![uni.lu UNIVERSITÉ DU LUXEMBOURG]

DISSERTATION

Defense held on 22/06/2017 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN BIOLOGIE

by

Stefanía MAGNÚSDÓTTIR
Born on 21 January 1988 in Reykjavík, Iceland

DEVELOPMENT AND ANALYSIS OF INDIVIDUAL-BASED GUT MICROBIOME METABOLIC MODELS

Dissertation defense committee
Dr Ines Thiele, dissertation supervisor
Luxembourg Centre for Systems Biomedicine
Associate Professor, Université du Luxembourg

Dr Christophe Lacroix
Professor, ETH Zürich, Switzerland

Dr Paul Wilmes, Chairman
Luxembourg Centre for Systems Biomedicine
Associate Professor, Université du Luxembourg

Dr Francisco Planes
CEIT and Tecnun
Associate Professor, University of Navarra, Spain

Dr Ronan Fleming, Vice Chairman
Luxembourg Centre for Systems Biomedicine
Collaborateur scientifique (Senior), Université du Luxembourg

II

UNIVERSITÉ DU
LUXEMBOURG

Molecular Systems Physiology

Luxembourg Centre for Systems Biomedicine

Faculty of Life Sciences, Technology and Communication

Doctoral School in Systems and Molecular Biomedicine

Luxembourg Centre
for Systems Biomedicine

Fonds National de la
Recherche Luxembourg

Disseration Defence Committee:
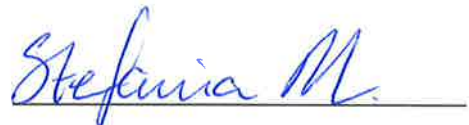
Committee members:	A-Prof. Dr. Paul Wilmes
	Dr. Ronan M. T. Fleming
	Prof. Dr. Christophe Lacroix
	A-Prof. Dr. Francisco Planes
Supervisor:	A-Prof. Dr. Ines Thiele

AFFIDAVIT

I hereby confirm that the PhD thesis entitled "Development and analysis of individual-based gut microbiome metabolic models" has been written independently and without any other sources than cited.

Luxembourg, 26.05.2017

Stefanía Magnúsdóttir

IV

Contents

List of Figures

List of Tables

List of Abbreviations

AGORA   Assembly of gut organisms through reconstruction and analysis
ATP   Adenine triphosphate
BOF   Biomass objective function
CoA   Coenzyme A
COBRA   Constraint-based reconstruction and analysis
DHF   Dihydrofolate
DNA   Deoxyribonucleic acid
DRI   Daily reference intake
FBA   Flux balance analysis
FNR   Fonds national de la recherche
FVA   Flux variability analysis
gDW   Grams dry weight
GENRE   Genome-scale network reconstruction
GI   Gastrointestinal
GPR   Gene-protein-reaction
HGM   Human gut microbiota
ORF   Open reading frame
PABA p-Aminobenzoic acid
PEG   Protein encoding gene
RNA   Ribonucleic acid
TCA   Tricarboxylic acid cycle
THF   Tetrahydrofolate
WGS   Whole-genome shotgun
VMH   Virtual metabolic human

Summary

The human gut microbiota plays a large role in the metabolism of our diet. These microorganisms can break down indigestible materials such as polysaccharides and convert them into metabolites that the human body can take up and utilize (e.g., vitamins, essential amino acids, and short-chain fatty acids). Disbalances in the gut microbiome have been associated with several diseases, including diabetes and obesity. However, little is known about the detailed metabolic crosstalk that occurs between individual organisms within the microbiome and between the microbiome and the human intestinal cells. Because of the complexity of the intestinal ecosystem, these interactions are difficult to determine using existing experimental methods. Constraint-based reconstruction and analysis (COBRA) can help identify the possible metabolic mechanisms at play in the human gut. By combining mathematical, computational, and experimental methods, we can generate hypotheses and design targeted experiments to elucidate the metabolic mechanisms in the gut microbiome.

In this thesis, I first applied comparative genomics to analyze the biosynthesis pathways of eight B-vitamins in hundreds of human gut microbial species. The results suggested that many gut microbes do not synthesize any B-vitamins, that is, they depend on the host's diet and neighboring bacteria for these essential nutrients. Second, I developed a semi-automatic reconstruction refinement pipeline that quickly generates biologically relevant genome-scale metabolic reconstructions (GENREs) of human gut microbes based on automatically generated metabolic reconstructions, comparative genomics data, and data extracted from biochemical experiments on the relevant organisms. The pipeline generated metabolically diverse reconstructions that maintain high accuracy with known biochemical data. Finally, the refined GENREs were combined with metagenomic data from individual stool samples to build personalized human gut microbiome metabolic reconstructions. The resulting large-scale microbiome models were both taxonomically and functionally diverse.

The work presented in this thesis has enabled the generation of biologically relevant human gut microbiome metabolic reconstructions. Metabolic models resulting from such reconstructions can be applied to study metabolism within the human gut microbiome and between the gut microbiome and the human host. Additionally, they can be used to study the effects of different dietary components on the metabolic exchanges in the gut microbiome and the metabolic differences between healthy and diseased microbiomes.

Chapter 1

Introduction: COBRA modeling of the human gut microbiome

Manuscript in preparation.

1.1   Human gut microbiome

The gut microbiome serves an important purpose in human metabolism and most intestinal microbes are beneficial; they break down large indigestible compounds from the diet and secrete essential nutrients. Human gut microbes can digest dietary fibers [59] and secrete fermentation products, including short-chain fatty acids, which can be taken up by the human cells and used as energy precursors [60]. In addition, gut microbes secrete essential amino acids and vitamins that the intestinal cells can take up [205].

Despite the known benefits of the intestinal microbes, studies have also reported connections between an imbalance in the microbiome ("dysbiosis") [39] and several diseases [129, 234, 56, 87, 169]. However, most of these studies investigated correlations between the microbiome and disease status, meaning that the detailed metabolic pathways that underlie these observations are not known. Present experimental methods are not capable of detecting such detailed mechanisms of metabolic exchanges among hundreds of microbes and how those exchanges affect the human metabolism.

With the rise of high-throughput genome sequencing methods, scientists have started to grasp the complexity of the human gut as an ecosystem. Metagenomics is presently the most

3

commonly used method when studying the microbial diversity in the human gut. In short, the microbial gene content of a given sample, usually a stool sample, is characterized either by identifying the 16S ribosomal RNA (rRNA) gene content or using whole genome shotgun (WGS) sequencing. The identification of rRNA genes has long been used to assign microbial organisms to taxonomic groups [244] and has been applied in many studies of the human gut microbiome [233, 37, 3, 106]. However, the 16S rRNA sequencing results cannot give a comprehensive view on the species or strain composition of a gut microbiome sample [239]. In contrast, WGS sequencing is more refined and can give information on individual strains when aligning the metagenomic sequence reads to a set of reference genomes [168, 41, 42, 24], identifying microbial clade marker genes [203], or using genome-specific markers [232].

In a WGS metagenomics study of the human gut in 2010, Qin et al. [168] found that every individual carries more than 160 microbial species in their intestine, and more than 1,000 microbes were identified among the cohort of only 124 people. In 2011, a study by Arumugam et al. [10] suggested that the human microbiome could be categorized based on the abundances of different taxonomic groups. However, others have challenged their hypothesis and suggest that the gut microbiome composition is highly variable among individuals [42].

The gut microbiome has been shown to respond to changes in several different factors. Claesson et al. (2012) [37] observed that microbiomes among elderly individuals varied more than the microbiomes among their young control subjects and that the composition correlated with both diet and health. That same year, Yatsunenko et al. [249] also observed age-related changes in the gut microbiomes of their subjects, as well as between the geographical locations of their cohorts. Several studies have shown that modifications to the host's diet can shift the microbiome composition [47, 217, 30]. However, David et al. (2014) [47] showed that while lifestyle changes made short-term alterations to individual microbiomes, the overall community structure was stable long-term. In addition, despite the high variability of the microbiome among individuals and the diverse influences to the microbiome composition, the functional annotations seem to be stable among diverse microbiomes in healthy individuals [42].

Despite what is known about the microbiome, there is much to be discovered about the detailed metabolic mechanisms by which the microbes interact, both among each other and with the human intestinal cells. Computational methods can help to guide the exploration

of the microbiome by forming hypotheses based on knowledge-based models and testing them in the laboratory. One such method is constraint-based reconstruction and analysis (COBRA) [201], which has been successfully applied in studies on metabolism and analysis of complex metabolic networks.

1.2 Constraint-based reconstruction and analysis

COBRA can be used to study metabolic pathways, individual species metabolism, and inter-species metabolic interactions [157, 164]. The work in this thesis builds on the use of genome-scale metabolic reconstructions (GENREs), which represent the full set of metabolic pathways that occur in that organism based on genomic and experimental knowledge.

1.2.1 Genome-scale metabolic reconstructions

GENREs are based on a collection of metabolic functions that can be inferred from the list of genes identified in an organism [226]. First, the genome sequence is annotated and the resulting set of genes encoding for metabolic enzymes is extracted. From the set of metabolic enzymes, we can infer the list of metabolic reactions that can occur in the organism. Today, GENREs can be created automatically through several different platforms; a few examples are listed in Table 1.1. Such automatic tools have made it possible to generate GENREs in only a few minutes to hours, a task that was previously very time-consuming [226]. However, these automatically created reconstructions require manual refinement to resolve various issues, including stoichiometric consistency [72], reaction directionality [73], gene mis-annotations [85], and known biological functions of the organism based on experimental knowledge [226]. Therefore, automatically generated GENREs are often referred to as "draft" reconstructions (Fig. 1.1a).

GENREs contain gene-protein-reaction (GPR) associations where the genes that are responsible for each metabolic enzyme are linked to the corresponding metabolic reactions [226]. Often, the relation between a gene and an enzyme is not straight-forward. In the case of isoenzymes, i.e., multiple enzymes that can perform the same metabolic reaction, we introduce an "OR" rule, meaning that at least one of the different genes must be present to allow flux through the reaction in question. In the case of enzyme complexes, where multiple

a) Genome-scale metabolic reconstruction

b) Flux balance analysis

Maximize/minimize an objective function $Z = \sum_{i=1}^{n} c_i v_i$

such that

and for every reaction i: $lb_i \leq v_i \leq ub_i$
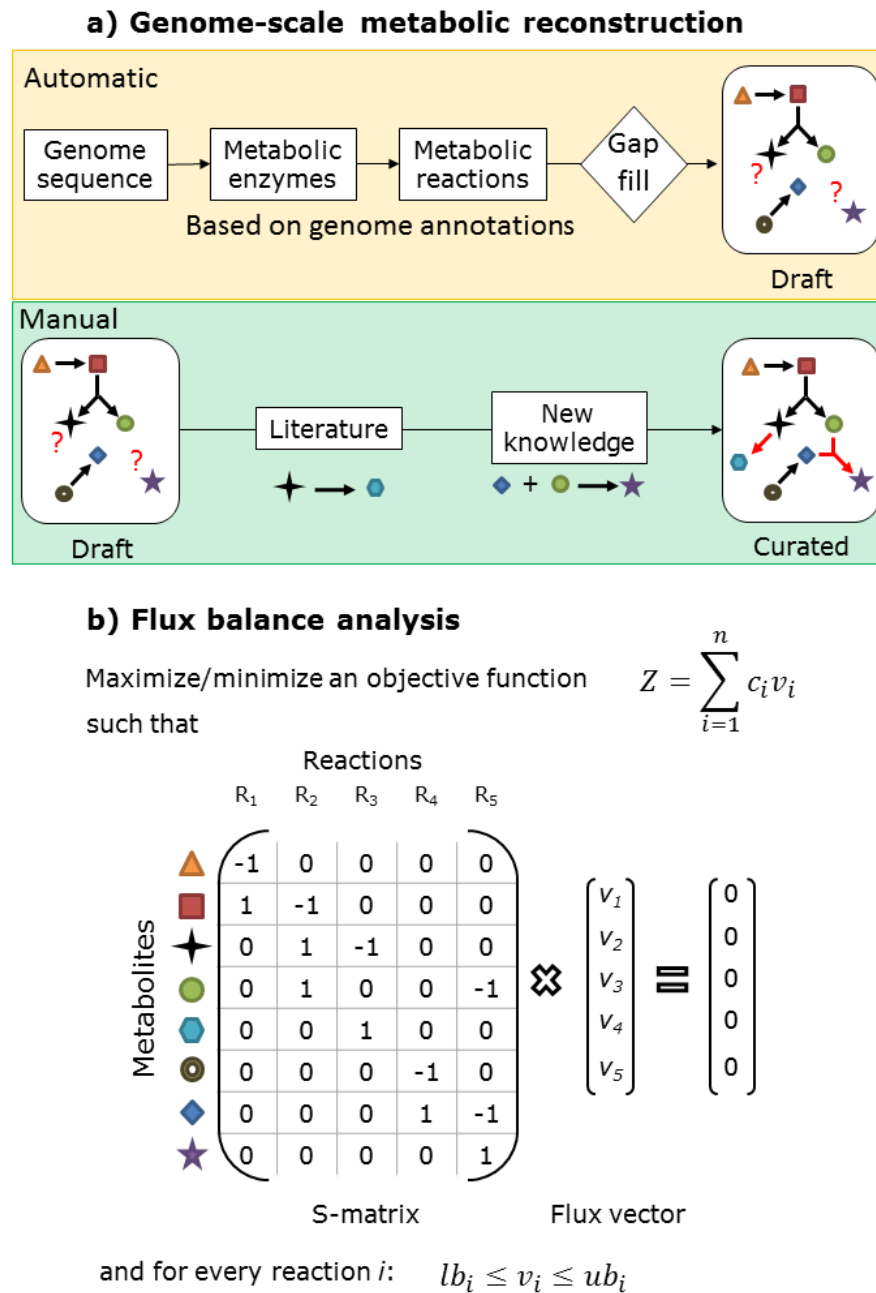
Figure 1.1: Illustration of genome-scale metabolic reconstruction and flux balance analysis. a) Genome-scale metabolic reconstructions are built in two steps. First, several platforms exist for the generation of a gap-filled draft metabolic reconstruction based on genome annotations (Table 1.1). Second, the draft reconstructions needs to be refined based on known experimental and biochemical data from literature. Novel experiments can be performed on the organism and the reconstruction refined accordingly. b) In flux balance analysis (FBA), the metabolic reconstruction containing n reactions and m metabolites is converted to a stoichiometric matrix S. FBA solves an optimization problem where an objective function Z is maximized or minimized. Z is formed by multiplying every reaction flux v_i with a predetermined constant, c_i, and adding up the resulting values. FBA solves a steady state, $\mathbf{S} \cdot \mathbf{v} = \mathbf{0}$. Every reaction i is bound by two values: an upper bound (ub_i) and lower bound (lb_i).

Table 1.1: Platforms for automatic generation of genome-scale metabolic reconstructions.

Platform	Author (year)
CellNetAnalyzer	Klamt et al. (2007) [118]
DOE Systems Biology Knowledgebase (KBase)	Arkin et al. (2016) [9]
Flux Analysis and Modeling Environment (FAME)	Boele et al. (2012) [27]
MicrobesFlux	Feng et al. (2012) [70]
ModelSEED	Henry et al., (2010) [102]
Pathway Tools	Karp et al. (2016) [113]
Reconstruction, Analysis and Visualization of Metabolic Networks (RAVEN)	Agren et al. (2013) [2]
SuBliMinaL Toolbox	Swainston et al. (2011) [218]

genes encode different subunits of a metabolic enzyme, we introduce an "AND" rule, meaning that all the responsible genes are required to enable flux through that reaction. The GPR associations are used when integrating gene expression data or performing gene knock-out simulations [164].

Draft GENREs can rarely be converted directly into functional metabolic models and must be refined to be used for biologically accurate modeling. Such manual curations can be very time consuming. Manual curations are data- and knowledge-based modifications to the draft GENREs. Per literature and experimental data, it is known that the organism performs certain biochemical functions that sometimes are not captured fully by the genome because of missing annotations or possible alternate pathways. This is also referred to as gap-filling the metabolic network. The gaps that are filled can be gaps that connect metabolic pathways or reactions that supply or remove side-substrates or products. Numerous algorithms have been developed to accelerate and facilitate the gap-filling process [159] (e.g., SMILEY [178], GapFind [196], fastGapFill [228], SONEC [25], and EnsemblFBA [26]. While each of these algorithms have different algorithms and thus result in potential alternate gap filling solutions, they all require experimental data to produce potential gap filling solutions and manual evaluation of the biological relevance of these solutions.

In microbial reconstructions, the cytosol and extracellular space are represented by separating metabolites into two compartments. This separation is symbolized using the suffixes [*c*] for the cytosol and [*e*] for the extracellular space. Metabolites that can be transported across the cell membrane are thus represented twice in the same reconstruction, once for each compartment. Some researchers choose to also represent the periplasmic space when

reconstructing Gram-negative microbes, e.g., the *Escherichia coli* reconstruction by Feist et al. (2007) [68]. Such reconstructions contain an additional compartment, [*p*], and separate transporters that carry metabolites from the extracellular space to the periplasm, and from the periplasm to the cytosol.

Traditionally, a biomass function [69] is defined in microbial *in silico* modeling. The substrates of the biomass function are metabolites that are necessary components of a cell and needed for cell replication. Many such components are universal among organisms [248] and include micro- and macromolecules such as nucleotides, amino acids, ATP, water, lipids, and vitamins. Platforms that automatically generate GENREs will usually include a biomass function that is defined based on metabolic network [102], and external platforms exist that can estimate the biomass composition of an organism based on genomic information [195]. However, such automatically generated biomass functions need to be manually refined and verified. Unfortunately, detailed biomass compositions have not been experimentally determined for many organisms, making it difficult to refine the biomass functions of many microbes.

1.2.2   Conversion to mathematical format

First, a metabolic reconstruction is assembled from the list of relevant biochemical reactions. Each reaction is written out as a chemical equation where substrates are consumed (left side) and products are produced (right side) [163]. The set of chemical reactions are represented by a stoichiometric matrix, **S**, in which rows represent the unique set of metabolites in the network and the columns represent the metabolic reactions (Fig. 1.1b). For every metabolite that appears in a reaction, the corresponding rows are filled with the stoichiometric coefficients from the chemical reaction. Metabolites on the left-hand side of the equation (substrates) get a negative coefficient, and right-hand side metabolites (products) have a positive coefficient.

1.2.3   Metabolic modeling

A GENRE can be converted into a condition-specific model with the application of constraints. Many different condition-specific models can be derived from a single reconstruc-

tion (Fig. 1.2). There are many different types of constraints to consider in biology, and COBRA metabolic modeling mainly accounts for three types of constraints: physicochemical, environmental, and regulatory [163].

Physicochemical constraints ensure that the laws of thermodynamics are met and that metabolite mass- and charge is balanced in the model. Thus, energy and mass cannot be created or destroyed by the model. For every reaction, the sum of every element and the sum of chemical charges must be equal on both the substrate and product sides. Environmental constraints represent the growth media of microbial cells, gases, and the media pH level. In human gut microbe simulations, the growth media is represented by the human diet. Environmental metabolites can enter the model through exchange reactions, which bring metabolites from an undefined outer source into the extracellular space [226]. The same exchange reactions are used to remove any secreted metabolites from the extracellular space. Regulatory constraints refer to the genomic regulation of an organism. Regulatory constraints are made necessary by a cell's topological constraints, e.g., crowding effects caused by the cell hosting large molecules, such as DNA and proteins. In addition, the organism has limited resources [153], e.g., nucleotides, amino acids, and energy, to regulate, translate, and transcribe genes, and thus cannot make use of its full metabolic network at a given time. The organism will express different sets of genes at any given moment, and to activate a new pathway, it must allocate resources to that task.

1.2.4 Flux balance analysis

The most commonly used method to analyze metabolic models is flux balance analysis (FBA) [159]. FBA is a biased method that seeks to identify a set of flux values that will satisfy a given objective. An objective function, Z, is defined as the dot product of a column vector, $\mathbf{c}$, which contains weight coefficients for all n reactions in the network, and the flux column vector, $\mathbf{v}$. Then the following linear programming problem is solved:

$$\text{min./max. } Z = \mathbf{c}^\mathsf{T} \cdot \mathbf{v}, \text{ such that}$$

$$\mathbf{S} \cdot \mathbf{v} = \mathbf{0}$$

METABOLIC RECTONSTRUCTION

CONDTION-SPECIFIC METABOLIC MODELS
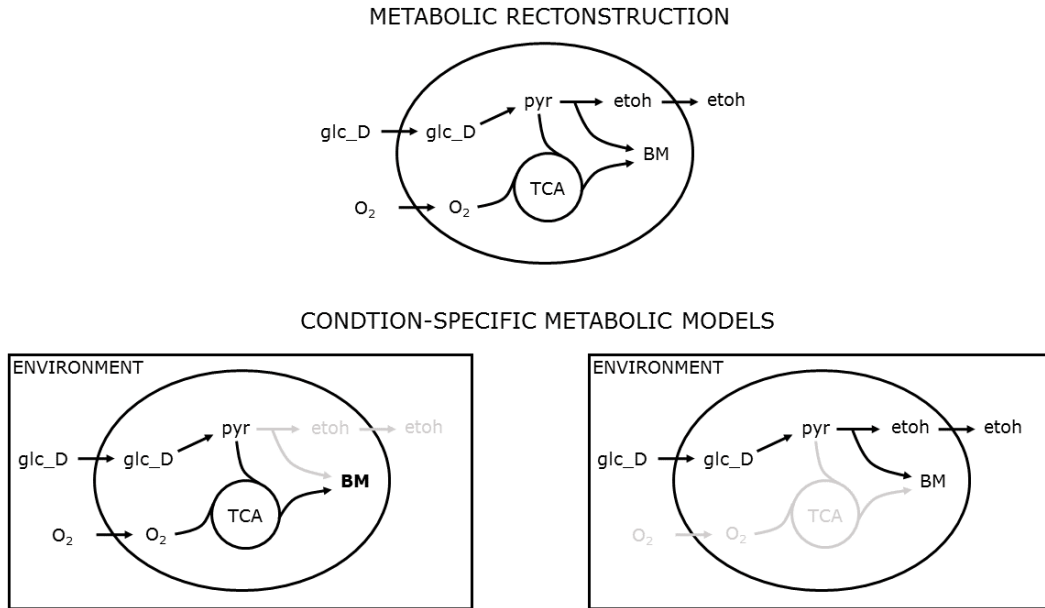
Figure 1.2: Illustration of the difference between a metabolic reconstruction and condition-specific metabolic models. Many models can be derived from a single reconstruction by applying different constraints. Here, a simplified metabolic reconstruction has the reactions to convert glucose (glc_D) to pyruvate (pyr). Pyruvate can either be converted to ethanol (etoh) through fermentation when no oxygen (O_2) is present in the environment, or enter the tricarboxylic acid cycle (TCA) when oxygen is present. In the latter case, ethanol is not produced by the model. Both models produce biomass (BM), but the amount of biomass produced is higher in the model in the aerobic environment.

$$\text{and for every reaction flux } i: lb_i \leq v_i \leq ub_i$$

where S is the network stoichiometric matrix, lb_i is the lower bound and ub_i is the upper bound of reaction flux i in the network. Traditionally in microbe modeling, the objective function is set as the biomass reaction [69]. However, it is also possible to set other reactions, or combination of reactions, from the network as objective functions [211]. For example, it can be an ATP demand reaction to investigate the ATP generating power of the model, or it can be an exchange reaction of a metabolite to simulate maximum uptake or secretion of the metabolite. FBA gives a non-unique flux vector as a solution. Metabolic reconstructions usually have a higher number of reactions than metabolites, making their S-matrix an underdetermined system. Therefore, there are many possible solution vectors that result in the same optimal value for the given problem. Flux variability analysis (FVA) [89] gives us the minimal and maximal flux values that every reaction in the network can have to achieve the optimal flux value for the objective function. However, FVA gives us information about the

extremes in the network, but does not give information about the likelihood of the flux values between the two extremes as can be done with sampling [200, 91].

1.3 Comparative genomics

Other computational methods can be used to investigate microbial metabolism. One such method is comparative genomics, which uses nucleotide and amino acid sequence alignments to analyze the genomic content in organisms [121]. Such studies rely on that nucleotide and amino acid sequences, and protein domain structures are widely conserved [142, 6]. Comparative genomics are commonly applied in microbial evolution studies [120, 166, 104, 16] and gene function discoveries [160, 177, 51, 52]. By combining sequence alignments with functional predictions on the gene or protein level [161, 162], it is possible to use comparative genomics to predict functional metabolic pathways in organisms.

Comparative genomics has been applied in several analyses of metabolic functions in microbes. In 2006, Gerdes et al. [81] used functional annotations of microbial genes to analyze NAD biosynthesis in cyanobacteria. In 2013, Ravcheev et al. [173] analyzed polysaccharide metabolism in *Bacteroidetes thetaiotaomicron* and reconstructed a regulatory network that could be used to refine polysaccharide metabolism in metabolic reconstructions of the bacterium. The following year, Ravcheev and Thiele (2014) [174] showed how aerobic and anaerobic reductases in the respiratory chain are distributed among 254 different human gut microbes. In 2016, the team applied the same analysis method to show the distribution of quinone (vitamin K) biosynthesis pathways in human gut microbes and propose novel biochemical reactions in menaquinone biosynthesis [175]. Additionally, Khoroshkin et al. (2016) analyzed the the carbohydrate utilization pathways in *Bifidobacteria* [116]. A recent study combined comparative genomics and experimental growth results to analyze metabolic enzyme cofactor dependencies in a microbiome community of 19 microbes and proposed the cofactor exchanges among the different microbes [185].

Chapter 2 in this thesis describes the application of comparative genomics to analyze the distribution of B-vitamin biosynthesis among human gut microbes. In Chapter 3, the comparative genomics-based analyses of B-vitamin biosynthesis [140], vitamin K biosynthesis [175], the respiratory chain [174], and several central metabolic pathways were used

in the refinement of the metabolic pathways and gene-protein-reaction (GPR) associations of hundreds of human gut genome-scale metabolic reconstructions (GENREs).

1.4 *In silico* metabolic modeling of the human gut microbiome

Until the work presented in this thesis, genome-scale metabolic reconstructions had been curated and published for only 15 human gut microbial species (Table 1.2). Even though many aspects of microbial and host-microbial metabolism could be studied using these reconstructions, having such a limited set available meant that we could not investigate detailed microbiomes that represent the diversity found in the human microbiome. However, previous studies have modeled microbiomes using small sets microbial GENREs as representatives of their respective taxonomic group [209, 208].

Table 1.2: Genome-scale metabolic reconstructions of human gut microbes.

Species	Author (year)
Bifidobacterium adolescentis	El-Semman et al. (2014) [63]
Bacteroides thetaiotaomicron	Heinken et al. (2013) [96], Shoaie et al. (2013) [209]
Enterococcus faecalis	Veith et al. (2015) [236]
Escherichia coli	Feist et al. (2007) [68], Baumler et al. (2011) [20]
Eubacterium rectale	Shoaie et al, (2013) [209]
Helicobacter pylori	Thiele et al. (2005) [229]
Faecalibacterium prausnitzii	Heinken et al. (2014) [95], El-Semman et al. (2014) [63]
Klebsiella pneumoniae	Liao et al. (2011) [133]
Lactobacillus plantarum	Teusink et al. (2006) [224]
Lactobacillus reuteri	Saulnier et al. (2011) [197]
Lactococcus lactis	Flahaut et al. (2013) [71]
Methanobrevibacter smithii	Shoaie et al. (2013) [209]
Pseudomonas aeruginosa	Oberhardt et al. (2008) [158], Bartell et al. (2017) [17]
Salmonella Typhimurium	Raghunathan et al. (2009) [171], Thiele et al. (2011) [225]
Streptococcus thermophilus	Pastink et al. (2009) [165]

1.5 Microbe community modeling

In the first microbial community COBRA model, Stolyar et al. (2007) [216] structured their pairwise community model based on the organelle compartmentalization of eukaryotic GEN-

REs. This community model structure has largely been maintained throughout multi-species modeling to date. Two years later, Taffs et al. (2009) [219] compared three different methods on modeling microbial communities and provided guidance of when each of the three methods is applicable. The three model set-ups were i) a compartmentalized model such as presented by Stolyar et al., ii) a pooled model that ignores the metabolic interactions within the community and focuses on the overall community metabolic potential, and iii) a nested model that uses an elementary node analysis on the microbial community using the results from the elementary node analyses on each microbe in the community. In 2012, Zomorrodi and Maranas [255] presented OptCom, a COBRA microbial community modeling framework that could model more than one objective function at the same time. OptCom allowed the user to model a system where both the overall community biomass and each individual microbe's biomass were optimized at the same time.

Klitgord and Segre (2010) [119] were the first to systematically investigate the effects of different environmental metabolites on the metabolic interactions between metabolic models. Using a compartmentalized approach, for every pair of seven microbial GENREs, they could identify growth media that could induce syntrophic interactions among the microbes. They hypothesized that environmental changes can more readily drive beneficial interactions between microbes than genomic changes. In turn, Chubiz et al. (2015) [33] performed gene knock-out studies using *in silico* microbial pairs and found that cooperating species were less affected by gene deletions than species that competed for resources.

Freilich et al. (2011) [76] were the first to look at competition between two microbial GENREs. They performed a large-scale pairwise interaction study using 118 GENREs and found that most their model pairs would compete for resources and impact each other negatively rather than cooperate. They found that cooperating GENRE pairs usually had few growth-requiring metabolites in common. Also using a large set of microbial reconstructions, Zelezniak et al. (2015) [251] found that community models of microbes that could cooperate required fewer metabolites in their growth medium than microbial communities with few cooperating species.

Microbiome communities are dynamic environment and using traditional COBRA methods alone these dynamics are often ignored. It is therefore also of interest to add a dynamics layer on top of the COBRA modeling techniques to capture these community prop-

erties. The first dynamic modeling of a GENRE community was presented by Zhuang et al. (2011) [253] as the "Dynamic Multi-species Metabolic Modeling" (DMMM) framework. The simulated growth ratios of two Fe(III)-reducing bacteria matched experimentally determined data. The model was also used to predict the metabolic mechanisms underlying observed behavior of the microbial pair under nitrogen-fixation. In 2014, Harcombe et al. [92] presented a modeling framework called "Computation of Microbial Ecosystems in Time and Space"(COMETS), enabling the dynamic modeling of metabolic exchanges between multi-species bacterial colonies. The same year, Zomorrodi et al. [254] presented an extension to the aforementioned OptCom framework, d-OptCom, which models dynamic changes in the extracellular metabolites and individual microbe biomass concentrations. Recently, an additional platform, BacArena, was published [19]. BacArena combines COBRA with agent-based modeling to simulate spatio-temporal dynamics of metabolic interactions in microbial communities. Additionally, since multiple copies of individual strains can be simulated simultaneously in BacArena, it can be used to explore different metabolic phenotypes of a single strain at different times or locations in a microbial community.

Taken together, significant efforts have been made to model metabolic exchanges among microbial communities. The studies have diverse model setups, including simulations between two or more coupled microbial GENREs and the coupling of COBRA modeling with other modeling methods to simulate dynamic changes in the microbiome metabolism. However, none of the aforementioned methods have been applied to microbial community sizes similar to the human gut microbiome.

1.5.1   COBRA studies of the gut microbiome

In the decade since the first microbial community COBRA model was published, very few community models have been used to study gut microbial interactions. The first multi-gut microbe metabolic model was published by Shoaie et al. (2013) [209] where they compared *in silico* simulations of small microbial communities consisting of two and three different bacteria with experimentally measured carbohydrate uptake rates and fermentation product secretion rates from germ-free mice colonized with the same microbes. In 2014, El-Semman et al. [63] applied the OptCom [255] framework to model the interactions between the gut

microbes *Bifidobacterium adolescentis* and *Faecalibacterium prausnitzii*. They found that the *F. prausnitzii* model could secrete more butyrate in the presence of *B. adolecentis* model through its acetate secretion. Later, Mardinoglu et al. (2015) [144] investigated how the gut microbiome could affect host amino acid metabolism by simulating the *in silico* co-culture of two microbes, *Bacteroides thetaiotamicron* and *Eubacterium rectale*, and comparing their predicted amino acid metabolism with metabolic measurements from germ-free and conventionally raised mice. That same year, Shoaie et al. (2015) [208] matched *in silico* simulations of microbial communities consisting of with fecal metabolomics data and metabolite serum levels in germ-free mice colonized with the same microbial communities. Finally, Heinken and Thiele (2015) [97], found that removing oxygen from the environment promoted mutualism in pairwise interaction metabolic models of 11 human gut microbes. Additionally, Heinken and Thiele coupled the paired microbial reconstructions with a reconstruction of the human small intestine [190], and found that the presence of a host cell drove the microbial pairs to competitive interactions due to host-derived carbohydrates.

1.5.2 Host-microbiome community models

The first study of a metabolic model between a mammalian host and a microbe was published by Bordbar et al. in 2010 [28]. To date, only two studies have been published on *in silico* modeling of a gut microbiome using host metabolic models. In the first study in 2013, Heinken et al. [96] combined a mouse metabolic reconstruction with a reconstruction of the human gut microbe *Bacteroidetes thetaiotaomicron*. They found that the presence of the microbe could rescue the phenotype of the mouse reconstruction simulated with an inborn error of metabolism. Two years later, Heinken and Thiele [98] published a study on the effects of different microbial communities on the metabolism of a human small intestinal cell [190]. They found that a community consisting of pathogenic microbes resulted in that the human cell could secrete less metabolites into its lumen compartment than a community of commensal microbes.

The largest host-microbiome model to date was the model presented by Heinken and Thiele (2015) using 11 human gut microbe GENREs coupled with a human small intestinal cell reconstruction. The work in this thesis has enabled the generation of large-scale human

gut microbiome metabolic reconstructions that can be applied in host-microbiome metabolic modeling. Chapter 3 describes the development of 773 GENREs of human gut microbial reconstructions that can be directly coupled with the human metabolic reconstruction Recon 2 [227] to simulate host-microbiome metabolism. Chapter 4 describes the generation and analysis of personalized human gut microbiome reconstructions based on metagenomic reads from individual stool samples. Such microbiome reconstructions can further be coupled and simulated with the human metabolic reconstruction.

1.6   Scope and aim of the thesis

The project described in this thesis was built on three main objectives. First, to generate a semi-automatic reconstruction tool that could be used to refine genome-scale metabolic reconstructions based on biochemical data. Second, to apply the aforementioned tool to a set of microbial representatives of the human gut microbiome. Last, to combine the microbial metabolic reconstructions into a microbiome model and investigate metabolic potentials of the microbial community.

This thesis describes my work on building up to personalized human gut microbiome metabolic models. First, I present a comparative genomics analysis of B-vitamin biosynthesis by human gut microbes (Chapter 2). Second, I show how data from comparative genomics of multiple metabolic pathways, in addition to data collected from literature, was used to GENREs of 773 human gut microbes (Chapter 3). Finally, I show how metagenomic reads can be mapped to the collection of microbial metabolic reconstructions to create personalized metabolic models of human gut microbiomes (Chapter 4). Below are short descriptions of each chapter and the detailed contributions of the collaborators involved in the different projects.

Chapter 2: Systematic genome assessment of B-vitamin biosynthesis suggests co-operation among gut microbes

Chapter 2 describes the comparative genomics analysis of B-vitamin biosynthesis in the human gut microbiome. The chapter is a full reprint of the paper published in Frontiers in

Genetics in April 2015 [140]. The B-vitamin biosynthesis analysis described in this chapter was further integrated into genome-scale metabolic reconstructions of hundreds of microbial reconstructions (Chapter 3).

Contributions

Stefanía Magnúsdóttir (S.M.) and Ines Thiele (I.T.) wrote the manuscript. All authors read and edited the manuscript. S.M., Valerie de Crécy-Lagard, and I.T. designed the study. Dmitry Ravcheev (D.R.) performed the subsystems analysis on the cobalamin subsystem. S.M. analyzed the data and designed the figures.

Chapter 3: Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota

Chapter 3 describes the large-scale curation effort of 773 genome-scale metabolic reconstructions of known human gut microbes. The resulting resource was named "assembly of gut organisms through reconstruction and analysis" (AGORA) and was published in Nature Biotechnology in January 2017 [139]. The chapter is a full reprint of the published article. The AGORA resource enabled the generation of large-scale personalized gut microbiome metabolic models described in Chapter 4.

Contributions

S.M. and I.T. wrote the manuscript. All authors read and edited the manuscript. S.M., Almut Heinken (A.H.), and I.T. designed the study. S.M. and A.H. reconstructed the pipeline. A.H. and Laura Kutt (L.K.) performed literature searches. S.M., A.H., and Eugen Bauer analyzed data. D.R. and L.K. performed genome analyses on metabolic pathways. Ronan M. T. Fleming performed quality analysis on all reconstructions. Alberto Noronha built the VMH database for the AGORA reconstructions. Kacy Greenhalgh, Joanna Baginska, and Paul Wilmes (P.W.) designed and performed *in vitro* experiments. Christian Jäger performed the metabolomic analysis. I.T. conceived and supervised the project.

Chapter 4: Individual-based gut microbiome models

Chapter 4 describes the generation of personalized metabolic models of human gut micro-biomes based on metagenomics data from individual stool samples. The results showed that the AGORA is a comprehensive resource of human gut GENREs and can be combined with unbiased metagenomic reads to create personalized microbiome metabolic reconstructions. The individual microbiome models are taxonomically and functionally diverse, and while showing functional similarities among models based on individuals from different groups, i.e., gender, family, and diabetic status.

Contributions

S.M. wrote the text, carried metabolic modeling simulations and analyses, and designed the figures. S.M. and I.T. designed the study. Anna Heintz-Buschart (A.H-B.) and P.W. provided the metagenomics data reads for the study. A.H-B. aligned the metagenomic reads to the AGORA microbial genomes and calculated the coverage depth and breadth, and the relative abundance.

Chapter 5: Concluding remarks

Chapter 5 contains the conclusions of the presented thesis and the author's personal outlook on the future directions of metabolic modeling of the human gut microbiome.

Contributions

The text was written in full by S.M.

Chapter 2

Systematic genome assessment of B-vitamin biosynthesis suggests co-operation among gut microbes

Abstract

The human gut microbiota supplies its host with essential nutrients, including B-vitamins. Using the PubSEED platform, we systematically assessed the genomes of 256 common human gut bacteria for the presence of biosynthesis pathways for eight B-vitamins: biotin, cobalamin, folate, niacin, pantothenate, pyridoxine, riboflavin, and thiamin. On the basis of the presence and absence of genome annotations, we predicted that each of the eight vitamins was produced by 40–65% of the 256 human gut microbes. The distribution of synthesis pathways was diverse; some genomes had all eight biosynthesis pathways, whereas others contained no *de novo* synthesis pathways. We compared our predictions to experimental data from 16 organisms and found 88% of our predictions to be in agreement with published data. In addition, we identified several pairs of organisms whose vitamin synthesis pathway pattern complemented those of other organisms. This analysis suggests that human gut bacteria actively exchange B-vitamins among each other, thereby enabling the survival of organisms that do not synthesize any of these essential cofactors. This result indicates the co-evolution of the gut microbes in the human gut environment. Our work presents the first comprehensive assessment of the B-vitamin synthesis capabilities of the human gut microbiota. We propose that in addition to diet, the gut microbiota is an important source of B-vitamins, and that changes in the gut microbiota composition can severely affect our dietary B-vitamin requirements.

2.1 Introduction

The human gut microbiota (HGM) supplies its host with severa lnutrients, including amino acids [151] and certain B-vitamins [105, 186]. B-vitamins are necessary cofactors for numerous aspects of human metabolism, including fat andc arbohydrate metabolism and DNA synthesis. Human cells are not capable of producing B-vitamins in sufficient amounts; thus, these cells must obtain such vitamins either from the human diet or the gut microbiota [188].

B-vitamins are found in many food products, but they are water-soluble andm any of them are temperature sensitive; thus, these vitamins can easily be removed or destroyed during the cooking process. B-vitamin deficiency is common in humans and supplements of these vitamins are often used because deficiencies in these vitamins can lead to several diseases, such as pellagra [188]. Therefore, there is great interest in the B-vitamin production by the gut microbiota, particularly because this production is thought to play a role in maintaining the vitamin homeostasis in colonocytes [188]. Because of the large number of gut microbial species, most of which have not been cultured, experimental validation of the HGM B-vitamin pathways is infeasible, at least with the current technologies. Genome annotations provide a method for both systematic and large-scale predictions of pathways for vitamin metabolism, enabling assessments of the potential of each species for vitamin biosynthesis.

Genome annotations can be obtained using, among other methods, a genomics-based approach implemented in the PubSEED platform [161, 57]. PubSEED is a multifunctional web platform, in which one can inspect defined sets of gene functional roles (annotations) for multiple genomes at once and assign annotations to protein encoding genes (PEGs), which refer to open reading frames (ORFs). A functional role can refer to any protein encoding gene and represents the protein function. Communities of experts define sets of functional roles, called subsystems. These subsystems usually contain functiona lroles that together form a metabolic pathway for a particular subset of metabolism, for example, the biosynthesis of vitamin B1 (thiamin). The subsystem can then be annotated over any number of genomes at once. In addition, the PubSEED platform offers multiple methods for the manual curation, discovery, and assignment of functional roles.

In this study, we used the PubSEED platform to analyze 256 HGM organisms for their ability to synthesize eight B-vitamins. The genomes were selected based on a published

collection of microorganisms commonly found in the human gut [168]. We used eight B-vitamin subsystems, which correspond to the biosynthesis of the active form of each vitamin from its known precursors, to predict the possible B-vitamin producers in the human gut as well as the potential competitors for resources from the vitamin pools in the gut. In addition, we compared our predictions to experimental data available in the literature on the B-vitamin requirements of 16 human gut microorganisms. This work presents a comprehensive assessment of the B-vitamin synthesis capabilities of the human gut microbiota.

2.2 Methods

The PubSEED platform (http://pubseed.theseed.org/) [161, 162, 57] contains fully annotated subsystems for eight B-vitamins: biotin, cobalamin, folate, niacin, pantothenate, pyridoxine, riboflavin, and thiamin (Table 2.1). The subsystems were populated with 256 HGM genomes. The genomes were selected based on a study by Qin et al. (2010) [168], which reported microorganisms commonly found in the human gut. A further selection criterion was that the microbes appeared in at least 50% of the participants in the aforementioned study and that their genomes were present in PubSEED (Supplementary Table A.1). For comparison, the subsystems were also populated with 257 non-HGM genomes that were isolated from human body sites other than the intestine (Supplementary Table A.1). The non-HGM genomes were chosen in the following way: (1) All genomes from the Human Microbiome Project (HMP, http://www.hmpdacc.org/HMRGD/) [41, 42] isolated from a human body site were selected; (2) HMP genomes with the body site "Gastrointestinal tract" (i.e., isolated from intestine) were excluded; (3) all genomes absent from PubSEED were excluded; (4) only one genome per species was selected; (5) among the multiple genomes for the same species, a genome with a minimal number of contigs (i.e., definitive) was selected; (6) if more than one genome with the minimal number of contigs existed, the genome with the maximal CDS number was selected. The non-HGM genomes were isolated from the following body sites: oral (127 genomes), urogenital tract (68 genomes), airways (28 genomes), skin (25 genomes), blood (7 genomes), liver (1 genome), and heart (1 genome).

Table 2.1: Subsystems analyzed for the studied B-vitamins.

Vitamin	Subsystem	Subsystem Curator
Biotin	"Biotin biosynthesis"	Dmitry Rodionov
Cobalamin	"Coenzyme B12 biosynthesis"	Dmitry Rodionov
Folate	"Folate Biosynthesis"	Valérie de Crécy-Lagard
Niacin	"NAD and NADP cofactor biosynthesis global"	Andrei Osterman
Pantothenate	"Coenzyme A Biosynthesis"	Andrei Osterman
Pyridoxine	"Pyridoxin (Vitamin B6) Biosynthesis"	Olga Zagnitko
Riboflavin	"Riboflavin, FMN, and FAD metabolism Extended"	Sveta Gerdes
Thiamin	"Thiamin biosynthesis"	Dmitry Rodionov

The subsystems can be accessed on the PubSEED Subsystems Editor website: http://pubseed.theseed.org/

Preparing and Populating the Subsystems

We first inspected the existing B-vitamin metabolism subsystems and determined whether the functional roles associated with the metabolism of each vitamin were consistent with most of the recent reports. To enable us to manually curate the functional roles involved in the biosynthesis of each vitamin, we copied the existing subsystems to new subsystems. We then compiled a list of HGM genomes by selecting a genome in the "Spreadsheet" tab and clicking "save genome selection". Using "Edit List", we selected our genomes and saved the list. Within the spreadsheet for each subsystem, we limited the display of genomes to our genome list in the window "UserSets". To simplify our analysis, we compiled functional roles with the same metabolic functions into subsets using the "Subsets" tab. Protein encoding genes (PEGs), which encode functional roles belonging to a certain pathway, often cluster together on a chromosome. In the tab "ColorSpreadsheet", we colored the genes by cluster to better identify these co-occurring genes. Finally, for each B-vitamin subsystem, we determined sets of essential functional roles that should be present in a genome for the corresponding organism to be considered a vitamin producer (Table 2.2). Missing essential roles in HGM genomes were manually curated where possible as described below. The 257 non-HGM genomes were subjected to the same criteria as those for the 256 HGM genomes.

Table 2.2: Combinations of essential functional roles for the biosynthesis of B-vitamins.

Pathway	Essential functional roles
Biotin	BioW + BioFADB
	BioC + BioFADB
Cobalamin	CbiL + CobG + CbiGF + (CobF or CbiD) + CbiECA + CobNST+ CobAT + CbiPB + CobUS
	(CbiKX or CysG) + CbiLGF + (CobF or CbiD) + CbiECA + CobAT + CbiPB + CobUS
Folate	FolEBKP + pabAc + DHFS + DHFR + FPGS
Niacin	ASPOX + QSYN + QAPRT + NaMNAT + NADS
Pantothenate	KPHMT + KPRED + ASPDC + PBAL + PANK + PPCS + PPCDC + DPCK
Pyridoxin	dxs + gapA + PdxBFAJH
	PdxTS
Riboflavin	GTPCH2 + PyrDR + DHBPS + DMRLS + RSA + RK + FMNAT
Thiamin	(ThiH or ThiO) + ThiGSF + ThiCDE
	Thi4 + ThiCDE

When more than one biosynthetic pathway for one vitamin is possible, several combinations of the essential roles are listed. Each abbreviation corresponds to the single functional role (Supplementary Table A.1).

Manual curation

The PubSEED platform offers several methods for manual curation. The function "Find candidates" uses four steps to search for a functional role in a target genome. First, it looks for the role in the target genome's existing gene annotations. Then, it searches for matching proteins in the genome by performing a similarity check with protein BLAST algorithm [6], where it compares the amino acid sequence of the gene from another genome with the amino acid sequences of all PEGs in the target genome. The third step searches for genes, which are co-localized on the chromosome with other genes from the subsystem in the target genome. Finally, a translated BLAST (tblastn) is performed, in which a query protein sequence is compared with the six-frame translation of the target genome .Another way to search for functional roles in the PubSEED platform is to examine genome annotation phylogenetic trees for sets of related proteins. Such trees show similarities among the amino acid sequences of proteins from different genome clusters together with the annotated functional role of each PEG. Trees often reveal mis-annotations of the target genes and also provide hints of the level of conservation of the sequence. For every protein encoding gen ein the PubSEED platform, it is possible to examine the NCBI Conserved Domains Database

(CDD) [143]. The database performs an RPS-BLAST algorithm (a stand-alone tool in the NCBI toolkit distribution), presents the resulting protein domain annotations on the query sequence, and lists the resulting E-values. High-confidence matches between the query sequence and conserved domains are listed as "specific hits". Once a candidate gene was found in a genome, the functional role of the PEG was changed manually on the condition that the existing annotation was not associated with another subsystem, because changing the annotation of a PEG affects all subsystems that list the current annotation can alterthe results of an existing subsystem without notifying its author. Either a new functional role was assigned to the PEG,or in the case of multiple domains or functions, the new annotation was appended to the existing ones.

Calculation of Vitamin Produced by Gut Microbes

We calculated the minimal number of microbes needed to supply a human host with the necessary B-vitamins using the following assumptions: (i) The number of bacterial cells in the colonic space is 10^{14} cells [198]. (ii) The dry weight of a single bacterial cell is $4.89 \times 10^{-13}\, gDW\, cell^{-1}$ [135]. (iii) Intracellular vitamins in bacteria become available to the host upon cell lysis. (iv) 31.7% of bacterial cells in fecal matter has been reported to consist of dead cells [23]. The ratio of dead cells in the colon could not be found, and we therefore use the measured ratio of dead cells in the feces in our calculations. (v) The intracellular concentration of vitamins was retrieved from the literature (Table 2.3) but represents only the capacity of those bacteria and not all gut bacteria in general. (vi) The volume of an *E.coli* cell is 1.1 μm^3 [123]. (vii) The dry weight of an *Escherichia coli* cell is $4.89 \times 10^{-13}\, gDW$ [135]. (viii) *E.coli* grows anaerobically at 0.26 h^{-1} [93]. The ratio of the human daily reference intake of each vitamin coming from bacteria could then be calculated in the following manner:

$$\% \text{ of DRI} = \frac{\text{Intracellular concentration} \times \text{Weight of bacteria} \times \text{AM} \times \text{HGM ratio}}{\text{DRI}} \times 100$$

(2.1)

where DRI is the recommended dietary reference intake in mg/day, AM is the atomic mass of the vitamin values from ChEBI [94], and HGM is the ratio of HGM producers as predicted in this study.

