Lehrstuhl für Ernährung und Immunologie
TUM School of Life Sciences
Technische Universität München

MiMiC
A computational method for automated generation of minimal microbial consortia based on functional metagenomic profiles

Neeraj Kumar

Vollständiger Abdruck der von der Fakultät TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

April 2021

TUM School of Life Sciences der
Technischen Universität München

MiMiC- A computational method for automated generation of minimal microbial consortia based on functional metagenomic profiles

Neeraj Kumar

Vollständiger Abdruck der von der promotionsführenden Einrichtung TUM School of Life Sciences der Technischen Universität München zur Erlangung des akademischen Grades eines **Doktors der Naturwissenschaften (Dr. rer. nat.)** genehmigten Dissertation.

Vorsitzende: Prof. Dr. Lindsay Hall

Prüfer der Dissertation: **1. Prof. Dr. Dirk Haller**
2. Prof. Dr. Thomas Clavel

Die Dissertation wurde am 14.04.2021 bei der Technischen Universität München eingereicht und durch die promotionsführende Einrichtung TUM School of Life Sciences am 25.08.2021 angenommen.

Abstract

Environmental and host-associated microbial communities are complex ecosystems, of which many members are still unknown. Hence, it is challenging to study community dynamics and thus important to create model systems of reduced complexity that mimic major community functions. However, comprehensive strategies for construction of such simplified communities are lacking. ***Here we present "MiMiC" (Minimal Microbial Consortia), a bioinformatic approach that derives the composition of the original metagenomic community into a minimal microbial consortium that represents main functionalities of the ecosystem***. MiMiC uses an iterative scoring system based on maximal match-to-mismatch ratios of Pfams (protein families) between single genomes and the input metagenome. To this end, all bacterial and archaeal genomes available (~156k) until February 2019 in the NCBI (National Center for Biotechnology information) RefSeq genome database were used.

First, I built a binary matrix (presence/absence) of Pfam (n = 17,929) for each genome assembly. Further, the redundancy of the genome assemblies was reduced using reference, representative and most completeness score for the prokaryotic species. This resulted in a Pfam binary vector of reference genome database containing 45 phyla, 95 classes, 213 orders, 481 families, 2,617 genera, and 22,627 species. This was complemented by three host-specific bacterial reference genome datasets of Pfam binary vectors from pig (111 species), mouse (104 species) and human (803 species) gut-derived isolates. ***By analyzing 937 metagenomes, I demonstrated that Pfam vectorization retained enough resolution to distinguish shotgun metagenomic profiles from various environments. Metagenomes from various locations in the mouse gut (colon, ileum, cecum) and their association with healthy or inflamed conditions (susceptibility to DSS-induced colitis) formed distinct clusters based on Pfam binary vectors***.

Furthermore, I evaluated MiMiC on the MBARC mock community (a reference mixture of known microbes) and found that the corresponding minimal community inferred from the entire reference database (n=22,806 genomes) represented 97% and 92% of the original functional and taxonomic profile, respectively. Also, I evaluated the MiMiC outcome using publicly available metagenomic datasets including human, mouse and pig gut, human tongue, soil and ocean. ***Pfam matches were significantly higher (p < 0.001) in MiMiC-selected consortia than in random sets of an equal number of species in all types of metagenomes.*** In addition, I predicted 23 species as a representative consortium for pig metagenomes (n=284) from the set of PiBAC (The Pig Intestinal Bacterial Collection) species (n=111). All in all, I demonstrated that MiMiC is able to propose the minimal microbial community functionally close to the native ecosystem and validated this approach on the mock community (MBARC). All scripts and data deployed during the work are open source and can be accessed at https://github.com/ClavelLab/MiMiC.

Zusammenfassung

Umwelt- und wirtsassoziierte Mikroorganismen-Gemeinschaften sind komplexe Ökosysteme, von denen zahlreiche Mitglieder bisher noch unbekannt sind. Untersuchungen der Gruppendynamik sind daher sehr anspruchsvoll, weshalb Modellsysteme mit reduzierter Komplexität erstellt werden müssen, welche die Hauptfunktionen der Gemeinschaft imitieren können. Bisher fehlt es jedoch noch an geeigneten Strategien zur Entwicklung dieser vereinfachten Gemeinschaften. Nachfolgend stellen wir ***"MiMiC" (Minimal Microbial Consortia)*** vor, einen bioinformatischen Ansatz, der ein minimales mikrobielles Konsortium, das die Hauptfunktionen des Ökosystems darstellt, von der ursprünglichen metagenomischen Gemeinschaft ableitet. MiMiC verwendet ein iteratives Bewertungssystem, das auf dem Verhältnis zwischen maximaler Konkordanz und Diskrepanz von Pfams (Proteinfamilien) zwischen einzelnen Genomen und dem Ausgangs-Metagenom basiert. Zu diesem Zweck wurden alle Bakterien- und Archaeengenome verwendet (~156k), die bis Februar 2019 in der RefSeq-Genomdatenbank des NCBI (National Center for Biotechnology Information) zur Verfügung standen.

Zunächst habe ich für jede Genomassemblierung eine binäre Matrix (Anwesenheit/Abwesenheit) aus Pfams (n = 17.929) erstellt. Des Weiteren wurde die Redundanz der Genomassemblierungen, für die prokaryotischen Spezies unter Verwendung von Referenz-, Repräsentativ- und Vollständigkeitswerten verringert. Daraus entstand eine Referenzgenom-Datenbank für Pfams aus Binärvektoren der 45 Phyla, 95 Klassen, 213 Ordnungen, 481 Familien, 2.617 Gattungen und 22.627 Arten enthielt. Dies wurde durch drei wirtsspezifische Bakterien-Referenzgenom-Datenbanken aus Darmisolaten von Schweinen (111 Arten), Mäusen (104 Arten) und Menschen (803 Arten) ergänzt. Durch die Analyse von 937 Metagenomen konnte ich zeigen, dass die Vektorisierung der Pfams für die Unterscheidung von shotgun-metagenomischen Profilen aus verschiedenen Umgebungen ausreichend ist. Es stellte sich heraus, dass Metagenome aus verschiedenen Bereichen des Mausdarms (Kolon, Ileum, Zäkum) und Bereichen, die mit gesunden oder entzündeten Zuständen (Anfälligkeit für DSS-induzierte Kolitis) assoziiert sind, eindeutige Gruppen auf der Basis von binären Pfam-Vektoren bildeten.

Darüber hinaus habe ich MiMiC an der MBARC-*Scheingemeinschaft* (einer Referenzmischung bekannter Mikroorganismen) getestet und festgestellt, dass die entsprechende minimale Gemeinschaft, die aus der gesamten Referenzdatenbank (n = 22.806 Genome) abgeleitet wurde, 97 % des ursprünglichen funktionellen beziehungsweise 92 % des ursprünglichen taxonomischen Profils betrug. Außerdem evaluierte ich das MiMiC-Ergebnis anhand öffentlich verfügbarer metagenomischer Datensätze, einschließlich Menschen-, Mause- und Schweinedarm, menschlicher Zunge, Boden- und Meerwasserproben. In den von MiMiC ausgewählten Konsortien fanden sich mehr signifikante (p< 0,001) Übereinstimmungen zwischen Pfams als in zufälligen Sätzen gleicher Artenzahl, in allen Metagenom-Typen. Die Diskrepanzen blieben, mit Ausnahme der Schweinedarm- und Meerwasserproben, niedriger als bei zufälligen Sätzen. Darüber hinaus prognostizierte ich 23 Arten als repräsentatives Konsortium für Schweine-Metagenome aus der Gruppe der PiBAC-Arten (The Pig Intestinal

Bacterial Collection, n = 111). Zusammenfassend konnte ich darlegen, dass MiMiC in der Lage ist, die minimale Mikroorganismen-Gemeinschaft aufzustellen, die dem ursprünglichen Ökosystem funktional am nächsten ist. Außerdem konnte ich diese Methode an der Scheingemeinschaft (MBARC) validieren. Alle während dieser Arbeit bereitgestellten Skripte und Daten sind quelloffen und können unter https://github.com/ClavelLab/MiMiC abgerufen werden.

