THE UNIVERSITY OF CHICAGO

A NOVEL FRAMEWORK FOR METABOLISM RECONSTRUCTION IDENTIFIES DETERMINANTS OF MICROBIAL RESILIENCE AND LIFESTYLE IN THE HUMAN GUT AND GLOBAL SURFACE OCEANS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
AND
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN BIOPHYSICAL SCIENCES

BY
IVA ANNAMARIA VESELI

CHICAGO, ILLINOIS
AUGUST 2023

To my father, who is my inspiration.

To my mother, who is my rock.

And to Florian, who is my love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Supplementary tables are available online at `https://doi.org/10.6084/m9.figshare.14138405` (Chapter 3.2), `https://doi.org/10.6084/m9.figshare.22579219` (Chapter 3.3), https://doi.org/10.6084/m9.figshare.22679080 (Chapter 4), and https://doi.org/10.1038/s41396-021-01135-1 (Chapter 5).

ACKNOWLEDGMENTS

The number of people who shaped this research and my personal development as a scientist are too many to enumerate here. I am grateful to all of them - my lab mates, my collaborators, my teachers, my friends, and my family. But I am especially thankful for my advisor, Meren, without whose support I would not be where I am today. Meren, thank you.

ABSTRACT

Microbes play a significant role in supporting life on our planet, and their metabolic capabilities mediate their interactions with each other and with their environments. In host-associated communities such as the human gut microbiome, microbes have been implicated in a variety of host physiological processes. Indeed, dysbiosis of the human gut microbiome is associated with several diseases and disorders. In the marine environment, microbes contribute to important biogeochemical cycles such as nitrogen fixation. The ability to predict metabolic capacity is thus critical to understanding microbial ecology in these systems. This thesis presents a novel software framework for estimating metabolic potential from ‘omics data and showcases its application to studies of the human gut microbiome and the marine microbiome. In the human gut, high metabolic independence emerges as a determinant of microbial fitness in the face of gut stress, as demonstrated by a longitudinal time-series analysis of colonization after fecal microbiota transplant (FMT) and a high-throughput meta-analysis of community metabolism in individuals with inflammatory bowel disease (IBD). In studies of the global surface oceans, this framework identifies an understudied yet abundant group of heterotrophic bacterial diazotrophs. Overall, this new tool facilitates diverse and flexible analyses of microbial metabolism from ‘omics datasets, leading to interesting insights into microbial ecology that are relevant to both human health and the health of the planet.

CHAPTER 1

INTRODUCTION

1.1 A microbial world

Microbes are practically everywhere on the planet's surface. They live in environments that are inhabitable to all other life forms (Shu and Huang, 2022), from thermal springs with extreme pH (Inskeep et al., 2013; Kozubal et al., 2012; Massello et al., 2020) to hypersaline habitats (Enache et al., 2012; Martínez et al., 2022; Wong et al., 2020) to hydrothermal vents at the bottom of the ocean (Gonnella et al., 2016; Davis and Moyer, 2008; Jebbar et al., 2015). They exist across the vast surfaces of the Earth, forming communities in soils and waters at most latitudes and longitudes (Xu et al., 2013; Amend et al., 2013; Sul et al., 2013; Fierer and Jackson, 2006). And there are multitudes of microbes in most spaces inside and on animals, including humans (Simon et al., 2019; Bordenstein and Theis, 2015; Gilbert et al., 2018).

The expansive prevalence of microbes matches their immense importance to many processes that support life. It was cyanobacteria that first oxygenated the planet (Sánchez-Baracaldo et al., 2022), enabling the evolution of larger, more complex life forms (Och and Shields-Zhou, 2012; Falkowski and Godfrey, 2008). Microbes contribute to global nutrient cycling, ensuring that all organisms have access to the key elements for cellular building blocks, especially carbon and nitrogen (Rousk and Bengtson, 2014; Kuypers et al., 2018; Falkowski et al., 2008). For example, marine microbes are responsible for about half of worldwide net primary production, thus serving as the basis of the global food web (Naselli-Flores and Padisák, 2022; Moran, 2015). They also play important roles in biogeochemical cycling of multiple nutrients, including nitrogen (Hutchins and Capone, 2022), sulfur (Jørgensen, 2021; Moran and Durham, 2019), phosphorus (Benitez-Nelson, 2000; Duhamel et al., 2021), and trace metals such as iron (Tortell et al., 1999; Morel and Price, 2003). These activities maintain the habitability of the biosphere by ensuring that critical elements for biological processes (especially energy generation) are continuously recycled (Madsen, 2011; Falkowski et al., 2008). The terrestrial microbiome is similarly impactful; it also contributes significantly to biogeochemical cycling (Crowther et al., 2019), and supports the growth and productivity of plants (Chaparro et al., 2012; van der Heijden et al., 2008). The best-studied example of plant-microbe interactions is the rhizosphere microbiome, which is particularly important to agriculture (Pantigoso et al., 2022; de Faria et al., 2021).

Just as plants rely on closely-associated microbes, animals have intimate relationships with microbiomes of their own. Insect species have co-evolved with their microbial endosymbionts, which produce essential nutrients for them (Moran, 2001) or modulate their behavior (Bi and Wang, 2020). Ocean invertebrates rely on microbes for food assimilation and bioluminescence (Osman and Weinnig, 2022). Ruminants digest plant matter with the help of a diverse consortia of cellulolytic microbes in their digestive tracts (Newbold and Ramos-Morales, 2020). Not all animals require microbial associations for survival (Hammer et al., 2019), but for many complex organisms, cohabitation with microbes is inescapable and these associations often provide important benefits to the host animal (Peixoto et al., 2021).

For humans in particular, several aspects of human health depend on the symbiotic relationship with the microbes living on and within us (Gilbert et al., 2018) – our microbiome primes the immune system (Ivanov and Littman, 2010), protects against pathogens (Khosravi and Mazmanian, 2013), helps us digest food (Hijova, 2019), influences brain development and behavior (Collins et al., 2012; Silva et al., 2020), and more. Of the several organ systems that benefit from microbial associations, by far the best-studied is the gastrointestinal tract. The gut microbiome impacts a wide variety of host physiological processes (Leser and Mølbak, 2009) and is a prominent research topic due to its potential use for diagnosis and treatment of a variety of diseases and disorders (Vijay and Valdes, 2022; Schupack et al., 2022).

Growing awareness of microbes' impact on life and technological improvements facilitating their study have spurred an ever-increasing body of research on microbiomes across different environments (Stulberg et al., 2016). The gut microbiome has been prominently featured in this work and has recently received a majority of public funding for host-associated microbiome research (NIH Human Microbiome Portfolio Analysis Team, 2019), which is a hallmark of its immense relevance to human health and disease.

1.2 The gut microbiome in humans: symbiosis is a two-way street

The gut microbiome is considered a human symbiont, being a community of organisms that collectively have a long-term relationship with their host despite the dynamic fluctuations of individual populations (Lloyd-Price et al., 2017; Caporaso et al., 2011; David et al., 2014a). It is a large community of about 10^{13} cells primarily concentrated in the colon (Sender et al., 2016) and composed mostly of bacteria, but also some archaea and yeasts (Woting and Blaut, 2016; Turnbaugh et al., 2009) (there is also a significant viral component, which is outside the scope of this dissertation). This typically diverse community is quite variable between individuals (Lloyd-Price et al., 2016a), as it is influenced by a number of host-specific factors (Hasan and Yang, 2019) such as age (Claesson et al., 2011), geography (Yatsunenko et al., 2012), environment (Turnbaugh et al., 2010), host genetics (Blekhman et al., 2015; Goodrich et al., 2014), diet (David et al., 2014b), and lifestyle (Jha et al., 2018). It is established early in human life (Jost et al., 2012, 2014; Bergström et al., 2014) and has a substantial effect on various aspects of host health (Ding et al., 2019; Shreiner et al., 2015).

Gut microbes have been implicated in a variety of physiological processes like immune system functioning (Bain and Cerovic, 2020; Ivanov and Littman, 2010), nutrient acquisition from otherwise indigestible substrates (Larsbrink et al., 2014; Portincasa et al., 2022; Hijova, 2019), and modulation of behavior (Silva et al., 2020; Collins et al., 2012). Part of their benefit to the host is that gut microbes provide a variety of useful metabolic capabilities, many of which are complementary to the functional capacity encoded in the human genome (Qin et al., 2010). Gut microbes synthesize a number of amino acids that are essential to humans (Metges, 2000; Hooper et al., 2002; Lin et al., 2017) and provide vitamins like biotin and cobalamin to their host (Hooper et al., 2002; Albert et al., 1980; Hill, 1997; Kau et al., 2011). They can make immunomodulatory molecules (Maslowski et al., 2009; Hoffman et al., 2022; Wang et al., 2021) and compounds that regulate host nutrient uptake and appetite (Bäckhed et al., 2004; Heiss and Olofsson, 2018). Some of these populations produce short-chain fatty acids such as acetate and butyrate, which are important energy sources for intestinal epithelial cells. These molecules also play a role in regulating gut barrier function and host immune responses (Martin-Gallausiaux et al., 2021; Zhang et al., 2022). Additionally, certain gut microbes conjugate host-derived bile acids into secondary bile acids, which are important regulatory molecules (Ridlon et al., 2014; Hylemon et al., 2009).

Importantly, human hosts can influence their gastrointestinal symbionts. Diet (David et al., 2014b), medication (especially antibiotics) (Palleja et al., 2018), and lifestyle (David et al., 2014a; Jha et al., 2018) can alter the composition and size of the gut community. Changes in host health can also influence the microbiota, and in turn, a number of diseases are associated with perturbations of the gut microbiome. These include inflammatory bowel diseases (Kostic et al., 2014; Nagalingam and Lynch, 2012; Knox et al., 2019a), diabetes (Karlsson et al., 2013; Qin et al., 2012), non-alcoholic fatty liver disease (Campo et al., 2019; Tokuhara, 2021), autoimmune disorders (Vaahtovuo et al., 2008; Hoffman et al., 2022), cardiovascular disease (Jie et al., 2017; Cui et al., 2017; Rath et al., 2017; Novakovic et al., 2020), cancer (Kostic et al., 2013; Raskov et al., 2017; Marchesi et al., 2011), and neurological disorders (Sorboni et al., 2022). A commonly-used (yet often vaguely-defined (Brüssow, 2020)) term for disease-related perturbations of the gut community is ‘dysbiosis’, which describes states of the microbiome that relate to one or several metrics (including composition, diversity and function) deviating from intestinal homeostasis (Lee and Chang, 2021).

In most cases, it is difficult to determine the relationship between dysbiosis in the gut microbiome and host disease states. Studies often take a host-centric view, focusing on the influence of the gut microbiota in disease pathogenesis and searching for ‘pathobionts’ that could be implicated as the root cause of disease (Jochum and Stecher, 2020; Clemente et al., 2012; Lee and Chang, 2021; Weiss and Hennet, 2017). Yet the reality is that the relationship between humans and their gut microbes is bidirectional – what impacts one affects the other. They are inextricably linked, and the complexity of this relationship makes it difficult to elucidate cause and effect (Shreiner et al., 2015). Thus, gut microbial dysbiosis may contribute to disease progression at the same time that disease pathophysiology exacerbates dysbiosis. Though many studies, including those described in this work, tend to discuss and interpret results with a focus on either the host side or the microbial side, it is important to keep in mind that these entities form an ecosystem.

1.3 Inflammatory bowel diseases and the gut microbiome

One human disease in which the gut microbiome is particularly relevant is inflammatory bowel disease (IBD). Though often referred to in the singular, IBD actually represents a heterogeneous group of chronic inflammatory disorders (Shan et al., 2022) which pose an increasingly common health risk around the globe and especially in industrialized countries (Kaplan, 2015). IBDs are characterized by chronic inflammation in the gastrointestinal (GI) tract that varies in location and level of continuity. The etiology of these diseases is complex and not yet well-understood, arising from a combination of host genetic risk factors, environmental and lifestyle factors, and negative interactions between the host and the gut microbiome. IBDs are thought to manifest from an overactive and misdirected immune response, and the gut microbiome represents one possible stimulus for this response (Knox et al., 2019a). The two main types of IBD, Crohn’s disease (CD) and ulcerative colitis (UC), are associated with dysbiosis of the human gut microbiome, particularly reduced microbial diversity (Kostic et al., 2014; Nagalingam and Lynch, 2012; Knox et al., 2019b).

Understanding the role of gut microbiota in IBD has been a prominent goal in human microbiome research. Studies focusing on individual microbial taxa that typically change in relative abundance in IBD patients have proposed a range of host-microbe interactions that may contribute to disease manifestation and progression, such as the reduction in butyrate-producing bacteria (Joossens et al., 2011; Schirmer et al., 2019; Henke et al., 2019; Machiels et al., 2014). However, even within well-constrained cohorts, inter-individual variability explains the majority of variance in all metrics yet explored to differentiate the microbiome of IBD patients from that of healthy individuals (Gevers et al., 2014; Schirmer et al., 2018b; Lloyd-Price et al., 2019; Khan et al., 2019). The focus on taxonomy, including the search for individual pathobionts and broad descriptions of compositional changes in the gut microbiome during disease, has not yielded a clear understanding of the microbiome's role in the etiology of this condition (Khan et al., 2019).

Given the inconsistency and unclear nature of taxonomy-related analyses, focus has slowly shifted to the functional potential of the microbiota in the context of this disease. Explorations of microbial gene content using reference genomes (or pangenomes) associated with 16S amplicon and/or metagenomic sequencing data have suggested that the IBD gut microbiome encodes fewer pathways for carbon metabolism and amino acid biosynthesis, as well as increased potential for oxidative stress response (Gevers et al., 2014; Morgan et al., 2012; Wlodarska et al., 2017; Ashton et al., 2017; Ananthakrishnan et al., 2017; Davenport et al., 2014; Tong et al., 2013; Lloyd-Price et al., 2019; Franzosa et al., 2019; Lewis et al., 2015; Vich Vila et al., 2018). However, such approaches may inaccurately reflect true functional potential due to the incomplete representation of genomic diversity in reference databases (Rodriguez-R et al., 2018; de la Cuesta-Zuluaga et al., 2020). Predicting functional potential directly from metagenomic assemblies would be more accurate, but is rarely done in the context of IBD (Knoll et al., 2017) (see, however, (Qin et al., 2010) for an example of this in healthy individuals). Studies leveraging metatranscriptomics for the study of gene expression in the IBD gut microbiome are similarly rare (Segal et al., 2019) and also have been limited to reference-based profiling (Schirmer et al., 2018b). In contrast, some studies have used metabolomics to analyze the compounds associated with IBD (Lloyd-Price et al., 2019; Franzosa et al., 2019; De Preter et al., 2015; Jansson et al., 2009). These have demonstrated the depletion of vitamins; short-, medium- and long-chain fatty acids; and certain secondary bile acids in the IBD gut, as well as the enrichment of polyunsaturated fatty acids, certain primary bile acids, and sphingolipids in this sample group. However, metabolomic data is difficult to interpret due to our inability to identify most compounds and to elucidate their biological sources – compounds associated with IBD might be related to host functional activity or external sources such as diet, and cannot be unequivocally linked to the gut microbiome (Shaffer et al., 2017).

Clearly, the quest to understand the gut microbiome's role in IBD will benefit from *de novo* estimation of microbial functional potential directly from metagenomic and metatranscriptomic data. Yet, signal at the individual gene level is extremely noisy and attempts to make sense of enzyme abundances in isolation of their metabolic contexts have not been fruitful (Gevers et al., 2014; Greenblum et al., 2012). Instead, a more insightful strategy would be to aggregate data from metabolically-related genes in order to draw conclusions on the presence, completeness, and/or throughput of metabolic pathways. Thus, metabolism reconstruction from 'omics data is a logical next step for assessing the links between IBD and the gut microbiome.

1.4 Strategies and challenges in metabolism reconstruction from 'omics data

Predicting the metabolic capabilities of microbes from genomic and metagenomic data is a culture-independent, high-throughput approach to investigating microbial functional capacity in environmental samples, including host-derived samples such as feces. Metabolism reconstruction, also referred to as estimation of metabolic potential, relies on the principle that known metabolic pathways can be defined in terms of particular gene functions – namely, the enzymes that catalyze each reaction in the pathway. Searching for evidence of these enzymes, typically in the form of gene annotations for the relevant enzyme families, allows inference of pathway presence, completeness, abundance, and/or throughput. While this computational estimation approach does not demonstrate the active use of these pathways *in vivo*, it does allow scientists to leverage the vast amount of sequencing data to generate hypotheses about metabolic capabilities that can then be verified in targeted experiments using methods such as transcriptomics, metabolomics, or growth assays.

There are two main strategies for estimation of metabolic potential from sequencing data. The first is metabolic modeling, in which genome-scale metabolic models (GSMMs) are built to computationally represent the network of available metabolic reactions for a particular organism (Fang et al., 2020; Gu et al., 2019). This strategy enables mathematical modeling of metabolic fluxes, typically with the linear programming technique known as flux-balance analysis (FBA) (Orth et al., 2010), which contextualizes the metabolic network within a set of constraints and thereby enables simulation of particular physiological conditions (Sen and Orešič, 2019). The second strategy is pathway prediction, which estimates the presence/absence and/or completeness of metabolic pathways to produce a summary of the metabolic capacity encoded in the input sequences. This technique has received less attention than metabolic modeling, but its results are more readily interpretable than models, and it is critical for understanding microbial functional roles without the need for a parameterized, *in silico* environment (Zhou et al., 2022). Both methods can be integrated with auxiliary information such as gene expression data or growth kinetics for validation of predicted metabolisms (Gu et al., 2019).

A variety of software tools exist for both types of metabolism reconstruction. Two early examples with basic approaches are the web-based server platforms KAAS (Moriya et al., 2007) and RAST (Aziz et al., 2008). KAAS simply highlights annotated enzymes within pathway maps from the KEGG database (Kanehisa et al., 2006), without producing any quantitative estimates. RAST similarly produces a limited summary of metabolism by categorizing enzymes into metabolic ‘subsystems’, but is also able to produce a metabolic model using the SEED infrastructure (DeJongh et al., 2007). There are a plethora of more contemporary modeling tools that generate GSMMs, including ModelSEED (Henry et al., 2010), RAVEN (Agren et al., 2013), merlin (Dias et al., 2015), CarveMe (Machado et al., 2018), and AuReMe (Aite et al., 2018). Several of these tools have been comprehensively reviewed (Faria et al., 2018; Mendoza et al., 2019; Gu et al., 2019), so the details of these platforms are not summarized here.

Software for pathway prediction include MinPath, DRAM, METABOLIC, and metaPathPredict. MinPath (Ye and Doak, 2009) uses integer programming to determine the minimum set of pathways that explain an input set of annotations. DRAM (Shaffer et al., 2020) and METABOLIC (Zhou et al., 2022) both integrate annotation of genes from various enzyme databases with estimation of pathway completeness; DRAM is specialized for working with metagenome-assembled genomes (MAGs) while METABOLIC focuses on biogeochemical cycles. The goal of metaPathPredict (Geller-McGrath et al., 2023) is to produce better estimations for incomplete genomes (especially MAGs reconstructed from environmental samples) using machine learning models trained on reference databases.

Though most of these tools specialize in one method of metabolism reconstruction, some software – such as Pathway Tools, KBase, gapseq, and KEMET – have the capacity for both reconstruction strategies. Pathway Tools (Karp et al., 2015) is a primarily web-based platform for numerous functional analyses based upon a custom ‘omics data format called a Pathway/Genome Database (PGDB), which can be used for both FBA and querying available metabolic capacity. KBase (Arkin et al., 2018a) is an online workspace for hosting scientific analyses on ‘omics datasets, and it contains apps for running existing metabolism software (such as DRAM, ModelSeed and Rast) on uploaded data. Both gapseq (Zimmermann et al., 2021) and KEMET (Palù et al., 2022) were designed to produce more accurate metabolic models by incorporating a gap-filling process into their model generation workflows, and their pathway prediction capabilities are a side effect of this strategy. Gapseq achieves this via a novel linear programming algorithm and by utilizing a highly-curated reaction database, while KEMET uses pathway prediction results for updating the metabolic models that it creates by internally running CarveMe (Machado et al., 2018).

A number of challenges affect the estimation of metabolic potential and the available tools differ in how they address these. Missing gene annotations, which can arise due to incomplete input sequencing data or poorly-matching annotation models built from limited reference databases, can cause false negatives – that is, underestimation of pathway completeness. Some tools, especially those for metabolic modeling (Machado et al., 2018; Palù et al., 2022; Zimmermann et al., 2021; Aite et al., 2018), use a process called 'gap-filling' to make educated guesses on which missing enzymes should be present and artificially add those back into the model. Pathway prediction tools (Zhou et al., 2022), on the other hand, often allow some genes in the pathway to be missing, using completeness score thresholds (usually around 75-80%) to determine pathway presence. They typically enable the user to adjust these thresholds according to their research question or data quality. Some software (Ye and Doak, 2009; Geller-McGrath et al., 2023) instead use math or machine learning to estimate the likelihood of pathway presence given the possibility of missing annotations (however, note that this doesn't necessarily eliminate the bias from incomprehensive reference databases, if these are used for training the models).

The opposite problem is overestimation of pathway completeness, which can happen due to extensive interconnectedness and overlap between metabolic pathways (i.e., sharing of multi-purpose enzymes between pathways) or incorrect homology-based annotation of paralogous proteins with different biological functions. This is a difficult problem to address without access to experimental validation data, and most reconstruction tools do not attempt to do so. Only MinPath implements a parsimonious solution for minimal pathway estimation from annotation (Ye and Doak, 2009).

Another issue with metabolism prediction is the relative lack of diversity represented in the sources of metabolic pathway definitions. Most repositories are biased towards well-studied metabolic pathways and their versions in model organisms, thus excluding alternatives and novel metabolisms from lesser-known clades of life (Ye et al., 2005). There are limited options for pathway definitions and many tools rely on freely-available databases like KEGG (Kanehisa, 2017) or MetaCyc (Caspi et al., 2014). These are highly-curated repositories that are excellent for summaries of general metabolism, but are consequently slow to be updated and may not comprehensively cover less-studied pathways, especially unusual metabolisms from across the vast diversity of microbial life. One solution to accommodate the study of these pathways is to allow software users to define their own pathways according to their expertise and research system, yet almost no tools have implemented this so far. DRAM (Shaffer et al., 2020), for instance, permits users to submit pathways for validation and curation by the developer team, but not direct estimation on user-defined pathways.

Finally, while estimation of metabolic potential from individual genomes has been implemented by all existing software for metabolism reconstruction, estimation from metagenomes still poses a challenge. Metagenomes contain multiple populations, and differentiating between these by binning individual populations into MAGs is time-consuming, error-prone and not always comprehensive (Chen et al., 2020). Most tools do not address this problem directly, instead relying on the user to process their metagenomes and input individual MAGs. METABOLIC (Zhou et al., 2022) goes one step further by allowing read-recruitment from metagenomes to the genomes used for estimation, but this still represents a reference-based analysis that will miss pathways not present in the input genomes. Similarly, community-level metabolic modeling is typically done by combining individual GSMMs (selected and/or quantified by read recruitment from a metagenome) into one model (Beura et al., 2022; Karp et al., 2022; Diener et al., 2020). Though some software (for instance, MinPath (Ye and Doak, 2009)) can technically process metagenomes simply by pooling all of their gene annotations (effectively treating a metagenome as one large genome), this strategy is only appropriate in certain use-cases as it tends to overestimate pathway completion. A far better solution for metagenome-level analyses is to estimate pathway redundancy. PathwayTools (Karp et al., 2015) does this by using read recruitment to calculate abundances of MetaCyc pathways within a given metagenome; however, no existing tool can calculate a redundancy metric (such as pathway copy number) directly from metagenome assemblies.

To address some of these challenges, I implemented a novel software framework for pathway prediction from genomes and metagenomes. In addition to being the first metabolism reconstruction tool that can estimate both completeness and redundancy (copy number) metrics for pathways, it also enables estimation on user-defined pathways.

1.5 A new open-source framework for microbial metabolism prediction in anvi'o

My implementation of metabolism reconstruction is designed to enable diverse and flexible analyses of metabolism to answer a variety of biological questions. The framework can be run on genomes, on entire metagenome assemblies, on each individual contig in an assembly, or on lists of enzymes, thereby supporting investigations at different levels of biological organization. It can derive metabolic pathway definitions from the commonly-used and highly-curated KEGG MODULE database (Kanehisa, 2017; Kanehisa et al., 2023) for analysis of general metabolism, and it can accept user-defined pathways with arbitrary functional annotation sources for more specialized questions. The tool implements two strategies for interpreting these pathway definitions, which are each suitable for different applications. In addition to providing completeness and copy number estimates for pathways, it can summarize a variety of useful auxiliary information – pathway substrates and products, gene coverages, unique or shared enzymes, and more – in various customizable output files. Furthermore, it can be used either as a standalone command-line program, or as a software library for development of other Python programs.

Beyond its flexibility, this novel metabolism reconstruction tool offers a number of benefits stemming from its design. As an object-oriented framework, it is easily extensible with new features and analysis capabilities. It is also open-source, which makes its methodology transparent, improves reproducibility, and enables community input for more targeted and rapid development. Finally, it is integrated within a larger software platform, enabling quick and easy access to supplementary analyses and data – for instance, integration with several annotation programs allows the use of multiple enzyme databases for defining and predicting metabolic pathways.

I chose to implement this framework as part of anvi'o (Eren et al., 2021b), a well-established and versatile software for analysis and interactive visualization of 'omics data, to benefit from its unique modular architecture and integration with related analysis capabilities on the same platform. Developed as an alternative to predefined 'omics workflows that allow little flexibility in the investigative strategies offered to researchers (Eren et al., 2015), anvi'o is a software ecosystem that supports interactive and fully integrated access to state-of-the-art 'omics strategies including genomics, genome-resolved metagenomics and metatranscriptomics, pangenomics, metapangenomics, phylogenomics and microbial population genetics. It is a community-driven software platform that currently stands upon more than 90,000 lines of open-source code, and it has been continuously growing and improving since its initial release. Anvi'o differs from existing bioinformatics software due to its modular architecture, which enables flexibility, interactivity, reproducibility and extensibility. To achieve this, the platform contains more than 100 interoperable programs, each of which performs individual tasks that can be combined to build new and unique analytical workflows. Anvi'o programs generate, modify, query, split and merge anvi'o projects, which are in essence a set of extensible, self-contained SQLite databases. The interconnected nature of anvi'o programs that are glued together by these common data structures yields a network (http://merenlab.org/nt) rather than predetermined, linear paths for analysis.

The metabolism reconstruction framework makes use of this modularity to separate logical steps in the reconstruction workflow into different programs for greater control of the analysis. For instance, annotation of input data is done independently of pathway prediction, enabling the user to adjust annotation parameters as needed or import custom annotations before running the estimation program. This separation also makes the prediction step quite fast, and the estimation program can be run repeatedly on the same data with different parameters. With the independent programs for creating local metabolism databases, multiple versions of these datasets (i.e., different versions of KEGG) can be generated for use in different contexts. Indeed, the interoperability of these programs with anvi'o data structures supports this modularity and makes collaboration easy. For example, one user can run various annotation steps and share the resulting database to another for pathway prediction, or users can share their metabolism databases with colleagues who wish to use the same data. The integration of this framework within anvi'o also provides the user with access to a variety of other programs, some of which can be used in conjunction with metabolism reconstruction. This includes the ability to interactively visualize predicted metabolic capacity.

This framework was designed with the goal of empowering scientists to explore and understand their data in the context of metabolism. In alignment with this goal, the software has been extensively documented online to increase its usability (for instance, `https://anvio.org/help/main/programs/anvi-estimate-metabolism/`). Subsequent chapters will further describe its functionality as well as its application to the study of microbial metabolism in several contexts, including the gut microbiome in inflammatory bowel disease.

1.6 Summary of thesis topics

This thesis details my implementation of a novel computational framework for metabolism reconstruction from ‘omics data, and showcases its application to the study of microbial ecology in the human gut environment as well as in the marine environment. It also showcases my work in supporting advanced ‘omics analyses and open science through the development of educational resources and a data integration strategy for collaborative research.

In Chapter 2, I describe the technical details of this software – including its architecture, data types, features, and calculation strategies. In Chapter 3, I introduce a study of colonization in the human gut, which uses fecal microbiota transplantation (FMT) as an *in natura* experimental model and applies my software to investigate the metabolic determinants of microbial survival in a new environment. This study proposes the concept of metabolic independence – the capacity for the production of critical metabolites that renders a microbe self-sufficient – and demonstrates its association with microbial colonization and fitness in the gut. It also suggests that high metabolic independence could determine the survival of gut microbes under the stress induced by host disease states, an idea that is further explored in Chapter 4 with a study of inflammatory bowel disease. A meta-analysis of hundreds of publicly-available human gut metagenomes using the metabolism reconstruction framework confirms high metabolic independence as a driver of microbial resilience in the IBD gut environment, and indeed as a marker of gut stress in general. It also demonstrates the power of combining pathway redundancy metrics with integrated normalization techniques for large-scale metagenomic data analyses. Finally, Chapter 5 features aspects of my work concerning the marine microbiome, which also demonstrate how the resources that I have developed facilitate advanced ‘omics research by other scientists (irrespective of their system or environment of interest). These include a recent study characterizing the high abundance of heterotrophic, nitrogen-fixing bacteria in the surface oceans; an example tutorial on targeted metagenomic binning in publicly-available polar metagenome samples; and a description of a framework for easily sharing reproducible, extensible, and integrated ‘omics datasets.

Overall, this dissertation offers insight into the metabolic determinants of microbial resilience in the human gut environment and into microbial lifestyle in the global oceans, facilitated by the development of novel computational methods for the analysis of metabolism in ‘omics datasets.

CHAPTER 2

RECONSTRUCTION OF MICROBIAL METABOLISM FROM GENOMES AND METAGENOMES IN ANVI'O

2.1 Introduction

There are two key gaps in the existing software landscape for metabolism reconstruction – the ability to predict pathways directly from metagenomic assemblies, thereby making high-throughput and non-reference-based analyses of large-scale data computationally tractable, and the ability to estimate on user-defined pathways rather than being limited to existing databases which are slow to include rare or novel metabolisms. My software framework implements both of these, in addition to offering numerous other features to support diverse and flexible investigations of metabolism with 'omics data.

In this chapter, I will discuss the technical workings of this software, starting from a summary of the data needed to estimate metabolism: enzyme annotations and pathway definitions. Then I will cover each step of the metabolism reconstruction workflow, from setting up the necessary local databases to running the estimation program. There are a variety of available output options, which will be discussed along with a strategy for visualizing the results as an interactive heatmap. At the end of the chapter, I will summarize a few opportunities for future improvement of the framework.

This framework is implemented as part of anvi'o (Eren et al., 2021b), and to make use of it, users can install anvi'o by following the instructions located at `https://anvio.org/install/`. It has been included in anvi'o since version 7.0, but continuously developed since then, such that a number of the features described herein (including user-defined metabolism) are available only in anvi'o version 8.0 (the latest release at the time of writing).

Extensive documentation for this framework is available online as part of the anvi'o 'help' pages at `https://anvio.org/help/main/`, which is a wiki describing each anvi'o-related concept (referred to as ‘artifacts’) and program on its own webpage. Much of the text in this chapter is derived from these pages, and I will link to each relevant page in the following sections. Since the documentation is included as part of the anvi’o codebase and is updated along with the code, following those links will eventually yield more up-to-date documentation than is provided in this thesis. The current version of the documentation, which applies to anvi’o version 8.0, is preserved at `https://anvio.org/help/8/`.

A basic familiarity with anvi’o, particularly its data structures, may be helpful in understanding this chapter. To briefly summarize the most critical concepts: anvi’o consolidates ‘omics data in the form of SQLite databases that serve as repositories of information that would otherwise be spread across multiple files of different formats. Sequence data and related information such as gene calls, functional annotations, and sequence statistics are stored in the ‘contigs database’, while read recruitment information such as coverage, detection, single-nucleotide variants, and indels are stored in the ‘profile database’. Anvi’o programs, including those in the metabolism reconstruction framework, operate on these databases. In fact, the metabolism workflow adds yet another database to the list – the ‘modules database’, which stores information on metabolic pathways. More detail can be found in the anvi’o help pages for interested readers.

2.2 Metabolism data: enzymes and pathways

Enzyme annotations and pathway definitions are the two basic data requirements for metabolic pathway prediction. The definitions describe each metabolic pathway in terms of the enzymes required to catalyze each reaction in the pathway, and the annotations indicate which of these enzymes are available to the organism(s) contributing to the input sequences. Matching these annotations to the pathway definitions enables the calculation of metrics like completeness and copy number. Therefore, a well-integrated source of both enzyme and pathway data is necessary for successful metabolism reconstruction efforts.

The KEGG resource is a commonly-used repository of metabolism information that provides both of these data types (Kanehisa, 2017; Kanehisa et al., 2023). It is manually curated with information reported in published scientific literature with a focus on model organisms, and updated fairly often (`https://www.genome.jp/kegg/docs/updnote.html`). As a system that interlinks various levels of biological information – from genes, gene families, and entire genomes to metabolic compounds, reactions, and pathways – it has set the standard for data integration in systems biology and bioinformatics. The two databases within KEGG that are most useful for pathway prediction are the protein family database, KEGG Orthology, and the database of metabolic pathway modules, KEGG MODULE. The KO database includes Hidden Markov model (HMM) profiles that can be used for annotating enzymes within sequencing data (Aramaki et al., 2020), while the KEGG MODULE database describes sets of genes that collectively characterize a metabolic pathway (referred to as a 'pathway module') or a phenotypic feature (referred to as a 'signature' module). Importantly, KEGG modules are defined in terms of KEGG Orthologs (KOs), which enables metabolic reconstruction via the simple matching of KO identifiers (Kanehisa, 2017).

Another database, MetaCyc (Caspi et al., 2014), is a potential alternative to KEGG. In addition to being highly-curated, it is extremely comprehensive and contains far more metabolic pathways than the KEGG MODULE database does (Altman et al., 2013). However, whereas KEGG has a straightforward correspondence between its module definitions and enzyme families and also provides the means to annotate these enzymes, MetaCyc's data requires multiple levels of iterative matching from pathways to reactions to enzymes and then to annotated genes. Furthermore, since MetaCyc does not include a means of annotating genes with these enzymes, there is no straightward link between its internal enzyme identifiers and externally-sourced gene annotations. PathwayTools, a software relying upon MetaCyc data for metabolism reconstruction, matches enzymes to reactions using their Enzyme Commission (EC) numbers, Gene Ontology (GO) terms, or as a last resort, their names, which requires a comprehensive search function to be implemented (Karp et al., 2015). KEGG therefore represents a more accessible metabolism resource for the purposes of pathway prediction.

As a result, I initially implemented my metabolism reconstruction framework to rely on KEGG data, specifically the KOfam profiles for enzyme annotation and KEGG MODULE for pathway definitions. The format of these data, as they have been implemented or provided by KEGG, will be detailed further in the next subsections. However, as discussed, KEGG is not comprehensive and it quickly became apparent that we could not exclusively use KEGG as a resource. To open the way for scientists to drive their own research questions with their unique expertise, I implemented user-defined metabolism so that users could create their own pathway definitions for estimation. These definitions currently utilize KEGG's format, but are not limited to using KOs as component enzymes. Rather, they can make use of arbitrary functional annotation sources, which is also described in a subsequent subsection.

For the metabolism reconstruction workflow to efficiently access enzyme and pathway data, it must store them locally on the user's computer. Thus, the first step in this workflow is the download and pre-processing of enzyme annotation profiles and KEGG module files. Pathway data from multiple files is consolidated into an SQLite database called the 'modules database' for easy querying and versioning. The technical details of this step are discussed in a later section.

2.2.1 Enzyme annotations

Two commonly-used methods for annotating genes are sequence alignment-based annotation (i.e., the method implemented by BLAST (Altschul et al., 1990)) and probabilistic Hidden Markov models (HMMs). Both rely on the homology, or conserved sequence structure, of proteins with similar function, and require the use of thresholds to differentiate between strong (likely) and weak (unlikely) matches, or 'hits', between the query gene sequence and the reference sequence or model (Loewenstein et al., 2009).

The KEGG Ortholog database provides HMM profiles (KOfams) for gene annotation (Aramaki et al., 2020). The enzyme families in this database are identified with a K number – an accession beginning with a ‘K’ and containing 5 unique integers – which is how they can be matched to pathway definitions in the KEGG MODULE database. Each profile has an associated sequence similarity score (bit score) threshold for distinguishing between hits of various likelihoods. These thresholds can either apply across the entire gene sequence, or within a given domain. They are used for eliminating weak hits, thereby ensuring high confidence in the annotation assigned to each gene. Note that KEGG’s thresholds tend to be rather conservative, occasionally leading to the removal of valid annotations from genes that are just different enough from the sequences used to generate the profile for their bit score to fall below the pre-computed threshold. A heuristic for restoring these lost hits will be discussed in a later section.

There are a number of databases offering profiles for enzyme family annotation beyond KEGG, including NCBI COGs (Galperin et al., 2021) and Pfam (Mistry et al., 2021). Furthermore, anyone with a set of homologous gene sequences can create an HMM for annotating other sequences in this family; for instance, by using the HMMER program ‘hmmbuild‘ (Eddy, 2011). Anvi’o allows users to annotate genes with arbitrary HMM profiles using the program ‘anvi-run-hmms‘ and the ‘-H‘ flag, and adding the ‘--add-to-functions-table‘ parameter allows these annotations to be stored as functions in the contigs database (as described on the page `https://anvio.org/help/main/programs/anvi-run-hmms/#adding-hmm-hits-as-a-functional-annotation-source`). Importing arbitrary functional annotations is also possible with ‘anvi-import-functions‘. For the user-defined metabolism feature in anvi’o, this enables any enzyme family to be used for defining metabolic pathways, as long as it can be labeled with an accession or other identifier for matching the annotation to the pathway definition.

2.2.2 KEGG pathway definitions

Since a metabolic pathway is simply a sequence of chemical reactions for converting an initial substrate compound into a final product compound, it can be defined in terms of the enzymes that catalyze each reaction (though spontaneous reactions that do not require a catalyst can be ignored for the purposes of metabolism estimation). The KEGG MODULE database does exactly this. Its module definitions are strings of KEGG Ortholog (KO) identifiers, each representing an enzyme family in the order of their corresponding reactions in a given pathway. The strings are formatted to distinguish between individual reactions (with spaces), alternative enzymes for the same reaction (with commas), essential components of an enzyme complex (with plus signs), and non-essential complex components (with minus signs). Parentheses are used to maintain the order of operations and thereby distinguish complex steps and branch points in the pathway.

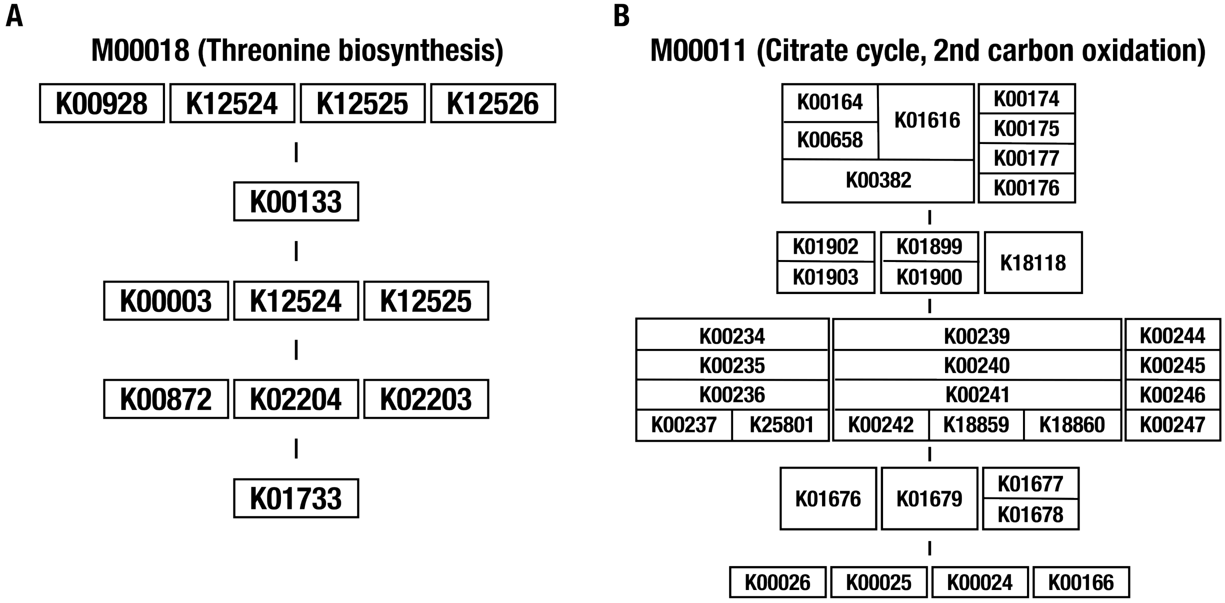


Figure 2.1: Pictorial representations of example metabolic pathways from the KEGG MODULE database. Boxes represent enzymes and are labeled with KO identifiers, with groups of adjacent boxes representing enzyme complexes. Images reproduced from the KEGG website. a) M00018, the KEGG module for the Threonine Biosynthesis pathway. b) M00011, the KEGG module for the second carbon oxidation phase of the citrate cycle.

KEGG's definition strategy is better understood when comparing the string definition to its pictorial representation. Figure 2.1a shows the example of module M00018, Threonine Biosynthesis. The definition string for this module is:

```
(K00928,K12524,K12525,K12526) K00133 (K00003,K12524,K12525)
(K00872,K02204,K02203) K01733.
```

This biosynthesis pathway has five major steps, or chemical reactions (in documentation, I refer to these major steps as 'top-level steps'). The first reaction in the pathway requires an aspartate kinase enzyme (also known as a homoserine dehydrogenase), and there are four possible orthologs known to encode this function: K00928, K12524, K12525, or K12526. Only one of these genes is required for an organism to be able to catalyze this step in the pathway, which is indicated by the separation of their KO identifiers with commas. In contrast, the second reaction can be fulfilled by only one known KO, the aspartate-semialdehyde dehydrogenase K00133. A more complicated example is shown in Figure 2.1b for module M00011, the second carbon oxidation phase of the citrate cycle. The definition string for this pathway is as follows:

```
(K00164+K00658+K00382,K00174+K00175-K00177-K00176)
(K01902+K01903,K01899+K01900,K18118)
(K00234+K00235+K00236+K00237,K00239+K00240+K00241-
(K00242,K18859,K18860),K00244+K00245+K00246-K00247)
(K01676,K01679,K01677+K01678) (K00026,K00025,K00024,K00116)
```

This pathway also has five steps, but this time, most of the reactions require an enzyme complex. Each KO within a multi-KO box is a component of a larger enzyme. For example, one option for the first reaction is 2-oxoglutarate dehydrogenase, a 3-component enzyme made up of K00164, K00658, and K00382. Yet, not all of the enzyme components are equally important. In the definition string the KO components of an enzyme complex are connected with either

‘+’ signs or ‘-‘ signs. The ‘+’ sign indicates that the following KO is an essential component of the enzyme, while the ‘-‘ sign indicates that it is non-essential. When estimating metabolism, ‘non-essential’ components can be ignored – that is, a reaction is considered to be possible if all its essential component KOs are annotated. For example, the first step in this pathway would be complete if just K00174 and K00175 were present. The presence or absence of either K00177 or K00176 would not affect the module completeness score at all.

Module definitions can be even more complex than this. Both of these examples have exactly five top-level steps, regardless of the KOs used to fulfill each reaction. However, in some modules, there can be branch points, or alternative sets of reactions and enzymes, with different numbers of steps. In addition, some modules (such as M00611, the module representing photosynthesis), are defined by other modules, in which case they can only be considered complete if their component modules are complete.

In KEGG, module data is provided in the form of flat text files containing the module definition as well as a variety of other information about the pathway, including its name, classification, enzymes, reactions, compounds, and any references that describe it. These files are parsed in the first step of the metabolism reconstruction workflow to establish the modules database for easy data access in later steps.

2.2.3 User-defined pathways and their enzymes

The current implementation for user-defined metabolism in anvi’o makes use of KEGG’s format for module definition strings. That is, to define a pathway, users must write a definition by putting enzyme accessions in the order of their corresponding reactions in the pathway. Different steps (reactions) in the pathway should be separated by spaces, and alternative enzymes that can catalyze the same reaction should be separated by commas, with parentheses to distinguish between alternatives with multiple steps. For enzyme complexes, all components should be in one string, with essential components separated by ‘+’ signs and non-essential components separated by '-' signs. The final definition should be placed into a flat text file describing the module, following the same format that is used by KEGG for its modules to make use of the same functions for processing this data into a database format.

An important difference between user-defined and KEGG pathways is that user-defined modules can make use of enzymes from arbitrary functional annotation sources, not just KOfam. As long as the enzyme identifier used in the pathway definition is the same as the accession that will be stored in the functions table of the contigs database when running gene annotation software, the metabolism estimation program will be able to match between the two. However, the source of each enzyme's annotation – whether that is KOfam, Pfam, NCBI COGs, or a custom source – must be indicated in the flat text file.

A detailed tutorial on generating user-defined metabolic pathways can be found at `https://anvio.org/help/main/artifacts/user-modules-data/`, and a description of the flat file format can be found on the page `https://anvio.org/help/main/programs/anvi-setup-user-modules/`.

2.3 The metabolism reconstruction workflow in anvi'o

Following the modular architecture of anvi'o, I implemented the metabolism reconstruction framework as a set of interoperable programs that perform independent tasks (Figure 2.2). For a first-time user, the typical workflow for using these programs will 1) set up function profiles and SQLite database(s) of metabolism data on the user's computer; 2) annotate input contigs with HMM hits to these function profiles; and 3) match functional annotations to pathway definitions in the database to estimate the completeness and copy number of each metabolic pathway in the database within each input genome, metagenome assembly, or metagenomic contig. Step (1) only has to be run once (unless updates to the local databases are desired), and later iterations of the workflow only require steps (2) and (3). The last step produces customizable tab-delimited output files containing various data about each metabolic pathway or functional annotation - including pathway completeness, pathway information like category and metabolites, accession IDs and locations of the genes that contribute to the pathway, read-recruitment coverage and detection of these contributing genes, and more.

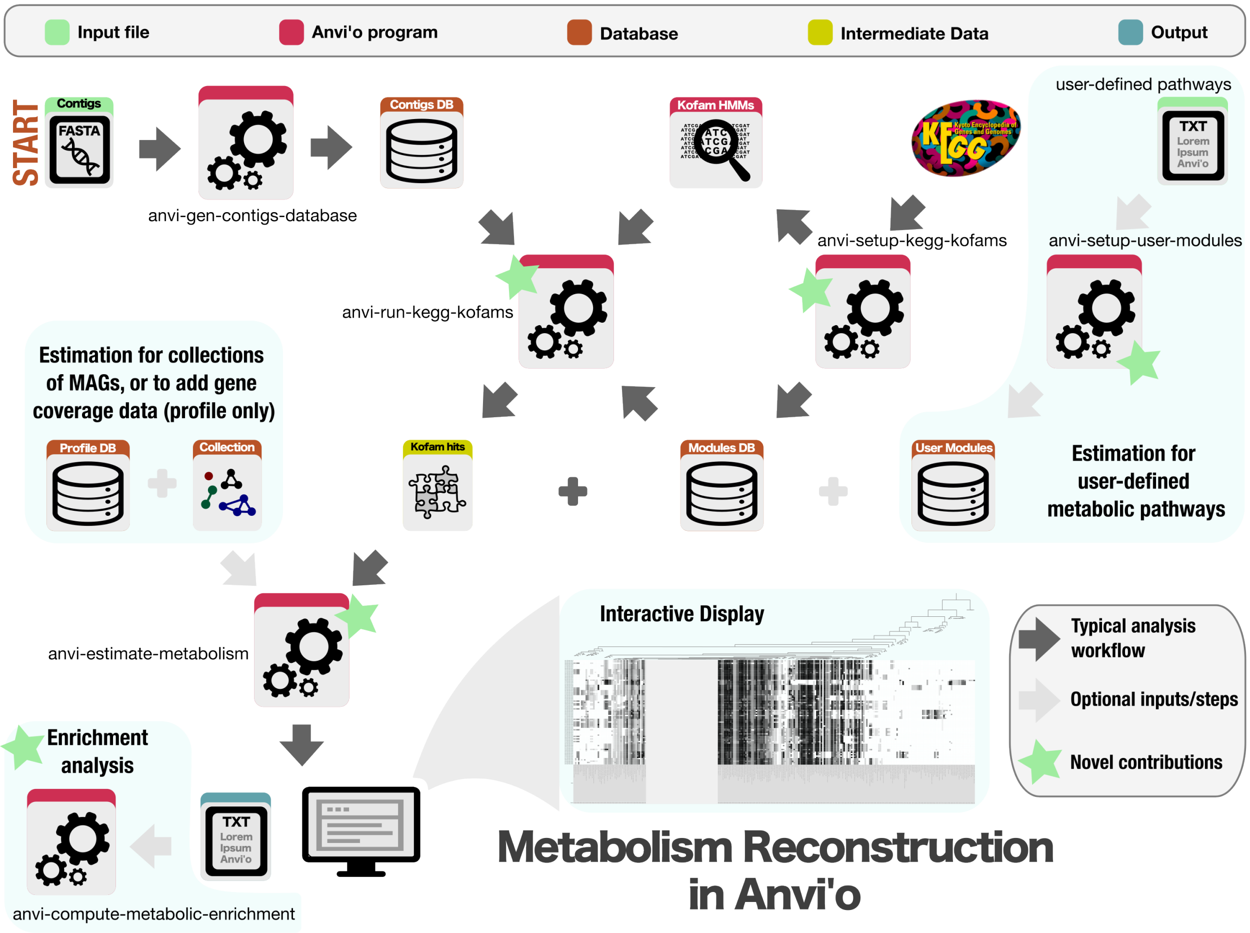


Figure 2.2: A schematic of the metabolism reconstruction framework within anvi'o. Dark arrows show the flow of a typical analysis workflow while light arrows indicate optional inputs or steps. Green stars highlight the novel programs implemented as part of this dissertation work.

Though this section describes the use of this framework via its standalone command-line programs, the methods and classes utilized by these programs are also accessible to Python developers wishing to incorporate them into their own code. By loading the anvi'o package and its ‘kegg‘ library, developers can access its functions for metabolism reconstruction as well as its internal data structures for metabolism information.

To begin any analysis within anvi'o, users first convert their input sequence data into an anvi'o-compatible format: the contigs database, in which sequences, gene calls and their functional annotations can be stored. The program ‘anvi-gen-contigs-database‘ generates this database from a FASTA file containing contig sequences. It also stores gene calls, by default running the gene-calling software Prodigal (Hyatt et al., 2010), or alternatively incorporating gene calls provided by the user. This is not a part of the metabolism reconstruction framework, but simply a prerequisite step to using it.

There are two programs in the framework that perform the step of data setup. ‘anvi-setup-kegg-kofams‘ downloads metabolism data (both enzyme annotation profiles and metabolic modules) from the KEGG website and pre-processes this data to make it readily accessible to later programs. ‘anvi-setup-user-modules‘ runs a similar pre-processing step on user-defined module files. Both of these programs generate a modules database from which pathway information can be queried.

The next step in the workflow is enzyme annotation. There are multiple programs for running (or importing) gene annotation in anvi'o, and which of these should be used is determined by which enzyme sources define the metabolic pathways. Since KEGG modules are defined by KEGG KOfams, I will discuss the program ‘anvi-run-kegg-kofams‘. Other programs may be necessary for running this workflow on user-defined pathways, and these will work in a similar fashion.

Finally, the program ‘anvi-estimate-metabolism‘ performs the metabolism reconstruction step by matching enzyme annotations to pathway definitions and calculating both a completeness score and a copy number for each pathway. It uses two different strategies for parsing the pathway definitions, which results in two values for each metric. The first is the ‘pathwise’ strategy, which considers each possible combination of enzymes that could be used to make the module complete. The second is the ‘stepwise’ strategy, which aggregates all possible alternative enzymes into major steps.

The output from ‘anvi-estimate-metabolism‘ can be sent to downstream programs for further analysis. Programs within the anvi'o network that can work with this data include ‘anvi-interactive‘ (which can visualize matrix-formatted metabolism output as heatmaps) and ‘anvi-compute-metabolic-enrichment‘ (which can use pathway completeness to compute enrichment scores for modules in groups of samples) (Shaiber et al., 2020a); however, these plain-text output files are designed to be both readable and easily parsable by ad-hoc scripts written by users.

2.3.1 anvi-setup-kegg-kofams

The program ‘anvi-setup-kegg-kofams‘ downloads metabolism data from the KEGG website and pre-processes them for later access by other programs. It generates a directory of KOfam profiles for enzyme annotation by ‘anvi-run-kegg-kofams‘ as well as the modules database containing definitions and auxiliary information for all KEGG modules. The webpage https://anvio.org/help/main/programs/anvi-setup-kegg-kofams/ serves documentation for this program. A description of the data directory that it creates can be found at https://anvio.org/help/main/artifacts/kegg-data/, and a description of the modules database can be found at https://anvio.org/help/main/artifacts/modules-db/.

Data download

KOfam profiles are downloadable from KEGG's FTP site (ftp://ftp.genome.jp/pub/db/kofam/) and all other KEGG data is accessible as flat text files through the KEGG API (https://www.kegg.jp/kegg/rest/keggapi.html). This program downloads a compressed archive of individual HMMs (one for each KOfam), a file describing the bit score thresholds and enzyme name for each KOfam (‘ko_list.txt‘), and one file for each pathway in KEGG MODULE. These collectively are saved into a directory on the local computer, henceforth referred to as the KEGG data directory, where they are processed.

Processing of KOfam profiles

To prepare the KOfam profiles for later homology searches, individual profiles are concatenated into one file, which is then indexed using the HMMER program ‘hmmpress‘ (Eddy, 2011). Only KOs which have a corresponding bit score threshold defined as described in (Aramaki et al., 2020) are included in the final set of concatenated profiles for annotation; those without such thresholds are moved to an alternate directory. This is done because the annotation program relies on bit scores for the removal of weak hits. The file containing these bit scores is maintained in the KEGG data directory for later use by ‘anvi-run-kegg-kofams‘.

Processing of module files

KEGG’s module files contain a wealth of information in a consistent, if unorthodox, format. This program parses the files to extract the information and store it in the modules database so that it is more easily (programmatically) accessible.

Each module file contains data related to a given pathway, including its identifier (‘ENTRY’), name (‘NAME’), definition string (‘DEFINITION’), enzyme information like name and EC number (‘ORTHOLOGY’), categorization (‘CLASS’), and its component reactions and compounds. The text file is formatted such that the initial column in a line describes what type of data is contained in the line (subsequent lines containing the same data type are not labeled), the second column contains that information or (in the case of data that can be linked to other KEGG databases, such as ‘ORTHOLOGY’) a KEGG identifier. In the latter case, the line also contains a third column with additional information, usually the name corresponding to the identifier in the second column. This format is human-readable, but not straightforward to programmatically parse.

The consequence of this is that ‘anvi-setup-kegg-kofams‘ has rigid expectations for the format of the KEGG data that it downloads. Extensive sanity checks were implemented for the parsing function, but future updates to KEGG may alter the formatting such that this data can no longer be directly downloaded. To mitigate this issue, I have built several archives, or ‘snapshots’, of the KEGG database that are already converted into an anvi’o-compatible format, which are downloadable directly by this program and can be found in the online data repository Figshare (`https://figshare.com/authors/Iva_Veseli/9014558`). Available snapshots are described in a YAML file in the anvi’o Github repository.

The modules database

The modules database stores the information parsed from individual module files in a programmatically queryable format. Two tables in this SQLite database are important for metabolism reconstruction: the ‘self’ table and the ‘modules’ table.

The ‘self’ table is a component of all anvi’o databases that describes the database itself. For the modules database, this table indicates the source of the metabolism data stored within (which is ‘KEGG’ for the database produced by ‘anvi-setup-kegg-kofams‘, and ‘USER’ for user-defined data). It describes the number of modules downloaded, and which enzyme annotation sources (i.e., KOfam) are used to define them. Most importantly, each modules database is given a unique identifier that is hashed from the contents of the database, such that databases containing the same metabolism data have identical hashes. This hash is critical for managing different versions of metabolism data, for ensuring that input sequences have been annotated with a compatible version of KOfam, and especially for reproducibility.

The ‘modules’ table contains the data parsed from the module files. In this table, the ‘module’ column indicates the module identifier. Similar to the KEGG flat file format, the ‘data_name’ column indicates what type of data the row contains. These data names are usually fairly self-explanatory - for instance, the ‘DEFINITION’ rows describe the module definition and the ‘ORTHOLOGY’ rows describe the enzymes belonging to the module – however, an official explanation can be found on the KEGG help page (`https://www.genome.jp/kegg/document/help_bget_module.html`). The ‘data_value’ and ‘data_definition’ columns hold the corresponding information; for ‘ORTHOLOGY’ fields these are the enzyme accession number and its functional annotation, respectively. Just like the flat files, not all rows have a ‘data_definition’ field. Finally, some rows of data originate from the same line in a module file (typically, KOs that share the same enzyme name and commission number); these rows will have the same number in the ‘line’ column.

Default usage: downloading a KEGG snapshot

By default, this program downloads a snapshot of the KEGG databases rather than downloading data directly from KEGG every time. The default snapshot is a ‘.tar.gz‘ archive of a KEGG data directory that was (usually) generated around the time of the latest anvi’o release. After the KEGG archive is downloaded, it is unpacked, checked that all the expected files are present, and moved into the KEGG data directory.

This strategy ensures that almost everyone (with the same version of anvi’o) uses the same version of KEGG data, which is good for reproducibility and makes it easy to share annotated datasets. Furthermore, it avoids the issues associated with continuous updates. The KEGG resources are updated fairly often, and updates to their metabolic pathways can be accompanied by updates to KOfam profiles, which necessitates keeping the two in sync by re-running the gene annotation step. The data download and re-annotation steps can both be time-consuming and are typically not worth repeating over short timescales. They also introduce the issue of data incompatibility between collaborators working on different machines. Therefore, the default usage of pre-processed KEGG snapshots is a compromise in favor of efficiency and reproducibility over having the most up-to-date data at all times. Of course, users can choose to download data directly from KEGG instead, using the ‘--download-from-kegg‘ parameter as described at `https://anvio.org/help/main/programs/anvi-setup-kegg-kofams/#getting-the-most-up-to-date-kegg-data-downloading-directly-from-kegg`.

There are multiple snapshots available containing different versions of KEGG, which are hosted on Figshare and summarized in a YAML file in the anvi'o Github repository. Users can choose amongst these snapshots; if none of them suffice, they can elect to download the data directly from KEGG. More information is available on the help page for this program, https://anvio.org/help/main/programs/anvi-setup-kegg-kofams/.

2.3.2 anvi-setup-user-modules

The program ‘anvi-setup-user-modules‘ creates a modules database out of metabolic pathways that have been defined by the user, thus enabling the study of arbitrary metabolisms and allowing novel metabolic insights to be generated based on scientists' expertise and research questions. The webpage https://anvio.org/help/main/programs/anvi-setup-user-modules/servesdocumentationforthisprogram.

Data for this program is generated by the user, who can create metabolic pathway definitions and compile this information in module files. Since this program processes user module files in the same way that ‘anvi-setup-kegg-kofams‘ does, it currently requires a formatting and definition strategy for modules similar to KEGG's. A comprehensive tutorial on generating these definitions and files can be found at https://anvio.org/help/main/artifacts/user-modules-data/. To reduce the tedium of this step and avoid formatting errors, there is a script to automate the process of generating these files (https://anvio.org/help/main/programs/anvi-script-gen-user-module-file/).

Module accessions must be unique identifiers because they are used as unique keys in data structures during the metabolism estimation process and in output files. Since user-defined modules can be used in conjunction with KEGG modules, during the creation of the user modules database, some sanity checking is performed to ensure that the modules' unique identifiers do not overlap with any KEGG module identifiers.

Incorporating arbitrary annotation sources

As discussed previously, user-defined modules can make use of any enzyme that is annotatable. There are a variety of protein sequence databases providing a means of homology-based gene annotation, whether those are HMM profiles or sequences for alignment, and a number of these are already implemented within anvi'o. Currently, in addition to KOfam (Aramaki et al., 2020), anvi'o can annotate genes with NCBI COGs (Galperin et al., 2021), Pfams (Mistry et al., 2021), and dbCAN CAZymes (Yin et al., 2012). There is also a program ('anvi-run-hmms') that enables users to annotate genes with arbitrary HMM profiles, and any source of gene annotations not covered by these options can be imported into a contigs database using 'anvi-import-functions'. To distinguish between functions from different sources, each gene annotation stored in an anvi'o contigs database is accompanied by a string indicating its source.

For metabolism reconstruction to work properly on user-defined modules, there must be a way to query the contigs database for enzymes annotated from multiple sources; therefore, user module files must describe the annotation sources relevant to the enzymes incorporated in the pathway definition. The 'ANNOTATION_SOURCE' field, which is not found in KEGG module files, matches each enzyme in the definition to its annotation source so that it can be searched for during the estimation process.

2.3.3 anvi-run-kegg-kofams

'anvi-run-kegg-kofams' is one example of a program that provides enzyme annotations to be used for metabolism reconstruction. It is required for estimation on KEGG modules (which are defined by KOs). The webpage `https://anvio.org/help/main/programs/anvi-run-kegg-kofams/` serves documentation for this program.

This program depends on the metabolism data downloaded by 'anvi-setup-kegg-kofams'. It takes gene sequences from the provided contigs database, identifies matches to the downloaded KOfam profiles, and stores the resulting gene annotations into the functions table of the same database. For KOs that belong to KEGG modules, the module and its classification are also added as annotations for the relevant genes. The contigs database is then labeled with the hash of the corresponding modules database so that only compatible KEGG modules will be used for metabolism estimation later. When running this program, users have the option of specifying which KEGG data directory they want to use (in order to manage annotation with multiple versions of KEGG); however, only one set of KOfam annotations (i.e., from one version) can be stored in a contigs database at a time.

Multi-threaded, homology-based annotation with KOfam profiles

The program internally runs HMMER (Eddy, 2011) to find matches between every gene sequence in the contigs database and every KOfam profile. There are over 30,000 KOfam profiles, so this process scales quite rapidly as the number of input gene sequences increases. This program can therefore be multi-threaded when greater computational resources are available to reduce the amount of processing time. Multi-threading consists of partitioning the gene sequences into smaller sets that can each be processed by one CPU core.

Elimination of weak hits

Once HMMER returns a set of hits, the program parses these to keep only high-confidence annotations. Weak hits will by default be eliminated according to the bit score thresholds provided by KEGG (Aramaki et al., 2020); that is, hits with bit scores below the threshold for a given KO profile will be discarded. The bit score thresholds are accessed from the ‘ko_list.txt‘ file downloaded from KEGG by ‘anvi-setup-kegg-kofams‘.

The user has the option to forgo this processing and keep all hits regardless of bit score. Bit scores for each annotation are not typically saved in the contigs database, but there is also an option to log these values as a text file.

A heuristic for relaxing stringent bit score thresholds

Due to the nature of homology-based annotation with HMMs, in which profiles are built from a set of reference sequences and designed to match with highly similar sequences, some KOfam bit score thresholds can be too stringent for hits from slightly more distant sequences in the same family to pass. This occasionally leads to the elimination of valid annotations with bitscores falling just below the pre-computed threshold.

To mitigate this problem, I implemented a heuristic for adding back weaker annotations when there is high confidence that these hits were eliminated due to a conservative bit score threshold and not because the gene in question represents a different protein family. This heuristic effectively relaxes the bit score threshold to a fraction of its pre-computed value, and uses an alternative measure of match probability, the hit's e-value, to keep a set of highly-probable matches for the gene. All of these matches must resolve to the same KO family for there to be high-confidence that annotating the gene with this KO would be valid.

Put more formally: for every gene without a KOfam annotation, we examine all the hits with an e-value below 'x' and a bit score above 'y' percent of the relevant KO's bit score threshold. If those hits are all to a unique KOfam profile, then the gene call is annotated with that KO. 'x' and 'y' are modifiable parameters, but by default the e-value threshold ('x') is 1e-05 and the bitscore fraction ('y') is 0.5.

This annotation heuristic is applied by default in a given run of 'anvi-run-kegg-kofams', but can be turned off (e.g., to use only KEGG-provided bit score thresholds for managing annotation quality) by providing the '--skip-bitscore-heuristic' flag.

2.3.4 anvi-estimate-metabolism

The program 'anvi-estimate-metabolism' is the workhorse of the metabolism reconstruction framework. It predicts the metabolic capabilities of organisms based on their genetic content by matching enzyme annotations in input sequence data to metabolic pathway definitions, and by computing completeness scores and copy numbers for each pathway. The webpage `https://anvio.org/help/main/programs/anvi-estimate-metabolism/` serves documentation for this program.

‘anvi-estimate-metabolism‘ relies upon gene annotations stored in a provided contigs database (i.e., by ‘anvi-run-kegg-kofams‘) or in a provided list of enzymes, as well as on the metabolism data prepared by either ‘anvi-setup-kegg-kofams‘ or ‘anvi-setup-user-modules‘. Users can choose to source pathway definitions from either or both of these modules databases.

Metabolic pathways can be complex – they are not always linear and may contain branch points, and due to biological redundancy, there may be numerous alternatives for each step in the pathway. Depending on their research goal, users may be interested in the specific set of enzymes used to catalyze each reaction, or they may simply need a summary of metabolic capacity agnostic to enzyme identity. To accommodate both of these needs, this program offers two strategies for interpreting the pathway definitions when estimating metabolic potential (Figure 2.3). There is a ‘stepwise’ strategy with equivalent treatment for alternative enzymes – i.e, enzymes that can catalyze the same reaction in a given metabolic pathway – and a ‘pathwise’ strategy that accounts for all possible variations of the pathway. These are discussed in detail below.

The program outputs metabolism reconstruction results in one or more tab-delimited text files, which are covered in the next section.

Input sequences: genomes, metagenome assemblies, and metagenomic contigs

This program typically derives gene annotations from a contigs database, which can represent different kinds of biological entities (anything that can be stored as a FASTA file). These can be individual genomes, unbinned metagenomes representing an entire community, or binned metagenomes (e.g., with metagenome-assembled genomes of individual microbial populations, or MAGs, already defined). Metabolism estimation can therefore be run in different 'modes' for proper contextualization of gene annotations in the database. When the input data contains multiple entities (i.e., several MAGs in a binned metagenome), metabolism estimation is run independently on each entity using only the gene annotations belonging to that entity.

During pathway prediction, the program considers a 'pool' of gene annotations within a given context. It loads the relevant annotations accordingly from the input contigs database. In 'genome mode', the contigs database typically contains an individual microbial genome; for instance, a reference genome downloaded from the NCBI, an assembly of an isolated population, or an individual MAG that has been split from its original metagenome assembly. In this case, all of the enzyme annotations in the database (from relevant annotation sources) are loaded at once. Note that multiple individual genomes can be provided at once to 'anvi-estimate-metabolism‘ for high-throughput processing.

If the contigs database instead contains a binned metagenome assembly, estimation can be run on each MAG individually. Gene annotations belonging to one MAG at a time are loaded from the database for use in the estimation algorithm.

Unbinned metagenomes can be processed in two different ways. In 'metagenome mode', this program will estimate metabolism for each contig in the metagenome separately. That is, gene annotations from one contig at a time will be loaded and used to calculate pathway completeness and copy numbers within that contig alone, resulting in a set of predicted pathways for each individual contig in the assembly. This will tend to underestimate module completeness because it is likely that many modules will be broken up across multiple contigs belonging to the same population. However, metabolism can also be predicted for the metagenomic community as a whole by running this program in 'genome mode' on the metagenomic assembly. This effectively treats all enzyme annotations in the metagenome as belonging to one collective genome (all annotations are loaded from the database at once), which will result in the opposite tendency to overestimate module completeness (as the enzymes will in reality be coming from multiple different populations). Nevertheless, copy number calculations are extremely relevant in this use case, as will be discussed later.

Enzyme list input

An alternative input option for this program is a list of enzyme accessions. These are provided to the program in a text file and serve as the annotation pool for metabolism estimation; that is, all enzymes in the list are considered available for catalyzing metabolic reactions. This is a more flexible input option to bypass the steps for generating and annotating a contigs database.

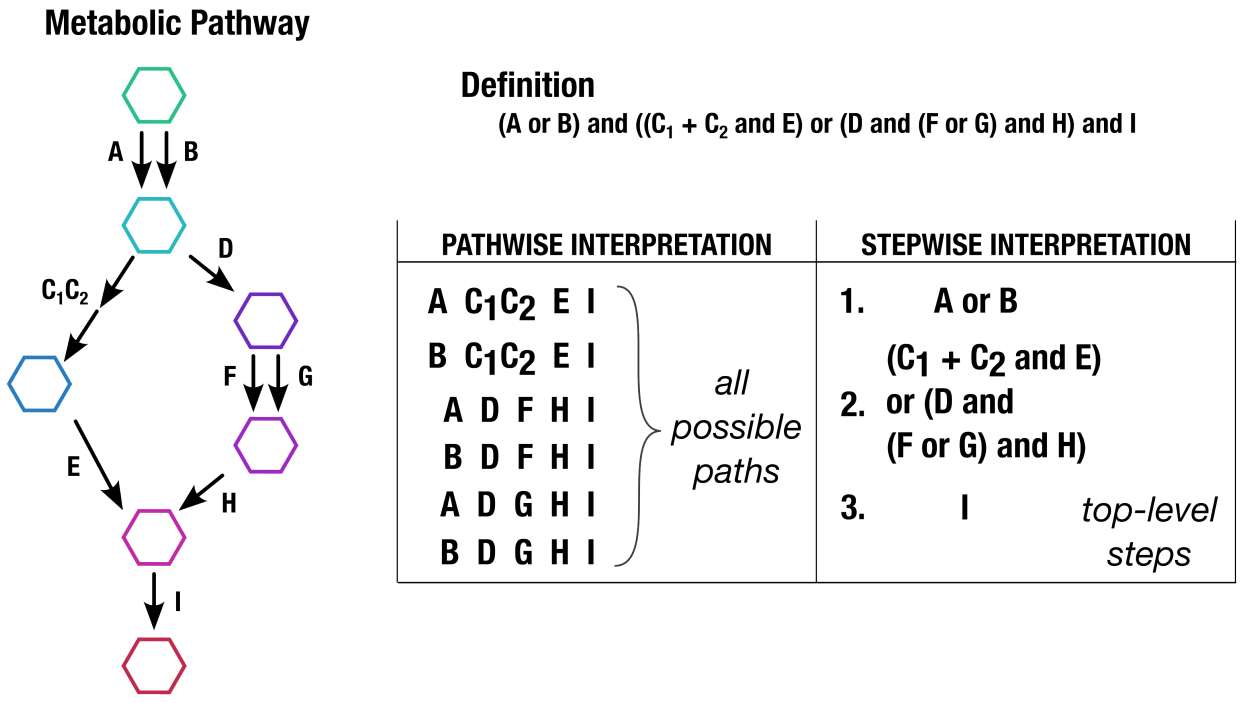
The file format for the enzyme list is described on the help page at `https://anvio.org/help/main/artifacts/enzymes-txt/`.

Completeness and copy number of metabolic pathways

The goal of ‘anvi-estimate-metabolism‘ is to compute a completeness score and a copy number for each metabolic pathway that is defined in the provided modules database(s). Completeness estimates refer to the percentage of steps (typically, reactions) in the pathway that are encoded in the genome or metagenome. Likewise, copy number summarizes the number of distinct sets of enzyme annotations that collectively encode the complete pathway.

Note that these metrics are each more appropriate for certain contexts. For individual genomes, particularly complete genomes, pathway completeness score is relevant for identifying metabolic capacity of the population, and copy number is likely to be meaningless. This is also the case for MAGs, though these tend to be less complete, composite genomes and those attributes may affect the pathway completeness scores. Lower-quality MAGs can also be contaminated, containing more than one microbial population, which could be reflected in pathway copy numbers. In metagenome assemblies (estimated via ‘genome mode’), there will be a lot of redundancy given the contributions of multiple populations to the enzyme annotation pool – copy numbers will be very relevant in this case, but completeness scores are likely to be uninformative as they will mostly tend to be artificially high. Finally, when estimating on individual contigs of a metagenome assembly, the values of both metrics will depend on the length of the contig, but will tend to be low given that typically few to no metagenomic contigs capture an entire microbial genome.

The calculation of both completeness and copy number depends on the strategy used to decompose a pathway definition into smaller parts (Figure 2.3). The 'pathwise' strategy considers all possible combinations of enzymes, and therefore the length of the pathway (the denominator of the completeness score) is set by the number of enzymes in each combination. The 'stepwise' strategy is less granular, considering alternative enzymes as equivalent contributors to the same step, which results in shorter pathway lengths (and smaller denominators). The copy number calculation is similarly affected.



Definition

(A or B) and ((C_1 + C_2 and E) or (D and (F or G) and H) and I

PATHWISE INTERPRETATION	STEPWISE INTERPRETATION
A C_1C_2 E I B C_1C_2 E I A D F H I B D F H I A D G H I B D G H I *all possible paths*	1. A or B 2. (C_1 + C_2 and E) or (D and (F or G) and H) 3. I *top-level steps*

Figure 2.3: Two strategies for interpreting metabolic pathway definitions. On the left is a theoretical metabolic pathway, in which shapes represent molecules, letters represent enzymes

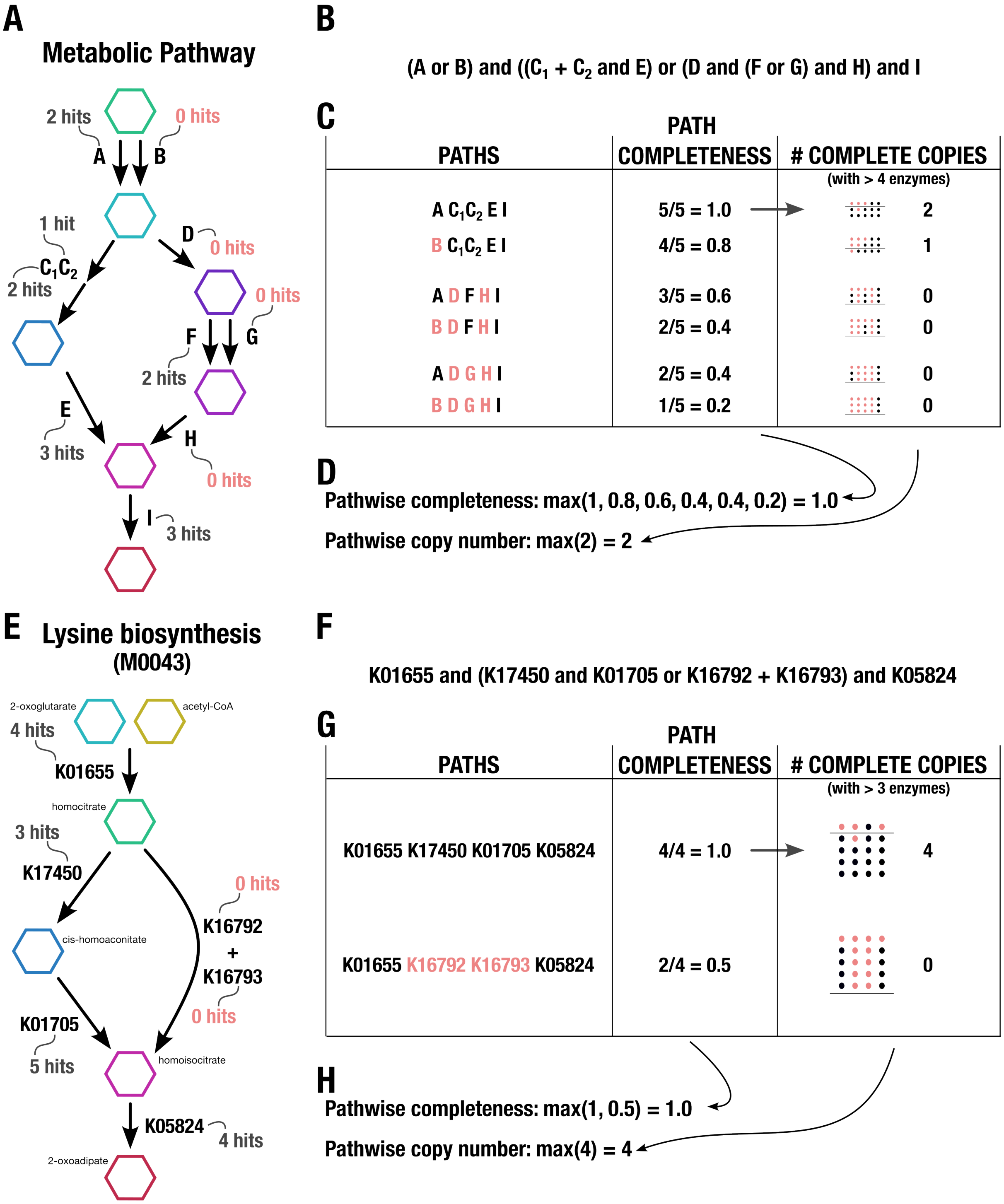
Figure 2.3 continued: (subscripts indicate enzyme components), and arrows represent reactions. The definition of the metabolic pathway is written at the top in terms of its required enzymes. The left box summarizes pathwise interpretation of the pathway definition, in which each possible path through the metabolic pathway is enumerated. The right box summarizes stepwise interpretation of the pathway definition, in which the pathway is broken down into its major, or 'top-level', steps.

Pathwise interpretation of pathways

The 'pathwise' strategy considers all possible 'paths' through the module – each alternative set of enzymes that could be used together to catalyze every reaction in the metabolic pathway. After calculating the percent completeness in all possible paths, it takes the maximum completeness to be the pathwise completeness score of the module as a whole. This is the most granular way of estimating module completeness because it considers all the possible alternatives. Similarly, path copy number is computed as the number of complete copies of a path through a module, and a module's pathwise copy number is then calculated as the maximum copy number of any of its paths that have the highest completeness score.

'anvi-estimate-metabolism' uses a recursive algorithm to 'unroll' the module definition string into a list of all possible paths. First, the definition string is split into its top-level steps (which are separated by spaces). Each step is either an atomic step, a protein complex (KO components separated by '+' or '-'), or a compound step (multiple alternatives, separated by commas). Compound steps and protein complexes are recursively broken down until only atomic steps remain. An atomic step can be a single KO, a module number, a non-essential KO starting with '-', or '--' (a string indicating that there is a reaction for which there is no available KOfam model). We use the atomic steps to build a list of alternative paths through the module definition. Protein complexes are split into their respective components using this strategy to find all possible alternative complexes, and then these complexes (with all their component KOs) are used to build the alternative paths.

Generating the set of alternative paths for each module is done once at the start of program execution to avoid repetitive processing in the case of multiple input datasets.



PATHS	PATH COMPLETENESS	# COMPLETE COPIES (with > 4 enzymes)
A C_1C_2 E I	5/5 = 1.0	2
B C_1C_2 E I	4/5 = 0.8	1
A D F H I	3/5 = 0.6	0
B D F H I	2/5 = 0.4	0
A D G H I	2/5 = 0.4	0
B D G H I	1/5 = 0.2	0

Pathwise completeness: max(1, 0.8, 0.6, 0.4, 0.4, 0.2) = 1.0

Pathwise copy number: max(2) = 2

PATHS	PATH COMPLETENESS	# COMPLETE COPIES (with > 3 enzymes)
K01655 K17450 K01705 K05824	4/4 = 1.0	4
K01655 K16792 K16793 K05824	2/4 = 0.5	0

Pathwise completeness: max(1, 0.5) = 1.0

Pathwise copy number: max(4) = 4

Figure 2.4: A demonstration of pathwise metric calculations for metabolic pathways (performed by the program ‘anvi-estimate-metabolism‘ is shown in panels a-d (for a theoretical pathway) and e-h (for a real pathway). a) Theoretical metabolic pathway, where hexagons represent metabolites, arrows represent chemical reactions, letters represent enzymes (subscripts

Figure 2.4 continued: indicate enzyme components), and the example number of gene annotation hits for each enzyme is written in gray. Enzymes with zero hits are highlighted in pink. b) The definition of the theoretical pathway from panel a, written in terms of the required enzymes. c) Table showing the paths through the module and example calculations of completeness and copy number for each path. Path completeness is calculated by taking the fraction of annotated enzymes in the path. Path copy number is calculated as the number of complete copies of the path. The number of enzymes required to have a complete copy with the default completeness threshold of 0.75 is given in smaller font. The collection of dots is a pictorial representation of the copies of each path. Each dot is an enzyme and is colored black if there is a hit for the enzyme in that copy or pink if there is no hit. The line separates complete copies of the path from incomplete copies. Thus, the height of the line is the number of complete copies of the path. d) Final calculations of completeness score (maximum completeness score taken over all possible paths through the module) and copy number (maximum copy number over all paths of highest completeness) for the theoretical metabolic pathway. e-h) Same as a-d, but for a real metabolic pathway.

Figure 2.4 demonstrates the pathwise calculations of completeness and copy number, as described in the following subsections.

Pathwise completeness

After the list of alternative paths is generated, the next task is to compute the completeness of each path in a given module. Each alternative path is a list of atomic steps or protein complexes. The program loops over every step in the path and uses the annotation pool of KOs to decide whether the step is complete (represented by a 1) or not (0). There are the following cases to handle:

1. A single KO. If this KO is annotated, then the step is complete (1).

2. A protein complex – that is, multiple KOs connected with '+' (if they are essential components) or '-' (if they are non-essential). For these steps, a fractional completeness score is computed based on the number of essential components that are present in the annotation pool. Non-essential KOs are ignored. For example, the complex 'K00174+K00175-K00177-K00176' would be considered 50% complete (a score of 0.5) if only 'K00174' were present in the annotation pool.

3. Non-essential KOs. Some KOs are marked as non-essential, with a minus sign in front of the KO identifier, even when they are not part of a protein complex. These steps are ignored for the purposes of computing module completeness.

4. Steps without associated KOs. Some reactions do not have a KO identifier, but instead there is the string '--' serving as a placeholder in the module definition. Since the genes required for these steps are not annotatable, they are always considered incomplete (0). Modules that have steps like this can therefore never be 100% complete. The program warns the user about these instances so that they can check manually for any missing steps.

5. A module. Some modules are defined by other modules. The completeness of these steps cannot be determined until the completeness of the component modules is known, so they are initially ignored. Later, the program adjusts the completeness scores for these steps according to the estimation results for the component modules.

To get the completeness score for a given path through the module, the program adds together the completeness of each essential step in the path. The resulting sum is then divided by the number of essential steps.

Once every possible path through the module has a completeness score (a fraction between 0 and 1), the maximum of all these completeness scores is taken as the completeness of the module overall. The assumption here is that the most complete set of enzymes in that pathway is the most likely to be used. This is certainly a questionable assumption, but some choices like this are necessary to summarize the data. It gets trickier to interpret this number when there is more than one path through the module that has the maximum completeness score. Identifying which path is biologically relevant requires additional data types or knowledge of the biological system.

Indeed, it is common for modules (especially those with a lot of alternative paths) to have more than one maximally-complete path. These are used later for calculating pathwise copy number, so the program stores all of the paths with the maximum completeness score (for each module).

Finally, there are some modules defined by other modules (not just enzymes). These are usually what KEGG calls ‘Signature Modules’, which are collections of enzymes that collectively encode some phenotype, rather than a typical pathway of chemical reactions. The program adjusts the completeness score of these modules after the completeness of its component modules is known. To do this, it reruns the previous tasks to recompute the number of complete steps in each path and the overall completeness of the module. This time, for ‘Module’ atomic steps (case 5), it takes that module’s fractional completeness score to be the completeness of the step.

To determine pathway presence, the module completeness score is checked against a parameter called the completeness score threshold (which is 0.75 by default, but adjustable by the user). If the module completeness score is greater than or equal to the threshold, the module is marked as ‘complete’. This Boolean value is meant primarily as a way to easily filter through the output files, which contain a lot more detail beyond the metrics for each pathway. Module presence/absence disguises a lot of nuance, and the philosophy of this program is to provide as much auxiliary data as possible for the user to interpret these summary metrics.

Pathwise copy number

Path copy numbers are computed simultaneously with the completeness score, by considering the number of annotations for each atomic step in the path. Once again, there are several cases to handle:

1. A single KO. The copy number of this atomic step is equal to the number of annotations (hits) of this enzyme.

2. A protein complex. The copy number for these is the number of complete instances of the protein complex. This is calculated by considering the number of annotations for each essential component of the complex (once again, non-essential components are ignored) and comparing that to the module completeness threshold. For instance, if the threshold is 50%, then 50% of the essential components is enough to consider the complex complete.

3. Non-essential KOs. These are ignored.

4. Steps without associated KOs (the ‘--‘ case). These always have a copy number of 0.

5. Modules. The copy numbers of these atomic steps are obtained later, after we've computed the copy number for every other module. There is an adjustment step for copy number just like there is one for completeness.

To get the copy number for a given path through the module, we determine the number of complete copies of the path. This is identical to the calculation of copy number of protein complexes, as described above, and it therefore depends on the module completeness threshold. Suppose a path has 4 essential atomic steps with the following copy numbers: 4,3,1, and 2. Using the default completeness threshold of 0.75, at least 3 out of 4 atomic steps must be present in order for a copy of the path to be considered complete. There is one copy of the path that has all 4 steps, one copy that has 3/4, one copy that has 2/4 and one that has 1/4. Therefore, there are 2 copies of the path with at least 3/4 atomic steps, which means that the path copy number is 2.

The copy number of the module overall is the maximum copy number taken over all paths with the maximum completeness score, which were identified while computing its completeness. If the module does not have any complete paths, then its copy number is 0. If it has one complete path, then its copy number is the copy number of that path. If there are multiple paths with highest completeness score, then its copy number is the maximum of the copy numbers of those paths – for example, suppose there are two paths, both of which are 90% complete. One of those paths has a copy number of 1 and the other has a copy number of 3. The overall module copy number would be 3, in this case. If a module is completely absent (e.g., it does not have any paths of highest completeness), we cannot compute the copy number. In this case, the copy number of the module will be reported as 'NA' in the program output file(s).