Table 2.3: Intracellular vitamin concentrations in selected gut microbes.

Vitamin	Organism	Reported value	References
Biotin	*Escherichia coli*	*40 μM*	[35]
Cobalamin	*Lactobacillus reuteri* CE	*50 μg/L*	[220]
Flavin	*E. coli*	*4 μM*	[242]
Folate derivatives	*Lactobacillus casei*	*22 μM*	[206]
Niacin	*E. coli*	1.12×10^6 molecules in a single cell	[149]
Pantothenate	*E. coli*	*< 1 μM*	[109]
Pyridoxine	*E. coli*	$1.5 \times 10^{-10}\ mol\ h^{-1}\ mgDW^{-1}$	[55]
Thiamin	*Lactobacillus fermenti*	*3 μg/gDW*	[212]

2.3 Results

2.3.1 Pathway Descriptions, Prediction Criteria, and Predictions

For each of the eight B-vitamins, we describe the known biosynthesis pathways and for each pathway, we present the frequency of the respective functional roles in the analyzed taxonomic groups. Based on the functional roles present in the subsystems, each HGM organism was predicted to be a producer or non-producer of the eight B-vitamins (Supplementary Table A.1).

Biotin (Vitamin B8)

Biotin can by synthesized *de novo* from two pimeloyl precursors, malonyl-ACP and pimelate (Fig. 2.1). There are five different synthesis routes [183, 134] but only three, via BioG, BioH, and BioW, could be identified in the 256 HGM genomes. We examined the prevalence of the BioC route using the BioG role in the 256 HGM genomes. Almost all genomes from the phylum Bacteroidetes (96%) were predicted to synthesize biotin through this route. All analyzed genomes from the phylum Fusobacteria were predicted to synthesize biotin through the BioC route with BioG. A single Fusobacteria genome, *Fusobacterium* sp. D11, was missing the essential role BioC, but it was still predicted to be a producer because the genome contained all other essential roles and all its related genomes contained BioC. In the phylum Proteobacteria, 84% of the genomes were predicted to synthesize biotin. Producers from the class Epsilonproteobacteria contained the BioG role, whereas the remaining producers in the phylum, which were mostly from the class Gammaproteobacteria, contained a BioH,

Only a single Gammaproteobacteria, *Escherichia* sp. 1_1_43, was a predicted non-producer.
We then examined the second biosynthesis route, in which BioW is used to convert salvaged

Abbr.	Metabolite	Abbr.	Metabolite
I	Pimelate	VI	Pimeloyl-ACP
II	Pimeloyl-CoA	VII	8-amino-7-oxononanoate
III	Malonyl-ACP	VIII	7,8-diaminononanoate
IV	Malonyl-ACP methyl ester	IX	Dethiobiotin
V	Pimeloyl-ACP methyl ester	X	Biotin

Actinobacteria
Bacteroidetes
Firmicutes
Proteobacteria
Fusobacteria

Ratio of predicted producers per phylum.

Figure 2.1: Biotin biosynthesis. The biotin biosynthesis subsystem contains 9 functional
roles (Supplementary Table A.1) and 10 metabolites. Functional roles are represented by
rectangles. The major biosynthesis routes are colored green, alternative routes are blue,
salvage routes are red, and known transporters are yellow. Metabolites are represented by
circles, and yellow circles represent metabolites that can be salvaged from the environment
in some cases. Only the core metabolites of the biosynthesis pathway are listed. Each bar
grpah represents the percentage of organisms in each phylum that contain the functional role.

pimelate to pimeloyl-CoA. The BioFADB pathway is then used to convert pimeloyl-CoA
to biotin. This route was only observed in the Firmicutes phylum in five genomes from
the Clostridia class and in *Bacillus subtilis* subsp. *subtilis* str. 168. During our analysis,
we observed that all Actinobacteria genomes lacked the essential biotin biosynthesis roles.
However, 19 of the 23 (83%) Actinobacteria genomes contained a BioY biotin transporter,
indicating a need for biotin. The remaining four genomes, all from the family Coriobacteri-
aceae, contained no functional roles in the pathway at all, but they did contain biotin-protein
ligases, suggesting that they require biotin.

Taken together, these findings show that the majority of Bacteroidetes, Fusobacteria, and Proteobacteria genomes contain the essential roles for biotin biosynthesis (Fig. 2.1). The synthesis of this vitamin is completely absent in Actinobacteria genomes and is very rare in genomes of the Firmicutes phylum. In total, biotin biosynthesis is present in 40% of the HGM genomes.

Cobalamin (Vitamin B12)

Cobalamin biosynthesis contains the longest pathway of eight vitamins, where adenosyl-cobalamin is synthesized from precorrin 2 (Fig. 2.2). The cobalamin can be synthesized either aerobically or anaerobically (blue and green pathways in (Fig. 2.2), respectively). No HGM microbe used the aerobic biosynthesis route, whereas some non-HGM producers contained the aerobic pathway specific roles CobG and CobNST. Only two Actinobacteria of the Coriobacteriaceae family were predicted cobalamin producers, *Collinsella aerofaciens* ATCC 25986 and *Gordonibacter pamelaeae* 7-10-1-b. Only half of the genomes in the Bacteroidetes phylum (26 of 51) were predicted to be a cobalamin producers being the lowest producer ratio observed for the Bacteroidetes of all eight vitamins. No taxonomic patterns were observed for the Bacteroidetes producers. Nearly half of the Firmicutes genomes (43%) were predicted to synthesize cobalamin. All Lactobacillales were predicted to be non-producers, except the six Lactobacillus reuteri strains that have been previously known to have this pathway [194]. Only a single other Bacilli, *Listeria monocytogenes* str. 1/2a F6854, was a predicted producer of this vitamin. The remaining cobalamin producers in the Firmicutes phylum belonged to the Clostridia class, but no specific taxonomic pattern was observed. All 14 Fusobacteria were predicted to be cobalamin producers. The Proteobacteria phylum contained only ten producers, three from Delta- and seven from Gammaproteobacteria.

Taken together, we find that cobalamin biosynthesis is present in 42% of the HGM genomes. The synthesis of cobalamin is present in all Fusobacteria, but is rare in Actinobacteria and Proteobacteria. Half of the Bacteroidetes genomes are missing the biosynthesis pathway, making cobalamin the least produced of the eight vitamins in the phylum.
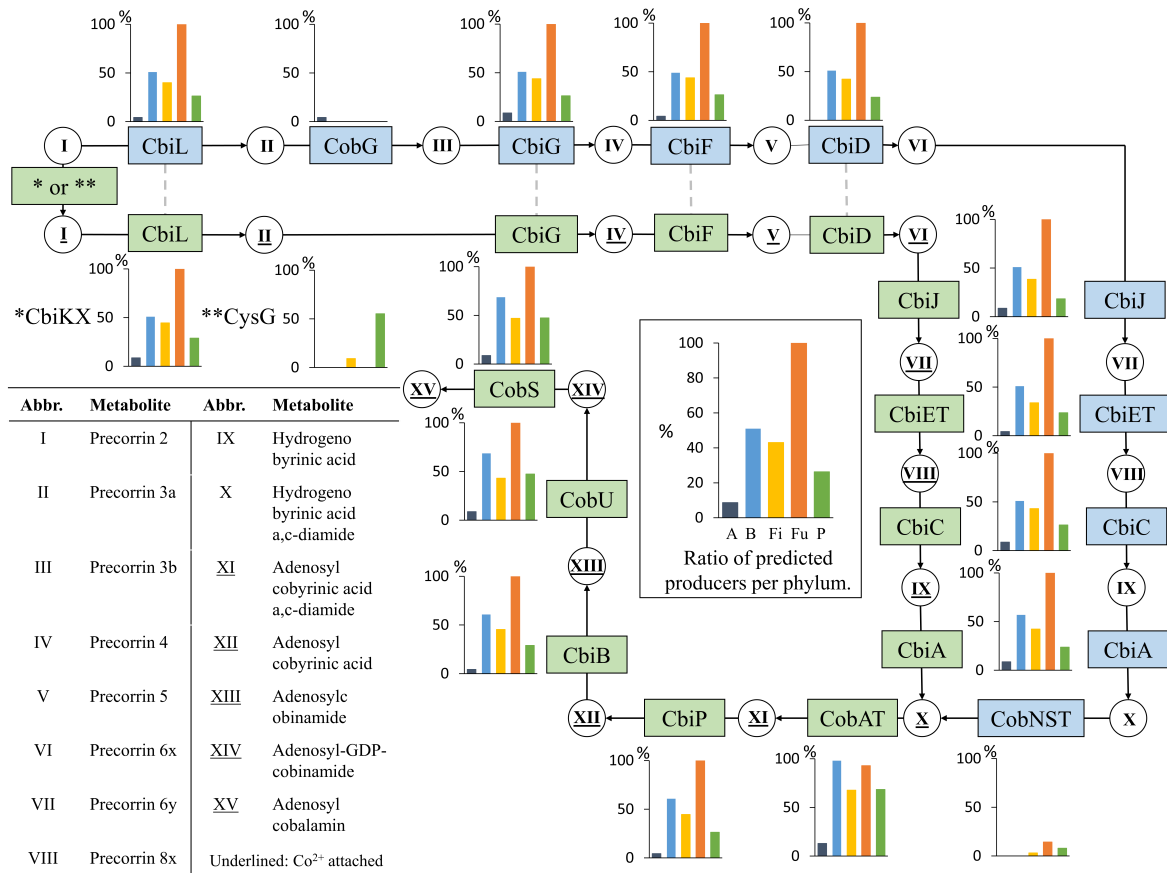
Figure 2.2: Cobalamin biosynthesis. The cobalamin biosynthesis subsystem contains 17 functional roles (Supplementary Table A.1) and 15 metabolites. Refer to Figure 2.1 for figure descriptions. Note that we show only CbiD since none of the HGM genomes contained CobF. The "*" or "**" text in the first green box on the top left corner of the figure refer to two functional roles, "*CbiKX" and "**CysG". The corresponding bar charts for the two functional roles are shown below the green box, just above the "Abbr./Metabolites" table.

Folate (Vitamin B9)

Folate biosynthesis includes the combination of two metabolic branches (Fig. 2.3). The first branch converts guanosine triphosphate (GTP) to 6-hydromethyl-7,8-dihydropterin and the second branch converts chorismate to *p*-aminobenzoic acid (PABA). Most living organisms contain more than one folate derivative. However, because these folates are derived from dihydrofolate (DHF) or tetrahydrofolate (THF), we considered the production of either metabolite to be sufficient to assign folate producer status to a genome. From our data, we identified 109 predicted folate producers, 86% of which included all the functional roles of the pathway or were only missing the FolQ role, which is commonly missing in genome annotations [80]. All producers were required to have the pabAc role of the PABA synthesis

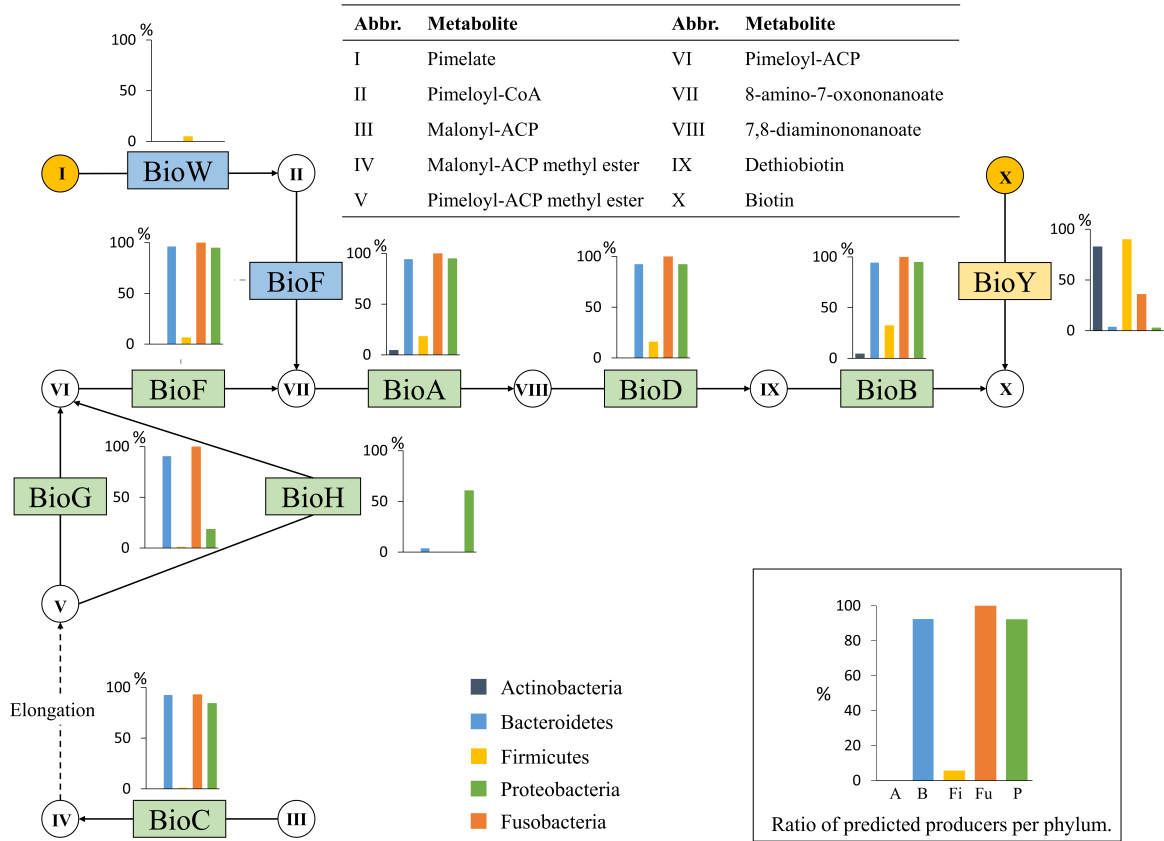| Abbr.	Metabolite	Abbr.	Metabolite
I	Guanosine triphosphate (GTP)	VIII	4-amino-4-deoxychorismate
II	7,8-dihydroneopterin triphosphate	IX	p-aminobenzoic acid (PABA)
III	7,8-dihydroneopterin monophosphate	X	Folic acid
IV	7,8-dihydroneopterin	XI	Dihydropteroate
V	6-hydromethyl-7,8-dihydropterin pyrophosphate	XII	Dihydrofolate
VI	6-hydromethyl-7,8-dihydropterin	XIII	Tetrahydrofolate
VII	Chorismate	XIV	Tetrahydrofolate polyglutamates
VIII	4-amino-4-deoxychorismate		

Figure 2.3: Folate biosynthesis. The folate biosynthesis subsystem contains 11 functional roles (Supplementary Table A.1) and 14 metabolites. Refer to Figure 2.1 for figure descriptions.

branch of the pathway. The two other roles, pabAa and pabAb, were commonly missing from our genomes, and the assignment of the roles from manual inspection was not trivial. Therefore, the presence of pabAa or pabAb was not considered essential, although these roles, along with pabAc, are needed for the synthesis of PABA. A PABA auxotroph must acquire the metabolite from its growth medium and was therefore not considered to be capable of producing folates *de novo*. Only six (26%) Actinobacteria were predicted producers: five from the order Bifidobacteriales and one from the order Actinomycetales. Although most genomes in this phylum had all the necessary functional roles for THF production, they did not contain pabAc. All but four (92%) Bacteroidetes were predicted to synthesize folates; the four exceptions lacked pabAc. Only 18 Firmicutes had complete synthesis pathways. Of the remaining 112 Firmicutes genomes, 49 contained a full biosynthesis pathway but were missing the PABA branch. The other genomes were predicted non-producers, but contained

roles for the conversion of dihydropteroate to THF, suggesting that they rely on folate up-take. Three Fusobacteria were predicted to be PABA auxotrophs, whereas the remaining 11 (79%) genomes contained all the essential roles for folate biosynthesis. The Proteobacteria phylum demonstrated more variability and class-specific patterns, with 71% of the organ-isms predicted as producers. The Betaproteobacteria included three producers, whereas two genomes from this class were missing PABA biosynthesis. All Deltaproteobacteria were predicted to be non-producers. The Gammaproteobacteria demonstrated the highest con-servation of the pathway; all functional roles were present. Only two non-producers were observed in the class: *Succinatimonas hippei* YIT 12066 was missing the PABA branch, and *Escherichia* sp. 1_1_43 was missing the PABA branch, FolE, DHFS, and FPGS.

Taken together, these findings show that the folate biosynthesis pathway is present in nearly all Bacteroidetes genomes as well as in most Fusobacteria and Proteobacteria. Folate synthesis is rare in the Actinobacteria and Firmicutes genomes, mostly because of the ab-sence of the PABA biosynthesis pathway. In total, folate biosynthesis is complete in 43% of the 256 HGM genomes tested.

Niacin (Vitamin B3)

Niacin is a group term for nicotinamide and nicotinic acid, and both of these metabolites are precursors for nicotinamide adenine dinucleotide (NAD). Nicotinamide and nicotinic acid can either be salvaged from the environment or produced through the recycling of NAD within the cell [84]. In this study, an organism was considered a niacin producer when it contained the *de novo* synthesis pathway of NAD [21] (Fig. 2.4). The first route uses the functional role NaMNAT to produce deamino-NAD, which in turn is converted to NAD by NADS. NADS requires ammonia, or alternatively, NADS is coupled to a glutaminase domain (GAT) that supplies ammonia [53]. In this study, the presence or absence of the GAT domain did not affect our predictions. All predicted producers among the 256 genomes contained both NaMNAT and NADS. In the case of Firmicutes, the Bacilli class contained only five producers and the Clostridia class included 44 predicted niacin producers. Most Fusobacteria were predicted to synthesize niacin, with only two predicted non-producers. Proteobacteria contained 29 predicted niacin synthesizers, but no taxa-specific patterns were observed.

| Abbr.	Metabolite	Abbr.	Metabolite
I	L-aspartate	VI	NAD
II	Iminoaspartate	VII	Nicotinamide mononucleotide
III	Quinolinic acid	VIII	Nicotinamide
IV	Ninotinate mononucleotide	IX	Nicotinic acid
V	Deamido-NAD	X	N-ribosylnicotinamide

Figure 2.4: Niacin biosynthesis. The niacin biosynthesis subsystem contains 14 functional roles (Supplementary Table A.1) and 10 metabolites. Refer to Figure 2.1 for figure descriptions.

Considering the prevalence of the first biosynthesis route in the gut microbiota, we wondered how common the alternative synthesis route was in these genomes. This route produces nicotinate D-ribonucleotide from quinolinic acid through NMNS and converts it to NAD using NMNAT. Only eight of the analyzed genomes contained both of these functional roles, and all contained the roles NaMNAT and NADS from the first route as well. All eight genomes belonged to the Firmicutes phylum, seven from the Clostridia class and one from Erysipelotrichia. Some prokaryotic cells share the eukaryotic NAD biosynthesis pathway from tryptophan [124]. However, we did not observe any evidence of an active tryptophan pathway in our list of HGM genomes. Because niacin is known to be salvaged from the environment, we investigated the functional roles associated with the uptake of three NAD precursors: nicotinamide, nicotinic acid, and N-ribosylnicotinamide. We found that the roles associated with the uptake of the nicotinamide and nicotinic acid (Fig. 2.4) were only present

in Actinobacteria, Firmicutes, and a single Proteobacteria. Two roles associated with the conversion of salvaged niacin to nicotinate mononucleotide (NAM and NAPRT) were present in all five phyla, but these roles are involved in the recycling of NAD [84] and therefore do not necessarily indicate the presence of niacin salvage. The salvage of N-ribosylnicotinamide through PnuC, RNK, and NMNAT was only present in four Firmicutes and 18 Proteobacteria.

Taken together, these findings show that niacin biosynthesis is present in the majority of the HGM genomes. The phyla Actinobacteria and Firmicutes contain lower ratios of *de novo* producers than the other three, and in these phyla, we observed the presence of the niacin salvage pathways, which are not present in the Fusobacteria and Bacteroidetes genomes. A total of 63% of all the investigated genomes contained the NAD biosynthesis pathways.

Pantothenate (Vitamin B5)

Pantothenate is a precursor for coenzyme A (CoA) and it can be synthesized *de novo* from 2-dihydropantoate and β-alanine (Fig. 2.5). Because CoA is the active form of pantothenate, we considered the *de novo* pantothenate biosynthesis pathway to be complete when a genome contained the functional roles needed for CoA biosynthesis. Although this pathway is well defined, it caused many uncertainties in our analysis. Many of the analyzed genomes were missing the roles KPHMT or ASPDC. When a genome was missing a single enzyme in the branched pathway, the organisms were still predicted to be a pantothenate producer. Some genomes were lacking the role PBAL, which is the final step of pantothenate biosynthesis. Hence, the absence of PBAL always resulted in a non-producer prediction.. According to our predictions, Fusobacteria, certain Actinobacteria, and Firmicutes do not possess the ability to synthesize pantothenate (Fig. 2.5). These predictions were mostly based on the absence of two or more of the functional roles KPHMT, KPRED, ASPDC, and PBAL. All Bacteroidetes and 36 (95%) of the Proteobacteria genomes were predicted to be CoA producers. Only three Actinobacteria were predicted CoA producers: two of the Actinomycetales order and one Coriobacteriales. None of the Bifidobacteriales contained a full CoA biosynthesis pathway, but all genomes in the order contained the pantothenate transporter PANF. The four non-producing Coriobacteriales genomes did not contain a PANF transporter. Firmicutes displayed certain class-level similarities, with 32% of the genomes containing the
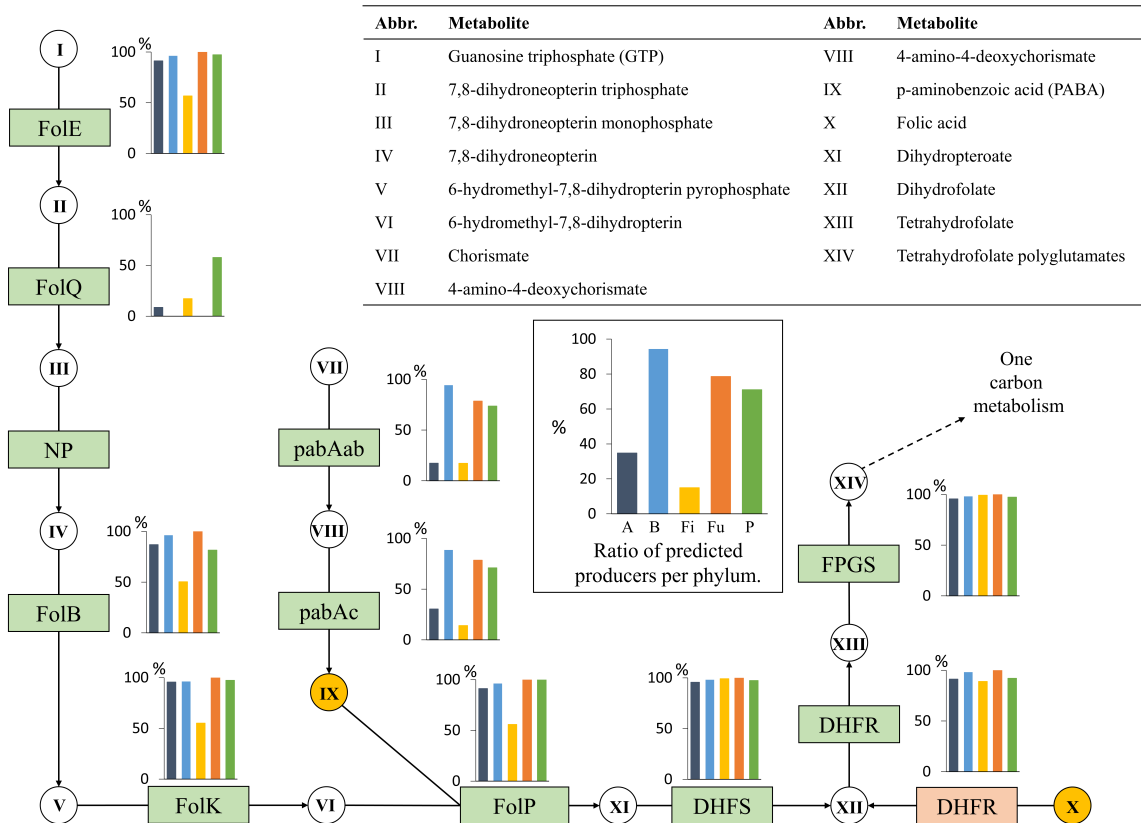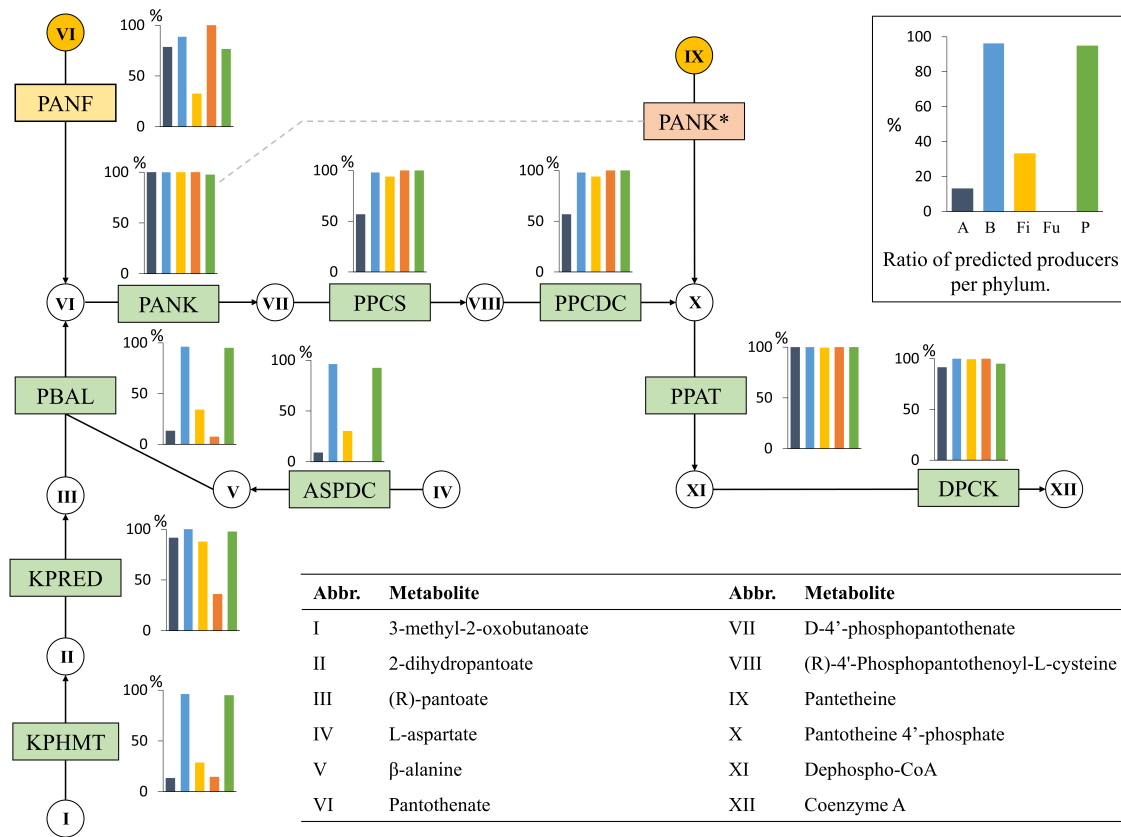
Figure 2.5: Pantothenate biosynthesis. The pantothenate biosynthesis subsystem contains 10 functional roles (Supplementary Table A.1) and 12 metabolites. Refer to Figure 2.1 for figure descriptions.

Abbr.	Metabolite	Abbr.	Metabolite
I	3-methyl-2-oxobutanoate	VII	D-4'-phosphopantothenate
II	2-dihydropantoate	VIII	(R)-4'-Phosphopantothenoyl-L-cysteine
III	(R)-pantoate	IX	Pantetheine
IV	L-aspartate	X	Pantotheine 4'-phosphate
V	β-alanine	XI	Dephospho-CoA
VI	Pantothenate	XII	Coenzyme A

essential roles for CoA biosynthesis. Only seven Bacilli, including three Bacillales and four Lactobacillales, contained all CoA-biosynthesis roles. The Clostridia class demonstrated considerable variability in the presence of roles, and no specific pattern could be observed for the lower taxa. In the phylum Proteobacteria, all organisms of the Betaproteobacteria class except one contained all essential functional roles for CoA biosynthesis, whereas none of the producers presented a PANF transporter. All Delta- and Epsilonproteobacteria were predicted to be producers. All Gammaproteobacteria contained all roles in the pathway, except a single genome of the order Aeromonadales was missing three out of the four necessary steps for pantothenate biosynthesis.

Taken together, these findings demonstrate that the synthesis of CoA from pantothenate is present in nearly all HGM genomes. However, pantothenate biosynthesis is not present in Fusobacteria and is only present in a few Actinobacteria and Firmicutes. Nearly all Bac-

teroidetes and Proteobacteria contain a full biosynthesis pathway for both pantothenate and CoA. CoA biosynthesis is present in 51% of the HGM genomes.

Pyridoxine (Vitamin B6)

The coenzyme form of pyridoxine is pyridoxal 5'-phosphate and is required for several enzymes in the cell, mainly those required for the metabolism of amino acids [46]. Pyridoxal 5'-phosphate can be synthesized *de novo* via two different routes (Fig. 2.6). We first analyzed

Abbr.	Metabolite	Abbr.	Metabolite
I	D-erythrose 4-phosphate	IX	Pyridoxine
II	4-phospho-D-erythronate	X	Pyridoxal 5'-phosphate
III	2-oxo-3-hydroxyl-4-phosphobutanoate	XI	Pyridoxal
IV	O-phospho-4-hydroxyl-L-threonine	XII	Pyridoxamin
V	3-amino-2-oxopropyl phosphate	XIII	Pyridoxamine 5'-phosphate
VI	D-glyceraldehyde 3-phosphate	XIV	L-glutamine
VII	1-deoxy-D-xylulose 5-phosphate	XV	L-ribulose 5-phosphate
VIII	Pyridoxine phosphate	XVI	Glyceraldehyde 3-phosphate

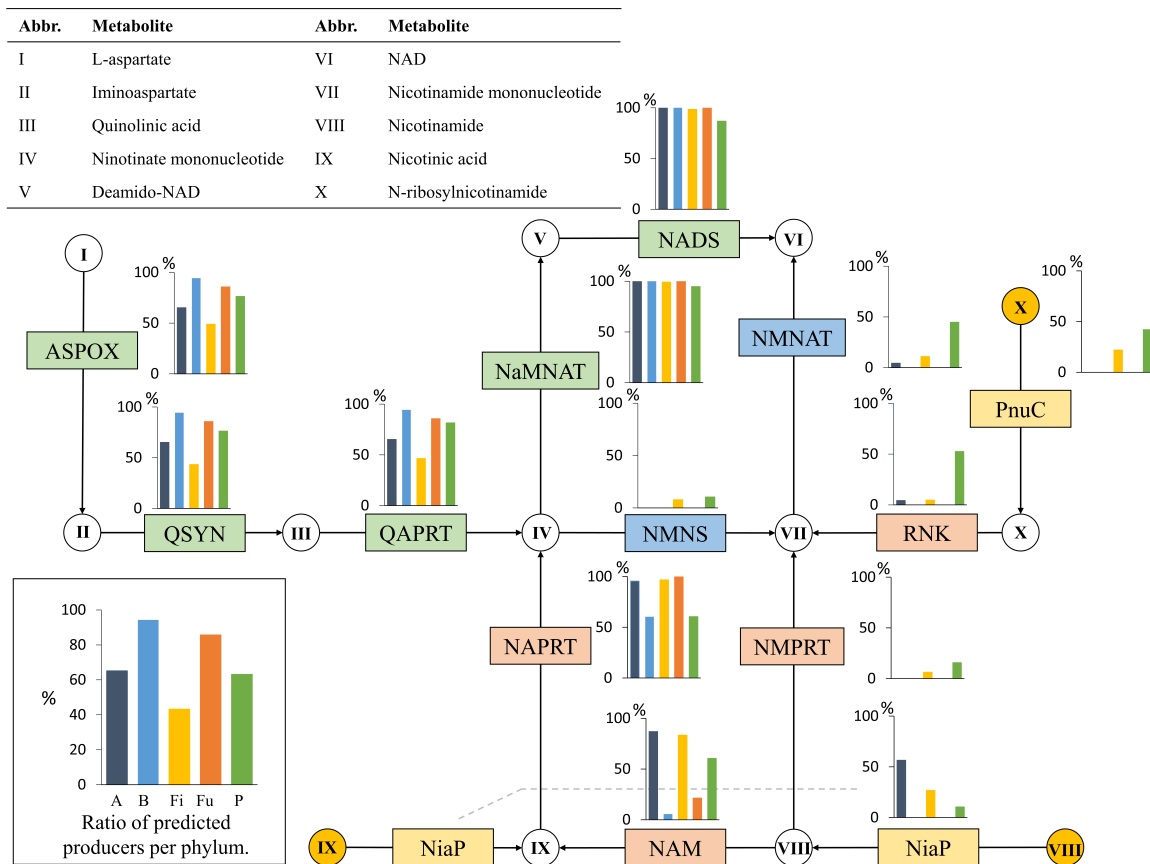Figure 2.6: Pyridoxine biosynthesis. The pyridoxine biosynthesis subsystem contains 11 functional roles (Supplementary Table A.1) and 16 metabolites. Refer to Figure 2.1 for figure descriptions.

the shorter biosynthesis route, which uses a single enzyme with two domains, represented by the functional roles PdxS and PdxT. This route joins glyceraldehyde 3-phosphate and D-ribulose 5-phosphate to produce pyridoxal 5'-phosphate. Both roles must be present for an organism to be a predicted producer. The 20 Actinobacteria and 33 Firmicutes genomes that were predicted to be pyridoxine producers contained the PdxTS roles. Only three Bac-

teroidetes producers contained the PdxTS route: two from the Prevotellaceae family and one Rikenellaceae, *A. indistinctus* YIT 12060, which also contained the alternative biosynthesis route described below. The three Fusobacteria producers contained PdxTS; four Proteobacteria producers (i.e., the three Deltaproteobacteria and a single Gammaproteobacteria, *A. junii* SH205) also contained PdxTS.

We then examined the alternative synthesis route, which requires seven functional roles for a complete pathway made of two branches. One branch converts D-erythrose-4-phosphate to 3-amino-2-oxopropyl phosphate in four steps. The other branch converts D-glyceraldehyde-3-phosphate to 1-deoxy-D-xylulose-5-phosphate with the functional role dxs (Fig. 2.6). The two end-metabolites are joined by PdxJ and converted to pyridoxal 5'-phosphate by PdxH. Both PdxJ and PdxH were required for an organism to be a predicted pyridoxine synthesizer. Only genomes from the Bacteroidetes and Proteobacteria phyla contained the roles PdxJ and PdxH in addition to the remaining essential functional roles of the pathway (Table 2.2); therefore, Bacteroidetes and Proteobacteria were the only phyla predicted to synthesize pyridoxine via this route. All Gammaproteobacteria, with the exception of *Escherichia* sp. 1_1_43, contained a full PdxJH biosynthesis route. Only three Bacteroidetes genomes–*Bacteroides coprocola* DSM 17136, *Bacteroides coprophilus* DSM 18228, and *Bacteroides plebeius* DSM 17135–were predicted to be non-producers. All three genomes were missing PdxH but contained all other functional roles of the pathway.

The results for the two routes showed that the majority of Actinobacteria, Bacteroidetes, and Proteobacteria have the ability to synthesize pyridoxal 5'-phosphate; most Bacteroidetes and Proteobacteria use the longer route, and all Actinobacteria contain the shorter synthesis route. Few Firmicutes and Fusobacteria can synthesize pyridoxal 5'-phosphate, but those that do use the shorter PdxTS route. Taken together, these findings reveal that the *de novo* biosynthesis of pyridoxal 5'-phosphate is present in 50% of the tested HGM genomes. The majority of genomes from all phyla, except for Fusobacteria, contained the role PdxK, which is involved in the salvage of pyridoxal 5'-phosphate and its two precursors.

Riboflavin (Vitamin B2)

Riboflavin can only be synthesized through one known pathway from GTP and D-ribulose-5-phosphate (Fig. 2.7). The role PyrP is commonly missing in our genomes and is com-

monly missing in plants [80]; thus, its absence did not affect our predictions. For an organism to be considered a predicted producer, its genome had to present all other functional roles to produce the redox cofactors FMN and FAD. All Bacteroidetes and Fusobacteria and

| Abbr.	Metabolite	Abbr.	Metabolite
I	Guanosine 5'-triphosphate (GTP)	VII	3,4-dihydroxybutanone 4-phosphate
II	2,5-diamino-6-ribosylamino-4(3H)-pyrimidinone 5'-phosphate	VIII	6,7-dimethyl-8-(D-ribityl)lumazine
III	5-amino-6-ribosylamino-2,4-(1H,3H)-pyrimidinedione 5'-phosphate	IX	Riboflavin
IV	5-amino-6-ribitylamino-2,4-(1H,3H)-pyrimidinone 5'-phosphate	X	Flavin mononucleotide (FMN)
V	5-amino-6-ribitylamino-2,4-(1H,3H)-pyrimidinedione	XI	Flavin adenine dinucleotide (FAD)
VI	D-ribulose 5-phosphate		

Figure 2.7: Riboflavin biosynthesis. The riboflavin biosynthesis subsystem contains 10 functional roles (Supplementary Table A.1) and 11 metabolites. Refer to Figure 2.1 for figure descriptions.

36 genomes (92%) of Proteobacteria contained all essential functional roles for riboflavin biosynthesis (Fig. 2.7, Table 2.2). Half of the Firmicutes were predicted to be riboflavin producers. This ratio is the largest in Firmicutes producers among all of the vitamins in our analysis. Although riboflavin biosynthesis was the most prominent in Firmicutes, no taxa-specific patterns were found. Only two genomes in the Actinobacteria phylum had all the essential roles for riboflavin biosynthesis: *Corynebacterium ammoniagenes* DSM 20306 and *Bifidobacterium longum* ATCC 15697. All Bacteroidetes were predicted producers. All but two genomes from Proteobacteria were predicted to synthesize riboflavin.

The riboflavin biosynthesis pathway was the most preserved of the eight vitamins we

examined, although it was absent in most of the Actinobacteria genomes and half of the Firmicutes genomes. However, all non-producing organisms from the two phyla contained the RibU riboflavin transporter role, indicating their need for the riboflavin-derived cofactors FMN and FAD. RibU was almost completely absent in the Bacteroidetes, Fusobacteria, and Proteobacteria, whereas the *de novo* synthesis pathway was found in nearly all genomes of the three phyla. The majority of all 256 genomes (65%) had the ability to produce riboflavin.

Thiamin (Vitamin B1)

The thiamin biosynthesis pathway consists of two branches that are joined in the final step to produce thiamin monophosphate (Fig. 2.8) [110]. All genomes in our analysis contained the

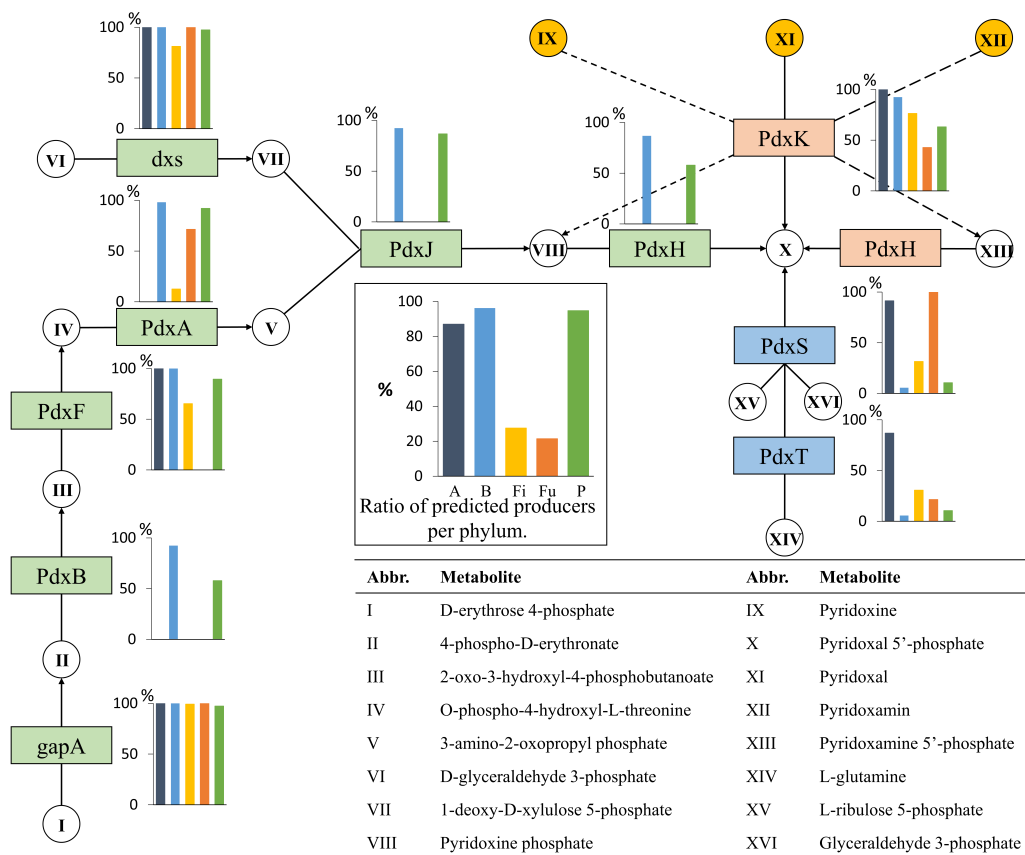| Abbr.	Metabolite
I	L-glycine
II	L-glycine imine
III	L-tyrosine
IV	Sulfur carrier protein ThiS
V	1-deoxy-D-xylulose-5-phosphate
VI	4-methyl-5-(2-phosphoethyl)-thiazole
VII	Aminoimidazole ribotide
VIII	4-amino-2-methyl-5-phosphomethylpyrimidine
IX	4-amino-2-methyl-5-diphosphomethylpyrimidine
X	Thiamin monophosphate
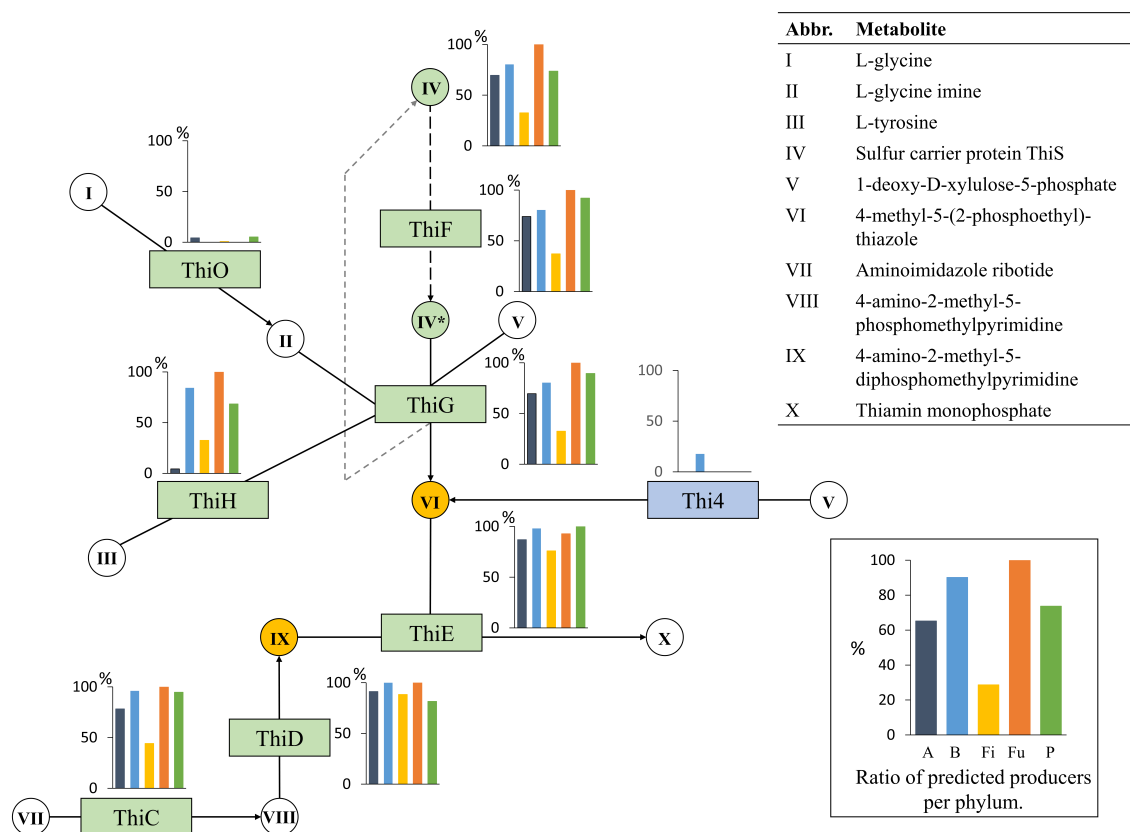
Figure 2.8: Thiamin biosynthesis. The thiamin biosynthesis subsystem contains 9 functional roles (Supplementary Table A.1) and 10 metabolites. One functional role, sulfur carrier protein ThiS, is drawn as a metabolite because it is combined with other metabolites and recycled in the pathway. Refer to Figure 2.1 for figure descriptions.

ThiCD functional roles. The Thi4 role was only found in nine genomes, which were all from

the Bacteroidetes phylum. The second HET-P synthesis route requires four functional roles: ThiS, ThiG, ThiF, and ThiO or ThiH. The roles ThiO and ThiH convert glycine and tyrosine, respectively, to glycine imine. The role ThiS carries a sulfur group to ThiG, which combines the sulfur with glycine imine and DXP to produce HET-P. Most predicted producers in our analysis used the ThiH role except for four genomes that contained ThiO (two from the Proteobacteria, one from the Actinobacteria, and one from the Firmicutes phylum). Only two Actinobacteria out of its 15 producers contained the ThiH or ThiO part of the pathway; because their genomes contained all other functional roles, they were predicted to be producers. In the final step, HET-P and HMP-PP are combined to produce thiamin monophosphate by ThiE. Although thiamin diphosphate is the functional version of thiamin, the production of thiamin monophosphate was considered sufficient for producer status. Only a single Bacteroidetes genome, *Prevotella salivae* DSM 15606, was missing essential functional roles for thiamin biosynthesis. In the Firmicutes phylum, no Bacilli received a producer status, whereas half of the Clostridia class genomes were predicted to synthesize thiamin. In Proteobacteria, all but one Gammaproteobacteria genome were predicted to be producers. The remaining seven producers from the phylum belonged to the Beta-, Delta-, and Epsilonproteobacteria classes.

Taken together, these results show that the synthesis of thiamin monophosphate is present in the majority of all phyla, except for Firmicutes, and this synthesis is most prevalent in Bacteroidetes and Fusobacteria. Only a few genomes in the Bacteroidetes phylum contain the role Thi4, which can replace the longer HET-P biosynthesis route. The ThiH role in the longer HET-P synthesis route is most commonly used by HGM organisms; in contrast, the ThiO role is rarely present and is only present in Actinobacteria and Proteobacteria genomes. In total, 56% of the HGM organisms have the ability to produce thiamin.

2.3.2   Comparison with experimental data

To assess the validity of our predictions, we compared our results with existing experimental data in the literature. Here, we considered only experimental studies that used defined media added specific B-vitamins for growth, or that tested for B-vitamin requirements or secretions specifically. We found experimental evidence of B-vitamin requirements or secretions for

11 of our HGM species. However, for four of these species, experimental data were not found for the specific species that we analyzed. In these cases, we compared the data to our analyzed strains while keeping in mind that strains within a species can differ substantially [132]. In total, we made eight predictions for each of the 16 strains, resulting in 128 predictions. We found that 113 of our predictions matched the experimental data (88%, Table 2.4). This result suggests that predictions concerning bacterial metabolism can be made based on well-annotated genomes.

Table 2.4: Comparison of genomic predictions and experimental evidence.

