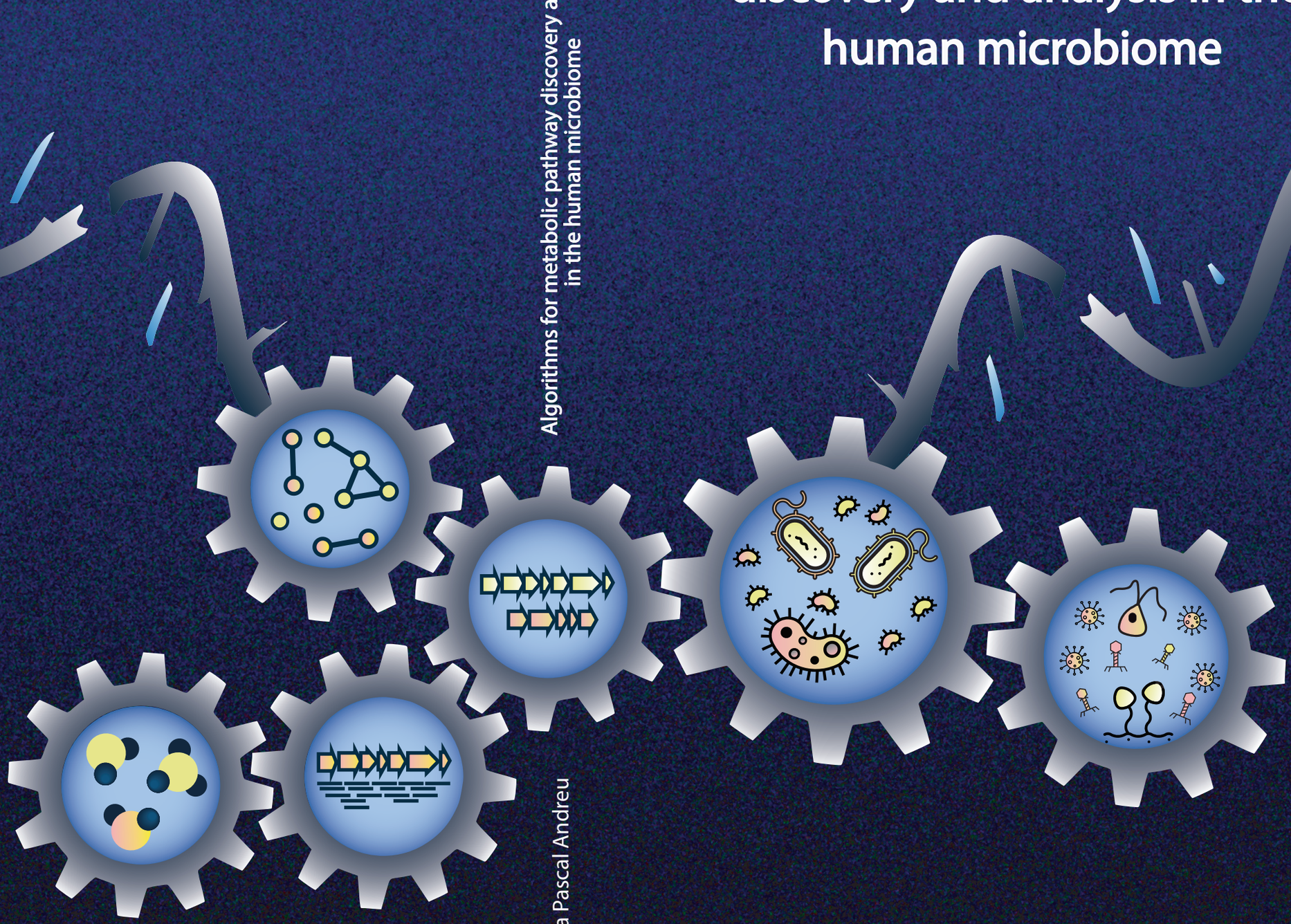
Algorithms for metabolic pathway discovery and analysis in the human microbiome

Victòria Pascal Andreu



Algorithms for metabolic pathway discovery and analysis in the human microbiome

Victòria Pascal Andreu

Propositions

1. Gene-centric approaches are better suited to accurately assess microbial functional potential than taxonomy-based ones.
   (this thesis)

2. Ignoring strain-level variation when profiling a microbial community implies having an incomplete snapshot of their overall metabolic potential.
   (this thesis)

3. Scientific discoveries are only fully worth the effort when they are effectively communicated to the general public.

4. Online participation in scientific meetings should always be facilitated, regardless of the circumstances.

5. Losing a language implies losing traditions and unique ways of understanding the world.

6. One of the best long-term contributions to solving major societal problems is to educate in critical thinking.

Propositions belonging to the thesis, entitled:

Algorithms for metabolic pathway discovery and analysis in the human microbiome

Victòria Pascal Andreu
Wageningen, 13th of September 2021

Algorithms for metabolic pathway discovery and analysis in the human microbiome

Victòria Pascal Andreu

Thesis committee

Promotor
Prof. Dr D. de Ridder
Professor of Bioinformatics
Wageningen University & Research

Co-promotors
Dr M.H. Medema
Assistant Professor, Bioinformatics
Wageningen University & Research

Dr M.A. Fischbach
Assistant Professor, Bioengineering
Stanford University, Stanford [CA], United States of America

Other members
Prof. Dr JM. Wells, Wageningen University & Research
Dr M.G.J. de Vos, University of Groningen
Dr M. Suárez Diez, Wageningen University & Research
Prof. Dr N. Segata University of Trento, Italy

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences

Algorithms for metabolic pathway discovery and analysis in the human microbiome

Victòria Pascal Andreu

Thesis

Submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 13 September 2021
at 4 p.m. in the Aula.

Victòria Pascal Andreu

Algorithms for metabolic pathway discovery and analysis in the human microbiome, 183 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2021).
With references, with summaries in English, Catalan and Spanish.

DOI: https://doi.org/10.18174/548302

ISBN: 978-94-6395-854-7

Table of contents

Chapter 1

General introduction

1.1 History of Microbiology and Bioinformatics

The discovery of the microscope had a great impact back in the 17th century. With a primitive version of this instrument, in 1665 Robert Hooke discovered the first microorganisms, at that time referred as 'microsco-pical mushrooms' and nowadays known as mold. A few years later, Antonie van Leeuwenhoek described microscopic protozoa and bacteria using magnifying lens that he built himself. He also observed microbial aggregates on the surface of teeth. These aggregates are microbial communities that live in symbiosis and interact with each other for their own benefit and for the community. These discoveries gave Hook and Van Leeuwenhoek the status of fathers of microbiology[1]. Despite revealing the microbial universe, follow-up investigations did not resume until more than a century later, when the importance of microorganisms in human and animal health was acknowledged[2]. Louis Pasteur and Robert Koch, two well-known names in medical microbiology in the 19th century, determined the role of bacteria in disease (pathogens) and demonstrated that specific pathogens are the causal agents of certain diseases. Moreover, Pasteur hypothesized that non-pathogenic microorganisms had an important role in our physiology. Elie Metchnikoff, another avowed microbiologist, hypothesized that bacteria and their interactions were important mediators of host health, important landmarks later in the human microbiome field. His beliefs were confirmed years later, when more advanced technology became available[3]. From that moment on, scientists began to develop laboratory techniques to study microbial diversity, their distribution across different habitats and their specific capabilities.

In 1921, Alexander Fleming by chance discovered the first drug, an antibiotic known as penicillin, produced by fungal colonies, which was later used to prevent and treat several bacterial infections[4]. During the 20th century, also known as the era of Molecular Biology, great advances were made to better understand the genetic code and its regula-

tion. These advances were in part possible due to the relative simplicity of microorganisms[5]. These were the first milestones in the era of the Human Gut Microbiome Research field, but it was not until 1950, when methods for culturing anaerobes became available, that the field really took off. From this moment on, the topic attracted increasing attention and researchers started to investigate which microorganisms were inhabiting our body and started researching their functional roles as well. However, it soon became obvious that the number of bacteria that can be grown using traditional culturing techniques was very limited[6]. Yet, when sequencing-based approaches came into play, around fifty years ago, together with other advances in molecular biology, it became possible to survey a vast number of otherwise "hidden" bacteria and obtain a more realistic snapshot of the communities that inhabit our bodies[7]. All these advances boosted the interest in comparing the bacterial communities of individuals that showed different phenotypes and to pinpoint potential candidates associated with specific host traits.

Simultaneously with the latest biotechnological advances, using computational tools has become a must to analyse all these data, a process otherwise very laborious and time-consuming. Bioinformatics emerged in the early 1960s with Margaret Dayhoff, one of the pioneers to apply computers on protein sequences analysis. It was not until later (Roger Staden, 1979) when the first software to analyse genomic sequences was published[8]. Another major advance in the mid-80s was the emergence of new and more intuitive scripting languages that simplified the task of coding new algorithms. Another important achievement was made during the 1980s, when joint forces helped creating standardized databases that unified information in specific formats that promoted, among others, reuse of data. It was not until the 90s when the internet became a reality when the field blasted. With time, computers became more sophisticated and other advances in biology such as sequencing the complete genome of *Haemophilus influenzae* and later on the assembly of the human genome allowed to gradually solve

more complex questions. During the past years, bioinformatics has become increasingly present in science, and while sequencing technologies kept improving and have greatly decreased in cost, data has started to mount exponentially, entering the era of big data. Currently, bioinformatics is key to processing biological information and its use is almost indispensable in every research topic.

Nowadays, to study microbial communities, transcriptomics, proteomics and metabolomics are increasing in popularity and are becoming great tools to shift the focus from *"who is there?"* to *"what are they doing?"*. This is propelling the understanding of the mechanisms of action of these communities in health and disease and enables to translate this knowledge into clinical interventions[9]. However, microbial bioinformatics still faces challenges: for instance, to design tools that properly handle the complexity of data and the integration of data from different sources[10].

1.2 The human microbiota: key players in our health

The human body harbors trillions of microorganisms that constitute a very complex and symbiotic microbial community also known as the human microbiota. Different body sites have been sampled and taxonomically characterized, revealing that differences in the composition of bacterial communities across anatomical regions are larger than inter-individual differences[11]. These body sites include the gastrointestinal tract, the nasal passages, the oral cavity, the skin and the urogenital tract. Some of these habitats comprise larger bacterial collections than others in terms of abundance and diversity: the gut for instance is the primary source of study, given the taxonomically diverse and dense populations that reside there. One of the reasons for studying the human microbiome is the fact that their gene repertoire vastly exceeds that of a human (by ~150 fold[12]), making it more flexible to adapt to changes and complement human enzyme functions that are involved in health and disease[13,14]. However, the human microbiome is not a stable

ecosystem but evolves throughout a human lifespan and changes occur according to the lifestyle, genetics, environmental conditions and diet of a person[15,13]. The latter is considered one of the most impactful community drivers, which determines the taxonomic and functional composition of the microbiome, and thus directly influences host homeostasis[16,17]. For instance, high-carbohydrate diets, which promote the growth of bacteria from the genus *Prevotella*, have been associated with several metabolic disorders such as type 2 diabetes, cardiovascular disease and inflammatory bowel disease (IBD), which can be countered by increasing fiber intake[18]. Thus, diet can shape microbial populations; what is more important, it can impact the pool of microbially derived metabolites, the mediators in the crosstalk between microbes and the host. Examples of metabolites synthesized in response to certain dietary components are short chain fatty acids (SCFA), bile acids or methylamines. These have fundamental roles in regulating and controlling important processes, such as triggering immune responses, serving as local energy sources and providing electron sources for secondary fermenters[18]. Another phenotype known to be promoted by bacteria is carcinogenesis. Disturbances in gut microbial populations can increase tissue permeability, which reinforces bacterial translocation and the production of inflammatory and DNA-damaging metabolites; this ultimately leads to a higher risk of suffering gastrointestinal cancer, for instance[17].

However, despite the fact that some of the mechanisms implicated in health and disease are known, the metabolism of anaerobes is still largely a black box[19]. Therefore, elucidating the molecular pathways by which bacteria interact with human cells or other microbes is crucial to obtain deeper insights into the microbial metabolism and be in a better position to engineer novel therapies to cure and prevent chronic diseases[20]. The current knowledge we have on host-microbe interactions, the constantly growing numbers of new datasets and the delivery of new technologies and tools to analyze data together pave the way to systematically characterize the molecular mechanisms behind disease states.

1.3 Targeting the microbiome as a treatment for some diseases

The tight and reciprocal relationship between the human microbiome and its host and the mounting evidence that bacteria, especially from the gut, are modulators of health and disease, open up new opportunities to prevent and treat chronic diseases[20]. There are effective approaches to restore dysbiotic microbial ecosystems that either aim to reduce overabundant bacteria, shift the community towards a more desired/beneficial one or replace the perturbed ecosystem with the community from a healthy donor. These strategies include the administration of prebiotics or probiotics, modulation of the microbiota via diet intervention, fecal microbiota transplantation (FMT), or selective depletion of specific bacteria by using antibiotic treatments[21]. Hence, good strategies exist to modulate the gut communities and are showing very promising results to ameliorate certain pathologies[22,23]. However, these are most of the time multifactorial diseases and highly variable between individuals, which makes it even more important to understand the specific mechanisms these targeted bacteria employ to influence health and disease. Understanding the metabolic role these bacteria exert in the gut could also help grasping why certain individuals are not responsive to some interventions. Thus, if we can identify which metabolic functions play a role in the onset or development of a pathology, we will be in a better position to use the gut microbiome for therapeutic means and more precisely determine the metabolic needs of the ecosystem to be restored.

1.4 Methods to study the microbiome: taxonomic profiling vs functional profiling

A decade ago, 16S rRNA sequencing began to be very popular to study microbial communities, due to the gradual drop in cost and the emergence of new methods to analyze such data. New bioinformatic tools were designed[24,25] in combination with reference databases[26,27] such as greengenes[28], SILVA[29] or PATRIC[30], which allowed obtaining deep-

er insights into the abundances of Operational Taxonomic Units (OTUs). These large databases of marker genes and taxonomy are used as references to putatively cluster and assign taxonomy to targeted amplicon data[31,32]. Among the most popular tools to taxonomically profile microbiomes are QIIME[25], Mothur[24] and VAMPS[33], pipelines that provide different clustering algorithms, reference databases and statistical analysis to further visualize the results.

Using this type of data, different studies elucidated links between specific taxa and human phenotypic traits[34], which cover a wide range of diseases as for instance IBD[35,36], obesity[37,38], Autism Spectrum Disorders[39,40] and Parkinson disease[41,42]. As an example, the 16S rRNA sequence analysis of stool from diabetes-prone and diabetes-resistant rats allowed correlating a higher abundance of *Lactobacillus* and *Bifidobacterium* with the onset of the disease[43]. However, the underlying molecular mechanisms by which these bacteria impact host health remains unknown based on 16S sequencing data alone. Moreover, microbial composition is known to vary more strongly between individuals in comparison to metabolic functions[37]. This is mainly due to functional redundancy, which helps the whole microbial community be more resilient in case of perturbation[44]. Another limitation of such methodologies is that in most of the cases, they are unable to resolve differences at the strain level and miss horizontal gene transfers events that commonly occur in these ecosystems[44]. Therefore, using taxonomy alone is generally not a good proxy to understand molecular mechanisms, nor to provide robust hypotheses regarding causation. Therefore, functional omics data such as metatranscriptomic, metaproteomic and metabolomics are required to provide more accurate information and are more suitable to study microbiomes from a functional point of view[45] with the ultimate goal of using the human microbiome as therapeutic target.

1.5 Microbiome-derived small molecules: types and functions

The human microbiome is a good source of natural products and other types of small molecules that can profoundly affect human homeostasis. In the last decade, several studies have elucidated and characterized the presence of certain metabolites at relevant concentrations, which, given their role in microbe-microbe and microbe-host interactions, can be used as biological markers[46]. The molecules that are known to date can be classified into two major groups based on their function; the ones that are involved in energy metabolism or primary metabolism and the ones used for protection against biotic and abiotic stresses[47] that derive from bacterial secondary metabolism. Examples of specialized primary metabolites are indole and trimethylamine (TMA). Indole, which is derived from tryptophan and is the precursor of indoxyl sulfate, has been associated with a decrease in bacterial pathogenicity when found at high concentrations[48]. TMA in contrast, is synthesized from carnitine or choline and is a marker for cardiovascular diseases[49]. The gut microbiome is also a good resource of structurally different natural products, such as, for instance, type II polyketides found in the oral, gut and skin human microbiome. The characterization of two of these BGCs allowed to elucidate the structure of five of these molecules, that have antibacterial activity against gram positive bacteria[50]. Another example is lactocillin, a thiopeptide antibiotic synthesized in the vagina by *Lactobacillus gasseri*, which inhibits the growth of some Gram-positive vaginal pathogens[51]. Oftentimes, the genes responsible for the synthesis of these molecules are clustered together in the genome, in loci also known as metabolic gene clusters (MGCs) or more specifically, biosynthetic gene clusters (BGCs) when they are responsible for secondary metabolite biosynthesis[52]. These gene clusters encode all the machinery to regulate, synthesize, modify and transport the resulting molecule (see Figure 1.1). In most cases, all these genes are under a common control mechanism, which allows to switch their transcription on and off when needed. How-

ever, there are also cases where certain genes involved in a pathway can be found elsewhere in the genome. One plausible explanation is that the de-localized coding gene is also used in another pathway and thus, the regulation of the pathway is different. For instance, the butyrate production via acetate involves the presence of a thiolase, crotonase, butyryl-CoA dehydrogenase, β-hydroxybutyryl-CoA dehydrogenase and electron transfer proteins α and β[53]. In many bacteria, these genes are found in synteny in the genome, as is the case in *Eubacterium hallii* and *Clostridium sporogenes*. However, there are two other genomic arrangements that imply the de-localization of the thiolase or the crotonase, as seen in *Clostridium acetobutylicum* and *Eubacterium rectale*, respectively[54].

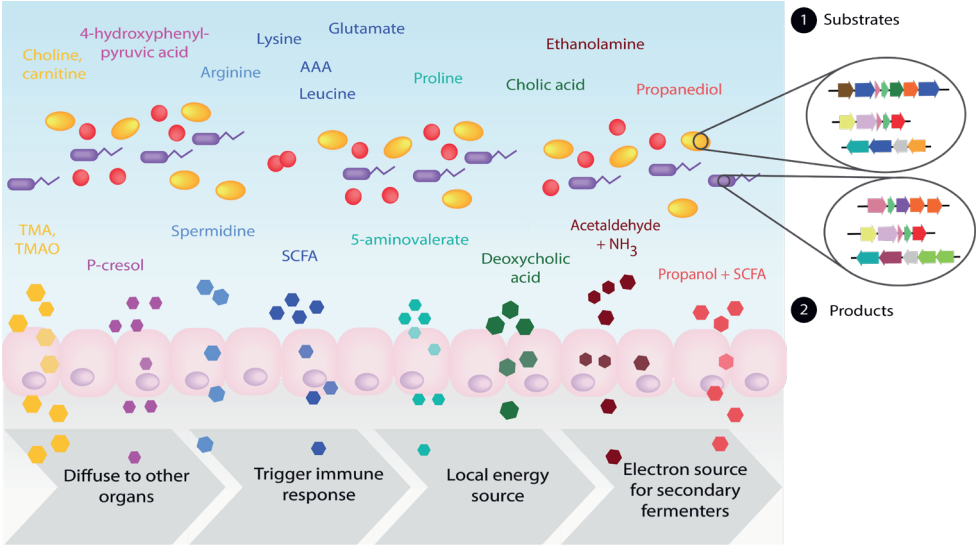


Figure 1.1. Representation of bacterial metabolic conversions encoded in MGCs that are typically found in human-associated bacteria. The resulting metabolites are also known to be involved in different functions as depicted at the bottom.

1.6 Small differences in the genomes can be involved in strain-specific traits

Within bacterial species, a multitude of strains exists. These are generally conceived as genetic variants that originate from a common ancestor over generations, diverging through mutations and acquisition/loss of genetic material. However, there is no consensus on what truly defines a strain[55,56]. One could assume that since most genomic content is shared between strains, they would have the same phenotypic traits, but this is not always the case. For instance, two single mutations in *gyrA* and *parC* genes confer resistance to quinolone to some *E. coli* strains[57], and mutations are the main mechanisms by which bacteria acquire antibiotic resistance[58]. Another example is the *Helicobacter pylori cagA*-positive strains, which produce the *cagA* protein associated with tumorigenic cell transformations[59] and with a higher risk of suffering gastric cancer[60]. Also, other strain-specific properties such as tolerance to acid and mucus adherence are known to play a role when selecting the best probiotics candidates[61]. These are examples of bacteria with very little differences in their genomic background that yet have a major impact on the host. Therefore, profiling the community at the species level can be sometimes insufficient. However, do we have the means to profile the gut microbiome at strain-level resolution? Certainly with the increasing numbers of metagenomes deposited in public repositories, the opportunities to effectively profile their genomes are greater. Binning contigs into accurate genomes from metagenome data is challenging though, and metagenome-assembled genomes (MAGs) tend to be less complete and to contain more chimeric sequences, thus of overall lesser quality than genomes of bacterial isolates[56]. Nonetheless, there are now tools that can accurately reconstruct bacterial strains from metagenomic data, such as DESMAN[62], metaSNV[63], ConStrains[64] and SGV-Finder (https://github.com/segalab/SGVFinder). These provide a promising avenue to profile microbiomes at strain-level resolution, despite their limitations (discussed more

in depth in reference 56). Several studies have recently used such approaches to elucidate structural variants from metagenomic data but also single-nucleotide variants (SNV). These variants are highly prevalent in the human microbiome and are person-specific[65,66]. For instance, Costea *et al.* assessed the variation within 71 abundant human gut-related bacteria. They observed that within *Eubacterium rectale* species*,* two subspecies harbour an additional set of genes that encode for a pro-inflammatory flagellum operon, associated with obesity and insulin resistance phenotypes[65]. Along the same line, a study by Zeevi *et al.* uncovered that a variable region found in *Anaerostipes hadrus* is in charge of the conversion of inostitol to butyrate, a health-promoting SCFA that colonocytes use as energy source[66]. Thus, focusing on the functional potential of the gut microbiome to develop personalized treatments will be important to analyse the community structure of each patient at strain-level resolution.

1.7 Bioinformatic tools to metabolically profile the human microbiome

Metabolites derived from specialized primary and secondary metabolism are key players in the crosstalk with the host and, as stated in the previous section, major modulators of host phenotypes. Traditionally, chemical diversity was screened by cultivating bacteria in specific media to then structurally characterize those that exert the desired bioactivity. However, this is time-consuming and very labor-intensive. For this reason, more recently, and taking advantage of the increasing number of microbial genomes in public repositories, new bioinformatics-powered approaches have been developed as an alternative to conventional methodologies for drug discovery[67]. Examples of tools built for such approaches are (1) SMURF; designed for fungal genome mining[68], (2) NP.searcher, specifically designed to predict NRPS and PKS BGCs[69], (3) ClustScan, also designed for the detection of NRPS, PKS or hybrids of them and their domain organization[70], (4) CLUSEAN, which also predicts NRPS/PKS gene cluster types found in bacterial

genomes[71], (5) PRISM also designed to predict NRPS and PKS BGCs[72], (6) antiSMASH, to date the tool capable of predicting the largest number of known types of gene clusters[73] and (7) ClusterFinder, which is able to predict BGCs from both known and novel classes[74]. Some of these tools allow downstream analysis and provide further information on the predicted gene clusters, such as structure predictions and predictions of substrate specificities. Complementary pipelines have more recently been introduced to overcome the limitations of the tools mentioned above and to provide further insights. This is the case for metaBGC, which identifies BGCs from uncultured bacteria by using metagenomic data as input[50]. Another useful tool that has been recently developed is biosyntheticSPAdes, which focuses on the challenge of predicting BGCs from fragmented genomic assemblies using assembly graphs[75]. In summary, there is a wide range of algorithms that can help elucidate novel compounds and advance the natural products discovery field.

In contrast, the options to systematically analyze microbial genomes for specialized primary metabolic gene clusters are much more limited. There are different algorithms that are able to predict bacterial operons by using the intergenic distances between contiguous genes and assessing their functional similarity using COG annotations[76] or other customized databases[77]. FMAP[78] uses metagenomic and metatranscriptomic data as input to infer the abundance of gene families and assess their differential abundance at the operon and pathway level using KOs. COG and KEGG are good resources to annotate genes into generic functional categories. One of the tools that uses these databases to infer functionality is, for instance, the HMP Unified Metabolic Analysis Network (HUMAnN) software, which uses metagenomic data to determine the presence of KEGG and MetaCyc pathways in the given community[79]. Despite the fact that these tools are able to generate interesting hypotheses on the association between microbiome-derived molecules and host phenotypes, as reported by Abubucker *et al.*[79], these normally do not refer to specific metabolites

but rather to more generally categorized pathways. Another newly designed tool is FishTaco[80], which allows to correlate taxon contributions with KEGG Orthology group (KO) abundances[81]; this can help identify potential taxonomic targets to modulate the human microbiome for therapeutic purposes. Although tools that rely on these databases are able to predict associations between microbiome-derived molecules and host phenotypes[79], they often lack resolution at the pathway level, making it difficult to pinpoint which particular pathways are involved in specific phenotypes.

Given the evidence that the production of specialized primary metabolites is encoded in metabolic gene clusters and that these metabolites have an important role in microbe-microbe and microbe-host interactions, the aim of this thesis is to provide new tools to functionally profile the human microbiome. This implies designing different methods not only to predict such MGCs, but also to assess the taxonomic distribution and architectural diversity of MGCs, and correlate their co-abundance and co-expression patterns in samples with specific phenotypes to better comprehend the roles they play in these complex ecosystems. Ultimately, the objective is to make use of all these tools and apply them to real datasets from public repositories, in order to make significant leaps in our understanding of the molecular mechanisms behind microbially derived phenotypic traits.

1.8 Thesis outline

The thesis has five different chapters that attempt to fill the gaps in knowledge in different research areas having, with as a central point the analysis and prediction of gene clusters.

In **Chapter 2**, the antiSMASH database version 2 is introduced. The database was provided with an updated infrastructure that stores compressed information of BGCs from a diverse bacterial genome collection chosen using average nucleotide identity on a large set of publicly available genomes

to remove redundant ones. This database is a comprehensive resource that allows to perform cross-genome searches. In **Chapter 3,** we highlight the usefulness of phylogenomics to uncover putative gene clusters reinforcing the accumulating evidence that large numbers of MGCs exist that encode yet-unknown metabolic pathways. Hence, this chapter shows a prospective potential to elucidate novel pathways and therefore how can help to functionally annotate novel genes. Based on this principle and tackling the need within the field to metabolically profile the gut microbiome, we built gutSMASH (**Chapter 4**). This is a new tool that not only allows to systematically profile specialized primary MGCs and bioenergetics-related gene clusters from anaerobes but also putative MGCs, which represent good candidates with unknown function to study further. Moreover, we also designed the gutSMASH web-server, a user-friendly platform that allows any researcher without bioinformatic background to run their analysis (**Chapter 5**). Finally, since analysing single "omic" layers (e.g. genomic functional potential) provides limited ability to fully establish causation between microbiome host-phenotypes, we designed BiG-MAP (**Chapter 6**). This tool represents a step forward to combine different omics data and obtain more biological insights at different molecular levels. More specifically, BiG-MAP allows to profile gene clusters' abundance and expression patterns that can ultimately help identify which gene clusters are most likely involved in conferring phenotypes of interest and prioritize them for further experimental characterization.

Chapter 2

The antiSMASH database version 2 – A comprehensive resource on secondary metabolite biosynthetic gene clusters

Kai Blin, Victòria Pascal Andreu, Emmanuel L.C. de los Santos, Francesco Del Carratore, Sang Yup Lee, Marnix H. Medema, Tilmann Weber

Published in 2019 in *Nucleic Acids Research*, Volume 47, Issue D1, Pages D625–D630

Abstract

Natural products originating from microorganisms are frequently used in antimicrobial and anticancer drugs, pesticides, herbicides, or fungicides. In the last years, the increasing availability of microbial genome data has made it possible to access the wealth of biosynthetic clusters responsible for the production of these compounds by genome mining. antiSMASH is one of the most popular tools in this field. The antiSMASH database provides pre-computed antiSMASH results for many publicly available microbial genomes and allows for advanced cross-genome searches. The current version 2 of the antiSMASH database contains annotations for 6,200 full bacterial genomes and 18,576 bacterial draft genomes and is available at https://antismash-db.secondarymetabolites.org/

2.1 Introduction

A majority of antibacterial and antifungal drugs, as well as drugs for many other indications, are derived from microbial natural products[82]. Traditionally, bioactive natural compounds were identified via classical isolation and analysis approaches. The increasing availability of genomic data in the last two decades allows us to complement these approaches with genome mining to identify and characterize biosynthetic pathways for natural products in genome and metagenome data[83]. Specialized software to support researchers in their search for natural products has been available for some years (for a comprehensive overview / list of such tools, please see[84,85,86]). Since its initial release in 2011, antiSMASH[73,87,88,89] has established itself as a standard tool for secondary metabolite genome mining and is currently the most widely used software pipeline for this task.

antiSMASH uses a rule-based cluster detection approach to identify 45 different types of secondary metabolite biosynthetic pathways via their core biosynthetic enzymes. For nonribosomal peptide synthases, type I

polyketides, terpenes, lanthipeptides, thiopeptides, sactipeptides and lassopeptides, antiSMASH can also provide more detailed predictions of the compounds produced by the respective biosynthetic gene clusters (BGCs). Identified clusters are compared to a database of clusters previously predicted by antiSMASH using the built-in ClusterBlast algorithm. A similar algorithm, KnownClusterBlast is used to compare the identified cluster against the manually curated set of known BGCs from the MIBiG[90] database. Secondary metabolite clusters of orthologous group (smCoG) classification is used to assign functions to gene products in the predicted BGCs.

As antiSMASH is a genome mining pipeline designed to analyze individual genomes, we developed the antiSMASH database[91] to provide interconnections and cross-genome search functionality based on antiSMASH results for many publicly available microbial genomes. Moreover, it provides users with instant access to full antiSMASH results of publicly available genome sequences. Here we present version 2 of the antiSMASH database. The database content of version 1, which was generated with version 3 of antiSMASH, was updated with annotation of the current antiSMASH 4.2.1 release. This implies that the antiSMASH database now includes updated detection rules, updated ClusterBlast database links, TTA codon prediction, NRPS-A domain predictions by the up-to-date Sandpuma software[92], classification of terpenes and improved links to MIBiG[90] (for details, please see reference 89). Furthermore, new sequences that became available after version 1 release were included. Version 2 of the antiSMASH database now contains genome mining results for 6,200 full bacterial genomes and 18,576 draft genomes from the NCBI RefSeq database[93]. The increased dataset is accompanied by improvements in the search functionality, data export options and the user interface of the antiSMASH database.

2.2 Materials and Methods

2.2.1 Selection of included genomes

Microbial genome resources are growing rapidly and, despite taxonomically novel genomes being released frequently, there is a lot of sequence redundancy in the NCBI genome databases, i.e. thousands of sequences of mostly pathogenic bacteria such as *Pseudomonas aeruginosa* or *Escherichia coli*. Therefore, with the objective of creating a representative set of genomes that are non-redundant, we designed an approach to effectively update the antiSMASH database, maintaining its high quality and adequately representing natural diversity without significantly decreasing the overall pipeline performance in terms of speed.

Genomes categorized as 'draft genomes' are fragmented in multiple contigs. As many secondary metabolite biosynthetic gene cluster contain repetitive sequences, this implies that many BGCs end up being split on multiple contigs without any linkage information, leading to low-quality BGC data. Consequently, in order to minimize this issue, we prioritized the inclusion of NCBI RefSeq genomes that were annotated with the assembly level 'complete genome' or 'chromosome' present in the database on April 2018 (10,863 genomes in total). We then estimated the distance between selected assemblies using fastANI (Average Nucleotide Identity) (https://github.com/ParBLiSS/FastANI). FastANI uses a hash-based algorithm to estimate the average nucleotide identity between pairs of genomic assemblies. A network was generated with each genome as a node, and weighted edges between nodes corresponding to the fastANI estimate between genomes. We used a fastANI similarity score of 99.6 as a cutoff for having an edge between nodes. Nodes were then assigned to communities using the multilevel community structure algorithm (https://arxiv.org/abs/0803.0476) in the igraph Python package (Csardi G, Nepusz T: The igraph software package

for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org). Finally, a representative genome from each community was chosen by prioritizing assemblies with the highest contig N50 and lowest contig L50. This resulted in a total of 6,200 complete genomes for the antiSMASH database.

In order to supplement the set of complete and chromosomal assemblies, we added a set of draft genomes to the antiSMASH database. To select draft genomes for addition to the database, we started with a previously published set of precomputed fastANI similarity scores of ninety thousand prokaryotic genomes (https://doi.org/10.1101/225342). We pre-filtered this set to remove poor quality genomes (N50 < 20kb and assembly anomalies). We then performed the same procedure as with the complete and chromosomal assemblies to group the draft genomes into communities. A representative genome from each community was chosen by prioritizing assemblies based on assembly level (scaffold > contig), and then selecting assemblies with the highest contig N50 and lowest contig L50. In order to maintain consistency with the complete and chromosomal set, only draft genomes that had corresponding Refseq assemblies were included in the database. The following resulted in an additional 18,576 draft genome entries that were added to the database.

2.2.2 antiSMASH annotations and data import

Based on the selection criteria mentioned above, the assemblies were downloaded from the NCBI servers in GenBank format using the ncbi-genome-download tool (https://github.com/kblin/ncbi-genome-download/). GNU parallel[94] was used to run multiple docker containers of antiSMASH 4.2.1 simultaneously. Different analysis para-meters were used for the full and partial genome set. For full genomes, ClusterBlast, KnownClusterBlast, SubClusterBlast, ActiveSiteFinder, TTA codon detection in automatic mode, secondary metabolite clusters of orthologous groups prediction, and clus-

ter-specific detailed annotations were run (command line flags: --clusterblast --knownclusterblast --subclusterblast --asf --tta-auto --smcogs-notree). For draft genomes, antiSMASH was run in fast mode, skipping the detailed annotations.
Additionally, KnownClusterBlast, TTA codon detection in automatic mode, and secondary metabolite clusters of orthologous groups prediction were run (command line flags: --minimal --knownclusterblast --tta-auto --smcogs-notree).

The SQL schema of the (https://github.com/antismash/db-schema/) antiSMASH database was updated to accommodate the annotation changes and additional features/predictions that were introduced by antiSMASH version 4. The antiSMASH results in GenBank format were loaded into the SQL schema using the import script available at https://github.com/antismash/db-import/.

2.3 Results and Discussion

With an update to the PGAP annotation pipeline used by the NCBI, the annotation issues causing us to use records from GenBank instead of RefSeq for version 1 of the antiSMASH database have largely been resolved. Hence, with version 2 of the database, we have switched to using RefSeq genomes to obtain more unified gene annotations.

The antiSMASH database 2 contains BGCs identified in 6,200 full genomes (an increase of 58%) and adds 18,576 draft genomes. Annotations in the database are generated by antiSMASH version 4.2.1, the most recent release of antiSMASH[89]. New in the antiSMASH 4.2.1 release are detection rules for N-acyl amino acids, polybrominated diphenyl ethers, and PPY-like pyrones. Detailed cluster product predictions have been added for lasso peptides, thiopeptides, sactipeptides (based on RODEO[95]), non-ribosomal peptide synthases (based on SANDPUMA[92]) and terpenes. The Cluster-

Blast and KnownClusterBlast databases have been updated.

The search builder has been extended to cover these new features. A new search field in the taxonomy browser makes it easier to navigate to species of interest in the much larger dataset.

The gene cluster data obtained in the queries can be downloaded. Depending on the type of search, different file formats are available. For gene cluster searches, the result table can be downloaded in tabular (CSV) format, alternatively it is possible to retrieve the DNA sequence of all matching clusters in FASTA format. Gene and protein domain searches offer a download the protein and nucleotide sequences of all matching genes or protein domains, respectively, or a tabular representation of the results. New options are provided to download specific chunks of the result data (for example only the first 1000 sequences) and to select between standard FASTA headers including the IDs and descriptive headers also including the query the hits were obtained with.

The selection of genomes available from NCBI still skews the perspective on the available diversity of biosynthetic gene clusters. While the antiSMASH database contains sequences from 33 different phyla, sequences from e.g. proteobacteria are vastly overrepresented due to their significance as pathogens. The database now contains 32,548 biosynthetic gene clusters from the full genome dataset, an increase of 46 % from version 1 (Table 2.1). Statistics from the 18,576 draft genomes certainly overpredict the number of identified clusters due to clusters being split over several contigs and counted multiple times, the fast-mode results still provide a good first estimate of the available biosynthetic diversity of the draft genomes. Of the 119,558 BGCs predicted on the draft genomes, over a third (41,482) are in contact with at least one contig edge and thus likely incomplete. In comparison, only about 1% of the clusters from the full genome dataset (390 in 32,548) are located on a contig edge. As the abundant fragmentation of clusters in

draft genomes is skewing the numbers, the following statistics only count the results from the full genomes. See Table 2.2 for detailed cluster counts by BGC type and a comparison with the cluster counts from version 1.

Table 2.1 Overview on BGC numbers in version 1 and version 2 of the antiSMASH database.

Overall database statistics	Version 1 counts	Version 2 counts	% change
Full (high quality) genomes	3907	6200	58
Number of BGCs in full genomes	22292	32548	46
Draft genomes	0	18576	new
Number of BGCs in draft genomes	0	119558	new
BGCs in total	**22292**	**152106**	**682**

Table 2.2 Changes in cluster counts of the different BGC types between version 1 and version 2 of the antiSMASH database (excluding data from draft genomes).

Gene cluster types (high quality genomes)	Version 1 counts	Version 2 counts	% change
NRPS			
Nonribosomal peptide	5878	7893	34
Terpenes			
Terpene	3362	5018	49
Polyketides			
Type I polyketide	2608	3302	27
Type III polyketide	742	1141	54
hglE-type polyketide	590	768	30
Trans-AT polyketide	512	623	22
Type II polyketide	173	307	77
PPY-like pyrone	0	13	new
RiPPs			
Bacteriocin/RiPP	3323	5198	56
Lanthipeptide	857	1121	31

Thiopeptide	122	1097	799
Lasso peptide	351	562	60
Sactipeptide	59	318	439
Microviridin	18	70	289
Head-to-tail cyclised (subtilosin-like)	22	52	136
Proteusin	13	39	200
Microcin	5	3	-40
Bottromycin-like	1	2	100
Other			
Other	1887	2322	23
Siderophore	1399	1745	25
Homoserine lactone	1084	1608	48
Aryl polyene	988	1595	61
Ectoine	424	794	87
Butyrolactone	189	392	107
Phosphonate	248	342	38
Resorcinol	184	261	42
Ladderane	113	217	92
Phenazine	152	210	38
Melanin	45	113	151
N-acyl amino acid cluster	0	110	new
Indole	48	104	117
Cyanobactin	30	77	157
Polyunsaturated fatty acid	45	61	36
Oligosaccharide	40	54	35
Aminoglycoside/aminocyclitol	26	51	96
Nucleoside	23	49	113
Linaridin	17	35	106
Beta-lactam	13	30	131
Aminocoumarin	3	10	233
Pheganomycin-like ligase	5	7	40
Phosphoglycolipid	1	4	300
Furan	2	3	50
Glycocin	14	3	-79
Polybrominated diphenyl ether	0	1	new

In order to get an accurate taxonomic overview, the identified BGCs were mapped to a phylogenetic tree displaying approximately half of the genomes (12,219 complete and draft) that conform the here presented database (Figure 2.1 A). The topology of the tree shows the microbial diversity chosen, ranging from well characterized phyla to unclassified bacteria found in diverse ecosystems. Proteobacteria, Actinobacteria, Bacteroidetes, Firmicutes, Spirochaetes, Tenericutes, Cyanobacteria and Deinococcus-Thermus, the eight most abundant bacterial divisions in our database, accounting for 97,6% of genomes and all vary in the number of harbored BGCs (Figure 2.1 B). High BGC numbers are characteristic features for some groups of bacteria such as Actinobacteria (containing 13 clusters on average (full genomes) while others rarely possess one, like Tenericutes. These bacteria exhibit different distributions in terms of encoded secondary metabolite types as defined by antiSMASH (Figure 2.1 C). For these statistics, the 45 BGC classes in antiSMASH have been condensed into five major groups: Non-Ribosomal Peptide Synthetase (NRPS), Polyketide, Ribosomally synthesized and post-translationally modified peptides (RiPP), terpenes and Others, clusters that do not belong to any of the aforementioned types. Terpenes, bacteriocins (a type of RiPP) and NRPS are the most common BGC types, all with higher number of representatives in the phylum Proteobacteria.