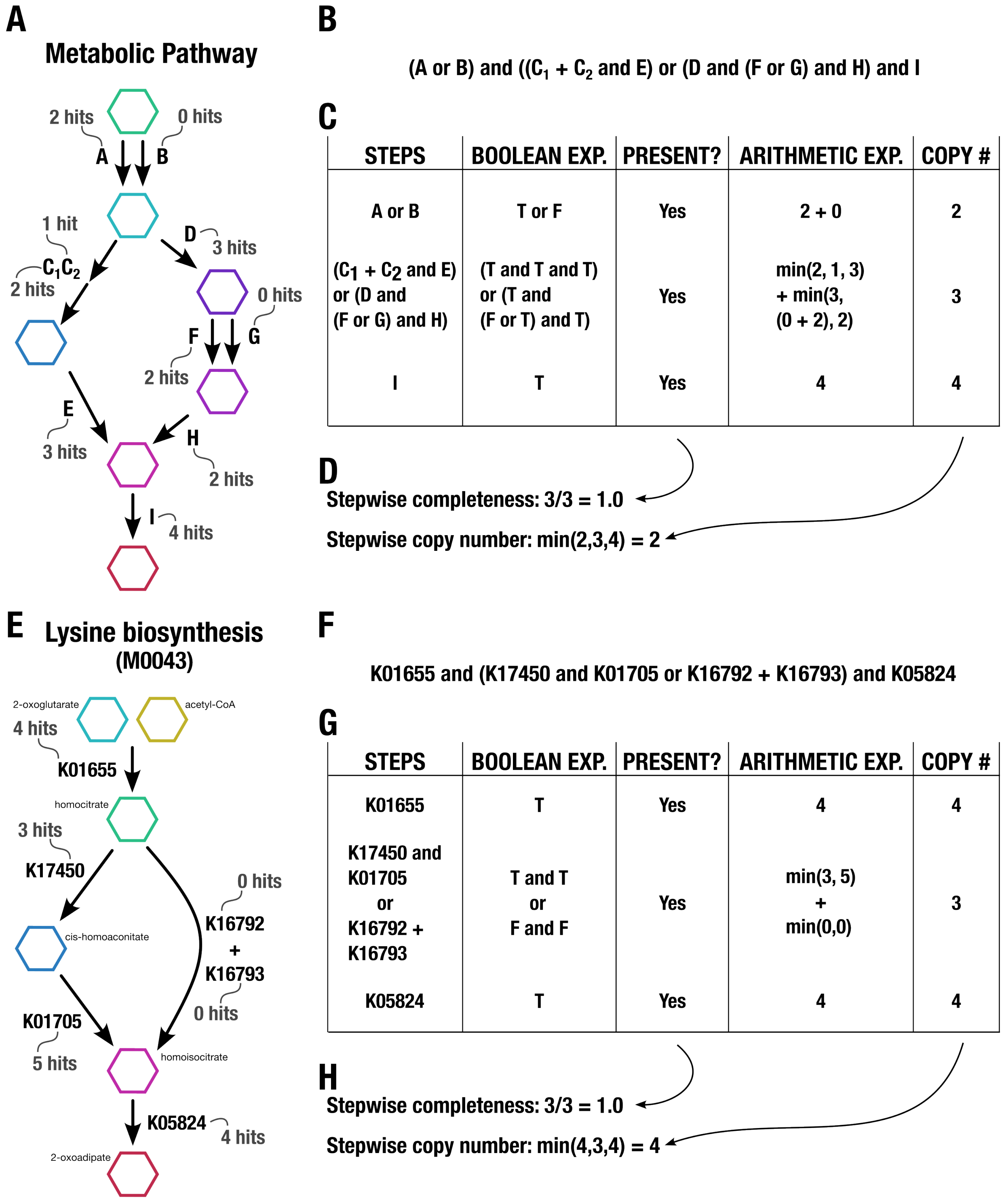
The final step is to recalculate the copy number for modules that are defined by other modules. We set the copy number of a module atomic step to be the previously-computed copy number of that module (if any). If this step has a copy number of 'NA', then the adjusted module copy number will be 'NA' as well; otherwise, the adjusted copy number is calculated as previously described.

Stepwise interpretation of pathways

In the ‘stepwise’ strategy, the program breaks down the module ‘DEFINITION’ into its major, or ‘top-level’, steps. Each ‘top-level’ step usually represents either one metabolic reaction or a branch point in the pathway, and is defined by one or more enzymes that either work together or serve as alternatives to each other to catalyze this reaction or set of reactions. The program uses the available enzyme annotations to determine whether each step can be catalyzed or not – obtaining a binary value representing whether the step is present or not. Then it computes the stepwise module completeness as the percent of present top-level steps. This is the least granular way of estimating module completeness because it does not distinguish between enzyme alternatives – these are all considered as one step which is either entirely present or entirely absent. Likewise, step copy number is the number of complete copies of a top-level step, and a module’s stepwise copy number is the minimum copy number of all of its top-level steps.

To get the top-level steps of a module, its ‘DEFINITION’ string is split on its spaces (not including any spaces within parentheses). Afterwards, the metrics for each step can be computed by converting the step definition into a Boolean expression (for completeness) or arithmetic expression (for copy number) and evaluating that expression.

Figure 2.5 demonstrates the stepwise calculations of completeness and copy number, as described in the following subsections.



(A or B) and ((C_1 + C_2 and E) or (D and (F or G) and H) and I

STEPS	BOOLEAN EXP.	PRESENT?	ARITHMETIC EXP.	COPY #
A or B	T or F	Yes	2 + 0	2
(C_1 + C_2 and E) or (D and (F or G) and H)	(T and T and T) or (T and (F or T) and T)	Yes	min(2, 1, 3) + min(3, (0 + 2), 2)	3
I	T	Yes	4	4

Stepwise completeness: 3/3 = 1.0

Stepwise copy number: min(2,3,4) = 2

K01655 and (K17450 and K01705 or K16792 + K16793) and K05824

STEPS	BOOLEAN EXP.	PRESENT?	ARITHMETIC EXP.	COPY #
K01655	T	Yes	4	4
K17450 and K01705 or K16792 + K16793	T and T or F and F	Yes	min(3, 5) + min(0,0)	3
K05824	T	Yes	4	4

Stepwise completeness: 3/3 = 1.0

Stepwise copy number: min(4,3,4) = 4

Figure 2.5: A demonstration of stepwise metric calculations for metabolic pathways (performed by the program ‘anvi-estimate-metabolism‘ is shown in panels a-d (for a theoretical pathway) and e-h (for a real pathway). a) Theoretical metabolic pathway, where hexagons represent

Figure 2.5 continued: metabolites, arrows represent chemical reactions, letters represent enzymes (subscripts indicate enzyme components), and the example number of gene annotation hits for each enzyme is written in gray. Enzymes with zero hits are highlighted in pink. b) The definition of the theoretical pathway from panel a, written in terms of the required enzymes. c) Table showing the major steps in the pathway and example calculations for step presence and copy number. Step presence is calculated by evaluating a Boolean expression created from the step definition in which enzymes with > 0 hits are replaced with True (T) and the others with False (F). Step copy number is calculated by evaluating the corresponding arithmetic expression in which the enzymes are replaced with their annotation counts. d) Final calculations of completeness score (fraction of present steps) and copy number for the theoretical metabolic pathway. e-h) Same as a-d, but for a real metabolic pathway.

Stepwise completeness

Unlike pathwise completeness, in which the program considers all possible alternatives and computes a fractional completeness for each path, a top-level step can only be entirely complete (1) or entirely incomplete (0). To compute this binary completeness for each top-level step, the step is converted into a Boolean expression by following this set of rules:

1. Enzyme accessions (ie, KOs) are replaced with 'True' if the enzyme is annotated in the sample, and otherwise are replaced with 'False'.

2. '--' steps do not have associated enzyme profiles, so these are always 'False'.

3. Commas represent alternative enzymes, meaning either one or the other is acceptable. The program converts commas into 'OR' relationships.

4. Spaces represent sequential enzymes, meaning that both are necessary (one after the other). The program converts spaces into AND relationships.

5. Plus signs ('+') represent essential enzyme components, meaning that both are necessary (at the same time). The program converts plus signs into AND relationships.

6. Minus signs ('-') represent nonessential enzyme components, meaning that they are unnecessary. These are ignored.

7. Parentheses are kept where they are to maintain proper order of operations.

After this conversion is done, the program evaluates the Boolean expression to determine whether or not the step is complete. After this is done for each top-level step in a module, it calculates the stepwise completeness of the module by taking the percentage of complete top-level steps. If a top-level step includes entire modules in its definition, its completeness is computed after all other modules have been processed.

Stepwise copy number

To compute a top-level step's copy number, the program converts its definition into an arithmetic expression, by following a new set of rules:

1. Enzyme accessions (ie, KOs) are replaced with the number of annotations this enzyme has in the given sample.
2. '--' steps are unknown, so they are replaced with a count of '0'.
3. Commas represent alternative enzymes, meaning either one or the other is acceptable. The program converts commas into addition operations.
4. Spaces represent sequential enzymes, meaning that both are necessary (one after the other). The program converts spaces into minimum, or 'min()', operations.
5. Plus signs ('+') represent essential enzyme components, meaning that both are necessary (at the same time). The program converts plus signs into 'min()' operations.
6. Minus signs ('-') represent nonessential enzyme components, meaning that they are unnecessary. These are ignored.
7. Parentheses are kept where they are to maintain proper order of operations.

This conversion from definition string to arithmetic expression is quite complex for a computer to do, and in the code for this program, it is implemented as a recursive function.

All of the top-level steps in the module have an AND relationship with each other – all are necessary in order to have the module complete. For this reason, the module's overall stepwise copy number is computed by taking the minimum copy number of all top-level steps. Once again, if a top-level step includes entire modules in its definition, its completeness is computed after all other modules have been processed.

Calculation summary

Pathwise module completeness in a given sample is calculated as the maximum fraction of essential enzymes that are annotated in the sample, where the maximum is taken over all possible combinations ('paths') of enzymes in the module definition. Likewise, pathwise module copy number is calculated as the maximum copy number of any path with the module's completeness score.

These values are difficult to interpret when considering metagenomes rather than the genomes of individual organisms. There could be lots of different paths through a module used by different populations in a metagenome, but the module completeness/copy number values would summarize only the most common path(s). In these situations, users can take advantage of the 'module paths' output mode (discussed in the next section) to look at the scores for all individual paths through each module.

Similarly, stepwise module completeness in a given sample is calculated as the percentage of complete top-level steps. Likewise, stepwise module copy number is calculated as the minimum copy number of all top-level steps in the module definition. To interpret these stepwise metrics for modules, it is useful to look at the 'module steps' output mode to see the scores for all individual top-level steps in a module.

2.3.5 Output options for metabolism estimation

'anvi-estimate-metabolism' stores pathway prediction results in tab-delimited text files, which can be either analyzed directly by the user or processed by downstream code. There are two types of output files – long-format output that provides a large variety of data fields, and matrix-format output (otherwise known as wide-format) that summarizes key information across multiple input samples. The long-format type includes several output 'modes' that differ in which data fields are stored in the file, including a customizable mode for pathway data permitting users to choose which fields to store. Multiple files can be generated by the same run of ‘anvi-estimate-metabolism‘, and the suffix of the file name indicates its type or mode. Detailed documentation on output format and options can be found on the help page `https://anvio.org/help/main/artifacts/kegg-metabolism/`.

Long-format output

In addition to storing the pathway prediction metrics (i.e., module completeness and copy number), long-format output files provide as much context as possible for the interpretation of these values. There are multiple ‘modes’ that each describes a different level of metabolic pathway organization. If multiple inputs are provided to ‘anvi-estimate-metabolism‘, their independent metabolism estimation results are stored in the same output file, with sample identifiers on each line to differentiate between each set of results.

‘Modules’ mode output describes the completeness score and copy number, for both the pathwise and stepwise pathway interpretation strategies, of each metabolic module in each input sample. Results are keyed by module accession number, and the module name, categorization, and definition are also provided. In addition, this file includes a number of fields to assist with interpreting the pathway metrics. It lists the enzymes from the module definition that were annotated in the input sample as well as their corresponding gene calls. This explains the completeness score value and allows users to analyze the topology of the annotated pathway components – for instance, to see which parts of the pathway are missing. The gene call list enables integration of these results with other gene call data stored in the database, such as sequences, lengths, and variants. The file also indicates which enzymes are unique to the module (that is, not belonging to other pathways in the modules database), the number of annotations for each unique enzyme, and the proportion of unique enzymes that were annotated. This can be useful information to interpret the completeness of modules with a high proportion of shared enzymes; if all of its annotated enzymes also belong to other pathways, it is less likely that the module is complete even if the completeness score is high. The file includes a ‘warnings’ column that lists any shared enzymes and their respective modules, as well as indicating when enzymes in the module definition do not have an annotation profile (and will therefore be missing due to a technical, rather than biological, reason). Including module copy number in this output is optional, but if it is included, the copy number of each top-level step will also be listed for a more nuanced interpretation of the overall stepwise copy number. Finally, users can elect to add substrate, intermediate, and product compounds to this output file for analysis of the metabolites involved in a given pathway.

Two output modes are useful for analyzing the nuances of pathway prediction metrics. ‘Module paths’ mode output provides information on each path through every module, as produced via pathwise interpretation of the module definition. In this file, each line describes one path, including the path definition, the path completeness score and (optionally) copy number, and a list of which enzymes in the path are annotated (or missing) for interpretation of those values. The pathwise completeness score of the module overall is also provided for context. ‘Module steps’ mode is the analogous mode for stepwise calculations, providing information on each top-level step in every module. In this file, each line describes one step, including its definition, its completeness score, and (optionally) its copy number. Again, the stepwise completeness score of the module overall is provided for context.

‘Hits’ mode output describes each enzyme annotation in detail, including enzymes from all of the annotation sources used for metabolism estimation regardless of whether they belong to a metabolic pathway. Since only a subset of these enzymes belong to modules, this output does not include pathway prediction data like paths and module completeness. Rather, it describes the enzyme function, which gene call it belongs to, which contig in the input sequences this gene is found on, and which metabolic modules the enzyme participates in (if any). The purpose of this mode is to allow a deeper investigation of functions of interest – for instance, transporter enzymes are often analyzed in conjunction with metabolic capacity.

If users are interested only in a subset of pathway prediction data, or the pre-defined output modes do not fit their needs, they can customize which data fields will appear in the output file. This ‘custom’ mode is currently only available for modules-related data (that is, any fields accessible to the ‘modules’, ‘module paths’, or ‘module steps’ modes).

It is occasionally useful to summarize the abundance and prevalence of metabolic pathways (and enzymes) across multiple metagenome samples by mapping these samples to a reference that encodes these pathways. If such read recruitment data is available (and provided to ‘anvi-estimate-metabolism‘ in the form of an anvi’o profile database), these can be used to compute coverage and detection of each annotated enzyme. Coverage describes the average number of sequencing reads that map to each nucleotide position in the gene, while detection describes the proportion of nucleotides in the gene sequence that are covered by at least one read. Gene-level coverage and detection information can be added to ‘modules’ mode and ‘hits’ mode output. In ‘modules’ mode, the average coverage and detection of all genes in the module is also provided, and can serve as alternative measures of pathway abundance and presence.

Matrix-format output

The purpose of matrix-format output is to generate matrices of key pathway prediction metrics for easy visualization, clustering, and downstream processing. It is available when running ‘anvi-estimate-metabolism‘ on multiple input samples (i.e., many individual genomes, or many MAGs in a binned metagenome). Each output matrix summarizes one value, such as pathwise module completeness, across every input sample and is therefore not as descriptive but much more concise than long-format output.

When generating output in this format, multiple matrix files are produced, and these are keyed by module, step, or enzyme accession depending on the summarized value. These currently include matrices for: pathwise completeness, stepwise completeness, binary module presence/absence for pathwise completeness and stepwise completeness, binary completeness of top-level steps in each module, and annotation counts for each enzyme. Copy number matrices are optional, and include a matrix for pathwise copy number, stepwise copy number, and copy number of top-level steps in each module. Users can also choose to generate enzyme annotation matrices that are specific to a given metabolic module for targeted analysis.

Output matrices by default include only accessions, sample names, and the value of interest so that they can be used for downstream applications like clustering without further processing. However, users can elect to include metadata for greater readability, such as module names and categories or enzyme functions.

Matrix-format output is useful for visualizing pathway prediction results as heatmaps, as described next.

2.3.6 Visualizing metabolism heatmaps

Pathway prediction results can be readily visualized as interactive heatmaps from matrix-format output of ‘anvi-estimate-metabolism‘ (Figure 2.6). ‘anvi-interactive‘ is an anvi’o program that serves interactive visualizations of data in a web browser. Though it is most often used for sequence data, this program can also visualize arbitrary data matrices. To produce an interactive heatmap of module completeness scores, for example, users can provide the completeness score matrix to ‘anvi-interactive‘ and modify the interface settings so that the data is shown as intensities. They can also incorporate clustering results to organize the rows and columns of the heatmap, such that modules are organized according to their similar distribution across input samples and input samples are organized according to similar metabolic capacity. This is achievable via two other anvi’o programs that can operate on matrices, ‘anvi-matrix-to-newick‘ (which clusters the input matrix row-wise) and ‘anvi-script-transpose-matrix‘ (which can flip the matrix so that the other dimension can be clustered). A detailed tutorial on visualizing metabolism heatmaps can be found at `https://merenlab.org/tutorials/infant-gut/#chapter-v-metabolism-prediction`. I also implemented a program to automate this process called ‘anvi-display-metabolism‘; however, it is still experimental and requires further development.

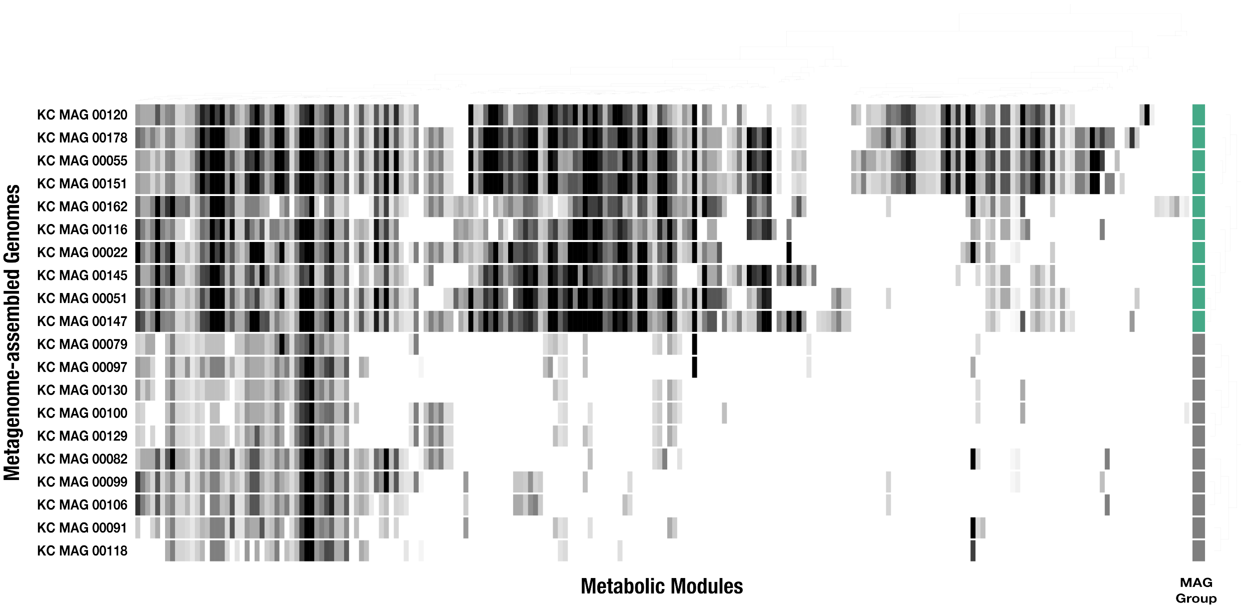


Figure 2.6: Example heatmap visualization for completeness of metabolic modules across multiple genomes. Each column holds data for one metabolic module and each row holds data for one genome, in this case metagenome-assembled genomes (MAGs). Each cell of the heatmap indicates module completeness score, from 0.0 (white) to 1.0 (black). The data is clustered in each dimension by a dendrogram that was calculated from the matrix of completeness scores. The data in this example is from the high- and low-independence MAGs described in Chapter 3.2, and the ’MAG Group’ column indicates which MAG has high metabolic independence (green) or low metabolic independence (gray).

2.4 Future work

The metabolism reconstruction framework is robust, includes a variety of features supporting diverse and flexible analyses of metabolism, and has already been used in a number of published studies. Regardless, there is always room for improvement and new features. Here I will discuss a few ideas for future implementations of this software that are of foremost importance to the study of microbial metabolism.

User-defined metabolism is still in its infancy. It works, but its implementation is clunky, having been limited by the initial framework for storing pathway data that was defined according to the data types and format provided by KEGG. It is indeed ironic that this critical feature for accessing metabolic pathways beyond KEGG modules is limited to KEGG's formatting strategies. A fresh and more generic method for defining modules, with easily-parsable file formats (such as JSON) and support for arbitrary data fields, would represent a significant improvement and is already being discussed by the anvi'o community (`https://github.com/merenlab/anvio/issues/1873`). This is a critical frontier for enabling cutting-edge research on novel and understudied metabolisms, and it should be easier to use. Improving user-defined metabolism would also allow me to generalize and streamline the codebase; for instance, by consolidating module definitions from multiple sources into a single modules database for easier sanity checking and data management.

There are several opportunities for extending this framework with additional analysis capabilities. For instance, the study of metabolism across pangenomes, or 'pan-metabolism', is particularly interesting. Pan-metabolism predicts the metabolic capabilities encoded by a set of closely-related genomes (typically of the same species or genus), and distinguishes between the 'core' metabolism present in all genomes and the 'accessory' metabolism encoded by subsets of the genomes. Analyses of this sort have only recently appeared in published scientific literature, and are typically done by combining all gene annotations from each genome for metabolism reconstruction (McCubbin et al., 2020; Lau Vetter et al., 2022; Mohite et al., 2022; Zhang et al., 2023). An opportunity remains for integrating genome-level pathway prediction results into a pan-metabolism summary that remains aware of the differences between individual genomes, and for interactively visualizing this data to facilitate its analysis and exploration by researchers (such as with the anvi'o pangenome visualization program 'anvi-display-pan'). Of course, pan-metabolism is not the only future direction available to the metabolism reconstruction framework; other opportunities for integrating analyses include prediction of community interactions, pathway structure studies (such as synteny analysis via ‘anvi-analyze-synteny‘ or operon identification), and validation of predicted pathways with transcriptomic and/or metabolomic data.

Finally, implementing interactive pathway visualization would be extremely useful for metabolism data exploration. By this, I mean the visualization of the interconnected components of each metabolic pathway, including enzymes and compounds, both for individual pathways and for networks of integrated pathways. These visualizations could be enriched with associated data from the metabolism reconstruction process, such as pathway prediction metrics and consolidated annotation information. They could even be interactively linked to auxiliary sequence data; for instance, by showing the distribution of all genes in the pathway across an input genome sequence. Several solutions for visualizing metabolic pathways exist, such as ESCHER (King et al., 2015) and KEGG pathway maps (Kanehisa et al., 2022), but the integration of my metabolism reconstruction framework within anvi’o makes it possible to substantially improve upon these with greater flexibility, data richness, and interactivity.

CHAPTER 3

METABOLIC INDEPENDENCE DETERMINES COLONIZATION SUCCESS FOR GUT MICROBES

3.1 Preface

This chapter introduces a study on microbial colonization in the human gut as well as a follow-up, targeted investigation of the colonization patterns within a set of closely-related populations. Both of these works rely on the metabolism reconstruction framework to generate insights into the metabolic potential of gut microbes.

The colonization study introduces the concept of metabolic independence, the degree to which microbial genomes encode complete metabolic modules for synthesizing critical metabolites, including amino acids, nucleotides, and vitamins. Microbes with high metabolic independence are self-sufficient, while those with low metabolic independence rely on their surrounding community to provide essential molecules. The study uses fecal microbiota transplantation (FMT) as an *in natura* experimental model to investigate the association between metabolic independence and resilience in stressed gut environments. It suggests that FMT serves as an environmental filter that favors populations with higher metabolic independence. Interestingly, we observed higher completion of the same biosynthetic pathways in microbes enriched in IBD patients. These observations suggest a general mechanism that underlies changes in diversity in perturbed gut environments, and reveal taxon-independent markers of 'dysbiosis' that may explain why widespread yet typically low abundance members of healthy gut microbiomes can dominate under inflammatory conditions without any causal association with disease.

A subsequent investigation highlights the problem of annotation bias; that is, the failure to capture functional annotations for populations that are less represented in publicly-available databases. This follow-up study explores the functional and metabolic differences between three clades of *Bifidobacterium* genomes with differential colonization success in the FMT recipients. The analysis indicated minor differences in functional capacity, paralleled by observed differences in genome length, that would seem to correlate with the degree of colonization success. However, deeper investigation revealed that bifidobacteria, which are poorly characterized in reference databases, suffer from annotation bias in a manner that is not systematic across different clades. Consequently, the functional data do not recapitulate some prior published observations on the functional capacity of these microbes, and it is difficult to trust the metabolism reconstruction results derived from these incomplete annotations. Overall, this investigation indicates that our understanding of metabolism is insufficient to differentiate between closely-related populations.

3.2 Metabolic independence drives gut microbial colonization and resilience in health and disease

This section is derived from the following publication:

Iva Veseli*, Andrea R Watson*, Jessika Fuessel*, Johanna Zaal DeLongchamp, Marisela Silva, Florian Trigodet, Karen Lolans, Alon Shaiber, Emily Fogarty, Christopher Quince, Michael K Yu, Arda Soylev, Hilary G Morrison, Sonny T M Lee, David T Rubin, Bana Jabri, Thomas Louie, and A Murat Eren. Metabolic independence drives gut microbial colonization and re-silience in health and disease. Genome Biol 24, 78 (2023). `https://doi.org/10.1186/s13059-023-02924-x`.

* *indicates co-first authors.*

3.2.1 Background

Understanding the determinants of microbial colonization is one of the fundamental aims of gut microbial ecology (Costello et al., 2012; Messer et al., 2017). The gradual maturation of the microbiome during the first months of life (Stewart et al., 2018), the importance of diet and lifestyle in shaping the gut microbiome (Koenig et al., 2011; Rothschild et al., 2018), and the biogeography of microbial populations along the gastrointestinal tract (Donaldson et al., 2016) strongly suggest the importance of niche-based interactions between the gut environment and its microbiota. Previous studies that described such interactions in the context of microbial colonization have focused on microbial succession in infant gut microbiomes (Stewart et al., 2018), or relied on model systems such as germ free mice conventionalized with a consortium of microbial isolates from infant stool (Feng et al., 2020). However, our understanding of the ecological underpinnings of secondary succession following a major ecosystem disturbance caused by complex environmental factors in the gut microbiome remains incomplete. A wide range of diseases and disorders are associated with such disturbances, (Almeida et al., 2020; Durack and Lynch, 2019; Lynch and Pedersen, 2016) however; mechanistic underpinnings of these associations have been difficult to resolve. This is in part due to the diversity of human lifestyles (David et al., 2014a), and the limited utility of model systems to make robust causal inferences for microbially mediated human diseases (Walter et al., 2020).

Inflammatory bowel disease (IBD), a group of increasingly common intestinal disorders that cause inflammation of the gastrointestinal tract (Baumgart and Carding, 2007), has been a model to study human diseases associated with the gut microbiota (Schirmer et al., 2019). The pathogenesis of IBD is attributed in part to the gut microbiome (Plichta et al., 2019), yet the microbial ecology of IBD-associated dysbiosis remains a puzzle. Despite marked changes in gut microbial community composition in IBD (Ott et al., 2004; Sokol and Seksik, 2010; Joossens et al., 2011), the microbiota associated with the disease lacks acquired infectious pathogens (Chow et al., 2011), and microbes that are found in IBD typically also occur in healthy individuals (Clooney et al., 2021), which complicates the search for robust functional or taxonomic markers of health and disease states (Lloyd-Price et al., 2019). One of the hallmarks of IBD is reduced microbial diversity during episodes of inflammation, when the gut environment is often dominated by microbes that typically occur in lower abundances prior to inflammation (Vineis et al., 2016b). The sudden increase in the relative abundance of microbes that are also common to healthy individuals suggests that the harsh conditions of IBD likely act as an ecological filter that eliminates some populations while allowing others to bloom. Yet, in the absence of an understanding of the genetic requirements for survival in IBD, critical insights into the functional drivers of microbial community succession in such disease states remains elusive.

Fecal microbiota transplantation (FMT), the transfer of stool from a donor into a recipient's gastrointestinal tract (Eiseman et al., 1958), represents an experimental middleground to capture complex ecological interactions that shape the microbial community during secondary succession of a disrupted gut environment. FMT is frequently employed in the treatment of recurrent *Clostridioides difficile* infection (CDI) (van Nood et al., 2013) that can cause severe diarrhea and intestinal inflammation. In addition to its medical utility, FMT offers a powerful framework to study fundamental questions of microbial ecology by colliding the microbiome of a healthy donor with the disrupted gut environment of the recipient. The process presents an ecological filter with the potential to reveal functional determinants of microbial colonization success and resilience in impaired gut environments (Schmidt et al., 2018).

Here we use FMT as an *in natura* experimental model to investigate the ecological and functional determinants of successful colonization of the human gut at the level of individual microbial populations using genome-resolved metagenomics. Our findings highlight the importance of environmental selection acting on the biosynthetic capacity for essential nutrients as a key driver of colonization outcome after FMT and resilience during inflammation, and demonstrate that metabolic independence serves as a taxonomy-independent determinant of colonization success in the human gut.

3.2.2 Results and Discussion

Study Design

Our study includes 109 gut metagenomes (Supplementary Table 3.1) from two healthy FMT donors (A and B) and 10 FMT recipients (five recipients per donor) with multiple recurrent CDI. We collected 24 Donor A samples over a period of 636 days and 15 Donor B samples over a period of 532 days to establish an understanding of the long-term microbial population dynamics within each donor microbiota. The FMT recipients received vancomycin for a minimum of 10 days to attain resolution of diarrheal illness prior to FMT. On the last day of vancomycin treatment, a baseline fecal sample was collected from each recipient, and their bowel contents were evacuated immediately prior to FMT. Recipients did not take any antibiotics on the day of transplant, or during the post-FMT sampling period (Supplementary Figure 3.4). We collected 5 to 9 samples from each recipient for a period of up to 336 days post-FMT. Deep sequencing of donor and recipient metagenomes using Illumina paired-end (2x150) technology resulted in a total of 7.7 billion sequences with an average of 71 million reads per metagenome (Figure 3.1, Supplementary Table 3.1, Supplementary Table 3.2). We employed genome-resolved metagenomics, microbial population genetics, and metabolic pathway reconstruction for an in-depth characterization of donor and recipient gut microbiotas, and we leveraged publicly available gut metagenomes to benchmark our observations.

Genome-resolved metagenomics show many, but not all, donor microbes colonized recipients and persisted long-term

We first characterized the taxonomic composition of each donor and recipient sample by analyzing our metagenomic short reads given a clade-specific k-mer database (Supplementary Table 3.2). The phylum-level microbial community composition of both donors reflected those observed in healthy individuals in North America (Human Microbiome Project Consortium, 2012b): a large representation of Firmicutes and Bacteroidetes, and other taxa with lower relative abundances including Actinobacteria, Verrucomicrobia, and Proteobacteria (Figure 3.1, Supplementary Table 3.2). In contrast, the vast majority of the recipient pre-FMT samples were dominated by Proteobacteria, a phylum that typically undergoes a drastic expansion in individuals treated with vancomycin (Isaac et al., 2017). After FMT, we observed a dramatic shift in recipient taxonomic profiles (Supplementary Table 3.2, Supplementary Figure 3.5, Supplementary Figure 3.6), a widely documented hallmark of this procedure (Khoruts et al., 2010; Grehan et al., 2010; Shahinas et al., 2012). Nearly all recipient samples post-FMT were dominated by Bacteroidetes and Firmicutes as well as Actinobacteria and Verrucomicrobia in lower abundances, resembling qualitatively, but not quantitatively, the taxonomic profiles of their donors (Supplementary Table 3.2). The phylum Bacteroidetes was over-represented in recipients: even though the median relative abundance of Bacteroidetes populations were 5% and 17% in donors A and B, their relative abundance in recipients post-FMT was 33% and 45%, respectively (Figure 3.1, Supplementary Table 3.2). A single genus, *Bacteroides*, made up 76% and 82% of the Bacteroidetes populations in the recipients of Donor A and B, respectively (Supplementary Table 3.2). The success of the donor *Bacteroides* populations in recipients upon FMT is not surprising given the ubiquity of this genus across geographically diverse human populations (Wexler and Goodman, 2017) and the ability of its members to survive substantial levels of stress (Swidsinski et al., 2005; Vineis et al., 2016b). This initial coarse taxonomic analysis demonstrates the successful transfer of only some populations, suggesting selective filtering of the transferred community.

To generate insights into the genomic content of the microbial community, we first assembled short metagenomic reads into contiguous segments of DNA (contigs). Co-assemblies of 24 Donor A and 15 Donor B metagenomes independently resulted in 53,891 and 54,311 contigs that were longer than 2,500 nucleotides, and described 0.70 and 0.79 million genes occurring in 179 and 248 genomes, as estimated by the mode of the frequency of bacterial

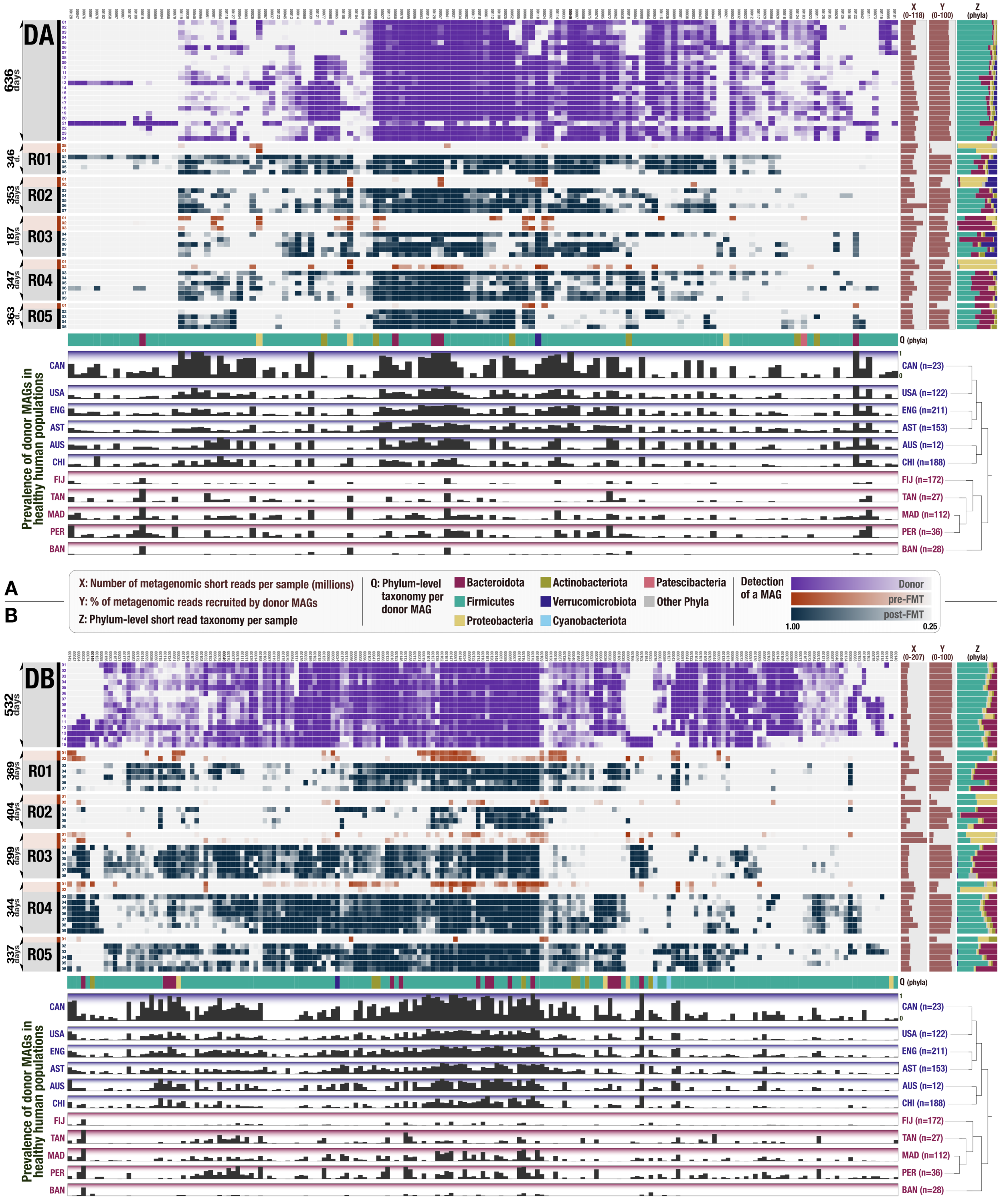


Figure 3.1: Detection of FMT donor genomes in FMT recipients and publicly available gut metagenomes. In both heat maps, each column represents a donor genome, each row represents a metagenome, and each data point represents the detection of a given genome in a

Figure 3.1 continued: given metagenome. Purple rows represent donor metagenomes from stool samples collected over 636 days for (A) Donor A and 532 days for (B) Donor B. Orange rows represent recipient pre-FMT metagenomes, and blue rows represent recipient post-FMT metagenomes. Rows are arranged in descending chronological order with respect to each subject. The intensity of purple, orange and blue color scales represent the detection value for each genome in each metagenome, with a minimum detection of 0.25. Genome columns are clustered according to their presence or absence in all metagenomes (Euclidean distance and Ward clustering). The three columns to the right of the heatmaps display, for each metagenome row: (X) the number of metagenomic short reads in millions, (Y) the percent of metagenomic short reads recruited by genomes, and (Z) the taxonomic composition of metagenomes (based on metagenomic short reads) at the phylum level. The first row below each heat map (Q) provides the phylum-level taxonomy for each donor genome. Finally, the 11 bottommost rows under each heat map show the fraction of healthy adult metagenomes from 11 different countries in which a given donor genome is detected (if a genome is detected in every individual from a country it is represented with a full bar and a value of 1). The dendrograms on the right-hand side of the country layers organize countries based on the detection patterns of genomes (Euclidean distance and Ward clustering). Purple and red shaded countries represent the two main clusters that emerge from this analysis, where purple layers are industrialized countries in which donor genomes are highly prevalent and red layers are less industrialized countries where the prevalence of donor genomes is low. A maximum resolution version of this figure is also available at https://doi.org/10.6084/m9.figshare.15138720.

single-copy core genes (Supplementary Table 3.2). On average, 80.8% of the reads in donor metagenomes mapped back to the assembled contigs from donor metagenomes, which suggests that the assemblies represented a large fraction of the donor microbial communities. Donor assemblies recruited only 43.4% of the reads on average from the pre-FMT recipient metagenomes. This number increased to 80.2% for post-FMT recipient metagenomes, and remained at an average of 76.8% even one year post-FMT (Supplementary Table 3.2). These results suggest that members of the donor microbiota successfully established in the recipient gut and persisted long-term.

To investigate functional determinants of microbial colonization by identifying donor populations that were successful at colonizing multiple individuals, we reconstructed microbial genomes from donor assemblies using sequence composition and differential coverage signal as previously described (Sharon et al., 2013; Lee et al., 2017). We manually refined metagenomic bins to improve their quality following previously described approaches (Delmont et al., 2018; Shaiber et al., 2020b) and only retained those that were at least 70% complete and had no more than 10% redundancy as predicted by bacterial single-copy core genes (Bowers et al., 2017; Chen et al., 2020). Our binning effort resulted in a final list of 128 metagenome-assembled genomes (MAGs) for Donor A and 183 MAGs for Donor B that included members of Firmicutes (n=265), Bacteroidetes (n=20), Actinobacteria (n=14), Proteobacteria (n=7), Verrucomicrobia (n=2), Cyanobacteria (n=2), and Patescibacteria (n=1) (Supplementary Table 3.2). The taxonomy of donor-derived genomes largely reflected the taxonomic composition of donor metagenomic short reads (Figure 3.1, Supplementary Table 3.2, Supplementary Table 3.3). While only 20 genomes (mostly of the genera *Bacteroides* and *Alistipes*) explained the entirety of the Bacteroidetes group, we recovered 265 genomes that represented lower abundance but diverse populations of Firmicutes (Figure 3.1, Supplementary Table 3.2, Supplementary Table 3.3).

Metagenomic read recruitment elucidates colonization events

Reconstructing donor genomes enabled us to characterize (1) population-level microbial colonization dynamics before and after FMT using donor and recipient metagenomes and (2) the distribution of each donor population across geographically distributed humans using 1,984 publicly available human gut metagenomes (Figure 3.1, Supplementary Table 3.4).

Our metagenomic read recruitment analysis showed that donor A and B genomes recruited on average 77.05% and 83.04%, respectively, of reads from post-FMT metagenomes, suggesting that the collection of donor genomes well represents the recipient metagenomes post-FMT (Figure 3.1). As expected, we detected each donor population in at least one donor metagenome (see Methods for ‘detection’ criteria). Yet, only 16% of Donor A populations were detected in every Donor A sample, and only 44% of Donor B populations were detected in every Donor B sample (Figure 3.1, Supplementary Table 3.3), demonstrating the previously documented dynamism of gut microbial community composition over time (David et al., 2014a). A marked increase in the detection of donor populations in recipients after FMT is in agreement with the general pattern of transfer suggested by the short-read taxonomy (Figure 3.1): while we detected only 38% of Donor A and 54% of Donor B populations in at least one recipient pre-FMT, these percentages increased to 96% for both donors post-FMT (Supplementary Table 3.3). We note that we observed a higher fraction of donor populations in recipients as a function of the FMT delivery method. Following the cases of FMT where donor stool was transplanted via colonoscopy, we detected 54.7% and 33.3% donor genomes in the recipients of donor A (n=3) and donor B (n=2), respectively. In contrast, in the cases of FMT where donor stool was transplanted via pills, we detected 69.5% and 61.6% donor genomes in the recipients of donor A (n=2) and donor B (n=3), respectively.

Overall, not every donor population in our dataset was detected in each recipient, but the emergence of donor populations in recipients did not appear to be random: while some donor populations colonized all recipients, others colonized none (Figure 3.1), providing us with an opportunity to quantify colonization success for each donor population in our dataset.

Succession of donor microbial populations in FMT recipients and their prevalence in publicly available metagenomes reveal good and poor colonizers

Of the populations that consistently occurred in donor metagenomes, some were absent in all or most recipient metagenomes after FMT, and others were continuously present throughout the sampling period in both donor and recipient metagenomes (Figure 3.1). To gain insights into the ecology of donor microbial populations beyond our dataset, we explored their occurrence in publicly available healthy gut metagenomes through metagenomic read recruitment. This analysis enabled us to consider the prevalence of donor populations in FMT recipients and global gut metagenomes, and define two groups of donor genomes that represented opposite colonization and prevalence phenotypes.

The ‘good colonizers’ comprise those microbial populations that colonized and persisted in all FMT recipients. Intriguingly, these populations were also the most prevalent in publicly available gut metagenomes from Canada. Overall, these donor microbial populations (1) systematically colonized the majority of FMT recipients, (2) persisted in these environments long-term regardless of host genetics or lifestyle, and (3) were prevalent in public gut metagenomes outside of our study. In contrast, the so-called ‘poor colonizers’ failed to colonize or persist in at least three FMT recipients. These populations were nevertheless viable in the donor gut environment: not only did they occur systematically in donor metagenomes but also they sporadically colonized some FMT recipients. Yet, unlike the good colonizers, the distribution patterns of poor colonizers were sparse within our cohort, as well as within the publicly available metagenomes. In fact, populations identified as poor colonizers were less prevalent than good colonizers in each of the 17 different countries we queried. In countries including the United States, Canada, Austria, China, England, and Australia, microbial populations identified as good colonizers occurred in 5 times more people than poor colonizers in the same country (Figure 3.1, Supplementary Table 3.3), which suggests that the outcomes of FMT in our dataset were unlikely determined by neutral processes. This observation is in contrast with previous studies that suggested ‘dose’ (i.e., the abundance of a given population in donor fecal matter) as a predominant force that determines outcomes of colonization after FMT (Smillie et al., 2018; Podlesny and Florian Fricke, 2020). However, our strain-resolved analysis of colonization events in our data in conjunction with the distribution of the same populations in publicly available metagenomes revealed (1) a significant correlation between the colonization success of donor populations and their prevalence across publicly available metagenomes, and (2) showed that the prevalence of a given population across global gut metagenomes can predict its colonization success after FMT better than its abundance in the donor stool sample (Wald test, p=6.3e-06 and p=9.0e-07) (Supplementary Information). Overall, these observations suggest a link between the colonization outcomes in our study and global prevalence of the same microbial populations, and that the succession of donor populations in our data were likely influenced by selective processes that influence colonization outcomes.

Next, we sought to investigate whether we can identify metabolic features that systematically differ between good colonizers and poor colonizers independent of their taxonomy. To conduct such a comparative analysis, we conservatively selected the top 20 populations from each group that best reflect their group properties by considering both their success after FMT and their prevalence across publicly available metagenomes (Supplementary Table 3.7). The 20 populations representative of good colonizers were dominated by Firmicutes (15 of 20) but also included Bacteroidetes and one Actinobacteria population. All populations identified as poor colonizers resolved to Firmicutes (Figure 3.2, Supplementary Table 3.7). Genome completion estimates did not differ between good and poor colonizers (Wilcoxon rank sum test, p=0.42) and averaged to 91% and 93%, respectively. But intriguingly, the genome sizes between the two groups differed dramatically (p=2.9e-06): genomes of good colonizers averaged to 2.8 Mbp while those of poor colonizers averaged to 1.6 Mbp. We considered that our bioinformatics analyses may have introduced biases to genome lengths, but found a very high correspondence between the lengths of the genomes and their best matching reference genomes in the Genome Taxonomy Database (GTDB) (R2=0.88, p=5e-14). Assuming that the generally larger genomes of good colonizers may be an indication of an increased repertoire of core metabolic competencies compared to poor colonizers, we next conducted a metabolic enrichment analysis for quantitative insights (see Methods).

Good colonizers are enriched in metabolic pathways for the biosynthesis of essential organic compounds

Our enrichment analysis between good and poor colonizers revealed 33 metabolic modules (out of 443 total in the KEGG module database) that were enriched in good colonizers and none that were enriched in poor colonizers (Figure 3.2, Supplementary Table 3.7). Of all enriched modules, 79% were related to biosynthesis, indicating an overrepresentation of biosynthetic capabilities among good colonizers as KEGG modules for biosynthesis only make up 55% of all KEGG modules (Figure 3.2, Supplementary Table 3.7). Of the 33 enriched modules, 48.5% were associated with amino acid metabolism, 21.2% with vitamin and cofactor metabolism, 18.2% with carbohydrate metabolism, 24.2% with nucleotide metabolism, 6% with lipid metabolism and 3% with energy metabolism (Supplementary Table 3.7). Metabolic modules that were enriched in the good colonizers included the biosynthesis of seven of nine essential amino acids, indicating the importance of high metabolic independence to synthesize essential compounds as a likely factor that increases success in colonizing new environments (Supplementary Table 3.7). This is further supported by the enrichment of biosynthesis pathways for the essential cofactor vitamin B12 (cobalamin), which occurred in 67.5% of the good colonizers and only 12.5% of the poor colonizers (Supplementary Table 3.7). Vitamin B12 is structurally highly complex and costly to produce, requiring expression of more than 30 genes that are exclusively encoded by bacteria and archaea (Martens et al., 2002). In addition to the biosynthesis of tetrahydrofolate, riboflavin, and cobalamin, the genomes of good colonizers had a larger representation of biosynthetic modules for vitamins including biotin, pantothenate, folate, and thiamine (Supplementary Table 3.7). These micronutrients are equally essential in bacterial and human metabolism and are important mediators of host-microbe interactions (Biesalski, 2016). Interestingly, enriched metabolic modules in our analysis partially overlap with those that Feng et al. identified as the determinants of microbial fitness using metatranscriptomics and a germ-free mouse model conventionalized with microbial isolates of human origin (Feng et al., 2020).

Even though these 33 metabolic modules were statistically enriched in populations identified as good colonizers, some of them also occurred in the genomes of poor colonizers (Figure 3.2). To identify whether the levels of completion of these modules could distinguish the good and poor colonizers, we matched six good colonizers that encoded modules enriched in these populations to six populations of poor colonizers from the same phylum (Figure 3.2). Bacte-

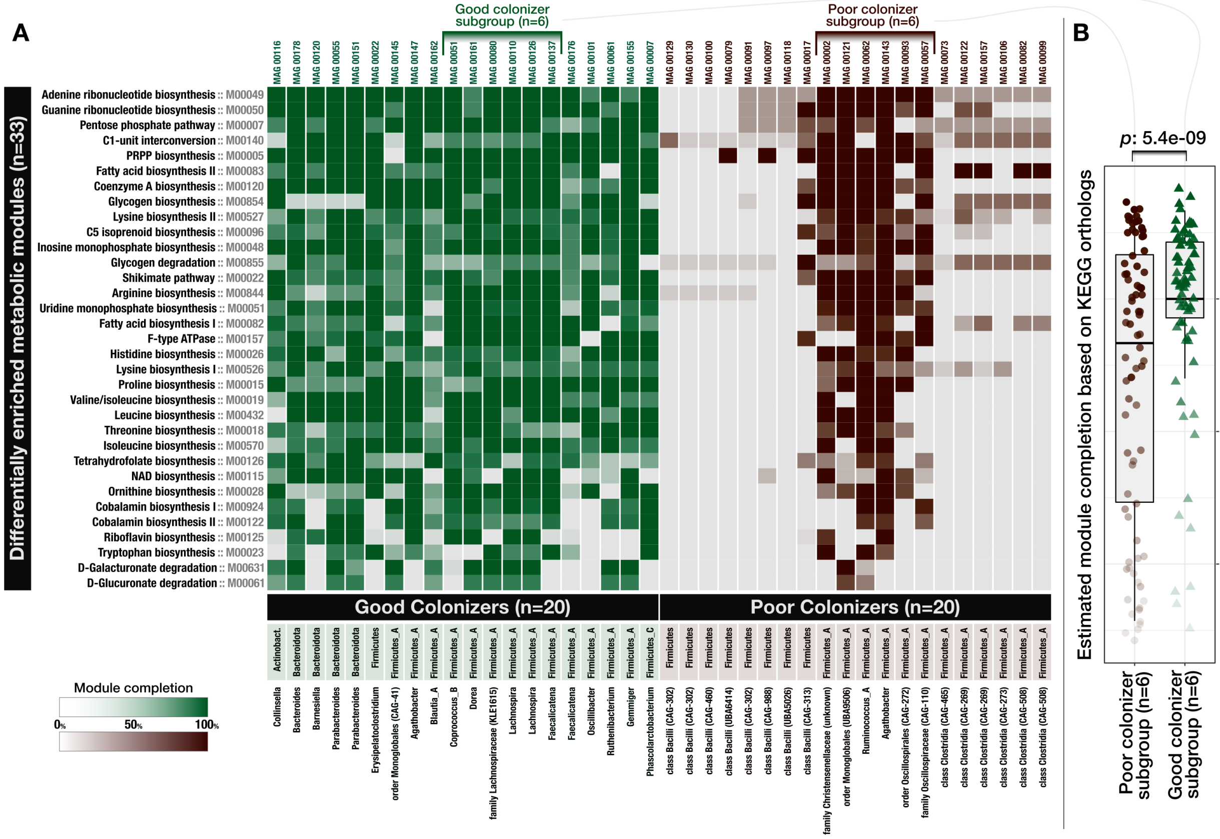


Figure 3.2: Distribution of metabolic modules across genomes of good and poor colonizers. Each data point in this heat map shows the level of completion of a given metabolic module (rows) in a given genome (columns). The box-plot on the right-side compares a subset of poor colonizer and good colonizer genomes, where each data point represents the level of completion of a given metabolic module in a genome and shows a statistically significant difference between the overall completion of metabolic modules between these subgroups (Wilcoxon rank sum test, p=5.4e-09). A high-resolution version of this figure is also available at `https://doi.org/10.6084/m9.figshare.15138720`.

rial single-copy core genes estimated that genomes in both subgroups were highly complete with a slight increase in average genome completion of poor colonizers (93.7%) compared to good colonizers (90.1%). Despite the higher estimated genome completion for populations of poor colonizers, estimated metabolic module completion values were slightly yet significantly lower in this group (Wilcoxon rank sum test with continuity correction, V=958, p=5e-09) (Figure 3.2, Supplementary Table 3.7). Thus, these modules were systematically missing genes in populations of poor colonizers, indicating their functionality was likely reduced, if not absent.

These observations suggest that the ability to synthesize cellular building blocks, cofactors and vitamins required for cellular maintenance and growth provides a substantial advantage during secondary succession, highlighting that the competitive advantages conferred by metabolic autonomy may outweigh the additional costs under certain conditions. For the remainder of our study, we use the term 'high metabolic independence' (HMI) to describe genomic evidence for a population's ability to synthesize essential compounds (that is, high completeness scores of biosynthesis pathways for these compounds indicating the presence of most, if not all, genes required to produce them), and 'low metabolic indepence' (LMI) to describe the absence of, or reduction in, such capacity.

While gut microbial ecosystems of healthy individuals include microbes with both low- and high-metabolic independence, IBD primarily selects for microbes with high-metabolic independence

Our results so far show that while the healthy donor environment could support both HMI and LMI populations (Figure 3.1, Supplementary Table 3.3), challenging microbes to colonize a new environment or to withstand ecosystem perturbation during FMT selects for HMI populations (Figure 3.2, Supplementary Table 3.7), suggesting that metabolic independence is a more critical determinant of fitness during stress than during homeostasis. Based on these observations, it is conceivable to hypothesize that (1) a gut environment in homeostasis will support a large variety of microbial populations with a wide spectrum of metabolic independence, and (2) a gut environment under stress will select for populations with high metabolic independence, potentially leading to an overall reduction in diversity.

To test these hypotheses, we compared genomes reconstructed from a cohort of healthy individuals (Pasolli et al., 2019) to genomes reconstructed from individuals who were diagnosed with inflammatory bowel disease (IBD). Our IBD dataset was composed of two cohorts: a set of patients with pouchitis (Vineis et al., 2016b), a form of IBD with similar pathology to ulcerative colitis (De Preter et al., 2009), and a set of pediatric Crohn's disease patients (Quince et al., 2015). The number of genomes per individual and the average level of genome completeness per group were similar between healthy individuals and those with IBD: overall, our analysis compared 264 genomes from 22 healthy individuals with an average completion of 90.4%, 44 genomes from 4 pouchitis patients with an average completion of 89.2% and 256 genomes from 12 Crohn's disease patients with an average completion of 94.1% (Supplementary Table 3.8). Intriguingly, similar to the size differences between genomes of HMI populations and LMI populations (2.8 Mbp versus 1.6 Mbp on average), genomes of microbial populations associated with IBD patients were larger compared to those of microbial populations in healthy people and averaged to 3.0 Mbp versus 2.6 Mbp, respectively (Supplementary Table 3.8). This suggests that the environmental filters created by FMT and gastrointestinal inflammation both select for microbial populations with larger genomes and potentially higher metabolic independence.

Next, we asked whether the completion of metabolic modules associated with colonization success and resilience during FMT differed between the genomes reconstructed from healthy and IBD individuals. The completion of the 33 metabolic modules was almost identical between the HMI populations revealed by FMT and microbial populations in IBD patients (Wilcoxon rank sum test, p=0.5) (Figure 3.3, Supplementary Table 3.8). In contrast, the completion of these metabolic modules was significantly reduced in microbial populations in healthy individuals (Wilcoxon rank sum test, $p < 1e\text{-}07$) (Figure 3.3, Supplementary Table 3.8). Metabolic modules with the largest differences in completion between genomes from healthy and IBD individuals included biosynthesis of cobalamin, arginine, ornithine, tryptophan, isoleucine as well as the Shikimate pathway (Figure 3.3, Supplementary Table 3.8), a seven step metabolic route bacteria use for the biosynthesis of aromatic amino acids (phenylalanine, tyrosine, and tryptophan) (Herrmann and Weaver, 1999).

Our findings show that the same set of biosynthetic metabolic modules that distinguish good and poor colonizers during FMT were also differentially associated with populations of IBD patients and healthy individuals. In particular, while healthy individuals harbored microbes with a broad spectrum of metabolic capacity, microbes from individuals who suffer from two different forms of IBD had significantly higher biosynthetic independence. It is conceivable that a stable gut microbial ecosystem is more likely to support LMI populations through metabolic cross-feeding, where vitamins, amino acids, and nucleotides are exchanged between microbes (D'Souza et al., 2018). In contrast, host-mediated environmental stress in IBD likely disrupts such interactions and creates an ecological filter that selects for metabolic independence, which subsequently leads to loss of diversity and the dominance of organisms with large genomes that are often not as abundant or as competitive in states of homeostasis.

These observations have implications for our understanding of the hallmarks of healthy gut microbial ecosystems. Defining the 'healthy gut microbiome' has been a major goal of human gut microbiome research (Bäckhed et al., 2012), which still remains elusive (Eisenstein, 2020). Despite comprehensive investigations that considered core microbial taxa (Arumugam et al., 2011; Lloyd-Price et al., 2016b) or guilds of microbes that represent coherent functional groups (Wu et al., 2021a), the search for 'biomarkers' of healthy gut microbiomes is ongoing (McBurney et al., 2019). Our findings indicate that beyond the taxonomic diversity of a microbial community, a broad range of metabolic independence represents a defining feature

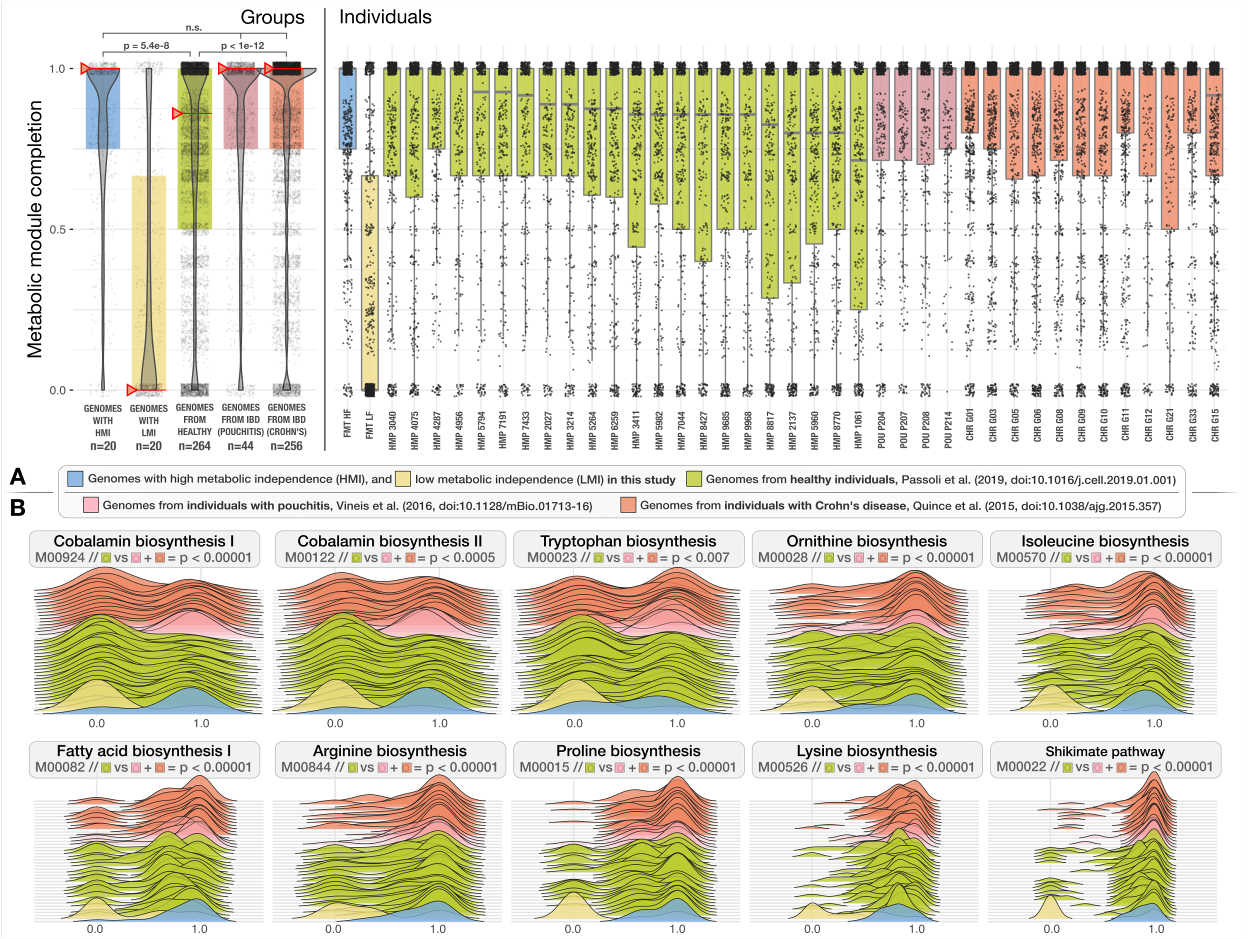


Figure 3.3: Distribution of metabolic modules in genomes reconstructed from healthy individuals and individuals with IBD. The boxplots in the top panels show the metabolic module completion values for (1) high- and (2) low-metabolic independence donor genomes identified in this study (blue and yellow), (3) genomes from healthy individuals (green), and (4) genomes from individuals with pouchitis (red) and Crohn's disease (orange). Each dot in a given box-plot represents one of 33 metabolic modules that were enriched in HMI FMT donor populations and the y-axis indicates its estimated completion. The leftmost top panel represents group averages and red whiskers indicate the median. The rightmost top panel shows the distribution of metabolic modules for individuals within each group. In the bottom panel the completion values for 10 of the 33 pathways are demonstrated as ridge-line plots. Each plot represents a single metabolic module where each layer corresponds to an individual, and the shape of the layer represents the completion of a given metabolic module across all genomes reconstructed from that individual. A high-resolution version of this figure is also available at `https://doi.org/10.6084/m9.figshare.15138720`.

of a healthy gut microbiome. Conversely, our findings also suggest that an enrichment of metabolically independent populations could serve as an indicator of environmental stress in the human gut. Detection of these metabolic markers is not influenced by fluctuations in taxonomic composition or diversity, and represents a quantifiable feature of microbial communities through genome-resolved metagenomic surveys.

Our findings offer a new, taxonomy-independent perspective on the determinants of microbial resilience in the human gut environment under stress. Yet, our study is limited to well-known metabolic pathways – which, given the extent of the unknown coding space in microbial genomes (Vanni et al., 2022), are likely far from complete – as well as by our ability to recognize gene function, which is determined by the sequences described in public databases that favor well-studied microbial organisms (Supplementary Information). Thus, conservatively put, the enrichment of biosynthetic modules in HMI populations suggests that the ability to synthesize essential biological compounds is necessary but likely insufficient to survive environmental stress in the gut. Nevertheless, the finding that the same metabolic modules that promote colonization success after FMT are also the hallmarks of resilience in IBD suggests the presence of unifying ecological principles that govern microbial diversity in distinct modes of stress, which warrants deeper investigation.

3.2.3 Conclusions

Our study identifies high metabolic independence conferred by the biosynthetic capacity for amino acids, nucleotides, and essential micronutrients as a distinguishing hallmark of microbial populations that colonize recipients of FMT and that thrive in IBD patients. These findings highlight the functional complexity of the human gut microbiome whose various interactions with the host are shaped through a network of microbial interactions such as cross-feeding of macro- and micro-nutrients. Our study offers a simple model that posits the following: microbial populations that are metabolically independent and those that lack the means to synthesize essential metabolites co-occur in a healthy gut environment in harmony, where their differential resilience to stress is indiscernible by their taxonomy or relative abundance. However, the challenges associated with the transfer to a new gut environment through FMT, or with host-mediated stress through IBD, initiate an ecological filter that selects for microbes that can self-sustain in the absence of ecosystem services associated with states of homeostasis. This model provides a hypothesis that explains the dominance of low-abundance members of healthy gut environments under stressful conditions, without any necessary direct causal association with disease state. If the association between particular microbial taxa and disease is solely driven by their superior metabolic independence, microbial therapies that aim to treat complex diseases by adding microbes associated with healthy individuals will be unlikely to compete with the adaptive processes that regulate complex gut microbial ecosystems.

3.2.4 Methods

Sample collection and storage. We selected our samples from a subset of individuals who participated in a randomized clinical trial (Kao et al., 2017). Our selection criteria took into consideration multiple factors that were not applicable to all participants of the clinical study. Briefly, we aimed to identify (1) donors that contributed a large number of fecal samples over long periods of time (to maximize the number and quality of genomes from metagenomes and to be able to identify the extent of intrapersonal variability of the microbiota and its potential impact on our results), (2) donors whose feces were transplanted to the largest number of recipients (to be able to discuss the colonization dynamics of the same donor populations in different individuals accurately), (3) multiple recipients for each donor that received FMT via different methods, such as colonoscopy versus pills (to be able to better understand the generalizability of our downstream observations independent of the delivery method), and (4) recipients that were followed the longest period of time after FMT (to be able to follow donor population dynamics accurately). We did not consider factors that may impact the microbial community composition (such as age, gender, or diet) to homogenize the recipient cohort to observe overarching microbial patterns after FMT that are beyond environmental factors dictated by the host. Based on these criteria we identified two donors (DA and DB), and 5 FMT recipients for each donor. All recipients received vancomycin for a minimum of 10 days pre-FMT at a dose of 125 mg four times daily. Three DA and two DB recipients received FMT via pill, and two DA and three DB recipients received FMT via colonoscopy. All recipients had recurrent *C. difficile* infection before FMT, and two DA recipients and one DB recipient were also diagnosed with ulcerative colitis (UC). 24 stool samples were collected from the DA donor over a period of 636 days, and 15 stool samples were collected from the DB donor over a period of 532 days. Between 5 and 9 stool samples were collected from each recipient over periods of 187 to 404 days, with at least one sample collected pre-FMT and 4 samples collected post-FMT. This gave us a total of 109 stool samples from all donors and recipients. Samples were stored at -80oC. (Supplementary Figure 3.4, Supplementary Table 3.1)

Metagenomic short-read sequencing. We extracted the genomic DNA from frozen samples according to the centrifugation protocol outlined in MoBio PowerSoil kit with the following modifications: cell lysis was performed using a GenoGrinder to physically lyse the samples in the MoBio Bead Plates and Solution (5–10 min). After final precipitation, the DNA samples were resuspended in TE buffer and stored at -20 ℃ until further analysis. Sample DNA concentrations were determined by PicoGreen assay. DNA was sheared to 400 bp using the Covaris S2 acoustic platform and libraries were constructed using the Nugen Ovation Ultralow kit. The products were visualized on an Agilent Tapestation 4200 and size-selected using BluePippin (Sage Biosciences). The final library pool was quantified with the Kapa Biosystems qPCR protocol and sequenced on the Illumina NextSeq500 in a 2 × 150 paired-end sequencing run using dedicated read indexing.

'Omics workflows. Whenever applicable, we automated and scaled our 'omics analyses using the bioinformatics workflows implemented by the program 'anvi-run-workflow' (Shaiber et al., 2020b) in anvi'o 7.1 (Eren et al., 2015, 2021b). Anvi'o workflows implement numerous steps of bioinformatics tasks including short-read quality filtering, assembly, gene calling, functional annotation, hidden Markov model search, metagenomic read-recruitment, metagenomic binning, and phylogenomics. Workflows use Snakemake (Köster and Rahmann, 2012) and a tutorial is available at the URL `http://merenlab.org/anvio-workflows/`. The following sections detail these steps.

Taxonomic composition of metagenomes based on short reads. We used Kraken2 v2.0.8-beta (Wood et al., 2019) with the NCBI's RefSeq bacterial, archaeal, viral and viral neighbors genome databases to calculate the taxonomic composition within short-read metagenomes.

Assembly of metagenomic short reads. To minimize the impact of random sequencing errors in our downstream analyses, we used the program 'iu-filter-quality-minoche' to process short metagenomic reads, which is implemented in illumina-utils v2.11 (Eren et al., 2013) and removes low-quality reads according to the criteria outlined by Minoche et al. (Minoche et al., 2011). IDBA_UD v1.1.2 (Peng et al., 2012) assembled quality-filtered short reads into longer contiguous sequences (contigs), although we needed to recompile IDBA_UD with a modified header file so it could process 150bp paired-end reads.

Processing of contigs. We use the following strategies to process both sequences we obtained from our assemblies and those we obtained from reference genomes. Briefly, we used (1) 'anvi-gen-contigs-database' on contigs to compute k-mer frequencies and identify open reading frames (ORFs) using Prodigal v2.6.3 (Hyatt et al., 2010), (2) 'anvi-run-hmms' to identify sets of bacterial (Campbell et al., 2013) and archaeal (Rinke et al., 2013) single-copy core genes using HMMER v3.2.1 (Eddy, 2011), (3) 'anvi-run-ncbi-cogs' to annotate ORFs with functions from the NCBI's Clusters of Orthologous Groups (COGs) (Tatusov et al., 2003), and (4) 'anvi-run-kegg-kofams' to annotate ORFs with functions from the KOfam HMM database of KEGG orthologs (KOs) (Aramaki et al., 2020; Kanehisa and Goto, 2000). To predict the approximate number of genomes in metagenomic assemblies we used the program ‘anvi-display-contigs-stats‘, which calculates the mode of the frequency of single-copy core genes as described previously (Delmont and Eren, 2016).

Metagenomic read recruitment, reconstructing genomes from metagenomes, determination of genome taxonomy and ANI. We recruited metagenomic short reads to contigs using Bowtie2 v2.3.5 (Langmead and Salzberg, 2012) and converted resulting SAM files to BAM files using samtools v1.9 (Li et al., 2009). We profiled the resulting BAM files using the program ‘anvi-profile‘ with the flag ‘--min-contig-length‘ set to 2500 to eliminate shorter sequences to minimize noise. We then used the program ‘anvi-merge‘ to combine all read recruitment profiles into a single anvi’o merged profile database for downstream visualization, binning, and statistical analyses (the DOI 10.6084/m9.figshare.14331236 gives access to reproducible data objects). We then used ‘anvi-cluster-contigs‘ to group contigs into 100 initial bins using CONCOCT v1.1.0 (Alneberg et al., 2014), ‘anvi-refine‘ to manually curate initial bins with conflation error based on tetranucleotide frequency and differential coverage signal across all samples, and ‘anvi-summarize‘ to report final summary statistics for each gene, contig, and bin. We used the program ‘anvi-rename-bins‘ to identify bins that were more than 70% complete and less than 10% redundant, and store them in a new collection as metagenome-assembled genomes (MAGs), discarding lower quality bins from downstream analyses. GTBD-tk v0.3.2 (Chaumeil et al., 2019) assigned taxonomy to each of our MAGs using GTDB r89 (Parks et al., 2018), but to assign species- and subspecies-level taxonomy for ‘DA_MAG_00057‘, ‘DA_MAG_00011‘, ‘DA_MAG_00052‘ and ‘DA_MAG_00018‘, we used ‘anvi-get-sequences-for-hmm-hits‘ to recover DNA sequences for bacterial single-copy core genes that encode ribosomal proteins, and searched them in the NCBI’s nucleotide collection (nt) database using BLAST (Altschul et al., 1990). Finally, the program ‘anvi-compute-genome-similarity‘ calculated pairwise genomic average nucleotide identity (gANI) of our genomes using PyANI v0.2.9 (Pritchard et al., 2016).

Criteria for MAG detection in metagenomes. Using mean coverage to assess the occurrence of populations in a given sample based on metagenomic read recruitment can yield misleading insights, since this strategy cannot accurately distinguish reference sequences that represent very low-abundance environmental populations from those sequences that do not represent an environmental population in a sample yet still recruit reads from non-target populations due to the presence of conserved genomic regions. Thus, we relied upon the ‘detection’ metric, which is a measure of the proportion of the nucleotides in a given sequence that are covered by at least one short read. We considered a population to be detected in a metagenome if anvi’o reported a detection value of at least 0.25 for its genome (whether it was a metagenome-assembled or isolate genome). Values of detection in metagenomic read recruitment results often follow a bimodal distribution for populations that are present and absent (see Supplementary Figure 2 in ref. (Utter et al., 2020)), thus 0.25 is an appropriate cutoff to eliminate false-positive signal in read recruitment results for populations that are absent.

Identification of MAGs that represent multiple subpopulations. To identify subpopulations of MAGs in metagenomes, we used the anvi’o command ‘anvi-gen-variability-profile‘ with the ‘--quince-mode‘ flag which exported single-nucleotide variant (SNV) information for all MAGs after read recruitment. We then used DESMAN v2.1.1 (Quince et al., 2017) to analyze SNVs to determine the number and distribution of subpopulations represented by a single genome. To account for non-specific mapping that can inflate the number of estimated subpopulations, we removed any subpopulation that made up less than 1% of the entire population explained by a single MAG. To account for noise due to low coverage, we only investigated subpopulations for MAGs for which the mean non-outlier coverage of single-copy core genes was at least 10X.

Criteria for colonization of a recipient by a MAG for colonization dynamics analyses (Supplementary Information). We applied the set of criteria described in Supplementary Figure 4 to determine whether or not a MAG successfully colonized a recipient, and to confidently assign colonization or non-colonization phenotypes to each MAG/recipient pair where the MAG was detected in the donor sample used for transplant into the recipient. If these criteria were met, we then determined whether the MAG was detected in any post-FMT recipient sample taken more than 7 days after transplant. If not, the MAG/recipient pair was considered a non-colonization event. If the MAG was detected in the recipient greater than 7 days post-FMT, we used subpopulation information to determine if any subpopulation present in the donor and absent in the recipient pre-FMT was detected in the recipient more than 7 days post-FMT. If this was the case, we considered this to represent a colonization event. See Supplementary Figure 4 for a complete outline of all possible cases.

Phylogenomic tree construction. To concatenate and align amino acid sequences of 46 single-copy core (Campbell et al., 2013) ribosomal proteins that were present in all of our *Bifidobacterium* MAGs and reference genomes, we ran the anvi'o command 'anvi-get-sequences-for-hmm-hits' with the '--return-best-hit', '--get-aa-sequence' and '--concatenate' flags, and the '--align-with' flag set to 'muscle' to use MUSCLE v3.8.1551 (Edgar, 2004) for alignment. We then ran 'anvi-gen-phylogenomic-tree' with default parameters to compute a phylogenomic tree using FastTree 2.1 (Price et al., 2010).

Analysis of metabolic modules and enrichment. We calculated the level of completeness for a given KEGG module (Kanehisa et al., 2014, 2017) in our genomes using the program 'anvi-estimate-metabolism', which leveraged previous annotation of genes with KEGG orthologs (KOs) (see the section 'Processing of contigs'). Then, the program 'anvi-compute-functional-enrichment' determined whether a given metabolic module was enriched in a group of genomes based on the output from the program 'anvi-estimate-metabolism'. The URL `https://anvio.org/m/anvi-estimate-metabolism` serves a tutorial for this program which details the modes of usage and output file formats. The statistical approach for enrichment analysis is defined elsewhere (Shaiber et al., 2020b), but briefly it computes enrichment scores for functions (or metabolic modules) within groups by fitting a binomial generalized linear model (GLM) to the occurrence of each function or complete metabolic module in each group, and then computing a Rao test statistic, uncorrected p-values, and corrected q-values. We considered any function or metabolic module with a q-value less than 0.05 to be 'enriched' in its associated group if it was also at least 75% complete and present in at least 50% of the group members.

Determination of MAGs representing good and poor colonizers for metabolic enrichment analysis. We classified MAGs as good colonizers if, in all 5 recipients, they were detected in the donor sample used for transplantation as well as the recipient more than 7 days post-FMT. We classified MAGs as poor colonizers as those that, in at least 3 recipients, were detected in the donor sample used for FMT but were not detected in the recipient at least 7 days post-FMT. We reduced the number of good colonizer MAGs to be the same as the number of poor colonizer MAGs for metabolic enrichment analysis by selecting only those populations that were the most prevalent in the Canadian gut metagenomes.

Classification of high metabolic independence. We developed a script to calculate the pathwise completeness of the 33 KEGG modules that were enriched in good colonizers in this study to determine whether a given genome resembles HMI or LMI populations. The URL `https://anvio.org/m/anvi-script-estimate-metabolic-independence` serves more information.

Ordination plots. We used the R vegan v2.4-2 package 'metaMDS' function to perform nonmetric multidimensional scaling (NMDS) with Horn-Morisita dissimilarity distance to compare taxonomic composition between donor, recipient, and global metagenomes. We visualized ordination plots using R ggplot2.

3.2.5 Availability of Data and Materials

Raw sequencing data for donor and recipient metagenomes are stored under the NCBI BioProject PRJNA701961 (see Supplementary Table 3.1 for accession numbers for each sample)(Watson et al., 2021). The geographically distributed human gut metagenomes were obtained from previously published datasets (Supplementary Table 3.4) (Zeevi et al., 2015; Le Chatelier et al., 2013; Li et al., 2014; of Sydney, 2016b; Feng et al., 2015; Raymond et al., 2016; David et al., 2015; Xie et al., 2016; Brito et al., 2016; Obregon-Tito et al., 2015; Rampelli et al., 2015; Liu et al., 2016; Wen et al., 2017; Qin et al., 2012; Human Microbiome Project Consortium, 2012a; Pasolli et al., 2019). The URL https://merenlab.org/data/fmt-gut-colonization serves a reproducible bioinformatics workflow and gives access to ad hoc scripts, usage instructions, and intermediate data objects to reproduce findings in our study. Supplementary tables are accessible also via doi:10.6084/m9.figshare.14138405.

3.2.6 Supplementary Figures

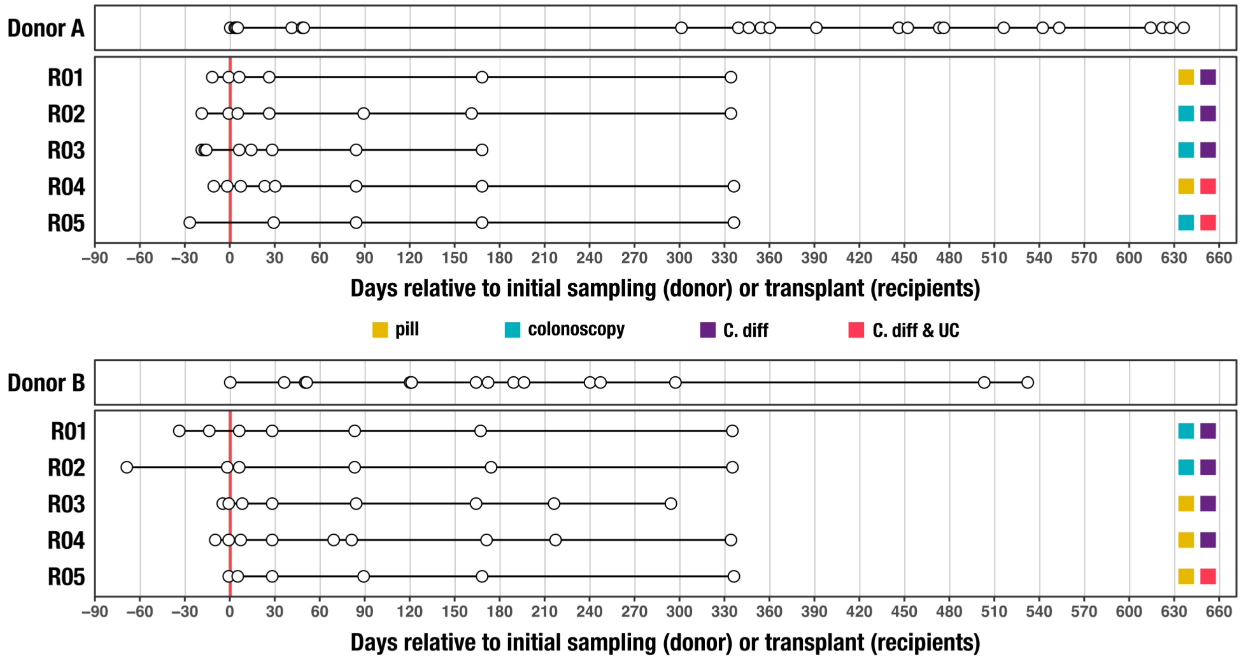


Figure 3.4: Timeline of stool samples collected from FMT study. Each circle represents a stool sample collected from either an FMT donor or FMT recipient. The thicker, red vertical line at day 0 represents the FMT event for each recipient. FMT method (pill or colonoscopy) and FMT recipient health and disease state (C. diff - chronic recurrent *Clostridium difficile* infection, UC - ulcerative colitis) are indicated on the right.

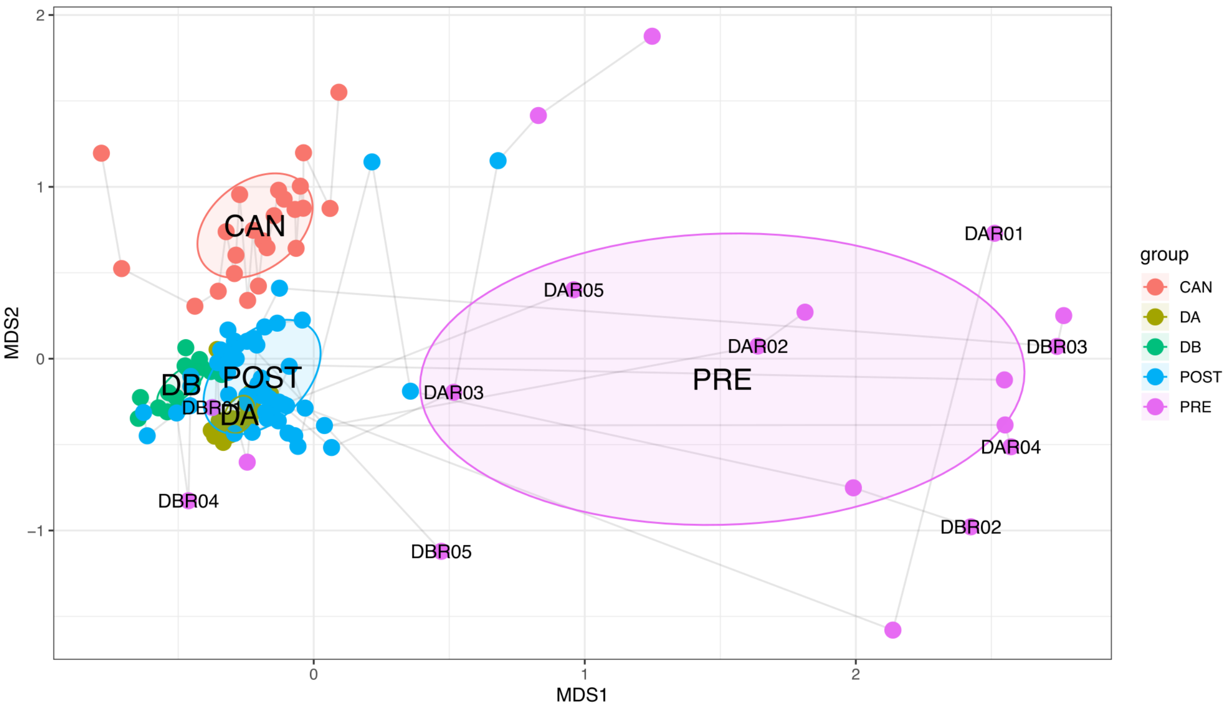


Figure 3.5: Nonmetric multidimensional scaling (NMDS) ordination of the taxonomic composition of donor, recipient, and Canadian gut metagenomes at the genus level based on Morisita-Horn dissimilarity. Samples from the same participant are joined by lines with the earliest time point labeled. CAN: Canadian gut metagenomes, DA: donor A, DB: donor B, POST: recipients post-FMT, PRE: recipients pre-FMT.

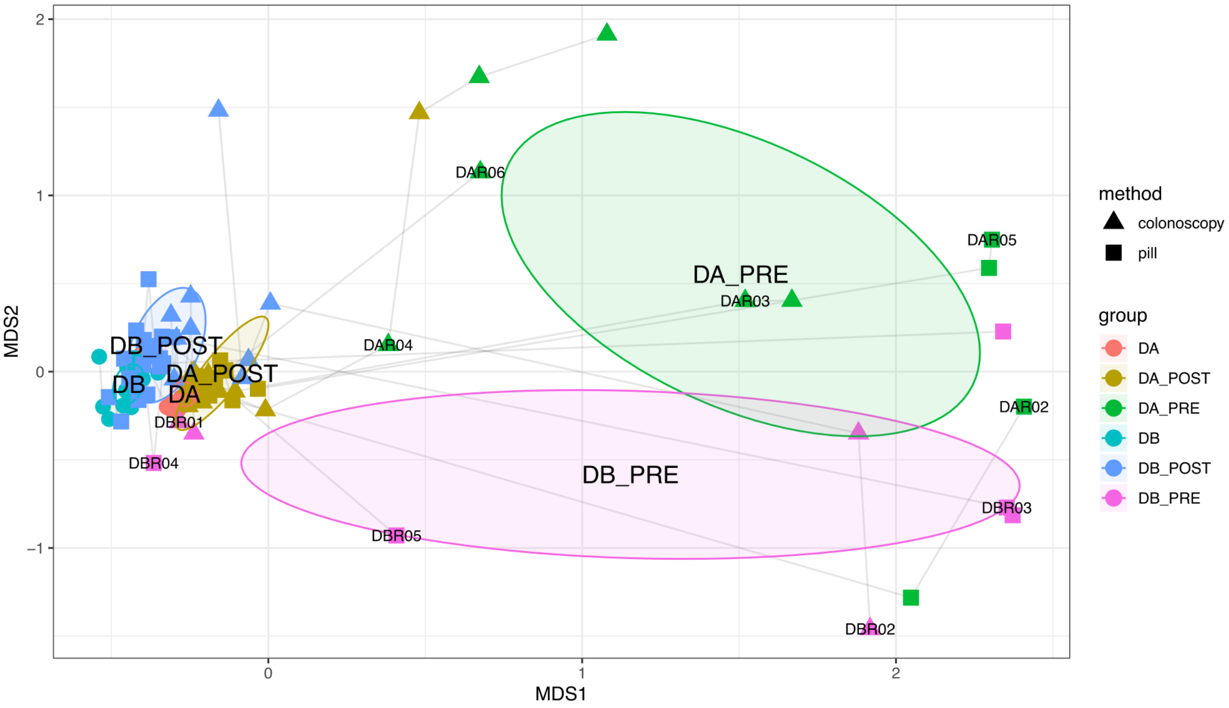


Figure 3.6: Nonmetric multidimensional scaling (NMDS) ordination of the taxonomic composition of the donor and recipient metagenomes at genus level based on Morisita-Horn dissimilarity. Samples from the same participant are joined by lines with the earliest time point labeled. DA_POST: donor A recipients post-FMT, DA_PRE: donor A recipients pre-FMT, DA: donor A, DB_POST: donor B recipients post-FMT, DB_PRE: donor B recipients pre-FMT, DB: donor B.