Organism	B	C	F	N	P	B6	R	T	Ref.
Bacteroides fragilis NCTC 9343 (272569.17)	●	□	●	●	●	●	●	●	[235]
Bacteroides thetaiotaomicron VPI-5482 (226186.1)	●	●	●	●	●	●	●	●	[235]
Bacteroides vulgatus ATCC 8482 (435590.6)	●	□	●	●	●	●	●	●	[235]
Clostridium difficile CD196 (645462.3)[a]	●	●	●	●	□	●	●	●	[112]
Clostridium difficile NAP07 (525258.3)[a]	●	●	●	●	□	●	●	●	[112]
Clostridium difficile NAP08 (525259.3)[a]	●	●	●	●	□	●	●	●	[112]
Escherichia coli str. K-12 substr. MG1655 (511145.6)	●	●	●	●	●	●	●	●	[49]
Faecalibacterium prausnitzii A2-165 (411483.3)	●	□	●	●	●	●	●	●	[95]
Faecalibacterium cf. prausnitzii KLE1256 (748224.3)[b]	□	●	●	●	●	●	□	●	[95]
Faecalibacterium prausnitzii L2-6 (718252.3)[b]	●	□	●	●	●	●	□	●	[95]
Faecalibacterium prausnitzii SL3/3 (657322.3)[b]	●	□	●	●	●	●	□	●	[95]
Helicobacter pylori 26695 (85962.1)	●	●	●	●	●	●	●	●	[222]
Klebsiella pneumoniae 1162281 (1037908.3)[c]	●	●	●	●	●	●	●	●	[7]
Lactobacillus plantarum WCFS1 (220668.1)	○	●	○	●	●	●	●	○	[240]
Listeria monocytogenes str. 1/2a F6854 (267409.1)[d]	●	●	●	●	●	●	●	●	[231]
Salmonella enterica subsp. *enterica* serovar Typhimurium str. TN061786 (946034.3)	●	●	●	●	●	●	●	□	[210]

B: biotin, C: cobalamin, F: folate, N: niacin, P: pantothenate, B6: pyridoxine, R: riboflavin, T: thiamin. Filled circles represent agreeing predictions and data, empty shapes represent mismatches. ●: predicted producers, ●: predicted non-producers, □: predicted producers, ○: predicted non-producers. Footnotes list the comparative strain in the cases where the predicted and experimental strains are different.
[a]*C. difficile* VPI 10463, KZ 1626, K 1630, KZ 1647, and KZ 1748.
[b]*F. prausnitzi* A2-165.
[c]*K. pneumonia* Kp1 and Kp102M.
[d]*L. monocytogenes* 10403, EGD-e, and L028.

2.3.3   B-vitamin synthesis Patterns in HGM and other Microbial Genomes

Examining the variation of the different synthesis pathways in our former analysis, we wondered how the combinations of the vitamins synthesized varied across the HGM genomes. Our data consist of binary information regarding the distribution of pathways in 256 HGM and 257 non-HGM genomes, i.e., the presence or absence of a vitamin biosynthesis pathway in a genome (Fig. 2.9). We investigated the 2^8 = 256 possible patterns of the eight studied pathways. Only 68 (27%) of the 256 possible pathway patterns were found in the 256 HGM genomes (Supplementary Table A.1), suggesting that the occurrences of these B-vitamin pathways may not be independent. To determine whether the pattern occurrence
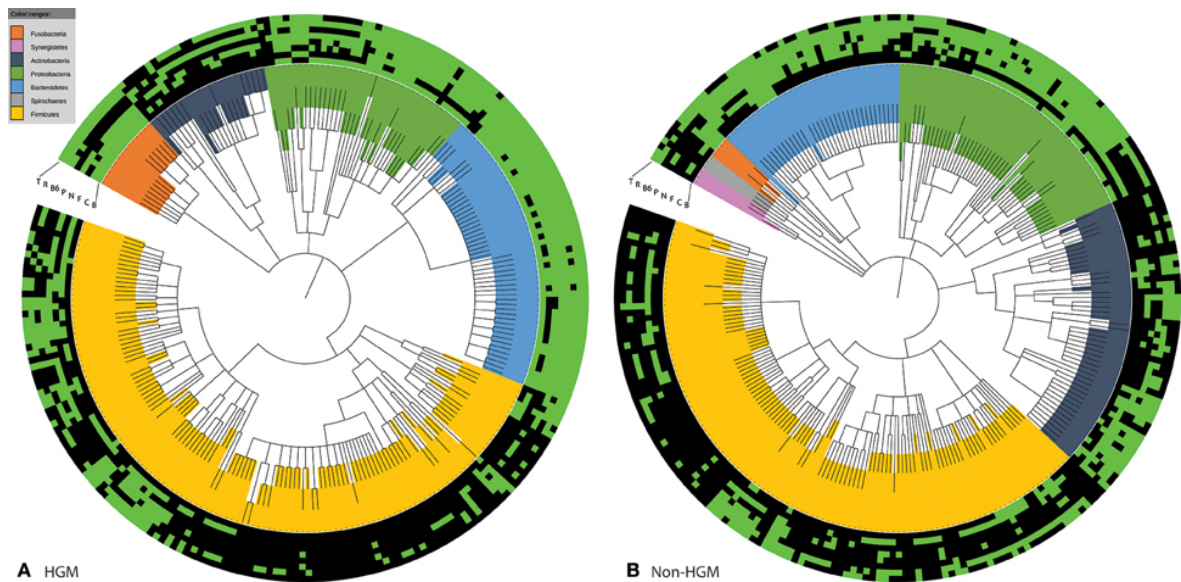
Figure 2.9: NCBI based taxonomic tree and the prsence or absence of the eight B-vitamin biosynthesis pathways. The two taxonomic trees show the (A) 256 HGM genomes and (B) the 257 non-HGM genomes, along with the heatmaps showing the presence (green) or absence (black) of each vitamin pathway. The taxonomic trees were produced using PhyloT: a tree generator (http://phylot.biobyte.de/index.html) and visualized through iTOL (http://itol.embl.de/) [127, 128].

and distribution was HGM specific or a general feature of bacteria, we inspected the B-vitamin metabolism distribution in 257 bacterial genomes (non-HGM). We identified four groups: (1) 32 HGM-specific patterns, (2) 36 patterns present in both HGM and non-HGM sets, (3) 36 patterns found only in non-HGM genomes, and (4) 152 patterns found in neither. The 32 HGM-specific patterns were present in 68 (27%) of the HGM genomes, and each pattern was found in eight or fewer genomes. Three HGM-specific patterns were found in

eight genomes each: (i) the presence of only niacin, pyridoxine, and thiamin pathways was found only in Actinobacteria, (ii) the absence of all pathways except for niacin was found in six Firmicutes and two Actinobacteria genomes, (iii) the presence of all pathways except for biotin and folate biosynthesis was found in five Firmicutes and three Proteobacteria genomes. The remaining HGM-specific patterns were found in four or fewer genomes each. The presence of these HGM-specific vitamin biosynthesis patterns suggests that human gut microbes have evolved to suit gut-specific conditions, resulting in complementary combinations of synthesized B-vitamins. Accordingly, we observed ten pairs of inversed patterns, in which pathways that were present or absent in one genome were absent or present in another, respectively (Fig. 2.10). Such pairs of organisms may be candidates for mutualistic or symbiotic interactions. For example, several Bacteroidetes and Proteobacteria genomes contained all biosynthesis pathways, except for cobalamin synthesis. In the complementary pattern, found in eight Firmicutes genomes, all biosynthesis pathways, except for cobalamin synthesis, were absent. In comparison, only five pairs of inversed patterns were observed in the analyzed non-HGM genomes. We analyzed the distribution of pathway patterns of the HGM genomes and noted similarities in the eight vitamin biosynthesis capabilities at the phylum, class, order, and family levels. Genomes in the phylum Actinobacteria demonstrated variations in pathway distributions; however, the most conserved pattern, the biosynthesis of niacin, pyridoxine, and thiamin, was present in eight (35%) genomes, and all of these genomes were from the order Bifidobacteriales and the family Bifidobacteriaceae (Fig. 2.9, Supplementary Table A.1). Pathway distribution was highly conserved in the phylum Bacteroidetes, which displayed only six different patterns. The two most common patterns in the Bacteroidetes phylum were the biosynthesis of all eight vitamins, which was present in 26 (51%) of the genomes in the phylum, and the synthesis of all vitamins except for cobalamin, which appeared in 17 (33%) of the Bacteroidetes genomes. Three genomes could not produce cobalamin and pyridoxin, and three did not meet the requirements for consideration as cobalamin and folate producers. A single organism in the Bacteroidetes phylum could produce all vitamins except for biotin and cobalamin. One genome of the family Prevotellaceae, *P. salivae* DSM 15606, was missing four biosynthesis pathways, and this organism only produced pantothenate, pyridoxine, and riboflavin. The phylum Firmicutes demonstrates the highest variability in pathway patterns, containing 43 patterns in 130 genomes.
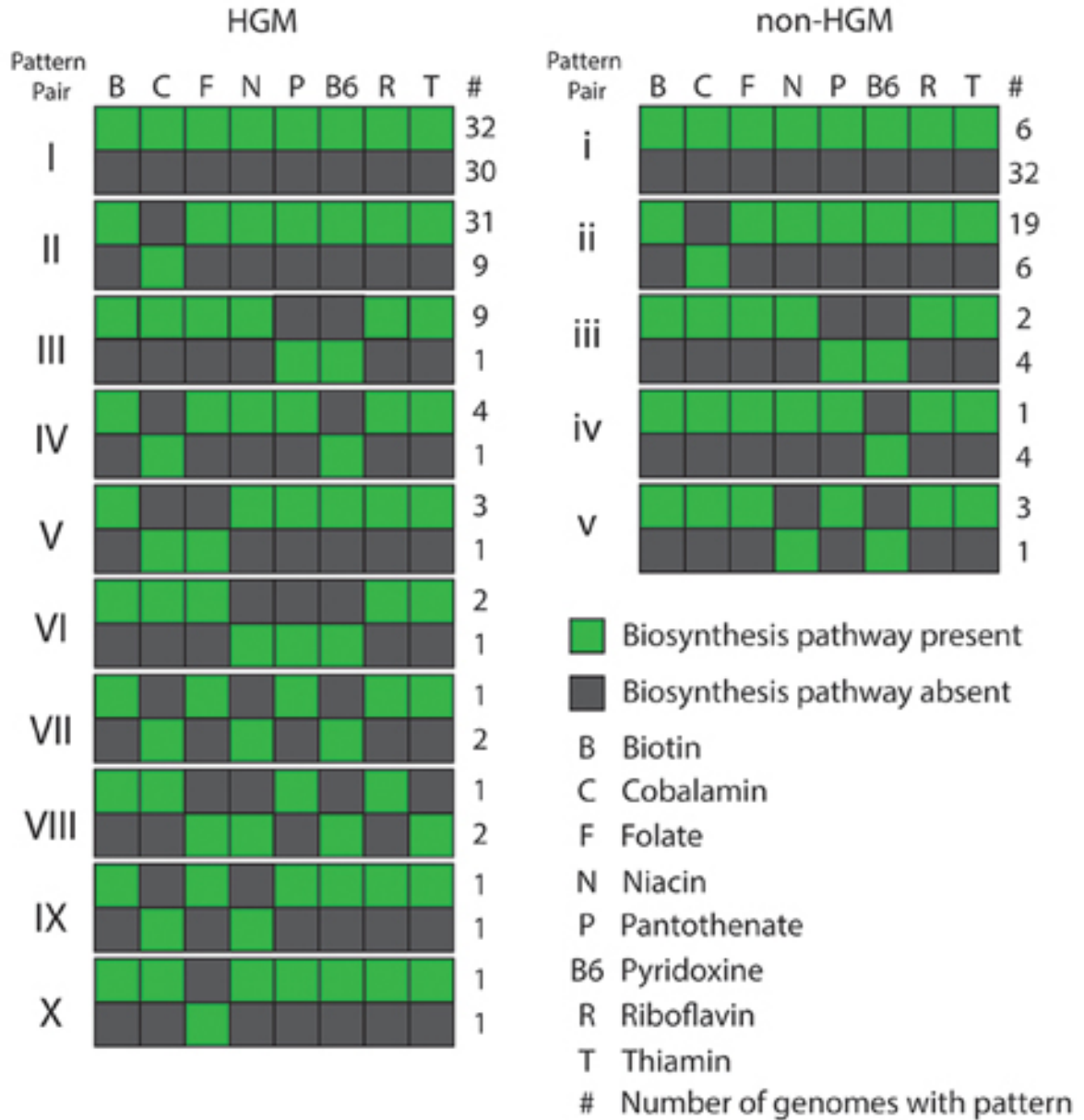
Figure 2.10: Inversed pattern pairs of HGM and non-HGM genomes. The HGM genomes showed ten pairs of inversed patterns, whereas the non-HGM contained 5 pairs.

However, this variability is not surprising in light of the taxonomic diversity within the phylum itself. The most common pattern was the absence of all biosynthesis pathways and was found in 30 genomes, 26 of which belong to the Lactobacillales order of the Bacilli class. The Lactobacillales order also contained 9 out of 10 Firmicutes genomes that contained only riboflavin biosynthesis. Genomes of the Clostridia class showed an even larger diversity of pathway patterns; these genomes contained 28 out of the 43 Firmicutes patterns. The most common Clostridia pattern, the absence of all synthesis pathways except cobalamin, was

found in seven organisms. All the analyzed Fusobacteria genomes belonged to the family Fusobacteriaceae and displayed conserved patterns of pathways. The most common pattern in nine of the 14 genomes was the absence of only the pantothenate and pyridoxine synthesis pathways. The pantothenate biosynthesis pathway was absent in all of the genomes, whereas all genomes contained the biosynthesis pathways for biotin, cobalamin, riboflavin, and thiamin. The phylum Proteobacteria showed surprisingly conserved patterns in light of the high divergence of the taxon. The class Gammaproteobacteria contained 12 out of the 13 Proteobacteria genomes that contained the biosynthesis of all vitamins except for cobalamin. The only other genome to show this pattern was the Betaproteobacteria *Ralstonia* sp. 5_7_47FAA. The remaining 14 patterns were only found in six or fewer genomes in the phylum, and ten patterns were observed in only a single organism each. Finally, few pathway patterns were shared between the five phyla. The presence of all pathways, except for cobalamin, was observed in Bacteroidetes, Firmicutes, and Proteobacteria in 17, 1, and 13 genomes, respectively. Only nine out of the 68 patterns observed in the HGM genomes were shared between two phyla, whereas 58 patterns appeared in a single phylum each.

Taken together, the distribution of pathway patterns was relatively conserved, with only 68 patterns appearing in the 256 HGM genomes. A third of these patterns (20 out of 68) were the inverse of another pattern, suggesting the existence of symbiotic relationships for B-vitamin metabolism in the gut microbiota.

2.3.4   Amount of HGM B-vitamins Available to the Gut

Considering that the B-vitamin biosynthesis pathways seem to be prevalent in the human gut microbiota genomes, we wondered whether gut bacteria have the collective capacity to produce sufficient B-vitamins for daily human requirements. The assumptions underlying the calculation are described in the Method Section. Overall, only four of the eight B-vitamins are predicted to be produced in amounts that could cover at least a quarter of the suggested dietary intake (Table 2.5). It must be noted that the calculated values are speculative and do not represent the true amount of B-vitamins provided by the human gut microbiota.

Table 2.5: Estimated maximal percentage of daily reference intake of the eight vitamins that could be provided by the human gut microbiota.

Vitamin	Intracellular concentration [mmol/gDW]	DRI[a] [mg/dayl]	HGM$_{ratio}$	% DRI from HGM
Biotin	9.0×10^{-7}	0.03	0.40	4.5
Cobalamin	8.5×10^{-8}	0.0024	0.42	31
Folate[b]	5.0×10^{-5}	0.4	0.43	37
Niacin[b]	3.3×10^{-3}	15	0.63	27
Pantothenate	2.3×10^{-6}	5	0.51	0.078
Pyridoxine[b]	5.8×10^{-4}	1.3	0.50	86
Riboflavin	9.0×10^{-6}	376.4	0.65	2.8
Thiamin[b]	8.7×10^{-6}	1.15	0.56	2.3

[a]Dietary reference intakes [214]. Values averaged for male and female references intakes (ages 19-50).
[b]Atomic mass for dihydrofolic acid, nicotinic acid, pyridoxine 5'-phosphate, and thiamine monophoshate.

2.4   Discussion

In this study, we predicted the B-vitamin biosynthesis of 256 known human gut microorganisms based on their genome annotations alone. Our key results are the following: (i) the majority of our genome-based predictions match published experimental data; (ii) B-vitamin biosynthesis is common in the HGM; and iii) inversed patterns of vitamin synthesis suggest symbiotic relationships among HGM organisms with regard to B-vitamins. Taken together, our data support the idea that the human gut microbiota has co-evolved relationships that are specific to the gut environment.

The predictions of B-vitamin synthesis capability agreed well with experimental data suggesting that the genomic analyses of well-defined pathways can provide much information regarding strain-specific metabolism. The prediction had a 12% error rate, which represents a lower bound for the other HGM capabilities (Table 2.4) as we compared against well-studied organisms with a higher genome completeness than expected for genomes assembled from metagenomic studies. One example of a wrong prediction is *L. plantarum* WCFS1, in which all necessary functional roles of the biotin pathway were missing. Because biotin is not required for its growth [240], either the organism does not need biotin or the manual curation of the pathway requires additional data. The *L. plantarum* WCFS1 genome contains the biotin-dependent protein biotin-protein ligase (Supplementary Table A.1), which

suggests a need for biotin. All analyzed *C. difficile* genomes were predicted to produce pantothenate as they contained all essential roles except ASPDC. However, pantothenate is required in the growth medium of the comparison strain [112]. The strain difference may explain the discrepancy but it is also possible that our essential role constraints are too loose and that the absence of ASPDC should lead to a negative prediction, even when it is the only missing functional role in the pathway. The four *F. prausnitzii* genomes were missing all essential roles for folate production and were therefore predicted to require folate in their growth media. In contrast, no significant folic acid consumption of *F. prausnitzii* A2-165 has been reported [95] and additionally, it can grow in folic acid free medium (personal communication, Dr. Delphine Saulnier). However, the genome of *F. prausnitzii* A2-165 contains an annotated folate transporter [95] and the folate-dependent methionyl-tRNA formyltransferase (UniProt ID: C7H5H7). We propose that *F. prausnitzii* A2-165 does not require folates for growth, although the presence of folate in the growth medium might enhance the growth rate. *L. plantarum* WCFS1 is missing all PABA functional roles while it has been shown to synthesize folate in the absence of folate precursors and PABA in the growth medium [240]. This contradiction suggests an unknown route for PABA biosynthesis, which is likely considering that two missing archaeal enzyme families in the folate biosynthesis pathway had recently been discovered [52].

The two most commonly synthesized vitamins of the human gut microbiota genomes were riboflavin and niacin, with 166 and 162 predicted producers, respectively. The gut-microbial production of riboflavin has been associated with the immune response through the activation of T-cells [117]. Riboflavin has also been shown to be involved in the extracellular electron transport chain in *F. prausnitzii* [115]. The conservation of the riboflavin biosynthesis pathway can be explained by the exclusive importance of its derivatives, FAD and FMN, because approximately 1–3% of cellular proteins are flavoproteins [50, 1]. Niacin, a precursor for NAD, is another essential cofactor. NAD and its reduced and phosphorylated derivatives (NADH, NADP, and NADPH) have many functions in cells, such as serving as hydride donors and acceptors in redox reactions [22, 107], participation in bacterial and DNA ligase reactions [241], and roles as molecules that signal the cellular redox status [107]. However, the two synthesis pathways are distributed differently over the five phyla. Riboflavin synthesis is mainly found in Bacteroidetes, Proteobacteria, and Fusobacteria, but it is only

found in half of the Firmicutes genomes and very few Actinobacteria (Fig. 2.7). In contrast, the niacin biosynthesis pathway is more evenly distributed over the genomes of the five phyla (Fig. 2.4). Such differences between the distributions of these two pathways can have various explanations. First, this variation may reflect their evolutionary history; riboflavin synthesis appears to be more ancient than the NAD biosynthesis pathway [78]. Second, the biosynthesis of riboflavin and its derivatives is a quite straightforward pathway [1], whereas the biosynthesis of NAD is very complex and includes numerous versions of salvage pathways in various bacterial taxa [124, 81, 77, 34].

The human gut microbiota is a large microbial community, in which metabolites are shared among individual microorganisms [245, 31]. The production of all vitamins and the absence of all biosynthesis pathways was one of the most common patterns in the HGM (Fig. 2.10). The absence of all biosynthesis pathways was only found in the Firmicutes phylum, whereas the presence of all eight pathways was found mainly in the Bacteroidetes and in several Proteobacteria (Fig. 2.9, Supplementary Table A.1).This distribution of vitamin biosynthesis patterns is interesting given that a higher ratio of Bacteroidetes to Firmicutes has been linked to a healthier gut microbiota compared with that in obese individuals [234]. Inversed pattern pairs in HGM was twice as frequent than in non-HGM, and 70% more genomes were associated with these pairs in the HGM compared to non-HGM (Fig. 2.10). These results suggest that the trait of sharing B-vitamins has evolved in the gut microbes.

HGM can supply the host with biotin, folate, and riboflavin [105, 186]. Human colonocytes contain transporters for several B-vitamins, such as RFT1 for riboflavin [250], FOLR1 for folate [45], and the multivitamin transporters SMCT1 [131] and SMVT [167]. It has therefore been speculated that the HGM contributes to B-vitamin homeostasis [191] because most of the dietary vitamin absorption occurs in the small intestine. The large intestine contains the highest density of microbes in the human gut; thus, it seems likely that bacteria provide human intestinal cells with sufficient B-vitamins to avoid deficiencies during short periods of vitamin-poor diet. According to our estimation, the gut microbiota would not be able to provide the host with the daily recommended intake of B-vitamins (Table 2.5) and a large portion of these vitamins may be taken up by non-producing gut microbes, which compete with the host. In fact, our results indicate that human gut microbes actively synthesize B-vitamins and provide them to their neighboring bacteria through symbiotic relationships.

A high fraction of the recommended pyridoxine intake and about a third of the recommended intake of cobalamin and folate could be available from the gut bacteria (Table 2.5). Consistently, these vitamins are thought to be produced by the human gut microbiota in addition to dietary uptake [103, 187, 126]. In fact, pyridoxine deficiency due to dietary deficiency is rarely observed [188]. Cobalamin and folate deficiencies are mainly reported in elderly people [103, 8, 38], who have been shown to have a less diverse and metabolically active microbiota [246, 36]. In addition, cobalamin and folate have several metabolites derived beyond their basic forms, and the host may not absorb these compounds. For example, it has been shown that little of the corrinoids found in human feces are gut-bacteria derived [5], and it has been proposed that the most of the corrinoids produced by the gut bacteria are taken up by non-producers in the gut [54]. Vitamin B12 has recently been suggested as a modulator of gut microbial ecology [54]. We propose that other B-vitamins could also have a gut microbiota-modulating role. In this study, we have not analyzed vitamin transporters because little is known about the B-vitamin export from bacterial cells [184]. However, if B-vitamins are released into the gut during cell lysis, any vitamin producer can be considered a vitaminsource, regardless of its transporter expression. Taken together, our results highlight the fact that the microbiota does indeed contribute to the B-vitamin pool of the gut and that the host can benefit to some extent from the B-vitamin biosynthesis of the microbiota.

Acknowledgements

Chapter 3

Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota

Abstract

Genome-scale metabolic models derived from human gut metagenomic data can be used as a framework to elucidate how microbial communities modulate human metabolism and health. We present AGORA (assembly of gut organisms through reconstruction and analysis), a resource of genome-scale metabolic reconstructions semi-automatically generated for 773 human gut bacteria. Using this resource, we identified a defined growth medium for *Bacteroides caccae* ATCC 34185. We also showed that interactions among modeled species depend on both the metabolic potential of each species and the nutrients available. AGORA reconstructions can integrate either metagenomic or 16S rRNA sequencing data sets to infer the metabolic diversity of microbial communities. AGORA reconstructions could provide a starting point for the generation of high-quality, manually curated metabolic reconstructions. AGORA is fully compatible with Recon 2, a comprehensive metabolic reconstruction of human metabolism, which will facilitate studies of host–microbiome interactions.

3.1 Introduction

Changes in the composition of the human gut microbiota have been associated with the development of chronic diseases including type 2 diabetes, obesity, and colorectal cancer [39]. Gut bacterial functions, such as synthesis of amino acids and vitamins [205], breakdown of indigestible plant polysaccharides [59], and production of metabolites involved in energy metabolism [60], have been linked to human health. The use of 'omics approaches to study human microbiome communities has led to the generation of enormous data sets whose interpretations require systems biology tools to shed light on the functional capacity of gut microbiomes and their interactions with the human host [66].

In order to infer the metabolic repertoire of a gut metagenome data set, researchers usually map sequenced genes or organisms onto metabolic networks derived from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [111], and functional annotations from KEGG ontologies [87]. However, this approach cannot identify the contribution of each bacterial species to the metabolic repertoire of the whole gut microbiome, nor can it infer the effects of different gut microbial communities on host metabolism.

A technique that can bridge this gap is constraint-based reconstruction and analysis (COBRA) [201] using genome-scale metabolic reconstructions (GENREs) of individual human gut microbes. GENREs are assembled using the genome sequence and experimental information [226]. These reconstructions form the basis for the development of condition-specific metabolic models whose functions are simulated and validated by comparison with experimental results. The models can be used to investigate genotype–phenotype relationships [148], microbe–microbe interactions [99], and host–microbe interactions [99]. Numerous tools can be used to automatically generate draft GENREs but such models contain errors [90] and are incomplete. Manual curation of draft reconstructions is time consuming because it involves an extensive literature review and experimental validation of metabolic functions [226].

To provide an extensive resource of GENREs for human gut microbes, we developed a comparative metabolic reconstruction method that enables any refinement to one metabolic reconstruction to be propagated to others. This accelerates reconstruction and improves model quality. We generated AGORA, which includes 773 gut microbes, comprising 205

genera and 605 species. All reconstructions were based on literature-derived experimental data and comparative genomics. The metabolic predictions of two AGORA reconstructions and their derived metabolic models were validated against experimental data.

3.2 Methods

Access to AGORA reconstructions.

AGORA is an ongoing resource of GENREs and that keeps abreast of human gut microbial knowledge and can be easily accessed via the Virtual Metabolic Human (VMH) database website (http://vmh.life), which allows querying each reconstructions' content and data gathered from the literature search performed in this study. Any AGORA reconstruction can be downloaded in SBML format from the VMH website on the "Downloads" page. The website also hosts the human metabolic reconstruction, Recon 2 (ref. [227]). New content can be added to an AGORA reconstruction manually or automatically, for example, using rBioNet [230], which is compatible with the COBRA toolbox [201] and ensures all QC/QA measures defined by the community as described by Thiele and Palsson (2010) [226]. The "Feedback" tab provides contact information to the VMH developers regarding improvements to any resources available on the VMH webpage, where additions to any resource will be upheld by the VMH developers.

Genome selection and draft reconstructions.

In a previous study [18], we retrieved 301 draft reconstructions from Model SEED [102]. Based on lists of human gut microbes reported by Qin et al. (2010) [168] and by Rajilić-Stojanović and de Vos (2014) [172], we obtained 472 additional draft reconstructions from Model SEED and KBase [9], both of which use the RAST annotation server [12] to annotate the genomes and build the draft metabolic networks [102]. All reconstructions were downloaded in SBML format and imported into Matlab (Mathworks, Inc., Natick, MA, USA) using the COBRA Toolbox [201]. Each reconstruction was refined using the rBioNet extension [230] to the COBRA Toolbox. We manually translated reaction and metabolite names into the VMH nomenclature (Supplementary Tables B.8 and B.9).

In silico simulations.

All simulations were performed in Matlab using the COBRA Toolbox [201] (https://opencobra.github.io/) and the linear programming solver CPLEX (IBM, Inc.) through the Tomlab interface (Tomlab, Inc.).

The curation process.

Note that we refer to a model, which was derived from the corresponding reconstruction, whenever simulations under a specified condition were carried out.

Reaction directionalities.

To ensure consistency with published reconstructions, the direction of each reaction in a draft reconstruction was set in agreement with the VMH database. This curation prevented fluxes from occurring in thermodynamically unfavorable directions. In several cases, the change of directionality resulted in blocked reactions and/or resulted in a nonzero flux through the biomass objective function (BOF) [69]. For example, the reaction alpha,alphatrehalose-phosphate synthase (VMH ID: TRE6PS) has been reported to be irreversible [29], but was reversible in each draft reconstruction and was required for the synthesis of UDP-glucose. We corrected this by adding the enzyme that synthesizes UDP-glucose (UTP-glucose-1-phosphate uridylyltransferase, VMH ID: GALU) after setting the reaction TRE6PS to be irreversible. In a similar manner, we manually identified solutions to the other blockages and added appropriate corrections to the reconstructions. Note that when an AGORA model is used to represent a bacterium within a particular part of the intestine, context-specific parameters (temperature, pH, ionic strength, cytoplasmic electrical potential difference, and metabolite concentrations) should be used when checking for consistency between the direction of each reaction and those obtained by context-specific thermodynamic estimates [155] (Supplementary Note B.2.1).

Anaerobic growth.

The majority of intestinal microbes are strict or facultative anaerobes, while strict aerobes colonizing the gut are rare [64]. A total of 192 draft-reconstruction-derived models could not

carry flux through the BOF on anaerobic rich medium (all exchange reactions open, aside from the oxygen exchange reaction). Anaerobic growth was enabled by adding oxygeninde-pendent reactions for gene products known to be functional under anaerobic conditions. For example, L-aspartate oxidase (EC 1.4.3.16) functions under both aerobic and anaerobic con-ditions [146]; therefore the anaerobic, fumarate-using L-aspartate oxidase reaction (VMH ID: ASPO5) was added to those reconstructions containing the oxygen-using L-aspartate oxidase reaction (VMH ID: ASPO6).

Removal of infeasible flux loops.

Futile cycles are sets of reactions that result in thermodynamically infeasible fluxes and are a common problem in reconstructions with many transport reactions [226]. Each draft model had an unfeasibly high export flux of protons from the cytosol, resulting in biologically implausibly high ATP production (average flux of 933 ± 229 mmol/gDW/h in the absence and of 933 ± 227 mmol/gDW/h in the presence of oxygen, on Western diet, Supplementary Table B.18). We identified futile cycles by constraining each reaction flux to zero individ-ually and computed the flux through the ATP demand reaction using flux balance analysis (FBA) [159]. No flux was forced through the BOF. Each deleted reaction lowering the ATP demand flux was inspected manually and replaced by an appropriate irreversible reaction in all reconstructions containing that futile cycle. If such change prevented the model from producing biomass, the change was reversed and another reaction eliminating the futile cy-cle was identified (Supplementary Fig. B.8). After the curation, the average ATP production flux was 19 ± 13 mmol/gDW/h in the absence and 38 ± 23 mmol/gDW/h in the presence of oxygen.

Curation of fermentation pathways.

An extensive literature search on the distribution and the structure of fermentation pathways in the considered gut microbes was performed on the genus level and, where possible, on the species level (Supplementary Note B.2.4). Drawing on information from two books and 112 publications, we curated 28 fermentation pathways, of which 20 were either absent from or nonfunctional in all draft reconstructions. Published information on fermentation products was available for 765 out of 773 reconstructed microbes (Supplementary Table B.1)

Curation of carbon source utilization pathways.

A thorough literature search was performed to assess the carbon source utilization on a species and strain level (Supplementary Note B.2.4). Pathways for 95 distinct carbon sources were added to the reconstructions based on evidence from two books [62, 122] and 199 publications (Supplementary Table B.1). The draft reconstructions captured 35 of the 95 carbon sources but did not contain any of the 31 oligo- and polysaccharides (Fig. 3.1e). Published information about carbon source utilization capabilities was available for 731 out of 773 reconstructed microbial strains (Supplementary Table B.1). To verify the functionality of new fermentation and carbon source utilization pathways, flux variability analysis (FVA) [89] was performed on each refined model with all exchange reactions open and no flux forced through the BOF.

Comparative genomics analysis.

We used the results of two previous studies on respiratory reductases [174] and B-vitamin synthesis pathways [140]. We extended the previous analyses to the 612 of our 773 organisms that were available in PubSEED [13]. In the case of missing annotations, a similarity search was performed with a BLAST algorithm implemented in PubSEED (cutoff = e^{-20}). In ambiguous situations, phylogenetic trees and genomic context for the corresponding proteins were analyzed as follows. The neighbor-joining approach implemented in PubSEED was used (default parameters) to construct the phylogenetic trees. Analysis of the genomic context was done using the tools available in PubSEED. Incorrect and inaccurate PubSEED annotations were edited and the corresponding genes were added to the relevant subsystems. Reactions were formulated for all pathways as needed and added to the reconstructions. Additionally, gene-protein-reaction associations (GPRs) of the corresponding reactions were corrected based on the refined organisms' gene annotations (Supplementary Note B.2.5). If a reaction was present in a draft reconstruction but the associated gene was not identified in the organism by comparative genomics, the reaction was removed if the deletion did not disable growth on rich medium.

Respiratory pathways and quinone biosynthesis.

The presence of biosynthetic pathways for three main respiratory quinones (ubiquinone, menaquinone, and 2-demethylmenaquinone) as well as for proton-driven ATP synthases were genomically analyzed (Supplementary Table B.10 and Supplementary Note B.2.6). Quinone-dependent reactions were added in agreement with the repertoire of quinones synthesized by the corresponding organism [175]. If no known quinone biosynthesis pathways were found in a genome, we assumed that the organism utilizes extracellular menaquinone and added the corresponding transport and exchange reactions [207]. All annotations are available in PubSEED [162] (`http://pubseed.theseed.org/`, under the subsystems "Respiration HGM", "Respiration HGM New", "Quinones biosynthesis HGM", "Quinones biosynthesis HGM New", "ATP Synthases HGM New", and "ATP Synthases HGM").

B-vitamin biosynthesis.

Eight B-vitamins (i.e., biotin, cobalamin (B12), folate, niacin, pantothenate (B6), pyridoxine, riboflavin, and thiamin) were considered. Based on the genomic predictions and available experimental data, biosynthesis pathways for the B-vitamins were curated in the reconstructions (Supplementary Tables B.10, B.19, and B.11). For the cases where genomic predictions contradicted published data, the experimental data on growth requirements was used for the curation of our reconstructions. If a vitamin biosynthesis pathway was present in a reconstruction that should not synthesize the vitamin based on genomic and experimental data, we added a transport and exchange reaction for the vitamin and removed the reaction(s) that had been gap-filled by the Model SEED or KBase pipelines.

Central metabolic pathways.

In order to close gaps and improve gene-proteinreaction associations in central metabolic pathways, a comparative genomic analysis was performed for (i) glycolysis and gluconeogenesis, (ii) the pentose phosphate pathway, (iii) the Entner-Doudoroff pathway, (iv) the citric acid cycle and the glyoxylate shunt, (v) the biosynthesis of purine and pyrimidine nucleotides, (vi) amino acid biosynthesis pathways, and (vii) the *N*-acetylglucosamine utilization for polysaccharide biosynthesis (Supplementary Table B.2). In total, one to 96 (median

34) reactions participating in the 28 central metabolic pathways, represented by 20 PubSEED subsystems, were added to 627 reconstructions.

Nutrient requirements.

To ensure that each model could grow on biologically plausible *in silico* growth media, manual curation of *in silico* growth requirements was systematically performed by (i) removing unlikely growth requirements, (ii) gap-filling biomass precursor biosynthesis pathways based on comparative genomics (see above), and (iii) curating and validating the models against experimentally determined growth requirements reported in the literature. A preliminary defined medium was identified serving as a starting point for curation of the nutrient requirements (Supplementary Note B.2.7). Microbial growth requirements for 64 nutrients, including amino acids, vitamins, and nucleobases, were identified for 244 bacteria from one book [122] and 72 peer-reviewed papers (Supplementary Table B.1). False-positive predictions (nutrients that were reported to be nonessential in literature, but required for *in silico* growth) and false-negative predictions (nutrients that were reported to be essential, but nonessential *in silico*) were inspected and eliminated where possible (Supplementary Note B.2.7). The curation resolved 413 false-positive and 245 false-negative predictions for 64 metabolites in 244 microbial reconstructions, increasing the prediction sensitivity from 0.32 to 0.68 and specificity from 0.92 to 0.98. All compounds identified as essential for at least one reconstruction after the curation were added to the *in silico* diets (Supplementary Table B.18).

Leak test.

We ensured that no metabolite in a model could be produced from nothing, which would indicate mass or charge imbalanced reactions. Therefore, the lower bounds for all exchange and sink reactions were set to zero to prevent any influx of metabolites. A demand reaction was then added for each metabolite in the reconstruction and maximized. A metabolite could be produced from nothing if the objective value was greater than zero and the responsible reaction(s) was corrected.

Mycoplasma **and** *Ureaplasma* **sp.**

The reconstructions of *Mycoplasma pneumoniae*, *Mycoplasma hominis*, *Ureaplasma parvum*, and *Ureaplasma urealyticum*, which do not contain a cell wall [122] required further refinement to the respective BOF. We removed cell wall components from the BOF of all four reconstructions. Since the *Mycoplasma* genus requires cholesterol for growth [122], we added cholesterol to the BOF substrate lists of both *Mycoplasma* sp. reconstructions as well as transport and exchange reactions for cholesterol. To the two *Ureaplasma* sp. reconstructions, urea transport and exchange reactions were added as this genus requires urea [122].

Organization of the pipeline.

The reconstruction curation, validation, and content expansion steps described above were integrated into one pipeline (Fig. 3.1a). As all draft reconstructions stemmed from Model SEED or KBase, issues in one draft reconstruction were thus systemic for a subset or all reconstructions, and could be corrected in a consistent manner by propagating curation and QC/QA insights gained for one draft reconstruction to the remainder (Fig. 3.1c). This 'comparative reconstruction' approach allowed for the curation of hundreds of draft reconstructions at once.

Stoichiometric and flux consistency.

The rank of a stoichiometric matrix is an objective measure of the comprehensiveness of a reconstruction as it represents the number of linearly independent constraints on a steady-state reaction flux. The matrix rank was computed with numerical linear algebra. A set of stoichiometrically (mass balanced) reactions is mathematically defined by the existence of at least one, such that, where is a vector of the molecular mass of molecular species and is a stoichiometric matrix. We computed the largest stoichiometrically consistent subset of each draft and AGORA stoichiometric matrix using numerical linear optimization [72]. We say a matrix S is net flux consistent if there exist matrices and such that, where each row contains at least one nonzero entry. This condition ensures that a reaction admits a nonzero net flux in some flux distribution. Flux consistency was tested using numerical linear optimization [238] as described in Fleming *et al.* [72].

Gene sequence acquisition.

We retrieved the nucleotide gene sequence of each microbe using the Perl API from the Model SEED and KBase web interface. We expanded these files with the appropriate gene sequences for respiration, quinone biosynthesis, B-vitamin synthesis, and central metabolic pathways were retrieved from the web interface of the PubSEED platform [13]. The final compiled gene sequences for each organism can be found in the VMH database.

Diet definitions.

We defined two different diets, a Western diet and a high fiber diet (Supplementary Table B.18). The diets varied in fat, simple sugar, starch, and fiber content. Additionally, both diets contained amino acids, vitamins, minerals, water, methanol [14], and other metabolites, each of which was required for a nonzero biomass reaction flux in at least one of the 773 models (Supplementary Note B.2.7).

Metabolic distance.

The metabolic distance (MD) between two microbes was calculated using the Jaccard distance, such that $MD = 1 - |R_i \cap R_j|/|R_i \cup R_j|$, where R_i is the list of reactions from reconstruction i and R_j is the list of reactions present in reconstruction j. MD of 1 means that the two reconstructions share no reactions, and MD of zero means that the two reconstructions have identical reaction lists.

Carbon source and fermentation phenotypes.

We set the lower bound of the BOF in a model to 0.001 h^{-1} to ensure its minimum growth, performed flux variability analysis (FVA) [89], and inspected minimal and maximal possible flux values through the exchange reactions. A model was considered to be able to take up a carbon source if the minimal possible flux through a carbon source exchange reaction was negative ($\leq -10^{-6}$ mmol/gDW/h). Similarly, a model was considered to be able to secrete a fermentation product if the maximum possible flux through the fermentation product exchange reaction was positive ($> 10^{-6}$ mmol/gDW/h). Even though several amino acids can

be used as carbon sources they were excluded from this analysis because amino acids are directly required in the BOFs.

Gene essentiality analysis.

Gene essentiality data were available for six AGORA microbes (Fig. 3.1d) [137]. We retrieved the gene sequences for the six draft reconstructions from the Model SEED and KBase platforms. Published reconstructions of the same strains were available for four of these microbes [229, 20, 225, 96]. Their gene sequences were retrieved from the NCBI database [43]. We used the BLAST search option of the database (`http://www.essentialgene.org/`) [137] to identify the *in vitro*–validated essential genes for each microbe. The *in silico* gene essentiality analysis was carried out under rich medium conditions for the draft, AGORA, and published models, by constraining the flux through all reactions associated with a deleted gene to 0 mmol/gDW/h and maximizing the BOF. A gene was considered essential *in silico* if its deletion resulted in a biomass reaction flux of zero.

Pairwise model simulations.

Pairwise simulations were performed on every possible pair of the 773 AGORA metabolic reconstructions (298,378 pairs). Microbial models were paired by introducing a common lumen compartment, as described elsewhere [97], in which each model could secrete or from which it could take up metabolites. Dietary compounds were added to the lumen and byproducts were removed. To prevent biologically implausible solutions, in which microbes benefit the paired microbe without producing any biomass, coupling constraints were added to the joined models [96]. Briefly, all reactions in a model were stoichiometrically coupled to its BOF, thereby enforcing a nonzero flux through the BOF if reaction fluxes were nonzero. Using FBA, growth of the microbial pair was maximized under both diets (Western and high-fiber diet), aerobically and anaerobically (Supplementary Table B.6). The maximal possible BOF of the individual microbe in the pair was determined by inactivating all reactions belonging to the other paired microbe. A minimal microbial BOF flux for each microbe of 0.001 h^{-1} was enforced. A model was considered to grow faster in the co-growth simulation when its paired growth rate was more than 10% higher than the individual growth

rate of the same microbe under the same condition (Supplementary Table B.18). A model was considered to grow slower in the co-growth simulation when its paired growth rate was more than 10% lower than the individual growth rate of the same microbe under the same condition.

Random order microbial assembly.

We selected randomly a reconstruction to obtain its reaction list and appended it to the growing unique reaction list until all 773 strain-resolved reconstructions were considered once (Fig. 3.6b). We monitored the number of reactions added by each new reconstruction. We repeated this procedure 1,000 times. On the species level, we randomly selected a species and obtained the unique reaction list of all strain-resolved reconstructions belonging to that species. We appended the list of reactions until all 605 species captured by AGORA were considered once (Fig. 3.6d). We monitored the number of reactions added by each new species. We repeated this procedure 1,000 times.

Comparison with Recon 2.

The human metabolic reconstruction, Recon 2.04 (ref. [227]), containing 5,063 metabolites and 7,440 reactions, was retrieved from http://vmh.life. A decompartmentalized Recon was created by placing all reactions occurring in one of the seven intracellular compartments into the cytosol. The extracellular compartment was retained. Duplicate reactions and reactions with identical sets of metabolites occurring on both sides of the chemical equation were removed, resulting in 6,256 unique metabolic reactions.

Metagenomic and 16S rRNA analysis.

Information on metagenomic reads from gastrointestinal tract samples mapped onto a set of reference genomes was downloaded ()http://hmpdacc.org/ for 149 healthy US individuals aged 18–40 (ref. [41]). We matched the strain names of the mapped reference genomes to 245 AGORA reconstructions (Supplementary Table B.12) and identified the unique reaction set (i.e., metabolic diversity) for each individual. Using the read depth information, we calculated the read number covered by AGORA.

For the ELDERMET [37] data, annotated species lists for 177 individuals from processed 16S rRNA data were retrieved from MG-RAST [152] (http://metagenomics.anl.gov/linkin.cgi?project=154). An individual's unique reaction set was obtained by mapping the reported species onto the pan-species reconstructions (Supplementary Table B.12). Principal coordinate analysis was performed on the metabolic distance between each individual's reaction set.

In vitro cell cultures.

B. caccae ATCC 43185 and *L. rhamnosus* GG ATCC 53103 (LGG) were precultured for 20 h in Brain Heart Infusion Broth (BHIS; Sigma), supplemented with 1% hemin under anaerobic conditions and shaking at 37°C (Supplementary Note B.2.3). After washing and resuspending in 10 ml of 0.9% w/v NaCl solution, they were inoculated in DMEM 6429 supplemented with 1% hemin and 3.33% vitamin K, with or without arabinogalactan (Sigma; 9.4 g/l), and maintained under anaerobic conditions. *B. caccae* and LGG were cultured for 33 and 44 h on average, respectively. Cells were harvested for cell counting by centrifugation (4,700g) and 750 μL aliquots of supernatant were removed for subsequent metabolite extraction. The aliquots were snap-frozen and placed at -80°C until dedicated analysis. The cultures were confirmed using 16S rRNA sequencing.

Metabolomic analysis.

The extraction and GC-MS measurement of shortchain fatty acids was based on a protocol from Moreau *et al.* [154] (Supplementary Note B.2.8 and Supplementary Table B.20). Extracellular polar metabolites from the supernatant samples and external concentration curves for each compound of interest were extracted applying a liquid–liquid extraction (Methanol/Water). GC-MS analysis was performed using an Agilent 7890A GC coupled to an Agilent 5975C inert XL Mass Selective Detector (Agilent Technologies). All GC-MS chromatograms were processed using MetaboliteDetector software, v3.020151231Ra72 (Supplementary Note B.2.8). In addition, absolute quantitative values for lactic acid, glutamine, glutamic acid, and glucose were acquired using a 2950D Biochemistry Analyzer (YSI) (Supplementary Note B.2.8).

In silico simulations of *B. caccae* and LGG.

The single models as well as the pairwise model of *B. caccae* and LGG were subjected to an *in silico* medium mimicking the supplemented DMEM 6429 medium without and with arabinogalactan (Supplementary Table B.17). Single and combined growth was predicted as described above. The potential of *B. caccae* and LGG to consume and produce metabolites *in silico* was computed using FVA [89] and compared with GC-MS measurements. To identify the cross-feeding between the two species at optimal growth, FBA was performed with simultaneous growth as the objective function while minimizing the sum of internal fluxes.

3.3 Results

3.3.1 Metabolic reconstruction pipeline

We devised a comparative metabolic reconstruction method (Fig. 3.1a,c), which is analogous to the comparative microbial genome annotation approach [161] that has enabled accelerated annotation by propagation of refinements to one genome to others. First, we downloaded draft GENREs using Model SEED [102] and KBase [9]. In both platforms, the genome sequence of an organism is automatically annotated and a metabolic reconstruction is assembled based on the identified metabolic functions. Gaps in the draft reconstruction are automatically filled, building a metabolic reconstruction whose condition-specific models can carry flux through a defined biomass objective function. We refined the draft reconstructions using rBioNet [230] and performed quality control and quality assurance (QC/QA) tests, including the verification of reaction directionality and mass and charge balance (Online Methods and Supplementary Note 1) to ensure that the reconstructions meet the quality standards set by Thiele and Palsson (2010) [226]. We expanded the reconstructions by refining gut-microbiota-specific and central metabolic subsystems, and curated all of the reconstructions by reference to 236 publications, two reference books (Supplementary Table B.1), and comparative genomics analyses (Online Methods and Supplementary Table B.2). Anaerobic growth was enabled for all genome-scale models (Fig. 3.1e) because the human gut is usually anaerobic or microaerobic [64]. We tested the metabolic capabilities of each model, as defined by published data, at the genus, species, and strain level. During the generation and

validation of each model, solutions to QC/QA problems or failure of the metabolic capability tests of a model were used to improve the quality of others. The propagation of such refinements among all the models was facilitated because all microbes share the same human gut environment.

Our pipeline increased both the proportion and total number of stoichiometrically and flux-consistent reactions [72], that is, mass balanced and admitting a nonzero steady state flux (Fig. 3.1f,g, Supplementary Table B.3 and Supplementary Figs. B.1 and B.2). The refinement process increased the predictive potential of AGORA models compared with the draft metabolic models; the AGORA models predicted gene essentiality more accurately than the draft models (Fig. 3.1d), and sensitivity to carbon source, fermentation product, and nutrient requirement data was greatly increased in the AGORA models with an average sensitivity of 1.00 ± 0.02 compared to 0.06 ± 0.09 for the draft models (Fig. 3.1h and Supplementary Table B.13).

3.3.2 Features of reconstructions

The 773 AGORA reconstructions contained an average of 771 ± 262 genes, $1,198 \pm 241$ reactions, and 933 ± 139 metabolites (Fig. 3.1b and Supplementary Table B.4). We found that taxonomic classes containing well-studied organisms, such as Gammaproteobacteria, had larger sets of genes and reactions than reconstructions of other classes. An AGORA reconstruction uses a generalized microbial biomass reaction, which summarizes the fractional contribution of a biomass precursor (e.g., amino acids and lipids) to the synthesis of a new cell, as provided in the draft reconstructions from Model SEED and KBase. The biomass reactions were not curated, because species-specific information is required for such refinements (Supplementary Note B.2.2). Qualitative growth predictions are not affected by generalized biomass equations, whereas information on species- and conditionspecific cellular composition is required for accurate quantitative prediction [69]. Nonetheless, the predicted average microbial doubling time of 2.3 h on a Western diet under anaerobic conditions (Fig. 3.1b and Supplementary Table B.5) was close to reported doubling times of single microbes in the mouse gastrointestinal tract [82].