Table of contents

1 Introduction

The present PhD work focuses on the selection of prokaryotic species in any given microbial ecosystem to design personalized simplified microbial consortium based on shotgun metagenomic data. This introduction highlights background information on microbiomes, metagenomic sequencing, bioinformatics, functional annotation of microbial species, and simplified microbial communities.

1.1 Background and recent development to study microbial species via sequencing

1.1.1 Early development in microbiology

“Microbiology” is a broad research discipline that studies the diversity and biology of microbes. The invention of the microscope opened the door to the new hidden world of microorganisms in the late 17th century. ***The field of microbiology has evolved tremendously since Antonie van Leeuwenhoek was the first to attempt looking at bacterial cells through a self-invented microscope in 1663*** *(Gest 2004)*. In the early phase of microbiology, only cultured bacteria were used for detailed studies. Louis Pasteur’s discoveries about vaccinations and fermentation provided an early ground for the impression and the role of bacteria in human health and disease (Lenormant 1861; Pasteur, Chamberland, and Roux 2002; Smith 2012). Robert Koch initiated the approach of isolating bacteria being responsible agents for infectious diseases. Initially, microorganisms were classified only based on morphology and Gram-staining behaviour (Coico 2005; “Ueber Die Isolirte Färbung Der Schizomyceten in Schnitt- Und Trockenpräparaten von Dr. Gram, Kopenhagen. — Fortschritte Der Medicin 1884 No. 6. Ref. Dr. Becker” 1884). Later in the 1970’s, Carl Richard Woese proposed the use of 16S rRNA molecules and genes there to classify and characterize bacteria via sequence comparison and phylogenetic analyses (Woese, Kandler, and Wheelis 1990; Woese and Fox 1977; Escobar-Zepeda, Vera-Ponce de León, and Sanchez-Flores 2015). The 16S rRNA is a conserved molecule within prokaryotic ribosomes that can be used as a phylogenetic clock and thus helps differentiating two species thanks to heterogeneity within variable regions of the molecule (Coenye and Vandamme 2003).

1.1.2 Moving from culture-based to molecular microbial ecology via sequencing

After the development of PCR by Kary Mullis in the 1980's (Kary B. Mullis et al. 1987; K. B. Mullis 1990; Garibyan and Avashia 2013) and the construction of first 16S rRNA gene libraries by Giovannoni (Britschgi and Giovannoni 1991), ***Handelsman proposed the term “metagenome”, in the context of collective genomes of microbial species in soil environment*** (J. Handelsman et al. 1998; Jo Handelsman 2004). Metagenomics can be defined as the area of microbiology dedicated to the study of complex microbial communities via sequencing the assemblage of all genomes from community members (Liebl 2011; Quince et al. 2017). The classification of microbes via 16S rRNA gene amplicons sheds light onto the dark matter within microbiomes not accessible by cultivation methods. In other words, 16S rRNA gene amplicon based studies allowed investigating the yet uncultured fraction of microbial communities. At the same time, Carl Woese used 16S rRNA genes to study evolutionary relationships between different bacterial species and defined the Archaea as a new domain of life in 1977 (Woese and Fox 1977). Recent development and cost reduction made it convenient to sequence the whole genetic material of microbial species from an environment (http://www.genome.gov/sequencingcostsdata). Whole shotgun metagenomics made it possible to study microbes at the level of functional potentials (gene catalogues) and nowadays with strain-level resolution (Dilthey et al. 2019). ***Sequencing 16S rRNA gene amplicons or whole DNA (shotgun metagenomics) has tremendously helped discovering a substantial number of microbial species and their functions, which was not possible via cultivation methods alone***.

1.2 Sequencing approaches in microbiology

Fredric Sanger and colleagues proposed the revolutionary chain termination method in 1977 (Sanger, Nicklen, and Coulson 1977). Since then, sequencing technologies have developed tremendously. High-throughput sequencing of DNA or RNA molecules enables rapid and comprehensive exploration of whole genomes and expressed functions. In 1995, the first complete genome was sequenced, namely the genome of the bacterium *Haemophilus influenzae Rd* (Fleischmann et al. 1995). Since then, the quality and throughput of sequencing have improved and costs have decreased exponentially, allowing scientists to obtain more insights at lower costs (http://www.genome.gov/sequencingcostsdata). Sequencing has become the most fundamental research approach to obtain evolutionary and functional insights into microbial communities. Sequencing habitat-specific microbial communities opens new avenues for the development of innovative strategies in the context of human health and environmental research. Sequencing 16S rRNA genes and whole genomes from complex microbial communities allows gathering valuable information on species members and their functions, even if many of them can not yet be cultured in laboratories.

Depending on the aims of a study, several sequencing approaches can be applied. ***Sequencing only 16S rRNA amplicons gives an idea about the overall taxonomic composition and phylogenetic diversity of communities. Shotgun sequencing gives more in-depth information about the functional potential and strain-level diversity within the ecosystem*** *(Hugenholtz, Goebel, and Pace 1998)****. Metatranscriptomics gives information about expressed functions under varying environmental conditions*** *(Aguiar-Pulido et al. 2016)****.***

1.2.1 High-throughput 16S rRNA gene amplicon sequencing

The very first question usually asked while studying a microbial ecosystem is "What is its composition?". This can be addressed by amplifying and sequencing 16S rRNA genes. Unlike eukaryotes, prokaryotic ribosomes are composed of two unequal subunits (40S and

30S). The small 30S subunit contains 16S rRNA molecules that play a structural role and thereby help the translation machinery (Schluenzen et al. 2000). 16S rRNA gene amplicon sequences are evolutionarily conserved. They contain hyper-variable regions that help distinguish between bacteria down to the species level (Gray, Sankoff, and Cedergren 1984). With this feature, 16S rRNA genes are used as chronological markers and to infer phylogenetic distances between microorganisms. 16S rRNA genes are approximately 1,500 bp-long (Clarridge 2004). Prior to sequencing, variable regions need to be amplified by PCR. This can be done by using a variety of more or less universal primers, which substantially influence the outcomes of analysis (Klindworth et al. 2013). Obtaining 16S rRNA profiles involve several steps from the wet lab to computational analysis, as follows:

- Sample collection and metagenomic DNA isolation. During these steps, freeze-thaw cycles are known to influence microbial profiles and the lysis of prokaryotic cells is a critical parameter.
- Amplifying hyper-variable regions followed by PCR cleanup and equimolar pooling: during this step, the choice of primers is important and the number of PCR cycles should be kept to a minimum (Bonnet 2002).
- Sequencing using next-generation machines. Technical parameters such as sequencing length, depth and coverage vary between technologies (Illumina, PacBio, Nanopore).
- Bioinformatic analysis of generated sequencing data: computation involves several steps from preprocessing of raw sequences to the generation of final readouts. Reads are usually first demultiplex (assigned to their sample of origin), quality checked and potentially assembled if paired-end sequencing was performed. Amplicon Sequence Variants or Operational Taxonomic Units are built and further analyzed in terms of alpha and beta-diversity as well as taxonomic composition (Edgar 2013; Callahan et al. 2016).

