neurons which are integrated into the neuronal network (Belluzzi et al., 2003; Carleton et al., 2003). Adult newborn neurons turn into OB interneurons; i.e., granule cells and periglomerular cells (Petreanu and Alvarez-Buylla, 2002). Granule cells of the OB shape the information passed from the projecting cells of the bulb (mitral and tufted cells) on to higher brain areas (Nissant and Pallotto, 2011). Thereby, they are able to spatiotemporally shape the mitral cell signal in a process called lateral inhibition. This process is supposed to facilitate odor encoding and discrimination (Yokoi et al., 1995; Urban, 2002; Tan et al., 2010; Ernst et al., 2014).

In the hippocampus, type I NSCs reside the inner layer of the DG, the subgranular zone (reviewed in (Yao et al., 2012a)). They generate self-amplifying type II intermediate progenitor cells, which migrate to the outer layers of the DG (Kuhn et al., 1996). Type II cells eventually give rise to type III neuroblasts that differentiate into glutamatergic dentate granule cells (DGCs) (Kuhn et al., 1996). Adult born DGCs integrate into the existing network and form synaptic connections with the entorhinal cortex and the CA3 subfield (van Praag et al., 2002; Toni et al., 2007; Yao et al., 2012a). This facilitates adult born neurons to contribute to pattern separation, a process that allows the distinct encoding of very similar stimuli (reviewed in (Vivar and Van Praag, 2015). Additionally, hippocampal adult neurogenesis has been shown to be involved in the regulation of cognition and mood (Zhao et al., 2008) as well as learning and memory (Stuchlik, 2014).

Adult born neurons display unique properties that clearly distinguish them from mature neurons (reviewed in (Ming and Song, 2011). After forming synaptic connections, newborn neurons in both, hippocampus and OB, show enhanced excitability as well as increased synaptic plasticity at certain stages of neuronal maturation (Nissant et al., 2009; Ming and Song, 2011). By this, they are able to modulate the existing local circuitry of established mature neurons and add specific properties to information processing (Bardy et al., 2010). Many studies aimed at investigating the behavioral functions of adult neurogenesis. In these studies a variety of different methods were used to alter cell turnover rates in the SVZ (reviewed in (Nissant and Pallotto, 2011). This methodological variation might explain that the results vary considerably with regards to effects on learning and memory. For example, differences in spontaneous odor discrimination, associative learning tasks as well as in short-term and in long-term memory have been reported (Lazarini and Lledo, 2011; Breton-Provencher and Saghatelyan, 2012).

The TRIM-NHL protein family has an evolutionary conserved function in neuronal cell fate specification in C. elegans, Drosophila, and mammals (Betschinger and Knoblich, 2004; Bello et al., 2006; Lee et al., 2006; Schwamborn et al., 2009; Hammell et al., 2009; Hillje et al., 2011)). During embryonic development of the neocortex in mice, the TRIM-NHL protein TRIM32 regulates cell fate decisions of newly born daughter cells (Schwamborn et al., 2009). In the adult brain, TRIM32 is upregulated during differentiation of SVZ generated neuroblasts and is necessary for the correct induction of neuronal differentiation of these cells (Hillje et al., 2013). Loss of TRIM32 results in an overproduction of adult generated OB neurons, which is the combined result of increased progenitor proliferation and decreased apoptosis.

On the molecular level, TRIM32 induces neuronal differentiation and suppresses self-renewal by ubiquitination of the transcription factor c-Myc and the activation of certain microRNAs (Schwamborn et al., 2009; Nicklas et al., 2012). TRIM32 has been linked to several human diseases including limb-girdle muscular dystrophy type 2H (Frosk et al., 2002, 2005; Kudryashova et al., 2005, 2012), Bardet−Biedl syndrome (Chiang et al., 2006), cancer; (Horn et al., 2004; Kano et al., 2008), autism spectrum disorder (Lionel et al., 2011, 2013), depression (Ruan et al., 2014), Alzheimer's disease (Yokota et al., 2006), obsessive compulsive disorder (Lionel et al., 2013), anxiety (Lionel et al., 2013), and attention deficit hyperactivity disorder (Lionel et al., 2011, 2013).

In the here presented study we show that the absence of the cell fate determinant TRIM32 alters the performance of mice in an olfactory habituation task without influencing the long-term olfactory memory. In order to control for inadvertent influence of other behavioral traits, we included a variety of tests investigating DG-related behavior in our study. Finally, we have hints that an impairment of adult neurogenesis caused by loss of TRIM32 results in the deregulation of metabolomic pathways that have been linked to depression and anxiety related behavior. Altogether, our study provides comprehensive data on how the impairment of neurogenesis caused by the loss of the cell fate determinant TRIM32 leads to decreased olfactory performance as well as changes in metabolomic profiles.

A.4 Material and Methods

A.4.1 Animals and Housing Conditions

A total of 14 male TRIM32 knockout mice and 21 male wildtype litter mates were used. All mice were born and tested in the Department of Behavioural Biology, University of Muenster. Their parents were derived from a locally bred colony at the Center for Molecular Biology of Inflammation that was founded from cryo-preserved spermatozoa derived from the Mutant Mouse Regional Research Centers, USA. Gene knockout of TRIM32 has been described earlier (Kudryashova et al., 2009). In detail, T32KO mice were generated using the BGA355 mouse embryonic stem cell line $[BayGenomics(formerweb - sitehttps : //www.mmrrc.org/catalog/sds.php?mmrrc_id = 11810;$ now available at International Gene Trap Consortium,$http : //www.genetrap.org/cgi - bin/annotation.py?cellline = BGA355)]$ carrying gene trap insertion in TRIM32, within exon 2. The position of the integration site was confirmed after nucleotide 278 starting from the ATG codon in exon 2 of the trim32 gene. Original founder mice were 129 SvEvBrd x C57 BL/6 chimeras, which were backcrossed to C57 BL/6J wt mice to obtain germ line transmission. Heterozygotes from this cross were interbred to produce ko and wt homozygotes (Kudryashova et al., 2009). All analyses were performed on interbred mice on a mixed 129 SvEvBrd x C57 BL/6J background. To ascertain a congenic background more than seven backcrosses were done.

All animals were housed in a temperature controlled room at 22 ℃ and a relative humidity of 45% $\pm 10\%$. A 12-h dark-light circle with lights on at 8.00 a.m. was installed. Offspring were weaned at postnatal day 22 and experimental mice were kept in standard cages (37 x 21 x 15 cm) in groups of 35 animals, preferably in groups of littermates. Tissue for genotyping was sampled by ear cuts and genotype specific DNA fragments were identified after PCR amplification and agarose gel electrophoresis. Ear-cuts also allowed individual discrimination of mice from the same cage. However, behavioral experiments were carried out with the experimenter being unaware of the genotypes of the subjects. Food (Altromin 1324, Altromin GmbH, Lage, Germany) and water were available ad libitum. A thin layer of wood shavings and paper towels served as bedding and nesting material that was changed weekly while transferring the mice to clean cages.

All procedures and protocols met the guidelines for animal care and animal experiments in accordance with national and European (86/609/EEC) legislation.

A.4.2 Health Check

To determine whether all mice included in the testing were healthy a health check was performed. In order to prevent interference of the behavioral tasks with testing experience gained during the health check, the test was conducted at postnatal day 101 and thus would have allowed retrospectively excluding animals with visible or detectable bodily defects. However, none of the tested animals had to be excluded. Health parameters were tested according to (Lewejohann et al., 2004) and included general appearance (e.g., fur, ears, eyes, vibrissae, extremities, and tail), sensory abilities (e.g., vision, startle response, tactile reaction), reflex functions (e.g., eyelid reflex, grasp reflex), and locomotor/coordinative abilities (climbing, balancing).
It was previously observed that male TRIM32 knockout mice weighted more than wildtype mice after reaching an age of 8 weeks (Kudryashova et al., 2009). We therefore measured weight development of all mice beginning at an age of 3 weeks until the age of 15 weeks (Supplementary Figure 1). However, under the here applied experimental conditions we were unable to detect a significant increase in weight in aging TRIM32 knockout mice.

A.4.3 Elevated Plus Maze (EPM)

The EPM was conducted with 14 TRIM32 knockout and 20 wildtype male mice at an age of 65 (2) days of age. The test measures anxiety related behavior by exploiting the tendency of mice to avoid exposed areas in favor of shielded areas. The apparatus consists of four 30 x 5 cm arms emerging from a central platform of 5 x 5 cm. Two opposing arms are open while the two orthogonal arms are enclosed by walls 20 cm of height. The apparatus was elevated ca. 50 cm above the ground. The runway of the open arms was surrounded by a low balustrade (0.5 cm), effectively preventing the mice from jumping or falling off. On the ceiling above the apparatus, at a height of 175 cm, a camera (Logitech Pro 9000, Freemont, USA) was installed, as well as a light bulb emitting ca. 100 lux. Videos of the mice performing the test were recorded and subsequently analyzed using an in-house programmed animal tracking software (Lewejohann et al., 2004). Before testing started, the runways and walls of the maze were cleaned with 70% ethanol to remove possible olfactory cues from preceding tested mice. In order to control for a similar level of alertness, the mice were placed in an empty cage for 1 min prior to testing. Each mouse was then placed on the central platform facing one of the closed arms. After 5 min of freely exploring the apparatus, the recording was stopped and the mouse was transported back into its home cage.

The parameters which were analyzed included the total path length in meters, time spent in closed and open alleys in seconds, as well as the number of entries into the open and closed arms.

A.4.4 Open Field Test (OF)

The OF was conducted with 14 TRIM32 knockout and 21 wildtype male mice at an age of 67 (± 2) days. The test evaluates locomotor activity and exploratory behavior in an open box measuring 80 by 80 cm with surrounding walls of 40 cm height. Comparable to the EPM test anxiety related behavior (avoidance of unprotected areas) can be observed by measuring the time spent in the center of the box in relation to the time spent in the more protected areas close to the walls of the box. The box was lit at ca. 100 lux and a camera was placed centrally above the apparatus. The apparatus was cleaned with 70% ethanol before testing and the mice were placed in an empty cage 1 min prior to being placed in the center of the open field. Videos of the mice performing the test were recorded for 5 min and subsequently analyzed (see EPM). The parameters analyzed included path length, time in center, time close to the wall, number of stops, and velocity. Stops were recognized by the tracking software when the velocity was 0 for at least 1 s.

A.4.5 Barnes Maze (BM)

The Barnes Maze was first developed by (Barnes, 1979) to compare spatial learning abilities of young vs. senescent rats. The maze itself consists of a circular platform of 1 m in diameter with 12 holes at the circumference drilled in an equal distance from each other. One of the holes is chosen to be the target hole for the tested individual and connected to the animal's home cage. The platform was brightly lit in order to create a mildly aversive environment that was escapable by learning the position of the correct hole during the course of repeated trials. Around the platform, visual cues were placed in order to facilitate spatial orientation.

During the training phase, one of the holes was connected to the animal's home cage via a wire-mesh tunnel, while all other holes were closed by a short tube made of wire mesh. Choosing the same ventilatable wire mesh for the rewarded as well as for the unrewarded holes guaranteed that it was not possible for the mouse to discriminate between open and closed holes from above. The maze was raised 125 cm above the floor and illuminated by a 100 W electric bulb located 110 cm above the center of the maze.

Two training trials with an inter-trial interval (ITI) of 1 h were conducted on four consecutive days with 14 TRIM32 knockout and 21 wildtype male mice starting at an age of 72 (± 2) days. On the fifth day a probe trial testing spatial memory was conducted with all holes being closed for 5 min. During this trial it was measured whether or not the animal spent significantly more time in the sixth of the BM were formally the escape hole was located. After the probe trial an additional training trial with the correct hole being connected to the home cage again was done in order to reinstall the memory for the correct position of the hole. One week later a single re-test was conducted in order to test for intermediate to long-term spatial memory. Before each trial the BM was cleaned with 70% ethanol and all mice were placed for 1 min in a starting cylinder placed in the middle of the platform. All trials were video recorded and analyzed using an animal tracking software (see EPM). For each trial, the time the mice needed to find the target holes, number of errors, as well as the path length traveled was measured.

A.4.6 Olfactory Habituation Test (OH)

To test olfactory discriminative abilities and short-term memory for different odors an olfactory habituation/dishabituation task was chosen. The odorants used were (A) isoamyl-acetate (SAFC, Hamburg, Germany), (B) 4-methylcyclohexanol (Sigma Aldrich, Steinheim, Germany), and (C) 3-octanol (Sigma Aldrich, Steinheim, Germany). All odorants were diluted 1:100 in paraffin oil (Sigma Aldrich, Steinheim, Germany). These odorants are described as being perceived dissimilar (Mandairon et al., 2006) and in a preliminary experiment, C57BL/6J wild-type mice did not show a preference for any of these odors when being exposed to them in an open field apparatus (data not shown).

The OH was conducted at an age of 80 ± 2) days with 14 TRIM32 knockout and 18 wildtype male mice accordingly to the protocol by (Yang and Crawley, 2009) with slight modifications. For each trial the mouse was placed into a clean cage (Macrolon type I, 19 x 10 x 13 cm) filled with fresh bedding. In the center of the cage lid, a wooden cotton swab (15 cm, PARAM GmbH, Hamburg, Germany) was attached so that its tip reached into the cage for ca. 7 cm. In order to acclimate to the new environment, the mouse was left undisturbed in the cage for 2530 min. The cage was then transferred to a chamber that was directly connected to the air ventilation system of the animal housing facility in order to minimize surrounding odors and also to eliminate any scents previously applied as fast as possible.

Each tested mouse was firstly exposed to distilled water and subsequently to three different odors (the order of odors was randomized for each mouse). Each substance was applied 3 times in a row with an inter trial interval of 30 s. The experimenter observed the behavior for 90 s and measured the time the mouse spent sniffing on the cotton tip (snout closer than 2 cm), the latency of the first sniff, and the number of sniffs. The test was repeated after 10 days in order to measure long-term olfactory memory.

A.4.7 Injection of Bromodeoxyuridine (BrdU)

Animals were injected with 50 mg/kg BrdU on 3 consecutive days and sacrificed after the indicated time. Sections were incubated with 2 M HCl in PBST (PBS+0.3% Triton) for 25 min at 37 ℃ for denaturation of the DNA, neutralized with 0.1 M sodium tetraborate (pH 8.5) for 7 min at room temperature and stained with an anti-BrdU antibody (AbDSerotec). On day 107109 at 12 a.m., the mice received a BrdU injection and at day 123, the animals were sacrificed. The brains were dissected and analyzed (Hillje et al., 2013). For quantification of BrdU+ cells in the DG, two sections each of 5 wt and 5 TRIM32 ko brains were analyzed, for quantification of BrdU+ cells in the OB, two sections of 4 wt and 4 TRIM32 ko mice were used. In each case, the mean of BrdU+ cells in TRIM32 ko tissue was normalized to the mean of BrdU+ cells in sections of wt brains.

A.4.8 Immunohistochemistry of Free-Floating Sections

Mice were deeply anesthetized by intraperitoneal injection of 0.017 ml of 2.5% Avertin (100% stock solution: 10 g 2, 2, 2-Tribromoethanolin 10 ml tert-Amylalcohol) per gram of body weight and sacrificed by perfusion. Brains were fixed overnight at 4 ℃ in 4% paraformaldehyde in phosphate buffer saline (PBS). Later, sections of 40μm were prepared using a vibratome (Leica, Wetzlar, Germany) and blocked for at least 1 h in TBS (0.1 M Tris,150 mM NaCl, pH 7.4) containing 0.5% Triton x 100, 0.1% Na-Azide, 0.1% Na-Citrate, and 5% normal goat serum. Immunostainings were performed by incubation of the sections with primary antibodies diluted in the blocking solution for 48 h at 4 ℃ on a shaker, followed by incubation with the secondary antibody diluted in the blocking solution for 2 h at room temperature. Finally, sections were mounted in AquaMount (DAKO, Glostrup, Denmark). The following primary antibodies were used for immunohistochemistry: anti-Neuronal Nuclei (NeuN) (mouse, Millipore), anti-Doublecortin x (guinea pig, Millipore), anti-TRIM32-1137 (Figure A.1A, rabbit, Gramsch Laboratories, Schwabhausen, Germany), anti-TRIM32-GS (Figure A.1B, rabbit, Gramsch Laboratories, Schwabhausen, Germany). As secondary antibodies Alexa goat anti-rabbit-568, Alexa goat anti-rabbit 568, Alexa goat anti-mouse 488, Alexa goat anti-mouse 568, and Alexa goat anti-guinea pig 568 (all from Invitrogen) were used. Nuclei were stained using Hoechst 33342 (Invitrogen). Images were collected by confocal microscopy using ZEN software (Zeiss, Jena, Germany); image analysis was performed with the ZEN software, Adobe Photoshop, Image J software, and Imaris software (Bitplane).

A.4.9 Terminal Deoxynucleotidyltransferase-Mediated Dutp Nick End Labeling (TUNEL)

TUNEL staining was used to detect DNA fragmentation in situ and performed with the In Situ Cell Death Detection Kit, TMR red (Roche, Cat.-Nr. 12156792910) according to manufacturer's instructions. In brief, 40μm brain sections of mouse brains were obtained as described above and blocked for 1 h at room temperature in TBS containing 0.5% Triton x-100, 0.1% Na-Azide, 0.1% Na-Citrate, and 5% normal goat serum. Sections were washed in PBS twice for 5 min each in PBS and incubated with the TUNEL labeling solution. Therefore, two brain sections were simultaneously incubated with 250 μl of TUNEL labeling solution in one well of a 24-well plate for 1 h at 37 ℃ covered with aluminum foil. Sections were once washed with PBS containing Hoechst for 10 min at room temperature to stain nuclei. Before mounting sections in AquaMount (DAKO, Glostrup, Denmark) they were once washed in PBS for 10 min at room temperature. TUNEL positive (TUNEL+) cells were counted.

A.4.10 Statistics

Graphics presented and statistics carried out were done using the statistical software "R" Version 2.15.0 (R Core Team, 2012). A significance-level (α) of 0.05 was selected. Data were analyzed using t-tests for comparisons between genotypes. The learning performance in the BM was analyzed using a repeated measures ANOVA with genotype as the between subject factor and trial as the repeated measure. Olfactory habituation within each genotype was analyzed by paired t-tests comparing the last trials of habituation with the respective first trial of a newly presented odor. Differences between genotypes were analyzed by unpaired t-tests comparing the first trials of each presented odor. In addition, Sigma Plot was used (Systat Software, Inc., San Jose, USA).

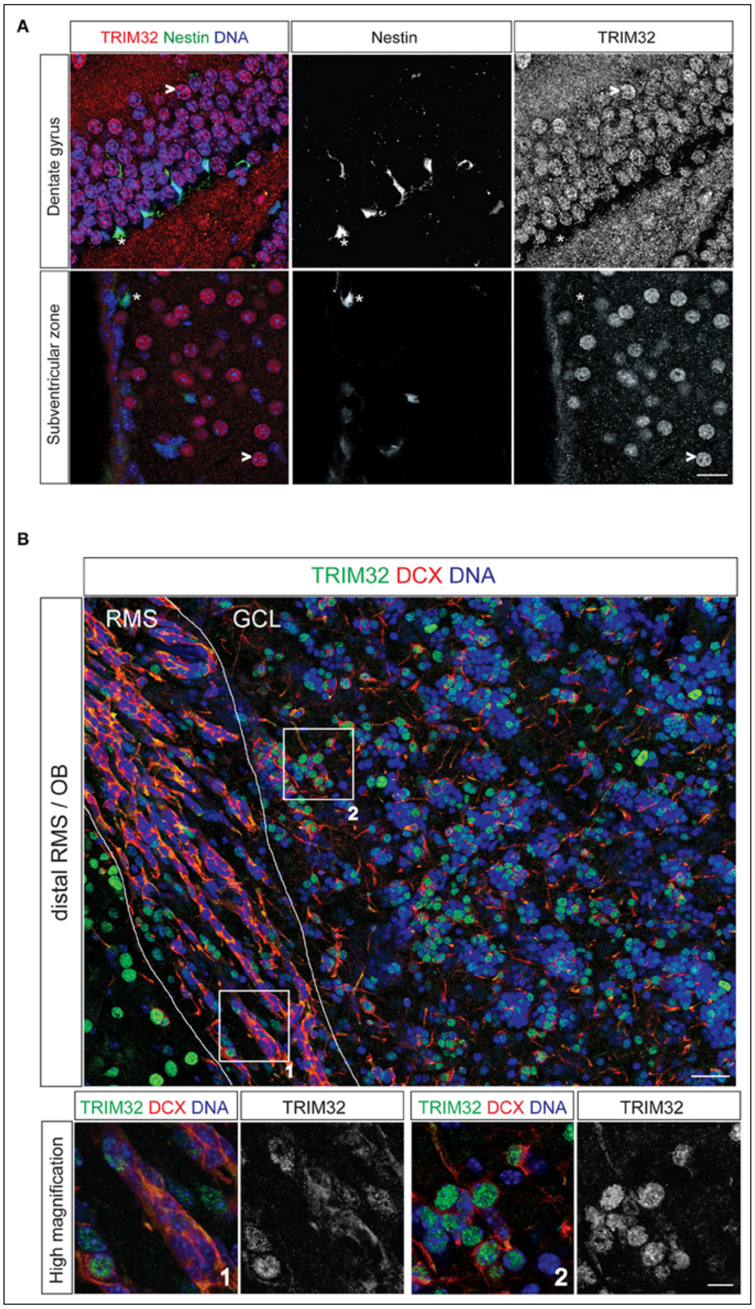


Figure A.1. TRIM32 is upregulated upon neuronal differentiation of subventricular zone (SVZ) and dentate gyrus (DG) stem cells. Free floating sections from adult mouse brain stained with the indicated antibodies. (A) Free floating sections from adult Nestin-GFP mice stained with the indicated antibodies. (*) highlights neural stem cells in the DG and SVZ, (>) marks mature neurons. Scale bar = 20μm(B) Free floating sections from wt mice stained with the indicated antibodies. Images in the lower panel represent high magnification of the indicated areas labeled in upper image. Scale bar = 30 Îijm for upper image, 10μm for lower panel. RMS, rostral migratory stream; GCL, granular cell layer; OB, olfactory bulb.

A.4.11 Metabolite Extraction

A.4.11.1 Brain Tissue

For quenching, dissected brain tissues (mainly consisting of striatum, cortex, RMS, SVZ, and Hippocampus) were directly snap-frozen in liquid nitrogen and stored at 80 ℃ until metabolite extraction. Pre-weighted brain tissues were transferred in 2 ml-Precellys tubes prefilled with 0.6 g ceramic beads (⊘ 1.4 mm, Peqlab, Germany) and the appropriate amount of extraction fluid (MeOH/H2O, 40+8.5, v/v) was added. For sample lysis, a Precellys24 (Bertin, France) homogenizer was used (30 s at 6000 rpm). The temperature was held at 0 ℃ by using the Cryolys cooling option (Bertin, France). Then water (200μl/100 mg tissue) was added to the homogenized tissue fluid, followed by chloroform (400μl/100 mg tissue). The homogenate was incubated for 20 min at 4 ℃ under continuous shaking. After the incubation period, the samples were centrifuged at 14,000 xg for 5 min at 4 ℃. Finally, 20μl of the upper aqueous phase were transferred into a sample glass vial with micro insert. Samples were evaporated using a CentriVap Concentrator (Labconco, USA) at 4 ℃.