When comparing eight AGORA reconstructions with the published genome-scale metabolic
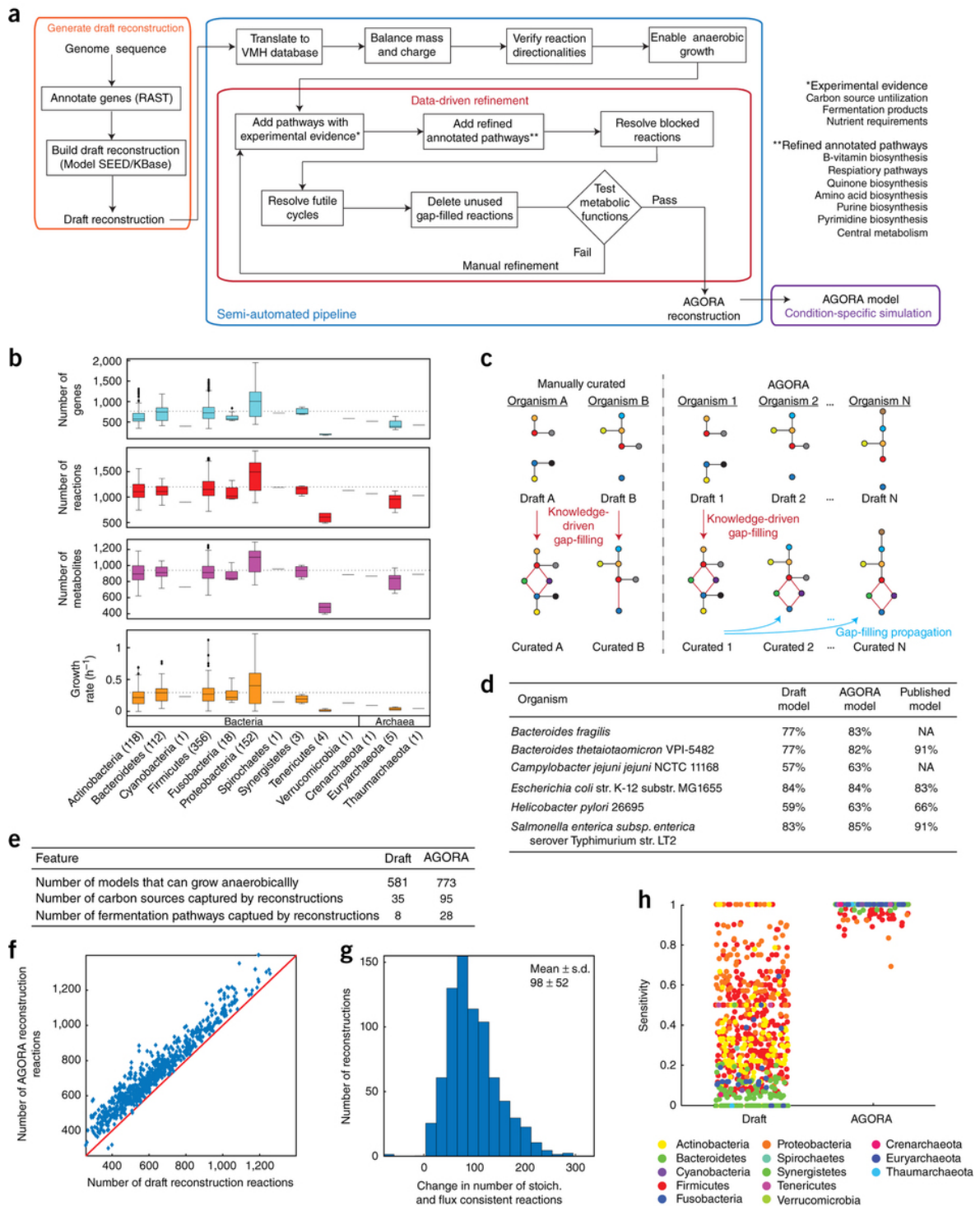
Figure 3.1

reconstructions for the same species, we found that the main differences in reaction content were in lipopolysaccharide biosynthesis and transport pathways (Supplementary Tables B.14 and B.15). Cell wall and lipopolysaccharide structures are species-specific and cannot easily

Figure 3.1: The reconstruction refinement pipeline and properties of the metabolic models, derived from the reconstructions. (a) The refinement pipeline starts with an automatic draft reconstruction generation from the online platforms Model SEED [102] and KBase [9], followed by the translation of reaction and metabolite identifiers to match those of the human metabolic reconstruction, Recon 2.04 (ref. [227]). The reconstructions are then tested for QC/QA measures and 162 metabolic functions are curated based on available knowledge and genomic evidence. A microbial reconstruction can be converted into a condition-specific model by the application of condition-specific constraints. All AGORA reconstructions are available at `http://vmh.life`. (b) Boxplots of the number of genes, reactions, and metabolites for reconstructions belonging to the 13 phyla captured in AGORA (Supplementary Table B.4). Dashed lines represent the average over all 773 reconstructions. The number of reconstructions per phylum is shown above the phylum name. Whiskers show the minimum and maximum values. Values below $Q1 - 1.5 \times IQR$ and above $Q3 + 1.5 \times IQR$ are plotted as outliers. Q1: first quartile, Q3: third quartile, IQR: interquartile range. (c) Differences in the gap-filling process of manually curated reconstructions and AGORA. In AGORA, gap filling of a certain pathway in one reconstruction is propagated to all N reconstructions that share the same gap and should perform the metabolic function in question. For manually curated reconstructions, every gap is filled with organism-specific reactions based on the available organism-specific data and experimental validations. (d) Gene essentiality accuracy predicted by the draft reconstructions, the AGORA reconstructions, and four published reconstructions (Supplementary Note B.2.2) when compared against *in vitro* data sets [137]. (e) Table showing the number of AGORA and draft models that grow anaerobically on rich medium and the number of carbon sources, and fermentation products captured by draft reconstructions and AGORA reconstructions. (f) The number of stoichiometrically- and flux-consistent reactions in each draft versus the corresponding AGORA reconstruction. (g) Change in number of stoichiometrically and flux consistent reactions. (h) Comparison of the predictive potential of the draft reconstructions and the AGORA reconstructions. The sensitivity (true-positive rate) of known fermentation products, carbon sources, and growth requirements captured by the corresponding metabolic models.

be derived from gene annotations alone [79, 15] (Supplementary Note B.2.2). The curation of such pathways requires experimental data that are currently not available for AGORA organisms. AGORA models were equivalent to published GENREs in terms of capturing gene essentiality, as reported in the literature (Fig. 3.1d). The AGORA models have been curated for carbon source utilization and fermentation product secretion; they outperformed the published models for those functions, as measured by a sensitivity analysis of the carbon source uptake and fermentation product secretion of both seven published models and AGORA models (Supplementary Fig. B.3). Until now, manually curated reconstructions have been refined to fit certain applications and, consequently, curated to different extents. As they are not comparable in scope, they may not be a good choice for investigating mi-

crobe interactions. In contrast, all AGORA reconstructions have been curated for the same metabolic pathways (Fig. 3.1a).

We determined the overlap of microbial metabolism with human metabolism. Two cellular compartments are common between human and microbes, the cytosol and extracellular space. To compare the metabolic functions, we decompartmentalized the human metabolic reactions by placing all metabolites that occur in an organelle (e.g., mitochondria) compartment in the human metabolic reconstruction Recon 2 (ref. [227]), into the cytosol, and removed duplicate reactions, resulting in 6,256 unique metabolic reactions. Collectively, the AGORA reconstructions account for 3,192 unique metabolic reactions, 695 of which are shared with Recon 2, including 162 (23%) exchange reactions. We found that 89% (5,561/6,256) of human metabolic reactions were unique to the human reconstruction, and 78% (2,495/3,192) of AGORA metabolic reactions were unique to the microbial reconstructions.

3.3.3   Metabolic diversity of AGORA reconstructions

The variety of AGORA reconstructions is shown in Figure 3.2a. To prove that our method produced metabolically distinct reconstructions for each organism, we computed the metabolic distance of every reconstruction pair (298,378 pairs; Supplementary Table B.6) using the Jaccard distance between the reaction lists from each reconstruction. Metabolic distances range from zero to 1 with identical reconstructions having a metabolic distance of zero and completely dissimilar reconstructions having a metabolic distance of 1. As expected, taxonomically related bacteria shared more reactions than taxonomically distant bacteria (Supplementary Fig. B.4). Bacilli and Clostridia had high metabolic distances, reflecting the metabolic and phenotypic differences between these two classes (Supplementary Fig. B.4). Notably, low metabolic distances were only observed within a class, but not between members of a phylum or between taxonomic classes. Overall, the average metabolic distance was 0.48, which is consistent with other reports of metabolic and functional distances between microbes based on the presence of KEGG enzymes and KEGG orthology annotations [147, 32]. Metabolic pathway enrichment was detected at different taxonomic levels (Fig. 3.2b); for example, plant polysaccharide degradation was mainly present in Bacteroidia,

O-glycan degradation in the genera *Akkermansia* spp., *Bifidobacterium* spp., and *Bacteroides* spp., and methane metabolism was unique to four archaea. As expected, lipopolysaccharide biosynthesis was found only in Gram-negative bacteria. Only 69 reactions were common to all 773 reconstructions. To assess whether the known functional diversity of the gut micro-
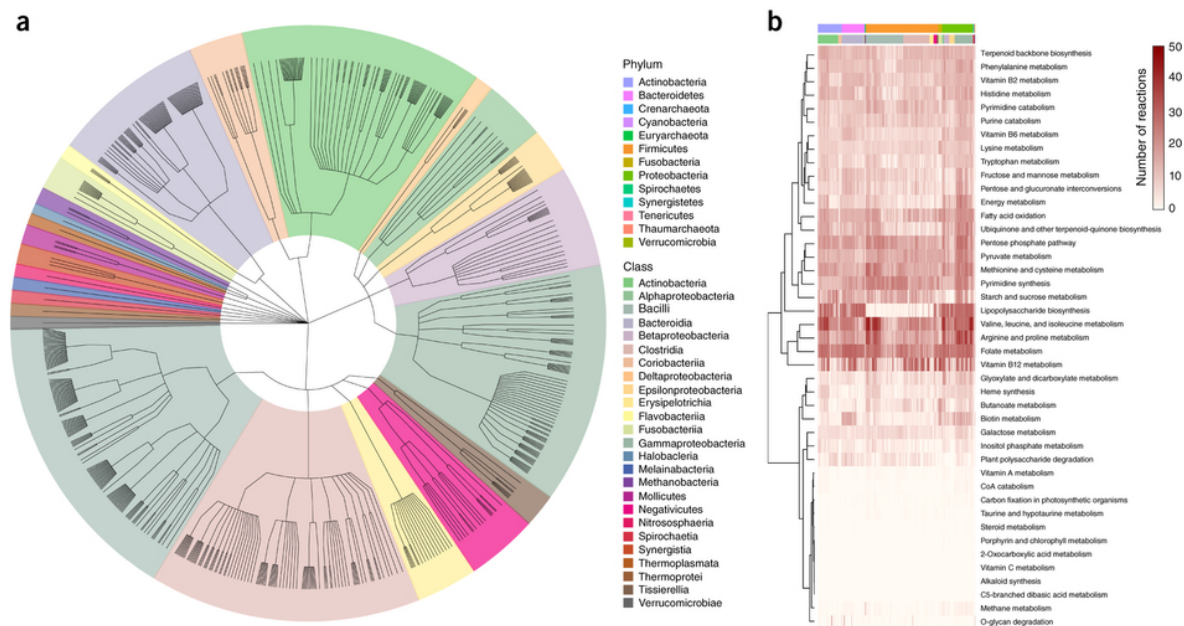
Figure 3.2: Taxonomic and metabolic diversity of the 773 genome-scale metabolic reconstructions. (a) Taxonomic tree of the 773 organisms showing the diversity of the AGORA resource. Colors correspond to the taxonomic classes. The tree was created using GraPhlAn [11]. (b) Hierarchical clustering of the average ratio of reactions per subsystems found in the reconstructions of the 25 taxonomic classes. The reconstructions are ordered based on phyla and taxonomic classes.

biota was captured in our models, we tested for the uptake of 74 different carbon sources and for the secretion of 18 fermentation products (Fig. 3.3) using flux variability analysis [89]. The known distribution of short-chain fatty acid production in the gut was wellrepresented in our models (Fig. 3.3); most models fermented sugars into short-chain fatty acids and organic acids, with acetate, succinate, formate, lactate, propionate, and ethanol being the most commonly produced metabolites. As expected, butyrate was secreted by the Fusobacteria models [237] and by many Firmicutes models [136], and methane secretion was specific to the Euryarchaeota. Carbon source utilization capabilities were found to be in agreement with the literature [74]. Notably, only certain genera (e.g., *Bacteroides* spp. and *Akkermansia* spp.) can utilize diet- and host-derived polysaccharides (Fig. 3.3).

3.3.4 Validation of AGORA model predictions

We demonstrated the predictive capability of AGORA models using two models, *Bacteroides caccae* ATCC 34185 (a fiber-degrading symbiont common in microbiomes of Western individuals [168, 150]) and *Lactobacillus rhamnosus* GG (LGG), a common human probiotic strain [204]. No chemically defined medium has been reported for *B. caccae*. We predicted that *B. caccae* should be able to grow on DMEM 6429 defined culture medium supplemented with vitamin K, hemin, and arabinogalactan under anaerobic conditions. In the laboratory, *B. caccae* was cultured on this medium in a flask under anaerobic conditions (Online Methods, Supplementary Note B.2.3 and Supplementary Table B.16). This medium would not support growth of LGG, according to our *in silico* predictions, and growth was unstable in flask cultures using this medium. The reported chemically defined growth medium for LGG contains all amino acids and most vitamins [199] but our *in silico* predictions suggested that these could be supplemented by growing LGG with *B. caccae* (Fig. 3.4). When modeled *in silico* in co-culture in DMEM 6429 defined culture medium supplemented with vitamin K, hemin, and arabinogalactan, both bacteria grew (Fig. 3.4d and Supplementary Table B.17). We predicted that *B. caccae* would supply LGG with alanine, asparagine, and nicotinic acid, while LGG would provide lactate to B. caccae (Fig. 3.4d). The addition of alanine and nicotinic acid to the defined medium was sufficient *in silico* to enable the single growth of the LGG model (Fig. 3.4c). Using gas chromatography–mass spectrometry (GC-MS)-based metabolomic analysis, we confirmed the secretion of numerous metabolites by the two strains grown individually, including alanine secretion by *B. caccae* (Fig. 3.4a and Supplementary Fig. B.5), thus supporting the predicted cross-feeding.

3.3.5 Pairwise interactions of models

We computed the pairwise growth interactions ('co-growth') of every pair of microbes in the AGORA resource (298, 378 pairs). Each model was grown *in silico* on its own and as part of a pair with every other model on two different diets with and without oxygen (Online Methods and Supplementary Tables B.6 and B.18). Under all conditions, the most commonly predicted pairwise co-growth relationships were parasitism or commensalism (Fig. 3.5a). Consistent with one previous *in silico* study [97], the presence of oxygen resulted in a de-

crease in commensalism and mutualism, especially for Gammaproteobacteria (Supplementary Fig. B.6). Competitive and amensal interactions increased in the presence of oxygen (Fig. 3.5a). The effects of a typical Western diet and a diet high in fibers, such as arabinogalactan and xylan, on pairwise interactions among models were also evaluated. The high fiber diet led to a higher proportion of commensal and mutualistic interactions (Fig. 3.5a), which, using flux balance analysis[159], was found to be due to cross-feeding of metabolites between species. Low-fiber diets are thought to modulate the microbiota composition by depleting commensal microbe-to-microbe interactions[58]. Based on hierarchical clustering of the ratio of pairwise interaction type per condition per taxonomical family (Fig. 3.5b), our simulations predict that the carbohydrate fermentation capacity defines the type of interaction between microbes when considering the family level. Cluster 1 contained saccharolytic microbes that do not produce butyrate and respiratory bacteria and was enriched in commensal interactions. Clusters 2a and 2b were enriched in butyrate and lactate fermenters and generalists related to Bacillaceae, respectively. Both subclusters were mostly negatively affected by parasitic interactions. Cluster 3 contained the majority of asaccharolytic and proteolytic families, which mainly benefitted from the pairwise interactions. The effect of carbohydrate fermentation capacity holds true for the genus level (Supplementary Fig. B.6). Even though most interactions were observed over a wide range of metabolic distances, positive interactions, in which the growth rates of one or both microbes are neutral or enhanced in the pairwise simulation, occurred only among metabolically distant organisms (Supplementary Fig. B.7), in agreement with previous computational studies[147, 32].

3.3.6 Integrating metagenomics and 16S rRNA with AGORA

We tested whether AGORA can be used to analyze metagenomic data. We retrieved strain-resolved metagenomic data from 149 healthy individuals from the human microbiome project (HMP; Fig. 3.6a) [41]. AGORA microbes mapped to, on average, 91% of the strains in the HMP individuals with comparable reaction diversity for all individuals (Fig. 3.6b), highlighting that AGORA is representative of the human gut microbiota.

We then mapped published species-resolved 16S rRNA data of 164 elderly and 13 young individuals ("ELDERMET") [37] onto AGORA (Fig. 3.6c,d). The 210 ± 39 species present

in each individual mapped to 108 ± 16 AGORA pan-species reconstructions, which are the union of reactions from all strain-specific reconstructions in AGORA of one species. Two clusters were observed (Fig. 3.6d). The cluster with more reactions was characterized by the presence of Gammaproteobacteria (Fig. 3.6d). The corresponding reconstructions contain on average more reactions compared to other taxonomic classes (Fig. 3.1b). Principal coordinate analysis, using each individual microbiota's metabolic reaction set, revealed that the clusters separated owing to reactions associated with glycerophospholipid and cell wall metabolism (Supplementary Table B.7), consistent with the Gram-negative nature of Gammaproteobacteria. The second principal coordinate was mainly associated with the presence of methanogenesis reactions that are unique to the methanogenic archaea.

3.4   Discussion

AGORA reconstructions were assembled using a comparative metabolic reconstruction approach that speeded up curation and provided knowledge-driven refinement of gut-specific metabolic microbial functions that were not present in the draft reconstructions (Fig. 3.1a). Our pipeline included extensive QC/QA and curation against available knowledge, which is not done by pipelines that automatically generate GENREs. The resulting models significantly outperformed the draft models in correctly capturing gene essentiality in addition to known carbon sources, fermentation products, and essential nutrients (Fig. 3.1d,h). Even though AGORA reconstructions do not cover all of the species-specific aspects of manually curated reconstructions (e.g., lipopolysaccharide biosynthesis), the performance of each of eight AGORA models was on par with that of their previously published, manually curated models of the same strain (Fig. 3.1d and Supplementary Fig. B.3). This resource of reconstructions helps to address the need for literature-curated GENREs to help to analyze gut metagenomic data [86].

GENREs have previously been used to design chemically defined growth media using an iterative *in silico*, *in vitro*, and metabolomics approach [95]. The predicted growth of *B. caccae* in extended DMEM medium supplemented with arabinogalactan under anaerobic conditions was validated. However, the *in vitro* growth of LGG was unstable, but our *in silico* co-culture simulations suggest that its growth could be enhanced by the presence of

B. caccae. This example highlights how metabolic models can serve as a starting point for generating experimentally testable hypotheses.

Bacteria in ecosystems engage in complex trophic webs based on interspecies interactions [88], which cannot be inferred from microbial abundance [67]. We explored pairwise interactions between gut microbiome models under four conditions. We found that central metabolic traits were predicted to define co-growth interactions as a function of diet composition and oxygen availability (Fig. 3.5 and Supplementary Figs. B.6 and B.7. Inflammation of the digestive tract can disrupt the intestinal cell barrier, thereby increasing the oxygen level in the normally anaerobic intestine [182] and reducing species variety in the gut microbiome [141]. A high-fiber diet might protect against the depletion of positive microbe–microbe interactions caused by the presence of oxygen. Negative interactions have been found to dominate stable microbial communities [44]. Based on our simulations, we propose that a small number of positive interactions may be sufficient to maintain a healthy microbial community.

Metabolic models have been used to map and analyze omics data for single organisms [202]. Here, we report that AGORA enables such analyses for a much larger set of human gut microbial communities, as most of metagenomic sequence reads and 16S rRNA data from a typical microbiome can be mapped onto our models (Fig. 3.6), resulting in metabolically diverse microbiota reconstructions that can be used to construct and simulate individual-specific metabolic models. The metabolic overlap between the human metabolic reconstruction and AGORA was enriched in exchange reactions, which supports the hypothesis that co-evolution driven by cross-feeding between the host and the microbes has occurred.

AGORA reconstructions are especially suited to studies in which multiple reconstructions are coupled together to simulate microbial interactions. However, the existence of consistent biases, owing to the semi-automated model generation means that additional strain-specific refinements may be necessary in order to use AGORA models in organism-specific applications, for example, bioengineering. AGORA reconstructions are missing non-dietary microbial functions, such as xenobiotic metabolism, which have not yet been extensively studied in human gut microbes. Because the focus of the curation has been on gut microbial functions, the resource is best suitable for studies on dietary effects on the human microbiota.

One open question in microbiome research is which functions microbes carry out and how those functions interface with host metabolism and affect host phenotypes. About a quarter of AGORA metabolic reactions were also present in Recon 2 (ref. [227]), the human metabolism reconstruction, highlighting the complementarity of host and microbial metabolisms. So far, a handful of studies have attempted to predict the metabolic effect of different microbiomes on host metabolism based on topological network approaches [213, 221], which cannot address functional links in the human-gut microbiota axis. AGORA enables superior simulations to address mechanistic questions about host–microbe co-metabolism. We envisage the combination of AGORA models into strain-resolved microbial community models and predictions of how those community models interact with human metabolic models [227] could be used to systematically investigate host–microbiome interactions [99].

Acknowledgements

Figure 3.3: Carbon source uptake and fermentation product secretion capabilities in AGORA. Number of AGORA models in each phylum that can consume the different carbon sources and secrete the tested fermentation products. The total number of models in each phylum is reported in parentheses. The models' capabilities to consume or secrete the different metabolites were determined using flux variability analysis.
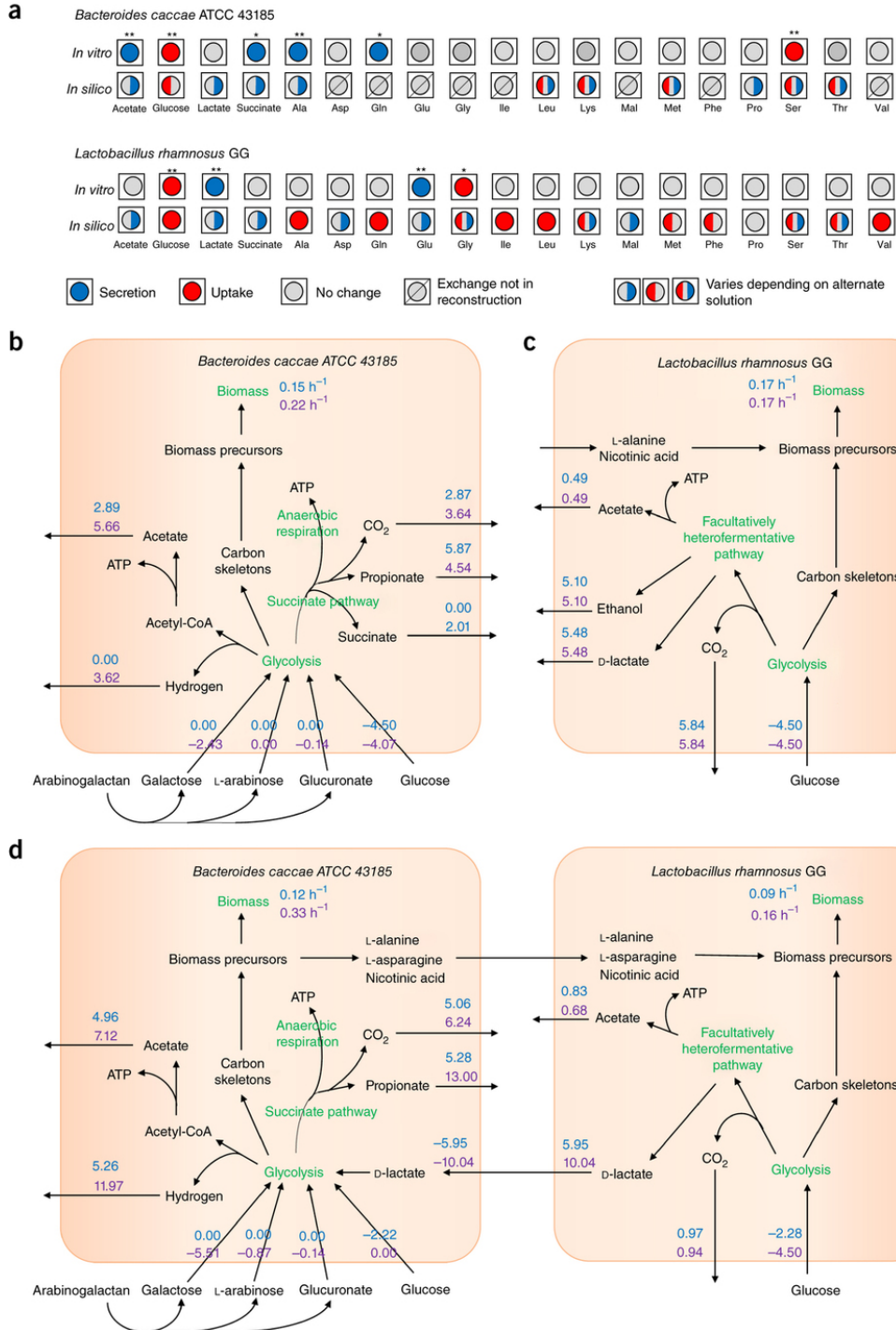
Figure 3.4

Figure 3.4: Comparison of *in vitro* experiments and *in silico* simulations for *Bacteroides caccae* ATCC 43185 and *Lactobacillus rhamnosus* GG ATCC 53103. (a) Both microbes were grown anaerobically on DMEM 6429 medium supplemented with hemin, vitamin K, and arabinogalactan. The composition of fresh medium and spent medium after cultivation was determined using GC-MS (Supplementary Fig. B.5). Only statistically significant metabolite uptake and secretion is shown. A paired t-test was performed for statistical significance. $**P < 0.005$ and $*P < 0.01$. The same medium composition was used for *in silico* simulations, and uptake and secretion capabilities were predicted using flux variability analysis. (b) *In silico* single culture fluxes of *B. caccae* on DMEM medium without and with arabinogalactan. Growth rates (h^{-1}) and predicted uptake and secretion fluxes (mmol/gDW/h) of major exchanged metabolites are shown in blue without arabinogalactan and in purple with arabinogalactan. (c) *In silico* single culture fluxes of LGG on DMEM medium without and with arabinogalactan. The *in silico* medium was supplemented with L-alanine and nicotinic acid, which were predicted to be essential for growth. Growth rates and uptake and secretion fluxes are depicted as described for b. (d) In silico co-culture fluxes of *B. caccae* and *L. rhamnosus* on DMEM medium without and with arabinogalactan. Growth rates, and uptake and secretion fluxes are depicted as described for panel b.
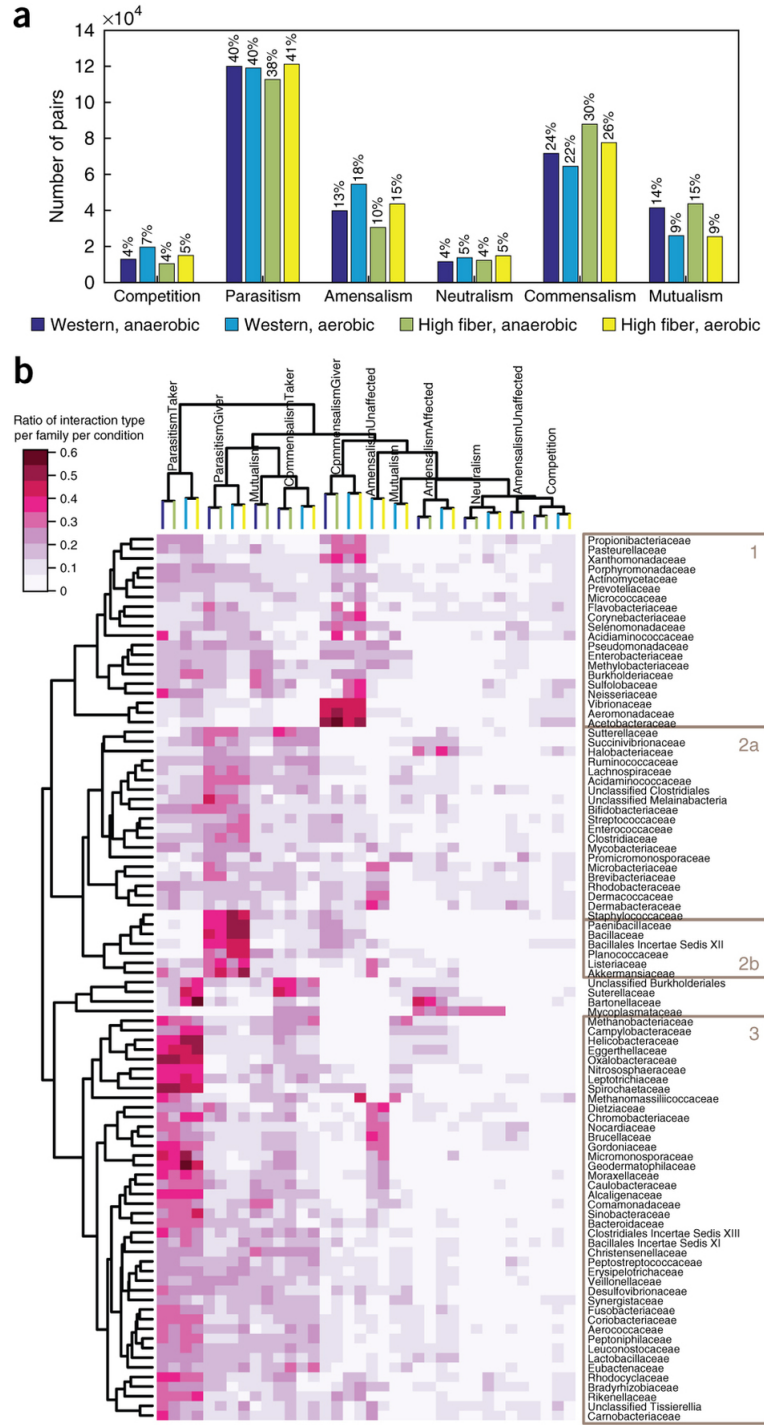
Figure 3.5

Figure 3.5: Pairwise interactions of all AGORA metabolic models. (a) The number and percentage of pairs that exhibit each of six different interaction types during *in silico* simulations under four different conditions; Western and high-fiber diets in the presence and absence of oxygen. Competition: both microbes' *in silico* growth rates (Supplementary Table B.6) are slower in the paired simulation compared with each microbe's *in silico* mono-culture growth rate on the same diet (Supplementary Table B.18). Parasitism: one microbe grows faster (Taker) in the paired simulation while the other microbe grows slower (Giver). Amensalism: one microbe grows slower (Affected) in the paired simulation while the other microbe's growth rate is unaffected (Unaffected). Neutralism: both microbes' growth rates are unaffected in the paired simulation. Commensalism: one microbe grows faster (Taker) in the paired simulation while the other microbe's growth is unaffected (Giver). Mutualism: both microbes grow faster in the paired simulation. (b) Hierarchical clustering (Euclidean distance) of the ratio of pairwise interaction types per condition (i.e., diet and oxygen presence) on the taxonomic family level. See a for a description of the interaction types. Three main clusters were identified, each enriched in microbes with different carbohydrate fermentation capabilities, belonging to the clustered families.
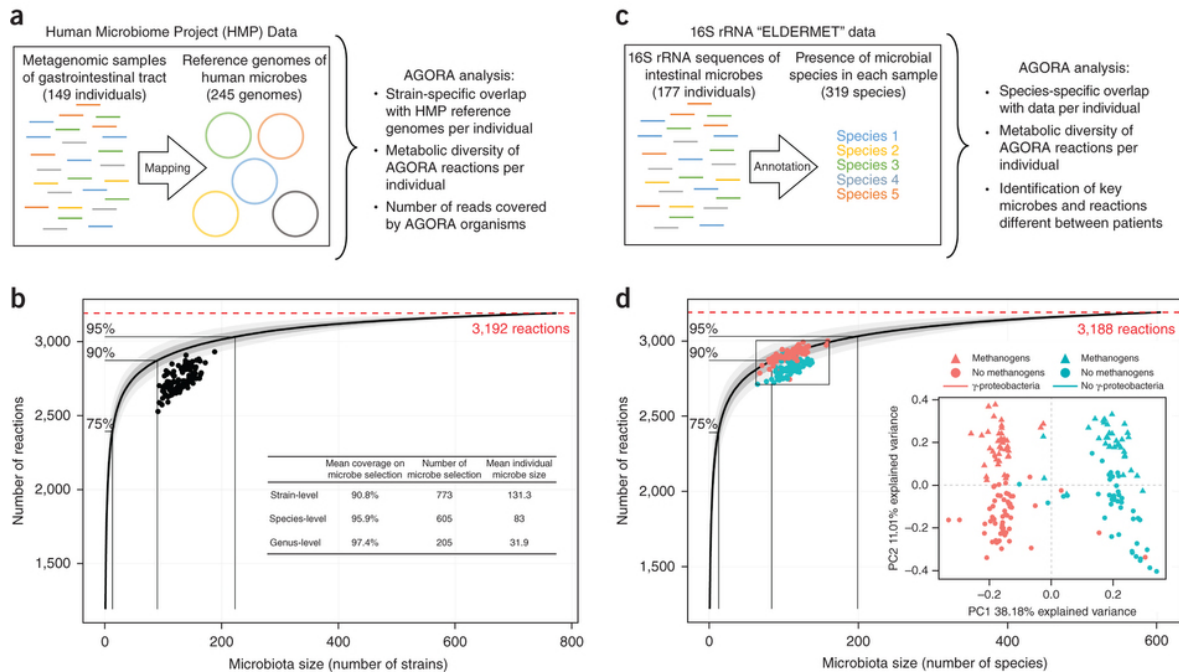
Figure 3.6: Metabolic diversity of individual human gut microbiomes. The average number of unique reactions was computed based on $1,000$ random orders of metabolic reconstruction additions. Dark and light gray areas correspond to one and two s.d., respectively. (a) Human samples of HMP, which have been mapped onto a set of reference genomes, were used to determine how many AGORA organisms are overlapping with each individual. (b) Mapped AGORA strains per HMP sample with their corresponding collective reaction diversity and the percentage of mapped reads they capture for three taxonomic groups. In AGORA, all microbial reconstructions were needed to achieve full metabolic diversity (i.e., $3,192$ unique reactions). Also shown are the number of reconstructions needed to capture 75%, 90%, and 95% of the full reaction diversity. To obtain 75% and 95% of the unique reactions, 12 and 123 reconstructions, respectively, were sufficient on average. (c) 16S rRNA results from Claesson *et al.* (2012) [37] ('ELDERMET') were used to determine how many pan-species reconstructions of AGORA are overlapping with each individual. (d) Mapped AGORA pan-species reconstructions per ELDERMET sample with their corresponding collective reaction diversity (i.e., $3,188$ unique reactions). Also shown are the number of reconstructions needed to capture 75%, 90%, and 95% of the full reaction diversity. The metabolic repertoire of the individual mapped microbiomes was analyzed by principal coordinate analysis in terms of the metabolic distance between the reaction sets of the mapped pan-species reconstructions. The presence and absence of methanogens and Gammaproteobacteria per individual mapped microbiome is shown.

Chapter 4

Individual-based gut microbiome models

Manuscript in preparation.

4.1   Abstract

By mapping metagenomic reads from individual stool samples to the genomes of the AGORA reconstructions, we created 53 personalized gut microbiome metabolic models based on metagenomic samples from 20 individuals. We showed that the resulting microbiome reconstructions are both taxonomically and functionally diverse. We found significant differences in maximum possible riboflavin secretion of the microbiomes of healthy and diabetic individuals, and a difference in the minimum thiamine uptake flux required by models based on individuals from different families. These differences in B-vitamin metabolism were driven by the presence of the microbe *Prevotella copri*, which can synthesize both thiamine and riboflavin. Gut microbiome metabolic models can be applied in a variety of studies, and can be further coupled with the human metabolic reconstruction Recon 2 to investigate host-microbiome metabolic exchanges.

4.2   Introduction

T1D is an autoimmune disease where the β-pancreatic cells are attacked by the host's own immune system resulting in a reduced production of insulin by the host cells. Several studies have investigated the link between the human gut microbiome and type 1 diabetes (T1D) [83, 61, 114, 145]. One such study by Heintz-Buschart et al. (2016) [100] used a variety of meta-omics analyses to investigate the microbial taxonomic and functional links with T1D. The study reported metagenomics, metatranscriptomic, and metaproteomic data of stool samples from 20 individuals, 12 of which were diagnosed with T1D. The study found that the functional transcript levels of the thiamine biosynthesis enzyme thiazole synthase (ThiG) were positively correlated with with relative protein abundances of the amylase proteins AMY2A and AMY2B, which were less abundant in individuals with T1D. The study failed to identify significant differences in thiamine plasma levels between healthy and diabetic individuals and did not correlate with *thiG* plasma levels. However, seven individuals in the cohort regularly ingested vitamin supplements that included thiamine, and this link between microbial thiamine biosynthesis and T1D would therefore be better investigated in a follow-up study where subjects would not ingest thiamine supplements.

Metagenomic studies have provided insight into several aspects of the human gut microbiome, including the compositional variability among individuals and metabolic functions occurring in the microbiome [42]. However, such data cannot provide information on the metabolic interactions among the different microbes found in the microbiome. A method that can help us investigate the metabolic exchanges among microbes in a community is constraint-based reconstruction and analysis (COBRA) [201] using genome-scale metabolic reconstructions (GENREs) [226] of the corresponding microorganisms. GENREs can be used to simulate the flow of metabolites in a system [226]. In short, based on the annotated genome of an organism, it is possible to derive the set of metabolic enzymes present in the cell. The corresponding metabolic reactions can be connected in a metabolic network, where the metabolites make up the nodes of the network, and the metabolic reactions link the nodes together [163]. Metabolic models derived from GENREs have been used to inspect the metabolic functions of microbial cells [148] as well as metabolic interactions between multiple microbes and host-microbe interactions [99].

Our aim was to map the different metagenomic read libraries from the study by Heintz-Buschart et al. to genomes of the AGORA collection of gut microbial GENREs [139] and create a personalized gut microbiome metabolic model for every sample. Because the AGORA GENREs were specifically curated for B-vitamin biosynthesis [140], we used metabolic modeling to investigate differences in thiamine biosynthesis between models based on healthy and diabetic microbiomes. We found no significant difference in thiamine uptake or secretion fluxes between microbiomes of healthy or diabetic individuals, but did observe differences between the four different families in the study. However, we identified a higher riboflavin biosynthesis capability in the microbiome models based on healthy individuals compared to those of diabetics. We found that the presence of the microbe *Prevotella copri* has the largest influence on the separation of the 53 samples based on their flux distributions when maximizing either thiamine or riboflavin exchange reaction in the models.

Taken together, this study illustrates that large and diverse microbiome metabolic reconstructions can be generated by mapping metagenomic reads from individual stool samples onto a large resource of GENREs [139]. The derived metabolic models show differences in their metabolic capabilities and such models could be applied in further studies to hypothesize about metabolic mechanisms underlying differences in gut microbiomes.

4.3 Methods

4.3.1 Metagenomic reads

We obtained metagenomic reads from 53 stool samples of 20 individuals from a multi-omics study on the human gut microbiome in type 1 diabetes by Heintz-Buschart, et al. (2016) [100]. In short, DNA was extracted from freshly frozen stool samples using 101 base pair (bp) paired-end sequencing of 350 bp inserts on a HiSeq2000 system (Illumina). The reads were quality-filtered and any reads mapping that mapped to a human genome were removed. The 53 individual metagenomic read libraries are available through NCBI BioProjects (PRJNA289586).

4.3.2 Mapping metagenomic reads to AGORA

We retrieved the genomes of 773 human gut microbial strains that were used to create a resource of 773 genome-scale metabolic reconstructions (AGORA v1.0) [139]. The genomes and corresponding reconstructions are available on the Virtual Metabolic Human database (http://vmh.life). All 773 AGORA genomes were concatenated while preserving the genome identifiers. For each of the 53 individual samples, the metagenomic reads were mapped to the concatenated genomes using Bowtie22 [125] with default parameters.

For every genomic contig, we calculated the coverage depth using SAMtools depth tool [130] and a perl script (adapted with correction from Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., & Nielsen, P. H. (2013)). Using a custom R [170] script, the depths of coverage and length of covered regions were aggregated for each genome. The relative abundance (RA) of every microbial strain i was estimated by normalizing the average coverage depth (CD) of the genome to the sum of all coverage depths in the sample and multiplying this by the proportion of metagenomic reads (MR) in the sample that mapped to any genome:

$$RA_i = \frac{CD_i}{CD_{total}} \cdot \frac{MR_{mapped}}{MR_{total}}$$

For every sample, we sorted the mapped AGORA genomes into three categories that represented the confidence level that the genome was found in the sample. Genomes that mapped with low confidence had a coverage breadth of minimum 10%. Genomes with good confidence had an average coverage depth of 1 and more than 60% coverage breadth. High confidence genomes had an average coverage depth higher than 8. All other genomes were defined as absent from the sample. For genomes with high confidence in a given sample, the coverage depths of all genes were extracted using the SAMtools bedcov [130] and an awk script [4].

4.3.3 Personalized gut microbiome reconstructions

For every individual sample, we extracted the list of AGORA reconstructions that mapped to the metagenomic reads with good confidence. Every AGORA reconstruction has two

compartments, the cytosol, $[c]$, and extracellular space, $[e]$. When combining the microbial reconstructions, we applied a compartmentalized approach [216, 219, 98] in which we define a new lumen compartment, $[u]$, that is shared by all the reconstructions. Dietary metabolites enter the lumen where they can be taken up by any of the AGORA reconstructions. The microbial reconstructions secrete metabolites into the lumen. These metabolites can either be taken up by the other reconstructions or removed from the lumen compartment as byproducts.

Every microbial reconstruction reserved its own biomass reaction where a placeholder metabolite called "Mic_biomass[c]" was produced, where "Mic" represents the microbe identifier in the reconstruction. To simulate the microbiome community biomass, we introduced a novel reaction whose substrates are each individual microbe reconstruction's biomass metabolite and the stoichiometric coefficients are the microbes' relative abundances in the sample. The "communityBiomass" function of a microbial community of n microbes is therefore represented by the following demand reaction:

$$RA_1BM1[c] + RA_2BM2[c] + \cdots + RA_nBMn[c] \rightarrow$$
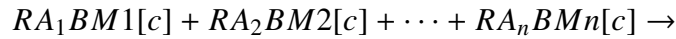
4.3.4 Coupling constraints

For every microbial strain in a microbiome model, we coupled the strain-specific metabolic reactions to the strain-specific biomass reaction as done by Heinken et al. (2013) [96]. This prevents one strain from utilizing the metabolic network of another strain without a non-zero flux through the biomass reaction of the second strain. In short, for all n strain-specific GEN-REs in the microbiome reconstruction, the flux through every reaction i (v_{in}) was coupled to the flux through the strain's biomass reaction (v_{Bn}). The ratio between the reactions was bound with a constant c such that $v_{in} - c \cdot v_{Bn} \geq u$ in the forward direction of all reactions. For reversible reactions, the reverse direction was bound with $v_{in} + c \cdot v_{Bn} \leq u$. The constant c was set to 400. The constant u represents a maintenance flux through the cell's reactions, i.e., when the cell maintains metabolic maintenance but does not actively divide. The constant u was set to 0 mmol/gDW/h.

4.3.5   Leak test

We performed a leak test to ensure that the microbiome community models could not "leak" metabolites, i.e., secrete metabolites without influx into the model. First, the lower bounds of all exchange, demand, and sink reactions were set to zero. Then, one by one, every exchange reaction was set as the objective function and the flux through the reaction optimized using flux balance analysis (FBA) [159]. If the maximum flux through the reaction was higher than the threshold 1E-6, the model was defined as "leaking" the corresponding metabolite. None of the 53 models leaked any metabolites.

4.3.6   Individual metadata

Metadata was gathered on the 20 individuals from the supplementary data form the study by Heintz-Buschart et al. [100] (Table 4.1). Eight of the 20 individuals reported taking dietary supplements containing either riboflavin or thiamine (Table 4.1).

4.3.7   Simulations

All simulations were performed using Matlab 2016a (Mathworks, Inc., Natick, MA, USA) using the COBRA Toolbox [201] (https://opencobra.github.io/) and the linear programming solver CPLEX (IBM, Inc.) through the Tomlab interface (Tomlab, Inc.).

Constraints

We simulated the metabolic potential of all 53 personalized microbe community models on a Western diet Table B.18 under anaerobic conditions. Based on Table I (column "Weight of faecal solids (g/day)") and Table II (column "Percentage of solids in fraction C") from a study by Stephen and Cummings (1980) [215], we calculated the hourly human fecal bacterial output and found it to be in the range of 0.44–0.73 gDW/h. We applied this range to the microbiome community biomass reaction in every model by setting the lower bound to 0.44 1/h and the upper bound to 0.73 1/h.

Table 4.1: Metadata on the 20 subjects from the study by Heintz-Buschart et al. [100]. Thiamine and riboflavin supplement amounts are reported on a weekly (w) or daily (d) basis. F: Female. M: Male. BMI: Body mass index. BMI is reported for all visits separately (V1-V3).

Individual	Diabetic	Age	Gender	BMI (V1 - V2 - V3)	Thiamine (mg)	Riboflavin (mg)
M1D1	Yes	62	M	25.2 - 27.19 - NA	1.4 (w)	1.6 (w)
M1D2	No	60	F	26.39 - 26.24 - NA	0	0
M1D3	No	57	F	37.12 - 37.21 - 36.99	0	0
M1D4	Yes	37	F	36.52 - 36.33 - 37.11	0	0
M1D5	No	6	M	NA - 18.31 - NA	0	0
M2D1	Yes	58	M	23.17 - 23.13 - 23.25	0	0
M2D2	Yes	24	F	19.54 - 19.44 - 19.22	0	0
M2D3	Yes	30	M	20.9 - 20.23 - 20.28	0	0
M2D4	No	58	F	20.76 - 20.59 - 20.57	60 (d)	0
M2D5	Yes	28	F	21.37 - 21.05 - NA	60 (d)	0
M3D1	No	50	F	NA - 16.93 - 17.03	1.375 (d)	1.61 (d)
M3D3	No	14	F	19.72 - 18.65 - 19.86	0.66 (d)	0.98 (d)
M3D4	Yes	17	F	21.07 - 20.57 - 20.51	1.5 (d)	1.7 (d)
M3D5	Yes	19	F	24.54 - NA - NA	1.375 (d)	1.61 (d)
M4D1	No	34	F	38.48 - NA - 37.96	0	0
M4D2	Yes	5	M	14.01 - 14.74 - 14.65	0	0
M4D3	Yes	9	M	24.47 - NA - 24.57	0	0
M4D4	No	12	F	29.9 - NA - 29.84	0	0
M4D5	No	16	F	29.32 - NA - 28.7	0	0
M4D6	No	43	M	35.35 - 34.89 - 35.82	0	0

Minimum thiamine uptake flux

For every microbiome model, we applied the constraints described in the "Constraints" section above. We set the thiamine exchange reaction ("EX_thm[u]") as the objective function of the models and maximized the flux through the objective function using FBA. Two microbiome models representing the samples "M2D1V3" and "M3D5V2' were infeasible under the set constraints. For the remaining models, the maximum possible flux through the objective function was negative, meaning that dietary thiamine uptake was essential in all the models and none of the microbiome community models can secrete thiamine.

Maximum riboflavin secretion flux

For every microbiome model, we applied the constraints described in the "Constraints" section above. We set the riboflavin exchange reaction ("EX_ribflv[u]") as the objective func-

tion of the models and maximized the flux through the objective function using FBA. Two microbiome models representing the samples "M2D4V1" and "M3D4V3" were infeasible under the set constraints. The remaining 51 models were feasible with a positive maximum flux through the objective function, meaning that all the models could secrete riboflavin.

Thiamine and riboflavin secreting microbe models

For all 773 microbial GENREs, we checked whether the corresponding metabolic model on unlimited media could secrete thiamine or riboflavin. We set the lower bounds on all exchange reactions in the model to -1000 mmol/gDW/h and set the lower bound of the microbe's biomass reaction to 0.001 1/h to ensure that minimal biomass could be produced. We then set the exchange reaction for either thiamine ("EX_thm[u]") or riboflavin ("EX_ribflv[u]") as the objective function and maximized the flux through the reaction using FBA. If the maximum flux through the objective reaction was higher than 1E-6 mmol/gDW/h, the microbe model was classified as a thiamine or riboflavin producer.

4.3.8   Statistics

P-values of differences in maximum secretion flux values between healthy and diabetic microbiomes were calculated using the Wilcoxon rank-sum test method implemented in Matlab with default parameters.

4.4   Results

First, we assessed the effects of different thresholds on the metagenomic reads mapping parameters on the resulting microbiome metabolic reconstruction sizes and content. By comparing the taxonomic and functional content of the microbiome reconstructions under different cut-off levels, we determined appropriate thresholds for deciding which genomes to include in the individual microbiome reconstructions. Second, we compared the reconstruction sizes with the subjects' age, body mass index, gender, and diabetic status. Finally, we simulated the individual models' biosynthesis capabilities of the two B-vitamins thiamine and riboflavin, and despite high variability among the different models, we saw significant

differences in B-vitamin biosynthesis of models based individuals belonging to different families, genders, and diabetic statuses.