Alpha-diversity is used to represent the diversity of species within a given environment (Whittaker 1960, 1972). Species richness is the number of observed molecular species found in a specific sample. Alpha-diversity parameters such as the Shannon-index and Simpson-index do take into account the abundance of individual species (Whittaker 1972; Shannon 1948; Spellerberg and Fedor 2003). Relative abundance is used to appreciate the level of occurrence of species and allows comparison between samples (Hubbell 2001; McGill et al. 2007). Beta-diversity is used to explain the level of relatedness between two ecosystems based on their microbial makeup (Whittaker 1960). Phylogenetic trees can be visualized based on 16S rRNA gene amplicon data. Evolutionary relationships between ecosystem members can be studied using phylogenetic trees.

16S rRNA gene amplicon sequencing made it possible to identify and estimate the occurrence (relative abundance) of species members in complex microbial ecosystems. Eztaxon (Yoon et al. 2017), SILVA (Pruesse et al. 2007), and RDP (Ribosomal Database Project) (Cole et al. 2014) are major platforms providing databases and tools for 16S rRNA gene amplicon-based analysis. QIIME2, Mothur, DADA2 are few important bioinformatic pipelines to analyze 16S rRNA gene amplicon data from row read to OTU table, taxonomy, and downstream statistical analysis (Bolyen et al. 2019; Schloss et al. 2009; Callahan et al. 2016). IMNGS (Integrated Microbial Next Generation Sequencing), is a web-based platform for the integration and processing of 16S rRNA gene amplicon sequencing data based on the UPARSE pipeline (Edgar 2013; Lagkouvardos, Joseph, et al. 2016). Rhea is a R based pipeline for statistical analysis of OTU tables generated by IMNGS or other programs. (Lagkouvardos et al. 2017).

1.2.2 Metagenomics

16S rRNA gene amplicon sequencing allows classifying prokaryotes, but the functions encoded within the communities are still unknown. In contrast, whole shotgun genome sequencing of individual species and subsequently collective sequencing of whole metagenomic DNA present in ecological systems give the opportunity to study community functions. Due to extensive development in the cultivation of anaerobes and in high-throughput sequencing techniques, hundreds of microbial species are being isolated and described every year. Whole genome sequencing of bacterial isolates has made it easier to understand the functional capacity of a bacterium and the evolutionary and functional relatedness with other species within an environment (Wooley, Godzik, and Friedberg 2010).

After the Sanger sequencing method (Sanger and Coulson 1989; Sanger, Nicklen, and Coulson 1977), many sequencing technologies have emerged such as pyrosequencing (Nyrén, Pettersson, and Uhlén 1993) (first NGS machine from Roche), reversible terminator sequencing by synthesis (Canard and Sarfati 1994) (Illumina), Nanopore sequencing (X. Sun et al. 2020) and Pacific Biosciences (Coupland et al. 2012; Sharon et al. 2013). ***Despite progress made in the ability to cultivate single species and to study their genomes, it is still challenging to obtain each and every single microbial species from a given ecosystem. Shotgun metagenomics can help here: using computational methods, it is possible to assemble short reads into species and even strain-specific individual genomes, also referred to as MAGs (metagenome-assembled genomes)*** *(Parks et al. 2018; Pasolli et al. 2019; Plaza Oñate et al. 2019)****.*** For shotgun approaches, protocols consist of several steps that must be tightly controlled in order to generate high-quality sequence data. (Costea et al. 2017). The standard protocol of metagenomic starts from isolating metagenomic DNA, preparing library sequencing, and eventually bioinformatic analysis. The latter steps involve preprocessing, assembly, taxonomic assignment of reads and functional annotation. The preprocessing step involves quality control, demultiplex, adapter removal, filtering host-specific reads.

Using computational methods, it is possible to determine genes, functional pathways and biological interactions between community members. However, proper metagenomic data processing is highly dependent on the quality of reference databases. In contrast to 16S rRNA gene amplicon sequencing reads, metagenomic data generate enormous amounts of reads which have to be processed on a unix/Macintosh based operating system with higher power of random access memory (RAM). Bioinformatic pipelines are combinations of bioinformatic tools used for step by step processing of sequencing data. Metagenomic pipelines can be customised based on the specific requirement and choice of different bioinformatic tools. Conda allows users to install different bioinformatic softwares in one package used for metagenomics under the channel named bioconda (Grüning et al. 2018). MetaPhaln2 and kraken2 are widely used tools for taxonomic classification (Truong et al. 2015; Wood and Salzberg 2014; Wood, Lu, and Langmead 2019). MetaPhaln2 uses predefined marker genes to assign reads to species whilst Kraken/kraken2 uses complete microbial genomes from NCBI. MEGAHIT and metaSPAdes are main tools to assemble metagenomic reads into contigs (Nurk et al. 2017; D. Li et al. 2015). MetaBAT2 is one of few tools to assign contigs into draft genomes, a process referred to as genome binning (Kang et al. 2015). The MOCAT2 pipeline provides a range of tools including preprocessing, quality control, assembly, taxonomy, and functional annotation (Kultima et al. 2016). HUMAnN2 and prokka are pipelines that provide a functional insights into a metagenome by assigning reads to conserved domain sequence, proteins, protein families, metabolic enzymes, cluster of orthologous genes (COGs) and metabolic pathways (KEGG) (Franzosa et al. 2018; Seemann 2014). MG-RAST has been used as a web-based tool for submitting, processing and annotating metagenomes (Meyer et al. 2008). Different tools and pipelines have their own pros and cons in terms of computational requirement, time required for processing, databases and algorithms used. As an example mapping reads to protein using standard BLASTX will cost a lot of computational timing and source compared to DIAMOND blast (Altschul et al. 1990; Buchfink, Xie, and Huson 2015). ***A benchmark of different tools designated for the same purpose can help to accommodate and choose the right tool for processing or analysing the data*** *(Ye et al. 2019; Fritz et al. 2019)*. Taxonomic abundance information from metagenomic data can be analyzed in a similar way as 16S rRNA amplicon data. Species richness, alpha-diversity and beta-diversity can also be calculated from metagenomic data.