A.4.11.2 Cell Culture

Cells were grown in six-well plates. For higher signal intensity two wells were pooled. First, the cells in all wells were washed with 1 ml 0.9% NaCl. After quenching with 0.4 ml cold methanol (20 ℃) and adding an equal volume of cold water (4 ℃), cells were collected with a cell scraper and transferred into the second well followed by cell scraping.

The cell extract was transferred into reaction tubes containing cold chloroform (20 ℃). The extracts were incubated at 4 ℃ for 20 min under shaking followed by centrifugation at 16,000 x g for 5 min at 4 ℃. 0.3 ml of the polar phase were transferred into sample glass vials with micro inserts and evaporated using a CentriVap Concentrator (Labconco, USA) at 4 ℃.

A.4.12 Derivatization and GC-MS Analysis

Metabolite derivatization was performed by using a multi-purpose sampler (GERSTEL, Germany). Dried samples were dissolved in 15μl pyridine, containing 20 mg/ml methoxyamine hydrochloride, at 40 ℃ for 60 min under shaking. After adding 15μ N$-$methyl$-$N$-$trimethylsilyl$-$triflouroacetamide (MSTFA) samples were incubated at 40 ℃ for 30 min under continuous shaking.

GC$-$MS analysis was performed by using an Agilent 7890A GC coupled to an Agilent 5975C inert xL MSD (Agilent Technologies, Germany). A sample volume of 1 mul was injected into a Split/Splitless inlet operating in splitless mode at 270 ℃. The gas chromatograph was equipped with a 30 m DB-35MS capillary column + 5 m DuraGuard capillary in front of the analytical column (Agilent J&W GC Column).

A.4.12.1 Brain Tissue Extracts

Helium was used as carrier gas with a constant flow rate of 1.2 ml/min. The GC oven temperature was held at 80 ℃ for 1 min and increased to 320 ℃ at 15 ℃ /min. The final temperature was held for 8 min. The total run time was 25 min.

A.4.12.2 Cell Culture Extracts

Helium was used as carrier gas with a constant flow rate of 1.0 ml/min. The GC oven temperature was held at 80 ℃ for 6 min and increased to 300 ℃ at 6 ℃ /min. After 10 min, the temperature was increased at 10 ℃ /min to 325 ℃ for 4 min. The total run time was 59.167 min.

The transfer line temperature was set constantly to 280 ℃. The MSD was operating under electron ionization at 70 eV. The MS source was held at 230 ℃ and the quadrupole at 150 ℃. Full scan mass spectra were acquired from m/z 70 to 800. The total run time was 25 min. Ion-chromatographic deconvolution, chromatogram alignment, identification and semi-quantification of metabolite amounts was done with the MetaboliteDetector software (Hiller et al., 2009). TIC normalization was performed to minimize systematic errors during measurement. In detail, the peak area of every compound in a sample was divided by the summed sample signal of all compounds in this sample.

A.4.13 Metabolic Network Modeling

Metabolic network models for wild type and TRIM32 mutated neuronal stem cells were built with the FASTCORMICS workflow (Pires Pacheco and Sauter, 2014) that allows the reconstruction of metabolic models based on microarray data. FASTCORMICS comprises a discretization step based on barcode (Zilliox and Irizarry, 2007), that computes for each probe set of the microarray a z-score for the measured intensity levels after frma normalization (McCall et al., 2011) against a standard intensity distribution of non-expressed genes obtained from a collection of thousands of arrays for the same platform stored in the mogene.1.0.st.v1frmavecs vector. The z-scores were then mapped to the reactions of the mouse model iSS1393 (Heinken et al., 2013) via the Gene-Protein Rules. Reactions with z-scores of zero and below in two out of three arrays were considered as non-expressed and removed from the model. Reactions associated to z-scores above 5 in two of three arrays constitute the set of core reactions. FASTCORMICS builds then consistent compact models that contain a maximal number of core reactions for the two different conditions while not including reactions regulated by non-expressed genes. Random sampling (Becker et al., 2007) of the possible solution space was then performed to obtain a qualitative estimate of the flux distribution allowed by the network topology and constraints of the two models when maximizing for the glycolysis pathway (PGM).

A.5 Results

A.5.1 TRIM32 is Upregulated upon Differentiation of Subgranular and Subventricular Zone Neural Stem Cells

To analyse TRIM32 protein expression by immunofluorescence stainings in NSCs of the adult brain, we used sections from Nestin-GFP mice. In these mice, NSCs in the SVZ and the DG express GFP driven by a Nestin promotor (Yamaguchi et al., 2000). Expression of TRIM32 protein in the SVZ-OB system has been described earlier (Hillje et al., 2013) and was repeatedly analyzed to compare expression levels of TRIM32 protein in progenitor cells of the SVZ with progenitor cells of the DG. Staining of sections from Nestin-GFP mice with an antibody against TRIM32, that has been shown to be specific before in immunohistochemical as well as biochemical approaches (Schwamborn et al., 2009; Hillje et al., 2011), revealed that TRIM32 protein is virtually absent from adult NSCs (type B cells) in the SVZ as well as DG NSCs (A.1A). NSCs and neuroblasts reside only the very inner layer of the DG. As soon as they become immature neurons and finally granule cells, they enter the outer layers of the DG. Cells in these layers show a strong nuclear TRIM32 expression (Figure A.1A), indicating that TRIM32 protein is upregulated upon differentiation of DG NSCs.

Stem cells of the SVZ give rise to transient amplifying cells, which in turn differentiate into neuroblasts. Neuroblasts are generated in the SVZ and migrate toward the OB along the RMS. Co-staining of the neuroblast marker doublecortin and TRIM32 in sections from adult wildtype mice (wt) brains indicate that neuroblasts located in the distal part of the RMS express TRIM32 in the cytoplasm as well as in the nucleus (Figure A.1B). Once these cells reach the OB, TRIM32 is strongly expressed in the nucleus. These data indicate that TRIM32 protein expression is upregulated upon neuronal differentiation of DG subgranular and SVZ NSCs. Furthermore, they are in good agreement with the TRIM32 expression pattern in the SVZ-OB system that we have shown previously (Hillje et al., 2013).

A.5.2 Loss of TRIM32 leads to more Newly Generated Neurons and Less Apoptosis in the SVZ OB System but not the DG

Since TRIM32 is upregulated during the critical period of differentiation of progenitors into neurons in the SVZ and DG, we analyzed rates of neurogenesis in the SVZ - OB system and DG of wt and TRIM32 deficient mice. BrdU was applied to mice from both genotypes on 3 consecutive days and the brains were fixed 14 days after the last injection. Compared to wt mice, we found a significant increase in the density of BrdU+ cells in the granule cell layer (GC) of the OB of TRIM32 ko mice (Hillje et al., 2013). In contrast, also there is a tendency toward more BrdU+ cells in the DG of TRIM32 ko mice, this tendency did not reach statistical significance (Figure A.2A,B). The higher density of BrdU+ cells in the OB GC of TRIM32 ko mice could either be due to an increase in proliferation of progenitor cells or increased survival of newly generated neurons in the OB. Recently, we have shown that the density of cell cycle active cells (Ki67+) is higher in the SVZ-OB system of TRIM32 ko mice (Hillje et al., 2013). However, the density of cell cycle active cells was unchanged in the DG (data not shown). Concerning the rate of cell death, the amount of Casp3+ as well as TUNEL+ cells was significantly reduced in the OB of TRIM32 deficient mice (Hillje et al., 2013). In the DG, we did not find significant changes in the amount of apoptotic cells (Supplementary Figure 1). Taken together, these data implicate that loss of TRIM32 leads to an overproduction of newly generated neurons and less apoptosis in the OB. No significant changes were observed in the DG.

A.5.3 Loss of TRIM32 does not Impair Exploratory Behavior, Anxiety Related Behavior and Spatial Learning but Leads to Increased Numbers of Stops in the Open Field Test

A health check did not reveal any significant differences between the genotypes. Furthermore, no severe deficits that would have excluded animals of either genotype from further analysis were observed. The weight development of individual mice was monitored during the course

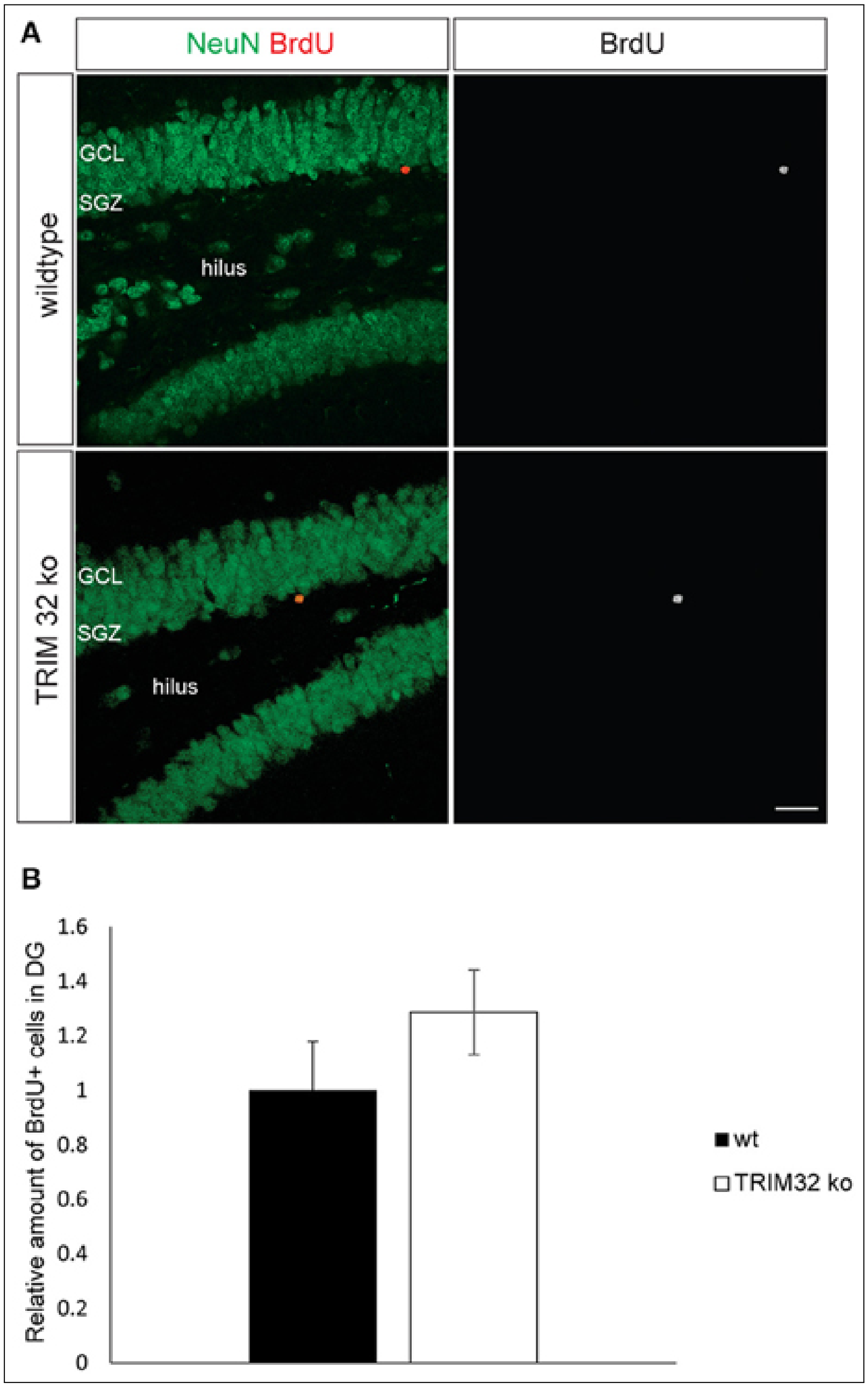


Figure A.2. Loss of TRIM32 is not influencing the rates of adult-born neurons in the dentate gyrus (DG) significantly. (A) Freefloating sections including the DG of wildtype (wt) and TRIM32 ko mice that were injected with BrdU and stained with the indicated antibodies. Scale bar = 20 μm. (B) shows the quantifications of (A). N = 5 mice ($p < 0.05$). GCL, granular cell layer; SGZ, subgranular zone.

of the study for 15 weeks. At first weighing at an age of 22 days, TRIM32 knockout mice weighed significantly less compared with wild-type conspecifics (mean ko: 6.78 g, wt: 7.96 g). This difference disappeared during the following weeks where both genotypes were virtually indistinguishable from each other (Supplementary Figure 2).

In the Elevated Plus Maze neither the percentage of time spent on open arms nor the percentage of entries into open arms did differ significantly between the genotypes (Figure A.3A). Additionally measured parameters for general activity did not reveal any significant genotype effects. In the Open Field Test no significant differences between the genotypes were detectable concerning the traveled path length (Figure A.3B). However, TRIM32 knockout mice were observed to show a significantly increased number of stops (Figure A.3B).

In order to analyze visual spatial memory a Barnes Maze test was conducted. Mice of both genotypes significantly learned to find the position of the correct hole during the course of the training phase of 4 consecutive days indicated by a significant trial effect revealed by a repeated measures ANOVA [F(1, 170) = 392.86, p < 2e-16] (Figure A.3C). The genotypes, however, did differ neither during the training phase nor in the probe trial or the re-trial. Thus, loss of TRIM32 does not impair exploratory behavior, anxiety-related behavior but leads to increased number of stops in the open field test.

A.5.4 TRIM32 Deficiency Impairs Olfactory Discrimination

To determine the influence of increased neurogenesis due to loss of TRIM32 on olfactory capabilities, olfactory memory was tested by means of an olfactory habituation test. Each tested mouse was firstly exposed to distilled water and subsequently to three different odors. Each substance was applied three times in a row with an inter trial interval of 30 s. Both genotypes habituated to the repeated presentation of distilled water on a cotton swab. After habituation to distilled water mice of both genotypes significantly increased sniffing time toward the new odor indicating general olfactory abilities (Figure A.4A). This was true for all three odors. Mice of both genotypes habituated to the odor comparable to the presentation of distilled water. But, the decrease of the sniffing time from the first to the second presentation was significantly lower for TRIM32 ko mice compared to wt mice for odor 2 and 3 (Figure A.4B). Compared to the initial sniffing time of the first trial, TRIM32 knockout mice spend less time sniffing at trial 2 and 3 without recognizing the already known odor. Thus, habituation levels were weaker for TRIM32 knockout mice.

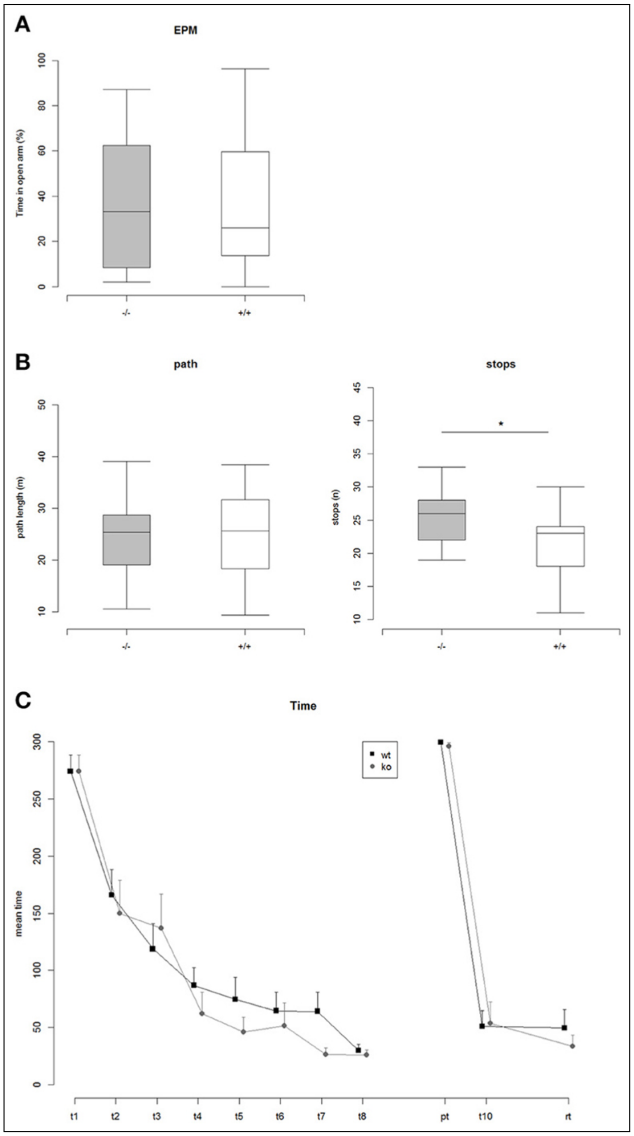


Figure A.3. (A) Percent time on open arms in the Elevated Plus Maze (EPM): ko vs. wt: Two Sample t-test, t = 0.17, p = 0.86 (n.s.), Nko = 14, Nwt = 20. (B) Left: Path length traveled in the Open Field Test: ko vs. wt: Two Sample t-test, t = 0.15, p = 0.88 (n.s.), Nko = 14, Nwt = 21. Right: Number of stops while exploring the Open Field arena: ko vs. wt: Two Sample t-test, t = 2.36, p = 0.025 (*), Nko = 14, Nwt = 21. (C) Mean time to find the correct hole on the Barnes maze. Repeated measures ANOVA revealed a highly significant effect of trial, indicating that both genotypes learned the position of the correct hole [F(1, 170) = 392.86, p < 2e-16]. There was no effect of genotype. In addition a comparison of the areas under the learning curves did not reveal any significant differences between ko vs. wt, AUC-Analysis: Two Sample t-test, t = 1.02, p = 0.32.

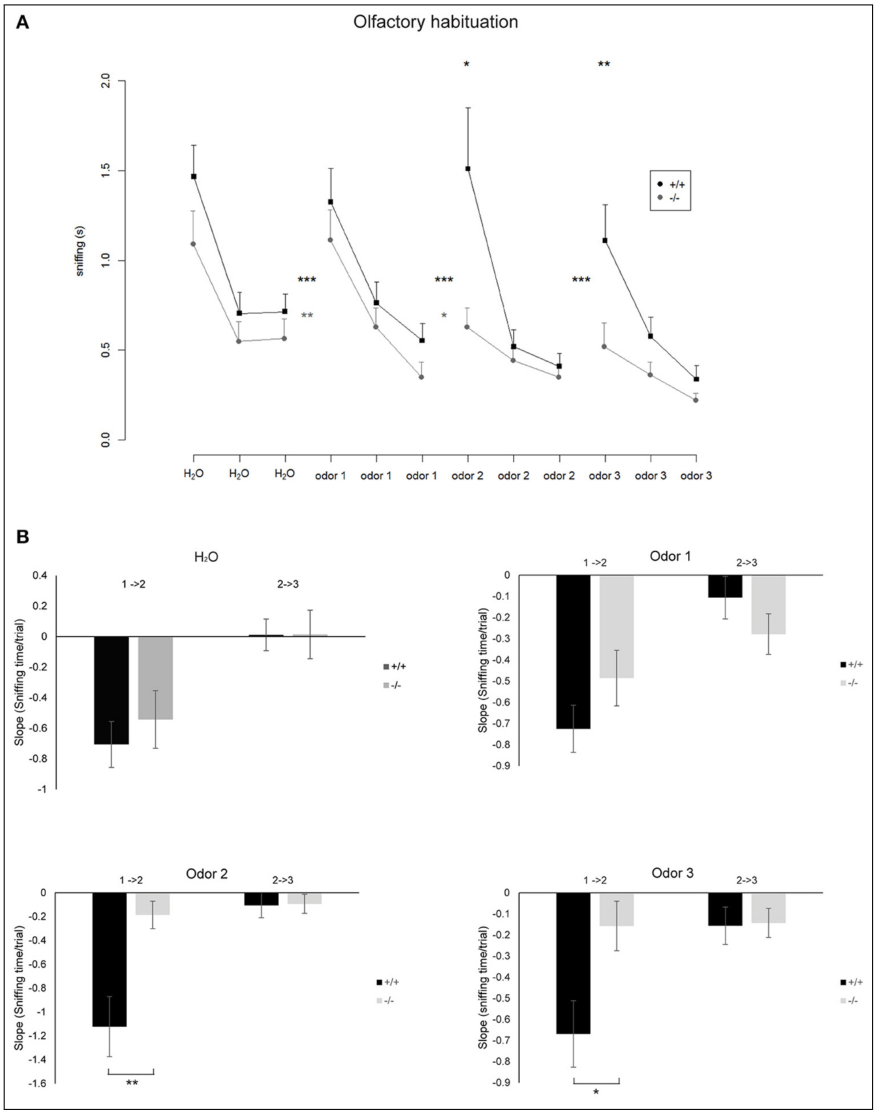


Figure A.4. (A) Mean time spent sniffing on different odors in the Olfactory Habituation Test. Odors were presented three times in a row and subsequently a new odor was presented. Significant differences between the time spent sniffing in the last trial of a known odor compared with the first presentation of a new odor are indicated by asterisks between the lines (paired t-tests, $*p < 0.05$, $**p < 0.01$, $***p < 0.001$). Differences between the genotypes regarding the time spent sniffing in the first trial of a newly presented odor are indicated by asterisks above the curves (Two sample unpaired t-tests). Nko = 14, Nwt = 18. (B) Bar diagrams representing the slope that indicates the rate at which the sniffing time decreases from the first to the second trial (1 → 2) and second to the third trial (2→ 3) for the indicated odors. Differences in genotypes are indicated by asterisks ($*p < 0.05$, $**p < 0.01$) (according to normal distribution of the values t-test for odor 1, MannâĂŞWhitney Rank Sum Test for odor 2 and 3). Nko = 14, Nwt = 18.

In general, compared to wt mice, TRIM32 ko mice spend shorter sniffing times at the cotton swap for all odors already from the very first presentation of the odor. Taken together, increased rates of neurogenesis due to loss of TRIM32 do not lead to an increased olfactory activity but in contrast even impair olfactory performance.

A.5.5 Loss of TRIM32 Leads to Deregulated Brain Metabolism

Accumulating evidence suggests that a deregulation of certain molecular pathways leads to the formation of brain disorders as well as anxiety and depression related phenotypes. To identify affected anxiety and depression related pathways, we performed metabolomic analyses from brain tissue of 3 wt and 4 TRIM32 ko mice. Two hundred and ten metabolites were detected by GC/MS of which 75 have been identified using our in-house mass spectral metabolite library (Supplementary Figure 3). Statistical analysis revealed that levels of nine out of these 210 metabolites differed significantly in levels of concentration between the two genotypes ($p < 0.05$). Of those, five have been be identified (Figure A.5A). These metabolites are phosphomonomethylester, 3-phosphoglyceric acid, cysteine, putrescine, and uracil. The concentration of 3-phosphoglyceric acid, which is a metabolic intermediate in glycolysis and is also a switching point to glycine and serine metabolism, was significantly higher in the tissue of TRIM32 ko mice (Figures A.5B,C). A similar significant increase was found for cysteine, which is related to serine, methionine, and glutathione metabolism. Concentrations of the degradation product putrescine, that has been linked to methionine metabolism as well, were elevated in a similar way (Figures A.5,C). The mouse brain tissue that was used for our metabolomics analysis not only contains NSCs but represents a mixture of multiple cell types. To investigate the relevance of these results for NSCs we performed the same analysis using a pure population of cultured NSCs (Supplementary Figure 4). From the five identified metabolites that significantly differed in their concentration from wt to TRIM32 ko brain tissue, putrescine and 3-phosphoglyceric acid were also significantly upregulated in TRIM32 ko NSCs.