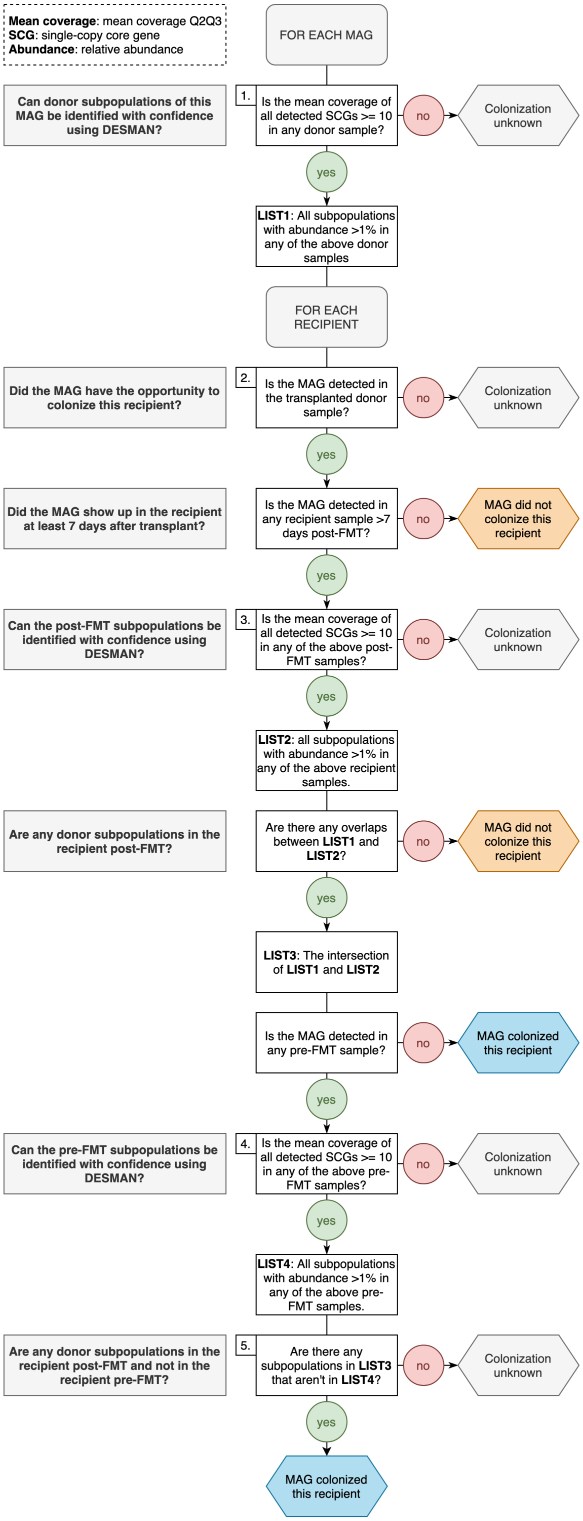

Mean coverage: mean coverage Q2Q3
SCG: single-copy core gene
Abundance: relative abundance
FOR EACH MAG
Can donor subpopulations of this MAG be identified with confidence using DESMAN?
1.
Is the mean coverage of all detected SCGs >= 10 in any donor sample?
no
Colonization unknown
yes
LIST1: All subpopulations with abundance >1% in any of the above donor samples
FOR EACH RECIPIENT
Did the MAG have the opportunity to colonize this recipient?
2.
Is the MAG detected in the transplanted donor sample?
no
Colonization unknown
yes
Did the MAG show up in the recipient at least 7 days after transplant?
Is the MAG detected in any recipient sample >7 days post-FMT?
no
MAG did not colonize this recipient
yes
Can the post-FMT subpopulations be identified with confidence using DESMAN?
3.
Is the mean coverage of all detected SCGs >= 10 in any of the above post-FMT samples?
no
Colonization unknown
yes
LIST2: all subpopulations with abundance >1% in any of the above recipient samples.
Are any donor subpopulations in the recipient post-FMT?
Are there any overlaps between LIST1 and LIST2?
no
MAG did not colonize this recipient
yes
LIST3: The intersection of LIST1 and LIST2
Is the MAG detected in any pre-FMT sample?
no
MAG colonized this recipient
yes
Can the pre-FMT subpopulations be identified with confidence using DESMAN?
4.
Is the mean coverage of all detected SCGs >= 10 in any of the above pre-FMT samples?
no
Colonization unknown
yes
LIST4: All subpopulations with abundance >1% in any of the above pre-FMT samples.
Are any donor subpopulations in the recipient post-FMT and not in the recipient pre-FMT?
5.
Are there any subpopulations in LIST3 that aren't in LIST4?
no
Colonization unknown
yes
MAG colonized this recipient

Figure 3.7: A flowchart outlining our method to assign successful colonization, failed colonization, or undetermined colonization phenotypes to donor-derived populations in the recipients of that donor's stool.

3.2.7 Supplementary Tables

This section's supplementary tables are accessible via doi:10.6084/m9.figshare.14138405.

Table 3.1: Description of FMT study and stool samples collected. a) Description of FMT donor stool samples and SRA accession numbers. b) Description of FMT recipient samples and SRA accession numbers. c) Description of transplantation events.

Table 3.2: Description of FMT metagenomes and co-assemblies. a) Metagenome SRA accession numbers and number of metagenomic short-reads sequenced and mapped to co-assemblies and MAGs. b) Phylum level taxonomic composition of metagenomes. c) Genus level taxonomic composition of metagenomes. d) Summary statistics for contigs from metagenome co-assemblies.

Table 3.3: Description of MAGs. a) Summary statistics and taxonomic assignments for MAGs. b) and c) Detection of Donor A and Donor B MAGs in FMT metagenomes, respectively. d) and e) Detection of Donor A and Donor B MAGs in global gut metagenomes, respectively. f) and g) Detection summary statistics of Donor A and Donor B MAGs in global gut metagenomes, respectively. h) and i) Mean non-outlier coverage of Donor A and Donor B MAG single-copy core genes in FMT metagenomes.

Table 3.4: Accession numbers of publicly-available gut metagenomes from 17 countries.

Table 3.5: MAG subpopulation information. a) and b) Number of Donor A and Donor B MAG subpopulations detected in FMT metagenomes, respectively. c) and d) Subpopulation composition of Donor A and Donor B MAGs in FMT metagenomes, respectively.

Table 3.6: MAG/recipient pair colonization outcomes and MAG mean coverage in the 2nd and 3rd quartiles in stool samples used for transplantation.

Table 3.7: Description of HMI vs. LMI populations. a) Taxonomic assignments and genome size estimates for high- and low-metabolic independence populations. b) KEGG module completeness information for high- and low-metabolic independence populations. c) Raw KEGG module enrichment information for high- and low-metabolic independence populations. d) KEGG module enrichment and categorical information for the 33 modules enriched in high-metabolic independence populations. e) and f) Completeness information for the 33 modules enriched in high-metabolic independence populations in all high- and low-metabolic independence populations.

Table 3.8: Description of genomes from healthy individuals and individuals with IBD. a) List of genomes from healthy individuals and individuals with IBD. b) Module completion values across genomes.

3.3 The more the better? Number of accessory functions predicts colonization success within the genus *Bifidobacterium*

3.3.1 Introduction

Host-microbe interactions influence human health and wellbeing by contributing to immune tolerance (Gensollen et al., 2016), gut barrier integrity (Tan et al., 2016; Schroeder et al., 2018) and cellular energy metabolism (Donohoe et al., 2011). Dysbiotic states of the gut microbiota are associated with diseases and disorders ranging from inflammatory bowel disease to neurological disorders (Durack and Lynch, 2019), and manipulation of the microbial composition and activity to improve health represents an area of intense research (Zuo and Ng, 2018; Aggeletopoulou et al., 2019). Administration of live microorganisms is a commonly used strategy, and microbial strains are generally chosen based on their taxonomic affiliation and prevalence with healthy adult microbiota for their potential to facilitate beneficial host-microbe interactions (Hill et al., 2014; Guslandi, 2022). However, most clinical trials do not observe restoration of homeostasis or substantial and reproducible health benefits through probiotics (Kristensen et al., 2016; Guslandi, 2022). Variable outcomes in clinical trials may be related to variable engraftment success of probiotics (Kristensen et al., 2016; Washburn et al., 2022), suggesting that the ability to engineer effective therapies critically depends on an understanding of the determinants of colonization success.

Members of the genus *Bifidobacterium* are widely used as probiotics (Sharma et al., 2021) as they are prevalent in the human gut microbiome and are intimately linked with immune system regulation (Arboleya et al., 2016). Yet, the ability to successfully establish in a new gut environment differs between members of the genus *Bifidobacterium*. In a recent fecal microbiota transplantation (FMT) study conducted with participants from Canada, we transferred the same human stool from a healthy donor to five unrelated individuals (Watson et al., 2023). In the donor, the genus *Bifidobacterium* was the second most abundant genus (14.1%) after

Bacteroides (15.8%), and was represented by three *Bifidobacterium* species: *B. longum*, *B. adolescentis* subsp. *adolescentis*, and *B. animalis* subsp. *lactis*. To investigate the reason for this discrepancy, we conducted a functional pangenomic analysis of the *Bifidobacterium* genus. We observed largely overlapping core metabolic capabilities, yet substantial differences in the size and content of the accessory genome that yield a striking correlation with the degree of colonization success and prevalence of each taxon across unrelated humans. However, comparison of these results to published literature revealed significant gaps in our analysis, which likely stem from annotation bias (due to the poor characterization of bifidobacteria in functional databases). This cautionary tale demonstrates the limitations of reference databases for homology-based annotation of non-model organisms, and indicates that our current understanding of metabolism is insufficient to differentiate between closely-related organisms.

3.3.2 Observations

For each of the *Bifidobacterium* populations, we reconstructed high-quality genomes from donor metagenomes (Supplementary Table 3.9). While each population was present in the donor gut environment over a 20-month period, they showed vastly different colonization efficiency in recipients after FMT (Figure 3.8). Populations of *B. longum* and *B. adolescentis* subsp. *adolescentis* (henceforth *B. adolescentis*) colonized most recipients, yet *B. animalis* subsp. *lactis* (henceforth *B. lactis*) did not persist in any of the recipients (Supplementary Table 3.9). Overall, *B. longum*, *B. adolescentis*, and *B. lactis* populations occurred in 83%, 79%, and 4% of all 24 post-FMT recipient metagenomes, respectively (Figure 3.8). Surprisingly, the patterns of colonization after FMT were reflected in the prevalence of these populations in publicly available gut metagenomes of healthy humans from 17 countries and *B. lactis* was also less prevalent than *B. longum* and *B. adolescentis* in these data (Supplementary Table 3.9). In Canada, *B. longum*, *B. adolescentis*, and *B. lactis* populations occurred in 74%, 39%, and 13% of unrelated individuals, demonstrating a positive relationship (Pearson's correlation of 0.9) between their colonization success and their prevalence (Supplementary Table 3.9). Intriguingly, the *B. lactis* genome we reconstructed was virtually identical (with over 99.99% sequence identity over 99.82% alignment, Supplementary Table 3.9) to *B. lactis* strains that are widely used as probiotics (Jungersen et al., 2014).

On a broad scale, colonization outcomes are influenced by various factors, including ecosystem state and niche availability (McFarland, 2014; Maldonado-Gómez et al., 2016), as well as the metabolic competencies of individual microbes (Watson et al., 2023). As higher ranks of taxonomy represent similar metabolic capabilities (Martiny et al., 2013) and ecological traits (Philippot et al., 2010) among closely related organisms, the distinct colonization success between closely related *Bifidobacterium* populations emerges as an opportunity to investigate more subtle factors that drive colonization outcomes. Here we extended our collection of three *Bifidobacterium* metagenome-assembled genomes (MAGs) with 31 complete isolate genomes from the NCBI (the within-group and across-groups average gANI estimates for all genomes were 99% and 77%, respectively) (Supplementary Table 3.9) to investigate gene-resolved differences within this group.

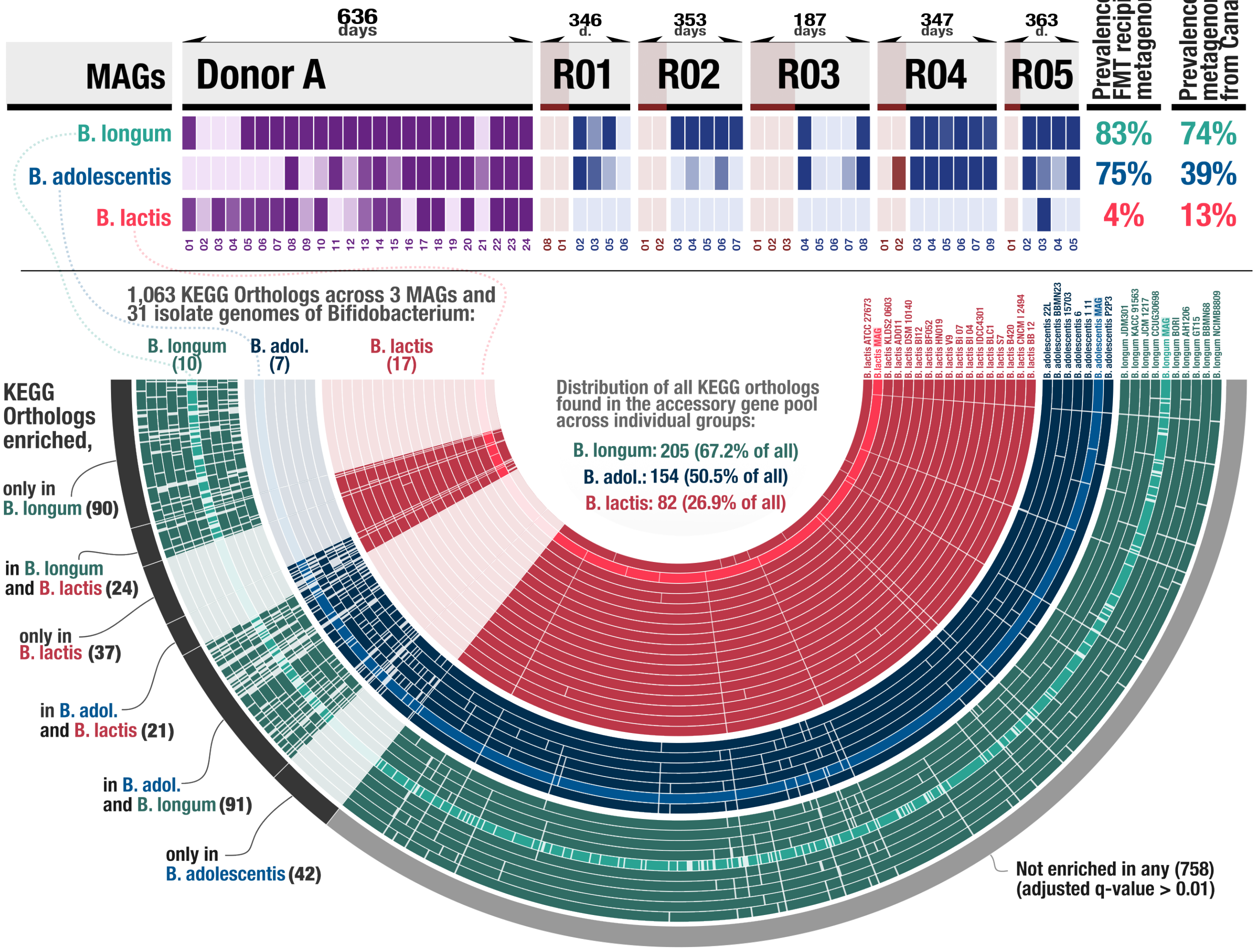


Figure 3.8: Characteristics of three *Bifidobacterium* species. Top panel shows the detection of Donor A MAGs that represent three distinct *Bifidobacterium* populations across donor and recipient metagenomes before and after FMT. The last two columns in this panel show the prevalence of these populations in post-FMT metagenomes, and publicly available gut metagenomes from Canada. The panel below displays the presence or absence of KEGG orthologs within the three *Bifidobacterium* MAGs and 31 high-quality *Bifidobacterium* isolate genomes from the NCBI. Each radii of the concentric semicircles represents a single function assigned by the database of KEGG Orthologs, and each layer of the semicircle is a distinct genome. The intensity of color indicates the presence of a given function in a given genome. The outermost circle indicates groups of functions that are enriched in various groups of *Bifidobacterium* genomes as well as those functions that are not enriched in any group as they are either in all genomes, or only a very small number of them. Maximum resolution version of this figure is also available at https://doi.org/10.6084/m9.figshare.15138720.

Genome length and accessory functions differentiate *Bifidobacterium* clades

The higher average genome lengths of *B. longum* (2.31 Mbp) and *B. adolescentis* (2.18 Mbp) compared to *B. lactis* (1.94 Mbp) concurs with previous observations of a positive correlation between genome size and colonization success (Watson et al., 2023). Nevertheless, there was no clear signal regarding differentially occurring metabolic pathways among the three groups (Supplementary Table 3.10). However, gene annotations with families of KEGG orthologs (KOfams) (Aramaki et al., 2020) and the Clusters of Orthologous Groups (COGs) (Galperin et al., 2021) revealed a large number of individual functions that differentially occurred between them, where 305 of the 1,168 unique KOfams in the *Bifidobacterium* pangenome were statistically enriched in either one or two groups (Figure 3.8). Of these accessory functions, *B. longum* encoded 205 (67.2%), *B. adolescentis* 154 (50.5%), and *B. lactis* 82 (26.9%) (Figure 3.8, Supplementary Table 3.10), showing a parallel between the fraction of the accessory functions enriched in a taxon and the extent of its colonization ability and prevalence.

Notably, several functions distinguishing *B. adolescentis* and *B. longum* from *B. lactis* were related to stress tolerance, including two multidrug resistance pumps of the ‘multidrug and toxin extrusion’ (MATE) type, three transporters of the major facilitator superfamily (MFS) involved in bile acid tolerance and macrolide efflux, two bile acid:natrium ion symporters, and one proton/chloride ion antiporter conferring acid tolerance (Supplementary Table 3.10). The reduced pool of stress tolerance-associated functions in *B. lactis* may at least in part explain its relatively lower success in colonization and prevalence in our data and predict an even lower colonization success of inflamed, antibiotic-treated or otherwise perturbed gut environments.

Bifidobacterium clades have only minor differences in metabolic capacity

Our analysis of the metabolic capacity of groups of genomes affiliated with *B. longum*, *B. adolescentis* and *B. lactis* identified 46 metabolic pathways that were present in at least one member (Supplementary Table 3.10), 40 of which were encoded by all members. Among all annotated pathways, 25 belonged to a set of 33 metabolic modules previously identified as markers for colonization success during FMT (Watson et al., 2023), only 22 of these were present in *B. lactis* while *B. adolescentis* and *B. longum* encoded 24 and 25 of these 33 pathways, respectively. *B. longum* and *B. adolescentis* encoded pathways for Tetrahydrofolate biosynthesis and the MEP pathway for isoprenoid biosynthesis. *B. longum* was the only species that encoded a pathway for the synthesis of Nicotinamide adenine dinucleotide (NAD), an essential cofactor, suggesting that the other two species may require an alternative way of acquiring NAD; however, of the two canonical NAD salvage genes *pncA* and *nadV* (Gazzaniga et al., 2009), only *pncA* (K08281) was found in all *B. longum* and 3 *B. adolescentis* genomes (Gazzaniga et al., 2009). *B. lactis* also encodes several metabolic pathways that are missing in one or both other species, including pathways for methionine degradation, thiamine salvage, initiation of fatty acid biosynthesis and pyrimidine deoxyribonucleotide biosynthesis. However, none of these pathways were enriched in successful colonizers during FMT (Watson et al., 2023).

Investigation of metabolic potential reveals annotation bias

In the process of investigating metabolic potential, we discovered that *B. lactis* genomes were missing annotations of several enzymes for which they indeed had genes, as indicated by the pangenome results. For example, two KOs in the histidine biosynthesis pathway, K01693 and K02501, were not annotated in these genomes, yet two core gene clusters contained gene sequences annotated with these enzymes. The *B. lactis* gene sequences in these clusters were unannotated despite having enough sequence similarity to resolve to the same cluster. This suggests that the *B. lactis* genomes are systematically missing enzyme annotations; that is, they suffer from annotation bias. The source of this bias can likely be attributed to the nature of gene annotation models, which are generated from a limited set of homologous reference sequences. Homologs from taxonomic clades not included in the set may be too different from the resulting model to match it with high similarity scores, thereby resulting in dropped annotations. Indeed, the bit score thresholds computed by KEGG for K01693 and K02501 were too high for the corresponding *B. lactis* genes to be annotated with these enzymes. This had the downstream effect of false negatives in our metabolism estimation results: the histidine biosynthesis pathway (which requires these two enzymes) was incorrectly predicted to be incomplete in *B. lactis*. In response to this discovery, we developed a heuristic for restoring valid annotations that nevertheless have bit scores lower than the pre-computed threshold (Methods). We then re-annotated all genomes using this heuristic and re-did all of the analyses described in this paper for greater accuracy; the observations given in the preceding sections result from the corrected annotations.

3.3.3 Limitations

Despite our attempt to mitigate the observed annotation bias in *B. lactis* genomes, our functional analysis results were still skewed by incomplete annotations and the limited diversity of metabolic pathways described by KEGG. We missed a number of important functions previously studied in bifidobacteria (as a reviewer kindly pointed out). For example, *B. lactis* has fewer glycan utilization capabilities than the other two clades (Milani et al., 2016, 2013), but this was not shown in our functional enrichment or metabolism estimation results. Furthermore, annotation bias can result in misleading metabolism estimates, as demonstrated by the pre-heuristic result that *B. lactis* lacks a histidine biosynthesis pathway. Our annotation heuristic for KOfams is not necessarily enough to make up for the lack of characterization of these genomes in KEGG, nor does it apply to other functional databases that likely suffer from similar biases.

3.3.4 Conclusion

This study highlights the inadequacy of our current characterization of non-model organisms in widely-used reference databases, which culminates in annotation bias that precludes the study of subtle differences in functional capacity between closely related populations. We were unable to reveal the secret to differential colonization success of *Bifidobacterium* species. However, this investigation provides an important, cautionary lesson for the study of poorly-characterized microbes: biased input results in biased output, and careful validation of results is paramount.

3.3.5 Methods

Genomes and metagenomes. Raw sequencing data for donor and recipient metagenomes are stored under the NCBI BioProject PRJNA701961 (see Supplementary Table 3.9 for accession numbers for each sample). The URL `https://merenlab.org/data/fmt-gut-colonization` serves a reproducible bioinformatics workflow and gives access to ad hoc scripts, usage instructions, and intermediate data objects to reproduce findings in our study. Supplementary datasets are also accessible via doi:10.6084/m9.figshare.14138405. For detailed information of study design, sample acquisition and processing, as well as genome reconstruction from metagenomic short reads, please refer to (Watson et al., 2023) and the reproducible bioinformatics workflow at `https://merenlab.org/data/fmt-gut-colonization/`.

Metabolism analysis. We estimated the metabolic capacities encoded in the MAGs and reference genomes using anvi'o. First, all genomes were converted into anvi'o contigs databases with 'anvi-gen-contigs-database', which included a gene-calling step using Prodigal (Hyatt et al., 2010). We annotated the genes with NCBI's Clusters of Orthologous Groups (COGs) (Galperin et al., 2021) and with KEGG KOfams (Aramaki et al., 2020), using 'anvi-run-ncbi-cogs' and 'anvi-run-kegg-kofams', respectively. Note that we used a version of KEGG downloaded in April 2022 (for reproducibility, the hash of the KEGG snapshot available via ‘anvi-setup-kegg-kofams‘ is 666feeac5de2). ‘anvi-run-kegg-kofams‘ includes a heuristic for annotating hits with bitscores that are just below the KEGG-defined threshold, which is described here: `https://anvio.org/help/main/programs/anvi-run-kegg-kofams/#how-does-it-work`. Finally, we estimated completeness of the metabolic pathways in each genome by running ‘anvi-estimate-metabolism‘, and we computed enrichment scores for pathways across the three *Bifidobacterium* species using ‘anvi-compute-metabolic-enrichment‘. For these analyses, we used the ‘pathwise’ completeness metric, which takes the maximum completeness over all possible combinations of enzymes for a given metabolic pathway.

Pangenomic analysis. We computed and visualized a functional pangenome of MAGs and reference genomes using anvi’o. Briefly, we stored all processed MAG and reference genome contigs in an anvi’o database using the command ‘anvi-gen-genomes-storage‘. To create and visualize the KOfam functional pangenome, we then passed that database to the command ‘anvi-display-functions‘, which uses function names to aggregate gene annotations into clusters and also computes the enrichment of functions within genome groups using the script ‘anvi-compute-functional-enrichment-across-genomes‘ with a lambda parameter of 0 (for more details about the enrichment calculation, see (Shaiber et al., 2020a)). We set the ‘anvi-display-functions‘ ‘--min-occurrence‘ flag to 3 to remove gene clusters only present in one (singletons) or two genomes. We computed functional enrichment for COGs in the same way.

3.3.6 Supplementary Tables

This section's supplementary tables are accessible via doi:10.6084/m9.figshare.22579219.

Table 3.9: Description of Bifidobacteria genomes. a) Accession numbers for Bifidobacteria reference genomes. b) Detection of Bifidobacteria MAGs in FMT metagenomes. c) Sample information and SRA accession numbers for publicly-available gut metagenomes used in this study. d) Detection of Bifidobacteria MAGs in global gut metagenomes. e) Prevalence of Bifidobacteria MAGs in global gut metagenomes. f) gANI percent identity between Bifidobacteria genomes. g) gANI percent alignment coverage between Bifidobacteria genomes. h) Bifidobacteria reference genomes from NCBI. i) Summary statistics for Bifidobacteria MAGs and reference genomes.

Table 3.10: Bifidobacteria functional analysis. a) List of KEGG modules that are complete in at least one *Bifidobacterium* genome. b) KEGG module completeness scores in each *Bifidobacterium* genome. c) KEGG modules enriched in different Bifidobacteria species. d) KOfam presence and absence in Bifidobacteria genomes. e) KOfams enriched in different Bifidobacteria species. f) COG function presence and absence in Bifidobacteria genomes. g) COG functions enriched in different Bifidobacteria species.

CHAPTER 4

MICROBES WITH HIGHER METABOLIC INDEPENDENCE ARE ENRICHED IN HUMAN GUT MICROBIOMES UNDER STRESS

4.1 Preface

This chapter is derived from a study investigating the metabolic potential of the human gut microbiome in individuals diagnosed with inflammatory bowel disease, or IBD. IBD is a class of gastrointestinal (GI) disorders characterized by chronic inflammation of the GI tract. These conditions are typically accompanied by a reduction in gut microbial diversity, but the determinants of microbial survival in the IBD gut environment are yet unknown. This study leverages the metabolism reconstruction framework in a high-throughput analysis of publicly-available human gut metagenomes to determine the relevance of metabolic capabilities to microbial resilience in this system. It introduces novel methodology for normalizing community-level estimates of pathway copy number with estimated community sizes to obtain per-population copy numbers, which provide an appropriate metric for comparing metabolic potential between communities of variable size without resorting to the time-intensive alternative of metagenomic binning followed by analysis of individual populations.

In accordance with the results from the previous chapter, this study finds that high metabolic independence represents a distinguishing characteristic of microbial populations associated with individuals diagnosed with IBD. Furthermore, a classifier, which is trained on metabolism data that captures the extent of metabolic independence in a metagenome, is not only able to reliably identify samples from IBD patients but also to track recovery of the gut microbiome following antibiotic treatment. These results suggest that high metabolic independence is a general hallmark of stressed gut environments and may be an interesting target for the dev-elopment of microbiota-based diagnostic tools and therapies.

This chapter is derived from the following publication:

Iva Veseli, Yiqun T. Chen, Matthew S. Schechter, Chiara Vanni, Emily C. Fogarty, Andrea R Watson, Bana Jabri, Ran Blekhman, Amy D. Willis, Michael K. Yu, Antonio Fernàndez-Guerra, Jessika Füssel, and A Murat Eren. Microbes with higher metabolic independence are enriched in human gut microbiomes under stress. *bioRxiv*. May 15, 2023. `https://doi.org/10.1101/2023.05.10.540289`

4.2 Introduction

The human gut is home to a diverse assemblage of microbial cells that form complex communities (Coyte et al., 2015). This gut microbial ecosystem is established almost immediately after birth and plays a lifelong role in human wellbeing by contributing to immune system maturation and functioning (Belkaid and Hand, 2014; Maynard et al., 2012), extracting dietary nutrients (Hijova, 2019), providing protection against pathogens (Khosravi and Mazmanian, 2013), metabolizing drugs (Zimmermann et al., 2019), and more (Knight et al., 2017). There is no universal definition of a healthy gut microbiome (Fan and Pedersen, 2021), but associations between host disease states and changes in microbial community composition have sparked great interest in the therapeutic potential of gut microbes (Cani, 2018; Sorbara and Pamer, 2022) and led to the emergence of hypotheses that directly link disruptions of the gut microbiome to non-communicable diseases of complex etiology (Byndloss and Bäumler, 2018).

Inflammatory bowel diseases (IBDs), which describe a heterogeneous group of chronic inflammatory disorders (Shan et al., 2022), represent an increasingly common health risk around the globe (Kaplan, 2015). Understanding the role of gut microbiota in IBD has been a major area of focus in human microbiome research. Studies focusing on individual microbial taxa that typically change in relative abundance in IBD patients have proposed a range of host-microbe interactions that may contribute to disease manifestation and progression (Joossens et al., 2011; Schirmer et al., 2019; Henke et al., 2019; Machiels et al., 2014). However, even within well-constrained cohorts, a large proportion of variability in the taxonomic composition of the microbiota is unexplained, and the proportion of variability explained by disease status is low (Gevers et al., 2014; Schirmer et al., 2018b; Lloyd-Price et al., 2019; Khan et al., 2019). As neither individual taxa nor broad changes in microbial community composition yield effective predictors of disease (Knox et al., 2019b; Lee and Chang, 2021), the role of gut microbes in the etiology of IBD – or the extent to which they are bystanders to disease – remains unclear (Khan et al., 2019).

The marked decrease in microbial diversity in IBD is often associated with the loss of Firmicutes populations and an increased representation of a relatively small number of taxa, such as Bacteroides, Enterococcaceae, and others (Prindiville et al., 2000; Saitoh et al., 2002; Sartor, 2006; Rhodes, 2007; Devkota et al., 2012; Machiels et al., 2014; Vineis et al., 2016a; Lloyd-Price et al., 2019). Why a handful of taxa that also typically occur in healthy individuals in lower abundances (Lee and Chang, 2021; Nishida et al., 2018) tend to dominate the IBD microbiome is a fundamental but open question to gain insights into the ecological underpinnings of the gut microbial ecosystem under IBD. Going beyond taxonomic summaries, a recent metagenome-wide metabolic modeling study revealed a significant loss of cross-feeding partners as a hallmark of IBD, where microbial interactions were disrupted in IBD-associated microbial communities compared to those found in healthy individuals (Marcelino et al., 2023). This observation is in line with another recent work that proposed that the extent of 'metabolic independence' (characterized by the genomic presence of a set of key metabolic modules for the synthesis of essential nutrients) is a determinant of microbial survival in IBD (Watson et al., 2023). It is conceivable that the disrupted metabolic interactions among microbes observed in IBD (Marcelino et al., 2023) indicates an environment that lacks the ecosystem services provided by a complex network of microbial interactions, and selects for those organisms that harness high metabolic independence (HMI) (Watson et al., 2023). This interpretation offers an ecological mechanism to explain the dominance of populations with specific metabolic features in IBD. However, this proposed mechanism warrants further investigation.

Here we implemented a high-throughput strategy to estimate metabolic capabilities of microbial communities directly from metagenomes and investigate whether the enrichment of populations with high metabolic independence predicts IBD in the human gut. We benchmarked our findings using representative genomes associated with the human gut and their distribution in healthy individuals and those who have been diagnosed with IBD. Our results suggest that high metabolic potential (indicated by a set of 33 largely biosynthetic metabolic pathways) provides enough signal to consistently distinguish gut microbiomes under stress from those that are in homeostasis, providing deeper insights into adaptive processes initiated by stress conditions that promote rare members of gut microbiota to dominance during disease.

4.3 Results and Discussion

We compiled 2,893 publicly-available stool metagenomes from 13 different studies, 5 of which explicitly studied the IBD gut microbiome (Supplementary Table 4.1a-c). The average sequencing depth varied across individual datasets (4.2 Mbp to 60.3 Mbp, with a median value of 21.4 Mbp, Supplementary Table 4.1c). To improve the sensitivity and accuracy of our downstream analyses that depend on metagenomic assembly, we excluded samples with less than 25 million reads, resulting in a set of 408 relatively deeply-sequenced metagenomes from 10 studies (26.4 Mbp to 61.9 Mbp, with a median value of 37.0 Mbp, Supplementary Table 4.1b, Supplementary Information, Methods), which we *de novo* assembled individually. The final dataset included individuals who were healthy (n=229), diagnosed with IBD (n=101), or suffered from other gastrointestinal conditions ("non-IBD", n=78). In accordance with previous observations of reduced microbial diversity in IBD (Kostic et al., 2014; Nagalingam and Lynch, 2012; Knox et al., 2019b), the estimated number of populations based on the occurrence of bacterial single-copy core genes present in these metagenomes was higher in healthy individuals than those diagnosed with IBD (Supplementary Figure 4.6, Supplementary Table 4.1).

4.3.1 Estimating normalized copy numbers of metabolic pathways from metagenomic assemblies

Gaining insights into microbial metabolism requires accurate estimates of pathway presence and completion. While a myriad of tools address this task for single genomes (Machado et al., 2018; Aziz et al., 2008; Arkin et al., 2018b; Palù et al., 2022; Shaffer et al., 2020; Geller-McGrath et al., 2023; Zorrilla et al., 2021; Zhou et al., 2022; Zimmermann et al., 2021), working with complex environmental metagenomes poses additional challenges due to the large number of organisms that are present in metagenomic assemblies. A few tools can estimate community-level metabolic potential from metagenomes without relying on the reconstruction of individual population genomes or reference-based approaches (Ye and Doak, 2009; Karp et al., 2021) (Supplementary Table 4.5). These high-level summaries of pathway presence and redundancy in a given environment are suitable for most surveys of metabolic capacity, particularly for microbial communities of similar richness. However, since the frequency of observed metabolic modules increases as microbial diversity increases, investigations of metabolic determinants of survival across environmental conditions with substantial differences in microbial richness requires quantitative insights into the extent of enrichment of metabolic capabilities in relation to microbial diversity. For instance, the estimated copy number of a given metabolic module may be identical between two metagenomes, but one metagenome can have a lower alpha diversity and thus have a higher selection for this module. To quantify the differential abundance of metabolic modules between metagenomes generated from healthy individuals and those from individuals diagnosed with IBD, we implemented a new software framework (https://anvio.org/m/anvi-estimate-metabolism) that reconstructs metabolic modules from genomes and metagenomes and then calculates the per-population copy number (PPCN) of modules in metagenomes (Methods, Supplementary Information). Briefly, the PPCN estimates the proportion of microbes in a community with a particular metabolic capacity (Figure 4.1, Supplementary Figure 4.7). We estimate the number of microbial populations using single-copy core genes (SCGs) instead of reconstructing individual genomes first, thus maximizing the *de novo* recovery of gene content.

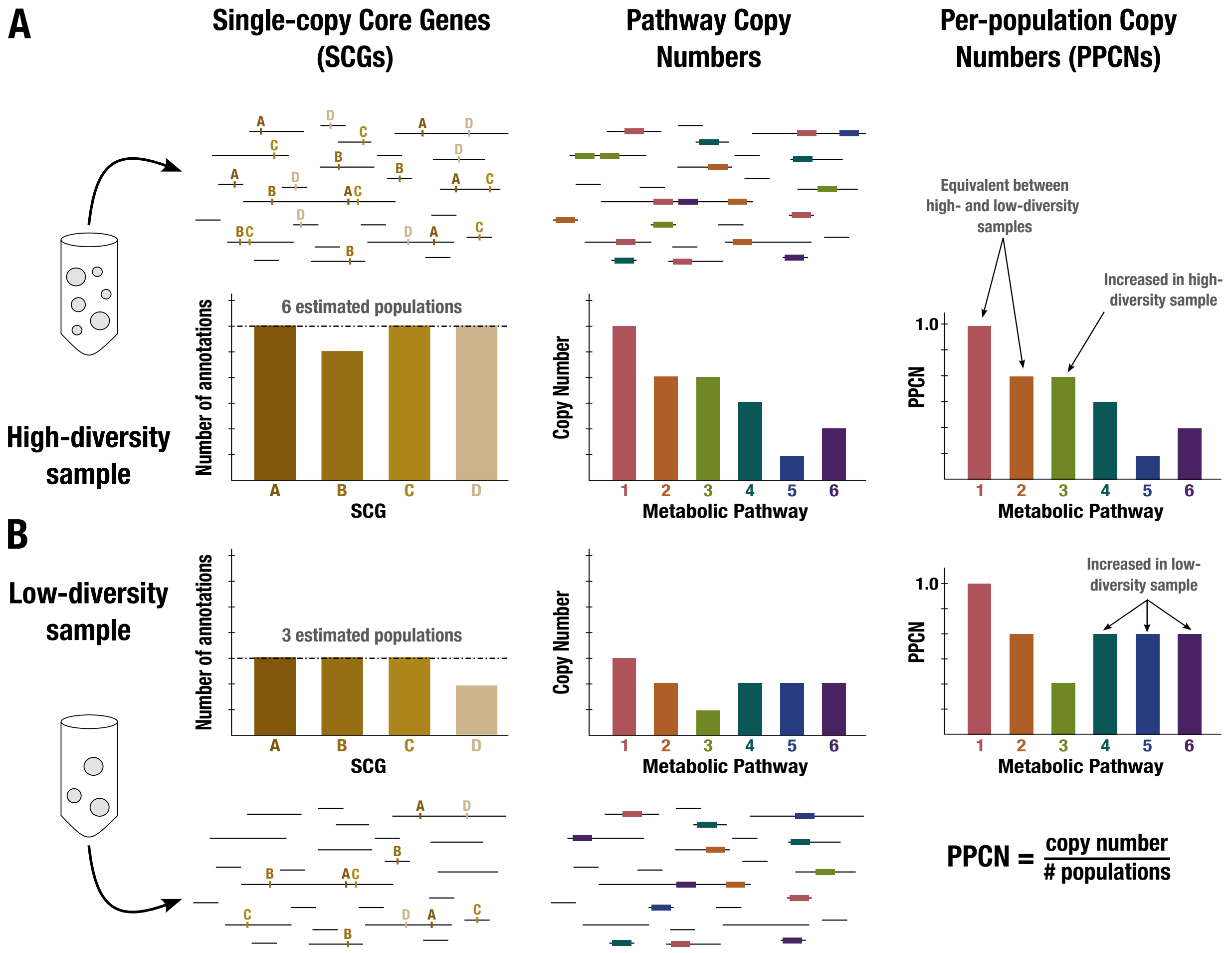

A
Single-copy Core Genes (SCGs)
Pathway Copy Numbers
Per-population Copy Numbers (PPCNs)
High-diversity sample
6 estimated populations
Number of annotations
A B C D
SCG
Copy Number
1 2 3 4 5 6
Metabolic Pathway
Equivalent between high- and low-diversity samples
Increased in high-diversity sample
1.0
PPCN
B
Low-diversity sample
3 estimated populations
Increased in low-diversity sample
PPCN = copy number / # populations

Figure 4.1: Conceptual diagram of per-population copy number (PPCN) calculation. Each step of the calculation is demonstrated in (A) for a sample with high diversity (6 microbial populations) and in (B) for a sample with low diversity (3 populations). Metagenome sequences are shown as black lines. The left panel shows the single-copy core genes annotated in the metagenome (indicated by letters), with a barplot showing the counts for different SCGs. The dashed black line indicates the mode of the counts, which is taken as the estimate of the number of populations. The middle panel shows the annotations of metabolic pathways (indicated by boxes and numerically labeled), with a barplot showing the copy number of each pathway (for more details on how this copy number is computed, see Supplementary Information and Supplementary Figure 4.7). The right panel shows the equation for per-population copy number (PPCN), with the barplots indicating the PPCN values for each metabolic pathway in each sample and arrows differentiating between different types of modules based on the comparison of their normalized copy numbers between samples.

4.3.2 Key biosynthetic pathways are enriched in microbial populations from IBD samples

To gain insight into potential metabolic determinants of microbial survival in the IBD gut environment, we assessed the distribution of metabolic modules within samples from each group (IBD and healthy) with and without using PPCN normalization. A set of 33 metabolic modules were significantly enriched in metagenomes obtained from individuals diagnosed with IBD when PPCN normalization was applied (Figure 4.2d, 4.2e). Each metabolic module had an FDR-adjusted $p < 2e-10$ and an effect size > 0.12 from a Wilcoxon Rank Sum Test comparing IBD and healthy samples. The set included 17 modules that were previously associated with high metabolic independence (Watson et al., 2023) (Figure 4.2f). However, without PPCN normalization, the signal was masked by the overall higher copy numbers in healthy samples, and the same analysis did not detect higher metabolic potential in microbial populations associated with individuals diagnosed with IBD (Figure 4.2a), showing weaker differential occurrence between cohorts (Figure 4.2b, 4.2c, Supplementary Figure 4.8). This result suggests that the PPCN normalization is an important step in comparative analyses of metabolisms between samples with disparate levels of diversity.

The majority of the metabolic modules that were enriched in the microbiomes of IBD patients encoded biosynthetic capabilities (23 out of 33) that resolved to amino acid metabolism (33%), carbohydrate metabolism (21%), cofactor and vitamin biosynthesis (15%), nucleotide biosynthesis (12%), lipid biosynthesis (6%) and energy metabolism (6%) (Supplementary Table 4.2a). In contrast to previous reports based on reference genomes (Gevers et al., 2014; Morgan et al., 2012), amino acid synthesis and carbohydrate metabolism were not reduced in the IBD gut microbiome in our dataset. Rather, our results were in accordance with a more recent finding that predicted amino acid secretion potential is increased in the microbiomes of individuals with IBD (Heinken et al., 2021).

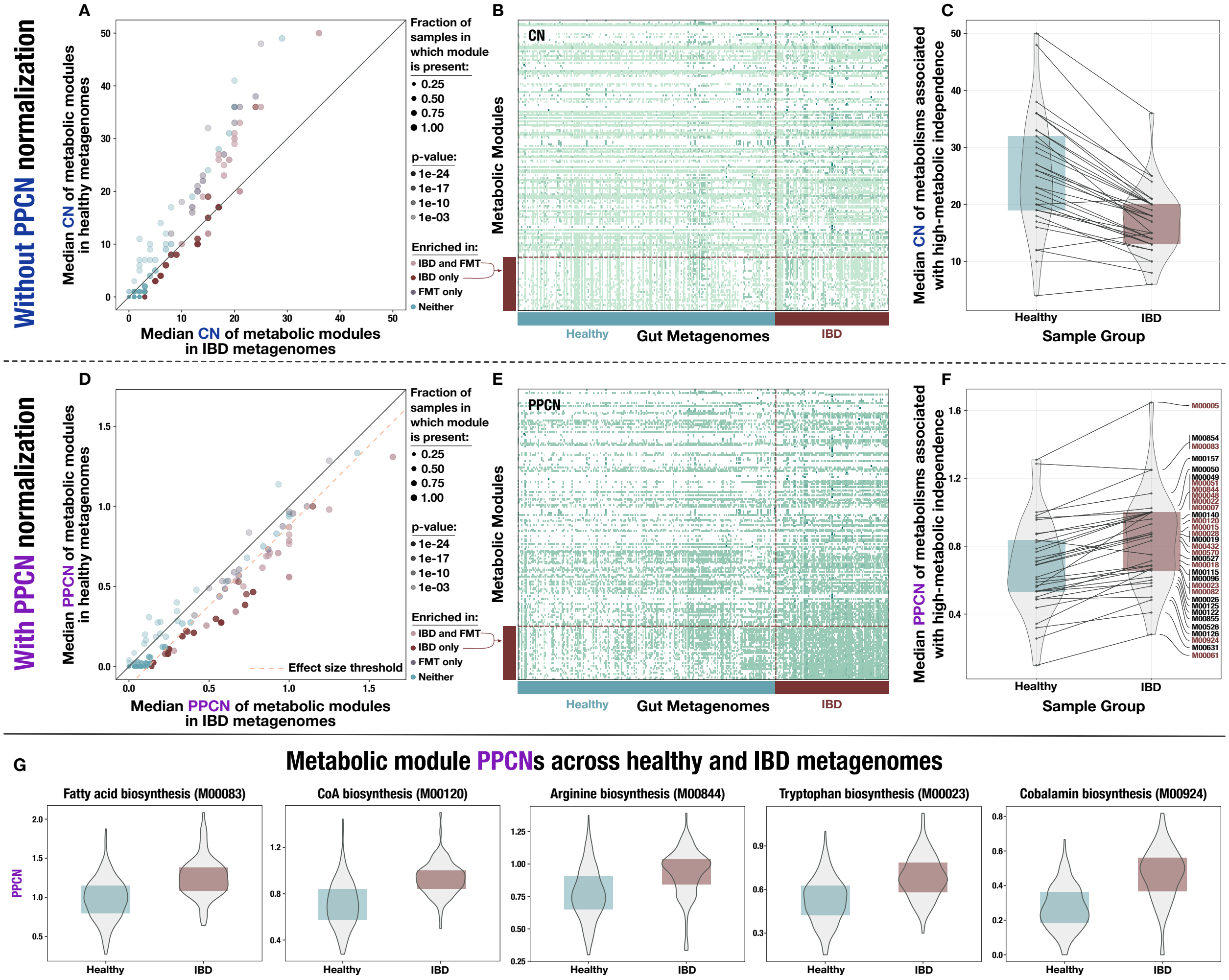


Figure 4.2: Comparison of metabolic potential across healthy and IBD cohorts. Panels A – C show unnormalized copy number data and the remaining panels show normalized per-population copy number (PPCN) data. A) Scatterplot of module copy number in IBD samples (x-axis) and healthy samples (y-axis). Transparency of points indicates the p-value of the module in a Wilcoxon Rank Sum test for enrichment (based on PPCN data), and color indicates whether the module is enriched in the IBD samples (in this study), enriched in the good colonizers from the fecal microbiota transplant (FMT) study (Watson et al., 2023), or enriched in both. B) Heatmap of unnormalized copy numbers for all modules. IBD-enriched modules are highlighted by the red bar on the left. Sample group is indicated by the blue (healthy) and red (IBD) bars on the bottom. C) Boxplots of median copy number for each module enriched in the FMT colonizers from (Watson et al., 2023) in the healthy samples (blue) and the IBD samples (red). Solid lines connect the same module in each plot. D) Scatterplot of module PPCN values in IBD samples (x-axis) and healthy samples (y-axis). Transparency and color of points are defined as in panel (A). The pink dashed line indicates the effect size threshold applied to modules when determining their enrichment in IBD. E) Heatmap of PPCN values for all modules. Side bars defined as in (B). F) Boxplots of median PPCN values for modules

Figure 4.2 continued: enriched in the FMT colonizers from (Watson et al., 2023) in the healthy samples (blue) and the IBD samples (red). Lines defined as in (D). Modules that were also enriched in the IBD samples (in this study) are highlighted in red. G) Boxplots of PPCN values for individual modules in the healthy samples (blue) and the IBD samples (red). All example modules were enriched in both this study and in (Watson et al., 2023).

The metagenome-level enrichment of several key biosynthesis pathways supports the hypothesis that high metabolic independence (HMI) is a determinant of survival for microbial populations in the IBD gut environment. We investigated whether biosynthetic capacity in general was enriched in IBD samples, and 62 out of 88 (70%) biosynthesis pathways described in the KEGG database had a significant enrichment in the IBD sample group at an FDR-adjusted 5% significance level (Supplementary Figure 4.10d) . However, a similar proportion of non-biosynthetic pathways, 63 out of 91 (69%), were also significantly increased in the IBD samples. While biosynthetic capacity is not over-represented in the IBD sample group compared to other types of metabolism, the high proportion of enriched pathways associated with biosynthesis suggests that biosynthetic capacity is important for microbial resilience.

Within our set of 33 pathways that were enriched in IBD, it is notable that all the biosynthesis and central carbohydrate pathways are directly or indirectly linked via shared enzymes and metabolites. Each enriched module shared on average 25.6% of its enzymes and 40.2% of metabolites with the other enriched modules, and overall 18.2% of enzymes and 20.4% of compounds across these pathways were shared (Supplementary Table 4.2a). Thus, modules may be enriched not just due to the importance of their immediate end products, but also because of their role in the larger metabolic network. The few standalone modules that were enriched included the efflux pump MepA and the beta-Lactam resistance system, which are associated with drug resistance. These capacities may provide an advantage since antibiotics are a common treatment for IBDs (Nitzan et al., 2016), but are not related to the systematic enrichment of biosynthesis pathways that likely provide resilience to general environmental stress rather than to a specific stressor such as antibiotics.

While so far we divided samples into two groups, our dataset also includes individuals who do not suffer from IBD, yet are not healthy either. A recent study using flux balance analysis to model metabolite secretion potential in the dysbiotic, non-dysbiotic, and control gut communities of Crohn's Disease patients has shown that several predicted microbial metabolic activities align with gradients of host health (Heinken et al., 2021). This observation suggests that the signal for HMI should also follow a similar gradient with the inclusion of the non-IBD group with other gastrointestinal conditions in our analyses. Our analysis of these data showed that the set of 78 samples classified as ‘non-IBD’ indeed represent an intermediate group between healthy individuals and those diagnosed with IBD (Supplementary Figure 4.10b). 75% of the pathways that were significantly enriched in the IBD group compared to the healthy group were also significantly enriched in the non-IBD group compared to the healthy group. We could further confirm this observation by sorting each individual cohort along a health gradient based on cohort descriptions in their respective studies (Supplementary Information), where the relative proportion of metabolic pathways indicative of HMI increased as a function of increasing disease severity (Supplementary Figure 4.11a). These findings suggest that the enrichment of HMI populations are proportional to gradients in host health, revealing a potential utility of the extent of HMI as a diagnostic tool to monitor changing stress levels in a single individual over time.

Microbiome data generated by different groups can result in systematic biases that may outweigh biological differences between otherwise similar samples (Lozupone et al., 2013; Sinha et al., 2017; Clausen and Willis, 2022). The potential impact of such biases constitutes an important consideration for meta-analyses such as ours that analyze publicly available metagenomes from multiple sources. To account for cohort biases, we conducted an analysis of our data on a per-cohort basis, which showed robust differences between the sample groups across multiple cohorts (Supplementary Figure 4.11bc). Another source of potential bias in our results is due to the representation of microbial functions in genomes in publicly available databases. For instance, we noticed that, independent of the annotation strategy, a smaller proportion of genes resolved to known functions in metagenomic assemblies of the healthy samples compared to the assemblies we generated from the IBD group (Supplementary Figure 4.9). This highlights the possibility that healthy samples merely appear to harbor less metabolic capabilities due to missing annotations. Indeed, we found that the normalized copy numbers of most metabolic modules were reduced in the healthy group, where 84% of KEGG modules (98 out of 118) have significantly lower median copy numbers (Supplementary Figure 4.10c, Supplementary Information). While the presence of a bias between the two cohorts is clear, the source of this bias and its implications are not as clear. One hypothesis that could explain this phenomenon is that the increased proportion of unknown functions in environments where populations with low metabolic independence (LMI) thrive is due to our inability to identify distant homologs of even well-studied functions in poorly studied novel genomes through public databases. If true, this would indeed impair our ability to annotate genes using state-of-the-art functional databases, and bias metabolic module completion estimates. Such a limitation would indeed warrant a careful reconsideration of common workflows and studies that rely on public resources to characterize gene function in complex environments. Another hypothesis that could explain our observation is that the general absence in culture of microbes with smaller genomes (that likely fare better in diverse gut ecosystems) had a historical impact on the characterization of novel functions that represent a relatively larger fraction of their gene repertoire. If true, this would suggest that the unknown functions are unlikely essential for well-studied metabolic capabilities. Furthermore, HMI and LMI genomes may be indistinguishable with respect to the distribution of such novel genes, but the increased number of genes in HMI genomes that resolve to well-studied metabolisms would reduce the proportion of known functions in LMI genomes, and thus in metagenomes where they thrive. While testing these hypotheses falls outside the scope of our work, we find the latter hypothesis more likely due to examples in existing literature that have successfully identified genes that belong to known metabolisms in some of the most obscure organisms via annotation strategies similar to those we have used in our work (Jaffe et al., 2020; Farag et al., 2020).

Taken together, these results (1) demonstrate that the PPCN normalization is an important consideration for investigations of metabolic enrichment in complex microbial communities as a function of microbial diversity, and (2) reveal that the enrichment of HMI populations in an environment offers a high resolution marker to resolve different levels of environmental stress.

4.3.3 Reference genomes with higher metabolic independence are over-represented in the gut metagenomes of individuals with IBD

So far, our findings demonstrate an overall, metagenome-level trend of increasing HMI within gut microbial communities as a function of IBD status without considering the individual genomes that contribute to this signal. Since the extent of metabolic independence of a microbial genome is a quantifiable trait, we considered a genome-based approach to validating our findings. Given the metagenome-level trends, we expected that the microbial genomes that encode a high number of metabolic modules associated with HMI should be more commonly detected in metagenomes from individuals diagnosed with IBD.

While publicly available reference genomes for microbial taxa will unlikely capture the diversity of individual gut metagenomes, we cast a broad net by surveying the ecology of 19,226 genomes in the Genome Taxonomy Database (GTDB) (Parks et al., 2022) that belonged to three major phyla associated with the human gut environment: Bacteroidetes, Firmicutes, and Proteobacteria (Woting and Blaut, 2016; Turnbaugh et al., 2009). We then used Human Microbiome Project data (Human Microbiome Project Consortium, 2012a) to characterize the distribution of these genomes across healthy human gut metagenomes. We used their single-copy core genes to identify genomes that were representative of microbial clades that are systematically detected in the healthy human gut (Figure 4.3a) and kept those that also occurred in at least 2% of samples from our set of 330 healthy and IBD metagenomes (see Methods). By selecting for genomes that are relatively well-detected in the HMP dataset, this filtering step effectively removed genomes representing taxa that primarily occur outside of the human gut. Of the final set of 338 reference genomes that passed our filters, 258 (76.3%) resolved to Firmicutes, 60 (17.8%) to Bacteroidetes, and 20 (5.9%) to Proteobacteria. Most of these genomes resolved to families common to the colonic microbiota, such as *Lachnospiraceae* (30.0%), *Ruminococcaceae* / *Oscillospiraceae* (23.1%), and *Bacteroidaceae* (10.1%) (Arumugam et al., 2011), while 5.9% belonged to poorly-studied families with temporary code names (Supplementary Table 4.3a). Finally, we performed a more comprehensive read recruitment analysis on this smaller set of genomes using all deeply-sequenced metagenomes from cohorts that included healthy, non-IBD, and IBD samples (Figure 4.3). This provided us with a quantitative summary of the detection patterns of GTDB genome representatives common to the human gut across our dataset.

We classified each genome as HMI if its average completeness of the 33 HMI-associated metabolic pathways was at least 80%, equivalent to a summed metabolic independence score of 26.4 (Methods). Across all genomes, the mean metabolic independence score was 24.0 (Q1: 19.9, Q3: 25.7). We identified 17.5% (59) of the reference genomes as HMI. HMI genomes were on average substantially larger (3.8 Mbp) than non-HMI genomes (2.9 Mbp) and encoded more genes (3,634 vs. 2,683 genes, respectively), which is in accordance with the reduced metabolic potential of non-HMI populations (Supplementary Table 4.3a). Our read recruitment analysis showed that HMI reference genomes were present in a significantly higher proportion of IBD samples compared to non-HMI genomes (Figure 4.3c, $p < 1e\text{-}5$, Wilcoxon Rank Sum test). Similarly, the fraction of HMI populations was significantly higher within a given IBD sample compared to 'non-IBD' samples and those from healthy individuals (Figure 4.3d, $p < 1e\text{-}24$, Kruskal-Wallis Rank Sum test). In contrast, the detection of HMI populations and non-HMI populations was similar in healthy individuals (Figure 4.3c, $p = 0.267$, Wilcoxon Rank Sum test). The intestinal environment of healthy individuals likely supports both HMI and non-HMI populations, wherein 'metabolic diversity' is maintained by metabolic interactions such as cross feeding. Indeed, loss of cross-feeding interactions in the gut microbiome appears to be associated with a number of human diseases, including IBD (Marcelino et al., 2023).

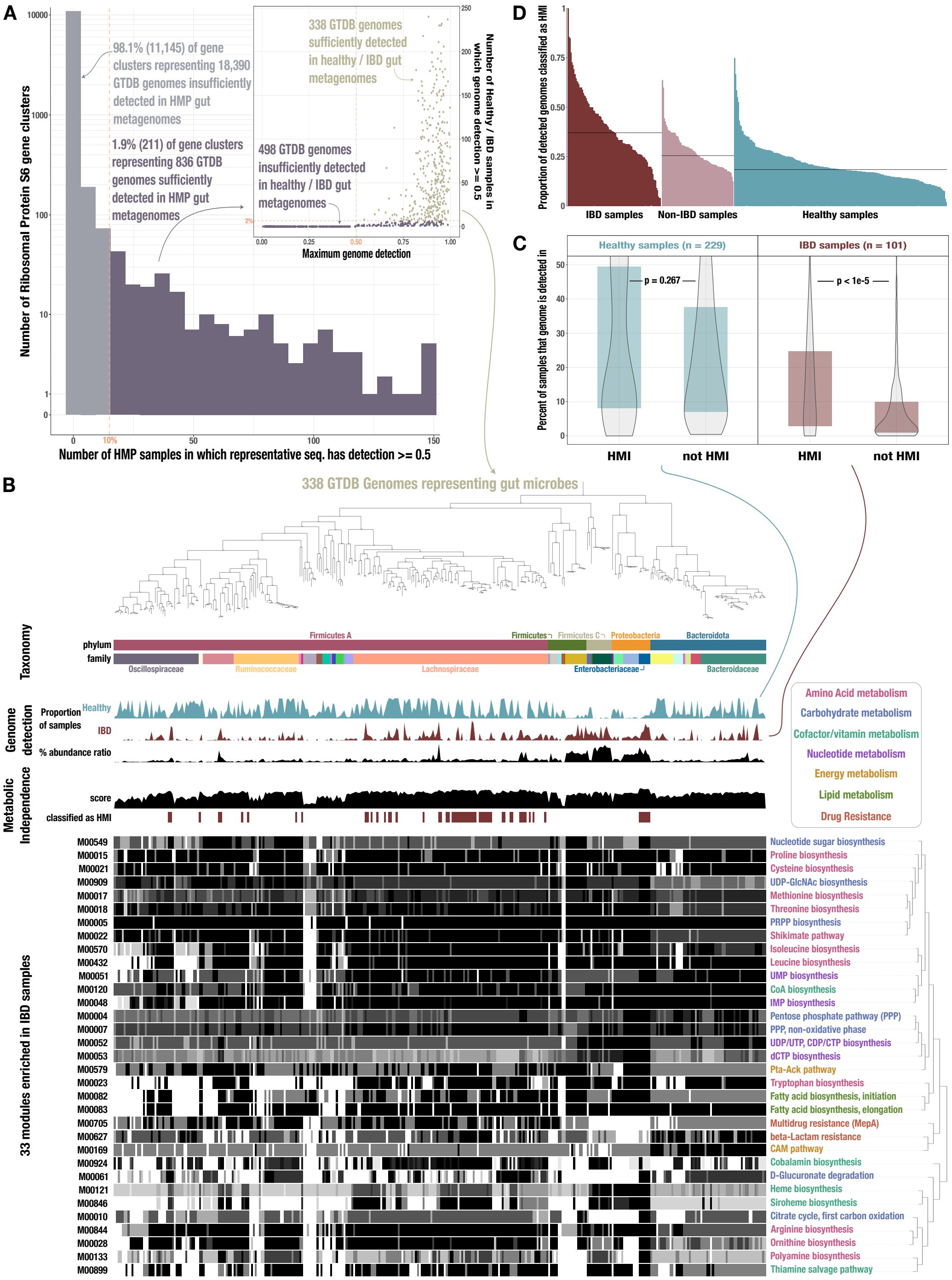


Figure 4.3: Identification of HMI genomes and their distribution across gut samples.

Figure 4.3 continued: A) Histogram of Ribosomal Protein S6 gene clusters (94% ANI) for which at least 50% of the representative gene sequence is covered by at least 1 read (>= 50% 'detection') in fecal metagenomes from the Human Microbiome Project (HMP) (Human Microbiome Project Consortium, 2012a). The dashed line indicates our threshold for reaching at least 50% detection in at least 10% of the HMP samples; gray bars indicate the 11,145 gene clusters that do not meet this threshold while purple bars indicate the 836 clusters that do. The subplot shows data for the 836 genomes whose Ribosomal Protein S6 sequences belonged to one of the passing (purple) gene clusters. The y-axis indicates the number of healthy/IBD gut metagenomes from our set of 330 in which the full genome sequence has at least 50% detection, and the x-axis indicates the genome's maximum detection across all 330 samples. The dashed line indicates our threshold for reaching at least 50% genome detection in at least 2% of samples; the 338 genomes that pass this threshold are tan and those that do not are purple. The phylogeny of these 338 genomes is shown in B) along with the following data, from top to bottom: taxonomic classification as assigned by GTDB; proportion of healthy samples with at least 50% detection of the genome sequence; proportion of IBD samples with at least 50% detection of the genome sequence; square-root normalized ratio of percent abundance in IBD samples to percent abundance in healthy samples; metabolic independence score (sum of completeness scores of 33 HMI-associated metabolic pathways); whether (red) or not (white) the genome is classified as having HMI with a threshold score of 26.4; heatmap of completeness scores for each of the 33 HMI-associated metabolic pathways (0% completeness is white and 100% completeness is black). Pathway name is shown on the right and colored according to its category of metabolism. C) Boxplot showing the proportion of healthy (blue) or IBD (red) samples in which genomes of each class are detected >= 50%, with p-values from a Wilcoxon Rank-Sum test on the underlying data. D) Barplot showing the proportion of detected genomes (with >= 50% genome sequence covered by at least 1 read) in each sample that are classified as HMI, for each group of samples. The black lines show the median for each group: 37.0% for IBD samples, 25.5% for non-IBD samples, and 18.4% for healthy samples.

The three HMI-associated pathways with the largest difference in average completion (>40%) between HMI and non-HMI reference genomes were siroheme biosynthesis, cobalamin biosynthesis, and tryptophan biosynthesis (Supplementary Table 4.3g). Siroheme and cobalamin biosynthesis represent complex pathways that require 6-8 and 11-13 enzymatic steps, respectively, and both compounds belong to the tetrapyrroles that are involved in various essential biological functions (Bryant et al., 2020). Siroheme is a cofactor required for nitrite and sulfite reduction and its biosynthetic pathway provides the precursors required for cobalamin biosynthesis. Genes belonging to biosynthetic pathways of siroheme and cobalamin had higher average relative abundance in infants diagnosed with neonatal necrotizing enterocolitis (Claud et al., 2013), an inflammatory bowel condition affecting premature newborns. The siroheme biosynthesis pathway is upregulated in some human pathogens in response to high nitric oxide (NO) levels likely in relation to the NO detoxification function of nitrite reductase (Porrini et al., 2021). Increased NO levels are commonly associated with active inflammation in IBD (Soufli et al., 2016).

While siroheme is central to sulfite and nitrite reduction in prokaryotes, cobalamin (vitamin B12) is essential not only for the majority of gut microbes (80%) (Kelly et al., 2019; Hossain et al., 2022; Degnan et al., 2014a) but also for the human host, and functions as a coenzyme in key metabolic pathways in humans and bacteria. However, only relatively few gut microbes (20-40%) encode the metabolic pathway for its synthesis (Degnan et al., 2014a; Magnúsdóttir et al., 2015; Kelly et al., 2019) and humans largely rely on cobalamin supplied via their diet. B12 deficiency in humans leads to reduced villi length (Berg et al. 1972) and may affect intestinal barrier functioning (Bressenot et al. 2013). However, microbially-produced cobalamin alone is insufficient to sustain the host's requirements (Magnúsdóttir et al. 2015). The high average completion of this complex pathway in reference genomes classified as HMI (86%) in contrast to non-HMI reference genomes (40%) demonstrates the importance of metabolic independence for the survival of microorganisms in stressed gut environments, whereas in a healthy gut environment cross-feeding of B-vitamins supports non-producers (Magnúsdóttir et al. 2015).

Tryptophan is an essential amino acid that serves as a precursor for a variety of microbial (Alkhalaf and Ryan, 2015) and human metabolites that play a potential role in IBD pathogenesis (Agus et al., 2018). Tryptophan metabolites mediate a variety of host microbe interactions in the human gut (Agus et al., 2018), contribute to gut barrier integrity, and exert anti-inflammatory functions (Bansal et al., 2010; Roager and Licht, 2018). While fecal tryptophan concentrations can be elevated in IBD patients (Jansson et al., 2009), tryptophan host metabolism via the Kynurenine pathway also appears to be elevated, resulting in decreased serum levels of the amino acid (Nikolaus et al., 2017). At the same time, a tryptophan-deficient diet in mice is linked to intestinal inflammation and alterations of the microbial community composition (Hashimoto et al., 2012; Yusufu et al., 2021). While it is not clear whether microbially-produced tryptophan contributes significantly to the host's tryptophan metabolism, the environmental pressure of tryptophan depletion may favor microbial populations with the biosynthetic capacity to produce this amino acid.

Overall, the classification of reference gut genomes as HMI and their enrichment in individuals diagnosed with IBD strongly supports the contribution of HMI to stress resilience of individual microbial populations. We note that survival in a disturbed gut environment will likely require a wide variety of additional functions that are not covered in the list of metabolic modules we consider to determine HMI status – for examples, see (Degnan et al., 2014a; Martens et al., 2014; Zong et al., 2020; Feng et al., 2020; Goodman et al., 2009; Powell et al., 2016). Indeed, there may be many ways for a microbe to be metabolically independent, and our strategy likely failed to identify some HMI populations. Nonetheless, these data suggest that HMI serves as a reliable proxy for the identification of microbial populations that are particularly resilient.

4.3.4 HMI-associated metabolic potential predicts general stress on gut microbes

Our analysis identified HMI as an emergent property of gut microbial communities associated with individuals diagnosed with IBD. This community-level signal translates to individual microbial populations and provides insights into the microbial ecology of stressed gut environments. HMI-associated metabolic pathways were enriched at the community level, and microbial populations encoding these modules were more prevalent in individuals with IBD than in healthy individuals. Furthermore, the copy number of these pathways and the proportion of HMI populations were higher in gut communities associated with more severe host health states, including in non-IBD samples compared to samples from healthy individuals (Supplementary Figure 4.10b, Figure 4.3d). The ecological implications of these observations suggest that HMI may serve as a predictor of general stress in the human gut environment.

So far, efforts to identify IBD using microbial markers have presented classifiers based on (1) taxonomy in pediatric IBD patients (Papa et al., 2012; Gevers et al., 2014), (2) microbial community composition in combination with clinical data (Halfvarson et al., 2017), (3) untargeted metabolomics and/or species-level relative abundance from metagenomes (Franzosa et al., 2019) and (4) k-mer-based sequence variants in metagenomes that can be linked to microbial genomes associated with IBD (Reiter et al., 2022). Performance varied both between and within studies according to the target classes and data types used for training and validation of each classifier (Supplementary Table 4.4a). For those studies reporting accuracy, a maximum accuracy of 77% was achieved based on either metabolite profiles (for prediction of IBD-subtype) (Franzosa et al., 2019) or k-mer-based sequence variants (for differentiating between IBD and non-IBD samples) (Reiter et al., 2022). Some studies reported performance as area under the receiver operating characteristic curve (AUROCC), a typical measure of classifier utility describing both sensitivity (ability to correctly identify the disease) and specificity (ability to correctly identify absence of disease). For this metric the highest value was 0.92, achieved by (Franzosa et al., 2019) when using metabolite profiles, with or without species abundance data, for classifying IBD vs non-IBD. However, the majority of these classifiers were trained and tested on a relatively small group of individuals that all come from the same region, i.e. clinical studies confined to a specific hospital. Though some had high performance, these were based on data that are inaccessible to most laboratories and clinics, considering that untargeted metabolomics analyses are difficult to reproduce (Koek et al., 2011; Lin et al., 2020), and that k-mer-based analyses do not provide specific clinical targets for intervention. These classifiers thus have limited translational potential across global clinical settings. For practical use as a diagnostic tool, a microbiome-based classifier for IBD should rely on an ecologically-meaningful, easy to measure, and high-level signal that is robust to host variables like lifestyle, geographical location, and ethnicity. High metabolic independence could potentially fill this gap as a metric related to the ecological filtering that defines microbial community changes in the IBD gut microbiome.

We trained a logistic regression classifier to explore the applicability of HMI as a non-invasive diagnostic tool for IBD. The classifier's predictors were the per-population copy numbers of IBD-enriched metabolic pathways in a given metagenome. Across the 330 deeply-sequenced IBD and healthy samples included in this analysis, the classifier had high sensitivity and specificity (Figure 4.4). It correctly identified (on average) 76.8% of samples from individuals diagnosed with IBD and 89.5% of samples representing healthy individuals, for an overall accuracy of 85.6% and an average AUROCC of 0.832 (Supplementary Table 4.4c). Our model outperforms (Gevers et al., 2014; Halfvarson et al., 2017; Reiter et al., 2022) or has comparable performance to (Franzosa et al. 2019; Papa et al. 2012) the previous attempts to classify IBD from fecal samples in more restrictively-defined cohorts. It also has the advantage of being a simple model, utilizing a relatively low number of features compared to the other classifiers. Thus, HMI shows promise as an accessible diagnostic marker of IBD. Of course, due to the lack of time-series studies that include individuals in the pre-diagnosis phase of IBD development, we cannot test the applicability of HMI as a predictive marker for early stages of this disease, as discussed in (Lloyd-Price et al., 2019).

Yet, the gradient of metabolic independence reflected by per-population pathway copy number and the proportion of detected HMI populations in non-IBD samples (Supplementary Figure 4.10b, Figure 4.3d) suggests that the degree of HMI in the gut microbiome may be predictive of general gut stress, such as that induced by antibiotic use. Antibiotics can cause long-lasting perturbations of the gut microbiome – including reduced diversity, emergence of opportunistic pathogens, increased microbial load, and development of highly-resistant strains – with potential implications for host health (Ramirez et al., 2020). We applied our metabolism classifier to a metagenomic dataset that reflects the changes in the microbiome of healthy people before, during and up to 6 months following a 4-day antibiotic treatment (Palleja et al., 2018). The resulting pattern of sample classification corresponds to the post-treatment decline and subsequent recovery of species richness documented in the study by (Palleja et al., 2018). All pre-treatment samples were classified as ‘healthy’ followed by a decline in the proportion of ‘healthy’ samples to a minimum 8 days post-treatment, and a gradual increase until 180 days post treatment, when over 90% of samples were classified as ‘healthy’ (Figure 4.5, Supplementary Table 4.4b). These observations support the role of HMI as an ecological driver of microbial resilience during gut stress caused by a variety of environmental perturbations and demonstrate its diagnostic power in reflecting gut microbiome state.

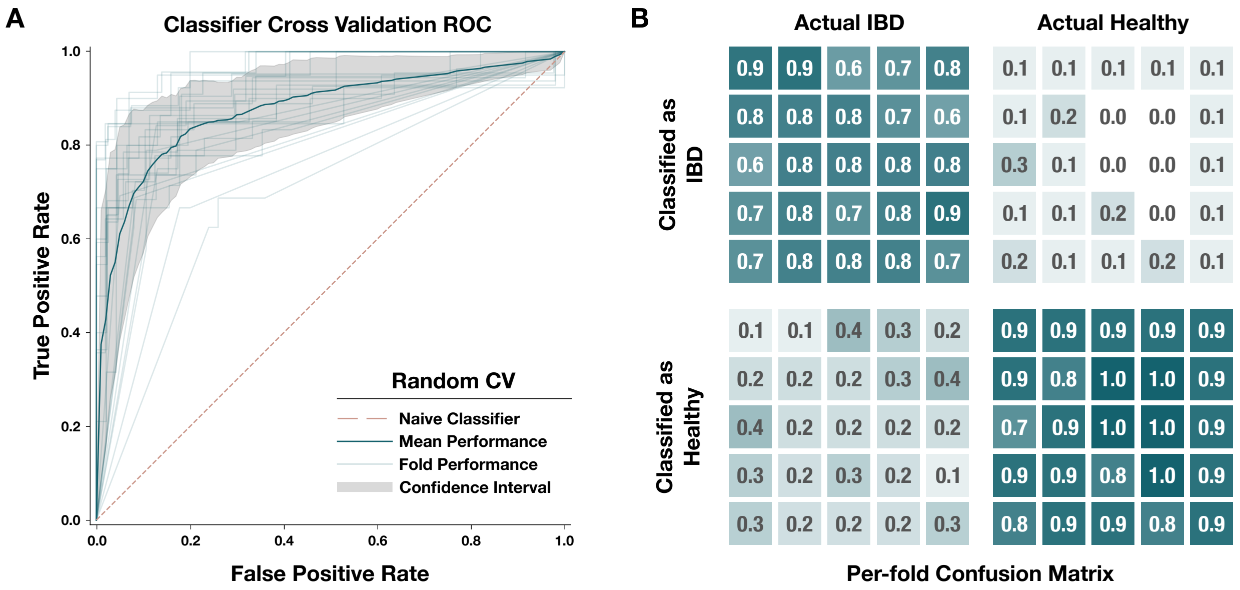


Figure 4.4: Performance of our metagenome classifier trained on per-population copy numbers of IBD-enriched modules. A) Receiver operating characteristic (ROC) curves for 25-fold cross-validation. Each fold used a random subset of 80% of the data for training and the other 20% for testing. In each fold, we calculated a set of IBD-enriched modules from the training dataset and used the PPCN of these modules to train a logistic regression model whose performance was evaluated using the test dataset. Light gray lines show the ROC curve for each fold, the dark blue line shows the mean ROC curve, the gray area delineates the confidence interval for the mean ROC, and the pink dashed line indicates the benchmark performance of a naive (random guess) classifier. B) Confusion matrix for each fold of the random cross-validation. Categories of classification, from top left to bottom right, are: true positives (correctly classified IBD samples), false positives (incorrectly classified Healthy samples), false negatives (incorrectly classified IBD samples), and true negatives (correctly classified Healthy samples). Each fold is represented by a box within each category. Opacity of the box indicates the proportion of samples in that category, and the actual proportion is written within the box with one significant digit. Underlying data for this matrix can be accessed in Supplementary Table 4.4d.

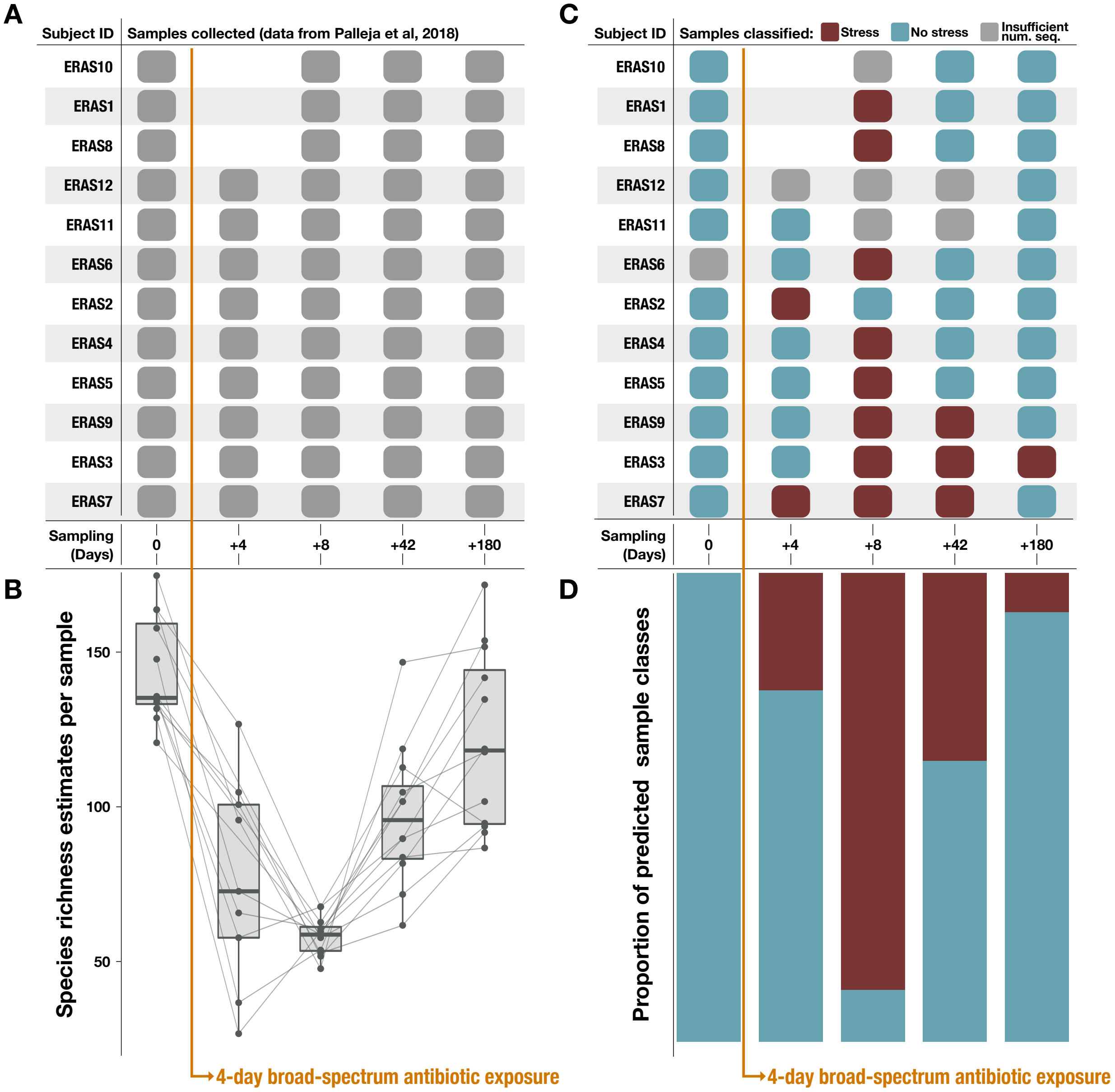


Figure 4.5: Classification results on an antibiotic time-series dataset from (Palleja et al., 2018). Note that antibiotic treatment was taken on days 1-4. A) Samples collected per subject during the time series. B) Species richness data (figure reproduced from (Palleja et al., 2018)). C) Classification of each sample by the metabolism classifier profiled in Figure 4.4. Samples with insufficient sequencing depth were not classified. D) Proportion of classes assigned to samples per day in the time series.

4.4 Conclusions

Overall, our observations that stem from the analysis of hundreds of reference genomes, deeply-sequenced gut metagenomes, and multiple categories of human disease states suggest that environmental stress in the human gut – whether it is associated with inflammation, cancer, or antibiotic use – promotes the survival and relative expansion of microbial populations with high metabolic independence. These results establish HMI as a high-level metric to classify gradients of human health states through the gut microbiota that is robust to ethnic, geographical or lifestyle factors. Taken together with recent evidence that models altered ecological relationships within gut microbiomes under stress due to disrupted metabolic cross-feeding (Heinken et al., 2021; Marcelino et al., 2023), our data support the hypothesis that the reduction in microbial diversity, or more generally 'dysbiosis', is an emergent property of microbial communities responding to disease pathogenesis or other external factors such as antibiotic use that disrupt the gut microbial ecosystem. This paradigm depicts microbes as bystanders by default, rather than perpetrators or drivers of noncommunicable human diseases, and provides an ecological framework to explain the frequently observed reduction in microbial diversity associated with IBD and other noncommunicable human diseases and disorders.

4.5 Methods

A bioinformatics workflow that further details all analyses described below and gives access to reproducible data products is available at the URL `https://merenlab.org/data/ibd-gut-metabolism/`.

A new framework for metabolism estimation. We developed a new program 'anvi-estimate-metabolism' (`https://anvio.org/m/anvi-estimate-metabolism`), which uses gene annotations to estimate 'completeness' and 'copy number' of metabolic pathways that are defined in terms of enzyme accession numbers. By default, this tool works on metabolic modules from the KEGG MODULE database (Kanehisa et al., 2012, 2023) which are defined by KEGG KOfams (Aramaki et al., 2020), but user-defined modules based on a variety of functional annotation sources are also accepted as input. Completeness estimates describe the percentage of steps (typically, enzymatic reactions) in a given metabolic pathway that are encoded in a genome or a metagenome. Likewise, copy number summarizes the number of distinct sets of enzyme annotations that collectively encode the complete pathway. This program offers two strategies for estimating metabolic potential: a 'stepwise' strategy with equivalent treatment for alternative enzymes – i.e, enzymes that can catalyze the same reaction in a given metabolic pathway – and a 'pathwise' strategy that accounts for all possible variations of the pathway. The Supplementary Information file includes more information on these two strategies and the completeness/copy number calculations. For the analysis of metagenomes, we used stepwise copy number of KEGG modules. Briefly, the calculation of stepwise copy number is done as follows: the copy number of each step in a pathway (typically, one chemical reaction or conversion) is individually evaluated by translating the step definition into an arithmetic expression that summarizes the number of annotations for each required enzyme. In cases where multiple enzymes or an enzyme complex are needed to catalyze the reaction, we take the minimum number of annotations across these components. In cases where there are alternative enzymes that can each catalyze the reaction individually, we sum the number of annotations for each alternative. Once the copy number of each step is computed, we then calculate the copy number of the entire pathway by taking the minimum copy number across all the individual steps. The use of minimums results in a conservative estimate of pathway copy number such that only copies of the pathway with all enzymes present are counted. For the analysis of genomes, we calculated the stepwise completeness of KEGG modules. This calculation is similar to the one described above for copy number, except that the step definition is translated into a boolean expression that, once evaluated, indicates the presence or absence of each step in the pathway. Then, the completeness of the modules is computed as the proportion of present steps in the pathway.

Metagenomic Datasets and Sample Groups. We acquired publicly-available gut metagenomes from 13 different studies (Le Chatelier et al., 2013; Feng et al., 2015; Franzosa et al., 2019; Lloyd-Price et al., 2019; Qin et al., 2012; Quince et al., 2015; Rampelli et al., 2015; Raymond et al., 2016; Schirmer et al., 2018b; Vineis et al., 2016a; of Sydney, 2016a; Wen et al., 2017; Xie et al., 2016). The studies were chosen based on the following criteria: (1) they included shotgun metagenomes of fecal matter (primarily stool, but some ileal pouch luminal aspirate samples (Vineis et al., 2016a) are also included); (2) they sampled from people living in industrialized countries (in the case where a study (Rampelli et al., 2015) included samples from hunter-gatherer populations, only the samples from industrialized areas were included in our analysis); (3) they included samples from people with IBD and/or they included samples from people without gastrointestinal (GI) disease or inflammation; and (4) clear metadata differentiating between case and control samples was available. A full description of the studies and samples can be found in Supplementary Table 4.1a-c. We grouped samples according to the health status of the sample donor. Briefly, the 'IBD' group of samples includes those from people diagnosed with Crohn's disease (CD), ulcerative colitis (UC), or pouchitis. The 'non-IBD' group contains non-IBD controls, which includes both healthy people presenting for routine cancer screenings as well as people with benign or non-specific symptoms that are not clinically diagnosed with IBD. Colorectal cancer patients from (Feng et al., 2015) were also put into the 'non-IBD' group on the basis that tumors in the GI tract may arise from local inflammation (Kraus and Arber, 2009) and represent a source of gut stress without an accompanying diagnosis of IBD. Finally, the 'HEALTHY' group contains samples from people without GI-related diseases or inflammation. Note that only control or pre-treatment samples were taken from the studies covering type 2 diabetes (Qin et al., 2012), ankylosing spondylitis (Wen et al., 2017), antibiotic treatment (Raymond et al., 2016), and dietary intervention (of Sydney, 2016a); these controls were all assigned to the 'HEALTHY' group. At least one study

(Le Chatelier et al., 2013) included samples from obese people, and these were also included in the ‘HEALTHY’ group.

Processing of metagenomes. We made single assemblies of most gut metagenomes using the anvi’o metagenomics workflow implemented in ‘anvi-run-workflow‘ (Shaiber et al., 2020a). This workflow uses Snakemake (Köster and Rahmann, 2012), and a tutorial is available at the URL `https://merenlab.org/2018/07/09/anvio-snakemake-workflows/`. Briefly, the workflow includes quality filtering using ‘iu-filter-quality-minoche‘ (Eren et al., 2013); assembly with IDBA-UD (Peng et al., 2012) (using a minimum contig length of 1000); gene calling with Prodigal v2.6.3 (Hyatt et al., 2010); tRNA identification with tRNAscan-SE v2.0.7 (Chan and Lowe, 2019); and gene annotation of ribosomal proteins (Seemann, 2017), single-copy core gene sets (Lee, 2019), KEGG KOfams (Aramaki et al., 2020), NCBI COGs (Galperin et al., 2021), and Pfam (release 33.1, (Mistry et al., 2021)). The aforementioned annotation was done with programs that relied on HMMER v3.3.2 (Eddy, 2011) as well as Diamond v0.9.14.115 (Buchfink et al., 2015). As part of this workflow, all single assemblies were converted into anvi’o contigs databases. Samples from (Vineis et al., 2016a) were processed differently because they contained merged reads rather than individual paired-end reads: no further quality filtering was run on these samples, we assembled them individually using MEGAHIT (Li et al., 2015), and we used the anvi’o contigs workflow to perform all subsequent steps described for the metagenomics workflow above. Note that we used a version of KEGG downloaded in December 2020 (for reproducibility, the hash of the KEGG snapshot available via ‘anvi-setup-kegg-kofams‘ is 45b7cc2e4fdc). Additionally, the annotation program ‘anvi-run-kegg-kofams‘ includes a heuristic for annotating hits with bitscores that are just below the KEGG-defined threshold, which is described at `https://anvio.org/m/anvi-run-kegg-kofams/`.

Genomic Dataset. We also analyzed microbial genomes from the Genome Taxonomy Database (GTDB), release 95.0 (Parks et al., 2018, 2020). We downloaded all reference genome sequences for the species cluster representatives.

Processing of GTDB genomes. We converted all GTDB genomes into anvi'o contigs databases and annotated them using the anvi'o contigs workflow, which is similar to the metagenomics workflow described above and uses the same programs for gene identification and annotation.

Estimation of the number of microbial populations per metagenome. We used single-copy core gene (SCG) sets belonging to each domain of microbial life (Bacteria, Archaea, Protista) (Lee, 2019) to estimate the number of populations from each domain present in a given metagenomic sample. For each domain, we calculated the number of populations by taking the mode of the number of copies of each SCG in the set. We then summed the number of populations from each domain to get a total number of microbial populations within each sample. We accomplished this using SCG annotations provided by 'anvi-run-hmms' (which was run during metagenome processing) and a custom script relying on the anvi'o class 'NumGenomesEstimator' (see reproducible workflow).

Removal of samples with low sequencing depth. We observed that, at lower sequencing depths, our estimates for the number of populations in a metagenomic sample were moderately correlated with sequencing depth (Supplementary Figure 4.6, R > 0.5). These estimates rely on having accurate counts of single-copy core genes (SCGs), so we hypothesized that lower-depth samples were systematically missing SCGs, especially from populations with lower abundance. Since accurate population number estimates are critical for proper normalization of pathway copy numbers, keeping these lower-depth samples would have introduced a bias into our metabolism analyses. To address this, we removed samples with low sequencing depth from downstream analyses using a sequencing depth threshold of 25 million reads, such that the remaining samples exhibited a weaker correlation (R < 0.5) between sequencing depth and number of estimated populations. We kept samples for which both the R1 file and the R2 file contained at least 25 million reads (and for the (Vineis et al., 2016a) dataset, we kept samples containing at least 25 million merged reads). This produced our final sample set of 408 metagenomes.

Estimation of normalized pathway copy numbers in metagenomes. We ran ‘anvi-estimate-metabolism‘, in genome mode and with the ‘–add-copy-number‘ flag, on each individual metagenome assembly to compute stepwise copy numbers for KEGG modules from the combined gene annotations of all populations present in the sample. We then divided these copy numbers by the number of estimated populations within each sample to obtain a per-population copy number (PPCN) for each pathway.