4.4.1   Metagenomic mapping to AGORA genomes

We mapped the metagenomic reads from 53 individual stool samples [100] onto the set of 773 microbial genomes underlying the AGORA reconstructions (Chapter 3). On average, half of the metagenomic reads from every sample could be mapped to the AGORA genomes ($53 \pm 12\%$, Fig. 4.1).
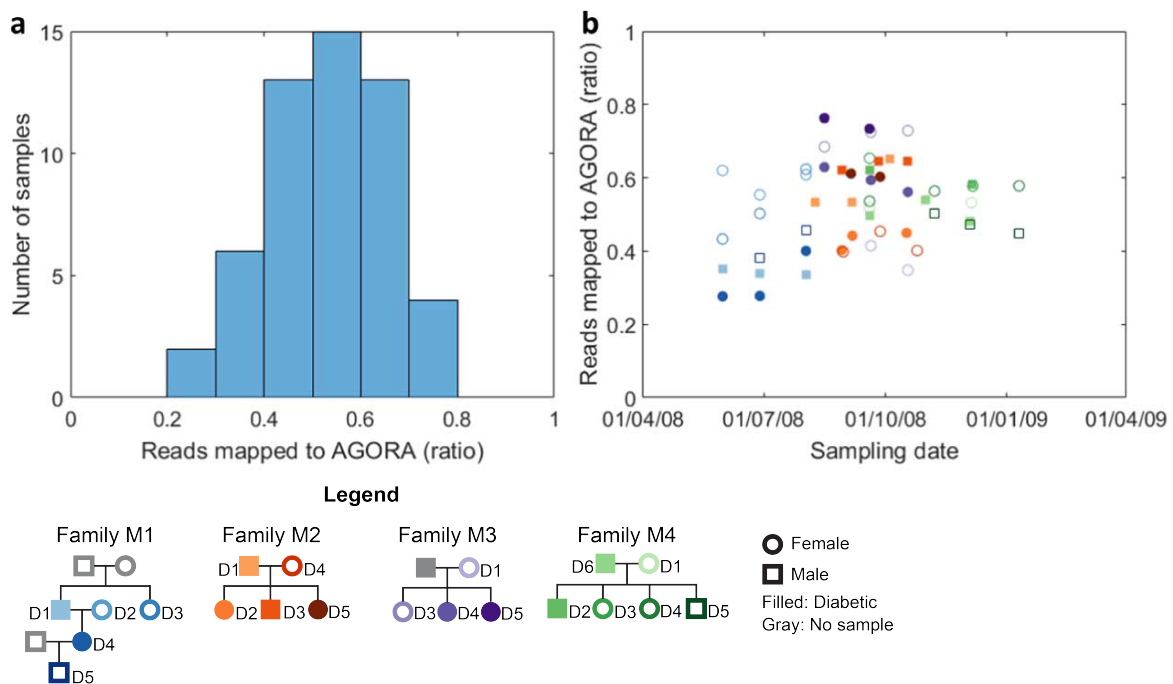
Figure 4.1: Ratio of metagenomic reads that mapped to any of the 773 AGORA genomes. a) Histogram of the ratio of metagenomic reads that mapped to the 53 samples. b) The ratio of metagenomic reads that mapped to the 53 samples as a function of their sampling date. The family trees, colors, and shapes are based on those in figures from the study by Heintz-Buschart et al. [100]. Individual identifiers correspond to those in Table 4.1.

For every genome sequence, we calculated the percentage of the sequence that was mapped by any read (coverage breadth) and how often each part of the genome was mapped by reads (coverage depth). Based on the average coverage depth of the genome and the ratio of mapped reads, we calculated the relative abundance of every mapped genome in the

sample (Fig. 4.2a).

To determine which microbes should be included in each microbiome model, we defined three confidence levels based on the average coverage depth and breadth: i) at least 10% of the genome is mapped to by metagenomic reads, ii) at least 60% of the genome is mapped to by metagenomic reads and the average coverage depth of the genome is one, and iii) the genome has an average coverage depth of eight. The confidence levels greatly affected the microbiome community size (Fig. 4.2b). As expected, the higher the thresholds, the fewer microbes could be mapped per sample. Additionally, we observed that adding microbes to the community increased number of metabolic functions that occur in any of the microbes in the community (Fig. 4.2c).

Based on the aforementioned results, we chose to simulate microbiome models based on genomes that were identified with the middle confidence level, i.e., with coverage breadth larger than or equal to 60% and a minimum average coverage depth of one. The average number of strains per microbiome model was 45 ± 9 (Fig. 4.2b). This confidence level was chosen because these microbiomes capture significantly more microbe strains (p=7E-19), reactions (p=1E-16) and subsystems (p=4E-16) than the higher confidence level (Figs. 4.2b-c), while still having high confidence that the microbes mapped are appropriate. The lower confidence level captures more microbes, but we expect a higher level of false-positive microbes in those microbiomes.

4.4.2   Microbiome model features

The average number of genomes that mapped to each sample was 45 ± 9 (Fig. 4.3, Table 4.2). The number of genomes mapped to each sample did not correlate with age, body mass index, diabetic status, or gender (Fig. 4.3). However, it was noted that the larger microbiomes, i.e., 55 or more strains, were only found in samples of young (<20 years) and older (50+ years) individuals.

The total relative abundance of mapped genomes in each sample was 0.43±0.11 (Fig. 4.4a, Table 4.2). The total relative abundance did not correlate with the number of genomes that mapped to each sample (Fig. 4.4b).

Table 4.2: Number of microbial strains and relative abundance (RA) in the 53 individual microbiome models at the three different coverage depth and breadth cut-off levels.

Individual	Visit 1		Visit 2		Visit 3		Average ± sd	
	Strains	RA	Strains	RA	Strains	RA	Strains	RA
M1D1	50	0.27	46	0.28	50	0.25	49 ± 2	0.27 ± 0.02
M1D2	63	0.52	50	0.48	54	0.53	56 ± 7	0.49 ± 0.06
M1D3	44	0.36	45	0.42	43	0.51	44 ± 1	0.43 ± 0.08
M1D4	33	0.20	47	0.21	34	0.32	38 ± 8	0.24 ± 0.07
M1D5	NA	NA	49	0.31	41	0.37	45 ± 6	0.34 ± 0.04
M2D1	42	0.43	41	0.42	52	0.56	45 ± 6	0.47 ± 0.08
M2D2	32	0.32	39	0.35	39	0.36	37 ± 4	0.34 ± 0.02
M2D3	36	0.52	39	0.54	41	0.54	39 ± 3	0.53 ± 0.01
M2D4	57	0.33	32	0.33	63	0.33	51 ± 16	0.33 ± 0.00
M2D5	41	0.48	37	0.47	NA	NA	39 ± 3	0.48 ± 0.01
M3D1	NA	NA	50	0.32	53	0.27	52 ± 2	0.30 ± 0.04
M3D3	57	0.55	66	0.61	68	0.63	64 ± 6	0.60 ± 0.04
M3D4	51	0.51	52	0.46	64	0.46	56 ± 7	0.48 ± 0.03
M3D5	38	0.62	50	0.58	NA	NA	44 ± 8	0.60 ± 0.03
M4D1	35	0.41	NA	NA	34	0.42	35 ± 1	0.42 ± 0.01
M4D2	37	0.36	51	0.44	31	0.33	40 ± 10	0.38 ± 0.06
M4D3	49	0.52	NA	NA	40	0.50	45 ± 6	0.51 ± 0.01
M4D4	52	0.55	NA	NA	42	0.50	47 ± 7	0.53 ± 0.04
M4D5	39	0.40	51	0.45	43	0.50	44 ± 6	0.45 ± 0.05
M4D6	40	0.44	33	0.41	43	0.39	39 ± 5	0.41 ± 0.03

4.4.3 Microbiome vitamin biosynthesis

None of the microbiome models could secrete thiamine and we therefore assessed the minimum thiamine requirements of each microbiome model. The microbiome models of healthy and diabetic individuals did not have significantly different thiamine requirements ($p=0.93$, Fig. 4.5a). However, when inspecting the thiamine requirements of different families, we found that families 1 and 4 require a significantly lower influx of thiamine than those of families 2 and 3 ($p=0.0037$–0.035). We inspected this difference among families and found that the major difference is that several individuals in families 1 and 4 have *Prevotella copri* strains presents, whereas none of the samples from families 2 and 3 do.

We observe a mild difference in the maximum possible riboflavin secretion between the diabetic and healthy microbiome models ($p=0.033$, Fig. 4.6a). However, when we split the samples further by gender, we see that the microbiomes based on healthy females can secrete higher fluxes of riboflavin than those based on diabetic females ($p=0.011$, Fig. 4.6b).

No difference in riboflavin secretion was observed among the male samples (p=0.82).

We further performed a principal component analysis (PCA) on the flux vectors that resulted in the differences in thiamine and riboflavin biosynthesis fluxes. In both cases, by looking at the separation of the samples by the first component, we saw that most samples from families 1 and 4 were separated from the others (Fig. 4.7a,b). We inspected the underlying scores, and found that the separation of the first component was mostly driven by reactions from the reconstruction *Prevotella copri* CB7 DSM 18205, which was only found in several samples from families 1 and 4, but not in any samples from families 2 and 3. The separation on the second component was found to be driven by exchange reactions of fermentation productions (Fig. 4.7c-d).

In light of these results, we investigated the effect of relative abundance of *P. copri* on the minimum possible thiamine uptake fluxes. Although we did not see a correlation between the relative abundance of *P. copri* and the minimum thiamine uptake fluxes (Fig. 4.8a), we saw a significant difference (p=2.1E-5) between microbiome models that contain *P. copri* in any abundance compared with those microbiome models that do not contain the microbe (Fig. 4.8b). No such effects were observed on the maximum riboflavin secretion fluxes (p=0.24, Fig. 4.8c-d).

Taken together, we have shown that it is possible to combine metagenomics with genome-scale metabolic reconstructions to create personalized microbiome metabolic models. The chosen cut-off parameters for selecting the mapped microbes have a large influence on the size and functionality of the resulting microbiome models. Finally, the different microbiome metabolic models showed variable metabolic capabilities when comparing maximum riboflavin secretion and the minimum thiamine uptake fluxes, while retaining differences between groups such as gender, families, and diabetic status.

4.5 Discussion

In this study, we mapped metagenomic reads from 53 individual stool samples to a reference library of microbial genomes to create personalized gut microbiome metabolic models. We illustrate the importance of choosing appropriate thresholds for metagenomic mapping coverage as the number of mapped microbes per sample depends on the chosen coverage depth

threshold. We observed that the microbiome models that were based on healthy individuals could secrete higher fluxes of riboflavin than those of diabetic individuals. Finally, we found that microbiomes containing the microbe *P. copri* required significantly lower uptake fluxes of thiamine than other microbiomes.

We mapped metagenomic reads from an unrelated data set to the genome sequences underlying the AGORA resource [139]. On average, more than half of the total metagenomic reads from each sample aligned to the AGORA genomes (Fig. 4.4), which shows that the AGORA microbes are suitable for applications in human gut microbiome studies. We observed high variability among the different samples, with the lowest mapping ratio of 28% and the highest mapping ratio of 76%. However, since 40% of the samples had less than half of the metagenomic reads mapped, the AGORA resource could be expanded to increase metagenomic read coverage for future studies.

We illustrated the importance of carefully choosing appropriate parameters when mapping metagenomic reads to a set of microbial genomes. Not only does the number of identified genomes highly depend on these coverage depth and breadth, but the taxonomic and functional diversity is largely affected as well (Fig. 4.2). Therefore, choosing a higher threshold will increase the confidence that the identified microbes are biologically correct, but at the same time will exclude biologically meaningful microbes that are in low abundance. Even though these microbes are in lower abundance, they can have important roles in the microbial community.

When creating gut microbiome metabolic models, in addition to parameter thresholds, we are limited by the fact that we can only map the metagenomic reads to the genomes of a set of microbes that have genome-scale metabolic reconstructions available. Currently, it is not possible to fully automatically generate biologically correct GENREs. However, the work presented in Chapter 3 has enabled us to generate individualized gut microbiome models using hundreds of genomes for our reference library.

The microbiome models of healthy individuals could secrete higher fluxes of riboflavin than the models of diabetic microbiomes (Fig. 4.6). This difference in riboflavin production capabilities is interesting in the light of a study by Cole et al. (1976) [40], which showed that riboflavin deficiency was more common in diabetic children than in non-diabetics. The riboflavin secretion of our *in silico* models did not correlate with the number or summed rela-

tive abundance of the microbial strains in the model, suggesting that the metabolic cross-talk between microbes community affects the maximum riboflavin secretion of the community. We also found a significant difference between the minimum thiamine uptake fluxes between models based individuals of different families (Fig. 4.5b).

We observed that the metabolic flux differences among the 53 samples when optimizing for minimum thiamine uptake and maximum riboflavin secretion was driven by the presence of *Prevotella copri* in the microbiomes. The *P. copri* genome suggests that the microbe can synthesize both riboflavin and thiamine (Chapter 2, [140]) and the proposed microbiome "enterotype" that is rich with organisms from the *Prevotella* genus was reported to have higher amounts of genes involved in thiamine biosynthesis [10]. The observed effect on minimum thiamine uptake of the models was not correlated with the relative abundance of *P. copri* (Fig. 4.8a). Interestingly, the original study by Heintz-Buschart et al. reported a correlation between the expression levels of thiamine biosynthesis gene *thiG* and amylase proteins, which were less abundant in diabetic individuals. The *thiG* of highest abundance were linked to *P. copri* contigs. However, it must be noted that the aforementioned results are based on non-unique flux balance analysis [159] flux vectors. It would be of value to repeat the analysis using the minimum and maximum flux values from a flux variability analysis [89] or flux value probabilities from sampling [91].

Taken together, we have shown that metagenomic reads from stool samples can be aligned with genomes of AGORA, a resource of gut microbial GENREs, to create personalized microbiome metabolic models. We have shown that the resulting microbiomes are taxonomically and functionally diverse and show significant differences in the maximum secretion and minimum uptake fluxes of two B-vitamins. Similar methods could be applied in various studies of the human gut microbiome, e.g., to investigate the effects of dietary metabolites on metabolic capabilities of different microbiome models or the metabolic exchanges among the individual microbe strains. By coupling the microbiome reconstructions to the human metabolic reconstruction Recon 2 [227], it would be possible to assess the metabolic effects of different microbiomes on human metabolism.
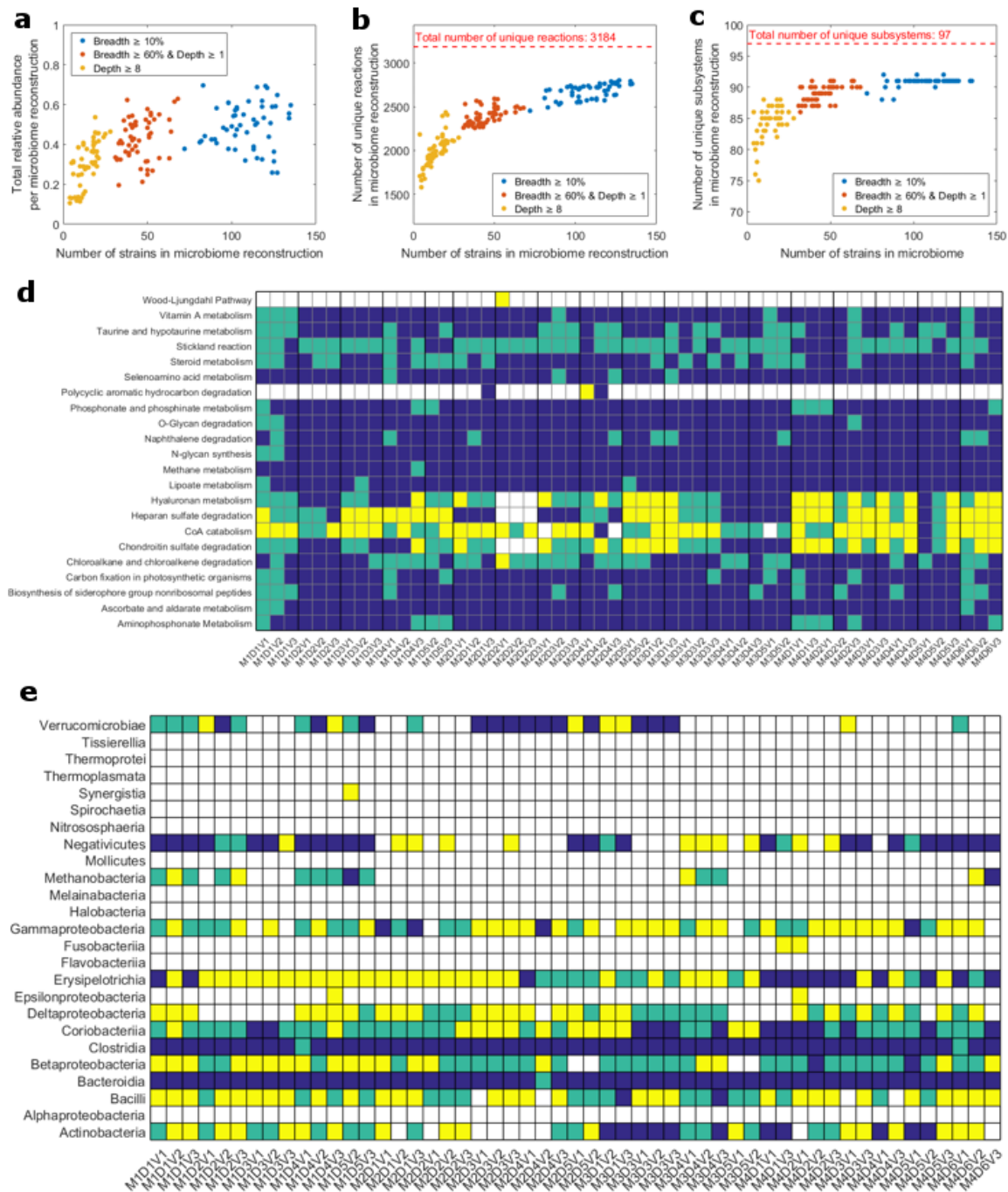
Figure 4.2: Relative abundance, unique reactions and subsystems, and taxonomic classes in each of the 53 microbiomes under the three confidence levels. a) Total relative abundance captured per sample plotted by the number of strains mapped per sample. b) The number of unique reactions in each microbiome model as a function of the number of microbial strains in the microbiomes. c) The number of unique subsystems in each microbiome model as a function of the number of microbial strains in the microbiomes. d) Presence of subsystems in each sample at the different confidence levels. e) Presence of taxonomic classes in each sample at the different confidence levels. Blue: Present in all confidence levels. Green: Present in confidence levels 1 (coverage breadth ≤ 10%) and 2 (coverage breadth ≤ 60%, coverage depth ≤ 1). Yellow: Present only in confidence level 1. White: Not present in any confidence levels.
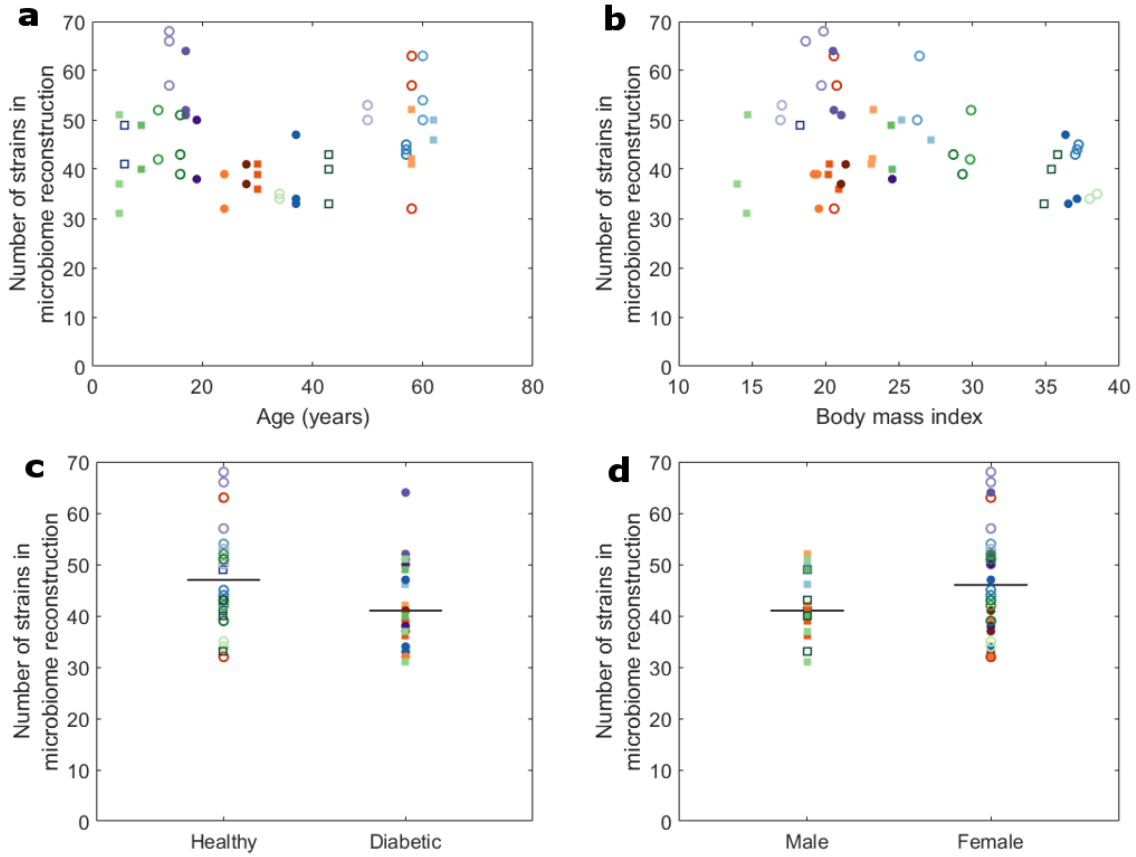
Figure 4.3: The effects of a) age, b) Body mass index (BMI), c) gender, and d) diabetic status on the number of microbial strains in the 53 microbiome reconstructions. For marker legend, refer to Figure 4.1.
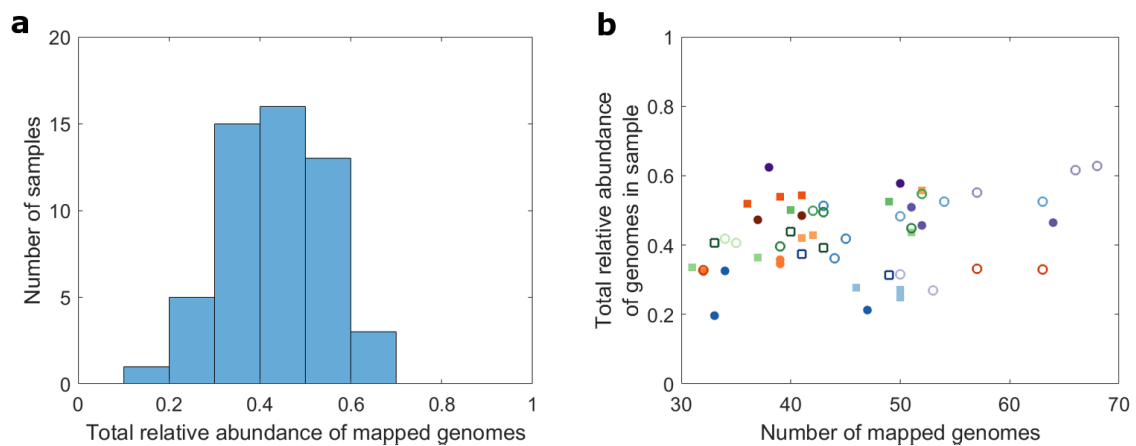
Figure 4.4: Total relative abundance captured in the 53 microbiomes based on the middle confidence level, i.e., min. coverage breadth of 60% and min. average coverage depth of one. a) Histogram of the total relative abundance per microbiome. b) Total relative abundance per microbiome plotted against the number of microbial strains. For marker legend, refer to Figure 4.1.
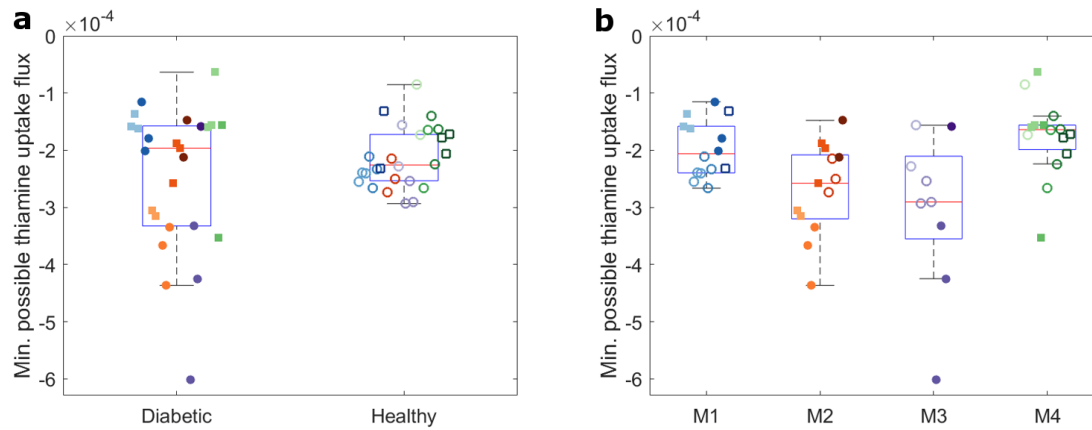
Figure 4.5: Minimum possible thiamine uptake fluxes (mmol/gDW/h) per microbiome. a) Comparison of min. thiamine uptake fluxes between healthy and diabetic individuals. b) Comparison of min. thiamine uptake fluxes between individuals from different families. The graphs show box plots based on the data from each group, overlaid with the data points showing the different samples. For marker legend, refer to Figure 4.1.
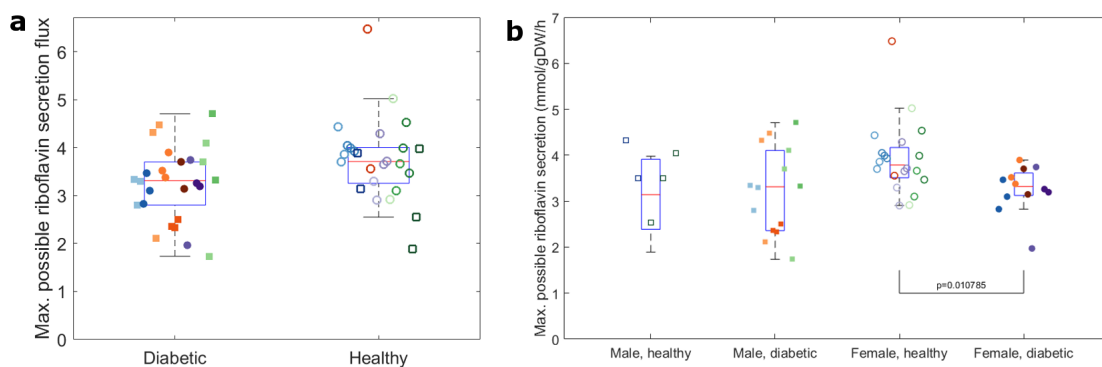
Figure 4.6: Maximum possible riboflavin secretion fluxes (mmol/gDW/h) per microbiome. a) Comparison of max. riboflavin secretion fluxes between healthy and diabetic individuals. b) Comparison of max. riboflavin secretion fluxes between healthy and diabetic individuals separated by gender. The graphs show box plots based on the data from each group, overlaid with the data points showing the different samples. For marker legend, refer to Figure 4.1.
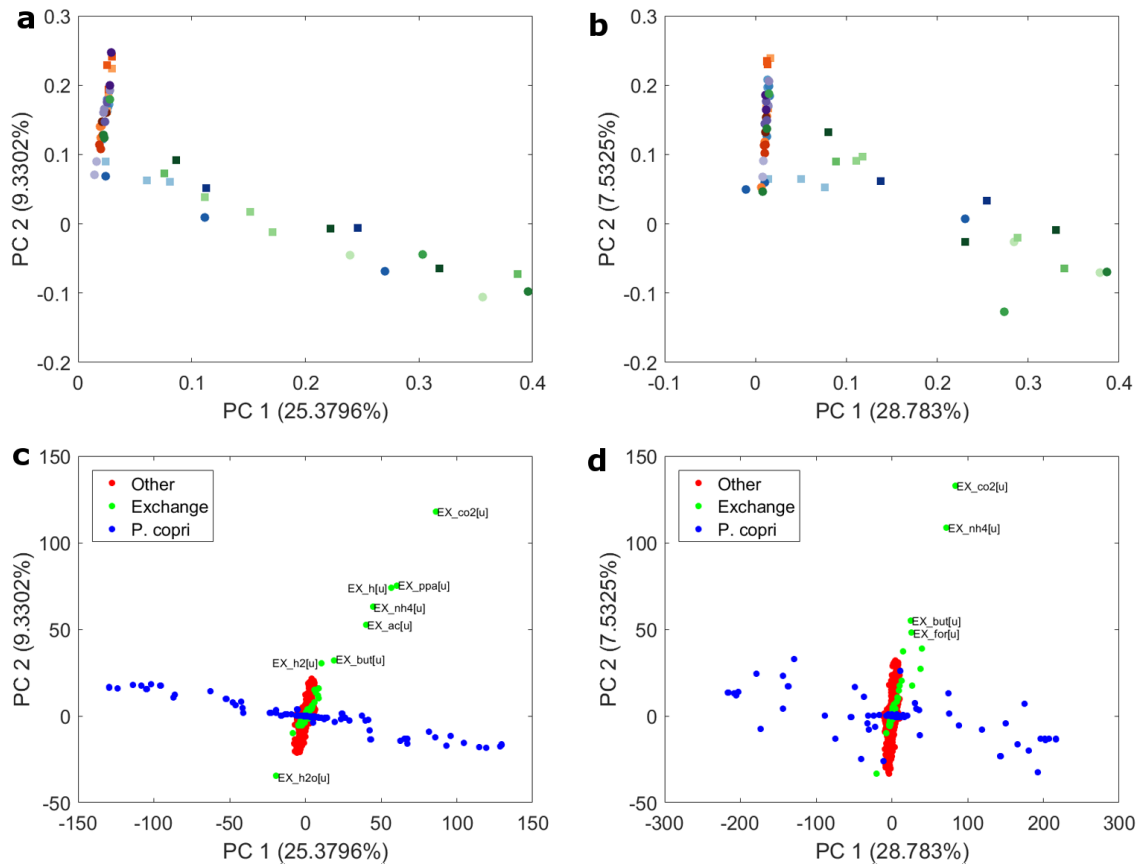
Figure 4.7: Principal component analysis (PCA) of FBA solution flux vectors from a) min. thiamine uptake simulations and b) max. riboflavin secretion. For marker legend, refer to Figure 4.1. Score plots of PCA results based on c) min. thiamine uptake simulations and d) max. riboflavin secretion. Reactions belonging to *Prevotella copri* are colored in blue, exchange reactions are marked with green.
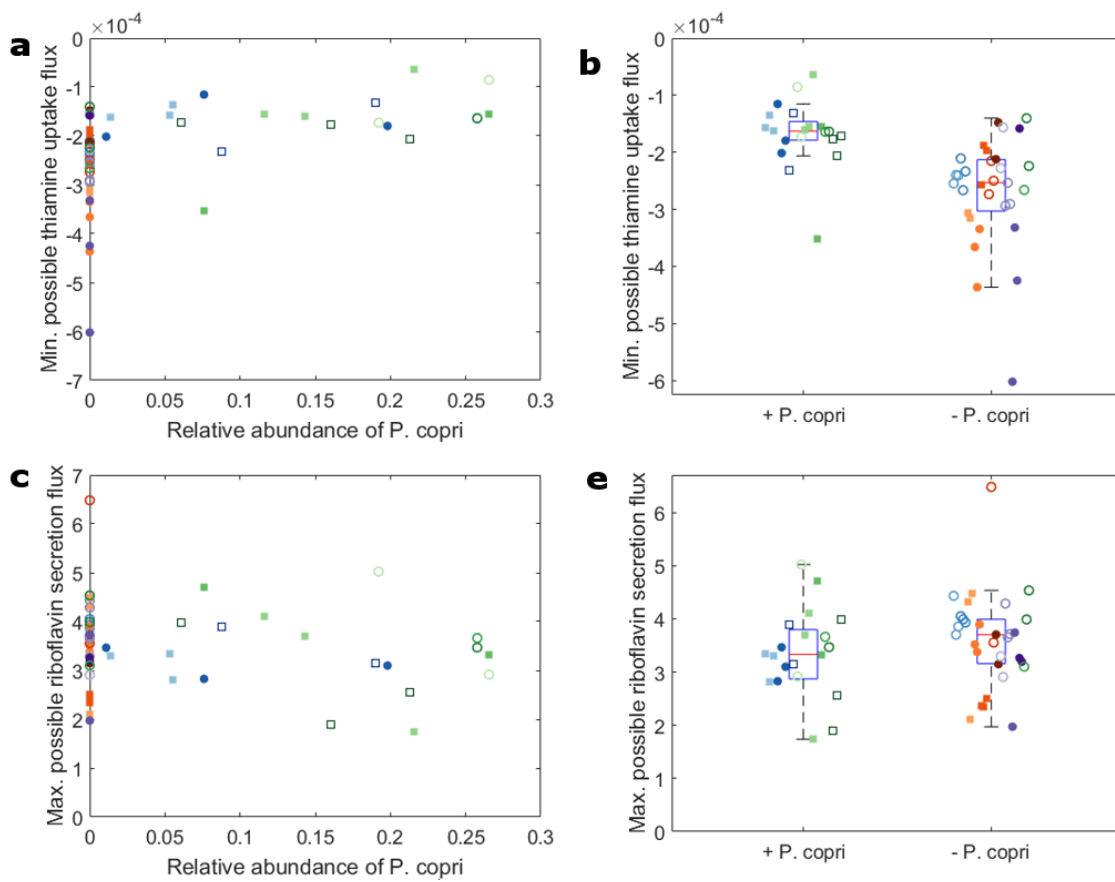
Figure 4.8: a) Min. possible thiamine uptake fluxes and c) max. possible riboflavin secretion fluxes plotted as a function of relative abundance of *P. copri*. b) Min. possible thiamine uptake fluxes and d) max. possible riboflavin secretion fluxes of microbiome models with (+) and without (-) *P. copri*. For marker legend, refer to Figure 4.1.

Chapter 5

Concluding remarks

Since the emergence of high-throughput methods that can be applied to study the various omics of the human gut microbiome, one of the largest challenges is to analyze the data that is being produced. Metagenomics analyses of individual gut microbiomes have identified microbial dysbiosis in several diseases, but the detailed mechanisms behind these implications have yet to be determined. A major factor contributing to this is the complexity of the gut microbiome. Every individual harbors hundreds of microbial species in their gut [168], and even more numerous strains. In addition, the microbial composition very diverse between individuals [42] and is affected by changes in dietary composition [47]. One method to elucidate the metabolic interactions within the gut microbiome is to use constraint-based reconstruction and analysis (COBRA). Before the work of this thesis, one of the issues with using COBRA modeling methods to analyze the human gut microbiome was that very few gut microbe GENREs had been built (Table 1.2), thus limiting the number of microbes that could be simulated simultaneously in a single microbiome metabolic model.

The work of this thesis describes the use of comparative genomics to analyze B-vitamin biosynthesis in human gut microbes (Chapter 2). The results showed that many intestinal microbes lack the ability to synthesize B-vitamins and since most dietary B-vitamins are absorbed through the small intestine [192], it is likely that these microbes acquire these essential nutrients from their neighboring bacteria, indicating cooperation among the microbes in the gut. The comparative genomics data of B-vitamin biosynthesis, in addition to similar analyses of multiple other vitamin and central metabolic pathways, was integrated into the development of "assembly of gut organisms through reconstruction and analysis" (AGORA),

a resource of 773 GENREs representing human gut microbes (Chapter 3). The AGORA reconstructions were further refined based on experimental data of carbohydrate uptake and fermentation pathways. The generation of AGORA enabled the creation of personalized human gut microbiome models that contained an average of 45 microbial strains per model (Chapter 4). These personalized models showed differences in taxonomic and metabolic subsystem composition, and we observed differences in B-vitamin metabolism between different groups of individuals based on gender, family, and diabetic status.

5.1   Comparative genomics to predict microbial metabolism

Even though comparative genomics studies can help us elucidate metabolic pathways in human gut microbes [173, 174, 140, 175, 176], the method has some limitations.

First, predicting the presence of a metabolic pathway in an organism requires a consistent set of rules to determine which functional roles should be present in the genome to define the pathway as present or absent. However, such rules require detailed knowledge about the metabolic pathway. For example, if a genome that is missing a single functional role from a pathway is automatically predicted to contain the metabolic function, it is still possible that the role in question has a key function in the pathway, and that without the particular enzyme the function cannot be performed.

Second, even though many human gut microbes have now been successfully cultured [172], detailed growth experiments, biochemical assays, and gene essentiality analyses have not been carried out for most of these microbes. As shown in Chapter 2, most of the comparative genomics predictions of B-vitamin biosynthesis were accurate (Table 2.4). However, this comparison was performed on few microbes since limited experimental data exists on the B-vitamin requirements of human gut microbes. Once more detailed growth experiments and biochemical knowledge is available on individual human gut microbial strains, it would be interesting to extend such evaluations of the comparative genomics results in order to assess how accurately they capture gut microbiome metabolism.

5.2 Gut microbial metabolic reconstructions

The generation of the AGORA resource has enabled us to build biologically relevant large-scale gut microbiome metabolic reconstructions. Such reconstructions were not possible before because of the lack of refined human gut microbe GENREs (Table 1.2).

The AGORA reconstructions were curated with a focus on metabolic functions concerning the human diet, i.e., simple sugar and polysaccharide uptake, carbohydrate fermentation, and vitamin metabolism. While these curations make AGORA suitable to study dietary effects on gut microbiome metabolism, we cannot model the microbiome effects on drug metabolism or other xenobiotics, which are also important factors in biomedicine. These pathways were left out of AGORA because xenobiotic and drug metabolism is not yet well studied in human gut microbes. These metabolic functions could be added to the AGORA reconstructions in future efforts when more data is available on the metabolism of such compounds by human gut microbes. Such an extension would make it possible to model microbial roles in human drug metabolism as efforts to include drug metabolism in the human metabolic reconstruction Recon 2 [227] have already been started by Sahoo et al. (2015) [189].

The AGORA resource should be continuously updated as growth experiments, gene studies, and information on nutrient requirements become available on the individual strains. The reconstructions were curated using a propagation method where refinements from one reconstructions were propagated to other reconstructions with the same gaps. This refinement method enabled the generation of hundreds of microbial GENREs, but such a method is only applicable in the absence of detailed biochemical evidence on the organisms. The propagation method ensures that the condition-specific metabolic models can perform the metabolic functions that were reported in literature. However, it is possible that the specific metabolic pathway that the individual strains utilize in nature are different than the propagated pathway, and such cases will need to be refined as strain-specific biochemical information becomes available.

5.3   Human gut microbiome metabolic reconstructions

When creating personalized gut microbiome models, we are limited by the number of biologically relevant gut microbial GENREs. The AGORA resource has made it possible to create such microbiome reconstructions. However, the human gut microbiome is very diverse, as demonstrated by more than 1,100 human gut microbial strains identified in a single study cohort [168]. Presently, it is not possible to automatically generate biologically feasible GENREs, and we are thus biased to include the microbial reconstructions that are available in AGORA. As observed in Chapter 4, the AGORA resource could be mapped to half of the metagenomic reads from an independent study [100] (Fig. 4.1). Future efforts could extend the AGORA resource with additional human gut microbial strains to capture a larger ratio of individual metagenomic samples.

5.4   COBRA in microbiome modeling

COBRA modeling is based on a steady-state assumption, meaning that metabolite concentrations in the microbial cell or microbiome are stable over time [163]. This approach has proven to be very useful when studying single cell metabolism and small microbial communities [157], but since it models snapshots in time, COBRA cannot be applied on its own to study dynamic changes in metabolism [163]. However, as previously mentioned, several methods have been developed that couple COBRA with other modeling methods in order to simulate dynamic changes in single cell or microbial community metabolism (see Chapter 1).

COBRA is a highly scalable modeling technique and is suitable for simulating large-scale microbiome reconstructions [101]. However, the some of the limitations we currently face in simulating such large models are computational power, data analysis, and experimental design. Simulations with large-scale metabolic model can generate large amounts of data; a single FBA optimization on a microbiome model from Chapter 4 will generate a flux vector with about 50,000 flux values. This amount doubles when looking at FVA data, and is multiplied further if looking at a data set from sampling. It is therefore necessary to have clear goals and analysis strategies before starting simulations on such large models.

Further analyses of large-scale microbiome metabolic models are needed in order to in-

vestigate which factors in the microbiome reconstructions can influence the results and in what way. For example, should the dietary input fluxes be scaled based the number of microbes in a microbiome community model or the relative abundances of the individual microbes? How do personalized microbiome models compare with randomly generated models of diverse sizes? How do the community biomass reaction flux limits affect the metabolic interactions within the microbiome? It is necessary to address such concerns before using such large-scale microbiome models to generate hypotheses on metabolic differences between gut microbiomes.

5.5 Future perspectives

Personalized human gut microbiome metabolic models have several possible applications. One application is to identify metabolic differences between groups of individuals as was done in Chapter 4. Such results could be useful when designing experimental studies, e.g., to help identify metabolites of interest that could be tested in the subjects. Another use would be to identify dietary metabolites or probiotic bacteria that could be ingested by one group to drive their metabolic features closer to the other group, e.g., bringing the metabolic regimes of diseased microbiomes towards those of healthy individuals. Gut microbiome metabolic models could also be used to identify "keystone" microbes in the human gut, e.g., by removing individual microbe reconstructions from the microbiome model and comparing the impact on the reduced microbiome's metabolism.

A future step in personalized metabolic modeling of the gut microbiome is to couple metegenomics-based microbiome models with human metabolism. Additionally, individual-specific genomic data could be integrated with the human-microbiome metabolic reconstructions, resulting in further individual-specific simulations of human-micobiome metabolism. Such models could provide insights into how different individuals respond differently to variations in the gut microbiome or diet. Personal nutrition could be applied to the different microbiomes to see how they could differently affect human metabolism under a variety of dietary regimes. However, as mentioned before, the gut microbiome composition is heavily influenced by the diet [47]. An individual's diet varies from day to day, as well as throughout the day. It might therefore be more applicable to design dietary model inputs based on a

weekly food journal and averaging the nutrient levels out instead of looking at a single day to assess the diet. Personalized metabolic models could also be applied to design personal diets for different purposes, e.g., to reduce the metabolic flux through a certain metabolic pathway, similar to the phenylalanine-reduced diets of phenylketonuria patients [247].

Bibliography

[1] Abbas, C. A. and Sibirny, A. A. (2011). Genetic control of biosynthesis and transport of riboflavin and flavin nucleotides and construction of robust biotechnological producers. *Microbiol Mol Biol Rev*, 75(2):321–60.

[2] Agren, R., Liu, L., Shoaie, S., Vongsangnak, W., Nookaew, I., and Nielsen, J. (2013). The raven toolbox and its use for generating a genome-scale metabolic model for penicillium chrysogenum. *PLoS Comput Biol*, 9(3):e1002980.

[3] Ahn, J., Sinha, R., Pei, Z., Dominianni, C., Wu, J., Shi, J., Goedert, J. J., Hayes, R. B., and Yang, L. (2013). Human gut microbiome and risk for colorectal cancer. *JNCI: Journal of the National Cancer Institute*, 105(24):1907.

[4] Aho, A. V., Kernighan, B. W., and Weinberger, P. J. (1987). *The AWK Programming Language*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[5] Allen, R. H. and Stabler, S. P. (2008). Identification and quantitation of cobalamin and cobalamin analogues in human feces. *Am J Clin Nutr*, 87(5):1324–35.

[6] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–10.

[7] Anderl, J. N., Franklin, M. J., and Stewart, P. S. (2000). Role of antibiotic penetration limitation in klebsiella pneumoniae biofilm resistance to ampicillin and ciprofloxacin. *Antimicrob Agents Chemother*, 44(7):1818–24.

[8] Andres, E., Loukili, N. H., Noel, E., Kaltenbach, G., Abdelgheni, M. B., Perrin, A. E., Noblet-Dick, M., Maloisel, F., Schlienger, J. L., and Blickle, J. F. (2004). Vitamin b12 (cobalamin) deficiency in elderly patients. *Cmaj*, 171(3):251–9.

[9] Arkin, A. P., Stevens, R. L., Cottingham, R. W., Maslov, S., Henry, C. S., Dehal, P., Ware, D., Perez, F., Harris, N. L., Canon, S., Sneddon, M. W., Henderson, M. L., Riehl, W. J., Gunter, D., Murphy-Olson, D., Chan, S., Kamimura, R. T., Brettin, T. S., Meyer, F., Chivian, D., Weston, D. J., Glass, E. M., Davison, B. H., Kumari, S., Allen, B. H., Baumohl, J., Best, A. A., Bowen, B., Brenner, S. E., Bun, C. C., Chandonia, J.-M., Chia, J.-M., Colasanti, R., Conrad, N., Davis, J. J., DeJongh, M., Devoid, S., Dietrich, E., Drake, M. M., Dubchak, I., Edirisinghe, J. N., Fang, G., Faria, J. P., Frybarger, P. M., Gerlach, W., Gerstein, M., Gurtowski, J., Haun, H. L., He, F., Jain, R., Joachimiak, M. P., Keegan, K. P., Kondo, S., Kumar, V., Land, M. L., Mills, M., Novichkov, P., Oh, T., Olsen, G. J., Olson, B., Parrello, B., Pasternak, S., Pearson, E., Poon, S. S., Price, G., Ramakrishnan, S., Ranjan, P., Ronald, P. C., Schatz, M. C., Seaver, S. M. D., Shukla, M.,

Sutormin, R. A., Syed, M. H., Thomason, J., Tintle, N. L., Wang, D., Xia, F., Yoo, H., and Yoo, S. (2016). The doe systems biology knowledgebase (kbase). *bioRxiv*.

[10] Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J. M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., de Vos, W. M., Brunak, S., Dore, J., Antolin, M., Artiguenave, F., Blottiere, H. M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K. U., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Huber, W., van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Merieux, A., Melo Minardi, R., M'Rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S. D., and Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–80.

[11] Asnicar, F., Weingart, G., Tickle, T. L., Huttenhower, C., and Segata, N. (2015). Compact graphical representation of phylogenetic data and metadata with graphlan. *PeerJ*, 3:e1029.

[12] 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. (2008). The rast server: rapid annotations using subsystems technology. *BMC Genomics*, 9:75.

[13] Aziz, R. K., Devoid, S., Disz, T., Edwards, R. A., Henry, C. S., Olsen, G. J., Olson, R., Overbeek, R., Parrello, B., Pusch, G. D., Stevens, R. L., Vonstein, V., and Xia, F. (2012). Seed servers: high-performance access to the seed genomes, annotations, and metabolic models. *PLoS One*, 7(10):e48053.

[14] Baldwin, E. A., Bai, J., Plotto, A., Cameron, R., Luzio, G., Narciso, J., Manthey, J., Widmer, W., and Ford, B. L. (2012). Effect of extraction method on quality of orange juice: hand-squeezed, commercial-fresh squeezed and processed. *J Sci Food Agric*, 92(10):2029–42.

[15] Barghash, A. and Helms, V. (2013). Transferring functional annotations of membrane transporters on the basis of sequence similarity and sequence motifs. *BMC Bioinformatics*, 14:343.

[16] Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D., Lenski, R. E., and Kim, J. F. (2009). Genome evolution and adaptation in a long-term experiment with escherichia coli. *Nature*, pages 1243–1247.

[17] Bartell, J. A., Blazier, A. S., Yen, P., Thogersen, J. C., Jelsbak, L., Goldberg, J. B., and Papin, J. A. (2017). Reconstruction of the metabolic network of pseudomonas aeruginosa to interrogate virulence factor synthesis. *Nat Commun*, 8:14631.

[18] Bauer, E., Laczny, C. C., Magnusdottir, S., Wilmes, P., and Thiele, I. (2015). Phenotypic differentiation of gastrointestinal microbes is reflected in their encoded metabolic repertoires. *Microbiome*, 3:55.

[19] Bauer, E., Zimmermann, J., Baldini, F., Thiele, I., and Kaleta, C. (2017). Bacarena: Individual-based metabolic modeling of heterogeneous microbes in complex communities. *PLOS Computational Biology*, 13(5):1–22.

[20] Baumler, D. J., Peplinski, R. G., Reed, J. L., Glasner, J. D., and Perna, N. T. (2011). The evolution of metabolic networks of e. coli. *BMC Syst Biol*, 5:182.

[21] Begley, T. P., Kinsland, C., Mehl, R. A., Osterman, A., and Dorrestein, P. (2001). The biosynthesis of nicotinamide adenine dinucleotides in bacteria. *Vitam Horm*, 61:103–19.

[22] Belenky, P., Bogan, K. L., and Brenner, C. (2007). Nad+ metabolism in health and disease. *Trends Biochem Sci*, 32(1):12–9.

[23] Ben-Amor, K., Heilig, H., Smidt, H., Vaughan, E. E., Abee, T., and de Vos, W. M. (2005). Genetic diversity of viable, injured, and dead fecal bacteria assessed by fluorescence-activated cell sorting and 16s rrna gene analysis. *Appl Environ Microbiol*, 71(8):4679–89.