Finally, we wanted to get a more systemic view of deregulated pathways that might lead to pathological changes in the brains of TRIM32 ko mice. Therefore, previously generated gene-expression data from TRIM32 ko mice (Hillje et al., 2013) were linked to the results that were obtained in the metabolomic analysis of wt and TRIM32 ko brains. Genome-scale metabolic network models were reconstructed via the FASTCORMICS workflow making use of the available microarray data (Pires Pacheco and Sauter, 2014). The wt and TRIM32 ko models contain

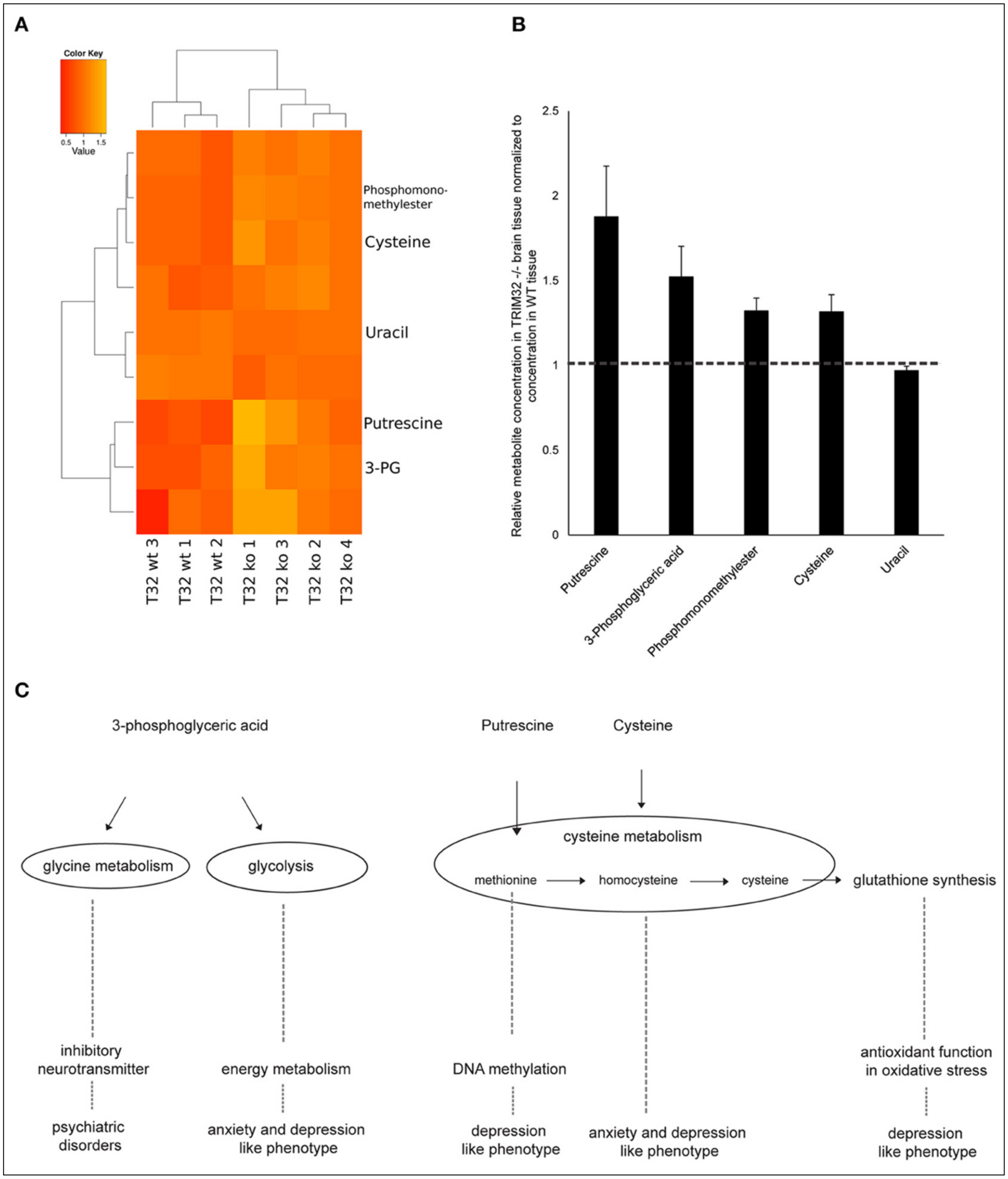


Figure A.5. (A) Heatmap showing metabolites that differ significantly between wt and TRIM32 ko mice (Student's t-test, p-value < 0.05). Medians of three technical replicates were used as basis for this data analysis. For visualization the individual intensities for each compound were divided by its mean intensity across all replicates. Colors represent metabolite levels in TRIM32 ko and wt brain tissue. Clustering was performed on Euclidean distances using Ward's minimum variance method. Three wt and 4 ko animals were used for analysis. (B) Relative concentration of metabolites that differ significantly in concentration between wt and TRIM32 ko mice. Data were calculated as means with standard error of the mean and values of TRIM32 ko mice were normalized to wt values. (C) Schematic overview depicting metabolomic pathways to which metabolites that were significantly different in concentration are linked to and their involvement in brain disorders and behavioral phenotypes.

735 and 635 reactions, respectively, with 516 reactions being shared between the two models, indicating some metabolic differences between the two conditions. Both models are available in SBML format as Supplementary Files. A qualitative estimate of the flux distribution of the pathways containing the significantly changed metabolites was then obtained by random sampling of the possible solution space (Figure A.6). It suggests in the ko model a decreased consumption of 3-phosphoglyceric acid by the phosphoglycerate dehydrogenase, the first enzyme in the serine biosynthesis pathway. This in turn leads to a slight accumulation of the glycolytic intermediate 3-phosphoglyceric acid, as observed in the metabolite measurements (Figure A.5). Taken together, we were able to show that there is a difference in metabolic profiles between wt and TRIM32 ko brains. In addition, modeling the estimation of flux distribution suggests that the increase of 3-phosphoglyceric acid, which was significantly deregulated in the TRIM32 ko brain tissue as well as in the TRIM32 ko NSC culture, might be due to a decreased consumption by phosphoglycerate dehydrogenase.

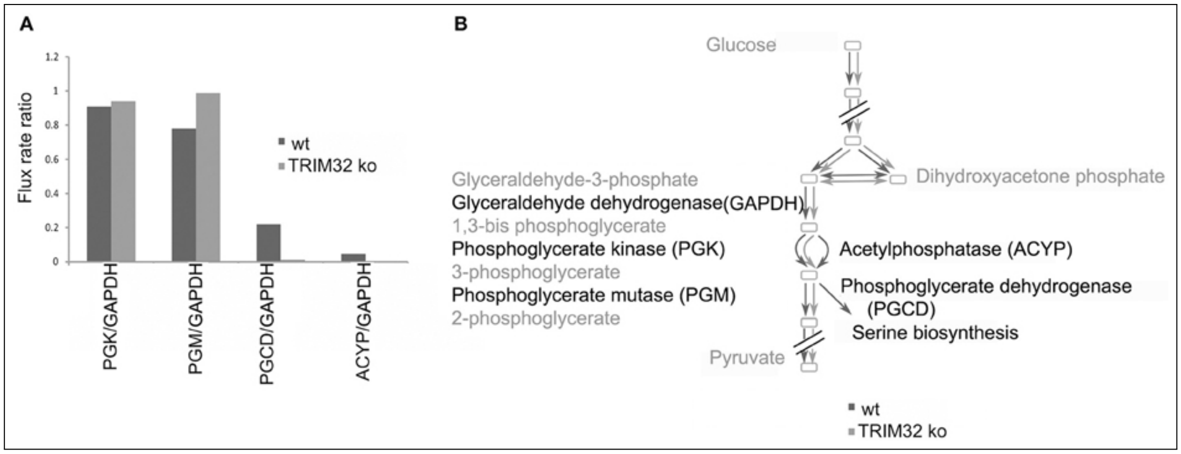


Figure A.6. Flux distribution estimated by random sampling for the wild type and the mutant models built via the FASTCORMICS workflow. (A) Random sampling results. Ratio of the flux rates for phosphoglycerate kinase (PGK), phosphoglycerate mutase (PGM), and acetylphosphatase (ACYP) over glyceraldehyde dehydrogenase (GADP) represented in blue for the wild type and in red for the mutant. (B) Schematic representation of the qualitative wild type (dark gray) and mutant (light gray) fluxes over the glycolysis pathway.

A.6 Discussion

We recently demonstrated that the absence of TRIM32 in knockout mice led to increased progenitor cell proliferation and less cell death, both effects accumulate in an overproduction of adult generated olfactory OB neurons of TRIM32 knockout mice (Hillje et al., 2013). Here, we show that such an increase does not necessarily lead to better olfactory performance but contrary, TRIM32 knockout mice even show impaired olfactory habituation.

The performed olfactory habituation assay evaluates the habituation to known odors as well as the detection of a new odor. Although it might be possible that TRIM32 ko mice merely lack interest or motivation in sniffing new odors, we believe that the increase of sniffing time for the first presentation of a new odor following the presentation of water or following the third presentation of an already presented odor indicates that there is general interest in a new odor. For both genotypes the sniffing time for the first presentation of odor 1 is significantly longer compared to the sniffing time of the last presentation of water. The same is true for the first presentation of odor 2. For the first presentation of odor 3 there is an increase even though not significant for TRIM32 ko mice. If TRIM32 ko mice would have no interest or motivation we would not expect to see this increase. In addition, these results point to the fact that TRIM32 ko mice are able to distinguish between the already known odor and a new odor. However, the lower decrease of sniffing time upon the second and third presentation of subsequently presented odors points to the fact that it takes TRIM32 ko mice longer to recognize the already known odor, meaning an impairment of olfactory memory and habituation. However, we cannot rule out that the lower decrease of sniffing time is a combinatorial effect of impaired memory (habituation), lower discrimination (olfactory capabilities), and lower interest. Anyways, since we tested a battery of non-olfactory based behavioral tests that showed no differences regarding emotional and motivational states (with respect to anxiety-related behavior tested in the EPM, motivation to gain access to the home cage and general learning performance tested in the Barnes Maze Test) we believe that lack of motivation is most probably not the main cause of the observed behavior. Furthermore, we also cannot fully exclude that TRIM32 deficient olfactory neurons are dysfunctional and this potential dysfunction contributes to the observed phenotypes.

Although there have been conflicting results in the past and its exact function could not be fully determined yet, adult neurogenesis seems to play an important role in olfactory processes including olfactory discrimination, memory and associative learning (for review see (Breton-Provencher and Saghatelyan, 2012)). Our data indicate that olfactory habituation was significantly impaired in TRIM32 ko mice. Thus, the mere number of newly generated neurons itself does not guarantee an improvement of olfactory information processing. These results are in line with findings by (Mechawar et al., 2004), who found that an increased number of granule cells due to a decreased apoptosis rate did not result in enhanced olfactory abilities, but lead to a declined short-term memory of odors. We hypothesize that due to the increased proliferation and decreased cell removal rates, the newborn interneurons are defectively integrated into the circuitry and therefore the animals show impaired olfactory functioning. Strikingly, long-term memory seemed not to be affected of this as the knockout mice showed similar habituation to the odors as their conspecifics after a second introduction of the same odor set after 1 week (data not shown). Most interestingly, other behavioral domains (such as exploratory behavior, anxiety-like behavior, and spatial learning) were not affected by the lack of TRIM32. The only remarkable difference we found was the number of stops conducted during the Open Field Test. Stopping while exploring a new environment is a typical behavior shown by mice. We speculate that the later shown olfactory deficits most likely have increased the demand for such back-pedaling behavior in order to gain information in the face of impaired olfactory capacities. There is growing evidence that deregulation in metabolite concentrations leads to the formation of brain disorders as well as anxiety- and depression-related phenotypes. Our statement that the cell fate determinant TRIM32 is required for a balanced activity of the adult neurogenesis process is supported by hints that point to a deregulation of several metabolic intermediates that are connected to glycolysis, glycine or cysteine metabolism in TRIM32 knockout mice brain tissue. Our data implicate that loss of TRIM32 leads to changes in the concentration of 3-phosphoglyceric acid, a metabolite linked to glycolysis. Rates of glycolysis have been shown to be deregulated in anxiety and depression-like phenotypes in systems biology approaches (Gormanns et al., 2011) as well as in different anxiety mouse models (Filiou et al., 2011; Zhang et al., 2011). In addition, our data show significant changes in the levels of cysteine due to loss of TRIM32. Cysteine is an intermediate of the trans-sulfuration pathway (methionine → homocysteine → cysteine) and thus glutathione synthesis. Glutathione is the major antioxidant of the brain and is of particular importance for defense mechanisms against oxidative damage. Methylation of DNA has been suggested to be involved in the cause of mood disorders and strongly relies on the availability of methyl groups from s-adenosyl methionine (SAM). After providing its methyl group, SAM is regenerated via homocysteine and methionine where the methyl group is provided either by trimethylglycine (betaine) or by 5-methyltetrahydrofolate. The latter mainly derives its methyl group from serine via 5,10-methylenetetrahydrofolate. Oxidative stress mechanisms and methylation have been implicated in remitted phases of major depressive disorders in humans (Kaddurah-Daouk et al., 2012) and deregulation in cysteine and methionine metabolism were linked to depression and anxiety in a systems biology approach as well as in a mouse model (Gormanns et al., 2011; Zhang et al., 2011). In addition to the above mentioned pathways, 3-phosphoglyceric acid functions in glycine and serine as well as cysteine metabolism. Glycine is an inhibitory neurotransmitter in the spinal cord and brain stem with a regulatory function in locomotor behavior (reviewed in (Legendre, 2001; Xu and Gong, 2010)). Glycine synaptic transmission was suggested to be involved in psychiatric disorders (Zhang et al., 2011). Since the used brain tissue contains multiple cell types, the metabolomics analysis was repeated using pure populations of cultured wt and TRIM32 ko NSCs. From the five identified metabolites that significantly differed in their concentration between the two genotypes in brain tissue, putrescine and 3-phosphoglyceric acid could be verified to be significantly differently abundant in wt and TRIM32 ko NSCs.

However, the metabolomics approach is limited by only taking a snapshot look at one static point. In order to get a more detailed understanding of metabolomics changes in TRIM32 ko brain, we modeled metabolomics fluxes by combining previously published gene expression (Hillje et al., 2013) and metabolic data. The ko model suggests a decreased consumption of 3-phosphoglycerate by the phosphoglycerate dehydrogenase, the first enzyme in the serine biosynthesis pathway, which leads to a slight accumulation of phosphoglycerate, as observed in the metabolite measurements of the brain tissue as well as the NSC cultures. The deficiency of 3-phosphoglycerate dehydrogenase (3-PGDH) is the most reported defect that causes serine deficiency disorders, a group of neurodevelopmental, neurometabolic disorders with congenital microcephaly, intractable seizures and severe psychomotor retardation (Van der Crabben et al., 2013), implicating the pathological relevance of this pathway. In addition to deregulated metabolic fluxes, loss of TRIM32 results in an overproduction of adult generated OB neurons, and it cannot be excluded that different cell population sizes might cause changes of metabolite levels in TRIM32 ko brains.

Even though no clear anxiety or depression like phenotype was found for TRIM32 knockout mice in the behavioral tests, shorter sniffing times in olfactory habituation tests as well as more stops in the Open Field tests might be hints for lower motivation or might even indicate slight depression like behavior. However, the function of TRIM32 in depression and anxiety is still controversial. Using a chronic unpredictable mild stress (CUMS) mouse model that generates anxiety- and depression-like behavior, Ruan and colleagues showed that TRIM32 protein levels are downregulated in the hippocampus under mild stress (Ruan et al., 2014). However, at the same time they demonstrate that a total loss of TRIM32 (knock-out mouse model) protects against depression. Depression in general is associated to reduced levels of neurogenesis, while in TRIM32 knock-out mice neurogenesis is increased. It seems tempting to speculate that these two effects, in the CUMS depression model in TRIM32 knock-out mice, balance each other leading to a normalized neurogenesis activity. Although in our TRIM32 ko mice we did not observe any anxiety related phenotypes, in our metabolomics approach we detected several metabolites to be deregulated which previously have been shown to be implicated in anxiety and depression. Hence, such a systematic omics approach might be even more sensitive in revealing fine changes in complex behaviors, which might be missed by conventional behavioral assays.

Our study provides comprehensive data on how the deregulation of adult neurogenesis caused by the loss of the cell fate determinant TRIM32 leads to a deregulation of metabolomic pathways and finally results in an impairment of olfactory capabilities. These results highlight that the function of the cell fate-determinant TRIM32 for a balanced activity of the adult neurogenesis process exceeds the cellular level and has far-reaching effects on metabolomics pathways that are linked to mood disorders as well as olfactory capabilities.

Review paper

B.1 Summary and contributions

Parts of the discussion were already published in the review: **Towards improved genome-scale metabolic network reconstructions: unification, transcript specificity and beyond** for which I share the co-authorship with Dr. Thomas Pfau. I wrote the chapters "Transcripts-the information lost in reconstructions" and "Non-specific co-factors can cause loops".

B.2 Abstract

Genome scale metabolic network reconstructions provide a basis for the investigation of the metabolic properties of an organism. There are reconstructions available for multiple organisms, from prokaryotes to higher organisms and methods for the analysis of a reconstruction. One example is the use of flux balance analysis to improve the yields of a target chemical, which has been applied successfully. However, comparison of results between existing reconstructions and models presents a challenge due to the heterogeneity of the available reconstructions, for example, of standards for presenting gene-protein-reaction associations, nomenclature of metabolites and reactions or selection of protonation states. The lack of comparability for gene identifiers or model specific reactions without annotated evidence often leads to the creation of a new model from scratch, as data cannot be properly matched otherwise. In this contribution, we propose to improve the predictive power of metabolic models by switching from gene-protein-reaction associations to transcript-isoform-reaction associations, thus taking advantage of the improvement of precision in gene expression measurements. To achieve this precision, we dis-

cuss available databases that can be used to retrieve this type of information and point at issues that can arise from their neglect. Further, we stress issues that arise from non-standardized building pipelines, like inconsistencies in protonation states. In addition, problems arising from the use of non-specific cofactors, e.g. artificial futile cycles, are discussed, and finally efforts of the metabolic modelling community to unify model reconstructions are highlighted.

B.3 Introduction

Over the last two decades, the increasing availability of genomic, proteomic and metabolomic information has led to the generation of a multitude of metabolic network reconstructions (Kim et al., 2012b). These reconstructions aim to represent our collective knowledge about the metabolism of the reconstructed organisms. They serve as a source of information on their target organism, and models derived from the reconstructions can be used to investigate its metabolic capabilities. The available reconstructions cover multiple types of organisms, ranging from microorganisms, like *Escherichia coli* (Reed et al., 2003; Orth et al., 2011; Keseler et al., 2013) and *Saccharomyces cerevisiae* (Förster et al., 2003; Aung et al., 2013), to complex multicellular organisms, like *Arabidopsis thaliana* (Poolman et al., 2009; de Oliveira Dal'Molin et al., 2010; Mintz-Oron et al., 2012) or *Homo sapiens* (Duarte et al., 2007; Ma et al., 2007; Thiele et al., 2013).

Despite the availability of high quality protocols for the reconstruction of a genome-wide network (Thiele and Palsson, 2010), efforts are far from consistent between different groups. The most common differences are multiple naming schemes for reactions, metabolites, and genes, along with different formats for reconstruction exchange. Some of the issues arising from these differences have been discussed in (Monk et al., 2014). The main challenge is to compare networks generated by different reconstruction tools, or using different naming schemes (Kumar et al., 2012). Furthermore, the lack of precise annotations leads to information being overlooked that could improve the models resulting from reconstruction efforts. With automation of model generation (Overbeek et al., 2005; Agren et al., 2013), in particular towards tissue specific submodels (Wang et al., 2012; Vlassis et al., 2014), it becomes ever more important that reconstructions are curated in a consistent way.

There have been attempts to establish databases that can help in generating consistent networks by providing links to multiple databases, like MetRxn or MetaNetX (Kumar et al., 2012; Ganter et al., 2013). These studies also highlighted the issues arising from the multitude of naming schemes used. While we know that there are multiple pathways which are shared between a multitude of organisms (like glycolysis or the Krebs cycle) finding these similarities in reconstructions is challenging. The authors of MetRxn report that by using simple string matching techniques only three reactions could be directly inferred as being identical in a set of over 30 models (Kumar et al., 2012). Thus unification is paramount to determine the novelty of new reconstructions.

Unified representation, however, is not the only issue with current reconstructions. Most reconstructions rely purely on genetic information for functional annotation, however recent advances in both microarray and RNA-seq technologies provide information about mRNA on a transcript level. Inclusion of this kind of information could potentially increase the accuracy of models. Another issue that can influence predictions is cofactor specificity, which has been shown to be influential in metabolic modelling (Cheung et al., 2013). In this paper we will highlight potential approaches to unify metabolic network representations, and highlight the importance of transcript specificity to metabolic networks. We will further elaborate on the issues arising from cofactor specificity in metabolic network analysis (e.g. sets of reactions using either NADPH or NADH, which can form futile cycles indicating those reactions as active while in truth they are disconnected from the network). Finally, we will provide an overview of projects aiming at improving the current lack of unification, by coordinating multiple reconstruction efforts for the same organism, or creating databases with compatible networks.