Selection of IBD-enriched pathways. We used a one-sided Mann-Whitney-Wilcoxon test with a FDR-adjusted p-value threshold of p <= 2e-10 on the per-sample PPCN values for each module individually to identify the pathways that were most significantly enriched in the IBD sample group compared to the healthy group. We calculated the median per-population copy number of each metabolic pathway in the IBD samples, and again in the healthy samples. After filtering for p-values <= 2e-10, we also applied a minimum effect size threshold based on the median per-population copy number in each group ($M_{IBD} - M_{Healthy} >= 0.12$) – this threshold was calculated by taking the mean effect size over all pathways that passed the p-value threshold. The set originally contained 34 pathways that passed both thresholds, but we removed one redundant module (M00006) which represents the first half of another module in the set (M00004).

Test for enrichment of biosynthesis pathways. We used a one-sided Fisher’s exact test (also known as hypergeometric test, see e.g., (Boyle et al., 2004)) for testing the independence between the metabolic pathways identified to be IBD-enriched (i.e., using the methods described in “Selection of IBD-enriched pathways) and functionality (i.e., pathways annotated to be involved in biosynthesis).

Pathway comparisons. Because the 33 IBD-enriched pathways were selected using PPCNs of healthy and IBD samples, statistical tests comparing PPCN distributions for these modules need to be interpreted with care, because the hypotheses were selected and tested on the same dataset (Fithian et al., 2014). Therefore, to assess the statistical validity of the identified IBD-enriched modules, we performed the following repeated sample-split analysis: we first randomly split the IBD and healthy samples into the equal-sized training and validation sets. We select IBD-enriched modules in the training set using the Mann-Whitney-Wilcoxon test, and then compute the p-values on the validation set. We repeat this sample split analysis 1,000 times with an FDR-adjusted p-value threshold of 1e-10 on the first split; most identified modules (89.4%; 95% CI: [87.5%, 91.3%]) on the training sets remain significant at a slightly less stringent threshold (1e-8) on the validation sets. This indicates that the approach we used to identify IBD-enriched modules yields stable and statistically significant results on this dataset.

Metagenome classification. We trained logistic regression models to classify samples as 'IBD' or 'healthy' using per-population copy numbers of IBD-enriched modules as features. We ran a 25-fold cross-validation pipeline on the set of 330 healthy and IBD metagenomes in our analysis, using an 80% train – 20% test random split of the data in each fold. The pipeline included selection of IBD-enriched pathways within the training samples using the same strategy as described above, followed by training and testing of a logistic regression model as implemented in the 'sklearn' Python package. We set the 'penalty' parameter of the model to "None" and the 'max_iter' parameter to 20,000 iterations, and we used the same random state in each fold to ensure changes in performance only come from differences in the training data rather than differences in model initialization. To summarize the overall performance of the classifier, we took the mean (over all folds) of each performance metric.

We trained a final classifier using the 33 IBD-enriched pathways selected earlier from the entire set of 330 healthy and IBD metagenomes. We then applied this classifier to the metagenomic samples from (Palleja et al., 2018), which we processed in the same way as the other samples in our analysis (including removal of samples with low sequencing depth and calculation of PPCNs of KEGG modules for use as input features to the classifier model).

Identification of gut microbial genomes from the GTDB. We took 19,226 representative genomes from the GTDB species clusters belonging to the phyla Firmicutes, Bacteroidetes, and Proteobacteria, which are most common in the human gut microbiome (Woting and Blaut, 2016). To evaluate which of these genomes might represent gut microbes in a computationally-tractable manner, we ran the anvi'o 'EcoPhylo' workflow (`https://anvio.org/m/ecophylo`) to contextualize these populations within 150 healthy gut metagenomes from the Human Microbiome Project (HMP) (Human Microbiome Project Consortium, 2012a). Briefly, the EcoPhylo workflow (1) recovers sequences of a gene family of interest from each genome and metagenomic sample in the analysis, (2) clusters resulting sequences and picks representative sequences using mmseqs2 (Steinegger and Söding, 2017), and (3) uses the representative sequences to rapidly summarize the distribution of each population cluster across the metagenomic samples through metagenomic read recruitment analyses. Here, we used the Ribosomal Pprotein S6 as our gene of interest, since it was the most frequently-assembled single-copy-core gene in our set of GTDB genomes. We clustered the Ribosomal Protein S6 sequences from GTDB genomes at 94% nucleotide identity.

To identify genomes that were likely to represent gut microbes, we selected genomes whose ribosomal protein S6 belonged to a gene cluster where at least 50% of the representative sequence was covered (i.e. detection >= 0.5x) in more than 10% of samples (i.e. n > 15). There are 100 distinct individuals represented in the 150 HMP gut metagenomes – 56 of which were sampled just once and 46 of which were sampled at 2 or 3 time points – so this threshold is equivalent to detecting the genome in 5% - 15% of individuals. From this selection we obtained a set of 836 genomes; however, these were not exclusively gut microbes, as some non-gut populations have similar ribosomal protein S6 sequences to gut microbes and can therefore pass this selection step. To eliminate these, we mapped our set of 330 healthy and IBD metagenomes to the 836 genomes using the anvi'o metagenomics workflow, and extracted genomes whose entire sequence was at least 50% covered (i.e. detection >= 0.5x) in over 2% (n > 6) of these samples. Our final set of 338 genomes was used in downstream analysis.

Genome phylogeny. To create the phylogeny, we identified the following ribosomal proteins that were annotated in at least 90% (n = 304) of the genomes: Ribosomal_S6, Ribosomal_S16, Ribosomal_L19, Ribosomal_L27, Ribosomal_S15, Ribosomal_S20p, Ribosomal_L13, Ribosomal_L21p, Ribosomal_L20, and Ribosomal_L9_C. We used the program ‘anvi-get-sequences-for-hmm-hits‘ to extract the amino acid sequences for these genes, align the sequences using MUSCLE v3.8.1551 (Edgar, 2004), and concatenate the alignments. We used trimAl v1.4.rev15 (Capella-Gutiérrez et al., 2009) to remove any positions containing more than 50% of gap characters from the final alignment. Finally, we built the tree with IQtree v2.2.0.3 (Minh et al., 2020), using the WAG model and running 1,000 bootstraps.

Determination of HMI status for genomes. We estimated metabolic potential for each genome with ‘anvi-estimate-metabolism‘ (in genome mode) to get stepwise completeness scores for each KEGG module, and then we used the script ‘anvi-script-estimate-metabolic-independence‘ to give each genome a metabolic independence score based on completeness of the 33 IBD-enriched pathways. Briefly, the latter script calculates the score by summing the completeness scores of each pathway of interest. Genomes were classified as having high metabolic independence (HMI) if their score was greater than or equal to 26.4. We calculated this threshold by requiring these 33 pathways to be, on average, at least 80% complete in a given genome.

Genome distribution across sample groups. We mapped the gut metagenomes from the healthy, non-IBD, and IBD groups to each genome using the anvi’o metagenomics workflow in reference mode. We used ‘anvi-summarize‘ to obtain a matrix of genome detection across all samples. We summarized this data as follows: for each genome, we computed the proportion of samples in each group in which at least 50% of the genome sequence was covered by at least 1 read (>= 50% detection). For each sample, we calculated the proportion of detected genomes that were classified as HMI. We also computed the percent abundance of each genome in each sample by dividing the number of reads mapping to that genome by the total number of reads in the sample.

Visualizations. We used ggplot2 (Wickham, 2016) to generate most of the initial data visualizations. The phylogeny and heatmap in Figure 4.3 were generated by the anvi'o interactive interface and the ROC curves in Figure 4.4 were generated using the pyplot package of matplotlib (Hunter, 2007). These visualizations were refined for publication using Inkscape, an open-source graphical editing software that is available at `https://inkscape.org/`.

4.6 Data Availability

Accession numbers for publicly available data are listed in our Supplementary Tables at `doi:10.6084/m9.figshare.22679080`. Our Supplementary Information file is also available at `doi:10.6084/m9.figshare.22679080`. Contigs databases of our assemblies for the 408 deeply-sequenced metagenomes can be accessed at `doi:10.5281/zenodo.7872967`, and databases for our assemblies of the (Palleja et al., 2018) metagenomes can be accessed at `doi:10.5281/zenodo.7897987`. Contigs databases of the 338 GTDB gut reference genomes are available at `doi:10.5281/zenodo.7883421`.

4.7 Supplementary Information

4.7.1 Low sequencing depth results in poor characterization of community richness

Within our dataset, we observed a correlation between the estimated number of distinct populations and sequencing depth, i.e. the number of short reads generated from a given sample.

When insufficient sequences are generated, genomes may not be entirely covered and not all single-copy core genes (SCGs) are detected in the assembly, resulting in an underestimation of the populations present. In contrast, higher sequencing depth increasingly fails to uncover more microbial populations, indicating that the estimate of the number of distinct microbial populations within these samples serves as a good approximation of the true number of genomes. Since an incomplete recovery of population genomes in metagenomic samples also interferes with a meaningful quantification of metabolic potential in a given sample, we set a minimum sequencing depth threshold of 25 million sequencing reads (Supplementary Figure 4.6). A set of 408 samples (101 IBD, 229 healthy, and 78 non-IBD) from 10 different studies passed our quality threshold to be utilized for further analysis.

Low sequencing depth disproportionately affects IBD metagenomes, thereby reducing our ability to effectively study this disease model in comparison with healthy controls. It also disproportionately affects some studies over others, which could allow cohort or study-specific effects to influence the differential signal between the groups. However, we concluded that the benefits of stringent thresholding outweigh the potential complications arising from imbalanced cohort sizes in our sample subset.

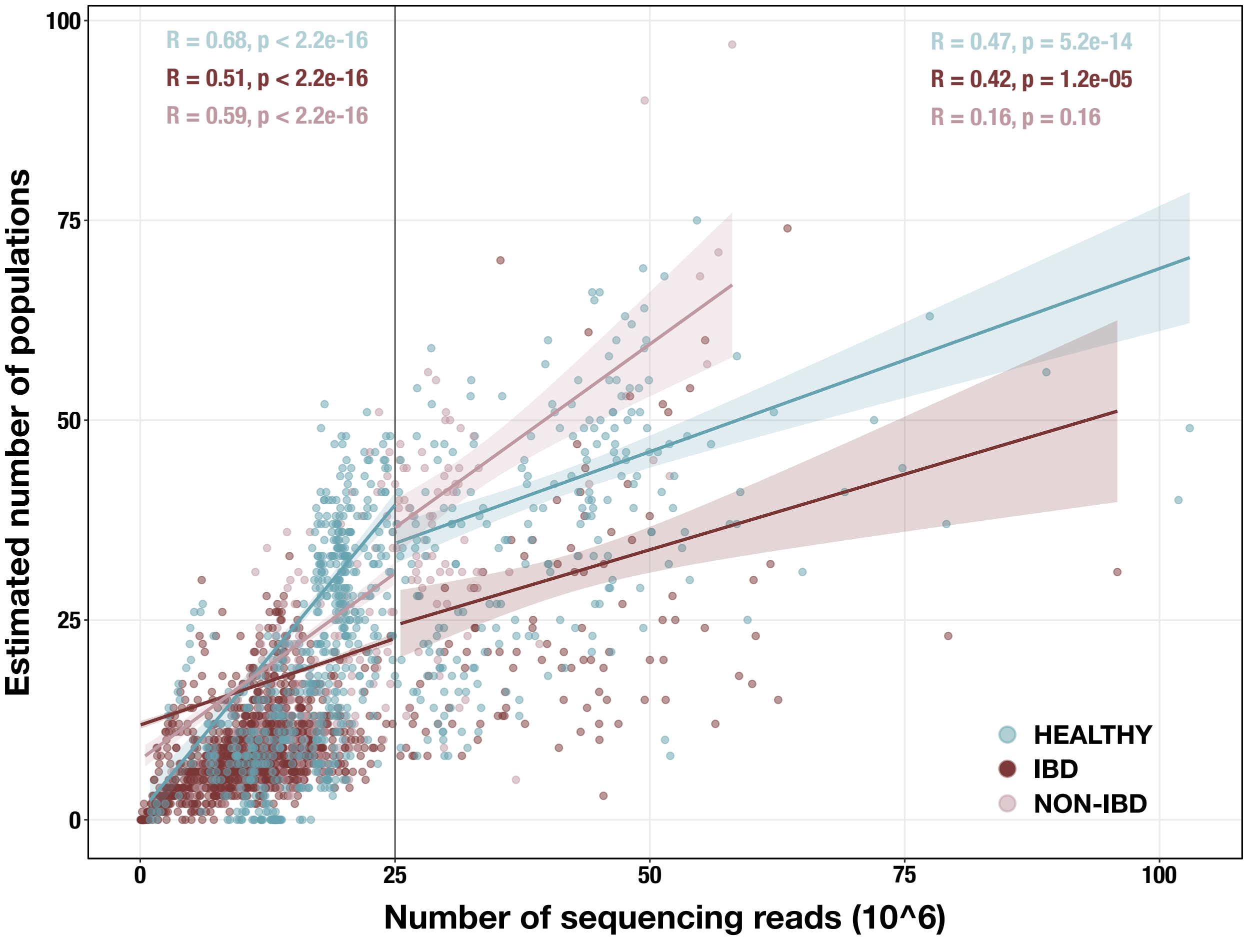


Figure 4.6: Scatterplot of sequencing depth vs estimated number of microbial populations in each of 2,893 stool metagenomes. Sequencing depth is represented by the number of R1 reads, except for (Vineis et al., 2016a) samples, in which case it is the number of merged paired-end reads. The vertical line indicates our sequencing depth threshold of 25 million reads. Per-group Spearman's correlation coefficients and p-values are shown for the subset of samples with depth < 25 million reads (top left) and for the subset with depth >= 25 million reads (top right). Regression lines are shown for each group in each subset, with standard error indicated by the colored background.

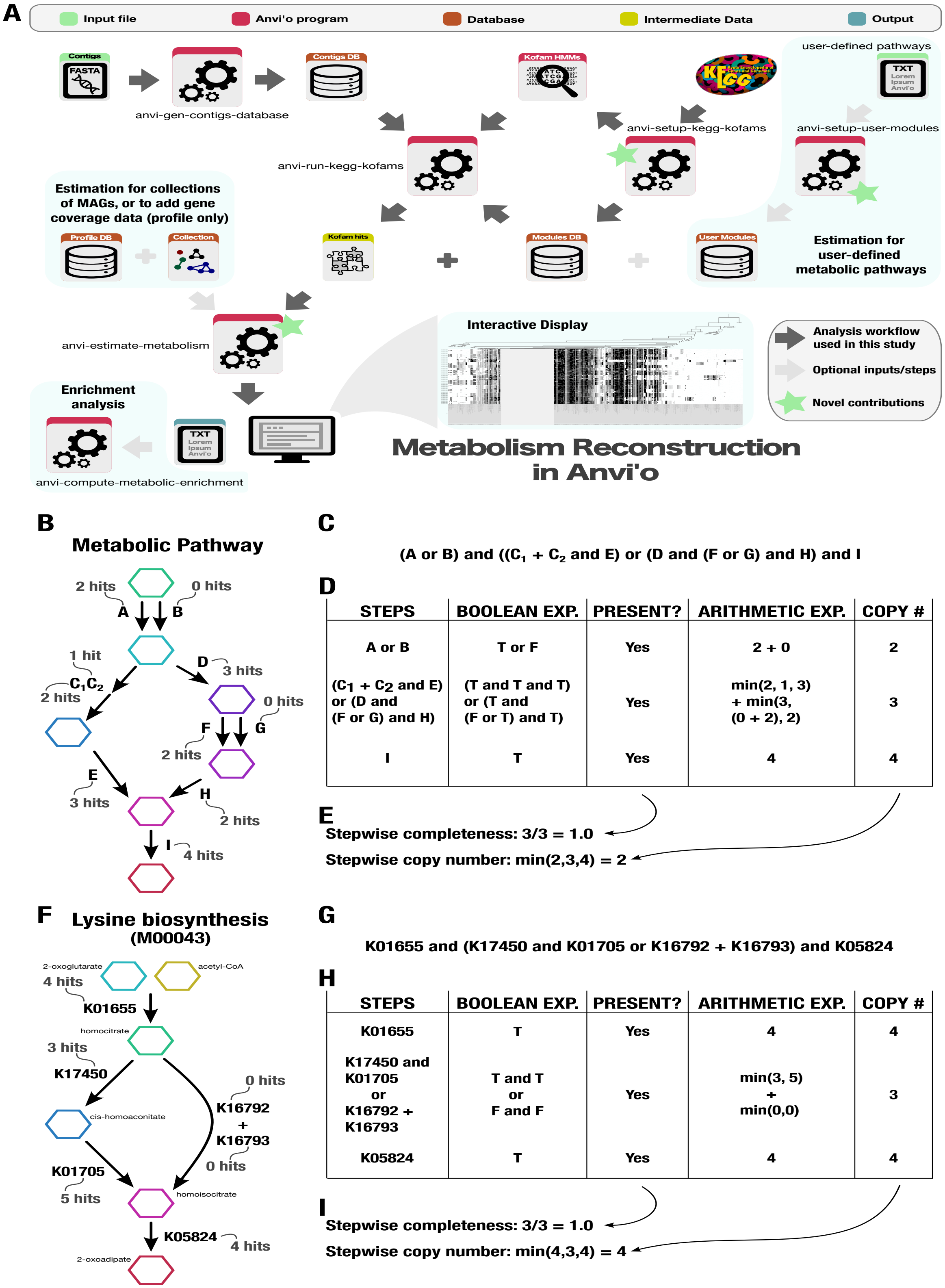
4.7.2 Technical details of metabolism estimation in anvi'o

This section describes technical details of the program 'anvi-estimate-metabolism', which is the main program in the metabolism reconstruction framework in anvi'o (Supplementary Figure 4.7a). Documentation for this program, including an extended and more up-to-date version of these technical details, can be found at `https://anvio.org/m/anvi-estimate-metabolism`.

Summary of program usage

The program 'anvi-estimate-metabolism' predicts the metabolic capabilities of organisms based on their genetic content. It relies upon enzyme annotations and metabolism information from KEGG, specifically using metabolic modules from the KEGG MODULE (Kanehisa et al., 2023) database, which are defined in terms of KEGG Orthologs (KOs) that can be annotated via the KOfam database of hidden Markov model (HMM) profiles (Aramaki et al., 2020). It can also work with user-defined metabolic pathways, as described in the documentation page `https://anvio.org/m/user-modules-data`.

The program determines which enzymes are annotated in an input sample and uses these functions to compute the completeness and copy number of each metabolic module within the sample. Input samples can be individual genomes, binned or unbinned metagenomes, or ad-hoc lists of enzyme accessions. The output of 'anvi-estimate-metabolism' is one or more tabular text files detailing the completeness and copy number scores per module as well as (customizable) information such as pathway metadata; shared/unique enzymes; gene coverage data; and pathway substrates, intermediates, and products. A detailed output description and examples can be found at `https://anvio.org/m/kegg-metabolism/`.



(A or B) and ((C_1 + C_2 and E) or (D and (F or G) and H) and I

STEPS	BOOLEAN EXP.	PRESENT?	ARITHMETIC EXP.	COPY #
A or B	T or F	Yes	2 + 0	2
(C_1 + C_2 and E) or (D and (F or G) and H)	(T and T and T) or (T and (F or T) and T)	Yes	min(2, 1, 3) + min(3, (0 + 2), 2)	3
I	T	Yes	4	4

Stepwise completeness: 3/3 = 1.0

Stepwise copy number: min(2,3,4) = 2

K01655 and (K17450 and K01705 or K16792 + K16793) and K05824

STEPS	BOOLEAN EXP.	PRESENT?	ARITHMETIC EXP.	COPY #
K01655	T	Yes	4	4
K17450 and K01705 or K16792 + K16793	T and T or F and F	Yes	min(3, 5) + min(0,0)	3
K05824	T	Yes	4	4

Stepwise completeness: 3/3 = 1.0

Stepwise copy number: min(4,3,4) = 4

Figure 4.7: Technical details of the metabolism reconstruction software framework in anvi'o.

Figure 4.7 continued: A) Workflow of metabolism reconstruction programs and their inputs/outputs. Dark arrows indicate the primary analysis path utilized in this study. Blue background indicates optional features in the framework. A demonstration of completeness score and copy number calculations for metabolic pathways (performed by the program 'anvi-estimate-metabolism' is shown using example enzyme annotation data in panels B – E (for a theoretical pathway) and F – I (for a real pathway). B) Theoretical metabolic pathway, where hexagons represent metabolites, arrows represent chemical reactions, letters represent enzymes (subscripts indicate enzyme components), and the example number of gene annotation hits for each enzyme is written in gray. C) The definition of the theoretical pathway from panel B, written in terms of the required enzymes. D) Table showing the major steps in the pathway and example calculations for step presence and copy number. Step presence is calculated by evaluating a boolean expression created from the step definition in which enzymes with > 0 hits are replaced with True (T) and the others with False (F). Step copy number is calculated by evaluating the corresponding arithmetic expression in which the enzymes are replaced with their annotation counts. E) Final calculations of completeness score (fraction of present steps) and copy number for the theoretical metabolic pathway. F – I) Same as panels B – E, but for KEGG module M00043.

Module definitions and interpretation strategies

Metabolic pathways are defined by the enzymes responsible for each reaction in the pathway, using the convention established by the KEGG MODULE database. In these definitions, commas separate alternative enzymes that can catalyze the same reaction, spaces separate subsequent reactions, plus signs indicate essential components of enzyme complexes, minus signs indicate non-essential components of complexes, and parentheses indicate the order of operations. These definitions can also be written in terms of the logical relationships between reactions, such that spaces and plus signs are converted into 'AND' relationships and commas are converted into 'OR' relationships (Supplementary Figure 4.7b-c and f-g).

'anvi-estimate-metabolism' has two strategies for interpreting module definition strings that treat alternative enzymes and pathway branches differently. One is the 'pathwise' strategy, which considers all possible combinations of enzymes. In this method, each alternative set of enzymes that could be used together to catalyze every reaction in the metabolic pathway is called a 'path' through the module. The program computes completeness and copy number metrics for each path separately, and then identifies the most complete path(s) as the most biologically-relevant representative of the module as a whole. Alternatively, with the 'stepwise' strategy the module definition is parsed into high-level 'steps' that each encompasses a set of alternative enzymes for a particular reaction or branch point. The presence and copy numbers of each step are respectively combined into a completeness score and copy number for the entire module.

Calculation of stepwise completeness and copy number

The analyses in this paper rely on the 'stepwise' metrics of module completeness and copy number, which are calculated as demonstrated in Supplementary Figure 4.7d-e and h-i. We divide each module into steps by splitting the definition string on the outermost 'AND' relationships (spaces not within parentheses). To determine whether each step is present, we convert the step definition into a Boolean expression in which ‘True’ represents annotated enzymes and ‘False’ represents enzymes without annotations. If the Boolean expression evaluates to ‘True’, then the step is considered present. The module completeness score is the number of present steps divided by the total number of steps. To determine the step copy number, we convert the step definition into an arithmetic expression wherein ‘AND’ relationships become minimum operations and ‘OR’ relationships become addition operations. We take the minimum of all per-step copy numbers obtained by evaluating these arithmetic operations to get the overall module copy number.

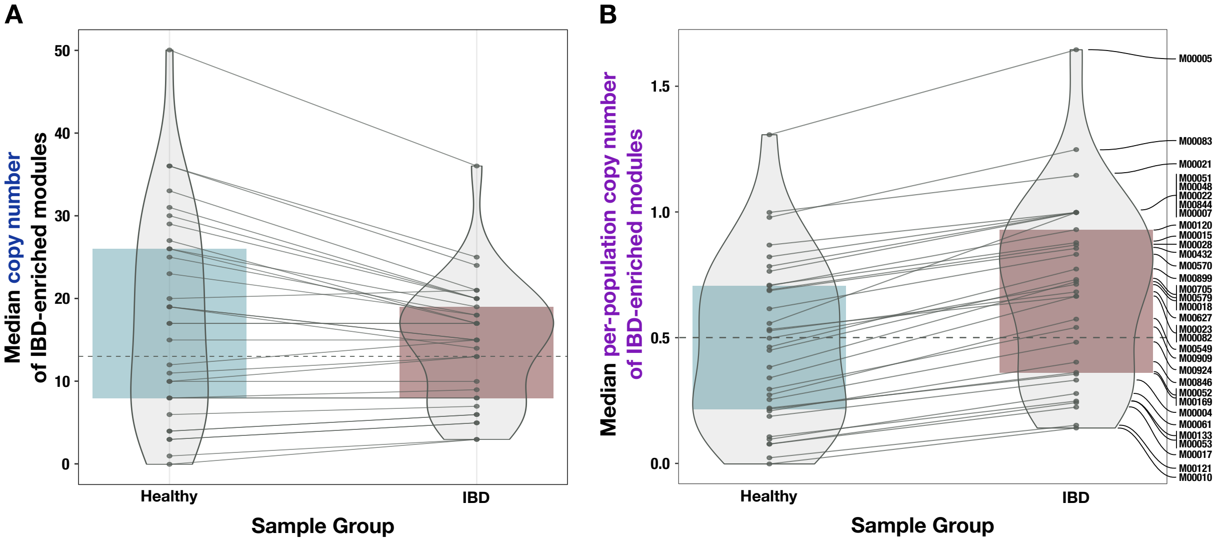


Figure 4.8: Comparison of unnormalized copy number data and normalized (per-population copy number, or PPCN) data for the IBD-enriched modules. A) Boxplot of median copy numbers for each module in the healthy samples (blue) and IBD samples (red). B) Boxplots of median PPCN for each module in the healthy samples (blue) and IBD samples (red). Lines connect data points for the same module in each plot. The gray dashed line in each plot indicates the overall median value.

4.7.3 Differential annotation efficiency between IBD and Healthy samples

We observed that the proportion of predicted genes with functional annotations was markedly less in healthy metagenomes than in IBD samples, for both sequence homology-based annotation methods (NCBI Clusters of Orthologous Groups, or COGs) and annotation with probabilistic models (KEGG KOfams and Pfams) (Supplementary Figure 4.9, Supplementary Table 4.1d). One possible interpretation of this that aligns with our metabolic competency hypothesis is that the (LMI) populations that thrive in the healthy gut environment are relatively less well-characterized than the (HMI) populations that are more likely to survive in the stressful conditions of IBD, resulting in an annotation bias against healthy samples. This interpretation is congruent with our observation that most uncharacterized gut microbial genomes from the GTDB, which have temporary code names in place of taxonomic assignments, were identified as non-HMI (Figure 4.3b).

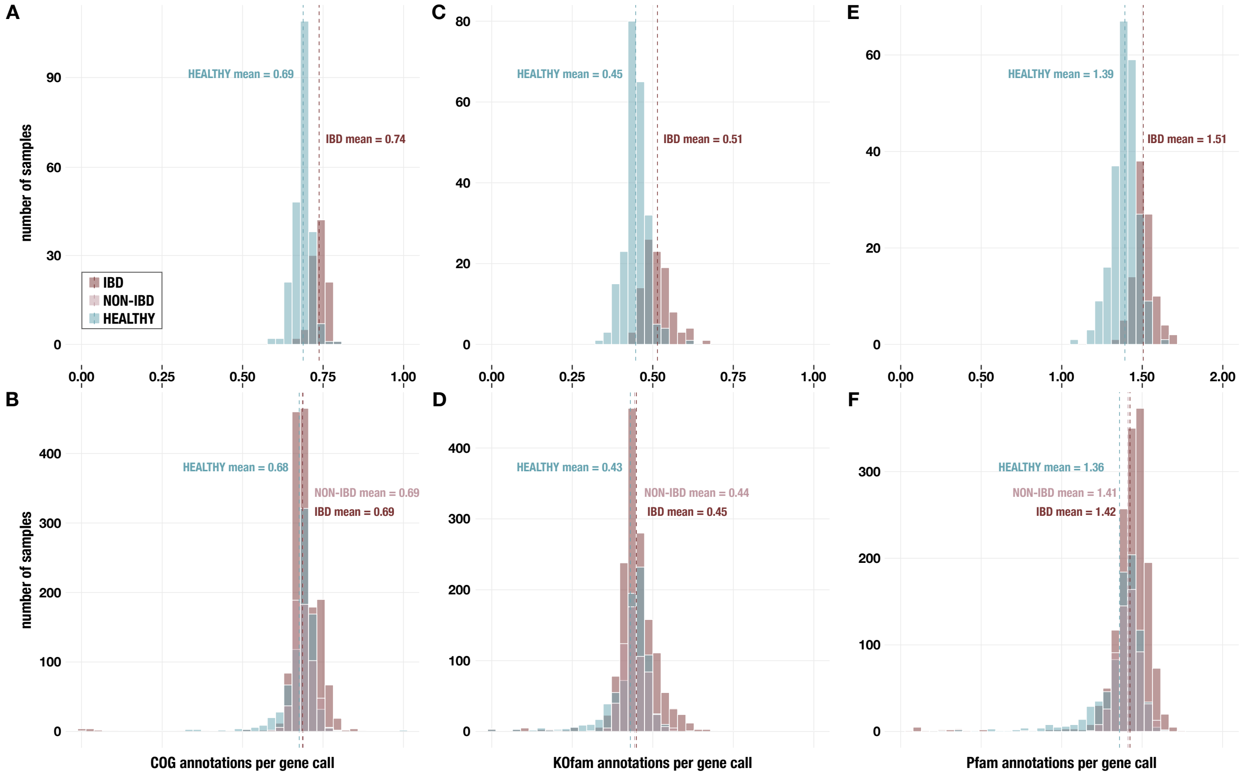


Figure 4.9: Histograms of annotations per gene call from A,B) NCBI COGs; C,D) KEGG KOfams; and E,F) Pfams. Panels A, C, and E show data for metagenomes in the subset of 330 deeply-sequenced samples from healthy people and people with IBD, and panels B, D, and F show data for all 2,893 samples including those from non-IBD controls.

Perhaps the reduced metabolic capacity of LMI microbes and their resulting reliance on robust community interactions (i.e., cross-feeding) makes them less easily culturable in vitro, especially since cultivation is typically attempted for isolated populations rather than communities. These microbes may rely on their surrounding community for access to key metabolites that they cannot synthesize for themselves, and our current cultivation practices may not be sufficient to make up for the lack of such interactions. In this case, their representation in common sequence databases would be limited to sequences from metagenomic surveys, which are often incomplete and/or composite and therefore not typically included in efforts to generate models and non-redundant sequence databases for gene annotation (Aramaki et al., 2020; Galperin et al., 2015; Sonnhammer et al., 1997). Therefore, one explanation for the reduced proportion of annotated genes in healthy metagenomes is that some annotations are simply missing due to lack of sufficiently-homologous sequences in state-of-the-art databases. Thus, the true reduction in metabolic potential in the healthy sample group may not be as extensive as we have observed in this study. However, this does not exclude the possibility that some reduction in metabolic capacity exists due to the inherent ecological differences between the healthy and IBD gut environments and irrespective of microbial characterization levels.

The discrepancy in annotation efficiency between the healthy and IBD groups disappeared when analyzing all 2,893 samples (Supplementary Figure 4.9). This suggests that the observed annotation bias does not strongly affect microbial populations that are readily assembled via shallow sequencing – likely, these are populations of high relative abundance in both healthy and IBD samples. Populations of lower abundance, which are less likely to be assembled from shallow metagenomes due to lack of sufficient coverage, are probably also less well-characterized as a result. For this to contribute to fewer annotations per gene in healthy samples would necessitate that healthy samples contain relatively more low-abundance populations than IBD samples. Indeed, this is the case: healthy samples contain an average of 86 detected genomes from our set of GTDB gut microbes, and those genomes have a low average percent abundance of 0.61% across these samples. Non-IBD samples are similar, having an average of 77 detected genomes per-sample with an average percent abundance of 0.79%.

IBD samples, meanwhile, contain 30 detected genomes on average, with a higher average percent abundance of 2.24%. Therefore, the lack of characterization of low-abundance populations may contribute to the relative reduction in gene annotations in the healthy samples.

4.7.4 Pathway enrichment without consideration of effect size leads to nonspecific results

Our analyses indicate that the majority of KEGG modules had higher per-population copy number in IBD metagenomes (Figure 4.2e, Supplementary Table 4.2b). Indeed, when we examine all modules with non-zero median per-population copy number in at least one group of samples (n = 117), their median normalized copy number is systematically higher in the IBD group than in the healthy group (Supplementary Figure 4.10c; 98 out of 117 modules have a higher normalized copy number in the IBD group than in the healthy group at 5% FDR-adjusted significance level using a one-sided Wilcoxon test). This result is likely a natural outcome of the differential distribution of HMI and non-HMI genomes in the two sample groups, as seen in our analysis of reference genomes (Figure 4.3c, 4.3d), where the overrepresentation of HMI populations with larger genomes that encode many more complete pathways in IBD samples leads to higher per-population copy numbers computed at the metagenome level. The consistent elevation in PPCN of metabolic modules in IBD could also be attributed, at least in part, to the aforementioned functional annotation biases that seem to disproportionally affect the characterization of healthy metagenomes. The lower annotation efficiency in healthy metagenomes could result in partial copies of pathways, which are ignored by our stringent copy number calculation that only counts complete copies. Therefore, to narrow down our results and identify which pathways are particularly important for microbial resilience in the IBD gut environment, we considered only those pathways with the largest difference in normalized copy number (‘effect size’) between the two groups to identify metabolic modules that are truly elevated in IBD metagenomes (see Methods).

With similar considerations we also investigated whether biosynthetic capacity in general was enriched in IBD samples. For this, we expanded our analysis to also consider biosynthesis pathways that did not meet the enrichment criteria we have used for inclusion in the final set of 33 IBD-enriched modules. As expected, we found that the majority of all biosynthesis pathways in the KEGG Module database (n = 88) have significantly higher normalized copy numbers in IBD samples (Supplementary Figure 4.10d; at a 5% FDR-adjusted significance level, 62 out of 88 (70%) biosynthesis pathways have a higher normalized copy number using a one-sided Wilcoxon test). This analysis also showed a similar increase for non-biosynthetic pathways: 63 out of 91 (69%) non-biosynthetic pathways showed significant increase in IBD samples (two-sample test for equality of proportion: 0.88). Overall, these data indicate that without the consideration of effect size, both biosynthetic and non-biosynthetic capacity appear to be increased in the IBD gut microbiome. In contrast, maintenance of a higher metabolic capacity for the biosynthesis of essential nutrients emerges as an important factor for microbial resilience in IBD through a strict enrichment criteria in addition to statistical significance scores calculated for differential abundance.

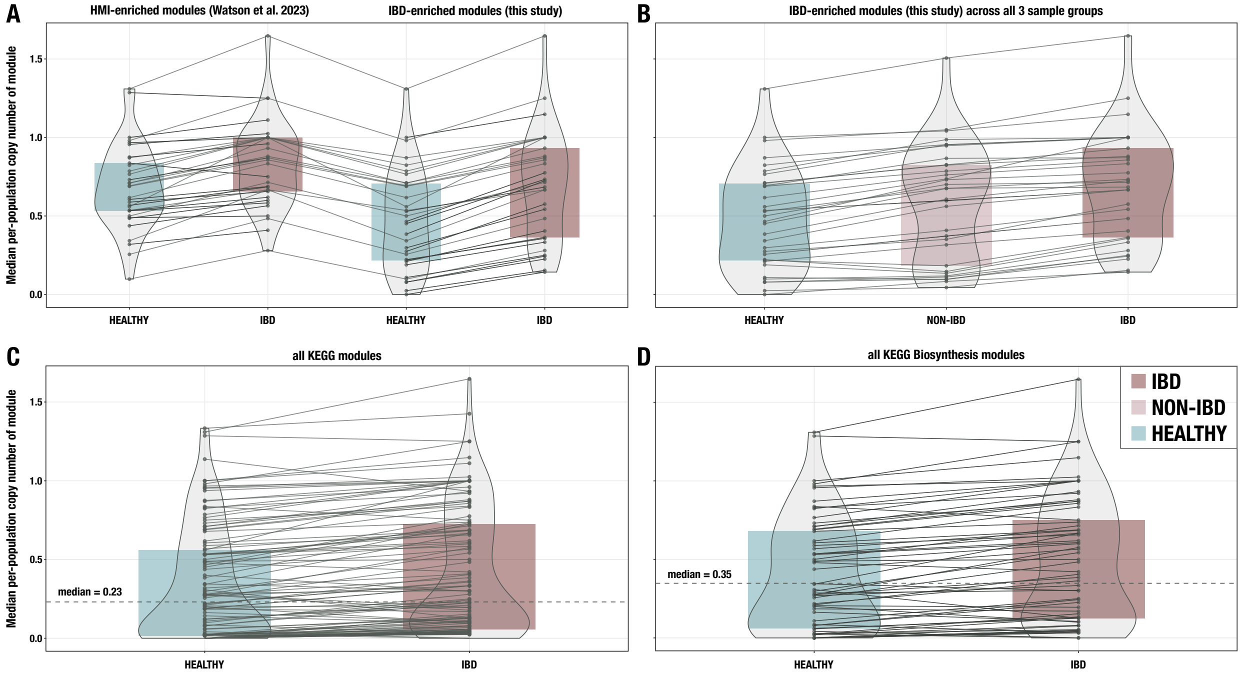


Figure 4.10: Additional boxplots of median per-population copy number for various subsets of metabolic pathways and metagenome samples. A) 33 modules enriched in HMI populations from (Watson et al., 2023) compared to the 33 IBD-enriched modules from this study, with medians computed in the set of deeply-sequenced healthy (n = 229) and IBD (n = 101) samples. B) The 33 IBD-enriched modules from this study, with medians computed in the set of deeply-sequenced healthy (n = 229), non-IBD (n = 78), and IBD (n = 101) samples. C) All KEGG modules (n = 117) with non-zero copy number in at least one sample, with medians computed in the set of deeply-sequenced healthy (n = 229) and IBD (n = 101) samples. D) All biosynthesis modules (n = 88) from the KEGG MODULE database, with medians computed in the set of deeply-sequenced healthy (n = 229) and IBD (n = 101) samples. Where applicable, dashed lines indicate the overall median for all modules, and solid lines connect the points for the same module in each sample group. The IBD sample group is highlighted in red, the NONIBD group in pink, and the HEALTHY group in blue.

4.7.5 A review of HMI-associated modules in the context of gut microbiome literature

The 33 pathways that are enriched in microbial communities associated with individuals with IBD likely provide competencies that are critical for survival in the stressed gut environment. In this section, we offer a review of IBD-enriched modules with existing gut microbiome scientific literature.

Amino acid pathways

The eleven amino acid pathways enriched in the IBD gut microbiome collectively encode for biosynthesis of 8 proteinogenic amino acids – tryptophan, cysteine, leucine, isoleucine, threonine, methionine, proline, and arginine. They also account for the production of chorismate (the precursor for aromatic amino acids) via the Shikimate pathway, the non-proteinogenic amino acid ornithine (which is a precursor for arginine), and polyamines such as spermidine and putrescine. Most of these pathways are interlinked, utilizing the same or similar intermediates and enzymes, and some modules are the successors of others. For instance, both threonine and methionine are produced from homoserine, and threonine can then be converted to isoleucine. The Shikimate pathway produces chorismate, which is an input to tryptophan biosynthesis. Both proline and ornithine are produced from glutamate, ornithine can be converted to arginine, and arginine is the precursor for the polyamines. The interdependence of these pathways may have influenced the enrichment signal, such that some of these modules may be enriched because they encode precursor molecules to critical metabolites, even if the precursor molecules themselves are not as essential to microbial survival in the IBD gut environment.

Of the eight proteinogenic amino acids that can be synthesized with IBD-enriched pathways, leucine, tryptophan, threonine, isoleucine, and methionine are essential amino acids for humans (Lopez and Mohiuddin, 2023), while cysteine and arginine are semi-essential (Rehman et al., 2020; Tong and Barbul, 2004). Furthermore, leucine, tryptophan, isoleucine, and cysteine have been shown to have a protective effect against oxidative stress for intestinal epithelial cells (Katayama and Mine, 2007), which may be relevant in the more aerobic gut environment that is characteristic of IBD (Rigottier-Gois, 2013). Several of these amino acids have been analyzed for their potential therapeutic effects in IBD (Liu et al., 2017). Regardless, it is unknown if the depleted gut microbiome in IBD would produce these amino acids in sufficient quantity to promote health benefits to the host, especially considering that the microbes themselves require these molecules for protein production and as nutrient sources – for instance, in proteolytic fermentation (Wu et al., 2021b; Lin et al., 2017).

The Shikimate pathway, which converts phosphoenolpyruvate (PEP) and erythrose 4-phosphate (E4P) to chorismate, is a prerequisite for tryptophan biosynthesis. This pathway is only present in microorganisms and plants, and it also produces intermediates for other metabolic pathways such as quinate degradation and antibiotic synthesis (Herrmann and Weaver, 1999). A recent analysis of paired fecal metagenomes and metatranscriptomes from the Human Microbiome Project using reference genome-based functional inference demonstrated that the Shikimate pathway is typically incomplete in gut microbes and only transcriptionally active in a few, suggesting that most gut microbes are auxotrophic for aromatic amino acids and therefore rely on dietary sources and potentially cross-feeding to obtain these molecules or their precursors (Mesnage and Antoniou, 2020). This offers a potential explanation for the enrichment of the Shikimate and tryptophan biosynthesis pathways in the IBD gut microbiome, where a depleted community may restrict the availability of cross-fed metabolites. Indeed, a tryptophan-deficient diet alters the composition of the gut microbiota in aged mice (Yusufu et al., 2021), providing auxiliary evidence that the loss of this amino acid impacts microbial survival. Furthermore, the serum levels of tryptophan are reduced in individuals with IBD due to high host metabolism rates (Nikolaus et al., 2017), which may exacerbate the lack of bioavailable tryptophan for gut microbes. On the host side, tryptophan and its derivatives influence a number of physiological processes, though it is unclear how much the microbial production of tryptophan contributes to these effects (Agus et al., 2018). Nevertheless, the lack of tryptophan appears to worsen intestinal inflammation while supplementation can attenuate it (Kim et al., 2010; Hashimoto et al., 2012).

Cysteine biosynthesis was previously found to be enriched in the IBD gut microbiome based on reference genome analysis of 16S amplicon data (Morgan et al., 2012). The authors of that study proposed that cysteine metabolism could be important to microbial management of oxidative stress via the production of glutathione, which is protective against reactive oxygen species (Sherrill and Fahey, 1998; Tepe et al., 2006), from cysteine and glutamate. Cysteine can also be converted into hydrogen sulfide (H2S) by host colonocytes and some intestinal microbes. Though H2S produced by colonocytes can help support their energy production, excess microbially-derived H2S in the lumen is a risk factor for gut mucosal inflammation and H2S may play a role in colorectal carcinogenesis (Blachier et al., 2019). Interestingly, cysteine biosynthesis is also enriched in the gut microbiomes of postmenopausal women, where it is thought to contribute to elevated homocysteine levels and therefore to increased risk of cardiovascular disease (Zhao et al., 2019).

Leucine and isoleucine, as branched-chain amino acids (BCAAs), are important nutrients and signaling molecules in humans (Gojda and Cahova, 2021). Gut microbial synthesis of these compounds does contribute to human BCAA pools, as evidenced by experiments with heavy isotope labeling and correlations between serum and fecal BCAA levels (Metges et al., 1999; Dhakan et al., 2019). The extent of this exchange has not been characterized in individuals with IBD. However, a study of individuals receiving anti-integrin therapy for Crohn's disease demonstrated that pathways for biosynthesis of L-isoleucine and arginine were enriched at baseline in the gut microbiomes of responders to the therapy (Ananthakrishnan et al., 2017). In a longitudinal study of mice, biosynthesis pathways for leucine and proline were more abundant in animals modeling IBD (Sharpton et al., 2017).

Threonine and proline are both important components of intestinal mucins (Johansson and Hansson, 2016; Faure et al., 2005) and thus contribute to mucosal barrier integrity, which is typically impaired in IBD (Johansson et al., 2010). For instance, threonine, proline, and cysteine supplementation has been shown to reduce symptoms and restore lactobacilli and bifidobacteria counts in rats with DSS-induced inflammation (Sprong et al., 2010; Faure et al., 2006). The latter observation suggests the importance of an external source of these three amino acids to the fitness of the lactobacilli and bifidobacterial populations and thereby supports the idea that they are community metabolites.

We also found methionine biosynthesis to be enriched in the IBD gut microbiome. In individuals with quiescent inflammatory bowel disease, reduced serum levels of methionine, proline, and tryptophan are correlated with changes in the gut microbiome that are associated with increased symptoms of fatigue (Borren et al., 2021), demonstrating a putative link between methionine bioavailability, microbial abundances, and host wellbeing. Indeed, L-methionine supplementation in piglets results in improved mucosal integrity and villus architecture (Chen et al., 2014), and the activated form of methionine, S-adenosylmethionine, can reverse colon lesions and cytoskeletal damage in intestinal cells in DSS-treated mice (Oz et al., 2005). Yet reducing methionine in high-fat diets given to mice was shown to improve intestinal barrier function, reduce inflammation, and increase the abundance of short-chain fatty acid-producing microbes (Yang et al., 2019), so the net impact of methionine on host health and microbial fitness remains unclear.

Arginine has been well-studied in the context of inflammatory bowel disease. It has been shown to reduce cytokine production, promote intestinal healing and improve intestinal barrier function in DSS-treated mice, perhaps by enhancing production of nitric oxide (NO) (Coburn et al., 2012; Gobert et al., 2004; Singh et al., 2019). NO is a free radical that has been implicated in regulating mucosal barrier integrity, gastrointestinal motility, and protection against oxidative stress, though overproduction of this compound can have detrimental effects (Kolios et al., 2004; Walker et al., 2018). Biosynthesis of ornithine, which is both a precursor and a derivative of arginine, was enriched in the IBD gut microbiome as well, in agreement with another study that reported an increase in ornithine biosynthesis in the gut microbiome of individuals with active ulcerative colitis (Hellmann et al., 2023). Finally, polyamines – which are derived from arginine and were also represented in the enriched pathways – promote intestinal barrier function (Liu et al., 2009); for instance, by regulating the growth of intestinal epithelial cells (McCormack and Johnson, 1991).

Carbohydrate pathways

Three KEGG modules describing the pentose phosphate pathway were enriched in IBD samples - the entire pentose phosphate cycle (M00004), the oxidative phase (M00006), and the non-oxidative phase (M00007). We removed the oxidative phase (M00006) from our set of IBD-enriched modules because it was an exact copy of the initial steps in M00004; however, we kept the non-oxidative phase (M00007) in our set because it is defined using slightly different enzymes than the non-oxidative portion of M00004. M00007 is defined in four steps and utilizes a ribulose-phosphate 3-epimerase and a ribose 5-phosphate isomerase in the last two steps, while the non-oxidative phase in M00004 is defined in three steps and utilizes a glucose-6-phosphate isomerase in the last step. The pentose phosphate pathway (PPP) is a ubiquitous pathway in most bacteria and eukaryotes, as it plays a central role in cellular metabolism. It produces the important cellular intermediates ribose 5-phosphate and erythrose 4-phosphate, which are used for synthesis of nucleotides and aromatic amino acids, respectively (Soderberg, 2005). In fact, erythrose 4-phosphate is one of the inputs to the Shikimate pathway, another IBD-enriched module discussed above. The PPP also produces NADPH, a reducing equivalent important for reductive reactions and prevention of oxidative stress (Kruger and von Schaewen, 2003; Christodoulou et al., 2018). Beyond its link to other enriched amino acid biosynthesis pathways, it is unusual that such a central pathway would have an increased copy number in the IBD gut microbiome rather than being equally distributed across all samples. Some gut microbes are known to lack the transaldolase gene in this pathway and may instead encode an alternative pathway for pentose degradation called the sedoheptulose 1,7-bisphosphate pathway (SBPP) (Garschagen et al., 2021); it is therefore possible that the enrichment of the more common PPP in IBD is related to an increased ratio of microbial populations that use the PPP rather than the SBPP in the less-diverse microbiome of IBD patients, though this requires further investigation to verify.

The first carbon oxidation of the citric acid cycle (TCA cycle), which is a three-step conversion from oxaloacetate to 2-oxoglutarate (alpha-ketoglutarate), is enriched in the IBD samples. Similar to the PPP, the citric acid cycle is a central metabolic pathway, especially with regards to generation of energy and key metabolites for other pathways (Akram, 2014). It is unclear why only this particular portion of the cycle would be enriched, though this could perhaps be attributed to the role of alpha-ketoglutarate in the production of glutamate, the precursor to proline, ornithine and arginine (three amino acids with enriched biosynthesis pathways in the IBD sample group, as discussed above). It has been said that 2-oxoglutarate is the most fundamental compound of this cycle, serving as the link between carbon and nitrogen metabolism and also as a critical element in the recovery of amine groups for amino acid and protein production (Pierzynowski and Pierzynowska, 2022; Huergo Luciano F. and Dixon Ray, 2015). Thus, the enrichment of 2-oxoglutarate production capacity in the IBD gut environment could be related to the enrichment of amino acid biosynthesis pathways.

Two nucleotide sugar biosynthesis pathways are enriched in the IBD gut microbiome. One of these is synthesis of UDP-glucose, which is an important molecule implicated in a variety of key cellular metabolisms. It is an intermediate in polysaccharide biosynthesis and pyrimidine metabolism, a precursor of lipopolysaccharides in the outer cell membrane of Gram-negative bacteria, and an extracellular signaling molecule (Ralevic, 2015). Additionally, as an agonist for P2Y-14 receptors, it could play a role in modulating host gastrointestinal functions like muscular contraction (Bassil et al., 2009), and in modulating host inflammatory responses by activating this receptor specifically in T-lymphocytes (Scrivens and Dickenson, 2005) and in immature monocyte-derived dendritic cells (MDDC) (Skelton et al., 2003). The other enriched nucleotide sugar pathway is UDP-GlcNAc biosynthesis. Flux through this pathway is linked to a multitude of other central metabolisms, including amino acid and fatty acid metabolism (Hardivillé and Hart, 2014). Furthermore, UDP-GlcNAc is an important substrate in protein glycosylation pathways (Hardivillé and Hart, 2014; Ryczko et al., 2016), and a precursor to critical cell wall components in bacteria (Liu and Breukink, 2016; Mikkola, 2020; van Dam et al., 2009). In the gut, this molecule has been implicated in regulation of nutrient uptake by the host (Ryczko et al., 2016).

D-Glucuronate (glucuronic acid) degradation into pyruvate and D-glyceraldehyde 3-phosphate is also enriched in the IBD gut microbiome. Some gut microbes are capable of growth on host-derived uronic acids (Lopez-Siles et al., 2012), so this pathway may serve as a source of energy to microbes living in the IBD gut environment. In mice, there is some evidence that derivatives of glycosaminoglycan degradation such as D-glucuronate can worsen colitis (Lee et al., 2009b), so it is possible that this pathway is relevant to modulation of inflammation in IBD patients.

Finally, the phosphoribosyl diphosphate (PRPP) biosynthesis pathway is important because PRPP is used in the formation of glycosidic bonds as well as in the biosynthesis of a number of cofactors, amino acids, and nucleotides (Hove-Jensen et al., 2017). It is discussed further below in the context of nucleotide metabolism.

Cofactor and vitamin pathways

Biosynthesis or salvage pathways for the following five cofactors and vitamins are enriched in IBD: heme, siroheme, thiamine (vitamin B1), cobalamin (vitamin B12), and coenzyme A (CoA).

Heme is required for aerobic respiration (Gruss et al., 2012) and the increase in this pathway may be related to elevated oxygen levels in the gut as a result of inflammation, which promotes the growth of aerotolerant microbes (Shah, 2016; Cevallos et al., 2019). Dietary heme has also been associated with gut dysbiosis, aggravated colitis, and increased cytotoxicity in the colon (Constante et al., 2017; Ijssennagger et al., 2015); and genes related to heme and siroheme biosynthesis have also been found with high abundance in infants with neonatal necrotizing enterocolitis (Claud et al., 2013).

Both thiamine and cobalamin are important cofactors that are commonly shared between gut microbes (Magnúsdóttir et al., 2015), suggesting that microbes incapable of synthesizing them are unable to thrive in the depleted community of the IBD gut environment. Neither of these vitamins is produced by host cells but they are typically acquired from dietary sources (cobalamin, in particular, is absorbed in the small intestine) (Seetharam and Alpers, 1982; Degnan et al., 2014b; Hossain et al., 2022), so the enrichment of these pathways is unlikely to have a large impact on host health.

Coenzyme A can be produced from pantothenate (vitamin B5) by most gut microbes (Magnúsdóttir et al., 2015) and its biosynthesis has been described as 'essential' considering that CoA is required for a large number of enzymatic reactions (Spry et al., 2008; Leonardi et al., 2005). It is therefore interesting that this pathway appears to be enriched in the IBD gut microbiome, which implies a relative deficiency of CoA biosynthesis in the healthy gut microbiome. It is possible that the module is spuriously enriched, despite its low p-value of 5.7e-21, given the short length of this pathway – it has 3 major steps when the KEGG module definition is interpreted in a 'stepwise' fashion by anvi-estimate-metabolism, though there are in fact 5 chemical conversions (Supplementary Table 4.2a). An alternative possibility is that the KEGG Ortholog hidden Markov models (HMMs) for the required enzymes do not sufficiently represent the diversity of these proteins across the gut microbiota, which could cause this pathway to be undercounted due to lack of proper annotations.

Nucleotide pathways

The IBD-enriched modules include pathways for synthesis of the first complete purine, inosine monophosphate (IMP) as well as a series of pyrimidine biosynthesis pathways encoding the conversion from uridine monophosphate (UMP) to ribonucleotides (UDP/UTP, CDP/CTP) and finally to the cytosine deoxyribonucleotide (dCTP). The phosphoribosyl diphosphate (PRPP) biosynthesis pathway is also included in this list; though it is classified as central carbohydrate metabolism in KEGG due to its role in glycosidic bond formation, this molecule is an important precursor for nucleotide biosynthesis (both purines and pyrimidines) and synthesis of some amino acids (namely, tryptophan and histidine) (Hove-Jensen et al., 2017). Though many microbes are capable of producing their own nucleotides, some – especially lactic acid bacteria – are not and rely on uptake of exogenous nucleosides and bases, which are converted to nucleotides via salvage pathways (Nygaard, 2014; Kilstrup et al., 2005). Notably, these salvage pathways are not enriched in the IBD gut microbiome, suggesting that self-sufficiency in nucleotide biosynthesis (especially in the early stages in this process) is selected for in these communities. This also implies the importance of pyrimidine and purine cross-feeding in the healthy gut environment, which is supported by evidence that some gut microbes (e.g. Bacteroides vulgatus) actively secrete nucleosides in the colon (Wong et al., 2023; Teng et al., 2023).

Lipid pathways

Two lipid biosynthesis pathways – initiation and elongation of fatty acids – are enriched in IBD. Fatty acids are essential components of cell membranes and also serve as signaling molecules (Brown et al., 2023); thus, the ability to synthesize them is an important fitness determinant. For example, gut Bacteroides species that are deficient in sphingolipid production capabilities are much less resilient to oxidative stress than wild-type species (An et al., 2011). Since oxidative stress is a hallmark of IBD, it is possible that this environment selects for microbes capable of fatty acid biosynthesis.

Energy pathways

The Pta-Ack pathway is important for microbial energy production and adaptation to different growth conditions via the ‘acetate switch’, which enables either production or consumption of acetate depending on available nutrients (Wolfe, 2005). Short-chain fatty acids (SCFAs) such as acetate serve as important nutrients to intestinal epithelial cells. They also play a role in regulating gut barrier function and host immune responses (Martin-Gallausiaux et al., 2021; Zhang et al., 2022), and impaired absorption and oxidation of SCFAs can contribute to the development of IBD (Zhang et al., 2022). Acetate promotes host intestinal IgA production and thereby has a protective effect against gut inflammation (Wu et al., 2017), but acetate levels are reduced in children with IBD (Treem et al., 1994). Further study is required to determine the flux direction of the Pta-Ack pathway and whether it contributes to the reduction of acetate in the IBD gut environment.

CAM metabolism is categorized as a carbon fixation pathway in the KEGG MODULE database yet is a short (2-step) pathway utilizing enzymes required in other common metabolisms. Its first step is catalyzed by phosphoenolpyruvate carboxylase (PEPCK), an enzyme that is involved in gluconeogenesis, serine biosynthesis, and carbon skeleton conversions in the citric acid cycle (Yang et al., 2009). Its second step is catalyzed by malate dehydrogenases, a ubiquitous class of enzymes that convert 2-hydroxy acids to 2-keto acids and are involved in gluconeogenesis, the TCA cycle, glyoxylate bypass, and amino acid synthesis (Minarik et al., 2002; Musrati et al., 1998). The increase in this pathway in IBD gut microbiomes could be attributed in part to the increase in aerobic respiration due to elevated oxygen levels (Shah, 2016; Cevallos et al., 2019) and in part to the increase in amino acid biosynthesis capacity as evidenced by the multiple amino acid pathways that are also enriched.

Drug resistance pathways

The use of antibiotics to treat IBD and its complications is known to increase antibiotic resistance in the gut microbiome (Nitzan et al., 2016; Ledder, 2019) and several studies have noted that individuals exposed to antibiotics are more likely to develop IBD (Kronman et al., 2012; Ungaro et al., 2014; Ledder, 2019; Shaw et al., 2011). This potentially explains the enrichment of two drug resistance pathways in the IBD microbiome: efflux pump MepA (conferring multidrug resistance) and the bla system (conferring beta-lactam resistance), as higher rates of antibiotic exposure in this sample group naturally leads to selection for resistance phenotypes (Levy, 2000; Alekshun and Levy, 2007). Beta-lactamases in particular have been found with higher frequency in people with IBD (Vich Vila et al., 2018; Leung et al., 2012; Vaisman et al., 2013). Increased microbial drug resistance can heighten the risk of a severe infection such as Clostridium difficile infection (CDI) (Llor and Bjerrum, 2014). CDI already occurs with higher frequency in individuals with IBD (Jodorkovsky et al., 2010), though the higher incidence of CDI is not necessarily linked to chronic antibiotic use in these individuals (at least in one retrospective study of Crohn's disease) (Roy and Lichtiger, 2016). Regardless, antibiotic resistance is a global health problem that affects everyone, not just those with IBD.

4.7.6 Characterizing cohort-specific metabolic capacity across the gradient of health and disease

We then sought to evaluate the cohort-specific trends in metabolic capacity. We computed the median per-population copy number of the 33 IBD-enriched modules within each sample from each study. Again considering the heterogeneity within each sample group, we ordered the studies from most healthy to least healthy, using the cohort description from each publication to approximate relative healthiness based on the number and types of exclusions listed for healthy or non-IBD controls, or on the diagnostic criteria for people with IBD (Supplementary

Table 4.1a).

This was difficult considering the variable amount of detail provided by each study as well as the variability in what kinds of conditions were considered by each study. We placed more emphasis on exclusionary conditions that are more likely to directly affect the gut microbiome. Whenever two studies appeared to cover individuals of roughly the same healthiness, we generally considered the cohort with more specific gut-related exclusions to be healthier. For instance, (Le Chatelier et al., 2013) and (Raymond et al., 2016) had the most exclusionary conditions out of all the studies that contributed deeply-sequenced samples to the healthy group in our analysis – both of these studies excluded patients with gastrointestinal-related conditions like disease, surgery, and medication; medications affecting the immune system; or antibiotics. Because (Le Chatelier et al., 2013) explicitly excluded type-2 diabetes while (Raymond et al., 2016) did not, we placed (Le Chatelier et al., 2013) first, but the two cohorts are roughly similar in health status.

For studies contributing to the other sample groups, the (Lloyd-Price et al., 2019), (Schirmer et al., 2018b), and (Franzosa et al., 2019) studies were conducted by the same group, recruited people from the same hospital systems, and therefore have similar exclusionary and diagnostic criteria. (Schirmer et al., 2018b) had more specific exclusions for their non-IBD controls than (Franzosa et al., 2019) and was therefore considered a healthier cohort within that group. For the IBD samples, we considered similar cohorts to be more unhealthy if their diagnosis was stated to be confirmed using more lines of evidence. For example, (Schirmer et al., 2018b) diagnosed IBD based on a screening colonoscopy and included existing patients with diagnoses lasting over 5 years, (Lloyd-Price et al., 2019) utilized both endoscopic and histopathologic evidence for diagnosis, and (Franzosa et al., 2019) required endoscopic, histopathologic, and radiographic criteria. Regardless, these cohorts are likely extremely similar in healthiness. Yet there is no doubt that the (Vineis et al., 2016a) cohort is the least healthy of the IBD sample group – this cohort is composed of total proctocolectomy patients with ileal pouches, some of which developed pouchitis.

Ordering the per-sample median PPCN values along this gradient of cohort health indicates that the HMI metric for gut microbial metabolic capacity increases as host health decreases (Supplementary Figure 4.11a). Therefore, HMI adequately captures the variability in gut environment conditions that challenge microbial survival.

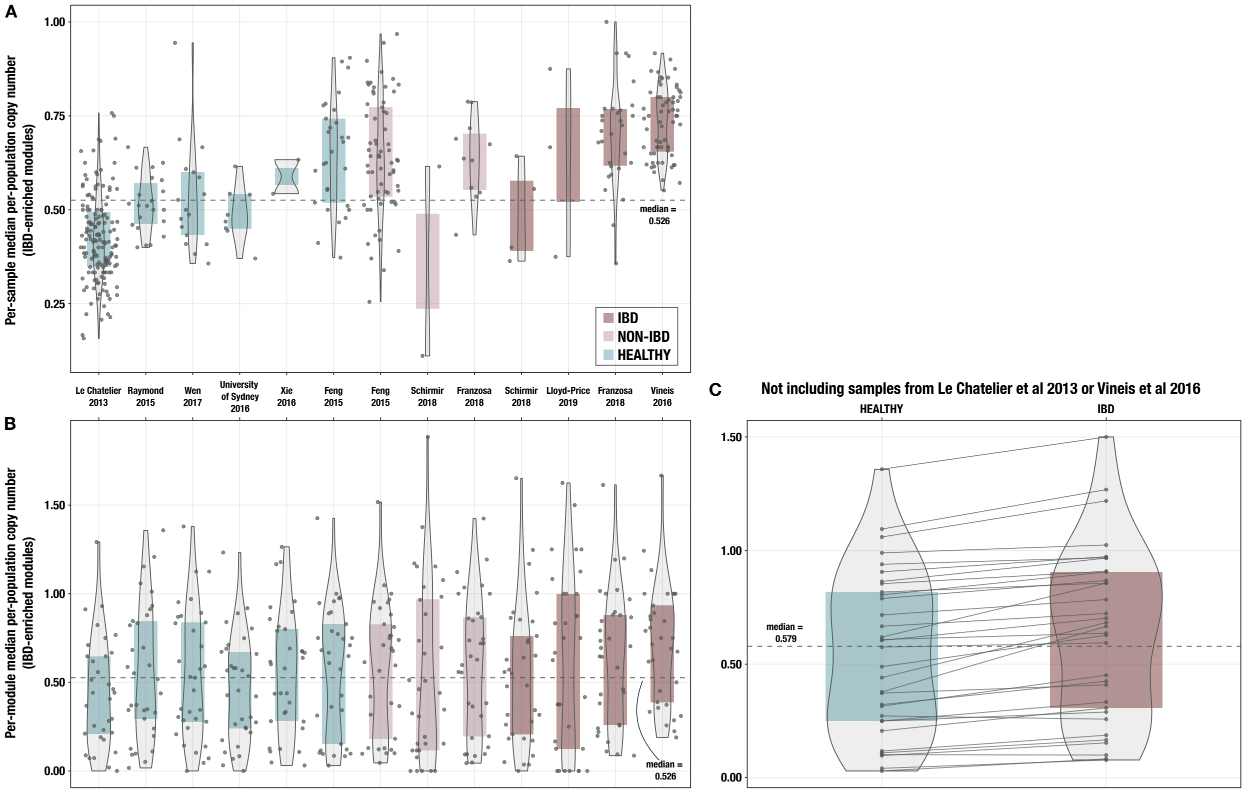


Figure 4.11: Boxplots of median per-population copy number of 33 IBD-enriched modules for samples from each individual cohort, A) with medians computed within each sample (ie, one point per sample) and B) with medians computed for each IBD-enriched module (ie, one point per module). The x-axis indicates study of origin. C) Boxplots of median per-population copy number of 33 IBD-enriched modules for the 115 samples in the deeply-sequenced set that are not from (Le Chatelier et al., 2013) or (Vineis et al., 2016a). The dashed line indicates the overall median for all 33 modules, and solid lines connect the points for the same module in each sample group.

4.7.7 Considering batch effect

One concern in comparing samples from multiple studies is that differential sample processing strategies could contribute to the signal between groups; in other words, batch effects could explain all or part of the trend we see between healthy, non-IBD, and IBD samples. This would likely be the case if we were comparing only one cohort per group. However, in this meta-analysis, samples are sourced from a wide variety of different studies, and it is unlikely that every study within a given group would be biased in the same direction. Furthermore, if we compute the median normalized copy number for each of the 33 IBD-enriched metabolic modules (summarized across all samples within a given study), these values are similarly distributed across the studies within a given sample group (Supplementary Figure 4.11b). Thus, the sample group explains more of the trend than study of origin.

However, two studies dominate our deeply-sequenced sample set: (Le Chatelier et al., 2013) contributes 151 (52.8%) of the healthy samples, and (Vineis et al., 2016a) contributes 64 (63.4%) of the IBD samples (Supplementary Table 4.1b). It is therefore still possible that cohort effect is responsible for the differential signal between the healthy and IBD group; that is, the IBD-enriched modules represent those that are primarily different between these two specific cohorts, rather than a general distinction between the overarching sample groups. To investigate this claim, we repeated the IBD-enrichment analysis on (i) (Le Chatelier et al., 2013) and (Vineis et al., 2016a) only; and (ii) the rest of samples. While the results obtained from the two larger studies tend to have smaller p-values, top IBD-enriched modules are broadly similar (Kendall correlation of Wilcoxon test p-values computed on two subsets: 0.59; see Supplementary Figure 4.12). This demonstrates that we are capturing generic signals across studies in our sample set.

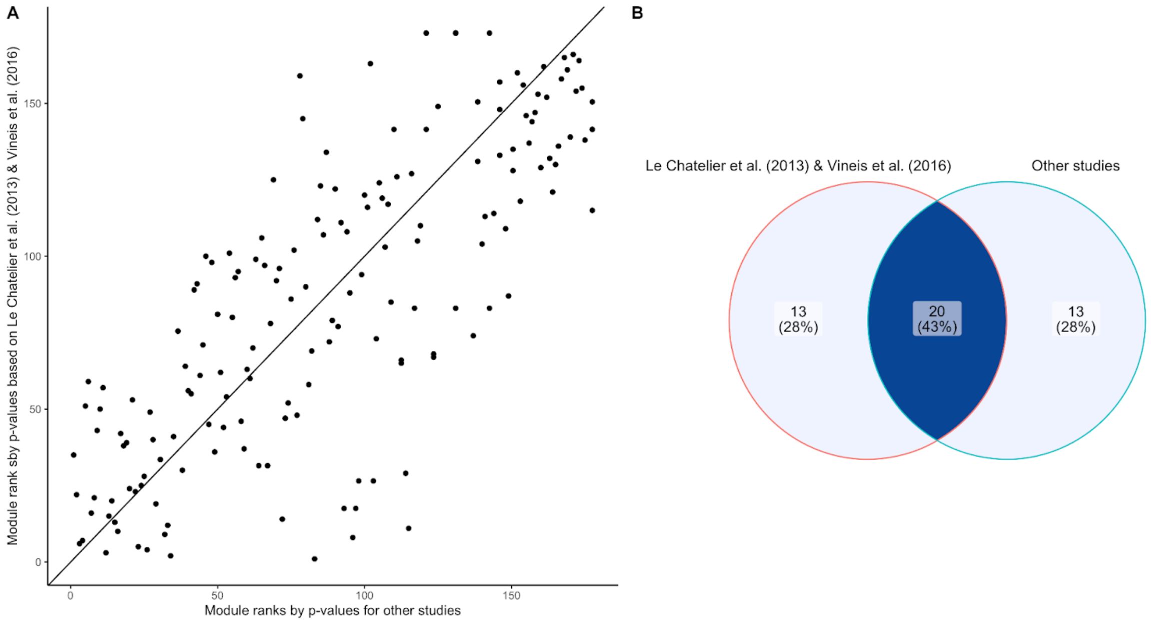


Figure 4.12: Assessing batch effect of the IBD-enrichment study. A) Scatter plot comparing the module ranks of Wilcoxon-Mann-Whitney p-values comparing IBD and healthy subjects on (Le Chatelier et al., 2013) and (Vineis et al., 2016a) (y axis) and the rest of our dataset (x axis). B) Venn diagram displaying the overlap of IBD-enriched modules identified by the 33 smallest p-values in (Le Chatelier et al., 2013) and (Vineis et al., 2016a) and the rest of our dataset. There is good agreement (20 out of 33) between the two sets of modules, indicating generalizability of the signals across studies used in our sample set.

4.7.8 Testing the generalizability of the metagenome classifier

To check whether performance of our logistic regression classifier was similar across the different studies in our sample set, we tested the model's performance using a leave-two-studies-out cross-validation strategy, whereby we trained the classifier on all samples except for those from one IBD study and one healthy study, and then tested it using samples from the two studies that were left out, for a total of 24 folds. Performance was quite variable across the different folds, as expected considering the large range of sample sizes from each study and the variability in health status of each cohort. The best overall performance occurred when testing on healthy samples from (Le Chatelier et al., 2013), with average accuracy of 89.9% across 3 folds. The worst performance occurred when testing on healthy samples from (Feng et al., 2015), with average accuracy of 43.1% across 4 folds. In the fold leaving out healthy samples from (Le Chatelier et al., 2013) and IBD samples from (Vineis et al., 2016a), no IBD-enriched modules had p-values below our FDR-adjusted significance threshold of 2e-10 and therefore no classifier was trained. As these two studies contributed the largest number of samples to our deeply sequenced subset (Le Chatelier et al. 2013: n = 151 out of 330 or 45.8%, all of which were healthy samples. Vineis et al. 2016: n = 64 out of 330 or 19.4%, all of which were IBD samples), we again considered that cohort-specific or study-specific effects could be driving the differential signal between healthy and IBD samples. To test this, we removed the samples from (Le Chatelier et al., 2013) and (Vineis et al., 2016a) and ran 10-fold cross-validation using an 80-20 train-test split of the remaining 115 samples (37 IBD, 78 healthy), using the 33 IBD-enriched modules (computed from the full sample set) as features. We found that the model performed better than a naive classifier, with an average fold accuracy of 66.5%, average true Healthy rate of 69.4%, and an average true IBD rate of 61%. Therefore, while a portion of the signal in our initial analysis is indeed attributable to the differences between samples from (Le Chatelier et al., 2013) and (Vineis et al., 2016a), we are still able to capture an IBD-specific signal across the other studies using this set of IBD-enriched pathways.

Furthermore, we note that the two dominating studies represent individuals at the extremes of the health gradient across our sample set, as described previously. The (Le Chatelier et al., 2013) cohort, with its numerous exclusionary conditions, contains the healthiest individuals, while the (Vineis et al., 2016a) cohort of proctocolectomy and pouchitis patients contains the unhealthiest. It is therefore unsurprising that there is a large contrast in the metabolic potential of the gut microbiome in these individuals, considering the biological differences in their respective gut environments. This is also supported by the aforementioned ability of HMI to resolve the variability in host health, as demonstrated in Supplementary Figures 4.10b and 4.11a.

4.8 Supplementary Tables

This section's supplementary tables are accessible via `doi:10.6084/m9.figshare.22679080`.

Table 4.1: Samples and cohorts used in this study. a) Description of studies/cohorts providing publicly-available gut metagenomes from healthy people, non-IBD controls, and people with IBD. For each study, we note the sample groups it contributes metagenomes to; whether or not those samples were sufficiently deeply sequenced to be included in the main analyses; the country of origin of the samples; the sample type (fecal metagenome or ileal pouch luminal aspirate); the number of samples it contributes to each group before and after applying the sequencing depth threshold; and cohort details/exclusions as described within the study. b) Description of 408 samples included in the primary analyses of this manuscript (ie, those with sufficient sequencing depth of >= 25 million reads), including their associated diagnosis (ulcerative colitis (UC), Crohn's disease (CD), non-IBD, healthy, colorectal cancer with adenoma (CRC_ADENOMA), or colorectal cancer with carcinoma (CRC_CARCINOMA)); study of origin; sample group; sequencing depth; and number of microbial populations estimated to be represented within the metagenome. c) Description of all samples initially considered and their SRA accession numbers. d) The number of gene calls and the number/proportion of annotations per gene call for KOfams, COGs, and Pfams in each sample. e) Description of the 57 antibiotic time-series gut metagenomes from (Palleja et al., 2018) used for classifier testing, including SRA accession number; sampling day in the time series; sequencing depth; and estimated numbers of microbial populations represented in the sample.

Table 4.2: Metabolism data in metagenomes. a) Description of the 33 KEGG modules enriched in IBD samples, including: module name, KEGG categorization, and definition; their median per-population copy numbers (PPCN) in the healthy sample group and IBD sample group; the p-value, FDR-adjusted p-value, and W statistic from the per-module Wilcoxon Rank Sum test used to determine enrichment in IBD; the difference between its median PPCN in IBD samples and median PPCN in healthy samples ('effect size'); the fraction of samples in which the module occurs with non-zero copy number; whether the module is also enriched in the HMI populations analyzed in (Watson et al., 2023); the number of total enzymes in the module; the number of total compounds in the modules; and the numbers and proportions of shared enzymes or compounds between this module and the other IBD-enriched modules. b) Description of all 179 KEGG modules with non-zero copy number in at least one metagenome. Most of the columns match the corresponding column in sheet (b) with the exception of the 'enrichment status' column, which indicates whether the module was found to be enriched in the IBD samples in this study ('IBD_ENRICHED'), in the high-metabolic independence genomes in (Watson et al., 2023) ('HMI_ENRICHED'), in both ('HMI_AND_IBD'), or in neither ('OTHER'). c) Matrix of stepwise copy number of each module in each deeply-sequenced gut metagenome. d) Per-population copy number of each module in each deeply-sequenced gut metagenome in the IBD, non-IBD and healthy sample groups. e) Per-population copy number of each module in each antibiotic time-series sample from (Palleja et al., 2018).

Table 4.3: GTDB genome data. a) List of 338 GTDB representative genomes identified as gut microbes, their taxonomy, metabolic independence score, classification as high metabolic independence ('HMI') or not ('non-HMI'), genome length in base pairs, and number of gene calls. b) Matrix of stepwise completeness of each module in each genome. c) Matrix of genome detection in each deeply-sequenced gut metagenome in the IBD, non-IBD, and healthy sample groups. d) Percent abundance of each genome in each deeply-sequenced gut metagenome. e) Per-genome proportion of samples from each sample group that the genome is detected in using a threshold of 50% (ie, at least half of the genome sequence is covered by at least one sequencing read in a given sample). f) Per-sample proportion of detected genomes that are classified as HMI. g) Average completion of each IBD-enriched module within the HMI genome group and the non-HMI genome group, as well as the difference between these values.

Table 4.4: Metagenome classifier information. a) Details and performance of previously-published classifiers for IBD and IBD subtypes. For each classifier, we summarize the cohort details as described by the study; the size of training datasets and validation datasets (if any); the type(s) of samples, data, and extracted features used for classification; the target classes (that is, what the samples were being classified as); the classifier type and training/validation strategy; and the performance metrics as reported by the study. b) Classification of each (Palleja et al., 2018) metagenome by our logistic regression model trained for distinguishing IBD vs healthy samples on the basis of PPCN data for IBD-enriched modules. This table describes whether the sample was classified as healthy ('HEALTHY') or stressed ('IBD', which we consider to be equivalent to an identification of gut stress), and also whether the sample had low sequencing depth (< 25 million reads) or not. c) Summary of the performance of our metagenome classifier across different training/validation strategies using the IBD and healthy metagenome samples. It also includes the details of our final classifier trained on all 330 samples, though performance data is not available for this model since there were no IBD/healthy samples left for validation – however, see manuscript for its performance on the (Palleja et al., 2018) antibiotic time-series dataset. The subsequent sheets include per-fold data and performance information for each train-test strategy: d) random split cross-validation (25-fold) on PPCN data; e) leave-two-studies-out cross-validation (24-fold); and f) (10-fold) cross-validation leaving out samples from the two dominating studies in our dataset, (Le Chatelier et al 2013) and (Vineis et al 2016).

Table 4.5: Details of available software for metabolism estimation. For each tool (including the one published in this study), we summarize: the software category (based upon the tool's architecture and mode of use); its metabolism reconstruction strategy (whether it is a pathway prediction tool or a modeling tool or both); the data source(s) it uses for enzyme and metabolic pathway information; how it calculates pathway completeness or generates models (depending on reconstruction strategy); what input and output types it accepts/generates; any additional capabilities as advertised by the tool's publication; whether or not the tool is open-source; the program type; and what language(s) it is developed in (if known). The reference publication and code repository or webpage for each tool is also included.

CHAPTER 5

A POWERFUL OPEN-SOURCE SOFTWARE FRAMEWORK PROPELS ADVANCED ‘OMICS RESEARCH

5.1 Preface

The previous chapters have detailed my work in the field of microbial ‘omics research with a focus on the study of metabolism in the human gut. Indeed, my implementation of the metabolism reconstruction framework and my work on metabolic independence as a determinant of microbial resilience in the face of gut stress represent my major intellectual contributions to this field. Now, I will take a step back and place my work in the context of a larger goal: facilitating advanced ‘omics research with open-source, integrated software solutions.

Of the numerous bioinformatics software for analyzing ‘omics data, the vast majority consist of predefined workflows and calculations for a small set of analysis tasks. While these programs are easy to use and offer a quick data-in, results-out solution for those with less computational training, these benefits come with the cost of limiting users’ options for original and flexible analyses. For ‘omics research to advance more effectively, scientists must have access to tools that are not only usable, but facilitate innovation by providing greater control over the direction of their research. Developing and providing access to such platforms is a key area of improvement for the ‘omics field.

The open-source software movement is critical to this effort. First, it improves transparency and reproducibility by exposing a software’s methodology to the public. Second, it facilitates community-driven software extensibility, enabling researchers to update the software with desired features – adapting the platform to their research needs rather than adapting their research to available tools. Yet, making a tool open-source does not necessarily make it usable; for that, clear and extensive documentation is required.

A significant part of my PhD work has been focused on improving researchers’ access to tools for advanced ‘omics research – not only by implementing software, but also by developing educational resources. I have already discussed my framework for metabolism reconstruction, which was designed (with input from numerous collaborators and colleagues) to facilitate flexible analyses of metabolic potential. I implemented this framework within anvi’o, an open-source software ecosystem, because it follows the philosophy of supporting varied, extensible, and integrated analyses of ‘omics data. As an anvi’o developer, I have been able to increase the accessibility of this platform by writing extensive documentation, tutorials, and blog posts describing how to use various aspects of the software (including, but not limited to, the metabolism framework), learn ‘omics vocabulary, get help from the anvi’o community, and extend the online help pages. Several of these resources are linked from the web page https://anvio.org/people/ivagljiva/. In addition, I have developed and participated in a number of workshops and seminars to teach users how to conduct analyses on their own data. For example, I lectured on metabolism reconstruction in the 2021 workshop on Emerging Bioinformatics Applications for Microbial Ecogenomics (https://maignienlab.gitlab.io/ebame6/), and I created hybrid workshops for teaching metagenomics (along with my colleague Matthew Schechter) to audiences of 50-70 participants from multiple career stages (https://microbiome.uchicago.edu/resources/tmc-news/anvio-omics-workshop-2023, https://microbiome.uchicago.edu/tmc-news/metagenomics-and-metabolomics-workshop).

This chapter describes a few of my efforts to promote advanced ‘omics science in systems beyond the human gut environment and areas beyond metabolism. The examples are drawn from my work on marine microbiomes. In section 5.2, I introduce a study utilizing the metabolism reconstruction framework to investigate nitrogen fixers in the global oceans. This study identifies heterotrophic bacterial diazotrophs as more abundant than cyanobacterial diazotrophs, challenging the current paradigm labeling cyanobacteria as the primary marine nitrogen fixers. In section 5.3, I demonstrate one of the several tutorials I have written to encourage advanced usage of the anvi'o platform; this particular example teaches people how to leverage the metabolism reconstruction framework for targeted binning of microbial population genomes with specific metabolic capabilities. The tutorial also includes an interesting scientific result: binning of a novel nitrogen-fixing population from the Arctic Ocean. Finally, section 5.4 describes my work in generating shareable, reproducible, integrated data packages for related (and especially large-scale) 'omics datasets on microbial populations. These 'digital microbes' were initially developed to fit the collaboration needs of a large research consortium studying marine carbon cycling, but are of general interest as a data-sharing tool for anyone in the 'omics field.

5.2 Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean

This section is derived from the following publication:

Tom O. Delmont, Juan José Pierella Karlusich, **Iva Veseli**, Jessika Fuessel, A. Murat Eren, Rachel A. Foster, Chris Bowler, Patrick Wincker and Eric Pelletier. Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. ISME J 16, 927–936 (2022). `https://doi.org/10.1038/s41396-021-01135-1`.

5.2.1 Introduction

Plankton communities in the sunlit ocean consist of numerous microbial lineages that influence global biogeochemical cycles and climate (Boyd, 2015; Charlson et al., 1987; Falkowski et al., 1998; Arrigo, 2005; Sanders et al., 2014; de Vargas et al., 2015). Phototrophic primary productivity is often constrained by the amount of bioavailable nitrogen (Moore et al., 2013; Tyrrell, 1999), a critical element for cellular growth and division. Only a few bacterial and archaeal populations within the large pool of marine microbial lineages are capable of performing nitrogen fixation, thereby providing an essential source of new nitrogen to phytoplankton (Dos Santos et al., 2012; Zehr and Capone, 2020; Zehr et al., 2003). These populations are known as diazotrophs and represent key marine players that sustain primary productivity in large oceanic regions (Zehr and Capone, 2020). Globally, marine nitrogen fixation is at least as important as the nitrogen fixation on land performed by *Rhizobium* bacteria in symbiosis with plants (Galloway et al., 2004).

Cyanobacterial diazotrophs are abundant in open ocean surface waters and provide a substantial portion of bioavailable nitrogen (Carpenter et al., 1992; Carpenter and Romans, 1991; Karl et al., 1997). They include populations within the genus *Trichodesmium* (Capone et al., 1997; Dyhrman et al., 2006; Pierella Karlusich et al., 2021) and several lineages that enter symbiotic associations with eukaryotes (e.g., *Richelia* (Gómez et al., 2005; Hilton et al., 2013), the *Candidatus* Atelocyanobacterium also labeled UCYN-A (Martínez-Pérez et al., 2016; Tripp et al., 2010)) or can exist as free-living cells such as *Crocosphaera watsonii* also labeled UCYN-B (Moisander et al., 2010; Montoya et al., 2004). A wide range of non-cyanobacterial diazotrophs has also been detected using amplicon surveys of the *nifH* gene required for nitrogen fixation. These molecular surveys showed non-cyanobacterial diazotrophs occurring in lower abundance compared to their cyanobacterial counterparts in various oceanic regions (e.g., (Church et al., 2005, 2008; Zehr et al., 2007; Fong et al., 2008; Moisander et al., 2008; Benavides et al., 2016; Langlois et al., 2005)) but could also be relatively abundant in some samples (e.g., (Man-Aharonovich et al., 2007; Bombar et al., 2016; Farnelid et al., 2011; Riemann et al., 2010; Moisander et al., 2017; Moreira-Coello et al., 2019)). Overall, decades of *Trichodesmium* cultivation, flow cytometry, molecular surveys, imaging, and in situ nitrogen fixation rate measurements have led to the emergence of a view depicting cyanobacterial diazotrophs as the principal marine nitrogen fixers (Luo et al., 2012).

Recently, a genome-resolved metagenomic survey exposed free-living heterotopic bacterial diazotrophs (HBDs) abundant in the surface waters of large oceanic regions (Delmont et al., 2018). This first set of genome-resolved HBDs from the open ocean was subsequently found to express their *nifH* genes in situ using metatranscriptomics (Salazar et al., 2019). However, the sole focus on free-living bacterial cells in this survey excluded not only key cyanobacterial players but also other diazotrophs that might occur under the form of aggregates, preventing a comprehensive investigation of diazotrophs in the sunlit ocean. Here we used nearly nine hundred *Tara* Oceans metagenomes (Sunagawa et al., 2020). to create a genomic database corresponding to free-living, as well as filamentous, colony-forming, particle-attached, and symbiotic bacterial and archaeal populations occurring in surface waters of the global ocean. Our genomic database includes dozens of previously unknown HBDs abundant in different size fractions and oceanic regions all of which express their *nifH* genes in situ. Most notably, we found HBDs to be more abundant compared to cyanobacterial diazotrophs in metagenomes covering most surface open oceans and seas, revealing their prevalence also under the form of putative large aggregates within plankton and suggesting they play a considerable role in the marine nitrogen balance.

5.2.2 Results and Discussion

Part one: Genome-wide metagenomic analyses

Nearly 2,000 manually curated bacterial and archaeal genomes from the 0.8-2,000 µm planktonic cellular size fractions in the surface oceans and seas. We performed a comprehensive genome-resolved metagenomic survey of bacterial and archaeal populations from the euphotic zone of polar, temperate, and tropical oceans using 798 metagenomes derived from the *Tara* Oceans expeditions. They correspond to surface waters and deep chlorophyll maximum (DCM) layers from 143 stations covering the Pacific, Atlantic, Indian, Arctic, and

Southern Oceans, as well as the Mediterranean and Red Seas, encompassing eight plankton size fractions ranging from 0.8 µm to 2000 µm (Supplementary Table 5.1). These 280 billion reads were already used as inputs for 11 metagenomic co-assemblies using geographically bounded samples to recover eukaryotic metagenome-assembled genomes (MAGs) (Delmont et al., 2022). Here, we recovered nearly 2,000 bacterial and archaeal MAGs from these 11 co-assemblies.

We combined these MAGs with 673 MAGs previously generated from the 0.2 µm to 3 µm size fraction (93 metagenomes) (Delmont et al., 2018) to create a culture-independent, non-redundant (average nucleotide identity <98%) genomic database for microbial populations consisting of 1,778 bacterial and 110 archaeal MAGs, all exhibiting >70% completion (average completion of 87.1% and redundancy of 2.5%; Supplementary Table 5.2). We manually characterized and curated these 1,888 MAGs using a holistic framework within anvi'o (Eren et al., 2015, 2021b) that relied heavily on differential coverage across metagenomes within the scope of their associated co-assembly. This genomic database has a total size of 4.8 Gbp, with MAGs affiliated to Proteobacteria (n = 916), Bacteroidetes (n = 314), Planctomycetes (n = 154), Verrucomicrobia (n = 128), Euryarchaeota (n = 105), Actinobacteria (n = 68), Cyanobacteria (n = 51), Chloroflexi (n = 36), *Candidatus* Marinimicrobia (n = 30), *Candidatus* Dadabacteria (n = 10) and 24 other phyla represented less than 10 times (Supplementary Table 5.1). We used their distribution and gene content to survey marine diazotrophs in the open ocean without relying on cultivation or *nifH* amplicon surveys.