2.4 Conclusions

Genome mining is a valuable method to assess the biosynthetic potential of microorganisms. Since 2011, antiSMASH has assisted researchers with their secondary metabolite genome mining projects. The public web service has processed around 400,000 jobs, and the standalone tool has been downloaded over 10,000 times. The antiSMASH database both allows instant access to antiSMASH results for many publicly available genomes instead of waiting several hours for a de-novo antiSMASH run and allows advanced cross-genome searches for BGCs with specific features of interest.

In comparison to version 1, the updated version 2 of the antiSMASH database provides antiSMASH 4.2.1 annotations for 6,200 full genomes, which is an increase by 58%, and newly introduces data for 18,576 draft genomes. The graphical query builder allows researchers to interactively formulate searches to answer cross-genome research questions, while the results are presented in the familiar antiSMASH output format.

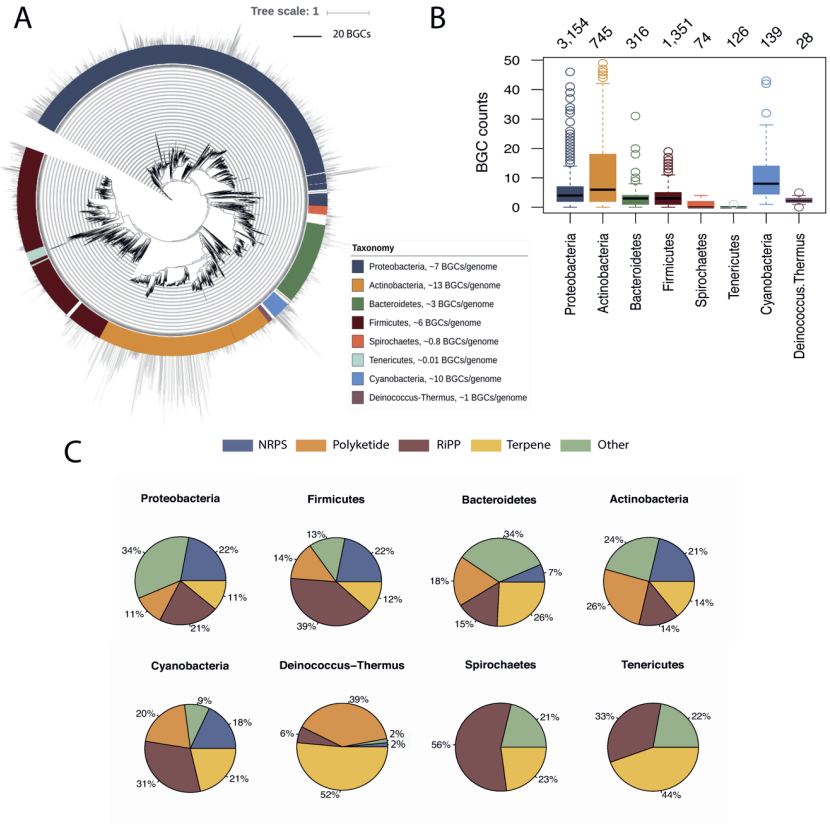


Figure 2.1 Statistic summary of the antiSMASH database version 2. (A) A phylogenetic tree constructed from the revised version of tree of life based on 120 conserved protein markers[96]. The original tree was pruned by genome assembly id using ETE Toolkit[97], to only keep leaves that belong to genomes of the antiSMASH database version 2. The visualization and customization of the tree was performed with iTOL[98]. As a result, 12 219 leaves from the total of 24,776 bacterial genomes

are shown in this phylogeny. The colored ring represents the eight most abundant phyla; 97.6% of the genomes, and the bar plots in the outer ring the number of BGCs per genome. (B) Boxplots of the BGCs counts per phylum, with the values on top showing the total number of complete genomes per phylum. (C) Pie charts of the five major BGC classes per phylum showing the diversity of natural products produced by each group of bacteria.

2.5 Availability

The antiSMASH database is available at https://antismash-db.secondary-metabolites.org/. There are no access restrictions for academic or commercial use of the web server. The source code components and SQL schema for the antiSMASH database are available on GitHub (https://github.com/antismash) under an OSI-approved Open Source license.

Chapter 3

Computational genomic discovery of diverse gene clusters harboring Fe-S flavoenzymes in anaerobic gut microbiota

Victòria Pascal Andreu, Michael A. Fischbach,
Marnix H. Medema

Published in 2020 in *Microbial Genomics*, Volume 6,
Issue D1, e000373

Abstract

The gut contains an enormous diversity of simple as well as complex molecules from highly diverse food sources as well as host-secreted molecules. This presents a large metabolic opportunity for the gut microbiota, but little is known on how gut microbes are able to catabolize this large chemical diversity. Recently, Fe-S flavoenzymes were found to be key in the transformation of bile acids, catalysing the key step in the 7alpha-dehydroxylation pathway that allows gut bacteria to transform cholic acid (CA) into deoxycholic acid (DCA), an exclusively microbe-derived molecule with major implications for human health. While this enzyme family has also been implicated in a limited number of other catalytic transformations, little is known about the extent to which it is of more global importance in gut microbial metabolism. Here, we use large-scale computational genomic analysis to show that this enzyme superfamily has undergone a remarkable expansion in Clostridiales, and occurs throughout a diverse array of >1,000 different families of putative metabolic gene clusters. Analysis of the enzyme content of these gene clusters suggests that they encode pathways with a wide range of predicted substrate classes, including saccharides, amino acids/peptides and lipids. Altogether, these results indicate a potentially important role of this protein superfamily in the human gut, and our dataset provides significant opportunities for the discovery of novel pathways that may have significant effects on human health.

3.1 Introduction

The gene set of the human gut microbiota vastly exceeds the human gene repertoire[14,99], which allows microbes to complement human metabolism by degrading undigested polysaccharides, lipids, and peptides that reach the large intestine[100]. For instance, saccharolytic bacteria can ferment carbohydrates to produce short chain fatty acids (SCFAs), beneficial metabolites that promote health[101]. Nevertheless, gut bacteria also produce many mol-

ecules involved in microbe-microbe interactions and microbe-host interactions that can have detrimental effects instead[52]. An example of a harmful diet-derived metabolite is trimethylamine (TMA), an amine that can be synthesized from choline or carnitine by certain gut bacteria, and which has been associated with cardiovascular and renal disease[102]. Thus, the identification of these molecules and the elucidation of their producing pathways is crucial to assess the causes and consequences of certain microbiome-associated phenotypes.

Another example of exclusively microbiome-derived molecules are the secondary bile acids (BA): deoxycholic acid (DCA) and lithocholic acid (LCA). While the primary bile acids cholic acid (CA) and chenodeoxycholic acid (CDCA) are synthesized by the liver[103], they are transformed into DCA or LCA by colonic bacteria during enterohepatic circulation[104]. These molecules have been proposed to act as inhibitors of *C. difficile* outgrowth[105], as well as to induce the development of colon cancer[106,107] and cholesterol gallstone disease[108]. The main bacterial pathway in charge of the 7alpha-dehydroxylation needed to produce them constitutes a multi-step biochemical reaction that can be accomplished by bacteria harboring the bile-acid-inducible (*bai*) operon[109]. Most of the bacteria capable of carrying out this reaction are anaerobes that are part of the *Clostridium* cluster XIVa[109]. The pathway encoded by the *bai* operon was recently elucidated by Funabashi & Grove *et. al*[110]**.** The authors showed that the key step is performed by the BaiCD enzyme, an Fe-S flavoenzyme that oxidizes 3-oxo-cholyl-CoA to 3-oxo-4,5-dehydrocholyl-CoA. Later, BaiH (also an Fe-S flavoenzyme), BaiCD and BaiA2 act again on the molecule to finally produce DCA. Importantly, the participation of Fe-S flavoenzymes in the key reductive steps of the pathway is consistent with a role for this pathway in using primary bile acids as terminal electron acceptors for an anaerobic electron transport pathway, constituting a unique metabolic niche within the gut community.

In addition to the BaiCD and BaiH enzymes, a few other members of the Fe-S flavoenzyme superfamily have been previously shown to play similar crucial roles in the redox metabolism of catalytic transformations. For instance, L-phenylalanine fermentation via a Stickland reaction, where cinnamate is reduced to 3-phenylpropionate, has been shown to be performed by a cinnamate reductase that is a member of the Fe-S flavoenzyme superfamily[111]. Additional Fe-S flavoenzyme representatives with experimentally characterized functions include a 2,4-dienoyl-CoA reductase implicated in fatty acid beta-oxidation[112] as well as a trimethylamine dehydrogenase involved in trimethylamine degradation[113].

The fact that these enzymes have been shown to facilitate the shuttling of electrons from the membrane to diverse organic terminal electron acceptors, enabling an anaerobic electron transport chain, led us to hypothesize that they might play a more widespread role in the microbial catabolism of the diversity of complex substrates available in the gut. Here, we provide an in-depth computational genomic study of the Fe-S flavoenzyme superfamily and its presence across genomes of gut microbiome-related bacteria. Using phylogenomic analyses, we show that this enzyme superfamily comprises a large sequence diversity and has particularly undergone strong evolutionary expansion in the class Clostridia, emphasizing its possible implication on different metabolic reactions in the gut. Large amounts of strain-level variation between genomes indicates that the pathways involved likely facilitate ecologically specialized functions. Finally, analysis of the enzyme content in their surrounding operons and gene clusters uncovers a wide array of putative catabolic gene clusters associated with the breakdown of diverse substrates, which provides a rich resource for uncovering novel pathways in the human microbiome and beyond.

3.2 Results & Discussion

3.2.1 The Fe-S flavoenzymes phylogeny includes many unexplored clades involved in many catalytic reactions

In order to assess the distribution of the Fe-S flavoenzymes along the bacterial kingdom, we scanned 111,651 bacterial genomes (see Methods section 3.4.1). Across this dataset, we identified 49,870 Fe-S flavoenzymes that belong to more than twenty bacterial phyla. Some bacterial genomes encode remarkable numbers of Fe-S flavoenzymes: e.g., the genome of the Firmicute *Sporobacter thermitidis* encodes no fewer than 18 Fe-S flavoenzymes. There are also representatives from Actinobacteria (including *Eggerthella* sp. YY7918 and *Rhodococcus* sp. SC4) that possess 9 different flavoenzymes. On average, Firmicutes have 1.81 ± 1.18 protein copies per genome, whereas Proteobacteria and Actinobacteria have 1.51 ± 0.83 and 1.10 ± 0.50 copies respectively. These numbers exclude genomes that do not encode for any flavoenzyme, representing 89%, 55.5% and 46.8% of the genomes from Firmicutes, Proteobacteria and Actinobacteria respectively. Together, these data suggest that Fe-S flavoenzymes may play important roles in multiple bacterial phyla, and particularly in Firmicutes.

To contextualize these quantities of Fe-S flavoenzyme genes with their evolutionary history and functional diversification, we then conducted a comprehensive phylogenetic analysis of this enzyme family. We obtained a collection of 8,097 non-redundant Fe-S flavoenzymes by selecting representatives from MMseqs2[114]-derived sequence clusters. From this dataset, we constructed an approximate maximum likelihood phylogeny (Figure 3.1; see also Methods section 3.4.2). Within this phylogeny, we were able to annotate eight clades based on experimentally characterized proteins documented in UniProt to date; the vast majority of clades have unknown functions. Given the size of this protein superfamily, the number of unexplored clades and its functional diversity, this phylogeny constitutes a road map to uncover novel clades of enzymes that could potentially aid in the characterization of yet undiscovered catabolic pathways.

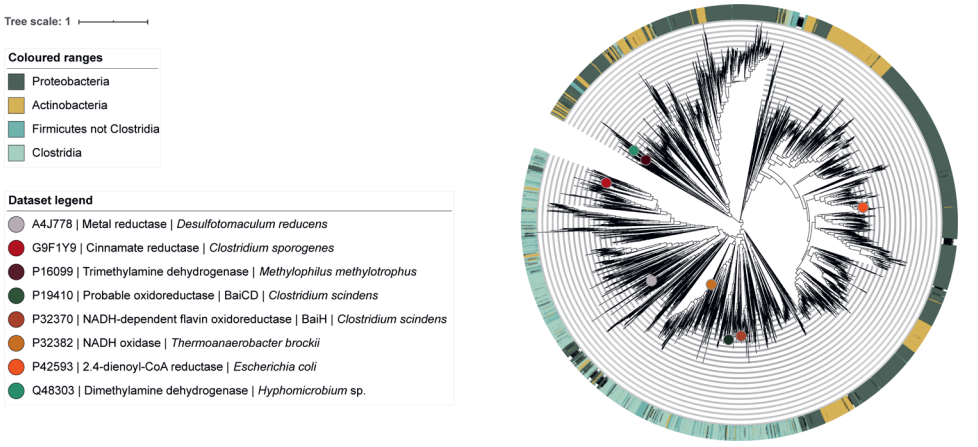


Figure 3.1 Fe-S Flavoenzyme superfamily phylogeny. The phylogeny was constructed with FastTree from 8,105 non-redundant protein sequences across the bacterial kingdom, including the eight experimentally characterized proteins from Uniprot: (https://www.uniprot.org/uniprot/?query=PF07992+and+PF00724+and+doreviewed%3Ayes&sort=score, accessed at 01/02/2020). All sequences used as input contain both the Oxidored_FMN and Pyr_redox_2 Pfam domains. The eight experimentally characterized proteins from Uniprot are indicated as circles on the tree to functionally annotate clades.

3.2.2 An evolutionary expansion of Fe-S flavoenzyme in Clostridia

As reported in the previous section, the phylum Firmicutes stands out for the high Fe-S flavoenzyme copy numbers found in genomes of the species that belong to it. For this reason, we investigated the flavoenzymes copy number variation in *Clostridium* along with other microbial taxa commonly found in the gut: Firmicutes, Bacteroidetes, Actinobacteria and Proteobacteria. The phylogenetic analysis already suggested a large evolutionary expansion of Fe-S flavoenzymes in Clostridia, with this bacterial class occupying about a third of the non-redundant superfamily phylogeny. Indeed, quantitative analysis of the Clostridia class confirms that it has undergone a massive expansion when comparing the Fe-S flavoenzyme copy number with other taxa

(Figure 3.2). The Actinobacteria and Bacteroidetes phyla seem to harbour considerably fewer Fe-S flavoenzymes. As expected, we found that among genomes from Clostridia, there is a high percentage that encode at least one Fe-S flavoenzyme (71.58%). In contrast, less than half of the genomes in our database belonging to Enterobacteriaceae, Bacilli and *Bacteroides* code for a flavoenzyme (49.92%, 7.64% and 3.19% respectively). Strikingly, *Eggerthella* seems to be an exception to this pattern, as 10 genomes from different *Eggerthella* species encode 38 Fe-S flavoenzymes, with 71.43% of the genomes with at least one count.

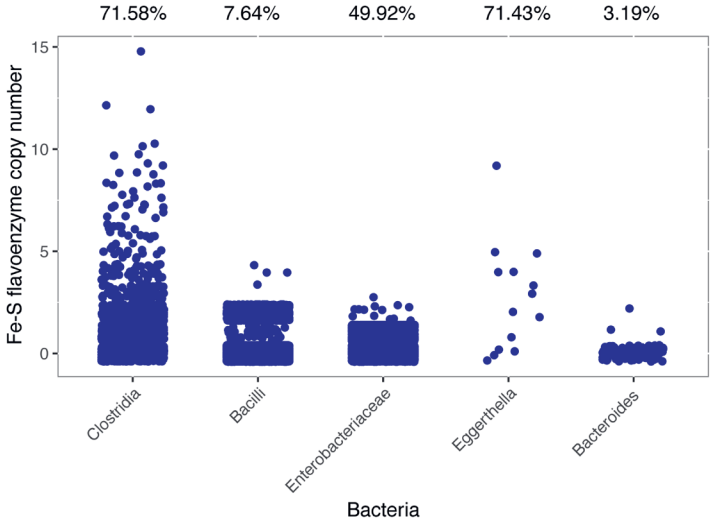


Figure 3.2 Expansion of the Fe-S flavoenzyme superfamily in Clostridia when compared to other gut-related taxa. Scatterplot representing the Fe-S flavoenzyme copy number across some gut-related bacteria. The taxonomic groups cover the five major phyla from the gut: Firmicutes (Clostridia and Bacilli), Proteobacteria (Enterobacteriaceae), Actinobacteria (Eggerthella) and Bacteroidetes (Bacteroides). The representative candidates used for this analysis were gathered by discarding genomes whose Genbank files were stated to be isolated from sources such as pig, insects, rumen, sludge, water or soil. The taxa picked are the ones found to possess at least one copy of flavoenzyme-encoding genes. The percentages on top represent how many of the genomes harbour at least one Fe-S flavoenzyme-coding gene out of the total number of genomes from that taxon.

3.2.3 Unexpected amounts of strain variation of flavoenzyme counts between genomes

In microbial ecosystems, strain-specific traits often confer different characteristics to otherwise almost identical bacteria, which allow them to thrive in specific conditions[115,60,116]. Therefore, we assessed Fe-S flavoenzyme copy number variation in order to know whether such enzymes are responsible for specialized functionalities that allow microbes to adapt to unique ecological niches. Indeed, we found that Fe-S flavoenzyme copy number distribution along Clostridia genomes often shows signs of strain-specificity, as can be seen in *Clostridioides difficile* clade and *Clostridia* clade II (Figure 3.3). In the Clostridia phylogeny (Figure 3.3), the number of copies ranges from 1 to 18 per genome, with a *Sporobacter thermitidis* genome being the one with highest counts. Another bacterium with a high Fe-S flavoenzyme copy number is *C. scindens* (Clostridia clade I), whose genome encodes up to 10 distinct members of the family (including BaiCD and BaiH). In comparison, other gut-related genera such as *Faecalibacterium* (clade IV, ~12 o'clock), *Roseburia* (clade I, ~5 o'clock) and *Anaerostipes* (clade I, ~4 o'clock) show quite different Fe-S flavoenzyme copy number profiles, harbouring on average 0.3, 0.6 and 2.4 copies per genome, respectively. Overall, these results confirm that indeed, for specific clades of Clostridia, Fe-S flavoenzymes are likely to play an important role in their primary metabolism. Moreover, they evince the importance of profiling the microbiome at high (or strain) resolution; 16S amplicon sequencing (and especially out-level clustering) would not be able to distinguish between bacteria with different metabolic repertoires. The differential metabolic capabilities of various strains are often due to very subtle changes in their genomes[55]. Therefore, studying these bacterial strains that differ in terms of their Fe-S flavoenzyme copy numbers, as well as the genomic contexts where these enzymes are encoded, could help to uncover strain-specific traits that play a role in conferring microbiome-associated host phenotypes.

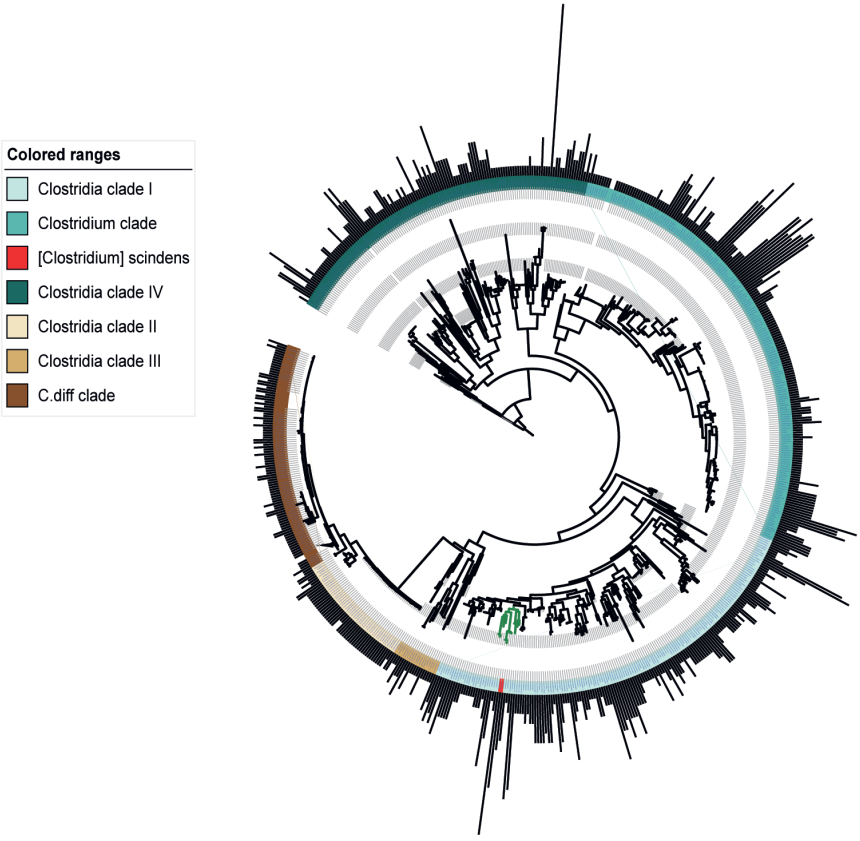


Figure 3.3 Strain- and species-specific copy numbers of Fe-S flavoenzymes across Clostridia. A circular phylogenetic tree showing bacterial strains in the class Clostridia. The outer ring of bars represents the copy number of Fe-S flavoenzymes in each strain, indicating a high degree of variability among strains and species. The phylogeny includes closely related strains whose genomes encode diverse numbers of Fe-S flavoenzymes, ranging from 1 to 18. Colored strips represent the taxonomic entities found within the class Clostridia. Two clades have been collapsed in order to remove almost identical genomes in the tree (shown as a discontinuity in the bar plots): the *Clostridium difficile* clade, Clostridia clade II and *Clostridium* clade.

3.2.4 Analysis of Fe-S flavoenzyme counts in reconstructed metagenome assembled genomes from the human gut microbiome

In order to evaluate the Fe-S flavoenzyme prevalence in the gut microbiota, we collected 1,952 uncultured MAGs belonging to different taxonomic groups and reconstructed from 11,850 human gut microbiomes by Almeida *et al.*[117] (see Methods section 3.4.4). From the protein sequences encoded in these MAGs, we identified a total of 1,602 Fe-S flavoenzymes that were found to be encoded across 771 of the genomes. Later, to determine which organisms encode at least one flavoenzyme, we ranked the MAGs assigned to a taxonomic lineage annotation provided by Almeida *et al.*[117] based on the number of predicted Fe-S flavoenzymes. The top 5 taxa with highest counts belong to either Actinobacteria, in particular *Collinsella sp.,* or Firmicutes, with 327 and 220 Fe-S flavoenzyme counts respectively (see Suppl. Table S1, section 3.5). This is in agreement with our previous findings that these taxa are two major contributors to the overall bacterial Fe-S flavoenzyme diversity as shown in Figure 3.1. Moreover, we also evaluated the Fe-S flavoenzyme copy number for each MAG, and surprisingly, an uncultured Clostridiales bacterium (GCA_900546485.1) codes for 22 different flavoenzymes, the highest copy number found in this study. This finding supports the evidence that the flavoenzyme superfamily has greatly expanded in the Clostridia class, but also indicates the remarkable role these enzymes may play in the gut microbiome and in particular in the bacteria with higher Fe-S flavoenzyme counts.

3.2.5 Analysis of gene neighborhoods leads to the identification of a wide range of families of putative catabolic gene clusters involved in breakdown of diverse biomolecules

We next explored the likely functionalities of flavoenzymes in Clostridia, by investigating the genomic contexts of the flavoenzyme-coding genes. With

that aim, we gathered a collection of 3,158 non-redundant gene clusters, all of them having in common the presence of a Fe-S flavoenzyme-coding gene (see Methods section 3.4.5). These gene clusters could be grouped into 1,052 Gene Cluster Families (GCFs). Interestingly, Fe-S flavoenzymes are found in a strikingly diverse array of GCFs. The examination of the corresponding gene cluster-coding proteins helped identifying more than 31 relevant Pfam domains that we categorized to be involved in 4 major types of metabolism (see Suppl. Table S2; see section 3.5). Thus, the profiling of this collection based on the presence of these protein domains within Fe-S-flavoenzyme-encoding gene clusters allowed to predict putative functions for 200 GCFs. Specifically, 80 distinct GCFs are predicted to function in saccharide catabolism, 83 in peptide/amino acid catabolism, 28 in nucleotide catabolism and 14 in lipid catabolism. In Figure 3.4, we highlight eleven gene clusters plus the *bai* operon to show examples of distinct cluster architectures that, based on their enzyme-coding gene contents, we predict to be capable of processing a diverse array of substrates. Importantly, most of the Fe-S-flavoenzyme-containing gene clusters also harbour a major facilitator superfamily transporter, indicating that they might process a diffusible substrate; also, many include proteins with flavodoxin or electron transfer flavoprotein domains, consistent with the possibility that these pathways are coupled to electron transport chains. Moreover, we also explored the genomic neighbourhood of the 22 Fe-S flavoenzyme-coding genes found in the uncultured Clostridiales bacterium from the MAGs collection, using the same methodology. One particularly interesting gene cluster is highlighted in Figure 3.4, which is putatively involved in carbohydrate catabolism, encoding an α-amylase (PF00128), a starch-binding domain (PF16738), an alanine dehydrogenase (PF01262) and a probable molybdopterin-binding domain (PF00994), among others.

Overall, these findings suggest that Fe-S flavoenzymes play a far more expansive role in anaerobic metabolism in the human gut than was previously known. Moreover, the predicted diversity in function indicates that these

oxidoreductases may be in charge of catalysing the transformation of a large variety of substrates, conferring to these bacteria specialized primary metabolic capabilities that may allow them to colonize specific micro-niches in the gut.

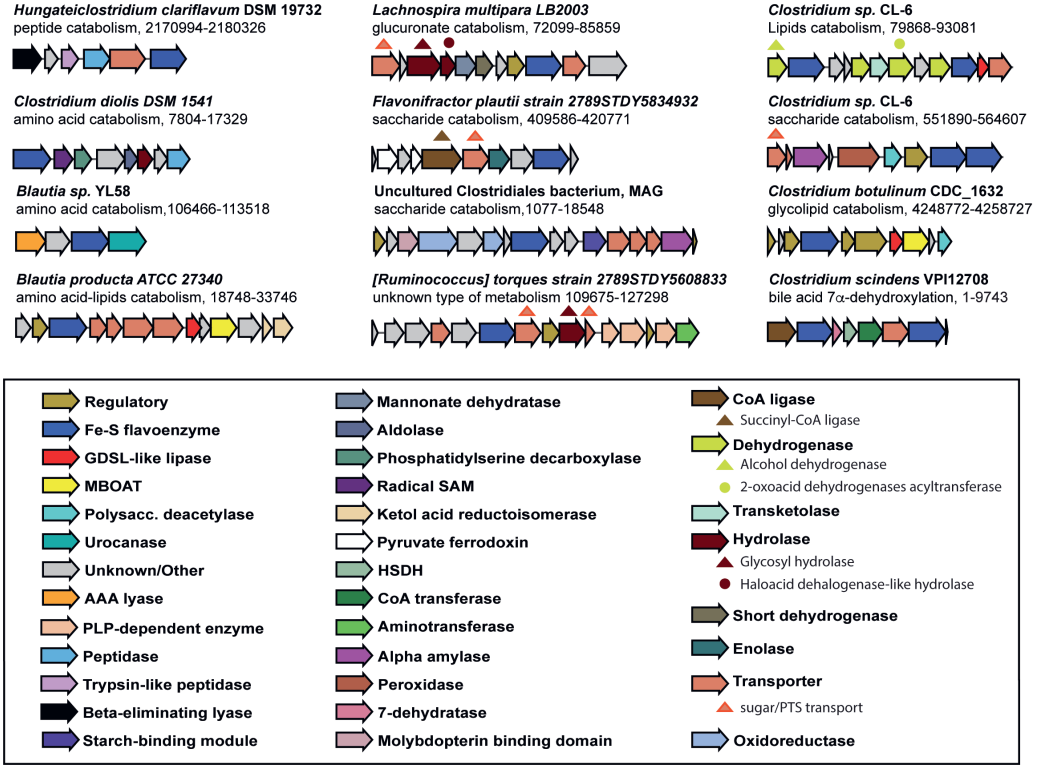


Figure 3.4 Diversity of Fe-S-flavoenzyme-encoding catabolic gene clusters from Clostridia. Gene clusters from Clostridia that contain a Fe-S flavoenzyme are numerous and diverse. Eleven examples predicted to be involved in peptides/amino acids catabolism, saccharide catabolism, lipid catabolism and unknown type of metabolism have been picked from a collection of 1,052 GCFs and one from the MAGs collection, plus the *bai* operon. Genes are colored by predicted function.

3.3 Conclusion

The metabolic potential of gut bacteria greatly exceeds the genetic potential of the human host. Therefore, bacteria can benefit from utilizing a diverse range of substrates that reach the digestive tract. Despite this, the mechanisms by which anaerobic bacteria catalyse these reactions are largely unknown. Here, we show that Fe-S flavoenzymes are likely to play a significant role in catalysing key redox steps within diverse catabolic pathways. Specifically, we found a remarkable expansion of Fe-S flavoenzyme copy numbers in the class Clostridia when compared to other gut-related bacteria. The strain-specificity of Fe-S-flavoenzyme-encoding gene cluster repertoires indicates that these enzymatic pathways may allow bacteria to specialize in the catabolism of specific dietary or host-derived molecules. We present a rich dataset of gene clusters that are candidates for detailed biochemical studies. Moreover, the presence or expression of these GCFs could be used as features alongside standard metabolic pathway annotations to assess whether their presence could explain variation in health/disease phenotypes, or whether they can explain variation observed in gut metabolomes[118,119]. All in all, given the importance of profiling the gut microbiome from a functional point of view, this study provides new ways of exploiting the genomic information present in public repositories, and provides a template for genomic exploration studies centred on key enzyme families to further understand the metabolic potential of the gut microbiome.

3.4 Material and Methods

3.4.1 Identification of members of the Fe-S flavoenzyme superfamily

BaiCD contains two Pfam domains, Oxidored_FMN (PF00724) and Pyr_redox_2 (PF07992). We used hmmsearch (HMMER 3.1b2, February 2015; http://hmmer.org/) to identify protein sequences harbouring both domains

from two databases: all bacterial sequences in GenBank (complete and draft genomes, >100,000 entries) and all sequences in RefSeq belonging to the class Clostridia (>2,500 entries), as some of the Clostridia genomes in GenBank lack gene coordinate annotations. Proteins with sequence e-value <= $1x10^{-5}$ for both domains were deemed hits. The resulting set of 49,437 Fe-S flavoenzyme superfamily members was used for the analyses described in the next three paragraphs.

3.4.2 Construction of an Fe-S flavoenzyme protein similarity network

The amino acid sequences of the 49,870 Fe-S flavoenzyme superfamily members were clustered using MMseqs2[114], setting the minimum identity to 0.9. From the 1,694 groups in the resulting protein similarity network, we picked five representatives of each cluster (randomly chosen when the cluster contained more than five nodes). ~2,500 singletons (families of size one) were also included in the subsequent analysis. We aligned the protein sequences to the Oxidored_FMN and Pyr_redox_2 Pfam domains using hmmalign (HMMER 3.1b2, February 2015; http://hmmer.org/), removed the unaligned and indel regions, merged the alignments of the two domains and reconstructed a phylogeny using FastTree[120]; the midpoint root of the tree was calculated using phytools package in R[121]. The resulting phylogenetic tree was annotated with interactive Tree of Life (iTOL[122]).

3.4.3 Computational analysis of the prevalence and phylogenetic distribution of Fe-S flavoenzymes

We investigated the taxonomic distribution of Fe-S flavoenzyme genes and their variability in copy number; for simplicity, the number of hits per genome assembly was used as a metric of the copy number per genome. We constructed a phylogenetic tree of genomes in the class Clostridia using the following procedure: Clostridia genome assemblies harbouring at least

one Fe-S flavoenzyme gene were downloaded and quality-filtered using the N50 statistic, setting the threshold at 50 kb. The 16S rRNA sequences from the high-quality scaffolds were predicted using Barrnap version 0.9 (https://github.com/tseemann/barrnap). 16S rRNA sequences were aligned with Clustal Omega[123], and FastTree was used to infer an approximately-maximum-likelihood phylogenetic tree. Finally, iTOL was used to display and annotate the phylogenetic tree. Two 16S sequences from *Bacillus subtilis* subsp. *subtilis* (NR_102783.2) and *Streptococcus agalactiae* DNF00839 (KU726685.1) were used as outgroups to root the tree.

3.4.4 Identification of Fe-S flavoenzymes from a collection of human gut MAGs

The 1,952 uncultured bacterial genomes from Almeida *et. al*[23] were used to evaluate the Fe-S flavoenzymes abundance in human gut bacteria communities. To this end, the genomic nucleotide sequences of the MAGs were downloaded using the ENA accession number ERP108418. Next, the protein sequences were predicted using Prodigal V2.6.3: February, 2016, to further use hmmsearch against the Oxidored_FMN and Pyr_redox_2 domains following the same procedure as stated above. Thus, the number of predicted Fe-S flavoenzymes (protein sequences with sequence e-value <= 1x10-5) for each MAG was calculated to evaluate their Fe-S flavoenzyme copy numbers. Moreover, in order to analyse the genomic context of the uncultured Clostridiales bacterium (GCA_900546485.1) Fe-S flavoenzymes coding genes, a genomic feature table was created from the Prodigal annotations. Subsequently, we used the same script to identify gene clusters as explained below. Finally, MGC protein sequences were scanned with hmmscan to identify Pfam domains.

3.4.5 Analysis of the genomic context of Fe-S flavoenzymes

Using the subset of Fe-S flavoenzymes from the class Clostridia found in the RefSeq database, we inspected genomic context by identifying 'neighboring genes' that met the following criteria: they were encoded on the same strand as the Fe-S flavoenzyme and located within a maximum intergenic distance of 400 bp between subsequent genes. A script was used to parse the feature tables from the genome assemblies found to harbour at least one copy of the Fe-S flavoenzyme gene, from which the starting and ending coordinates of the cluster were extracted. Subsequently, the corresponding Genbank file of the flanking region was downloaded. The resulting Genbank file collection was used as an input for BiG-SCAPE[124], which groups metabolic gene clusters (MGCs) into families. The output networks were visualized using the BiG-SCAPE interactive visualization tool.

3.5 Data summary

Supporting information for this article can be found in Zenodo repository (https://zenodo.org/) under the following DOI: 10.5281/zenodo.3746013. The authors confirm all supporting data, code and protocols have been provided within the article or through supporting information files.

Chapter 4

A systematic analysis of metabolic pathways in the human gut microbiota

Victòria Pascal Andreu, Hannah E. Augustijn, Lianmin Chen
Alexandra Zherkanova, Jingyuan Fu, Michael A. Fischbach,
Dylan Dodd, Marnix H. Medema.

Slightly modified from the preprint available in bioRxiv:
https://www.biorxiv.org/content/10.1101/2021.02.25.432841v1

Abstract

The gut microbiota produce hundreds of small molecules, many of which modulate host physiology. Although efforts have been made to identify biosynthetic genes for secondary metabolites, the chemical output of the gut microbiome consists predominantly of primary metabolites. Here, we introduce the gutSMASH algorithm for identification of primary metabolic gene clusters and use it to systematically profile gut microbiome metabolism, identifying 19,885 gene clusters in 4,240 high-quality microbial genomes. We find marked differences in pathway distribution among phyla, reflecting distinct strategies for energy capture. These data explain taxonomic differences in short-chain fatty acid production and suggest a characteristic metabolic niche for each taxon. Analysis of 1,135 subjects from a Dutch population-based cohort shows that the level of 14 microbiome-derived metabolites in plasma is almost completely uncorrelated with the metagenomic abundance of the corresponding biosynthetic genes, revealing a crucial role for pathway-specific gene regulation and metabolite flux. This work is a starting point for understanding differences in how bacterial taxa contribute to the chemistry of the microbiome.

The pathways encoding the production of microbial metabolites are often physically clustered in the genome, in regions known as metabolic gene clusters (MGCs). Current tools for computational prediction of metabolic pathways focus on gene clusters for natural product biosynthesis[125] or generic primary metabolism[126,79]. Here, we introduce a new algorithm, gutSMASH, to profile known and predicted novel primary metabolic gene clusters from the gut microbiome. We use this tool to perform a systematic analysis of primary metabolic gene clusters in bacterial strains from the gut microbiome, and identify the prevalence and abundance of each of these pathways across a large population-based cohort.

Algorithms that identify physically clustered genes have become a main-

stay of bacterial pathway identification; taking into account the conserved physical clustering of genes prevents false positive hits based on sequence similarity alone. This principle has been widely applied in the field of natural product biosynthesis, e.g. in antiSMASH[125], which predicts biosynthetic gene clusters (BGCs) by detecting physically clustered protein domains using profile hidden Markov Models (pHMMs). Here, we tailored this gene cluster detection framework to detect MGCs involved in primary metabolism and bioenergetics.

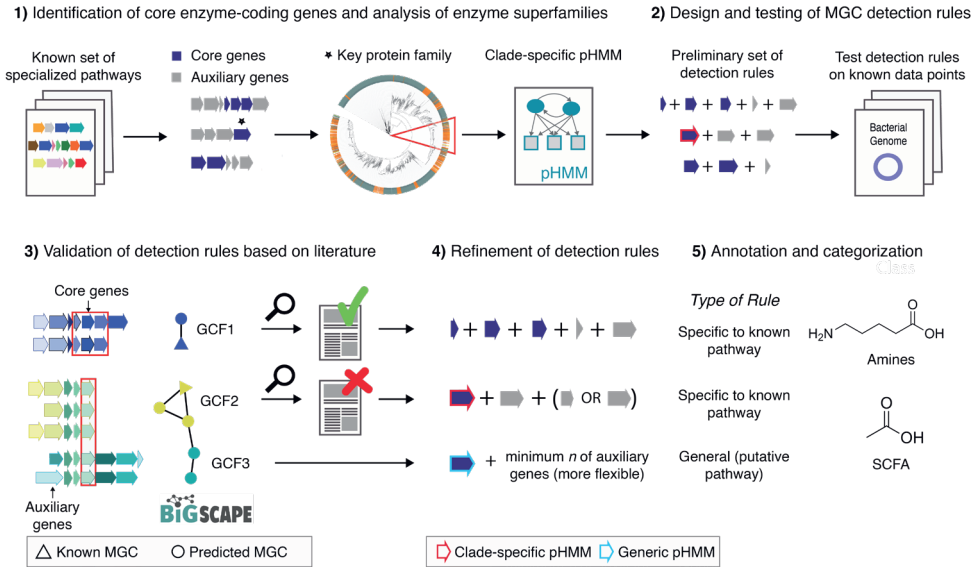


Figure 4.1 Development and design of detection rules for gutSMASH. (1) A set of known and characterized MGC-encoded pathways were curated from the literature. Protein domains were identified across all MGCs and core enzymatic domains were manually identified. For enzymatic domains belonging to broad multifunctional enzyme families, protein superfamily phylogenies were built to create clade-specific pHMMs. (2) These domains were incorporated in the initial detection rules. The detection rules were run on a test set, and all the MGC predicted by the same rule were grouped together and (3) run through BiG-SCAPE, which grouped the MGCs into gene cluster families (GCFs). (4) Based on literature analysis of GCF members, detection rules were manually fine-tuned to either include or exclude MGC architectures that were either related to specialized primary metabolism or not. (5) Finally, fine-tuned detection rules were annotated and categorized into different MGC classes based on their metabolic end products.