B.4 Steps towards a unification of model representation

Metabolites and reactions linking them form the core of a metabolic network. Additional information is often provided in the form of genes which are coding for enzymes catalysing a specific reaction. These can be simply lists of genes associated with a reaction, or they can form gene-protein-reaction association (GPR) rules representing protein complex formation. To provide this information multiple different types of formats have been used (see Table B.1). Some, like the Systems Biology Markup Language (SBML, (Hucka et al., 2003)) or spreadsheets are platform-independent while others, like MATLAB structs, depend on a specific software. The advantage of SBML over other formats is its versatility, and general usability by almost all current

Model Style	Description	Advantages // Disadvantages	Examples
SBML/COBRA	SBML with additional information in the notes sections of entries (Schellenberger et al., 2011a)	Models are usable in any SBML capable tool but the additional information needs explicit parsers. Tool independent. // There is no clear definition of used fields in the SBML format and different groups use multiple different data fields.	BiGG models (Schellenberger et al., 2011b), MetaCyc SBMLs (Caspi et al., 2014), iJO1366 (Orth et al., 2011)
SBML/Mod	SBML using *ModifierSpecies* to define GPRs	Models are usable in any SBML capable tool. Genes can be linked to multiple sources. Proteins can be encoded and linked explicitly. Tool independent. // Needs parsers that make use of these properties. Lacks a defined standard how *ModifierSpecies* have to be defined.	HMR (Mardinoglu et al., 2014a), yeast consensus (Aung et al., 2013)
SBML/FBC	SBML with FBC extension for flux balance analysis specific information (Olivier and Bergmann., b)	Uses SBML defined fields (from the FBC extension) to provide FBA specific information. Proteins can be encoded (and identified) explicitly. Tool independent. // FBC extension not yet processed by many tools.	BiGG2 Database (Systems Biology Research Group, 2015)
Toolbox specific formats	Formats specific to one modelling tool e.g. COBRA MATLAB files (Schellenberger et al., 2011a) or ScrumPy .spy files (Poolman, 2006)	Files can directly be used in the respective toolbox and can contain additional information. // Not easily loaded into other tools.	Recon2 (Thiele et al., 2013), iMM1415 (Sigurdsson et al., 2010) (Poolman et al., 2009)
Spread sheets	Commonly multiple sheets or files with compounds, reactions and genes	Easily accessible for non computational users. Tool independent. // Difficult to parse for further analysis, due to the lack of a standard format.	HepatoNet (Gille et al., 2010), (Oh et al., 2007), iNJ661 (Jamshidi and Palsson, 2007)

Table B.1. Different formats for the exchange of metabolic models. Annotation of the SBML is either achieved by COBRA notes fields (e.g. for Database links), or using bio qualifiers (BQ) and the annotation class of SBML. Both types have been used in combination with SBML/Mod and SBML/COBRA, even though commonly SBML/COBRA models do not include BQ annotations, as they rely on the COBRA annotations.

software tools specific to metabolic modelling (for recent reviews on these tools see (Lakshmanan et al., 2014) or (Dandekar et al., 2014)). Nowadays, most models are indeed published in the SBML format(Schatschneider et al., 2013; Mardinoglu et al., 2014a; Dias et al., 2014; Larocque et al., 2014). In addition, many software tools, even if they have an alternative internal storage format, like ScrumPy(Poolman, 2006), COBRA (Schellenberger et al., 2011a), RAVEN (Agren et al., 2013), or Pathway Tools (Karp et al., 2010), provide some type of import and export functionality to read and generate SBML files that can be used as input into other tools. However, there are still models like the latest versions of the popular metabolic network reconstruction of *Homo sapiens*, Recon2, which are only available as a MATLAB export specific to the COBRA toolbox environment (Schellenberger et al., 2011a). Beyond the common general file format models tend to diverge substantially.

B.4.1 Flux balance specific information

Gene-protein-reaction (GPR) association rules, which are commonly used to link gene expression or proteomics data to metabolic networks (Becker and Palsson, 2008; Jerby et al., 2010; Agren et al., 2012; Yizhak et al., 2014a), are inconsistently represented in different models. While some reconstructions provide those GPRs in supplemental spreadsheets (Gille et al., 2010), the COBRA toolbox defines additional fields in the SBML *Notes* section of a reaction, that contain the GPR rules (Schellenberger et al., 2011a). Recently some reconstructions, like the yeast consensus model (Aung et al., 2013) or the Human Metabolic Reconstruction (HMR) (Mardinoglu et al., 2014a), provide *ModifierSpecies* which are annotated as being encoded by specific genes using bio-qualifiers (Li et al., 2010). The COBRA toolbox also added further information into the *Notes* section, including metabolite formula and charge information, or information on pathways that include a given reaction. While this information is useful for network analysis, it lacks a clear definition of which fields can be used or should be present. Thus, multiple different fields have been used across models, with some fields remaining undefined in some models. However, this information could also be provided within the annotation field of a metabolite or reaction using biomodel qualifiers (BQ) (Li et al., 2010), e.g. a reaction *isPartOf* a specific pathway, without the necessity of additional field definition. Another specification made by the COBRA toolbox was to use the *kineticLaw* field to define flux constraints, thus using a structure that is not designed to hold this information but is supposed to be used for real kinetic information. Since SBML is a general systems biology representation, this could lead to confusion if the structure of a stoichiometric model is imported into a kinetic tool. These inconsistencies in the use of SBML, in addition to the increasing amount of available reconstructions, have prompted the development of the ’FBC’ extension (Olivier and Bergmann., a) to SBML, which covers many aspects specific to flux balance analysis (FBA). While initially only providing support for flux bounds and providing additional SBML fields for charge and formula within the *Species* class, the latest version (Version2, Release 1 (Olivier and Bergmann., b)) also provides facilities to handle GPRs, including the option to add gene products (thus directly adding protein identifiers to the model along with gene/transcript identifiers). FBC further allows the inclusion of specific settings for simulations in *FluxObjectives*. The clear definition of the FBC extension along with its direct link to the SBML specification makes it an ideal choice for data provision.

B.4.2 Naming conventions and comparability

While the ’FBC’ extension handles many of the aspects specific to flux balance models, there is still wide diversity in naming schemes used for metabolite or reaction identification and the choice of gene representation. Until now, there are no generally accepted naming conventions for metabolites or reactions, and thus the choice of identifiers strongly depends on the database used as a basis for the reconstruction, or how the researchers choose to define their system. Naming schemes have included custom abbreviations (Feist et al., 2007; Flahaut et al., 2013), consecutive numberings (Gille et al., 2010), or extracted identifiers from databases (Poolman et al., 2009).

Newer reconstructions tend to make extensive use of the SBML annotation field, Systems Biology Ontology (SBO) identifiers (see http://www.ebi.ac.uk/sbo/) and biomodel qualifiers. Usage of these qualifiers in addition to adherence to standards defined as the “Minimum Information Required In the Annotation of Models” (MIRIAM) (Le Novère et al., 2005) will make it possible to create universally applicable interpreters and tools. However, even when trying to adhere to the MIRIAM standards, it is important to select a proper set of resources to annotate the model components. There are multiple databases for compounds (e.g. CHEBI (Hastings et al., 2013), PubChem (Bolton et al., 2008), KEGG (Kanehisa et al., 2014), MetaCyc (Caspi et al., 2014)), reactions (KEGG, MetaCyc, BRENDA (Schomburg et al., 2013), GO (Ashburner et al., 2000)), proteins (BRENDA, UniProt (The UniProt Consortium, 2014), PDB (Berman et al., 2000), ENZYME (Bairoch, 2000)) and genes (NCBI - Gene (Maglott et al., 2005), UniProt, GeneDB (Logan-Klumpler et al., 2012), GeneCards (Safran et al., 2010)) with some (like KEGG and MetaCyc) catering primarily to metabolism, while others are more comprehensive.

As new models are commonly accompanied by novel functionalities or entities, databases that allow the deposition of new entries would be preferable. While the most popular metabolic databases (MetaCyc and KEGG) do contain entry types on the most relevant entities, they do not allow a direct deposition of new entries. They are therefore unsuitable for deposition of newly developed models, as this would lead to new identifiers that cannot be directly used by others. Using multiple databases to solve this issue can introduce new sources of errors. For metabolites, one database might consider all compounds to be present at a certain pH (like MetaCyc), while other databases represent the same compound as fully protonated (like BRENDA). Thus, when trying to determine charge balance or hydrogen balance, issues arise if inconsistent sources are used, and one source might not provide the required protonation state for all compounds in the reconstruction. If novel compounds, proteins, or genes are introduced in a reconstruction, we would recommend using CHEBI, UniProt and NCBI - Gene to directly deposit the novel entries and use them to annotate the entities in the model. For known compounds a selection of consistent sources (e.g. the same protonation state as in the reconstruction) would, in our opinion, be more suitable than a large selection of databases, with different definitions, to avoid confusion.

B.5 Transcripts - the information lost in reconstructions

As mentioned above, GPRs are informationally important in metabolic reconstructions, in particular when trying to integrate omics data into metabolic networks, e.g. to extract context-specific models from a generic genome-wide reconstruction. The GPRs annotated in metabolic reconstructions mostly consider only genes, completely neglecting the fact that one locus can be translated in different variants through alternative splicing.

Alternative splicing (as shown in Figure B.1) allows increased diversity and regulatory complexity of an organism without requiring a massive increase in genome size (Ladd and Cooper, 2002). It is particularly important in humans, with splicing variants affecting 95% of the genes (Buck et al., 1992; Pan et al., 2008). Even if the different variants have mostly similar functions, in some cases the alternative variants have opposing effects, like the FLICE isoforms that are anti- and pro-apoptotic (Djerbi et al., 2001); provide insufficient activity, as in the instance of the TAZ gene (Vaz et al., 2003); or inhibit the main isoform. An example for the latter is isoform i2 of UGT1A that negatively modulates the glucuronosyltransferase activity of isoform i1 (Girard et al., 2007; Bellemare et al., 2010).

In general, several splice variants are simultaneously expressed, although usually one variant dominates the others, accounting for on average 85% of the protein-coding mRNA at a given loci (Rodriguez et al., 2013). The dominant variant is usually highly conserved during evolution, But the expression pattern is constantly changing to meet cell- and condition-specific requirements (Lewis et al., 2010b). Not only do different celltypes have a different set of variants, but also different individuals show different splicing. Furthermore, switch-like effects, where variants lose their dominant position in favour of other variants, were observed for hundreds of genes during differentiation (Trapnell et al., 2010; Gonzàlez-Porta et al., 2013), demonstrating the plasticity of a tightly regulated process. Alterations of the latter are implicated in numerous pathologies, especially in cancer, and several splice variants are even considered as biomarkers, like PRKC-ζ-PrC for prostate cancer, Nek2C for breast cancer and CD-44 splice variants for colon cancer (Yao et al., 2012b; Liu et al., 2012; Wielenga et al., 1993).

B.5.1 Current use of transcripts

Most metabolic models do not consider transcript variants as functional information is often only available at the gene or protein level. Even metabolic reconstructions that introduced transcript identifiers in their gene-protein-reactions association rules (GPR) based on bibliographic research, like Recon1 (Duarte et al., 2007), do not allow mapping of the transcripts identifiers of the model to transcript identifiers used by databases. This issue arises from the lack of direct matching between the reconstruction identifier and available databases' identifiers. Therefore, in practice, the information related to splicing variants is simply ignored. GPRs are gene-oriented and, as a consequence, the intensity levels of the transcripts variants are usually simply summed up or the maximal intensity values are mapped to the reactions of the model.

Alternative splicing was shown to be altered in a wide range of diseases (Tazi et al., 2009; Nissim-Rafinia and Kerem, 2002). In cancer, usually minor isoforms get overexpressed and dominate the main splice form. For example, the alternative splice form pyruvate kinase isoform 2 (PKM2) favours aerobic glycolysis whereas the main form promotes oxydative phosphorylation. The expression of PKM2, which is the embryonic isoform, is restricted in adults to cancer cells that do not express PKM1 (Mazurek et al., 2005). A model with gene-oriented GPRs cannot differentiate between the two isoforms and will therefore consider the same set of target reactions as active for both isoforms .

The existence of tissue- or context-specific alternative exons involved in the same pathways, and regulated by common mechanisms as e.g. the neural-specific splicing regulator nSR100, was demonstrated in several studies (Calarco et al., 2011; Kalsotra and Cooper, 2011; Licatalosi and Darnell, 2010; Ellis et al., 2012; Xu et al., 2002). Although alternative exons were mostly studied for their impact on protein-protein interaction networks, it is probable that alternative exons have a similar role in metabolic modelling, controlling the activation of tissue-specific metabolic sub-pathways. In this case, a model with gene-oriented GPRs would fail to capture tissue-specific activation patterns.

B.5.2 Sources for transcript specific information

The prevalent barrier to the inclusion of transcript variants in metabolic network reconstruction is the lack of knowledge about the alternative splice forms in most organisms. Databases collecting information on alternative splicing are mainly dedicated to humans, mice and other vertebrates, since splicing is most important in eukaryotic organisms. The largest benefit of this endeavour is therefore expected for human models, where the inclusion of transcript information could explain pathologies linked to alternative splice forms e.g. in cancer (Pal et al., 2012; Chen and Weiss, 2015), neurodegenerative diseases (Mills and Janitz, 2012; Alarcón et al., 2013; Beyer and Ariza, 2013) or autosomal dominant retinitis pigmentosum (Ishunina and Swaab, 2012). The inclusion of information on alternative splice forms will increase the capacity of cell-specific and context-specific models to capture the variability in metabolism of different cell types. However, even for the organisms with the highest information content on alternative splicing, the functional activity of most splice forms remains unknown. To address this problem, several databases have been dedicated for a decade to collecting transcript information. These include ASAP II (Kim et al., 2007), ECGene (Lee et al., 2007), ASTD (Koscielny et al., 2009), HOLLYWOOD (Holste et al., 2006), H-DBAS (Takeda et al., 2007), FAST DB (De La Grange et al., 2005) and FANTOM 3 (Maeda et al., 2006), which try to supplement generic gene databases (ENSEMBL (Hubbard et al., 2002; Flicek et al., 2013), Pfam (Bateman et al., 2002; Finn et al., 2008), Uniprot/Swiss-Prot (Bairoch et al., 2005)). A more intensive review of these databases can be found in (Kelemen et al., 2013) or (Taneri and Gaasterland, 2013).

Problems of automated annotation pipelines The increased amount of data on alternative splicing obtained through deep-sequencing technologies outpaces the capacity of databases to completely annotate the transcripts manually, and therefore nearly all databases use semi-automated or automated pipelines.Automated annotation process are more prone to errors than manual curation. The rate of wrong annotation in GenBank (Bilofsky and Burks, 1988), NR (Benson et al., 2004), TrEMBL (Bairoch et al., 2005) and KEGG (Kanehisa and Goto, 2000) was assessed by (Schnoes et al., 2009), who tested 37 enzyme families. They found misannotation rates ranging from 5% up to 63% for the automated databases, whereas Swissprot, which performs manual curation, had a misannotation rate close to 0 (Schnoes et al., 2009). A similar misannotation rate due to the automated pipeline is expected for alternative splice forms. Several strategies can be used in order to identify the function of a new alternative splice form. The most common compares the sequence of the transcripts or the isoforms to species already present in the databases using tools like BLAST. The reliability of the annotation depends equally on the quality of the algorithms uses and the correctness of the annotations of species already present in the databases. Although algorithms do create errors in the identification of open reading frames, the database entries themselves might be more problematic, as erronous entries can propagate quickly through automated methods. For example, one of the most used databases (Salzberg, 2007), GenBank, only allows the sequence submitter to correct or update the submitted annotation. This leads to very few corrections and updates thus accumulating errors in a database that shares its entries with several other databases (Salzberg, 2007). In addition, the prediction of function based on the amino acid sequence, taking advantage of massive high-throughput data, is getting more popular. The different tools used by the databases have very different accuracy levels and the characteristics of the annotation tools must be taken into consideration when selecting a reference database.

Transcript databases suitable for metabolic model annotation The GENCODE collaboration (Harrow et al., 2012) tries to annotate genes and splice variants discovered by the ENCODE consortium (The ENCODE Project Consortium and others, 2004) using a combination of manual curation, automated annotation pipelines and targeted validation approaches. Within the GENCODE collaboration, the APPRIS database (Rodriguez et al., 2013) is dedicated to the annotation of principal and alternative splice isoforms. The aim of APPRIS is to validate manually annotated isoforms with functional data and protein structures. APPRIS selects the major isoform that is present in most cells and contexts and compares that isoform to all other isoforms. APPRIS could identify the dominant variants of 85% of the protein coding transcripts of the GENCODE 7 release for ENSEMBL (Hubbard et al., 2002; Flicek et al., 2013).

Vega (Wilming et al., 2008), a database for vertebrate genomes that contains a section with annotations for alternative splicing information is another useful source of transcript information. The HAVANA team is actively participating in these annotation efforts and it was incorporated into the set of ENSEMBL databases (Hubbard et al., 2002; Flicek et al., 2013). The aim is to systematically annotate all experimentally validated ESTs or mRNAs from ENCODE (The ENCODE Project Consortium and others, 2004) and the 1000 Genomes loss-of-function project (The 1000 Genomes Project Consortium and others, 2010), without prior filtering based e.g. on the tissue of origin. This unbiased approach allows the annotation of transcripts that do not yet have an obvious function.

The ASPicDB database (Martelli et al., 2010) considers the human isoforms that result from alternative splicing events. Annotation is then performed by machine-learning approaches that categorize the proteins by function, localization, transmembrane domains, signal peptides, gpi- and coiled-coil domains, and similarity to known peptide sequences. The ADPicDB database employs the ASPIC algorithm (Bonizzoni et al., 2009) to perform multi-alignments to the genome. The alignment that minimizes the splicing events is then retained.

H-DBAS II (Takeda et al., 2010) is the successor of H-DBAS (Takeda et al., 2007), a database that collects information on human alternative splice forms, with the focus on alternative splicing events altering protein functions. The H-DBAS database was mainly based on cDNA libraries. H-DBAS II now takes advantage of the RNA-seq technology to improve the annotation of splicing variants.

The SASD database (Zhang and Drabier, 2013) predicts alternative splice forms expressed in different contexts e.g. during disease, under drug effects, or in different organs. Data extracted from ENSEMBL (Flicek et al., 2013) and from the Integrated Pathway Analysis database is used to create artificial transcripts and peptides.

While all databases mentioned above are focussing on different vertebrates, the ASIP database is specialized to plants (Wang and Brendel, 2006). It allows the visualization of alternative splice forms in plants like *A. thaliana* or *Oryza sativa*. To obtain the annotations the ASIP database uses an automated approach based on alignment tools.

Table B.2 gives an overview of these databases which, along with further information provided

Name	Species	Method of annotation	Reference	Link
GENCODE	human and mouse	manual and automated	(Harrow et al., 2012)	http://www.gencodegenes.org/
ASPicDB	human	automated	(Martelli et al., 2010)	http://srv00.ibbe.cnr.it/ASPicDB/
Vega	human, zebrafish, pig, mouse and rat	manual annotation	(Wilming et al., 2008)	http://vega.sanger.ac.uk/index.html
H-DBAS	(human, mouse, rat, chimpanzee, macaque and dog	manual	(Takeda et al., 2010)	http://www.h-invitational.jp/h-dbas/
SASD	human	prediction	(Zhang and Drabier, 2013)	http://bioinfo.hsc.unt.edu/sasd/
ASIP	plants	automated	(Wang and Brendel, 2006)	http://www.plantgdb.org/

Table B.2. Databases for transcript specific genome annotations of multiple species.

by RNA-seq experiments, represent a valuable source of data that could increase the predictive capabilities of metabolic models. Besides automated pipelines to map the correct transcripts to known metabolic reactions, data mining approaches and bibliographic research similar to those performed by the Recon1 project would be required to unravel the function of the variants. It would, however, be important to use these resources to implement a common nomenclature that would prevent information loss and create consistency between models.

B.6 Non-specific cofactors can cause infeasible loops

Another issue commonly observed when reconstructing metabolic networks is the difficulty of selecting the right cofactors for reactions, specifically the right redox pairs. The assignment of cofactors to reactions is complicated by the fact that the cofactor requirement is organism- and cell-specific, explaining at least partially that the cofactors requirements vary between databases (Radrich et al., 2010). Furthermore, gene matching algorithms used to reconstruct networks will often find reactions using all potential cofactors and include them in the reconstruction. The discrepancies are further accentuated by the fact that in the case of missing electron transfer pair information, NAD^+/ NADH is most often the default transfer cofactor used (Henry et al., 2010). The reason for this default choice is that finding organism-specific information is not trivial and can necessitate extensive literature research even for well studied organisms. Furthermore, several enzymes have different isoformes that do not exhibit the same cofactors requirements. One example is aldehyde dehydrogenases, which may use both NADH and NADPH. In the cytoplasm of *S. cerevisiae*, the main isoform uses NADP^+, whereas stress-induced isoforms prefer NAD^+ as cofactor (Remize et al., 2000). Unfortunately, databases tend to either provide inspecific reactions (using $\mathrm{NAD(P)}^+$), only one variant, or often both variants associated with both genes in these instances, which makes it challenging to assign the correct reaction to the respective isoform. In addition, several enzymes are able to catalyze various reactions and the catalysed reactions depend on the availability of a specific cofactor. This leads to the incorporation of all potentially catalysed reactions that vary only by their cofactor requirements (Förster et al., 2003), which is likely to cause loops or cycles that are thermodynamically infeasible if one or more of the reactions are reversible. Loops carry a non-zero flux, even in the absence of an input and output flux, if no thermodynamical constraints are added. These loops violate the loop law, a law similar to Kirchhoff's second law for electrical circuits. There have been attempts to eliminate the presence of thermodynamically infeasible loops from FBA calculations and it has been shown that their presence can diminish the predictive power of models (Schellenberger et al., 2011b). However, the use of loopless FBA converts the simple linear problem into a mixed integer linear problem which can lead to long computational times, particularly if multiple rounds of the problem have to be solved. Other approaches to solving this issue show similar characteristics with respect to computational requirements (De Martino et al., 2013) and are therefore often not included in the analysis of metabolic models.

B.7 Community efforts to improve metabolic models

There have been attempts to create collections of metabolic networks, e.g. Model SEED (Overbeek et al., 2005) or BiGG (Schellenberger et al., 2011b), and unify identifiers like MetRxn (Kumar et al., 2012) or MetaNetX (Ganter et al., 2013) (listed in Table B.3).

Model SEED is aimed at providing a platform for model reconstruction based on automated genome annotation using RAST (Aziz et al., 2008). While this is sufficient for the analysis tools provided on the website, the exportable model formats lack unification information. They do adhere to the COBRA toolbox standard, but as mentioned earlier, that definition itself lacks a lot of information. BiGG was introduced to allow comparison between different networks, but

Resource	Unification	Description
BiGG (Schellenberger et al., 2011b)	SBML/COBRA	Database containing multiple genome scale metabolic networks in the COBRA format.
BiGG2 (Systems Biology Research Group, 2015)	SBML/COBRA, SBML/FBC	Update to BiGG, currently in a beta version, providing multiple models annotated using FBC.
MetaCyc (Caspi et al., 2014)	SBML/COBRA, biocyc flat files	Large collection of metabolic reconstructions. Flat File format contains additional details not included in the provided SBMLs.
SEED (Overbeek et al., 2005)	SEED IDs, Partial SBML/COBRA format	System for construction of metabolic reconstructions and analysis. Export of reconstructions is available in SBML format (with minimal annotations) and Excel sheets.
MetaNetX (Ganter et al., 2013)	MNXRef IDs, SBML/COBRA, bioql information for metabolites	Repository of unified metabolic reconstructions linking to multiple external databases. Offers tools for network analysis and modifications. SBML files contain additional yeast-style annotations for species.
MetRxn (Kumar et al., 2012)	MetRxn ID, SBML/COBRA	Database matching multiple metabolite and reaction databases aiming at providing a curated basis for network reconstruction.

Table B.3. Databases aiming at providing functional metabolic models that are directly comparable.

relied on all deposited networks adhering to the same nomenclature, and is restricted by the limited number of deposited reconstructions. The database is currently being updated however and a beta version of BiGG2, comprising lots of additional models and providing well annotated models, has recently been made available online.

In contrast to this approach, MetRxn and MetaNetX aim at identifying common reactions by combining multiple pieces of information. (Bernard et al., 2014) give a good overview of the issues arising when trying to match metabolites, and how different databases try to address them. The biggest issues arise from stereoisomers and difference in protonation states. While most often protonation states can be ignored (as long as they are consistent within a model), there might be issues when different compartments exhibit different pH. This could become particularly important for energetic considerations if different protonation states are assumed for mitochondria and cytosol. The same problems can potentially arise from considering equality of stereoisomers, with different stereoisomers being processed at different efficiencies (Abelö et al., 2000). Both MetRxn and MetaNetX can be a great help to overcome most of these issues, with MetaNetX being the more comprehensive approach. Using an extensive set of external databases it tries to match similar external compounds to its namespace. To address issues of stereoisomers and protonation states, it provides a distinction between identical structures, structures with the same tautomeric form at pH 7.3, and inferred similarities. Even though this information is not directly visible on the website, it can be retrieved from the data export files. However useful these tools become, it is even more important that they are actually used, and that the community works in concert to improve models, avoiding the creation of multiple distinct reconstructions for the same organism. While the exchange of models in a common language would be an important step, as it would make the combination of models easier, we also want to highlight two recent collaborative efforts that lead to the development of more comprehensive reconstructions.