[24] Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2013). Genbank. *Nucleic Acids Res*, 41(Database issue):D36–42.

[25] Biggs, M. B. and Papin, J. A. (2016). Metabolic network-guided binning of metagenomic sequence fragments. *Bioinformatics*, 32(6):867.

[26] Biggs, M. B. and Papin, J. A. (2017). Managing uncertainty in metabolic network structure and improving predictions using ensemblefba. *PLoS Comput Biol*, 13(3):e1005413.

[27] Boele, J., Olivier, B. G., and Teusink, B. (2012). Fame, the flux analysis and modeling environment. *BMC Syst Biol*, 6:8.

[28] Bordbar, A., Lewis, N. E., Schellenberger, J., Palsson, B. Ø., and Jamshidi, N. (2010). Insight into human alveolar macrophage and m. tuberculosis interactions via metabolic reconstructions. *Molecular Systems Biology*, 6(1).

[29] Cabib, E. and Leloir, L. F. (1958). The biosynthesis of trehalose phosphate. *J Biol Chem*, 231(1):259–75.

[30] Carmody, R. N., Gerber, G. K., Luevano, J. M., J., Gatti, D. M., Somes, L., Svenson, K. L., and Turnbaugh, P. J. (2015). Diet dominates host genotype in shaping the murine gut microbiota. *Cell Host Microbe*, 17(1):72–84.

[31] Chassard, C., Gaillard-Martinie, B., and Bernalier-Donadille, A. (2005). Interaction between h2-producing and non-h2-producing cellulolytic bacteria from the human colon. *FEMS Microbiol Lett*, 242(2):339–44.

[32] Chiu, H. C., Levy, R., and Borenstein, E. (2014). Emergent biosynthetic capacity in simple microbial communities. *PLoS Comput Biol*, 10(7):e1003695.

[33] Chubiz, L. M., Granger, B. R., Segre, D., and Harcombe, W. R. (2015). Species inter-actions differ in their genetic robustness. *Frontiers in microbiology*, 6:271.

[34] Cialabrini, L., Ruggieri, S., Kazanov, M. D., Sorci, L., Mazzola, F., Orsomando, G., Osterman, A. L., and Raffaelli, N. (2013). Genomics-guided analysis of nad recycling yields functional elucidation of cog1058 as a new family of pyrophosphatases. *PLoS One*, 8(6):e65595.

[35] Cicmanec, J. F. and Lichstein, H. C. (1978). Uptake of extracellular biotin by es-cherichia coli biotin prototrophs. *Journal of Bacteriology*, 133(1):270–278.

[36] Claesson, M. J., Cusack, S., O'Sullivan, O., Greene-Diniz, R., de Weerd, H., Flannery, E., Marchesi, J. R., Falush, D., Dinan, T., Fitzgerald, G., Stanton, C., van Sinderen, D., O'Connor, M., Harnedy, N., O'Connor, K., Henry, C., O'Mahony, D., Fitzgerald, A. P., Shanahan, F., Twomey, C., Hill, C., Ross, R. P., and O'Toole, P. W. (2011). Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc Natl Acad Sci U S A*, 108 Suppl 1:4586–91.

[37] Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O/'Connor, E. M., Cusack, S., Harris, H. M. B., Coakley, M., Lakshminarayanan, B., O/'Sullivan, O., Fitzgerald, G. F., Deane, J., O/'Connor, M., Harnedy, N., O/'Connor, K., O/'Mahony, D., van Sinderen, D., Wallace, M., Brennan, L., Stanton, C., Marchesi, J. R., Fitzgerald, A. P., Shanahan, F., Hill, C., Ross, R. P., and O/'Toole, P. W. (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature*, 488(7410):178–184.

[38] Clarke, R., Grimley Evans, J., Schneede, J., Nexo, E., Bates, C., Fletcher, A., Prentice, A., Johnston, C., Ueland, P. M., Refsum, H., Sherliker, P., Birks, J., Whitlock, G., Breeze, E., and Scott, J. M. (2004). Vitamin b12 and folate deficiency in later life. *Age Ageing*, 33(1):34–41.

[39] Clemente, J. C., Ursell, L. K., Parfrey, L. W., and Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. *Cell*, 148(6):1258–70.

[40] Cole, H. S., Lopez, R., and Cooperman, J. M. (1976). Riboflavin deficiency in children with diabetes mellitus. *Acta Diabetol Lat*, 13(1-2):25–9.

[41] Consortium, T. H. M. P. (2012a). A framework for human microbiome research. *Nature*, 486(7402):215–21.

[42] Consortium, T. H. M. P. (2012b). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–14.

[43] Coordinators, N. R. (2016). Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 44(D1):D7–d19.

[44] Coyte, K. Z., Schluter, J., and Foster, K. R. (2015). The ecology of the microbiome: Networks, competition, and stability. *Science*, 350(6261):663–6.

[45] Crott, J. W., Liu, Z., Keyes, M. K., Choi, S. W., Jang, H., Moyer, M. P., and Mason, J. B. (2008). Moderate folate depletion modulates the expression of selected genes involved in cell cycle, intracellular signaling and folate uptake in human colonic epithelial cell lines. *J Nutr Biochem*, 19(5):328–35.

[46] Dakshinamurti, S. and Dakshinamurti, K. (2007). *Handbook of Vitamins*. Taylor & Francis Group, Boca Raton, FL, 4 edition.

[47] David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., Biddinger, S. B., Dutton, R. J., and Turnbaugh, P. J. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–63.

[48] Davila, A. M., Blachier, F., Gotteland, M., Andriamihaja, M., Benetti, P. H., Sanz, Y., and Tome, D. (2013). Re-print of "intestinal luminal nitrogen metabolism: role of the gut microbiota and consequences for the host". *Pharmacol Res*, 69(1):114–26.

[49] Davis, B. D. and Mingioli, E. S. (1950). Mutants of escherichia coli requiring methionine or vitamin b12. *J Bacteriol*, 60(1):17–28.

[50] De Colibus, L. and Mattevi, A. (2006). New frontiers in structural flavoenzymology. *Curr Opin Struct Biol*, 16(6):722–8.

[51] de Crécy-Lagard, V., El Yacoubi, B., de la Garza, R. D., Noiriel, A., and Hanson, A. D. (2007). Comparative genomics of bacterial and plant folate synthesis and salvage: predictions and validations. *BMC Genomics*, 8(1):245.

[52] de Crecy-Lagard, V., Phillips, G., Grochowski, L. L., El Yacoubi, B., Jenney, F., Adams, M. W., Murzin, A. G., and White, R. H. (2012). Comparative genomics guided discovery of two missing archaeal enzyme families involved in the biosynthesis of the pterin moiety of tetrahydromethanopterin and tetrahydrofolate. *ACS Chem Biol*, 7(11):1807–16.

[53] De Ingeniis, J., Kazanov, M. D., Shatalin, K., Gelfand, M. S., Osterman, A. L., and Sorci, L. (2012). Glutamine versus ammonia utilization in the nad synthetase family. *PLoS One*, 7(6):e39115.

[54] Degnan, P. H., Taga, M. E., and Goodman, A. L. (2014). Vitamin b12 as a modulator of gut microbial ecology. *Cell Metab*, 20(5):769–78.

[55] Dempsey, W. B. (1971). Role of vitamin b6 biosynthetic rate in the study of vitamin b6 synthesis in escherichia coli. *Journal of Bacteriology*, 108(3):1001–1007.

[56] Dicksved, J., Halfvarson, J., Rosenquist, M., Jarnerot, G., Tysk, C., Apajalahti, J., Engstrand, L., and Jansson, J. K. (2008). Molecular analysis of the gut microbiota of identical twins with crohn's disease. *ISME J*, 2(7):716–727.

[57] Disz, T., Akhter, S., Cuevas, D., Olson, R., Overbeek, R., Vonstein, V., Stevens, R., and Edwards, R. A. (2010). Accessing the seed genome databases via web services api: tools for programmers. *BMC Bioinformatics*, 11:319.

[58] Dore, J. and Blottiere, H. (2015). The influence of diet on the gut microbiota and its consequences for health. *Curr Opin Biotechnol*, 32:195–9.

[59] Duncan, S. H., Belenguer, A., Holtrop, G., Johnstone, A. M., Flint, H. J., and Lobley, G. E. (2007). Reduced dietary intake of carbohydrates by obese subjects results in decreased concentrations of butyrate and butyrate-producing bacteria in feces. *Appl Environ Microbiol*, 73(4):1073–8.

[60] Duncan, S. H., Scott, K. P., Ramsay, A. G., Harmsen, H. J., Welling, G. W., Stewart, C. S., and Flint, H. J. (2003). Effects of alternative dietary substrates on competition between human colonic bacteria in an anaerobic fermentor system. *Appl Environ Microbiol*, 69(2):1136–42.

[61] Dunne, J. L., Triplett, E. W., Gevers, D., Xavier, R., Insel, R., Danska, J., and Atkinson, M. A. (2014). The intestinal microbiome in type 1 diabetes. *Clinical & Experimental Immunology*, 177(1):30–37.

[62] Dworkin, M. and Falkow, S. (2006). *The prokaryotes: a handbook on the biology of bacteria*. Springer, New York.

[63] El-Semman, I. E., Karlsson, F. H., Shoaie, S., Nookaew, I., Soliman, T. H., and Nielsen, J. (2014). Genome-scale metabolic reconstructions of bifidobacterium adolescentis l2-32 and faecalibacterium prausnitzii a2-165 and their interaction. *BMC Systems Biology*, 8(1):41.

[64] Espey, M. G. (2013). Role of oxygen gradients in shaping redox relationships between the human intestine and its microbiota. *Free Radic Biol Med*, 55:130–40.

[65] Fallingborg, J. (1999). Intraluminal ph of the human gastrointestinal tract. *Dan Med Bull*, 46(3):183–96.

[66] Faust, K. and Raes, J. (2012). Microbial interactions: from networks to models. *Nat Rev Microbiol*, 10(8):538–50.

[67] Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol*, 8(7):e1002606.

[68] 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. (2007). A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Molecular systems biology*, 3:121.

[69] Feist, A. M. and Palsson, B. O. (2010). The biomass objective function. *Curr Opin Microbiol*, 13(3):344–9.

[70] Feng, X., Xu, Y., Chen, Y., and Tang, Y. J. (2012). Microbesflux: a web platform for drafting metabolic models from the kegg database. *BMC Syst Biol*, 6:94.

[71] 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. (2013). Genome-scale metabolic model for lactococcus lactis mg1363 and its application to the analysis of flavor formation. *Applied microbiology and biotechnology*, 97:8729–39.

[72] Fleming, R., Vlassis, N., Thiele, I., and Saunders, M. (2016). Conditions for duality between fluxes and concentrations in biochemical networks. *Journal of Theoretical Biology*, 409:1–10.

[73] Fleming, R. M. T., Thiele, I., and Nasheuer, H. P. (2009). Quantitative assignment of reaction directionality in constraint-based models of metabolism: Application to escherichia coli. *Biophysical Chemistry*, 145(2–3):47–56.

[74] Flint, H. J. (2012). The impact of nutrition on the human microbiome. *Nutr Rev*, 70 Suppl 1:S10–3.

[75] Flint, H. J., Scott, K. P., Duncan, S. H., Louis, P., and Forano, E. (2012). Microbial degradation of complex carbohydrates in the gut. *Gut Microbes*, 3(4):289–306.

[76] Freilich, S., Zarecki, R., Eilam, O., Segal, E. S., Henry, C. S., Kupiec, M., Gophna, U., Sharan, R., and Ruppin, E. (2011). Competitive and cooperative metabolic interactions in bacterial communities. *Nat Commun*, 2:589.

[77] Galeazzi, L., Bocci, P., Amici, A., Brunetti, L., Ruggieri, S., Romine, M., Reed, S., Osterman, A. L., Rodionov, D. A., Sorci, L., and Raffaelli, N. (2011). Identification of nicotinamide mononucleotide deamidase of the bacterial pyridine nucleotide cycle reveals a novel broadly conserved amidohydrolase family. *J Biol Chem*, 286(46):40365–75.

[78] Gazzaniga, F., Stebbins, R., Chang, S. Z., McPeek, M. A., and Brenner, C. (2009). Microbial nad metabolism: lessons from comparative genomics. *Microbiol Mol Biol Rev*, 73(3):529–41, Table of Contents.

[79] Gelfand, M. S. and Rodionov, D. A. (2008). Comparative genomics and functional annotation of bacterial transporters. *Physics of Life Reviews*, 5(1):22–49.

[80] Gerdes, S., Lerma-Ortiz, C., Frelin, O., Seaver, S. M., Henry, C. S., de Crecy-Lagard, V., and Hanson, A. D. (2012). Plant b vitamin pathways and their compartmentation: a guide for the perplexed. *J Exp Bot*, 63(15):5379–95.

[81] Gerdes, S. Y., Kurnasov, O. V., Shatalin, K., Polanuyer, B., Sloutsky, R., Vonstein, V., Overbeek, R., and Osterman, A. L. (2006). Comparative genomics of nad biosynthesis in cyanobacteria. *J Bacteriol*, 188(8):3012–23.

[82] Gibbons, R. J. and Kapsimalis, B. (1967). Estimates of the overall rate of growth of the intestinal microflora of hamsters, guinea pigs, and mice. *J Bacteriol*, 93(1):510–2.

[83] Giongo, A., Gano, K. A., Crabb, D. B., Mukherjee, N., Novelo, L. L., Casella, G., Drew, J. C., Ilonen, J., Knip, M., Hyoty, H., Veijola, R., Simell, T., Simell, O., Neu, J., Wasserfall, C. H., Schatz, D., Atkinson, M. A., and Triplett, E. W. (2011). Toward defining the autoimmune microbiome for type 1 diabetes. *ISME J*, 5(1):82–91.

[84] Gossmann, T. I., Ziegler, M., Puntervoll, P., de Figueiredo, L. F., Schuster, S., and Heiland, I. (2012). Nad(+) biosynthesis and salvage–a phylogenetic perspective. *Febs j*, 279(18):3355–63.

[85] Green, M. L. and Karp, P. D. (2005). Genome annotation errors in pathway databases due to semantic ambiguity in partial ec numbers. *Nucleic Acids Res*, 33(13):4035–9.

[86] Greenblum, S., Chiu, H. C., Levy, R., Carr, R., and Borenstein, E. (2013). Towards a predictive systems-level model of the human microbiome: progress, challenges, and opportunities. *Curr Opin Biotechnol*, 24(4):810–20.

[87] Greenblum, S., Turnbaugh, P. J., and Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A*, 109(2):594–9.

[88] Grosskopf, T. and Soyer, O. S. (2014). Synthetic microbial communities. *Curr Opin Microbiol*, 18:72–7.

[89] Gudmundsson, S. and Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC Bioinformatics*, 11:489.

[90] Hamilton, J. J. and Reed, J. L. (2014). Software platforms to facilitate reconstructing genome-scale metabolic networks. *Environmental Microbiology*, 16(1):49–59.

[91] Haraldsdottir, H. S., Cousins, B., Thiele, I., Fleming, R. M., and Vempala, S. (2017). Chrr: Coordinate hit-and-run with rounding for uniform sampling of constraint-based models. *Bioinformatics*.

[92] Harcombe, W. R., Riehl, W. J., Dukovski, I., Granger, B. R., Betts, A., Lang, A. H., Bonilla, G., Kar, A., Leiby, N., Mehta, P., Marx, C. J., and Segrè, D. (2014). Metabolic resource allocation in individual microbes determines ecosystem interactions and spatial dynamics. *Cell Reports*, 7(4):1104–1115.

[93] Hasona, A., Kim, Y., Healy, F. G., Ingram, L. O., and Shanmugam, K. T. (2004). Pyruvate formate lyase and acetate kinase are essential for anaerobic growth of escherichia coli on xylose. *J Bacteriol*, 186(22):7593–600.

[94] Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., and Steinbeck, C. (2013). The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res*, 41(Database issue):D456–63.

[95] Heinken, A., Khan, M. T., Paglia, G., Rodionov, D. A., Harmsen, H. J., and Thiele, I. (2014). Functional metabolic map of faecalibacterium prausnitzii, a beneficial human gut microbe. *J Bacteriol*, 196(18):3289–302.

[96] Heinken, A., Sahoo, S., Fleming, R. M., and Thiele, I. (2013). Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. *Gut Microbes*, 4(1):28–40.

[97] Heinken, A. and Thiele, I. (2015a). Anoxic conditions promote species-specific mutualism between gut microbes in silico. *Appl Environ Microbiol*, 81(12):4049–61.

[98] Heinken, A. and Thiele, I. (2015b). Systematic prediction of health-relevant human-microbial co-metabolism through a computational framework. *Gut Microbes*, 6(2):120–30.

[99] Heinken, A. and Thiele, I. (2015c). Systems biology of host-microbe metabolomics. *Wiley Interdiscip Rev Syst Biol Med*, 7(4):195–219.

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

[101] Heirendt, L., Thiele, I., and Fleming, R. M. T. (2017). Distributedfba.jl: high-level, high-performance flux balance analysis in julia. *Bioinformatics*, 33(9):1421.

[102] Henry, C. S., DeJongh, M., Best, A. A., Frybarger, P. M., Linsay, B., and Stevens, R. L. (2010). High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotech*, 28(9):977–982.

[103] Herbert, V. (1963). A palatable diet for producing experimental folate deficiency in man. *Am J Clin Nutr*, 12:17–20.

[104] Herring, C. D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M. K., Joyce, A. R., Albert, T. J., Blattner, F. R., van den Boom, D., Cantor, C. R., and Palsson, B. . (2006). Comparative genome sequencing of escherichia coli allows observation of bacterial evolution on a laboratory timescale. *Nature Genetics*, 38:1406–1412.

[105] Hill, M. J. (1997). Intestinal flora and endogenous vitamin synthesis. *Eur J Cancer Prev*, 6(2):S43–5.

[106] Hill-Burns, E. M., Debelius, J. W., Morton, J. T., Wissemann, W. T., Lewis, M. R., Wallen, Z. D., Peddada, S. D., Factor, S. A., Molho, E., Zabetian, C. P., Knight, R., and Payami, H. (2017). Parkinson's disease and parkinson's disease medications have distinct signatures of the gut microbiome. *Movement Disorders*.

[107] Houtkooper, R. H., Canto, C., Wanders, R. J., and Auwerx, J. (2010). The secret life of nad+: an old metabolite controlling new metabolic signaling pathways. *Endocr Rev*, 31(2):194–223.

[108] Huynen, M. A., Dandekar, T., and Bork, P. (1999). Variation and evolution of the citric-acid cycle: a genomic perspective. *Trends Microbiol*, 7(7):281–91.

[109] Jackowski, S. and Alix, J. H. (1990). Cloning, sequence, and expression of the pantothenate permease (panf) gene of escherichia coli. *Journal of Bacteriology*, 172(7):3842–3848.

[110] Jurgenson, C. T., Begley, T. P., and Ealick, S. E. (2009). The structural and biochemical foundations of thiamin biosynthesis. *Annu Rev Biochem*, 78:569–603.

[111] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., and Hirakawa, M. (2006). From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res*, 34(Database issue):D354–7.

[112] Karasawa, T., Ikoma, S., Yamakawa, K., and Nakamura, S. (1995). A defined growth medium for clostridium difficile. *Microbiology*, 141 (Pt 2):371–5.

[113] Karp, P. D., Latendresse, M., Paley, S. M., Krummenacker, M., Ong, Q. D., Billington, R., Kothari, A., Weaver, D., Lee, T., Subhraveti, P., Spaulding, A., Fulcher, C., Keseler, I. M., and Caspi, R. (2016). Pathway tools version 19.0 update: software for pathway/genome informatics and systems biology. *Brief Bioinform*, 17(5):877–90.

[114] Kemppainen, K. M., Ardissone, A. N., Davis-Richardson, A. G., Fagen, J. R., Gano, K. A., León-Novelo, L. G., Vehik, K., Casella, G., Simell, O., Ziegler, A. G., Rewers, M. J., Lernmark, Å., Hagopian, W., She, J.-X., Krischer, J. P., Akolkar, B., Schatz, D. A., Atkinson, M. A., Triplett, E. W., and TEDDY Study Group, . (2015). Early childhood gut microbiomes show strong geographic differences among subjects at high risk for type 1 diabetes. *Diabetes Care*, 38(2):329–332.

[115] Khan, M. T., Browne, W. R., van Dijl, J. M., and Harmsen, H. J. (2012). How can faecalibacterium prausnitzii employ riboflavin for extracellular electron transfer? *Antioxid Redox Signal*, 17(10):1433–40.

[116] Khoroshkin, M. S., Leyn, S. A., Van Sinderen, D., and Rodionov, D. A. (2016). Transcriptional regulation of carbohydrate utilization pathways in the bifidobacterium genus. *Front Microbiol*, 7:120.

[117] Kjer-Nielsen, L., Patel, O., Corbett, A. J., Le Nours, J., Meehan, B., Liu, L., Bhati, M., Chen, Z., Kostenko, L., Reantragoon, R., Williamson, N. A., Purcell, A. W., Dudek, N. L., McConville, M. J., O'Hair, R. A., Khairallah, G. N., Godfrey, D. I., Fairlie, D. P., Rossjohn, J., and McCluskey, J. (2012). Mr1 presents microbial vitamin b metabolites to mait cells. *Nature*, 491(7426):717–23.

[118] Klamt, S., Saez-Rodriguez, J., and Gilles, E. D. (2007). Structural and functional analysis of cellular networks with cellnetanalyzer. *BMC Syst Biol*, 1:2.

[119] Klitgord, N. and Segre, D. (2010). Environments that induce synthetic microbial ecosystems. *PLoS Comput Biol*, 6(11):e1001002.

[120] Koonin, E. V., Aravind, L., and Kondrashov, A. S. (2000). The impact of comparative genomics on our understanding of evolution. *Cell*, 101(6):573 – 576.

[121] Koonin, E. V. and Galperin, M. Y. (2003). *Sequence - Evolution - Function: Computational Approaches in Comparative Genomics*. Kluwer Academic Kluwer Academic., Boston.

[122] Krieg, N. R. (2010). *Bergey's manual of systematic bacteriology*, volume 4. Springer, New York, 2 edition.

[123] Kubitschek, H. E. and Friske, J. A. (1986). Determination of bacterial cell volume with the coulter counter. *J Bacteriol*, 168(3):1466–7.

[124] Kurnasov, O., Goral, V., Colabroy, K., Gerdes, S., Anantha, S., Osterman, A., and Begley, T. P. (2003). Nad biosynthesis: identification of the tryptophan to quinolinate pathway in bacteria. *Chem Biol*, 10(12):1195–204.

[125] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat Methods*, 9(4):357–9.

[126] LeBlanc, J. G., Milani, C., de Giori, G. S., Sesma, F., van Sinderen, D., and Ventura, M. (2013). Bacteria as vitamin suppliers to their host: a gut microbiota perspective. *Curr Opin Biotechnol*, 24(2):160–8.

[127] Letunic, I. and Bork, P. (2007). Interactive tree of life (itol): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1):127–8.

[128] Letunic, I. and Bork, P. (2011). Interactive tree of life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res*, 39(Web Server issue):W475–8.

[129] Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., and Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31):11070–11075.

[130] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009a). The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–9.

[131] Li, H., Myeroff, L., Smiraglia, D., Romero, M. F., Pretlow, T. P., Kasturi, L., Lutterbaugh, J., Rerko, R. M., Casey, G., Issa, J. P., Willis, J., Willson, J. K., Plass, C., and Markowitz, S. D. (2003). Slc5a8, a sodium transporter, is a tumor suppressor gene silenced by methylation in human colon aberrant crypt foci and cancers. *Proc Natl Acad Sci U S A*, 100(14):8412–7.

[132] Li, W., Raoult, D., and Fournier, P. E. (2009b). Bacterial strain typing in the genomic era. *FEMS Microbiol Rev*, 33(5):892–916.

[133] Liao, Y. C., Huang, T. W., Chen, F. C., Charusanti, P., Hong, J. S., Chang, H. Y., Tsai, S. F., Palsson, B. O., and Hsiung, C. A. (2011). An experimentally validated genome-scale metabolic reconstruction of klebsiella pneumoniae mgh 78578, iyl1228. *J Bacteriol*, 193(7):1710–7.

[134] Lin, S. and Cronan, J. E. (2011). Closing in on complete pathways of biotin biosynthesis. *Mol Biosyst*, 7(6):1811–21.

[135] Loferer-Krößbacher, M., Klima, J., and Psenner, R. (1998). Determination of bacterial cell dry mass by transmission electron microscopy and densitometric image analysis. *Applied and Environmental Microbiology*, 64(2):688–694.

[136] Louis, P. and Flint, H. J. (2009). Diversity, metabolism and microbial ecology of butyrate-producing bacteria from the human large intestine. *FEMS Microbiol Lett*, 294(1):1–8.

[137] Luo, H., Lin, Y., Gao, F., Zhang, C. T., and Zhang, R. (2014). Deg 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. *Nucleic Acids Res*, 42(Database issue):D574–80.

[138] Macfarlane, G. T. and Macfarlane, S. (2012). Bacteria, colonic fermentation, and gastrointestinal health. *J AOAC Int*, 95(1):50–60.

[139] Magnusdottir, S., Heinken, A., Kutt, L., Ravcheev, D. A., Bauer, E., Noronha, A., Greenhalgh, K., Jager, C., Baginska, J., Wilmes, P., Fleming, R. M., and Thiele, I. (2017). Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. *Nat Biotechnol*, 35(1):81–89.

[140] Magnusdottir, S., Ravcheev, D., de Crecy-Lagard, V., and Thiele, I. (2015). Systematic genome assessment of b-vitamin biosynthesis suggests co-operation among gut microbes. *Front Genet*, 6:148.

[141] Manichanh, C., Borruel, N., Casellas, F., and Guarner, F. (2012). The gut microbiota in ibd. *Nat Rev Gastroenterol Hepatol*, 9(10):599–608.

[142] Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., Geer, R. C., He, J., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., and Bryant, S. H. (2015). Cdd: Ncbi's conserved domain database. *Nucleic Acids Res*, 43(Database issue):D222–6.

[143] Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Lu, S., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., and Bryant, S. H. (2013). Cdd: conserved domains and protein three-dimensional structure. *Nucleic Acids Res*, 41(Database issue):D348–52.

[144] Mardinoglu, A., Shoaie, S., Bergentall, M., Ghaffari, P., Zhang, C., Larsson, E., Bäckhed, F., and Nielsen, J. (2015). The gut microbiota modulates host amino acid and glutathione metabolism in mice. *Molecular Systems Biology*, 11(10).

[145] Marino, E., Richards, J. L., McLeod, K. H., Stanley, D., Yap, Y. A., Knight, J., McKenzie, C., Kranich, J., Oliveira, A. C., Rossello, F. J., Krishnamurthy, B., Nefzger, C. M., Macia, L., Thorburn, A., Baxter, A. G., Morahan, G., Wong, L. H., Polo, J. M., Moore, R. J., Lockett, T. J., Clarke, J. M., Topping, D. L., Harrison, L. C., and Mackay, C. R. (2017). Gut microbial metabolites limit the frequency of autoimmune t cells and protect against type 1 diabetes. *Nat Immunol*, 18(5):552–562.

[146] Mattevi, A., Tedeschi, G., Bacchella, L., Coda, A., Negri, A., and Ronchi, S. (1999). Structure of l-aspartate oxidase: implications for the succinate dehydrogenase/fumarate reductase oxidoreductase family. *Structure*, 7(7):745–56.

[147] Mazumdar, V., Amar, S., and Segre, D. (2013). Metabolic proximity in the order of colonization of a microbial community. *PLoS One*, 8(10):e77617.

[148] McCloskey, D., Palsson, B. O., and Feist, A. M. (2013). Basic and applied uses of genome-scale metabolic network reconstructions of escherichia coli. *Mol Syst Biol*, 9:661.

[149] McLaren, J., Ngo, D. T. C., and Olivera, B. M. (1973). Pyridine nucleotide metabolism in escherichia coli : Iii. biosynthesis from alternative precursors in vivo. *Journal of Biological Chemistry*, 248(14):5144–5149.

[150] McNulty, N. P., Wu, M., Erickson, A. R., Pan, C., Erickson, B. K., Martens, E. C., Pudlo, N. A., Muegge, B. D., Henrissat, B., Hettich, R. L., and Gordon, J. I. (2013). Effects of diet on resource utilization by a model human gut microbiota containing bacteroides cellulosilyticus wh2, a symbiont with an extensive glycobiome. *PLoS Biol*, 11(8):e1001637.

[151] Metges, C. C. (2000). Contribution of microbial amino acids to amino acid homeostasis of the host. *J Nutr*, 130(7):1857s–64s.

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

[153] Molenaar, D., van Berlo, R., de Ridder, D., and Teusink, B. (2009). Shifts in growth strategies reflect tradeoffs in cellular economics. *Molecular systems biology*, 5:323.

[154] Moreau, N. M., Goupry, S. M., Antignac, J. P., Monteau, F. J., Le Bizec, B. J., Champ, M. M., Martin, L. J., and Dumon, H. J. (2003). Simultaneous measurement of plasma concentrations and 13c-enrichment of short-chain fatty acids, lactic acid and ketone bodies by gas chromatography coupled to mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci*, 784(2):395–403.

[155] Noor, E., Haraldsdottir, H. S., Milo, R., and Fleming, R. M. (2013). Consistent estimation of gibbs energy using component contributions. *PLoS Comput Biol*, 9(7):e1003098.

[156] Nowicka, B. and Kruk, J. (2010). Occurrence, biosynthesis and function of isoprenoid quinones. *Biochim Biophys Acta*, 1797(9):1587–605.

[157] Oberhardt, M. A., Palsson, B. O., and Papin, J. A. (2009). Applications of genome-scale metabolic reconstructions. *Mol Syst Biol*, 5:320.

[158] Oberhardt, M. A., Puchalka, J., Fryer, K. E., Martins dos Santos, V. A., and Papin, J. A. (2008). Genome-scale metabolic network analysis of the opportunistic pathogen pseudomonas aeruginosa pao1. *J Bacteriol*, 190(8):2790–803.

[159] Orth, J. D., Thiele, I., and Palsson, B. O. (2010). What is flux balance analysis? *Nat Biotechnol*, 28(3):245–8.

[160] Osterman, A. and Overbeek, R. (2003). Missing genes in metabolic pathways: a comparative genomics approach. *Current Opinion in Chemical Biology*, 7(2):238 – 251.

[161] Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., de Crecy-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., Ruckert, C., Steiner, J., Stevens, R., Thiele, I., Vassieva, O., Ye, Y., Zagnitko, O., and Vonstein, V. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*, 33(17):5691–702.

[162] Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., Edwards, R. A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A. R., Xia, F., and Stevens, R. (2014). The seed and the rapid annotation of microbial genomes using subsystems technology (rast). *Nucleic Acids Res*, 42(Database issue):D206–14.

[163] Palsson, B. (2006). *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, Cambridge, New York.

[164] Palsson, B. (2015). *Systems Biology: Constraint-based Reconstruction and Analysis*. Cambridge University Press, Cambridge, New York.

[165] Pastink, M. I., Teusink, B., Hols, P., Visser, S., de Vos, W. M., and Hugenholtz, J. (2009). Genome-scale model of streptococcus thermophilus lmg18311 for metabolic comparison of lactic acid bacteria. *Appl Environ Microbiol*, 75(11):3627–33.

[166] Pál, C., Papp, B., and Lercher, M. J. (2005). Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genetics*, pages 1372–1375.

[167] Prasad, P. D., Wang, H., Huang, W., Fei, Y. J., Leibach, F. H., Devoe, L. D., and Ganapathy, V. (1999). Molecular and functional characterization of the intestinal na+-dependent multivitamin transporter. *Arch Biochem Biophys*, 366(1):95–106.

[168] Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J. M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Dore, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Bork, P., Ehrlich, S. D., and Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65.

[169] Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., Peng, Y., Zhang, D., Jie, Z., Wu, W., Qin, Y., Xue, W., Li, J., Han, L., Lu, D., Wu, P., Dai, Y., Sun, X., Li, Z., Tang, A., Zhong, S., Li, X., Chen, W., Xu, R., Wang, M., Feng, Q., Gong, M., Yu, J., Zhang, Y., Zhang, M., Hansen, T., Sanchez, G., Raes, J., Falony, G., Okuda, S., Almeida, M., LeChatelier, E., Renault, P., Pons, N., Batto, J. M., Zhang, Z., Chen, H., Yang, R., Zheng, W., Li, S., Yang, H., Wang, J., Ehrlich, S. D., Nielsen, R., Pedersen, O., Kristiansen, K., and Wang, J. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature*, 490(7418):55–60.

[170] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[171] Raghunathan, A., Reed, J., Shin, S., Palsson, B., and Daefler, S. (2009). Constraint-based analysis of metabolic capacity of salmonella typhimurium during host-pathogen interaction. *BMC Systems Biology*, 3(1):38.

[172] Rajilic-Stojanovic, M. and de Vos, W. M. (2014). The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiol Rev*, 38(5):996–1047.

[173] Ravcheev, D. A., Godzik, A., Osterman, A. L., and Rodionov, D. A. (2013). Polysaccharides utilization in human gut bacterium bacteroides thetaiotaomicron: comparative genomics reconstruction of metabolic and regulatory networks. *BMC Genomics*, 14(1):873.

[174] Ravcheev, D. A. and Thiele, I. (2014). Systematic genomic analysis reveals the complementary aerobic and anaerobic respiration capacities of the human gut microbiota. *Front Microbiol*, 5:674.

[175] Ravcheev, D. A. and Thiele, I. (2016). Genomic analysis of the human gut microbiome suggests novel enzymes involved in quinone biosynthesis. *Frontiers in Microbiology*, 7.

[176] Ravcheev, D. A. and Thiele, I. (2017). Comparative genomic analysis of the human gut microbiome reveals a broad distribution of metabolic pathways for the degradation of host-synthetized mucin glycans.

[177] Reader, J. S., Metzgar, D., Schimmel, P., and de Crécy-Lagard, V. (2004). Identification of four genes necessary for biosynthesis of the modified nucleoside queuosine. *Journal of Biological Chemistry*, 279(8):6280–6285.

[178] 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. (2006). Systems approach to refining genome annotation. *Proceedings of the National Academy of Sciences*, 103(46):17480–17484.

[179] Reichardt, N., Duncan, S. H., Young, P., Belenguer, A., McWilliam Leitch, C., Scott, K. P., Flint, H. J., and Louis, P. (2014). Phylogenetic distribution of three pathways for propionate production within the human gut microbiota. *Isme j*, 8(6):1323–35.

[180] Rey, F. E., Faith, J. J., Bain, J., Muehlbauer, M. J., Stevens, R. D., Newgard, C. B., and Gordon, J. I. (2010). Dissecting the in vivo metabolic potential of two human gut acetogens. *J Biol Chem*, 285(29):22082–90.

[181] Rey, F. E., Gonzalez, M. D., Cheng, J., Wu, M., Ahern, P. P., and Gordon, J. I. (2013). Metabolic niche of a prominent sulfate-reducing human gut bacterium. *Proc Natl Acad Sci U S A*, 110(33):13582–7.

[182] Rigottier-Gois, L. (2013). Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis. *Isme j*, 7(7):1256–61.

[183] Rodionov, D. A., Mironov, A. A., and Gelfand, M. S. (2002). Conservation of the biotin regulon and the bira regulatory signal in eubacteria and archaea. *Genome Res*, 12(10):1507–16.

[184] Rodionova, I. A., Li, X., Plymale, A. E., Motamedchaboki, K., Konopka, A. E., Romine, M. F., Fredrickson, J. K., Osterman, A. L., and Rodionov, D. A. (2014). Genomic distribution of b-vitamin auxotrophy and uptake transporters in environmental bacteria from the chloroflexi phylum. *Environ Microbiol Rep*.

[185] Romine, M. F., Rodionov, D. A., Maezato, Y., Osterman, A. L., and Nelson, W. C. (2017). Underlying mechanisms for syntrophic metabolism of essential enzyme cofactors in microbial communities. *Isme j*.

[186] Rossi, M., Amaretti, A., and Raimondi, S. (2011). Folate production by probiotic bacteria. *Nutrients*, 3(1):118–34.

[187] Roth, K. S. (1981). Biotin in clinical medicine–a review. *Am J Clin Nutr*, 34(9):1967–74.

[188] Rucker, R., Zempleni, J., Suttie, J., and McCormick, D. (2007). *Handbook of Vitamins, Fourth Edition*. CRC Press.

[189] Sahoo, S., Haraldsdóttir, H. S., Fleming, R. M. T., and Thiele, I. (2015). Modeling the effects of commonly used drugs on human metabolism. *FEBS Journal*, 282(2):297–317.

[190] Sahoo, S. and Thiele, I. (2013). Predicting the impact of diet and enzymopathies on human small intestinal epithelial cells. *Hum Mol Genet*, 22(13):2705–22.

[191] Said, H. M. (2004). Recent advances in carrier-mediated intestinal absorption of water-soluble vitamins. *Annu Rev Physiol*, 66:419–46.

[192] Said, H. M. (2011). Intestinal absorption of water-soluble vitamins in health and disease. *Biochemical Journal*, 437(3):357–372.

[193] Samuel, B. S., Hansen, E. E., Manchester, J. K., Coutinho, P. M., Henrissat, B., Fulton, R., Latreille, P., Kim, K., Wilson, R. K., and Gordon, J. I. (2007). Genomic and metabolic adaptations of methanobrevibacter smithii to the human gut. *Proc Natl Acad Sci U S A*, 104(25):10643–8.

[194] Santos, F., Vera, J. L., van der Heijden, R., Valdez, G., de Vos, W. M., Sesma, F., and Hugenholtz, J. (2008). The complete coenzyme b12 biosynthesis gene cluster of lactobacillus reuteri crl1098. *Microbiology*, 154(Pt 1):81–93.

[195] Santos, S. and Rocha, I. (2016). Estimation of biomass composition from genomic and transcriptomic information. *J Integr Bioinform*, 13(2):285.

[196] Satish Kumar, V., Dasika, M. S., and Maranas, C. D. (2007). Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 8(1):212.

[197] Saulnier, D. M., Santos, F., Roos, S., Mistretta, T.-A., Spinler, J. K., Molenaar, D., Teusink, B., and Versalovic, J. (2011). Exploring metabolic pathway reconstruction and genome-wide expression profiling in lactobacillus reuteri to define functional probiotic features. *PloS one*, 6:e18783.

[198] Savage, D. C. (1977). Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol*, 31:107–33.

[199] Savijoki, K., Suokko, A., Palva, A., and Varmanen, P. (2006). New convenient defined media for [(35)s]methionine labelling and proteomic analyses of probiotic lactobacilli. *Lett Appl Microbiol*, 42(3):202–9.

[200] Schellenberger, J. and Palsson, B. O. (2009). Use of randomized sampling for analysis of metabolic networks. *J Biol Chem*, 284(9):5457–61.

[201] Schellenberger, J., Que, R., Fleming, R. M., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R., and Palsson, B. O. (2011). Quantitative prediction of cellular metabolism with constraint-based models: the cobra toolbox v2.0. *Nat Protoc*, 6(9):1290–307.

[202] Schmidt, B. J., Ebrahim, A., Metz, T. O., Adkins, J. N., Palsson, B. O., and Hyduke, D. R. (2013). Gim3e: condition-specific models of cellular metabolism developed from metabolomics and expression data. *Bioinformatics*, 29(22):2900–8.

[203] Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 9:811–814.

[204] Segers, M. E. and Lebeer, S. (2014). Towards a better understanding of lactobacillus rhamnosus gg–host interactions. *Microb Cell Fact*, 13 Suppl 1:S7.

[205] Shafquat, A., Joice, R., Simmons, S. L., and Huttenhower, C. (2014). Functional and phylogenetic assembly of microbial communities in the human microbiome. *Trends Microbiol*, 22(5):261–6.

[206] Shane, B. and Stokstad, E. L. (1976). Transport and utilization of methyltetrahydrofolates by lactobacillus casei. *Journal of Biological Chemistry*, 251(11):3405–10.

[207] Shearer, M. J. and Newman, P. (2014). Recent trends in the metabolism and cell biology of vitamin k with special reference to vitamin k cycling and mk-4 biosynthesis. *J Lipid Res*, 55(3):345–62.

[208] Shoaie, S., Ghaffari, P., Kovatcheva-Datchary, P., Mardinoglu, A., Sen, P., Pujos-Guillot, E., de Wouters, T., Juste, C., Rizkalla, S., Chilloux, J., Hoyles, L., Nicholson, J. K., Dore, J., Dumas, M. E., Clement, K., Backhed, F., and Nielsen, J. (2015). Quantifying diet-induced metabolic changes of the human gut microbiome. *Cell Metab*, 22(2):320–31.

[209] Shoaie, S., Karlsson, F., Mardinoglu, A., Nookaew, I., Bordel, S., and Nielsen, J. (2013). Understanding the interactions between bacteria in the human gut through metabolic modeling. *Scientific reports*, 3:2532.

[210] Skovran, E. and Downs, D. M. (2000). Metabolic defects caused by mutations in the isc gene cluster in salmonella enterica serovar typhimurium: implications for thiamine synthesis. *J Bacteriol*, 182(14):3896–903.

[211] Sánchez, C. E. G. and Sáez, R. G. T. (2014). Comparison and analysis of objective functions in flux balance analysis. *Biotechnology Progress*, 30(5):985–991.

[212] Sompolinsky, D. and Neujahr, H. Y. (1971). Studies of thiamine uptake in growing cultures and in cell fragments of lactobacillus fermenti. *Acta chem. scand*, 25:3054–3066.

[213] Sridharan, G. V., Choi, K., Klemashevich, C., Wu, C., Prabakaran, D., Pan, L. B., Steinmeyer, S., Mueller, C., Yousofshahi, M., Alaniz, R. C., Lee, K., and Jayaraman, A. (2014). Prediction and quantification of bioactive microbiota metabolites in the mouse gut. *Nat Commun*, 5:5492.

[214] Standing Committee on the Scientific Evaluation of Dietary Reference Intakes and its Panel on Folate, other B Vitamins, and Choline (1998). Dietary reference intakes for thiamin, riboflavin, niacin, vitamin b6, folate, vitamin b12, pantothenic acid, biotin, and choline. Report, National Academies Press (US).

[215] Stephen, A. M. and Cummings, J. H. (1980). The microbial contribution to human faecal mass. *J Med Microbiol*, 13(1):45–56.

[216] Stolyar, S., Van Dien, S., Hillesland, K. L., Pinel, N., Lie, T. J., Leigh, J. A., and Stahl, D. A. (2007). Metabolic modeling of a mutualistic microbial community. *Molecular Systems Biology*, 3(1).

[217] Suez, J., Korem, T., Zeevi, D., Zilberman-Schapira, G., Thaiss, C. A., Maza, O., Israeli, D., Zmora, N., Gilad, S., Weinberger, A., Kuperman, Y., Harmelin, A., Kolodkin-Gal, I., Shapiro, H., Halpern, Z., Segal, E., and Elinav, E. (2014). Artificial sweeteners induce glucose intolerance by altering the gut microbiota. *Nature*, 514(7521):181–6.

[218] Swainston, N., Smallbone, K., Mendes, P., Kell, D., and Paton, N. (2011). The subliminal toolbox: automating steps in the reconstruction of metabolic networks. *J Integr Bioinform*, 8(2):186.

[219] Taffs, R., Aston, J. E., Brileya, K., Jay, Z., Klatt, C. G., McGlynn, S., Mallette, N., Montross, S., Gerlach, R., Inskeep, W. P., Ward, D. M., and Carlson, R. P. (2009). In silico approaches to study mass and energy flows in microbial consortia: a syntrophic case study. *BMC Syst Biol*, 3:114.

[220] Taranto, M. P., Vera, J. L., Hugenholtz, J., De Valdez, G. F., and Sesma, F. (2003). Lactobacillus reuteri crl1098 produces cobalamin. *Journal of Bacteriology*, 185(18):5643–5647.

[221] Taxis, T. M., Wolff, S., Gregg, S. J., Minton, N. O., Zhang, C., Dai, J., Schnabel, R. D., Taylor, J. F., Kerley, M. S., Pires, J. C., Lamberson, W. R., and Conant, G. C. (2015). The players may change but the game remains: network analyses of ruminal microbiomes suggest taxonomic differences mask functional similarity. *Nucleic Acids Research*.

[222] Testerman, T. L., McGee, D. J., and Mobley, H. L. (2001). Helicobacter pylori growth and urease detection in the chemically defined medium ham's f-12 nutrient mixture. *J Clin Microbiol*, 39(11):3842–50.

[223] Teusink, B., van Enckevort, F. H., Francke, C., Wiersma, A., Wegkamp, A., Smid, E. J., and Siezen, R. J. (2005). In silico reconstruction of the metabolic pathways of lactobacillus plantarum: comparing predictions of nutrient requirements with those from growth experiments. *Appl Environ Microbiol*, 71(11):7253–62.

[224] Teusink, B., Wiersma, A., Molenaar, D., Francke, C., de Vos, W. M., Siezen, R. J., and Smid, E. J. (2006). Analysis of growth of lactobacillus plantarum wcfs1 on a complex medium using a genome-scale metabolic model. *J Biol Chem*, 281(52):40041–8.

[225] 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., De Keersmaecker, S. C., Liao, Y. C., Marchal, K., Mo, M. L., Ozdemir, E., Raghunathan, A., Reed, J. L., Shin, S. I., Sigurbjornsdottir, S., Steinmann, J., Sudarsan, S., Swainston, N., Thijs, I. M., Zengler, K., Palsson, B. O., Adkins, J. N., and Bumann, D. (2011). A community effort towards a knowledge-base and mathematical model of the human pathogen salmonella typhimurium lt2. *BMC Syst Biol*, 5:8.

[226] Thiele, I. and Palsson, B. O. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, 5(1):93–121.

[227] Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., Thorleifsson, S. G., Agren, R., Bolling, 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., Le Novere, N., Malys, N., Mazein, A., Papin, J. A., Price, N. D., Selkov, E., S., Sigurdsson, M. I., Simeonidis, E., Sonnenschein, N., Smallbone, K., Sorokin, A., van Beek, J. H., Weichart, D., Goryanin, I., Nielsen, J., Westerhoff, H. V., Kell, D. B., Mendes, P., and Palsson, B. O. (2013). A community-driven global reconstruction of human metabolism. *Nat Biotechnol*, 31(5):419–25.

[228] Thiele, I., Vlassis, N., and Fleming, R. M. (2014). fastgapfill: efficient gap filling in metabolic networks. *Bioinformatics*, 30(17):2529–31.

[229] Thiele, I., Vo, T. D., Price, N. D., and Palsson, B. O. (2005). Expanded metabolic reconstruction of helicobacter pylori (iit341 gsm/gpr): an in silico genome-scale characterization of single- and double-deletion mutants. *J Bacteriol*, 187(16):5818–30.

[230] Thorleifsson, S. G. and Thiele, I. (2011). rbionet: A cobra toolbox extension for reconstructing high-quality biochemical networks. *Bioinformatics*, 27(14):2009–10.

[231] Tsai, H. N. and Hodgson, D. A. (2003). Development of a synthetic minimal medium for listeria monocytogenes. *Appl Environ Microbiol*, 69(11):6943–5.

[232] Tu, Q., He, Z., and Zhou, J. (2014). Strain/species identification in metagenomes using genome-specific markers. *Nucleic Acids Research*, 42(8):e67.

[233] Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R., and Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457:480–484.

[234] Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., and Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–31.

[235] Varel, V. H. and Bryant, M. P. (1974). Nutritional features of bacteroides fragilis subsp. fragilis. *Appl Microbiol*, 28(2):251–7.

[236] Veith, N., Solheim, M., van Grinsven, K. W. A., Olivier, B. G., Levering, J., Grosseholz, R., Hugenholtz, J., Holo, H., Nes, I., Teusink, B., and Kummer, U. (2015). Using a genome-scale metabolic model of enterococcus faecalis v583 to assess amino acid uptake and its impact on central metabolism. *Applied and environmental microbiology*, 81:1622–33.

[237] Vital, M., Howe, A. C., and Tiedje, J. M. (2014). Revealing the bacterial butyrate synthesis pathways by analyzing (meta)genomic data. *MBio*, 5(2):e00889.

[238] Vlassis, N., Pacheco, M. P., and Sauter, T. (2014). Fast reconstruction of compact context-specific metabolic network models. *PLoS Comput Biol*, 10(1):e1003424.

[239] Větrovský, T. and Baldrian, P. (2013). The variability of the 16s rrna gene in bacterial genomes and its consequences for bacterial community analyses. *PLOS ONE*, 8(2):1–10.

[240] Wegkamp, A., Teusink, B., de Vos, W. M., and Smid, E. J. (2010). Development of a minimal growth medium for lactobacillus plantarum. *Lett Appl Microbiol*, 50(1):57–64.

[241] Wilkinson, A., Day, J., and Bowater, R. (2001). Bacterial dna ligases. *Mol Microbiol*, 40(6):1241–8.

[242] Wilson, A. and Pardee, A. (1962). Regulation of flavin synthesis by escherichia coli. *Journal of General Microbiology*, 28(2):283–303.

[243] Wissenbach, U., Ternes, D., and Unden, G. (1992). An escherichia coli mutant containing only demethylmenaquinone, but no menaquinone: effects on fumarate, dimethyl-sulfoxide, trimethylamine n-oxide and nitrate respiration. *Arch Microbiol*, 158(1):68–73.

[244] Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87:4576–4579.