As a starting point, we constructed a dataset of 51 primary metabolic pathways from the gut microbiome with biochemical or genetic literature support (including MGCs as well as pathways encoded by a single gene) and identified core enzymes (i.e., required for pathway function) to serve as a signature for the detection rules (Figure 4.1, Table S4.1; see *Methods* for details). To more accurately predict MGCs of interest, we performed three computational procedures. First, for core enzymes belonging to 12 of the protein superfamilies that are known to catalyze diverse types of reactions and were most commonly found across a wide range of pathways, we constructed phylogenies and used them to create clade-specific pHMMs to detect specific subfamilies (see S.4.1.1). Second, we designed pathway-specific rules for each MGC type in our dataset (see *Methods*). These rules were validated and optimized by detailed visual inspection and analysis of MGC sequence similarity networks made using BiG-SCAPE[124], generated from gutSMASH results on a set of 1,621 microbial genomes (Online Data: https://gutsmash.bioinformatics.nl/help.html#Validation; see S.4.1.2) (Suppl. Table 2 & 3). Third, despite the fact that most specialized primary metabolic pathways are encoded in MGCs, there are also single-protein pathways that are in charge of the secretion of key specialized primary metabolites in the gut microbial ecosystem, such as serine dehydratase, which produces ammonia and pyruvate from serine[127]. For this reason, we also built 10 clade-specific pHMMs to detect these (see section S.4.2.6). The above procedures led to the design of a robust set of detection rules to identify both known and putative MGCs that are potentially relevant for metabolite-mediated microbiome-associated phenotypes.

To profile the metabolic capacity of strains from the human gut microbiome, we selected a set of 4,240 unique high-quality reference genomes consisting of 1,520 genomes from the Culturable Genome Reference (CGR) collection[128], 2,308 genomes from the Microbial Reference Genomes collection of the Human Microbiome Project (HMP) consortium[129] and 414 additional genomes from the class Clostridia to account for their metabolic

versatility[130] (Table S4). We refrained from including metagenome-assembled genomes in this analysis, as they often lack the taxon-specific genomic islands[131] on which many specialistic metabolic functions are encoded. In total, gutSMASH predicted 19,885 MGCs across these genomes that are clear homologues of MGCs for our set of known pathway types (See Methods section S.4.2.7).

The combined results of the gutSMASH MGC scanning and the single-protein pHMM detection across the three reference collections provide unique insights into the metabolic traits encoded by the genomes of human gut bacteria. While some genera harbor a small set of highly conserved pathways, (e.g., *Akkermansia*, *Faecalibacterium*), other genera contain much larger interspecies differences (Figure 4.2A). The genus *Clostridium* displays remarkable metabolic versatility, with 42 distinct metabolic pathways present across members of this genus (Figure 4.2A). Clostridial strains that are indistinguishable by 16S sequencing often harbor distinct gene cluster ensembles (Suppl. Figure 4.1), suggesting that specialization in primary metabolism leads to functional differentiation even among closely related strains. *Clostridium* is a clear outlier: by comparison, the next most numerous set of metabolic pathways are found within the Enterobacteriaceae (e.g., *Salmonella, Escherichia, Enterobacter,* and *Klebsiella*) with 22-25 metabolic pathways. Intriguingly, many of the metabolic pathways encoded by *Clostridium* and members of the Enterobacteriaceae are non-overlapping (with 23/42 *Clostridium* pathways not being identified among Enterobacteriaceae), highlighting the distinct metabolic strategies these microbes employ within the gut (Figure 4.2A). The *Bacteroides*, Actinobacteria (*Eggerthella* and *Collinsella*) and Verrucomicrobia (*Akkermansia*) harbor a more restricted set of primary metabolic pathways, likely reflecting versatility in upstream components of their metabolism (i.e., glycan foraging and other forms of substrate utilization).

Our results provide insights into the metabolic strategies that microbes use to produce short chain fatty acids (SCFAs). As expected, butyrate production is found exclusively in certain Firmicutes and Fusobacteria, whereas propionate production is largely confined to (and conserved in) the Bacteroidetes. However, the phylogenetic distribution of pathways that generate acetate -- the most concentrated molecule produced in the gut[132] -- has not yet been described. Two pathways for the conversion of pyruvate to acetate -- pyruvate formate-lyase (pyruvate to acetate/formate) and pyruvate:ferredoxin oxidoreductase (PFOR) -- are widely distributed across microbial strains from diverse phyla (Figure 4.2B). Two observations suggest that these two pathways are the most prolific source of acetate in the gut. First, some strains known to produce large quantities of acetate rely entirely on one or both of the pathways. Second, each one uses pyruvate as a substrate, consistent with a model in which these pathways are the primary conduit through which carbohydrate-derived carbon is converted to acetate. Additional taxon-specific pathways for acetate include the CO_2 to acetate pathway and the glycine to acetate pathway (each specific to a subset of Firmicutes), as well as the choline and ethanolamine utilization pathways (widespread among Enterobacteriaceae and each found in different clades of Firmicutes) (Figure 4.2A).

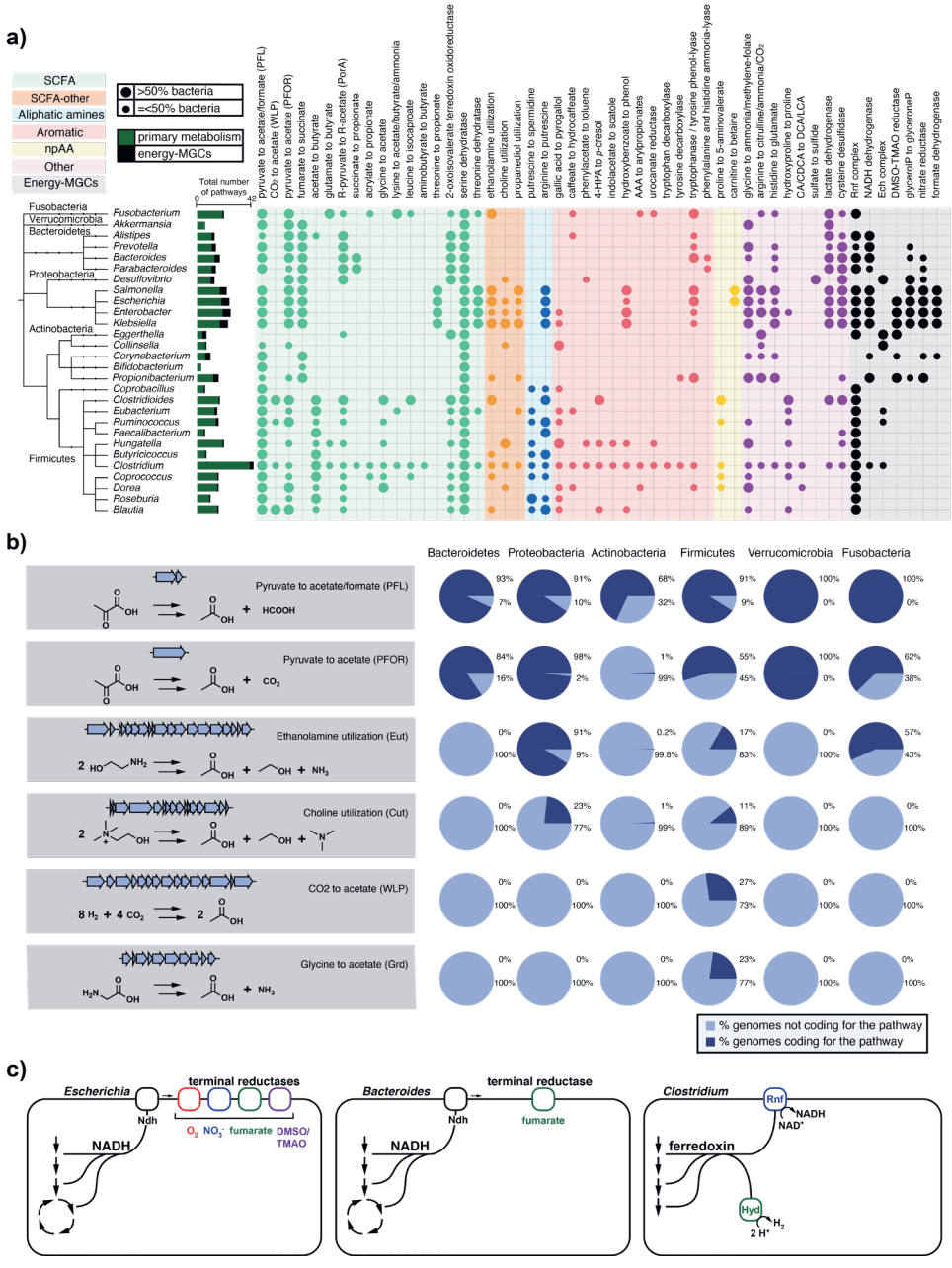


Figure 4.2 Distribution of known pathways across most representative genera in the human gut. (A) Circles represent the absence/presence of known pathways in each genus. Larger circles indicate cases in which more than 50% of the genomes for a genus encode the pathway, while smaller circles indicate cases in which 50% or fewer of the genomes encode it. Colored ranges indicate a categori-

zation of MGCs by chemical class of their product, in which npAA represents non-proteinogenic amino acids and SCFA represents short-chain fatty acids. Taxonomic assignments were applied using the Genome Taxonomy Database (GTDB)[96]. The tree was generated using phyloT (https://phylot.biobyte.de/) and visualized using iTOL[122]. Raw data are available in Table S4.5. (B) Distribution of the main acetate synthesis pathways at phylum level. Some of the pathways are ubiquitous across the five major phyla (e.g. pyruvate to acetate/formate [PFL]), while others are only found in Firmicutes (CO2 to acetate [WLP]). Raw data for the pie charts is available in Table S6. Genes and gene clusters depicted are representatives from *Bacteroides thetaiotaomicron* (PFL & PFOR), *Salmonella enterica* (Eut), *Clostridium sporogenes* (Cut), *Clostridium difficile* (WLP) and *Clostridium sticklandii* (Grd). (C) Bioenergetic strategies in *Escherichia* that has a variety of alternate electron acceptors to choose from compared to *Bacteroides* and *Clostridium*. Abbreviations: PFL, pyruvate formate-lyase; PFOR, pyruvate:ferredoxin oxidoreductase; Eut, ethanolamine utilization; Cut, choline utilization; WLP, Wood-Ljungdahl Pathway; Grd, glycine reductase; CA, cholic acid; CDCA, chenodeoxycholic acid; DCA, deoxycholic acid; LCA, lithocholic acid; TMAO, trimethylamine N-oxide; DMSO, dimethylsulfoxide; SCFA, short-chain fatty acid; Ndh, NADH dehydrogenase, Rnf, Rhodobacter nitrogen fixation like complex; Hyd, hydrogenase.

Our results demonstrate a striking difference in mechanisms for energy capture by three of the major bacterial genera in the gut: *Bacteroides*, *Escherichia*, and *Clostridium*. When growing aerobically with glucose, *E. coli* generates most of its energy by channelling electrons through membrane bound cytochromes using oxygen as the terminal electron acceptor (Figure 4.2C). However, oxygen is limiting in the gut. Under anaerobic conditions, bacteria from the genus *Escherichia* employ alternate terminal electron acceptors such as nitrate, DMSO, TMAO, and fumarate by substituting alternate terminal reductases into their electron transport system (Figure 4.2C). However, in the healthy gut these alternate electron acceptors are either absent or available in limited amounts, likely explaining why these facultative anaerobes represent a small proportion of the healthy microbiome[133]. In contrast to the diversity of terminal reductases used by the *Escherichia*, *Bacteroides* genomes encode only fumarate reductase (Figure 4.2C). They use a unique pathway, carboxylating PEP to form fumarate, which they use

as a terminal electron acceptor to run an anaerobic electron transport chain involving NADH dehydrogenase and fumarate reductase, ultimately forming propionate. Thus, the metabolic strategy employed by *Bacteroides* ensures a steady stream of electron acceptor to fuel their metabolism. The *Clostridium* do not utilize similar mechanisms for energy capture as the *Escherichia* and the *Bacteroides*. Recent analyses suggest that they use the Rnf complex for generating a proton motive force. Several pathways encoded by the genomes of *Clostridium* (e.g., acetate to butyrate, AAA to arylpropionates, leucine to isocaproate) (Figure 4.2A) consist of an electron bifurcating acyl-CoA dehydrogenase enzyme. This complex bifurcates electrons from NADH to the low potential electron carrier ferredoxin which can then donate electrons to the RNF complex which functions as a proton or sodium pump, generating an ion motive force. Although much still is to be learned about Clostridial metabolism, our findings suggest that their metabolism operates at a different scale of the redox tower compared to *Bacteroides* and Enterobacteriaceae, using low potential electron carriers to fuel their metabolism.

Next, we set out to determine the prevalence and abundance of each pathway in a cohort of human samples. We used BiG-MAP[134] to profile the relative abundance of each MGC class across 1,135 metagenomes from the population-based LifeLines DEEP cohort[135], by mapping metagenomic reads against a collection of 6,836 non-redundant MGCs detected in our set of reference genomes (Figure 4.3A,B). Some pathways, such as CO_2 to acetate (acetogenesis) and butyrate production from acetate or glutamate, as well as polyamine-forming pathways, were found in >99% of microbiomes. Others, such as 1,2-propanediol utilization and *p*-cresol production, both associated with negative effects on gut health[136,137], were observed at detectable levels in only half of the samples. In terms of abundance, it is striking that for example the bile acid-induced (*bai*) operon for the formation of the secondary bile acids deoxycholic acid and lithocholic acid, which has been characterized from very low-abundance *Clostridium scindens* strains[110], was still shown to be present in relatively high abundance

across a subset of subjects. Analysis of the mapped reads showed that the vast majority of these mapped to a homologous MGC from the genus *Dorea* instead (Suppl. Figure 4.2), for which the physiological relevance remains to be established. It is also interesting to see that, while two of the three acetate-forming pathways (PFL and WLP) were consistently found at high abundance levels, the abundance of all butyrate-forming pathways is highly variable across subjects, with a >20-fold difference between lower and upper quartiles in the abundance distribution of the glutamate-to-butyrate pathway, and a >440-fold difference between the 10th percentile and the 90th percentile.

The wide variability in the metagenome abundance of each pathway raises the question of whether metagenomic abundance of a pathway correlates with the level of its small molecule product in the host. To address this question, we systematically compared the level of each pathway with the quantity of the corresponding metabolite as determined by plasma metabolomics. We find a striking lack of correlation between pathway and metabolite levels (*r* ranging from -0.04 to 0.24, Figure 4.3C). These data indicate that gene abundances in metagenomes are not (on their own) a useful predictor of metabolic outputs. This finding has important implications for analyses that make metabolic inferences from gene abundances[119] or the abundances of individual strains[138]. We speculate that a more detailed understanding of the influence of diet, differences in gene regulation, characteristic pathway flux (turnovers per unit time per protein copy), and pharmacokinetic characteristics (e.g., absorption, distribution, metabolism, and excretion) could ultimately enable the prediction of metabolite abundance from metagenome abundance. The systematic detection of the relevant genes and gene clusters by gutSMASH will provide a technological foundation for future studies in this direction, by allowing mapping of metatranscriptomic data to these accurately defined and categorized sets of genomic loci in order to understand which conditions and interactions are driving the expression of these pathways and the accumulation of their products.

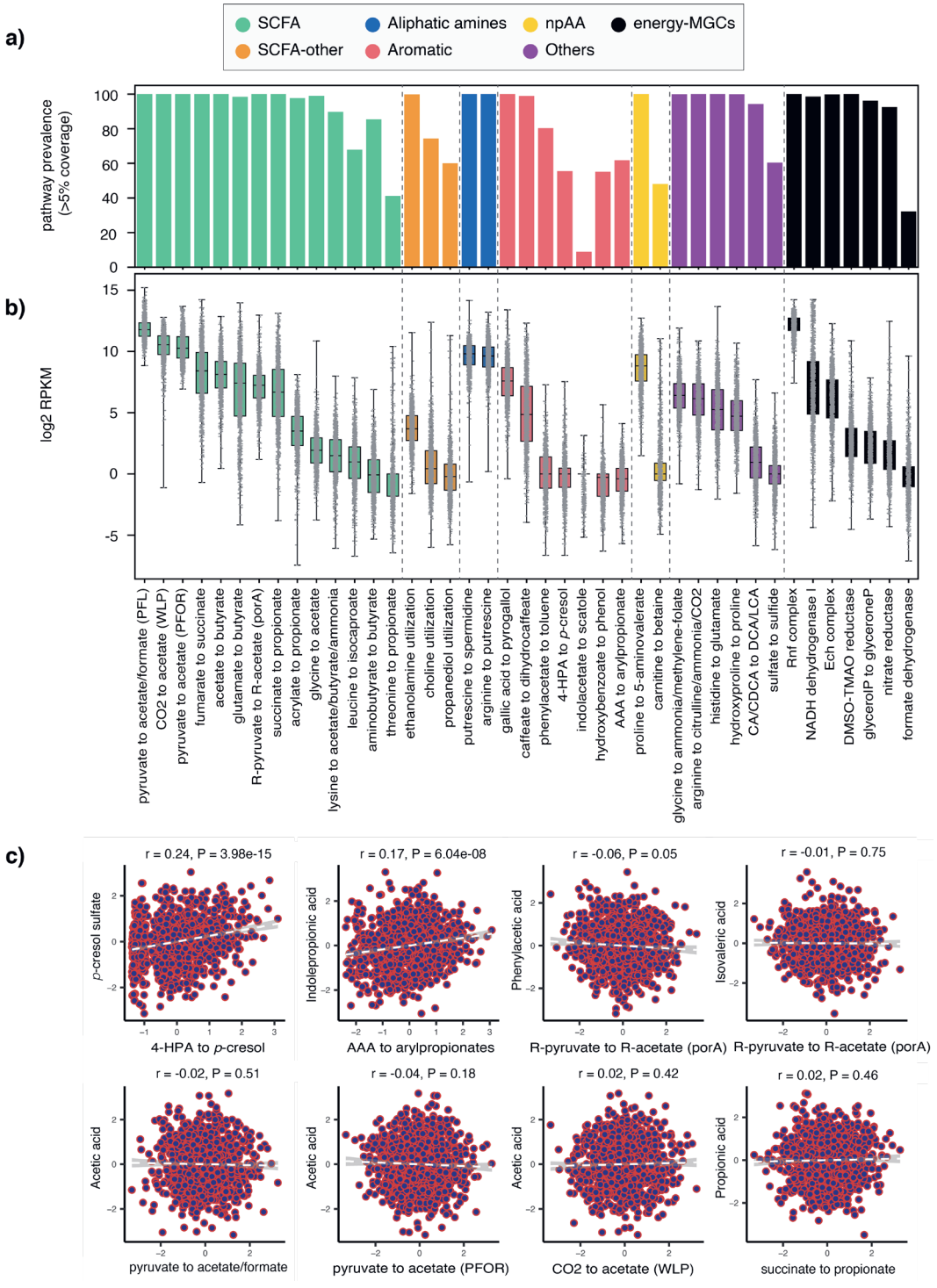

a)
SCFA
SCFA-other
Aliphatic amines
Aromatic
npAA
Others
energy-MGCs
pathway prevalence (>5% coverage)
b)
log2 RPKM
pyruvate to acetate/formate (PFL)
CO2 to acetate (WLP)
pyruvate to acetate (PFOR)
fumarate to succinate
acetate to butyrate
glutamate to butyrate
R-pyruvate to R-acetate (porA)
succinate to propionate
acrylate to propionate
glycine to acetate
lysine to acetate/butyrate/ammonia
leucine to isocaproate
aminobutyrate to butyrate
threonine to propionate
ethanolamine utilization
choline utilization
propanediol utilization
putrescine to spermidine
arginine to putrescine
gallic acid to pyrogallol
caffeate to dihydrocaffeate
phenylacetate to toluene
4-HPA to p-cresol
indolacetate to scatole
hydroxybenzoate to phenol
AAA to arylpropionate
proline to 5-aminovalerate
carnitine to betaine
glycine to ammonia/methylene-folate
arginine to citrulline/ammonia/CO2
histidine to glutamate
hydroxyproline to proline
CA/CDCA to DCA/LCA
sulfate to sulfide
Rnf complex
NADH dehydrogenase I
Ech complex
DMSO-TMAO reductase
glycerolP to glyceroneP
nitrate reductase
formate dehydrogenase
c)
r = 0.24, P = 3.98e-15
p-cresol sulfate
4-HPA to p-cresol
r = 0.17, P = 6.04e-08
Indolepropionic acid
AAA to arylpropionates
r = -0.06, P = 0.05
Phenylacetic acid
R-pyruvate to R-acetate (porA)
r = -0.01, P = 0.75
Isovaleric acid
R-pyruvate to R-acetate (porA)
r = -0.02, P = 0.51
Acetic acid
pyruvate to acetate/formate
r = -0.04, P = 0.18
Acetic acid
pyruvate to acetate (PFOR)
r = 0.02, P = 0.42
Acetic acid
CO2 to acetate (WLP)
r = 0.02, P = 0.46
Propionic acid
succinate to propionate

Figure 4.3 Prevalence and abundance of specialized primary metabolic pathways across 1,135 human microbiome samples. (A) Prevalence of each of the 41 known pathway classes across all microbiomes, measured as the percentage of samples in which core enzyme-coding genes of at least one reference MGC belonging to a given class were covered by metagenomic reads across >5% of their sequence length. This cutoff was kept low to avoid false negatives due to limited sequencing depth for low-abundance taxa (raw data available at Table S8). (B) Distributions of log2 RPKM relative abundance values of all 41 known pathway classes, categorized by product class, across all LifeLines DEEP metagenomes (raw data available at Table S9). All samples are represented by a dot in the box plot, representing the log2 RPKM value for a given sample. The box limits indicate the quartiles of the dataset while the whiskers extend to 1.5x the interquartile range; center line denotes the median. (C) Limited correlation of genetic pathway abundance with abundance of metabolites in blood plasma.

The gutSMASH software constitutes, to our knowledge, the first comprehensive automated tool designed to identify niche-defining primary metabolic pathways from genome sequences or metagenomic contigs—even a full-fledged metabolic network reconstruction software like PathwayTools[139] (which uses the extensive MetaCyc database[140]) lacks detection capabilities for 3 out of the 41 MGC-encoded pathways detected by gutSMASH (Table S7). Moreover, the identification of MGCs provides considerably increased confidence that detected homologues for a given pathway are truly working together. Downstream, detected MGCs can be used as input for read-based tools such as HUMAnN[141] or BiG-MAP[134] to measure abundance or expression levels of the encoded pathways. On top of these functionalities, the gutSMASH framework also facilitates identifying new (i.e., uncharacterized) pathways in the microbiome. To this end, we designed an additional set of rules to detect primary metabolic gene clusters of unknown function that harbor Fe-S flavoenzymes[142], glycyl-radical enzymes, 2-hydroxyglutaryl-CoA-dehydratase-related enzymes, and/or enzymes involved in oxidative decarboxylation. From this analysis of putative MGCs (see section S.4.2.8), we found 12,259 putative MGCs from 760 different species, that,

after redundancy filtering at 90% sequence similarity, were classified into 932 GCFs. Within these, we manually prioritized a range of gene clusters with unprecedented enzyme-coding gene content (see Figure 4.4 and Suppl. Figures 4.4 and section S.4.1.2). These can be a potential new source to discover new pathways and metabolites.

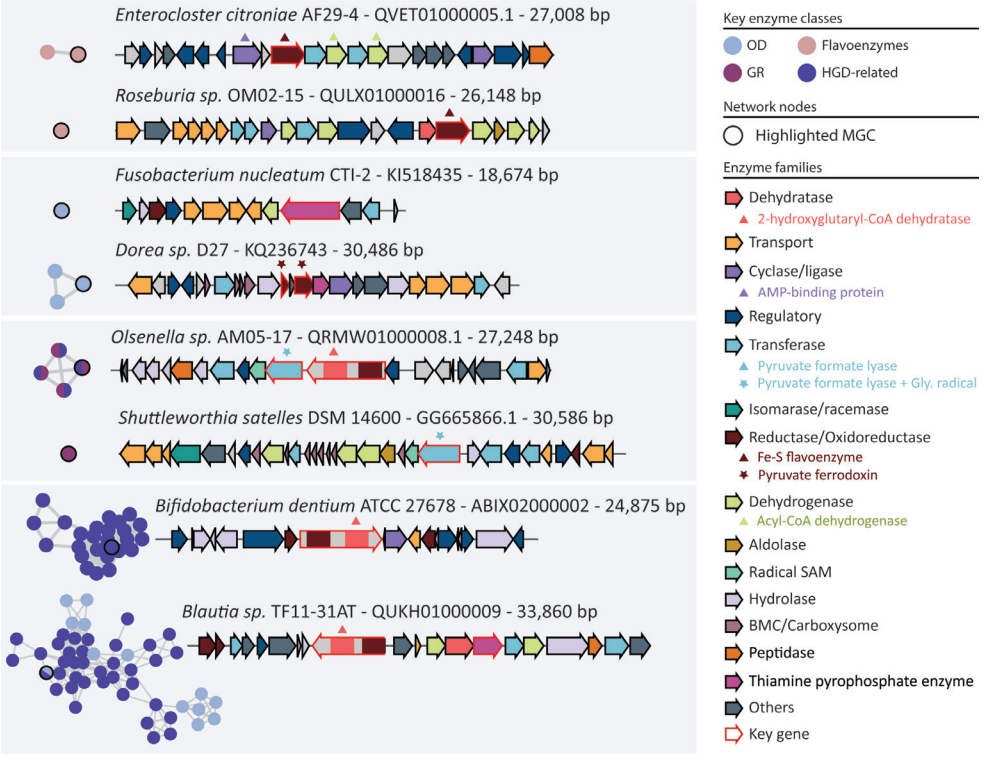


Figure 4.4 Subset of unknown MGCs predicted by gutSMASH manually picked. The network/nodes present in the left side of the figure represent the subnetwork extracted from the complete network in Supplementary Figure 4.4. The arrows have been coloured-coded based on the Pfam domains found in the protein-coding sequences and the functional annotations of these proteins.

4.1 Data Availability

The LifeLines-DEEP metagenomic and metabolomic data are available at the European Genome-phenome Archive under accession EGAS00001001704.

4.2 Code Availability

The gutSMASH source code is available freely under an open-source AGPL-3.0 license from https://github.com/victoriapascal/gutsmash/ .

S.4.1 Supplementary Results

S.4.1.1 Phylogenetic analysis of protein superfamilies to identify pathway-specific clades

Performing hmmscan searches on the protein sequences helped identify the presence of keystone domains (Pfams) between pathways that share an enzymatic core. However, some of these enzymes are part of multifunctional enzyme families that are recognised by very broad Pfam domains. Thus, in order to more accurately identify relevant functional subgroups for keystone enzyme families, we used protein phylogenetic analysis to pinpoint pathway-specific clades and discern, for instance, specialized primary metabolism-related enzymes from housekeeping related ones (Figure 1). For each protein family, we created a non-redundant set of representatives by gathering protein sequences from four different sources: the proteins from the representative pathway, the reference proteome available in Pfam[143], the respective proteins from the MGC collection and the experimentally characterized proteins available in UniProt, to assign functionality to clades (see S.4.2.2). In total, we performed this phylogenetic analysis to 12 major protein superfamilies. For instance, phenyllactate dehydratase homologues are involved in the degradation of aromatic amino acids into propionate, in the

acrylate to propionate pathway and in the leucine reductive branch pathway; in contrast, 2-hydroxyglutaryl-CoA dehydratase allows the transformation of glutamate into butyrate. Despite the fact that these reactions are different, both key enzymes harbour the 2-hydroxyglutaryl-CoA dehydratase, D-component domain (HGD-D Pfam domain, PF06050). The HGD-D protein family phylogenetic tree (see Suppl. Figure 4.3) revealed that 10 clades were implicated in these three pathways. Subsequently, 10 profile hidden Markov models (pHMMs) specific to these clades were built and, after assessing the sensitivity of the models (see Methods), 9 out of 10 were selected for being sensitive enough as to correctly identify the subdomains of this protein family involved in these 3 pathways. Consequently, these 9 pHMMs were included in the corresponding detection rule. The same procedure was followed for the other protein superfamilies, creating a total of 43 pHMMs (see Table S10). As a result, gutSMASH uses newly built pHMMs in combination with the ones included in the Pfam database to identify proteins families of interest. Therefore, gutSMASH competitively scores hmmsearch hits and assigns to the sequence the domain with a higher score. Altogether, this procedure allowed to define a preliminary set of detection rules using the newly built pHMMs to predict close homologues of known gene clusters.

S.4.1.2 Validation of gutSMASH detection rules by evaluating their predictive performance

Evaluating gutSMASH performance implies having a set of bacteria whose genomes are known to encode a given pathway. However, the false positive rate is unknown (as it is not feasible to experimentally verify large numbers of diverse putative annotations) and the false negative rate is difficult to determine (as only in few species, MGCs with experimentally proven functions are described in literature). Hence, as a first step in validation, we decided that the best course of action would be to perform detailed manual analysis of large numbers of diverse predicted MGCs. For this reason, gutSMASH was run on a test set created from 1,632 bacterial genomes. All the MGCs

predicted by the same detection rule were grouped together to further run BiG-SCAPE on each subset (see section S.4.2.3). In this manner, we could manually evaluate the range of gene clusters predicted by the same detection rule and quickly find out if any distantly-related gene cluster should not be picked by the rule based on, e.g., having divergent enzyme-coding composition. Also, this procedure allowed us to acquire an overview of the bacterial taxa predicted to possess a given gene cluster type and identify if any MGCs from taxa referenced in primary literature were missing from this set. Thus, this system allowed to fine-tune the detection rules and evaluate their predicting potential. After several iterations of adjusting rules, performing new predictions and creating new sequence similarity networks, we froze gutSMASH version 1.0 with 41 specific-to-known-pathway detection rules (see Table S11) to accurately and comprehensively predict MGCs of known function.

An additional validation step was performed using PaperBLAST[144], which was used to look for genomes encoding any experimentally characterized homologues of the key proteins involved in gutSMASH-predicted pathways; all corresponding 19 identified MGCs (representing 16 different pathways, and which can be treated as proven true positives) were successfully detected using gutSMASH (see Table S2). When compared to the reference MGCs, these detected clusters showed an average amino acid sequence identity between 54 and 100% and an overall gene cluster similarity (percentage of homologues detected in KnownClusterBlast) ranging from 44 to 100%.

S.4.1.2 Analysis of putative clusters and distant homologues: relevant candidates to study further

Next, we evaluated the potential of gutSMASH to predict putative MGCs of interest and to explore the metabolic landscape covered by them. For this reason, the 12,259 putative clusters predicted from the HMP, CGR and

Clostridioides genomes were used and subjected to a redundancy filtering of 90% similarity at the protein sequence level (see Methods section S.4.2.7) to investigate their functional diversity. To create a non-redundant collection, two random representative clusters of each set of highly similar clusters were picked; all representatives were then clustered together into 932 GCFs using BiG-SCAPE[124] (see Methods: *Analysis of distant homologues and putative MGCs from CGR, HMP and Clostridioides dataset*). From the resulting network (see Suppl. Figure 4.4), we made three main observations. First, we identified several distant homologues of the known MGCs; these are picked by specific pathway rules but classified as putative when for instance the MGC is from a distantly related taxonomic group and therefore shares low sequence similarity with the reference gene cluster. Second, we found previously characterized MGCs that were not included in the original training sets of known MGC types, as for instance a region from *Ruminococcus gnavus strain AM22-7AC* (accession number QRIA01000012.1) involved in the metabolism of rhamnose and fucose described in 2013 by E.Petit & W. Latouf *et al.*[145]; this validated the capability of gutSMASH to identify real MGCs that were not included in the initial list of known pathways. Third, we observed vast numbers of novel clusters of unknown function that represent good candidates for further experimental characterization. The examples shown in Suppl. Figure 4.4 are metabolically diverse MGCs that encode flavoenzymes, oxidative decarboxylation (OD), glycyl radical (GR), 2-hydroxyglutaryl-CoA dehydratase (HGD-D) related or hybrids of them, which were found in phylogenetically diverse genomes from Firmicutes, Fusobacteria and Actinobacteria. Moreover, all these MGCs presented plausible architectures as to be real gene clusters, since they included all the genetic elements to regulate, synthesize and transport the resulting molecule (or its substrates). The first MGC for instance, which is found in *Enterocloster citronae* AF29-4, shares some similarity with the acetate to butyrate MGCs, since it encodes an acyl-CoA dehydrogenase and two electron transfer flavoproteins as well as a distant *baiH* homologue. Another interesting example is the *Dorea sp.* D27 MGC, a pathway involving a pyruvate:ferre-

doxin oxidoreductase, which might be encapsulated given the presence of bacterial microcompartment genes (BMC-encoding genes) in the cluster. *Blautia sp.* TF11-31AT also presents an unprecedented gene cluster architecture, encoding a combination of enzymes involved in oxidative decarboxylation (thiamine pyrophosphate-related) and related to HGD-D. This overview highlights the ability of gutSMASH to systematically predict novel and interesting gene clusters from a diverse range of bacteria that could help associating function to unknown genes and predict novel pathways. Additionally, the expression of families of MGCs of unknown function could potentially be correlated to microbiome-associated phenotypes to prioritize them for experimental characterization based on physiological and ecological relevance.

S.4.1.3 Assessing pathway abundance and prevalence across metagenomes

From the predicted pool of known pathways, we aimed to assess their abundance and prevalence by mapping the metagenomic reads of the LifeLines-DEEP cohort[135,146]. To compute both, certain assumptions had to be made and thresholds needed to be chosen. In both cases, the numbers of reads mapping to the core regions are key to assess whether a given pathway is present or not and how abundant it is. In order to account for spurious mapping, we designed an approach to assess the pathway abundance by using the lower quartile number of reads mapping to 2kb long regions for each pathway and sample (see more in section S.4.2.9). In contrast, to evaluate the prevalence of all the pathways across samples, the core coverage score estimated by BiG-MAP was used. Ideally, with unlimited sequencing depth, one would expect to find reads mapping evenly to the whole gene cluster, thus, implying a relatively high core coverage score compared to the chosen one (>5% coverage, see Figure 5) and also account for spurious mapping. However, when raising the minimum coverage value from 10-80% some of the pathways that are known to be present in healthy individuals

showed a prevalence of 0 (see Suppl. Figure 4.5). Assessing the minimum sequence identity of a read mapping to a gene cluster showed 78% similarity (at nucleotide level), which confirms that even with low coverage scores, the reads are mapping very specifically to the gene clusters and hence, the lack of coverage of some gene clusters is very likely due to the sequence depth of a sample rather than the absence of the pathway. For this reason, we set the minimum coverage used in the analyses presented in the main text at 5%, to avoid undue false negatives.

S.4.2 Methods

gutSMASH is a Python-based pipeline that has been built from antiSMASH version 5.0 source code. The latest command line version is freely available and can be downloaded and installed from here: https://github.com/victoriapascal/gutsmash/tree/gutsmash

S.4.2.1 Finding pathway signatures for a collection of known and characterized MGCs

To create a new set of detection rules, 41 known and characterized MGCs were gathered from literature and used as positive controls. The protein sequences of these MGCs were searched using hmmscan (HMMER suit version 3.1b2, February 2015; http://hmmer.org/). From the resulting pHMM profile hits, auxiliary and core domains were manually identified for each pathway, to ultimately determine the pathway signature and specify it in the corresponding detection rule. To discern and more precisely identify key enzymes of interest sharing a keystone domain, we used custom-made pHMMs following a procedure described in section S.4.2.2. Altogether, the knowledge on the core enzyme coding-genes and the newly-built pHMMs helped to construct a preliminary set of detection rules to predict known pathways.

S.4.2.2 Towards a more robust MGC identification by building new HMM profiles

Certain core domains are shared across diverse pathways, including the PFL-like domain and the HGD-D domain. In total, 13 keystone domains were found to be ubiquitous in multiple pathways (see Suppl. Table 10). Hence, to increase gutSMASH precision and discern between enzyme subfamilies of interest, 12 protein superfamily phylogenies were constructed by aligning the protein sequences harbouring the domain of interest from the MGC collection (described in section S.4.2.5), the respective reference proteome[143] at a 15% or 35% co-membership threshold (the latter only for the domains Gly_radical and Acyl-CoA_dh_1) and any experimentally characterized UniProt representatives. After aligning the sequences with Clustal Omega[123], approximately-maximum-likelihood phylogenetic trees using FastTree 2.1[120] were inferred to further annotate the tree with iTOL[122]. Thus, from the desired and functionally relevant clades, specific pHMMs were built by extracting the amino acid sequence of the clade-specific proteins, aligning them with Clustal Omega, trimming the edges of the multiple sequence alignment using Jalview[147], re-aligning all the sequences with Clustal Omega and finally building a pHMM using hmmbuild (HMMER suite version 3.1b2, February 2015; http://hmmer.org/). Subsequently, for all the newly created pHMMs, sensitivity was assessed using 10-fold jackknife cross-validation. Each clade was divided randomly into training and testing sets. The protein sequences from the training set were aligned using Clustal Omega and used to create a pHMM. Next, the protein sequences of the test set were hmmscanned (HMMER suit version 3.1b2, February 2015; http://hmmer.org/) against the newly built testing pHMMs. When a sequence scored positively for multiple domains in the same region, only the domain with a higher bit score was picked out. Sensitivity then accounted for the number of sequences positively associated with the correct pHMM out of the total number of sequences in the testing set. The same procedure was repeated 10 times.

The pHMMs with a true positive rate higher than 0.85 across the 10 rounds were included in the detection rules. In total, 43 newly built pHMMs were included in the corresponding detection rules. Moreover, a pHMM to capture succinate dehydrogenase/fumarate reductase was built by aligning 10 protein sequences of such enzymes and building the model from this alignment using a hmmbuild. To also competitively score similar Pfam domains, HHsearch pre-computed results obtained from the Pfam FTP (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/current_release/database_files/) were parsed and included in the gutSMASH code.

S.4.2.3 Testing and validating gutSMASH specific-to-known-pathway detection rules

To evaluate the performance of the preliminary set of detection rules, a total of 1,621 bacterial genomes, including 1,520 genomes from the CGR collection[128] and 101 manually selected genomes from the most representative bacterial genera in the human gut, were used as input for gutSMASH (see Table S3). The predicted MGCs were classified based on the detection rule they were predicted from, to later run BiG-SCAPE on each sub-collection. The resulting networks were screened individually to evaluate the taxonomic and architectural diversity, to assess if any architectural variant or taxon (based on literature) was missing from the MGC pool or was incorrectly predicted by the detection rule. Hence, this procedure ultimately helped to tweak the detection rules to predict true homologues of the known pathways. After two iterations of fine-tuning and testing, all detection rules were performing as intended and constituted the new set of detection rules of gutSMASH version 1.0.

S.4.2.4 gutSMASH customized databases and output visualization

The antiSMASH version 5.0 source code was further tailored to meet gutSMASH functionality. The 32,144 predicted MGCs obtained from running gutSMASH on the CGR, HMP and Clostridiales collections (see section S.4.2.7), were used to create the ClusterBlast database. In a similar way, 59 positive controls carrying the known pathways (from which we created the specific-to-known-pathway detection rules) were used to create the KnownClusterBlast database. These databases facilitate comparative gene cluster analysis using BLAST[148]. Thus, they allow assessing how broadly distributed an MGC is across bacteria (in the case of ClusterBlast) or evaluating the similarity between the predicted MGC and a known and functionally characterized MGC (when using KnownClusterBlast).

Another functionality of antiSMASH is to classify coding genes based on the domains into five major functional categories: core biosynthetic, additional biosynthetic, transport-related, regulatory, resistance and other, using the pmCOG (primary metabolism Clusters of Orthologous Groups) tool, which is embedded in antiSMASH (there originally named smCOG for 'secondary' metabolism Clusters of Orthologous Groups). Thus, the pHMM library pmCOG uses was updated to include relevant domains found in specialized primary metabolism. Also, two other important functional categories were added: electron transport-related genes and encapsulation genes.