1.3 NCBI sequence database

NCBI is a division of the National institute of Health (NIH) and hosts several databases, bioinformatic tools. GenBank is the primary sequence database of NCBI which interchange sequence data on daily bases with two other collaborative centers European Molecular Biology laboratory (EMBL) and DNA Databank of Japan (DDBJ) (Benson et al. 2012). Sequence Read Archive (SRA) is the database for high-throughput short read sequencing data hosted by NCBI. High throughput sequence data have to be submitted to SRA or European Nucleotide Archive (ENA) prior to publishing the study. GenBank does not accept nucleotide sequences less than 200 base pairs. Being as a primary database, GenBank allows multiple entries for the same sequence (loci). Whole genome shotgun (WGS) database accepts the genome sequence for prokaryotic and eukaryotic genomes. ***To overcome the redundancy of the same species and quality standard, a separate reference sequence database was built and named as NCBI RefSeq database. NCBI RefSeq allows one sequence per loci filtering from GenBank.*** NCBI RefSeq (Pruitt, Tatusova, and Maglott 2005) contains well-curated whole genome sequences of bacterial and archaeal species from several hosts and ecological environments. The number of full genomes of bacteria, archaea and fungi has increased exponentially in recent years, which allows more detailed functional characterization and annotation of microbial species (Land et al. 2015).

From the human gut alone, genomes from thousands of bacterial isolates are available and the number is increasing each year (Bilen et al. 2018; Zou et al. 2019). Despite the advancements of culture and sequencing techniques, there is still a long way to uncover all still unknown microbial species and to understand the functional contribution of each in complex ecosystems (Peisl, Schymanski, and Wilmes 2017). NCBI RefSeq genome database makes all microbial genomes available as open source. Microbial genomes in NCBI RefSeq are annotated using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) (Haft et al. 2018; Tatusova et al. 2016). ***RefSeq follows a defined quality control standard and maintains high-quality genome assemblies and annotations. Refseq excludes the assemblies from metagenomes, mixed cultures, single-cell genomes, undefined genera, low-quality sequence and not meeting a certain cutoff of genome length (<15Mb), assembly quality (L50 >500, N50 <5000 or N50 > 2000)***. ***Only genome sequences from pure cultures are taken into account to reduce the risks of any biological contaminations***. Moreover, Refseq provides the list of assemblies as reference or representative genome assemblies following certain criterias. Reference genomes are chosen based on experimental backup, sequence and annotation quality. Representative genomes are assigned based on manual curated information about the assembly, lowest standard deviation from the average assembly length of the available assemblies of the species. Non-reference and non-representative genome assemblies cover a majority of databases and were kept to observe the strain-level variations within the species. ***Following all these criteria and strict quality standards, RefSeq provides the best quality genomes of microbial species, which were utilized in the present work to build a reference genome database and thereby predict minimal microbial consortia.***

1.4 Computational methods to define microbial function

The functional profile of a given metagenome provides insights into what microbes are capable of doing in an ecosystem. The functional profile of metagenomes can be assessed by several computational methods, including coding sequence (CDS prediction), gene ontology (GO) (T. G. O. Consortium and The Gene Ontology Consortium 2008), KEGG pathway (M. Kanehisa 1997; M. Kanehisa and Goto 2000; Minoru Kanehisa et al. 2006), protein family (Pfam) (Finn et al. 2014), cluster of orthologs (COGs) (Tatusov, Koonin, and Lipman 1997) and EggNOG (Jensen et al. 2008).

1.4.1 Gene prediction and protein annotation

The very first step in the process of assigning functions is to identify microbial genes. The basic principle involves identifying open reading frames flanked by start and stop codons. A number of tools like Glimmer (Delcher et al. 2007), GenemarkHMM (Lukashin 1998), EASY genes (Schou Larsen and Krogh 2003), MED (Zhu et al. 2007), Prodigal (Hyatt et al. 2010) and Prokka (Seemann 2014) exist for gene prediction. Prodigal provides an improved and better prediction in terms of finding translation initiation sites, and less false positives ORFs. Protein annotations from gene prediction tools are further used to map on existing HMM models of COGs or Pfams or metabolic pathways.

1.4.2 Cluster of Orthologs

COGs are clusters of orthologous proteins based on pairwise alignment of protein sequences among the genomes from different clades. Orthologs are genes which are similar by structure and oftenly have the similar functions in an evolutionary diverse species. Most often, orthologs occur due to speciation during evolution by retaining the similar function from the common ancestral genes (Fitch 1970; Fang et al. 2010). COGs are useful to characterize functions of newly sequenced genomes of species and proteins by mapping them to the models of known ortholog clusters. The basic rule of COGs is that a conserved gene should be orthologous in at least 3 distantly related species which can be expanded further to find a match to another group until there is no further hit found (Tatusov, Koonin, and Lipman 1997). Each cluster of orthologs associated with conserved and specific function which consequently lead to predict function prediction of a protein to be compared. COGs are assigned into 20 functional categories, which makes it easier to interpret results. Initially, COGs were developed using seven complete genomes, *Escherichia coli, H. influenzae*, *Mycoplasma genitalium, Mycoplasma pneumoniae*, *Synechocystis sp*., *Methanococcus jannaschii*, and *Saccharomyces cerevisiae* from 5 different phylogenetically distant related phyla (Tatusov, Koonin, and Lipman 1997), which eventually developed further into a COG database and is accessible via the NCBI database (Tatusov and Fedorova 2003).

COGs are the first initiative of its kind which were later used by several developers and a number of databases were developed based on the similar approach (Galperin et al. 2015, 2020; Kristensen, Wolf, and Koonin 2017; Makarova, Wolf, and Koonin 2015). Based on a concept similar to COGs, the EggNOG (evolutionary genealogy of genes of Non-supervised Orthologous Group) database was created using a higher number of complete genomes from bacteria and domains like eukaryotes and archaea (Huerta-Cepas et al. 2016; Jensen et al. 2008; Muller et al. 2010; Powell et al. 2012, 2014; Huerta-Cepas et al. 2019). Each cluster of orthologous proteins is created as a Hidden Markov Model (HMM) (Eddy 2004).

1.4.3 Protein family

Pfam profiles are HMMs of evolutionarily related proteins sharing similar functional domains (Sonnhammer et al. 1998; Sonnhammer, Eddy, and Durbin 1997). The Pfam represents multiple sequence alignment of well curated proteins with known functions and HMM profile to identify or assign function to a protein with unknown function. An extensive development has been taken over the time in the Pfam database and it’s number of protein families (Bateman et al. 2002; Finn et al. 2006, 2008, 2014; El-Gebali et al. 2019). The HMM profiles are made by HMMER using seed alignment (Durbin et al. 1998). The seed alignment is generated and updated based on experimentally validated proteins available in the latest updated version of uniprotKB database (“UniProt: A Hub for Protein Information” 2015; Finn et al. 2016). Pfams are classified a step further as clans (Sammut, Finn, and Bateman 2008). Clan is a group of similar Pfam which share similar functional domains among different protein families and have emerged from a single evolutionary origin (Sonnhammer, Eddy, and Durbin 1997). About 75% of Pfams are classified to a clan by 2019 (El-Gebali et al. 2019; Finn et al. 2016). Sequence motifs and tertiary structure similarities are the main criteria used to assign Pfams into clans (“UniProt: A Hub for Protein Information” 2015).

An enormous number of sequence data is being generated and it is not possible to validate each single protein's function experimentally. ***Pfam models are very handy to assign functions based on similar functional domains present in the uncharacterized or unannotated new proteins. At the same time, it is computationally quicker to annotate protein using Pfam models than mapping protein sequences against each other within a comprehensive database*** *(Sonnhammer, Eddy, and Durbin 1997; El-Gebali et al. 2019)***.** Assigning more weight to the match in conserved regions helps to detect distant related homology proteins and thus able to assign function to the poorly annotated genomes. Prosite (Sigrist et al. 2010), Print (group of conserved motifs) (Scordis, Flower, and Attwood 1999; Attwood et al. 2003), SMART (based on domains from signaling, extracellular and chromatin related proteins) (Letunic, Doerks, and Bork 2009), InterPro (combined information from multiple database) (R. Apweiler et al. 2001), CDD (conserved domain database including 3-D structure information of protein) (Marchler-Bauer et al. 2013) are some of the databases that used the similar concept of conserved domain in protein families.