[245] Wolin, M. J. (1974). Metabolic interactions among intestinal microorganisms. *Am J Clin Nutr*, 27(11):1320–8.

[246] Woodmansey, E. J., McMurdo, M. E., Macfarlane, G. T., and Macfarlane, S. (2004). Comparison of compositions and metabolic activities of fecal microbiotas in young adults and in antibiotic-treated and non-antibiotic-treated elderly subjects. *Appl Environ Microbiol*, 70(10):6113–22.

[247] Woolf, L. I., Griffiths, R., and Moncrieff, A. (1955). Treatment of phenylketonuria with a diet low in phenylalanine. *British Medical Journal*, 1(4905):57–64.

[248] Xavier, J. C., Patil, K. R., and Rocha, I. (2017). Integration of biomass formulations of genome-scale metabolic models with experimental data reveals universally essential cofactors in prokaryotes. *Metab Eng*, 39:200–208.

[249] Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., Knight, R., and Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–7.

[250] Yonezawa, A., Masuda, S., Katsura, T., and Inui, K. (2008). Identification and functional characterization of a novel human and rat riboflavin transporter, rft1. *Am J Physiol Cell Physiol*, 295(3):C632–41.

[251] Zelezniak, A., Andrejev, S., Ponomarova, O., Mende, D. R., Bork, P., and Patil, K. R. (2015). Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc Natl Acad Sci U S A*, 112(20):6449–54.

[252] Zhi, X.-Y., Yao, J.-C., Tang, S.-K., Huang, Y., Li, H.-W., and Li, W.-J. (2014). The futalosine pathway played an important role in menaquinone biosynthesis during early prokaryote evolution. *Genome Biology and Evolution*, 6(1):149.

[253] Zhuang, K., Izallalen, M., Mouser, P., Richter, H., Risso, C., Mahadevan, R., and Lovley, D. R. (2011). Genome-scale dynamic modeling of the competition between rhodoferax and geobacter in anoxic subsurface environments. *Isme j*, 5(2):305–16.

[254] Zomorrodi, A. R., Islam, M. M., and Maranas, C. D. (2014). d-optcom: Dynamic multi-level and multi-objective metabolic modeling of microbial communities. *ACS Synthetic Biology*, 3(4):247–57.

[255] Zomorrodi, A. R. and Maranas, C. D. (2012). Optcom: a multi-level optimization framework for the metabolic modeling and analysis of microbial communities. *PLoS Comput Biol*, 8(2):e1002363.

Appendix A

Supplementary Information for Chapter 2

A.1 Supplementary Data

The file "PEG.fasta" contains amino acid sequences of all protein encoding genes (PEGs) that appear in the eight subsystems for the 256 human gut microbiota (HGM) genomes. Direct download: http://journal.frontiersin.org/file/downloadfile/129714_supplementary-materials_datasheets_pe stream/PEG.FASTA/1/129714.

A.2 Supplementary Tables

The following tables are too large to be displayed in text and are available via the publisher's website. All tables can be found in the file "SupplementaryTables.xlsx". Direct download: http://journal.frontiersin.org/file/downloadfile/129714_supplementary-materials_tables_supplementarytables_ stream/SupplementaryTables.XLSX/1/129714.

Table A.1: List of all peer-reviewed references and books used for knowledge-driven refinement of the AGORA reconstructions.   Direct download: http://www.nature.com/nbt/journal/v35/n1/extref/nbt.3703-S2.xlsx

Sheet	Description
Table S1	Presence (1) and absence (0) of the eight B-vitamin biosynthesis pathways in the 256 HGM and 257 non-HGM genomes.
Table S2	All PubSEED functional roles associated with each abbreviation used in Table 2.2 and Figures 2.1, 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, and 2.8.
Table S3	Biotin biosynthesis subsystem, HGM genomes. PEG numbers of each functional role found in the subsystem for the 256 HGM genomes.
Table S4	Cobalamin biosynthesis subsystem, HGM genomes. PEG numbers of each functional role found in the subsystem for the 256 HGM genomes.
Table S5	Folate biosynthesis subsystem, HGM genomes. PEG numbers of each functional role found in the subsystem for the 256 HGM genomes.
Table S6	Niacin biosynthesis subsystem, HGM genomes. PEG numbers of each functional role found in the subsystem for the 256 HGM genomes.
Table S7	Pantothenate biosynthesis subsystem, HGM genomes. PEG numbers of each functional role found in the subsystem for the 256 HGM genomes.
Table S8	Pyridoxine biosynthesis subsystem, HGM genomes. PEG numbers of each functional role found in the subsystem for the 256 HGM genomes.
Table S9	Riboflavin biosynthesis subsystem, HGM genomes. PEG numbers of each functional role found in the subsystem for the 256 HGM genomes.
Table S10	Thiamin biosynthesis subsystem, HGM genomes. PEG numbers of each functional role found in the subsystem for the 256 HGM genomes.
Table S11	Biotin biosynthesis subsystem, non-HGM genomes. PEG numbers of each functional role found in the subsystem for the 257 non-HGM genomes.
Table S12	Cobalamin biosynthesis subsystem, non-HGM genomes. PEG numbers of each functional role found in the subsystem for the 257 non-HGM genomes.
Table S13	Folate biosynthesis subsystem, non-HGM genomes. PEG numbers of each functional role found in the subsystem for the 257 non-HGM genomes.
Table S14	Niacin biosynthesis subsystem, non-HGM genomes. PEG numbers of each functional role found in the subsystem for the 257 non-HGM genomes.
Table S15	Pantothenate biosynthesis subsystem, non-HGM genomes. PEG numbers of each functional role found in the subsystem for the 257 non-HGM genomes.
Table S16	Pyridoxine biosynthesis subsystem, non-HGM genomes. PEG numbers of each functional role found in the subsystem for the 257 non-HGM genomes.
Table S17	Riboflavin biosynthesis subsystem, non-HGM genomes. PEG numbers of each functional role found in the subsystem for the 257 non-HGM genomes.
Table S18	Thiamin biosynthesis subsystem, non-HGM genomes. PEG numbers of each functional role found in the subsystem for the 257 non-HGM genomes.

Appendix B

Supplementary Information for Chapter 3

B.1 Online Supplementary Material

The following tables are too large to be displayed in text and are available via the publisher's website.

Table B.1: List of all peer-reviewed references and books used for knowledge-driven refinement of the AGORA reconstructions. Direct download: http://www.nature.com/nbt/journal/v35/n1/extref/nbt.3703-S2.xlsx

Sheet	Description
CarbonSourceReferences	Summary of the carbon source utilization pathways included in the 773 gut microbe reconstructions, and supporting references. (1) = presence, (0) = absence.
FermentationReferences	Summary of the fermentation pathways included in the 773 gut microbe reconstructions, and supporting references. (1) = presence, (0) = absence.
NutrientRequirementReferences	Summary of nutrient requirements of the microbes reported in literature that were used for reconstruction curation and validation. (1)=essential *in vivo*, (-1)=nonessential *in vivo*, (0)=no information available.

Table B.2: List of the central metabolic pathways that were curated using a comparative genomics approach. Comparative genomics analysis was performed using PubSEED subsystems [13]. Direct download: http://www.nature.com/nbt/journal/v35/n1/extref/nbt.3703-S3.xlsx

Sheet names		
1. Comments	8. Glu, Gln, As, & Asp	15. Phe & Tyr biosynthesis
2. Purine biosynthesis	9. Cys biosynthesis	16. Pro biosynthesis
3. Pyrimidine biosynthesis	10. Gly biosynthesis	17. Ser biosynthesis
4. TCA and glyoxylate pathway	11. His bioysnthesis	18. Thr biosynthesis
5. Glycolytic pathways	12. Leu, Ile, & Val biosynthesis	19. Trp biosynthesis
6. Ala biosynthesis	13. Lys biosynthesis	20. GlcNAc for polysaccharides
7. Arg biosynthesis	14. Met biosynthesis	

Table B.3: Comparison of stoichiometric consistency and flux consistency between the draft reconstructions and the resulting curated AGORA reconstructions. Direct download: http://www.nature.com/nbt/journal/v35/n1/extref/nbt.3703-S4.xlsx

Sheet	Description
StoichiometricConsistency	Number of reactants (metabolites) and reactions, stoichiometrically and flux consistent reactions and metabolites, and rank of the draft reconstructions (ReconstructionName_B) and the resulting curated AGORA reconstructions (ReconstructionName_A).

Table B.4: Description of each reconstructed strain including taxonomy, biological traits, and reconstruction size. Direct download: http://www.nature.com/nbt/journal/v35/n1/extref/nbt.3703-S5.xlsx

Sheet
OrganismInformation

Table B.5: Predicted *in silico* growth rates (1/h) of AGORA models on Western and high fiber diets in the absence and presence of oxygen. Direct download: http://www.nature.com/nbt/journal/v35/n1/extref/nbt.3703-S6.xlsx

Sheet	Description
GrowthRatesDiets	The input fluxes of Western and high fiber diets are described in Supplementary Table B.18.

Table B.6: Metabolic distance and pairwise growth rates of all AGORA microbe-microbe pairs under four conditions. Direct download: http://www.nature.com/nbt/journal/v35/n1/extref/nbt.3703-S7.xlsx

Sheet	Description
Sheet1	Pairwise *in silico* growth rates calculated for all 298,378 microbe-microbe pairs on each diet (Supplementary Table B.18) and each pair's metabolic distance (Chapter 3: Methods).

Table B.7: Eigenvalues and reactions from a principal coordinate analysis using each Eldermet individual's pan-species reconstruction reaction set. Direct download: http://www.nature.com/nbt/journal/v35/n1/extref/nbt.3703-S8.xlsx

Sheet	Description
PCoA_eigenvalues	The pan-species reconstruction reaction set is obtained by taking the union of reactions from all strain-resolved reconstructions of a certain species. The Eldermet dataset refers to the paper Claesson et al. (2012) [37].

Table B.8: Detailed information on the unique reactions and metabolites present in the 773 reconstructions. Direct download: http://www.nature.com/nbt/journal/v35/n1/extref/nbt.3703-S9.xlsx

Sheet	Description
Unique_reactions	Reaction VMH abbreviation, full name, formula, EC number, Subsystem, KEGG Orthology ID, KEGG reaction ID, and COG ID.
Unique_metabolites	Metabolite abbreviations, full description, charged formula, and charge.

Table B.9: Translation of draft reconstruction reaction and metabolite IDs to the corresponding VMH IDs. Direct download: http://www.nature.com/nbt/journal/v35/n1/extref/nbt.3703-S10.xlsx

Sheet	Description
RxnMetTranslation	Model SEED [102] and KBase [9] draft reconstructions use the same reaction and metabolite IDs. We translated these to the corresponding VMH IDs (http://vmh.uni.lu/) for compatability with the human metaboli reconstruction Recon 2 [227].

Table   B.10:   Reactions   added   to   or   deleted   from   each   of   the   773 draft   reconstructions   during   the   curation   process.   Direct   download: http://www.nature.com/nbt/journal/v35/n1/extref/nbt.3703-S11.xlsx

Sheet	Description
GapfilledReactions	Summary of reactions that were gapfilled manually to enable biomass production aerobically and anaerobically after translating reaction directionalities to BiGG and VMH database standards.
DeletedReactions	Summary of reactions that were gap-filled by the Model SEED and KBase pipelines and were removed from the AGORA reconstructions.
AddedReactions_GenomicAnalysis	Summary of respiration, vitamin biosynthesis, central metabolism, pyrimidine, purine, and amino acid biosynthesis reactions added based on comparative genomic analysis.
FermentationReactions	Summary of fermentation pathways reactions added based on a thorough literature review (see Supplementary Table B.1).
CarbonSourceReactions	Summary of carbon source utilization reactions added based on a thorough literature review (see Supplementary Table B.1).

Table   B.11:   Predicted   presence   or   absence   of   eight   B-vitamin   biosynthesis   pathways   in   the   773   AGORA   organisms.   Direct   download: http://www.nature.com/nbt/journal/v35/n1/extref/nbt.3703-S12.xlsx

Sheet	Description
VitaminPredictions	Predicted presence (1) and absence (0) of eight B-vitamin biosynthesis pathways based on genome analyses (Magnusdottir et al. (2015) [140]) and published experimental data (green cells).

Table   B.12:   The   list   of   AGORA   reconstructions   that   were   mapped   to each   of   the   HMP   and   ELDERMET   individual   samples.   Direct   download: http://www.nature.com/nbt/journal/v35/n1/extref/nbt.3703-S13.xlsx

Sheet	Description
HMP_Mapping	AGORA reconstructions mapped to 149 HMP individual metagenomic samples (The Human Microbiome Project Consortium (2012) [42]). 1: mapped to sample, 0: not mapped to sample.
Eldermet_Mapping	AGORA species mapped to 177 Eldermet individual 16S rRNA samples (Claesson et al. (2012) [37]). 1: mapped to sample, 0: not mapped to sample.

B.2 Supplementary Notes

B.2.1 Description of QC/QA and data-driven curation efforts.

We have compared and evaluated the predictive accuracy of AGORA against numerous resources, including genome-scale single gene deletion data (Fig. 3.1d), which are generally accepted as gold standard for the evaluation of the predictive potential of microbial metabolic models. The resources, which we used to assess and demonstrate the prediction accuracy, include (i) Genome-scale gene essentiality data (taken from `http://www.essentialgene.org/`). (ii) Experimental data on the utilization of all main groups of carbon sources. (iii) Experimental data on the major fermentation pathways in the human gut microbiota. (iv) Experimentally determined defined media/ nutrient requirements. The latter three data have been collected from more than 170 primary research papers (Supplementary Table B.1) and the books: "Bergey's manual of systematic bacteriology" and "The prokaryotes: a handbook on the biology of bacteria". As such, we demonstrated the predictive sensitivity of AGORA using information generally accepted as gold standards by the field. The results can be found in Figures 3.1 and 3.3 in the main text, Supplementary Table B.13, and Supplementary Fig. B.3. We would like to highlight that inter-organism propagations were performed on the species level and in limited cases on the genus level. Hence, the metabolic functions were not inferred from distant organisms. It is to be noted that molecular systems biology is characterized by an iterative cycle of model predictions and experimental validation. It is natural in such a process for either prediction or validation to be one step ahead of the other. Compared to other biological domains, our experimental knowledge of the gut microbiome lags behind. Genome-scale models, such as those in AGORA, can be used to provide novel insight into the biology of the considered gut microbe and to drive the design of experimental projects.

Reaction directionality was propagated from biochemically curated reactions in the VMH database (`http://vmh.life`) and each was checked for consistency with thermodynamic estimates of maximum and minimum standard transformed Gibbs energy for each reaction using our state of the art Component Contribution method1. As documented on the VMH database, the following conditions were assumed: (i) temperature 310.15K; (ii) cytosol pH 7.2, extracellular pH 7.4; (iii) ionic strength 0.15 mol/L; (iv) 30 mV electrical potential difference between extracellular space and cytosol; and (v) minimum and maximum concentration range from 10^{-7} to 10^{-2} mol/L. The transformed Gibbs energy of formation of a metabolite is the sum of standard transformed Gibbs energy of formation and an $RTlog(x)$ term, where R is the gas constant, T is temperature and x is the absolute concentration of the metabolite. Note using this approach, the activity coefficient is absorbed into the standard transformed Gibbs energy of formation. If we assume that the conditions above do vary, let us consider them one by one. (i) Temperature varies only slightly, and there are not enough enthalpic data to make a temperature adjustment beyond what we do. (ii) Our choice of pH 7.4 is representative of the terminal ileum. According to Fallingborg J. [65] "The intraluminal pH is rapidly changed from highly acid in the stomach to about pH 6 in the duodenum. The pH gradually increases in the small intestine from pH 6 to about pH 7.4 in the terminal ileum. The pH drops to 5.7 in the caecum, but again gradually increases, reaching pH 6.7 in the rectum." This slight variation in small and large intestinal pH might change the direction of some cytoplasmic transport reactions. (iii) The ionic strength is not known ac-

curately, for any cell. We use 0.15 mol/L as is standard in the biochemical thermodynamic literature. Experimental evidence would be required to justify the use of a different value. (iv) The electrical potential difference between extracellular space and cytosol could indeed vary, however, then one would be modeling the same organism in two different conditions so one would expect the model predictions to be different. (v) There is insufficient data on the absolute concentrations of the metabolome in the cytoplasm in any cell. With gut microbial species, the paucity of data is even more acute. Without further data, one could not safely assume a tighter concentration range.

B.2.2   Comparison with published reconstructions.

We compared all 3,192 unique reactions included in the 773 AGORA reconstructions with the 4,608 unique reactions included in 11 published gut microbe reconstructions previously used to simulate a model gut community [98] after unifying the reaction namespaces. Additionally, the reconstructions of eight strains in AGORA overlapped with published reconstructed strains [96, 165, 224, 225, 229, 20, 133] and were compared against each other directly. Subsystems were assigned to every reaction in the AGORA and the published reconstructions based on the subsystem nomenclature in Recon2 [227]. The results are shown in Supplementary Tables B.14 and B.15

Between 310 and 1,058 (mean 682 ± 290) reactions overlapped between the respective AGORA and published reconstructions. Between 305 and 934 (mean 770 ± 227) reactions were unique to the curated reconstructions, and between 229 and 1,776 (mean 956 ± 678) reactions were unique to the published reconstructions (Supplementary Table B.14). Most of the reactions that were unique to the published reconstructions belonged to exchange and transport subsystems, particularly for *Escherichia coli* K-12 substr. MG1655, *Escherichia coli* O157H7 substr. Sakai, *Klebsiella pneumoniae* MGH78578, and *Salmonella enterica* sv. typhimurium LT-2 (Supplementary Table B.15). These reconstructions used the *Escherichia coli* reconstruction iAF1260 [68] as a template, which includes a periplasm compartment. Thus, the majority of reaction content unique to these reconstructions could be attributed to the presence of the additional periplasm compartment, which requires additional transport reactions. For example, the *Escherichia coli* K-12 substr. MG1655 reconstruction iEco1339_MG1655 [20] contains 1,573 reactions not found in the AGORA reconstruction. Of those, 913 reactions take place in the periplasmic compartment. Of the 934 reactions found only in the AGORA *Escherichia coli* K-12 substr. MG1655 reconstruction, 197 are cytosolic versions of reactions that take place in the periplasm in the published reconstruction. Another major reason for discrepancies between the AGORA and published reconstruction was the captured cell wall and lipopolysaccharide biosynthesis pathways. Cell wall and lipopolysaccharide structures are species-specific and generally poorly annotated, and hence difficult to curate. Consequently, between 30 and 299 reactions from cell wall and lipopolysaccharide biosynthesis were unique to the published reconstructions (Supplementary Table B.15). Curating for cell wall and lipopolysaccharides structures and including accurate transport mechanisms would require experimental data, which is not available for most AGORA organisms. Thus, these curation steps were not performed in the present study, with the exception of reconstructions from two genera, *Mycoplasma* and *Ureoplasma* (Chapter 3: Methods). At the same time, 161 reactions from cell wall and lipopolysaccharide biosynthesis were present in the *E. coli* draft reconstruction, and thus also present in

its AGORA version, but not in the published reconstruction iEco1339_MG1655 (Supplementary Table B.15). These reactions are involved in fatty acid biosynthesis, as the ones catalyzed by the 3-oxoacyl-acyl-carrier-protein reductase. Since the presence of these enzymes in *E. coli* is supported by genome annotation, these reactions should also be included in the published reconstruction. Another example is dipeptide degradation, which is supported by genome annotation in the AGORA reconstruction but is absent from the published *E. coli* reconstruction.

The eight AGORA reconstructions had on average a higher number of blocked reactions that the respective published reconstructions ($31 \pm 9\%$ compared with $13 \pm 6\%$). This was mainly due to the deletion of reactions that were added during the automated gap-filling step of the draft reconstructing pipelines and were found to no longer be required for biomass production. Since the presence of these reactions is not supported by gene annotation or experimental data, corresponding to a confidence score of 1 [226], their inclusion is hypothetical. Another cause for blocked reactions in AGORA reconstructions was the adjustment of reaction reversibilities to VMH standards. This led to some pathways that were carrying flux in an infeasible direction in the draft reconstructions to be blocked in the resulting AGORA reconstructions. For example, the vitamin B12 biosynthesis pathway was reversible in many draft reconstructions and often allowed to produce downstream biomass precursors from cobalamin. Making this pathway irreversible caused it to be blocked in many reconstructions due to missing steps, in agreement with many microbes being unable to synthesize cobalamin [140] (Supplementary Table B.11). Nonetheless, the QC/QA curation effort led to an increase of the overall stoichiometric and flux-consistent reactions in the AGORA reconstructions compared to the draft reconstructions (Fig. 3.1, Supplementary Figs. B.1 and B.2).

B.2.3 *In vitro* cell cultures and cell counting.

In vitro cell culture: Pre-cultures of *Bacteroides caccae* ATCC 43185 (*B. caccae*) and *Lactobacillus rhamnosus* GG ATCC 53103 (LGG) were prepared using Brain Heart Infusion Broth (BHIS; Sigma) supplemented with 1% hemin. The pre-cultures were run for 20 hours under anaerobic conditions and while shaking at 37°C. The volume of the cell suspension was adjusted to a maximum optical density (OD) at 600 nm of 0.8 using sterile 0.9% w/v NaCl solution to obtain reliable OD measurements. Following pre-culturing, centrifugation was carried out at 4,700 x g for 10 min at room temperature. The resulting cell pellets were washed twice in 0.9% w/v NaCl solution and subsequently resuspended in 10 ml of 0.9% w/v NaCl solution. Subsequently, 1 ml of the cell suspensions were inoculated in DMEM 6429 supplemented with 1% hemin and 3.33% vitamin with a starting OD of 0.1 and K maintained under anaerobic conditions. Culture was carried out in media with or without the addition of arabinogalactan (Sigma; 9.4 g/l). *B. caccae* and LGG were cultured for 33 and 44 hours of culture for *B. caccae* and LGG, respectively. Cells were harvested for cell counting by centrifugation (4,700 g) and 750 μL aliquots of supernatant were removed for subsequent metabolite extraction. The aliquots were snap-frozen and placed at $-80°C$ until analysis. The measured ODs and pH values are listed in Supplementary Table B.16.

Cell counting: Bacterial pellets were thawed and subsequently stained with the Texas Red®-X dye–labeled Wheat Germ Aglutinin component which selectively binds to the surface of gram-positive bacteria, effectively distinguishing them from gram-negative bacteria.

Cells were washed, resuspended in sterile 0.9% w/v NaCl solution and quantified by flow cytometry (BD Fortessa) using negative beads (Thermo Fischer) as a standard for the volume of suspension. The resulting data were analyzed using the DIVA 8.0.1 software (BD Biosciences). The cell counts are shown in Supplementary Table B.16.

Real-time PCR (RT-PCR): The presence of *B. caccae* cells in the cultures were validated using qPCR. Microbial genomic DNA was extracted using the PowerSoil DNA Isolsation Kit from MoBio accordingly to the manufacturer's protocol. qPCR was performed using *B. caccae*-specific primers (Eurogentec). The *B. caccae*-specific primers were 5'-CCC GGA GTT GGA AAA CAA TG-3' (forward) and 3'-CCT CTT CAG AAA TGA GCT TTT GC-3' (reverse). 5 ng of DNA were used in a 20 μl PCR reaction mixture containing 10 μl iQ$^{\text{TM}}$ SYBR®Green Supermix (Bio-Rad) and 500 nM of each primer. PCR amplifications were performed on a LightCycler®480 Instrument (40 cycles at 95°C for 10 sec and 55°C for 20 sec; Roche).

B.2.4   Fermentation and carbon source utilization pathways captured by AGORA.

A thorough literature search was performed for the reconstruction of the main fermentation pathways in the human gut and their distribution across phyla. The main products of carbohydrate and protein fermentation by the human gut microbiota are the short-chain fatty acids (SCFAs) acetate, propionate, and butyrate [138]. The different routes leading to these products were reconstructed for all AGORA organisms reported to carry these pathways and are briefly described below. Acetate is produced via the widespread acetate kinase [122]. Moreover, bacteria in the *Bifidobacterium* genus produce acetate via the bifid shunt [122], and the genera *Blautia* and *Marvinbryantia* convert CO_2 and hydrogen to acetate via acetogenesis [180]. Three pathways exist for the conversion of carbohydrates and amino acids to propionate, with succinate, propane-1,2-diol, and lactate, respectively, as intermediates [179]. Their distribution in the human gut microbiota has been analyzed by comparative genomics and experimentally validated [179]. Carbohydrate and amino acid fermentation to butyrate is carried out via five routes, with acetyl-CoA, glutarate, lysine, 4-aminobutyrate, and succinate, respectively, as intermediates [237]. A genomic analysis of their distribution across several phyla has been performed [237]. Moreover, gut microbes produce lactate, formate, butanol, acetoin, 2,3-butanediol, and ethanol, as well as carbon dioxide and hydrogen [122]. For most gut microbes, the capabilities to produce these acids and gases are well described in the literature [122]. Produced hydrogen is cross-fed to other species in three ways: methanogenesis by archaeal representatives resulting in methane production [193], dissimilatory sulfate reduction to sulfide [181], and acetogenesis yielding acetate [180]. Amino acid fermentation results in the production of not only SCFAs and gases, but also the branched-chain fatty acids isobutyrate and isovalerate, as well as phenols (e.g., phenylacetate) and indoles [138, 48]. The pathways described above, resulting in 16 fermentation products and two gases in total, were included in the respective AGORA reconstructions.

The available literature was also searched for the utilization of carbon sources by AGORA organisms. The gut microbiota utilizes a variety of diet-and host-derived carbon sources, including simple sugars, starch, fiber, host-derived polysaccharides, protein, and organic acids [74]. The potential to exploit carbon sources is species-specific. While the ability to utilize mono- and disaccharides is widely spread [122], the capability to break down diet-

and host-derived polysaccharides is limited to certain genera, e.g., *Bacteroides*, *Bifidobacterium*, *Roseburia*, and *Ruminococcus* [75]. Some species utilize amino acids as carbon and energy sources, e.g., *Clostridium difficile*, *Pseudomonas aeruginosa*, and *Porphyromonas gingivalis* [122], and some utilize intermediates of central metabolism and organic acids, e.g., *Bacillus cereus*, *Citrobacter* sp., *Oxalobacter formigenes*, and *Veillonella* sp. [122]. A thorough literature search was performed for the distribution of these four groups of carbon sources utilized by the AGORA organisms: (i) simple sugars and oligosaccharides, (ii) polysaccharides and fibers, (iii) amino acids, and (iv) organic acids and other metabolic intermediates. In total, information on the utilization of 95 carbon sources was gathered and the corresponding pathways were included in the respective AGORA reconstructions.

B.2.5   Definition of sub-pathways.

A sub-pathway was determined as any set of reactions that converts an initial substrate of the pathway into the final product(s) of the pathway. For example, the biosynthesis of a purine nucleotide has one initial substrate, phosphoribosyl pyrophosphate, and two final products, AMP and GMP. Thus, this pathway includes two sub-pathways. A sub-pathway was considered complete if all genes required for all the reactions in the sub-pathway were present in the subsystem. A sub-pathway was gap-filled if the length of a gap in the pathway did not exceed one reaction. In this case, the gaps were filled by reactions not associated with GPRs. Sub-pathways with gaps longer than one reaction were considered incomplete and reactions for these subpathways were not included in reconstructions. These criteria were applied to all the reconstructed pathways except the citric acid cycle, because the presence of an incomplete citric acid cycle has been confirmed for multiple microbial genomes [108]. These incomplete versions of the citric acid cycle are used for the biosynthesis of various compounds, such as fatty and amino acids [108]. Thus, no gap-filling was performed for the citric acid cycle and reactions for this pathway were included in the reconstructions regardless of the pathway's completeness in the genome.

B.2.6   Curation of respiration and quinone biosynthesis in AGORA.

The genomes of the reconstructed gut bacteria contain multiple aerobic reductases as well as anaerobic reductases for tetrathionate, thiosulfate, polysulfide, sulfite, adenylyl sulfate, heterodisulfides, fumarate, trimethylamine N-oxide, dimethyl sulfoxide, nitrate, nitrate, nitrogen oxide, nitrous oxide, selenate, and arsenate [174]. Nonetheless, the reactions for reduction of respiratory electron acceptors should include two half-reactions, one for a reduction of the electron acceptor itself and another for an oxidation of the corresponding quinone. Since the repertoire of synthesized quinones varies among bacterial taxa [156] and quinones demonstrate specificity for their electron acceptors [243], the inclusion of respiratory reactions into the models was preceded by the reconstruction of the quinone biosynthetic pathways.

In bacteria, one pathway has been described for ubiquinone (UQ) biosynthesis. For menaquinone (MK) biosynthesis, two different pathways are known, the first one through O-succinylbenzoate and the second one through futalosine [156]. The last steps of MK biosynthesis through futalosine are unknown, but are proposed to be catalyzed by a polyprenyltransferase, a carboxy-lyase, and a methyltransferase [252]. All three steps of polyprenylation,

carboxyl elimination, and methylation are present also in UQ biosynthesis and in MK biosynthesis O-succinylbenzoate biosynthesis in the same order as listed above. So, we proposed that this reaction mechanism should be conserved in MK through futalosine biosynthesis. Thus, in analogy with the corresponding steps in the UQ and MK via O-succinylbenzoate biosynthesis pathways, we predicted the three last steps of the MK through futalosine pathway in the analyzed genomes:

- Polyprenyltransferase

$$1,4\text{-dihydroxy-6-naphthoate + polyprenyl-pyrophosphate} \rightarrow$$
$$3\text{-polyprenyl-1,4-dihydroxy-6-naphthoate + pyrophosphate}$$

- Carboxy-lyase

$$3\text{-polyprenyl-1,4-dihydroxy-6-naphthoate} \rightarrow 2\text{-demethyl menaquinol} + CO_2$$

- Methyltransferase

$$2\text{-demethyl menaquinol + S-adenosyl-L-methionine} \rightarrow$$
$$\text{menaquinol + S-adenosyl-L-homocysteine} + H^+$$
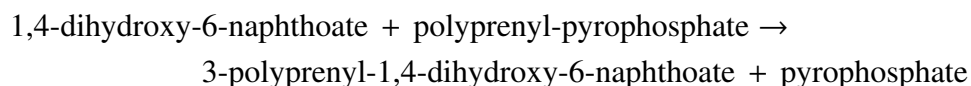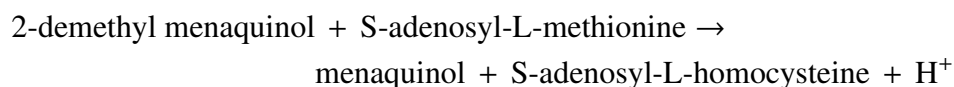
The reactions for the respiratory reduction of electron acceptors were constructed in agreement with the following features, (1) the presence of the quinone biosynthetic pathways in the analyzed genome, (2) specificity of the electron acceptor to quinones, and (3) subcellular localization of an active center of the corresponding reductase. For example, the *Bacteroides thetaiotaomicron* genome contains the biosynthesis pathway for MK and 2-demethylmenaquinone (DMK), one aerobic reductase with the cytoplasmic active center, and two anaerobic reductases: a nitrite reductase with an extracellular active center and a fumarate reductase with a cytoplasmic active center. Because oxygen can be reduced with MK, whereas nitrite and fumarate can be reduced with both MK and DMK [243], the respiratory reduction of electron acceptors in *Bacteroides thetaiotaomicron* is carried out by the following five reactions:

- Cyd: aerobic reductase

$$0.5\,O_2[c] + \text{menaquinol}[c] + 2\,H^+[c] \rightarrow H_2O[c] + \text{menaquinone}[c] + 2\,H^+[e]$$

- Nrf: nitrite reductase

$$NO_2[e] + 3\,\text{menaquinol}[c] + 2\,H^+[c] \rightarrow NH_4^+[c] + 3\,\text{menaquinone}[c] + H_2O[c]$$
$$NO_2[e] + 3\,2\text{-demethyl menaquinol}[c] + 2\,H^+[c] \rightarrow$$
$$NH_4^+[c] + 3\,\text{menaquinone}[c] + H_2O[c]$$

- Frd: fumarate reductase

$$\text{fumarate}[c] + \text{menaquinol}[c] \rightarrow \text{succinate}[c] + \text{menaquinone}[c]$$
$$\text{fumarate}[c] + 2\text{-demethyl menaquinol}[c] \rightarrow$$
$$\text{succinate}[c] + 2\text{-demethyl menaquinone}[c]$$
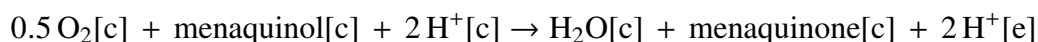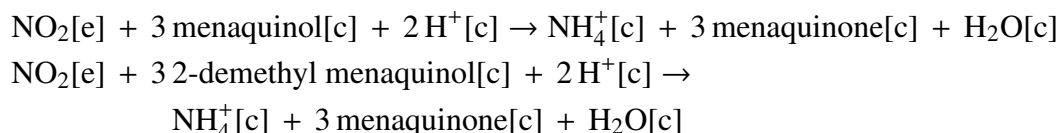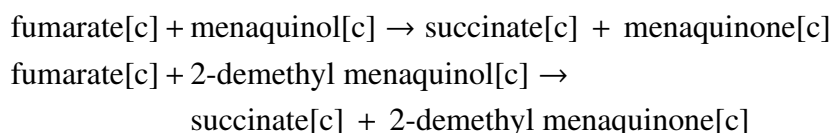
Reactions for ATP synthesis via proton-driven ATP synthases were also added to the reconstructions. All analyzed genomes have genes for F-type or V-type ATP synthases. All added respiration, quinone biosynthesis, and ATP synthase reactions are listed in Supplementary Table B.10.

B.2.7   Curation of nutrient requirements.

The *in silico* growth requirements were computed by setting the lower bounds for all exchange reactions to zero individually and predicting if the model could still produce biomass. The analysis revealed essential metabolites for *in silico* growth that are unlikely to be found in the human gut, such as end products of coenzyme biosynthesis, including CoA and NAD(P)H. Bacteria take up precursors of these coenzymes in the form of vitamins, such as pantothenic acid or nicotinic acid. Exchange and transport reactions for vitamins were added where necessary and we ensured that CoA synthesis from pantothenic acid, as well as NADH biosynthesis from nicotinic acid, were unblocked in all metabolic reconstructions. In some cases, a draft reconstructions required dimers or oligomers in the *in silico* growth medium to fulfill a monomer requirement. As it can be expected that uptake of the monomer itself would also satisfy the growth requirement, exchange and transport reactions for the corresponding monomers were added. For example, the galactose-containing oligosaccharides stachyose, melibiose, and lactose were in some cases essential for a model to satisfy its requirement for galactose. Similarly, some draft models depended on dipeptides for certain amino acids, so we enabled the uptake of amino acid monomers.

False negative predictions were resolved in the following ways: (i) for essential nutrients, if they were not yet present in the reconstruction, exchange and transport reactions were added, (ii) for metabolites not included in the biomass reaction, a demand reaction was added for metabolites required *in vitro* and a minimal flux through the demand reaction was enforced, and (iii) for metabolites that were false positives due to gap-filled reactions in the biosynthesis pathways, those gap-filled reactions were removed. False positive predictions were resolved by manually inspecting and gap-filling the corresponding pathways to enable production or consumption of dead-end metabolites. This curation was performed for 244 models.

After curation, 173 false positive predictions remained, which were due to multiple gaps in the networks that could not be resolved without further experimental and genomic evidence. For most of the 219 remaining false negative predictions, the compounds were essential *in vitro* although the biosynthesis pathways were completely annotated in the bacterial genomes. This has been previously observed, e.g., in *Lactobacillus plantarum* WCFS1, and may be explained by feedback inhibition of the biosynthesis pathways *in vitro* [223].

B.2.8   Metabolite extraction.

Short chain fatty acid extraction, derivatization, and GC-MS measurement

The extraction of short chain fatty acids was based on a protocol from Moreau et al. [154]. Briefly, 10 μL of the internal standard (2-Ethylbutyric acid, c = 200 mmol/L) were added to 190 μL of medium. After acidification with 10 μL of hydrochloric acid (c = 1 mol/L), 1 mL of diethyl ether was added and the samples were vortexed for 10 min at 1400 rpm at room

temperature (Eppendorf Thermomixer). The upper organic phase was separated by centrifugation (5 min, 21,000 xg) and 900 μL were collected in a new reaction tube. Again, 1 mL of diethyl ether were added to the medium, incubated and separated by centrifugation. 900 μL of the organic phase were combined with the first extract. Then, 250 μl were transferred into a GC glass vial with micro insert (5-250 μL) in triplicates. For derivatization, 25 μL of N-tert-Butyldimethylsilyl-Nmethyltrifluoroacetamide with 1% tert-Butyldimethylchlorosilane (MTBSTFA + 1% TBDMSCI, Restek) was added and the samples were incubated for a minimum of 1 hour at room temperature. To determine retention times and evaluate separation efficiency, a Volatile Free Acid Mix (Sigma-Aldrich) including all compounds of interest was prepared, extracted, and derivatized as described before.

GC-MS analysis was performed by using an Agilent 7890A GC coupled to an Agilent 5975C inert XL Mass Selective Detector (Agilent Technologies). A sample volume of 1 μl was injected into a Split/Splitless inlet, operating in split mode (20:1) at 270°C. The gas chromatograph was equipped with a 20 m (I.D. 180 μm, film 0.18 μm) DB-1MS capillary column (Agilent J&W GC Column). Helium was used as carrier gas with a constant flow rate of 1.0 ml/min. The GC oven temperature was held at 80°C for 0.75 min and increased to 150°C at 15°C/min. After 2 min, the temperature was increased at 50°C/min to 280°C and held for 2 min. The total run time was 12.017 min. The transfer line temperature was set to 280°C. The mass selective detector (MSD) was operating under electron ionization at 70 eV. The MS source was held at 230°C and the quadrupole at 150°C. The detector was switched off during elution of MTBSTFA. For quantification, measurements of the compounds of interest were performed in selected ion monitoring mode. Dwell times as well as quantification and qualification ions (m/z) are shown in Supplementary Table B.20.

Absolute quantification of medium components using the YSI Biochemistry Analyzer

Prior to measurement, media samples were filtrated (PHENEX-RC 4mm, 0.2 μm; Phenomenex) to remove particles. Absolute quantitative values for lactic acid, glutamine, glutamic acid, and glucose were acquired using the 2950D Biochemistry Analyzer (YSI). In addition, for a precise and reliable quantification, external concentration curves for each compound of interest were prepared and measured in triplicate.

Polar metabolite extraction, derivatization, and GC-MS measurement

Extracellular metabolites from media samples were extracted in triplicate using ice-cold extraction fluid (5:1 methanol/water, v/v) containing the internal standards [U-13C]ribitol (c = 10 μg/mL; Omicron Biochemicals, Inc) and pentanedioic acid-D6 (c = 4 μg/mL; C/D/N Isotopes Inc.). 20 μL of medium was added to 180 μL ice-cold extraction fluid, vortexed for 15 min at 4°C and 1,400 rpm (Eppendorf Thermomixer), then, centrifuged at 21,000 xg for 5 min at 4°C. 50 μL of medium extracts were transferred to GC glass vial with micro insert (5-250 μL) and evaporated under vacuum to dry at −4°C. For absolute metabolite quantification, a dilution series of a standard mixture containing all metabolites of interest was included in the extraction procedure and measured in triplicates. Metabolite derivatization was performed by using a multipurpose sampler (Gerstel). Dried medium extracts were dissolved in 15 μl pyridine, containing 20 mg/ml methoxyamine hydrochloride (Sigma-Aldrich), at 55°C for 90 min under shaking. After adding 15 μl MTBSTFA + 1% TBDMSCI (Restek), samples were incubated at 55°C for 60 min under continuous shaking.

GC-MS analysis was performed by using an Agilent 7890A GC coupled to an Agilent 5975C inert XL Mass Selective Detector (Agilent Technologies). A sample volume of 1 μl was injected into a Split/Splitless inlet, operating in split mode (10:1) at 270°C. The gas chromatograph was equipped with a 30 m (I.D. 250 μm, film 0.25 μm) DB-35MS capillary column + 5 m DuraGuard column in front of the analytical column (Agilent J&W GC Column). Helium was used as carrier gas with a constant flow rate of 1.0 ml/min. The GC oven temperature was held at 100°C for 2 min and increased to 300°C at 10°C/min and held for 4 min. The total run time was 26 min. The transfer line temperature was set to 280°C. The MSD was operating under electron ionization at 70 eV. The MS source was held at 230°C and the quadrupole at 150°C. For precise quantification, GC-MS measurements of the derivatives of interest were performed in selected ion monitoring mode. Dwell times as well as quantification and qualification ions (m/z) are shown in Supplementary Table B.20.

Data normalization and data processing

All GC-MS chromatograms were processed using MetaboliteDetector, v3.020151231Ra32. The software package supports automatic deconvolution of all mass spectra. Compounds were annotated by retention time and mass spectrum. The internal standards were added at the same concentration to every medium sample to correct for uncontrolled sample losses and analyte degradation during metabolite extraction. The data set was normalized by using the response ratio of the QI_Analyte and the QI_Internal Standard (peak area of the analyte divided by the peak area of the internal standard). Absolute concentrations were determined using calibration curves from external standards. To evaluate the variability of independent cultivations, mean values of three technical replicates have been calculated for each biological replicate.

Figure B.1

Figure B.1: Comparison of metabolite stoichiometric and flux consistency [72] of draft and AGORA reconstructions (Supplementary Table B.3). The figures highlight two groups in the draft reconstruction set. We identified the time of download as the key separating factor. Those reconstructions that have been obtained from Model SEED before summer 2015 had a smaller reaction (and metabolite) content, than those ones downloaded from model SEED or KBase afterwards. Model SEED/KBase updated and expanded their underlying database substantially in 2015 to include secondary metabolite, xenobiotics, and plant metabolism. Consequently, the average size of the draft reconstructions increased. These metabolic pathways were outside the scope of the current reconstruction effort and thus were not retained in AGORA reconstructions of the black group. Further experimental data and comparative genomic efforts will be required to establish that those out-of-scope reactions do indeed occur in the respective gut microbes. All reconstructions of the black group thus fall below the red line, meaning that the AGORA counterpart is smaller in terms of metabolites. However, it is notable and thanks to the QC/QA effort applied to all AGORA reconstructions that the there is no observable difference between the two groups when comparing the stoichiometric and flux consistent metabolites. In all cases, we improved the quality of the reconstructions when considering this measure over the draft reconstructions. (a) The number of metabolites in each AGORA reconstruction versus the corresponding draft reconstruction. The red line shows the line $y = x$, where the number of metabolites in the AGORA reconstruction is the same as the number of metabolites in the draft reconstruction. (b) Histograms showing the change in number of metabolites after the curation of the draft reconstructions of the two groups. The mean and standard deviation of the change in number of metabolites is shown for both groups. A similar separation was observed for (c-d) the rank of the draft and AGORA stoichiometric matrices and (e-f) the number of stoichiometrically consistent metabolites in AGORA versus the draft reconstructions. (g-h) In most organisms of both groups, the number of stoichiometrically and flux consistent metabolites was increased by the AGORA curation process.
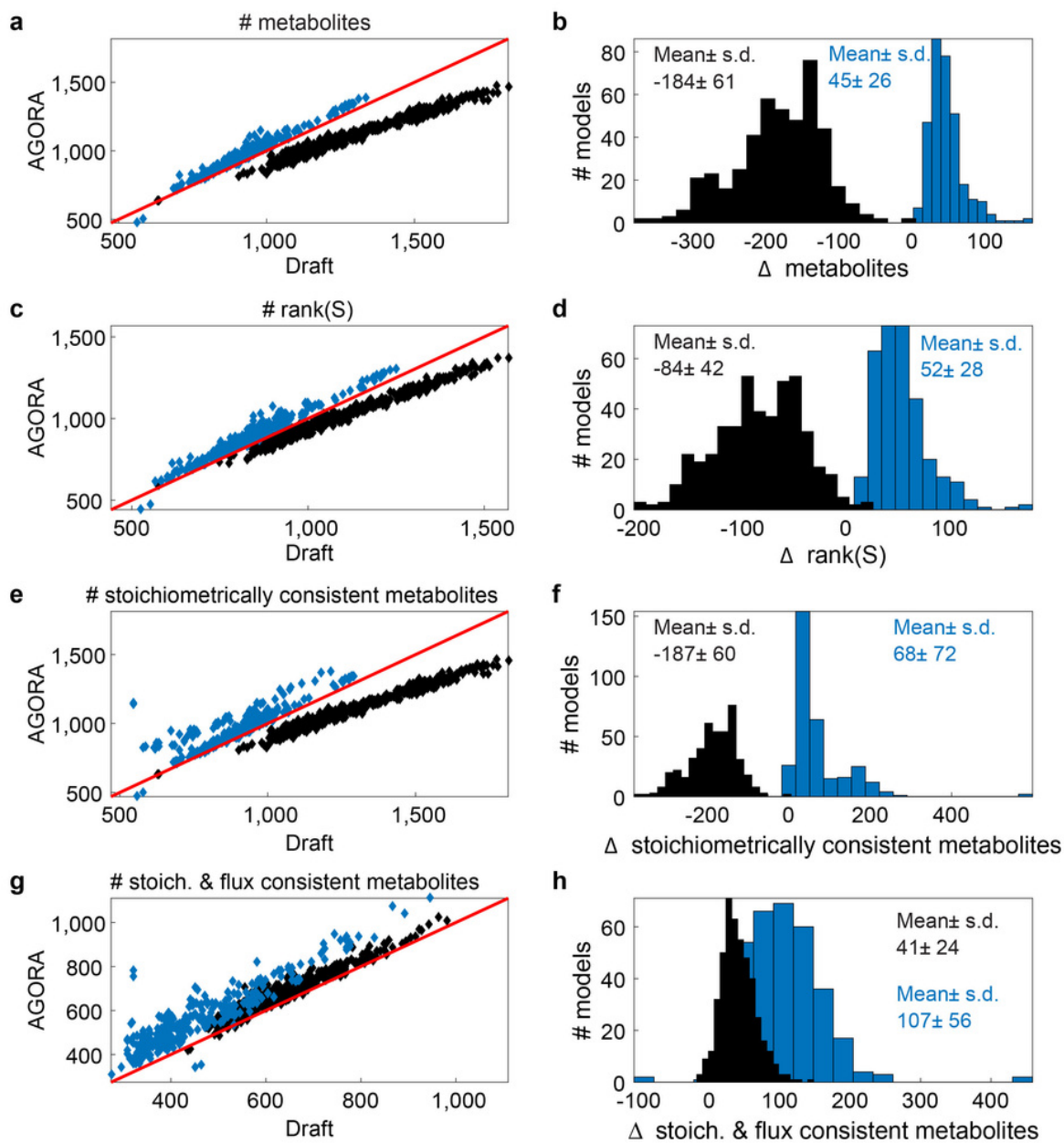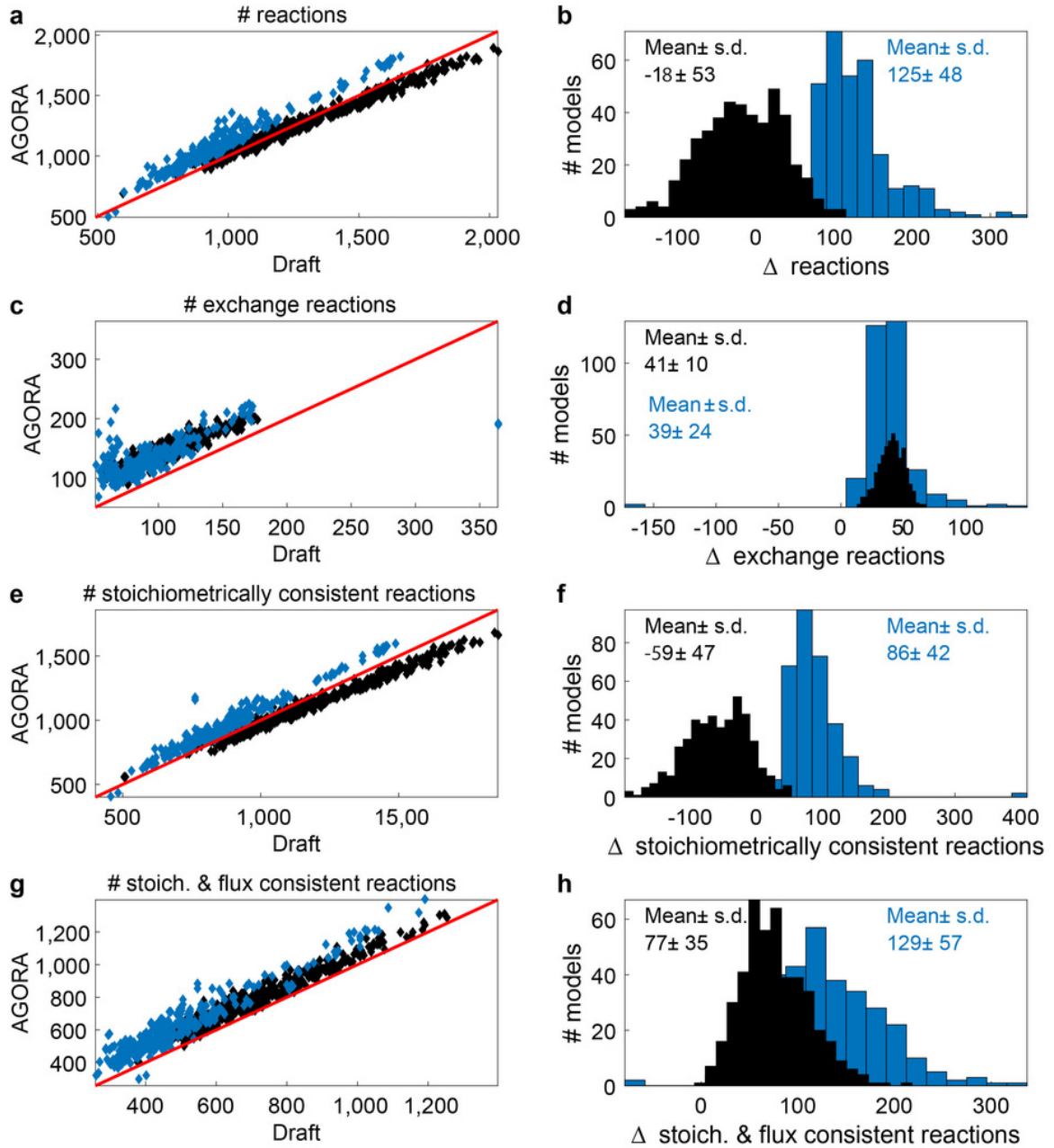
Figure B.2