S.4.2.5 Exploring the yet unknown metabolic diversity by creating general detection rules

With the objective of creating general detection rules to predict putative MGCs, a similar approach used to screen the surrounding genes around a Fe-S flavoenzyme coding gene was used[142]. Some of the representative known pathways share proteins with biochemically similar functions; these

include, for instance, pyruvate formate lyase-like enzymes that are found in the threonine-to-propionate pathway, the choline utilization pathway and the pyruvate-to-acetate pathways. In order to cover a large amount of sequence diversity, we created a database that included 11,000 complete genomes and 98,886 draft genomes available in Genbank (in February 2017) in order to use clusterTools[149], a software to find remote homologues of known MGCs. As input, a subset of the known pathways used to design the detection rules for known pathways were used as input (see Table S12). The output of several iterated clusterTools searches were grouped to acquire a collection of over 29,000 clusters. For visualization and manual scoring purposes, MultiGeneBlast[150] was run using the clusterTools output as input. Thus, MGCs harbouring at least half of the genes from the query gene list and with a cumulative BLAST score higher than 1,000 were included in the MGC collection. In order to filter out redundant sequences, we used MMseqs2[114] at a 95% similarity cut-off. From the resulting network of 1,599 groups, a maximum of 1 random representative plus singletons were picked creating a 'non-redundant' set of almost 3,200 clusters. This collection was screened for gene clusters harbouring the *baiCD* or *baiH* coding gene (Oxidored_FMN and Pyr_redox_2), pyruvate-formate lyase (PFL-like or Gly_radical), pyruvate ferredoxin (POR, POR_N or PFOR_II), thiamine pyrophosphate enzyme (TPP_enzyme_C) and 2-hydroxyglutaryl-CoA dehydratase (HGD-D), each of which are keystone domains in charge of important anaerobic reactions. This helped creating general detection rules, by identifying which other enzyme-coding Pfam domains are found around these 'anchor' domains in flanking regions; this was systematically analyzed per gene cluster family to make sure that the general rules captured all major families of homologous MGCs of interest. Also, when validating the specific-to-known pathway detection rules, whenever a specific rule predicted interesting MGCs that were variants of the representative pathway with likely differing functions, a general rule was created out of the specific one by loosening up the Pfam requirements. The full list of general rules can be found in Table S13.

S.4.2.6 Assessing single-protein pathway abundance within representative human gut bacteria

To include single-protein pathways in our analysis to assess the overall abundance of specialized primary metabolic pathways, 10 enzyme families were selected for downstream analysis. Following the same procedure as described in section S.4.2.2, protein phylogenies were built for each protein superfamily. Similarly, from the pathway-specific monophyletic clades, we built new pHMMs. A bitscore threshold for each newly built pHMM was calibrated in order to identify with high confidence proteins belonging to the same functional clades. To this end, the protein sequences that composed the superfamily phylogeny were subjected to an hmmsearch run with the new pHMM. The bitscore reported by hmmsearch for the most distantly related protein within the pathway-specific clade was chosen as the threshold for that specific pHMM. Next, the protein sequences from the CGR, HMP and Clostridiales collections (further information in section S.4.2.7) were scanned using the newly built pHMMs. Finally, the hmmsearch output tables for each pHMM were parsed so that the proteins with a bitscore equal or higher to the chosen threshold were deemed hits. In those cases in which the single-protein sequence codes for two Pfam domains, as for instance the serine dehydratase (SDH_alpha and SDH_beta), one of the Pfam domains was selected to create a protein phylogeny to further build a clade-specific pHMM, in this case SDH_alpha. Then, the protein sequences from the three collections were subjected to hmmsearch runs with both the clade-specific pHMM and the other co-occurring Pfam domain (in this case SDH_beta). The sequences that harbour both the specific pHMM at the chosen threshold and the co-occurring domain with an e-value =< 10^{-05} were deemed hits.

S.4.2.7 Evaluating the functional potential of the human microbiome using gutSMASH

To evaluate the metabolic potential of the human microbiome, gutSMASH was run on three different genome collections: (1) the CGR collection, with 1,520 CGR genomes deposited under the PRJNA482748, (2) the HMP reference genomes, with 2,146 HMP bacterial genomes downloaded in September 2019 from here: https://www.hmpdacc.org/hmp/catalog/grid.php?-dataset=genomic and (3) 414 Clostridiales complete genomes under the taxid 186802. The genomic FASTA sequence of these genomes was used as input for gutSMASH, which used Prodigal[151] to annotate genes across all of them in a consistent way. Moreover, in order to assess which MGC belonged to known pathways, the KnownClusterBlast (see S.4.2.4) option was enabled. Thus, from the KnownClusterBlast output, the predicted regions were classified as known when the following two requirements were met: (1) an overall pathway similarity of at least 50% and at least half of the genes with a minimum protein sequence similarity of 40% or (2) an overall similarity of 60% and half of the genes with protein sequence similarity higher than 30%. However, in order not to penalize MGCs with similar domain profiles but substantially larger sizes, the requirements to be considered "known" slightly changed for the KnownClusterBlast MGCs longer than 17 coding genes. In this case, the same requirements as described above were used but instead of considering candidates with at least half of the coding genes having either 30 or 40% minimum sequence identity, one third of the genes were required to be present with the same minimum sequence identity. This was the case for the ethanolamine utilization operon, the *bai* operon characterized from *C. scindens* ATCC35704 (CA/CDCA to DCA/LCA pathway), the acetyl-CoA pathway (CO2 to acetate (WLP)), the tetrathionate to thiosulfate pathway and the NADH dehydrogenase I complex. Thus, all the MGCs that did not satisfy these conditions were classified as putative MGCs. The phylogenetic tree in Figure 2 was generated using phyloT v2 (https://phylot.biobyte.de/). The GDTB database[96] was used to assign the taxonomy to the genomes of the three collections (when present) and those taxonomic identifiers were the ones used for the subsequent pathway absence/presence analysis. Finally, the tree was annotated using iTOL[122].

S.4.2.8 Analysis of distant homologues and putative MGCs from CGR, HMP and Clostridioides dataset

The putative MGCs predicted from the CGR, HMP and Clostridiales genome collections were selected following the definition of "known" and "putative" gene clusters stated in section S.4.2.7. To account for redundant MGCs, protein sequences extracted from all gene clusters were subjected to a redundancy filtering of 90% sequence similarity using MMseqs2. From the resulting clustering, two random representatives were chosen from each group, including the singletons. The resulting non-redundant collection of 3,040 putative MGCs was used as input for BiG-SCAPE using the default thresholds. The network in Suppl. Figure 4.4 was constructed and annotated using Cytoscape[152].

S.4.2.9 Mapping metagenomics reads from healthy samples to the known gutSMASH predicted MGCs

The HMP, CGR and Clostridiales-predicted MGCs were used as input for BiG-MAP[134], a tool that assesses gene cluster abundance or expression across metagenomics or metatranscriptomics data, respectively, by mapping the genomic reads onto the gene cluster sequences. The BiG-MAP family module grouped the 32,144 MGCs into 6,836 GCFs. Next, the reads of 1,135 participants of the population-based cohort LifeLines-DEEP[146] were mapped onto the resulting 6,836 Mash[153] representative MGCs by using BiG-MAP.map module. To assess the abundance of known pathways, the RPKM values from the known MGCs (following the definition of "known" stated in section S.4.2.7) were pulled out. The RPKM values of all the MGCs predicted by the same detection rule were merged. The pathway abundance (RPKM) was computed by dividing the gene clusters in 2kb-sized bins, and assessing the lower quartile number of reads mapping the 2kb bins for each gene cluster and sample. In contrast, a pathway was annotated as present in a sample when reads from that sample were found to be mapping

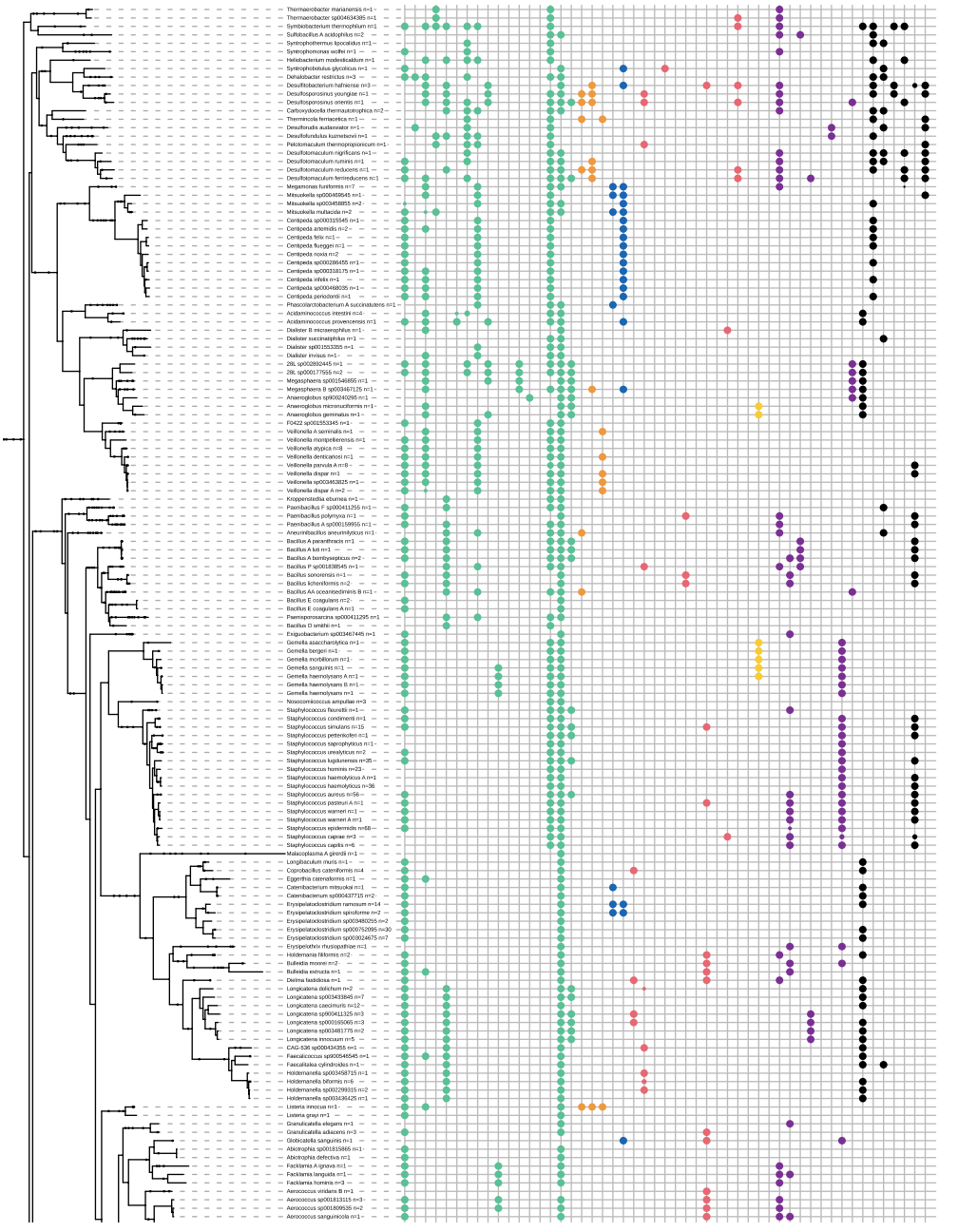
to at least 5% of the core region of that MGC. This threshold was kept low to enable detection of MGCs from low-abundant microbes and avoid false negatives due to limited sequencing depth. The lowest percentage identity of reads mapped to MGCs was 78% at the nucleotide level, which instilled confidence that finding multiple reads mapping to different locations within a MGC provides sufficient evidence for its presence in a sample. The pathway prevalence was also computed using 10%, 20%, 30%, 40%, 50%, 60%, 70% and 80% core coverage thresholds (Suppl. Figure 4.5), and results for increasing thresholds were consistent with gradual loss of detection capability for pathways known to be associated with low-abundance bacteria, such as the AAA to arylpropionate pathway (aromatic amino acid reductive branch).

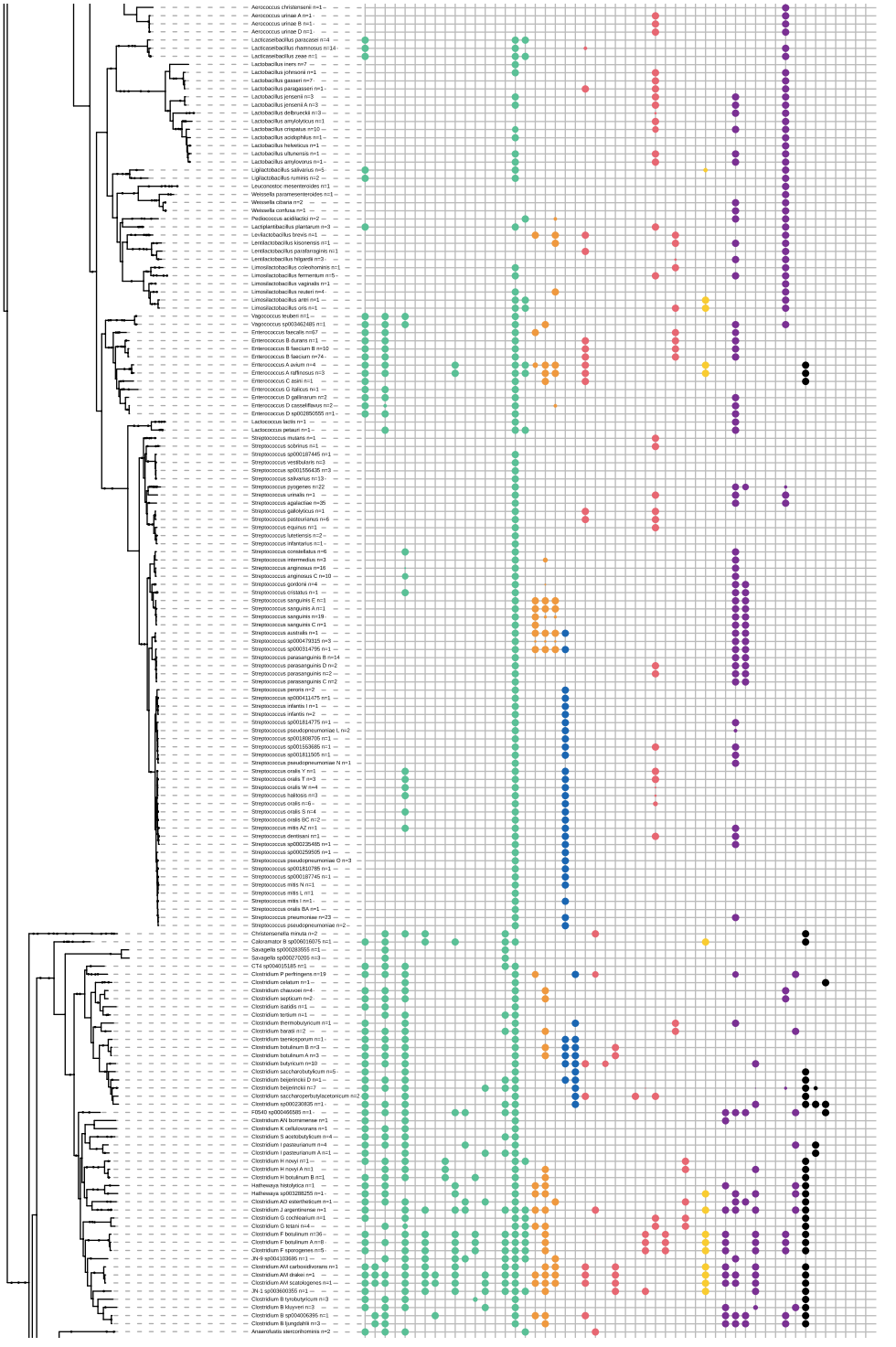
S.4.2.9 Correlating pathway abundance with metabolite concentrations in plasma

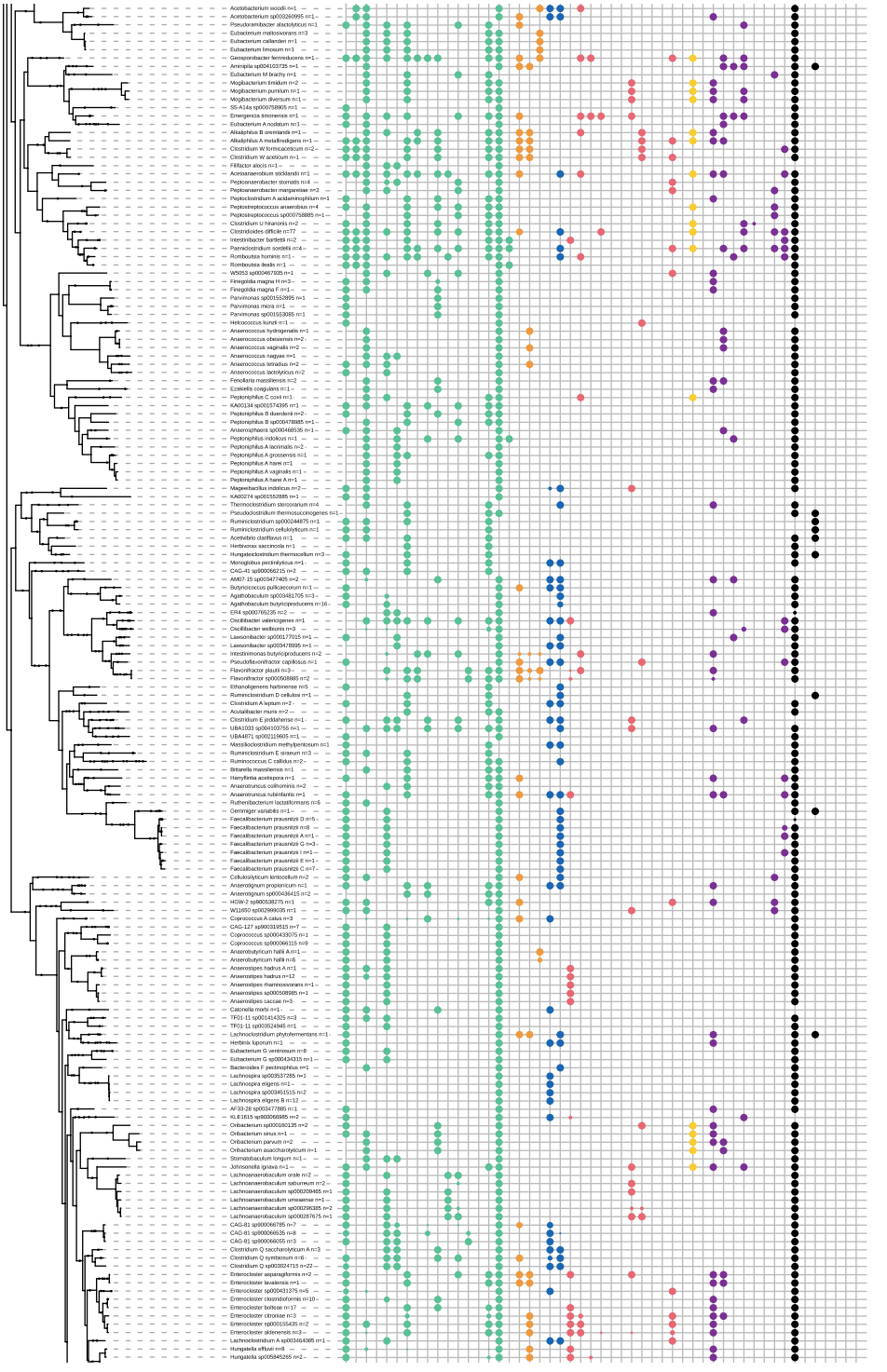
To evaluate the correlation between the gene cluster abundance and metabolite concentrations, the masses of 7 metabolites derived from several gutSMASH predicted gene clusters could be found in the Mass Spectrometry (MS) data of the plasma measured for 1,054 paired samples from LifeLines DEEP[135,146]. These metabolites included acetic acid, indolepropionic acid, isovaleric acid, *p*-cresol, *p*-cresol sulfate, phenylacetic acid and propionic acid (see Figure 3c and Suppl. Figure 4.6). Both metabolite and pathway abundance (RPKM counts) were inverse-rank-transformed and the linear regression was applied to adjust covariates including age, sex and metagenomic sequencing depth (only for pathway abundance). Metabolite and pathway abundance residuals from the linear regression model were then used to perform Spearman correlation test. Finally, the Benjamini Hochberg method was applied to control for false discovery rate (FDR).

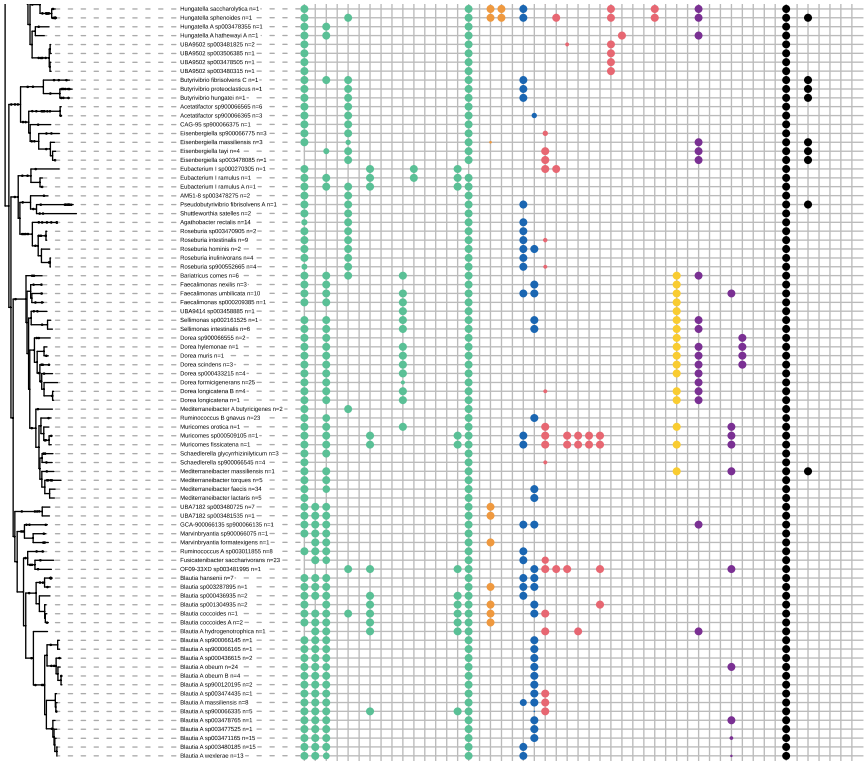
The LifeLines-DEEP study has complied with all relevant ethical regulations and has been approved by the Ethical Committee of the University Medical Center Groningen. All participants provided written informed consent.

Supplementary Figures

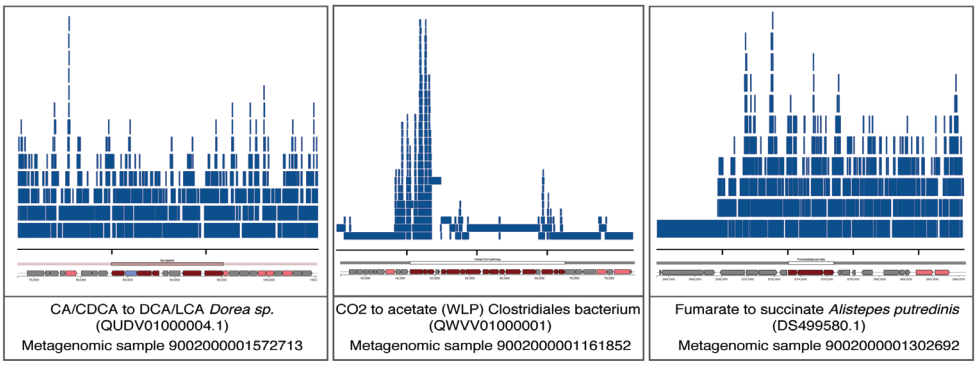




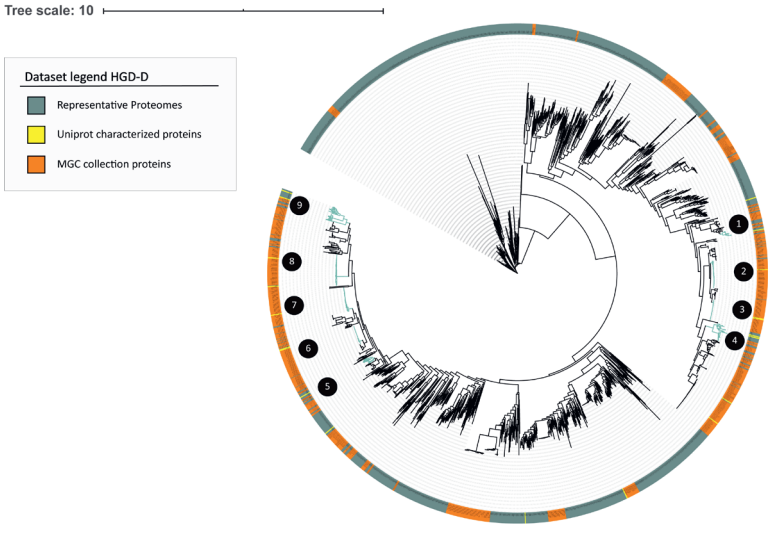




Supplementary Figure 4.1 Pathway distribution across a phylogenetic tree of 557 Firmicutes present in the HMP, CGR and Clostridiales datasets. The taxonomic assignments were performed using the GTDB database[96] and the phylogeny was produced using phyloT (https://phylot.biobyte.de/). Each column represents the presence/absence of the 51 metabolic pathways including single gene ones), which are color-coded based on the pathway's end product (metabolic classes). The full-sized circle implies that all strains in the node code for the pathway (see species label for information on the species group size), while smaller circles represent the relative number of species that encode the pathway. The pathways annotations were visualized using iTOL[122].

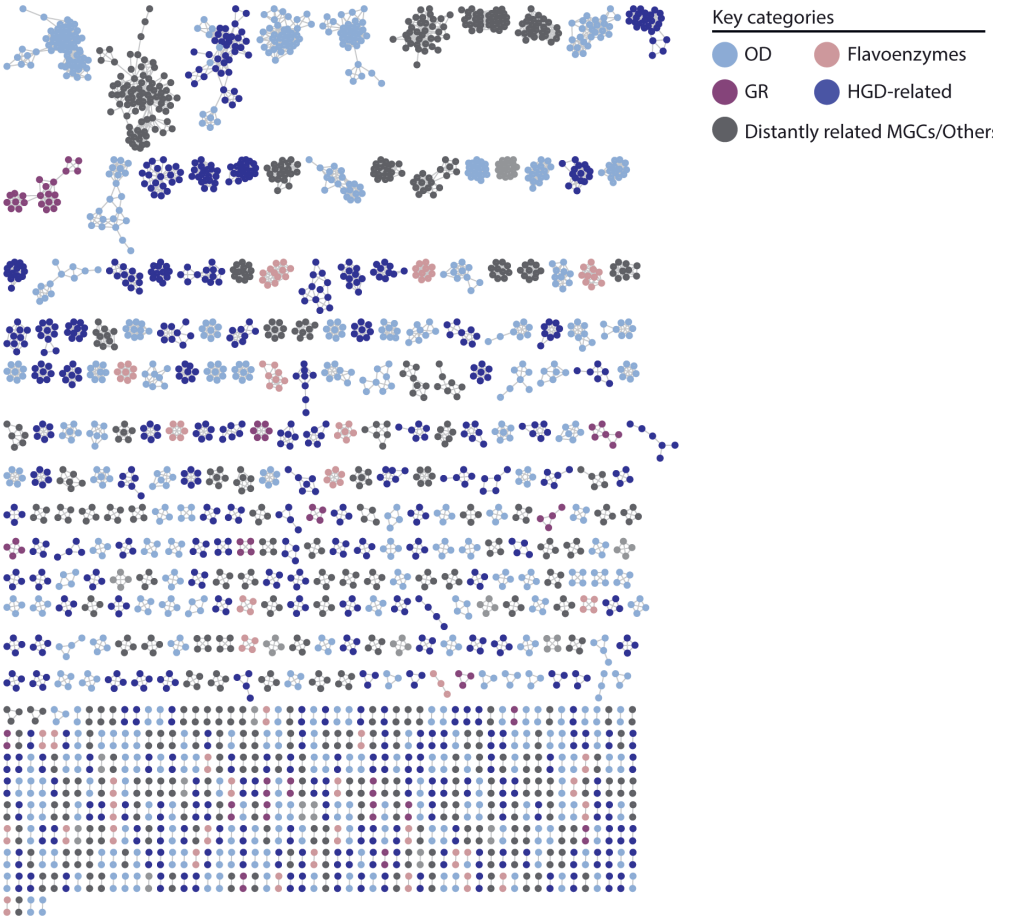


Supplementary Figure 4.2 Read coverage of three random metagenomic samples mapping to three gene clusters of interest: the *bai* operon (CDA/CDCA to DCA/LCA), a gene cluster encoding the acetyl-CoA pathway (CO_2 to acetate) and a gene cluster encoding the fumarate-to-succinate pathway. Reads are represented by blue lines, which are distributed along the x-axis based on the bedgraphs output by BiG-MAP. The plots have been produced using the Sushi R package (version 3.5.1)[154] and show how, despite the fact that some regions of the MGC attract more reads, the whole gene cluster is covered. They also illustrate the rationale for using the lower quartile of read coverage across 2kb regions, as this will avoid basing MGC abundance (partially) on outliers.



Supplementary Figure 4.3 HGD-D protein superfamily phylogeny. The phylogeny contains 2,054 protein sequences gathered from three different sources as

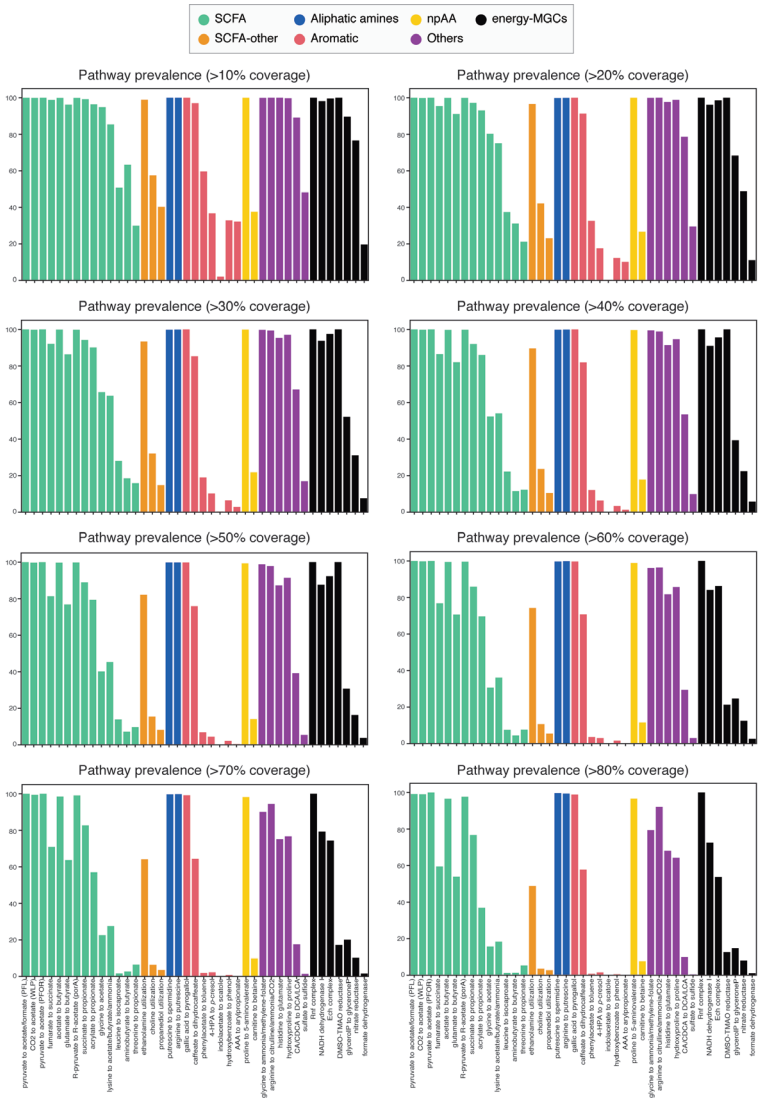
shown with different colours in the outer ring. The highlighted clades with numbers associated are the pathway-specific clades used to create the 9 pHMMs used in the AAA to aryl propionates (AAA reductive branch), leucine to isocaproate (leucine reductive branch), glutamate to butyrate and acrylate to propionate MGC detection rules



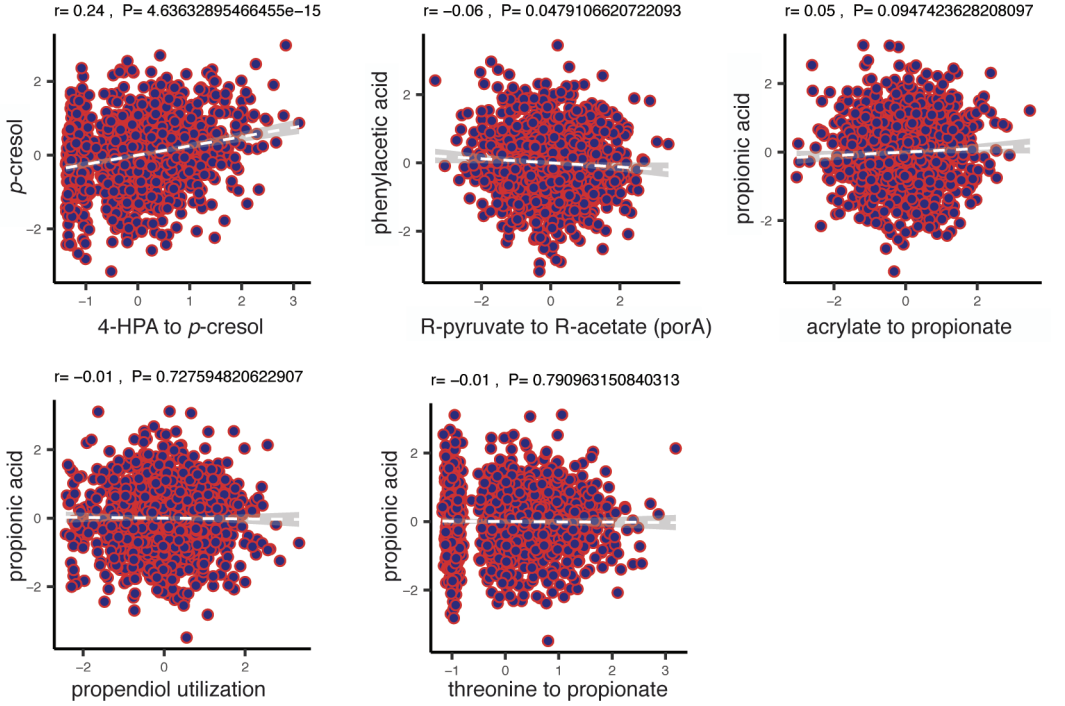
.

Supplementary Figure 4.4 Network of putative non-redundant MGCs predicted by gutSMASH. From all the unknown predicted MGCs, a redundancy filtering of 0.9 sequence similarity was applied using MMseqs2. From each cluster, two representatives were picked, and all representatives were used as input for BiG-SCAPE using the default cut-offs. The network contains 2,921 nodes and 7,474 edges. The MGCs have been classified into four different categories based on the key

enzyme classes they code for. The GR (glycyl-radical) category is composed of MGCs that include pyruvate formate-lyase (PFL-like) and/or glycyl radical (Gly_radical), OD (oxidative decarboxylation) involves MGCs with at least one of the following Pfam domains: Pyruvate ferredoxin/flavodoxin oxidoreductase (POR), Pyruvate flavodoxin/ferredoxin oxidoreductase, thiamine diP-bdg (POR_N), Pyruvate:ferredoxin oxidoreductase core domain II (PFOR_II) and Thiamine pyrophosphate enzyme, C-terminal TPP binding domain (TPP_enzyme_C). The Flavoenzymes category is a combination of MGCs harbouring at least one of the custom-made BaiCD and BaiH pHMMs. HGD-D-related MGCs, as the name states, include enzymes matching any of the 2-hydroxyglutaryl-CoA dehydratase, D-component (HGD-D)-related pHMM domains.



Supplementary Figure 4.5 Pathway prevalence using different core coverage thresholds. Pathway prevalence was computed by assessing the number of reads (per sample) mapping to known gene clusters at a certain core coverage cut-off. The figure illustrates how the pathway prevalence gradually changes when increasing the core coverage cut-off from 10 to 80%.



Supplementary Figure 4.6 Limited correlation of genetic pathway abundance with metabolites abundance in blood plasma. This figure shows correlation plots for additional metabolites not shown in Figure 4.3c.

Supplementary tables are available online at: https://www.biorxiv.org/content/10.1101/2021.02.25.432841v1.supplementary-material

Chapter 5

The gutSMASH webserver – automated identification of primary metabolic gene clusters from the gut microbiota

Victòria Pascal Andreu, Jorge Roel-Touris, Dylan Dodd, Michael A. Fischbach, Marnix H. Medema.

Published in 2021 in *Nucleic Acids Research*, gkab353

Abstract

Anaerobic bacteria from the human microbiome produce a wide array of molecules at high concentrations that can directly or indirectly affect the host. The production of these molecules, mostly derived from their primary metabolism, is frequently encoded in metabolic gene clusters (MGCs). Important efforts have been made to design tools to predict gene clusters involved in secondary metabolism, such as antiSMASH, the gold standard algorithm in the natural discovery field. However, despite the importance of microbiome-derived primary metabolites in the host, no tool existed to predict the gene clusters responsible for their production. For this reason, and to support efforts to elucidate the metabolic potential of the human microbiota, we recently introduced gutSMASH. This tool is based on the antiSMASH version 5 framework, which identifies MGCs by detecting protein domain combinations that are defining for specific pathways. gutSMASH can predict 41 different known pathways, including MGCs involved in bioenergetics, but also putative ones that are good candidates for novel pathway discovery. To make the tool more user-friendly and accessible, here we present the gutSMASH webserver, hosted at https://gutsmash.bioinformatics.nl/. The user can either input the GenBank assembly accession or upload a genome file in FASTA or GenBank format. Optionally, the user can enable additional set(s) of analysis to acquire further insights on the predicted MGCs. An interactive HTML output (viewable online or downloadable for offline use) provides a user-friendly way to browse functional gene annotations and sequence comparisons with reference gene clusters as well as gene clusters predicted in other genomes. Thus, this web server provides the community with a streamlined and user-friendly tool to analyze the metabolic potential of gut microbiomes.

5.1 Introduction

Microbiome research has received considerable attention in the last decade, charting the taxonomic and functional diversity found in complex ecosystems and the effects on their host. Many microbiome-associated phenotypes in the gut microbiome are derived from small molecules synthesized by anaerobic bacteria[52]. These compounds are mostly products of their primary metabolic pathways that allow them to either colonize specific micro-niches in the gut or interact with other microbes. Despite the fact that some of these molecules are produced by low-abundant bacteria, they can reach high concentrations in the gut as well as in blood plasma, sometimes comparable to those of therapeutic drugs[110]. This is of great interest because these molecules can profoundly modulate host metabolism, immunity, and homeostasis. An example of such an end product is trimethylamine, derived from carnitine and choline metabolism, which is known to be associated with increased risk of cardiovascular disease[102]. Another interesting pathway is ethanolamine utilization, which has been indirectly associated with bacterial pathogenesis by enabling these bacteria to bloom, using this molecule as carbon and nitrogen source[155]. Finally, short chain fatty acids (SCFAs) have been found to positively impact human health, their most abundant representatives are acetate, butyrate and propionate. Among other functions, butyrate is used by colonocytes as their main energy source and promote cell differentiation[156].

Characterizing the metabolic potential of the human microbiome will enable the research community to define how metabolic interactions among microbes and with the host influences human health. Although tools exist to predict the metabolic potential of bacteria, they are either focused on predicting gene clusters encoding the biosynthesis of secondary metabolites (e.g., antiSMASH[73]) or they identify only individual genes instead of MGCs. Moreover, these tools strongly depend on generic primary metabolic databases such as KEGG[157] or BioCyc[126] that do not always include such specialized primary

metabolic pathways and only cover known pathways[158]. Examples of such tools for studying primary metabolism are the HUMAnN pipeline[141], MetaPath[159] or MetaTrans[160]. Given the fact that genes encoding for specialized primary metabolic pathways are often found clustered together and evidence exist that the metabolic potential of such bacteria is far from fully uncovered[142], we recently introduced gutSMASH[158], a tool that identifies known as well as potentially novel MGCs, based on the antiSMASH version 5.0 framework[125].