1.4.4 Metabolic pathways

Metabolic pathway is a series of biochemical reactions interconnected with the enzymatic reaction within a living cell (Fisher 2001). Use or consumption of energy in bacterial cells depends on the metabolic pathways occurring in the cell. Metabolic pathways are catalysts by enzymes that are proteins itself. Predicted proteins which are enzymes can be annotated to metabolic pathways which provide an insight of microbial species responsible for biochemical reactions. KEGG (Kyoto Encyclopedia of Genes and Genomes) provides the detailed information of interconnected metabolic pathways and pathways involved in transcription, translations, replication, membrane transport, signal transduction, cell growth, cell membrane, immune system, human disease and drug development (M. Kanehisa 1997; M. Kanehisa and Goto 2000; Minoru Kanehisa et al. 2006; Xiao et al. 2015). The BioCyc database is another major source of metabolic pathways which are further categorized into several branches. MetaCyc is the major branch of BioCyc database comprising a large number (2,749) of metabolic pathways (Caspi et al. 2020).

1.5 Microbiome: profiling the whole microbial community from an ecosystem

A microbiome is defined as the overall community of bacteria, archaea, fungi and viruses within a specific habitat together with their genomes and surrounding environmental factors (Highlander 2013; Marchesi and Ravel 2015). ***It represents a symbiotic system of microorganisms in a specific environment or host*** *(Lee Lerner and Lerner 2003)***. *Microbiomes are environment-specific and tend to be relatively stable at high taxonomic***

levels (e.g. phyla, families) without marked changes in environmental conditions, whilst more variability has been observed at the species level. For instance, although various species of the genus *Bacteroides* are found universally among healthy human individual's colon such as *Bacteroides fragilis*, *Bacteroides oralis*, *Bacteroides melaninogenicus,* relative abundance profiles of the same species may vary depending on environmental conditions. However, the functional profiles of an ecosystem has been observed to have the least variance in spite of varying taxonomic profiles, partly due to functional redundancy but also our inability to properly annotate substantial parts of sequence data (Human Microbiome Project Consortium 2012).

Until recently, it was considered that the human microbiome represents 10 times more cells than own eukaryotic cells in the human body itself (Luckey 1972). This assumption was revisited in recent years, suggesting rather comparable cell numbers (Sender, Fuchs, and Milo 2016). In a recent study it was found that there are a total of 45,666,334 non-redundant genes present in the human microbiome (Tierney et al. 2019). Microbiota is an inevitable part of human health. Microbial composition also plays a major role at molecular level such as bile acid metabolism (Ramírez-Pérez et al. 2017).

1.6 Roles of the Microbiome in human health and diseases

Microbial communities are an intrinsic part of human health. The human gut harbors complex communities of commensal microbial species which benefit humans in a mutualistic way. ***The human gut microbiota helps training the immune system, prevents colonization by pathogenic strains, and can metabolize indigestible food components such as carbohydrates and thereby produce short-chain fatty acids*** *(Cummings JH Cummings 1981 JH)*. For instance, several members of the genera *Bacteroides* and *Bifidobacterium* are specialized in the degradation of dietary fibers and complex carbohydrates (Larsbrink et al. 2014; Morowitz, Carlisle, and Alverdy 2011). Other bacteria are known to produce certain vitamins such as vitamin B or vitamin K (B. Wang et al. n.d.2017).

Shifts in the gut microbiota have been associated with specific health conditions. Obesity or inflammatory bowel diseases (IBD), including Crohn's disease and ulcerative colitis, have been linked for instance to dysbiosis of the gut microbiota, defined as structural or functional changes in the ecosystem which were shown in some cases to play a causative role in disease (DiBaise et al. 2008; Mukhopadhya et al. 2012). ***Firmicutes and Bacteroidetes are major phyla within the human gut microbiome, the occurrence of which is altered in IBD*** *(Sha et al. 2013; Nemoto et al. 2012; Lloyd-Price et al. 2019)*. Alterations of the gut microbiota can influence the immune system at the intestinal epithelial surface, thereby affecting mucosal barrier and favoring uncontrolled immune responses to microbial antigens (Coskun 2014; Frank et al. 2011; Maloy and Powrie 2011).

The gut microbiota also plays a role as mediator between the gastrointestinal tract and the nervous system, also known as the gut-brain axis (Y. Wang and Kasper 2014; Sudo et al. 2004; Mayer et al. 2014). The taxonomic compositions of the gut microbiota differ significantly between healthy individuals and patients with major depressive disorder (MDD) (Zheng et al. 2016). Changes in the gut microbiota were also reported in the case of autism and Parkinson's disease, albeit without causal relationships demonstrated yet (Buie 2015; Sarkar et al. 2016; Dinan and Cryan 2015).

In terms of intervention, a healthy gut microbiota can be used to beneficially alter the disturbed ecosystem via transplanting it to a subject with disease. ***Fecal microbiota transplantation (FMT) has been used to successfully treat patients with chronic infection by the Gram-positive pathogen Clostridioides difficile*** *(van Nood et al. 2013; Moayyedi et al. 2017)*. Other types of interventions include prebiotics that may act by increasing the production of short-chain fatty acids (Macfarlane, Macfarlane, and Cummings 2006).

1.7 Simplified microbial community

Figure 1: Mimicking an ecosystem with a minimum number of species. Concept of MiMiC representing the complex microbial ecosystem (left) with the minimum bacterial community (right).

A synthetic microbial consortium (or community) can help to understand the symbiotic relationships between species in a defined environment. The term "consortium" was used for the first time by Johannes Reinkee in 1872 (Reinke 1872; Kull 2010). A few computational methods have been developed to design and study minimal microbial consortia using different approaches like dynamic models and flux balance analysis. Dynamic models can predict the richness of species over the time in a consortium (McCarty and Ledesma-Amaro 2019). ***Flux balance analysis (FBA) simulates metabolic networks of microbial communities based on genome-scale reconstruction*** *(Francke, Siezen, and Teusink 2005)*. FBA can predict the growth of bacteria under a certain source of energy and vice-versa. The genomes scale metabolic (GSM) reconstruction are built based on genome sequence of the organism using genes, metabolic pathways, enzymes and their targeted biochemical reaction, and metabolites. GSM predicts the growth, molecular mechanism, and metabolic pathways interaction of the respective organism and it can be constructed on prokaryotes and eukaryotes (Thiele and Palsson 2010). Information from many databases are integrated into one to build the metabolic reconstruction of an organism. KEGG, BioCyc, MetaCyc, ENZYMES (Bairoch 2000), BRENDA (Jeske et al. 2019), Bigg (King et al. 2016) are major databases from which the information about pathways, corresponding enzymes for a protein and organism, and already predicted GSM models are merged together into a mathematical model. Pathway Tools (Karp et al.

2010), ModelSeed (Henry et al. 2010), CoReco (Pitkänen et al. 2014; Castillo et al. 2016), are few major bioinformatic tools used to reconstruct GSM models.