The first example of a successful community effort for organism specific reconstruction is the creation of the consensus model of *S. cerevisiae*. Several models of yeast had been published (Förster et al., 2003; Duarte et al., 2004; Kuepfer et al., 2005) until, in 2007, a combined effort was undertaken to merge these models and bring them into a more standardized format (Herrgård et al., 2008). This early combined effort now led the seventh iteration of the model (Aung et al., 2013), which inspired the formulation of GPRs as suggested above.

Another example of community efforts to merge models is the human metabolic reconstruction Recon 2 (Thiele et al., 2013). The first human genome scale metabolic reconstruction, HumanCyc, was published in 2005 (Romero et al., 2005). Soon after, two refined genome-scale reconstructions were published; Recon 1 by (Duarte et al., 2007), and the Edinburgh Human Metabolic Network (EHMN) by (Ma et al., 2007). These competing models along, with HepatoNet (Gille et al., 2010) and further information from the literature, were combined into Recon 2 (Thiele et al., 2013) in an effort to unify the different sources. While the attempt led to a more complete knowledge source, it reinforced the problems of incompatibility between different networks. For example, Recon 1 used Entrez gene identifiers with transcript specific details as gene IDs, while HepatoNet used gene symbols leading to mixed identifiers in Recon 2, which makes simulations more challenging. In addition, the transcript-specific information from Recon 1 got mostly lost since it, unfortunately, was not traceable to databasesSection B.5), and neither EHMN nor HepatoNet contained similar information. This again highlights the importance of linking information to databases since otherwise great efforts can be lost or have to be repeated. Still, Recon 2 is an important step in the development of human metabolic reconstructions and only in its second iteration, and there remains competing reconstructions or knowledgebases like the HMR, which will hopefully be merged in the future.

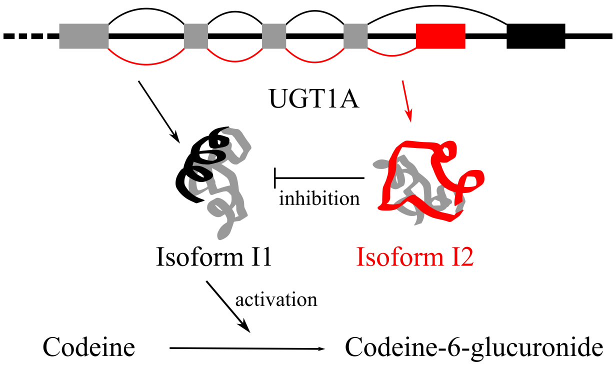


Figure B.1. Alternative splice forms are created by removal and addition of exons during the splicing process. This example shows two the alternate splice forms i1 (depicted in black) and i2 (depicted in red) of a human glucuronosyltransferase (UGT1A). The main isoform, i1, is implicated in the metabolism and excretion of toxic compounds e.g. drugs like codeine while isoform i2 inhibits the activity of the main isoform.

Bibliography

Abelö, A., Andersson, T. B., Antonsson, M., Naudot, A. K., Skånberg, I., and Weidolf, L. Stereoselective metabolism of omeprazole by human cytochrome P450 enzymes. *Drug Metabolism & Disposition.*, 28(8):966–972, 2000. (Cited on page 208.)

AbuOun, M., Suthers, P. F., Jones, G. I., Carter, B. R., Saunders, M. P., Maranas, C. D., Woodward, M. J., and Anjum, M. F. Genome scale reconstruction of a salmonella metabolic model comparison of similarity and differences with a commensal escherichia coli strain. *Journal of Biological Chemistry*, 284(43):29480–29488, 2009. (Cited on page 3.)

Acuña, V., Chierichetti, F., Lacroix, V., Marchetti-Spaccamela, A., Sagot, M., and Stougie, L. Modes and cuts in metabolic networks: Complexity and algorithms. *Biosystems*, 95(1):51–60, 2009. (Cited on pages 36, 37, 38 and 42.)

Adams, C. C. and Workman, J. L. Binding of disparate transcriptional activators to nucleosomal dna is inherently cooperative. *Molecular and Cellular Biology*, 15(3):1405–1421, 1995. (Cited on page 25.)

Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I., and Nielsen, J. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Computational Biology*, 8(5):e1002518, 2012. (Cited on pages xiii, 9, 11, 17, 21, 32, 35, 36, 47, 49, 76, 115, 118, 122, 123, 129 and 199.)

Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., and Nielsen, J. The RAVEN toolbox and its use for generating a genome-scale metabolic model for *Penicillium chrysogenum*. *PLoS Computational Biology*, 9(3):e1002980, 2013. (Cited on pages 196 and 199.)

Agren, R., Mardinoglu, A., Asplund, A., Kampf, C., Uhlen, M., and Nielsen, J. Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Molecular Systems Biology*, 10(3), 2014. (Cited on pages 11, 20, 21, 22, 24, 76, 117 and 122.)

Ajioka, R. S., Phillips, J. D., and Kushner, J. P. Biosynthesis of heme in mammals. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1763(7):723–736, 2006. (Cited on page 139.)

Åkesson, M., Förster, J., and Nielsen, J. Integration of gene expression data into genome-scale metabolic models. *Metabolic Engineering*, 6(4):285–293, 2004. (Cited on pages 122, 123, 129 and 152.)

Alarcón, M. A., Medina, M. A., Hu, Q., Avila, M. E., Bustos, B. I., Pérez-Palma, E., Peralta, A., Salazar, P., Ugarte, G. D., Reyes, A. E., et al. A novel functional low-density lipoprotein receptor-related protein 6 gene alternative splice variant is associated with Alzheimer's disease. *Neurobiology of Aging*, 34(6):1709–e9, 2013. (Cited on page 203.)

Antoniewicz, M. R., Kelleher, J. K., and Stephanopoulos, G. Elementary metabolite units (emu): a novel framework for modeling isotopic distributions. *Metabolic Engineering*, 9(1):68–86, 2007. (Cited on page 17.)

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, 2000. (Cited on page 200.)

Aung, H. W., Henry, S. A., and Walker, L. P. Revising the Representation of Fatty Acid, Glycerolipid, and Glycerophospholipid Metabolism in the Consensus Model of Yeast Metabolism. *Industrial Biotechnology*, 9(4):215–228, 2013. (Cited on pages 9, 196, 198, 199 and 209.)

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., and Zagnitko, O. The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, 9:75, 2008. (Cited on page 207.)

Bailey, J. E. Complex biology with no parameters. *Nature Biotechnology*, 19(6):503–504, 2001. (Cited on page 4.)

Bairoch, A. The ENZYME database in 2000. *Nucleic Acids Research*, 28(1):304–305, Jan 2000. (Cited on page 200.)

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L.-S. L. The Universal Protein Resource (UniProt). *Nucleic Acids Research*, 33 (Database issue):D154–D159, 2005. (Cited on pages 203 and 204.)

Bardy, C., Alonso, M., Bouthour, W., and Lledo, P.-M. How, when, and where new inhibitory neurons release neurotransmitters in the adult olfactory bulb. *The Journal of Neuroscience*, 30(50):17023–17034, 2010. (Cited on page 171.)

Barnes, C. A. Memory deficits associated with senescence: a neurophysiological and behavioral study in the rat. *Journal of Comparative and Physiological Psychology*, 93(1):74, 1979. (Cited on page 175.)

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. Ncbi geo: archive for functional genomics data sets–update. *Nucleic Acids Research*, 41(Database issue):D991–D995, Jan 2013. (Cited on page 118.)

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. The Pfam protein families database. *Nucleic Acids Research*, 30(1):276–280, 2002. (Cited on page 203.)

Baumuratova, T., Dobre, S., Bastogne, T., and Sauter, T. Switch of sensitivity dynamics revealed with dyglosa toolbox for dynamical global sensitivity analysis as an early warning for system's critical transition. *PLoS One*, 8, 2013. (Cited on page 91.)

Becker, S. A. and Palsson, B. Ø. Context-specific metabolic networks are consistent with experiments. *PLoS Computational Biology*, 4(5):e1000082, 2008. (Cited on pages 11, 23, 33, 35, 36, 47, 56, 57, 63, 66, 76, 98, 115, 117, 118, 122, 129, 136, 151, 156, 162 and 199.)

Becker, S. A. and Palsson, B. Ø. Genome-scale reconstruction of the metabolic network in staphylococcus aureus n315: an initial draft to the two-dimensional annotation. *BMC Microbiology*, 5(1):1, 2005. (Cited on page 3.)

Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø., and Herrgard, M. J. Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox. *Nature Protocols*, 2(3):727–738, 2007. (Cited on pages 6 and 182.)

Bellemare, J., Rouleau, M., Harvey, M., and Guillemette, C. Modulation of the human glucuronosyltransferase UGT1A pathway by splice isoform polypeptides is mediated through protein-protein interactions. *The Journal of Biological Chemistry*, 285(6):3600–3607, 2010. (Cited on page 202.)

Bello, B., Reichert, H., and Hirth, F. The brain tumor gene negatively regulates neural progenitor cell proliferation in the larval central brain of drosophila. *Development*, 133(14):2639–2648, 2006. (Cited on page 172.)

Belluzzi, O., Benedusi, M., Ackman, J., and LoTurco, J. J. Electrophysiological differentiation of new neurons in the olfactory bulb. *The Journal of Neuroscience*, 23(32):10411–10418, 2003. (Cited on page 171.)

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B*, 57, 1995. (Cited on page 106.)

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. GenBank: update. *Nucleic Acids Research*, 32(Database issue):D23–D26, Jan 2004. (Cited on page 204.)

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000. (Cited on page 200.)

Bernard, T., Bridge, A., Morgat, A., Moretti, S., Xenarios, I., and Pagni, M. Reconciliation of metabolites and biochemical reactions for metabolic networks. *Briefings in Bioinformatics*, 15 (1):123–135, 2014. (Cited on page 208.)

Bernstein, B. E., Humphrey, E. L., Erlich, R. L., Schneider, R., Bouman, P., Liu, J. S., Kouzarides, T., and Schreiber, S. L. Methylation of histone h3 lys 4 in coding regions of active genes. *Proceedings of the National Academy of Sciences*, 99(13):8695–8700, 2002. (Cited on page 26.)

Beste, D. J., Hooper, T., Stewart, G., Bonde, B., Avignone-Rossa, C., Bushell, M. E., Wheeler, P., Klamt, S., Kierzek, A. M., and McFadden, J. Gsmn-tb: a web-based genome-scale network model of mycobacterium tuberculosis metabolism. *Genome Biology*, 8(5):R89, 2007. (Cited on page 3.)

Betschinger, J. and Knoblich, J. A. Dare to be different: asymmetric cell division in drosophila, c. elegans and vertebrates. *Current Biology*, 14(16):R674–R685, 2004. (Cited on page 172.)

Beyer, K. and Ariza, A. Alpha-synuclein posttranslational modification and alternative splicing as a trigger for neurodegeneration. *Molecular Neurobiology*, 47(2):509–524, 2013. (Cited on page 203.)

Bilofsky, H. S. and Burks, C. The GenBank genetic sequence data bank. *Nucleic Acids Research*, 16(5):1861–1863, Mar 1988. (Cited on page 204.)

Björkhem, I. Mechanism of degradation of the steroid side chain in the formation of bile acids. *The Journal of Lipid Research*, 33, 1992. (Cited on page 91.)

Björkhem, I., Andersson, O., Diczfalusy, U., Sevastik, B., Xiu, R. J., Duan, C., and Lund, E. Atherosclerosis and sterol 27-hydroxylase: evidence for a role of this enzyme in elimination of cholesterol from human macrophages. *Proc Natl Acad Sci*, 91, 1994. (Cited on page 91.)

Björnson, E., Mukhopadhyay, B., Asplund, A., Pristovsek, N., Cinar, R., Romeo, S., Uhlen, M., Kunos, G., Nielsen, J., and Mardinoglu, A. Stratification of hepatocellular carcinoma patients based on acetate utilization. *Cell Reports*, 13(9):2014–2026, 2015. (Cited on page 22.)

Blazier, A. and Papin, J. Integration of expression data in genome-scale metabolic network reconstructions. *Frontiers in Physiology*, 3, 2012. (Cited on pages 36 and 47.)

Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. In Wheeler, R. A. and Spellmeyer, D. C., editors, *Annual Reports in Computational Chemistry*, volume 4 of *Annual Reports in Computational Chemistry*, chapter 12, pages 217 – 241. Elsevier, 2008. (Cited on page 200.)

Bonizzoni, P., Mauri, G., Pesole, G., Picardi, E., Pirola, Y., and Rizzi, R. Detecting alternative gene structures from spliced ESTs: a computational approach. *Journal of Computational Biology*, 16(1):43–66, 2009. (Cited on page 205.)

Bordbar, A., Lewis, N., Schellenberger, J., Palsson, B., and Jamshidi, N. Insight into human alveolar macrophage and m. tuberculosis interactions via metabolic reconstructions. *Molecular Systems Biology*, 6(1), 2010. (Cited on pages 23, 67, 69 and 103.)

Bordbar, A., Mo, M., Nakayasu, E., Schrimpe-Rutledge, A., Kim, Y., Metz, T., Jones, M., Frank, B., Smith, R., Peterson, S., et al. Model-driven multi-omic data analysis elucidates metabolic immunomodulators of macrophage activation. *Molecular Systems Biology*, 8(1), 2012. (Cited on pages 10, 11, 23, 36 and 56.)

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004. (Cited on pages 40 and 45.)

Brenner, S. Sequences and consequences. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365(1537):207–212, 2010. (Cited on page 2.)

Breton-Provencher, V. and Saghatelyan, A. Newborn neurons in the adult olfactory bulb: unique properties for specific odor behavior. *Behavioural Brain Research*, 227(2):480–489, 2012. (Cited on pages 172 and 191.)

Brown, M. S. and Goldstein, J. L. *A receptor-mediated pathway for cholesterol homeostasis*. 1985. (Cited on page 97.)

Buck, K., Vanek, M., Groner, B., and Ball, R. K. Multiple forms of prolactin receptor messenger ribonucleic acid are specifically expressed and regulated in murine tissues and the mammary cell line HC11. *Endocrinology*, 130(3):1108–1114, 1992. PMID: 1537278. (Cited on page 201.)

Bulger, M. and Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell*, 144(3):327–339, 2011. (Cited on page 26.)

Cairns, R. A., Harris, I. S., and Mak, T. W. Regulation of cancer cell metabolism. *Nature Reviews Cancer*, 11, 2011. (Cited on page 75.)

Calarco, J. A., Zhen, M., and Blencowe, B. J. Networking in a global world: Establishing functional connections between neural splicing regulators and their target transcripts. *RNA*, 17 (5):775–791, 2011. (Cited on page 203.)

Cali, J. J., Hsieh, C. L., Francke, U., and Russell, D. W. Mutations in the bile acid biosynthetic enzyme sterol 27-hydroxylase underlie cerebrotendinous xanthomatosis. *J Biol Chem*, 266, 1991. (Cited on page 95.)

Calo, E. and Wysocka, J. Modification of enhancer chromatin: What, how, and why? *Mol Cell*, 49, 2013. (Cited on page 75.)

Carey, M. The enhanceosome and transcriptional synergy. *Cell*, 92(1):5–8, 1998. (Cited on page 26.)

Carleton, A., Petreanu, L. T., Lansford, R., Alvarez-Buylla, A., and Lledo, P.-M. Becoming a new neuron in the adult olfactory bulb. *Nature Neuroscience*, 6(5):507–518, 2003. (Cited on page 171.)

Caspi, R., Altman, T., Dale, J. M., Dreher, K., Fulcher, C. A., Gilham, F., Kaipa, P., Karthikeyan, A. S., Kothari, A., and Krummenacker, M. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 38, 2010. (Cited on page 76.)

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Weerasinghe, D., Zhang, P., and Karp, P. D. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 42(Database issue):D459–D471, 2014. (Cited on pages 8, 198, 200 and 208.)

Chandrasekaran, S. and Price, N. Probabilistic integrative modeling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, 107(41):17845–17850, 2010. (Cited on page 36.)

Chang, R., Xie, L., Xie, L., Bourne, P., and Palsson, B. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Computational Biology*, 6(9):e1000938, 2010. (Cited on page 35.)

Chapuy, B., McKeown, M. R., Lin, C. Y., Monti, S., Roemer, M. G., Qi, J., Rahl, P. B., Sun, H. H., Yeda, K. T., Doench, J. G., et al. Discovery and characterization of super-enhancer-associated dependencies in diffuse large b cell lymphoma. *Cancer Cell*, 24(6):777–790, 2013. (Cited on page 26.)

Chen, J. and Weiss, W. Alternative splicing in cancer: implications for biology and therapy. *Oncogene*, 34(1):1–14, 2015. (Cited on page 203.)

Cheung, C. Y. M., Williams, T. C. R., Poolman, M. G., Fell, D. A., Ratcliffe, R. G., and Sweetlove, L. J. A method for accounting for maintenance costs in flux balance analysis improves the prediction of plant cell metabolic phenotypes under stress conditions. *Plant Journal*, 2013. (Cited on page 197.)

Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K.-Y. A., Huang, J., et al. Homozygosity mapping with snp arrays identifies trim32, an e3 ubiquitin ligase, as a bardet–biedl syndrome gene (bbs11). *Proceedings of the National Academy of Sciences*, 103(16):6287–6292, 2006. (Cited on page 172.)

Christian, N., May, P., Kempa, S., Handorf, T., and Ebenhöh, O. An integrative approach towards completing genome-scale metabolic networks. *Molecular BioSystems*, 5(12):1889–1903, 2009. (Cited on page 35.)

Chvátal, V. A greedy heuristic for the set-covering problem. *Mathematics of operations research*, 4(3):233–235, 1979. (Cited on page 45.)

Colijn, C., Brandes, A., Zucker, J., Lun, D. S., Weiner, B., Farhat, M. R., Cheng, T.-Y., Moody, D. B., Murray, M., and Galagan, J. E. Interpreting expression data with metabolic flux models: predicting Mycobacterium tuberculosis mycolic acid production. *PLoS Computational Biology*, 5(8):e1000489, August 2009. ISSN 1553-7358. (Cited on pages 125, 152, 156 and 162.)

Consortium, E. P. et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012. (Cited on page 26.)

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., et al. Histone h3k27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50):21931–21936, 2010. (Cited on pages 26, 27, 75, 84 and 163.)

Dandekar, T., Fieselmann, A., Majeed, S., and Ahmed, Z. Software applications toward quantitative metabolic flux analysis and modeling. *Briefings in Bioinformatics*, 15(1):91–107, Jan 2014. (Cited on page 199.)

De La Grange, P., Dutertre, M., Martin, N., and Auboeuf, D. FAST DB: a website resource for the study of the expression regulation of human gene products. *Nucleic Acids Research*, 33 (13):4276–4284, 2005. (Cited on page 203.)

De Martino, D., Capuani, F., Mori, M., De Martino, A., and Marinari, E. Counting and correcting thermodynamically infeasible flux cycles in genome-scale metabolic networks. *Metabolites*, 3 (4):946–966, 2013. (Cited on page 207.)

de Oliveira Dal'Molin, C. G., Quek, L.-E., Palfreyman, R. W., Brumbley, S. M., and Nielsen, L. K. AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiology*, 152(2):579–589, 2010. (Cited on pages 9 and 196.)

DeBerardinis, R. and Thompson, C. Cellular metabolism and disease: what do metabolic outliers teach us? *Cell*, 148(6):1132–1144, 2012. (Cited on page 35.)

Dias, O., Pereira, R., Gombert, A. K., Ferreira, E. C., and Rocha, I. iOD907, the first genome-scale metabolic model for the milk yeast Kluyveromyces lactis. *Biotechnol Journal*, 9(6): 776–790, Jun 2014. (Cited on page 199.)

Djerbi, M., Darreh-Shori, T., Zhivotovsky, B., and Grandien, A. Characterization of the Human FLICE-Inhibitory Protein Locus and Comparison of the Anti-Apoptotic Activity of Four Different FLIP Isoforms. *Scandinavian Journal of Immunology*, 54(1-2):180–189, 2001. ISSN 1365-3083. (Cited on page 201.)

Doetsch, F., Garcia-Verdugo, J. M., and Alvarez-Buylla, A. Cellular composition and three-dimensional organization of the subventricular germinal zone in the adult mammalian brain. *The Journal of Neuroscience*, 17(13):5046–5061, 1997. (Cited on page 171.)

Doetsch, F., Caille, I., Lim, D. A., García-Verdugo, J. M., and Alvarez-Buylla, A. Subventricular zone astrocytes are neural stem cells in the adult mammalian brain. *Cell*, 97(6):703–716, 1999. (Cited on page 170.)

Dowen, J. M., Fan, Z. P., Hnisz, D., Ren, G., Abraham, B. J., Zhang, L. N., Weintraub, A. S., Schuijers, J., Lee, T. I., Zhao, K., and Young, R. A. Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell*, 159, 2014. (Cited on page 76.)

Dreyfuss, J. M. *Algorithms for reconstruction and analysis of metabolic networks, with an application to Neurospora crassa*. PhD thesis, BOSTON UNIVERSITY, 2014. (Cited on page 42.)

Duarte, N. C., Herrgård, M. J., and Palsson, B. Ø. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Research*, 14(7):1298–1309, 2004. (Cited on page 209.)

Duarte, N., Becker, S., Jamshidi, N., Thiele, I., Mo, M., Vo, T., Srivas, R., and Palsson, B. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777–1782, 2007. (Cited on pages xiii, 9, 32, 35, 51, 76, 121, 158, 196, 202 and 209.)

Edgar, R., Domrachev, M., and Lash, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002. (Cited on page 122.)

Edwards, J. S., Covert, M., and Palsson, B. Metabolic modelling of microbes: the flux-balance approach. *Environmental microbiology*, 4(3):133–140, 2002. (Cited on page 7.)

Ellis, J. D., Barrios-Rodiles, M., ï£¡olak, R., Irimia, M., Kim, T., Calarco, J. A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P. M., Wrana, J. L., and Blencowe, B. J. Tissue-Specific Alternative Splicing Remodels Protein-Protein Interaction Networks. *Molecular Cell*, 46(6):884 – 892, 2012. ISSN 1097-2765. (Cited on page 203.)

Ernst, A., Alkass, K., Bernard, S., Salehpour, M., Perl, S., Tisdale, J., Possnert, G., Druid, H., and Frisén, J. Neurogenesis in the striatum of the adult human brain. *Cell*, 156(5):1072–1083, 2014. (Cited on page 171.)

Escher, G., Krozowski, Z., Croft, K. D., and Sviridov, D. Expression of sterol 27-hydroxylase (cyp27a1) enhances cholesterol efflux. *J Biol Chem*, 278, 2003. (Cited on page 95.)

Estévez, S. R. and Nikoloski, Z. Generalized framework for context-specific metabolic model extraction methods. *Frontiers in plant science*, 5, 2014. (Cited on pages 10 and 12.)

Estévez, S. R. and Nikoloski, Z. Context-specific metabolic model extraction based on regularized least squares optimization. *PloS one*, 10(7):e0131875, 2015. (Cited on pages 156 and 162.)