Here, we present the gutSMASH web server, available at https://gutsmash.bioinformatics.nl/, which is designed to mine anaerobic bacterial genomes for primary specialized metabolic gene cluster (MGCs) in a user-friendly manner. The server is able to predict not only known MGCs but also putative gene clusters that may aid in discovering novel molecules of importance for human (or animal) health. Moreover, gutSMASH also predicts gene clusters related to energy acquisition. This platform runs the algorithm on any given correctly formatted genome the user inputs, and outputs an interactive visualization that provides information on the predicted gene cluster. Additionally, it performs comparative genomic analysis using two customized databases (KnownClusterBlast and ClusterBlast) to assess the novelty of the predicted MGCs, assess their taxonomic distribution and identify architectural variants. Also, it annotates MGC genes into functional categories and highlights these with specific colors. In order to provide examples on the use of gutSMASH, we present an analysis of several genomes to showcase the diversity of gene clusters that this tool can predict, their annotations and their similarity to any gene cluster previously predicted by gutSMASH.

5.2 Methods and implementation

5.2.1 Overview of the gutSMASH workflow

The gutSMASH algorithm is based on the antiSMASH version 5 framework. As in antiSMASH, detection rules are used for MGC identification, consist-

ing of Pfam combinations that constitute the signature of a given metabolic pathway. The design and validation of the detection rules are described in detail in Pascal Andreu *et al.*[158]. Figure 5.1 illustrates the different steps followed by gutSMASH. From a bacterial genome specified by the user, gutSMASH first identifies the core metabolic genes by iterating over the detection rules. Once the core genes are identified, each protocluster is extended from each flank to include accessory genes. Then, if the "KnownClusterBlast" and/or "ClusterBlast" options are enabled, gutSMASH performs an MGC comparative genomic analysis by blasting the predicted gene clusters to a collection of known and characterized MGCs and/or to a broader collection of gutSMASH-predicted MGCs respectively. Next, if desired, gutSMASH can functionally annotate genes into eight different categories: core biosynthetic, additional biosynthetic, transport-related, regulatory, resistance, and others (already found in antiSMASH) and including encapsulation- and electron-transport-related genes as new additions. Once all the analysis has finished, gutSMASH writes the results and displays the interactive output. Also, the webserver gives the option to download all results as a ZIP file.

5.2.2 Interactive input and output

The ideal input for gutSMASH is an annotated nucleotide file in Genbank or EMBL format. The user can either upload a GenBank/EMBL file manually, or simply enter the GenBank assembly accession number, upon which gutSMASH will automatically use the annotated assembled genome from the NCBI FTP. Alternatively, the user can provide a FASTA file containing one or more sequences. In this case, gutSMASH will predict the genes and annotate the genome using Prodigal[151], and use those annotations to run the rest of the analysis.

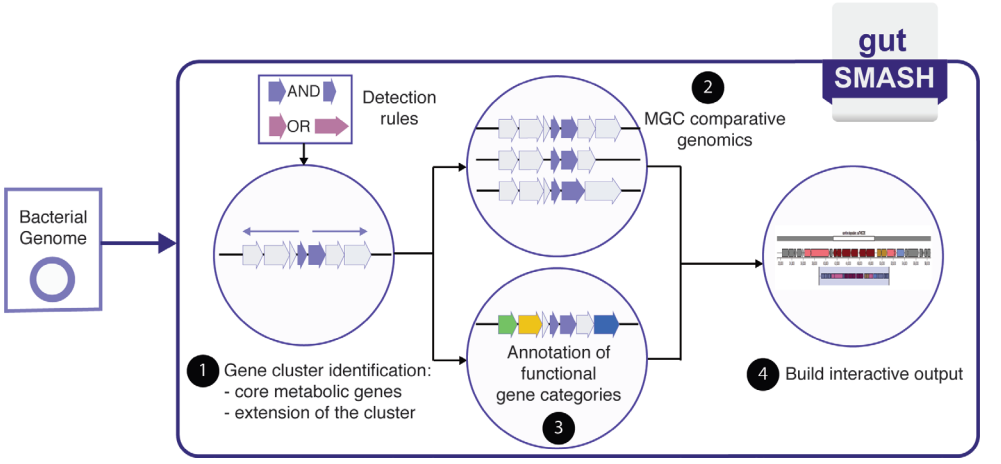


Figure 5.1 Overall gutSMASH workflow. gutSMASH takes a bacterial genome sequence as input either in GenBank, EMBL or FASTA format. First, the program iterates over the detection rules to identify the gene clusters. Next, if enabled, the predicted MGCs are compared to the databases specified to evaluate the similarity to any known pathways, or to assess the similarity to gene clusters that were predicted by gutSMASH from publicly available whole-genome sequences.

The gutSMASH results can either be visualized online in a browser or downloaded locally. The output consists of several interactive HTML pages that allow the user to explore the results further. The overview page gives information on all the predicted MGCs, including their location in the genome and the functional class to which each of the MGCs belong. The main page also contain links to the gutSMASH documentation page (https://gutsmash-documentation.readthedocs.io/en/latest/) for more details. Moreover, each predicted MGC can be visualized individually for further inspection of additional MGC-specific results, depending on the options enabled before submitting the job. Besides the HTML pages, gutSMASH also generates plain text files with the KnownClusterBlast/ClusterBlast results (more details in the next section) and a GenBank file for each predicted region for further processing.

5.2.3 Comparative genomic analysis to identify distant homologs and assess MGC taxonomic distribution

GutSMASH uses two different databases, KnownClusterBlast and ClusterBlast, to find MGCs homologous to the query. This comparative analysis can provide good indications of the distribution of the MGC among bacterial taxa from the human microbiome, provide insights into the extant variation of MGC architectures (gene content) and provide clues regarding MGC function (using homology-based inference).

The KnownClusterBlast module aims to identify similarities between the predicted MGCs and a reduced set of genetically and biochemically characterized gene clusters. In order to design the detection rules for these known pathways, a reduced set of well-known pathways was analysed (as described in Pascal Andreu *et al.*[158]). The sequences of these MGCs of known function were included in the KnownClusterBlast database, which currently contains 59 entries. Therefore, ticking the KnownClusterBlast button allows the user to identify which MGCs are homologous to and likely share the same function with those reference MGCs, and to study the similarities and differences in detail. Given its usefulness, this option is enabled by default.

To build the ClusterBlast database, the Culturable Genome Reference (CGR) collection[128], the Human Microbiome Project (HMP, https://www.hmpdacc.org/reference_genomes/reference_genomes.php) reference genomes and 414 Clostridiales complete genomes available in October 2019 under taxid 186802 were used as input for gutSMASH. Based on the output, both the known and putative predicted-gene clusters, a total of 30,883 MGC selected sequences, were combined to form the ClusterBlast database. When the ClusterBlast option is enabled, DIAMOND[161] is used to compare the predicted MGC protein sequences to those in the ClusterBlast MGC database to identify close homologs, using the same procedure as in antiSMASH. After

ranking the gene clusters based on the numbers of pairs of homologous proteins and the cumulative bit scores, the top 10 most similar gene clusters (based on highest bit scores) are displayed in the ClusterBlast HTML tab of each predicted region. The complete list of homologs that includes similarity scores can be retrieved from the ClusterBlast output folder in the downloadable ZIP output.

5.2.3 Annotation of functional gene categories: additional pmDBFA categories

The module "primary metabolite domain-based functional annotations" (pmDBFA) (analogous to the secondary metabolite Clusters of Orthologous Groups [smCOGs] in antiSMASH) facilitates the functional annotation of accessory genes within a predicted gene cluster into different categories based on the presence of key Pfams in the gene-coding sequences. To tailor these annotations towards a more meaningful classification for gutSMASH output, two extra functional categories were included: encapsulation and electron transport-related genes. There are several Pfams that belong to these categories, such as the Electron transfer flavoprotein FAD-binding domain (PF00766) or the BMC domain (PF00936), respectively, which are regularly found in pathways predicted by gutSMASH. For instance, in the reductive metabolism of aromatic amino acids to arylpropionates, an electron transfer protein encoded by *etfA,* which harbours an electron-transferring flavoprotein domain (PF00766), is required to reduce the substrate[162]. The encapsulation category in contrast, aims to include genes involved in bacterial microcompartmentalization, which have been found to be important for some reactions, such as propanediol utilization operon (*pdu*), to encapsulate pathway intermediates that would be toxic for the cell at high concentrations[163]. Nevertheless, in some cases, these domains are part of the detection rules of known pathways, and thus are also annotated as core genes. When found in the flanking regions, they are annotated in the corresponding category.

5.2.4 Code development and server implementation

The gutSMASH server, hosted at https://gutsmash.bioinformatics.nl/, is based on the Python3 Flask web framework (https://flask.palletsprojects.com) for server-side logic combined with Jinja templating language (https://jinja.palletsprojects.com) and JavaScript for client-side logic, and implements the gutSMASH software, for which the source code can be found at https://github.com/victoriapascal/gutsmash. The submission interface requires different (some of them optional) inputs from the user, including a mandatory valid sequence file or assembly accession ID. For the job handling, different statuses were defined as:

- *Submitted*: The job has been successfully submitted.
- *Queued*: The job is waiting to be processed.
- *Running*: gutSMASH analysis have started.
- *Finished*: The job has successfully finished.
- *Failed*: An error has occurred, and it is displayed for troubleshoot ing.
- *Notified*: The job has successfully finished, and the user has been notified.
- *Failed-notified*: An error has occurred, and the user has been notified.

To handle all the job statuses, the Advanced Python Scheduler (APScheduler) library was used. Please note that the status *notified* and *failed-notified* are only applicable if the user provides an email address. Finally, the communication between the web interface and the application is done via a Redis database that stores the job's information. As reference, the antiSMASH[125] (https://github.com/antismash/websmash/tree/master/websmash) and plantiSMASH[164] (https://github.com/plantismash/webserver) webservers were used for inspiration for the main layout and content

5.3 Results

In the following sections, we provide several examples to illustrate how gutSMASH works and how to interpret the results.

5.3.1 gutSMASH detects known and putative gene clusters from prominent human gut pathogens

Escherichia albertii belongs to the family *Enterobacteriaceae* and is an emergent human enteropathogenic microbe[165]. As a close-relative of the well-studied *E. coli*, certain metabolic functions are expected to be shared with it, while others may differ. To uncover its specialized primary metabolism and bioenergetics, we used the GenBank assembly accession (GCA_002285455.1) as input to the gutSMASH server and found that this genome contains 24 MGCs that belong to different MGC classes (see Figure 5.2). From these, 19 gene clusters have 50% or higher overall gene cluster similarity to reference MGCs when using the KnownClusterBlast database as reference. Among them, gutSMASH identifies homologs of the known carnitine (*cai*) degradation operon[166] (100% identical), the propanediol utilization (*pdu)* operon[167] (91% identical), the ethanolamine utilization (*eut)* operon[155] (94% identical) and the threonine to propionate degradation pathway[168] (90% identical). Enabling the ClusterBlast option also allowed to check whether other bacteria shared similar gene clusters or not. As expected, most of the predicted MGCs are found in other *Escherichia* genomes. However, the leucine reductive branch gene cluster (62% identical to the leucine reductive branch reference MGC) is rarely found among other close relatives, as shown in the ClusterBlast results in Figure 5.2.

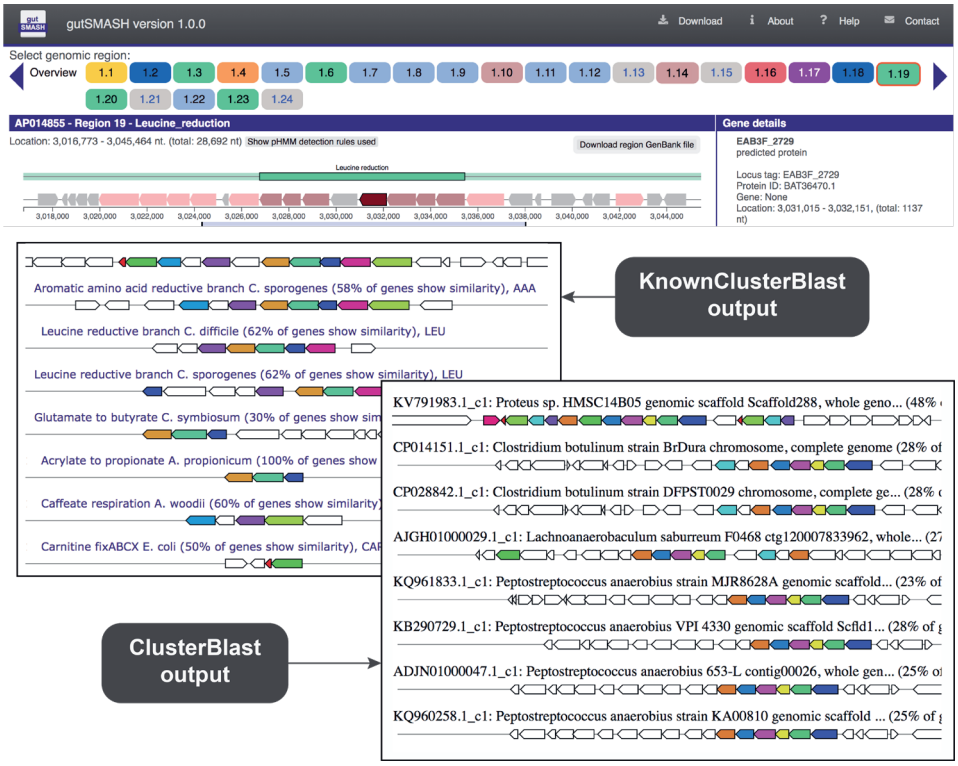
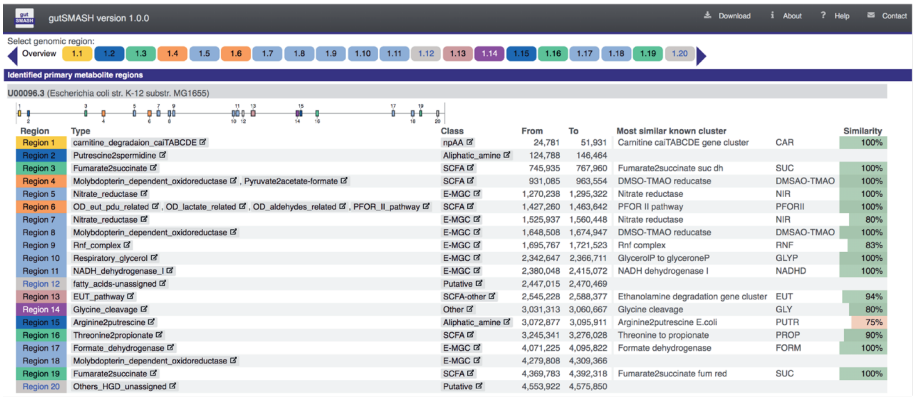


Figure 5.2 gutSMASH run for *Escherichia albertii* (GCA_002285455.1). From 24 predicted MGCs, the leucine reductive branch MGC is highlighted as an example (region 1.19). The KnownClusterBlast results show that five coding genes of this MGC share similarity with the leucine reductive branch reference gene cluster genes (overall 62% similarity) and seven with the aromatic amino acid reductive branch ones (overall 58% similarity). The ClusterBlast output shows that this gene cluster does not have homologous MGCs among other *Escherichia* members present in the database.

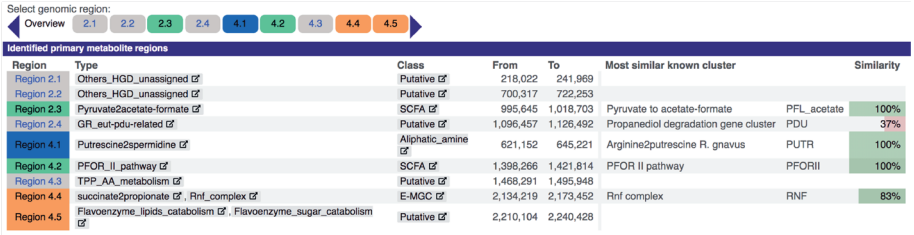
For comparison, we also analysed the genome of *Escherichia coli* K-12 (GCA_000005845.2) and found that this genome contains 20 MGCs, 16 of them with overall gene cluster identity of >= 75% when compared to the KnownClusterBlast database (see Figure 5.3A). Among the predicted MGCs, we find the *cai* and *eut* operon, but it lacks the *pdu* and the leucine reductive branch MGC, among others, when compared to *Escherichia albertii*.

A) Overview of the predicted MGCs using the *E. coli* genome (GCA_000005845.2)



Region	Type	Class	From	To	Most similar known cluster		Similarity
Region 1	carnitine_degradaion_caiTABCDE	npAA	24,781	51,931	Carnitine caiTABCDE gene cluster	CAR	100%
Region 2	Putrescine2spermidine	Aliphatic_amine	124,788	146,464			
Region 3	Fumarate2succinate	SCFA	745,935	767,960	Fumarate2succinate suc dh	SUC	100%
Region 4	Molybdopterin_dependent_oxidoreductase , Pyruvate2acetate-formate	SCFA	931,085	963,554	DMSO-TMAO reducatse	DMSAO-TMAO	100%
Region 5	Nitrate_reductase	E-MGC	1,270,238	1,295,322	Nitrate reductase	NIR	100%
Region 6	OD_eut_pdu_related , OD_lactate_related , OD_aldehydes_related , PFOR_II_pathway	SCFA	1,427,260	1,463,642	PFOR II pathway	PFORII	100%
Region 7	Nitrate_reductase	E-MGC	1,525,937	1,560,448	Nitrate reductase	NIR	80%
Region 8	Molybdopterin_dependent_oxidoreductase	E-MGC	1,648,508	1,674,947	DMSO-TMAO reducatse	DMSAO-TMAO	100%
Region 9	Rnf_complex	E-MGC	1,695,767	1,721,523	Rnf complex	RNF	83%
Region 10	Respiratory_glycerol	E-MGC	2,342,647	2,366,711	GlycerolP to glyceroneP	GLYP	100%
Region 11	NADH_dehydrogenase_I	E-MGC	2,380,048	2,415,072	NADH dehydrogenase I	NADHD	100%
Region 12	fatty_acids-unassigned	Putative	2,447,015	2,470,469			
Region 13	EUT_pathway	SCFA-other	2,545,228	2,588,377	Ethanolamine degradation gene cluster	EUT	94%
Region 14	Glycine_cleavage	Other	3,031,313	3,060,667	Glycine cleavage	GLY	80%
Region 15	Arginine2putrescine	Aliphatic_amine	3,072,877	3,095,911	Arginine2putrescine E.coli	PUTR	75%
Region 16	Threonine2propionate	SCFA	3,245,341	3,276,028	Threonine to propionate	PROP	90%
Region 17	Formate_dehydrogenase	E-MGC	4,071,225	4,095,822	Formate dehydrogenase	FORM	100%
Region 18	Molybdopterin_dependent_oxidoreductase	E-MGC	4,279,808	4,309,366			
Region 19	Fumarate2succinate	SCFA	4,369,783	4,392,318	Fumarate2succinate fum red	SUC	100%
Region 20	Others_HGD_unassigned	Putative	4,553,922	4,575,850			

B) Overview of the predicted MGCs using the *R. gnavus* genome (GCF_000526735.1)



Region	Type	Class	From	To	Most similar known cluster		Similarity
Region 2.1	Others_HGD_unassigned	Putative	218,022	241,969			
Region 2.2	Others_HGD_unassigned	Putative	700,317	722,253			
Region 2.3	Pyruvate2acetate-formate	SCFA	995,645	1,018,703	Pyruvate to acetate-formate	PFL_acetate	100%
Region 2.4	GR_eut-pdu-related	Putative	1,096,457	1,126,492	Propanediol degradation gene cluster	PDU	37%
Region 4.1	Putrescine2spermidine	Aliphatic_amine	621,152	645,221	Arginine2putrescine R. gnavus	PUTR	100%
Region 4.2	PFOR_II_pathway	SCFA	1,398,266	1,421,814	PFOR II pathway	PFORII	100%
Region 4.3	TPP_AA_metabolism	Putative	1,468,291	1,495,948			
Region 4.4	succinate2propionate , Rnf_complex	E-MGC	2,134,219	2,173,452	Rnf complex	RNF	83%
Region 4.5	Flavoenzyme_lipids_catabolism , Flavoenzyme_sugar_catabolism	Putative	2,210,104	2,240,428			

Figure 5.3 Overview of the gutSMASH runs for the a) *E. coli* K-12 and b) *R. gnavus* genomes. In the overview, various pieces of information regarding the predicted MGCs can be seen, such as the number of predicted MGCs, type and class, and genomic coordinates. If there is similarity to any MGC from the KnownClusterBlast database, the overall percentage similarity (percent of genes in the reference MGC with a homolog in the query) can also be seen.

To also show an example from a different phylum, we analysed a genome of *Ruminococcus gnavus,* a firmicute species that has been previously associated with Crohn's disease for the secretion of a complex polysaccharide that promotes inflammation[169]. In this case, the genomic FASTA file of the strain AGR2154 (GCF_000526735.1) was used as input for gutSMASH, which predicted 10 MGCs (see Figure 5.3B), 5 of them classified as putative, since they either do not have matches to any entry in the KnownClusterBlast database or the overall sequence identity is very low. Interestingly, this microbe produces the aliphatic amine spermidine that acts as an anti-inflammatory agent by suppressing immune reactions[170].

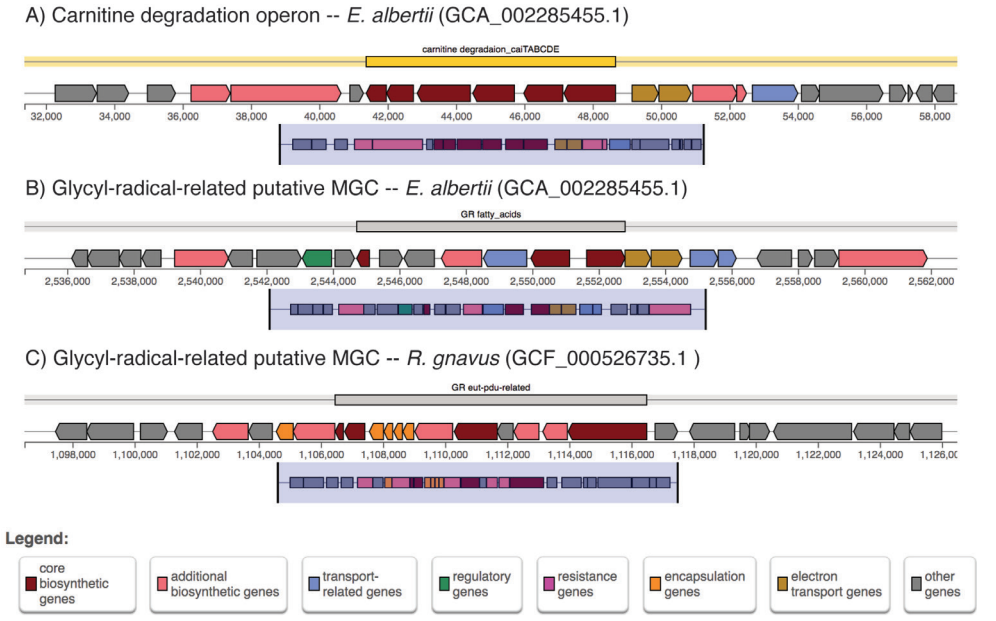


Figure 5.4 Known and putative MGCs predicted by gutSMASH with functionally annotated genes. A and B are predicted from *E. albertii* and C from *R. gnavus*. Gene cluster A belongs to the carnitine degradation operon, thus a known MGC, which contains core (dark red arrows) and auxiliary genes (salmon) but also transport-related (blue) and electron transport-related genes (dark yellow). Gene cluster B is a putative gene cluster that encodes a glycyl-radical enzyme. The MGC also contains genes belonging to the following categories: auxiliary, electron transport and regulatory. Gene cluster C is a glycyl-radical putative MGC with several encapsulation-related genes.

5.3.2 Gene functional annotations help to get more insights on the reaction

In the three gutSMASH runs, we enabled the gene functional annotations using the pmCOGs module. From the results, we could confirm that known MGCs such as the *cai* operon encode for two electron-transfer proteins in *E. albertii* (BAT33784.1 and BAT33785.1), which may participate in electron transfer reactions to fuel the electron transport chain. However, we also find these coding genes in other MGCs of unknown function, such as the

predicted glycyl-radical (GR) fatty-acids gene cluster (see Figure 5.4A & 5.4B). The latter, belongs to the GR class since it encodes for a pyruvate formate-lyase but also for an acyl-CoA dehydrogenase. The gene color annotations help to visualize and show that this putative MGC not only codes for the core enzyme-coding genes but also for transport, regulatory and electron-transport related genes. Similarly, for the *Ruminococcus gnavus* genome, we find a putative MGC that belongs to the GR class since the coding sequence of *ctg2_1090* harbours a Gly_radical and PFL-like domain (PF01228 and PF02901 respectively, see Figure 5.4C). One of the genes also codes for an aldolase (PF00596) and, because of the presence of several BMC-coding genes, it indicates that this pathway might be encapsulated into the so-called bacterial microcompartments.

5.4 Conclusion and future perspectives

Currently, gutSMASH is able to predict a wide range of known and putative gene clusters that are interesting to functionally profile gut microbiomes, and, in principle, any other microbiomes where the profiled pathways regularly occur, including skin and oral. The set-up of this user-friendly webserver makes this tool accessible to researchers that are not familiar with the command line. Moreover, the examples provided in this article show how, by enabling the comparative genomics analysis and the functional gene annotation options, the user can extract useful information to interpret and further analyse the predicted regions.

gutSMASH version 1.0 presents a first step to mine anaerobic genomes for primary metabolic gene clusters and bioenergetics, but we already anticipate that, in the near future, a more updated version of the tool will have to include new detection rules to predict newly characterized gene clusters. For this reason, users are advised to regularly check the online gutSMASH documentation for a description of the most up-to-date list of detected gene cluster types. In the future, a comprehensive database of gutSMASH-pre-

dicted gene clusters, similar to the antiSMASH database for antiSMASH-detected gene clusters[171], also would be useful to allow users to query the database based on taxonomic entities of interest (e.g., at the species or genus level) or retrieve gene clusters with specific combinations of protein families of interest. Overall, we believe that with this work we provide the community with a useful asset to profile microbiomes for specialized primary MGCs and bioenergetics to better understand the chemistry in this ecosystem with the benefit of a ready-to-use streamlined protocol.

5.5 Data Availability Statement

The source code for the gutSMASH software hosted by the web server can be found at https://github.com/victoriapascal/gutsmash.

5.6 Acknowledgements

The authors acknowledge the help provided by the Medema group at Wageningen University & Research at large, and specially Satria A. Kautsar and Mohammad Alanjary. Also, we acknowledge our antiSMASH collaborators at Technical University of Denmark (Copenhagen), Kai Blin, Simon Shaw and Tilmann Weber for their technical support and guidance on the design of the gutSMASH framework and also this webserver.

Chapter 6

BiG-MAP: an automated pipeline to profile metabolic gene cluster abundance and expresion in microbiomes

Victòria Pascal Andreu, Hannah E. Augustijn, Koen van den Berg, Justin J. J. van der Hooft, Michael A. Fischbach, Marnix H. Medema.

Updated from the preprint available at bioRxiv:
https://www.biorxiv.org/content/10.1101/2020.12.14.422671v1

Abstract

Microbial gene clusters encoding the biosynthesis of primary and secondary metabolites play key roles in shaping microbial ecosystems and driving microbiome-associated phenotypes. Although effective approaches exist to evaluate the metabolic potential of such bacteria through identification of metabolic gene clusters in their genomes, no automated pipelines exist to profile the abundance and expression levels of such gene clusters in microbiome samples to generate hypotheses about their functional roles and to find associations with phenotypes of interest. Here, we describe BiG-MAP, a bioinformatic tool to profile abundance and expression levels of gene clusters across metagenomic and metatranscriptomic data and evaluate their differential abundance and expression between different conditions. To illustrate its usefulness, we analyzed 96 metagenomic samples from healthy and caries-associated human oral microbiome samples and identified 252 gene clusters, including unreported ones, that were significantly more abundant in either phenotype. Among them, we found the *muc* operon, a gene cluster known to be associated to tooth decay. Additionally, we found a putative reuterin biosynthetic gene cluster from a *Streptococcus* strain to be enriched but not exclusively found in healthy samples; metabolomic data from the same samples showed masses with fragmentation patterns consistent with (poly)acrolein, which is known to spontaneously form from the products of the reuterin pathway and has been previously shown to inhibit pathogenic *Streptococcus mutans* strains. Thus, we show how BiG-MAP can be used to generate new hypotheses on potential drivers of microbiome-associated phenotypes and prioritize the experimental characterization of relevant gene clusters that may mediate them.

6.1 Introduction

Bacteria can produce diverse sets of small molecules that interact with other microbes or with their host. These metabolites include members of both primary and secondary metabolism and cover a wide chemical diversity[52,172]. Importantly, the pathways responsible for their production are often specific to certain strains or species and help them to compete for space and resources[173], e.g. through antimicrobial, nutrient-scavenging or immunomodulatory activities[174]. The genes that encode these pathways are often physically clustered and are also known as Biosynthetic Gene Clusters (BGCs) or Metabolic Gene Clusters (MGCs)[155,175] — the latter being a broader definition that also includes catabolic pathways. Several studies have indicated metabolites produced from such gene clusters to be the major drivers of specific phenotypic traits; for instance, pseudomonads in the rhizosphere of sugar beet plants were shown to produce the antifungal non-ribosomal peptide (NRP) thanamycin, which protects plants from fungal infections[176]. Another example from primary metabolism is trimethylamine, a diet derived-molecule that is processed by bacteria harboring the *cut* gene cluster, and has been associated with an increased risk of suffering from cardiovascular disease[177]. Therefore, mining genomes for MGCs enables moving the field towards a deeper understanding of function at the molecular level and determine the role a given microbe plays in the ecosystem[46].

Several tools have been developed to mine genomes for these gene clusters, like antiSMASH[125], gutSMASH[158] or DeepBGC[178]. In contrast to other tools for functional profiling of microbial communities, such as HUMAnN2[141], MetaPath[159], FMAP[78] and Metatrans[160], these do not depend on pathways that are present in reference databases like KEGG[81] or MetaCyc[179], which only include pathways for which most or all enzymatic steps have been elucidated. In fact, the majority of gene clusters identified by antiSMASH and many gene clusters predicted by gutSMASH encode pathways for which the catalytic steps, intermediates, and final products are yet unknown.

However, known pathways that are encoded by gene clusters can also be reliably detected. The detection of complete gene clusters instead of individual enzyme-coding genes likely decreases false positive detections of enzymes that show sequence similarity to reference enzyme sequences but are part of different functional contexts. For these reasons, identification of gene clusters of known and unknown function provides a useful basis to look for functional explanations of microbiome-associated phenotypes of interest. As phenotypes are often triggered by metabolites at physiologically relevant concentrations, while samples without the phenotype lack these metabolites or have them at lower concentrations, assessing gene cluster abundance and expression levels across samples is crucial to predict associations with the phenotype in question. Another significant advantage of profiling the community by combining different omics data is to prioritize the characterization of putative gene clusters that are highly abundant or expressed in samples of interest, in order to elucidate the structures and functions of the most relevant novel compounds and their biosynthetic pathways.

Here, we present BiG-MAP (the Biosynthetic Gene cluster Meta'omics Abundance Profiler), which provides a streamlined and automated process to determine BGC/MGC abundance and expression in bacterial communities by mapping metagenomic and metatranscriptomic reads to gene cluster sequences from reference genomes or metagenomic assemblies. BiG-MAP uses MinHash-based redundancy filtering and groups BGCs into families with BiG-SCAPE[124] to avoid ambiguous mapping, and uses these to output and visualize profiles of MGC abundance or expression levels across samples. Additionally, it calculates differential abundance or expression using either parametric or nonparametric tests. We validate the tool using simulated metagenomic data and show how MGC abundance and expression levels are accurately recapitulated. Finally, to showcase its usefulness, we applied BiG-MAP on a large publicly available metagenome dataset from the human oral microbiome and describe how it successfully identified dif-

ferential abundance of gene clusters related to bacteria's specialized primary and secondary metabolism that are (potentially) relevant for caries development. Among others, this collection includes a *pdu* and cobalamin gene cluster involved in reuterin biosynthesis and the *muc* operon involved in reutericyclin/mutanocyclin biosynthesis. Thus, BiG-MAP suggests new lines to further explore the onset and development of oral cavities.

6.2 Results and discussion

6.2.1 An approach to map metagenomics and metatranscriptomic reads to gene clusters

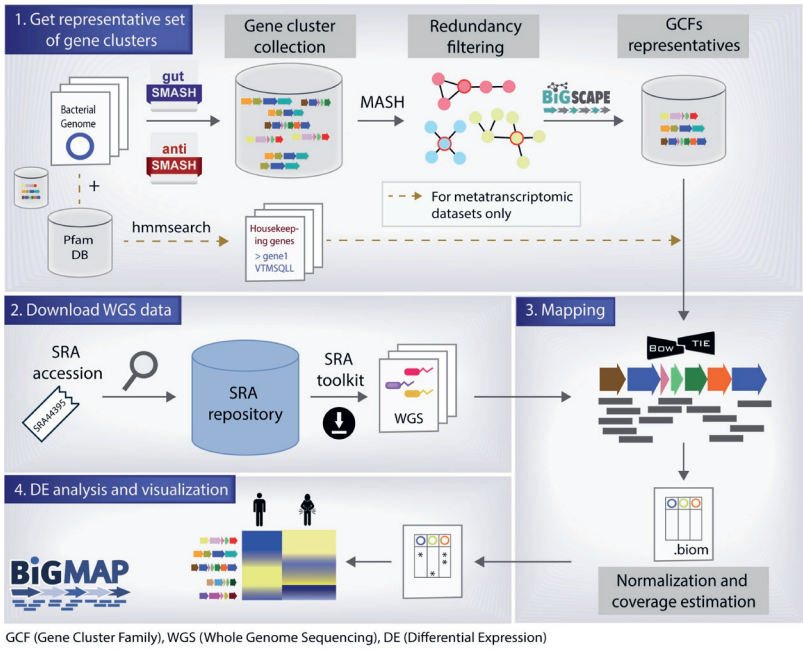
BiG-MAP maps shotgun sequencing reads onto gene clusters that have been either predicted by antiSMASH[125] or gutSMASH[158]. It is a Python-based pipeline, which allows downloading datasets from SRA, mapping metagenomic or metatranscriptomic reads to gene clusters detected in reference genome collections or in a metagenomic assembly, providing normalized counts across samples, performing differential analyses, and visualizing the results. It requires three main inputs: (1) a gene cluster collection obtained from running any "SMASH-based" algorithm, (2) the meta'omic dataset in FASTQ or FASTA format or, alternatively, the Sequence Read Archive (SRA) accession numbers to download it, and (3) a metadata file with sample information to segregate them into groups and compare their gene cluster content.

BiG-MAP is composed of four different modules (see Figure 6.1): (1) BiG-MAP.family, which performs redundancy filtering on the input collection of predicted gene clusters and provides a set of representative gene clusters for the mapping process. (2) BiG-MAP.download, which uses a list of SRA accession ids to download the shotgun data if present in the SRA database (this step is optional). (3) BiG-MAP.map, which maps reads from the metagenomic or metatranscriptomic samples onto the set of representative gene

clusters obtained from BiG-MAP.family. (4) BiG-MAP.analyse, which normalizes the counts for sparsity and sequencing depth, performs differential abundance/expression analysis and visualizes the output.

The BiG-MAP.family module performs a redundancy analysis on the gene cluster collection to remove almost-identical sequences, in order to reduce the computing time and avoid problems with ambiguous read mapping. To achieve this, the protein sequences encoded in each of the gene clusters are used as input for MASH[153], a MinHash-based algorithm to estimate sequence distance. Next, a representative gene cluster is selected based on a medoids calculation. The resulting representatives are then clustered into Gene Cluster Families (GCFs) using BiG-SCAPE[124], an algorithm that uses three different distance metrics to group MGCs into families based on sequence and architectural similarity. This step helps to group more distantly related homologous gene clusters that likely have the same chemical products but that are encoded in more distantly related organisms. In such cases, BiG-MAP maps reads to the family representatives separately, but also allows reporting combined abundance or expression levels per family to find associations with phenotypes at a higher level. In order to set an expression baseline when using metatranscriptomic data, BiG-MAP screens bacterial genomes whose gene clusters have been included in the non-redundant representative set of gene clusters for five house-keeping genes known to have stable expression levels using HMMer (for details, see Methods section 6.4.3). Next, the reads are mapped to the representative gene clusters using the short-read aligner Bowtie2[180]. The obtained raw read counts are then converted to RPKM (Reads Per Kilobase Million) values, which are summed across all representative MGCs within a GCF when reporting family abundances (raw counts are also outputted for each representative MGC). In the last module, RPKM values are then normalized using Cumulative Sum Scaling[181] (CSS) to account for sparsity. Moreover, for each aligned gene cluster, we assess its coverage to control for gene clusters that are only partially mapped to by meta'omic reads. We report two coverage values in the

intermediate files; one for the whole gene cluster and the other considering only the core genes of the BGC/MGC; showing both these numbers is often insightful in cases where borders of gene clusters called by antiSMASH or gutSMASH are imprecise and reads may be mapped to regions flanking the actual gene cluster. Subsequently, BiG-MAP detects differentially abundant or differentially expressed gene clusters by using either zero-inflated gaussian distribution mixture models (ZIG-models) or using a Kruskal-Wallis model. Finally, all the generated results are displayed in a plot that includes a heatmap for the gene clusters abundance/expression values, a bar plot for the log fold change, the coverage values and finally another heatmap for the housekeeping gene expression values when analyzing metatranscriptomes (see Suppl. Figure. S6.2). The output folders contain various intermediate and final results, including the BiG-SCAPE results, the resulting bedgraphs, the raw and normalized RPKM counts for each sample (in BIOM format[182]) and the results of the fitZIG and Kruskal Wallis tests in tab-separated tables, and mapping coverage values for each gene cluster and sample. Altogether, this tool presents a streamlined method to functionally profile meta'omics data by mapping reads to known or putative gene clusters.



GCF (Gene Cluster Family), WGS (Whole Genome Sequencing), DE (Differential Expression)

Figure 6.1 BiG-MAP workflow. BiG-MAP is composed of four different modules: (1) BiG-MAP.family, which, given a set of predicted gene clusters by either gutSMASH or antiSMASH, returns a representative set of non-redundant gene clusters based on sequence similarity (this module also identifies the protein sequences of 5 housekeeping genes from the bacterial genomes that encode the representative gene clusters when metatranscriptomes are used); (2) BiG-MAP.download, to download a set of metagenomes/metatranscriptomes given their SRA accessions; (3) BiG-MAP.map, to align omics reads to the representative set of gene clusters using Bowtie and (4) BiG-MAP.analyse, to normalize and perform differential abundance/expression analysis of gene clusters across different conditions and visualize the results (see Suppl. Figure S6.1 and S6.2 as an example).

6.2.2 Assessing and validating BiG-MAP performance using simulated data

In order to evaluate the overall performance of BiG-MAP and in particular, all the default parameters chosen as defaults, such as the Bowtie alignment mode and the MASH similarity score cut-off, we designed a mock microbial community for metagenome simulation. From the Culturable Genome Reference (CGR) genome collection[128], we randomly chose 101 CGR genomes to simulate metagenome reads from and to use as input for gutSMASH. To assess the impact of different sequencing depths (coverage of 0.5x and 0.05x) and community structure (uniform, linear, power-law and exponential), we simulated eight different metagenomic libraries. Since the gene cluster content and their abundance levels in simulated data are known (ground truth), this allowed us to assess the recall and precision of the BiG-MAP abundance calculations using MASH dissimilarity scores ranging from 10-100 and the eight different alignment modes available in Bowtie across the eight different simulated data libraries. From these results we computed the F1-score or harmonic mean of precision and recall (see Figure 6.2), which showed that the community structure slightly affects BiG-MAP results. Moreover, since the highest F1 scores were obtained when using a MASH score cut-off (similarity) of 0.8 and using the “fast” alignment mode (end-to-

end) of Bowtie2, we set these parameters as defaults. Still, the user is able to change them as desired by indicating this with the appropriate flag.