Symbiotic relationships within microbial communities depends on the syntrophy of metabolites between two species (Schink and Stams 2013). Stoichiometric metabolic models of metabolic networks provide an insight of mutualism and adaptation of species to an environment by predicting metabolic fluxes (Stolyar et al. 2007). ***Synthetic microbial consortia have applications in bioremediation, biofuels and biotechnology to predict the effect of certain species and metabolites in metabolic reaction.*** However, these models can predict the behaviour of microbial species in a consortium whilst selecting the species at first place for a synthetic consortium is still a challenge. Most of the computational tools have been developed to model a consortium of knowledge based selected species. Which means these methods are applied in the second step of designing microbial consortia. At the first step, a set of species are needed with a well known function mimicking corresponding functions of corresponding environment.

Due to advancement in the sequencing field it is possible to sequence the whole microbial community from an ecosystem at once by high-throughput technologies instead of only relying on conventional culture-based methods. This opens the door to an uncultured microbial world whose function and abundance can be known via metagenomics (Hugenholtz, Goebel, and Pace 1998). This provides an opportunity of more options to choose microbial species for a microbial consortium mimicking or modeling a corresponding microbiota.

A microbiota provides a key to model an ecosystem by selecting the species depending on their cumulative functions based on its metagenomic profile. ***It is well-known that, in an ecosystem, a single species does not carry all the functions by itself, but rather the functional potential is distributed among the multiple species in an interdependent manner*** *(Johns et al. 2016)*. Due to the inherent complexity of native microbial communities and the still substantial dark matter (Filee et al. 2005; Lloyd et al. 2018) (i.e. unknown fraction) within them, it is very difficult to study microbiomes in comprehensive manners. Hence, there is a great value in building representative yet simplified models of complex native ecosystems (Babaei et al. 2018; Baldini et al. 2019; Sen and Orešič 2019). This is where minimal microbial consortia consisting of a limited number of known species are useful and allow deciphering functional interactions between microbes and their environment. A minimal microbial consortium can also be designed for specific purposes choosing microbes with certain metabolic functional properties (Brenner, You, and Arnold 2008; Song et al. 2014). ***Our approach (MiMiC) provides an essential step of selecting the minimum number of species that represent the maximum functional profile of a corresponding environment***. This can be used further for studying species behaviour and interaction via metabolic networks or genome scale metabolic reconstruction using existing methods like flux balance analysis.

1.8 Example of existing minimal microbial consortia

Several minimal microbial consortia of gut microbes have been defined, including the Altered Schaedler Flora (ASF) (Dewhirst et al. 1999) with 8 species, the Simplified Human Intestinal Microbiota (SIHUMI) (Becker et al. 2011) with 7 species, and most recently the Oligo-MM (Brugiroux et al. 2016) with 12 species in its basic conformation. Whilst the ASF consists of isolates from the mouse gut and can be used to model murine microbiota, SIHUMI can be used for the human intestinal microbiota. The OligoMM is helpful for understanding mechanisms underlying resistance to enteric infection by *Salmonella enterica* serovar Typhimurium (Brugiroux et al. 2016). Simplified microbial communities were proposed even for plant roots (Zhang et al. 2019) and soil to understand biochemical interactions among the microbial communities (Zhalnina et al. 2018) and how beneficial they could be for the growth of certain plants. Several studies had been carried out in an attempt to understand the microbe-plant interaction. It is essential to understand the mechanism of community dynamics for the microbe's benefits for the betterment of plants (Niu et al. 2017). Nie *el al 2017* proposed 7 strains and studied the effect of individual species in this community to see how important the keystone species could be in certain communities. Other simplified communities have been proposed in sugarcane (Armanhi et al. 2017) and *Arabidopsis thaliana* (Herrera Paredes et al. 2018).

Although useful, the minimal microbial consortia mentioned above were designed by hand-picking species according to expert knowledge about ecosystems and the functions carried out by the respective species and driven by the ability to grow strains in the lab. Hence, considering the tremendous diversity between individual microbiomes, they can not be used for the purpose of individual studies. For instance, it has been observed that mice from different facilities tend to be colonized by different microbiota with varying functional profiles (Alexander et al. 2006).

1.9 Mock communities

A mock community is a predefined mixture of microbial species used for standardization purposes and to simulate defined microbiota in vitro *(Highlander 2015)*. Since microbial composition, abundance are not known in an ecosystem prior to sequence, it is difficult to estimate the accuracy of *in vitro* and *in silico* methods used to predict the microbiota. For example, if a species is very low abundant in an ecosystem, there are few chances to get the whole genome level information from metagenomic sequences unless the sequencing depth and coverage is high. Mock communities help optimize analysis parameters and estimate the reliability of predictions made from metagenomic data. The human microbiome project (HMP) has provided a number of microbial mock communities which can be used for streamlining bioinformatic analysis pipelines (Human Microbiome Project Consortium 2012; “The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease” 2014).Mock microbiota is another database consisting of 11 mock communities including bacterial, archaeal and eukaryotic communities (Bokulich et al., n.d.2016). MBARC-26 (Mock Bacteria ARchaea Community

MBARC-26) is a mixture of 23 bacteria and 3 archaeal strains with complete reference genomes available (Singer et al. 2016). These species cover 10 different phyla and 14 classes, including different abundances with very high coverage and long reads using Illumina and PacBio sequencing. This mock community has been used in the present work to validate computed prediction of minimal microbial consortium.

2 Hypothesis and Aim of the dissertation

The primary aim of this dissertation was to simplify the inherent complexity within microbiota in order to generate detailed functional insights and being able to generate individualized minimal microbial consortia that depict the majority of the functional profile within a given metagenome. There are several computational models that can simulate interactions between microbes in a preselected set of microbial species. In most of the cases, these models of synthetic microbial consortia are built based on the taxonomic composition of species or common knowledge on their general functions. In contrast, there is a lack of computational methods to select species for a minimal microbial consortium that represents the functional profile of the entire ecosystem. ***We hypothesized that, despite the high diversity of microbial species that coexist in an ecosystem, a minimum number of them which cover the maximum functionality of the respective ecosystem can be used as functionally representative and thus as backbone for minimal microbial consortia. We thus explored approaches on how a minimal microbial consortium can be designed based on the functional profile of an input metagenome.*** The second goal was to make the method freely available as a bioinformatic tool to the scientific community.

Accordingly, the work within this dissertation was carried in a series of specific aims. ***The first objective*** was to develop a reference genome database usable as a comprehensive reference to construct minimal consortia as individual sets of microbial species. In this part of the dissertation, I explored the genomes availability of microbial species (bacteria and archaea) in NCBI and developed a curated reference database with the functional profile as protein family (Pfam) binary vector (presence and absence). At the same time, I made a separate dataset of host-specific microbial species such as pig (PiBAC), mice (miBC) and human gut. ***Following this,*** I investigated how good Pfams were to distinguish different ecosystems and established it as a fundamental unit of the method to generate function-based minimal microbial consortia. **In the third phase,** I explored different ways of scoring the species for their selection in order to obtain the best possible set that covers the original metagenomic functions. ***For validation***, I applied the final version of the scoring method on a mock community (i.e. a mixture of known composition) and 937 metagenomes from diverse ecosystems. ***In the last section,*** I applied MiMiC to generate minimal microbial consortia for two animal species of great importance in research: pigs and mice. MiMiC is proposed as an open source bioinformatic pipeline providing a series of shell, perl and R scripts to generate minimal microbial consortia from metagenomic dataset.