A genomic collection of 48 marine diazotrophs abundant in the open ocean. While none of the 110 archaeal MAGs indicated a diazotrophic lifestyle, a total of 48 bacterial MAGs contained genes encoding the catalytic (*nifHDK*) and biosynthetic (*nifENB*) proteins required for nitrogen fixation (Supplementary Table 5.3). Among these, only one MAG (Gammaproteobacterial) lacked the *nifH* gene, which is likely a result of the limitations inherent to genome-resolved metagenomics. Based on the taxonomic signal and the occurrence or absence of genes required for a photosynthetic lifestyle, these MAGs could be categorized into eight cyanobacterial diazotrophs and 40 HBDs. Their estimated completion averaged 93.4%, suggesting they correspond to near-complete environmental genomes (Figure 5.1 and Supplementary Table 5.4).

The reconstructed cyanobacterial MAGs recapitulated findings of major marine diazotrophs previously discovered within this phylum and for which a genome (partial or complete) had been characterized previously using either available cultures or sorted cells from flow cytometry: UCYN-A1 (ANI of 99.3%) and UCYN-A2 (ANI of 99.6%), *Crocosphaera watsonii* (strain WH-8501; ANI of 99.4%), *Richelia intracellularis* (strain RintHH01; ANI of 99.5%), *Trichodesmium erythraeum* (strain IMS101; ANI of 99%), and *Trichodesmium thiebautii* (strain H9-4; n = 2 with ANI of 98.7% and 98%). Interestingly, while the two *Trichodesmium thiebautii* populations displayed high genomic similarity (ANI of 97.9%) and correlated across 81 metagenomes with signal for this lineage (R^2 = 0.93), the mean coverage ratio revealed one population that was dominant at three sites of the North Atlantic Ocean while the second population was relatively more abundant in the Indian Ocean, Pacific Ocean and Red Sea (Supplementary Figure 5.5). In addition, one MAG corresponded to an unknown population we tentatively named '*Candidatus* Richelia exalis' given its close evolutionary relationship with *R. intracellularis* (e.g., ANI of 87.3% when compared to the strain RintHH01; see Supplementary Table 5.3 for more comparisons) (Figure 5.1). The strong signal of '*Candidatus* Richelia exalis' in the large size fractions, similar to *R. intracellularis*, and their comparable functional traits (see following section) suggest this species also leads a symbiotic lifestyle.

Compared to the cyanobacterial diazotrophs that were already well characterized prior to this genome-resolved metagenomic survey, the HBDs we recovered substantially increase the number of known diazotrophic populations. In addition to eight previously characterized HBDs reconstructed from the 0.2–3 µm size fraction (Delmont et al., 2018) (five of which were replaced by MAGs characterized from the larger size fractions that displayed improved completion statistics), the genomic database includes 32 additional HBDs belonging to the phyla Deltaproteobacteria (eight HBDs; six new *nifH* genes when compared to a comprehensive set of reference databases (Pierella Karlusich et al., 2021), see methods), Gammaproteobacteria (16 HBDs; four new *nifH* genes), Planctomycetes (three HBDs; one new *nifH* gene), Alphaproteobacteria (eight HBDs; three new *nifH* genes), Epsilonproteobacteria (2 HBDs; two new *nifH* genes), and Verrucomicrobia (three HBDs; three new *nifH* genes) (Figure 5.1 and Supplementary Table 5.5). Interestingly, some of the newly identified *nifH* gene sequences are incompatible with the design of several primers frequently used in *nifH* gene amplicon surveys (Supplementary Figure 5.6 and Supplementary Table 5.6). This was especially true of the “nifH4” primer (round one of widely utilized nested primers (Farnelid et al., 2011; Gaby and Buckley, 2012; Turk-Kubo et al., 2014; Zehr and Turner, 2001)) (Figure 5.1) that appears incompatible with most HBDs identified in this study.

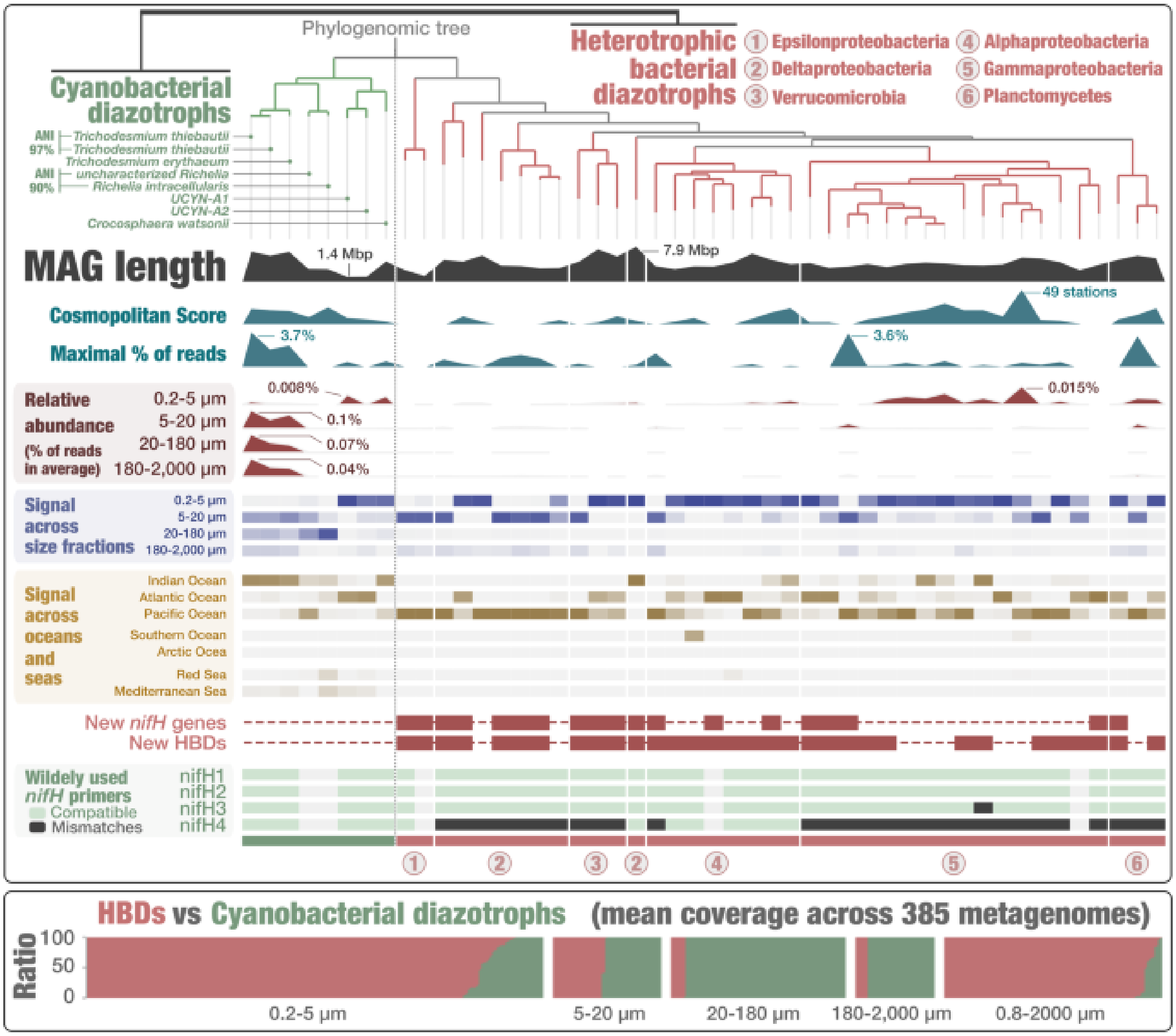


Figure 5.1: The phylogeny of 48 marine bacterial diazotrophs. Top panel displays a phylogenomic tree of the 48 diazotroph MAGs using 37 gene markers and visualized with anvi'o (A. Murat Eren et al. 2015). Additional layers of information display the length of MAGs alongside environmental signal computed using genome-wide metagenomic read recruitments across 937 metagenomes, and *nifH* primer compatibilities (only full length and non-fragmented *nifH* genes were considered). For each MAG, the "maximal percent of mapped reads" layer displays the percent of mapped reads corresponding to the sample for which this metric was the highest among all 937 metagenomes. Thus, this sample is MAG dependent. In contrast, the "relative abundance" layers display for each MAG the average number of mapped reads across samples corresponding to the same size fraction. Bottom panel displays the ratio of cumulative genome-scale mean coverage between eight cyanobacterial diazotrophs (green) and 40 HBDs (red) across 385 metagenomes we organized into five size fractions.

The emergence of three main functional groups for marine HBDs. In order to provide a global view of functional capabilities among the 48 diazotrophs, we accessed functions in their gene content using COG20 functions, categories and pathways (Galperin et al., 2021), KOfam (Aramaki et al., 2020), KEGG modules, and classes (Kanehisa et al., 2017) from within the anvi'o genomic workflow (Eren et al., 2015) (Supplementary Table 5.7). Genomic clustering based on the completeness of 322 functional modules exposed four distinct groups: (1) the cyanobacterial diazotrophs, (2) HBDs dominated by Alphaproteobacteria, (3) HBDs associated with Gammaproteobacteria, and finally (4) HBDs organized in closely related subgroups corresponding to Deltaproteobacteria, Epsilonproteobacteria, Verrucomicrobia and Planctomycetes (Figure 5.2). Several HBDs have the metabolic capacity to generate energy using pathways other than aerobic respiration. One population associated with Alphaproteobacteria (genus *Marinibacterium*) for example encodes anoxygenic photosystem II as well as all pathways required for aerobic respiration, thiosulfate oxidation and dissimilatory nitrate reduction to ammonia. Within the HBD group affiliated with Alphaproteobacteria, the majority of populations encode the SOX complex necessary for thiosulfate oxidation (Supplementary Table 5.7) and one population encodes the genes required for denitrification. Among the HBDs affiliated with Deltaproteobacteria, a large majority encodes the pathway for dissimilatory sulfate reduction and mostly lack metabolic pathways required for aerobic respiration. Four representatives of the Gammaproteobacteria have the metabolic potential for denitrification and one population can generate energy via thiosulfate oxidation, a capacity that is also encoded in one of the HBDs affiliated with Epsilonproteobacteria. The metabolic pathway for dissimilatory nitrate reduction to ammonia can be found in all taxonomic groups (occurrence: 20–100%) (Supplementary Table 5.7). This intriguing metabolic diversity among HBDs indicates their potential importance in major biogeochemical cycles. All deltaproteobacterial HBDs encode the complex biosynthesis pathway for cobalamin, also found in a majority of cyanobacterial diazotrophs (including the symbionts) (Supplementary Table 5.7). Only the final 5–6 steps of cobalamin synthesis are also encoded in HBD populations associated with Gamma- and Alphaproteobacteria (Supplementary Table 5.7). Overall, we found the HBDs to be functionally more diverse compared to their cyanobacterial counterparts.

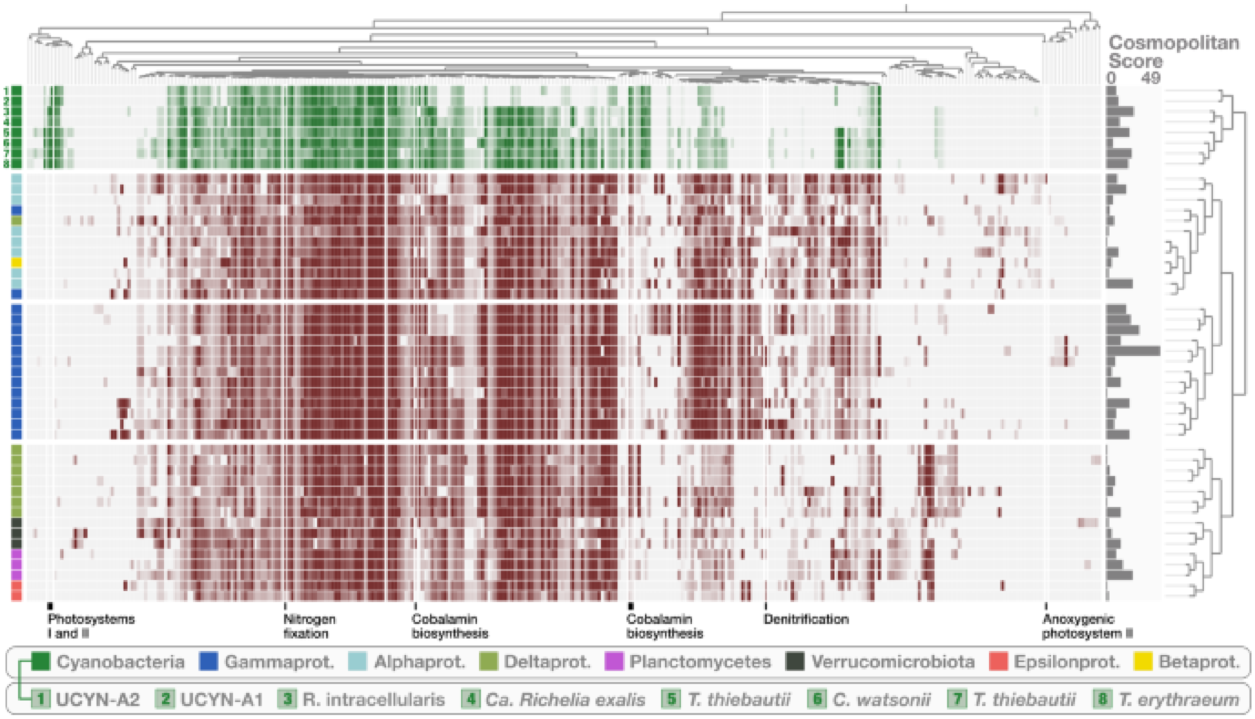


Figure 5.2: Functional lifestyle of marine diazotrophs. The figure displays a heatmap of the completeness of 322 functional modules across the 48 diazotrophic MAGs. Clustering of MAGs and modules is based on completeness values (Euclidean distance and ward linkage) and the data were visualized using anvi'o (A. Murat Eren et al. 2015). The cosmopolitan score corresponds to the number of stations in which a given MAG was detected (cut-off: >25% of the MAG is covered by metagenomic reads).

HBDs are generally more abundant compared to cyanobacterial diazotrophs. The 48 diazotrophs occurred at up to 49 stations (out of 119 stations considered to compute this cosmopolitan score) and recruited up to 3.7% of metagenomic reads (Figures 5.1, 5.2, and Supplementary Table 5.2) when considered individually. Yet, the locally most abundant diazotrophs were not the most widespread (R^2 of 0.007 when comparing the maximal number of recruited reads and cosmopolitan score). We detected no diazotrophs in the Arctic Ocean or the Red Sea, only a single HBD in the Southern Ocean (Delmont et al., 2018) and very few representatives in the Mediterranean Sea. Within temperate and tropical open ocean regions, marine diazotrophs affiliated with Epsilonproteobacteria, Deltaproteobacteria and Verrucomicrobia mostly occurred in the Pacific Ocean. The remaining diazotrophic lineages occurred in the Pacific, Indian, and Atlantic Oceans. Within the group of cyanobacterial diazotrophs, the two populations of *Trichodesmium thiebautii* were highly abundant in some of the large size fractions and generally prevailed in the Indian Ocean (Figure 5.1). The overall geographic distribution of diazotrophs indicates that the Pacific Ocean is dominated by HBDs, corroborating previously observed trends (Pierella Karlusich et al., 2021; Farnelid et al., 2011; Delmont et al., 2018).

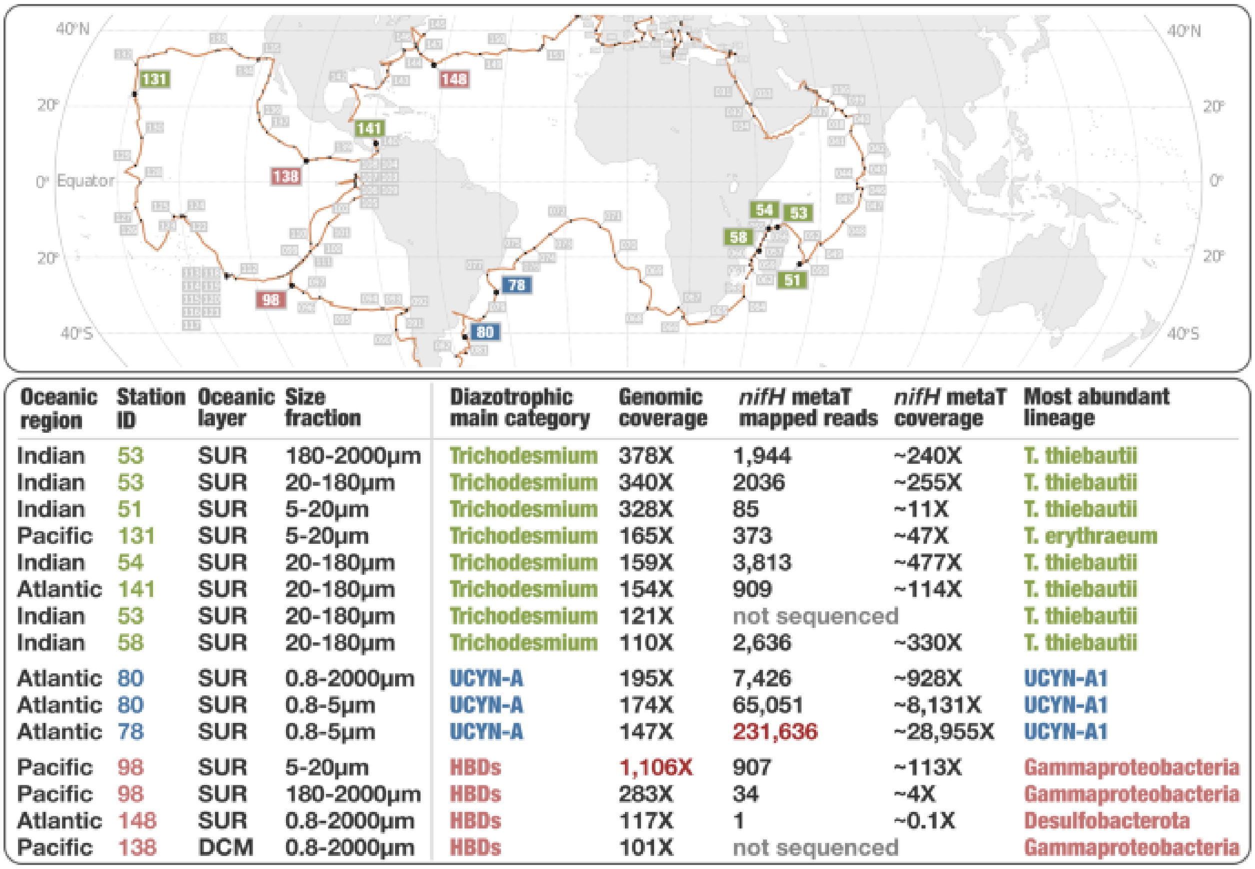
The majority of the 48 diazotrophs were associated with the 0.2–5 µm size fraction that covers most of the free-living bacterial cells, while the remaining diazotrophs were detected in the 5–20 µm (n = 15) and 20–180 µm (n = 2; *Richelia intracellularis* and '*Candidatus* Richelia exalis') size fractions (Figure 5.1, Supplementary Table 5.4). We then computed the ratio of cumulative mean coverage (i.e., number of times a genome is sequenced) between the eight cyanobacterial diazotrophs and 40 HBDs across 385 metagenomes organized by size fraction (552 metagenomes with no signal for any of the 48 diazotrophs were not considered here). Overall, HBDs displayed a cumulative mean coverage superior to that of cyanobacterial diazotrophs in 250 metagenomes, compared to 135 for the latter. Furthermore, a clear signal emerged in which HBDs were more abundant in most metagenomes representing the 0.2–5 µm (86.5%) and 0.8–2000 µm (92.6%) size fractions while cyanobacterial diazotrophs predominated in the 20–180 µm (92.3%) and 180–2000 µm (86.2%) size fractions (Figure 5.1, bottom panel). Finally, the 5–20 µm size fraction was more balanced between HBDs and cyanobacterial diazotrophs.

The 0.8–2000 µm size fraction was not collected in the Mediterranean Sea, Red Sea and Indian Ocean, but became an integral part of *Tara* Oceans sampling efforts in the other oceans

(Pesant et al., 2015). This broad size range fraction provides a valuable metric to compare the relative abundance of diazotrophs that otherwise would be separated between the different size fractions. In other words, this size fraction could be used to effectively compare the genomic signal of diazotrophs corresponding to free-living, particle-attached, filamentous, colony-forming, and symbiotic cells, provided they (or their hosts) pass through 2 mm filter pores, either undamaged or fragmented (e.g., *Trichodesmium* colonies are known to be fragile). While uncertainty remains in the Indian Ocean, trends from metagenomes corresponding to the 0.8–2000 µm size fraction in other regions largely mirrored the free-living size fraction and were typically dominated by HBD signal. Metagenomes representing microbial populations from the 0.2–3 µm and 0.8–2000 µm size fractions indicate that HBDs are more abundant compared to their cyanobacterial counterparts in most of the surface oceans investigated here.

Co-occurrence of HBDs in large size fractions from a Pacific Ocean station. We detected a considerable metagenomic signal for HBDs at Station 98 in the South Pacific Ocean (Figure 5.3; Supplementary Table 5.4), which was also found using reference *nifH* genes (Pierella Karlusich et al., 2021). Station 98 includes five surface and three DCM metagenomes covering all size fractions except for 0.8–2000 µm. The only cyanobacterial diazotroph we detected in this metagenomic set was ‘*Candidatus* Richelia exalis’ with a mean coverage of just 0.4X in the 20–180 µm size fraction of the surface layer. The 40 HBDs remained undetected in the DCM and only two HBDs were marginally detected in the 0.2–3 µm size fraction of the surface layer. In marked contrast, 14 HBDs were detected in the 5–20 µm, 20–180 µm, and 180–2000 µm size fractions of surface waters with a cumulative mean coverage reaching 1,106X (i.e., their genomes were sequenced cumulatively more than one thousand times in this particular metagenome), 15X and 283X, respectively. Such a high genomic coverage for bacterial populations in large size fractions is unusual and exceeded the maximum signal associated with UCYN-A and *Trichodesmium* in this study (Figure 5.3; Supplementary Table

5.4). The 14 HBDs were affiliated with Deltaproteobacteria (n = 5), Alphaproteobacteria (n = 2), Gammaproteobacteria (n = 2), Epsilonproteobacteria (n = 2), Planctomycetes (n = 2) and Verrucomicrobia (n = 1). Surface waters at Station 98 were nitrogen depleted (nitrate near the detection limit at 0.001 µm; Supplementary Table 5.1), likely providing favorable conditions for a diverse assemblage of HBDs that were particularly abundant within the large size fractions. Lack of signal in the small size fraction suggests that similar populations might be missed in oceanic sampling that typically restricts bacterial analyses to free-living cells. Mechanisms maintaining diazotrophs in large plankton size fractions have yet to be fully elucidated (Farnelid et al., 2011, 2010, 2019; Foster et al., 2006; Scavotto et al., 2015; Zani et al., 2000). Our results nonetheless support recent observations in coast and estuary linking active HBDs to large aggregates (Geisler et al., 2019; Martínez-Pérez et al., 2018). Exopolymer particles and aggregates might create low-oxygen microenvironments favorable for nitrogen fixation in marine environments (Rahav et al., 2013), as observed in laboratory cultures (Martínez-Pérez et al., 2018; Bentzon-Tilia et al., 2015). Thus, we suggest that HBDs formed a considerable number of large aggregates (up to >180 µm in size) at Station 98 in order to optimize their nitrogen fixation capabilities.



Oceanic region	Station ID	Oceanic layer	Size fraction	Diazotrophic main category	Genomic coverage	*nifH* metaT mapped reads	*nifH* metaT coverage	Most abundant lineage
Indian	53	SUR	180-2000µm	Trichodesmium	378X	1,944	~240X	T. thiebautii
Indian	53	SUR	20-180µm	Trichodesmium	340X	2036	~255X	T. thiebautii
Indian	51	SUR	5-20µm	Trichodesmium	328X	85	~11X	T. thiebautii
Pacific	131	SUR	5-20µm	Trichodesmium	165X	373	~47X	T. erythraeum
Indian	54	SUR	20-180µm	Trichodesmium	159X	3,813	~477X	T. thiebautii
Atlantic	141	SUR	20-180µm	Trichodesmium	154X	909	~114X	T. thiebautii
Indian	53	SUR	20-180µm	Trichodesmium	121X	not sequenced		T. thiebautii
Indian	58	SUR	20-180µm	Trichodesmium	110X	2,636	~330X	T. thiebautii
Atlantic	80	SUR	0.8-2000µm	UCYN-A	195X	7,426	~928X	UCYN-A1
Atlantic	80	SUR	0.8-5µm	UCYN-A	174X	65,051	~8,131X	UCYN-A1
Atlantic	78	SUR	0.8-5µm	UCYN-A	147X	231,636	~28,955X	UCYN-A1
Pacific	98	SUR	5-20µm	HBDs	1,106X	907	~113X	Gammaproteobacteria
Pacific	98	SUR	180-2000µm	HBDs	283X	34	~4X	Gammaproteobacteria
Atlantic	148	SUR	0.8-2000µm	HBDs	117X	1	~0.1X	Desulfobacterota
Pacific	138	DCM	0.8-2000µm	HBDs	101X	not sequenced		Gammaproteobacteria

Figure 5.3: Oceanic stations with highest metagenomic signal for diazotrophs. The world map provides coordinates for 15 *Tara* Oceans metagenomes (10 stations) displaying cumulative genomic coverage >100X for MAGs affiliated to diazotrophic *Trichodesmium*, UCYN-A or the HBDs. The bottom panel summarizes multi-omic signal (including at the level of *nifH* genes) statistics for those 15 metagenomes.

Part two: Gene-centric multi-omic analyses (*nifH* gene)

48 diazotrophic MAGs may cover >90% of cells containing known *nifH* genes. In order to analyze the significance of 48 diazotrophic MAGs with regard to other marine diazotrophic populations, we combined their *nifH* gene sequences with a comprehensive set of *nifH* sequences obtained from cultures, metagenomic assemblies, clones and amplicon surveys (see Methods). We used this extended *nifH* database (n = 328; redundancy removal at 98% identity over 90% of the length) to recruit metagenomic reads from *Tara* Oceans (Supplementary Table 5.8). Strikingly, *nifH* genes corresponding to the eight cyanobacterial diazotrophs and

40 HBDs recruited 42.3% and 49.1% of all mapped metagenomic reads, respectively, with just 8.7% of the signal corresponding to 280 orphan *nifH* genes for which the genomic content within plankton has not yet been characterized (Figure 5.4 and Supplementary Table 5.8). These include a well-known diazotroph that awaits genomic characterization, the Gamma-A lineage (Cornejo-Castillo and Zehr, 2021), which accounted for 0.4% of mapped reads. Overall, this *nifH* centric metagenomic survey indicates that the 48 bacterial diazotrophic MAGs characterized in this study encapsulate 90% of read recruitment signal for known *nifH* genes in the surface oceans and seas investigated during *Tara* Oceans. One remaining uncertainty is the extent of abundant marine heterotrophic bacterial *nifH* genes that have yet to be discovered. These might further swell the ranks of HBDs in years to come.

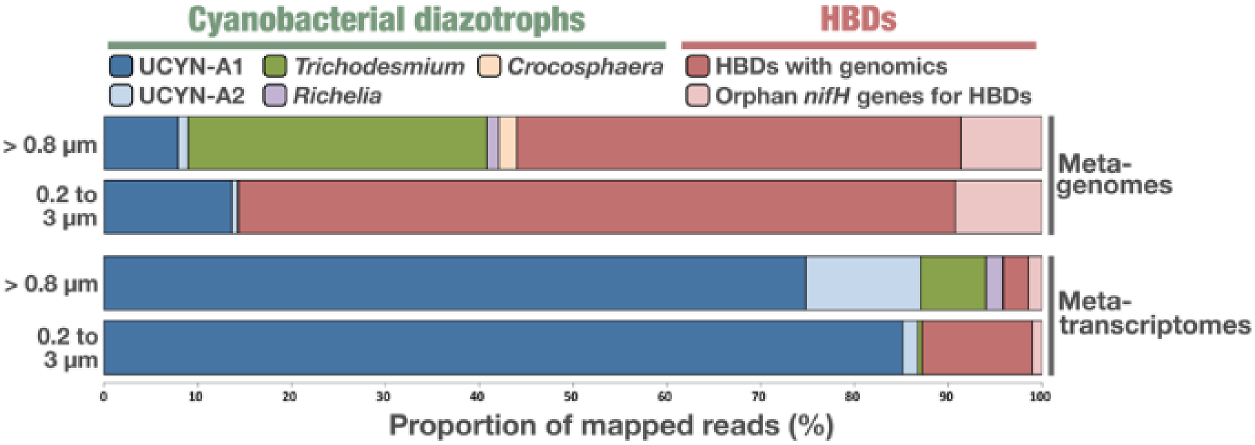


Figure 5.4: Detection of *nifH* genes across marine metagenomes and metatranscriptomes. The figure displays the proportion of metagenomic and metatranscriptomic reads mapping onto *nifH* genes as a function of ranges in two size fractions. Target genes correspond to the extended *nifH* gene database of 328 sequences including 280 orphan genes. The mapped samples (781 metagenomes and 520 metatranscriptomes) correspond to the surface and deep chlorophyll maximum layers of all oceans and two seas. For each size fraction range, the number of cumulated mapped reads represents each diazotrophic lineage (seven categories) across all samples. Results are displayed in relative proportion. The >0.8 µm size fraction range includes up to five size fractions: 0.8–5 µm, 5–20 µm, 20–180 µm, 180–2000 µm, and 0.8–2000 µm.

HBD populations express their *nifH* genes. We mapped hundreds of *Tara* Oceans metatranscriptomes against the extended *nifH* database to gain some insights into the potential for nitrogen fixation activity of cyanobacterial diazotrophs and HBDs. Specifically, we recruited "bacteria-compatible" metatranscriptomic reads from the free-living bacterial size fraction (0.2–3 µm), as well as poly-A enriched metatranscriptomic reads from larger size fractions ranging from 0.8 µm to 2,000 µm that was produced primarily to explore the transcriptomic diversity of microbial eukaryotes (Carradec et al., 2018). Bacterial transcripts are rarely polyadenylated, and even when it occurs, polyadenylation is often a degradation signal (Güell et al., 2011). Importantly, all of the HBD *nifH* genes recruited reads, indicating at the very least a basal expression of genes encoding the nitrogen fixation apparatus (Supplementary Table 5.8). Furthermore, the considerable genomic signal for HBDs at station 98 was reflected in the metatranscriptomic signal, demonstrating the expression of *nifH* genes by HBDs in these waters.

Given the methodological differences in RNA sequencing and other factors that may influence the observed signal (e.g., RNA stability across the bacterial tree of life, time intervals from sampling to RNA storage across stations and size fractions), we present global trends for the free-living bacterial size fraction (0.2–3 µm) and the larger size fractions as a combined pool (Figure 5.4). When considering the extended *nifH* database as a whole, most of the signal among metatranscriptomes corresponded to UCYN-A1, followed by UCYN-A2, HBDs, and *Trichodesmium* (Supplementary Table 5.8). The predominance of UCYN-A signal (including in the 0.2–3 µm size fraction) was driven by the high nitrogen fixation activity for UCYN-A1 at Stations 78 and 80 in the South West region of the Atlantic Ocean in which hundreds of thousands of metatranscriptomic reads corresponded to its *nifH* gene alone (Figure 5.3, Supplementary Table 5.8), as reported previously (Cornejo-Castillo et al., 2016). Metatranscriptomic read recruitments suggest that the UCYN-A1 symbiont drives a substantial portion of nitrogen fixation at the critical interface between oceans and atmosphere, which is quantitatively not reflected in the metagenomic signal (this genome was detected in just 13 stations). This metatranscriptomic analysis at large scale substantiates the importance of UCYN-A as previously observed with in situ nitrogen fixation surveys (e.g., (Martínez-Pérez et al., 2016)). A trend emerged in which the *nifH* genes for symbiotic diazotrophs (UCYN-A, *Richelia*) were more significantly detected relative to their metagenomic signal compared to non-symbiotic diazotrophs, corroborating previous studies (e.g., (Needoba et al., 2007; Foster et al., 2009)). These symbiotic relationships appear highly successful, and likely have an improved nitrogen-fixing capacity in contrast to free-living cells (Tripp et al., 2010; Cornejo-Castillo et al., 2016; Thompson et al., 2012). At the same time, the high abundance of *nifH* transcripts related to diazotrophic symbionts may partially reflect a protective effect of the host cell resulting in a sampling bias. Given that bacterial RNA molecules are highly unstable, marine metatranscriptomes should be interpreted with caution. Nevertheless, the relatively low signal for *Trichodesmium* and HBDs was surprising but might partially be related to the exclusion of bacterial transcripts from the larger size fractions.

For now, the nitrogen fixation activity of HBDs versus cyanobacterial diazotrophs remains unclear. HBDs may contribute very little to nitrogen fixation rates among plankton, in particular as compared to UCYN-A, *Richelia*, and *Trichodesmium* populations. For instance, the streamlined genomes of UCYN-A populations and beneficial interactions with their hosts have created highly effective nitrogen fixation machineries (Tripp et al., 2010; Cornejo-Castillo et al., 2016; Thompson et al., 2012) compared to what HBDs can do by themselves and without ATP production from photosynthesis. Yet metatranscriptomic surveys cannot be trusted to the same extent as metagenomes for semi-quantitative investigations, and do not equate to activity. Our only certitude at this point is that HBDs (1) are widespread and sufficiently abundant to make a real difference in the oceanic nitrogen balance, and (2) regularly transcribe their *nifH* gene in the sunlit ocean, including when co-occurring in large size fractions. These environmental genomic insights indicate that HBDs should not be excluded from the restricted list of most relevant marine nitrogen fixers (currently only represented by cyanobacterial lineages (Zehr and Capone, 2020)), at least until extensive studies of putative aggregates in the field as well as culture conditions shed light on their functional lifestyle and metabolic activities.

A simple nomenclature to keep track of genome-resolved marine HBDs. As an effort to maintain some continuity between studies, here we suggest applying a simple nomenclature to name with a numerical system the non-redundant HBD MAGs with sufficient completion statistics as a function of their phylum-level affiliation (historic NCBI naming). For example, HBDs affiliated to Alphaproteobacteria and discovered thus far were named HBD Alpha 01 to HBD Alpha 08. Supplementary Table 5.3 describes the 40 HBDs using this nomenclature, which could easily be expanded moving forward. To this point, only MAGs with completion >70% are part of this environmental genomic database, and the redundancy removal was set to ANI of 98%. Their genomic content can be accessed from `https://figshare.com/articles/dataset/Marine_diazotrophs/14248283`.

5.2.3 Conclusion

Our genome-resolved metagenomic survey of plankton in the surface of five oceans and two seas covering organismal sizes ranging from 0.2 µm to 2,000 µm has allowed us to go beyond cultivation and *nifH* amplicon surveys to characterize the genomic content and geographic distribution of key diazotrophs in the ocean. Briefly, we identified eight cyanobacterial diazotrophs, seven of which were already known at the species level, and 40 HBDs, 32 of which were first characterized in this study. The 40 HBDs are functionally diverse and expand the known diversity of abundant marine nitrogen fixers within Proteobacteria and Planctomycetes while also covering Verrucomicrobia. Overall, the collection of 48 diazotrophs we characterized here encapsulates 90% of metagenomic signal for known *nifH* genes in the sunlit ocean. In other words, the genomic search for the most abundant diazotrophs at the surface of the open ocean may be nearing completion.

Nitrogen fixers in the sunlit ocean have long been categorized into two main taxonomic groups: few cyanobacterial diazotrophs contributing most of the fixed nitrogen input (Carpenter and Romans, 1991; Gómez et al., 2005; Martínez-Pérez et al., 2016; Zehr et al., 2001), and a wide range of non-cyanobacterial diazotrophs considered to have little impact on the marine nitrogen balance, in part due to their very low abundances within plankton as seen from several *nifH* based amplicon surveys (Church et al., 2005, 2008; Zehr et al., 2007; Fong et al., 2008; Moisander et al., 2008; Benavides et al., 2016; Langlois et al., 2005; Man-Aharonovich et al., 2007). Here we provide three results contrasting with this paradigm. First, we found that a wide range of HBDs can occasionally co-occur under nitrate-depleted conditions in large size fractions, with metagenomic signals exceeding what was observed for UCYN-A and *Trichodesmium* lineages in other oceanic regions. Critically, insights from estuaries (Geisler et al., 2019; Bentzon-Tilia et al., 2015) may offer an explanation for the presence of HBDs in large size fractions of the open ocean, indicating their ability to form aggregates that provide low-oxygen microenvironments favorable for nitrogen fixation. These insights could explain, at least to some extent, high nitrogen fixation rates previously observed in parts of the Pacific Ocean that are depleted in cyanobacterial diazotrophs, which at the time was referred to as a paradox (Turk-Kubo et al., 2014). But most importantly, genome-wide metagenomic read recruitments for the 48 diazotrophs indicated that HBDs are more abundant than their cyanobacterial counterparts in most regions of the surface ocean. Metagenomes covering a wide size range of plankton (the 0.8–2000 µm size fraction) were critical to reach this conclusion. Mismatches between the widely used “nifH4” primer and the *nifH* genes of most HBDs might partially explain the growing gap between prior *nifH* based sequence surveys and genome-resolved metagenomics studies. Finally, we found that all HBDs express their *nifH* genes, including when co-occurring in large size fractions, expanding on previous observations based on a subset of the lineages in the 0.2–3 µm size fraction (Salazar et al., 2019).

As a result, a new understanding is emerging from large-scale multi-omic surveys that depict nitrogen fixers in the sunlit ocean as the sum of few cyanobacterial diazotrophs and a wide range of HBDs, all capable of using their nitrogen fixation machinery while thriving in specific size fractions and oceanic regions. Surveying HBD aggregates, including their nitrogen-fixing activity, might represent a new key asset in understanding the marine nitrogen cycle and its balance.

Now that genome-resolved metagenomics has shed light on dozens of abundant marine HBDs, first within the scope of free-living cells (Delmont et al., 2018), and now by covering a much wider plankton size range of plankton, it becomes apparent how little we know about their ecology and role in supporting oceanic primary productivity via nitrogen fixation. As a starting point, genomic analyses exposed three main functional groups of HBDs that might denote distinct diazotrophic lifestyles. Moving forward, it will be critical to enrich or cultivate these HBDs, as done for some of the key cyanobacterial diazotrophs decades ago (Ohki et al., 1992) or HBDs from the coast or estuaries more recently (Martínez-Pérez et al., 2018; Bentzon-Tilia et al., 2015). Experiments with HBDs in cell culture conditions and in situ investigations could shed light on HBD nitrogen fixation rates and elucidate the conditions that elicit nitrogen-fixing activity by these populations. These lines of research should strongly benefit our understanding of nitrogen budgets in the open ocean.

5.2.4 Material and Methods

***Tara* Oceans metagenomes.** We analyzed a total of 937 *Tara* Oceans metagenomes available at the EBI under project PRJEB402. Supplementary Table 5.1 reports general information (including the number of reads and environmental metadata) for each metagenome.

Genome-resolved metagenomics. The 798 metagenomes corresponding to size fractions ranging from 0.8 µm to 2 mm were previously organized into 11 'metagenomic sets' based upon their geographic coordinates (Delmont et al., 2022). Those 0.28 trillion reads were used as inputs for 11 metagenomic co-assemblies using MEGAHIT (Li et al., 2015) v1.1.1, and the scaffold header names were simplified in the resulting assembly outputs using anvi'o (Eren et al., 2015) v.6.1. Co-assemblies yielded 78 million scaffolds longer than 1,000 nucleotides for a total volume of 150.7 Gbp. Here, we performed a combination of automatic and manual binning on each co-assembly output, focusing only on the 11.9 million scaffolds longer than 2,500 nucleotides, which resulted in 1,925 manually curated bacterial and archaeal metagenome-assembled genomes (MAGs) with a completion >70%. Briefly, (1) anvi'o profiled the scaffolds using Prodigal (Hyatt et al., 2010) v2.6.3 with default parameters to identify an initial set of genes, and HMMER (Eddy, 2011) v3.1b2 to detect genes matching to bacterial and archaeal single-copy core gene markers, (2) we used a customized database including both NCBI's NT database and METdb to infer the taxonomy of genes with a Last Common Ancestor strategy (Carradec et al., 2018) (results were imported as described in `http://merenlab.org/2016/06/18/importing-taxonomy`), (3) we mapped short reads from the metagenomic set to the scaffolds using BWA v0.7.15 (Li and Durbin, 2009) (minimum identity of 95%) and stored the recruited reads as BAM files using samtools (Li et al., 2009), (4) anvi'o profiled each BAM file to estimate the coverage and detection statistics of each scaffold, and combined mapping profiles into a merged profile database for each metagenomic set. We then clustered scaffolds with the automatic binning algorithm CONCOCT (Alneberg et al., 2014) by constraining the number of clusters per metagenomic set to a number ranging from 50 to 400 depending on the set. Each CONCOCT clusters (n = 2,550, 12 million scaffolds) was manually binned using the anvi'o interactive interface. The interface considers the sequence composition, differential coverage, GC-content, and taxonomic signal of each scaffold. Finally, we individually refined each bacterial and archeal MAG with >70% completion as outlined in Delmont and Eren (Delmont and Eren, 2016), and renamed scaffolds they contained according to their MAG ID. Supplementary Table 5.2 reports the genomic features (including completion and redundancy values) of the bacterial and archaeal MAGs.

MAGs from the 0.2–3 µm size fraction. We incorporated into our database 673 bacterial and archaeal MAGs with completion >70% and characterized from the 0.2–3 µm size fraction (Delmont et al., 2018), providing a set of MAGs corresponding to bacterial and archaeal populations occurring in size fractions ranging from 0.2 µm to 2 mm.

Characterization of a non-redundant database of SMAGs. We determined the average nucleotide identity (ANI) of each pair of MAGs using the dnadiff tool from the MUMmer package (Delcher et al., 2002) v.4.0b2. MAGs were considered redundant when their ANI was >98% (minimum alignment of >25% of the smaller SMAG in each comparison). We then selected the MAG with the best statistics (highest value when computing completion minus redundancy) to represent a group of redundant MAGs. This analysis provided a non-redundant genomic database of 1,888 MAGs.

Taxonomical inference of MAGs. We determined the taxonomy of MAGs using both CheckM (Parks et al., 2015) and GTDB version 86 (Chaumeil et al., 2019). However, we used NCBI taxonomy from the GTDB output to describe the phylum of MAGs in the results and discussion sections, in order to be in line with the literature.

Biogeography of MAGs. We performed a final mapping of all metagenomes to calculate the mean coverage and detection of the MAGs. Briefly, we used BWA v0.7.15 (minimum identity of 90%) and a FASTA file containing the 1,888 non-redundant MAGs to recruit short reads from all 937 metagenomes. We considered MAGs were detected in a given filter when >25% of their length was covered by reads to minimize non-specific read recruitments (Delmont et al., 2018). The number of recruited reads below this cut-off was set to 0 before determining vertical coverage and percent of recruited reads.

Cosmopolitan score. Using metagenomes from the Station subset 1 (n = 757; excludes the 0.8–2000 µm size fraction lacking in the first leg of the *Tara* Oceans expeditions), MAGs were assigned a "cosmopolitan score" based on their detection across 119 stations, as previously quantified for eukaryotes (Delmont et al., 2022).

Identification of diazotroph MAGs. In a first step, we used three HMM models from Pfam (Bateman et al., 2002) within anvi'o (e-value cutoff of e^{-15}) and targeting the catalytic genes (*nifH*, *nifD*, *nifK*) and biosynthetic genes (nifE, nifN, nifB) for nitrogen fixation. We then ran Interproscan (Zdobnov and Apweiler, 2001) on genes with a HMM hit and used TIGRFAMs (Haft et al., 2003) results (we found those to be the most relevant for nitrogen fixation) to identify diazotroph MAGs. Finally, we used RAST (Aziz et al., 2008) as a complementary approach to identify nitrogen-fixing genes the HMM/Interproscan approach failed to characterize. Among the 48 diazotroph MAGs, only one single gene (*nifH*) was not recovered with this approach. The most likely explanation is that the gene is simply missing from the MAG.

Functional inferences of diazotroph MAGs. We inferred functions among the genes of diazotrophic MAGs using COG20 functions, categories, and pathways (Galperin et al., 2021), KOfam (Aramaki et al., 2020), KEGG modules, and classes (Kanehisa et al., 2017) within the anvi'o genomic workflow (Eren et al., 2015). Regarding the KOfam modules, we calculated their level of completeness in each genomic database using the anvi'o program "anvi-estimate-metabolism" with default parameters. The URL `https://anvio.org/m/anvi-estimate-metabolism` describes this program in more detail.

Sequence novelty for the *nifH* genes. The 47 *nifH* genes identified in the MAGs were considered novel if their sequence identity scores never exceeded 98% identity over an alignment of at least 200 nucleotides, when compared to a recently built *nifH* gene catalog by Pierella Karlusich et al. (Pierella Karlusich et al., 2021) using blast (Altschul et al., 1990). Briefly, the *nifH* gene catalog consists of sequences from Zehr laboratory (mostly diazotroph isolates and environmental clone libraries; `https://www.jzehrlab.com`), sequenced genomes, and additional sequences retrieved from *Tara* Oceans metagenomic assemblies co-assemblies (Delmont et al., 2018) and the OM-Reference Gene Catalog version 2 (Salazar et al., 2019)).

A new database of *nifH* genes including diazotroph MAGs. We created a database of *nifH* genes covering the diazotroph MAGs as well as a few hundred sequences from Pierella Karlusich et al. (Pierella Karlusich et al., 2021) with signal in *Tara* Oceans metagenomes. We removed redundancy (cut-off=98% identity) between the diazotroph MAGs and the Pierella Karlusich database, except for *Trichodesmium thiebautii* due to the occurrence of multiple populations (and slight differences between MAGs and culture representatives) that stressed the need to further explore *nifH* gene microdiversity within this species. We performed a mapping of metagenomes and metatranscriptomes to calculate the mapped reads and mean coverage of sequences in the extended *nifH* gene database. Briefly, we used BWA v0.7.15 (minimum identity of 90%) and a FASTA file containing the sequences to recruit short reads.

Phylogenetic analyses of diazotroph MAGs. We used PhyloSift (Darling et al., 2014) v1.0.1 with default parameters to infer associations between MAGs in a phylogenomic context. Briefly, PhyloSift (1) identifies a set of 37 marker gene families in each genome, (2) concatenates the alignment of each marker gene family across genomes, and (3) computes a phylogenomic tree from the concatenated alignment using FastTree (Price et al., 2010) v2.1. We used anvi'o to visualize the phylogenomic tree in the context of additional information and root it at the level of the phylum Cyanobacteria.

Metatranscriptomic read recruitment for *nifH* genes. We performed a mapping of 587 *Tara* Oceans metatranscriptomes to calculate the mean coverage of sequences in the extended *nifH* gene database. Briefly, we used BWA v0.7.15 (minimum identity of 90%) and a FASTA file containing the *nifH* gene sequences to recruit short reads from all 587 metatranscriptomes.

5.2.5 Data Availability

All data our study generated are publicly available at `http://www.genoscope.cns.fr/tara/` (metagenomic co-assemblies, FASTA files) or `https://figshare.com/articles/dataset/Marine_diazotrophs/14248283` for the supplemental tables and information, as well as the genomic content of 48 marine diazotrophs using the new nomenclature (diazotrophic genomic database).

5.2.6 Supplementary Figures

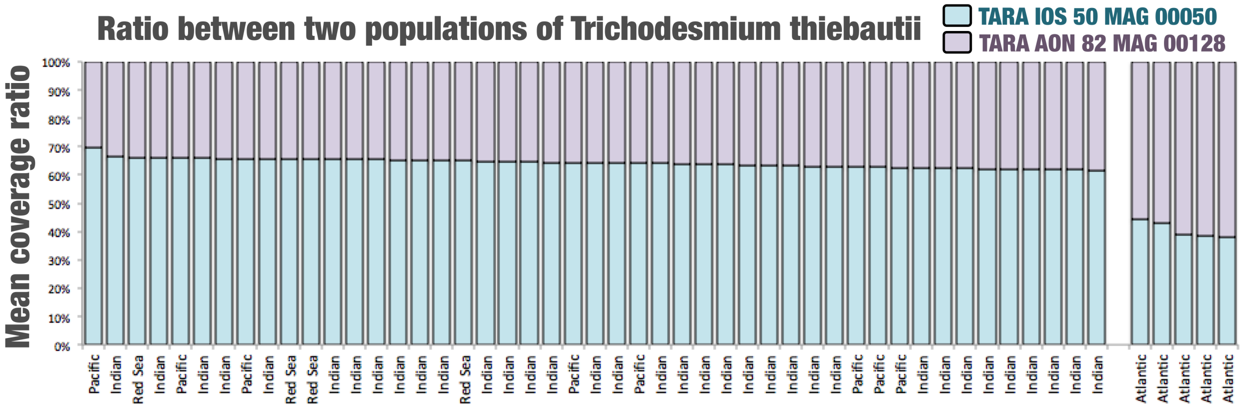


Figure 5.5: Mean coverage ratio for the two *Trichodesmium thiebautii* MAGs across 52 *Tara* Oceans metagenomes displaying a cumulative coverage >2X. Station Ids and associated data are available in the Supplementary Table 5.4.

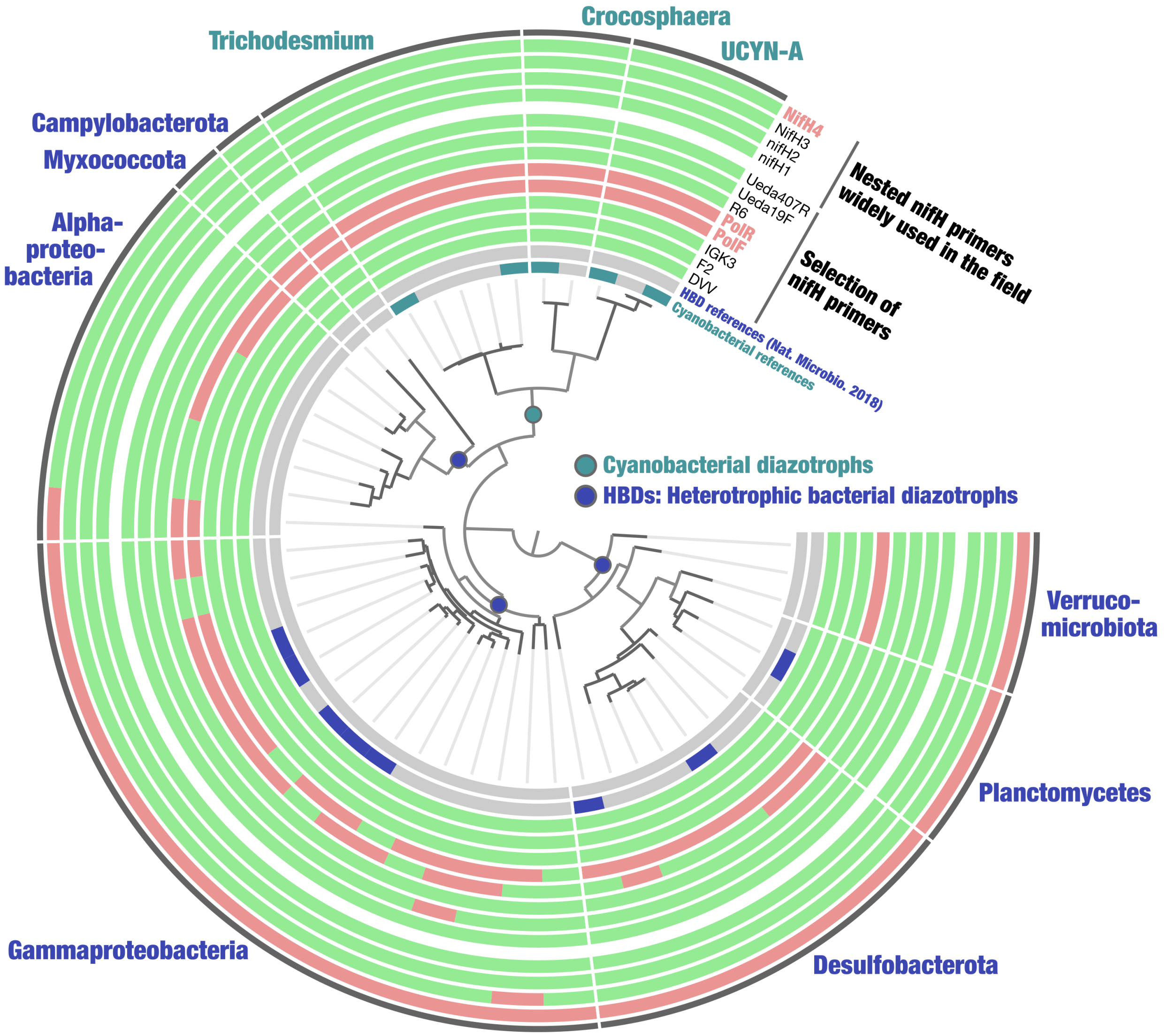


Figure 5.6: Interplay between the phylogeny and primer compatibility of *nifH* genes. The inner tree represents a phylogenetic tree of *nifH* genes from MAGs in our survey plus cyanobacterial references (built at the amino acid level with fastree (Price et al., 2010) within Genomenet (`https://www.genome.jp/tools-bin/ete`). Layers represent the compatibility (green) or incompatibility (red) of specific *nifH* primers used in the field (including for large-scale amplicon surveys).

5.2.7 Supplementary Tables

This section's supplementary tables are accessible via doi:10.1038/s41396-021-01135-1 or at `https://figshare.com/articles/dataset/Marine_diazotrophs/14248283`.

Table 5.1: Statistics for 937 *Tara* Oceans metagenomes organized by depth, size fraction and oceanic region. The table also contains environmental conditions across *Tara* Oceans stations.

Table 5.2: Statistics for the 1,888 bacterial and archaeal MAGs. The table contains genomic statistics (e.g., completion and length), taxonomic information, general mapping trends such as the cosmopolitan score, as well as additional information regarding the 48 diazotroph MAGs.

Table 5.3: Occurrence of genes that encode the catalytic (*nifHDK*) and biosynthetic (*nifENB*) across 48 diazotrophic MAGs.

Table 5.4: Genome-wide metagenomic read recruitment statistics for the 48 diazotroph MAGs. The table contains mean genomic coverage values across the 937 *Tara* Oceans metagenomes described in Table S01.

Table 5.5: The *nifH* genes of 48 diazotroph MAGs. The table contains blast results when comparing *nifH* gene sequences from the diazotrophic MAGs to a reference *nifH* gene catalog.

Table 5.6: Compatibility between *nifH* primers and diazotrophic MAGs.

Table 5.7: KEGG MODULE functional completeness of the 48 diazotrophic MAGs.

Table 5.8: Metagenomic and metatranscriptomic mapping for the extended *nifH* database. The table also includes *nifH* gene sequences.

5.3 Targeted binning of a novel nitrogen-fixing population from the Arctic Ocean: an anvi'o tutorial

This section provides an example tutorial written to support the public's usage of the metabolism reconstruction framework for targeted binning of metagenome-assembled genomes. It is derived from the following blog post: `https://anvio.org/blog/targeted-binning/`. Some edits and omissions have been made to improve its readability in this format.

In this blog post, I will demonstrate how to use ‘anvi-estimate-metabolism‘ to find and bin a novel, nitrogen-fixing population from a set of publicly-available Arctic Ocean metagenomes. Of course, nitrogen fixation is just an example here, and the same technique can be applied to survey metagenomic datasets for microbial populations with other characteristic metabolic capabilities. So if you are interested in learning about how to leverage anvi'o's metabolism estimation capabilities to go fishing through your data, or if you just can't get enough of cool marine nitrogen fixation stories, keep on reading!

The commands in this post were written for anvi'o v7.1.

5.3.1 Setting up our story

Exciting things are happening right now in the world of marine microbiology. Our friend and colleague Tom Delmont has recently published a cool story about heterotrophic bacterial diazotrophs, or HBDs (Delmont et al., 2021). For anyone who doesn't know, a diazotroph is a microbe that fixes nitrogen. Nitrogen fixation is a very important process that supports all forms of life on Earth by yanking nitrogen atoms from one of the most recalcitrant molecules on earth, N_2 gas, and putting them into biologically-usable molecules such as ammonia. It happens quite a lot in the global oceans, so to marine microbiologists, nitrogen-fixing microbes - that is, marine diazotrophs - are a Big Deal (Zehr and Capone, 2020).

Marine diazotrophs come in many forms, but not all of them have been getting equal amounts of attention. For instance, cyanobacterial diazotrophs have long been thought to be the most abundant type of nitrogen-fixing microbes in the ocean, and therefore have been the focus of most research related to the subject. Though the existence of non-cyanobacterial, heterotrophic diazotrophs has long been known, their presence was measured through amplicon surveys that targeted a single gene in the nitrogen fixation operon: *nifH*. These amplicon surveys were also the only way to discuss the relative abundance of heterotrophic diazotrophs in oceans compared to their cyanobacterial counterparts. But there were no actual microbial genomes, which prevented our ability to study their ecology without primer sequences and PCR amplification biases.

A relatively recent study, which took years in the making and shaped large parts of anvi'o (Murat Eren , Meren), changed this by giving access to the first-ever genomes of heterotrophic diazotrophs (Delmont et al., 2018). It showed that they were much more abundant than what we could survey with *nifH* primers and came from taxa that were not previously considered in the context of nitrogen fixation (such as Planctomycetes). The saga of discovering new heterotrophic diazotrophs continues with this new paper by Tom Delmont et al. (Delmont et al., 2021), in which Tom manually curates almost 2,000 metagenome-assembled genomes (MAGs) from co-assemblies of almost 800 ocean metagenomes, and then mines this MAG set for nitrogen fixation genes to find an additional 48 diazotrophs.

The benefits of characterizing all genomes found in ocean metagenomes are obvious. Yet, it is not as obvious how one would survey these metagenomes for targeted genome-resolved insights – for instance, to recover particular genomes that encode metabolic modules of interest, such as nitrogen fixation.

One might suggest that we could automatically bin all genomes from metagenomes, and then survey the *nifH* genes in them to find diazotrophs (as *nifH* is the standard marker for nitrogen fixation). But it may not be that simple. Not only there are many things that can go wrong with automatic binning (Eren and Scott, 2020a), but perhaps more critically, the presence of a key function from a metabolic module does not necessarily indicate the presence of the metabolic capacity. For instance, the *nifH* gene can occur in a variety of different contexts, so it is actually necessary to find 6 different *nif* genes to be sure of a microbe's nitrogen-fixing capabilities (Dos Santos et al., 2012). As a clear example, when I survey a metagenome from the Southern Ocean, an environment where no microbes that fix nitrogen have ever been found, I find 12 COG annotations of *nifH* genes. In the case of *nifH*, one can look for other genes nearby in the same contig, such as *nifK*, *nifD*, etc. But what if you are targeting a more complex metabolic capability for which the required genes are more diverse?

This is precisely what ‘anvi-estimate-metabolism‘ enables you to do. Without having to implement a lot of *ad hoc* steps, you can identify contigs in your metagenomes – prior to binning – that may belong to a specific population of interest that encodes a desired metabolic capacity. ‘anvi-estimate-metabolism‘ is a program that predicts the metabolic capabilities of microbes from genomic or metagenomic data. It combines functional annotations from the KOfam database (Aramaki et al., 2020) with KEGG definitions of metabolic pathways (Kanehisa, 2017; Kanehisa et al., 2023) to estimate completeness of these pathways and produce easily-parsable output. If you are interested in finding a microbe that has a particular metabolic capability, it is much easier to find it using this output rather than parsing through annotations for individual genes.

And that is exactly how we are going to do it today. As I mentioned at the beginning of this post, we'll be using ‘anvi-estimate-metabolism‘ to look for a previously-uncharacterized marine diazotroph in a set of Arctic Ocean metagenomes.

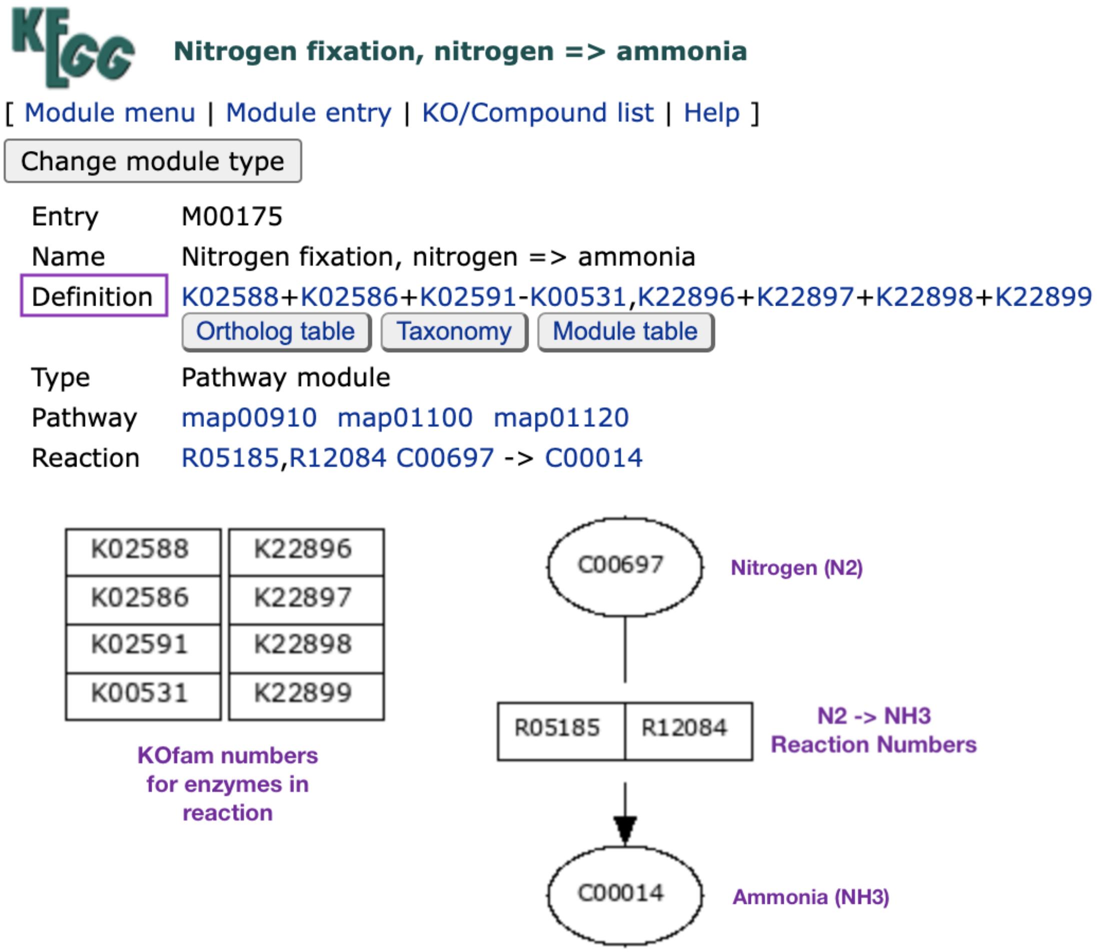
5.3.2 A bit of background

Before we get started, let's talk a bit more about nitrogen fixation. I am by no means an expert in this (or in any particular metabolism, really), so here is a very light summary of the background. Experts, feel free to roll your eyes at me and skip to the good stuff.

Nitrogen fixation is the process of converting gaseous nitrogen (N_2) into the more biologically-usable form ammonia (NH_3). The resulting ammonia can then be further converted into other bioavailable compounds like nitrate and nitrite. Since nitrogen is an essential component of many biological molecules, nitrogen fixation is a fairly important process that supports life, in general. Only some ocean microbes, called marine diazotrophs, have the genes that allow them to fix nitrogen, which are (usually) encoded in the *nif* operon (Sohm et al., 2011).

The nitrogen fixation pathway – KEGG vs reality

The KEGG module for nitrogen fixation is M00175 (Figure 5.7). The part of the module that you'll want to focus on here is the "Definition" line containing the KO numbers for each enzyme required in the pathway (these KOs are also arranged in the rectangular boxes in the bottom-left). You need a nitrogenase enzyme complex to convert nitrogen gas to ammonia, and there are currently only two major versions of this complex: the "molybdenum-dependent nitrogenase" protein complex encoded by genes *nifH* (K02588), *nifD* (K02586), and *nifK* (K02591) of the *nif* operon; or the "vanadium-dependent nitrogenase" protein complex encoded by genes *vnfD* (K22896), *vnfK* (K22897), *vnfG* (K22898), and *vnfH* (K22899), of the *vnf* operon. The latter complex has been isolated from soil bacteria and is known to be an alternative nitrogenase that is expressed when molybdenum is not available (Lee et al., 2009a; Bishop et al., 1980). We're going to ignore it, because I have yet to see it in any ocean samples.



Nitrogen fixation, nitrogen => ammonia

[Module menu | Module entry | KO/Compound list | Help]
Change module type
Ortholog table
Taxonomy
Module table

Entry	M00175
Name	Nitrogen fixation, nitrogen => ammonia
Definition	K02588+K02586+K02591-K00531,K22896+K22897+K22898+K22899
Type	Pathway module
Pathway	map00910 map01100 map01120
Reaction	R05185,R12084 C00697 -> C00014

K02588	K22896
K02586	K22897
K02591	K22898
K00531	K22899

KOfam numbers for enzymes in reaction
C00697
Nitrogen (N2)
R05185
R12084
N2 -> NH3 Reaction Numbers
C00014
Ammonia (NH3)

Figure 5.7: KEGG module M00175 for nitrogen fixation. Image taken from KEGG website and modified with labels (purple).

You might have noticed that I left K00531 out of the above discussion. That is because this KO is not part of the *nif* operon - rather, it is the gene anfG, which is part of the alternative nitrogen fixation operon *anf*. *anf* encodes an alternate nitrogenase enzyme made up of the components anfHDKG, but *anfHDK* are very similar to the *nifHDK* components (Joerger et al., 1989). My best guess as to why *anfHDK* don't have their own KOfam profiles is that the *nifHDK* KOfams can match to these genes. But since *anfG* is an additional component that is not required for the *nif* operon, it has its own KO and is labeled as non-essential to the enzyme complex in this module (that is what the minus sign in front of K00531 in the module definition means). This is a very long-winded way of saying that we don't have to worry about looking for K00531 in our data.

So that means we effectively care about only *nifHDK* in this module. But wait. While *nifHDK* represent the catalytic components of the nitrogenase enzyme, it turns out that there are a few other genes required to produce an essential FeMo-cofactor and incorporate it into this complex. At minimum, the extra genes required are *nifE* (K02587), *nifN* (K02592), and *nifB* (K02585) (Dos Santos et al., 2012). Figure 5.8 shows a diagram of the *nif* operon in the *Azotobacter vinelandii* genome sequence, courtesy of (Dos Santos et al., 2012).

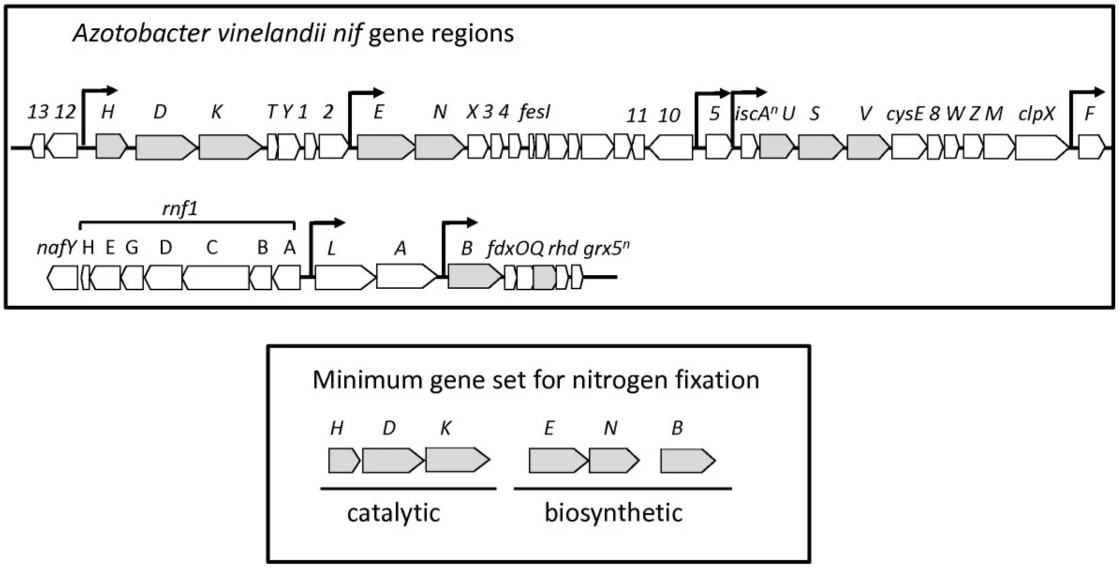

Azotobacter vinelandii nif gene regions
13 12 H D K T Y 1 2 E N X 3 4 fesI 11 10 5 iscA^n U S V cysE 8 W Z M clpX F
rnf1
nafY H E G D C B A L A B fdxOQ rhd grx5^n
Minimum gene set for nitrogen fixation
H D K E N B
catalytic
biosynthetic

Figure 5.8: Diagram of the *nif* operon in the *Azotobacter vinelandii* genome sequence. Figure credited to (Dos Santos et al., 2012).

The catalytic genes - those from module M00175 - are located next to each other on the bacterial chromosome. The other required biosynthetic genes are located farther along, with *nifE* and *nifN* expressed under the same promoter and *nifB* isolated from the rest of the genes and expressed under its own promoter.

What this means is that we need to find six genes - preferably located on the same contig - within a metagenome assembly in order to be confident that the metagenome includes a nitrogen-fixing population. We expect to see the same structure as in the diagram above reflected in our metagenome assemblies, meaning that gene groups *nifHDK* and *nifEN* are the most likely to end up on the same contig. If you keep reading, you will see that this is indeed the case!

Nitrogen fixation in the polar oceans, and an awesome polar ocean dataset

One aspect of this analysis that makes it a bit more interesting is that we will be working with polar ocean metagenomes. The polar oceans – which are the Arctic Ocean and the Southern Ocean around Antarctica – are not typically associated with nitrogen fixation. The vast majority of nitrogen-fixing microbes have been found in non-polar oceans, perhaps because the diazotrophic cyanobacteria that are typically studied tend to be found in warmer waters (Stal, 2009), or perhaps because the polar oceans are just not as well-studied as the other oceans. Regardless, in recent times, there have been several reports of nitrogen fixation happening in the Arctic and Antarctic Oceans (Harding et al., 2018; Shiozaki et al., 2020; von Friesen and Riemann, 2020). So it is certainly not an unusual or fruitless choice to be searching for nitrogen-fixing microbes in the colder waters of our planet.

We will be using a recently published dataset of Arctic and Antarctic ocean metagenomes by (Cao et al., 2020). These brave scientists faced the cold to bring the marine science community 60 new samples from 28 different locations in the polar oceans. Their comparative analyses demonstrated that polar ocean microbial communities are distinct from non-polar ones, both in terms of their taxonomic diversity and their gene content. They also reconstructed 214 metagenome-assembled genomes, 32 of which were enriched in the polar oceans according to read recruitment analyses. In their paper, the authors analyzed some metabolic pathways in these MAGs, but notably did not check for nitrogen fixation – maybe because they didn't expect to find it. But as you will see, it is there to be found.

So without further ado, let's go through this analysis together.

5.3.3 Estimating metabolism in Arctic Ocean metagenomes

```
# Download tutorial datapack (the unzipped datapack is 3.4 GB in size)
wget -O NIF_MAG_DATAPACK.tar.gz \
    https://figshare.com/ndownloader/files/31119277

# unzip and cd into working directory
tar -xvf NIF_MAG_DATAPACK.tar.gz && cd NIF_MAG_DATAPACK
```

Listing 5.1: Datapack Download

To start, we need metagenome assemblies of the Arctic Ocean samples from Cao et al.'s dataset. I am fortunate to be colleagues with Matt Schechter, an awesome microbiologist who knows way more about oceans than I do, and who also happened to be interested in this dataset. He downloaded the samples and made single assemblies of them using the software IDBA-UD (Peng et al., 2012) as part of the anvi'o metagenomic workflow (Shaiber and Eren, 2018). We are all benefiting from his hard work today - thanks, Matt!

We won't look at all 60 samples from the Cao et al. paper, only 16 of their surface Arctic Ocean samples (taken from a depth of 0 m). When we downloaded these samples, we assigned different (shorter) names to them, so the sample names I will discuss below are different from the ones in the Cao et al. paper. If you want to know the correspondence between our sample names and those in the paper, check out the 'cao_sample_metadata.txt' file in the datapack. You will find their sample names in the 'sample_name_cao_et_al' column.

The first thing that I did with those 16 assemblies was run ‘anvi-estimate-metabolism‘ in metagenome mode. I will show you the commands that I used to do this, but I won’t ask you to do it yourself, because it takes quite a long time (and currently requires an obscene amount of memory, for which I deeply apologize). I created a metagenomes file called ‘metagenomes.txt‘ which contains the names and contigs database paths of each sample, and I wrote a bash loop to estimate metabolism individually on each sample:

```
while read name path; \
do \
    anvi-estimate-metabolism -c $path \
    --metagenome-mode \
    -O $name \
    --kegg-output-modes modules,kofam_hits; \
done < <(tail -n+2 metagenomes.txt)
```

Listing 5.2: Metabolism estimation loop

What this loop does is read each line of the ‘metagenomes.txt‘ file, except for the first one (the ‘tail -n+2‘ command skips the first line). Each non-header line in the file contains the name of the metagenome sample (which gets placed into the ‘$name‘ variable) and the path to its contigs database (which gets placed into the ‘$path‘ variable). Therefore, ‘anvi-estimate-metabolism‘ gets run on each contigs database in metagenome mode, and the resulting output files (two per sample) are prefixed with the sample name.

It is possible to run ‘anvi-estimate-metabolism‘ on more than one contigs database at a time, using multi-mode, which you can read about on the ‘anvi-estimate-metabolism‘ help page. However, I did not do this here because I wanted the output for each sample to be printed to a separate output file, for purely organizational purposes.

You will find the resulting output files in the datapack, which you should have downloaded at the beginning of this post. Notice that there are 32 text files, two for each metagenome assembly, in the ‘METABOLISM_ESTIMATION_TXT‘ folder. Let’s take a look at the first few lines of the modules file for sample N02:

```
cd METABOLISM_ESTIMATION_TXT/

head -n 4 N02_modules.txt
```

Listing 5.3: Examining the metabolism output for sample N02

You should see something like this:

unique_id	**contig_name**	**kegg_module**	**...**	**module_completeness**	**...**
0	c_000000008738	M00546	...	0.5	...
1	c_000000000052	M00001	...	0.4	...
2	c_000000000052	M00002	...	0.5	...
...	...	...	...	...	...

Table 5.9: Modules mode output for sample N02.

This is a modules mode output file from ‘anvi-estimate-metabolism‘ (which is the default output type). Since we ran the program in metagenome mode, each row of the file describes the completeness of a metabolic module within one contig of the metagenome. What this means is that every KOfam hit belonging to this pathway (listed in the ‘kofam_hits_in_module‘ column) was present on the same contig in the metagenome assembly. This is important, because metagenomes contain the DNA sequences of multiple organisms, so the only time that we can be sure two genes go together within the same population genome is when they are assembled together onto the same contig sequence.

If right now you are thinking, “But wait... if we only focus on the genes within the same contig, many metabolic pathways will have completeness scores that are too low,” then you are exactly correct. It is likely that most metabolic pathways from the same genome will be split across multiple contigs, and their components will therefore end up in different lines of this file. In the example above, contig ‘c_000000008738‘ contains 50% of the KOs required for the purine degradation module M00546, but perhaps the other KOs in the pathway (such as K01477) also belong to whatever microbial population this is, just on a different contig. Putting many contigs together to match up the different parts of the pathway, while making sure that you are not producing a chimeric population, is a task that requires careful binning.

Luckily for us, the nitrogen fixation module from KEGG (as described earlier in the post) has a couple of helpful characteristics. First, it contains only the 3 catalytic genes, and second, it is encoded in an operon, so those genes are located close together in any given genome sequence. These two things make it much more likely that the entire module will end up within a single contig in our metagenome assemblies, which means it will be relatively easier to find a complete nitrogen fixation module in our metabolism estimation output files.

5.3.4 Looking for evidence of a nitrogen-fixing population

Though M00175 only contains the catalytic portion of our required *nif* gene set, it is a good starting point for our search. If we look for this module in our metabolism estimation results, we can find out which contig(s) it is located on and use that to guide our search for the remaining genes.

Using modules mode output to find M00175

You can use the following BASH code to search for lines describing M00175 in all metabolism estimation ‘modules mode’ outputs. The code filters the output so that it contains only those lines which have a score of 1.0 in the ‘module_completeness‘ column, meaning that all 3 *nifHDK* genes are located on the same contig in the assembly. It further filters the output to contain only the columns describing 1) the file name and line number in the file where M00175 was found, 2) the contig name, 9) the completeness score, 11) the list of KO hits that we found from this module, and 12) the corresponding gene caller IDs of these hits.

```
grep M00175 *_modules.txt | \
    awk -F’\t’ ’$9 == 1.0’ | \
    cut -f 1,2,9,11,12 | \
column -t
```

Listing 5.4: Filtering the metabolism estimation output

Your output should look like this:

```
N06_modules.txt:7398    c_000000000415   1.0   K02586,K02588,K02591   35121,35120,35122
N07_modules.txt:7413    c_000000004049   1.0   K02586,K02591,K02588   94224,94225,94223
N07_modules.txt:31467   c_000000000073   1.0   K02586,K02591,K02588   14638,14637,14639
N22_modules.txt:44057   c_000000000122   1.0   K02586,K02591,K02588   16856,16857,16855
N25_modules.txt:11798   c_000000000104   1.0   K02586,K02591,K02588   13919,13920,13918
```

Table 5.10: Complete instances of M00175 in modules output.

These are promising results! The complete M00175 module was found in 4 different Arctic Ocean samples (there are two instances in sample N07).

I encourage you to look through the other, less complete instances of this module in the output files. If you do this, you will see that some metagenomes appear to have all 3 of these genes split across multiple contigs (could they be contigs from the same genome?). For instance, here is a pair of contigs from sample N22:

```
N22-contigs_modules.txt:35879   c_000000000861   0.333   K02588          43430
N22-contigs_modules.txt:49457   c_000000003717   0.666   K02591,K02586   84130,84129
```

Table 5.11: Partial instance of M00175 in modules output.

nifH is on one contig and *nifDK* are on the other. I think it is likely that these two contigs go together, because it seems unlikely that a genome would have one of these genes from this operon and not the rest (though it could happen, of course. Microbial genomes are incredibly plastic, and operons are not immune to genome reorganization).

All in all, as you examine the estimation results for these 16 metagenomes, you should find that 9 of them have at least a partial copy of M00175, and 5 of those contain at least one complete set of *nifHDK* (though not necessarily all on the same contig).

Of course, as we discussed earlier, there are 3 other genes that we need to find alongside *nifHDK* in order to be sure that we have a microbial population capable of fixing nitrogen. KEGG may not have put these genes in M00175, but it does have a KOfam profile for each one of *nifENB* - those KOs are K02587, K02592, and K02585. To search for these, we turn to our ‘KOfam hits’ mode output files.

Using ‘KOfam hits’ mode output to find the other *nif* genes

We will focus on the 5 samples that contain *nifHDK*, which are N06, N07, N22, N25, and N38. Let’s look at their ‘kofam_hits‘ output files one at a time, starting with sample N06.

```
# print the header line, then run a search loop
head -n 1 N06_kofam_hits.txt; \
for k in K02587 K02592 K02585; \
do \
  grep $k N06_kofam_hits.txt; \
done
```

Listing 5.5: Loop to search for *nifENB* in sample N06

The loop above searches for each KO of *nifENB* in this file. When you run it, you should see output that looks like this:

unique_id	**contig_name**	**ko**	**gene_caller_id**	**modules_with_ko**	**ko_definition**
70353	c_000000000415	K02587	35136	None	NifE
70352	c_000000000415	K02592	35137	None	NifN
82427	c_000000001170	K02585	58423	None	NifB

Table 5.12: Search results for *nifENB* in sample N06.

In sample N06, we previously found a complete M00175 module on ‘c_000000000415‘. From the ‘kofam_hits‘ output, we can see that *nifE* and *nifN* are on the same contig, while *nifB* is on a different one (contig ‘c_000000001170‘). This arrangement makes sense based on the *A. vinelandii* genome we looked at earlier, in which *nifB* was the farthest gene from the start of the *nifHDK* operon. Since all six of the required *nif* genes are present, it seems likely that this metagenome contains a legitimate nitrogen-fixing population! These contigs would be an excellent starting point for binning.

If we use the same code to search in file ‘N07_kofam_hits.txt‘, we get:

unique_id	**contig_name**	**ko**	**gene_caller_id**	**modules_with_ko**	**ko_definition**
3729	c_000000000256	K02587	29649	None	NifE
8116	c_000000000073	K02587	14636	None	NifE
3727	c_000000000256	K02592	29650	None	NifN
8110	c_000000000073	K02592	14635	None	NifN
8118	c_000000000073	K02585	14642	None	NifB
122901	c_000000000095	K02585	17048	None	NifB

Table 5.13: Search results for *nifENB* in sample N07.

Recall from earlier that in sample N07, one complete M00175 module was on contig ‘c_000000000073‘, and another was on contig ‘c_000000004049‘. The ‘kofam_hits‘ file shows that there is one copy each of *nifENB* on ‘contig c_000000000073‘, which means that we have found all six *nif* genes on the same contig! This is excellent. There is a nitrogen-fixing population here for sure (and there may even be two different ones, considering that contig ‘c_000000004049‘ also contains a complete M00175 and there is a second set of the *nifENB* genes spread across two different contigs).

What does sample N22 have in store for us? Earlier, we found a complete M00175 on contig ‘c_000000000122‘ in this sample.

unique_id	**contig_name**	**ko**	**gene_caller_id**	**modules_with_ko**	**ko_definition**
83218	c_000000000122	K02587	16870	None	NifE
120563	c_000000003718	K02587	84133	None	NifE
83216	c_000000000122	K02592	16871	None	NifN
120562	c_000000003718	K02592	84134	None	NifN
2217	c_000000000860	K02585	43377	None	NifB
90602	c_000000000014	K02585	5285	None	NifB

Table 5.14: Search results for *nifENB* in sample N22.

Since there is a K02587 and a K02592 on contig ‘c_000000000122‘, 5 out of 6 *nif* genes appear on the same contig in this metagenome. N22 also appears to have a second set of these genes spread across multiple contigs, just as in N07.

You can take a look at N25 and N38 yourself. N25 should have at least one copy of all six genes (and 5/6 on contig ‘c_000000000104‘), but N38 should be missing *nifN*.

```
# we are done here
cd ..
```

Listing 5.6: Go back to parent directory

At this point, we can be fairly confident that there are nitrogen-fixing populations in samples N06, N07, N22, and N25. The natural question to ask next is - what are they (and are they worth binning)?

5.3.5 Determining population identity

Since we are working with individual contigs and not full genomes right now, a good strategy to figure out what these populations could be is to use BLAST (Altschul et al., 1990) to see if there is anything similar to these contigs in the NCBI database.

I extracted the relevant contig sequences from these 4 metagenome assemblies for you. You will find them in the ‘FASTA/contigs_of_interest.fa‘ file in the datapack. Each contig name is prefixed by the name of the sample it came from, as in ‘N06_c_000000000415‘.

```
# go to folder with sequences:
cd FASTA/

# see what contigs are in this file
grep '>' contigs_of_interest.fa
```

Listing 5.7: Viewing the contigs of interest

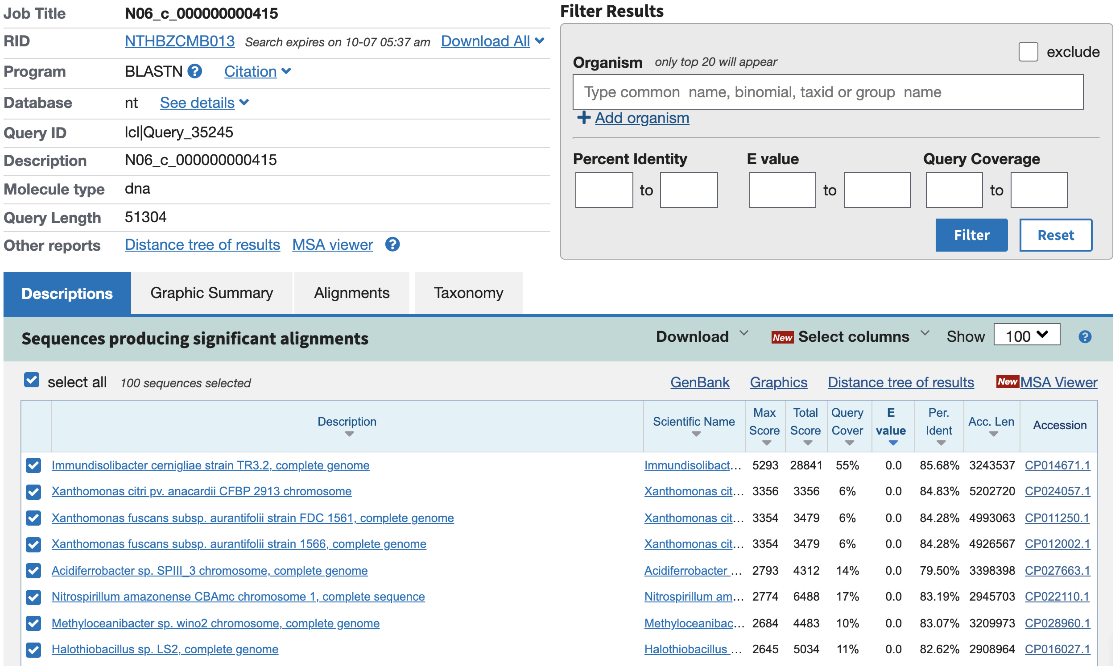
You do not have to BLAST every sequence that is in that file (unless you want to). I recommend at least looking at the contig that contains the most *nif* genes in each metagenome, namely: ‘N06_c_000000000415‘, ‘N07_c_000000000073‘, ‘N22_c_000000000122‘, and ‘N25_c_000000000104‘.

Go ahead and BLAST those contigs.

Did you do it? Great. Your results will of course depend on what is currently in the NCBI database at the time you are BLASTing (or the version of that database that you have on your computer, if you are running it locally instead of on their web service), but I will show you what I got at the time I was writing this post. I used the 'blastn' suite with all default parameters, which searches the NR/NT databases using Megablast.

BLAST results for sample N06

First, let's look at contig 'c_000000000415' from sample N06, which had 5/6 of the *nif* genes we were looking for (Figure 5.9).