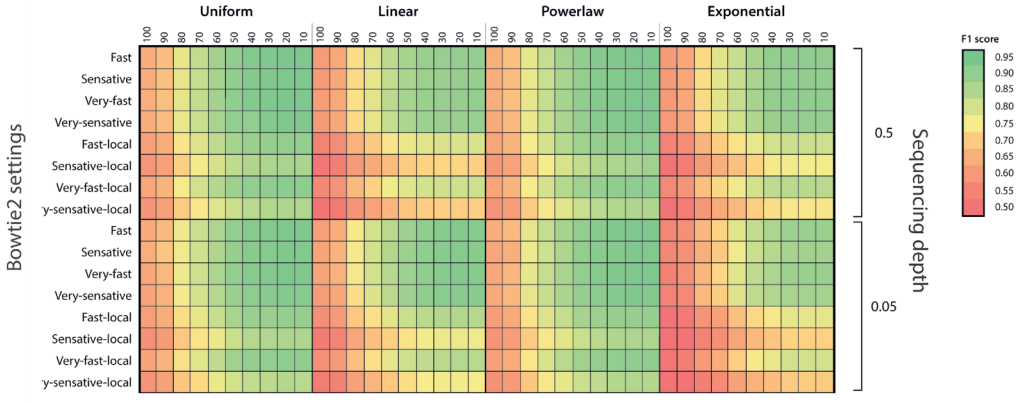


Figure 6.2 BiG-MAP validation using a mock community. F1 score heatmap using eight different synthetic communities built to assess the best MASH dissimilarity cut-off across four different community structures, two different sequencing depth values and eight different Bowtie alignment modes.

6.2.3 Analysis of the oral microbiome: revealing the presence of gene clusters associated with health and disease

The oral cavity is a natural habitat for many bacteria that reside in or on the gingival sulcus, tongue, teeth and cheeks, among other surfaces. These bacteria take part in important processes such as initial digestion of food, but are also associated with several oral diseases such as caries[183] and periodontitis[184]. It is known that these bacteria can organize themselves to form biofilms, which can play a causal role in the development of these diseases[185]. There are different functional and metabolic pathway alterations that have been associated with the onset of disease via the production of small molecules[186,187,188,50]. For instance, tetramic acids produced by the caries-associated bacterium *Streptococcus mutans* have been linked to tooth decay[189]. For this reason, in order to functionally profile these oral communities and acquire further insights into the MGCs that might be involved,

we studied a dataset of 47 oral microbiome samples[188] using BiG-MAP, for which paired metagenomics and metabolomics data are publicly available (see Methods sections 6.4.7 and 6.4.8).

To evaluate possible molecular mechanisms underpinning caries formation, we first analyzed the available MS/MS data together with the metabolite feature abundance table using Pathway Activity Level Scoring (PALS)[190], which uses molecular families obtained using molecular networking[191] to group similar metabolites, and PLAGE[192] to find differentially expressed metabolite groups between two conditions. PALS showed a very consistent and strong differential abundance between healthy and caries volunteers of a number of features in a metabolite group that we could annotate with polymer-like structures based on their C_3H_4O mass differences. With MASST searches[193] across all public data present in GNPS-MassIVE, we could confirm the occurrence of these differential features in various microbial, human, and environmental-related public datasets (see Methods and Supplementary Methods for further information on the metabolomics data analysis). Based on this, we concluded that these polymer-like structures could well represent molecules called polyacroleins (metabolite identification level 3 - annotated compound class), which are known to spontaneously form from a component of the antimicrobial set of molecules called reuterin[194]. The formation of (poly)acrolein has been shown to contribute strongly to the antimicrobial activity of reuterin[194]. Reuterin is produced by lactobacilli from a genomic island containing a *pdu-like* operon together with a cobalamin biosynthetic gene cluster[195]. Of note, acrolein is an ubiquitous compound that can be found in the human body for various other reasons as well, such as endogenous production, the ingestion of food sources or due to exposure to certain environmental conditions[196]. There are various known routes that can converge into the formation of acrolein, as it can be formed spontaneously from glycerol and 3-hydroxypropionaldehyde[194]. Furthermore, glycerol metabolism from gut bacteria has also been found to produce this molecule[197]. Typically, the acrolein polymerization occurs under alkaline conditions[198],

thus, it is more likely to accumulate in saliva from healthy samples, as caries typically acidifies the oral cavity. Indeed, our results show that the possible polyacroleins are more abundant in samples of healthy volunteers. Interestingly, the presence of acrolein has been linked to inhibition of *Streptococcus mutans*, a well-known species of cariogenic bacteria[199,200].

Based on these findings, we were motivated to look for the presence of the *pdu* operon in the metagenomics samples, in order to identify candidate MGCs that might be involved in acrolein formation. To this end, we ran gutSMASH on the 1,440 genomes from the Human Microbiome Oral Database (HMOD, http://www.homd.org/) available in April 2020. Interestingly, gutSMASH identified a *pdu*-like operon in the genome of *Streptococcus sp. F0442* that also includes a cobalamin (vitamin B12) biosynthetic region and is architecturally similar (cumulative Blast bit score of 13,271) to the *Lactobacillus reuteri* one (see Fig. 3A). Therefore, to assess the abundance of the predicted gene clusters in the oral microbiome we used our gutSMASH run, which predicted 3,352 gene clusters, as input for the BiG-MAP.family module, to filter out redundant MGCs. Next, the reads of the 47 oral metagenomes (24 healthy and 23 caries-related) were mapped onto the 1,544 representative gene clusters using BiG-MAP.map and the counts were further normalized and parsed with BiG-MAP.analyse. We found that 59 gene clusters predicted by gutSMASH were significantly differentially abundant between caries-related and healthy samples when using Kruskal Wallis. Despite the fact that the *pdu* operon was not among these, we could see that it was still somewhat more abundant in healthy samples (mean: 6.014 RPKM counts/sample) when compared to the diseased group (mean: 4.81 RPKM counts/sample). Motivated by this, we sought to assess its presence in a larger oral microbiome dataset by using 48 paired publicly available paired-end metagenome samples, which also included metagenomes from samples suffering from periodontitis and plaque formation, all considered as disease-related samples. These were used in combination with the already analyzed ones, making a total of 96 samples; 33 caries-related, 34 healthy,

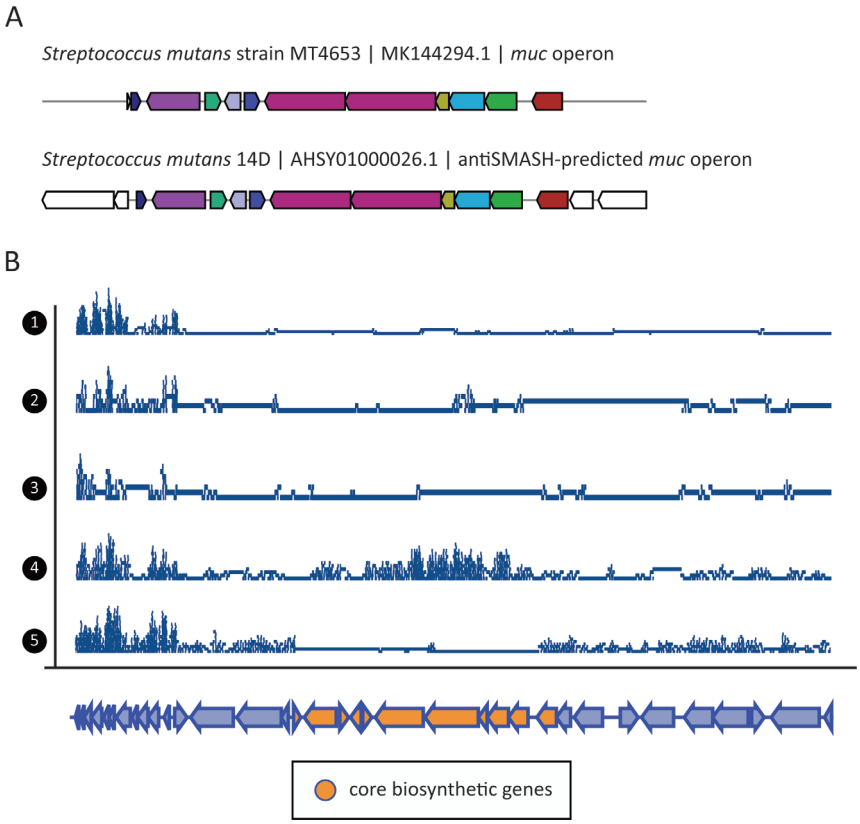
10 periodontitis-related and 19 involved in plaque development and all were used as input for BiG-MAP (see Methods section 6.4.7). From this run, we found 246 gene clusters differentially abundant between groups (using a Kruskal Wallis test) and 173 gene clusters when only considering the core region encoding the key enzymes used to detect the MGC. Among the significantly differentially abundant gene clusters when mapping reads to the core region, we found the *pdu* operon. While healthy samples on average have 5.62 RPKM counts/sample mapping to this gene cluster, diseased ones have 3.38 (p-value= 0.00049 using Kruskal Wallis). We also evaluated the coverage of the read mapping the core genomic region within the expanded metagenomic datasets and found that within healthy samples, not all samples contain this gene cluster. For instance, from 34 healthy samples in the extended dataset, we could find 13 of them that appear not to have the *Streptococcus* sp. F0442 *pdu* operon (coverage below 0.5), while the rest had fairly high coverage scores with a mean coverage value of 0.77 (selecting the samples with coverage values of at least 0.5), implying the presence of this operon or a close homologue of it (see Figure 6.3B). Overall, this MGC constitutes a potential source for polyacrolein production, and might be involved in inhibition of *Streptococcus mutans* strains in non-acidic conditions. As logically, expression of the MGC would be required for conferring a metabolic and potentially disease-suppressive phenotype, metatranscriptomics analysis of samples where polyacrolein accumulation is observed could be an interesting follow-up analysis in the future to test the hypothesis of the involvement of this MGC in its production. Altogether, our study illustrates how BiG-MAP analysis, especially when combined with complimentary omics data such as metabolomics, can generate concrete hypotheses about microbiome-associated phenotypes that can be tested in the laboratory.

Figure 6.3 *Pdu* and cobalamin operon abundance in healthy oral metagenomes. (A) MultiGeneBlast comparison between the *pdu* operon found in *Streptococcus sp.* F0442 by gutSMASH and the characterized one from *Lactobacillus reuteri* (AP007281). (B) Read coverage of five randomly chosen healthy metagenomes along the gutSMASH-predicted *pdu* gene cluster. The coverage graphs, which were plotted using the Sushi R package (version 3.5.1)[154], show that some samples (3 and 4) contain reads that cover the whole gene cluster, while in other samples, reads hardly cover the cluster (1 and 2) or only part of it (5).

Another example of a gene cluster that has been found relevant in the oral cavity is the *muc* operon, which has been shown to be responsible for the production of tetramic acids, which are known to inhibit the colonization of commensal bacteria in the oral cavity. This gene cluster encodes a hybrid between a polyketide synthase and nonribosomal peptide synthetase (PKS/

NRPS)[189]. In order to further test this association and assess whether there is a difference in abundance of the *muc* operon in the oral cavity between healthy and diseased samples, a collection of 170 *Streptococcus mutans* genomes collected from Tang *et al*[189] and Liu *et al*[201] was run through antiSMASH[125], which predicted a total of 1,849 BGCs. After obtaining 41 representative gene clusters with the BiG-MAP.family module, reads from the 47 oral microbiome metagenomes were mapped onto the predicted gene clusters and further processed using BiG-MAP.map and BiG-MAP.analyse subsequently. From the results, three gene clusters were found to be significantly differentially abundant between healthy and diseased samples when using the fitZIG model and one when using Kruskal-Wallis. However, among them, the *muc* operon from *Streptococcus mutans* 14D was not found. When we analyzed an extended dataset (96 samples, see Methods section 6.4.8), we did find that in this larger dataset, the *muc* operon is differentially abundant between healthy and disease samples when using the fitZIG model (but not using Kruskal-Wallis), together with five other gene clusters. The *muc* operon from this strain shows high similarity to the one characterized by Tang *et al.*[189] (see Figure 6.4A). However, the mean read coverage of the MGC core in both groups was low; 0.274 in healthy and 0.161 in caries-associated samples, which implies a relatively low abundance of the *muc* operon in many samples and/or partial coverage of the *muc* gene cluster locus with reads mapping only to flanking regions (see Figure 6.4B). Nonetheless, within both groups we see that some samples have reads mapping to the complete gene cluster, with coverage values close to 1. When filtering out samples with coverage values < 0.5, leaving only 8 samples in the healthy group and 6 in the disease group, the mean coverage amounts to 0.848 in healthy and 0.997 in disease. The fitZIG BiG-MAP output heatmap (Suppl. Figure S6.3) shows that, despite the fact that the *muc* operon is significantly more abundant in diseased samples according to the test, the abundance of this gene cluster across all samples looks generally very similar. In addition, the fold change between the averages for the healthy and diseased phenotypic states is minimal, both with and

without filtering. Therefore, despite finding this operon to be slightly more abundant in caries-prone samples when applying the fitZIG model, the oral microbiota from healthy donors seem to also harbor this PKS/NRPS MGC at very similar levels. Hence, the microbiota from healthy samples may have a mechanism to counteract the inhibiting effect of tetramic acids, or there might be a difference in expression of the gene cluster between healthy and diseased subjects.



Figue 6.4 *Muc* operon overrepresentation in caries-related samples. (A) Multi-GeneBlast comparison between the *muc* operon characterized from *Streptoccocus mutans* strain MT4653.1 and the antiSMASH predicted one from *Streptococcus mutans* 14D. (B) Read coverage of five randomly chosen caries-related metagenomes along the antiSMASH predicted *muc* gene cluster. The coverage graphs, which were plotted using the Sushi R package (version 3.5.1)[154] show that despite the

fact that the *muc* operon is generally not very highly covered by reads from the randomly picked samples, some seem to truly contain this operon, such as sample 4, where the core biosynthetic genes appear abundant at sufficient levels. Full data on all samples can be found in Figure S6.3. It is important to note that the results found in this study come from analyzing a large collection of metagenomes that may largely differ at the community structure level. The 96 samples are not only healthy or caries-associated metagenomes but also metagenomes from patients suffering from periodontitis and samples from a study that observes how a biofilm evolves over time. Therefore, it might be that all these samples differ quite a lot in terms of MGC content but also regarding the presence of *Streptococcus mutans*, influencing the signal of significance reported here. All in all, our results suggest that the abundance of the *muc* operon is not very predictive for a healthy or disease state of the microbiome by itself, and other factors likely play (more) important roles.

6.3 Conclusions

Overall, combining different omics data is a very useful approach to understand which microbes are doing what and poses a promising avenue to better understand complex biological processes. Here, we presented BiG-MAP, a command-line tool that it is able to profile the abundance and expression of a collection of gene clusters across metagenomic and metatranscriptomic data. Each of the steps in the BiG-MAP pipeline is robust, as demonstrated using a mock community. Indeed, BiG-MAP can discover interesting and relevant potential associations between genomic regions and phenotypes, which can guide experimental efforts to test MGC function. It is worth noting the usefulness of the gene cluster mapping coverage values, since they allow the user to discern between the real presence of predicted gene clusters of interest and spurious read mapping. Also, the associations that can be found using BiG-MAP strongly depend on the WGS data sequencing depth and sample size. For instance, in the examples described in our study, we found both gene clusters (the *pdu*-like operon and *muc*) only significant in either dataset (reduced or extended one). Both examples illustrate how users would do well to study expression or abundance values of individual samples and should not draw conclusions too quickly based

on the results of a single statistical test. From the BiG-MAP output folders, which include raw and processed results, it is possible to extract valuable information to this end, such as the differences within groups, coverage distribution of reads across a gene cluster, etc. In combination with detailed sample metadata, this can help provide insights into the microbially derived phenotypes. Overall, we believe that BiG-MAP will help researchers solving biologically complex questions by integrative multi-omics approaches, to obtain deeper insights into the relationships between microbial metabolic capacities and microbiome-associated phenotypes.

6.4 Methods

6.4.1 Code availability

BIG-MAP is implemented in Python 3 as a command line package. It consists of four modules: BIG-MAP.download, BiG-MAP.family, BiG-MAP.map, and BiG-MAP.analyse. The code is available at: https://github.com/medema-group/BiG-MAP together with documentation on how to install BiG-MAP and its dependencies and a short tutorial on how to run it.

6.4.2 BiG-MAP.download: Data collection

This module allows to retrieve sequencing data present in the SRA database using the SRA toolkit (https://github.com/ncbi/sra-tools). To initially develop, test and validate this tool, we used an IBD cohort that contains metagenomic and metatranscriptomic data from 78 individuals, 21 suffering from UC, 46 individuals with CD, and 11 healthy samples[202]. These samples were retrieved using the SRA accession IDs under BioProject PRJNA389280.

6.4.3 BiG-MAP.family: Creating a non-redundant MGC representative collection

The family module uses as input a directory that contains GenBank files of gene clusters identified by the antiSMASH[73] or gutSMASH algorithms (https://github.com/victoriapascal/gutsmash). The predicted gene clusters are then subjected to a redundancy filtering step based on their mutual sequence similarity. For that, the protein sequences of the gene clusters are extracted and used as input for MASH[153] sketch, which creates sketches from the raw sequences. The sketches are then used to calculate the distances between sequences using MASH dist. The resulting tab-delimited file with the pairwise distance comparisons is used to group together gene clusters with above a 0.8 default similarity cut-off (see Figure 2). Next, to pick the best representative of each group, medoids are computed (see formula below). For this, a distance matrix is created comparing all distances between pairs of gene clusters; the one with minimal cumulative distance value is picked as representative of that group. Additionally, the selected gene clusters are subjected to another round of clustering using BiG-SCAPE[124], to cluster gene clusters into GCFs at a 0.3 similarity cut-off (default value), from which a random representative is picked.

$$x_{medoid} = argmin_{y \in \{x_1, x_2, \ldots x_n\}} \sum_{i=1}^{n} d(y, x_i)$$

If metatranscriptomes are used in the BiG-MAP.map module, an additional step is performed to set an expression baseline. For this, the protein sequences of the genomes whose gene clusters form the non-redundant representative gene cluster collection are scanned using hmmsearch (hmmsearch version 3.1b2) for five housekeeping-coding proteins: DNA gyrase A (PF00521), DNA gyrase B (PF00204), Recombinase A (PF00154), DNA directed RNA polymerase A (PF01000), and DNA directed RNA poly-

merase B (PF00562). The selection of these Pfam domains was based on the findings by Rocha *et al.*[203] that these housekeeping genes show highly stable expression across samples. Next, the gathered protein sequences are also used as queries in the mapping module to align metatranscriptomic reads to gene clusters.

6.4.4 BiG-MAP.map: mapping reads to a non-redundant gene cluster collection

This module relies on Bowtie2[180] (version 2.3.4.3) to align reads to a given sequence. From the reference gene cluster sequences selected by the medoid calculation, Bowtie index files are created. Next, Bowtie2 aligns reads to these index files, using the fast alignment mode by default. The resulting alignment is stored in SAM format and converted to BAM format to later be parsed by SAMtools[204] (version 1.9). The alignments are then sorted by their leftmost coordinates, the aligned reads are counted and either summed by GCF or averaged over the GCF size. Later, the corrected raw counts are converted to TPM counts (Transcripts Per Kilobase Million) and consecutively to RPKM (Reads Per Kilobase Million) counts to account for sequencing depth.

Another functionality that was added in this module was to compute the read coverage of each gene cluster using the coordinates in the sorted BAM files. To do so, the sorted alignment files are converted to bedgraphs using BEDtools[205] (v2.28.0), which allow estimating the number of covered bases for each cluster (*coverage*) by subtracting the number of non-covered bases (*ncb*) to the length of each cluster (*cl*) as indicated in the formula below.

$$coverage = \frac{cl - ncb}{cl}$$

The same procedure is followed to compute the RPKM counts and the coverage of the core genes within a gene cluster, which strictly considers the core metabolic genes within each gene cluster. This information is taken from the antiSMASH/gutSMASH (or any other "SMASH" related algorithm) Genbank output files that flag the key coding genes that are needed for the synthesis of a given molecule. Once the core genes are identified, their alignment information is retrieved using SAMtools. Next, in the same manner as RPKM values are computed for the whole gene clusters, reads aligned to the core region are extracted, counted and corrected to finally get the RPKM counts. To perform the coverage calculation, the locations of the core genes are extracted from the bedgraph to evaluate the coverage.

6.4.5 BiG-MAP.analyse: Normalization of RPKM counts and finding differentially expressed/abundant MGCs

In order to account for sparse high-throughput sequencing, RPKM values are normalized using Cumulative Sum Scaling (CSS) from the R Bioconductor package MetagenomeSeq[181]. BiG-MAP offers two different statistics to account for differentially abundant/expressed gene clusters: the parametric zero-inflated gaussian distribution mixture model (ZIG-model) or the non-parametric Kruskal-Wallis test. ZIG-model values are adjusted with log2 fold-change that ultimately helps fitting the model to a log-normal distribution, thus when the abundance/expression values are expected to follow normal distribution, ZIG-mode is more appropriate to use. Alternatively, Kruskal-Wallis can be run on the normalized RPKM counts, which allows assessing whether the distribution of ranks for one group significantly differs from the distribution of ranks for the other group. Additionally, FDR correction is applied to correct for multiple hypothesis testing. Finally, heatmaps are produced to visualize the results using the Seaborn Python package (https://github.com/mwaskom/seaborn).

6.4.6 Testing BiG-MAP performance using a mock community

To test BiG-MAP performance, 101 bacterial genomes were randomly chosen from the CGR collection[128] (PRJNA482748 BioProject id). Thus, the gutSMASH-predicted MGCs from each genome were used as ground truth (https://github.com/victoriapascal/gutsmash, version 0.8, github commit stamp: 569e860). Next, paired-end reads were generated with a mean read length of 100 bp from the 101 CGR bacterial genomes using Grinder v0.5.3[206]. Two different read coverage thresholds were used (0.5x and 0.05x) in combination with four different community structures: uniform, linear, power-law and exponential. Both the MGCs and the simulated reads were used as input for BiG-MAP, which was run ranging the MASH similarity thresholds between 10-100% in intervals of 10% along the eight different Bowtie2 alignment modes. From each individual run, true positive, false positive and false negatives rates were calculated to evaluate the precision and recall, which was ultimately used to compute the harmonic mean of precision and recall, also known as the F1-score. The results were plotted in a heatmap using the ComplexHeatmap package in R[207].

6.4.7 Assessing the *pdu* operon abundance by surveying different oral metagenomic samples.

To find possible leads on metabolic perturbances between healthy and caries-related samples, the processed mass spectra (MGF format) and metabolomics feature tables from Aleti *et al.*[188] were downloaded from GNPS-MassIVE[191] accession ID MSV000081832 to perform reanalysis. Feature-based Molecular Networks[208] were run using GNPS release version 21 (https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e4f64542ab24a7b0802ceacbcfa071, https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9c95754d1fdc42b4a43b16919c398ecd).

The resulting molecular family information together with the metabolite feature tables and sample information (metadata) were loaded into PALS (https://pals.glasgowcompbio.org/app/)[190], to identify metabolite families differing in activity between 25 healthy and 24 caries-related samples. From the results, three out of seven candidate metabolites in one differentially expressed molecular family showing clear different abundance patterns between healthy and caries samples were further examined using GNPS MASST (https://masst.ucsd.edu)[193], the ChemCalc MF finder[209], and PubChem[210], leading to the putative annotation of polyacrolein-related metabolites in healthy samples, which may be produced from a *pdu*-like operon that requires the presence of the cobalamin biosynthetic genes (see Supplementary material for further information).

For the analysis of the *pdu* operon and its presence in the oral microbiome, 1,440 oral bacteria genomes were downloaded from the HOMD collection (http://www.homd.org/?name=GenomeList&link=GenomeList&type=all_oral). Next, these genomes were used as input for gutSMASH (version 0.8). The comparison between the two *pdu* operons from *Lactobacillus reuteri* (AP007281) and *Streptococcus* sp. F0442 (GCA_000314795.2) was done using MultiGeneBlast[150]. Next, all predicted clusters were used as input for the BiG-MAP family module. At the same time, the oral metagenomics datasets were downloaded using the BiG-MAP.download module by providing the SRA accession IDs associated to the PRJNA478018, PRJNA396840, and PRJNA398963 BioProject IDs. Once the metagenomes were downloaded, BiG-MAP.map was run using the output of the family module and the metagenomic reads in FASTQ format. Finally, the RPKM counts were normalized, processed and visualized using BiG-MAP.analyse.

6.4.8 Evaluating the presence of the *muc* operon in caries-associated metagenomes

AntiSMASH was used to predict BGCs from a total of 170 *Streptococcus*

mutans genomes reported in Tang *et al*[189] and Liu *et al*[201]. Within the predicted BGCs, the *muc* operon was found and compared to the *muc* operon characterized by Hao *et al*.[211] using MultiGeneBlast[150]. The predicted BGCs were then used as input for the BiG-MAP.family module. Both, the representative BGCs and metagenomic reads were then used as input in the subsequent BiG-MAP.map mapping module using the metagenomes from the following three BioProjects: PRJNA478018, PRJNA396840, and PRJNA398963. Finally, the raw mapping counts were normalized and further processed and visualized using BiG-MAP.analyse.

6.5 Data availability

The supporting information for this article can be found in the supplementary material of this article and in the Zenodo repository (https://zenodo.org/) with the following DOI: 10.5281/zenodo.4656097. The metabolomics data used for reanalysis is available from GNPS-MassIVE accession ID MSV000081832.

6.6 Acknowledgements

We particularly thank Daria Zuzanna Świgoń, Arno Hagenbeek, Sarah van den Broek, Jeanine Boot and Robert Koetsier for preliminary results on the *pdu* operon, which provided us the lead to further explore these datasets. We also acknowledge the guidance provided by Rens Holmer in the early stage of this study and Dr. Madeleine Ernst for her help in locating the relevant files of the relevant metabolomics data files from the Aleti *et al*. study.

6.7 Supplementary Methods

6.7.1 Metabolomics analyses

GNPS Molecular Networking was done as follows: A molecular network was

created with the Feature-Based Molecular Networking (FBMN) workflow on GNPS (https://gnps.ucsd.edu). The data was filtered by removing all MS/MS fragment ions within +/- 17 Da of the precursor m/z. MS/MS spectra were window filtered by choosing only the top 6 fragment ions in the +/- 50 Da window throughout the spectrum. The precursor ion mass tolerance was set to 0.02 Da and the MS/MS fragment ion tolerance to 0.02 Da. A molecular network was then created where edges were filtered to have a cosine score above 0.65 and more than 5 matched peaks. Further, edges between two nodes were kept in the network if and only if each of the nodes appeared in each others respective top 10 most similar nodes. Finally, the maximum size of a molecular family was set to 100, and the lowest scoring edges were removed from molecular families until the molecular family size was below this threshold. The analogue search mode was used by searching against MS/MS spectra with a maximum difference of 200.0 in the precursor ion value. The library spectra were filtered in the same manner as the input data. All matches kept between network spectra and library spectra were required to have a score above 0.7 and at least 6 matched peaks.

GNPS Molecular Networking job of mass spectrometry data: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9c95754d1fdc42b4a43b16919c398ecd MASST searches were performed for a number of features that showed typical C3H4O mass differences using the default settings. A single spectrum search was completed using the online workflow (https://ccms-ucsd.github.io/GNPSDocumentation/) on the GNPS website (http://gnps.ucsd.edu). The data was filtered by removing all MS/MS fragment ions within +/- 17 Da of the precursor m/z. MS/MS spectra were window filtered by choosing only the top 6 fragment ions in the +/- 50 Da window throughout the spectrum. The precursor ion mass tolerance was set to 2.0 Da and a MS/MS fragment ion tolerance of 0.5 Da. The library spectra were filtered in the same manner as the input data. All matches kept between input spectra and library spectra were required to have a score above 0.7 and at least 6 matched peaks.

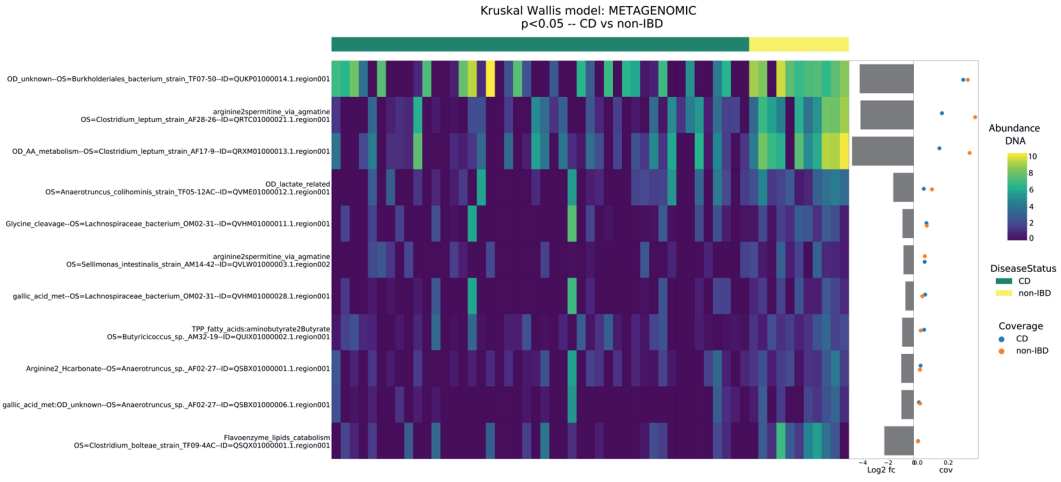
Feature ID 3779, precursor m/z 680.4799:
https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=71831f592f27496faf-16c62bc16b1b69&view=view_all_datasets_matched

Feature ID 153, precursor m/z 722.4900:

https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=79bc074e6e274f05ac-4d3a273c50681a&view=view_all_datasets_matched

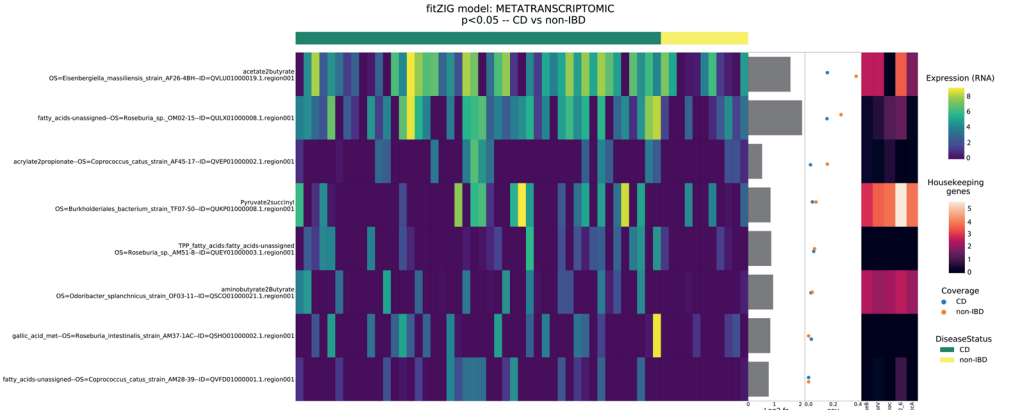
Feature ID 126, precursor m/z 663.4528 :

https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=8742b10dda2a4e29b-c405fdc34aa1aa6&view=view_all_datasets_matched

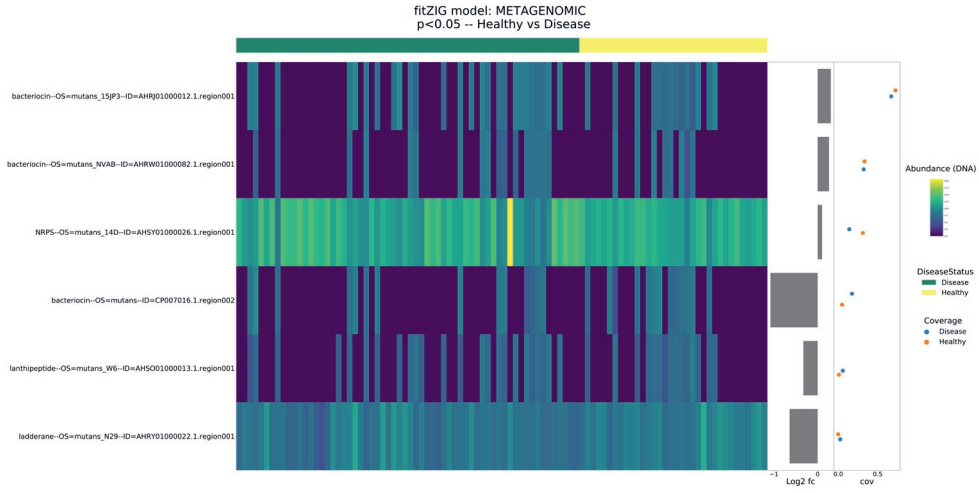


Supplementary Figure S6.1 Illustration of BiG-MAP analysis module output when using metagenomics data. In this case, for testing and developing purposes, the CGR genomes were used as input for gutSMASH. The resulting predictions were used as input together with the metagenomics samples from Schirmer *et al.* (PRJNA389280). The significant differential abundance of 11 gene clusters across Crohn disease (CD) samples and healthy (non-IBD) using Kruskal Wallis are shown in the heatmap. The more abundant a gene cluster is, the more yellow it is shown in the heatmap. Next to the heatmap, the bar chart represents the abundance log2 fold-change values. On the far most right, the dots represent the coverage values

computed by BiG-MAP to show how evenly the reads map along the whole gene cluster, in blue for the CD samples and in orange for the healthy.



Supplementary Figure S6.2 Illustration of BiG-MAP analysis module output when using metatranscriptomic data. In this case, for testing and developing purposes, the CGR genomes were used as input for gutSMASH. The resulting predictions were used as input together with the metatranscriptomes from Schirmer *et al*. The significant differential expression of 8 gene clusters across Crohn disease (CD) samples and healthy (non-IBD) using Kruskal Wallis are displayed in the heatmap. The higher the expression of a gene cluster is, the more yellow it is shown in the heatmap. Similarly to Figure S6.1, the log2 fold-change and coverage values are also depicted. When analyzing metatranscriptomic data, it is also included a heatmap with the expression values of five housekeeping genes that allows to set an expression baseline.



Supplementary Figure S6.3 The *muc* operon (AHSY01000026.1.region001), predicted by antiSMASH, is significantly enriched in disease-associated samples when applying the FitZIG model. Despite the significance, the difference in abundance between both groups is minimal, as the similar coloring depicts and the log2 fold-change bar chart evinces. The plot shows another five gene clusters enriched in either group. The heatmap has been produced by the BiG-MAP.analyse module using the 96 oral microbiome metagenomes.

Chapter 7

General discussion

Developments in sequencing techniques and bioinformatic tools to analyse genomic data enable the study of the human microbiome from multiple perspectives. In the past decade, great advances have been made towards understanding its composition in relation to external factors such as diet, how they modulate our immune system and also regarding underlying mechanisms by which certain communities trigger specific microbe-derived phenotypes. Hence, the human microbiota can be regarded as another organ that exerts a remarkable influence on the host, but the intrinsic dynamics and functioning remain mostly unknown. This thesis aims to contribute to this field by providing various computational methods to predict, analyse and mine genomes for metabolic gene clusters responsible for the synthesis of small molecules, which are important mediators of microbe-microbe and microbe-host interactions. Moreover, we have utilized these tools to profile the human microbiome, revealing important aspects of its metabolism by assessing the pathway taxonomic distribution, uncovering a large collection of putative gene clusters classified into different classes and assessing the abundance and prevalence of known pathways in a healthy population-based cohort. After presenting and showcasing the use of these tools and reporting new leads on how microbes contribute to the overall pool of excreted metabolites, the aim of this discussion is to provide new ideas on how these tools could be updated and improved in the near future. In addition, we evaluate how all the data generated in this thesis could be stored in public databases for users to search and mine in different ways and describe which strategies could be used to integrate knowledge from different “omic layers” using some of the tools and data here presented.

7.1 Meeting the needs: future directions

Despite the challenges that profiling microbial community functions poses, it is essential to understand which causal agents (metabolites) are driving the specific host phenotypes. In this thesis we have presented new tools and approaches to better understand the molecular mechanisms behind

certain phenotypes. gutSMASH can be a valuable tool in the field to guide enzyme/pathway discovery, link metabolites to genes and identify genes responsible for microbiome-associated phenotypes. As described in Chapter 4, gutSMASH uses detection rules designed based on the profiling of a manually selected set of pathways. Despite the fact that this set was meant to be as complete as possible, other relevant pathways from literature might have been missed. This is a limitation that is partially alleviated by the inclusion of general rules, as shown in the supplementary material of Chapter 4. Nonetheless, it is a matter of time until new studies uncover novel pathways to synthesize new metabolites or alternative ones that are linked to already known molecules. This could also affect the proposed MGC classification, requiring the inclusion of additional categories or expanding the current ones. For this reason, gutSMASH should stay in development to account for novel gene clusters, in order to have an up-do-date prediction tool. This has also been the case for antiSMASH, the predecessor tool that has been updated every two years by the scientific community.

Like many other tools, gutSMASH could benefit a lot from community-driven efforts and crowdsourcing. As a community asset, a wiki-style platform could be devised so that users could help pinpoint relevant gene clusters that are missed in the current version. In this way, special efforts could also be made to better cover the metabolic capacities found in other body sites. As a specific example, gutSMASH predicted a varied array of MGCs from the oral microbiota, including a *pdu*-like MGC that protects the oral cavity from cariogenic bacteria (reported in Chapter 6). he metabolic profiles of such communities are still largely unannotated[187] but are known to be defining for the different oral microhabitats that exist in the oral cavity[212]. Thus, to account for this variability, a more exhaustive literature search could be done to check if any new detection rules could be designed to predict primary MGCs typically found in the oral microbiome.

Besides new detection rules that could be included in further gutSMASH versions, community efforts could also help design new functionalities of interest based on user experience and preferences. For instance, it could be interesting to show for each predicted MGC its abundance and expression profiles in a given dataset. For this, the MGC collection derived from running gutSMASH on the three genome collections described in Chapter 4 could be used as input for BiG-MAP, which would also need the data of a population-based cohort as a second input. These pre-calculated results could be then linked in the gutSMASH output whenever the same MGC type is predicted. The association could be made by calculating sequence similarity between the predicted MGCs and the ClusterBlast entries, displaying the corresponding BiG-MAP results whenever this exceeds a pre-set cutoff. Linking results obtained from surveying different omics layers can provide information on how common it is to find a given pathway in the human microbiome and how abundant it is, but also to see if the pathway plays an active role in that ecosystem or not based on metatranscriptomic data.

Another possible improvement of the tool could be to enhance the interface and make it more suitable for the analysis of metagenome assemblies. Metagenomes are a gold mine to uncover and annotate novel genes, and gutSMASH could be a very useful approach to predict the metabolic potential of a whole community and help annotate functions to yet unknown genomic regions. Currently, gutSMASH is able to predict MGCs given a metagenomics sample, but the interactive output is not the most appropriate to visually screen all the results. An alternative could be to group the predicted MGCs based on either MGC types or taxonomic assignments provided by the user. Thus, instead of scrolling through all the MGCs predicted from the inputted metagenomic sequence, the user could choose to (1) overview the metabolic potential of the microbial community or (2) specifically evaluate the metabolic potential of each taxon. In option 1, all MGCs could be classified based on the detection rules used and, similarly to what we have done to fine tune the gutSMASH version 1.0 detection rules, BiG-

SCAPE could then be run on each sub-collection (predicted MGCs classified by type). In this way, the BiG-SCAPE output could be integrated in the gutSMASH visualization. This would be useful for different reasons. First, it would provide a summary and some statistics of the MGCs belonging to each sub-collection. Second, there would no longer be a need to display identical MGCs multiple times; and third, it would allow users to decide which MGCs are worth looking into in more detail. Thus, each node of the BiG-SCAPE network could redirect the user to the actual gutSMASH page corresponding to that MGC, which contains information on the coordinates, genes and annotations, as well as the KnownClusterBlast and ClusterBlast comparative genomic analyses.