3 Materials and Methods

Sequencing the whole microbial DNA present in an ecological system provides detailed information about the microbes and their functions. Reads generated via high-throughput sequencing need to be processed to eliminate artifacts which may cause biases in the generation of taxonomic as well as functional profiles. Here, I present a method to propose a minimal microbial consortium from a given metagenome. The publicly available MBARC mock community was used to validate the approach. Several publicly available metagenomic datasets were taken from human, mice, pig, soil and ocean to investigate the Pfam-based profiling and inferred minimal microbial consortia variations within and between the ecosystems. The sources and further information on all these data are provided in detail below. This method section is divided into following parts.

- Description of metagenomic datasets used in the dissertation.
- Pfam-based reference genome database.
- Processing of metagenomes and genomes.
- Scoring strategies.
    - MiMiC-score.
    - knee point calculation to determine the diversity of minimal microbial consortia.
    - Summary statistics generated from for minimal microbial consortia.
- Methods for visualization of Pfam-based classifications.

3.1 Metagenomic datasets

Metagenomes from different hosts, environmental habitats, and mock community were selected from public repositories or published studies. A diverse collection of metagenomes was considered to see if the functional profile of microbial communities based on Pfam could be distinguished. The prediction of minimal microbial consortia based on metagenomic Pfam profiles were expected to be specific to the respective ecosystem. The following section provides detailed information on the final set of 937 metagenomes studied (Table 1).

3.1.1 Mock community

A mock community is a mixture of genomic DNA from already known microbial species with complete genome; it can be used to benchmark bioinformatic and *in vitro* methods. I selected the MBARC-26 (Mock Bacteria Archaea Community) (Singer et al. 2016) mock community to test the performance of MiMiC (https://www.ebi.ac.uk/ena/browser/view/SRX1836716). It contains the species with complete reference genome sequence available in the database, which makes it more reliable to benchmark this approach.

3.1.2 Host-derived metagenomes

- Human

Metagenomes were selected from the Integrative Human Microbiome Project (iHMP) (“The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease” 2014). The iHMP project was initiated to establish and characterize human microbiota and its role in human health and disease. The iHMP data portal provides access to metagenomic assemblies from different parts of the human body. Already processed Metagenomic assemblies from different body sites were downloaded (https://portal.hmpdacc.org/). Metagenomes were considered only from healthy humans. The data was downloaded for feces and tongue.

- Mice

In research, mice are extensively used as a model organism to mimic the function, disease, and drug responses in humans (Rosenthal and Brown 2007). Mouse gut metagenomes were selected from the study of mouse gut gene catalog (Xiao et al. 2015) which included 184 mouse metagenomic samples from diverse strains, genetic background, providers and housing laboratories. The samples were sequenced with Illumina HISeq 2000 as single and paired-end layouts. The metagenomic files (paired-end) were downloaded from the ENA under the project ID PRJEB7759 (https://www.ebi.ac.uk/ena/browser/view/PRJEB7759).

- Pig

Pig is also a model organism often used for human health and disease related studies (Bassols et al. 2014). Pig gut metagenomes were taken from gut gene catalog study. There are 287 pig samples sequenced by Illumina HiSeq 2000 under project ID PRJEB11755 in ENA (Xiao et al. 2016). Paired-end libraries were downloaded from the link “Submitted FTP” (https://www.ebi.ac.uk/ena/browser/view/PRJEB11755). Metagenomes were sequenced from three countries China, Denmark, and France.

3.1.3 Environmental metagenomes

In order to test the approach on diverse ecosystems, environmental metagenomic samples from ocean and soil were also used in this study.

- Ocean

Ocean data was collected from the study of global ocean atlas of eukaryotic genes (Carradec et al. 2018). The samples were collected from different sites world wide from the surface water and deep surface chlorophyll maximum (DCM). Samples were sequenced with Illumina HiSeq 2000 in paired-end mode. The assemblies of processed data were downloaded from ENA under the project ID “PRJEB4352” (https://www.ebi.ac.uk/ena/browser/view/PRJEB4352).

- Soil

I performed a search in NCBI assemblies for soil metagenomes and downloaded all 100 pre-assembled metagenomes from different studies (Tringe et al. 2005; Meier, Paterson, and Lambert 2016; Johnston et al. 2016; Mitchell et al. 2018; Van Goethem et al. 2018; Carlos, Fan, and Currie 2018).

3.1.4 Healthy vs disease susceptible metagenome

Metagenomes from DSS-colitis susceptible and non-susceptible mice were used to generate minimal microbial consortia (Roy et al. 2017). Mice (76) were taken from several providers (CharlesRivers (ChR), Taconic (Tac), National Cancer Institute (NCI), Helmholtz Centre for Infection Research (HZI), Dys/N6 B6.Cg5Nlrp6tm1Flv from Yale, Harlan (Har)) and kept at the facility of Helmholtz Center for Infection Research in Braunschweig. Mice were treated with DSS (Dextran Sodium Sulphate) for 7 days and 3 days with normal water. If mice developed colitis following DSS treatment, the mice from the same provider and housing facility were considered DSS-colitis susceptible. In contrast, mice that did not develop colitis were considered as DSS-resistant. Samples were collected from different gut locations (ileum, ceacum, colon) and sequenced with illumina HiSeq 2000 with read size 100 in paired-end mode. Preprocessed metagenomic reads were provided by colleagues at HZI (Prof. Dr. Till Strowig).

Table 1: Description of the final datasets used in this study

Metagenome	**Numbers of metagenome**	**Source of data**
Human gut	117	The Integrative Human Microbiome Project (iHMP)
Human tongue	121	The Integrative Human Microbiome Project (iHMP)
Mouse gut	170	(Xiao et al. 2015)
Mouse gut (Health vs disease susceptible)	76	HZI (Prof. Dr. Till Strowig)
Pig gut	284	(Xiao et al. 2016)
Ocean	186	(Carradec et al. 2018)
soil	72	NCBI
Mock community	1	(Singer et al. 2016)

3.2 Reference genome database

MiMiC calculates the best functionally representative species from a reference genome dataset. ***This approach is reference dependent and will always find the best functionally representative species whose genomes were available in the reference dataset**. **Considering this I made reference datasets based on bacterial and archaeal genomes available in NCBI RefSeq genome database and host specific isolated bacterial genomes from published studies.***

3.2.1 NCBI RefSeq genome database

- Bacterial genome

NCBI RefSeq filters out the bad quality genome assemblies and makes a seperate data repository from GenBank after carrying out certain quality controls. All annotated bacterial genome files (translated_cds.faa.gz) were downloaded from NCBI RefSeq genome database updated until February 2019. There were multiple assemblies for the same species in NCBI RefSeq resulting in a huge redundancy for the annotated proteins as well. Some widely studied species and pathogens had comparatively more redundancy. ***I developed a non-redundant***

reference database containing mostly one assembly per species. Multiple assemblies were retained if they were marked as reference or representative genome.

- Archaeal genome

All archaeal annotated (translated_cds.faa.gz) files were downloaded from NCBI RefSeq genome database updated until February 2019 (n=977) and Pfam binary vector was generated for each annotated assembly accession. A non-redundant version of the database was prepared based on the reference genome, representative genome and best checkM (Parks et al. 2015) completeness score.