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V., and Palsson, B. O. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Molecular Systems Biology*, 3:121, 2007. (Cited on pages 3 and 200.)

Filiou, M. D., Zhang, Y., Teplytska, L., Reckow, S., Gormanns, P., Maccarrone, G., Frank, E., Kessler, M. S., Hambsch, B., Nussbaumer, M., et al. Proteomics and metabolomics analysis of a trait anxiety mouse model reveals divergent mitochondrial pathways. *Biological psychiatry*, 70(11):1074–1082, 2011. (Cited on page 192.)

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L., et al. The Pfam protein families database. *Nucleic Acids Research*, 36(suppl 1):D281–D288, 2008. (Cited on page 203.)

Flahaut, N. A. L., Wiersma, A., van de Bunt, B., Martens, D. E., Schaap, P. J., Sijtsma, L., Dos Santos, V. A. M., and de Vos, W. M. Genome-scale metabolic model for Lactococcus lactis MG1363 and its application to the analysis of flavor formation. *Applied Microbiology and Biotechnology*, 97(19):8729–8739, Oct 2013. (Cited on page 200.)

Fleming, R., Maes, C., Ye, Y., Saunders, M., and Palsson, B. A variational principle for computing nonequilibrium fluxes and potentials in genome-scale biochemical networks. *Journal of Theoretical Biology*, 292:71–77, 2012. (Cited on page 36.)

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., García-Girón, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kähäri, A. K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W. M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T. J. P., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A., and Searle, S. M. J. Ensembl 2013. *Nucleic Acids Research*, 41(Database issue):D48–D55, 2013. (Cited on pages 203 and 205.)

Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., and Shlomi, T. Predicting selective drug targets in cancer through metabolic networks. *Molecular Systems Biology*, 7(501), 2011. (Cited on pages 10, 20, 36, 57, 63, 65, 66, 69, 76, 98, 102, 103, 119 and 156.)

Fondi, M. and Liò, P. Multi-omics and metabolic modelling pipelines: challenges and tools for systems microbiology. *Microbiological research*, 171:52–64, 2015. (Cited on page 20.)

Förster, J., Famili, I., Fu, P., Palsson, B. Ø., and Nielsen, J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research*, 13(2):244–253, 2003. (Cited on pages 9, 196 and 207.)

Förster, J., Famili, I., Palsson, B. Ø., and Nielsen, J. Large-scale evaluation of in silico gene deletions in *Saccharomyces cerevisiae*. *OMICS*, 7(2):193–202, 2003. (Cited on page 209.)

Frezza, C., Zheng, L., Folger, O., Rajagopalan, K. N., MacKenzie, E. D., Jerby, L., Micaroni, M., Chaneton, B., Adam, J., Hedley, A., et al. Haem oxygenase is synthetically lethal with the tumour suppressor fumarate hydratase. *Nature*, 477(7363):225–228, 2011. (Cited on page 20.)

Frosk, P., Weiler, T., Nylen, E., Sudha, T., Greenberg, C. R., Morgan, K., Fujiwara, T. M., and Wrogemann, K. Limb-girdle muscular dystrophy type 2h associated with mutation in trim32, a putative e3-ubiquitin–ligase gene. *The American Journal of Human Genetics*, 70(3):663–672, 2002. (Cited on page 172.)

Frosk, P., Greenberg, C. R., Tennese, A. A., Lamont, R., Nylen, E., Hirst, C., Frappier, D., Roslin, N. M., Zaik, M., Bushby, K., et al. The most common mutation in fkrp causing limb girdle muscular dystrophy type 2i (lgmd2i) may have occurred only once and is present in hutterites and other populations. *Human mutation*, 25(1):38–44, 2005. (Cited on page 172.)

Gage, F. H. Mammalian neural stem cells. *Science*, 287(5457):1433–1438, 2000. (Cited on page 170.)

Gagneur, J. and Klamt, S. Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics*, 5(1):175, 2004. (Cited on pages 36, 37 and 44.)

Galhardo, M., Berninger, P., Nguyen, T.-P., Sauter, T., and Sinkkonen, L. Cell type-selective disease-association of genes under high regulatory load. *Nucleic Acids Research*. (Cited on page 96.)

Galhardo, M., Sinkkonen, L., Berninger, P., Lin, J., Sauter, T., and Heinäniemi, M. Integrated analysis of transcript-level regulation of metabolism reveals disease-relevant nodes of the human metabolic network. *Nucleic Acids Research*, 42, 2014. (Cited on pages 84, 95 and 164.)

Ganter, M., Bernard, T., Moretti, S., Stelling, J., and Pagni, M. MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics*, 29 (6):815–816, 2013. (Cited on pages 152, 197, 207 and 208.)

Gatto, F., Miess, H., Schulze, A., and Nielsen, J. Flux balance analysis predicts essential genes in clear cell renal cell carcinoma metabolism. *Scientific reports*, 5, 2015. (Cited on page 67.)

Ge, X., Yamamoto, S., Tsutsumi, S., Midorikawa, Y., Ihara, S., Wang, S. M., and Aburatani, H. Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics*, 86(2):127–141, 2005. (Cited on page 126.)

Gille, C., Bölling, C., Hoppe, A., Bulik, S., Hoffmann, S., Hübner, K., Karlstädt, A., Ganeshan, R., König, M., Rother, K., et al. Hepatonet1: a comprehensive metabolic reconstruction of the human hepatocyte for the analysis of liver physiology. *Molecular Systems Biology*, 6(1): 411, 2010. (Cited on pages 9, 23, 115, 120, 121, 126, 127, 136, 148, 157, 198, 199, 200 and 209.)

Girard, H., Lévesque, E., Bellemare, J., Journault, K., Caillier, B., and Guillemette, C. Genetic diversity at the UGT1 locus is amplified by a novel 3' alternative splicing mechanism leading to nine additional UGT1A proteins that act as regulators of glucuronidation activity. *Pharmacogenetics and Genomics*, 17(12):1077–1089, 2007. (Cited on page 202.)

Gluckman, P. D., Hanson, M. A., Buklijas, T., Low, F. M., and Beedle, A. S. Epigenetic mechanisms that underpin metabolic and cardiovascular diseases. *Nat Rev Endocrinol*, 5, 2009. (Cited on page 75.)

Goldstein, I., Yizhak, K., Madar, S., Goldfinger, N., Ruppin, E., and Rotter, V. p53 promotes the expression of gluconeogenesis-related genes and enhances hepatic glucose production. *Cancer Metab*, 1(9), 2013. (Cited on page 21.)

Gonzàlez-Porta, M., Frankish, A., Rung, J., Harrow, J., and Brazma, A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology*, 14:R70, 2013. (Cited on page 202.)

Gormanns, P., Mueller, N. S., Ditzen, C., Wolf, S., Holsboer, F., and Turck, C. W. Phenome-transcriptome correlation unravels anxiety and depression related pathways. *Journal of psychiatric research*, 45(7):973–979, 2011. (Cited on pages 192 and 193.)

Gras, G., Porcheray, F., Samah, B., and Leone, C. The glutamate-glutamine cycle as an inducible, protective face of macrophage activation. *J Leukoc Biol*, 80, 2006. (Cited on page 97.)

Gudmundsson, S. and Thiele, I. Computationally efficient flux variability analysis. *BMC Bioinformatics*, 11(1):489, 2010. (Cited on pages xxvii, 39 and 51.)

Halaris, A. Inflammation, heart disease, and depression. *Curr Psychiatry Rep.*, 400, 2013. (Cited on page 97.)

Hammell, C. M., Lubin, I., Boag, P. R., Blackwell, T. K., and Ambros, V. nhl-2 modulates microrna activity in caenorhabditis elegans. *Cell*, 136(5):926–938, 2009. (Cited on page 172.)

Hao, T., Ma, H., Zhao, X., and Goryanin, I. Compartmentalization of the Edinburgh human metabolic network. *BMC Bioinformatics*, 11(1):393, 2010. (Cited on page 35.)

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, 22(9):1760–1774, 2012. (Cited on pages 204 and 206.)

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., and Steinbeck, C. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research*, 41(Database issue):D456–D463, 2013. (Cited on page 200.)

Hatzimanikatis, V., Li, C., Ionita, J. A., Henry, C. S., Jankowski, M. D., and Broadbelt, L. J. Exploring the diversity of complex metabolic networks. 21(8):1603–1609, 2005. (Cited on page 155.)

Heinemann, M., Kümmel, A., Ruinatscha, R., and Panke, S. In silico genome-scale reconstruction and validation of the staphylococcus aureus metabolic network. *Biotechnology and Bioengineering*, 92(7):850–864, 2005. (Cited on page 3.)

Heinken, A., Sahoo, S., Fleming, R. M., and Thiele, I. Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut microbes*, 4(1):28–40, 2013. (Cited on pages 168 and 182.)

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, 2009. (Cited on pages 26, 27 and 163.)

Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnology*, 28(9):977–982, 2010. ISSN 1087-0156. (Cited on pages 152 and 206.)

Herrgård, M. J., Swainston, N., Dobson, P., Dunn, W. B., Arga, K. Y., Arvas, M., Blüthgen, N., Borger, S., Costenoble, R., Heinemann, M., Hucka, M., Le Novère, N., Li, P., Liebermeister, W., Mo, M. L., Oliveira, A. P., Petranovic, D., Pettifer, S., Simeonidis, E., Smallbone, K., Spasić, I., Weichart, D., Brent, R., Broomhead, D. S., Westerhoff, H. V., Kirdar, B., Penttilä, M., Klipp, E., Palsson, B. Ø., Sauer, U., Oliver, S. G., Mendes, P., Nielsen, J., and Kell, D. B. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nature Biotechnology*, 26(10):1155–1160, 2008. (Cited on page 209.)

Hiller, K. and Metallo, C. Profiling metabolic networks to study cancer metabolism. *Current Opinion in Biotechnology*, 24:60–68, 2013. (Cited on page 35.)

Hiller, K., Hangebrauk, J., JaÌĹger, C., Spura, J., Schreiber, K., and Schomburg, D. Metabolitedetector: comprehensive analysis tool for targeted and nontargeted gc/ms based metabolome analysis. *Analytical chemistry*, 81(9):3429–3439, 2009. (Cited on page 181.)

Hillje, A.-L., Worlitzer, M., Palm, T., and Schwamborn, J. C. Neural stem cells maintain their stemness through protein kinase c ζ-mediated inhibition of trim32. *Stem Cells*, 29(9):1437–1447, 2011. (Cited on pages 172 and 182.)

Hillje, A.-L., Pavlou, M., Beckmann, E., Worlitzer, M., Bahnassawy, L., Lewejohann, L., Palm, T., and Schwamborn, J. C. Trim32-dependent transcription in adult neural progenitor cells regulates neuronal differentiation. *Cell death & disease*, 4(12):e976, 2013. (Cited on pages xiv, 172, 177, 182, 183, 188, 191 and 193.)

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., and Young, R. A. Super-enhancers in the control of cell identity and disease. *Cell*, 155, 2013. (Cited on pages 26, 75, 84, 95 and 163.)

Holste, D., Huo, G., Tung, V., and Burge, C. B. HOLLYWOOD: a comparative relational database of alternative splicing. *Nucleic Acids Research*, 34(suppl 1):D56–D62, 2006. (Cited on page 203.)

Horn, E. J., Albor, A., Liu, Y., El-Hizawi, S., Vanderbeek, G. E., Babcock, M., Bowden, G. T., Hennings, H., Lozano, G., Weinberg, W. C., et al. Ring protein trim32 associated with skin carcinogenesis has anti-apoptotic and e3-ubiquitin ligase properties. *Carcinogenesis*, 25(2): 157–167, 2004. (Cited on page 172.)

Huang, D. W., Sherman, B. T., and Lempicki, R. A. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc*, 4, 2009. (Cited on page 83.)

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Clamp, M. The Ensembl genome database project. *Nucleic Acids Research*, 30(1):38–41, 2002. (Cited on pages 203 and 205.)

Huber, R., Pietsch, D., Günther, J., Welz, B., Vogt, N., and Brand, K. Regulation of monocyte differentiation by specific signaling modules and associated transcription factor networks. *Cell Mol Life Sci*, 71, 2014. (Cited on page 81.)

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J.-H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Novère, N. L., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J., and Forum, S. B. M. L. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003. (Cited on page 197.)

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4, 2003. (Cited on page 106.)

Ishibashi, Y., Kohyama-Koganeya, A., and Hirabayashi, Y. New insights on glucosylated lipids: Metabolism and functions. *Biochim Biophys Acta*, 1831, 2013. (Cited on page 96.)

Ishunina, T. A. and Swaab, D. F. Decreased alternative splicing of estrogen receptor-α mRNA in the Alzheimer's disease brain. *Neurobiology of Aging*, 33(2):286–296, 2012. (Cited on page 203.)

Jain, M., Nilsson, R., Sharma, S., Madhusudhan, N., Kitami, T., Souza, A. L., Kafri, R., Kirschner, M. W., Clish, C. B., and Mootha, V. K. Metabolite profiling identifies a key role for glycine in rapid cancer cell proliferation. *Science*, 336(6084):1040–1044, 2012. (Cited on pages 68 and 163.)

Jamshidi, N. and Palsson, B. Ø. Investigating the metabolic capabilities of mycobacterium tuberculosis h37rv using the in silico strain inj 661 and proposing alternative drug targets. *BMC Systems Biology*, 1(1):1–20, 2007. ISSN 1752-0509. (Cited on pages 3 and 198.)

Jensen, P. A. and Papin, J. A. Functional integration of a metabolic network model and expression data without arbitrary thresholding. *Bioinformatics*, 27(4):541–547, 2011. (Cited on pages 36, 76, 152, 156 and 162.)

Jerby, L. and Ruppin, E. Predicting drug targets and biomarkers of cancer via genome-scale metabolic modeling. *Clinical Cancer Research*, 18(20):5572–5584, 2012. (Cited on page 57.)

Jerby, L., Shlomi, T., and Ruppin, E. Computational reconstruction of tissue-specific metabolic models: Application to human liver metabolism. *Molecular Systems Biology*, 6(401), 2010. (Cited on pages xxvii, 10, 11, 21, 23, 33, 35, 36, 37, 39, 47, 49, 51, 52, 53, 55, 56, 66, 76, 115, 117, 119, 120, 122, 152, 155, 161 and 199.)

Jerby, L., Wolf, L., Denkert, C., Stein, G. Y., Hilvo, M., Oresic, M., Geiger, T., and Ruppin, E. Metabolic associations of reduced proliferation and oxidative stress in advanced breast cancer. *Cancer research*, 72(22):5712–5720, 2012. (Cited on page 21.)

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes,

T. R., Lemaire, P., Ukkonen, E., Kivioja, T., and Taipale, J. Dna-binding specificities of human transcription factors. *Cell*, 152, 2013. (Cited on page 110.)

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G., Wu, G., Matthews, L., et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(suppl 1):D428–D432, 2005. (Cited on page 8.)

Julius, A., Imielinski, M., and Pappas, G. Metabolic networks analysis using convex optimization. In *47th IEEE Conference on Decision and Control*, pages 762–767, 2008. (Cited on page 45.)

Kaddurah-Daouk, R., Yuan, P., Boyle, S. H., Matson, W., Wang, Z., Zeng, Z. B., Zhu, H., Dougherty, G. G., Yao, J. K., Chen, G., et al. Cerebrospinal fluid metabolome in mood disorders-remission state has a unique metabolic profile. *Scientific reports*, 2, 2012. (Cited on page 193.)

Kalsotra, A. and Cooper, T. A. Functional consequences of developmentally regulated alternative splicing. *Nature Reviews Genetics*, 12(10):715–729, 2011. (Cited on page 203.)

Kanehisa, M. and Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, Jan 2000. (Cited on page 204.)

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(suppl 1):D355–D360, 2010. (Cited on page 8.)

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Research*, 42 (Database issue):D199–D205, 2014. (Cited on page 200.)

Kano, S., Miyajima, N., Fukuda, S., and Hatakeyama, S. Tripartite motif protein 32 facilitates cell growth and migration via degradation of abl-interactor 2. *Cancer Research*, 68(14):5572–5580, 2008. (Cited on page 172.)

Karp, P. D., Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I. M., and Caspi, R. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, 11(1):40–79, Jan 2010. (Cited on page 199.)

Kaufman, D. E. and Smith, R. L. Direction choice for accelerated convergence in hit-and-run sampling. *Operations Research*, 46(1):84–95, 1998. (Cited on page 7.)

Kelemen, O., Convertini, P., Zhang, Z., Wen, Y., Shen, M., Falaleeva, M., and Stamm, S. Function of alternative splicing. *Gene*, 514(1):1–30, 2013. (Cited on page 203.)

Keseler, I. M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martínez, C., Fulcher, C., Huerta, A. M., Kothari, A., Krummenacker, M., Latendresse, M., Muñiz-Rascado, L., Ong, Q., Paley, S., Schröder, I., Shearer, A. G., Subhraveti, P., Travers, M., Weerasinghe, D., Weiss, V., Collado-Vides, J., Gunsalus, R. P., Paulsen, I., and Karp, P. D. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Research*, 41(Database issue):D605–D612, 2013. (Cited on pages 9 and 196.)

Kim, H. U., Kim, T. Y., and Lee, S. Y. Genome-scale metabolic network analysis and drug targeting of multi-drug resistant pathogen acinetobacter baumannii aye. *Molecular BioSystems*, 6(2):339–348, 2010. (Cited on page 20.)

Kim, J., Reed, J. L., et al. Relatch: relative optimality in metabolic networks explains robust metabolic and regulatory responses to perturbations. *Genome Biol*, 13(9):R78, 2012a. (Cited on pages 152 and 156.)

Kim, N., Alekseyenko, A. V., Roy, M., and Lee, C. The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Research*, 35 (suppl 1):D93–D98, 2007. (Cited on page 203.)

Kim, T. Y., Sohn, S. B., Kim, Y. B., Kim, W. J., and Lee, S. Y. Recent advances in reconstruction and applications of genome-scale metabolic models. *Current Opinion in Biotechnology*, 23 (4):617–623, 2012b. (Cited on pages 162 and 196.)

Kitano, H. Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. *Current genetics*, 41(1):1–10, 2002. (Cited on page 2.)

Klipp, E., Herwig, R., Kowald, A., Wierling, C., and Lehrach, H. Systems biology in practice: Concepts, implementation and application. *Hoboken, New Jersey: John Wiley & Sons*, 2008. (Cited on pages 96 and 164.)

Koscielny, G., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Riethoven, J.-J., Nardone, F., Stanley, E., Fallsehr, C., Hofmann, O., Kull, M., Harrington, E., Boué, S., Eyras, E., Plass, M., Lopez, F., Ritchie, W., Moucadel, V., Ara, T., Pospisil, H., Herrmann, A., G Reich, J., Guigó, R., Bork, P., Doeberitz, M. v. K., Vilo, J., Hide, W., Apweiler, R., Thanaraj, T. A., and Gautheret, D. ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics*, 93(3):213–220, 2009. (Cited on page 203.)

Kudryashova, E., Kudryashov, D., Kramerova, I., and Spencer, M. J. Trim32 is a ubiquitin ligase mutated in limb girdle muscular dystrophy type 2h that binds to skeletal muscle myosin and ubiquitinates actin. *Journal of molecular biology*, 354(2):413–424, 2005. (Cited on page 172.)

Kudryashova, E., Wu, J., Havton, L. A., and Spencer, M. J. Deficiency of the e3 ubiquitin ligase trim32 in mice leads to a myopathy with a neurogenic component. *Human molecular genetics*, 18(7):1353–1367, 2009. (Cited on pages 173 and 174.)

Kudryashova, E., Kramerova, I., and Spencer, M. J. Satellite cell senescence underlies myopathy in a mouse model of limb-girdle muscular dystrophy 2h. *The Journal of clinical investigation*, 122(5):1764–1776, 2012. (Cited on page 172.)

Kuepfer, L., Sauer, U., and Blank, L. M. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Research*, 15(10):1421–1430, 2005. (Cited on page 209.)

Kuhn, H. G., Dickinson-Anson, H., and Gage, F. H. Neurogenesis in the dentate gyrus of the adult rat: age-related decrease of neuronal progenitor proliferation. *The Journal of Neuroscience*, 16(6):2027–2033, 1996. (Cited on page 171.)

Kumar, A., Suthers, P. F., and Maranas, C. D. MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinformatics*, 13:6, 2012. (Cited on pages 152, 153, 196, 197, 207 and 208.)

Kumar, V. S., Dasika, M. S., and Maranas, C. D. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 8(1):1, 2007. (Cited on page 155.)

Ladd, A. N. and Cooper, T. A. Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biology*, 3(11):1–16, 2002. (Cited on page 201.)

Lakshmanan, M., Koh, G., Chung, B. K. S., and Lee, D.-Y. Software applications for flux balance analysis. *Briefings in Bioinformatics.*, 15(1):108–122, Jan 2014. (Cited on page 199.)

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10, 2009. (Cited on page 109.)

Larocque, M., Chénard, T., and Najmanovich, R. A curated C. difficile strain 630 metabolic network: prediction of essential targets and inhibitors. *BMC Systems Biology*, 8:117, 2014. (Cited on page 199.)

Lazarini, F. and Lledo, P.-M. Is adult neurogenesis essential for olfaction? *Trends in Neurosciences*, 34(1):20–30, 2011. (Cited on page 172.)

Le Novère, N., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., Crampin, E. J., Halstead, M., Klipp, E., Mendes, P., Nielsen, P., Sauro, H., Shapiro, B., Snoep, J. L., Spence, H. D., and Wanner, B. L. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nature Biotechnology*, 23(12):1509–1515, 2005. (Cited on page 200.)

Lee, B., Yu, H., Jahoor, F., O'Brien, W., Beaudet, A. L., and Reeds, P. In vivo urea cycle flux distinguishes and correlates with phenotypic severity in disorders of the urea cycle. *Proceedings of the National Academy of Sciences*, 97(14):8021–8026, 2000. (Cited on pages 55 and 56.)

Lee, C.-Y., Robinson, K. J., and Doe, C. Q. Lgl, pins and apkc regulate neuroblast self-renewal versus differentiation. *Nature*, 439(7076):594–598, 2006. (Cited on page 172.)

Lee, C.-K., Shibata, Y., Rao, B., Strahl, B. D., and Lieb, J. D. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics*, 36(8):900–905, 2004. (Cited on page 26.)

Lee, D., Smallbone, K., Dunn, W., Murabito, E., Winder, C., Kell, D., Mendes, P., and Swainston, N. Improving metabolic flux predictions using absolute gene expression data. *BMC Systems Biology*, 6(1):73, 2012. ISSN 1752-0509. (Cited on pages 118, 125, 129, 152, 156 and 162.)

Lee, D.-S., Burd, H., Liu, J., Almaas, E., Wiest, O., Barabási, A.-L., Oltvai, Z. N., and Kapatral, V. Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple staphylococcus aureus genomes identify novel antimicrobial drug targets. *Journal of Bacteriology*, 191(12):4015–4024, 2009. (Cited on page 3.)

Lee, Y., Lee, Y., Kim, B., Shin, Y., Nam, S., Kim, P., Kim, N., Chung, W.-H., Kim, J., and Lee, S. ECgene: an alternative splicing database update. *Nucleic Acids Research*, 35(suppl 1): D99–D103, 2007. (Cited on page 203.)