Given that the utility of the tools presented in this thesis directly depends on the quality of the input sequences, another important aspect to take into account is how new assembly algorithms could affect their performance. Next-generation sequencing has already yielded important insights in the identification and characterization of the human microbiome. However, limitations in tools to analyse these data sometimes affects downstream analyses that make use of these genomes. Reconstructing the genome from the sequenced reads can be done by using prior knowledge (reference genomes to map the reads onto) or without any prior scaffolding information (*de novo* assembly using overlapping sequencing reads)[213]. Depending on which approach is chosen, the output may substantially differ. Obtaining high quality genomes using *de novo* assembly is time-consuming and challenging due to (amongst others) short reads and repetitive regions[214]. Thus, using metagenome-assembled genomes (MAGs) implies being more cautious when making certain claims[215] including gutSMASH predictions. As reported in Chapter 4, bacterial strains have different metabolic potential due to small differences in the genomes. Hence, the more complete and accurately the genome is assembled, the better predictions gutSMASH can yield. While genome quality is important, it is worth noting that to assess MGC abundance and expression using BiG-MAP (Chapter 6), it is also crucial to

choose the right collection of genomes to map the reads. Whole genome sequencing data is the most realistic representation of microbial communities that live in a given ecosystem and may (or may not) include some of the available cultured isolates. Thus, depending on the taxonomic coverage of these metagenomes and on whether relevant complete genomes have been metabolically profiled with antiSMASH or gutSMASH, it will be more or less likely to find homologous gene clusters in the MGC collection. Hence, it is important to have some prior knowledge on the "omics" datasets to be analysed to choose the right genome collection (e.g., not only reference genomes but also MAGs sequenced from similar environments) to increase the chances of finding relevant associations with BiG-MAP.

7.2 Integration of the knowledge: combining data into databases

During the last decade, much attention has focused on the human microbiome[216], and this is still increasing at a vertiginous pace. A quick search (14th of March 2021) in PubMed for research articles published in 2020 using the "human microbiome" tag yields 16,882 results, while the same search for 2010 retrieves only 1,114 articles, an almost 15-fold increase. With such large numbers of studies piling up, staying up-to-date with the latest findings as a researcher becomes tedious. For this reason, integrative databases can be very useful. To this end, not only the results retrieved by gutSMASH could be stored in a searchable database as done for antiSMASH (Chapter 2), but also the genetic and biochemical evidence for the function of (new) MGCs as potential drivers of phenotypes could be linked. An example of the latter is the GWAS Catalog, which links all the genome-wide association studies of certain traits to better understand the mechanisms of the onset of diseases, pinpoints the causal variants and highlights potential therapeutic targets[217]. Similar to what MIBiG provides for secondary metabolic gene clusters[218], primary MGCs involved in specific traits could be used to create a new database with the aim of gathering all the published information in one place. The aim of such a database is to allow researchers to find

all information related to a specific query in one go; thus, powerful search engines would have to be used to relate and retrieve comprehensively all relevant (meta)data associated to a query. In addition to the MGC id, the database should provide information on its substrate and end product, the MGC class, the bacteria in which it was characterized, and whether any phenotypic data has been linked to it, all linked to the corresponding scientific evidence. Another useful addition is whether the MGC is predicted by gutSMASH or not. If so, there could be a link to the gutSMASH output and an additional column showing its taxonomic distribution, information that could be obtained from the ClusterBlast analysis, for instance. ClusterBlast is a very useful tool for comparative genomic analysis and to assess the degree of novelty a given predicted MGCs has. However, the overall gutSMASH running time depends on the number of comparisons between the predicted MGC and the number of entries in the ClusterBlast database. Thus, to not compromise the overall gutSMASH performance, the number of entries this database can contain will always be limited. For this reason, and given the number of newly sequenced genomes, a more complete version of the ClusterBlast database could be integrated in the gutSMASH database with the purpose of allowing users to acquire a more global overview of the MGC diversity in the human microbiome. For instance, Almeida *et al.* recently published a large genome catalogue, containing 204,938 bacterial and archaeal metagenome-assembled reference genomes from the human gut, that could be very interesting to analyse with gutSMASH. This run could easily yield hundreds of thousands of MGCs that could be compiled in a database for further exploration by grouping them in GCFs, similarly to what is implemented in the BiG-FAM database for secondary metabolic gene clusters. As highly similar MGCs are likely to be involved in the synthesis of very similar molecules (if not the same), this can help prioritize the characterization of putative MGCs and identify truly novel MGCs and/or molecules, as it has done in the natural products discovery field[219].

The list of potential MGCs (included in this database) not represented by

any detection rule in gutSMASH could be easily used for the inclusion of new rules in future gutSMASH versions. Moreover, it could be very useful to report the abundance and expression profiles of such MGCs in public metagenomes and metatranscriptomes from various body sites, if available, which could be obtained by running BiG-MAP. These results could provide more information on ecological or physiological functions of MGCs. In this way, users could have a complete overview of the characteristics of each MGC, including (lack of) evidence for their roles in health and disease. As at the moment there is no database that contains such information, the proposed one could easily become the new standard to annotate new specialized primary MGCs found in the human microbiome. Contributors from the scientific community could directly update the database by submitting relevant information on a new MGC, which could be incorporated later on as new entries after a round of manual curation.

Currently gutSMASH uses the ClusterBlast database, which contains 30,883 MGCs predicted from running gutSMASH on 4,220 high quality genomes (Chapter 5), to perform homology searches. This is a comprehensive database that at the moment is only accessible by downloading gutSMASH. Another functionality that could be enabled is to allow the user to query it using a gene or protein ID or sequence to see if it is part of any predicted MGC. With this purpose, a second database could be built, the gutSMASH database, that would include information on the gutSMASH predicted MGCs predicted from the 4,220 genomes. For incremental updates, to not compromise the speed of querying the database, a redundancy filtering based on genome sequence identity could be performed, similarly to the proposed method for antiSMASH database 2 (Chapter 2). However, since we have seen that specialized primary metabolism often shows strain-specific taxonomic distribution, additional information regarding whether a specific MGC is found in closely related strains or not could be included as well.

7.3 Towards integrative omics approaches

The ultimate goal of microbiome research is to better understand the role of the microbiota in human health and disease and to provide means to more easily prevent, identify and treat microbiome-related diseases[220]. Hence, it is crucial to focus on causation and seek to understand the genes and molecular mechanisms underpinning genes behind microbiome-driven phenotypes[46]. This is why nowadays integrative omics analyses are becoming more and more popular in functional genomic studies, given that they allow gaining biological insights at different molecular levels[221,222]. The results shown in Chapter 4 are good indicators that gutSMASH can not only be used to find known gene clusters, but also provides a good source of unexplored regions awaiting to be characterized. We have already highlighted a few interesting examples. However, given the high numbers of unannotated genes, thousands of cases could lead to the discovery of novel molecules that may be involved in yet undiscovered pathways with relevant implications for the human host. For this reason, BiG-MAP could be a very useful tool to prioritize the study of MGCs that may be associated with specific phenotypes and to assess their abundance and/or expression in a given microbial community. This could shed light on the functional characteristics of a given community, which traits/MGCs are more common among bacteria and ideally provide information on which of the traits are selected and/or more expressed when the environment is subjected to a given perturbation. The associations that BiG-MAP returns could then be validated by using heterologous expression in combination with building synthetic communities, or by for instance *in vivo* inoculation in mice. Depending on which approach and setup is chosen, one could get information on the minimum set of genes required for the production of a metabolite and on how the system responds to the addition or deletion of the genomic regions of interest in terms of the metabolites being produced[223,224]. More specifically, certain reactions can favor cooperation between microbes by producing molecules that serve as substrates for secondary fermenters; or the opposite,

drive a population to extinction due to physicochemical changes promoted by the presence of certain substrates that make the ecosystem no longer favorable to some community members. This could facilitate engineering bacteria to have specific functional profiles of interest (e.g., bacteria with specific MGCs identified by gutSMASH and selected by their abundance/expression patterns). Then, by using synthetic communities, one could test their behaviour in more real and complex environments as found in nature. The next logical step to scale up in complexity then would be to evaluate the direct impact of inoculating these microbial communities in germ-free mice to study the potential microbe-host interactions from a mechanistic point of view. Ultimately, this could open up the door for microbiome-derived treatments by delivering defined communities of naturally occurring bacteria or engineered bacteria that provide the host with mechanisms to alleviate or revert (some of) the symptomatology. Different studies have already proven the efficacy of administering engineered living microorganisms to reduce the size of solid tumors[225] and other diseases such as phenylketonuria[226]. Patients with phenylketonuria are unable to convert phenylalanine to tyrosine due to a mutation in the phenylalanine hydroxylase coding gene. In mice it has been proven that the oral administration of an engineered *E. coli* able to transform phenylalanine to phenylpyruvic acid (that includes the phenylalanine ammonia lyase and L-amino acid deaminase coding genes) reduces the concentration of phenylalanine in blood. At the moment, there are several engineered bacterial therapeutics in clinical development, including an *E. coli* specifically designed to treat phenylketonuria. Despite the fact that none of these has been approved yet, they provide a promising avenue to cure diseases[227]. As shown in Chapter 4 (Figure 4.2A), phenylalanine and histidine ammonia lyase are only found in Bacteroides and Parabacteroides. Thus, this particular case evinces the importance of identifying, annotating and characterizing such coding genes and MGCs from the human microbiome that can open up new opportunities to genetically modify bacteria to influence the host biology *in situ*.

Genomic content does not seem to be strongly correlated with metabolite concentrations, as shown in Chapter 4, because of the complexity of regulation of gene expression. For this reason, metatranscriptomic and metabolomic data could play a crucial role in mechanistic studies in the coming years. Metratranscriptomics can reveal very useful information not included in any other type of data. For instance, evaluating the RNA-seq gene expression profile can help understand which bacteria are active and how they adapt to specific fluctuations in the ecosystem. However, it is important to consider certain limitations associated with this technology. As mRNA is not very stable, evaluating expression shifts that occur in a rather fast pace can be hard to capture[228]. Moreover, detecting a transcript does not always imply the presence of the associated protein[229] due to different phenomena such as post-translational modifications, miRNA-guided degradation, etc. On the other hand, metabolomics allows directly grasping the diversity and abundance of molecules released in specific environments. Hence, integrative approaches including the combination of metatranscriptomic and metabolomics data can elucidate the implications for the host of the expression and circulation of the metabolites synthesized from differentially expressed pathways. For instance, one logical approach to follow could be to use paired omics datasets that include metagenomic, metatranscriptomic and metabolomic data. From the metagenomic data, specialized primary MGCs could be predicted using gutSMASH or, alternatively, a comprehensive set of reference genomes could be used as input. The resulting MGC collection could then be used as input for BiG-MAP together with the RNA-seq reads that can help identifying differentially expressed MGCs in the conditions tested. Finally, the RPKM values of the differentially expressed gene clusters (if any) could be correlated to the metabolite concentrations, in order to either validate the findings or to identify potential links between putative gene clusters and their molecules. Correlating the expression values of known gene clusters with the corresponding metabolite concentration could serve as an extra validation step that the expression of the pathway indeed results in the synthesis of the corresponding molecule. However, a lack of correlation does not necessarily indicate that the pathway does not produce

the molecule, as this may be caused by imperfect measurement of metabolite or transcript levels[230]. Alternatively, the correlation between expression levels of orphan gene clusters (i.e. lacking an associated metabolite) and metabolite concentrations can help formulate hypotheses on which MGCs are responsible for the production of which molecules. These hypotheses would then have to be tested experimentally, which could ultimately aid in the elucidation of novel pathways and discovery of new compounds. Overall, following this procedure we would gain insights into the community structure, its metabolic potential and its dynamics.

7.4 Closing remarks

Many human microbiome researchers have as ultimate goal to obtain a better understanding of these microbial communities in order to improve human health. For this reason, it is important to discern between mere associations (between bacteria and disease) and causation (the actual mechanisms behind microbiome-derived host phenotypes). Certainly, there is a need to uncover the metabolic potential of human gut bacteria, that can serve as a proxy to understand the underlying molecular mechanisms triggering specific traits. As shown in this thesis, computational genomics is a very promising resource to generate new knowledge towards the functional role microbes can exert. To overcome some of the limitations, though, it is important to survey and combine different types of data that allow to comprehensively analyse complex biological systems. Still, to understand our microbiome, joint efforts around the globe from different disciplines are required, not only to generate new hypotheses but also to validate them, involving both dry-lab and wet-lab expertise. Understanding the whole ecosystem implies looking beyond bacteria, as viruses, fungi and archaea also play relevant roles in the community structure and dynamics. In summary, with the amount of attention the human microbiome is receiving and the fast pace at which technology is being developed I foresee very exciting breakthroughs in the coming years that will pave the way to using the human microbiome in personalized medicine.

References

1. Gest, H. The discovery of microorganisms by Robert Hooke and Antoni van Leeuwenhoek, Fellows of The Royal Society. R. Soc. 58, 187–201 (2004).

2. Pelczar, Rita M. & Pelczar, M. J. Microbiology. Encycl. Br. 1–21 (2020).

3. Merien, F. A Journey with elie Metchnikoff: From innate cell mechanisms in infectious diseases to quantum biology. Front. public Heal. 4, 1–5 (2016).

4. Tan, S. Y. & Tatsumura, Y. Alexander Fleming (1881–1955): Discoverer of penicillin. Singapore Med. J 56, 366–367 (2015).

5. Davidson, A. L. et al. Structure, function, and evolution of bacterial ATP-binding cassette systems. Microbiol. Mol. Biol. Rev. 72, 317–64 (2008).

6. Pariente, N. A field is born. Nat. milestones, 3–4 (2019).

7. Garza, D. R. & Dutilh, B. E. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. Cell. Mol. Life Sci. 72, 4287–4308 (2015).

8. Gauthier, J. et al. A brief history of bioinformatics. Brief. Bioinform. 20, 1981–1996 (2018).

9. The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 486, 207–214 (2012).

10. Li, Y. & Chen, L. Big biological data: Challenges and opportunities. Genomics Proteomics Bioinformatics 12, 187–189 (2014).

11. Costea, P.I. et al. Enterotypes in the landscape of gut microbial community composition. Nat. Microbiol. 3, 8–16 (2018).

12. Ziesemer, K. et al. Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. Sci. Rep. 5, 1–20 (2015).

13. Gilbert, J. et al. Current understanding of the human microbiome. Nat. Med 24, 392–400 (2018).

14. Rowland I. et al. Gut microbiota functions: metabolism of nutrients and other food components. Eur. J. Nutr. 57, 1–24 (2018).

15. Davenport, E.R. et al. The human microbiome in evolution. BMC Biol. 15, 1–12 (2017).

16. Ercolini, D. & Fogliano, V. Food design to feed the human gut microbiota. J. Agric. food Chem. 66, 3754–3758 (2018).

17. Shahid, M. et al. Interaction between diet composition and gut microbiota and its impact on gastrointestinal tract health. Food Sci. Hum. Wellness 6, 121–130 (2017).

18. Seo, Y. S. et al. Dietary carbohydrate constituents related to gut dysbiosis and health. Microorganisms 6, 1–11 (2020).

19. Martinez, K. B. et al. Microbial metabolites in health and disease: Navigating the unknown in search of function. J Biol Chem. 292, 8553–8559 (2017).

20. Durack, J. & Lynch, S. V. The gut microbiome: Relationships with disease and opportunities for therapy. J. Exp. Med. 216, 20–40 (2018).

21. Gagliardi, A. et al. Rebuilding the gut microbiota ecosystem. Int. J. Environ. Res. Public Health 15, 1679 (2018).

22. Kim, K. O. & Gluck, M. Fecal microbiota transplantation: An update on clinical practice. Clin. Endosc. 52, 137–143 (2019).

23. Quigley, E. M. M. & Gajula, P. Recent advances in modulating the microbiome. F1000Research 9, 1–12 (2020).

24. Schloss, P. D. et al. Introducing Mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537–7541 (2009).

25. Caporaso, J. G. et al. QIIME allows analysis of high- throughput community sequencing data. Nature 7, 335–336 (2010).

26. Quast, C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41, 590–596 (2013).

27. Wattam, A. R. et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res 45, 535–542 (2017).

28. Desantis, T. Z. et al. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl. Environ. Microbiol. 72, 5069–5072 (2006).

29. Pruesse, E. et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res 35, 7188–7196 (2007).

30. Wattam, A. R. et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 42, 581–591 (2014).

31. Hamady, M. & Knight, R. Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. Genome Res. 19, 1141–1152 (2009).

32. Kuczynski, J. et al. Experimental and analytical tools for studying the human microbiome. Nat. Rev. Genet. 13, 47–58 (2012).

33. Huse, S. M. et al. VAMPS: a website for visualization and analysis of microbial population structures. BMC Bioinformatics 15, 1–6 (2014).

34. Langille, M. G. I. Exploring linkages between taxonomic and functional profiles of the human microbiome. mSystems 3, 1–4 (2018).

35. Pascal, V. et al. A microbial signature for Crohn's disease. Gut 66, 1–10 (2017).

36. Halfvarson, J. et al. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat. Microbiol. 2, 1–7 (2017).

37. Turnbaugh, P. J. et al. A core gut microbiome in obese and lean twins. Nature 457, 480–485 (2009).

38. Turnbaugh, P. J. et al. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444, 1027–1031 (2006).

39. Kang, D. et al. Reduced incidence of prevotella and other fermenters in intestinal microflora of autistic children. PLoS One 8, e68322 (2013).

40. Kushak, R. I. et al. Analysis of the duodenal microbiome in autistic individuals: Association with carbohydrate digestion. Gastroenterology 64, 110–116 (2017).

41. Hopfner, F. et al. Gut microbiota in Parkinson disease in a northern German cohort. Brain Res. 1667, 41–45 (2017).

42. Pietrucci, D. et al. Parkinsonism and related disorders dysbiosis of gut microbiota in a selected population of Parkinson' s patients. Park. Relat. Disord. 65, 124–130 (2019).

43. Roesch, L. F. W. et al. Culture-independent identification of gut bacteria correlated withthe onset of diabetes in a rat model. ISME.J. 3, 536–548 (2010).

44. Lozupone, C. A. et al. Diversity, stability and resilience of the human gut microbiota. Nature 489, 220–230 (2012).

45. Heintz-buschart, A. & Wilmes, P. Human gut microbiome: Function matters. Trends Microbiol. 7, 563–574 (2017).

46. Fischbach, M. A. Microbiome: Focus on causation and mechanism. Cell 174, 785–790 (2018).

47. Keller, N. P. Translating biosynthetic gene clusters into fungal armor and weaponry. Nat. Chem. Biol. 11, 671–677 (2015).

48. Signaling, I. & Kumar, A. Indole signaling at the host-microbiota-pathogen interface. MBio 10, e01031-19 (2019).

49. Tang, W. H. W. & Hazen, S. L. The contributory role of gut microbiota in cardiovascular disease. J. Clin. Investigtion 124, 4204–4211 (2014)

50. Sugimoto, Y. et al. A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. Science 366, 1–17 (2019).

51. Donia, M. S. et al. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell 158, 1402–1414 (2014).

52. Donia, M. S. & Fischbach, M. A. Small molecules from the human microbiota. Science 349, 395–406 (2015).

53. Louis, P. et al. Organization of butyrate synthetic genes in human colonic bacteria: phylogenetic conservation and horizontal gene transfer. FEMS Microbiol. Lett. 183, 240–247 (2007).

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

55. Segata, N. On the road to strain-resolved comparative metagenomics. mSystems 3, 1–6 (2018).

56. Rossum, T. Van. et al. Diversity within species: interpreting strains in microbiomes. Nat. Rev. Microbiol. 18, 491–505 (2020).

57. Wiedemann, B. et al. Impact of gyrA and parC mutations on quinolone resistance, doubling time, and supercoiling degree of Escherichia coli. Antimicrob. Agents Chemother. 43, 868–875 (1999).

58. Woodford, N. & Ellington, M. J. The emergence of antibiotic resistance by mutation. Clin. Microbiol. Infect. 13, 5–18 (2006).

59. Atakeyama, B. M. H. Structure and function of Helicobacter pylori CagA, the first idntified bacterial protein involved in human cancer. Proc. Jpn. Acad. Ser. B Phys. Biol. Sci. 93, 196–219 (2017).

60. Blaser, M. J. et al. Infection with Helicobacter pylori strains possessing cagA is associated with an increased risk of developing adenocarcinoma of the stomach. Cancer Res. 55, 2111–2116 (1995).

61. Marteau, P. Evidence of Probiotic Strain Specificity Makes Extrapolation of Results Impossible From a Strain to Another, Even From the Same Species. Ann. Gastroentol. Hepatol. 000, 0–2 (2011).

62. Quince, C. et al. DESMAN: a new tool for de novo extraction of strains from metagenomes. Genome Biol. 18, 1–22 (2017).

63. Costea, P. I. et al. metaSNV: A tool for metagenomic strain level analysis. PLoS One 12, 1–9 (2017).

64. Luo C. et al. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotech 33, 1045–1052 (2016).

65. Costea, P. I. et al. Subspecies in the global human gut microbiome. Mol. Syst. Biol. 13, 1–11 (2017).

66. Zeevi, D. et al. Structural variation in the gut microbiome associates with host health. Nature 568, 43–48 (2019).

67. Luo, Y. et al. Recent advances in natural product discovery. Curr. Opin. Biotechnol. 30, 230–237 (2014).

68. Khaldi, N. et al. SMURF: genomic mapping of fungal secondary metabolite clusters. Fungal Genet Biol. 47, 736–741 (2011).

69. Li, M. H. et al. Automated genome mining for natural products. BMC Bioinformatics 10, 1–10 (2009).

70. Starcevic, A. et al. ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucleic Acids Res. 36, 6882–6892 (2008).

71. Weber, T. et al. CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. J. Biotechnol. 140, 13–17 (2009).

72. Skinnider, M. A. et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Res. 43, 9645–62 (2015).

73. Medema, M. H. et al. AntiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res. 39, 339–346 (2011).

74. Cimermancic, P. et al. Insights into secondary metabolism from a global analysis ofprokaryotic biosynthetic gene clusters. Cell 158, 412–421 (2015).

75. Meleshko, D. et al. BiosyntheticSPAdes: Reconstructing biosynthetic gene clusters from assembly graphs. Genome Res. 29, 1352–1362 (2019).

76. Price, M. N. et al. A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res. 33, 880–892 (2005).

77. Taboada, B. et al. High accuracy operon prediction method based on STRING database scores. Nucleic Acids Res. 38, e130 (2010).

78. Kim, J. et al. FMAP: Functional Mapping and Analysis Pipeline for metagenomics and metatranscriptomics studies. BMC Bioinformatics 17, 1–8 (2016).

79. Abubucker, S. et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput. Biol. 8, e1002358 (2012).

80. Manor, O. & Borenstein, E. Systematic characterization and analysis of the taxonomic drivers of functional shifts in the human resource. Cell Host Microbe 21, 254–267 (2017).

81. Kanehisa, M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40, 109–114 (2012).

82. Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs from 1981 to 2014. J. Nat. Prod. 79, 629–66 (2016).

83. Ziemert, N. et al. The evolution of genome mining in microbes – a review. Nat. Prod. Rep. 33, 988–1005 (2016).

84. Medema, M. H. & Fischbach, M. A. Computational approaches to natural product discovery. Nat Chem. Biol. 11, 639–648 (2016).

85. Weber, T. & Uk, H. The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. Synth. Syst. Biotechnol. 1, 69–79 (2016).

86. Weber, T. In silico tools for the analysis of antibiotic biosynthetic pathways. Int. J. Med. Microbiol. 304, 230–235 (2014).

87. Blin, K. et al. AntiSMASH 2.0 - A versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 41, W204–W211 (2013).

88. Weber, T. et al. AntiSMASH 3.0 - A comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. 43, W237–W243 (2015).

89. Blin, K. et al. AntiSMASH 4.0 - Improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 45, W36–W41 (2017).

90. Medema, M. et al.. Minimum information about a biosynthetic gene cluster. Nat. Chem. Biol 11, 625–631 (2015).

91. Blin, K. et al. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. Nucleic Acids Res. 45, D555–D559 (2017).

92. Chevrette, M. G. et al. Sequence analysis SANDPUMA: Ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. Bioinformatics 33, 3202–3210 (2017).

93. Leary, N. A. O. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733–D745 (2016).

94. Tange O. GNU Parallel: The command-line power tool. USENIX Mag. 3, 42–47 (2011).

95. Tietz, J. I. et al. A new genome-mining tool redefines the lasso peptide biosynthetic landscape. Nat. Chem. Biol. 13, 470–478 (2017).

96. Parks, D. H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. 36, (2018).

97. Huerta-Cepas, J. et al. ETE: A python Environment for Tree Exploration. BMC Bioinformatics 11, 1–7 (2010).

98. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44, W242–W245 (2016).

99. Koppel, N. & Balskus, E. P. Exploring and understanding the biochemical diversity of the human microbiota. Cell Chem. Biol. 23, 18–30 (2016).

100. Rooks, M. G. & Garrett, W. S. Gut microbiota, metabolites and host immunity. Nat. Publ. Gr. 16, 341–352 (2016).

101. Koh, A. et al. From dietary fiber to host physiology: Short-chain fatty acids as key bacterial metabolites. Cell 165, 1332–1345 (2016).

102. Rath, S. et al. Potential TMA-producing bacteria are ubiquitously found in mammalia. Front. Microbiol. 10, 1–10 (2020).

103. Wahlström, A. et al. Intestinal crosstalk between bile acids and microbiota and its impact on host metabolism. Cell Metab. 24, 41–50 (2016).

104. Ridlon, J. M. et al. Bile salt biotransformations by human intestinal bacteria. J. Lipid Res. 47, 241–259 (2006).

105. Buffie, C. G. et al. Precision microbiome restoration of bile acid-mediated resistance to Clostridium difficile. Nature 517, 205–208 (2015).

106. Bayerdörffer E. et al. Increased serum deoxycholic acid levels in men with colorectal adenomas. Gastroenterology 104, 145–151 (1993).

107. Bernstein, H. et al. Bile acids as carcinogens in human gastrointestinal cancers. Mutat. Res. 589, 47–65 (2005).

108. Berr F. et al. 7-alpha-Dehydroxylating Bacteria Enhance Deoxycholic Acid Input and Cholesterol Saturation of Bile in Patients With Gallstones. Gastroenterology 111, 1611–1620 (1996).

109. Ridlon, J. M., Kang, D. & Hylemon, P. B. Isolation and characterization of a bile acid Iinducible 7α-dehydroxylating operon in Clostridium hylemonae TN271. Anaerobe 16, 137–146 (2018).

110. Funabashi, M. et al. A metabolic pathway for bile acid dehydroxylation by the gut microbiome. Nature 582, 566–570 (2020).

111. Dickert, S. et al. The involvement of coenzyme A esters in the dehydration of (R)-phenyllactate to (E)-cinnamate by Clostridium sporogenes. Eur J Biochem. 3884, 3874–3884 (2000).

112. Hubbard, P. A. et al. The crystal structure and reaction mechanism of Escherichia coli 2,4-Dienoyl-CoA reductase. J Biol Chem. 278, 37553–37560 (2003)

113. Trickey P. et al. Structural and biochemical characterization of recombinant wild type and a C30A mutant of trimethylamine dehydrogenase from Methylophilus. Biochemistry 39, 7678–7688 (2000).

114. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. 35, 1–3 (2017).

115. Tett, A. et al. Unexplored diversity and strain-level structure of the skin microbiome associated with psoriasis. npj Biofilms Microbiomes 3, 1–11 (2017).

116. Askar, M. et al. Epidemic profile of Shiga-toxin–producing Escherichia coli O104:H4 Outbreak in Germany. N Engl J Med. 365, 1771-1780 (2011).

117. Almeida, A. et al. A new genomic blueprint of the human gut microbiota. Nature 568, 499–504 (2019).

118. Morton, J. T. et al. Learning representations of microbe–metabolite interactions. Nat. Methods 16, 1306–1314 (2019).

119. Mallick, H. et al. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nat. Commun. 10, 1–11 (2019).

120. Price, M. N. et al. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490 (2010).

121. Revell, L. J. Phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 217–223 (2012).

122. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47, 256–259 (2019).

123. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539–539 (2014).

124. Navarro-Muñoz, J. C. et al. A computational framework to explore large-scale biosynthetic diversity. Nat. Chem. Biol. 16, 60–68 (2019).

125. Blin, K. et al. AntiSMASH 5.0: Updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 47, 81–87 (2019).

126. Karp, P. D. et al. The BioCyc collection of microbial genomes and metabolic pathways. 20, 1085–1093 (2019).

127. Kitamoto, S. et al. Dietary l-serine confers a competitive fitness advantage to Enterobacteriaceae in the inflamed gut. Nat. Microbiol. 5, 116–125 (2020).

128. Zou, Y. et al. 1,520 reference genomes from cultivated human gut bacteria enable functional microbiome analyses. Nat. Biotechnol. 37, 179–187 (2019).

129. Lloyd-price, J. et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature 550, 61–66 (2017).

130. Tracy, B. P. et al. Clostridia: The importance of their exceptional substrate and metabolite diversity for biofuel and biorefinery applications. Curr. Opin. Biotechnol. 23, 364–381 (2012).

131. Maguire, F. et al. Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic Islands. Microb. genomics 6, 1–12 (2020).

132. Pomare, E. W. et al. Short chain fatty acids in human large intestine, portal, hepatic and venous blood. Gut 10, 1221–1227 (1987).

133. Jones, S. A. et al. Anaerobic respiration of Escherichia coli in the mouse intestine. Infect. Immun. 79, 4218–4226 (2011).

134. Andreu, V. P. et al. BiG-MAP: An automated pipeline to profile metabolic gene cluster abundance and expression in microbiomes. bioRxiv (2020).

135. Tigchelaar, E. F. et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. BMJ Open 5, e006772 (2015).

136. Faber, F. et al. Respiration of Microbiota-Derived 1,2-propanediol Drives Salmonella Expansion during Colitis. PLoS Pathog. 3, e1006129 (2017).

137. Andriamihaja, M. et al. The deleterious metabolic and genotoxic effects of the bacterial metabolite p-cresol on colonic epithelial cells. Free Radic. Biol. Med. 85, 219–227 (2015).

138. Gavin, D. M. et al. PICRUSt2 for prediction of metagenome functions. Nat. Biotech. 38, 669–688 (2013).

139. Karp, P. D. et al. Pathway Tools version 23.0 update: Software for pathway/genome informatics and systems biology. Briefings Bioinforma. 22, 109–126 (2021).

140. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes - a 2019 update. Nucleic Acids Res 48, 445–453 (2020).

141. Franzosa, E. A. et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat. Methods 15, 962–968 (2018).

142. Andreu, V. P. et al. Computational genomic discovery of diverse gene clusters harbouring Fe-S flavoenzymes in anaerobic gut microbiota. Microb. genomics 6, 1–8 (2020).

143. Chen C. et al. Representative proteomes: A stable, scalable and unbiased proteome set for sequence analysis and functional annotation. PLoS One 6, e18910 (2011).

144. Price, M. N. & Arkin, A. P. PaperBLAST: Text mining papers for information about homologs. mSystems 2, 1–10 (2017).

145. Petit, E. et al. Involvement of a bacterial microcompartment in the metabolism of fucose and Rhamnose by Clostridium phytofermentans. PLoS One 8, 1–12 (2013).

146. Zhernakova, A. et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 352, 565–569 (2016).

147. Waterhouse, A. M. et al. Jalview version 2 - A multiple sequence alignment editor and analysis workbench. 25, 1189–1191 (2009).

148. Camacho, C. et al. BLAST+: Architecture and applications. BMC Bioinformatics 10, 1–9 (2009).

149. Santos, E. L. C. D. L. & Challis, G. L. clusterTools: Proximity searches for functional elements to identify putative biosynthetic gene clusters. bioRxiv (2017).

150. Medema, M. H. et al. Detecting sequence homology at the gene cluster level with multigeneblast. Mol. Biol. Evol. 30, 1218–1223 (2013).

151. Hyatt, D. et al. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 1–11 (2010).

152. Shannon, P. et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2498–2504 (2003).

153. Ondov, B. D. et al. Mash: Fast genome and metagenome distance estimation using MinHash. Genome Biol. 17, 1–14 (2016).

154. Phanstiel, D. H. et al. Sushi.R: Flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. Bioinformatics 30, 2808–2810 (2014).

155. Garsin, D. A. Ethanolamine utilization in bacterial pathogens: Roles and regulation. Nat. Rev. Microbiol 8, 290–295 (2010).

156. Wong, J. M. W., Souza, R. De, Kendall, C. W. C., Emam, A. & Jenkins, D. J. A. Colonic health: Fermentation and short chain fatty acids. J. Clin. Gastroenterol. 40, 235–243 (2006).

157. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30 (2000).

158. Andreu, V. P. et al. A systematic analysis of metabolic pathways in the human gut microbiota. bioRxiv (2021).

159. Liu, B. & Pop, M. MetaPath: Identifying differentially abundant metabolic pathways in metagenomic datasets. BMC Proc. 5, 1–12 (2011).

160. Martinez, X. et al. MetaTrans: An open-source pipeline for metatranscriptomics. Sci. Rep. 6, 1–12 (2016).

161. Buchfink, B. et al. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–63 (2015).

162. Dodd, D. et al. A gut bacterial pathway metabolizes aromatic amino acids into nine circulating metabolites. Nat. Publ. Gr. 551, 648–652 (2017).

163. Yeates, T. O. et al. The protein shells of bacterial microcompartment organelles. Curr. Opin. Struct. Biol. 21, 223–231 (2011).

164. Kautsar, S. A. et al. PlantiSMASH: Automated identification, annotation and expression analysis of plant biosynthetic gene clusters. Nucleic Acids Res. 45, W55-W63 (2017).

165. Ooka, T. et al. Defining the genome features of Escherichia albertii, an emerging enteropathogen closely related to Escherichia coli. Genome Biol. Evol. 7, 3170–3179 (2015).

166. Eichler, K. & Bourgls, F. Molecular characterization of the cai operon necessary for carnitine metabolism in Escherichia coli. Mol. Microbiol. 13, 775–786 (1994).

167. Engels, C. et al. The common gut microbe Eubacterium hallii also contributes to intestinal propionate formation. Front. Microbiol. 7, 1–12 (2016).

168. Heßlinger, C. & Fairhurst, S. A. Novel keto acid formate-lyase and propionate kinase enzymes are components of an anaerobic pathway in Escherichia coli that degrades L-threonine to propionate. Molecular microbiology 27, 477–492 (1998).

169. Henke, M. T. et al. Ruminococcus gnavus, a member of the human gut microbiome associated with Crohn 's disease, produces an inflammatory polysaccharide. PNAS 116, 12672–12677 (2019).

170. Weiss, T. S. et al. Intracellular polyamine levels of intestinal epithelial cells in Inflammatory Bowel Disease. Inflamm. Bowel Dis. 10, 529–535 (2004).

171. Blin, K. et al. The antiSMASH database version 3: Increased taxonomic coverage and new query features for modular enzymes. Nucleic Acids Res. 49, 639–643 (2021).

172. Berendsen, R. L. et al. The rhizosphere microbiome and plant health. Trends Plant Sci. 17, 478–486 (2012).

173. Hibbing, M. E. et al. Bacterial competition: Surviving and thriving in the microbial jungle. Nat. Rev. Microbiol 8, 15–25 (2010).

174. Pickard, J. M. et al. Gut microbiota: Role in pathogen colonization, immune responses and inflammatory disease. Immunol. Rev. 279, 70–89 (2017).

175. Tracanna, V. et al. Mining prokaryotes for antimicrobial compounds: From diversity to function. FEMS Microbiol. Rev. 41, 417–429 (2017).

176. Mendes, R. et al. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science 911, 1–5 (2011).

177. Brial, F. et al. Implication of gut microbiota metabolites in cardiovascular and metabolic diseases. Cell. Mol. Life Sci. 75, 3977–3990 (2018).

178. Hannigan, G. D. et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res. 47, 110–123 (2019).

179. Caspi, R. et al. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res 46, 633–639 (2018).

180. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–360 (2012).

181. Paulson, J. N. et al. Differential abundance analysis for microbial marker-gene surveys. Nat. Methods 10, 1–6 (2013).

182. Mcdonald, D. et al. The Biological Observation Matrix (BIOM) format or: How I learned to stop worrying and love the ome-ome. Gigascience 1, 1–6 (2012).

183. Deo, P. N. & Deshmukh, R. Oral microbiome: Unveiling the fundamentals. J. Oral Maxillofac. Pathol. 23, 122–128 (2019).

184. Graves, D. T. et al. The oral microbiota is modified by systemic diseases. J. Dent. Res. 98, 148 –156 (2019).

185. Dewhirst, F. E. et al. The human oral microbiome. J. Bacteriol. 192, 5002–5017 (2010).

186. Garcia, S. S. et al. Targeting of Streptococcus mutans biofilms by a novel small molecule prevents dental caries and preserves the oral microbiome. Dent. Res. 96, 807–814 (2017).

187. Edlund, A. et al. Metabolic fingerprints from the human oral microbiome reveal a vastknowledge gap of secreted small peptidic molecules. mSystems 2, 1–16 (2017).

188. Aleti, G. et al. Identification of the bacterial biosynthetic gene clusters of the oral microbiome illuminates the unexplored social language of bacteria during health and disease. MBio 10, 1–19 (2019).

189. Tang, X. et al. Cariogenic Streptococcus mutans produces tetramic acid strain-specific antibiotics that impair commensal colonization. ACS Infect. Dis. 6, 563–571 (2020).

190. Mcluskey, K. et al. Decomposing metabolite set activity levels with PALS. bioRxiv (2020).

191. Wang, M. Perspective sharing and community curation of mass spectrometry data with global natural products social molecular networking. Nat. Biotechnol. 34, 828–837 (2016).

192. Tomfohr, J. et al. Pathway level analysis of gene expression using singular value decomposition. BMC Bioinformatics 11, 1–11 (2005).

193. Wang, M. et al. Mass spectrometry searches using MASST. Nat. Biotechnol. 38, 23–26 (2020).

194. Engels, C. et al. Acrolein contributes strongly to antimicrobial and heterocyclic amine transformation activities of reuterin. Mol. Nutr. Food Res. 6, 1–13 (2016).

195. Orita, H. M. et al. Comparative genome analysis of Lactobacillus reuteri and Lactobacillus fermentum reveal a genomic island for reuterin and cobalamin production. DNA Res. 15, 151–161 (2008).

196. Stevens, J. F. & Maier, C. S. Acrolein: Sources, metabolism, and biomolecular interactions relevant to human health and disease. Mol Nutr Food Res. 52, 7–25 (2008).

197. Zhang, J. et al. Gut Microbial Glycerol Metabolism as an Endogenous Acrolein. MBio 9, 1–6 (2018).

198. Shlomo M. & Erika W. Acrolein polymerization: Monodisperse, homo, and hybrido microspheres, synthesis, mechanism, and reactions. J. Polym. Sci. Polym. Chem. Ed. 22, 145–158 (1984).

199. Nikawa H. et al. Lactobacillus reuteri in bovine milk fermented decreases the oral carriage of mutans streptococci. Int. J. Food Microbiol. 95, 219–223 (2004).

200. Kang M.S. et al. Inhibitory effect of Lactobacillus reuteri on periodontopathic and cariogenic bacteria. J. Microbiol. 49, 193–199 (2011).

201. Liu, L. et al. Genome mining unveils widespread natural product biosynthetic capacity in human oral microbe Streptococcus mutans. Sci. Rep. 6, 1–10 (2016).

202. Schirmer, M. et al. Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. Nat. Microbiol. 3, 337–346 (2018).

203. Rocha D.J. et al. Bacterial reference genes for gene expression studies by RT-qPCR: survey and analysis. Antonie Van Leeuwenhoek 108, 685–693 (2015).

204. Li H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079 (2009).

205. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010).

206. Angly, F. E. et al. Grinder: A versatile amplicon and shotgun sequence simulator. Nucleic Acids Res. 40, 1–8 (2012).