Job Title	N06_c_000000000415
RID	NTHBZCMB013 *Search expires on 10-07 05:37 am* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_35245
Description	N06_c_000000000415
Molecule type	dna
Query Length	51304
Other reports	Distance tree of results MSA viewer

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Immundisolibacter cernigliae strain TR3.2, complete genome	Immundisolibact...	5293	28841	55%	0.0	85.68%	3243537	CP014671.1
☑	Xanthomonas citri pv. anacardii CFBP 2913 chromosome	Xanthomonas cit...	3356	3356	6%	0.0	84.83%	5202720	CP024057.1
☑	Xanthomonas fuscans subsp. aurantifolii strain FDC 1561, complete genome	Xanthomonas cit...	3354	3479	6%	0.0	84.28%	4993063	CP011250.1
☑	Xanthomonas fuscans subsp. aurantifolii strain 1566, complete genome	Xanthomonas cit...	3354	3479	6%	0.0	84.28%	4926567	CP012002.1
☑	Acidiferrobacter sp. SPIII_3 chromosome, complete genome	Acidiferrobacter ...	2793	4312	14%	0.0	79.50%	3398398	CP027663.1
☑	Nitrospirillum amazonense CBAmc chromosome 1, complete sequence	Nitrospirillum am...	2774	6488	17%	0.0	83.19%	2945703	CP022110.1
☑	Methyloceanibacter sp. wino2 chromosome, complete genome	Methyloceanibac...	2684	4483	10%	0.0	83.07%	3209973	CP028960.1
☑	Halothiobacillus sp. LS2, complete genome	Halothiobacillus ...	2645	5034	11%	0.0	82.62%	2908964	CP016027.1

Figure 5.9: Screenshot of BLAST results for sample N06.

There aren't any good hits here. The best one covers only 55% of the contig sequence (though it does so with a decently-high percent identity). If we look at the graphic summary, you will see that the alignment is sporadic (Figure 5.10).

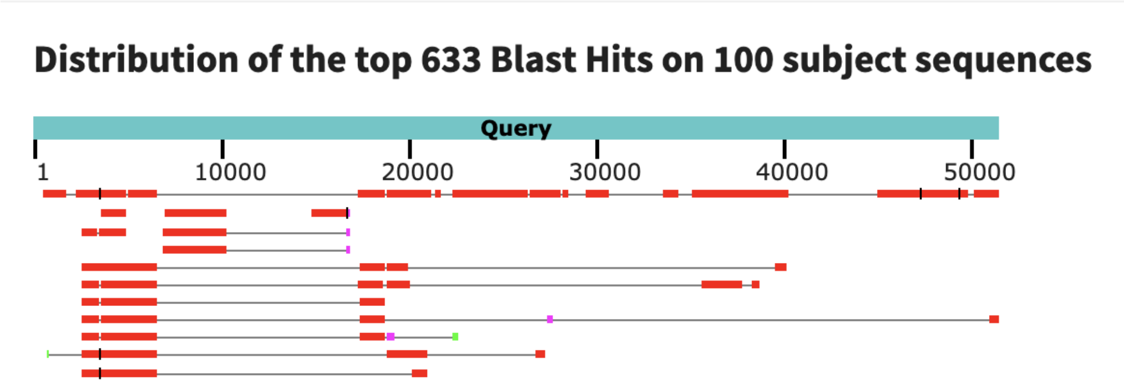
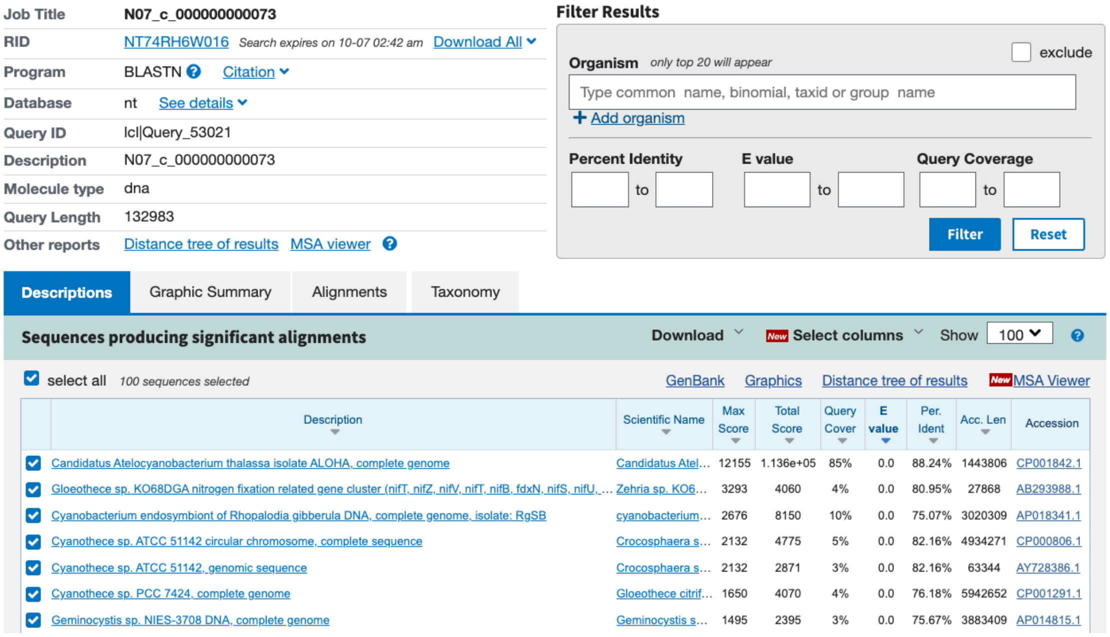


Figure 5.10: Graphic alignment of BLAST hits to the contig from sample N06.

Possibly, the top hit is matching only to the genes of this contig. According to (Corteselli et al., 2017), *Immundisolibacter cernigliae* is a soil microbe, so we wouldn't really expect to find it in the ocean. Based on these results, it seems like this nitrogen-fixing population in N06 could be a novel microbe! At the very least, it is not that similar to anything in this database. We are drawing this conclusion based on only one contig sequence from its genome, but even if the rest of its (yet unbinned) genome was similar to that of another microbe in the NCBI database, the fact that this population contains a contig with a near-complete set of *nif* genes means that it is already substantially different from that hypothetical similar population.

BLAST results for sample N07

Next, we will view the BLAST results for contig 'c_000000000073' from sample N07. This contig had all 6 of our *nif* genes on it (Figure 5.11).



Job Title	N07_c_000000000073
RID	NT74RH6W016 *Search expires on 10-07 02:42 am* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_53021
Description	N07_c_000000000073
Molecule type	dna
Query Length	132983
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity □ to □ E value □ to □ Query Coverage □ to □

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Candidatus Atelocyanobacterium thalassa isolate ALOHA, complete genome	Candidatus Atel...	12155	1.136e+05	85%	0.0	88.24%	1443806	CP001842.1
☑	Gloeothece sp. KO68DGA nitrogen fixation related gene cluster (nifT, nifZ, nifV, nifT, nifB, fdxN, nifS, nifU, ...	Zehria sp. KO6...	3293	4060	4%	0.0	80.95%	27868	AB293988.1
☑	Cyanobacterium endosymbiont of Rhopalodia gibberula DNA, complete genome, isolate: RgSB	cyanobacterium...	2676	8150	10%	0.0	75.07%	3020309	AP018341.1
☑	Cyanothece sp. ATCC 51142 circular chromosome, complete sequence	Crocosphaera s...	2132	4775	5%	0.0	82.16%	4934271	CP000806.1
☑	Cyanothece sp. ATCC 51142, genomic sequence	Crocosphaera s...	2132	2871	3%	0.0	82.16%	63344	AY728386.1
☑	Cyanothece sp. PCC 7424, complete genome	Gloeothece citrif...	1650	4070	4%	0.0	76.18%	5942652	CP001291.1
☑	Geminocystis sp. NIES-3708 DNA, complete genome	Geminocystis s...	1495	2395	3%	0.0	75.67%	3883409	AP014815.1

Figure 5.11: Screenshot of BLAST results for sample N07.

It has a much better hit in the NCBI database than the previous contig – 85% query coverage with 88% identity. *Atelocyanobacterium thalassa* is actually a well-known cyanobacterial marine diazotroph (Thompson et al., 2012). Judging by the alignment, N07's nitrogen-fixing population is extremely similar to this one (Figure 5.12).

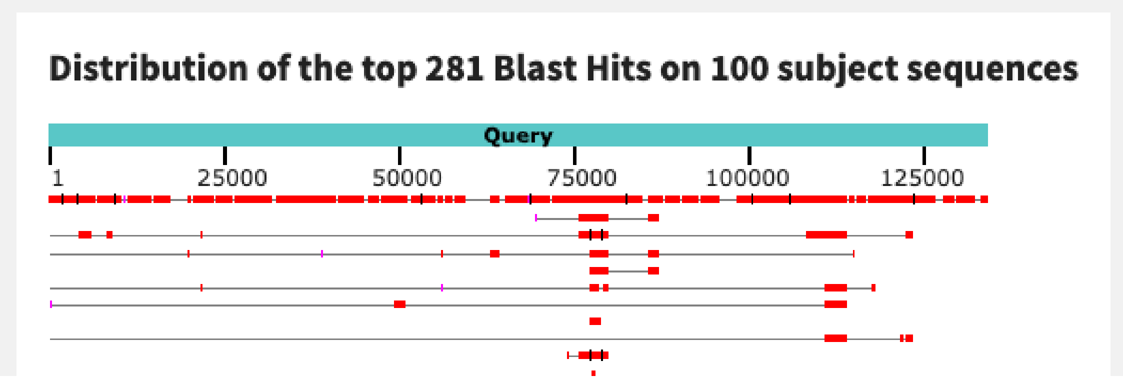
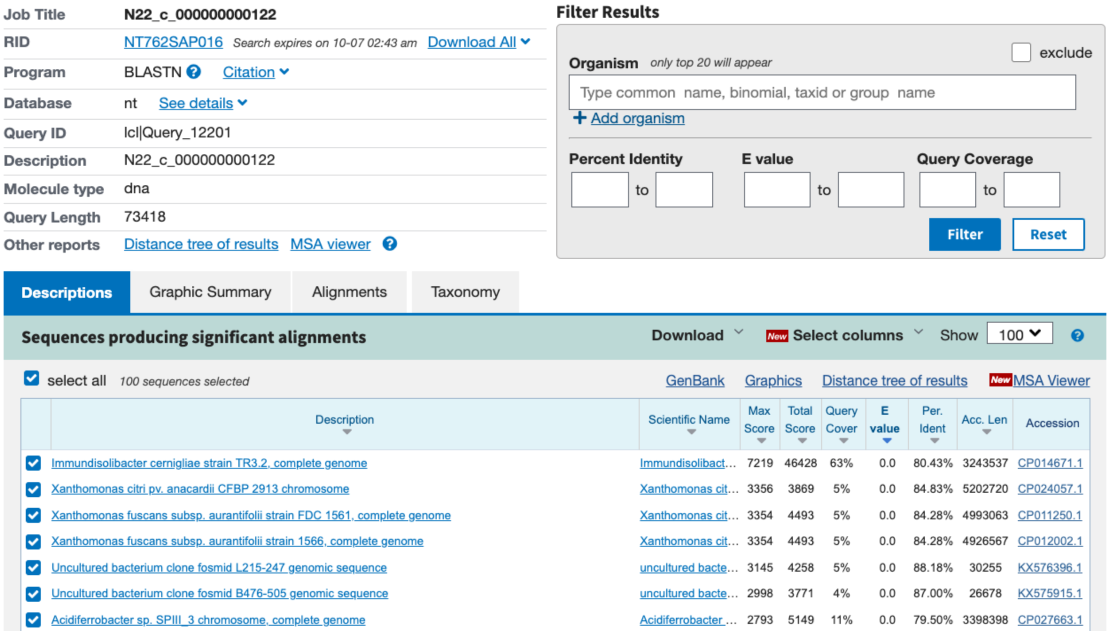


Figure 5.12: Graphic alignment of BLAST hits to the contig from sample N07.

This does not mean that the N07 population resolves to the same taxonomy as *A. thalassa* – we would need to bin the population and look at the whole genome average nucleotide identity (ANI) as well as other evidence to verify that. But it is similar enough to indicate that this population is not entirely novel. You might recall that sample N07 had another set of these genes split across a few different contigs. I wonder what you would find if you blasted those?

BLAST results for sample N22

In sample N22, the contig with the most *nif* genes was 'c_000000000122'. The BLAST results for this contig are shown in Figure 5.13, and the alignment in Figure 5.14.



Job Title	N22_c_000000000122
RID	NT762SAP016 *Search expires on 10-07 02:43 am* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_12201
Description	N22_c_000000000122
Molecule type	dna
Query Length	73418
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity □ to □ E value □ to □ Query Coverage □ to □

Filter Reset

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Immundisolibacter cernigliae strain TR3.2, complete genome	Immundisolibact...	7219	46428	63%	0.0	80.43%	3243537	CP014671.1
☑	Xanthomonas citri pv. anacardii CFBP 2913 chromosome	Xanthomonas cit...	3356	3869	5%	0.0	84.83%	5202720	CP024057.1
☑	Xanthomonas fuscans subsp. aurantifolii strain FDC 1561, complete genome	Xanthomonas cit...	3354	4493	5%	0.0	84.28%	4993063	CP011250.1
☑	Xanthomonas fuscans subsp. aurantifolii strain 1566, complete genome	Xanthomonas cit...	3354	4493	5%	0.0	84.28%	4926567	CP012002.1
☑	Uncultured bacterium clone fosmid L215-247 genomic sequence	uncultured bacte...	3145	4258	5%	0.0	88.18%	30255	KX576396.1
☑	Uncultured bacterium clone fosmid B476-505 genomic sequence	uncultured bacte...	2998	3771	4%	0.0	87.00%	26678	KX575915.1
☑	Acidiferrobacter sp. SPIII_3 chromosome, complete genome	Acidiferrobacter ...	2793	5149	11%	0.0	79.50%	3398398	CP027663.1

Figure 5.13: Screenshot of BLAST results for sample N22.

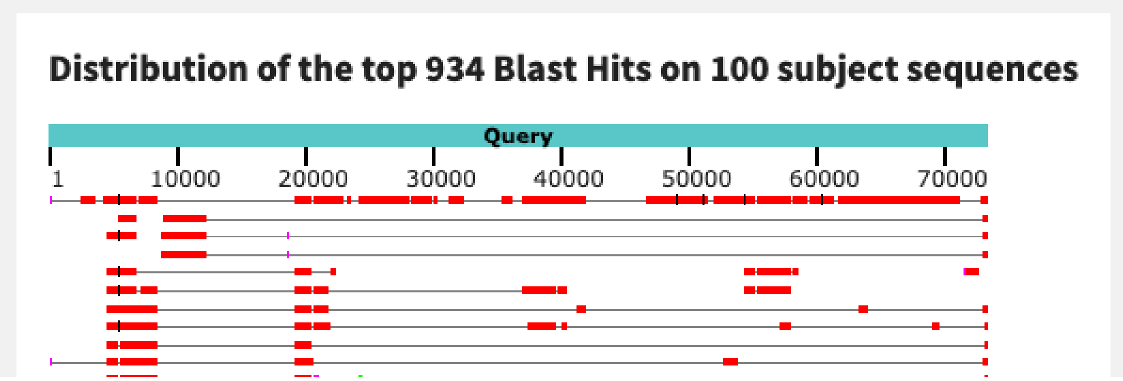
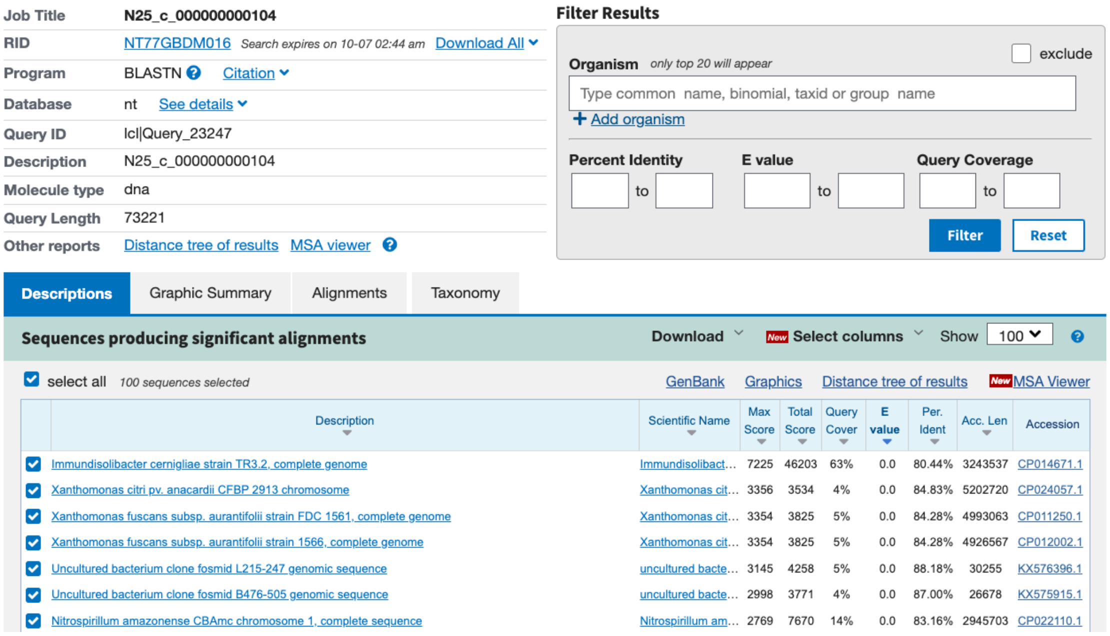


Figure 5.14: Graphic alignment of BLAST hits to the contig from sample N22.

Huh. Just like in N06, the best hit is to the *I. cernigliae* genome, with somewhat sporadic alignment.

BLAST results for sample N25

The contig from sample N25 gives us extremely similar BLAST results as the one from sample N22 (Figure 5.15 and 5.16).



Job Title	N25_c_000000000104
RID	NT77GBDM016 *Search expires on 10-07 02:44 am* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_23247
Description	N25_c_000000000104
Molecule type	dna
Query Length	73221
Other reports	Distance tree of results MSA viewer

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	Immundisolibacter cernigliae strain TR3.2, complete genome	Immundisolibact...	7225	46203	63%	0.0	80.44%	3243537	CP014671.1
☑	Xanthomonas citri pv. anacardii CFBP 2913 chromosome	Xanthomonas cit...	3356	3534	4%	0.0	84.83%	5202720	CP024057.1
☑	Xanthomonas fuscans subsp. aurantifolii strain FDC 1561, complete genome	Xanthomonas cit...	3354	3825	5%	0.0	84.28%	4993063	CP011250.1
☑	Xanthomonas fuscans subsp. aurantifolii strain 1566, complete genome	Xanthomonas cit...	3354	3825	5%	0.0	84.28%	4926567	CP012002.1
☑	Uncultured bacterium clone fosmid L215-247 genomic sequence	uncultured bacte...	3145	4258	5%	0.0	88.18%	30255	KX576396.1
☑	Uncultured bacterium clone fosmid B476-505 genomic sequence	uncultured bacte...	2998	3771	4%	0.0	87.00%	26678	KX575915.1
☑	Nitrospirillum amazonense CBAmc chromosome 1, complete sequence	Nitrospirillum am...	2769	7670	14%	0.0	83.16%	2945703	CP022110.1

Figure 5.15: Screenshot of BLAST results for sample N25.

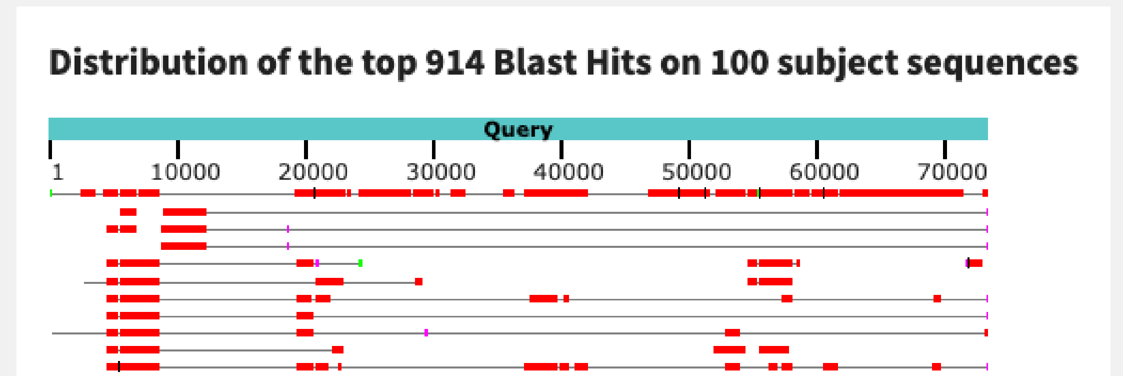


Figure 5.16: Graphic alignment of BLAST hits to the contig from sample N25.

There is a pattern emerging here. Three of the contigs that we've looked at thus far have hits to *I. cernigliae* with similar alignment coverage and identity. It is possible that these three sequences could belong to the same microbial population, in different samples.

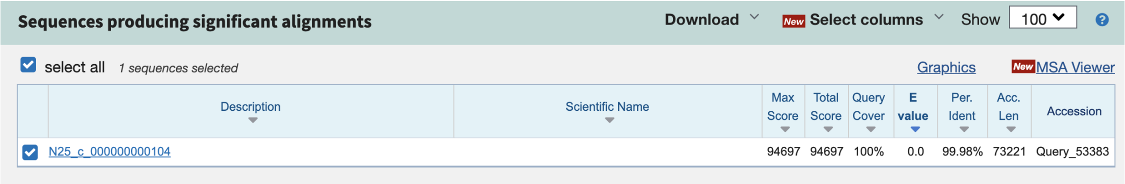
To verify their similarity, let's align the contig sequences to each other.

Aligning the contigs from different samples

The BLAST results for the contigs from N22 and N25 were so similar that we don't really need to align these two sequences, but the contig from N06 was somewhat different, with only 55% query coverage to the *I. cernigliae* genome. Let's align 'c_000000000415' from N06 and 'c_000000000104' from N25 to see whether they are similar enough to belong to the same population genome.

I again used the BLAST web service for this, just so I could show you the nice graphical alignment, but feel free to use whatever local sequence alignment program you want. If you are using the online 'blastn' suite, however, you should check the box that says 'Align two or more sequences' on the input form so that it will do this instead of searching for your sequences in the NCBI database.

Figure 5.17 shows the BLAST hit that I got when I aligned 'N25_c_000000000104' (the longer contig) to 'N06_c_000000000415'. The contigs are extremely similar, with near-100% identity! And the graphical summary shows a long, unbroken alignment (Figure 5.18).



Sequences producing significant alignments Download New Select columns Show 100

☑ select all *1 sequences selected* Graphics New MSA Viewer

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
☑	N25_c_000000000104		94697	94697	100%	0.0	99.98%	73221	Query_53383

Figure 5.17: BLAST alignment results for two contigs from different metagenomes.

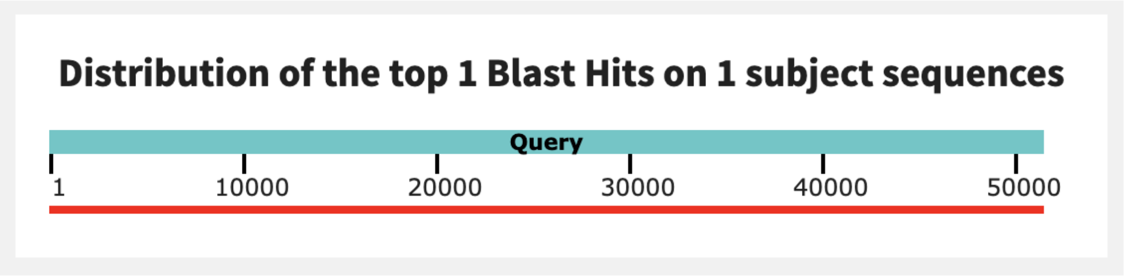


Figure 5.18: Graphical summary of alignment for two contigs from different metagenomes.

If you were to flip the order of the alignment (aligning the shorter contig from N06 to the longer one from N25), you would get a smaller query coverage value but a similar percent identity. I think these sequences are likely coming from the same microbial population, after all.

Three samples with the same population

This means that at least three of our samples (N06, N22, and N25) have the same nitrogen-fixing microbial population in them. Therefore, if we do read-recruitment of the metagenomes against any one of these samples, we'll be able to use differential coverage to bin this population.

If you are curious about where these samples are located geographically, Figure 5.19 shows the sampling map from Figure 1 of the Cao et al. paper (Cao et al., 2020), with our three samples highlighted and labeled in purple.

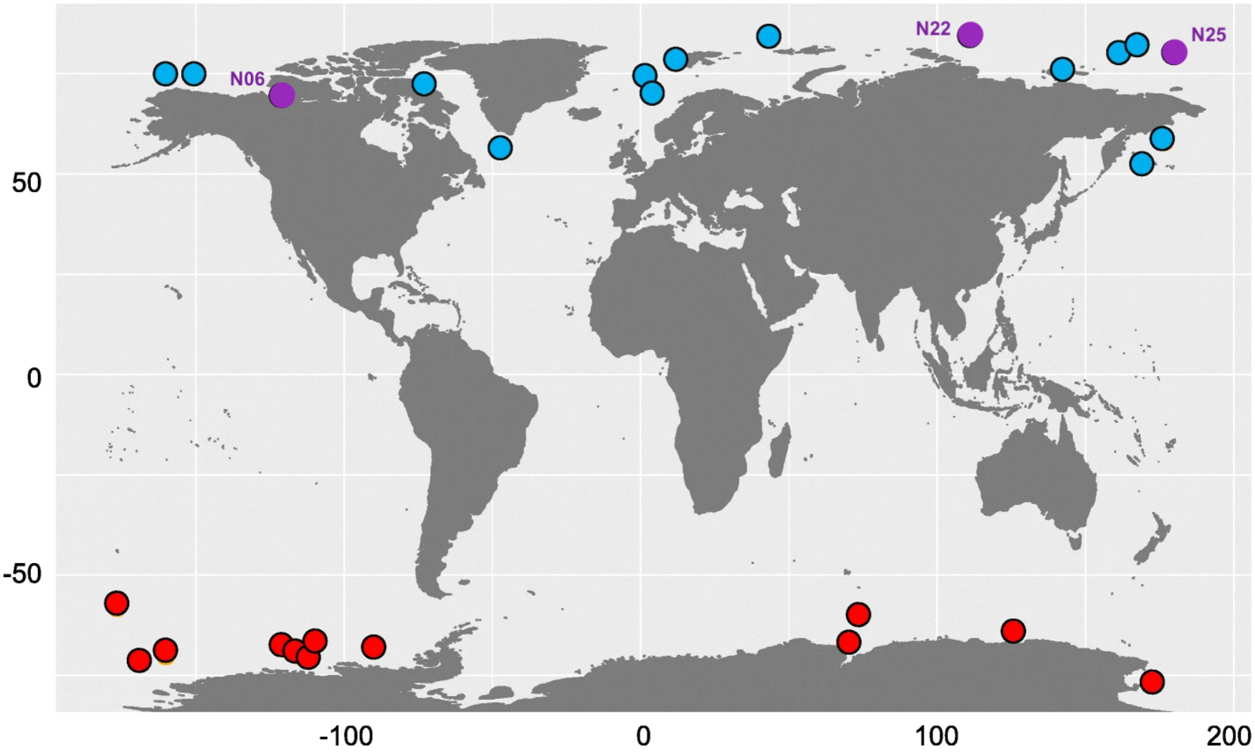


Figure 5.19: Geographic location of metagenome samples. Three samples containing a similar nitrogen-fixing population are highlighted in purple. Image modified from (Cao et al., 2020).

Clearly, this microbial population is widespread in the Arctic Ocean since it is found in both the Eastern and Western hemispheres. It also makes sense that the sequences from N22 and N25 are more similar to each other than to the one from N06, since those two samples are geographically closer together.

Comparison of *nifH* genes

We've found a nitrogen-fixing population that appears to be novel, based on its lack of good matches in NCBI. But NCBI is by no means the only source of publicly-available genomic data, so this perhaps does not mean as much as we want it to. To further verify the novelty of this population (while keeping the workload reasonably easy for us), we're going to check its *nifH* alignment against other known *nifH* genes.

When I was doing this analysis, I got a great deal of help from Tom Delmont. He kindly took the *nifH* gene from contig ‘N25_c_000000000104‘ and placed it on a phylogeny (Figure 5.20) of known *nifH* sequences from all around the world (most of them, as you may tell from the phylogeny, come from the *Tara* Oceans dataset (Sunagawa et al., 2015).

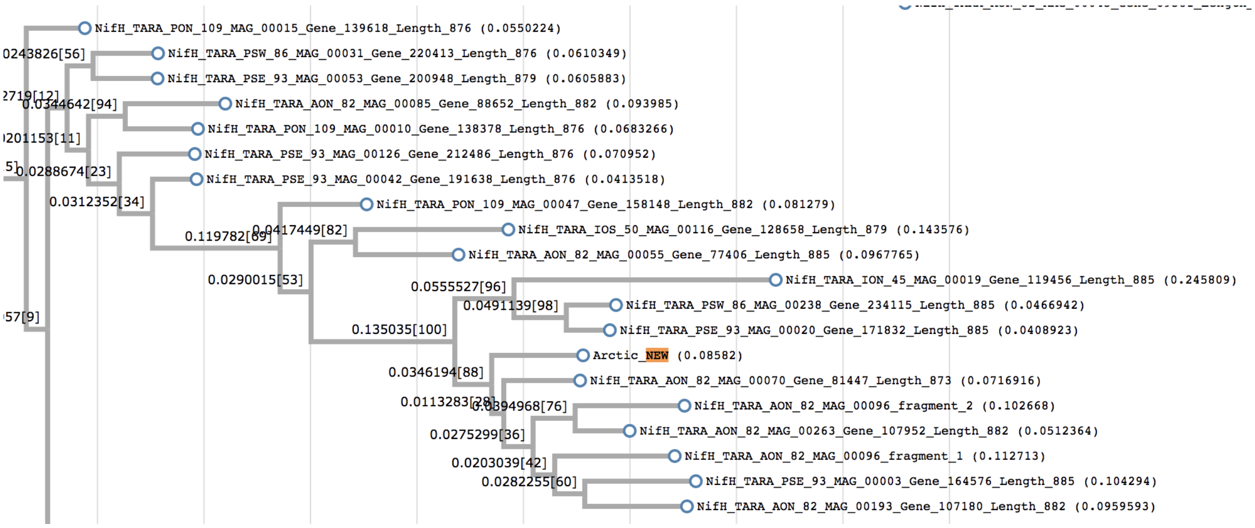

NifH_TARA_PON_109_MAG_00015_Gene_139618_Length_876 (0.0550224)
0243826[56]
NifH_TARA_PSW_86_MAG_00031_Gene_220413_Length_876 (0.0610349)
NifH_TARA_PSE_93_MAG_00053_Gene_200948_Length_879 (0.0605883)
2719[12]
0.0344642[94]
NifH_TARA_AON_82_MAG_00085_Gene_88652_Length_882 (0.093985)
NifH_TARA_PON_109_MAG_00010_Gene_138378_Length_876 (0.0683266)
201153[11]
NifH_TARA_PSE_93_MAG_00126_Gene_212486_Length_876 (0.070952)
5]
0.0288674[23]
NifH_TARA_PSE_93_MAG_00042_Gene_191638_Length_876 (0.0413518)
0.0312352[34]
NifH_TARA_PON_109_MAG_00047_Gene_158148_Length_882 (0.081279)
0.119782[69]
0.0417449[82]
NifH_TARA_IOS_50_MAG_00116_Gene_128658_Length_879 (0.143576)
NifH_TARA_AON_82_MAG_00055_Gene_77406_Length_885 (0.0967765)
0.0290015[53]
0.0555527[96]
NifH_TARA_ION_45_MAG_00019_Gene_119456_Length_885 (0.245809)
0.0491139[98]
NifH_TARA_PSW_86_MAG_00238_Gene_234115_Length_885 (0.0466942)
57[9]
0.135035[100]
NifH_TARA_PSE_93_MAG_00020_Gene_171832_Length_885 (0.0408923)
Arctic_NEW (0.08582)
0.0346194[88]
NifH_TARA_AON_82_MAG_00070_Gene_81447_Length_873 (0.0716916)
0.0113283[28]
0.0394968[76]
NifH_TARA_AON_82_MAG_00096_fragment_2 (0.102668)
0.0275299[36]
NifH_TARA_AON_82_MAG_00263_Gene_107952_Length_882 (0.0512364)
0.0203039[42]
NifH_TARA_AON_82_MAG_00096_fragment_1 (0.112713)
0.0282255[60]
NifH_TARA_PSE_93_MAG_00003_Gene_164576_Length_885 (0.104294)
NifH_TARA_AON_82_MAG_00193_Gene_107180_Length_882 (0.0959593)

Figure 5.20: Phylogeny of *nifH* sequences from around the world. The *nifH* gene from the Arctic population is highlighted. Image courtesy of Tom Delmont.

Our population’s *nifH* was most closely related to *nifH* genes from the north Atlantic Ocean, but on its own branch, indicating that there are no *nifH* genes in Tom’s collection that are exactly like it.

However, Tom found that it was most similar (with 95% identity) to the *nifH* gene from the genome of “*Candidatus* Macondimonas diazotrophica”, a crude-oil degrader isolated from a beach contaminated by the Deepwater Horizon oil spill (Karthikeyan et al., 2019).

We’re going to check how similar our population is to this “*Ca.* M. diazotrophica” genome by aligning the ‘N25_c_000000000104‘ contig against it.

```
# download the genome
wget http://enve-omics.ce.gatech.edu/data/public_macondimonas/
    Macon_spades_assembly.fasta.gz
gunzip Macon_spades_assembly.fasta.gz


# extract N25_c_000000000104 sequence into its own file
```

```
grep -A 1 "N25_c_000000000104" contigs_of_interest.fa > N25-c_000000000104
    .fa

# make a blast database for the genome
makeblastdb -in Macon_spades_assembly.fasta \
            -dbtype nucl \
            -title M_diazotrophica \
            -out M_diazotrophica

# run the alignment
blastn -db M_diazotrophica \
       -query N25-c_000000000104.fa \
       -evalue 1e-10 \
       -outfmt 6 \
       -out c_000000000104-M_diazotrophica-6.txt
```

Listing 5.8: Aligning to the “*Ca.* M. diazotrophica” genome

Looking at the ‘c_000000000104-M_diazotrophica-6.txt‘ file, you should see that the alignments are not very long (the contigs are far longer) and that the percent identities, while high, are not that high.

Here are the top 10 hits in this file:

qseqid	**sseqid**	**pident**	**length**	**mismatch**	**...**
c_000000000104	NODE_14_length_74635_cov_31.4532	90.375	3761	313	...
c_000000000104	NODE_14_length_74635_cov_31.4532	97.212	1578	43	...
c_000000000104	NODE_14_length_74635_cov_31.4532	77.627	3902	763	...
c_000000000104	NODE_14_length_74635_cov_31.4532	78.256	814	141	...
c_000000000104	NODE_14_length_74635_cov_31.4532	96.970	264	8	...
c_000000000104	NODE_14_length_74635_cov_31.4532	87.831	189	23	...
c_000000000104	NODE_11_length_97838_cov_34.3382	92.602	2379	163	...
c_000000000104	NODE_11_length_97838_cov_34.3382	93.967	1558	92	...
c_000000000104	NODE_11_length_97838_cov_34.3382	81.016	748	130	...
c_000000000104	NODE_20_length_33832_cov_39.6157	82.974	417	51	

Table 5.15: Best alignments to “*Ca.* M. diazotrophica”.

While their *nifH* genes may be very similar, this is certainly not the same population as the one we found.

There is one more set of genes that we should check. In July 2021, (Pierella Karlusich et al., 2021) published a paper containing, among other things, a set of 10 novel *nifH* genes. You will find these genes in the datapack, in the file ‘FASTA/Karlusich_novel_nifH.fa‘. Make a BLAST database out of the contig from N25 (which you extracted above), and align these *nifH* genes against that database.

```
makeblastdb -in N25-c_000000000104.fa \
            -dbtype nucl \
            -title N25-c_000000000104 \
            -out N25-c_000000000104
blastn -db N25-c_000000000104 \
       -query Karlusich_novel_nifH.fa \
       -evalue 1e-10 \
       -outfmt 6 \
       -out novel_nifH-N25_c_000000000104-6.txt
```

Listing 5.9: BLASTing against novel *nifH* genes

There are only three hits in the resulting file, and their maximum percent identity is about 86%, so none of them originate from our Arctic Ocean diazotroph.

qseqid	sseqid	pident	length	mismatch	...
ENA MW590317	c_000000000104	85.000	320	45	...
ENA MW590318	c_000000000104	84.211	323	51	...
ENA MW590319	c_000000000104	85.802	324	46	...

Table 5.16: Hits to novel *nifH* genes.

```
# navigate out of the FASTA/ folder
cd ..
```

Listing 5.10: Return to parent directory

5.3.6 Identifying the associated Cao et al. MAG

At this point, we've verified (to the best of our current knowledge), that we've identified an uncharacterized diazotrophic population in these Arctic Ocean metagenomes. Since this novel nitrogen-fixing population is present in multiple samples from the Cao et al. paper, it is extremely likely that the authors have already binned it in some form. So before we bin this population ourselves, we are going to see what else we can learn about it from their data.

Cao et al. did their binning iteratively by running first MaxBin2 (Wu et al., 2016) and then MetaBAT (Kang et al., 2015) on the contigs of individual MEGAHIT (Li et al., 2015) assemblies of these samples, and they got 214 MAGs out of this process.

We're going to find out which one of those MAGs represents the nitrogen-fixing population that we have identified in samples N06, N22, and N25. First, download their MAG set, which is hosted on FigShare (https://figshare.com/s/fd5f60b5da7a63aaa74b). You'll need to unzip the folder, and probably re-name it something sensible (I called the folder 'Cao_et_al_MAGs', and you'll see it referred to this way in the code snippets below).

Each MAG is in a FASTA file that is named according to the MAG number. We will run BLAST against all of these MAGs at the same time, so each MAG's contig sequences need to have the corresponding MAG number in the contig name. That way we will be able to determine which MAG each BLAST hit belongs to. 'anvi-script-reformat-fasta' is the perfect tool for this job.

The loop in the following code learns the MAG number from its FASTA file name and runs 'anvi-script-reformat-fasta', which will simplify the contig names and make sure each one is prefixed with the MAG number. The reformatted FASTA files will end in '*reformat.fa' and the text file matching the original contig name to its new one will end in '*reformat_report.txt'.

```
# download Cao et al MAG set
mkdir Cao_et_al_MAGs
cd Cao_et_al_MAGs/
wget https://figshare.com/ndownloader/articles/10302425?private_link=
    fd5f60b5da7a63aaa74b \
     -O Cao_et_al_MAGs.zip
unzip Cao_et_al_MAGs.zip
cd ..

# reformat contig names to contain MAG number
for g in Cao_et_al_MAGs/*.fasta; do \
  mag=$(basename $g | sed 's/.fasta//g'); \
  filename=$(echo $g | sed 's/.fasta//g'); \
  anvi-script-reformat-fasta  -o ${filename}_reformat.fa \
        --simplify-names \
        --prefix $mag \
        -r ${mag}_reformat_report.txt $g; \
done

# organize the resulting files into sensible folders
mkdir REFORMAT_REPORTS
mv *reformat_report.txt REFORMAT_REPORTS/
mkdir CAO_MAGS_REFORMATTED
mv Cao_et_al_MAGs/Genome*reformat.fa CAO_MAGS_REFORMATTED/
```

Listing 5.11: Downloading and reformatting the Cao et al. MAGs

After that finishes, you can concatenate all of the MAG FASTAs into one big FASTA file, and make a BLAST database out of it:

```
# concatenate all MAG contigs into one file
cat CAO_MAGS_REFORMATTED/*.fa > all_Cao_MAGs.fa

# make database for mapping against these contigs
makeblastdb -in all_Cao_MAGs.fa \
            -dbtype nucl \
            -title all_Cao_MAGs \
            -out all_Cao_MAGs
```

Listing 5.12: Make a BLAST database

Since we know that contigs ‘N06_c_000000000415‘, ‘N22_c_000000000122‘, and ‘N25_c_000000000104‘ are all similar, we only need to BLAST one of them against this database. I chose ‘N25_c_000000000104‘ arbitrarily, but feel free to try one of the others if you’d like.

```
# extract sequence into its own file (if you haven't done this already)
grep -A 1 "N25_c_000000000104" \
    FASTA/contigs_of_interest.fa > FASTA/N25-c_000000000104.fa


# blast this contig against all Cao et al MAGs
    # standard output format
blastn -db all_Cao_MAGs \
       -query FASTA/N25-c_000000000104.fa \
       -evalue 1e-10 \
       -out c_000000000104-all_Cao_MAGs-0.txt
    # tabular output format
blastn -db all_Cao_MAGs \
       -query FASTA/N25-c_000000000104.fa \
       -evalue 1e-10 \
       -outfmt 6 \
       -out c_000000000104-all_Cao_MAGs-6.txt
```

Listing 5.13: BLAST against the Cao et al. MAG database

If you look at the tabular output file, you will see that there is really only one good match for contig ‘N25_c_000000000104‘, and that is a hit against ‘Genome_122_000000000019‘ (or, contig 19 from ‘Genome_122‘. The reformat report for this MAG indicates that contig 19 was originally named ‘k141_74885‘. In case that matters to anyone.). It has almost 100% identity over nearly the entire contig (you can see the alignment in the standard output file, if you are curious about that).

It seems like ‘Genome_122‘ is the nitrogen-fixing MAG that we have been looking for. In fact, supplementary table S5 from the Cao et al. paper indicates that (according to GTDB-Tk) this MAG belongs to the Immundisolibacter genus. Well, we have seen enough of the alignments to know that this taxonomy is probably not correct, but it was the closest match on NCBI. This is enough to verify that we found the correct MAG.

So a MAG of our population of interest has already been binned, as expected. If this weren’t a blog post on how to do targeted binning, you might think that we were done here. But it is a blog post about targeted binning, so we are not done just yet. We will do our own binning in just a moment. In the meantime, we can still use the Cao et al. MAG to learn things about our population of interest that will help us bin this population ourselves - namely, the distribution of this population in different parts of the ocean, which will help us know what to expect in terms of differential coverage patterns.

In addition, this ‘Genome_122‘ MAG was binned automatically without any manual refinement, which as we know can be problematic (Eren and Scott, 2020b). Is this MAG complete? Is it chimeric? As we investigate it further in the next section, keep these questions in mind, because their answers will further motivate us to bin this population ourselves.

5.3.7 Distribution of ‘Genome_122‘ in the global oceans

Thus far, we’ve 1) identified a nitrogen-fixing population in the Arctic Ocean, 2) inferred its novelty from its lack of matches to NCBI and a collection of known *nifH* genes, and 3) found its corresponding MAG in the Cao et al. data. Our next question is - where does this population occur across the world? Can it be found only in the Arctic, or is it a globally-distributed population (that for some reason has not yet been characterized in temperate oceans)? Is it limited to the surface ocean, or can it live in the deep?

To answer this question, I mapped four different datasets of ocean metagenomes to the

‘Genome_122‘ MAG using the anvi’o metagenomic workflow. Those datasets are: the current one from Cao et al. (including all samples, from the Arctic and Antarctic), the ACE dataset of Southern Ocean metagenomes (which is yet unpublished, but is a sampling expedition led by our collaborator Lois Maignien at the IUEM-Brest), and the vast global ocean sampling efforts *Tara* (Sunagawa et al., 2015) and *Tara2* (Salazar et al., 2019). We’re going to look at these mapping results. You will find the relevant databases in your datapack, in the ‘GENOME_122_DBS‘ folder.

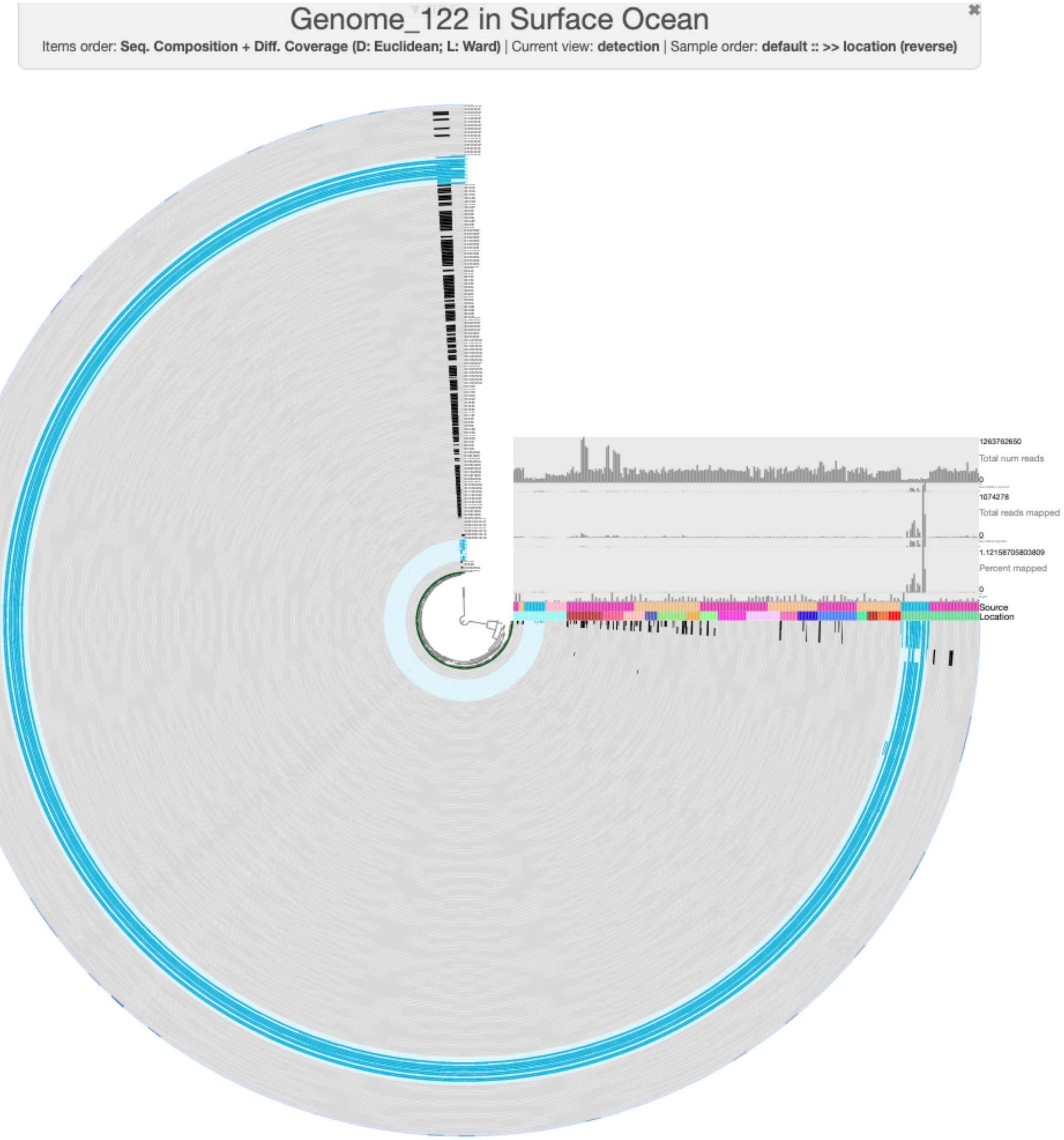
Surface ocean distribution

First, open up the anvi’o interactive interface and take a look at the distribution of this MAG in the surface ocean (which includes metagenomes sampled at depths 0-100m from these four datasets):

```
cd GENOME_122_DBS/
anvi-interactive -c Genome_122-contigs.db \
                 -p SURFACE/SURFACE_PROFILE.db \
                 --title "Genome_122 in Surface Ocean"
```

Listing 5.14: Opening the interactive interface for surface ocean distribution

Figure 5.21 shows the view of the interface that you should see.



Genome_122 in Surface Ocean

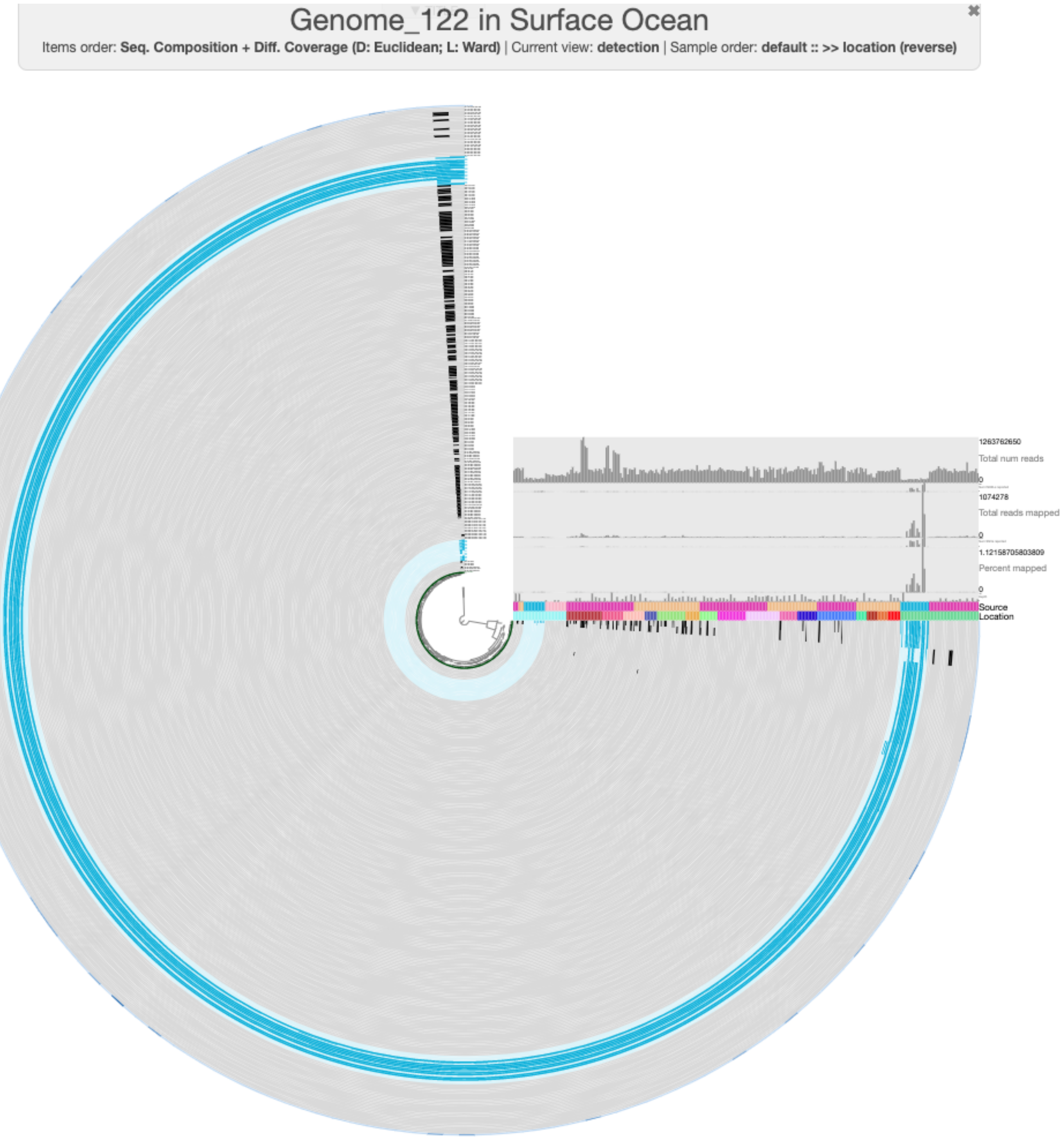
Items order: **Seq. Composition + Diff. Coverage (D: Euclidean; L: Ward)** | Current view: **detection** | Sample order: **default :: >> location (reverse)**

Figure 5.21: Distribution of ‘Genome_122‘ MAG from Cao et al. across the global surface oceans. Each radial bar (organized via the dendrogram) represents one contig from the MAG and each concentric barplot represents the contig’s detection (proportion of bases covered by at least one short read) in a single metagenome sample. Samples from Cao et al. are colored in light blue.

The default view in the interface should show log-normalized detection of this MAG in all of the ocean metagenomes. Each concentric circle in the figure is one metagenome sample, and each spoke of the wheel is a contig from ‘GENOME_122‘. Samples from Cao et al. have been marked in light blue to distinguish them from the rest. You can hover over the ‘Source’ layer to see which dataset each sample comes from, and the ‘Location’ layer to see which ocean region it was sampled from.

There are a few things we can immediately see from the mapping results. First, it is clear that this population is geographically isolated, as it is detected only in the Arctic Ocean samples from the Cao et al. dataset. There are some Arctic Ocean samples from *Tara2* (the darkest green in the ‘Location’ layer), but this population is not detected in these. Second, this MAG must have been binned from the Cao et al. assembly of sample N07, since that sample has the highest proportion of mapping reads. Though we didn’t discuss it earlier, our nitrogen-fixing population is also present in sample N07 (which you may already have deduced if you took a look at the BLAST results for the second set of *nif* contigs in N07).

Finally, there are several splits in ‘Genome_122‘ that appear to be contamination. For instance, Figure 5.22 shows a zoomed view on three splits that have different detection values across samples than the rest of the MAG.



Genome_122 in Surface Ocean

Items order: **Seq. Composition + Diff. Coverage (D: Euclidean; L: Ward)** | Current view: **detection** | Sample order: **default :: >> location (reverse)**

Figure 5.22: Three pieces of MAG ‘Genome_122‘ with a different detection pattern. This figure is a zoomed-in view of Figure 5.21.

One of those splits is missing detection in samples N22 and N25 (where we know our population exists) - this split is marked with an arrow in the figure above. The other two are detected in a variety of samples from the other datasets as well as a different detection pattern across the other Cao et al. Arctic Ocean samples. There are also a couple more splits at the top of the circular phylogram that seem problematic.

A quick aside to look for *nif* genes

Well, I'm sure you found contig 19. But you couldn't find our *nif* genes, could you? In fact, if you search for functions with "nitrogen fixation", you will find several annotated *nif* genes but not the ones that we were looking for – except for *nifB*, which is not on contig 19 (as expected) but on contig 27. This is extremely curious. How could this happen? Previously, contig 'N25_c_000000000104', which contains 5 out of 6 of our *nif* genes, matched with almost 100% identity against the entirety of contig 19 - so what is missing?

It turns out that contig 19 is quite a bit shorter – only 51,626 bp – compared to 73,221 bp for N25_c_000000000104. You might have noticed this if you checked the standard output file from BLAST. Clearly, the part of contig 104 that contains those 5 *nif* genes was not the part that matched to contig 19. We are working with different assemblies of these metagenomes than the ones created by Cao et al., so some differences are to be expected.

Furthermore, we now know that 'Genome_122' was binned from sample N07. In N07, the second set of *nif* genes was split across 3 contigs ('c_000000004049', 'c_000000000256', and 'c_000000000095'), so it is likely that a similar situation occurred in the Cao et al. assembly of this sample. Which means it is certainly possible that only the contig containing *nifB* was binned into this MAG. Contig 27 from 'Genome_122' is probably the counterpart to 'c_000000000095' from our assembly.

You can check this, if you want, by BLASTing those three N07 contigs against the Cao et al. MAGs:

```
# go back to the previous folder
cd ..

# extract just these 3 contigs from N07 into a separate file
for c in c_000000004049 c_000000000256 c_000000000095; \
do \
  grep -A 1 $c FASTA/contigs_of_interest.fa >> FASTA/N07_second_set.fa; \
done

# align against the MAG set
blastn -db all_Cao_MAGs \
       -query FASTA/N07_second_set.fa \
       -evalue 1e-10 \
       -outfmt 6 \
       -out N07_second_set-all_Cao_MAGs-6.txt
```

Listing 5.15: BLASTing N07 contigs against Cao et al. MAGs

I'll paste the relevant hits from the output below. These are the best hits for each contig query (meaning that they have the highest percent identity, the longest alignment lengths, and the smallest e-value of all hits from that contig).

qseqid	**sseqid**	**pident**	**length**	**mismatch**	**...**
N07_c_000000000095	Genome_122_000000000027	99.982	116075	5	...
N07_c_000000004049	Genome_022_000000000007	99.729	7751	20	...
N07_c_000000004049	Genome_022_000000000007	99.203	753	6	...
N07_c_000000000256	Genome_122_000000000019	99.992	51584	4	...

Table 5.17: Best hits from N07 contigs to Cao et al. MAGs.

First of all, contig ‘N07_c_000000000095‘ (the one with the *nifB* gene) indeed matches extremely well to contig 27 from ‘Genome_122‘, as expected. Contig ‘N07_c_000000004049‘, which contained a copy of the M00175 module, does not match to anything in ‘Genome_122‘ at all (which explains why those three genes were missing from the MAG). Instead it matches to a contig from the MAG named ‘Genome_022‘, but the alignment length is rather small.

However, contig ‘N07_c_000000000256‘, which contained *nifE* and *nifN* in our assembly, matches to contig 19 of ‘Genome_122‘! It is 64,963 bp long, so here we have the same situation as contig 104 from sample N25 - it is a much longer sequence than contig 19, and the *nifE* and *nifN* genes must be on the part that does not match to contig 19. (Indeed, if you blast ‘N07_c_000000000256‘ against ‘N25_c_000000000104‘, it will match to one end of that contig.)

The long story short is that our *nif* genes of interest were not binned into the ‘Genome_122‘ MAG, but we have plenty of evidence that they do belong to this population, considering that those genes were assembled together in other samples. Sadly, that means that ‘Genome_122‘ is incomplete, and it is missing the genes we care most about. But more on this later.

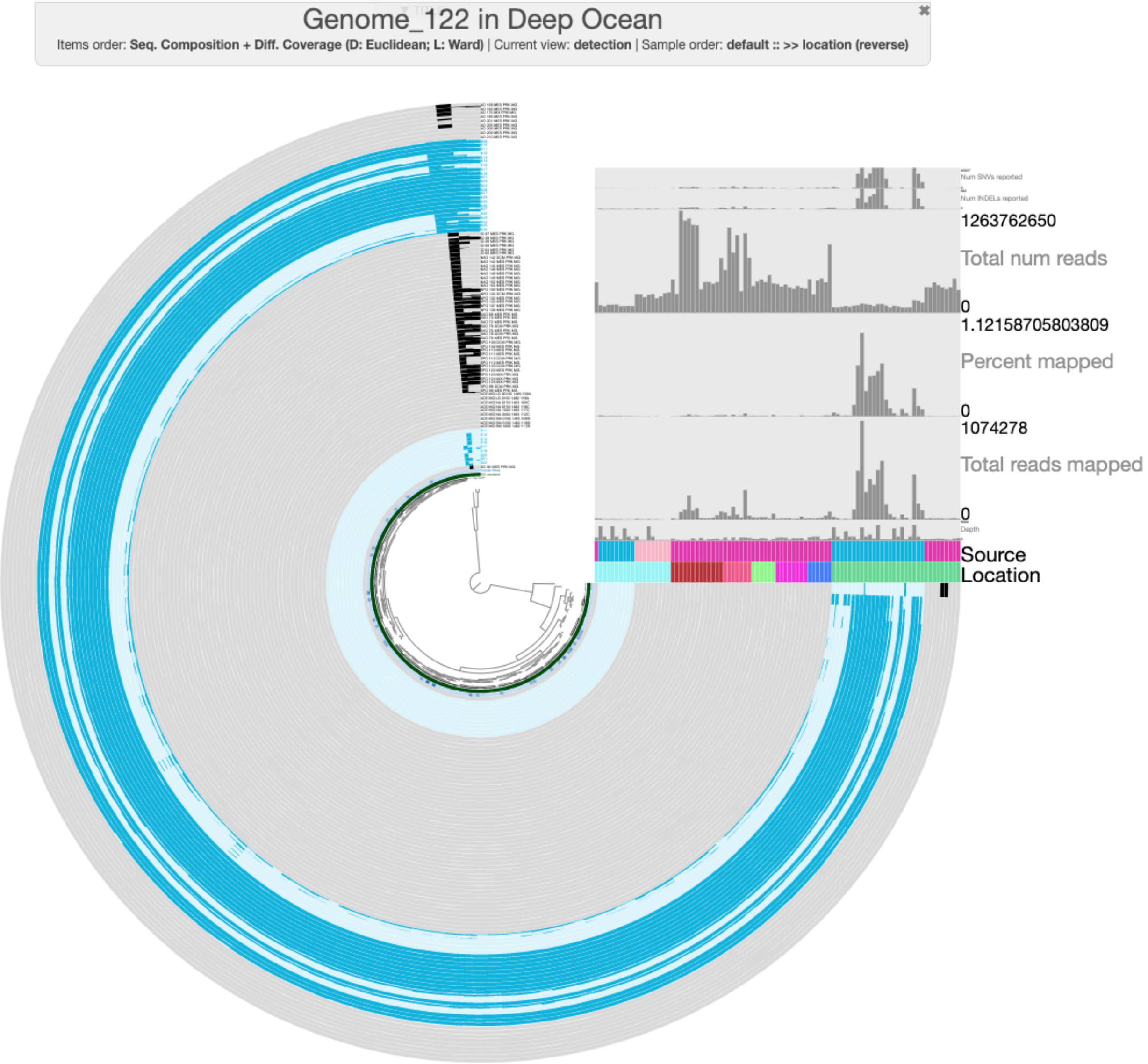
Deep ocean distribution

The next thing to view is the distribution of this MAG in deeper samples (100m < depth <= 3800 m).

```
cd GENOME_122_DBS/
anvi-interactive -c Genome_122-contigs.db \
                 -p DEEP/DEEP_PROFILE.db \
                 --title "Genome_122 in Deep Ocean"
```

Listing 5.16: Opening the interactive interface for deep ocean distribution

A screenshot from the interface is shown in Figure 5.23. The samples are color-coded in the same way as before. You should be able to see that this MAG is present in deeper waters (even those as deep as 3800m), though it is still geographically limited to the Arctic Ocean. And once again, there are several splits that just don’t seem to fit with the rest and most likely represent contamination.



Genome_122 in Deep Ocean

Items order: **Seq. Composition + Diff. Coverage (D: Euclidean; L: Ward)** | Current view: **detection** | Sample order: **default :: >> location (reverse)**

Figure 5.23: Distribution of ‘Genome_122‘ MAG from Cao et al. across the global deep oceans. Each radial bar (organized via the dendrogram) represents one contig from the MAG and each concentric barplot represents the contig’s detection (proportion of bases covered by at least one short read) in a single metagenome sample. Samples from Cao et al. are colored in light blue.

```
# we are done here
cd ..
```

Listing 5.17: Return to parent directory

5.3.8 Targeted binning of the nitrogen-fixing population

We've seen above that the ‘Genome_122‘ MAG appears to have some contamination, which is a normal thing to see in MAGs (particularly automatically-generated ones), because binning is hard. We've also seen that it does not contain the *nif* genes that belong to this nitrogen-fixing population. But since we have time on our hands, a particular interest in just this one nitrogen-fixing population, and the knowledge of which *nif* gene-containing contigs belong to this population, we can make a better MAG. It's time for some targeted binning.

We know that our population of interest is present in samples N06, N07, N22, and N25. We could use any of these assemblies for binning, though N07 is not the best choice because the *nif* genes are split across more contigs in that one. I once again made the completely arbitrary choice to use sample N25 for this. I ran a read recruitment workflow to map all 60 Cao et al. metagenomes against our assembly of N25 so that we can look at differential coverage across different metagenomes – our population of interest should only be present in the Arctic Ocean samples, but we will be able to use its absence from the Antarctic samples to help guide our binning. You'll find the contigs database for the N25 assembly and the profile database containing these mapping results in the datapack (in the ‘N25_DBS‘ folder). You can open them up in anvi-interactive:

```
cd N25_DBS/
anvi-interactive -c N25-contigs.db \
                 -p PROFILE.db \
                 --title "Cao et al Read Recruitment to N25" \
            --state-autoload binning
```

Listing 5.18: Opening the interactive interface for binning

The databases are rather large, and may take some time to load, but once they do you should see a display similar to Figure 5.24.

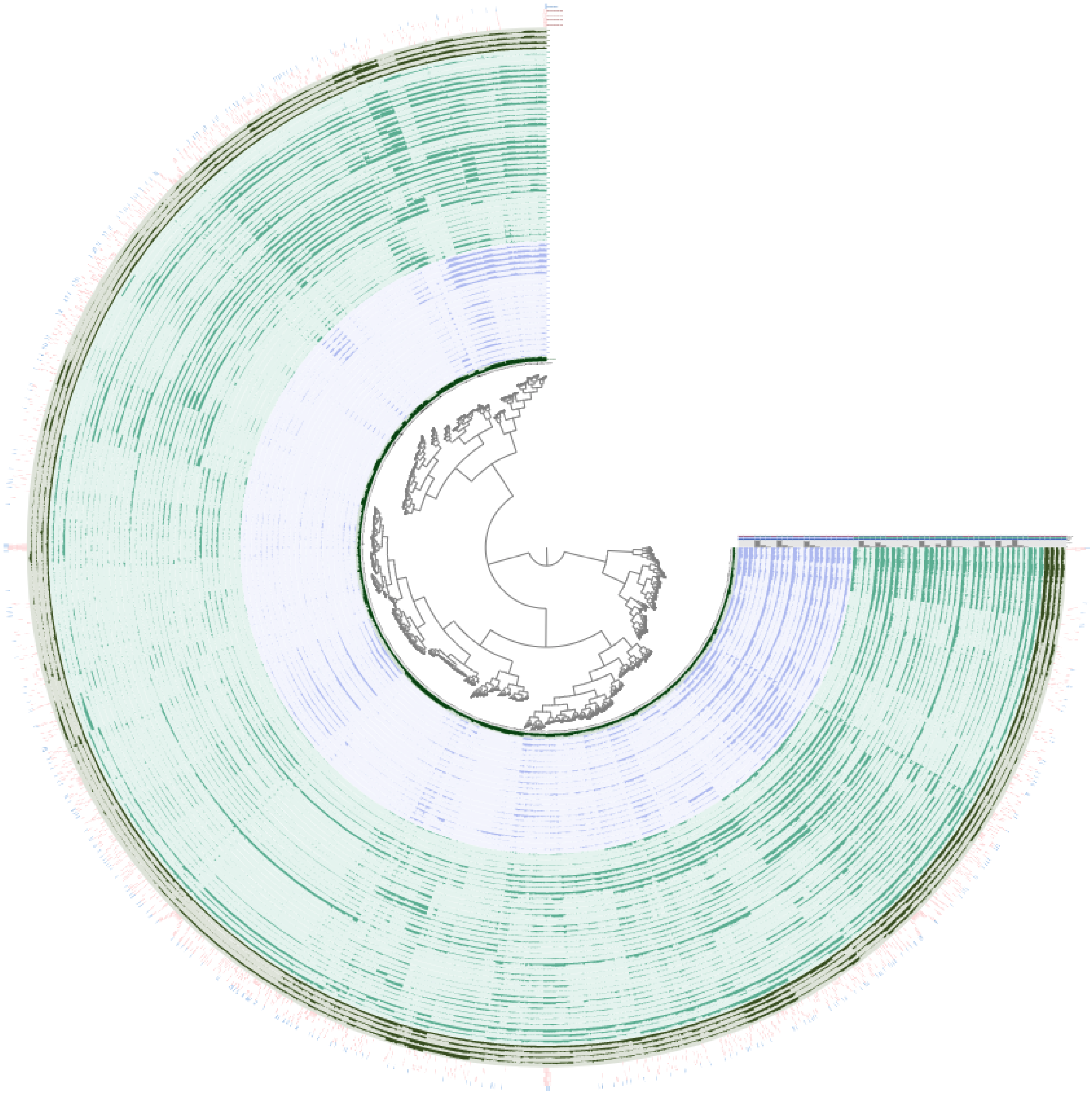


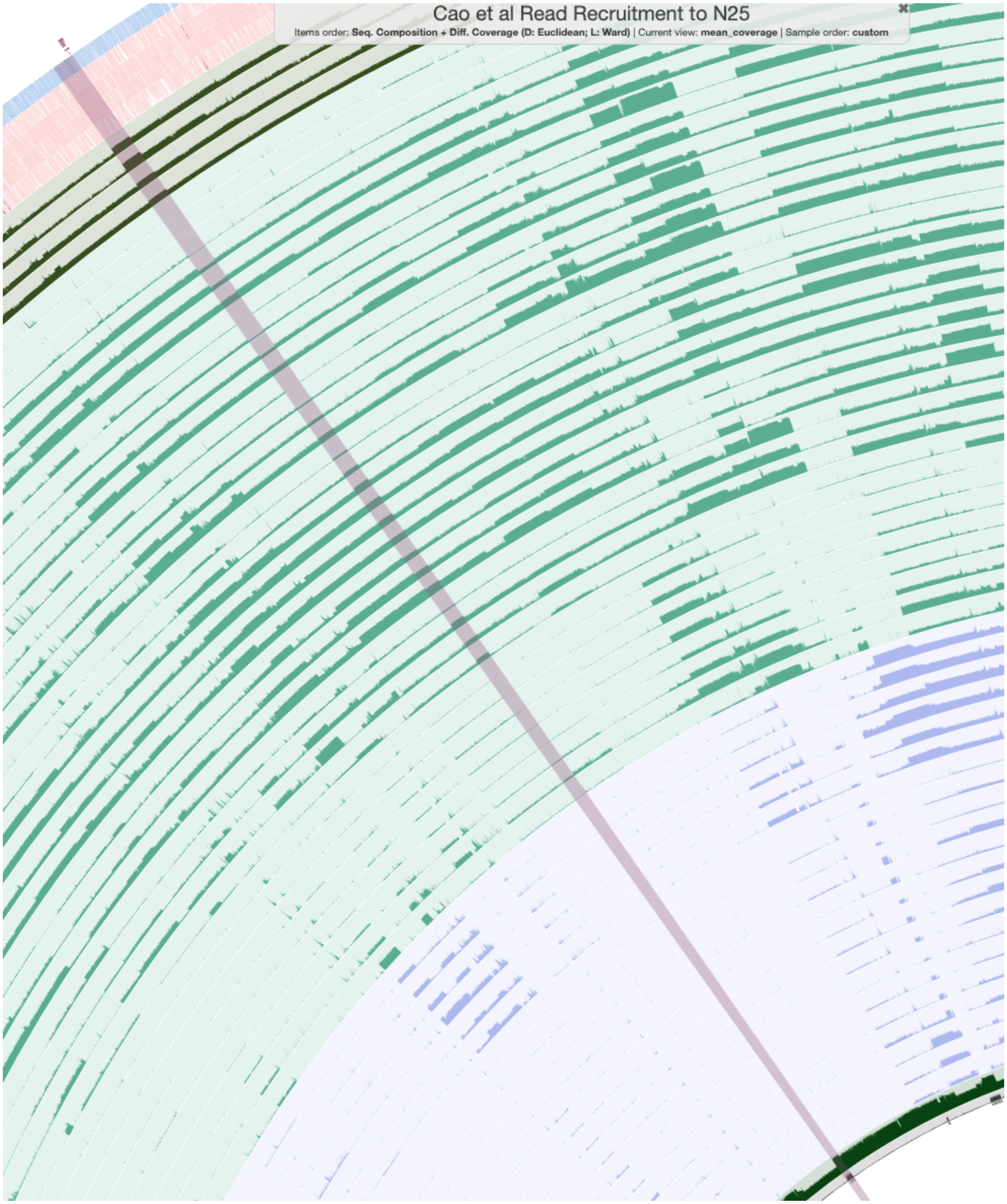
Figure 5.24: Read recruitment of Cao et al. metagenomes to the N25 assembly. Each radial bar (organized via the dendrogram) represents one contig from the N25 assembly and each concentric barplot represents the contig's coverage (number of reads mapping to it). Samples from the Antarctic are blue and samples from the Arctic are green. Samples N06, N07, N22, and N25 are darker green.

The Arctic Ocean samples are green, and the four samples we expect to find our population in are the outermost, darker green layers so that we can more easily focus on those. The blue samples are the Antarctic ones.

We will start our binning with the contig that contains the most *nif* genes in N25, which is ‘000000000104‘. You can search for this contig in the ‘Search’ tab of the ‘Settings’ panel, and add its splits to a bin.

The contigs in this assembly are clustered according to their sequence composition and their differential coverage (across all Cao et al. samples), so the other contigs that belong to our nitrogen-fixing population should be located next to contig ‘000000000104‘ in the circular phylogram. These contigs should also appear in all four of our samples of interest (dark green), have zero coverage in the Antarctic samples (blue), and have similar GC content (the green layer below the Antarctic samples). If you zoom to the location of the splits you just binned, you should see a set of splits that fit this criteria.

Did you find it? Figure 5.25 shows the splits I am talking about, so that you can check your work.



Cao et al Read Recruitment to N25

Items order: **Seq. Composition + Diff. Coverage (D: Euclidean; L: Ward)** | Current view: **mean_coverage** | Sample order: **custom**

Figure 5.25: Contigs belonging to the nitrogen-fixing population in sample N25. A zoomed-in view of Figure 5.24. The binned contigs are highlighted in light pink.

These were the splits that I binned (there are 168 of them). You can bin them yourself, or just load the collection called 'Nif_MAG' to see the same bin on your own screen. The anvi'o estimates of completion and redundancy (based on bacterial single-copy core genes) for this bin are 100% and 0%, respectively, which is great news. Furthermore, if you check the box for real-time taxonomy estimation on the "Bins" tab, you will see that this bin is labeled as *Immundisolibacter cernigliae*, the same microbe that we kept getting BLAST hits to previously. So we've certainly binned the correct population, and it is a high-quality MAG at that.

Bonus activity: Recall that there were 3 copies of *nifB* in sample N25, on three separate contigs. Which one belongs to this population?

5.3.9 Estimating metabolism for our new MAG

Now that we have a complete MAG for our nitrogen-fixing population, let's see what else it can do. We are going to run metabolism estimation on this population.

There are a couple of different ways we can go about this. Since the bin is saved as a collection, you can directly estimate its metabolism from the current set of databases for the entire N25 assembly, just like this:

```
anvi-estimate-metabolism -c N25-contigs.db \
                         -p PROFILE.db \
                         -C Nif_MAG \
                         -O Nif_MAG \
                         --kegg-output-modes kofam_hits,modules
```

Listing 5.19: Estimating metabolism on our new MAG

Or, you could split this MAG into its own set of (smaller) contig/profile databases, and then run metabolism estimation in genome mode:

```
anvi-split -c N25-contigs.db \
           -p PROFILE.db \
           -C Nif_MAG \
          -o Nif_MAG
anvi-estimate-metabolism -c Nif_MAG/Nif_MAG/CONTIGS.db \
                         -O Nif_MAG \
                         --kegg-output-modes kofam_hits,modules
```

Listing 5.20: A different way to estimate metabolism on the MAG

You can pick whichever path you like. I went with the latter option because I wanted a stand-alone database for the MAG so I could do other things with it, but the former is less work for you (and for your computer). Regardless of how you do it, you should end up with a ‘Nif_MAG_modules.txt‘ file containing the module completeness scores for this population, and a ‘Nif_MAG_kofam_hits.txt‘ file containing its KOfam hits.

You are free to explore these results according to your interests, but one of my remaining questions about this population is whether it is a cyanobacteria or a heterotroph. Cyanobacteria have photosynthetic and carbon fixation capabilities, while heterotrophs have ABC transporters for carbohydrate uptake (Cheung et al., 2021). I looked for modules related to each of these things and checked their completeness scores.

Here is my search code. I once again clipped the output so that it shows only relevant fields.

```
head -n 1 Nif_MAG_modules.txt | cut -f 3,4,7,9; \
grep -i "carbon fixation" Nif_MAG_modules.txt | cut -f 3,4,7,9
```

Listing 5.21: Parsing for carbon fixation pathways

kegg_module	module_name	module_subcategory	module_completeness
M00165	Reductive pentose phosphate cycle (Calvin cycle)	Carbon fixation	0.8181818181818182
M00166	Reductive pentose phosphate cycle, ribulose-5P =>glyceraldehyde-3P	Carbon fixation	0.75
M00167	Reductive pentose phosphate cycle, glyceraldehyde-3P =>ribulose-5P	Carbon fixation	0.8571428571428571
M00168	CAM (Crassulacean acid metabolism), dark	Carbon fixation	0.5
M00173	Reductive citrate cycle (Arnon-Buchanan cycle)	Carbon fixation	0.8
M00376	3-Hydroxypropionate bi-cycle	Carbon fixation	0.4423076923076923
M00375	Hydroxypropionate-hydroxybutylate cycle	Carbon fixation	0.14285714285714285
M00374	Dicarboxylate-hydroxybutyrate cycle	Carbon fixation	0.38461538461538464
M00377	Reductive acetyl-CoA pathway (Wood-Ljungdahl pathway)	Carbon fixation	0.2857142857142857
M00579	Phosphate acetyltransferase-acetate kinase pathway, acetyl-CoA =>acetate	Carbon fixation	0.5
M00620	Incomplete reductive citrate cycle, acetyl-CoA =>oxoglutarate	Carbon fixation	0.35714285714285715

Table 5.18: Carbon fixation modules in our MAG.

Several of the reductive pentose phosphate cycle pathways look near-complete. However, these results must be taken with a grain of salt because many of these pathways share a large number of their KOs with other pathways. We can confirm whether or not this is a cyanobacteria by looking for photosynthesis capabilities:

```
head -n 1 Nif_MAG_modules.txt | cut -f 3,4,7,9; \
grep -i "photo" Nif_MAG_modules.txt | cut -f 3,4,7,9
```

Listing 5.22: Parsing for photosynthesis pathways

kegg_module	module_name	module_subcategory	module_completeness
M00532	Photorespiration	Other carbohydrate metabolism	0.475
M00611	Oxygenic photosynthesis in plants and cyanobacteria	Metabolic capacity	0.4090909090909091
M00612	Anoxygenic photosynthesis in purple bacteria	Metabolic capacity	0.4090909090909091
M00613	Anoxygenic photosynthesis in green nonsulfur bacteria	Metabolic capacity	0.22115384615384615
M00614	Anoxygenic photosynthesis in green sulfur bacteria	Metabolic capacity	0.4

Table 5.19: Photosynthesis modules in our MAG.

As you can see, none of these modules are complete (including the pathway specifically for cyanobacteria), so this doesn't appear to be a cyanobacterial population. A huge caveat here is that our MAG could simply be missing the genes relevant to this pathway (or, it has them, but they are not homologous enough to their corresponding KO families to be annotated). This is a possibility with any MAG. But if we choose to trust these estimations (given the high completeness score of our bin), the current evidence points to this population being heterotrophic.

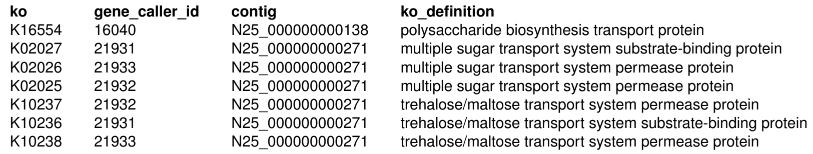
There is no module for carbohydrate transporters, since these are individual proteins rather than a metabolic pathway, but we can look for KOfam hits that are annotated as transporters instead.

```
head -n 1 Nif_MAG_kofam_hits.txt | cut -f 3-5,7; \
grep -i 'transport' Nif_MAG_kofam_hits.txt | cut -f 3-5,7
```

Listing 5.23: Parsing for transporter enzymes

There are plenty of hits, including several specifically for carbohydrates:

So it looks like this microbe is indeed a heterotroph, which would make it a heterotrophic bacterial diazotroph, or HBD. It can join its temperate ocean relatives in Tom's hard-earned collection (Delmont et al., 2021).



ko	gene_caller_id	contig	ko_definition
K16554	16040	N25_000000000138	polysaccharide biosynthesis transport protein
K02027	21931	N25_000000000271	multiple sugar transport system substrate-binding protein
K02026	21933	N25_000000000271	multiple sugar transport system permease protein
K02025	21932	N25_000000000271	multiple sugar transport system permease protein
K10237	21932	N25_000000000271	trehalose/maltose transport system permease protein
K10236	21931	N25_000000000271	trehalose/maltose transport system substrate-binding protein
K10238	21933	N25_000000000271	trehalose/maltose transport system permease protein

Table 5.20: Transporter enzymes in our MAG.

5.3.10 Final Words

So there you have it - a novel, heterotrophic nitrogen-fixing population from the Arctic Ocean, binned directly from public metagenomes with guidance from anvi-estimate-metabolism. We went fishing, and we caught something interesting. And it wasn't all that hard.

If you ever have to look for a microbe of interest in metagenomic data, and you know it has something unique in terms of its metabolic capabilities, you can try out this technique in your search. Perhaps you'll find what you are looking for!

5.4 'Digital microbe': a data integration framework for genomics and pangenomics

As the amount and variety of 'omics data increases at a rapid pace, a mechanism for effective sharing of integrated datasets has become an emerging need in the 'omics field. There is now a wide range of methods for characterizing biological systems – i.e., genomics, transcriptomics, proteomics, metabolomics – each of which produces complementary information that can be integrated to yield a comprehensive understanding of a given organism. Yet, data integration remains a challenging task that begins at the earliest stages of data acquisition with technical bottlenecks that hamper analysis workflows. This problem is exacerbated in a collaborative research environment because the team's primary data may be downloaded from independent repositories with divergent records; for example, different versions of a microbe's genome, or different annotations applied to the same gene, transcript, or protein. This limits downstream efforts to build synergistic insights into living systems.

In my work as part of a large research consortium, the Center for Chemical Currencies of a Microbial Planet (C-CoMP, `https://ccomp-stc.org/`), I have started to address this problem by implementing a framework for collaborative data integration and sharing. C-CoMP is a center spanning multiple research groups from thirteen different institutions, and its primary scientific goal is to understand microbial carbon cycling in the global surface ocean by combining experiments on model organisms, modeling of metabolic networks, and field observations. To accomplish this task, it requires a means of passing integrated ‘omics datasets (versioned for reproducibility and consistency) between its members. The C-CoMP Data Integration Working Group has devised a strategy for doing this via a ‘digital microbe’ framework that is broadly useful to any ‘omics scientist needing to share data.

A ‘digital microbe’ is a curated and versioned public data package that combines multiple datasets related to a particular organism (i.e., genome) or group of related organisms (i.e., pangenome). It is both self-contained (it can explain itself and its contents) and extensible (others can extend a digital microbe data package with additional layers of information coming from new experiments). The datasets in the package are organized and linked through reference to the microbial genome(s), consolidating a variety of information including gene annotations, read-mapping data such as coverage statistics, and sample metadata. Digital microbes are flexible in scope, being suitable for single microbial genomes as well as clade-level genomic collections. Further, their extensibility via the programmatic addition of new ‘omics data types makes them future-proof. The databases are easy to share and directly usable as inputs to further analyses, making them an efficient collaboration strategy that eliminates the problems of sharing individual datasets and then requiring collaborators to integrate independently. Importantly, the databases can be versioned as different team members add datasets and analyses. This allows for data management between collaborators and reproducibility by external scientists.

In C-CoMP, we implemented this concept using anvi’o (Eren et al., 2021a) to ensure integrated access to its analysis and visualization tools. That is, each of our ‘digital microbes’ is composed of one or more anvi’o SQLite databases, which are both programmatically-queryable and accessible via the vast network of interoperable anvi’o programs. C-CoMP makes these databases available, to both collaborators in the center and to the public, on the online data repository Zenodo (`https://zenodo.org/`). This particular strategy offers several benefits, such as the ability to share multiple datasets by providing a single link, and the ability to directly analyze and interactively visualize these data. However, researchers could implement a similar framework using alternative tools and data-sharing platforms, if desired.

The utility of this framework is exemplified by a digital microbe that I generated for one of C-CoMP’s primary model organisms, *Ruegeria pomeroyi* DSS-3. *R. pomeroyi* is a member of the *Rhodobacteraceae* family, a clade of metabolically active bacterial cells in algal blooms and coastal environments (Munson-McGee et al., 2022). This r-strategist species has a large genome encoding a variety of catabolic genes that control the fate of climate-relevant organic sulfur gasses (Durham et al., 2015; Howard et al., 2006; Landa et al., 2019) as well as numerous pathways for synthesis of cofactors and vitamins that support chemical interactions with marine phytoplankton (Durham et al., 2017). It was isolated and sequenced to obtain a complete reference genome (Moran et al., 2004), and since then numerous studies on this model organism have yielded a wealth of data, from transcriptomes to a TnSeq mutant library (publication in progress by Moran et al.).

In our ‘digital microbe’ for *R. pomeroyi* DSS-3, I integrated its complete genome and megaplasmid sequence (Moran et al., 2004) with a manually-curated set of gene function annotations provided by Zac Cooper, Christa Smith and William Schroer in the Moran Lab; automatically-generated functional annotations from sources including Pfam (Mistry et al., 2021), NCBI COGs (Galperin et al., 2021), and KEGG KOfam/BRITE (Aramaki et al., 2020; Kanehisa et al., 2023); read recruitment data from 133 (meta)transcriptome samples taken spanning 6 publications (and 2 manuscripts in preparation) (Durham et al., 2015; Landa et al., 2019; Ferrer-González et al., 2021; Nowinski and Moran, 2021; Olofsson et al., 2022; Uchimiya et al., 2022); and annotations for genes with available mutants in the Moran Lab's TnSeq library. The databases (along with a reproducible workflow describing how they were generated) are publicly-available on Zenodo (Veseli and Cooper, 2022), and have been documented in a blog post that also provides examples of how to visualize and analyze the data further using anvi'o (Veseli, 2023) (`https://ccomp-stc.org/rpom-digital-microbe/`). Version tracking in Zenodo has enabled the seamless sharing of several iterations of this data package, and C-CoMP members are already using the integrated data as a base for further research on this important marine microbe.

Overall, the significance of the 'digital microbe' data integration framework lies in its ability to facilitate scientific progress, collaboration, and accountability. It enables the synergistic investigation of multiple integrated datasets for systems-level biological insights into microbial life. The databases are easily shareable, extensible, and reproducible; thus, they are an ideal platform for collaborative analyses and open science practices. Though C-CoMP's digital microbes make use of anvi'o, SQLite databases, and the Zenodo data repository, the same benefits would apply to alternative implementations using other software and platforms; that is, the concept of this framework is more important than the implementation. Regardless of their form, digital microbes represent a major breakthrough for the future of collaborative research in the 'omics field.

CHAPTER 6

CONCLUSIONS

6.1 Summary of contributions

Microbes are of immense ecological and medical importance, and with the advancement of high-throughput sequencing technologies, ‘omics data and analysis software have become critical for microbiome research (White et al., 2016; Callahan et al., 2018). In many cases, it is now more effective to analyze microbes from environmental samples rather than isolating and culturing them for *in vitro* experimental work (which is indeed not yet possible for many recalcitrant clades) (Lloyd et al., 2018; Steen et al., 2019; Wang et al., 2020). Thus, high-throughput analysis workflows and data integration techniques are necessary to make sense of the wealth of sequencing data and to better understand microbial ecology in a variety of contexts. During my graduate studies, I have advanced this field of research by developing accessible, flexible computational strategies for integration and analysis of ‘omics data, foremost a novel tool for the study of microbial metabolism. I applied my tool in a number of studies of metagenomic data to further our understanding of microbial functional potential, especially in the gut microbiome, which is extremely relevant to human health. In particular, my work in the human gut environment has shed light on the determinants of gut microbial resilience in the context of inflammatory bowel disease.