Legendre, P. The glycinergic inhibitory synapse. *Cellular and Molecular Life Sciences CMLS*, 58(5-6):760–793, 2001. (Cited on page 193.)

Lelli, K. M., Slattery, M., and Mann, R. S. Disentangling the many layers of eukaryotic transcriptional regulation. *Annual review of genetics*, 46:43, 2012. (Cited on page 26.)

Lewejohann, L., Skryabin, B., Sachser, N., Prehn, C., Heiduschka, P., Thanos, S., Jordan, U., Dell Omo, G., Vyssotski, A., Pleskacheva, M., et al. Role of a neuronal small non-messenger rna: behavioural alterations in bc1 rna-deleted mice. *Behavioural Brain Research*, 154(1): 273–289, 2004. (Cited on page 174.)

Lewis, N. E., Hixson, K. K., Conrad, T. M., Lerman, J. A., Charusanti, P., Polpitiya, A. D., Adkins, J. N., Schramm, G., Purvine, S. O., Lopez-Ferrer, D., Weitz, K. K., Eils, R., König, R., Smith, R. D., and Palsson, B. Ø. Omic data from evolved e. coli are consistent with computed optimal growth from genome-scale models. *Mol Syst Biol*, 6, 2010a. (Cited on page 75.)

Lewis, N., Schramm, G., Bordbar, A., Schellenberger, J., Andersen, M., Cheng, J., Patel, N., Yee, A., Lewis, R., Eils, R., et al. Large-scale in silico modeling of metabolic interactions between cell types in the human brain. *Nature Biotechnology*, 28(12):1279–1285, 2010b. (Cited on pages 35 and 202.)

Li, C., Courtot, M., Le Novère, N., and Laibe, C. BioModels.net Web Services, a free and integrated toolkit for computational modelling software. *Briefings in Bioinformatics*, 11(3): 270–277, 2010. (Cited on page 199.)

Licatalosi, D. D. and Darnell, R. B. RNA processing and its regulation: global insights into biological networks. *Nature Reviews Genetics*, 11(1):75–87, 2010. (Cited on page 203.)

Lin, M. T. and Beal, M. F. Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature*, 443, 2006. (Cited on page 75.)

Lionel, A. C., Crosbie, J., Barbosa, N., Goodale, T., Thiruvahindrapuram, B., Rickaby, J., Gazzellone, M., Carson, A. R., Howe, J. L., Wang, Z., et al. Rare copy number variation discovery and cross-disorder comparisons identify risk genes for adhd. *Science translational medicine*, 3(95):95ra75–95ra75, 2011. (Cited on page 172.)

Lionel, A. C., Tammimies, K., Vaags, A. K., Rosenfeld, J. A., Ahn, J. W., Merico, D., Noor, A., Runke, C. K., Pillalamarri, V. K., Carter, M. T., et al. Disruption of the astn2/trim32 locus at 9q33. 1 is a risk factor in males for autism spectrum disorders, adhd and other neurodevelopmental phenotypes. *Human molecular genetics*, page ddt669, 2013. (Cited on page 172.)

Liu, Z., Wang, Y., Wang, S., Zhang, J., Zhang, F., and Niu, Y. Nek2C functions as a tumor promoter in human breast tumorigenesis. *International Journal of Molecular Medicine*, 30(4): 775, 2012. (Cited on page 202.)

Logan-Klumpler, F. J., De Silva, N., Boehme, U., Rogers, M. B., Velarde, G., McQuillan, J. A., Carver, T., Aslett, M., Olsen, C., Subramanian, S., Phan, I., Farris, C., Mitra, S., Ramasamy, G., Wang, H., Tivey, A., Jackson, A., Houston, R., Parkhill, J., Holden, M., Harb, O. S., Brunk, B. P., Myler, P. J., Roos, D., Carrington, M., Smith, D. F., Hertz-Fowler, C., and Berriman, M. GeneDB–an annotation database for pathogens. *Nucleic Acids Research*, 40(Database issue):D98–108, 2012. (Cited on page 200.)

Lovén, J., Hoke, H. A., Lin, C. Y., Lau, A., Orlando, D. A., Vakoc, C. R., Bradner, J. E., Lee, T. I., and Young, R. A. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, 153(2):320–334, 2013. (Cited on page 26.)

Luger, K. and Richmond, T. J. The histone tails of the nucleosome. *Current Opinion in Genetics & development*, 8(2):140–146, 1998. (Cited on page 25.)

Luo, B., Cheung, H. W., Subramanian, A., Sharifnia, T., Okamoto, M., Yang, X., Hinkle, G., Boehm, J. S., Beroukhim, R., Weir, B. A., et al. Highly parallel identification of essential genes in cancer cells. *Proceedings of the National Academy of Sciences*, 105(51):20380–20385, 2008. (Cited on pages 63, 65, 66, 77 and 102.)

Ma, H., Sorokin, A., Mazein, A., Selkov, A., Selkov, E., Demin, O., and Goryanin, I. The Edinburgh human metabolic network reconstruction and its functional analysis. *Molecular Systems Biology*, 3:135, 2007. (Cited on pages 9, 76, 196 and 209.)

Mabbott, N. A., Baillie, J. K., Brown, H., Freeman, T. C., and Hume, D. A. An expression atlas of human primary cells: inference of gene function from coexpression networks. *BMC Genomics*, 14, 2013. (Cited on pages 78, 83, 87 and 103.)

Machado, D. and Herrgård, M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Computational Biology*, 10, 2014. (Cited on pages 23, 99, 117, 118, 119, 125, 132, 156, 162 and 163.)

Machanick, P. and Bailey, T. L. Meme-chip: Motif analysis of large dna datasets. *Bioinformatics*, 27, 2011. (Cited on page 110.)

Maeda, N., Kasukawa, T., Oyama, R., Gough, J., Frith, M., Engström, P. G., Lenhard, B., Aturaliya, R. N., Batalov, S., Beisel, K. W., et al. Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genetics*, 2(4):e62, 2006. (Cited on page 203.)

Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research*, 33(Database issue):D54–D58, 2005. (Cited on page 200.)

Mahadevan, R. and Schilling, C. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metabolic Engineering*, 5(4):264–276, 2003. (Cited on pages 7, 11, 39, 156 and 162.)

Mahadevan, R., Bond, D. R., Butler, J. E., Esteve-Núñez, A., Coppi, M. V., Palsson, B. O., Schilling, C. H., and Lovley, D. Characterization of metabolism in the fe (iii)-reducing organism geobacter sulfurreducens by constraint-based modeling. *Applied and Environmental Microbiology*, 72(2):1558–1568, 2006. (Cited on page 3.)

Mandairon, N., Sacquet, J., Garcia, S., Ravel, N., Jourdan, F., and Didier, A. Neurogenic correlates of an olfactory discrimination task in the adult olfactory bulb. *European journal of Neuroscience*, 24(12):3578–3588, 2006. (Cited on page 176.)

Maniloff, J. The minimal cell genome: “on being the right size”. *Proceedings of the National Academy of Sciences*, 93(19):10004, 1996. (Cited on page 57.)

Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Nookaew, I., Jacobson, P., Walley, A. J., Froguel, P., Carlsson, L. M., Uhlen, M., et al. Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Molecular Systems Biology*, 9(1):649, 2013. (Cited on pages 9, 20, 22, 118 and 121.)

Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Uhlen, M., and Nielsen, J. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nature communications*, 5, 2014a. (Cited on pages xiii, 9, 20, 22, 32, 76, 115, 198 and 199.)

Mardinoglu, A., Kampf, C., Asplund, A., Fagerberg, L., Hallstrom, B. M., Edlund, K., BluÌĹher, M., Pontén, F., Uhlen, M., and Nielsen, J. Defining the human adipose tissue proteome to reveal metabolic alterations in obesity. *Journal of proteome research*, 13(11):5106–5119, 2014b. (Cited on page 22.)

Martelli, P. L., D'Antonio, M., Bonizzoni, P., Castrignanò, T., D'Erchia, A. M., De Meo, P. D., Fariselli, P., Finelli, M., Licciulli, F., Mangiulli, M., et al. ASPicDB: a database of annotated transcript and protein variants generated by alternative splicing. *Nucleic Acids Research*, page gkq1073, 2010. (Cited on pages 205 and 206.)

Maston, G. A., Landt, S. G., Snyder, M., and Green, M. R. Characterization of enhancer function from genome-wide analyses. *Annu Rev Genomics Hum Genet*, 13, 2012. (Cited on page 75.)

Mazurek, S., Boschek, C. B., Hugo, F., and Eigenbrodt, E. Pyruvate kinase type M2 and its role in tumor growth and spreading. *Seminars in Cancer Biology*, 15(4):300–308, 2005. (Cited on page 202.)

McCall, M. N., Bolstad, B. M., and Irizarry, R. A. Frozen robust multiarray analysis (frma). *Biostatistics*, 11, 2010. (Cited on page 101.)

McCall, M. N., Uppal, K., Jaffee, H. A., Zilliox, M. J., and Irizarry, R. A. The gene expression barcode: leveraging public data repositories to begin cataloging the human and murine transcriptomes. *Nucleic Acids Research*, 39(suppl 1):D1011–D1015, 2011. (Cited on pages 19, 60, 66, 77, 121, 123, 124, 156, 160 and 181.)

Mechawar, N., Saghatelyan, A., Grailhe, R., Scoriels, L., Gheusi, G., Gabellec, M.-M., Lledo, P.-M., and Changeux, J.-P. Nicotinic receptors regulate the survival of newborn neurons in the adult olfactory bulb. *Proceedings of the National Academy of Sciences of the United States of America*, 101(26):9822–9826, 2004. (Cited on page 192.)

Megchelenbrink, W., Huynen, M., and Marchiori, E. optgpsampler: An improved tool for uniformly sampling the solution-space of genome-scale metabolic networks. *PloS one*, 9(2): e86587, 2014. (Cited on page 7.)

Merrill Jr, A. H., Henderson, J. M., Wang, E., McDonald, B. W., and Millikan, W. J. Metabolism of vitamin b-6 by human liver. *The Journal of nutrition*, 114(9):1664–1674, 1984. (Cited on page 129.)

Mills, J. D. and Janitz, M. Alternative splicing of mRNA in the molecular pathology of neurodegenerative diseases. *Neurobiology of Aging*, 33(5):1012–e11, 2012. (Cited on page 203.)

Ming, G.-l. and Song, H. Adult neurogenesis in the mammalian brain: significant answers and significant questions. *Neuron*, 70(4):687–702, 2011. (Cited on page 171.)

Mintz-Oron, S., Meir, S., Malitsky, S., Ruppin, E., Aharoni, A., and Shlomi, T. Reconstruction of *Arabidopsis* metabolic network models accounting for subcellular compartmentalization and tissue-specificity. *Proceedings of the National Academy of Sciences*, 109(1):339–344, 2012. (Cited on pages 9 and 196.)

Mo, M. L., Palsson, B. Ø., and Herrgård, M. J. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Systems Biology*, 3(1):1, 2009. (Cited on page 3.)

Monk, J., Nogales, J., and Palsson, B. Ø. Optimizing genome-scale network reconstructions. *Nature Biotechnology*, 32(5):447–452, 2014. (Cited on page 196.)

Morowitz, H. J. The completeness of molecular biology. *Israel journal of medical sciences*, 20 (9):750, 1984. (Cited on page 57.)

Najafi-Shoushtari, S. H., Kristo, F., Li, Y., Shioda, T., Cohen, D. E., Gerszten, R. E., and Näär, A. M. Microrna-33 and the srebp host genes cooperate to control cholesterol homeostasis. *Science*, 328, 2010. (Cited on page 81.)

Navid, A. and Almaas, E. Genome-level transcription data of yersinia pestis analyzed with a new metabolic constraint-based approach. *BMC Systems Biology*, 6(1):150, 2012. (Cited on page 152.)

Nicklas, S., Otto, A., Wu, X., Miller, P., Stelzer, S., Wen, Y., Kuang, S., Wrogemann, K., Patel, K., Ding, H., et al. Trim32 regulates skeletal muscle stem cell differentiation and is necessary for normal adult muscle regeneration. *PloS one*, 7(1):e30445, 2012. (Cited on page 172.)

Nissant, A. and Pallotto, M. Integration and maturation of newborn neurons in the adult olfactory bulb–from synapses to function. *European Journal of Neuroscience*, 33(6):1069–1077, 2011. (Cited on page 171.)

Nissant, A., Bardy, C., Katagiri, H., Murray, K., and Lledo, P.-M. Adult neurogenesis promotes synaptic plasticity in the olfactory bulb. *Nature Neuroscience*, 12(6):728–730, 2009. (Cited on page 171.)

Nissim-Rafinia, M. and Kerem, B. Splicing regulation as a potential genetic modifier. *Trends Genetics*, 18(3):123–127, 2002. (Cited on page 202.)

Norris, P. C. and Dennis, E. A. A lipidomic perspective on inflammatory macrophage eicosanoid signaling. *Adv Biol Regul*, 54, 2014. (Cited on page 97.)

Oh, Y.-K., Palsson, B. O., Park, S. M., Schilling, C. H., and Mahadevan, R. Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *The Journal of Biological Chemistry*, 282(39):28791–28799, Sep 2007. (Cited on page 198.)

Olivier, B. G. and Bergmann., F. T. Flux Balance Constraints, Version 1 Release 1., a. (Cited on page 200.)

Olivier, B. G. and Bergmann., F. T. Flux Balance Constraints, Version 2 Release 1., b. (Cited on pages 198 and 200.)

Orth, J. D., Thiele, I., and Palsson, B. O. What is flux balance analysis? *Nat Biotech*, 28(3): 245–248, March 2010. ISSN 1087-0156. (Cited on pages 4, 6, 53 and 54.)

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. a., Nam, H., Feist, A. M., and Palsson, B. O. A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Molecular Systems Biology*, 7(535):535, January 2011. ISSN 1744-4292. (Cited on pages 9, 36, 51, 196 and 198.)

Ostuni, R., Piccolo, V., Barozzi, I., Polletti, S., Termanini, A., Bonifacio, S., Curina, A., Prosperini, E., Ghisletti, S., and Natoli, G. Latent enhancers activated by stimulation in differentiated cells. *Cell*, 152(1):157–171, 2013. (Cited on page 27.)

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T., Edwards, R., Fonstein, M., Frank, E. D., Gerdes, S., Glass, E. M., Goesmann, A., Hanson, A., Iwata-Reuyl, D., Jensen, R., Jamshidi, N., Krause, L., Kubal, M., Larsen, N., Linke, B., McHardy, A. C., Meyer, F., Neuweger, H., Olsen, G., Olson, R., Osterman, A., Portnoy, V., Pusch, G. D., Rodionov, D. A., Rückert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., and Vonstein, V. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research*, 33(17):5691–5702, 2005. (Cited on pages 196, 207 and 208.)

Pal, S., Gupta, R., and Davuluri, R. V. Alternative transcription and alternative splicing in cancer. *Pharmacology & therapeutics*, 136(3):283–294, 2012. (Cited on page 203.)

Palsson, B. et al. The challenges of in silico biology. *Nature Biotechnology*, 18(11):1147–1150, 2000. (Cited on pages 4 and 6.)

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, 2008. (Cited on page 201.)

Parker, S. C. J., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., Bueren, K. L., Chines, P. S., Narisu, N., Black, B. L., Visel, A., Pennacchio, L. A., and Collins, F. S. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences*, 110, 2013. (Cited on pages 75 and 84.)

Pasquali, L., Gaulton, K. J., Rodríguez-Seguí, S. A., Mularoni, L., Miguel-Escalada, I., Akerman, I., Tena, J. J., Morán, I., Gómez-Marín, C., Bunt, M., Ponsa-Cobas, J., Castro, N., Nammo, T., Cebola, I., García-Hurtado, J., Maestro, M. A., Pattou, F., Piemonti, L., Berney, T., Gloyn, A. L., Ravassard, P., Skarmeta, J. L. G., Müller, F., McCarthy, M. I., and Ferrer, J. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet*, 46, 2014. (Cited on pages 84 and 91.)

Patil, K. R. and Nielsen, J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(8):2685–2689, 2005. (Cited on page 20.)

Petreanu, L. and Alvarez-Buylla, A. Maturation and death of adult-born olfactory bulb granule neurons: Role of olfaction. *The Journal of Neuroscience*, 22(14):6106–6113, 2002. (Cited on page 171.)

Pfau, T., Pacheco, M. P., and Sauter, T. Towards improved genome-scale metabolic network reconstructions: unification, transcript specificity and beyond. *Briefings in Bioinformatics*, 2015. (Cited on page xiv.)

Pfister, T. D., Reinhold, W. C., Agama, K., Gupta, S., Khin, S. A., Kinders, R. J., Parchment, R. E., Tomaszewski, J. E., Doroshow, J. H., and Pommier, Y. Topoisomerase i levels in the nci-60 cancer cell line panel determined by validated elisa and microarray analysis and correlation with indenoisoquinoline sensitivity. *Mol Cancer Ther*, 8, 2009. (Cited on pages 63 and 77.)

Pham, T. . H., Benner, C., Lichtinger, M., Schwarzfischer, L., Hu, Y., Andreesen, R., Chen, W., and Rehli, M. Dynamic epigenetic enhancer signatures reveal key transcription factors associated with monocytic differentiation states. *Blood*, 119, 2012. (Cited on page 81.)

Pires Pacheco, M. and Sauter, T. Fast reconstruction of compact context-specific metabolic networks via integration of microarray data. *arXiv preprint arXiv:1407.6534*, 2014. (Cited on pages xiv, 59, 181 and 188.)

Pires Pacheco, M., John, E., Kaoma, T., Heinäniemi, M., Nicot, N., Vallar, L., Bueb, J.-L., Sinkkonen, L., and Sauter, T. Integrated metabolic modelling reveals cell-type specific epigenetic control points of the macrophage metabolic network. *BMC Genomics*, 16(1):1–24, 2015a. ISSN 1471-2164. (Cited on pages xiv, 10, 12, 23, 59, 72, 117, 119, 122, 124, 129, 136, 149, 158, 161 and 163.)

Pires Pacheco, M., Pfau, T., and Sauter, T. Benchmarking procedures for high-throughput context specific reconstruction algorithms. *Frontiers in Physiology*, 6(410), 2015b. ISSN 1664-042X. (Cited on pages 22, 156, 157, 161 and 162.)

Poolman, M. G. ScrumPy: metabolic modelling with Python. *Systems Biology*, 153(5):375–378, 2006. (Cited on pages 198 and 199.)

Poolman, M. G., Miguet, L., Sweetlove, L. J., and Fell, D. A. A genome-scale metabolic model of Arabidopsis and some of its properties. *Plant Physiology*, 151(3):1570–1581, 2009. (Cited on pages 9, 196, 198 and 200.)

Pourfar, M., Niethammer, M., and Eidelberg, D. Metabolic networks in Parkinson's disease. In Grimaldi, G. and Manto, M., editors, *Mechanisms and Emerging Therapies in Tremor Disorders*, Contemporary Clinical Neuroscience, pages 403–415. Springer New York, 2013. ISBN 978-1-4614-4026-0. (Cited on page 35.)

Price, N. D., Papin, J. A., Schilling, C. H., and Palsson, B. O. Genome-scale microbial in silico models: the constraints-based approach. *Trends in biotechnology*, 21(4):162–169, 2003. (Cited on page 20.)

Price, N. D., Reed, J. L., and Palsson, B. Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews Microbiology*, 2(11):886–897, 2004. (Cited on pages 4, 6, 36 and 55.)

Puryear, W. B., Yu, X., Ramirez, N. P., Reinhard, B. M., and Gummuluru, S. Hiv-1 incorporation of host-cell-derived glycosphingolipid gm3 allows for capture by mature dendritic cells. *Proc Natl Acad Sci*, 109, 2012. (Cited on page 97.)

Quek, L.-E., Dietmair, S., Hanscho, M., Martínez, V. S., Borth, N., and Nielsen, L. K. Reducing recon 2 for steady-state flux analysis of hek cell culture. *J Biotechnol*, 184:172–178, Aug 2014. (Cited on page 115.)

Queralt-Rosinach, N. and Furlong, L. I. Disgenet rdf: A gene-disease association linked open data resource. In *SWAT4LS*, 2013. (Cited on pages xxvii, 67, 69, 78 and 103.)

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., and Wysocka, J. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, 470(7333):279–283, 2011. (Cited on pages 26, 27, 75, 84 and 163.)

Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A., Swainston, N., Baart, G., and Schwartz, J.-M. Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Systems Biology*, 4:114, 2010. (Cited on page 206.)

Raghunathan, A., Reed, J., Shin, S., Palsson, B., and Daefler, S. Constraint-based analysis of metabolic capacity of salmonella typhimurium during host-pathogen interaction. *BMC Systems Biology*, 3(1):38, 2009. (Cited on page 3.)

Ranganathan, S., Suthers, P. F., and Maranas, C. D. Optforce: an optimization procedure for identifying all genetic manipulations leading to targeted overproductions. *PLoS Computational Biology*, 6(4):e1000744, 2010. (Cited on page 3.)

Reed, J. L., Vo, T. D., Schilling, C. H., and Palsson, B. Ø. An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). *Genome Biology*, 4(9):R54, 2003. (Cited on pages 3, 9 and 196.)

Reed, J. L., Patel, T. R., Chen, K. H., Joyce, A. R., Applebee, M. K., Herring, C. D., Bui, O. T., Knight, E. M., Fong, S. S., and Palsson, B. O. Systems approach to refining genome annotation. *Proceedings of the National Academy of Sciences*, 103(46):17480–17484, 2006. (Cited on page 155.)

Remize, F., Andrieu, E., and Dequin, S. Engineering of the Pyruvate Dehydrogenase Bypass in *Saccharomyces cerevisiae*: Role of the Cytosolic Mg2+ and Mitochondrial K+ Acetaldehyde Dehydrogenases Ald6p and Ald4p in Acetate Formation during Alcoholic Fermentation. *Applied and Environmental Microbiology*, 66(8):3151–3159, 2000. (Cited on page 207.)

Resendis-Antonio, O., Checa, A., and Encarnación, S. Modeling core metabolism in cancer cells: surveying the topology underlying the warburg effect. *PloS one*, 5(8):e12383, 2010. (Cited on page 21.)

Robaina Estévez, S. and Nikoloski, Z. Generalized framework for context-specific metabolic model extraction methods. *Front Plant Sci*, 5:491, 2014. (Cited on pages 116 and 118.)

Robaina Estévez, S. and Nikoloski, Z. Context-specific metabolic model extraction based on regularized least squares optimization. *PLoS One*, 10(7):e0131875, 2015. (Cited on pages 116, 118, 122, 123, 129 and 136.)

Rodriguez, J. M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.-J., Lopez, G., Valencia, A., and Tress, M. L. APPRIS: annotation of principal and alternative splice isoforms. *Nucleic Acids Research*, 41(Database issue):D110–D117, 2013. (Cited on pages 202 and 204.)

Romero, P., Wagg, J., Green, M. L., Kaiser, D., Krummenacker, M., and Karp, P. D. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biology*, 6(1):R2, 2005. (Cited on pages 76 and 209.)

Rosenthal, M. and Glew, R. *Medical biochemistry: human metabolism in health and disease*, 2009. (Cited on pages 129 and 139.)

Ross, R. Atherosclerosis–an inflammatory disease. *N Engl J Med*, 340, 1999. (Cited on pages 97 and 98.)