207. Gu, Z. et al. Genome analysis Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32, 2847–2849 (2016).

208. Nothias, L.F. et al. Feature-based molecular networking in the GNPS analysis environment. Nat. Methods 17, 905–908 (2020).

209. Patiny, L. & Borel, A. ChemCalc: A building block for tomorrow's chemical infrastructure. J. Chem. Inf. Model. 53, 1223–1228 (2012).

210. Sunghwan K. et al. PubChem 2019 update: Improved access to chemical data. Nucleic Acids Res. 47, 1102–1109 (2019).

211. Hao T. et al. An anaerobic bacterium host system for heterologous expression of natural product biosynthetic gene clusters. Nat. Commun. 10, 1–13 (2019).

212. Edlund A. et al. Meta-omics uncover temporal regulation of pathways across oral microbiome genera during in vitro sugar metabolism. ISME J. 9, 2605–2619 (2015).

213. Dark, M. J. Whole-genome sequencing in bacteriology: State of the art. Infect. Drug Resist. 3, 115–123 (2013).

214. Salzberg, S. L. et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 22, 557–567 (2012).

215. Chen, L. et al. Accurate and complete genomes from metagenomes. Genome Res. 30, 315–333 (2020).

216. NIH human Microbiome Portfolio Analysis Team. A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016. Microbiome 7, 1–19 (2019).

217. Buniello, A. et al. The NHGRI-EBI GWAS: Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 47, 1005–1012 (2019).

218. Kautsar, S. A. et al. MIBiG 2.0: A repository for biosynthetic gene clusters. Nucleic Acids Res 48, 454–458 (2020).

219. Kautsar, S. A. et al. BiG-FAM: The biosynthetic gene cluster families database. Nucleic Acids Res 49, 490–497 (2021).

220. Zhang, X. et al. Advancing functional and translational microbiome research using meta-omics approaches. Microbiome 7, 1–12 (2019).

221. Subramanian, I. et al. Multi-omics data integration, interpretation, and its application. Bioinform. Biol. Insights 14, 1–24 (2020).

222. Yan, J. et al. Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. Brief. Bioinform. 19, 1370–1381 (2018).

223. Dantas, G. et al. Experimental approaches for defining functional roles of microbes in the human gut. Annu. Rev. Microbiol. 67, 459–475 (2013).

224. Ivanov, I. I. et al. Induction of intestinal Th17 cells by segmented filamentousbacteria. Cell 139, 485–498 (2009).

225. Chowdhury, S. et al. Programmable bacteria induce durable tumor regression and systemic antitumor immunity. Nat. Med. 25, 1057–1063 (2019).

226. Isabella, V. M. et al. Development of a synthetic live bacterial therapeutic for the human metabolic disease phenylketonuria. Nat. Biotechnol. 36, 857–864 (2018).

227. Charbonneau, M. R. et al. Developing a new class of engineered live bacterial therapeutics to treat human diseases. Nat. Commun. 11, 1–11 (2020).

228. Bikel, S. et al. Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: Towards a systems-level understanding of human microbiome. CSBJ 13, 390–401 (2015).

229. Bashiardes, S. et al. Use of metatranscriptomics in microbiome research. Bioinform. Biol. Insights 10, 19–25 (2016).

230. Narayanasamy, S. et al. Integrated omics for the identification of key functionalities in biological wastewater treatment microbial communities. Microb. Biotechnol. 8, 363–368 (2015).

Summary

The human body is colonized by trillions of microorganisms including bacteria, fungi and viruses that inhabit our cavities and surfaces, such as the skin, the oral cavity, the respiratory and the gastrointestinal tract. Each body site is characterized by specific microbial communities, also referred to as microbiota, that altogether weigh around 2 kg and can be regarded as another organ. The microbiota include a diverse range of microbes, predominantly bacteria, which help us to digest food, synthesize vitamins, develop the immune system and prevent pathogen infections, among other functions. However, imbalances in these microbial communities have also been linked to some diseases, such as colon cancer, inflammatory bowel disease, cardiovascular disease, obesity and diabetes. Therefore, studying the composition and function of the human microbiota can help understanding these diseases better from a molecular point of view.

Bacteria outnumber human cells by a ratio of 10 to 1, and the bacterial gene repertoire vastly exceeds the human one. The genomic diversity of the microbiota enables them to respond to different stimuli by excreting small molecules that play a relevant role in microbe-microbe and microbe-host interactions. Most of the enzymatic pathways producing these molecules are encoded in Metabolic Gene Clusters (MGCs) and belong to specialized primary metabolism. Ultimately, the circulation of these molecules can result in different phenotypes, beneficial or detrimental for the host. In order to better understand how microbes influence health, it is therefore important to profile their metabolic potential and elucidate the molecular mechanisms behind these phenotypes.

Several bioinformatic tools have been designed to predict the metabolic potential of these bacteria, such as antiSMASH. This is the gold-standard tool to predict and analyse Biosynthetic Gene Clusters (BGCs) that encode secondary metabolites. Running antiSMASH on a large collection of genomes generates invaluable information, but it is difficult to fully exploit using current bioinformatic tools. For this reason, a method to store and

organise the massive amounts of data is presented in **Chapter 2**: the antiSMASH database version 2, a platform that allows the user to interactively perform cross-genome searches from pre-computed antiSMASH runs. Besides antiSMASH, there are other tools to predict microbial metabolism that rather focus on more generic primary metabolism. As no tool existed to predict specialized primary MGCs, in **Chapter 3** we combined different computational methods to investigate how large the hidden diversity of primary metabolic gene clusters is in gut microbiota, using a case study on flavoenzyme-associated pathways. We revealed a large collection of putative MGCs capable of transforming a wide range of substrates including saccharides, peptides and lipids. These results evinced that many MGCs with relevant implications for the host remain to be characterized. Moreover, it served as a proof-of-concept that targeting coding genes (Fe-S flavoenzyme coding genes in this case) in a relevant genomic context is a useful approach to mine genomes for specialized primary MGCs. From this analysis, we learned that Fe-S flavoenzymes might play a more important role in anaerobic metabolism than previously anticipated and that specialized primary metabolism is strain-specific. Motivated by these findings and by the lack of a method to mine bacterial genomes for specialized primary metabolic MGCs, we designed a new tool to functionally profile the human gut microbiome, called gutSMASH (**Chapter 4**). By using this tool, which not only predicts known MGCs involved in primary metabolism and bioenergetics but also putative ones, we made a step forward in our understanding of how bacteria from different taxonomic groups contribute to microbiome-derived chemistry. To make the tool globally accessible, we built the gutSMASH web server (**Chapter 5**), which predicts and annotates MGCs given any genomic sequence. Finally, to fully understand these complex communities and establish causation between microbial functions and host phenotypes, it is important to obtain more biological insights at different molecular levels. This can only be achieved by combining different types of data or "-omic" layers. Since metabolic potential does not (always) correlate with molecule abundance (Chapter 4), it is important to evaluate if

gene clusters are expressed or not. To this end, we designed BiG-MAP, a method to assess (differential) abundance and expression profiles of gene clusters across metagenomic and metatranscriptomic samples (**Chapter 6**). BiG-MAP can be useful to pinpoint MGCs that are expressed under certain conditions, thus providing more evidence that their metabolic products are actually present in the host. Moreover, as there are still many MGCs with unknown function (Chapter 3 and 4), this tool can help prioritize their characterization based on the most interesting abundance/expression profiles, by, e.g., correlating these to health/disease phenotypes.

Resum

El cos humà està colonitzat per bilions de microorganismes. Entre ells, bacteris, fongs i virus que habiten a les nostres cavitats i superfícies, com per exemple la pell, la cavitat oral, el tracte respiratori i gastrointestinal. En cada lloc del cos hi resideixen comunitats microbianes específiques, també conegudes com a microbiota. En total pesen al voltant de 2 Kg, considerant-se per aquest motiu com un òrgan més. La microbiota inclou una àmplia gamma de microorganismes, principalment bacteris, que ens ajuden a digerir els aliments, sintetitzar vitamines, madurar el sistema immunitari i prevenir les infeccions de patògens, entre altres funcions. Els desequilibris en aquestes comunitats microbianes també s'han relacionat amb varies malalties, com ara el càncer de còlon, la malaltia inflamatòria intestinal, les malalties cardiovasculars, l'obesitat i la diabetis. Per tant, estudiar la composició i la funció de la microbiota humana pot ajudar a comprendre millor aquestes malalties des del punt de vista molecular.

El número de cèl·lules bacterianes són superiors a les cèl·lules humanes en una proporció d'1 a 10. La diversitat genòmica de la microbiota humana els permet respondre de diferents maneres a estímuls mitjançant l'excreció de petites molècules, que poden mediar les interaccions entre microorganismes o microorganisme-humà. Per aquest motiu, és important saber quins bacteris poden sintetitzar aquestes molècules, però també com i quan les sintetitzen. Algunes d'aquestes molècules els serveixen per competir amb altres bacteris (metabolisme secundari) o estan implicades en el creixement i desenvolupament de la cèl·lula (metabolisme primari). Quan aquests compostos tenen un paper més especialitzat en l'adaptació, també es coneixen com a metabòlits primaris especialitzats. Els gens responsables de la regulació, síntesi i transport d'aquestes molècules es troben sovint agrupats en el genoma, i aquesta disposició gènica es coneix com a agrupacions gèniques metabòliques, en anglès Metabolic Gene Clusters (MGCs). Més concretament, quan intervenen metabòlits secundaris, aquestes regions genòmiques es coneixen com a agrupacions de gens biosintètics (Biosynthetic Gene Clusters, BGCs). En definitiva, la circulació d'aquestes

molècules pot donar lloc a diferents fenotips (característiques visibles), beneficiosos o perjudicials per a l'humà. Per tant, per entendre millor com influeixen els microorganismes en la salut, és important definir el seu potencial metabòlic i caracteritzar els mecanismes moleculars que hi ha darrere d'aquests fenotips.

S'han dissenyat diverses eines bioinformàtiques per predir el potencial metabòlic d'aquests bacteris, com antiSMASH. Aquest és l'algoritme de referència per predir i analitzar BGCs que codifiquen metabòlits secundaris. antiSMASH pot ser una gran font d'informació quan s'utilitzen molts genomes com a entrada, però és difícil explotar aquestes dades completament mitjançant les eines bioinformàtiques actuals. Així en el **Capítol 2** es presenta un mètode per emmagatzemar i organitzar totes aquestes dades massives en forma d'una base de dades anomenada antiSMASH versió 2. És una plataforma que permet a l'usuari realitzar cerques inter-genòmiques de manera interactiva a partir de resultats pre-computats. A part d'antiSMASH, hi ha altres eines per predir el metabolisme microbià però es centren en el metabolisme primari més genèric. Com que no existia cap eina per predir MGC primaris especialitzats, en el **Capítol 3** es combinen diferents mètodes computacionals per investigar l'extensió i diversitat oculta d'aquests MGCs en la microbiota intestinal mitjançant l'estudi d'un tipus d'enzims en concret: el flavoenzims. En aquest estudi, revelem una gran col·lecció de MGCs no caracteritzats capaços de transformar una àmplia gamma de substrats, inclosos sacàrids (sucres), pèptids (petites proteïnes) i lípids (greix). Aquests resultats demostren que encara hi ha molts MGCs per estudiar que poden tenir implicacions rellevants per a l'ésser humà. A més, serveix com a prova de concepte que aquest tipus d'eines que identifiquen gens codificants d'interès (en aquest cas gens que codifiquen per flavoenzims) en contextos genòmics rellevants es un mètode útil per predir MGCs bacterians. A partir d'aquesta anàlisi, aprenem que els flavoenzims podrien tenir un paper més important en el metabolisme anaeròbic fins ara desconegut i que alguns d'aquests MGCs no es troben en totes les soques

bacterianes. Motivats per aquestes resultats i per la manca d'un mètode per predir MGCs sistemàticament, dissenyem una nova eina per descriure el funcionament del microbioma intestinal humà, anomenada gutSMASH (**Capítol 4**). Mitjançant l'ús d'aquesta eina, que no només prediu MGC coneguts implicats en el metabolisme primari i la bioenergètica, sinó també MGCs de funció desconeguda i potencialment interessants, fem un pas endavant en la nostra comprensió de com els bacteris de diferents grups taxonòmics contribueixen a la química derivada del microbioma. Per fer l'eina accessible a tothom, hem creat el servidor web gutSMASH (**Capítol 5**), que prediu i anota MGCs donada qualsevol seqüència genòmica. Finalment, per comprendre plenament aquestes comunitats complexes i establir la causalitat entre les funcions microbianes i els fenotips de l'humà, és important obtenir més coneixements biològics a diferents nivells moleculars i això només es pot aconseguir combinant diferents tipus de dades. Com que el potencial metabòlic no (sempre) es correlaciona amb l'abundància de molècules (Capítol 4), és important avaluar si els grups de gens s'expressen (manifestació del fenotip) o no. Amb aquesta finalitat, dissenyem BiG-MAP, un mètode per avaluar diferencialment l'abundància i els perfils d'expressió dels MGCs a través de mostres metagenòmiques (tots els gens d'una comunitat microbiana) i metatranscriptòmiques (perfil d'expressió dels gens de tota la comunitat microbiana) (**Capítol 6**). BiG-MAP pot ser útil per identificar MGCs que s'expressen en determinades condicions, proporcionant així més evidències que els seus productes metabòlics estan realment presents a l'humà, o qualsevol altre organisme com per exemple en plantes. A més, com que encara hi ha molts MGCs amb funció desconeguda (Capítols 3 i 4), aquesta eina pot ajudar a prioritzar la seva caracterització en funció dels perfils d'abundància / expressió més interessants, per exemple, correlacionant-los amb fenotips de salut / malaltia.

Resumen

El cuerpo humano está colonizado por billones de microorganismos. Entre ellos, encontramos bacterias, hongos y virus, que habitan en nuestras cavidades y superficies, como por ejemplo la piel, la cavidad oral, el tracto respiratorio y gastrointestinal. En cada lugar del cuerpo residen comunidades microbianas específicas, también conocidas como microbiota, que suman un peso total de alrededor de 2 Kgs, considerándose por este motivo como un órgano más. La microbiota incluye una amplia gama de microorganismos, principalmente bacterias, que nos ayudan a digerir los alimentos, sintetizar vitaminas, madurar el sistema inmunitario y prevenir las infecciones causadas por patógenos, entre otras funciones. Los desequilibrios en estas comunidades microbianas se han relacionado con diversas enfermedades, como cáncer de colon, enfermedad inflamatoria intestinal, enfermedades cardiovasculares, obesidad y diabetes. Por lo tanto, estudiar la composición y la función de la microbiota humana puede ayudar a comprender mejor estas enfermedades desde el punto de vista molecular.

El número de células bacterianas es superior al de células humanas en una proporción de 1 a 10. La diversidad genómica de la microbiota humana les permite responder de diferentes maneras a estímulos mediante la excreción de pequeñas moléculas, que pueden mediar la interacción entre microorganismos o microorganismo-humano. Por este motivo, es importante saber qué bacterias pueden sintetizar estas moléculas, pero también cómo y cuándo las sintetizan. Algunas de estas moléculas les sirven para competir con otras bacterias (metabolismo secundario) o están implicadas en el crecimiento y desarrollo de la célula (metabolismo primario). Cuando estos compuestos tienen un papel más especializado en la adaptación, también se conocen como metabolitos primarios especializados. Los genes responsables de la regulación, síntesis y transporte de estas moléculas se encuentran a menudo agrupados en el genoma, y esta disposición se conoce como agrupaciones génicas metabólicas, en inglés: Metabolic Gene Clusters (MGCs). Más concretamente, cuando intervienen metabolitos secundarios, estas regiones genómicas se conocen como agrupaciones de genes biosintéticos (Biosynthetic Gene Clusters, BGCs). En definitiva, la

circulación de estas moléculas puede dar lugar a diferentes fenotipos (características visibles), beneficiosos o perjudiciales para el humano. Por lo tanto, para entender mejor cómo influyen los microorganismos en la salud, es importante definir su potencial metabólico y caracterizar los mecanismos moleculares que hay detrás de estos fenotipos.

Se han diseñado diversas herramientas bioinformáticas para predecir el potencial metabólico de estas bacterias, como antiSMASH. Este es el algoritmo de referencia para predecir y analizar BGCs que codifican metabolitos secundarios. antiSMASH puede ser una gran fuente de información cuando se utilizan muchos genomas de entrada, pero es difícil explotar estos datos completamente mediante las herramientas bioinformáticas actuales. De este modo, en el **Capítulo 2** se presenta un método para almacenar y organizar todos estos datos masivos en forma de una base de datos llamada antiSMASH versión 2. Es una plataforma que permite al usuario realizar búsquedas inter-genómicas de manera interactiva a partir de resultados pre-computados. Aparte de antiSMASH, hay otras herramientas para predecir el metabolismo microbiano, pero se centran en el metabolismo primario más genérico. Como no existía ninguna herramienta para predecir MGCs primarios especializados, en el **Capítulo 3** se combinan diferentes métodos computacionales para investigar la extensión y diversidad oculta de estos MGCs en la microbiota intestinal mediante el estudio de un tipo de enzimas en concreto: las flavoenzimas. En este estudio, revelamos una gran colección de MGCs no caracterizados capaces de transformar una amplia gama de sustratos, incluidos sacáridos (azúcares), péptidos (pequeñas proteínas) y lípidos (grasas). Estos resultados demuestran que todavía hay muchos MGCs por estudiar que pueden tener implicaciones relevantes para el ser humano. Además, estos análisis sirven como prueba de concepto y demuestran que este tipo de herramientas que identifican genes codificantes de interés (en este caso genes que codifican para flavoenzimas) en contextos genómicos relevantes es un método útil para predecir MGCs bacterianos. A partir de estos análisis, aprendemos que las flavoenzimas podrían tener un papel más importante en el metabolismo

anaeróbico, hasta ahora desconocido, y que algunos de estos MGCs no se encuentran en todas las cepas bacterianas. Motivados por estos resultados y por la falta de un método para predecir MGCs de una forma sistemática, diseñamos una nueva herramienta para describir el funcionamiento del microbioma intestinal humano, llamada gutSMASH (**Capítulo 4**). Mediante el uso de esta herramienta, que no sólo predice MGCs conocidos implicados en el metabolismo primario y la bioenergética, sino también MGCs de función desconocida y potencialmente interesantes, damos un paso adelante en nuestra comprensión de cómo las bacterias de diferentes grupos taxonómicos contribuyen a la química derivada del microbioma. Para hacer la herramienta accesible a la investigación, hemos creado el servidor web gutSMASH (**Capítulo 5**), que predice y anota MGCs dada cualquier secuencia genómica. Sin embargo, para comprender plenamente estas comunidades complejas y establecer la causalidad entre las funciones microbianas y los fenotipos de lo humano, es importante obtener más conocimientos biológicos a diferentes niveles moleculares y esto sólo se puede conseguir combinando diferentes tipos de datos. Como el potencial metabólico no (siempre) se correlaciona con la abundancia de moléculas (Capítulo 4), es importante evaluar si los grupos de genes se expresan (manifestación del fenotipo) o no. Con este fin, diseñamos hemos diseñado BiG-MAP, un método para evaluar la abundancia diferencial y los perfiles de expresión de los MGCs a través de muestras metagenómicas (todos los genes de una comunidad microbiana) y metatranscriptómicas (perfil de expresión de los genes de toda la comunidad microbiana) (**Capítulo 6**). BiG-MAP puede ser útil para identificar MGCs que se expresan en determinadas condiciones, proporcionando así más evidencias de que sus productos metabólicos están realmente presentes en el humano, o cualquier otro organismo, como por ejemplo plantas. Además, como todavía hay muchos MGCs con función desconocida (Capítulos 3 y 4), esta herramienta puede ayudar a priorizar su caracterización en función de los perfiles de abundancia / expresión más interesante, por ejemplo, correlacionando-con fenotipos de salud / enfermedad.

Acknowledgements

A PhD can be described as the highest level of academic qualification one can achieve after conducting an original research in a specific field that is worth compiling into a PhD thesis. While I agree with this technical definition, I can now confirm that doing a PhD means a lot more than that. In my case, during these years, I did not only learn and expand my knowledge as a researcher but I did also witness a significant personal growth. For this reason, I will always be thankful to all the people that have been involved into this journey in one way or another.

To start off, I would like to thank Marnix Medema, my supervisor and co-promotor for giving me this opportunity. Since the very first day you have proven to be committed to science but also to truly care about people. I have been always amazed for how much you can achieve within 24 hours; you are incredibly fast at replying to emails, providing input on reports, manuscripts, posters, etc., and coming up with solutions to any given problem and still have some time to ask how things are going and not only project-wise. In fact, I remember that when sharing my experience with other PhD students, I oftentimes was stared at suspiciously because of my positivity and happy experience. I shortly understood that enjoying your time while doing a PhD is more an exception rather than the rule and I believe one of the key components of such a great experience was you. Thanks again for your patience, dedication, understanding and caring, you have been a true role model for me that inspires me now and will for sure shape my future self.

Of course, this project would not have been the same without Michael Fischbach. My other supervisor that did his best mentoring me from the US, which helped us excelling at online meetings long before COVID times. I would like to thank you, first, for making our (almost) weekly Skypes much more fun and entertaining but also for sharing your knowledge with me. Indeed, certain times I felt extremely privileged for having direct access to your “private lectures”. Second, thank you for the warm welcome I received during my one-month stay at Stanford University by you and your friendly group

and third, for teaching me to avoid conformism. Now seeing the result of our master piece (Chapter 4 of this thesis) I am very happy to have perfected the figures countless times and prior to that, do tons of other analysis that at the end were not included in the article but that helped me acquire new technical skills and better inspect the data from other perspectives. From the same university, I would like to also dedicate few words to Dylan Dodd. He joined the team towards the end of the project but when most needed. I would like to thank you for involving yourself so much in the project and so rapidly and always finding time to get your hands dirty. Thanks for your time and effort, for sure the end result would not have been the same without you.

I also want to thank Dick de Ridder, my promotor. Dick, I have been very lucky to have you by my side during these 4 years but also in important and complicated times during my PhD. I admire your practicality and how you handle problems. In fact, I feel that things I perceived to be quite concerning, for you were only small bumps on the road. I hope I learned this lesson from you to be a bit more practical also in future experiences. Also, another important learning is to always be in your PubQuiz team, where winning is almost guaranteed!

Special thanks to the all the co-authors of the research projects described in this thesis, your support in a way or in another has been key for me to finalize this chapter of my career. These include Kai, Simon and Tilmann from the Technical University of Denmark, the other members of the Medema group and Lianmin and Jing from the University of Groningen. Also, I would like to thank the rest of the Bioinformatics department for the nice times we have spent together during lunch, borrels, barbecues and several retreats. It has been nice to meet you all also outside the academic environment. I also want to thank the bachelor and master students that decided to do their thesis with me. Specially, thanks a lot to Petros, Koen and Hannah for giving me the opportunity to guide your projects but also to work more as a team.

Definitely, there have been times when pressure to finish up things has been high and Hannah, your predisposition to help made me feel a lot more accompanied. I wish you the best and a very successful PhD!

Last but not least from my working environment, I would also like to address to you, Geertjan (kind of from work right?) and Eef, my paranymphs. Thanks for accepting the invitation of being by my side not only on this special occasion but also during these years. It is curious that my two paranymphs are both Dutch because I have met more internationals than locals but there is truly something special with you. It has been a pleasure to also get to know the Netherlands and your culture better from your perspective. I hope you have enjoyed as much as I did around you and have learned indispensable skills in life such as playing padel (Geertjan) or making sangria (Eef) that will take you very far!

Many thanks also to the people from the MMB group at the UMC Utrecht for your warm welcome. Definitely, you have made my transition from one position to another a lot easier but also very exciting. I hope to have a lot of fun working together the coming two years.

Fora de l'àmbit professional, també m'agradaria agrair els meus pares el seu suport incondicional. Sempre m'heu dit que he sigut una bona estudiant però també perquè m'heu ensenyat a superar-me, ser valenta i sobre tot creure més en mi mateixa. De fet, sense tots aquests valors no se si hagués pres moltes de les decisions que he pres fins ara, incloent la de canviar laboratori per un ordinador, perquè no va ser gens trivial. Gràcies pel vostre esforç i dedicació (sobretot això, perquè no m'imagino arribar aquí sense saber-me les taules de multiplicar), per estar sempre al meu costat tant en els moments fàcils com els més difícils i estimar-me tant. Sentir que no estàs sol no té preu i espero que amb aquestes paraules jo també us pugui transmetre tot el meu agraïment i estima, farem sempre pinya passi el que passi! Es clar que aquesta pinya no seria el mateix sense la Clàudia i el Nil

(i com oblidar al Boyete), els meus petits deixebles/germans. Ara que ja som grans i tenim una relació més adulta (sobretot amb tu Nil), espero que sempre puguem comptar l'un amb l'altre i ajudar-nos en el que fagi falta, us estimo molt! Esther y Jose, como veis os incluyo en la sección de familia porque verdaderamente así os considero. Gracias por escucharme en los momentos que lo necesitaba, por preocuparos y por estar cerca nuestro siempre que habéis podido. Me siento muy privilegiada de poder contar con vuestro apoyo y compañía. Aritz que no me olvido de ti, pero no sé si ponerte en el sector de familia vasca o catalana! Gracias por cuidar de mi hermana como lo haces y aunque no nos veamos muy a menudo siempre es un placer hablar y estar contigo, espero coincidir más contigo ahora que vivimos más cerca. Tiques i avis, gràcies per ser el pal de paller d'aquesta família, en especial a tu avi, que m'hagués agradat molt poder compartir aquest dia amb tu perquè sé que n'hauries estat molt orgullós.

També m'agradaria agrair als amics, tant els que han sigut allà des de ja fa molt temps com a la gent nova que he conegut més recentment. Rebeca, el temps ha demostrat que siguem on siguem, fem el que fem o estem amb qui estiguem sempre puc comptar amb tu, així que gràcies per tot i per fer la meva vida una mica més emocionant (això sobretot!). Brian, moltes gràcies a tu també pel suport tècnic però també per la teva companyia durant aquest temps, un plaer haver conegut l'Anthony! También a todos los españoles e internacionales que he conocido en Holanda, que sois unos cuantos repartidos por todo el país, por haberme hecho sentir menos lejos de casa y más acompañada durante este tiempo. Gracias especialmente a la gente de *Droef* y la *Farm* por contar conmigo en muchas ocasiones. También al *dream team* de padeleros que no solo me habéis hecho mejorar en el padel (o eso creo yo) sino también pasar momentos inolvidables. I do not forget neither the Pint of Science NL team, I had a lot of fun with you but also learned a lot.

I falta per agrair-li a una de les persones més importants per mi, el meu company d'aventures. Jorge, antes de mudarnos a Holanda no teníamos muy claro que nos depararía el futuro, lo que sabíamos era que queríamos vivirlo juntos. Sin duda durante este tiempo hemos evolucionado como personas, pero también como pareja, no sin mucho esfuerzo y paciencia, mucha comunicación y frustración, pero también con muchas ganas de entendernos y conocernos mejor. No sé cuántas veces debería escribir "gracias" para agradecerte todo lo que has hecho por mí, no solo a nivel personal (que ya es mucho) sino también profesional, porque sin ti hoy no podría estar defendiendo la tesis. Acabamos esta etapa con nota para empezar otra más emocionante si cabe y seguir sumando aventuras juntos, porque somos un equipo.

11. **Pascal Andreu V.**, Roel-Touris J., Dodd D., Fischbach M.A., Medema M.H. The gutSMASH web server: automated identification of primary metabolic gene clusters from the gut microbiota. Nucleic Acids Research, gkab353 (2021)

10. Bickhart D.M., Kolmogorov M., Tseng E., Portik D.M., Korobeynikov A., Tolstoganov I., Uritskiy G., Liachko I., Sullivan S.T., Bong Shin S., Zorea A., **Pascal Andreu V**., Panke-Buisse K., Marnix M.H., Mizrahi I., Pevzner P.A., Smith T.P.L. Generation of lineage-resolved complete metagenome-assembled genomes by precision phasing. bioRxiv, https://doi.org/10.1101/2021.05.04.442591 (2021)

9. **Pascal Andreu V.**, Augustijn H. E., Chen L., Zhernakova A., Fu Y., Fischbach M.A., Dodd D., Medema M.H. A systematic analysis of metabolic pathways in the human gut microbiota. bioRxiv, https://doi.org/10.1101/2021.02.25.432841 (2021)

8. **Pascal Andreu V.**, Augustijn H. E., van den Berg, K., van der Hooft J.J.J, Fischbach M.A., Medema M.H. BiG-MAP: an automated pipeline to profile metabolic gene cluster abundance and expression in microbiomes. bioRxiv, https://doi.org/10.1101/2020.12.14.422671 (2020)

7. **Pascal Andreu V.**, Fischbach M.A., Medema M.H. Computational genomic discovery of diverse gene clusters harbouring Fe-S flavoenzymes in anaerobic gut microbiota. Microbial Genomics 6, e000373 (2020)

6. Machiels K., Pozuelo del Río M., Martinez-De la Torre A., Xie Z., **Pascal Andreu V.**, Sabino J., Santiago A., Campos D., Wolthuis A., D'Hoore A., De Hertogh G., Ferrante M., Manichanh C., Vermeire S., Early Postoperative Endoscopic Recurrence in Crohn's Disease Is Characterised by Distinct Microbiota Recolonisation. Journal of Crohn's and Colitis 14, 1535–1546 (2020)

5. Kautsar S.A., Blin K., Shaw S., Navarro-Muñoz J.C., Terlouw B.R., van der Hooft J.J.J, van Santen J.A., Tracanna V., Suarez Duran H.S, **Pascal Andreu V.**, Selem-Mojica N., Alanjary M., Robinson S.L, Lund G., Epstein S.C., Sisto A.C, Charkoudian L. K., Collemare J., Linington R. G., Weber T, Medema M. H, MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Research 48, D454–D458 (2019).

4. Hanachi M, Manichanh C, Schoenenberger A, **Pascal V**, Levenez F, Cournède N, Doré J, Melchior JC. Altered host-gut microbes symbiosis in severely malnourished anorexia nervosa (AN) patients undergoing enteral nutrition: An explicative factor of functional intestinal disorders?. Clinical Nutrition 38, 2304-2310 (2019).

3. Blin K., **Pascal Andreu, V.**, L C de los Santos, E., Del Carratore F., Lee SY, Medema M. H, Weber T., The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. Nucleic Acids Research 47, D625–D630 (2019).

2. **Pascal V**, Pozuelo M, Borruel N, Casellas F, Campos D, Santiago A, Martinez X, Varela E, Sarrabayrouse G, Machiels K, Vermeire S, Sokol H, Guarner F, Manichanh C. A microbial signature for Crohn's disease. Gut 66, 813-822 (2017).

1. Martinez, X., Pozuelo, M., **Pascal, V.**, Campos D, Gut I., Gut M., Azpiroz F., Guarner F, Manichanh C. MetaTrans: an open-source pipeline for metatranscriptomics. Scientific Reports 6, 1-12 (2016).

Education Statement of the Graduate School

Experimental Plant Sciences

Issued to: Victòria Pascal Andreu
Date: 13 September 2021
Group: Bioinformatics
University: Wageningen University & Research

1) Start-Up Phase	date	cp
► **First presentation of your project**		
Algorithms for biosynthetic pathway discovery and analysis in the human microbiome	20 Apr 2017	1.5
► **Writing or rewriting a project proposal**		
Algorithms for biosynthetic pathway discovery and analysis in the human microbiome	09 Mar 2017	6.0
► **MSc courses**		
Subtotal Start-Up Phase		7.5

2) Scientific Exposure	date	cp
► **EPS PhD student days**		
Annual EPS PhD Student Days 'Get2Gether' 2018, Soest, NL	15-16 Feb 2018	0.6
Annual EPS PhD Student Days 'Get2Gether' 2019, Soest, NL	11-12 Feb 2019	0.6
► **EPS theme symposia**		
EPS Theme 4 "Genome Biology", Wageningen, NL	13 Dec 2019	0.3
EPS Theme 3 "Metabolism and Adaptation", Nijmegen, NL	21 Oct 2019	0.3
EPS Theme 4 "Genome Biology", online	11 Dec 2020	0.2
► **Lunteren Days and other national platforms**		
The 3rd Dutch Bioinformatics & Systems Biology conference (BioSB 2017), Lunteren, NL	04-05 Apr 2017	0.6
The 5th Dutch Bioinformatics & Systems Biology conference (BioSB 2019), Lunteren, NL	02-03 Apr 2019	0.6
► **Seminars (series), workshops and symposia**		
Seminar: Dr. Peer Bork, Deciphering function and evolution of biological systems", followed by masterclass on career advice, publishing, bioinformatics, big data, and science in general	28 Mar 2017	0.1
Seminar: Eric Brown, NCOH Lecture	12 Dec 2017	0.1
Seminar: Andres Aranda-Día, The ecology of poop teas	01 Jun 2018	0.1
Seminar, Dylan Dodd, A mechanistic approach to define the biology of microbiome-derived metabolites	08 Jun 2018	0.1
B-Wise Seminar: Judith Risse & Ruben van Heck	10 Jan 2017	0.2
B-Wise Seminar: Richard Notebaart & Fleur Gawehns-Bruning	07 Feb 2017	0.2
B-Wise Seminar: Egon Willighagen & Sabrina Simon	07 Mar 2017	0.2
B-Wise Seminar: Hesham Gibriel & Eduardo Saccenti	11 Apr 2017	0.2
B-Wise Seminar: Yang Li & Satria Kautsar	02 May 2017	0.2
B-Wise Seminar: Berend Snel	06 Jun 2017	0.2
B-Wise Seminar: Jens Allmer & Jesse van Dam	05 Sep 2017	0.2
B-Wise Seminar: Katy Wolstencroft & Dennis van Muijen	03 Oct 2017	0.2
B-Wise Seminar: Purva Kulkarni & Twan America	07 Nov 2017	0.2
B-Wise Seminar: Mathijs Nieuwenhuis & Jorge Navarro	05 Dec 2017	0.2
B-Wise Seminar: Anton Feenstra & Ehsan Motazedi	09 Jan 2018	0.2
B-Wise Seminar: Justin van der Hooft & Victor Carrion	06 Feb 2018	0.2
B-Wise Seminar: Martijn Derks & Rik Kooke	06 Mar 2018	0.2
B-Wise Seminar: Jeroen de Ridder & Miguel Correa	03 Apr 2018	0.2
B-Wise Seminar: Sumanth Mutte & Hernando Suarez Duran	01 May 2018	0.2
B-Wise Seminar: Joana Gonçalves & Jasper Depotter	05 Jun 2018	0.2
B-Wise Seminar: Gurnoor Singh & Janani Durairaj	04 Sep 2018	0.2
B-Wise Seminar: Christian Gilissen & Mohammad Alanjary	02 Oct 2018	0.2
B-Wise Seminar: Erik van den Bergh & Willem Kruijer	06 Nov 2018	0.2
B-Wise Seminar: Rachel Cavill & Mehmet Akdel	04 Dec 2018	0.2
B-Wise Seminar: Rik van Rosmalen & Sevgin Demirci	08 Jan 2019	0.2
B-Wise Seminar: Gerben Hermes & Pariya Behrouzi	05 Feb 2019	0.2
B-Wise Seminar: Martijn Huijnen & Mark Sterken	05 Mar 2019	0.2
B-Wise Seminar: Sven Warris & Vittorio Tracanna	02 Apr 2019	0.2
B-Wise Seminar: Jorge Roel & Victoria Pascal Andreu	04 Jun 2019	0.1
B-Wise Seminar: Veronika Laine & Raul Wijfjes	03 Sep 2019	0.2
B-Wise Seminar: Eliana Papoutsoglou & Roeland Voorrips	01 Oct 2019	0.2
B-Wise Seminar: Jingyuan Fu & Catarina Sales e Santos Loureiro	05 Nov 2019	0.2
B-Wise Seminar: Simon Rogers & Barbara Terlouw	03 Dec 2019	0.2
B-Wise Seminar: Mario Calus & Eef Jonkheer	07 Jan 2020	0.2
B-Wise Seminar: Chaozhi Zheng & Carlos Lannoy	04 Feb 2020	0.2
B-Wise Seminar: Age Smilde & Cristina Furlan	03 Mar 2020	0.2
Workshop: Hadley Wickham, Make your R code purrr with functional programming, Stanford, USA	24 May 2018	0.2
Workshop The Brave New World of Smart Data & Semantics in the Life Sciences, Wageningen, NL	24 Jan 2019	0.2
Symposium: Microbiomes and Metagenomics: Analysis and challenges, Utrecht, NL	23 Nov 2018	0.3
Symposium: Infant & Pregnancy Microbiome Experts Meeting-- BaseClear, online	14 May 2020	0.3
Symposium: The Barcelona Debates on the Human Microbiome 2020, online	26-27 Jun 2020	0.3
► **Seminar plus**		
► **International symposia and congresses**		
International VAAM Workshop 'Biology of Bacteria Producing Natural Products', Tübingen, DE	27-29 Sep 2017	0.7
EMBO-EMBL Symposium 'The Human Microbiome', Heidelberg, DE	16-19 Sep 2018	1.1
16th Copenhagen Bioscience Conference 'Natural Products - Discovery, Synthesis and Application', Copenhagen, DK	03-08 May 2019	1.2
► **Presentations**		
Poster at BioSB	04 Apr 2017	1.0
Poster at VAAM Workshop 'Biology of Bacteria Producing Natural Products'	27 Sep 2017	1.0
Talk about my project in Christian Hansen annual meeting	25 Apr 2018	1.0
Poster at the EMBO-EMBL Symposium 'The Human Microbiome'	17 Sept 2018	1.0
Talk at BioSB 2019	03 April 2019	1.0
Poster at Copenhagen Bioscience Conference 'Natural products - Discovery, Synthesis and Application'	06 May 2019	1.0
Talk CMCB Leiden	26 Jun 2020	1.0
► **3rd year interview**		
► **Excursions**		
Subtotal Scientific Exposure		21.2

3) In-Depth Studies	date	cp
► **Advanced scientific courses & workshops**		
VLAG postgraduate course 'The Intestinal Microbiome and Diet in Human and Animal Health', Wageningen, NL	06-08 Feb 2017	0.9
Course 'High Performance Computing Cluster Agrogenomics', Wageningen, NL	07 Jun 2017	0.3
► **Journal club**		
Participation in a literature discussion group	2017-2020	3.0
► **Individual research training**		
Period abroad - Stanford University (California), USA	19 May-15 Jun 2018	3.0
Subtotal In-Depth Studies		7.2

4) Personal Development	date	cp
► **General skill training courses**		
Workshop Research Data Management, Utrecht, NL	17 Mar 2017	0.2
RSG Netherlands-Soft Skills: Networking and presentation skills in science, Utrecht, NL	22 Mar 2017	0.2
Course Career Perspectives, Wageningen, NL	21 Mar - 18 Apr 2019	1.5
Workshop Scientific Paper Writing, Wageningen, NL	24 Oct 2019	0.2
Course Scientific Publishing, Wageningen, NL	15 Oct 2019	0.3
EPS Postdoc Career Day, Wageningen, NL	07 Feb 2020	0.3
► **Organisation of meetings, PhD courses or outreach activities**		
Pint of Science 2018	14 May 2018	1.5
Pint of Science 2019	20-22 May 2019	1.5
► **Membership of EPS PhD Council**		
Subtotal Personal Development		5.7

5) Teaching & Supervision Duties	date	cp
► **Courses**		
Practical Computing for Biologists	Sep 2017	0.5
Practical Computing for Biologists	Sep 2018	0.5
Practical Computing for Biologists	Sep 2019	0.5
► **Supervision of BSc/MSc students**		
BSc student Koen van den Berg	2018	
BSc student Lucas Schuit	2019	
MSc student Petros Skiadas	2018	3.0
MSc student Koen van den Berg	2019	
MSc student Hannah Augustijn	2019/2020	
Subtotal Teaching & Supervision Duties		4.5

TOTAL NUMBER OF CREDIT POINTS*	**46.1**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.

** A credit represents a normative study load of 28 hours of study.*

The research described in this thesis was financially supported by U.S. Defense Advanced Research Projects Agency's Living Foundries program.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by Victòria Pascal Andreu
Printed by Proefschriftmaken