Figure B.2: Comparison of reaction stoichiometric and flux consistency [72] of draft and AGORA reconstructions (Supplementary Table B.3). The figures highlight two groups in the draft reconstruction set. We identified the time of download as the key separating factor. Those reconstructions that have been obtained from Model SEED before summer 2015 had a smaller reaction (and metabolite) content, than those ones downloaded from model SEED or KBase afterwards. Model SEED/KBase updated and expanded their underlying database substantially in 2015 to include secondary metabolite, xenobiotics, and plant metabolism. Consequently, the average size of the draft reconstructions increased. These metabolic pathways were outside the scope of the current reconstruction effort and thus were not retained in AGORA reconstructions of the black group. Further experimental data and comparative genomic efforts will be required to establish that those out-of-scope reactions do indeed occur in the respective gut microbes. All reconstructions of the black group thus fall below the red line, meaning that the AGORA counterpart is smaller in terms of reactions. However, it is notable and thanks to the QC/QA effort applied to all AGORA reconstructions that the there is no observable difference between the two groups when comparing the stoichiometric and flux consistent reactions. In all cases, we improved the quality of the reconstructions when considering this measure over the draft reconstructions. (a) The number of reactions in each AGORA reconstruction versus the corresponding draft reconstruction. The red line shows the line $y = x$, where the number of reactions in the AGORA reconstruction is the same as the number of metabolites in the draft reconstruction. (b) Histograms showing the change in number of reactions after the curation of the draft reconstructions of the two groups. The mean and standard deviation of the change in number of reactions is shown for both groups. A similar separation was observed for (c-d) the number of exchange reactions and (e-f) the number of stoichiometrically consistent reactions in AGORA versus the draft reconstructions. (g-h) In most organisms of both groups, reaction stoichiometric and flux consistency was improved by the AGORA curation process.
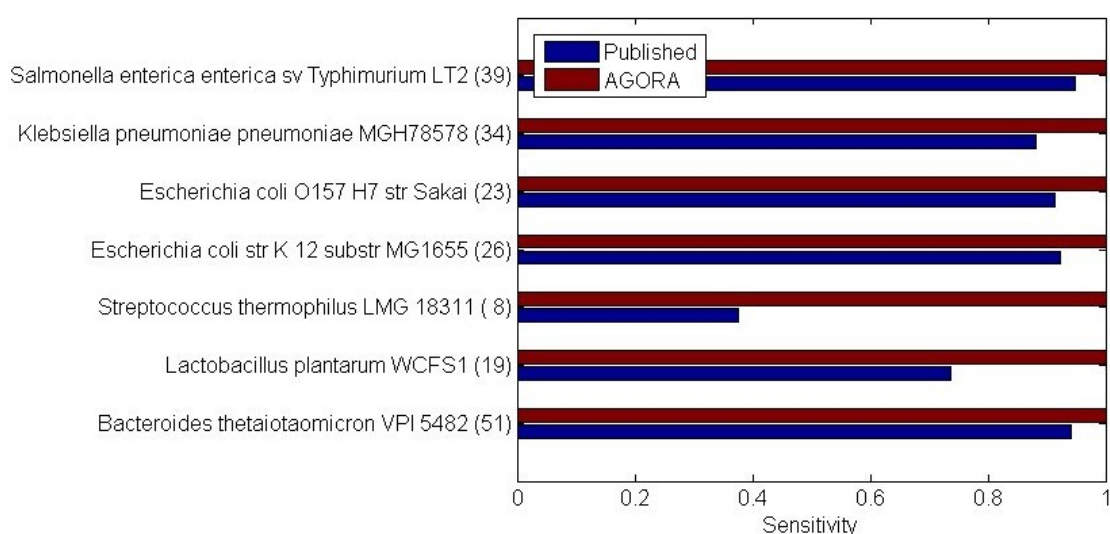
Figure B.3: Sensitivity of carbon source uptake and fermentation product secretion of seven published models [165, 224, 225, 20, 133] and the corresponding AGORA models. In the parentheses are the number of carbon sources or fermentation products that the models should take up or secrete, respectively, according to data from literature (Supplementary Table B.1). Uptake and secretion capabilities were determined using flux variability analysis. All exchange reactions had unlimited upper and lower bounds and a minimum flux of 0.001 1/h through the biomass objective function enforced.
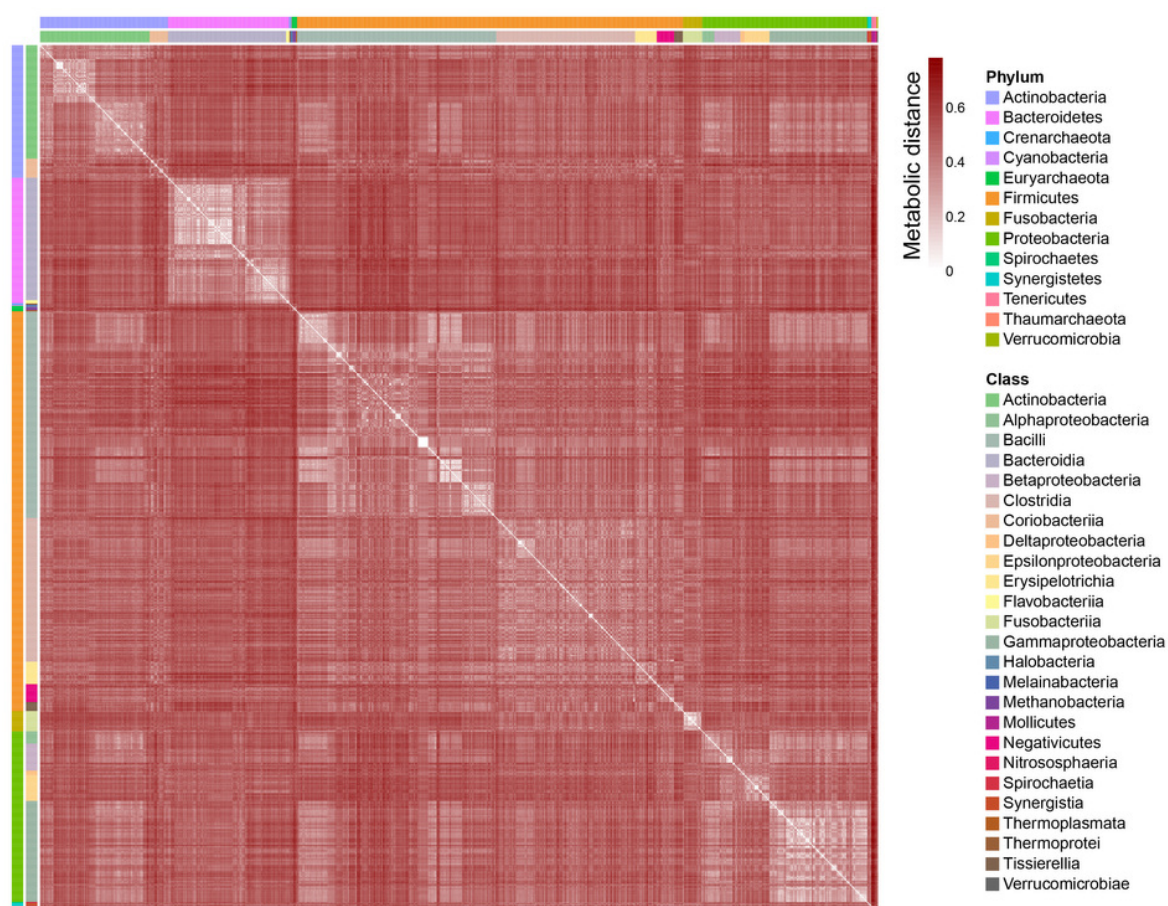
Figure B.4: Metabolic distances between the 773 AGORA reconstructions. Reconstructions with identical reaction content have a metabolic distance of zero, while reconstructions having no overlap have a metabolic distance of 1 (Supplementary Table B.6). Reconstructions are ordered based on phyla and taxonomic classes (Supplementary Table B.4).
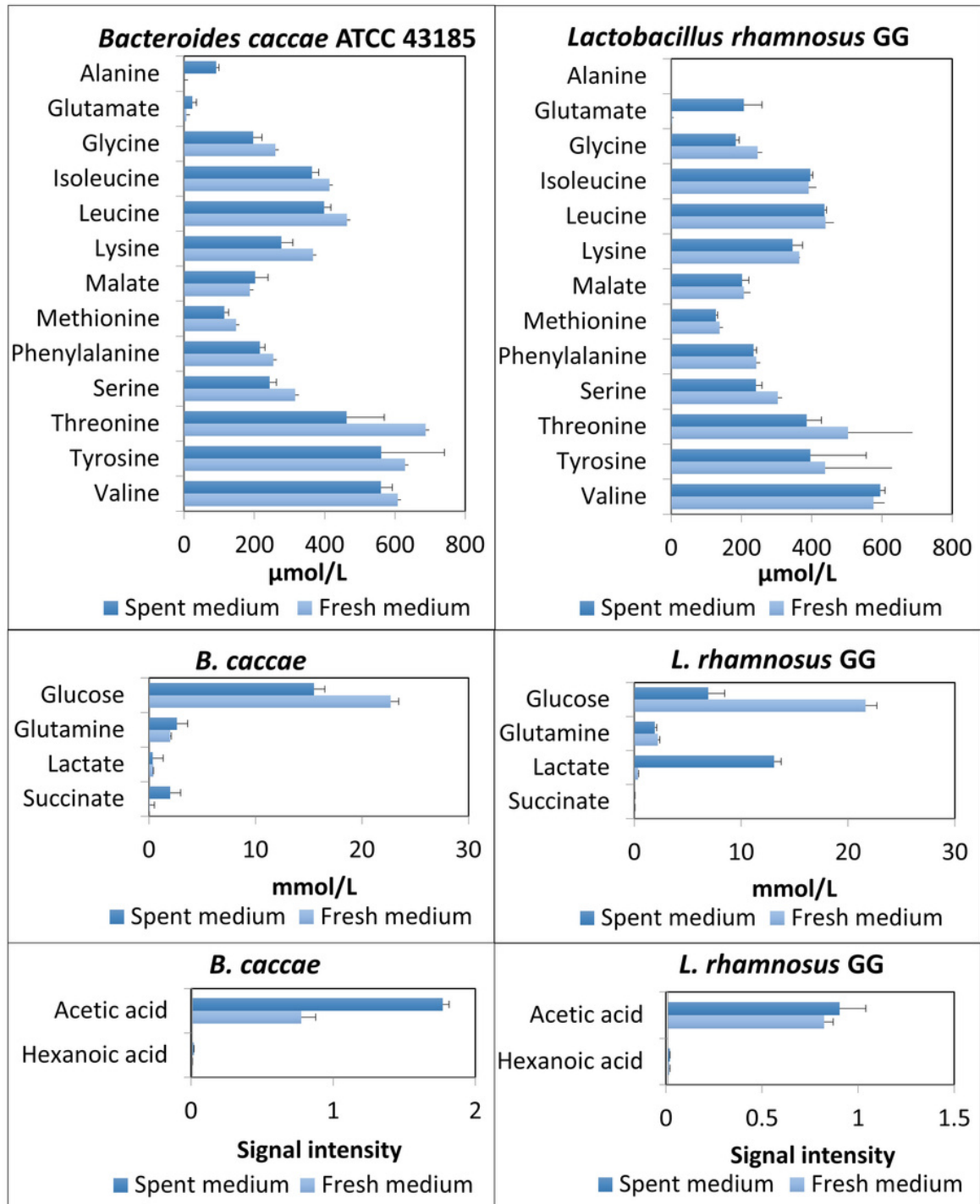
Figure B.5: Metabolomic measurements for two bacterial strains grown *in vitro*. Bar graphs showing the average metabolomic measurements determined for *Bacteroides caccae* ATCC 43185 and *Lactobacillus rhamnosus* GG ATCC 53103 during growth on DMEM 6429 medium supplemented with arabinogalactan (Supplementary Note B.2.3). Error bars show the standard deviation. Statistically significant uptake and secreted is shown and compared with *in silico* predictions in Fig. 3.4a.

Figure B.6

Figure B.6: Clustering of the ratio of pairwise interaction types on the genus level per growth condition. The growth of all AGORA microbe pairs (298,378 pairs) was simulated under four different conditions; using either Western diet or a high fiber diet (Supplementary Table B.18) with and without oxygen. Depending on whether each microbe grew faster or slower in the co-culture simulation compared with by itself under the same condition (Supplementary Tables B.5 and B.6), the interaction between the two microbes was categorized as i) mutualism: both microbes grow faster, ii) commensalism: one microbe grows faster (Taker) while the other's growth rate is not affected (Giver), iii) neutralism: neither microbe's growth rate is affected, iv) amensalism: one microbe grows slower (Affected) while the other's growth rate is not affected (Unaffected), v) parasitism: one microbe grows faster (Taker) while the other grows slower (Giver), and vi) competition: both microbes grow slower. The heatmap shows the ratio of microbes per genus that have each interaction type per condition, e.g., the cell in the first row and first column shows that about 40% of microbes belonging to the *Parvimonas* genus have the interaction type "ParasitismTaker" on Western diet under anaerobic conditions.
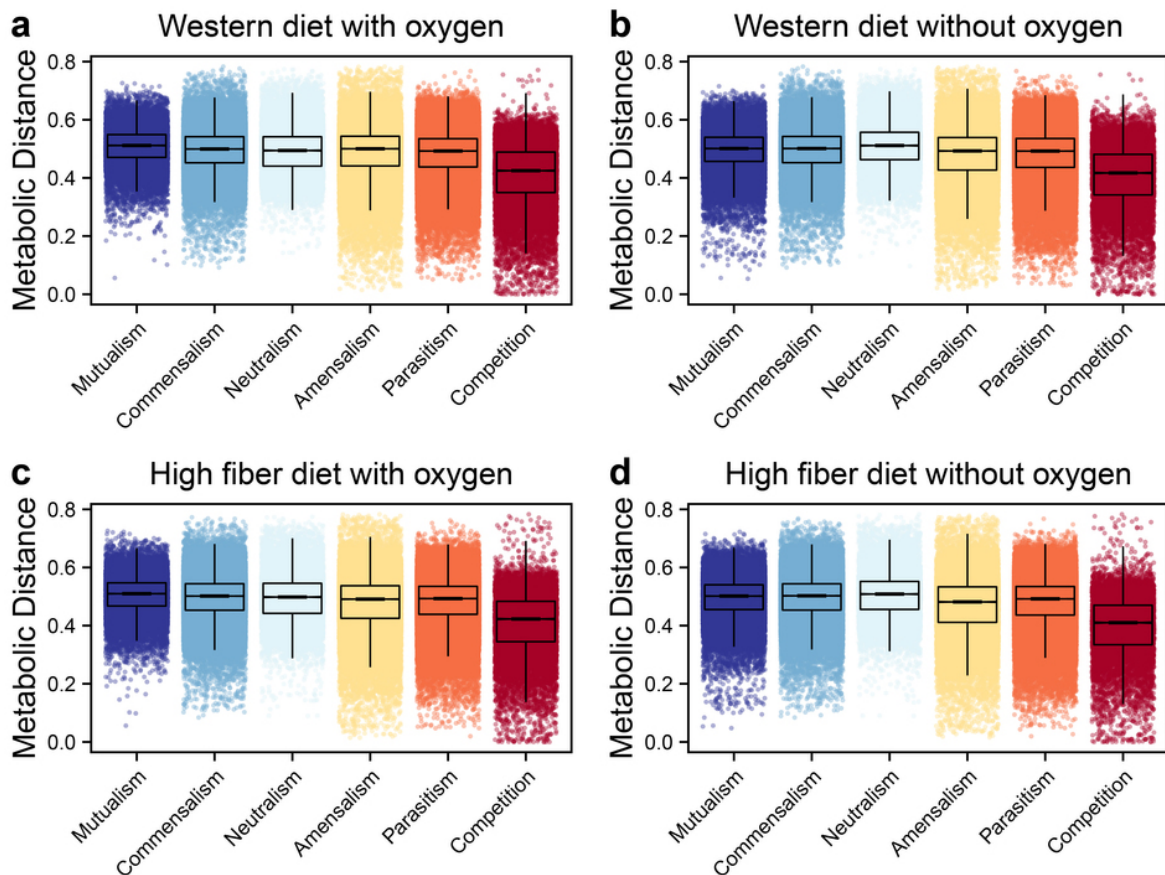
Figure B.7: Metabolic distances plotted by the six types of interactions between the 298,378 microbe-microbe pairs by diet. The growth of all AGORA microbe pairs (298,378 pairs) was simulated under four different conditions; using either Western diet or a high fiber diet (Supplementary Table B.18) with and without oxygen. Depending on whether each microbe grew faster or slower in the co-culture simulation compared with by itself under the same condition (Supplementary Tables B.5 and B.6), the interaction between the two microbes was categorized as: mutualism (both microbes grow faster), commensalism (one microbe grows faster while the other's growth rate is not affected), neutralism (neither microbe's growth rate is affected), amensalism (one microbe grows slower while the other's growth rate is not affected), parasitism (one microbe grows faster while the other grows slower), and competition (both microbes grow slower). Based on the Jaccard index between the reaction content of the reconstructions of the two microbes, the metabolic distance between them was calculated (Online Methods, Supplementary Table B.6). Reconstructions with identical reaction content have a metabolic distance of zero, whereas reconstructions sharing no reactions have a metabolic distance of 1.
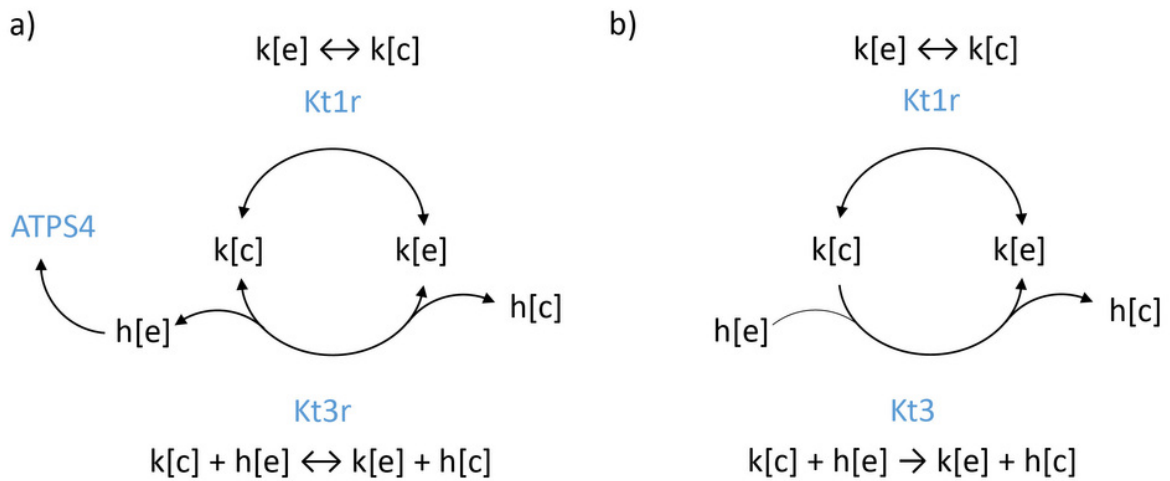
Figure B.8: Example of a typical futile cycle resolved during curation. a) When the ATP demand reaction is optimized, flux in the reverse direction through the potassium uniporter (Kt1r) and the potassium antiporter (Kt3r) leads to an unfeasibly high secretion flux of proton into the extracellular space. This leads to a corresponding flux through the ATP synthase (ATPS4), resulting in unfeasible ATP production. b) The futile cycle is resolved by replacing Kt3r with the irreversible version Kt3. This prevents infeasible ATP production while still allowing potassium transport in both directions.

Table B.13: List of tests that the AGORA reconstructions were subjected to evaluate the curation effort. The number of reconstructions that could be curated and number of reconstructions that passed the tests are shown. Several genomes were missing from the PubSEED platform [12] and could thus not be curated based on the comparative genomic analyses (Supplementary Table B.2). Similarly, a few organisms were not captured in the literature-driven curation on carbon sources and fermentation pathways (Supplementary Table B.1). Less than a third of the AGORA organisms had literature information on essential nutrients (Supplementary Table B.1) and were curated based on the available information.

Test	# of reconstructions that could be curated	# of reconstructions that passed the tests
Reaction and metabolite nomenclature standardized with VMH database	773	773
Reaction constraints standardized with VMH database	773	773
Can grow anaerobically	773	773
No metabolites are produced from nothing	773	773
ATP production rates from the available carbon sources are feasible	773	773
Metabolite formulas are defined and mass-charge balanced	773	773
Number of gap-filling reactions that were included for modeling purposes only is minimized	773	773
In silico growth rates on the defined diets are in realistic ranges	773	773
Gene-protein-reaction associations and reactions in aerobic and anaerobic respiration, B-vitamin biosynthesis pathways, central carbon metabolism, amino acid biosynthesis and/or pyrimidine and purine biosynthesis determined by a comparative genomics approach are implemented	612	612
Carbon source utilization pathways supported by evidence from literature are present and can carry flux	732	732
Fermentation pathways supported by evidence from literature are present and can carry flux	765	765
Species' capabilities/ incapabilities to synthesize essential biomass precursors are captured	244	112 passed all tests, the remaining 132 have one or more false positive or false negative predictions.

Table B.14: Overview of the comparison of the reaction content between AGORA reconstructions and published reconstructions targeting the same strain. Also shown is the comparison between all 773 AGORA reconstructions and 11 published gut microbe reconstructions that were previously used to construct a simplified gut microbe community model [98]. Eight strains overlapped between both reconstruction collections and were compared directly. Shown is the comparison between the eight overlapping strains individually, an between the pooled reactions of all 11 published and all 773 AGORA reconstructions. recon. = reconstruction.

Reconstructed Species	Total reactions in AGORA recon.	Total reactions in published recon.	Overlapping reactions	Reactions only in AGORA recon.	Reactions only in published recon.	Ref. for published recon.
Bacteroides thetaiotaomicron VPI 5482	1,362	1,528	1,058	305	470	[96]
Lactobacillus plantarum WCFS1	1,213	777	395	818	382	[224]
Streptococcus thermophilus LMG 18311	927	556	327	600	229	[165]
Escherichia coli str. K12 substr. MG1655	1,786	2,426	853	934	1,573	[20]
Escherichia coli O157 H7 str. Sakai	1,742	2,372	821	922	1,551	[20]
Helicobacter pylori 26695	1,014	555	310	705	245	[229]
Klebsiella pneumoniae MGH78578	1,801	2,262	843	959	1,419	[133]
Salmonella enterica sv. Typhimurium LT2	1,765	2,623	847	919	1,776	[225]
All reconstructions (11 vs . 773)	3,192	4,608	2,066	1,127	1,540	[20, 96, 133, 165, 224, 229, 225, 95, 71]

Table B.15: Subsystem coverage of reactions that are overlapping between curated reconstructions and published reconstructions targeting the same strain. Also shown are differences in subsystem coverage between all 773 reconstructions and 11 published gut microbe reconstructions. Shown is the comparison between the eight overlapping strains individually, and between the pooled reactions of all 11 published and all 773 AGORA reconstructions. For references for the published reconstructions, refer to Supplementary Table B.14. PPP = pentose phosphate pathway.

		Amino acid metabolism	Biomass	Carbohydrate metabolism	Cell wall/ lipopolysaccharide biosynthesis	Central metabolism	Exchange/demand	Fatty acid metabolism	Nitrogen, sulfur, and iron metabolism	Other	PPP and amino sugars	Polysaccharide degradation	Purine and pyrimidine metabolism	Transport	tRNA charging	Vitamin and cofactor biosynthesis
Bacteroides	Overlapping	151	0	28	118	63	181	118	2	6	46	65	100	73	0	108
thetaiotaomicron	AGORA only	36	1	13	61	22	34	2	1	7	7	3	18	49	1	49
VPI 5482	Published only	47	1	18	59	22	107	13	0	4	12	35	14	104	0	33
Lactobacillus	Overlapping	70	0	14	17	38	70	1	0	5	25	3	67	41	1	43
plantarum WCFS1	AGORA only	82	1	23	193	43	86	119	5	7	16	6	51	116	0	70
	Published only	56	2	15	57	24	44	15	6	7	4	0	31	83	18	20
Streptococcus	Overlapping	69	0	11	11	26	35	0	0	2	13	0	82	41	1	35
thermophilus LMG	AGORA only	77	1	11	135	39	73	82	2	8	8	1	47	69	0	46
18311	Published only	33	1	5	52	16	9	17	2	2	3	0	16	40	18	16
Escherichia coli str	Overlapping	131	0	35	133	89	160	45	5	8	44	1	100	0	2	100
K 12 substr.																
MG1655																

	AGORA only	93	1	14	161	73	59	107	16	11	23	3	58	238	0	77
	Published only	72	2	79	261	65	146	66	37	15	11	1	64	686	21	47
Escherichia coli O157 H7 str. Sakai	Overlapping	123	0	34	131	85	151	45	5	7	38	1	99	0	2	100
	AGORA only	97	1	15	160	72	57	107	16	12	18	4	59	227	0	77
	Published only	72	2	75	254	66	154	66	36	15	12	1	65	665	21	47
Helicobacter pylori 26695	Overlapping	59	0	6	20	19	55	4	0	2	13	0	47	29	2	54
	AGORA only	64	1	1	208	38	63	111	3	11	11	1	39	82	0	73
	Published only	39	1	5	30	34	23	7	2	2	5	0	21	48	0	28
Klebsiella pneumoniae MGH78578	Overlapping	141	0	31	127	83	144	43	4	10	38	1	99	2	2	111
	AGORA only	109	1	15	165	73	70	101	11	19	23	8	50	227	0	83
	Published only	73	1	63	238	52	149	61	12	12	9	1	61	629	21	44
Salmonella enterica sv. Typhimurium LT2	Overlapping	131	0	33	132	73	169	45	5	9	44	1	100	0	2	103
	AGORA only	92	1	14	160	79	39	107	17	11	23	3	59	213	0	102
	Published only	70	3	63	299	73	198	65	59	18	7	1	61	778	21	60
All reconstructions (11 vs. 773)	Overlapping	271	0	66	286	163	332	165	11	20	74	69	149	243	2	214
	AGORA only	130	1	26	85	134	91	42	38	29	30	10	50	192	0	65
	Published only	158	12	102	443	115	252	64	70	40	17	38	83	1,033	24	88

Table B.16: Cell count per ml, optical density (OD), and pH values before and after cell culture of *Bacteroides caccae* ATCC 43185 (*B. caccae*) and *Lactobacillus rhamnosus* GG ATCC 53103 (LGG). All samples were grown on DMEM 6429 medium supplemented with 1% haemin, 3.33% vitamin K, and 9.4 g/L arabinogalactan.

Sample	OD initial (time=0)	OD end-point	Cell count at end-point	Time (hours) of culture	pH before	pH after
B. caccae (1)	0.09	0.49	2.86E+07	31	8.05	NA
B. caccae (2)	0.05	0.56	2.79E+06	27	8.05	7.16
B. caccae (3)	0.09	0.42	3.09E+05	27	8.05	7.27
LGG (1)	0.09	0.55	5.38E+07	31	8.2	5.72
LGG (2)	0.12	0.1	1.27E+05	48	8.2	7.1
LGG (3)	0.1	0.14	1.76E+05	48	8.2	6.64

Table B.17: Uptake rates (mmol/gDW/h) implemented to simulate DMEM 6429 medium.

Metabolite ID	Exchange reaction ID	Metabolite name	Uptake rate
ala_L	EX_ala_L(e)	L-alanine	1
arg_L	EX_arg_L(e)	L-arginine	1
asn_L	EX_asn_L(e)	L-asparagine	1
asp_L	EX_asp_L(e)	L-aspartate	1
cys_L	EX_cys_L(e)	L-cysteine	1
gln_L	EX_gln_L(e)	L-glutamine	1
glu_L	EX_glu_L(e)	L-glutamate	1
gly	EX_gly(e)	Glycine	1
his_L	EX_his_L(e)	L-histidine	1
ile_L	EX_ile_L(e)	L-isoleucine	1
leu_L	EX_leu_L(e)	L-leucine	1
lys_L	EX_lys_L(e)	L-lysine	1
met_L	EX_met_L(e)	L-methionine	1
phe_L	EX_phe_L(e)	L-phenylalanine	1
pro_L	EX_pro_L(e)	L-proline	1
ser_L	EX_ser_L(e)	L-serine	1
thr_L	EX_thr_L(e)	L-threonine	1
trp_L	EX_trp_L(e)	L-tryptophan	1
tyr_L	EX_tyr_L(e)	L-tyrosine	1
val_L	EX_val_L(e)	L-valine	1
glc_D	EX_glc(e)	Glucose	4.5
pyr	EX_pyr(e)	Pyruvate	1
ca2	EX_ca2(e)	Calcium(2+)	1
chol	EX_chol(e)	Choline	1
cl	EX_cl(e)	Chloride	1
cobalt2	EX_cobalt2(e)	Co2+	1
cu2	EX_cu2(e)	Cu2+	1
fe2	EX_fe2(e)	Fe2+	1

fe3	EX_fe3(e)	Fe3+	1
fol	EX_fol(e)	Folate	1
k	EX_k(e)	Potassium	1
h2o	EX_h2o(e)	Water	10
h2s	EX_h2s(e)	Hydrogen sulfide	1
inost	EX_inost(e)	Inositol	1
mg2	EX_mg2(e)	Magnesium	1
mn2	EX_mn2(e)	Manganese	1
ncam	EX_ncam(e)	Nicotinamide	1
pi	EX_pi(e)	Hydrogenphosphate	10
pnto_R	EX_pnto_R(e)	Pantothenate	1
pydxn	EX_pydxn(e)	Pyridoxine	1
ribflv	EX_ribflv(e)	Riboflavin	1
so4	EX_so4(e)	Sulfate	1
thm	EX_thm(e)	Thiamin	1
zn2	EX_zn2(e)	Zinc	1
mqn7	EX_mqn7(e)	Menaquinone 7	1
mqn8	EX_mqn8(e)	Menaquinone 8	1
pheme	EX_pheme(e)	Protoheme	1
arabinogal	EX_arabinogal(e)	Larch arabinogalactan	0.0094
q8	EX_q8(e)	Ubiquinone-8	1
sheme	EX_sheme(e)	Siroheme	1

Table B.18: Uptake rates (mmol/gDW/h) for dietary compounds implemented to simulate Western and high fiber diet.

Metabolite ID	Exchange reaction ID	Metabolite name	Western diet	High fiber diet
arab_L	EX_arab_L(e)	L-arabinose	0.17878295	0.04736842
cellb	EX_cellb(e)	Cellobiose	0.07449289	0.01973684
drib	EX_drib(e)	2-deoxy-D-ribose	0.17878295	0.04736842
fru	EX_fru(e)	D-Fructose	0.14898579	0.03947368
fuc_L	EX_fuc_L(e)	L-fucose	0.14898579	0.03947368
gal	EX_gal(e)	D-Galactose	0.14898579	0.03947368
glc_D	EX_glc(e)	D-glucose	0.14898579	0.03947368
glcn	EX_glcn(e)	D-gluconate	0.14898579	0.03947368
lcts	EX_lcts(e)	Lactose	0.07449289	0.01973684
malt	EX_malt(e)	Maltose	0.07449289	0.01973684
man	EX_man(e)	D-Mannose	0.14898579	0.03947368
melib	EX_melib(e)	Melibiose	0.07449289	0.01973684
mnl	EX_mnl(e)	D-Mannitol	0.14898579	0.03947368
oxa	EX_oxa(e)	Oxalate(2-)	0.44695737	0.11842105
rib_D	EX_rib_D(e)	D-ribose	0.17878295	0.04736842
rmn	EX_rmn(e)	L-Rhamnose	0.14898579	0.03947368
sucr	EX_sucr(e)	Sucrose	0.07449289	0.01973684
tre	EX_tre(e)	Trehalose	0.07449289	0.01973684

xyl_D	EX_xyl_D(e)	D-xylose	0.17878295	0.04736842
strch1	EX_strch1(e)	Starch	0.25733909	0.06818182
amylopect900	EX_amylopect900(e)	Amylopectin	0.00001567	0.00034722
amylose300	EX_amylose300(e)	Amylose	0.00004702	0.00104167
arabinan101	EX_arabinan101(e)	Arabinan	0.00016628	0.00368369
arabinogal	EX_arabinogal(e)	Larch arabinogalactan	0.00002191	0.00048550
arabinoxyl	EX_arabinoxyl(e)	Arabinoxylan	0.00030665	0.00679348
bglc	EX_bglc(e)	Beta-glucan	0.00000007	0.00000156
cellul	EX_cellul(e)	Cellulose	0.00002821	0.00062500
dextran40	EX_dextran40(e)	Dextran 40, 1,6-alpha-DGlucan	0.00017632	0.00390625
galmannan	EX_galmannan(e)	Carob galactomannan	0.00001411	0.00031250
glcmannan	EX_glcmannan(e)	Konjac glucomannan	0.00003288	0.00072844
homogal	EX_homogal(e)	Homogalacturonan	0.00012823	0.00284091
inulin	EX_inulin(e)	Chicory inuli n 0.00047019		0.01041667
kestopt	EX_kestopt(e)	Kestopentaose	0.00282117	0.06250000
levan1000	EX_levan1000(e)	Levan, 1000 fructose units	0.00001411	0.00031250
lichn	EX_lichn(e)	Lichenin from Icelandic moss	0.00008298	0.00183824
lmn30	EX_lmn30(e)	Laminarin	0.00047019	0.01041667
pect	EX_pect(e)	Pectin	0.00003339	0.00073964
pullulan1200	EX_pullulan1200(e)	Pullulan	0.00001175	0.00026042
raffin	EX_raffin(e) Raffinose	0.00470194		0.10416667
rhamnogalurI	EX_rhamnogalurI(e)	Potato rhamnogalacturonan I	0.00001449	0.00032106
rhamnogalurII	EX_rhamnogalurII(e)	Wine rhamnogalacturonan II	0.00026699	0.00591483
starch1200	EX_starch1200(e)	Resistant starch	0.00001175	0.00026042
xylan	EX_xylan(e)	Oat spelt xylan	0.00003206	0.00071023
xyluglc	EX_xyluglc(e)	Xyluglucan	0.00001315	0.00029124
arachd	EX_arachd(e)	Arachidonate	0.00332813	0.00166406
chsterol	EX_chsterol(e)	Cholesterol	0.00495795	0.00247898
glyc	EX_glyc(e)	Glycerol	1.79965486	0.89982743
hdca	EX_hdca(e)	Hexadecanoate (n-C16:0)	0.39637090	0.19818545
hdcea	EX_hdcea(e)	Hexadecenoate (n-C16:1)	0.03651697	0.01825848
lnlc	EX_lnlc(e)	Linoleate	0.35910921	0.17955461
lnlnca	EX_lnlnca(e)	Alpha-linolenate	0.01756512	0.00878256
lnlncg	EX_lnlncg(e)	Gamma-linolenate	0.01756512	0.00878256
ocdca	EX_ocdca(e)	Octadecanoate (n-C18:0)	0.16928260	0.08464130

ocdcea	EX_ocdcea(e)	Octadecenoate (n-C18:1)	0.68144465	0.34072233
octa	EX_octa(e)	Octanoate (n-C8:0)	0.01294272	0.00647136
ttdca	EX_ttdca(e)	Tetradecanoate (n-C14:0)	0.06867567	0.03433784
ala_L	EX_ala_L(e)	L-alanine	1	1
arg_L	EX_arg_L(e)	L-arginine	0.15	0.15
asn_L	EX_asn_L(e)	L-asparagine	0.225	0.225
asp_L	EX_asp_L(e)	L-aspartate	0.225	0.225
cys_L	EX_cys_L(e)	L-cysteine	1	1
gln_L	EX_gln_L(e)	L-glutamine	0.18	0.18
glu_L	EX_glu_L(e)	L-glutamate	0.18	0.18
gly	EX_gly(e)	Glycine	0.45	0.45
his_L	EX_his_L(e)	L-histidine	0.15	0.15
ile_L	EX_ile_L(e)	L-isoleucine	0.15	0.15
leu_L	EX_leu_L(e)	L-leucine	0.15	0.15
lys_L	EX_lys_L(e)	L-lysine	0.15	0.15
met_L	EX_met_L(e)	L-methionine	0.18	0.18
phe_L	EX_phe_L(e)	L-phenylalanine	1	1
pro_L	EX_pro_L(e)	L-proline	0.18	0.18
ser_L	EX_ser_L(e)	L-serine	1	1
thr_L	EX_thr_L(e)	L-threonine	0.225	0.225
trp_L	EX_trp_L(e)	L-tryptophan	0.08181818	0.08181818
tyr_L	EX_tyr_L(e)	L-tyrosine	1	1
val_L	EX_val_L(e)	L-valine	0.18	0.18
12dgr180	EX_12dgr180(e)	1,2-Diacyl-sn-glycerol (dioctadecanoyl, n-C18:0)	1	1
26dap_M	EX_26dap_M(e)	meso-2,6-Diaminoheptanedioate	1	1
2dmmq8	EX_2dmmq8(e)	2-Demethylmenaquinone 8	1	1
2obut	EX_2obut(e)	2-Oxobutanoate	1	1
3mop	EX_3mop(e)	3-methyl-2-oxopentanoate	1	1
4abz	EX_4abz(e)	4-Aminobenzoate	1	1
4hbz	EX_4hbz(e)	4-hydroxybenzoate	1	1
ac	EX_ac(e)	Acetate	1	1
acgam	EX_acgam(e)	N-acetyl-D-glucosamine	1	1
acmana	EX_acmana(e)	N-acetyl-D-mannosamine	1	1
acnam	EX_acnam(e)	N-acetylneuraminate	1	1
ade	EX_ade(e)	Adenine	1	1
adn	EX_adn(e)	Adenosine	1	1

adocbl	EX_adocbl(e)	Adenosylcobalamin	1	1
ala_D	EX_ala_D(e)	D-alanine	1	1
amp	EX_amp(e)	AMP	1	1
arab_D	EX_arab_D(e)	D-Arabinose	1	1
btn	EX_btn(e)	Biotin	1	1
ca2	EX_ca2(e)	Calcium(2+)	1	1
cbl1	EX_cbl1(e)	Cob(I)alamin	1	1
cgly	EX_cgly(e)	L-cysteinylglycine	1	1
chol	EX_chol(e)	Choline	1	1
chor	EX_chor(e)	Chorismate	1	1
cit	EX_cit(e)	Citrate	1	1
cl	EX_cl(e)	Chloride	1	1
cobalt2	EX_cobalt2(e)	Co2+	1	1
csn	EX_csn(e)	Cytosine	1	1
cu2	EX_cu2(e)	Cu2+	1	1
dad_2	EX_dad_2(e)	2-deoxyadenosine	1	1
dcyt	EX_dcyt(e)	Deoxycytidine	1	1
ddca	EX_ddca(e)	Laurate	1	1
dgsn	EX_dgsn(e)	Deoxyguanosine	1	1
fe2	EX_fe2(e)	Fe2+	1	1
fe3	EX_fe3(e)	Fe3+	1	1
fe3dcit	EX_fe3dcit(e)	Fe(III)dicitrate	1	1
fald	EX_fald(e)	Formaldehyde	1	1
fol	EX_fol(e)	Folate	1	1
for	EX_for(e)	Formate	1	1
fum	EX_fum(e)	Fumarate	1	1
gam	EX_gam(e)	D-Glucosamine	1	1
glu_D	EX_glu_D(e)	D-Glutamate	1	1
glyc3p	EX_glyc3p(e)	Glycerol 3-phosphate	1	1
gthox	EX_gthox(e)	Oxidized glutathione	1	1
gthrd	EX_gthrd(e)	Reduced glutathione	1	1
gua	EX_gua(e)	Guanine	1	1
h	EX_h(e)	Proton	1	1
H2	EX_h2(e)	Hydrogen	1	1
h2o	EX_h2o(e)	Water 10 10		
h2s	EX_h2s(e)	Hydrogen sulfide	1	1
hxan	EX_hxan(e)	Hypoxanthine	1	1
indole	EX_indole(e)	Indole	1	1
k	EX_k(e)	Potassium	1	1
lanost	EX_lanost(e)	lanosterol	1	1
meoh	EX_meoh(e)	Methanol	10	10
metsox_S_L	EX_metsox_S_L(e)	L-Methionine Sulfox-ide	1	1
mg2	EX_mg2(e)	Magnesium	1	1
mn2	EX_mn2(e)	Mn2+	1	1
mobd	EX_mobd(e)	Molybdate	1	1

mqn7	EX_mqn7(e)	Menaquinone 7	1	1
mqn8	EX_mqn8(e)	Menaquinone 8	1	1
na1	EX_na1(e)	Sodium	1	1
nac	EX_nac(e)	Nicotinate	1	1
ncam	EX_ncam(e)	Nicotinamide	1	1
nmn	EX_nmn(e)	NMN	1	1
no2	EX_no2(e)	Nitrite	1	1
no3	EX_no3(e)	Nitrate	1	1
orn	EX_orn(e)	Ornithine	1	1
pheme	EX_pheme(e)	Protoheme	1	1
pi	EX_pi(e)	Hydrogenphosphate	1	1
pime	EX_pime(e)	Pimelate	1	1
pnto_R	EX_pnto_R(e)	(R)-Pantothenate	1	1
ptrc	EX_ptrc(e)	Putrescine	1	1
pydam	EX_pydam(e)	Pyridoxamine	1	1
pydx	EX_pydx(e)	Pyridoxal	1	1
pydx5p	EX_pydx5p(e) P	yridoxal 5-phosphate	1	1
pydxn	EX_pydxn(e)	Pyridoxine	1	1
q8	EX_q8(e)	Ubiquinone-8	1	1
ribflv	EX_ribflv(e)	Riboflavin	1	1
sel	EX_sel(e)	Selenate	1	1
sheme	EX_sheme(e)	Siroheme	1	1
so4	EX_so4(e)	Sulfate	1	1
spmd	EX_spmd(e)	Spermidine	1	1
thm	EX_thm(e)	Thiamin	1	1
thymd	EX_thymd(e)	Thymidine	1	1
ura	EX_ura(e)	Uracil	1	1
uri	EX_uri(e)	Uridine	1	1
xan	EX_xan(e)	Xanthine	1	1
zn2	EX_zn2(e)	Zinc	1	1

Table B.19: Reactions associated with functional roles of the eight different B-vitamin biosynthesis pathways. The functional roles involved in the eight B-vitamin biosynthesis pathways are the same as described in a study by Magnusdottir et al. [140].

B-vitamin	Functional role	VMH Reaction(s)
Biotin	BioH / BioG	PMACPME
Biotin	BioW	EX_pime(e), PIMEtr
Biotin	BioC	ACS, ACCOAC, MACPMT, MAL-COACD, 3OAACPR1, 3HACPR1, EACPR1, GACPCD, 3OAACPR2, 3HACPR2, EACPR2
Biotin	BioF	AOXSr2
Biotin	BioA	AMAOTr
Biotin	BioD	DBTS
Biotin	BioB	BTS4

Cobalamin	CysG / CbiKX	SHCHCC
Cobalamin	CbiL	CPC2MT
Cobalamin	CbiG	CPC3MT
Cobalamin	CbiF	CPC4MT
Cobalamin	CobF	CPC5MT
Cobalamin	CbiJ	CPC6R
Cobalamin	CbiE / CbiT	CPC6MT
Cobalamin	CbiC	CPC8MM
Cobalamin	CbiA	CYRDAS
Cobalamin	CobAT	CYRDAR, CYRDAAT
Cobalamin	CbiP	ADCYRS
Cobalamin	CbiB	ADCPS2
Cobalamin	CobU	ACBIPGT
Cobalamin	CobS	ADOCBLS
Folate	FolE1 / FolE2	GTPCI
Folate	folQ2 / folQ3	DNTPPA
Folate	FolB / ptpS-III	DHNPA2
Folate	FolK	HPPK2
Folate	FolP	DHPS2
Folate	pabAc / pabAb	ADCS
Folate	pabAa	ADCL
Folate	FolCDHFS	DHFS
Folate	FolCFPGS / FolC2	FPGS
Folate	Dhfr0 / Dhfr1 / Dhfr2	DHFR
Niacin	ASPOX	ASPO7
Niacin	ASPDH	ASPO2
Niacin	QSYN	QULNS
Niacin	QAPRT	NNDPR
Niacin	NaMNAT_D	NNATr
Niacin	NADS	NADS1
Niacin	NMNS	R0527
Niacin	NMNAT / NMNAT_R / NMNAT_M	NMNAT
Niacin	NADK	NADK
Pantothenate	ASPDC	ASP1DC
Pantothenate	KPHMT	MOHMT
Pantothenate	KPRED / KARED	DPR
Pantothenate	PBAL	PANTS
Pantothenate	PANK / PANK2 / PANK3	PNTK
Pantothenate	PPCS	PPNCL3
Pantothenate	PPCDC	PPCDC

Pantothenate	PPAT	PTPAT
Pantothenate	DPCK	DPCOAK
Pyridoxine	PdxT / PdxS	PLPS
Pyridoxine	Dxs	DXPS
Pyridoxine	gapA	E4PD
Pyridoxine	PdxB	PERD
Pyridoxine	PdxF / PdxA	OHPBAT
Pyridoxine	PdxJ	PDX5PS
Pyridoxine	PdxH	PDX5PO2, PYAM5POr
Riboflavin	GTPCH2	GTPCII2
Riboflavin	PyrD / PyrD-a	DHPPDA2
Riboflavin	PyrR	APRAUR
Riboflavin	PyrP	PMDPHT
Riboflavin	DHBPS	DB4PS
Riboflavin	DMRLS	RBFSa
Riboflavin	RSAe / RSAalpha	RBFSb
Riboflavin	RK	RBFK
Riboflavin	FMNAT	FMNAT
Thiamin	ThiG	THZPSN
Thiamin	ThiC	AMPMS2
Thiamin	ThiD / ThiD_alt	PMPK, HMPK1
Thiamin	TMP-Pase(ThiE)	TMPPP

Table B.20: GC-MS dwell times and quantification and qualification ions (m/z) for the measured short-chain fatty acids and amino acids.

Derivatives	Quant-Ion (m/z)	Qual-Ion (m/z)	Qual-Ion (m/z)	Dwell Time (ms)
Short-chain fatty acids				
Formic acid 1TBDMS	103.0	75.0	99.0	20
Acetic acid 1TBDMS	117.0	75.0	99.0	20
Butyric acid 1TBDMS	145.1	75.0	115.1	20
Isobutyric acid 1TBDMS	145.1	75.0	115.1	20
Valeric acid 1TBDMS	159.1	75.0	201.1	20
IS 2-Ethylbutyric acid 1TBDMS	173.1	115.1	99.0	20
4-Methylvaleric acid 1TBDMS	173.1	215.1	99.0	20
Hexanoic acid 1TBDMS	173.1	75.0	131.0	20
Polar metabolites				
Alanine 2TBDMS	260.2	158.1	232.1	70
Glycine 2TBDMS	246.1	189.1	218.1	50
Valine 2TBDMS	288.6	186.1	260.2	70
Leucine 2TBDMS	302.2	200.0	274.2	70
Isoleucine 2TBDMS	302.2	200.2	274.2	70
Threonine 2TBDMS	290.2	159.1	303.2	50
Proline 2TBDMS	286.2	184.1	258.2	50
Succinic acid 2TBDMS	289.1	215.1	331.2	50
IS Glutaric acid-D6 2TBDMS	309.2	235.2	351.3	70
Serine 3TBDMS	390.2	302.2	362.2	50
Threonine 3TBDMS	404.2	376.3	417.3	70
Methionine 2TBDMS	320.2	218.1	292.2	50
Malic acid 3TBDMS	419.2	287.1	403.2	50
Phenylalanine 2TBDMS	336.2	308.2	combined	50
Aspartic acid 3TBDMS	418.2	316.2	(coeluting)	
Ornithine 3TBDMS	474.4	184.1	286.2	50
Glutamic acid_3TBDMS	432.3	330.2	404.3	50
Lysine 3TBDMS	431.3	300.2	488.4	70
Tyrosine 3TBDMS	466.3	302.2	438.3	50
Histidine 3TBDMS	440.3	338.3	196.1	70