Ruan, C.-S., Wang, S.-F., Shen, Y.-J., Guo, Y., Yang, C.-R., Zhou, F. H., Tan, L.-T., Zhou, L., Liu, J.-J., Wang, W.-Y., et al. Deletion of trim32 protects mice from anxiety-and depression-like behaviors under mild stress. *European Journal of Neuroscience*, 40(4):2680–2690, 2014. (Cited on pages 172 and 194.)

Ryu, J. Y., Kim, H. U., and Lee, S. Y. Reconstruction of genome-scale human metabolic models using omics data. *Integr. Biol.*, Mar 2015. (Cited on page 116.)

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., Nativ, N., Bahir, I., Doniger, T., Krug, H., Sirota-Madi, A., Olender, T., Golan, Y., Stelzer, G., Harel, A., and Lancet, D. GeneCards Version 3: the human gene integrator. *Database (Oxford)*, 2010: baq020, 2010. (Cited on page 200.)

Salzberg, S. L. Genome re-annotation: a wiki solution? *Genome Biology*, 8(1):102, 2007. (Cited on page 204.)

Schatschneider, S., Persicke, M., Watt, S. A., Hublik, G., Pühler, A., Niehaus, K., and Vorhölter, F.-J. Establishment, in silico analysis, and experimental verification of a large-scale metabolic network of the xanthan producing Xanthomonas campestris pv. campestris strain B100. *Journal Biotechnol*, 167(2):123–134, Aug 2013. (Cited on page 199.)

Schellenberger, J., Que, R., Fleming, R., Thiele, I., Orth, J., Feist, A., Zielinski, D., Bordbar, A., Lewis, N., Rahmanian, S., Kang, J., Hyduke, D., and Palsson, B. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat Protoc*, 6 (9):1290–1307, Sep 2011a. (Cited on pages 39, 50, 51, 66, 122, 198 and 199.)

Schellenberger, J., Park, J. O., Conrad, T. M., and Palsson, B. Ø. Bigg: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11(1):213, 2010. (Cited on page 152.)

Schellenberger, J., Lewis, N. E., and Palsson, B. Ø. Elimination of thermodynamically infeasible loops in steady-state metabolic models. *Biophysical Journal*, 100(3):544–553, 2011b. (Cited on pages 198, 207 and 208.)

Schilling, C. H., Letscher, D., and Palsson, B. Ø. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of theoretical biology*, 203(3):229–248, 2000. (Cited on page 6.)

Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S., and Palsson, B. O. Genome-scale metabolic model of helicobacter pylori 26695. *Journal of Bacteriology*, 184 (16):4582–4593, 2002. (Cited on page 3.)

Schmidt, B. J., Ebrahim, A., Metz, T. O., Adkins, J. N., Palsson, B. Ø., and Hyduke, D. R. Gim3e: condition-specific models of cellular metabolism developed from metabolomics and expression data. *Bioinformatics*, 29(22):2900–2908, 2013. (Cited on page 11.)

Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, 5(12):e1000605, 2009. (Cited on page 204.)

Schomburg, I., Chang, A., Placzek, S., Söhngen, C., Rother, M., Lang, M., Munaretto, C., Ulas, S., Stelzer, M., Grote, A., Scheer, M., and Schomburg, D. BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA. *Nucleic Acids Research*, 41(Database issue):D764–D772, 2013. (Cited on pages 8, 118 and 200.)

Schultz, A. and Qutub, A. A. Reconstruction of tissue-specific metabolic networks using corda. *PLoS Computational Biology*, 12(3):e1004808, 2016. (Cited on pages 152 and 161.)

Schuster, S. and Hilgetag, C. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 2(2):165–182, 1994. (Cited on pages 6, 36 and 37.)

Schwamborn, J. C., Berezikov, E., and Knoblich, J. A. The trim-nhl protein trim32 activates micrornas and prevents self-renewal in mouse neural progenitors. *Cell*, 136(5):913–925, 2009. (Cited on pages 172 and 182.)

Schwarcz, R., Bruno, J. P., Muchowski, P. J., and Wu, H. . Q. Kynurenines in the mammalian brain: when physiology meets pathology. *Nat Rev Neurosci*, 13, 2012. (Cited on pages 96 and 98.)

Sebé-Pedrós, A., Ballaré, C., Parra-Acero, H., Chiva, C., Tena, J. J., Sabidó, E., Gómez-Skarmeta, J. L., Di Croce, L., and Ruiz-Trillo, I. The dynamic regulatory genome of capsaspora and the origin of animal multicellularity. *Cell*, 165(5):1224–1237, 2016. (Cited on page 25.)

Segre, D., Vitkup, D., and Church, G. M. Analysis of optimality in natural and perturbed metabolic networks. *Proceedings of the National Academy of Sciences*, 99(23):15112–15117, 2002. (Cited on pages 7 and 21.)

Shanahan, C. M., Carpenter, K. L. H., and Cary, N. R. B. A potential role for sterol 27-hydroxylase in atherogenesis. *Atherosclerosis*, 154, 2001. (Cited on page 95.)

Shankavaram, U. T., Varma, S., Kane, D., Sunshine, M., Chary, K. K., Reinhold, W. C., Pommier, Y., and Weinstein, J. N. Cell miner: a relational database and query tool for the nci-60 cancer cell lines. *BMC Genomics*, 10, 2009. (Cited on pages 63, 77 and 101.)

Shlomi, T., Berkman, O., and Ruppin, E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7695–7700, 2005. (Cited on page 7.)

Shlomi, T., Eisenberg, Y., Sharan, R., and Ruppin, E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Molecular Systems Biology*, 3(1):101, 2007. (Cited on pages 9 and 10.)

Shlomi, T., Cabili, M. N., Herrgård, M. J., Palsson, B. Ø., and Ruppin, E. Network-based prediction of human tissue-specific metabolism. *Nat Biotechnol*, 26(9):1003–1010, Sep 2008. (Cited on pages 36, 118 and 119.)

Shlomi, T., Cabili, M. N., and Ruppin, E. Predicting metabolic biomarkers of human inborn errors of metabolism. *Molecular Systems Biology*, 5(1):263, 2009. (Cited on pages 20 and 120.)

Shlomi, T., Benyamini, T., Gottlieb, E., Sharan, R., and Ruppin, E. Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the warburg effect. *PLoS Computational Biology*, 7(3):e1002018, 2011. (Cited on page 21.)

Siersbæk, R., Rabiee, A., Nielsen, R., Sidoli, S., Traynor, S., Loft, A., Poulsen, L. L. C., Rogowska-Wrzesinska, A., Jensen, O. N., and Mandrup, S. Transcription factor cooperativity in early adipogenic hotspots and super-enhancers. *Cell Rep*, 7, 2014. (Cited on pages 76 and 96.)

Sigurdsson, M. I., Jamshidi, N., Steingrimsson, E., Thiele, I., and Palsson, B. Ø. A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. *BMC Systems Biology*, 4:140, 2010. (Cited on page 198.)

Smith, J. C. A tutorial guide to mixed-integer programming models and solution techniques. (Cited on pages 13 and 154.)

Smith, R. L. Efficient monte carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984. (Cited on page 6.)

Smyth, G. (Cited on page 106.)

Starr, T., Abbott, K., Nyre, E., Abrahante, J., Ho, Y.-Y., and Isaksson Vogel, R. Candidate cancer gene database. (Cited on page 67.)

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., and Gilles, E. D. Metabolic network structure determines key aspects of functionality and regulation. *Nature*, 420(6912):190–193, 2002. (Cited on page 6.)

Stephanopoulos, G., Aristidou, A., and Nielsen, J. *Metabolic engineering: principles and methodologies*. Academic Press, 1998. (Cited on page 36.)

Strahl, B. D. and Allis, C. D. The language of covalent histone modifications. *Nature*, 403(6765): 41–45, 2000. (Cited on pages 25 and 26.)

Stuchlik, A. Dynamic learning and memory, synaptic plasticity and neurogenesis: an update. *Front Behav Neurosci*, 8(106):1–6, 2014. (Cited on page 171.)

Suckow, J., Markiewicz, P., Kleina, L. G., Miller, J., Kisters-Woike, B., Müller-Hill, B., et al. Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. *Journal of molecular biology*, 261(4):509, 1996. (Cited on page 57.)

Sun, J., Sayyar, B., Butler, J. E., Pharkya, P., Fahland, T. R., Famili, I., Schilling, C. H., Lovley, D. R., and Mahadevan, R. Genome-scale constraint-based modeling of geobacter metallireducens. *BMC Systems Biology*, 3(1):1, 2009. (Cited on page 3.)

Suzuki, E., Williams, S., Sato, S., Gilkeson, G., Watson, D. K., and Zhang, X. K. The transcription factor fli-1 regulates monocyte, macrophage and dendritic cell development in mice. *Immunology*, 139, 2013. (Cited on page 81.)

Suzuki, H., Forrest, A. R. R., Nimwegen, E., Daub, C. O., Balwierz, P. J., Irvine, K. M., Lassmann, T., Ravasi, T., Hasegawa, Y., Hoon, M. J. L., Katayama, S., Schroder, K., Carninci, P., Tomaru, Y., Kanamori-Katayama, M., Kubosaki, A., Akalin, A., Ando, Y., Arner, E., Asada, M., Asahara, H., Bailey, T., Bajic, V. B., Bauer, D., Beckhouse, A. G., Bertin, N., Björkegren, J., Brombacher, F., Bulger, E., and Chalk, A. M. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet*, 41, 2009. (Cited on page 93.)

Systems Biology Research Group. BiGG 2, Aug 2015. URL bigg.ucsd.edu. (Cited on pages 198 and 208.)

Takeda, J.-i., Suzuki, Y., Nakao, M., Kuroda, T., Sugano, S., Gojobori, T., and Imanishi, T. H-DBAS: alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational. *Nucleic Acids Research*, 35(suppl 1):D104–D109, 2007. (Cited on pages 203 and 205.)

Takeda, J.-i., Suzuki, Y., Sakate, R., Sato, Y., Gojobori, T., Imanishi, T., and Sugano, S. H-DBAS: human-transcriptome database for alternative splicing: update 2010. *Nucleic Acids Research*, 38(suppl 1):D86–D90, 2010. (Cited on pages 205 and 206.)

Tan, J., Savigner, A., Ma, M., and Luo, M. Odor information processing by the olfactory bulb analyzed in gene-targeted mice. *Neuron*, 65(6):912–926, 2010. (Cited on page 171.)

Tan, L. and Yu, J. T. The kynurenine pathway in neurodegenerative diseases: Mechanistic and therapeutic considerations. *J Neurol Sci*, 323, 2012. (Cited on page 98.)

Taneri, B. and Gaasterland, T. *Genome and Transcriptome Sequence Databases for Discovery, Storage, and Representation of Alternative Splicing Events*, pages 1–34. John Wiley & Sons, Inc., Hoboken, New Jersey., 2013. doi: 10.1002/9781118617151.ch01. (Cited on page 203.)

Tazi, J., Bakkour, N., and Stamm, S. Alternative splicing and disease. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1792(1):14–26, 2009. (Cited on page 202.)

The 1000 Genomes Project Consortium and others. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010. (Cited on page 205.)

The ENCODE Project Consortium and others. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004. (Cited on pages 204 and 205.)

The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, 42(Database issue):D191–D198, 2014. doi: 10.1093/nar/gkt1140. URL `http://dx.doi.org/10.1093/nar/gkt1140`. (Cited on page 200.)

Thiele, I. and Palsson, B. Ø. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature Protocols*, 5(1):93–121, 2010. ISSN 1754-2189. (Cited on pages 35 and 196.)

Thiele, I., Vo, T. D., Price, N. D., and Palsson, B. Ø. Expanded metabolic reconstruction of helicobacter pylori (iit341 gsm/gpr): an in silico genome-scale characterization of single- and double-deletion mutants. *Journal of Bacteriology*, 187(16):5818–5830, 2005. (Cited on page 3.)

Thiele, I., Hyduke, D. R., Steeb, B., Fankam, G., Allen, D. K., Bazzani, S., Charusanti, P., Chen, F.-C., Fleming, R. M., Hsiung, C. A., et al. A community effort towards a knowledge-base and mathematical model of the human pathogen salmonella typhimurium lt2. *BMC Systems Biology*, 5(1):8, 2011. (Cited on page 3.)

Thiele, I., Swainston, N., Fleming, R. M. T., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., Thorleifsson, S. G., Agren, R., Bölling, C., Bordel, S., Chavali, A. K., Dobson, P., Dunn, W. B., Endler, L., Hala, D., Hucka, M., Hull, D., Jameson, D., Jamshidi, N., Jonsson, J. J., Juty, N., Keating, S., Nookaew, I., Novère, N. L., Malys, N., Mazein, A., Papin, J. A., Price, N. D., Selkov, E., Sigurdsson, M. I., Simeonidis, E., Sonnenschein, N., Smallbone, K., Sorokin, A., van Beek, J. H. G. M., Weichart, D., Goryanin, I., Nielsen, J., Westerhoff, H. V., Kell, D. B., Mendes, P., and Palsson, B. Ø. A community-driven global reconstruction of human metabolism. *Nature Biotechnology*, 31(5):419–425, 2013. (Cited on pages xiii, 9, 32, 35, 51, 76, 115, 120, 121, 126, 136, 157, 158, 159, 196, 198 and 209.)

Thiele, I., Vlassis, N., and Fleming, R. M. fastgapfill: efficient gap filling in metabolic networks. *Bioinformatics*, 30(17):2529–2531, 2014. (Cited on pages 23, 155 and 158.)

Toni, N., Teng, E. M., Bushong, E. A., Aimone, J. B., Zhao, C., Consiglio, A., van Praag, H., Martone, M. E., Ellisman, M. H., and Gage, F. H. Synapse formation on neurons born in the adult hippocampus. *Nature Neuroscience*, 10(6):727–734, 2007. (Cited on page 171.)

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, 2010. (Cited on page 202.)

Trewavas, A. A brief history of systems biology every object that biology studies is a system of systems. francois jacob (1974). *The Plant Cell Online*, 18(10):2420–2430, 2006. (Cited on page 2.)

Trinh, C. T., Unrean, P., and Srienc, F. Minimal escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses. *Applied and environmental microbiology*, 74(12): 3634–3643, 2008. (Cited on page 6.)

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. Towards a knowledge-based human protein atlas. *Nature Biotechnology*, 28(12):1248–1250, 2010. (Cited on pages 17 and 24.)

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J., and Pontén, F. Proteomics. tissue-based map of the human proteome. *Science*, 347(6220):1260419, Jan 2015. (Cited on pages 26, 117, 118, 121, 136, 149 and 158.)

Urban, N. N. Lateral inhibition in the olfactory bulb and in olfaction. *Physiology & behavior*, 77 (4):607–612, 2002. (Cited on page 171.)

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., and Sidow, A. Genome-wide analysis of transcription factor binding sites based on chip-seq data. *Nat Methods*, 5, 2008. (Cited on page 109.)

Van der Crabben, S., Verhoeven-Duif, N., Brilstra, E., Van Maldergem, L., Coskun, T., Rubio-Gozalbo, E., Berger, R., and De Koning, T. An update on serine deficiency disorders. *Journal of inherited metabolic disease*, 36(4):613–619, 2013. (Cited on page 193.)

Van Holde, K. E. *Chromatin*. Springer Science & Business Media, 2012. (Cited on page 25.)

van Praag, H., Schinder, A. F., Christie, B. R., Toni, N., Palmer, T. D., and Gage, F. H. Functional neurogenesis in the adult hippocampus. *Nature*, 415(6875):1030–1034, 2002. (Cited on page 171.)

Van Vliet, A. H. Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS microbiology letters*, 302(1):1–7, 2010. (Cited on page 19.)

Varma, A. and Palsson, B. O. Metabolic flux balancing: Basic concepts, scientific and practical use. *BIO/Techology*, 12:994–998, 1994. (Cited on pages 4 and 6.)

Varrette, S., Bouvry, P., Cartiaux, H., and Georgatos, F. Management of an academic hpc cluster: The ul experience. In *Proc. of the 2014 Intl. Conf. on High Performance Computing & Simulation (HPCS 2014)*, Bologna, Italy, July 2014. IEEE. (Cited on page 128.)

Vaz, F. M., Houtkooper, R. H., Valianpour, F., Barth, P. G., and Wanders, R. J. A. Only one splice variant of the human TAZ gene encodes a functional protein with a role in cardiolipin metabolism. *The Journal of Biological Chemistry*, 278(44):43089–43094, 2003. (Cited on page 201.)

Vazquez, A., Liu, J., Zhou, Y., and Oltvai, Z. N. Catabolic efficiency of aerobic glycolysis: the warburg effect revisited. *BMC systems biology*, 4(1):58, 2010. (Cited on page 21.)

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. Chip-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858, 2009. (Cited on page 26.)

Vitkin, E. and Shlomi, T. Mirage: a functional genomics-based approach for metabolic network model reconstruction and its application to cyanobacteria networks. *Genome Biology*, 13(11): R111, 2012. (Cited on pages 23, 51 and 155.)

Vivar, C. and Van Praag, H. Functional circuits of new neurons in the dentate gyrus. *Frontiers in neural circuits*, 7, 2015. (Cited on page 171.)

Vlassis, N., Pires Pacheco, M., and Sauter, T. Fast reconstruction of compact context-specific metabolic network models. *PLoS Computational Biology*, 10, 2014. (Cited on pages xiii, 10, 11, 23, 32, 62, 76, 100, 116, 118, 122, 123, 124, 129, 155, 158, 159 and 196.)

Wang, B.-B. and Brendel, V. Genomewide comparative analysis of alternative splicing in plants. *Proceedings of the National Academy of Sciences of the United States of America*, 103(18): 7175–7180, May 2006. (Cited on pages 205 and 206.)

Wang, Y., Eddy, J. A., and Price, N. D. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Systems Biology*, 6(1):153, 2012. (Cited on pages 11, 21, 36, 47, 48, 52, 57, 66, 76, 103, 116, 117, 118, 120, 122, 129, 139, 152 and 196.)

Wang, Z., Gerstein, M., and Snyder, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009. (Cited on page 19.)

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319, 2013. (Cited on pages 75, 84 and 163.)

Wiback, S. J., Famili, I., Greenberg, H. J., and Palsson, B. Ø. Monte carlo sampling can be used to determine the size and shape of the steady-state flux space. *Journal of Theoretical Biology*, 228(4):437–447, 2004. (Cited on page 6.)

Wielenga, V. J., Heider, K.-H., Johan, G., Offerhaus, A., Adolf, G. R., van den Berg, F. M., Ponta, H., Herrlich, P., and Pals, S. T. Expression of CD44 variant proteins in human colorectal cancer is related to tumor progression. *Cancer Research*, 53(20):4754–4756, 1993. (Cited on page 202.)

Wilming, L. G., Gilbert, J. G., Howe, K., Trevanion, S., Hubbard, T., and Harrow, J. L. The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Research*, 36(suppl 1):D753–D760, 2008. (Cited on pages 205 and 206.)

Wolffe, A. P. and Hayes, J. J. Chromatin disruption and modification. *Nucleic Acids Research*, 27(3):711–720, 1999. (Cited on page 25.)

Xu, Q., Modrek, B., and Lee, C. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Research*, 30(17):3754–3766, 2002. (Cited on page 203.)

Xu, T.-L. and Gong, N. Glycine and glycine receptor signaling in hippocampal neurons: diversity, function and regulation. *Progress in neurobiology*, 91(4):349–361, 2010. (Cited on page 193.)

Yamaguchi, M., Saito, H., Suzuki, M., and Mori, K. Visualization of neurogenesis in the central nervous system using nestin promoter-gfp transgenic mice. *Neuroreport*, 11(9):1991–1996, 2000. (Cited on page 182.)

Yang, M. and Crawley, J. N. Simple behavioral assessment of mouse olfaction. *Current protocols in Neuroscience*, pages 8–24, 2009. (Cited on page 176.)

Yao, J., Mu, Y., and Gage, F. H. Neural stem cells: mechanisms and modeling. *Protein & cell*, 3 (4):251–261, 2012a. (Cited on page 171.)

Yao, S., Ireland, S., Bee, A., Beesley, C., Forootan, S., Dodson, A., Dickinson, T., Gerard, P., Lian, L., Risk, J., et al. Splice variant PRKC-ζ-PrC is a novel biomarker of human prostate cancer. *Br. J. Cancer*, 107(2):388–399, 2012b. (Cited on page 202.)

Yizhak, K., Gaude, E., Le Dévédec, S., Waldman, Y. Y., Stein, G. Y., van de Water, B., Frezza, C., and Ruppin, E. Phenotype-based cell-specific metabolic modeling reveals metabolic liabilities of cancer. *eLife*, 3:e03641, 2014a. (Cited on pages 21, 66, 116, 120, 122, 152, 157 and 199.)

Yizhak, K., Le Dévédec, S. E., Rogkoti, V. M., Baenke, F., de Boer, V. C., Frezza, C., Schulze, A., van de Water, B., and Ruppin, E. A computational study of the warburg effect identifies metabolic targets inhibiting cancer migration. *Molecular Systems Biology*, 10(8):744, 2014b. (Cited on pages 21 and 66.)

Yokoi, M., Mori, K., and Nakanishi, S. Refinement of odor molecule tuning by dendrodendritic synaptic inhibition in the olfactory bulb. *Proceedings of the National Academy of Sciences*, 92(8):3371–3375, 1995. (Cited on page 171.)

Yokota, T., Mishra, M., Akatsu, H., Tani, Y., Miyauchi, T., Yamamoto, T., Kosaka, K., Nagai, Y., Sawada, T., and Heese, K. Brain site-specific gene expression analysis in alzheimer's disease patients. *European journal of clinical investigation*, 36(11):820–830, 2006. (Cited on page 172.)

Zhang, F. and Drabier, R. SASD: the Synthetic Alternative Splicing Database for identifying novel isoform from proteomics. *BMC Bioinformatics*, 14(Suppl 14):S13, 2013. ISSN 1471-2105. (Cited on pages 205 and 206.)

Zhang, Y., Filiou, M. D., Reckow, S., Gormanns, P., Maccarrone, G., Kessler, M. S., Frank, E., Hambsch, B., Holsboer, F., Landgraf, R., et al. Proteomic and metabolomic profiling of a trait anxiety mouse model implicate affected pathways. *Molecular & Cellular Proteomics*, 10(12): M111–008110, 2011. (Cited on pages 192 and 193.)

Zhao, C., Deng, W., and Gage, F. H. Mechanisms and functional implications of adult neurogenesis. *Cell*, 132(4):645–660, 2008. (Cited on page 171.)

Zilliox, M. J. and Irizarry, R. A. A gene expression bar code for microarray data. *Nature Methods*, 4(11):911–913, 2007. (Cited on pages 19, 60, 63, 77, 98, 99, 123, 156, 160 and 181.)

Zomorrodi, A. and Maranas, C. Improving the imm904 s. cerevisiae metabolic model using essentiality and synthetic lethality data. *BMC Systems Biology*, 4(1):178, 2010. ISSN 1752-0509. (Cited on page 51.)

Zur, H., Ruppin, E., and Shlomi, T. imat: an integrative metabolic analysis tool. *Bioinformatics*, 26(24):3140–3142, 2010. (Cited on pages 21, 47, 48, 63, 66, 76, 98, 115, 118, 122, 129, 136, 151, 152, 156 and 162.)