My most important contributions to the microbial ‘omics field are three-fold. The first is technical: I implemented a software framework for metabolism reconstruction, specifically pathway prediction, from genomes and metagenomes. This was a significant undertaking that required thoughtful design, based on the input of multiple collaborators, to incorporate features necessary for it to be a useful addition to the existing repertoire of metabolism reconstruction software. Over the past few years my framework underwent several incremental improvements to become a comprehensive, flexible, and accessible tool that has already been used in a number of scientific publications – for example, (Watson et al., 2022; Delmont et al., 2021; Miranda et al., 2022; Runde et al., 2023; Delmont, 2021; Weigel et al., 2022; Rasmussen et al., 2023; Becken et al., 2021; Castro-Severyn et al., 2021; Johnson et al., 2021; Choudoir et al., 2023; Komova et al., 2022; Nuppunen-Puputti et al., 2022; Giacomini et al., 2023; Modin et al., 2022; Yang et al., 2023; Eberhard et al., 2022; Breusing et al., 2022; Busi et al., 2022). One of its novel features is user-defined metabolism, which addresses the problem of our current dependency on metabolism databases that are slow to incorporate novel metabolisms from across the vast diversity of microbial life. Another key novelty of this tool is its pathway redundancy metrics, which are useful for studying community-level metabolic potential from metagenomes without requiring binning of individual populations, especially when applying the normalization technique I developed to take community size into account.

My second contribution is scientific in nature. By applying my software framework to study the metabolic potential of gut microbes (at both the genome level and the community level), I have illustrated that metabolic independence is a fitness determinant in gut communities under stress. The high completeness of several biosynthesis pathways for crucial metabolites in populations that successfully colonize FMT recipients (following several rounds of antibiotics) indicates the importance of microbial self-sufficiency for survival in depleted gut communities (Watson et al., 2022). This result also suggests that metabolic independence could be relevant in the context of inflammatory bowel disease, given that IBDs are associated with reduced diversity in the gut microbiome. I therefore leveraged a large dataset of publicly-available fecal metagenomes, as well as numerous reference genomes associated with the human gut, to verify that microbes living in the IBD gut environment are indeed more metabolically independent than microbes hosted by healthy individuals. Thus, high metabolic independence serves as a potential mechanism for microbial survival during the reduction of gut communities that occurs during disease progression. Furthermore, I showed that the pattern of metabolic independence levels mirrors the depletion and recovery of these communities following antibiotic treatment. Metabolic independence is therefore, more generally, an indicator of the level of stress faced by human gut microbes and has potential to be used for diagnosis of gastrointestinal conditions (Veseli et al., 2023). Going forward, these observations may shape our perspective on the microbiome's role in human diseases associated with microbial dysbiosis and could lead to the development of more effective diagnostic tools. Importantly, it would not have been possible to obtain these results without the metabolism framework and my state-of-the-art method for normalizing metagenomic pathway copy numbers with estimates of community size.

Finally, I believe I have made a philosophical contribution to this field by supporting and advocating for open science practices and collaborative research, especially via open-source software solutions for advanced data integration and analysis. In addition to being open-source, my metabolism reconstruction framework was developed openly, which exposed it to early scrutiny and allowed community input to shape its progression. I created a wealth of educational resources to describe the tool's methodology and increase its usability. When conducting research, I not only made use of publicly-available metagenomes, but I also generated public data (see `https://figshare.com/authors/Iva_Veseli/9014558`, `https://zenodo.org/record/7439166#.ZCHDo-zMJQ0`, and `https://www.ncbi.nlm.nih.gov/bioproject/PRJNA767321`). My primary scientific publications have been accompanied with reproducible workflows and posted on preprint servers for greater transparency and enhanced community access to the results of my work. For example, the digital microbe data integration framework that I have contributed to as part of C-CoMP combines all of these strategies. It is by design a tool for enhancing scientific collaboration on 'omics data; the databases we have generated are publicly-available on Zenodo; I wrote a blog post describing our *Ruegeria pomeroyi* digital microbe, how to use it, and the reproducible workflow for generating it (Veseli, 2023); and the manuscript (in preparation) for describing this framework will be posted on a preprint server. By practicing open science, I hope that I have demonstrated the benefits of this approach to the scientific community as a whole.

I would like to conclude this dissertation by highlighting a few of the unsolved challenges and future directions encompassed by my research. In the following section, I offer my perspectives on the technical, scientific, and philosophical aspects of my work introduced above.

6.2 Concluding remarks and perspectives

6.2.1 The problem of annotation bias and its effect on metabolism prediction

The overrepresentation of model organisms, easily cultured microbes, and organisms of particular medical or industrial significance in genomic databases leads to annotation bias, whereby gene sequences that are similar to reference genes are more readily annotated with protein functions than more distant (yet still homologous) sequences. This insidious problem results from the very nature of homology-based gene annotation strategies, and causes a systematic lack of gene annotations in less-well-studied clades of organisms and for less-characterized gene families (Lobb et al., 2020). It occurs despite the continuous addition of genomic data from novel organisms to reference databases, because experimental validation of gene function occurs at a much slower pace than generation of new sequence data, because models used for annotation are not frequently updated to incorporate this new data, and because the e-value and bit score thresholds for assessing match quality must be stringent to avoid mistakenly annotating a gene with an incorrect function (e.g., false positives). Annotation bias is not often scrutinized in academic literature, considering that analysis workflows often rely on automated annotation strategies and that downstream, the focus is generally on the annotations that are found rather than those that are missing. Yet it can result in partial or even erroneous functional conclusions in published studies.

Annotation bias remains a significant technical challenge in metabolism reconstruction since enzyme annotations are a fundamental data source for this type of analysis. Draft genome-scale metabolic models are typically incomplete and require a gap-filling step to restore critical missing reactions before the models can be used for simulation (Orth and Palsson, 2010), yet automated gap-filling methods also rely on reference databases and can make mistakes (Karp et al., 2018). In pathway prediction, missing enzyme annotations can result in false negatives, or underestimation of pathway completeness. When the lack of annotations is systematic, affecting a particular group of organisms or samples more than another, it can produce the appearance of a biological signal, in the form of false metabolic differences between sample groups.

Annotation bias was directly observed in two of the studies featured in this dissertation, and it likely goes undetected in many more. An egregious example of annotation bias was discussed as part of the bifidobacterial study in Chapter 3. *Bifidobacteria* are relatively poorly-characterized in functional databases, which resulted in missing annotations and a falsely incomplete histidine biosynthesis pathway, until this bias was partially corrected by implementing an novel annotation heuristic to reduce the removal of annotations for distant homologs. Regardless, the heuristic was not sufficient to restore all missing annotations and discrepancies between our results and published literature on these organisms remained. As none of the authors were experts in bifidobacteria, these discrepancies went unnoticed until they were kindly pointed out by a scrupulous reviewer during a publication attempt. Thankfully, we avoided publishing incorrect conclusions in this case, but similar issues due to annotation bias must surely slip through the cracks in the peer-review process to end up in the published literature.

In Chapter 4's study of metabolic capacity in the IBD gut microbiome, I analyzed the proportion of gene calls with functional annotations across a large set of publicly-available gut metagenomes, and found that samples from healthy individuals had fewer annotations than those from individuals with IBD regardless of annotation source (Figure 6.1). The bias between the groups was visible in deeply-sequenced metagenomes, but disappeared after I expanded the set to include shallow metagenomes. Annotation bias provides one explanation for these observations: metagenomes from healthy people contain a more diverse community of microbes, and these populations (especially those of lower abundance that are only detected in deeply-sequenced samples) seem to be less well-characterized than the microbes living in the IBD gut environment, leading to the discrepancy in annotation efficiency between these two sample groups. This bias confounds the metabolic signal between the sample groups in this study, and it is difficult to determine how extensively it, rather than real biological signal, contributes to the group differences.

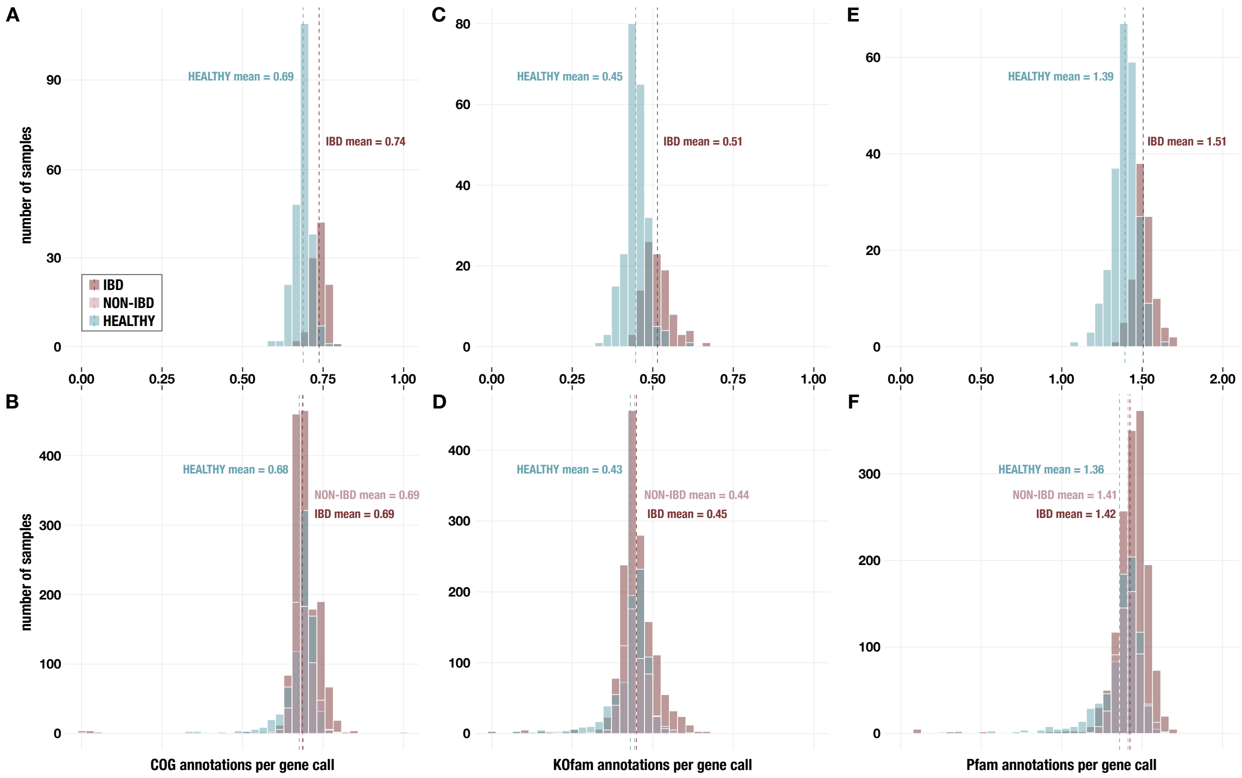


Figure 6.1: Histograms of annotations per gene call from A,B) NCBI COGs; C,D) KEGG KOfams; and E,F) Pfams. Panels A, C, and E show data for metagenomes in the subset of 330 deeply-sequenced samples from healthy people and people with IBD, and panels B, D, and F show data for all 2,893 samples including those from non-IBD controls.

With its low rate of detection and potentially drastic consequences for functional analysis, annotation bias is an insidious problem affecting a wide range of ‘omics research that relies on homology-based annotation from reference databases. Addressing it remains a major challenge, especially for automated metabolism reconstruction efforts. Curation of public genomic data and generation of better annotation models will continue to be slow, so technical advancements to improve the annotation process remain the best option, for now. Strategies such as the annotation heuristic I implemented in ‘anvi-run-kegg-kofams‘ or the extension of annotations from a curated genome to less-characterized, taxonomically-related genomes using pangenomic gene clusters are only the first step in addressing this issue, and more comprehensive solutions are needed. In the meantime, great care must be taken when assessing metabolism reconstruction results to avoid drawing incorrect conclusions on microbial functional potential.

6.2.2 Implications of metabolic independence in gut microbiome research

The gut microbiome is so often studied due to its relevance to human health, with a focus on the ecosystem services microbes collectively provide to their host as well as their influence on disease pathogenesis. Yet gut microbes also provide extensive ecosystem services to each other. Our observations of high metabolic independence in disrupted gut communities are consistent with this notion and underscore the immense importance of cross-feeding and microbe-microbe interactions to the robustness of the gut microbiome (Wang et al., 2019; van Hoek and Merks, 2017; Gutiérrez and Garrido, 2019). These findings suggest a mechanism whereby stress-related perturbations in the gut disrupt these mutualistic interactions to create a ‘snowball’ effect that further deteriorates the community. Notably, another recent investigation (Marcelino et al., 2023) supports this interpretation.

Many studies of diseases associated with dysbiosis of the human gut microbiome operate under the assumption that these changes reflect a microbial basis of the disease (Clemente et al., 2012; Lee and Chang, 2021; Weiss and Hennet, 2017), though causal relationships have often been elusive (Ni et al., 2017; Janssen and Kersten, 2015; Dinan and Dinan, 2022;

Bielka et al., 2022; Lynch et al., 2019). For example, in IBD, the lack or attenuation of disease symptoms in gnotobiotic animal models indicates the requirement for a microbial stimulus to promote disease development (Sellon et al., 1998), leading to the suggestion that the IBD gut microbiota is “functionally defective” (Nagao-Kitamoto et al., 2016) or contains some problematic microbes that potentiate disease. However, there are no specific clades of microbes that are universally associated with IBD, and the characteristic reduction of gut diversity instead follows broad taxonomic patterns (Knox et al., 2019b; Lane et al., 2017). This raises the possibility that no causal pathobionts exist; rather, inflammation arises from an overactive immune response to nonspecific microbes, and the observed gut dysbiosis is a secondary effect of intestinal inflammation (Tamboli et al., 2004). The inconclusive efficacy of antibiotics to treat this condition (Ledder, 2019) as well as the limitations of animal models for human disease (Arrieta et al., 2016) further complicate the picture. And more generally, the sheer amount of research proposing a causal relationship between the gut microbiome and diseases is implausible, suggesting that we may have overstated the importance of the gut microbiome to disease pathogenesis (Walter et al., 2020; Lynch et al., 2019).

The impulse to label microbes as perpetrators of disease – if not as pathogens, then as pathobionts (Jochum and Stecher, 2020) – is tempting because it offers a simple mechanism for disease pathogenesis as well as the promise of microbially-based treatments. Yet the concept of metabolic independence suggests a different paradigm that could apply to some of these diseases and disorders: that gut microbes are bystanders to the onset of health conditions and ‘dysbiosis’ represents their reaction to disease-related changes in their environment. The disruption of the microbial community may eventually have side effects that worsen disease symptoms or progression; for instance, subsequent infection with opportunistic pathogens such as adherent-invasive *Escherichia coli*, which is often found in individuals with IBD (Darfeuille-Michaud et al., 2004). However, under this hypothesis, the initial impetus for pathogenesis is not caused by a set of specific, problematic populations or “functionally-defective" microbes – rather, the role of the microbiota in disease progression is primarily reactive. Clearly, the human-microbiota holobiont is a very complex system, and it is difficult to study the intricacies of their interactions, especially in humans and even more so from samples representing singular time points in advanced stages of disease. But perhaps we would benefit from a more holistic, 'ecosystem'-level perspective on the human microbiome, in which we consider that both humans and their microbes are jointly affected by disease progression, and that a microbial scapegoat does not necessarily exist for every disease.

Regardless of how involved gut microbes are in the pathogenesis of various gastrointestinal disorders, the observation that these conditions result in the survival of only a subset of highly competent populations suggests an opportunity to leverage the gut microbiome as an indicator of host disease states. If a high proportion of metabolically-independent populations in the gut microbiome is consistently found in individuals with disease and not in healthy individuals, then estimation of metabolic potential from metagenomes could be a viable diagnostic tool. That is, metabolic independence could represent a reproducible and detectable microbial signature of gastrointestinal stress conditions. This notion is supported by the accurate classification of post-antibiotic treatment samples using the high metabolic independence metric that was described in Chapter 4. Further study across a variety of gastrointestinal conditions is necessary to verify whether this concept is widely-applicable, but if so, it could be especially significant for improving the detection of diseases like IBD for which no taxonomic biomarker has been identified (Knox et al., 2019b; Lee and Chang, 2021).

Finally, microbial taxonomy has long been a significant focus of gut microbiome research, with many studies elucidating the contributions of specific taxa to ecosystem services (Leylabadlo et al., 2020; Salyers et al., 1977; Kovatcheva-Datchary et al., 2009) and profiling the changes in overall community composition related to various conditions (Machiels et al., 2014; Petrov et al., 2017; Schirmer et al., 2018a; Zhu et al., 2018). The underlying assumption of this focus is the idea that populations of similar taxonomy have conserved functional roles and will behave similarly in a given environment. Yet there is evidence against this assumption. For instance, in a recent time-series study of exclusive enteral nutrition therapy for Crohn's disease, highly-similar microbial populations (with average nucleotide identity >= 98%, even more closely-related than taxonomic species) exhibited variable and even contradictory responses to the treatment in different subjects (Runde et al., 2023). Local environmental context is therefore a significant determinant of microbial behavior, more so than taxonomic identity. Furthermore, a number of biological characteristics undermine the idea that shared taxonomy necessitates shared functional capacity: the well-known plasticity of microbial genomes that permits the exchange of genetic information even between distant clades (Gogarten et al., 2002; Koonin, 2016; Frye et al., 2011), the phenomena of gene deletions (Zhu et al., 2015), and the variable sizes of accessory genomes (Mira et al., 2010). These characteristics prevent the unequivocal attribution of functional capacity to microbial genomes based on their taxonomy alone, and highlight the utility of functional inference directly from sequence data as opposed to reference-based approaches. Thus, an over-reliance on taxonomy may be limiting our mechanistic understanding of the microbiome's role in human health (Armour et al., 2019). Accordingly, the 'omics field has been gradually shifting towards functional investigations of microbial ecology (Armour et al., 2019; Doolittle and Booth, 2017). The investigations of metabolic independence described in this thesis provide examples of entirely taxonomy-independent analyses that yield significant insights into microbial ecology, thus demonstrating the power of functional approaches in characterizing the dynamics of the gut microbiome.

6.2.3 The importance of open science practices, accessible open-source software, and public data to scientific advancement

The growing open science movement offers a number of benefits to the scientific community. Conducting research openly and collaboratively advances science faster, encourages reproducibility and accountability, improves scientific accuracy by exposing early results to public scrutiny and feedback, and indeed could be considered our social responsibility (Munafò et al., 2017; Ramachandran et al., 2021; McKiernan et al., 2016). The 'omics field has long been at the forefront of open data practices (Byrd et al., 2020; Perez-Riverol et al., 2019) and is uniquely suited to lead the adoption of other open science practices. We publish our data in public repositories for sequences, proteins, metabolites, and models (O'Leary et al., 2016; Parks et al., 2022; UniProt Consortium, 2023; Drula et al., 2022; Haug et al., 2020; Karp et al., 2019; Aramaki et al., 2020; Galperin et al., 2021; Bateman et al., 2002); we develop and use software that relies extensively on this public data (Beghini et al., 2021; Chaumeil et al., 2019; Zdobnov and Apweiler, 2001); and our computational analysis workflows have the potential to be highly reproducible (more so than wet lab protocols) if we carefully manage our computational environments and record software commands and parameters. A large number of 'omics software is open-source, and several labs and organizations have spearheaded the practice of publishing reproducible analysis workflows alongside data (i.e., https://hypocolypse.github.io; https://merenlab.org/data/; (Arkin et al., 2018a)).

In my research, I have experienced first-hand the benefits of open science. I primarily use publicly-available metagenomes rather than generating new sequence data; in fact, the work described in Chapter 4 relies almost exclusively on public metagenomes, as does the (Delmont et al., 2021) study from Chapter 5 and several analyses in Chapter 3. In all of these studies, the existing datasets led to new insights and provided critical context for any original data. Repurposing public data is an extremely efficient way to do science, and as a community, we can make it even more efficient by reducing the number of steps required for others to use our public data. The FAIR data principles (Wilkinson et al., 2016) provide a framework for increasing reusability of public data. One of the most critical areas of improvement, in my experience, is publishing clear and comprehensive metadata that allows downstream users to find the samples they need and to integrate them with other datasets. The types of metadata required to properly describe a sample are often field-specific (for instance, gut metagenomes need metadata on host diet and health conditions, while marine metagenomes need metadata on filter size and sampling location) and thus difficult to standardize. This means that the onus is currently on individual researchers to ensure their public data is accompanied by enough metadata to accommodate a reasonable range of use-cases. Another strategy to increase data reusability is to share pre-integrated datasets, thereby reducing the need for other researchers to repeat the integration process. The digital microbe framework discussed in Chapter 5 is one example of this.

Open science practices were also critical to the development of the metabolism reconstruction framework, and have helped it to become a widely-used tool. By developing my software in a public Github repository and sharing it with scientists in my network, I was able to rapidly fix bugs and add helpful features thanks to public feedback from its early users, making it a truly community-driven effort. For example, the option to provide a list of enzymes as input to ‘anvi-estimate-metabolism‘ arose from a collaborator’s need to make it compatible with output from his BLASTx filtering software (`https://github.com/merenlab/anvio/pull/1890`). Moreover, the extensive documentation and tutorials associated with this tool enabled other researchers to use it independently of my help, as evidenced by the large number of publications referencing my framework long before I published any formal journal article about it – for examples, see (Rasmussen et al., 2023; Becken et al., 2021; Castro-Severyn et al., 2021; Johnson et al., 2021; Choudoir et al., 2023; Komova et al., 2022; Nuppunen-Puputti et al., 2022; Giacomini et al., 2023; Modin et al., 2022; Yang et al., 2023; Eberhard et al., 2022; Breusing et al., 2022; Busi et al., 2022). However, one remaining bottleneck in the open-source ‘omics software movement is the relative lack of community contributions to the development and documentation of software, which are most often maintained and extended by a small number of people typically originating from the same lab or organization. Expanding the pool of developers to include people from the wider community with relevant experience and ideas would increase the pace of progress, prevent stagnation of the codebase and reduce issue buildup; this is indeed one of the supposed benefits of developing on public code repositories like Github. However, the barrier for ‘outsiders’ – that is, people outside of the original development team – to contribute code remains high. Primary software developers need to lower this barrier in order to benefit from community involvement, whether that is by providing guidelines on how to contribute, writing understandable and properly documented code, or establishing a welcoming digital presence for their tool to encourage newcomers to its development process. For example, I wrote a blog post including step-by-step instructions for updating documentation on the anvi’o codebase to help anvi’o users share their expertise with the rest of the community (Veseli, 2022).

Despite its benefits, significant obstacles remain before open science practices can be widely and freely adopted by all (Dominik et al., 2022). Many scientists are constrained by their available resources and energy; not everyone has the time, funding, institutional support, technology, or knowledge necessary to follow open science practices. Yet small actions can tip the scale in favor of systemic change. I was fortunate enough to have enough resources and support to do my science openly, and I chose to do that because I think it is an important direction for the scientific community to strive for. I hope that by doing so, I was able to positively influence research in the microbial ‘omics field.

REFERENCES

Aggeletopoulou, I., Konstantakis, C., Assimakopoulos, S. F., and Triantos, C. (2019). The role of the gut microbiota in the treatment of inflammatory bowel diseases. *Microb. Pathog.*, 137:103774.

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

Agus, A., Planchais, J., and Sokol, H. (2018). Gut microbiota regulation of tryptophan metabolism in health and disease. *Cell Host Microbe*, 23(6):716–724.

Aite, M., Chevallier, M., Frioux, C., Trottier, C., Got, J., Cortés, M. P., Mendoza, S. N., Carrier, G., Dameron, O., Guillaudeux, N., Latorre, M., Loira, N., Markov, G. V., Maass, A., and Siegel, A. (2018). Traceability, reproducibility and wiki-exploration for “à-la-carte” reconstructions of genome-scale metabolic models. *PLoS Comput. Biol.*, 14(5):e1006146.

Akram, M. (2014). Citric acid cycle and role of its intermediates in metabolism. *Cell Biochem. Biophys.*, 68(3):475–478.

Albert, M. J., Mathan, V. I., and Baker, S. J. (1980). Vitamin B12 synthesis by human small intestinal bacteria. *Nature*, 283(5749):781–782.

Alekshun, M. N. and Levy, S. B. (2007). Molecular mechanisms of antibacterial multidrug resistance. *Cell*, 128(6):1037–1050.

Alkhalaf, L. M. and Ryan, K. S. (2015). Biosynthetic manipulation of tryptophan in bacteria: pathways and mechanisms. *Chem. Biol.*, 22(3):317–328.

Almeida, C., Oliveira, R., Soares, R., and Barata, P. (2020). Influence of gut microbiota dysbiosis on brain function: a systematic review. *Porto Biomed J*, 5(2).

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

Altman, T., Travers, M., Kothari, A., Caspi, R., and Karp, P. D. (2013). A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, 14:112.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410.

Amend, A. S., Oliver, T. A., Amaral-Zettler, L. A., Boetius, A., Fuhrman, J. A., Horner-Devine, M. C., Huse, S. M., Welch, D. B. M., Martiny, A. C., Ramette, A., Zinger, L., Sogin, M. L., and Martiny, J. B. H. (2013). Macroecological patterns of marine bacteria on a global scale. *J. Biogeogr.*, 40(4):800–811.

An, D., Na, C., Bielawski, J., Hannun, Y. A., and Kasper, D. L. (2011). Membrane sphingolipids as essential molecular signals for *Bacteroides* survival in the intestine. *Proc. Natl. Acad. Sci. U. S. A.*, 108 Suppl 1(Suppl 1):4666–4671.

Ananthakrishnan, A. N., Luo, C., Yajnik, V., Khalili, H., Garber, J. J., Stevens, B. W., Cleland, T., and Xavier, R. J. (2017). Gut microbiome function predicts response to anti-integrin biologic therapy in inflammatory bowel diseases. *Cell Host Microbe*, 21(5):603–610.e3.

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., and Ogata, H. (2020). KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252.

Arboleya, S., Watkins, C., Stanton, C., and Ross, R. P. (2016). Gut bifidobacteria populations in human health and aging. *Front. Microbiol.*, 7:1204.

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., Dehal, P., Ware, D., Perez, F., Canon, S., and Others (2018a). KBase: the united states department of energy systems biology knowledgebase. *Nat. Biotechnol.*, 36(7):566–569.

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., Dehal, P., Ware, D., Perez, F., Canon, S., Sneddon, M. W., Henderson, M. L., Riehl, W. J., Murphy-Olson, D., Chan, S. Y., Kamimura, R. T., Kumari, S., Drake, M. M., Brettin, T. S., Glass, E. M., Chivian, D., Gunter, D., Weston, D. J., Allen, B. H., Baumohl, J., Best, A. A., Bowen, B., Brenner, S. E., Bun, C. C., Chandonia, J.-M., Chia, J.-M., Colasanti, R., Conrad, N., Davis, J. J., Davison, B. H., DeJongh, M., Devoid, S., Dietrich, E., Dubchak, I., Edirisinghe, J. N., Fang, G., Faria, J. P., Frybarger, P. M., Gerlach, W., Gerstein, M., Greiner, A., Gurtowski, J., Haun, H. L., He, F., Jain, R., Joachimiak, M. P., Keegan, K. P., Kondo, S., Kumar, V., Land, M. L., Meyer, F., Mills, M., Novichkov, P. S., Oh, T., Olsen, G. J., Olson, R., Parrello, B., Pasternak, S., Pearson, E., Poon, S. S., Price, G. A., Ramakrishnan, S., Ranjan, P., Ronald, P. C., Schatz, M. C., Seaver, S. M. D., Shukla, M., Sutormin, R. A., Syed, M. H., Thomason, J., Tintle, N. L., Wang, D., Xia, F., Yoo, H., Yoo, S., and Yu, D. (2018b). KBase: The united states department of energy systems biology knowledgebase. *Nat. Biotechnol.*, 36(7):566–569.

Armour, C. R., Nayfach, S., Pollard, K. S., and Sharpton, T. J. (2019). A metagenomic meta-analysis reveals functional signatures of health and disease in the human gut microbiome. *mSystems*, 4(4).

Arrieta, M.-C., Walter, J., and Finlay, B. B. (2016). Human Microbiota-Associated mice: A model with challenges. *Cell Host Microbe*, 19(5):575–578.

Arrigo, K. R. (2005). Marine microorganisms and global nutrient cycles. *Nature*, 437(7057):349–355.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., de Vos, W. M., Brunak, S., Doré, J., MetaHIT Consortium, Antolín, M., Artiguenave, F., Blottiere, H. M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K. U., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Huber, W., van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Mérieux, A., Melo Minardi, R., M'rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S. D., and Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180.

Ashton, J. J., Colquhoun, C. M., Cleary, D. W., Coelho, T., Haggarty, R., Mulder, I., Batra, A., Afzal, N. A., Beattie, R. M., Scott, K. P., and Ennis, S. (2017). 16S sequencing and functional analysis of the fecal microbiome during treatment of newly diagnosed pediatric inflammatory bowel disease. *Medicine*, 96(26):e7347.

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

Bäckhed, F., Ding, H., Wang, T., Hooper, L. V., Koh, G. Y., Nagy, A., Semenkovich, C. F., and Gordon, J. I. (2004). The gut microbiota as an environmental factor that regulates fat storage. *Proc. Natl. Acad. Sci. U. S. A.*, 101(44):15718–15723.

Bäckhed, F., Fraser, C. M., Ringel, Y., Sanders, M. E., Sartor, R. B., Sherman, P. M., Versalovic, J., Young, V., and Finlay, B. B. (2012). Defining a healthy human gut microbiome: current concepts, future directions, and clinical applications. *Cell Host Microbe*, 12(5):611–622.

Bain, C. C. and Cerovic, V. (2020). Interactions of the microbiota with the mucosal immune system. *Clin. Exp. Immunol.*, 199(1):9–11.

Bansal, T., Alaniz, R. C., Wood, T. K., and Jayaraman, A. (2010). The bacterial signal indole increases epithelial-cell tight-junction resistance and attenuates indicators of inflammation. *Proc. Natl. Acad. Sci. U. S. A.*, 107(1):228–233.

Bassil, A. K., Bourdu, S., Townson, K. A., Wheeldon, A., Jarvie, E. M., Zebda, N., Abuin, A., Grau, E., Livi, G. P., Punter, L., Latcham, J., Grimes, A. M., Hurp, D. P., Downham, K. M., Sanger, G. J., Winchester, W. J., Morrison, A. D., and Moore, G. B. T. (2009). UDP-glucose modulates gastric function through P2Y14 receptor-dependent and -independent mechanisms. *Am. J. Physiol. Gastrointest. Liver Physiol.*, 296(4):G923–30.

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

Baumgart, D. C. and Carding, S. R. (2007). Inflammatory bowel disease: cause and immunobiology. *Lancet*, 369(9573):1627–1640.

Becken, B., Davey, L., Middleton, D. R., Mueller, K. D., Sharma, A., Holmes, Z. C., Dallow, E., Remick, B., Barton, G. M., David, L. A., McCann, J. R., Armstrong, S. C., Malkus, P., and Valdivia, R. H. (2021). Genotypic and phenotypic diversity among human isolates of akkermansia muciniphila. *MBio*, 12(3).

Beghini, F., McIver, L. J., Blanco-Míguez, A., Dubois, L., Asnicar, F., Maharjan, S., Mailyan, A., Manghi, P., Scholz, M., Thomas, A. M., Valles-Colomer, M., Weingart, G., Zhang, Y., Zolfo, M., Huttenhower, C., Franzosa, E. A., and Segata, N. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *Elife*, 10.

Belkaid, Y. and Hand, T. W. (2014). Role of the microbiota in immunity and inflammation. *Cell*, 157(1):121–141.

Benavides, M., Moisander, P. H., Daley, M. C., Bode, A., and Arístegui, J. (2016). Longitudinal variability of diazotroph abundances in the subtropical north atlantic ocean. *J. Plankton Res.*, 38(3):662–672.

Benitez-Nelson, C. R. (2000). The biogeochemical cycling of phosphorus in marine systems. *Earth-Sci. Rev.*, 51(1):109–135.

Bentzon-Tilia, M., Severin, I., Hansen, L. H., and Riemann, L. (2015). Genomics and ecophysiology of heterotrophic Nitrogen-Fixing bacteria isolated from estuarine surface water. *MBio*, 6(4):e00929.

Bergström, A., Skov, T. H., Bahl, M. I., Roager, H. M., Christensen, L. B., Ejlerskov, K. T., Mølgaard, C., Michaelsen, K. F., and Licht, T. R. (2014). Establishment of intestinal microbiota during early life: a longitudinal, explorative study of a large cohort of danish infants. *Appl. Environ. Microbiol.*, 80(9):2889–2900.

Beura, S., Kundu, P., Das, A. K., and Ghosh, A. (2022). Metagenome-scale community metabolic modelling for understanding the role of gut microbiota in human health. *Comput. Biol. Med.*, 149:105997.

Bi, J. and Wang, Y.-F. (2020). The effect of the endosymbiont wolbachia on the behavior of insect hosts. *Insect Sci.*, 27(5):846–858.

Bielka, W., Przezak, A., and Pawlik, A. (2022). The role of the gut microbiota in the pathogenesis of diabetes. *Int. J. Mol. Sci.*, 23(1). pre-print.

Biesalski, H. K. (2016). Nutrition meets the microbiome: micronutrients and the microbiota. *Ann. N. Y. Acad. Sci.*, 1372(1):53–64.

Bishop, P. E., Jarlenski, D. M., and Hetherington, D. R. (1980). Evidence for an alternative nitrogen fixation system in azotobacter vinelandii. *Proc. Natl. Acad. Sci. U. S. A.*, 77(12):7342–7346.

Blachier, F., Beaumont, M., and Kim, E. (2019). Cysteine-derived hydrogen sulfide and gut health: a matter of endogenous or bacterial origin. *Curr. Opin. Clin. Nutr. Metab. Care*, 22(1):68–75.

Blekhman, R., Goodrich, J. K., Huang, K., Sun, Q., Bukowski, R., Bell, J. T., Spector, T. D., Keinan, A., Ley, R. E., Gevers, D., and Clark, A. G. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.*, 16(1):191.

Bombar, D., Paerl, R. W., and Riemann, L. (2016). Marine Non-Cyanobacterial diazotrophs: Moving beyond molecular detection. *Trends Microbiol.*, 24(11):916–927.

Bordenstein, S. R. and Theis, K. R. (2015). Host biology in light of the microbiome: Ten principles of holobionts and hologenomes. *PLoS Biol.*, 13(8):e1002226.

Borren, N. Z., Plichta, D., Joshi, A. D., Bonilla, G., Peng, V., Colizzo, F. P., Luther, J., Khalili, H., Garber, J. J., Janneke van der Woude, C., Sadreyev, R., Vlamakis, H., Xavier, R. J., and Ananthakrishnan, A. N. (2021). Alterations in fecal microbiomes and serum metabolomes of fatigued patients with quiescent inflammatory bowel diseases. *Clin. Gastroenterol. Hepatol.*, 19(3):519–527.e5.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., Tringe, S. G., Ivanova, N. N., Copeland, A., Clum, A., Becraft, E. D., Malmstrom, R. R., Birren, B., Podar, M., Bork, P., Weinstock, G. M., Garrity, G. M., Dodsworth, J. A., Yooseph, S., Sutton, G., Glöckner, F. O., Gilbert, J. A., Nelson, W. C., Hallam, S. J., Jungbluth, S. P., Ettema, T. J. G., Tighe, S., Konstantinidis, K. T., Liu, W.-T., Baker, B. J., Rattei, T., Eisen, J. A., Hedlund, B., McMahon, K. D., Fierer, N., Knight, R., Finn, R., Cochrane, G., Karsch-Mizrachi, I., Tyson, G. W., Rinke, C., Genome Standards Consortium, Lapidus, A., Meyer, F., Yilmaz, P., Parks, D. H., Eren, A. M., Schriml, L., Banfield, J. F., Hugenholtz, P., and Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.*, 35(8):725–731.

Boyd, P. W. (2015). Toward quantifying the response of the oceans' biological pump to climate change. *Front. Mar. Sci.*, 2.

Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., and Sherlock, G. (2004). GO::TermFinder–open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715.

Breusing, C., Klobusnik, N. H., Hauer, M. A., and Beinart, R. A. (2022). Genome assembly of the chemosynthetic endosymbiont of the hydrothermal vent snail alviniconcha adamantis from the mariana arc. *G3*, 12(10).

Brito, I. L., Yilmaz, S., Huang, K., Xu, L., Jupiter, S. D., Jenkins, A. P., Naisilisili, W., Tamminen, M., Smillie, C. S., Wortman, J. R., Birren, B. W., Xavier, R. J., Blainey, P. C., Singh, A. K., Gevers, D., and Alm, E. J. (2016). Mobile genes in the human microbiome are structured from global to individual scales. *Nature*, 535(7612):435–439.

Brown, E. M., Clardy, J., and Xavier, R. J. (2023). Gut microbiome lipid metabolism and its impact on host physiology. *Cell Host Microbe*, 31(2):173–186.

Brüssow, H. (2020). Problems with the concept of gut microbiota dysbiosis. *Microb. Biotechnol.*, 13(2):423–434.

Bryant, D. A., Neil Hunter, C., and Warren, M. J. (2020). Biosynthesis of the modified tetrapyrroles—the pigments of life. *J. Biol. Chem.*, 295(20):6888–6925.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, 12(1):59–60.

Busi, S. B., Bourquin, M., Fodelianakis, S., Michoud, G., Kohler, T. J., Peter, H., Pramateftaki, P., Styllas, M., Tolosano, M., De Staercke, V., Schön, M., de Nies, L., Marasco, R., Daffonchio, D., Ezzat, L., Wilmes, P., and Battin, T. J. (2022). Genomic and metabolic adaptations of biofilms to ecological windows of opportunity in glacier-fed streams. *Nat. Commun.*, 13(1):2168.

Byndloss, M. X. and Bäumler, A. J. (2018). The germ-organ theory of non-communicable diseases. *Nat. Rev. Microbiol.*, 16(2):103–110.

Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X., and Greene, C. S. (2020). Responsible, practical genomic data sharing that accelerates research. *Nat. Rev. Genet.*, 21(10):615–629.

Callahan, A., Winnenburg, R., and Shah, N. H. (2018). U-Index, a dataset and an impact metric for informatics tools and databases. *Sci Data*, 5:180043.

Campbell, J. H., O'Donoghue, P., Campbell, A. G., Schwientek, P., Sczyrba, A., Woyke, T., Söll, D., and Podar, M. (2013). UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc. Natl. Acad. Sci. U. S. A.*, 110(14):5540–5545.

Campo, L., Eiseler, S., Apfel, T., and Pyrsopoulos, N. (2019). Fatty liver disease and gut microbiota: A comprehensive update. *J Clin Transl Hepatol*, 7(1):56–60.

Cani, P. D. (2018). Human gut microbiome: hopes, threats and promises. *Gut*, 67(9):1716–1725.

Cao, S., Zhang, W., Ding, W., Wang, M., Fan, S., Yang, B., Mcminn, A., Wang, M., Xie, B.-B., Qin, Q.-L., Chen, X.-L., He, J., and Zhang, Y.-Z. (2020). Structure and function of the arctic and antarctic marine microbiota as revealed by metagenomics. *Microbiome*, 8(1):47.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973.

Capone, D. G., Zehr, J. P., Paerl, H. W., Bergman, B., and Carpenter, E. J. (1997). Trichodesmium , a globally significant marine cyanobacterium. *Science*, 276(5316):1221–1229.

Caporaso, J. G., Lauber, C. L., Costello, E. K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N., Gordon, J. I., and Knight, R. (2011). Moving pictures of the human microbiome. *Genome Biol.*, 12(5):R50.

Carpenter, E. J., Capone, D. G., and Reuter, J. G. (1992). *Marine Pelagic Cyanobacteria: Trichodesmium and other Diazotrophs*. Springer Science & Business Media.

Carpenter, E. J. and Romans, K. (1991). Major role of the cyanobacterium trichodesmium in nutrient cycling in the north atlantic ocean. *Science*, 254(5036):1356–1358.

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Tara Oceans Coordinators, Jaillon, O., Aury, J.-M., Karsenti, E., Sullivan, M. B., Sunagawa, S., Bork, P., Not, F., Hingamp, P., Raes, J., Guidi, L., Ogata, H., de Vargas, C., Iudicone, D., Bowler, C., and Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nat. Commun.*, 9(1):373.

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Weerasinghe, D., Zhang, P., and Karp, P. D. (2014). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome databases. *Nucleic Acids Res.*, 42(Database issue):D459–71.

Castro-Severyn, J., Pardo-Esté, C., Mendez, K. N., Fortt, J., Marquez, S., Molina, F., Castro-Nallar, E., Remonsellez, F., and Saavedra, C. P. (2021). Living to the high extreme: Unraveling the composition, structure, and functional insights of bacterial communities thriving in the Arsenic-Rich salar de huasco altiplanic ecosystem. *Microbiol Spectr*, 9(1):e0044421.

Cevallos, S. A., Lee, J.-Y., Tiffany, C. R., Byndloss, A. J., Johnston, L., Byndloss, M. X., and Bäumler, A. J. (2019). Increased epithelial oxygenation links colitis to an expansion of tumorigenic bacteria. *MBio*, 10(5).

Chan, P. P. and Lowe, T. M. (2019). tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol. Biol.*, 1962:1–14.

Chaparro, J. M., Sheflin, A. M., Manter, D. K., and Vivanco, J. M. (2012). Manipulating the soil microbiome to increase soil health and plant fertility. *Biol. Fertil. Soils*, 48(5):489–499.

Charlson, R. J., Lovelock, J. E., Andreae, M. O., and Warren, S. G. (1987). Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature*, 326(6114):655–661.

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2019). GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*.

Chen, L.-X., Anantharaman, K., Shaiber, A., Eren, A. M., and Banfield, J. F. (2020). Accurate and complete genomes from metagenomes. *Genome Res.*, 30(3):315–333.

Chen, Y., Li, D., Dai, Z., Piao, X., Wu, Z., Wang, B., Zhu, Y., and Zeng, Z. (2014). L-methionine supplementation maintains the integrity and barrier function of the small-intestinal mucosa in post-weaning piglets. *Amino Acids*, 46(4):1131–1142.

Cheung, S., Zehr, J. P., Xia, X., Tsurumoto, C., Endo, H., Nakaoka, S.-I., Mak, W., Suzuki, K., and Liu, H. (2021). Gamma4: a genetically versatile gammaproteobacterial nifh phylotype that is widely distributed in the north pacific ocean. *Environ. Microbiol.*, 23(8):4246–4259.

Choudoir, M. J., Narayanan, A., Rodriguez-Ramos, D., Simoes, R., Efroni, A., Sondrini, A., and DeAngelis, K. M. (2023). Pangenomes reveal genomic signatures of microbial adaptation to chronic soil warming. pre-print.

Chow, J., Tang, H., and Mazmanian, S. K. (2011). Pathobionts of the gastrointestinal microbiota and inflammatory disease. *Curr. Opin. Immunol.*, 23(4):473–480.

Christodoulou, D., Link, H., Fuhrer, T., Kochanowski, K., Gerosa, L., and Sauer, U. (2018). Reserve flux capacity in the pentose phosphate pathway enables escherichia coli's rapid response to oxidative stress. *Cell Syst*, 6(5):569–578.e7.

Church, M. J., Björkman, K. M., Karl, D. M., Saito, M. A., and Zehr, J. P. (2008). Regional distributions of nitrogen-fixing bacteria in the pacific ocean. *Limnol. Oceanogr.*, 53(1):63–77.

Church, M. J., Short, C. M., Jenkins, B. D., Karl, D. M., and Zehr, J. P. (2005). Temporal patterns of nitrogenase gene (nifh) expression in the oligotrophic north pacific ocean. *Appl. Environ. Microbiol.*, 71(9):5362–5370.

Claesson, M. J., Cusack, S., O'Sullivan, O., Greene-Diniz, R., de Weerd, H., Flannery, E., Marchesi, J. R., Falush, D., Dinan, T., Fitzgerald, G., Stanton, C., van Sinderen, D., O'Connor, M., Harnedy, N., O'Connor, K., Henry, C., O'Mahony, D., Fitzgerald, A. P., Shanahan, F., Twomey, C., Hill, C., Ross, R. P., and O'Toole, P. W. (2011). Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Natl. Acad. Sci. U. S. A.*, 108 Suppl 1(Suppl 1):4586–4591.

Claud, E. C., Keegan, K. P., Brulc, J. M., Lu, L., Bartels, D., Glass, E., Chang, E. B., Meyer, F., and Antonopoulos, D. A. (2013). Bacterial community structure and functional contributions to emergence of health or necrotizing enterocolitis in preterm infants. *Microbiome*, 1(1):20.

Clausen, D. S. and Willis, A. D. (2022). Evaluating replicability in microbiome data. *Biostatistics*, 23(4):1099–1114.

Clemente, J. C., Ursell, L. K., Parfrey, L. W., and Knight, R. (2012). The impact of the gut microbiota on human health: an integrative view. *Cell*, 148(6):1258–1270.

Clooney, A. G., Eckenberger, J., Laserna-Mendieta, E., Sexton, K. A., Bernstein, M. T., Vagianos, K., Sargent, M., Ryan, F. J., Moran, C., Sheehan, D., Sleator, R. D., Targownik, L. E., Bernstein, C. N., Shanahan, F., and Claesson, M. J. (2021). Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. *Gut*, 70(3):499–510.

Coburn, L. A., Gong, X., Singh, K., Asim, M., Scull, B. P., Allaman, M. M., Williams, C. S., Rosen, M. J., Washington, M. K., Barry, D. P., Piazuelo, M. B., Casero, Jr, R. A., Chaturvedi, R., Zhao, Z., and Wilson, K. T. (2012). L-arginine supplementation improves responses to injury and inflammation in dextran sulfate sodium colitis. *PLoS One*, 7(3):e33546.

Collins, S. M., Surette, M., and Bercik, P. (2012). The interplay between the intestinal microbiota and the brain. *Nat. Rev. Microbiol.*, 10(11):735–742.

Constante, M., Fragoso, G., Calvé, A., Samba-Mondonga, M., and Santos, M. M. (2017). Dietary heme induces gut dysbiosis, aggravates colitis, and potentiates the development of adenomas in mice. *Front. Microbiol.*, 8:1809.

Cornejo-Castillo, F. M., Cabello, A. M., Salazar, G., Sánchez-Baracaldo, P., Lima-Mendez, G., Hingamp, P., Alberti, A., Sunagawa, S., Bork, P., de Vargas, C., Raes, J., Bowler, C., Wincker, P., Zehr, J. P., Gasol, J. M., Massana, R., and Acinas, S. G. (2016). Cyanobacterial symbionts diverged in the late cretaceous towards lineage-specific nitrogen fixation factories in single-celled phytoplankton. *Nat. Commun.*, 7:11071.

Cornejo-Castillo, F. M. and Zehr, J. P. (2021). Intriguing size distribution of the uncultured and globally widespread marine non-cyanobacterial diazotroph Gamma-A. *ISME J.*, 15(1):124–128.

Corteselli, E. M., Aitken, M. D., and Singleton, D. R. (2017). Description of immundisolibacter cernigliae gen. nov., sp. nov., a high-molecular-weight polycyclic aromatic hydrocarbon-degrading bacterium within the class gammaproteobacteria, and proposal of immundisolibacterales ord. nov. and immundisolibacteraceae fam. nov. *Int. J. Syst. Evol. Microbiol.*, 67(4):925–931.

Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. M., and Relman, D. A. (2012). The application of ecological theory toward an understanding of the human microbiome. *Science*, 336(6086):1255–1262.

Coyte, K. Z., Schluter, J., and Foster, K. R. (2015). The ecology of the microbiome: Networks, competition, and stability. *Science*, 350(6261):663–666.

Crowther, T. W., van den Hoogen, J., Wan, J., Mayes, M. A., Keiser, A. D., Mo, L., Averill, C., and Maynard, D. S. (2019). The global soil community and its influence on biogeochemistry. *Science*, 365(6455).

Cui, L., Zhao, T., Hu, H., Zhang, W., and Hua, X. (2017). Association study of gut flora in coronary heart disease through High-Throughput sequencing. *Biomed Res. Int.*, 2017:3796359.

Darfeuille-Michaud, A., Boudeau, J., Bulois, P., Neut, C., Glasser, A.-L., Barnich, N., Bringer, M.-A., Swidsinski, A., Beaugerie, L., and Colombel, J.-F. (2004). High prevalence of adherent-invasive escherichia coli associated with ileal mucosa in crohn's disease. *Gastroenterology*, 127(2):412–421.

Darling, A. E., Jospin, G., Lowe, E., Matsen, 4th, F. A., Bik, H. M., and Eisen, J. A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, 2:e243.

Davenport, M., Poles, J., Leung, J. M., Wolff, M. J., Abidi, W. M., Ullman, T., Mayer, L., Cho, I., and Loke, P. (2014). Metabolic alterations to the mucosal microbiota in inflammatory bowel disease. *Inflamm. Bowel Dis.*, 20(4):723–731.

David, L. A., Materna, A. C., Friedman, J., Campos-Baptista, M. I., Blackburn, M. C., Perrotta, A., Erdman, S. E., and Alm, E. J. (2014a). Host lifestyle affects human microbiota on daily timescales. *Genome Biol.*, 15(7):R89.

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., Biddinger, S. B., Dutton, R. J., and Turnbaugh, P. J. (2014b). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559–563.

David, L. A., Weil, A., Ryan, E. T., Calderwood, S. B., Harris, J. B., Chowdhury, F., Begum, Y., Qadri, F., LaRocque, R. C., and Turnbaugh, P. J. (2015). Gut microbial succession follows acute secretory diarrhea in humans. *MBio*, 6(3):e00381–15.

Davis, R. E. and Moyer, C. L. (2008). Extreme spatial and temporal variability of hydrothermal microbial mat communities along the mariana island arc and southern mariana back-arc system. *J. Geophys. Res.*, 113(B8).

de Faria, M. R., Costa, L. S. A. S., Chiaramonte, J. B., Bettiol, W., and Mendes, R. (2021). The rhizosphere microbiome: functions, dynamics, and role in plant protection. *Trop. Plant Pathol.*, 46(1):13–25.

de la Cuesta-Zuluaga, J., Ley, R. E., and Youngblut, N. D. (2020). Struo: a pipeline for building custom databases for common metagenome profilers. *Bioinformatics*, 36(7):2314–2315.

De Preter, V., Bulteel, V., Suenaert, P., Geboes, K. P., De Hertogh, G., Luypaerts, A., Geboes, K., Verbeke, K., and Rutgeerts, P. (2009). Pouchitis, similar to active ulcerative colitis, is associated with impaired butyrate oxidation by intestinal mucosa. *Inflamm. Bowel Dis.*, 15(3):335–340.

De Preter, V., Machiels, K., Joossens, M., Arijs, I., Matthys, C., Vermeire, S., Rutgeerts, P., and Verbeke, K. (2015). Faecal metabolite profiling identifies medium-chain fatty acids as discriminating compounds in IBD. *Gut*, 64(3):447–458.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horák, A., Jaillon, O., Lima-Mendez, G., Lukeš, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans Coordinators, Acinas, S. G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M. E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., and Karsenti, E. (2015). Ocean plankton. eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237):1261605.

Degnan, P. H., Barry, N. A., Mok, K. C., Taga, M. E., and Goodman, A. L. (2014a). Human gut microbes use multiple transporters to distinguish vitamin b12 analogs and compete in the gut. *Cell Host Microbe*, 15(1):47–57.

Degnan, P. H., Taga, M. E., and Goodman, A. L. (2014b). Vitamin B12 as a modulator of gut microbial ecology. *Cell Metab.*, 20(5):769–778.

DeJongh, M., Formsma, K., Boillot, P., Gould, J., Rycenga, M., and Best, A. (2007). Toward the automated generation of genome-scale metabolic networks in the SEED. *BMC Bioinformatics*, 8:139.

Delcher, A. L., Phillippy, A., Carlton, J., and Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.*, 30(11):2478–2483.

Delmont, T. O. (2021). Discovery of nondiazotrophic *Trichodesmium* species abundant and widespread in the open ocean. *Proc. Natl. Acad. Sci. U. S. A.*, 118(46).

Delmont, T. O. and Eren, A. M. (2016). Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ*, 4:e1839.

Delmont, T. O., Gaia, M., Hinsinger, D. D., Frémont, P., Vanni, C., Fernandez-Guerra, A., Eren, A. M., Kourlaiev, A., d'Agata, L., Clayssen, Q., Villar, E., Labadie, K., Cruaud, C., Poulain, J., Da Silva, C., Wessner, M., Noel, B., Aury, J.-M., Tara Oceans Coordinators, de Vargas, C., Bowler, C., Karsenti, E., Pelletier, E., Wincker, P., and Jaillon, O. (2022). Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genom*, 2(5):100123.

Delmont, T. O., Karlusich, J. J. P., Veseli, I., Fuessel, J., Murat Eren, A., Foster, R. A., Bowler, C., Wincker, P., and Pelletier, E. (2021). Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. *ISME Journal*, pages 927—936.

Delmont, T. O., Quince, C., Shaiber, A., Esen, Ö. C., Lee, S. T., Rappé, M. S., McLellan, S. L., Lücker, S., and Eren, A. M. (2018). Nitrogen-fixing populations of planctomycetes and proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol*, 3(7):804–813.

Devkota, S., Wang, Y., Musch, M. W., Leone, V., Fehlner-Peach, H., Nadimpalli, A., Antonopoulos, D. A., Jabri, B., and Chang, E. B. (2012). Dietary-fat-induced taurocholic acid promotes pathobiont expansion and colitis in il10-/- mice. *Nature*, 487(7405):104–108.

Dhakan, D. B., Maji, A., Sharma, A. K., Saxena, R., Pulikkan, J., Grace, T., Gomez, A., Scaria, J., Amato, K. R., and Sharma, V. K. (2019). The unique composition of indian gut microbiome, gene catalogue, and associated fecal metabolome deciphered using multi-omics approaches. *Gigascience*, 8(3).

Dias, O., Rocha, M., Ferreira, E. C., and Rocha, I. (2015). Reconstructing genome-scale metabolic models with merlin. *Nucleic Acids Res.*, 43(8):3899–3910.

Diener, C., Gibbons, S. M., and Resendis-Antonio, O. (2020). MICOM: Metagenome-Scale modeling to infer metabolic interactions in the gut microbiota. *mSystems*, 5(1).

Dinan, K. and Dinan, T. G. (2022). Gut microbes and neuropathology: Is there a causal nexus? *Pathogens*, 11(7).

Ding, R.-X., Goh, W.-R., Wu, R.-N., Yue, X.-Q., Luo, X., Khine, W. W. T., Wu, J.-R., and Lee, Y.-K. (2019). Revisit gut microbiota and its impact on human health and disease. *J. Food Drug Anal.*, 27(3):623–631.

Dominik, M., Nzweundji, J. G., Ahmed, N., Carnicelli, S., Mat Jalaluddin, N. S., Fernandez Rivas, D., Narita, V., Enany, S., and Rios Rojas, C. (2022). Open science – for whom? *Data Sci. J.*, 21.

Donaldson, G. P., Lee, S. M., and Mazmanian, S. K. (2016). Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.*, 14(1):20–32.

Donohoe, D. R., Garge, N., Zhang, X., Sun, W., O'Connell, T. M., Bunger, M. K., and Bultman, S. J. (2011). The microbiome and butyrate regulate energy metabolism and autophagy in the mammalian colon. *Cell Metab.*, 13(5):517–526.

Doolittle, W. F. and Booth, A. (2017). It's the song, not the singer: an exploration of holobiosis and evolutionary theory. *Biol. Philos.*, 32(1):5–24.

Dos Santos, P. C., Fang, Z., Mason, S. W., Setubal, J. C., and Dixon, R. (2012). Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes.

Drula, E., Garron, M.-L., Dogan, S., Lombard, V., Henrissat, B., and Terrapon, N. (2022). The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.*, 50(D1):D571–D577.

D'Souza, G., Shitut, S., Preussger, D., Yousif, G., Waschina, S., and Kost, C. (2018). Ecology and evolution of metabolic cross-feeding interactions in bacteria. *Nat. Prod. Rep.*, 35(5):455–488.

Duhamel, S., Diaz, J. M., Adams, J. C., Djaoudi, K., Steck, V., and Waggoner, E. M. (2021). Phosphorus as an integral component of global marine biogeochemistry. *Nat. Geosci.*, 14(6):359–368.

Durack, J. and Lynch, S. V. (2019). The gut microbiome: Relationships with disease and opportunities for therapy. *J. Exp. Med.*, 216(1):20–40.

Durham, B. P., Dearth, S. P., Sharma, S., Amin, S. A., Smith, C. B., Campagna, S. R., Armbrust, E. V., and Moran, M. A. (2017). Recognition cascade and metabolite transfer in a marine bacteria-phytoplankton model system. *Environ. Microbiol.*, 19(9):3500–3513.

Durham, B. P., Sharma, S., Luo, H., Smith, C. B., Amin, S. A., Bender, S. J., Dearth, S. P., Van Mooy, B. A. S., Campagna, S. R., Kujawinski, E. B., Armbrust, E. V., and Moran, M. A. (2015). Cryptic carbon and sulfur cycling between surface ocean plankton. *Proc. Natl. Acad. Sci. U. S. A.*, 112(2):453–457.

Dyhrman, S. T., Chappell, P. D., Haley, S. T., Moffett, J. W., Orchard, E. D., Waterbury, J. B., and Webb, E. A. (2006). Phosphonate utilization by the globally important marine diazotroph trichodesmium. *Nature*, 439(7072):68–71.

Eberhard, F. E., Klimpel, S., Guarneri, A. A., and Tobias, N. J. (2022). Exposure to trypanosoma parasites induces changes in the microbiome of the chagas disease vector rhodnius prolixus. *Microbiome*, 10(1):45.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comput. Biol.*, 7(10):e1002195.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797.

Eiseman, B., Silen, W., Bascom, G. S., and Kauvar, A. J. (1958). Fecal enema as an adjunct in the treatment of pseudomembranous enterocolitis. *Surgery*, 44(5):854–859.

Eisenstein, M. (2020). The hunt for a healthy microbiome. *Nature*, 577(7792):S6–S8.

Enache, M., Popescu, G., Itoh, T., and Kamekura, M. (2012). Halophilic microorganisms from man-made and natural hypersaline environments: Physiology, ecology, and biotechnological potential. In Stan-Lotter, H. and Fendrihan, S., editors, *Adaption of Microbial Life to Environmental Extremes: Novel Research Results and Application*, pages 173–197. Springer Vienna, Vienna.

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

Eren, A. M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S. E., Schechter, M. S., Fink, I., Pan, J. N., Yousef, M., Fogarty, E. C., and Others (2021a). Community-led, integrated, reproducible multi-omics with anvi'o. *Nature microbiology*, 6(1):3–6.

Eren, A. M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S. E., Schechter, M. S., Fink, I., Pan, J. N., Yousef, M., Fogarty, E. C., Trigodet, F., Watson, A. R., Esen, Ö. C., Moore, R. M., Clayssen, Q., Lee, M. D., Kivenson, V., Graham, E. D., Merrill, B. D., Karkman, A., Blankenberg, D., Eppley, J. M., Sjödin, A., Scott, J. J., Vázquez-Campos, X., McKay, L. J., McDaniel, E. A., Stevens, S. L. R., Anderson, R. E., Fuessel, J., Fernandez-Guerra, A., Maignien, L., Delmont, T. O., and Willis, A. D. (2021b). Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol*, 6(1):3–6.

Eren, A. M. and Scott, J. J. (2020a). Visualizing the fate of contigs across metagenomic binning algorithms. `https://merenlab.org/2020/01/02/visualizing-metagenomic-bins/`. Accessed: 2023-3-22.

Eren, A. M. and Scott, J. J. (2020b). Visualizing the fate of contigs across metagenomic binning algorithms. `https://merenlab.org/2020/01/02/visualizing-metagenomic-bins/`. Accessed: 2023-4-3.

Eren, A. M., Vineis, J. H., Morrison, H. G., and Sogin, M. L. (2013). A filtering method to generate high quality short reads using illumina paired-end technology. *PLoS One*, 8(6):e66643.

Falkowski, P. G., Barber, R. T., and Smetacek, V, V. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science*, 281(5374):200–207.

Falkowski, P. G., Fenchel, T., and Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science*, 320(5879):1034–1039.

Falkowski, P. G. and Godfrey, L. V. (2008). Electrons, life and the evolution of earth's oxygen cycle. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 363(1504):2705–2716.

Fan, Y. and Pedersen, O. (2021). Gut microbiota in human metabolic health and disease. *Nat. Rev. Microbiol.*, 19(1):55–71.

Fang, X., Lloyd, C. J., and Palsson, B. O. (2020). Reconstructing organisms in silico: genome-scale models and their emerging applications. *Nat. Rev. Microbiol.*, 18(12):731–743.

Farag, I. F., Biddle, J. F., Zhao, R., Martino, A. J., House, C. H., and León-Zayas, R. I. (2020). Metabolic potentials of archaeal lineages resolved from metagenomes of deep costa rica sediments. *ISME J.*, 14(6):1345–1358.

Faria, J. P., Rocha, M., Rocha, I., and Henry, C. S. (2018). Methods for automated genome-scale metabolic model reconstruction. *Biochem. Soc. Trans.*, 46(4):931–936.

Farnelid, H., Andersson, A. F., Bertilsson, S., Al-Soud, W. A., Hansen, L. H., Sørensen, S., Steward, G. F., Hagström, Å., and Riemann, L. (2011). Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria. *PLoS One*, 6(4):e19223.

Farnelid, H., Tarangkoon, W., Hansen, G., Hansen, P. J., and Riemann, L. (2010). Putative n2-fixing heterotrophic bacteria associated with dinoflagellate–cyanobacteria consortia in the low-nitrogen indian ocean. *Aquat. Microb. Ecol.*, 61(2):105–117.

Farnelid, H., Turk-Kubo, K., Ploug, H., Ossolinski, J. E., Collins, J. R., Van Mooy, B. A. S., and Zehr, J. P. (2019). Diverse diazotrophs are present on sinking particles in the north pacific subtropical gyre. *ISME J.*, 13(1):170–182.

Faure, M., Mettraux, C., Moennoz, D., Godin, J.-P., Vuichoud, J., Rochat, F., Breuillé, D., Obled, C., and Corthésy-Theulaz, I. (2006). Specific amino acids increase mucin synthesis and microbiota in dextran sulfate sodium-treated rats. *J. Nutr.*, 136(6):1558–1564.

Faure, M., Moënnoz, D., Montigon, F., Mettraux, C., Breuillé, D., and Ballèvre, O. (2005). Dietary threonine restriction specifically reduces intestinal mucin synthesis in rats. *J. Nutr.*, 135(3):486–491.

Feng, L., Raman, A. S., Hibberd, M. C., Cheng, J., Griffin, N. W., Peng, Y., Leyn, S. A., Rodionov, D. A., Osterman, A. L., and Gordon, J. I. (2020). Identifying determinants of bacterial fitness in a model of human gut microbial succession. *Proc. Natl. Acad. Sci. U. S. A.*, 117(5):2622–2633.

Feng, Q., Liang, S., Jia, H., Stadlmayr, A., Tang, L., Lan, Z., Zhang, D., Xia, H., Xu, X., Jie, Z., Su, L., Li, X., Li, X., Li, J., Xiao, L., Huber-Schönauer, U., Niederseer, D., Xu, X., Al-Aama, J. Y., Yang, H., Wang, J., Kristiansen, K., Arumugam, M., Tilg, H., Datz, C., and Wang, J. (2015). Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat. Commun.*, 6:6528.

Ferrer-González, F. X., Widner, B., Holderman, N. R., Glushka, J., Edison, A. S., Kujawinski, E. B., and Moran, M. A. (2021). Resource partitioning of phytoplankton metabolites that support bacterial heterotrophy. *ISME J.*, 15(3):762–773.

Fierer, N. and Jackson, R. B. (2006). The diversity and biogeography of soil bacterial communities. *Proc. Natl. Acad. Sci. U. S. A.*, 103(3):626–631.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. pre-print.

Fong, A. A., Karl, D. M., Lukas, R., Letelier, R. M., Zehr, J. P., and Church, M. J. (2008). Nitrogen fixation in an anticyclonic eddy in the oligotrophic north pacific ocean. *ISME J.*, 2(6):663–676.

Foster, R. A., Carpenter, E. J., and Bergman, B. (2006). Unicellular cyanobionts in open ocean dinoflagellates, radiolarians, and tintinnids: Ultrastructural characterization and immunolocalization of phycoerythrin and nitrogenase1. *J. Phycol.*, 42(2):453–463.

Foster, R. A., Paytan, A., and Zehr, J. P. (2009). Seasonality of N2 fixation andnifhgene diversity in the gulf of aqaba (red sea). *Limnol. Oceanogr.*, 54(1):219–233.

Franzosa, E. A., Sirota-Madi, A., Avila-Pacheco, J., Fornelos, N., Haiser, H. J., Reinker, S., Vatanen, T., Hall, A. B., Mallick, H., McIver, L. J., Sauk, J. S., Wilson, R. G., Stevens, B. W., Scott, J. M., Pierce, K., Deik, A. A., Bullock, K., Imhann, F., Porter, J. A., Zhernakova, A., Fu, J., Weersma, R. K., Wijmenga, C., Clish, C. B., Vlamakis, H., Huttenhower, C., and Xavier, R. J. (2019). Gut microbiome structure and metabolic activity in inflammatory bowel disease. *Nat Microbiol*, 4(2):293–305.

Frye, J. G., Lindsey, R. L., Meinersmann, R. J., Berrang, M. E., Jackson, C. R., Englen, M. D., Turpin, J. B., and Fedorka-Cray, P. J. (2011). Related antimicrobial resistance genes detected in different bacterial species co-isolated from swine fecal samples. *Foodborne Pathog. Dis.*, 8(6):663–679.

Gaby, J. C. and Buckley, D. H. (2012). A comprehensive evaluation of PCR primers to amplify the nifh gene of nitrogenase. *PLoS One*, 7(7):e42149.

Galloway, J. N., Dentener, F. J., Capone, D. G., Boyer, E. W., Howarth, R. W., Seitzinger, S. P., Asner, G. P., Cleveland, C. C., Green, P. A., Holland, E. A., Karl, D. M., Michaels, A. F., Porter, J. H., Townsend, A. R., and Vöosmarty, C. J. (2004). Nitrogen cycles: Past, present, and future. *Biogeochemistry*, 70(2):153–226.

Galperin, M. Y., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2015). Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, 43(Database issue):D261–9.

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., and Koonin, E. V. (2021). COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.*, 49(D1):D274–D281.

Garschagen, L. S., Franke, T., and Deppenmeier, U. (2021). An alternative pentose phosphate pathway in human gut bacteria for the degradation of C5 sugars in dietary fibers. *FEBS J.*, 288(6):1839–1858.

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

Geisler, E., Bogler, A., Rahav, E., and Bar-Zeev, E. (2019). Direct detection of heterotrophic diazotrophs associated with planktonic aggregates. *Sci. Rep.*, 9(1):9288.

Geller-McGrath, D., Konwar, K. M., Edgcomb, V. P., Pachiadaki, M., Roddy, J. W., Wheeler, T. J., and McDermott, J. E. (2023). MetaPathPredict: A machine learning-based tool for predicting metabolic modules in incomplete bacterial genomes. pre-print.

Gensollen, T., Iyer, S. S., Kasper, D. L., and Blumberg, R. S. (2016). How colonization by microbiota in early life shapes the immune system. *Science*, 352(6285):539–544.

Gevers, D., Kugathasan, S., Denson, L. A., Vázquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S. J., Yassour, M., Morgan, X. C., Kostic, A. D., Luo, C., González, A., McDonald, D., Haberman, Y., Walters, T., Baker, S., Rosh, J., Stephens, M., Heyman, M., Markowitz, J., Baldassano, R., Griffiths, A., Sylvester, F., Mack, D., Kim, S., Crandall, W., Hyams, J., Huttenhower, C., Knight, R., and Xavier, R. J. (2014). The treatment-naive microbiome in new-onset crohn's disease. *Cell Host Microbe*, 15(3):382–392.

Giacomini, J. J., Torres-Morales, J., Dewhirst, F. E., Borisy, G. G., and Mark Welch, J. L. (2023). Site specialization of human oral veillonella species. *Microbiol Spectr*, 11(1):e0404222.

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., and Knight, R. (2018). Current understanding of the human microbiome. *Nat. Med.*, 24(4):392–400.

Gobert, A. P., Cheng, Y., Akhtar, M., Mersey, B. D., Blumberg, D. R., Cross, R. K., Chaturvedi, R., Drachenberg, C. B., Boucher, J.-L., Hacker, A., Casero, Jr, R. A., and Wilson, K. T. (2004). Protective role of arginase in a mouse model of colitis. *J. Immunol.*, 173(3):2109–2117.

Gogarten, J. P., Doolittle, W. F., and Lawrence, J. G. (2002). Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.*, 19(12):2226–2238.

Gojda, J. and Cahova, M. (2021). Gut microbiota as the link between elevated BCAA serum levels and insulin resistance. *Biomolecules*, 11(10).

Gómez, F., Furuya, K., and Takeda, S. (2005). Distribution of the cyanobacterium richelia intracellularis as an epiphyte of the diatom chaetoceros compressus in the western pacific ocean. *J. Plankton Res.*, 27(4):323–330.

Gonnella, G., Böhnke, S., Indenbirken, D., Garbe-Schönberg, D., Seifert, R., Mertens, C., Kurtz, S., and Perner, M. (2016). Endemic hydrothermal vent species identified in the open ocean seed bank. *Nat Microbiol*, 1(8):16086.

Goodman, A. L., McNulty, N. P., Zhao, Y., Leip, D., Mitra, R. D., Lozupone, C. A., Knight, R., and Gordon, J. I. (2009). Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host Microbe*, 6(3):279–289.

Goodrich, J. K., Waters, J. L., Poole, A. C., Sutter, J. L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J. T., Spector, T. D., Clark, A. G., and Ley, R. E. (2014). Human genetics shape the gut microbiome. *Cell*, 159(4):789–799.

Greenblum, S., Turnbaugh, P. J., and Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc. Natl. Acad. Sci. U. S. A.*, 109(2):594–599.

Grehan, M. J., Borody, T. J., Leis, S. M., Campbell, J., Mitchell, H., and Wettstein, A. (2010). Durable alteration of the colonic microbiota by the administration of donor fecal flora. *J. Clin. Gastroenterol.*, 44(8):551–561.

Gruss, A., Borezée-Durant, E., and Lechardeur, D. (2012). Chapter three - environmental heme utilization by Heme-Auxotrophic bacteria. In Poole, R. K., editor, *Advances in Microbial Physiology*, volume 61, pages 69–124. Academic Press.

Gu, C., Kim, G. B., Kim, W. J., Kim, H. U., and Lee, S. Y. (2019). Current status and applications of genome-scale metabolic models. *Genome Biol.*, 20(1):121.

Güell, M., Yus, E., Lluch-Senar, M., and Serrano, L. (2011). Bacterial transcriptomics: what is beyond the RNA horiz-ome? *Nat. Rev. Microbiol.*, 9(9):658–669.

Guslandi, M. (2022). Chapter 17 - probiotics and intestinal health. In Brandelli, A., editor, *Probiotics*, pages 343–353. Academic Press.

Gutiérrez, N. and Garrido, D. (2019). Species deletions from microbiome consortia reveal key metabolic interactions between gut microbes. *mSystems*, 4(4).

Haft, D. H., Selengut, J. D., and White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Res.*, 31(1):371–373.

Halfvarson, J., Brislawn, C. J., Lamendella, R., Vázquez-Baeza, Y., Walters, W. A., Bramer, L. M., D'Amato, M., Bonfiglio, F., McDonald, D., Gonzalez, A., McClure, E. E., Dunklebarger, M. F., Knight, R., and Jansson, J. K. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol*, 2:17004.

Hammer, T. J., Sanders, J. G., and Fierer, N. (2019). Not all animals need a microbiome. *FEMS Microbiol. Lett.*, 366(10).

Harding, K., Turk-Kubo, K. A., Sipler, R. E., Mills, M. M., Bronk, D. A., and Zehr, J. P. (2018). Symbiotic unicellular cyanobacteria fix nitrogen in the arctic ocean. *Proc. Natl. Acad. Sci. U. S. A.*, 115(52):13371–13375.

Hardivillé, S. and Hart, G. W. (2014). Nutrient regulation of signaling, transcription, and cell physiology by O-GlcNAcylation. *Cell Metab.*, 20(2):208–213.

Hasan, N. and Yang, H. (2019). Factors affecting the composition of the gut microbiota, and its modulation. *PeerJ*, 7:e7502.

Hashimoto, T., Perlot, T., Rehman, A., Trichereau, J., Ishiguro, H., Paolino, M., Sigl, V., Hanada, T., Hanada, R., Lipinski, S., Wild, B., Camargo, S. M. R., Singer, D., Richter, A., Kuba, K., Fukamizu, A., Schreiber, S., Clevers, H., Verrey, F., Rosenstiel, P., and Penninger, J. M. (2012). ACE2 links amino acid malnutrition to microbial ecology and intestinal inflammation. *Nature*, 487(7408):477–481.

Haug, K., Cochrane, K., Nainala, V. C., Williams, M., Chang, J., Jayaseelan, K. V., and O'Donovan, C. (2020). MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res.*, 48(D1):D440–D444.

Heinken, A., Hertel, J., and Thiele, I. (2021). Metabolic modelling reveals broad changes in gut microbial metabolism in inflammatory bowel disease patients with dysbiosis. *NPJ Syst Biol Appl*, 7(1):19.

Heiss, C. N. and Olofsson, L. E. (2018). Gut Microbiota-Dependent modulation of energy metabolism. *J. Innate Immun.*, 10(3):163–171.

Hellmann, J., Ta, A., Ollberding, N. J., Bezold, R., Lake, K., Jackson, K., Dirksing, K., Bonkowski, E., Haslam, D. B., and Denson, L. A. (2023). Patient-Reported outcomes correlate with microbial community composition independent of mucosal inflammation in pediatric inflammatory bowel disease. *Inflamm. Bowel Dis.*, 29(2):286–296.

Henke, M. T., Kenny, D. J., Cassilly, C. D., Vlamakis, H., Xavier, R. J., and Clardy, J. (2019). Ruminococcus gnavus, a member of the human gut microbiome associated with crohn's disease, produces an inflammatory polysaccharide. *Proc. Natl. Acad. Sci. U. S. A.*, 116(26):12672–12677.

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

Herrmann, K. M. and Weaver, L. M. (1999). THE SHIKIMATE PATHWAY. *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, 50:473–503.

Hijova, E. (2019). Gut bacterial metabolites of indigestible polysaccharides in intestinal fermentation as mediators of public health. *Bratisl. Lek. Listy*, 120(11):807–812.

Hill, C., Guarner, F., Reid, G., Gibson, G. R., Merenstein, D. J., Pot, B., Morelli, L., Canani, R. B., Flint, H. J., Salminen, S., Calder, P. C., and Sanders, M. E. (2014). Expert consensus document. the international scientific association for probiotics and prebiotics consensus statement on the scope and appropriate use of the term probiotic. *Nat. Rev. Gastroenterol. Hepatol.*, 11(8):506–514.

Hill, M. J. (1997). Intestinal flora and endogenous vitamin synthesis. *Eur. J. Cancer Prev.*, 6 Suppl 1:S43–5.

Hilton, J. A., Foster, R. A., Tripp, H. J., Carter, B. J., Zehr, J. P., and Villareal, T. A. (2013). Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont. *Nat. Commun.*, 4:1767.

Hoffman, K., Brownell, Z., Doyle, W. J., and Ochoa-Repáraz, J. (2022). The immunomodulatory roles of the gut microbiome in autoimmune diseases of the central nervous system: Multiple sclerosis as a model. *J. Autoimmun.*, page 102957.

Hooper, L. V., Midtvedt, T., and Gordon, J. I. (2002). How host-microbial interactions shape the nutrient environment of the mammalian intestine. *Annu. Rev. Nutr.*, 22:283–307.

Hossain, K. S., Amarasena, S., and Mayengbam, S. (2022). B vitamins and their roles in gut health. *Microorganisms*, 10(6).

Hove-Jensen, B., Andersen, K. R., Kilstrup, M., Martinussen, J., Switzer, R. L., and Willemoës, M. (2017). Phosphoribosyl diphosphate (PRPP): Biosynthesis, enzymology, utilization, and metabolic significance. *Microbiol. Mol. Biol. Rev.*, 81(1).

Howard, E. C., Henriksen, J. R., Buchan, A., Reisch, C. R., Bürgmann, H., Welsh, R., Ye, W., González, J. M., Mace, K., Joye, S. B., Kiene, R. P., Whitman, W. B., and Moran, M. A. (2006). Bacterial taxa that limit sulfur flux from the ocean. *Science*, 314(5799):649–652.

Huergo Luciano F. and Dixon Ray (2015). The emergence of 2-oxoglutarate as a master regulator metabolite. *Microbiol. Mol. Biol. Rev.*, 79(4):419–435.

Human Microbiome Project Consortium (2012a). A framework for human microbiome research. *Nature*, 486(7402):215–221.

Human Microbiome Project Consortium (2012b). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214.

Hunter (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9:90–95.

Hutchins, D. A. and Capone, D. G. (2022). The marine nitrogen cycle: new developments and global change. *Nat. Rev. Microbiol.*, 20(7):401–414.

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

Hylemon, P. B., Zhou, H., Pandak, W. M., Ren, S., Gil, G., and Dent, P. (2009). Bile acids as regulatory molecules. *J. Lipid Res.*, 50(8):1509–1520.

Ijssennagger, N., Belzer, C., Hooiveld, G. J., Dekker, J., van Mil, S. W. C., Müller, M., Kleerebezem, M., and van der Meer, R. (2015). Gut microbiota facilitates dietary heme-induced epithelial hyperproliferation by opening the mucus barrier in colon. *Proc. Natl. Acad. Sci. U. S. A.*, 112(32):10038–10043.

Inskeep, W. P., Jay, Z. J., Tringe, S. G., Herrgård, M. J., Rusch, D. B., and YNP Metagenome Project Steering Committee and Working Group Members (2013). The YNP metagenome project: Environmental parameters responsible for microbial distribution in the yellowstone geothermal ecosystem. *Front. Microbiol.*, 4:67.

Isaac, S., Scher, J. U., Djukovic, A., Jiménez, N., Littman, D. R., Abramson, S. B., Pamer, E. G., and Ubeda, C. (2017). Short- and long-term effects of oral vancomycin on the human intestinal microbiota. *J. Antimicrob. Chemother.*, 72(1):128–136.

Ivanov, I. I. and Littman, D. R. (2010). Segmented filamentous bacteria take the stage. *Mucosal Immunol.*, 3(3):209–212.

Jaffe, A. L., Castelle, C. J., Matheus Carnevali, P. B., Gribaldo, S., and Banfield, J. F. (2020). The rise of diversity in metabolic platforms across the candidate phyla radiation. *BMC Biol.*, 18(1):69.

Janssen, A. W. F. and Kersten, S. (2015). The role of the gut microbiota in metabolic health. *FASEB J.*, 29(8):3111–3123.

Jansson, J., Willing, B., Lucio, M., Fekete, A., Dicksved, J., Halfvarson, J., Tysk, C., and Schmitt-Kopplin, P. (2009). Metabolomics reveals metabolic biomarkers of crohn's disease. *PLoS One*, 4(7):e6386.

Jebbar, M., Franzetti, B., Girard, E., and Oger, P. (2015). Microbial diversity and adaptation to high hydrostatic pressure in deep-sea hydrothermal vents prokaryotes. *Extremophiles*, 19(4):721–740.

Jha, A. R., Davenport, E. R., Gautam, Y., Bhandari, D., Tandukar, S., Ng, K. M., Fragiadakis, G. K., Holmes, S., Gautam, G. P., Leach, J., Sherchand, J. B., Bustamante, C. D., and Sonnenburg, J. L. (2018). Gut microbiome transition across a lifestyle gradient in himalaya. *PLoS Biol.*, 16(11):e2005396.

Jie, Z., Xia, H., Zhong, S.-L., Feng, Q., Li, S., Liang, S., Zhong, H., Liu, Z., Gao, Y., Zhao, H., Zhang, D., Su, Z., Fang, Z., Lan, Z., Li, J., Xiao, L., Li, J., Li, R., Li, X., Li, F., Ren, H., Huang, Y., Peng, Y., Li, G., Wen, B., Dong, B., Chen, J.-Y., Geng, Q.-S., Zhang, Z.-W., Yang, H., Wang, J., Wang, J., Zhang, X., Madsen, L., Brix, S., Ning, G., Xu, X., Liu, X., Hou, Y., Jia, H., He, K., and Kristiansen, K. (2017). The gut microbiome in atherosclerotic cardiovascular disease. *Nat. Commun.*, 8(1):845.

Jochum, L. and Stecher, B. (2020). Label or concept - what is a pathobiont? *Trends Microbiol.*, 28(10):789–792.

Jodorkovsky, D., Young, Y., and Abreu, M. T. (2010). Clinical outcomes of patients with ulcerative colitis and co-existing clostridium difficile infection. *Dig. Dis. Sci.*, 55(2):415–420.

Joerger, R. D., Jacobson, M. R., Premakumar, R., Wolfinger, E. D., and Bishop, P. E. (1989). Nucleotide sequence and mutational analysis of the structural genes (anfHDGK) for the second alternative nitrogenase from azotobacter vinelandii. *J. Bacteriol.*, 171(2):1075–1086.

Johansson, M. E. V., Gustafsson, J. K., Sjöberg, K. E., Petersson, J., Holm, L., Sjövall, H., and Hansson, G. C. (2010). Bacteria penetrate the inner mucus layer before inflammation in the dextran sulfate colitis model. *PLoS One*, 5(8):e12238.

Johansson, M. E. V. and Hansson, G. C. (2016). Immunological aspects of intestinal mucus and mucins. *Nat. Rev. Immunol.*, 16(10):639–649.

Johnson, M. D., Scott, J. J., Leray, M., Lucey, N., Bravo, L. M. R., Wied, W. L., and Altieri, A. H. (2021). Rapid ecosystem-scale consequences of acute deoxygenation on a caribbean coral reef. *Nat. Commun.*, 12(1):4522.

Joossens, M., Huys, G., Cnockaert, M., De Preter, V., Verbeke, K., Rutgeerts, P., Vandamme, P., and Vermeire, S. (2011). Dysbiosis of the faecal microbiota in patients with crohn's disease and their unaffected relatives. *Gut*, 60(5):631–637.

Jørgensen, B. B. (2021). SULFUR BIOGEOCHEMICAL CYCLE OF MARINE SEDIMENTS. *Geochemical Perspectives*, 10(2):145–146.

Jost, T., Lacroix, C., Braegger, C. P., and Chassard, C. (2012). New insights in gut microbiota establishment in healthy breast fed neonates. *PLoS One*, 7(8):e44595.

Jost, T., Lacroix, C., Braegger, C. P., Rochat, F., and Chassard, C. (2014). Vertical mother-neonate transfer of maternal gut bacteria via breastfeeding. *Environ. Microbiol.*, 16(9):2891–2904.

Jungersen, M., Wind, A., Johansen, E., Christensen, J. E., Stuer-Lauridsen, B., and Eskesen, D. (2014). The science behind the probiotic strain bifidobacterium animalis subsp. lactis BB-12(®). *Microorganisms*, 2(2):92–110.

Kanehisa, M. (2017). Enzyme annotation and metabolic reconstruction using KEGG. *Methods Mol. Biol.*, 1611:135–145.

Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., and Ishiguro-Watanabe, M. (2023). KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.*, 51(D1):D587–D592.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, 45(D1):D353–D361.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30.

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

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, 40(Database issue):D109–14.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, 42(Database issue):D199–205.

Kanehisa, M., Sato, Y., and Kawashima, M. (2022). KEGG mapping tools for uncovering hidden features in biological data. *Protein Sci.*, 31(1):47–53.

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

Kao, D., Roach, B., Silva, M., Beck, P., Rioux, K., Kaplan, G. G., Chang, H.-J., Coward, S., Goodman, K. J., Xu, H., Madsen, K., Mason, A., Wong, G. K.-S., Jovel, J., Patterson, J., and Louie, T. (2017). Effect of oral capsule- vs Colonoscopy-Delivered fecal microbiota transplantation on recurrent clostridium difficile infection: A randomized clinical trial. *JAMA*, 318(20):1985–1993.

Kaplan, G. G. (2015). The global burden of IBD: from 2015 to 2025. *Nat. Rev. Gastroenterol. Hepatol.*, 12(12):720–727.

Karl, D., Letelier, R., Tupas, L., Dore, J., Christian, J., and Hebel, D. (1997). The role of nitrogen fixation in biogeochemical cycling in the subtropical north pacific ocean. *Nature*, 388(6642):533–538.

Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., Nielsen, J., and Bäckhed, F. (2013). Gut metagenome in european women with normal, impaired and diabetic glucose control. *Nature*, 498(7452):99–103.

Karp, P. D., Billington, R., Caspi, R., Fulcher, C. A., Latendresse, M., Kothari, A., Keseler, I. M., Krummenacker, M., Midford, P. E., Ong, Q., Ong, W. K., Paley, S. M., and Subhraveti, P. (2019). The BioCyc collection of microbial genomes and metabolic pathways. *Brief. Bioinform.*, 20(4):1085–1093.

Karp, P. D., Midford, P. E., Billington, R., Kothari, A., Krummenacker, M., Latendresse, M., Ong, W. K., Subhraveti, P., Caspi, R., Fulcher, C., Keseler, I. M., and Paley, S. M. (2021). Pathway tools version 23.0 update: software for pathway/genome informatics and systems biology. *Brief. Bioinform.*, 22(1):109–126.

Karp, P. D., Paley, S., Krummenacker, M., Kothari, A., Wannemuehler, M. J., and Phillips, G. J. (2022). Pathway tools management of Pathway/Genome data for microbial communities. *Front Bioinform*, 2:869150.

Karp, P. D., Paley, S. M., Midford, P. E., Krummenacker, M., Billington, R., Kothari, A., Ong, W. K., Subhraveti, P., Keseler, I. M., and Caspi, R. (2015). Pathway tools version 24.0: Integrated software for Pathway/Genome informatics and systems biology. pre-print.

Karp, P. D., Weaver, D., and Latendresse, M. (2018). How accurate is automated gap filling of metabolic models? *BMC Syst. Biol.*, 12(1):73.

Karthikeyan, S., Rodriguez-R, L. M., Heritier-Robbins, P., Kim, M., Overholt, W. A., Gaby, J. C., Hatt, J. K., Spain, J. C., Rosselló-Móra, R., Huettel, M., Kostka, J. E., and Konstantinidis, K. T. (2019). “candidatus macondimonas diazotrophica”, a novel gammaproteobacterial genus dominating crude-oil-contaminated coastal sediments. *ISME J.*, 13(8):2129–2134.

Katayama, S. and Mine, Y. (2007). Antioxidative activity of amino acids on tissue oxidative stress in human intestinal epithelial cell model. *J. Agric. Food Chem.*, 55(21):8458–8464.

Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., and Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature*, 474(7351):327–336.

Kelly, C. J., Alexeev, E. E., Farb, L., Vickery, T. W., Zheng, L., Eric L, C., Kitzenberg, D. A., Battista, K. D., Kominsky, D. J., Robertson, C. E., Frank, D. N., Stabler, S. P., and Colgan, S. P. (2019). Oral vitamin B12 supplement is delivered to the distal gut, altering the corrinoid profile and selectively depleting bacteroides in C57BL/6 mice. *Gut Microbes*, 10(6):654–662.

Khan, I., Ullah, N., Zha, L., Bai, Y., Khan, A., Zhao, T., Che, T., and Zhang, C. (2019). Alteration of gut microbiota in inflammatory bowel disease (IBD): Cause or consequence? IBD treatment targeting the gut microbiome. *Pathogens*, 8(3).

Khoruts, A., Dicksved, J., Jansson, J. K., and Sadowsky, M. J. (2010). Changes in the composition of the human fecal microbiome after bacteriotherapy for recurrent clostridium difficile-associated diarrhea. *J. Clin. Gastroenterol.*, 44(5):354–360.

Khosravi, A. and Mazmanian, S. K. (2013). Disruption of the gut microbiome as a risk factor for microbial infections. *Curr. Opin. Microbiol.*, 16(2):221–227.

Kilstrup, M., Hammer, K., Ruhdal Jensen, P., and Martinussen, J. (2005). Nucleotide metabolism and its control in lactic acid bacteria. *FEMS Microbiol. Rev.*, 29(3):555–590.

Kim, C. J., Kovacs-Nolan, J. A., Yang, C., Archbold, T., Fan, M. Z., and Mine, Y. (2010). l-tryptophan exhibits therapeutic function in a porcine model of dextran sodium sulfate (DSS)-induced colitis. *J. Nutr. Biochem.*, 21(6):468–475.

King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., and Palsson, B. O. (2015). Escher: A web application for building, sharing, and embedding Data-Rich visualizations of biological pathways. *PLoS Comput. Biol.*, 11(8):e1004321.

Knight, R., Callewaert, C., Marotz, C., Hyde, E. R., Debelius, J. W., McDonald, D., and Sogin, M. L. (2017). The microbiome and human biology. *Annu. Rev. Genomics Hum. Genet.*, 18:65–86.

Knoll, R. L., Forslund, K., Kultima, J. R., Meyer, C. U., Kullmer, U., Sunagawa, S., Bork, P., and Gehring, S. (2017). Gut microbiota differs between children with inflammatory bowel disease and healthy siblings in taxonomic and functional composition: a metagenomic analysis. *Am. J. Physiol. Gastrointest. Liver Physiol.*, 312(4):G327–G339.

Knox, N. C., Forbes, J. D., Peterson, C.-L., Van Domselaar, G., and Bernstein, C. N. (2019a). The gut microbiome in inflammatory bowel disease: Lessons learned from other Immune-Mediated inflammatory diseases. *Am. J. Gastroenterol.*, 114(7):1051–1070.

Knox, N. C., Forbes, J. D., Van Domselaar, G., and Bernstein, C. N. (2019b). The gut microbiome as a target for IBD treatment: Are we there yet? *Curr. Treat. Options Gastroenterol.*, 17(1):115–126.

Koek, M. M., Jellema, R. H., van der Greef, J., Tas, A. C., and Hankemeier, T. (2011). Quantitative metabolomics based on gas chromatography mass spectrometry: status and perspectives. *Metabolomics*, 7(3):307–328.

Koenig, J. E., Spor, A., Scalfone, N., Fricker, A. D., Stombaugh, J., Knight, R., Angenent, L. T., and Ley, R. E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proc. Natl. Acad. Sci. U. S. A.*, 108 Suppl 1:4578–4585.

Kolios, G., Valatas, V., and Ward, S. G. (2004). Nitric oxide in inflammatory bowel disease: a universal messenger in an unsolved puzzle. *Immunology*, 113(4):427–437.

Komova, A. V., Bakhmutova, E. D., Izotova, A. O., Kochetova, E. S., Toshchakov, S. V., Namsaraev, Z. B., Golichenkov, M. V., and Korzhenkov, A. A. (2022). Nitrogen fixation activity and genome analysis of a moderately haloalkaliphilic anoxygenic phototrophic bacterium rhodovulum tesquicola. *Microorganisms*, 10(8).

Koonin, E. V. (2016). Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Res.*, 5.

Köster, J. and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522.

Kostic, A. D., Chun, E., Robertson, L., Glickman, J. N., Gallini, C. A., Michaud, M., Clancy, T. E., Chung, D. C., Lochhead, P., Hold, G. L., El-Omar, E. M., Brenner, D., Fuchs, C. S., Meyerson, M., and Garrett, W. S. (2013). Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell Host Microbe*, 14(2):207–215.

Kostic, A. D., Xavier, R. J., and Gevers, D. (2014). The microbiome in inflammatory bowel disease: current status and the future ahead. *Gastroenterology*, 146(6):1489–1499.

Kovatcheva-Datchary, P., Egert, M., Maathuis, A., Rajilić-Stojanović, M., de Graaf, A. A., Smidt, H., de Vos, W. M., and Venema, K. (2009). Linking phylogenetic identities of bacteria to starch fermentation in an in vitro model of the large intestine by RNA-based stable isotope probing. *Environ. Microbiol.*, 11(4):914–926.

Kozubal, M. A., Macur, R. E., Jay, Z. J., Beam, J. P., Malfatti, S. A., Tringe, S. G., Kocar, B. D., Borch, T., and Inskeep, W. P. (2012). Microbial iron cycling in acidic geothermal springs of yellowstone national park: integrating molecular surveys, geochemical processes, and isolation of novel fe-active microorganisms. *Front. Microbiol.*, 3:109.

Kraus, S. and Arber, N. (2009). Inflammation and colorectal cancer. *Curr. Opin. Pharmacol.*, 9(4):405–410.

Kristensen, N. B., Bryrup, T., Allin, K. H., Nielsen, T., Hansen, T. H., and Pedersen, O. (2016). Alterations in fecal microbiota composition by probiotic supplementation in healthy adults: a systematic review of randomized controlled trials. *Genome Med.*, 8(1):52.

Kronman, M. P., Zaoutis, T. E., Haynes, K., Feng, R., and Coffin, S. E. (2012). Antibiotic exposure and IBD development among children: a population-based cohort study. *Pediatrics*, 130(4):e794–803.

Kruger, N. J. and von Schaewen, A. (2003). The oxidative pentose phosphate pathway: structure and organisation. *Curr. Opin. Plant Biol.*, 6(3):236–246.

Kuypers, M. M. M., Marchant, H. K., and Kartal, B. (2018). The microbial nitrogen-cycling network. *Nat. Rev. Microbiol.*, 16(5):263–276.

Landa, M., Burns, A. S., Durham, B. P., Esson, K., Nowinski, B., Sharma, S., Vorobev, A., Nielsen, T., Kiene, R. P., and Moran, M. A. (2019). Sulfur metabolites that facilitate oceanic phytoplankton–bacteria carbon flux. *ISME J.*, 13(10):2536–2550.

Lane, E. R., Zisman, T. L., and Suskind, D. L. (2017). The microbiota in inflammatory bowel disease: current and therapeutic insights. *J. Inflamm. Res.*, 10:63–73.

Langlois, R. J., LaRoche, J., and Raab, P. A. (2005). Diazotrophic diversity and distribution in the tropical and subtropical atlantic ocean. *Appl. Environ. Microbiol.*, 71(12):7910–7919.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359.

Larsbrink, J., Rogers, T. E., Hemsworth, G. R., McKee, L. S., Tauzin, A. S., Spadiut, O., Klinter, S., Pudlo, N. A., Urs, K., Koropatkin, N. M., Creagh, A. L., Haynes, C. A., Kelly, A. G., Cederholm, S. N., Davies, G. J., Martens, E. C., and Brumer, H. (2014). A discrete genetic locus confers xyloglucan metabolism in select human gut bacteroidetes. *Nature*, 506(7489):498–502.

Lau Vetter, M. C. Y., Huang, B., Fenske, L., and Blom, J. (2022). Metabolism of the genus guyparkeria revealed by pangenome analysis. *Microorganisms*, 10(4).

Le Chatelier, E., Nielsen, T., Qin, J., Prifti, E., Hildebrand, F., Falony, G., Almeida, M., Arumugam, M., Batto, J.-M., Kennedy, S., Leonard, P., Li, J., Burgdorf, K., Grarup, N., Jørgensen, T., Brandslund, I., Nielsen, H. B., Juncker, A. S., Bertalan, M., Levenez, F., Pons, N., Rasmussen, S., Sunagawa, S., Tap, J., Tims, S., Zoetendal, E. G., Brunak, S., Clément, K., Doré, J., Kleerebezem, M., Kristiansen, K., Renault, P., Sicheritz-Ponten, T., de Vos, W. M., Zucker, J.-D., Raes, J., Hansen, T., MetaHIT consortium, Bork, P., Wang, J., Ehrlich, S. D., and Pedersen, O. (2013). Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464):541–546.

Ledder, O. (2019). Antibiotics in inflammatory bowel diseases: do we know what we're doing? *Transl Pediatr*, 8(1):42–55.

Lee, C. C., Hu, Y., and Ribbe, M. W. (2009a). Unique features of the nitrogenase VFe protein from *Azotobacter vinelandii*. *Proc. Natl. Acad. Sci. U. S. A.*, 106(23):9209–9214.

Lee, H.-S., Han, S.-Y., Ryu, K.-Y., and Kim, D.-H. (2009b). The degradation of glycosaminoglycans by intestinal microflora deteriorates colitis in mice. *Inflammation*, 32(1):27–36.

Lee, M. and Chang, E. B. (2021). Inflammatory bowel diseases (IBD) and the Microbiome—Searching the crime scene for clues. *Gastroenterology*, 160(2):524–537.

Lee, M. D. (2019). GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics*, 35(20):4162–4164.

Lee, S. T. M., Kahn, S. A., Delmont, T. O., Shaiber, A., Esen, Ö. C., Hubert, N. A., Morrison, H. G., Antonopoulos, D. A., Rubin, D. T., and Eren, A. M. (2017). Tracking microbial colonization in fecal microbiota transplantation experiments via genome-resolved metagenomics. *Microbiome*, 5(1):50.

Leonardi, R., Zhang, Y.-M., Rock, C. O., and Jackowski, S. (2005). Coenzyme a: back in action. *Prog. Lipid Res.*, 44(2-3):125–153.

Leser, T. D. and Mølbak, L. (2009). Better living through microbial action: the benefits of the mammalian gastrointestinal microbiota on the host. *Environ. Microbiol.*, 11(9):2194–2206.

Leung, W., Malhi, G., Willey, B. M., McGeer, A. J., Borgundvaag, B., Thanabalan, R., Gnanasuntharam, P., Le, B., Weizman, A. V., Croitoru, K., Silverberg, M. S., Steinhart, A. H., and Nguyen, G. C. (2012). Prevalence and predictors of MRSA, ESBL, and VRE colonization in the ambulatory IBD population. *J. Crohns. Colitis*, 6(7):743–749.

Levy, S. B. (2000). The future of antibiotics: facing antibiotic resistance. *Clin. Microbiol. Infect.*, 6 Suppl 3:101–106.

Lewis, J. D., Chen, E. Z., Baldassano, R. N., Otley, A. R., Griffiths, A. M., Lee, D., Bittinger, K., Bailey, A., Friedman, E. S., Hoffmann, C., Albenberg, L., Sinha, R., Compher, C., Gilroy, E., Nessel, L., Grant, A., Chehoud, C., Li, H., Wu, G. D., and Bushman, F. D. (2015). Inflammation, antibiotics, and diet as environmental stressors of the gut microbiome in pediatric crohn's disease. *Cell Host Microbe*, 18(4):489–500.

Leylabadlo, H. E., Ghotaslou, R., Feizabadi, M. M., Farajnia, S., Moaddab, S. Y., Ganbarov, K., Khodadadi, E., Tanomand, A., Sheykhsaran, E., Yousefi, B., and Kafil, H. S. (2020). The critical role of faecalibacterium prausnitzii in human health: An overview. *Microb. Pathog.*, 149:104344.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., and Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., Arumugam, M., Kultima, J. R., Prifti, E., Nielsen, T., Juncker, A. S., Manichanh, C., Chen, B., Zhang, W., Levenez, F., Wang, J., Xu, X., Xiao, L., Liang, S., Zhang, D., Zhang, Z., Chen, W., Zhao, H., Al-Aama, J. Y., Edris, S., Yang, H., Wang, J., Hansen, T., Nielsen, H. B., Brunak, S., Kristiansen, K., Guarner, F., Pedersen, O., Doré, J., Ehrlich, S. D., MetaHIT Consortium, Bork, P., Wang, J., and MetaHIT Consortium (2014). An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.*, 32(8):834–841.

Lin, R., Liu, W., Piao, M., and Zhu, H. (2017). A review of the relationship between the gut microbiota and amino acid metabolism. *Amino Acids*, 49(12):2083–2090.

Lin, Y., Caldwell, G. W., Li, Y., Lang, W., and Masucci, J. (2020). Inter-laboratory reproducibility of an untargeted metabolomics GC-MS assay for analysis of human plasma. *Sci. Rep.*, 10(1):10918.

Liu, L., Guo, X., Rao, J. N., Zou, T., Xiao, L., Yu, T., Timmons, J. A., Turner, D. J., and Wang, J.-Y. (2009). Polyamines regulate e-cadherin transcription through c-myc modulating intestinal epithelial barrier function. *Am. J. Physiol. Cell Physiol.*, 296(4):C801–10.

Liu, W., Zhang, J., Wu, C., Cai, S., Huang, W., Chen, J., Xi, X., Liang, Z., Hou, Q., Zhou, B., Qin, N., and Zhang, H. (2016). Unique features of ethnic mongolian gut microbiome revealed by metagenomic analysis. *Sci. Rep.*, 6:34826.

Liu, Y. and Breukink, E. (2016). The membrane steps of bacterial cell wall synthesis as antibiotic targets. *Antibiotics (Basel)*, 5(3).

Liu, Y., Wang, X., and Hu, C.-A. A. (2017). Therapeutic potential of amino acids in inflammatory bowel disease. *Nutrients*, 9(9).

Llor, C. and Bjerrum, L. (2014). Antimicrobial resistance: risk associated with antibiotic overuse and initiatives to reduce the problem. *Ther Adv Drug Saf*, 5(6):229–241.

Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J., and Crosby, L. (2018). Phylogenetically novel uncultured microbial cells dominate earth microbiomes. *mSystems*, 3(5).

Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. (2016a). The healthy human microbiome.

Lloyd-Price, J., Abu-Ali, G., and Huttenhower, C. (2016b). The healthy human microbiome. *Genome Med.*, 8(1):51.

Lloyd-Price, J., Arze, C., Ananthakrishnan, A. N., Schirmer, M., Avila-Pacheco, J., Poon, T. W., Andrews, E., Ajami, N. J., Bonham, K. S., Brislawn, C. J., Casero, D., Courtney, H., Gonzalez, A., Graeber, T. G., Hall, A. B., Lake, K., Landers, C. J., Mallick, H., Plichta, D. R., Prasad, M., Rahnavard, G., Sauk, J., Shungin, D., Vázquez-Baeza, Y., White, 3rd, R. A., IBDMDB Investigators, Braun, J., Denson, L. A., Jansson, J. K., Knight, R., Kugathasan, S., McGovern, D. P. B., Petrosino, J. F., Stappenbeck, T. S., Winter, H. S., Clish, C. B., Franzosa, E. A., Vlamakis, H., Xavier, R. J., and Huttenhower, C. (2019). Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*, 569(7758):655–662.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., Brady, A., Creasy, H. H., McCracken, C., Giglio, M. G., McDonald, D., Franzosa, E. A., Knight, R., White, O., and Huttenhower, C. (2017). Strains, functions and dynamics in the expanded human microbiome project. *Nature*, 550(7674):61–66.

Lobb, B., Tremblay, B. J.-M., Moreno-Hagelsieb, G., and Doxey, A. C. (2020). An assessment of genome annotation coverage across the bacterial tree of life. *Microb Genom*, 6(3).

Loewenstein, Y., Raimondo, D., Redfern, O. C., Watson, J., Frishman, D., Linial, M., Orengo, C., Thornton, J., and Tramontano, A. (2009). Protein function annotation by homology-based inference. *Genome Biol.*, 10(2):207.

Lopez, M. J. and Mohiuddin, S. S. (2023). *Biochemistry, Essential Amino Acids*. StatPearls Publishing.

Lopez-Siles, M., Khan, T. M., Duncan, S. H., Harmsen, H. J. M., Garcia-Gil, L. J., and Flint, H. J. (2012). Cultured representatives of two major phylogroups of human colonic faecalibacterium prausnitzii can utilize pectin, uronic acids, and host-derived substrates for growth. *Appl. Environ. Microbiol.*, 78(2):420–428.

Lozupone, C. A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., Vázquez-Baeza, Y., Jansson, J. K., Gordon, J. I., and Knight, R. (2013). Meta-analyses of studies of the human microbiota. *Genome Res.*, 23(10):1704–1714.

Luo, Y.-W., Doney, S. C., Anderson, L. A., Benavides, M., Berman-Frank, I., Bode, A., Bonnet, S., Boström, K. H., Böttjer, D., Capone, D. G., Carpenter, E. J., Chen, Y. L., Church, M. J., Dore, J. E., Falcón, L. I., Fernández, A., Foster, R. A., Furuya, K., Gómez, F., Gundersen, K., Hynes, A. M., Karl, D. M., Kitajima, S., Langlois, R. J., LaRoche, J., Letelier, R. M., Marañón, E., McGillicuddy, Jr, D. J., Moisander, P. H., Moore, C. M., Mouriño-Carballido, B., Mulholland, M. R., Needoba, J. A., Orcutt, K. M., Poulton, A. J., Rahav, E., Raimbault, P., Rees, A. P., Riemann, L., Shiozaki, T., Subramaniam, A., Tyrrell, T., Turk-Kubo, K. A., Varela, M., Villareal, T. A., Webb, E. A., White, A. E., Wu, J., and Zehr, J. P. (2012). Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates. *Earth Syst. Sci. Data*, 4(1):47–73.

Lynch, K. E., Parke, E. C., and O'Malley, M. A. (2019). How causal are microbiomes? a comparison with the helicobacter pylori explanation of ulcers. *Biol. Philos.*, 34(6):62.

Lynch, S. V. and Pedersen, O. (2016). The human intestinal microbiome in health and disease. *N. Engl. J. Med.*, 375(24):2369–2379.

Machado, D., Andrejev, S., Tramontano, M., and Patil, K. R. (2018). Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res.*, 46(15):7542–7553.

Machiels, K., Joossens, M., Sabino, J., De Preter, V., Arijs, I., Eeckhaut, V., Ballet, V., Claes, K., Van Immerseel, F., Verbeke, K., Ferrante, M., Verhaegen, J., Rutgeerts, P., and Vermeire, S. (2014). A decrease of the butyrate-producing species roseburia hominis and faecalibacterium prausnitzii defines dysbiosis in patients with ulcerative colitis. *Gut*, 63(8):1275–1283.

Madsen, E. L. (2011). Microorganisms and their roles in fundamental biogeochemical cycles. *Curr. Opin. Biotechnol.*, 22(3):456–464.

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

Maldonado-Gómez, M. X., Martínez, I., Bottacini, F., O'Callaghan, A., Ventura, M., van Sinderen, D., Hillmann, B., Vangay, P., Knights, D., Hutkins, R. W., and Walter, J. (2016). Stable engraftment of bifidobacterium longum AH1206 in the human gut depends on individualized features of the resident microbiome. *Cell Host Microbe*, 20(4):515–526.

Man-Aharonovich, D., Kress, N., Zeev, E. B., Berman-Frank, I., and Béjà, O. (2007). Molecular ecology of nifh genes and transcripts in the eastern mediterranean sea. *Environ. Microbiol.*, 9(9):2354–2363.

Marcelino, V. R., Welsh, C., Diener, C., Gulliver, E. L., Rutten, E. L., Young, R. B., Giles, E. M., Gibbons, S. M., Greening, C., and Forster, S. C. (2023). Disease-specific loss of microbial cross-feeding interactions in the human gut. pre-print.

Marchesi, J. R., Dutilh, B. E., Hall, N., Peters, W. H. M., Roelofs, R., Boleij, A., and Tjalsma, H. (2011). Towards the human colorectal cancer microbiome. *PLoS One*, 6(5):e20447.

Martens, E. C., Kelly, A. G., Tauzin, A. S., and Brumer, H. (2014). The devil lies in the details: how variations in polysaccharide fine-structure impact the physiology and evolution of gut microbes. *J. Mol. Biol.*, 426(23):3851–3865.

Martens, J. H., Barg, H., Warren, M. J., and Jahn, D. (2002). Microbial production of vitamin B12. *Appl. Microbiol. Biotechnol.*, 58(3):275–285.

Martin-Gallausiaux, C., Marinelli, L., Blottière, H. M., Larraufie, P., and Lapaque, N. (2021). SCFA: mechanisms and functional importance in the gut. *Proc. Nutr. Soc.*, 80(1):37–49.

Martínez, G. M., Pire, C., and Martínez-Espinosa, R. M. (2022). Hypersaline environments as natural sources of microbes with potential applications in biotechnology: The case of solar evaporation systems to produce salt in alicante county (spain). *Curr Res Microb Sci*, 3:100136.

Martínez-Pérez, C., Mohr, W., Löscher, C. R., Dekaezemacker, J., Littmann, S., Yilmaz, P., Lehnen, N., Fuchs, B. M., Lavik, G., Schmitz, R. A., LaRoche, J., and Kuypers, M. M. M. (2016). The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nat Microbiol*, 1(11):16163.

Martínez-Pérez, C., Mohr, W., Schwedt, A., Dürschlag, J., Callbeck, C. M., Schunck, H., Dekaezemacker, J., Buckner, C. R. T., Lavik, G., Fuchs, B. M., and Kuypers, M. M. M. (2018). Metabolic versatility of a novel N2 -fixing alphaproteobacterium isolated from a marine oxygen minimum zone. *Environ. Microbiol.*, 20(2):755–768.

Martiny, A. C., Treseder, K., and Pusch, G. (2013). Phylogenetic conservatism of functional traits in microorganisms. *ISME J.*, 7(4):830–838.

Maslowski, K. M., Vieira, A. T., Ng, A., Kranich, J., Sierro, F., Yu, D., Schilter, H. C., Rolph, M. S., Mackay, F., Artis, D., Xavier, R. J., Teixeira, M. M., and Mackay, C. R. (2009). Regulation of inflammatory responses by gut microbiota and chemoattractant receptor GPR43. *Nature*, 461(7268):1282–1286.

Massello, F. L., Chan, C. S., Chan, K.-G., Goh, K. M., Donati, E., and Urbieta, M. S. (2020). Meta-Analysis of microbial communities in hot springs: Recurrent taxa and complex shaping factors beyond ph and temperature. *Microorganisms*, 8(6).

Maynard, C. L., Elson, C. O., Hatton, R. D., and Weaver, C. T. (2012). Reciprocal interactions of the intestinal microbiota and immune system. *Nature*, 489(7415):231–241.

McBurney, M. I., Davis, C., Fraser, C. M., Schneeman, B. O., Huttenhower, C., Verbeke, K., Walter, J., and Latulippe, M. E. (2019). Establishing what constitutes a healthy human gut microbiome: State of the science, regulatory considerations, and future directions. *J. Nutr.*, 149(11):1882–1895.

McCormack, S. A. and Johnson, L. R. (1991). Role of polyamines in gastrointestinal mucosal growth. *Am. J. Physiol.*, 260(6 Pt 1):G795–806.

McCubbin, T., Gonzalez-Garcia, R. A., Palfreyman, R. W., Stowers, C., Nielsen, L. K., and Marcellin, E. (2020). A Pan-Genome guided metabolic network reconstruction of five propionibacterium species reveals extensive metabolic diversity. *Genes*, 11(10).

McFarland, L. V. (2014). Use of probiotics to correct dysbiosis of normal microbiota following disease or disruptive events: a systematic review. *BMJ Open*, 4(8):e005047.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., Spies, J. R., Thaney, K., Updegrove, A., Woo, K. H., and Yarkoni, T. (2016). How open science helps researchers succeed. *Elife*, 5.

Mendoza, S. N., Olivier, B. G., Molenaar, D., and Teusink, B. (2019). A systematic assessment of current genome-scale metabolic reconstruction tools. *Genome Biol.*, 20(1):158.

Mesnage, R. and Antoniou, M. N. (2020). Computational modelling provides insight into the effects of glyphosate on the shikimate pathway in the human gut microbiome. *Curr Res Toxicol*, 1:25–33.

Messer, J. S., Liechty, E. R., Vogel, O. A., and Chang, E. B. (2017). Evolutionary and ecological forces that shape the bacterial communities of the human gut. *Mucosal Immunol.*, 10(3):567–579.

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

Metges, C. C., El-Khoury, A. E., Henneman, L., Petzke, K. J., Grant, I., Bedri, S., Pereira, P. P., Ajami, A. M., Fuller, M. F., and Young, V. R. (1999). Availability of intestinal microbial lysine for whole body lysine homeostasis in human subjects. *Am. J. Physiol.*, 277(4):E597–607.

Mikkola, S. (2020). Nucleotide sugars in chemistry and biology. *Molecules*, 25(23).

Milani, C., Duranti, S., Lugli, G. A., Bottacini, F., Strati, F., Arioli, S., Foroni, E., Turroni, F., van Sinderen, D., and Ventura, M. (2013). Comparative genomics of bifidobacterium animalis subsp. lactis reveals a strict monophyletic bifidobacterial taxon. *Appl. Environ. Microbiol.*, 79(14):4304–4315.

Milani, C., Turroni, F., Duranti, S., Lugli, G. A., Mancabelli, L., Ferrario, C., van Sinderen, D., and Ventura, M. (2016). Genomics of the genus bifidobacterium reveals Species-Specific adaptation to the Glycan-Rich gut environment. *Appl. Environ. Microbiol.*, 82(4):980–991.

Minarik, P., Tomaskova, N., Kollarova, M., and Antalik, M. (2002). Malate dehydrogenases-structure and function. *Gen. Physiol. Biophys.*, 21(3):257–266.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.*, 37(5):1530–1534.

Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on illumina HiSeq and genome analyzer systems. *Genome Biol.*, 12(11):R112.

Mira, A., Martín-Cuadrado, A. B., D'Auria, G., and Rodríguez-Valera, F. (2010). The bacterial pan-genome: a new paradigm in microbiology. *Int. Microbiol.*, 13(2):45–57.

Miranda, K., Weigel, B. L., Fogarty, E. C., Veseli, I. A., Giblin, A. E., Eren, A. M., and Pfister, C. A. (2022). The diversity and functional capacity of microbes associated with coastal macrophytes. *mSystems*, 7(5):e0059222.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., and Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res.*, 49(D1):D412–D419.

Modin, O., Fuad, N., Abadikhah, M., I'Ons, D., Ossiansson, E., Gustavsson, D. J. I., Edefell, E., Suarez, C., Persson, F., and Wilén, B.-M. (2022). A relationship between phages and organic carbon in wastewater treatment plant effluents. *Water Res X*, 16:100146.

Mohite, O. S., Lloyd, C. J., Monk, J. M., Weber, T., and Palsson, B. O. (2022). Pangenome analysis of enterobacteria reveals richness of secondary metabolite gene clusters and their associated gene sets. *Synth Syst Biotechnol*, 7(3):900–910.

Moisander, P. H., Beinart, R. A., Hewson, I., White, A. E., Johnson, K. S., Carlson, C. A., Montoya, J. P., and Zehr, J. P. (2010). Unicellular cyanobacterial distributions broaden the oceanic N2 fixation domain. *Science*, 327(5972):1512–1514.

Moisander, P. H., Beinart, R. A., Voss, M., and Zehr, J. P. (2008). Diversity and abundance of diazotrophic microorganisms in the south china sea during intermonsoon. *ISME J.*, 2(9):954–967.

Moisander, P. H., Benavides, M., Bonnet, S., Berman-Frank, I., White, A. E., and Riemann, L. (2017). Chasing after non-cyanobacterial nitrogen fixation in marine pelagic environments. *Front. Microbiol.*, 8:1736.

Montoya, J. P., Holl, C. M., Zehr, J. P., Hansen, A., Villareal, T. A., and Capone, D. G. (2004). High rates of N2 fixation by unicellular diazotrophs in the oligotrophic pacific ocean. *Nature*, 430(7003):1027–1032.

Moore, C. M., Mills, M. M., Arrigo, K. R., Berman-Frank, I., Bopp, L., Boyd, P. W., Galbraith, E. D., Geider, R. J., Guieu, C., Jaccard, S. L., Jickells, T. D., La Roche, J., Lenton, T. M., Mahowald, N. M., Marañón, E., Marinov, I., Moore, J. K., Nakatsuka, T., Oschlies, A., Saito, M. A., Thingstad, T. F., Tsuda, A., and Ulloa, O. (2013). Processes and patterns of oceanic nutrient limitation. *Nat. Geosci.*, 6(9):701–710.

Moran, M. A. (2015). The global ocean microbiome. *Science*, 350(6266):aac8455.

Moran, M. A., Buchan, A., González, J. M., Heidelberg, J. F., Whitman, W. B., Kiene, R. P., Henriksen, J. R., King, G. M., Belas, R., Fuqua, C., Brinkac, L., Lewis, M., Johri, S., Weaver, B., Pai, G., Eisen, J. A., Rahe, E., Sheldon, W. M., Ye, W., Miller, T. R., Carlton, J., Rasko, D. A., Paulsen, I. T., Ren, Q., Daugherty, S. C., Deboy, R. T., Dodson, R. J., Durkin, A. S., Madupu, R., Nelson, W. C., Sullivan, S. A., Rosovitz, M. J., Haft, D. H., Selengut, J., and Ward, N. (2004). Genome sequence of silicibacter pomeroyi reveals adaptations to the marine environment. *Nature*, 432(7019):910–913.

Moran, M. A. and Durham, B. P. (2019). Sulfur metabolites in the pelagic ocean. *Nat. Rev. Microbiol.*, 17(11):665–678.

Moran, N. A. (2001). The coevolution of bacterial endosymbionts and Phloem-Feeding insects. *Ann. Mo. Bot. Gard.*, 88(1):35–44.

Moreira-Coello, V., Mouriño-Carballido, B., Marañón, E., Fernández-Carrera, A., Bode, A., Sintes, E., Zehr, J. P., Turk-Kubo, K., and Varela, M. M. (2019). Temporal variability of diazotroph community composition in the upwelling region off NW iberia. *Sci. Rep.*, 9(1):3737.

Morel, F. M. M. and Price, N. M. (2003). The biogeochemical cycles of trace metals in the oceans. *Science*, 300(5621):944–947.

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., Reyes, J. A., Shah, S. A., LeLeiko, N., Snapper, S. B., Bousvaros, A., Korzenik, J., Sands, B. E., Xavier, R. J., and Huttenhower, C. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.*, 13(9):R79.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C., and Kanehisa, M. (2007). KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, 35(Web Server issue):W182–5.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nat Hum Behav*, 1:0021.

Munson-McGee, J. H., Lindsay, M. R., Sintes, E., Brown, J. M., D'Angelo, T., Brown, J., Lubelczyk, L. C., Tomko, P., Emerson, D., Orcutt, B. N., Poulton, N. J., Herndl, G. J., and Stepanauskas, R. (2022). Decoupling of respiration rates and abundance in marine prokaryoplankton. *Nature*, 612(7941):764–770.

Murat Eren (Meren), A. (2018). Microbiologists vs. shotgun metagenomes: The surface ocean edition. `https://microbiologycommunity.nature.com/posts/34040-microbiologists-vs-shotgun-metagenomes-surface-ocean`. Accessed: 2023-3-22.

Musrati, R. A., Kollárová, M., Mernik, N., and Mikulásová, D. (1998). Malate dehydrogenase: distribution, function and properties. *Gen. Physiol. Biophys.*, 17(3):193–210.

Nagalingam, N. A. and Lynch, S. V. (2012). Role of the microbiota in inflammatory bowel diseases. *Inflamm. Bowel Dis.*, 18(5):968–984.

Nagao-Kitamoto, H., Shreiner, A. B., Gillilland, 3rd, M. G., Kitamoto, S., Ishii, C., Hirayama, A., Kuffa, P., El-Zaatari, M., Grasberger, H., Seekatz, A. M., Higgins, P. D. R., Young, V. B., Fukuda, S., Kao, J. Y., and Kamada, N. (2016). Functional characterization of inflammatory bowel Disease-Associated gut dysbiosis in gnotobiotic mice. *Cell Mol Gastroenterol Hepatol*, 2(4):468–481.

Naselli-Flores, L. and Padisák, J. (2022). Ecosystem services provided by marine and freshwater phytoplankton. *Hydrobiologia*, pages 1–16.

Needoba, J. A., Foster, R. A., Sakamoto, C., Zehr, J. P., and Johnson, K. S. (2007). Nitrogen fixation by unicellular diazotrophic cyanobacteria in the temperate oligotrophic north pacific ocean. *Limnol. Oceanogr.*, 52(4):1317–1327.

Newbold, C. J. and Ramos-Morales, E. (2020). Review: Ruminal microbiome and microbial metabolome: effects of diet and ruminant host. *Animal*, 14(S1):s78–s86.

Ni, J., Wu, G. D., Albenberg, L., and Tomov, V. T. (2017). Gut microbiota and IBD: causation or correlation? *Nat. Rev. Gastroenterol. Hepatol.*, 14(10):573–584.

NIH Human Microbiome Portfolio Analysis Team (2019). A review of 10 years of human microbiome research activities at the US national institutes of health, fiscal years 2007-2016. *Microbiome*, 7(1):31.

Nikolaus, S., Schulte, B., Al-Massad, N., Thieme, F., Schulte, D. M., Bethge, J., Rehman, A., Tran, F., Aden, K., Häsler, R., Moll, N., Schütze, G., Schwarz, M. J., Waetzig, G. H., Rosenstiel, P., Krawczak, M., Szymczak, S., and Schreiber, S. (2017). Increased tryptophan metabolism is associated with activity of inflammatory bowel diseases. *Gastroenterology*, 153(6):1504–1516.e2.

Nishida, A., Inoue, R., Inatomi, O., Bamba, S., Naito, Y., and Andoh, A. (2018). Gut microbiota in the pathogenesis of inflammatory bowel disease. *Clin. J. Gastroenterol.*, 11(1):1–10.

Nitzan, O., Elias, M., Peretz, A., and Saliba, W. (2016). Role of antibiotics for treatment of inflammatory bowel disease. *World J. Gastroenterol.*, 22(3):1078–1087.

Novakovic, M., Rout, A., Kingsley, T., Kirchoff, R., Singh, A., Verma, V., Kant, R., and Chaudhary, R. (2020). Role of gut microbiota in cardiovascular diseases. *World J. Cardiol.*, 12(4):110–122.

Nowinski, B. and Moran, M. A. (2021). Niche dimensions of a marine bacterium are identified using invasion studies in coastal seawater. *Nat Microbiol*, 6(4):524–532.

Nuppunen-Puputti, M., Kietäväinen, R., Raulio, M., Soro, A., Purkamo, L., Kukkonen, I., and Bomberg, M. (2022). Epilithic microbial community functionality in deep oligotrophic continental bedrock. *Front. Microbiol.*, 13:826048.

Nygaard, P. (2014). Purine and pyrimidine salvage pathways. In *Bacillus subtilis and Other Gram-Positive Bacteria*, <i>Bacillus subtilis</i>and Other Gram-Positive Bacteria, pages 359–378. ASM Press, Washington, DC, USA.

Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P. M., Spicer, P., Lawson, P., Marin-Reyes, L., Trujillo-Villarroel, O., Foster, M., Guija-Poma, E., Troncoso-Corzo, L., Warinner, C., Ozga, A. T., and Lewis, C. M. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.*, 6:6505.

Och, L. M. and Shields-Zhou, G. A. (2012). The neoproterozoic oxygenation event: Environmental perturbations and biogeochemical cycling. *Earth-Sci. Rev.*, 110(1):26–57.

of Sydney, U. (2016a). BioProject. `https://www.ncbi.nlm.nih.gov/bioproject/PRJEB6092/`. Accessed: 2022-9-23.

of Sydney, U. (2016b). EKmeta. public dataset without associated publication.

Ohki, K., Zehr, J. P., and Fujita, Y. (1992). Trichodesmium: Establishment of culture and characteristics of N2-Fixation. In Carpenter, E. J., Capone, D. G., and Rueter, J. G., editors, *Marine Pelagic Cyanobacteria: Trichodesmium and other Diazotrophs*, pages 307–318. Springer Netherlands, Dordrecht.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, 44(D1):D733–45.

Olofsson, M., Ferrer-González, F. X., Uchimiya, M., Schreier, J. E., Holderman, N. R., Smith, C. B., Edison, A. S., and Moran, M. A. (2022). Growth-stage-related shifts in diatom endometabolome composition set the stage for bacterial heterotrophy. *ISME Communications*, 2(1):1–9.

Orth, J. D. and Palsson, B. Ø. (2010). Systematizing the generation of missing metabolic knowledge. *Biotechnol. Bioeng.*, 107(3):403–412.

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What is flux balance analysis? *Nat. Biotechnol.*, 28(3):245–248.

Osman, E. O. and Weinnig, A. M. (2022). Microbiomes and obligate symbiosis of Deep-Sea animals. *Annu Rev Anim Biosci*, 10:151–176.

Ott, S. J., Musfeldt, M., Wenderoth, D. F., Hampe, J., Brant, O., Fölsch, U. R., Timmis, K. N., and Schreiber, S. (2004). Reduction in diversity of the colonic mucosa associated bacterial microflora in patients with active inflammatory bowel disease. *Gut*, 53(5):685–693.

Oz, H. S., Chen, T. S., McClain, C. J., and de Villiers, W. J. S. (2005). Antioxidants as novel therapy in a murine model of colitis. *J. Nutr. Biochem.*, 16(5):297–304.

Palleja, A., Mikkelsen, K. H., Forslund, S. K., Kashani, A., Allin, K. H., Nielsen, T., Hansen, T. H., Liang, S., Feng, Q., Zhang, C., Pyl, P. T., Coelho, L. P., Yang, H., Wang, J., Typas, A., Nielsen, M. F., Nielsen, H. B., Bork, P., Wang, J., Vilsbøll, T., Hansen, T., Knop, F. K., Arumugam, M., and Pedersen, O. (2018). Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nat Microbiol*, 3(11):1255–1265.

Palù, M., Basile, A., Zampieri, G., Treu, L., Rossi, A., Morlino, M. S., and Campanaro, S. (2022). KEMET - a python tool for KEGG module evaluation and microbial genome annotation expansion. *Comput. Struct. Biotechnol. J.*, 20:1481–1486.

Pantigoso, H. A., Newberger, D., and Vivanco, J. M. (2022). The rhizosphere microbiome: Plant-microbial interactions for resource acquisition. *J. Appl. Microbiol.*, 133(5):2864–2876.

Papa, E., Docktor, M., Smillie, C., Weber, S., Preheim, S. P., Gevers, D., Giannoukos, G., Ciulla, D., Tabbaa, D., Ingram, J., Schauer, D. B., Ward, D. V., Korzenik, J. R., Xavier, R. J., Bousvaros, A., and Alm, E. J. (2012). Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. *PLoS One*, 7(6):e39242.

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for bacteria and archaea. *Nat. Biotechnol.*, 38(9):1079–1086.

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., and Hugenholtz, P. (2022). GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, 50(D1):D785–D794.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, 36(10):996–1004.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, 25(7):1043–1055.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M. C., Rice, B. L., DuLong, C., Morgan, X. C., Golden, C. D., Quince, C., Huttenhower, C., and Segata, N. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 176(3):649–662.e20.

Peixoto, R. S., Harkins, D. M., and Nelson, K. E. (2021). Advances in microbiome research for animal health. *Annu. Rev. Anim. Biosci.*, 9(1):289–311.

Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428.

Perez-Riverol, Y., Zorin, A., Dass, G., Vu, M.-T., Xu, P., Glont, M., Vizcaíno, J. A., Jarnuczak, A. F., Petryszak, R., Ping, P., and Hermjakob, H. (2019). Quantifying the impact of public omics data. *Nat. Commun.*, 10(1):3512.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., Iudicone, D., Karsenti, E., Speich, S., Troublé, R., Dimier, C., and Searson, S. (2015). Open science resources for the discovery and analysis of tara oceans data. *Scientific Data*, 2(1):1–16.

Petrov, V. A., Saltykova, I. V., Zhukova, I. A., Alifirova, V. M., Zhukova, N. G., Dorofeeva, Y. B., Tyakht, A. V., Kovarsky, B. A., Alekseev, D. G., Kostryukova, E. S., Mironova, Y. S., Izhboldina, O. P., Nikitina, M. A., Perevozchikova, T. V., Fait, E. A., Babenko, V. V., Vakhitova, M. T., Govorun, V. M., and Sazonov, A. E. (2017). Analysis of gut microbiota in patients with parkinson's disease. *Bull. Exp. Biol. Med.*, 162(6):734–737.

Philippot, L., Andersson, S. G. E., Battin, T. J., Prosser, J. I., Schimel, J. P., Whitman, W. B., and Hallin, S. (2010). The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev. Microbiol.*, 8(7):523–529.

Pierella Karlusich, J. J., Pelletier, E., Lombard, F., Carsique, M., Dvorak, E., Colin, S., Picheral, M., Cornejo-Castillo, F. M., Acinas, S. G., Pepperkok, R., Karsenti, E., de Vargas, C., Wincker, P., Bowler, C., and Foster, R. A. (2021). Global distribution patterns of marine nitrogen-fixers by imaging and molecular methods. *Nat. Commun.*, 12(1):4160.

Pierzynowski, S. and Pierzynowska, K. (2022). Alpha-ketoglutarate, a key molecule involved in nitrogen circulation in both animals and plants, in the context of human gut microbiota and protein metabolism. *Adv. Med. Sci.*, 67(1):142–147.

Plichta, D. R., Graham, D. B., Subramanian, S., and Xavier, R. J. (2019). Therapeutic opportunities in inflammatory bowel disease: Mechanistic dissection of Host-Microbiome relationships. *Cell*, 178(5):1041–1056.

Podlesny, D. and Florian Fricke, W. (2020). Microbial strain engraftment, persistence and replacement after fecal microbiota transplantation. *medRxiv*, page 2020.09.29.20203638.

Porrini, C., Guérin, C., Tran, S.-L., Dervyn, R., Nicolas, P., and Ramarao, N. (2021). Implication of a key region of six bacillus cereus genes involved in siroheme synthesis, nitrite reductase production and iron cluster repair in the bacterial response to nitric oxide stress. *Int. J. Mol. Sci.*, 22(10).

Portincasa, P., Bonfrate, L., Vacca, M., De Angelis, M., Farella, I., Lanza, E., Khalil, M., Wang, D. Q.-H., Sperandio, M., and Di Ciaula, A. (2022). Gut microbiota and short chain fatty acids: Implications in glucose homeostasis. *Int. J. Mol. Sci.*, 23(3).

Powell, J. E., Leonard, S. P., Kwong, W. K., Engel, P., and Moran, N. A. (2016). Genome-wide screen identifies host colonization determinants in a bacterial gut symbiont. *Proc. Natl. Acad. Sci. U. S. A.*, 113(48):13887–13892.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, 5(3):e9490.

Prindiville, T. P., Sheikh, R. A., Cohen, S. H., Tang, Y. J., Cantrell, M. C., and Silva, Jr, J. (2000). Bacteroides fragilis enterotoxin gene sequences in patients with inflammatory bowel disease. *Emerg. Infect. Dis.*, 6(2):171–174.

Pritchard, L., Glover, R. H., Humphris, S., Elphinstone, J. G., and Toth, I. K. (2016). Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal. Methods*, 8(1):12–24.

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

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

Quince, C., Delmont, T. O., Raguideau, S., Alneberg, J., Darling, A. E., Collins, G., and Eren, A. M. (2017). DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biol.*, 18(1):181.

Quince, C., Ijaz, U. Z., Loman, N., Eren, A. M., Saulnier, D., Russell, J., Haig, S. J., Calus, S. T., Quick, J., Barclay, A., Bertz, M., Blaut, M., Hansen, R., McGrogan, P., Russell, R. K., Edwards, C. A., and Gerasimidis, K. (2015). Extensive modulation of the fecal metagenome in children with crohn's disease during exclusive enteral nutrition. *Am. J. Gastroenterol.*, 110(12):1718–29; quiz 1730.

Rahav, E., Bar-Zeev, E., Ohayon, S., Elifantz, H., Belkin, N., Herut, B., Mulholland, M. R., and Berman-Frank, I. (2013). Dinitrogen fixation in aphotic oxygenated marine environments. *Front. Microbiol.*, 4:227.

Ralevic, V. (2015). UDP-Glucose. In *Reference Module in Biomedical Sciences*. Elsevier.

Ramachandran, R., Bugbee, K., and Murphy, K. (2021). From open data to open science. *Earth Space Sci.*, 8(5).

Ramirez, J., Guarner, F., Bustos Fernandez, L., Maruy, A., Sdepanian, V. L., and Cohen, H. (2020). Antibiotics as major disruptors of gut microbiota. *Front. Cell. Infect. Microbiol.*, 10:572912.

Rampelli, S., Schnorr, S. L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A. N., Henry, A. G., and Candela, M. (2015). Metagenome sequencing of the hadza Hunter-Gatherer gut microbiota. *Curr. Biol.*, 25(13):1682–1693.

Raskov, H., Burcharth, J., and Pommergaard, H.-C. (2017). Linking gut microbiota to colorectal cancer. *J. Cancer*, 8(17):3378–3395.

Rasmussen, J. A., Kiilerich, P., Madhun, A. S., Waagbø, R., Lock, E.-J. R., Madsen, L., Gilbert, M. T. P., Kristiansen, K., and Limborg, M. T. (2023). Co-diversification of an intestinal mycoplasma and its salmonid host. *ISME J.*

Rath, S., Heidrich, B., Pieper, D. H., and Vital, M. (2017). Uncovering the trimethylamine-producing bacteria of the human gut microbiota. *Microbiome*, 5(1):54.

Raymond, F., Ouameur, A. A., Déraspe, M., Iqbal, N., Gingras, H., Dridi, B., Leprohon, P., Plante, P.-L., Giroux, R., Bérubé, È., Frenette, J., Boudreau, D. K., Simard, J.-L., Chabot, I., Domingo, M.-C., Trottier, S., Boissinot, M., Huletsky, A., Roy, P. H., Ouellette, M., Bergeron, M. G., and Corbeil, J. (2016). The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J.*, 10(3):707–720.

Rehman, T., Shabbir, M. A., Inam-Ur-Raheem, M., Manzoor, M. F., Ahmad, N., Liu, Z.-W., Ahmad, M. H., Siddeeg, A., Abid, M., and Aadil, R. M. (2020). Cysteine and homocysteine as biomarker of various diseases. *Food Sci Nutr*, 8(9):4696–4707.

Reiter, T. E., Irber, L., Gingrich, A. A., Haynes, D., Tessa Pierce-Ward, N., Brooks, P. T., Mizutani, Y., Moritz, D., Reidl, F., Willis, A. D., Sullivan, B. D., and Titus Brown, C. (2022). Meta-analysis of metagenomes via machine learning and assembly graphs reveals strain switches in crohn's disease. pre-print.

Rhodes, J. M. (2007). The role of escherichia coli in inflammatory bowel disease. *Gut*, 56(5):610–612.

Ridlon, J. M., Kang, D. J., Hylemon, P. B., and Bajaj, J. S. (2014). Bile acids and the gut microbiome. *Curr. Opin. Gastroenterol.*, 30(3):332–338.

Riemann, L., Farnelid, H., and Steward, G. F. (2010). Nitrogenase genes in non-cyanobacterial plankton: prevalence, diversity and regulation in marine waters. *Aquat. Microb. Ecol.*, 61(3):235–247.

Rigottier-Gois, L. (2013). Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis. *ISME J.*, 7(7):1256–1261.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., Darling, A., Malfatti, S., Swan, B. K., Gies, E. A., Dodsworth, J. A., Hedlund, B. P., Tsiamis, G., Sievert, S. M., Liu, W.-T., Eisen, J. A., Hallam, S. J., Kyrpides, N. C., Stepanauskas, R., Rubin, E. M., Hugenholtz, P., and Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437.

Roager, H. M. and Licht, T. R. (2018). Microbial tryptophan catabolites in health and disease. *Nat. Commun.*, 9(1):3294.

Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R., and Konstantinidis, K. T. (2018). Nonpareil 3: Fast estimation of metagenomic coverage and sequence diversity. *mSystems*, 3(3).

Rothschild, D., Weissbrod, O., Barkan, E., Kurilshikov, A., Korem, T., Zeevi, D., Costea, P. I., Godneva, A., Kalka, I. N., Bar, N., Shilo, S., Lador, D., Vila, A. V., Zmora, N., Pevsner-Fischer, M., Israeli, D., Kosower, N., Malka, G., Wolf, B. C., Avnit-Sagi, T., Lotan-Pompan, M., Weinberger, A., Halpern, Z., Carmi, S., Fu, J., Wijmenga, C., Zhernakova, A., Elinav, E., and Segal, E. (2018). Environment dominates over host genetics in shaping human gut microbiota. *Nature*, 555(7695):210–215.

Rousk, J. and Bengtson, P. (2014). Microbial regulation of global biogeochemical cycles. *Front. Microbiol.*, 5:103.

Roy, A. and Lichtiger, S. (2016). Clostridium difficile infection: A rarity in patients receiving chronic antibiotic treatment for crohn's disease. *Inflamm. Bowel Dis.*, 22(3):648–653.

Runde, J., Veseli, I., Fogarty, E. C., Watson, A. R., Clayssen, Q., Yosef, M., Shaiber, A., Verma, R., Quince, C., Gerasimidis, K., Rubin, D. T., and Eren, A. M. (2023). Transient suppression of bacterial populations associated with gut health is critical in success of exclusive enteral nutrition for children with crohn's disease. *J. Crohns. Colitis*.

Ryczko, M. C., Pawling, J., Chen, R., Abdel Rahman, A. M., Yau, K., Copeland, J. K., Zhang, C., Surendra, A., Guttman, D. S., Figeys, D., and Dennis, J. W. (2016). Metabolic reprogramming by hexosamine biosynthetic and golgi N-Glycan branching pathways. *Sci. Rep.*, 6:23043.

Saitoh, S., Noda, S., Aiba, Y., Takagi, A., Sakamoto, M., Benno, Y., and Koga, Y. (2002). Bacteroides ovatus as the predominant commensal intestinal microbe causing a systemic antibody response in inflammatory bowel disease. *Clin. Diagn. Lab. Immunol.*, 9(1):54–59.

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.-J., Cuenca, M., Field, C. M., Coelho, L. P., Cruaud, C., Engelen, S., Gregory, A. C., Labadie, K., Marec, C., Pelletier, E., Royo-Llonch, M., Roux, S., Sánchez, P., Uehara, H., Zayed, A. A., Zeller, G., Carmichael, M., Dimier, C., Ferland, J., Kandels, S., Picheral, M., Pisarev, S., Poulain, J., Tara Oceans Coordinators, Acinas, S. G., Babin, M., Bork, P., Bowler, C., de Vargas, C., Guidi, L., Hingamp, P., Iudicone, D., Karp-Boss, L., Karsenti, E., Ogata, H., Pesant, S., Speich, S., Sullivan, M. B., Wincker, P., and Sunagawa, S. (2019). Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell*, 179(5):1068–1083.e21.

Salyers, A. A., West, S. E., Vercellotti, J. R., and Wilkins, T. D. (1977). Fermentation of mucins and plant polysaccharides by anaerobic bacteria from the human colon. *Appl. Environ. Microbiol.*, 34(5):529–533.

Sánchez-Baracaldo, P., Bianchini, G., Wilson, J. D., and Knoll, A. H. (2022). Cyanobacteria and biogeochemical cycles through earth history. *Trends Microbiol.*, 30(2):143–157.

Sanders, R., Henson, S. A., Koski, M., De La Rocha, C. L., Painter, S. C., Poulton, A. J., Riley, J., Salihoglu, B., Visser, A., Yool, A., Bellerby, R., and Martin, A. P. (2014). The biological carbon pump in the north atlantic. *Prog. Oceanogr.*, 129:200–218.

Sartor, R. B. (2006). Mechanisms of disease: pathogenesis of crohn's disease and ulcerative colitis. *Nat. Clin. Pract. Gastroenterol. Hepatol.*, 3(7):390–407.

Scavotto, R. E., Dziallas, C., Bentzon-Tilia, M., Riemann, L., and Moisander, P. H. (2015). Nitrogen-fixing bacteria associated with copepods in coastal waters of the north atlantic ocean. *Environ. Microbiol.*, 17(10):3754–3765.

Schirmer, M., Denson, L., Vlamakis, H., Franzosa, E. A., Thomas, S., Gotman, N. M., Rufo, P., Baker, S. S., Sauer, C., Markowitz, J., Pfefferkorn, M., Oliva-Hemker, M., Rosh, J., Otley, A., Boyle, B., Mack, D., Baldassano, R., Keljo, D., LeLeiko, N., Heyman, M., Griffiths, A., Patel, A. S., Noe, J., Kugathasan, S., Walters, T., Huttenhower, C., Hyams, J., and Xavier, R. J. (2018a). Compositional and temporal changes in the gut microbiome of pediatric ulcerative colitis patients are linked to disease course. *Cell Host Microbe*, 24(4):600–610.e4.

Schirmer, M., Franzosa, E. A., Lloyd-Price, J., McIver, L. J., Schwager, R., Poon, T. W., Ananthakrishnan, A. N., Andrews, E., Barron, G., Lake, K., Prasad, M., Sauk, J., Stevens, B., Wilson, R. G., Braun, J., Denson, L. A., Kugathasan, S., McGovern, D. P. B., Vlamakis, H., Xavier, R. J., and Huttenhower, C. (2018b). Dynamics of metatranscription in the inflammatory bowel disease gut microbiome. *Nat Microbiol*, 3(3):337–346.

Schirmer, M., Garner, A., Vlamakis, H., and Xavier, R. J. (2019). Microbial genes and pathways in inflammatory bowel disease. *Nat. Rev. Microbiol.*, 17(8):497–511.

Schmidt, T. S. B., Raes, J., and Bork, P. (2018). The human gut microbiome: From association to modulation. *Cell*, 172(6):1198–1215.

Schroeder, B. O., Birchenough, G. M. H., Ståhlman, M., Arike, L., Johansson, M. E. V., Hansson, G. C., and Bäckhed, F. (2018). Bifidobacteria or fiber protects against Diet-Induced Microbiota-Mediated colonic mucus deterioration. *Cell Host Microbe*, 23(1):27–40.e7.

Schupack, D. A., Mars, R. A. T., Voelker, D. H., Abeykoon, J. P., and Kashyap, P. C. (2022). The promise of the gut microbiome as part of individualized treatment strategies. *Nat. Rev. Gastroenterol. Hepatol.*, 19(1):7–25.

Scrivens, M. and Dickenson, J. M. (2005). Functional expression of the P2Y14 receptor in murine t-lymphocytes. *Br. J. Pharmacol.*, 146(3):435–444.

Seemann, T. (2017). barrnap: Bacterial ribosomal RNA predictor.

Seetharam, B. and Alpers, D. H. (1982). Absorption and transport of cobalamin (vitamin b12). *Annu. Rev. Nutr.*, 2:343–369.

Segal, J. P., Mullish, B. H., Quraishi, M. N., Acharjee, A., Williams, H. R. T., Iqbal, T., Hart, A. L., and Marchesi, J. R. (2019). The application of omics techniques to understand the role of the gut microbiota in inflammatory bowel disease. *Therap. Adv. Gastroenterol.*, 12:1756284818822250.

Sellon, R. K., Tonkonogy, S., Schultz, M., Dieleman, L. A., Grenther, W., Balish, E., Rennick, D. M., and Sartor, R. B. (1998). Resident enteric bacteria are necessary for development of spontaneous colitis and immune system activation in interleukin-10-deficient mice. *Infect. Immun.*, 66(11):5224–5231.

Sen, P. and Orešič, M. (2019). Metabolic modeling of human gut microbiota on a genome scale: An overview. *Metabolites*, 9(2).

Sender, R., Fuchs, S., and Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS Biol.*, 14(8):e1002533.

Shaffer, M., Armstrong, A. J. S., Phelan, V. V., Reisdorph, N., and Lozupone, C. A. (2017). Microbiome and metabolome data integration provides insight into health and disease. *Transl. Res.*, 189:51–64.

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L., Solden, L. M., Liu, P., Narrowe, A. B., Rodríguez-Ramos, J., Bolduc, B., Gazitúa, M. C., Daly, R. A., Smith, G. J., Vik, D. R., Pope, P. B., Sullivan, M. B., Roux, S., and Wrighton, K. C. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.*, 48(16):8883–8900.

Shah, Y. M. (2016). The role of hypoxia in intestinal inflammation. *Mol Cell Pediatr*, 3(1):1.

Shahinas, D., Silverman, M., Sittler, T., Chiu, C., Kim, P., Allen-Vercoe, E., Weese, S., Wong, A., Low, D. E., and Pillai, D. R. (2012). Toward an understanding of changes in diversity associated with fecal microbiome transplantation based on 16S rRNA gene deep sequencing. *MBio*, 3(5):e00338–12.

Shaiber, A. and Eren, A. M. (2018). Anvi'o snakemake workflows. `https://merenlab.org/2018/07/09/anvio-snakemake-workflows/`. Accessed: 2023-3-22.

Shaiber, A., Willis, A. D., Delmont, T. O., Roux, S., Chen, L.-X., Schmid, A. C., Yousef, M., Watson, A. R., Lolans, K., Esen, Ö. C., Lee, S. T. M., Downey, N., Morrison, H. G., Dewhirst, F. E., Mark Welch, J. L., and Eren, A. M. (2020a). Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol.*, 21(1):292.

Shaiber, A., Willis, A. D., Delmont, T. O., Roux, S., Chen, L.-X., Schmid, A. C., Yousef, M., Watson, A. R., Lolans, K., Esen, Ö. C., Lee, S. T. M., Downey, N., Morrison, H. G., Dewhirst, F. E., Mark Welch, J. L., and Eren, A. M. (2020b). Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol.*, 21(1):292.

Shan, Y., Lee, M., and Chang, E. B. (2022). The gut microbiome and inflammatory bowel diseases. *Annu. Rev. Med.*, 73:455–468.

Sharma, M., Wasan, A., and Sharma, R. K. (2021). Recent developments in probiotics: An emphasis on bifidobacterium. *Food Bioscience*, 41:100993.

Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., and Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.*, 23(1):111–120.

Sharpton, T., Lyalina, S., Luong, J., Pham, J., Deal, E. M., Armour, C., Gaulke, C., Sanjabi, S., and Pollard, K. S. (2017). Development of inflammatory bowel disease is linked to a longitudinal restructuring of the gut metagenome in mice. *mSystems*, 2(5).

Shaw, S. Y., Blanchard, J. F., and Bernstein, C. N. (2011). Association between the use of antibiotics and new diagnoses of crohn's disease and ulcerative colitis. *Am. J. Gastroenterol.*, 106(12):2133–2142.

Sherrill, C. and Fahey, R. C. (1998). Import and metabolism of glutathione by streptococcus mutans. *J. Bacteriol.*, 180(6):1454–1459.

Shiozaki, T., Fujiwara, A., Inomura, K., Hirose, Y., Hashihama, F., and Harada, N. (2020). Biological nitrogen fixation detected under antarctic sea ice. *Nat. Geosci.*, 13(11):729–732.

Shreiner, A. B., Kao, J. Y., and Young, V. B. (2015). The gut microbiome in health and in disease. *Curr. Opin. Gastroenterol.*, 31(1):69–75.

Shu, W.-S. and Huang, L.-N. (2022). Microbial diversity in extreme environments. *Nat. Rev. Microbiol.*, 20(4):219–235.

Silva, Y. P., Bernardi, A., and Frozza, R. L. (2020). The role of Short-Chain fatty acids from gut microbiota in Gut-Brain communication. *Front. Endocrinol.*, 11:25.

Simon, J.-C., Marchesi, J. R., Mougel, C., and Selosse, M.-A. (2019). Host-microbiota interactions: from holobiont theory to analysis. *Microbiome*, 7(1):5.

Singh, K., Gobert, A. P., Coburn, L. A., Barry, D. P., Allaman, M., Asim, M., Luis, P. B., Schneider, C., Milne, G. L., Boone, H. H., Shilts, M. H., Washington, M. K., Das, S. R., Piazuelo, M. B., and Wilson, K. T. (2019). Dietary arginine regulates severity of experimental colitis and affects the colonic microbiome. *Front. Cell. Infect. Microbiol.*, 9:66.

Sinha, R., Abu-Ali, G., Vogtmann, E., Fodor, A. A., Ren, B., Amir, A., Schwager, E., Crabtree, J., Ma, S., Microbiome Quality Control Project Consortium, Abnet, C. C., Knight, R., White, O., and Huttenhower, C. (2017). Assessment of variation in microbial community amplicon sequencing by the microbiome quality control (MBQC) project consortium. *Nat. Biotechnol.*, 35(11):1077–1086.

Skelton, L., Cooper, M., Murphy, M., and Platt, A. (2003). Human immature monocyte-derived dendritic cells express the G protein-coupled receptor GPR105 (KIAA0001, P2Y14) and increase intracellular calcium in response to its agonist, uridine diphosphoglucose. *J. Immunol.*, 171(4):1941–1949.

Smillie, C. S., Sauk, J., Gevers, D., Friedman, J., Sung, J., Youngster, I., Hohmann, E. L., Staley, C., Khoruts, A., Sadowsky, M. J., Allegretti, J. R., Smith, M. B., Xavier, R. J., and Alm, E. J. (2018). Strain tracking reveals the determinants of bacterial engraftment in the human gut following fecal microbiota transplantation. *Cell Host Microbe*, 23(2):229–240.e5.

Soderberg, T. (2005). Biosynthesis of ribose-5-phosphate and erythrose-4-phosphate in archaea: a phylogenetic analysis of archaeal genomes. *Archaea*, 1(5):347–352.

Sohm, J. A., Webb, E. A., and Capone, D. G. (2011). Emerging patterns of marine nitrogen fixation. *Nat. Rev. Microbiol.*, 9(7):499–508.

Sokol, H. and Seksik, P. (2010). The intestinal microbiota in inflammatory bowel diseases: time to connect with the host. *Curr. Opin. Gastroenterol.*, 26(4):327–331.

Sonnhammer, E. L., Eddy, S. R., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–420.

Sorbara, M. T. and Pamer, E. G. (2022). Microbiome-based therapeutics. *Nat. Rev. Microbiol.*, 20(6):365–380.

Sorboni, S. G., Moghaddam, H. S., Jafarzadeh-Esfehani, R., and Soleimanpour, S. (2022). A comprehensive review on the role of the gut microbiome in human neurological disorders. *Clin. Microbiol. Rev.*, 35(1):e0033820.

Soufli, I., Toumi, R., Rafa, H., and Touil-Boukoffa, C. (2016). Overview of cytokines and nitric oxide involvement in immuno-pathogenesis of inflammatory bowel diseases. *World J. Gastrointest. Pharmacol. Ther.*, 7(3):353–360.

Sprong, R. C., Schonewille, A. J., and van der Meer, R. (2010). Dietary cheese whey protein protects rats against mild dextran sulfate sodium-induced colitis: role of mucin and microbiota. *J. Dairy Sci.*, 93(4):1364–1371.

Spry, C., Kirk, K., and Saliba, K. J. (2008). Coenzyme a biosynthesis: an antimicrobial drug target. *FEMS Microbiol. Rev.*, 32(1):56–106.

Stal, L. J. (2009). Is the distribution of nitrogen-fixing cyanobacteria in the oceans related to temperature? *Environ. Microbiol.*, 11(7):1632–1645.

Steen, A. D., Crits-Christoph, A., Carini, P., DeAngelis, K. M., Fierer, N., Lloyd, K. G., and Cameron Thrash, J. (2019). High proportions of bacteria and archaea across most biomes remain uncultured. *ISME J.*, 13(12):3126–3130.

Steinegger, M. and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, 35(11):1026–1028.

Stewart, C. J., Ajami, N. J., O'Brien, J. L., Hutchinson, D. S., Smith, D. P., Wong, M. C., Ross, M. C., Lloyd, R. E., Doddapaneni, H., Metcalf, G. A., Muzny, D., Gibbs, R. A., Vatanen, T., Huttenhower, C., Xavier, R. J., Rewers, M., Hagopian, W., Toppari, J., Ziegler, A.-G., She, J.-X., Akolkar, B., Lernmark, A., Hyoty, H., Vehik, K., Krischer, J. P., and Petrosino, J. F. (2018). Temporal development of the gut microbiome in early childhood from the TEDDY study. *Nature*, 562(7728):583–588.

Stulberg, E., Fravel, D., Proctor, L. M., Murray, D. M., LoTempio, J., Chrisey, L., Garland, J., Goodwin, K., Graber, J., Harris, M. C., Jackson, S., Mishkind, M., Porterfield, D. M., and Records, A. (2016). An assessment of US microbiome research. *Nature Microbiology*, 1(1):1–7.

Sul, W. J., Oliver, T. A., Ducklow, H. W., Amaral-Zettler, L. A., and Sogin, M. L. (2013). Marine bacteria exhibit a bipolar distribution. *Proc. Natl. Acad. Sci. U. S. A.*, 110(6):2342–2347.

Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Tara Oceans Coordinators, Eveillard, D., Gorsky, G., Guidi, L., Iudicone, D., Karsenti, E., Lombard, F., Ogata, H., Pesant, S., Sullivan, M. B., Wincker, P., and de Vargas, C. (2020). Tara oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.*, 18(8):428–445.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans coordinators, Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G., and Bork, P. (2015). Ocean plankton. structure and function of the global ocean microbiome. *Science*, 348(6237):1261359.

Swidsinski, A., Weber, J., Loening-Baucke, V., Hale, L. P., and Lochs, H. (2005). Spatial organization and composition of the mucosal flora in patients with inflammatory bowel disease. *J. Clin. Microbiol.*, 43(7):3380–3389.

Tamboli, C. P., Neut, C., Desreumaux, P., and Colombel, J. F. (2004). Dysbiosis in inflammatory bowel disease. *Gut*, 53(1):1–4.

Tan, T. G., Sefik, E., Geva-Zatorsky, N., Kua, L., Naskar, D., Teng, F., Pasman, L., Ortiz-Lopez, A., Jupp, R., Wu, H.-J. J., Kasper, D. L., Benoist, C., and Mathis, D. (2016). Identifying species of symbiont bacteria from the human gut that, alone, can induce intestinal th17 cells in mice. *Proc. Natl. Acad. Sci. U. S. A.*, 113(50):E8141–E8150.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., and Natale, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41.

Teng, H., Wang, Y., Sui, X., Fan, J., Li, S., Lei, X., Shi, C., Sun, W., Song, M., Wang, H., Dong, D., Geng, J., Zhang, Y., Zhu, X., Cai, Y., Li, Y., Li, B., Min, Q., Wang, W., and Zhan, Q. (2023). Gut microbiota-mediated nucleotide synthesis attenuates the response to neoadjuvant chemoradiotherapy in rectal cancer. *Cancer Cell*, 41(1):124–138.e6.

Tepe, N., Shimko, L. A., and Duran, M. (2006). Toxic effects of thiol-reactive compounds on anaerobic biomass. *Bioresour. Technol.*, 97(4):592–598.

Thompson, A. W., Foster, R. A., Krupke, A., Carter, B. J., Musat, N., Vaulot, D., Kuypers, M. M. M., and Zehr, J. P. (2012). Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science*, 337(6101):1546–1550.

Tokuhara, D. (2021). Role of the gut microbiota in regulating non-alcoholic fatty liver disease in children and adolescents. *Front Nutr*, 8:700058.

Tong, B. C. and Barbul, A. (2004). Cellular and physiological effects of arginine. *Mini Rev. Med. Chem.*, 4(8):823–832.

Tong, M., Li, X., Wegener Parfrey, L., Roth, B., Ippoliti, A., Wei, B., Borneman, J., McGovern, D. P. B., Frank, D. N., Li, E., Horvath, S., Knight, R., and Braun, J. (2013). A modular organization of the human intestinal mucosal microbiota and its association with inflammatory bowel disease. *PLoS One*, 8(11):e80702.

Tortell, P. D., Maldonado, M. T., Granger, J., and Price, N. M. (1999). Marine bacteria and biogeochemical cycling of iron in the oceans. *FEMS Microbiol. Ecol.*, 29(1):1–11.

Treem, W. R., Ahsan, N., Shoup, M., and Hyams, J. S. (1994). Fecal short-chain fatty acids in children with inflammatory bowel disease. *J. Pediatr. Gastroenterol. Nutr.*, 18(2):159–164.

Tripp, H. J., Bench, S. R., Turk, K. A., Foster, R. A., Desany, B. A., Niazi, F., Affourtit, J. P., and Zehr, J. P. (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature*, 464(7285):90–94.

Turk-Kubo, K. A., Karamchandani, M., Capone, D. G., and Zehr, J. P. (2014). The paradox of marine heterotrophic nitrogen fixation: abundances of heterotrophic diazotrophs do not account for nitrogen fixation rates in the eastern tropical south pacific. *Environ. Microbiol.*, 16(10):3095–3114.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. A., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R., and Gordon, J. I. (2009). A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484.

Turnbaugh, P. J., Quince, C., Faith, J. J., McHardy, A. C., Yatsunenko, T., Niazi, F., Affourtit, J., Egholm, M., Henrissat, B., Knight, R., and Gordon, J. I. (2010). Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc. Natl. Acad. Sci. U. S. A.*, 107(16):7503–7508.

Tyrrell, T. (1999). The relative influences of nitrogen and phosphorus on oceanic primary production. *Nature*, 400(6744):525–531.

Uchimiya, M., Schroer, W., Olofsson, M., Edison, A. S., and Moran, M. A. (2022). Diel investments in metabolite production and consumption in a model microbial system. *ISME J.*, 16(5):1306–1317.

Ungaro, R., Bernstein, C. N., Gearry, R., Hviid, A., Kolho, K.-L., Kronman, M. P., Shaw, S., Van Kruiningen, H., Colombel, J.-F., and Atreja, A. (2014). Antibiotics associated with increased risk of new-onset crohn's disease but not ulcerative colitis: a meta-analysis. *Am. J. Gastroenterol.*, 109(11):1728–1738.

UniProt Consortium (2023). UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res.*, 51(D1):D523–D531.

Utter, D. R., Borisy, G. G., Eren, A. M., Cavanaugh, C. M., and Mark Welch, J. L. (2020). Metapangenomics of the oral microbiome provides insights into habitat adaptation and cultivar diversity. *Genome Biol.*, 21(1):293.

Vaahtovuo, J., Munukka, E., Korkeamäki, M., Luukkainen, R., and Toivanen, P. (2008). Fecal microbiota in early rheumatoid arthritis. *J. Rheumatol.*, 35(8):1500–1505.

Vaisman, A., Pivovarov, K., McGeer, A., Willey, B., Borgundvaag, B., Porter, V., Gnanasuntharam, P., Wei, Y., and Nguyen, G. C. (2013). Prevalence and incidence of antimicrobial-resistant organisms among hospitalized inflammatory bowel disease patients. *Can. J. Infect. Dis. Med. Microbiol.*, 24(4):e117–21.

van Dam, V., Olrichs, N., and Breukink, E. (2009). Specific labeling of peptidoglycan precursors as a tool for bacterial cell wall studies. *Chembiochem*, 10(4):617–624.

van der Heijden, M. G. A., Bardgett, R. D., and van Straalen, N. M. (2008). The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecol. Lett.*, 11(3):296–310.

van Hoek, M. J. A. and Merks, R. M. H. (2017). Emergence of microbial diversity due to cross-feeding interactions in a spatial model of gut microbial metabolism. *BMC Syst. Biol.*, 11(1):56.

van Nood, E., Vrieze, A., Nieuwdorp, M., Fuentes, S., Zoetendal, E. G., de Vos, W. M., Visser, C. E., Kuijper, E. J., Bartelsman, J. F. W. M., Tijssen, J. G. P., Speelman, P., Dijkgraaf, M. G. W., and Keller, J. J. (2013). Duodenal infusion of donor feces for recurrent clostridium difficile. *N. Engl. J. Med.*, 368(5):407–415.

Vanni, C., Schechter, M. S., Acinas, S. G., Barberán, A., Buttigieg, P. L., Casamayor, E. O., Delmont, T. O., Duarte, C. M., Murat Eren, A., Finn, R. D., Kottmann, R., Mitchell, A., Sanchez, P., Siren, K., Steinegger, M., Glöckner, F. O., and Fernandez-Guerra, A. (2022). Unifying the global coding sequence space enables the study of genes with unknown function across biomes. *eLife*, page e67667.

Veseli, I. (2022). Updating anvi'o documentation and putting those changes online. `https://anvio.org/blog/anvio-docs-tutorial/`. Accessed: 2023-4-3.

Veseli, I. (2023). The ruegeria pomeroyi digital microbe. `https://ccomp-stc.org/rpom-digital-microbe/`. Accessed: 2023-4-3.

Veseli, I., Chen, Y. T., Schechter, M. S., Vanni, C., Fogarty, E. C., Watson, A. R., Jabri, B. A., Blekhman, R., Willis, A. D., Yu, M. K., Fernandez-Guerra, A., Fussel, J., and Murat Eren, A. (2023). Microbes with higher metabolic independence are enriched in human gut microbiomes under stress. preprint.

Veseli, I. and Cooper, Z. (2022). Ruegeria pomeroyi digital microbe databases.

Vich Vila, A., Imhann, F., Collij, V., Jankipersadsing, S. A., Gurry, T., Mujagic, Z., Kurilshikov, A., Bonder, M. J., Jiang, X., Tigchelaar, E. F., Dekens, J., Peters, V., Voskuil, M. D., Visschedijk, M. C., van Dullemen, H. M., Keszthelyi, D., Swertz, M. A., Franke, L., Alberts, R., Festen, E. A. M., Dijkstra, G., Masclee, A. A. M., Hofker, M. H., Xavier, R. J., Alm, E. J., Fu, J., Wijmenga, C., Jonkers, D. M. A. E., Zhernakova, A., and Weersma, R. K. (2018). Gut microbiota composition and functional changes in inflammatory bowel disease and irritable bowel syndrome. *Sci. Transl. Med.*, 10(472).

Vijay, A. and Valdes, A. M. (2022). Role of the gut microbiome in chronic diseases: a narrative review. *Eur. J. Clin. Nutr.*, 76(4):489–501.

Vineis, J. H., Ringus, D. L., Morrison, H. G., Delmont, T. O., Dalal, S., Raffals, L. H., Antonopoulos, D. A., Rubin, D. T., Eren, A. M., Chang, E. B., and Others (2016a). Patient-specific bacteroides genome variants in pouchitis. mbio 7.

Vineis, J. H., Ringus, D. L., Morrison, H. G., Delmont, T. O., Dalal, S., Raffals, L. H., Antonopoulos, D. A., Rubin, D. T., Eren, A. M., Chang, E. B., and Sogin, M. L. (2016b). Patient-Specific *Bacteroides* Genome Variants in Pouchitis. *MBio*, 7(6):e01713–16, /mbio/7/6/e01713–16.atom.

von Friesen, L. W. and Riemann, L. (2020). Nitrogen fixation in a changing arctic ocean: An overlooked source of nitrogen? *Front. Microbiol.*, 11:596426.

Walker, M. Y., Pratap, S., Southerland, J. H., Farmer-Dixon, C. M., Lakshmyya, K., and Gangula, P. R. (2018). Role of oral and gut microbiome in nitric oxide-mediated colon motility. *Nitric Oxide*, 73:81–88.

Walter, J., Armet, A. M., Finlay, B. B., and Shanahan, F. (2020). Establishing or exaggerating causality for the gut microbiome: Lessons from human Microbiota-Associated rodents. *Cell*, 180(2):221–232.

Wang, J., Xu, W., Wang, R., Cheng, R., Tang, Z., and Zhang, M. (2021). The outer membrane protein amuc_1100 of akkermansia muciniphila promotes intestinal 5-HT biosynthesis and extracellular availability through TLR2 signalling. *Food Funct.*, 12(8):3597–3610.

Wang, M., Noor, S., Huan, R., Liu, C., Li, J., Shi, Q., Zhang, Y.-J., Wu, C., and He, H. (2020). Comparison of the diversity of cultured and total bacterial communities in marine sediment using culture-dependent and sequencing methods. *PeerJ*, 8:e10060.

Wang, T., Goyal, A., Dubinkina, V., and Maslov, S. (2019). Evidence for a multi-level trophic organization of the human gut microbiome. *PLoS Comput. Biol.*, 15(12):e1007524.

Washburn, R. L., Sandberg, D., and Gazdik Stofer, M. A. (2022). Supplementation of a single species probiotic does not affect diversity and composition of the healthy adult gastrointestinal microbiome. *Human Nutrition & Metabolism*, 28:200148.

Watson, A. R., Füssel, J., Veseli, I., DeLongchamp, J. Z., Silva, M., Trigodet, F., Lolans, K., Shaiber, A., Fogarty, E., Runde, J. M., Quince, C., Yu, M. K., Söylev, A., Morrison, H. G., Lee, S. T. M., Kao, D., Rubin, D. T., Jabri, B., Louie, T., and Eren, A. M. (2023). Metabolic independence drives gut microbial colonization and resilience in health and disease. *Genome Biol.*, 24(1):78.

Watson, A. R., Füssel, J., Veseli, I., DeLongchamp, J. Z., Silva, M., Trigodet, F., Lolans, K., Shaiber, A., Fogarty, E., Runde, J. M., Quince, C., Yu, M. K., Söylev, A., Morrison, H. G., Lee, S. T. M., Kao, D., Rubin, D. T., Jabri, B., Louie, T., and Murat Eren, A. (2021). Donor and recipient stool metagenomes from a fecal microbiota transplantation study. `https://www.ncbi.nlm.nih.gov/bioproject/prjna701961`.

Watson, A. R., Füssel, J., Veseli, I., DeLongchamp, J. Z., Silva, M., Trigodet, F., Lolans, K., Shaiber, A., Fogarty, E., Runde, J. M., Quince, C., Yu, M. K., Söylev, A., Morrison, H. G., Lee, S. T. M., Kao, D., Rubin, D. T., Jabri, B., Louie, T., and Murat Eren, A. (2022). Metabolic independence drives gut microbial colonization and resilience in health and disease. preprint.

Weigel, B. L., Miranda, K. K., Fogarty, E. C., Watson, A. R., and Pfister, C. A. (2022). Functional insights into the kelp microbiome from Metagenome-Assembled genomes. *mSystems*, 7(3):e0142221.

Weiss, G. A. and Hennet, T. (2017). Mechanisms and consequences of intestinal dysbiosis. *Cell. Mol. Life Sci.*, 74(16):2959–2977.

Wen, C., Zheng, Z., Shao, T., Liu, L., Xie, Z., Le Chatelier, E., He, Z., Zhong, W., Fan, Y., Zhang, L., Li, H., Wu, C., Hu, C., Xu, Q., Zhou, J., Cai, S., Wang, D., Huang, Y., Breban, M., Qin, N., and Ehrlich, S. D. (2017). Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol.*, 18(1):142.

Wexler, A. G. and Goodman, A. L. (2017). An insider's perspective: Bacteroides as a window into the microbiome.

White, R. A., Callister, S. J., Moore, R. J., Baker, E. S., and Jansson, J. K. (2016). The past, present and future of microbiome analyses. *Nat. Protoc.*, 11(11):2049–2053.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci Data*, 3:160018.

Wlodarska, M., Luo, C., Kolde, R., d'Hennezel, E., Annand, J. W., Heim, C. E., Krastel, P., Schmitt, E. K., Omar, A. S., Creasey, E. A., Garner, A. L., Mohammadi, S., O'Connell, D. J., Abubucker, S., Arthur, T. D., Franzosa, E. A., Huttenhower, C., Murphy, L. O., Haiser, H. J., Vlamakis, H., Porter, J. A., and Xavier, R. J. (2017). Indoleacrylic acid produced by commensal peptostreptococcus species suppresses inflammation. *Cell Host Microbe*, 22(1):25–37.e6.

Wolfe, A. J. (2005). The acetate switch. *Microbiol. Mol. Biol. Rev.*, 69(1):12–50.

Wong, C. C., Fong, W., and Yu, J. (2023). Gut microbes promote chemoradiotherapy resistance via metabolic cross-feeding. *Cancer Cell*, 41(1):12–14.

Wong, H. L., MacLeod, F. I., White, 3rd, R. A., Visscher, P. T., and Burns, B. P. (2020). Microbial dark matter filling the niche in hypersaline microbial mats. *Microbiome*, 8(1):135.

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with kraken 2. *Genome Biol.*, 20(1):257.

Woting, A. and Blaut, M. (2016). The intestinal microbiota in metabolic disease. *Nutrients*, 8(4):202.

Wu, G., Zhao, N., Zhang, C., Lam, Y. Y., and Zhao, L. (2021a). Guild-based analysis for understanding gut microbiome in human health and diseases. *Genome Med.*, 13(1):22.

Wu, L., Tang, Z., Chen, H., Ren, Z., Ding, Q., Liang, K., and Sun, Z. (2021b). Mutual interaction between gut microbiota and protein/amino acid metabolism for host mucosal immunity and health. *Anim Nutr*, 7(1):11–16.

Wu, W., Sun, M., Chen, F., Cao, A. T., Liu, H., Zhao, Y., Huang, X., Xiao, Y., Yao, S., Zhao, Q., Liu, Z., and Cong, Y. (2017). Microbiota metabolite short-chain fatty acid acetate promotes intestinal IgA response to microbiota which is mediated by GPR43. *Mucosal Immunol.*, 10(4):946–956.

Wu, Y.-W., Simmons, B. A., and Singer, S. W. (2016). MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, 32(4):605–607.

Xie, H., Guo, R., Zhong, H., Feng, Q., Lan, Z., Qin, B., Ward, K. J., Jackson, M. A., Xia, Y., Chen, X., Chen, B., Xia, H., Xu, C., Li, F., Xu, X., Al-Aama, J. Y., Yang, H., Wang, J., Kristiansen, K., Wang, J., Steves, C. J., Bell, J. T., Li, J., Spector, T. D., and Jia, H. (2016). Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst*, 3(6):572–584.e3.

Xu, X., Thornton, P. E., and Post, W. M. (2013). A global analysis of soil microbial biomass carbon, nitrogen and phosphorus in terrestrial ecosystems. *Glob. Ecol. Biogeogr.*, 22(6):737–749.

Yang, J., Kalhan, S. C., and Hanson, R. W. (2009). What is the metabolic role of phosphoenolpyruvate carboxykinase? *J. Biol. Chem.*, 284(40):27025–27029.

Yang, S., Tang, R., Han, S., Xie, C.-J., Narsing Rao, M. P., Liu, G.-H., and Zhou, S.-G. (2023). Fundidesulfovibrio agrisoli sp. nov., a Nitrogen-Fixing bacterium isolated from rice field. *Curr. Microbiol.*, 80(2):68.

Yang, Y., Zhang, Y., Xu, Y., Luo, T., Ge, Y., Jiang, Y., Shi, Y., Sun, J., and Le, G. (2019). Dietary methionine restriction improves the gut microbiota and reduces intestinal permeability and inflammation in high-fat-fed mice. *Food Funct.*, 10(9):5952–5968.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., Knight, R., and Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227.

Ye, Y. and Doak, T. G. (2009). A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLoS Comput. Biol.*, 5(8):e1000465.

Ye, Y., Osterman, A., Overbeek, R., and Godzik, A. (2005). Automatic detection of subsystem/pathway variants in genome analysis. *Bioinformatics*, 21 Suppl 1:i478–86.

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., and Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.*, 40(Web Server issue):W445–51.

Yusufu, I., Ding, K., Smith, K., Wankhade, U. D., Sahay, B., Patterson, G. T., Pacholczyk, R., Adusumilli, S., Hamrick, M. W., Hill, W. D., Isales, C. M., and Fulzele, S. (2021). A Tryptophan-Deficient diet induces gut microbiota dysbiosis and increases systemic inflammation in aged mice. *Int. J. Mol. Sci.*, 22(9).

Zani, S., Mellon, M. T., Collier, J. L., and Zehr, J. P. (2000). Expression of nifh genes in natural microbial assemblages in lake george, new york, detected by reverse transcriptase PCR. *Appl. Environ. Microbiol.*, 66(7):3119–3124.

Zdobnov, E. M. and Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, 17(9):847–848.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., Suez, J., Mahdi, J. A., Matot, E., Malka, G., Kosower, N., Rein, M., Zilberman-Schapira, G., Dohnalová, L., Pevsner-Fischer, M., Bikovsky, R., Halpern, Z., Elinav, E., and Segal, E. (2015). Personalized nutrition by prediction of glycemic responses. *Cell*, 163(5):1079–1094.

Zehr, J. P. and Capone, D. G. (2020). Changing perspectives in marine nitrogen fixation. *Science*, 368(6492).

Zehr, J. P., Jenkins, B. D., Short, S. M., and Steward, G. F. (2003). Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Environ. Microbiol.*, 5(7):539–554.

Zehr, J. P., Montoya, J. P., Jenkins, B. D., Hewson, I., Mondragon, E., Short, C. M., Church, M. J., Hansen, A., and Karl, D. M. (2007). Experiments linking nitrogenase gene expression to nitrogen fixation in the north pacific subtropical gyre. *Limnol. Oceanogr.*, 52(1):169–183.

Zehr, J. P. and Turner, P. J. (2001). Nitrogen fixation: Nitrogenase genes and gene expression. In *Methods in Microbiology*, volume 30, pages 271–286. Academic Press.

Zehr, J. P., Waterbury, J. B., Turner, P. J., Montoya, J. P., Omoregie, E., Steward, G. F., Hansen, A., and Karl, D. M. (2001). Unicellular cyanobacteria fix N2 in the subtropical north pacific ocean. *Nature*, 412(6847):635–638.

Zhang, Y., Zhang, H., Zhang, Z., Qian, Q., Zhang, Z., and Xiao, J. (2023). ProPan: a comprehensive database for profiling prokaryotic pan-genome dynamics. *Nucleic Acids Res.*, 51(D1):D767–D776.

Zhang, Z., Zhang, H., Chen, T., Shi, L., Wang, D., and Tang, D. (2022). Regulatory role of short-chain fatty acids in inflammatory bowel disease. *Cell Commun. Signal.*, 20(1):64.

Zhao, H., Chen, J., Li, X., Sun, Q., Qin, P., and Wang, Q. (2019). Compositional and functional features of the female premenopausal and postmenopausal gut microbiota. *FEBS Lett.*, 593(18):2655–2664.

Zhou, Z., Tran, P. Q., Breister, A. M., Liu, Y., Kieft, K., Cowley, E. S., Karaoz, U., and Anantharaman, K. (2022). METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome*, 10(1):33.

Zhu, A., Sunagawa, S., Mende, D. R., and Bork, P. (2015). Inter-individual differences in the gene content of human gut bacterial species. *Genome Biol.*, 16(1):82.

Zhu, Q., Gao, R., Zhang, Y., Pan, D., Zhu, Y., Zhang, X., Yang, R., Jiang, R., Xu, Y., and Qin, H. (2018). Dysbiosis signatures of gut microbiota in coronary artery disease. *Physiol. Genomics*, 50(10):893–903.

Zimmermann, J., Kaleta, C., and Waschina, S. (2021). gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol.*, 22(1):81.

Zimmermann, M., Zimmermann-Kogadeeva, M., Wegmann, R., and Goodman, A. L. (2019). Mapping human microbiome drug metabolism by gut bacteria and their genes. *Nature*, 570(7762):462–467.

Zong, X., Fu, J., Xu, B., Wang, Y., and Jin, M. (2020). Interplay between gut microbiota and antimicrobial peptides. *Anim Nutr*, 6(4):389–396.

Zorrilla, F., Buric, F., Patil, K. R., and Zelezniak, A. (2021). metaGEM: reconstruction of genome scale metabolic models directly from metagenomes. *Nucleic Acids Res.*, 49(21):e126.

Zuo, T. and Ng, S. C. (2018). The gut microbiota in the pathogenesis and therapeutics of inflammatory bowel disease. *Front. Microbiol.*, 9:2247.