3.2.2 Host specific reference genome datasets

I made three host specific reference genome datasets collecting genome from three different cultivable collections, human, mouse and pig.

- Human

I collected translated_cds.faa.gz from NCBI of 1500 isolated bacteria from fecal samples of healthy humans (Zou et al. 2019). This collection covered major phyla including Firmicutes, Bacteroidetes, Actinobacteria, Proteobacteria and Fusobacteria of human gut microbiota profile. On the basis of the accession numbers provided in the study, I downloaded the annotations from RefSeq database making up to 1486 assemblies representing 803 species. In an effort to make a non-redundant database I calculated the completeness score (checkM) and selected the assembly with the best completeness score. ***Though these species were already present in the complete bacterial reference genome database, this subset could be used specifically for human gut microbiota related studies.***

- Mice

The MiMiC reference dataset for mice was made based on the genomes from cultivable species proposed as the Mouse Intestinal Bacterial Culture Collection (miBC) (Lagkouvardos, Pukall, et al. 2016). Mice are the most extensively used model organism to understand the physiology and functional characterization for human related disease and health (Park and Im 2020). Still there is a large number of bacterial species that are yet to be uncovered or characterized from mouse gut. The miBC culture collection covered a large number of diversity in terms of phyla, family, genus and species level. The miBC culture collection of 78 species provides a platform to explore the functions and possibly cultivable representative minimal microbial consortia for different mouse models. The genome sequence and isolates are available at www.dsmz.de/miBC. The expanded version of miBC culture collection including 104 cultivable species from the mouse intestine was used for MiMiC prediction of mouse gut minimal microbial consortia.

- Pig

MiMiC reference dataset for pig was prepared based on the culture collection in Pig Intestinal Bacterial Culture Collection (PiBAC) (Wylensek et al. 2020). The PiBAC culture collection covers a huge diversity of pig gut microbiota expanding to 9 phyla, 39 families and 111 species. The culture collection includes 21 novel species. Out of 111 species 65 genomes were sequenced in house and 46 genomes were picked as a close match in NCBI RefSeq in this study. Reference or representative genomes were available for 36 species. There were 10 bacterial species that did not have the representative or reference genomes. For these species the assembly with maximum completeness score using checkM was selected. The genome sequence and isolates are available at www.dsmz.de/pibac.

3.3 Data preprocessing for making Pfam profile of metagenome/genome

3.3.1 Processing of metagenomes

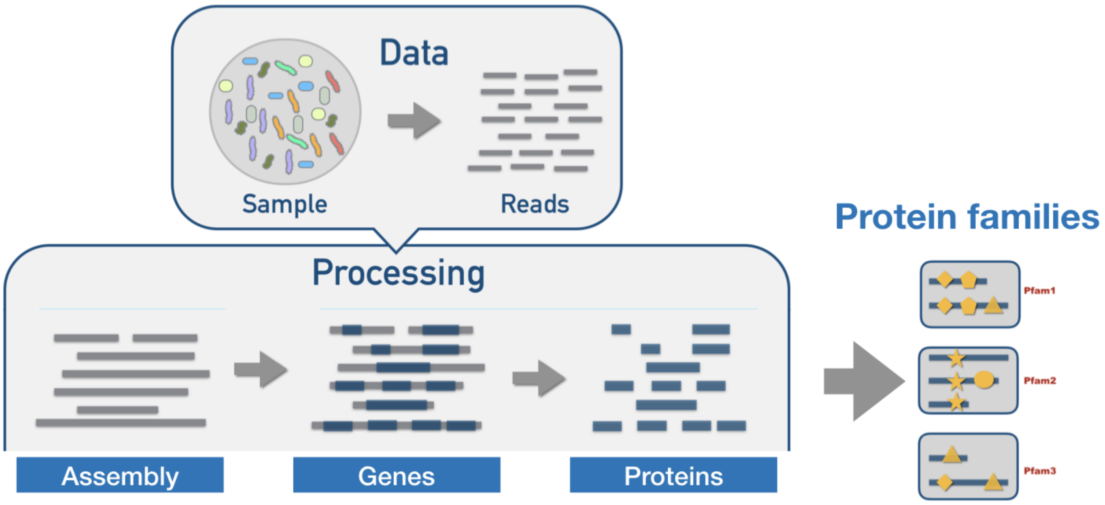

Figure 2: An overview of generating a Pfam profile for an ecosystem. Basic steps of scanning Pfam domains on a sequenced metagenomic library. Starting from sequencing the library, read assembly, gene and protein prediction and scanning of Pfams on the resulting protein sequences. The sequenced reads and assembled genomic sequences are shown as grey lines and predicted genes/proteins as dark blue. The yellow colored shapes depict the Pfam domains on protein sequences.

I developed an inhouse metagenomic pipeline to process the metagenomic raw reads into assembly, protein annotation and Pfam (protein families) scanning. ***Figure 2 depicts a conceptual representation of the functional metagenomic approach used in this study.*** I have collected metagenomes from several sources. If the public datasets used in this study had the contig assemblies available, I proceed from the protein annotation step. I had applied the complete preprocessing pipeline only on pig and mice gut metagenomes.

Following methods and tools were applied on the metagenomic datasets and assembled together as a bioinformatic pipeline.

- The quality control involves removing the reads with low sequencing quality below defined threshold and removing adapter sequences. This step was carried out by fastp [version 0.14.1] (Chen et al. 2018). Adapter sequences were removed with default

option. Reads were filtered based on percentage of unqualified bases. The default phred quality of a base should be more than 15(q) (Ewing et al. 1998). On average 40% of unqualified base pairs were allowed in a sequence. Reads were trimmed at 3'end (-3) with the average mean quality 20 and window size 16(-W). Reads less than 35 (-l) base pairs were filtered out.
- It is important to know the species present in the metagenome that helps to understand the taxonomic level abundance and richness in an ecosystem. Kraken was used to define the read level taxonomic classification (Wood and Salzberg 2014).
- Host specific reads were filtered from the metagenomic data if the contamination is expected really high (Bushnell 2014).
- Processed reads were assembled into contigs with default options (-meta) using metaSPAdes tools version 3.13.0 (Bankevich et al. 2012).
- Contings were filtered based on the minimum length of 500 base pairs using seqkit version 0.10.0 (Shen et al. 2016).
- The assembled contigs were then annotated for proteins using prodigal version 2.6.3 (Hyatt et al. 2010). Gene modeling parameters (-c, -m, -q, -f sco, -d) were used. Partial genes were not considered at the edge of sequence with the option -c. The option -m does not allow to predict the genes from gaps, -q is an output option to suppress logging output, -f sco sets the outfile format to a simple coordinated file and -d is used to assign the input file.
- The proteins were then scanned for the presence of Pfams (database version 17.0) (Finn et al. 2014) using hmmscan version 3.2.1 (http://hmmer.org/) with the threshold cut-ga (Carradec et al. 2018).

3.3.2 Processing of genomes

The bacterial and archaeal genome assemblies from NCBI RefSeq database were downloaded in the form of annotated proteins. NCBI RefSeq genomes were already annotated using the PGAP annotation pipeline (Tatusova et al. 2016). The annotated file with the suffix "translated_cds.faa.gz" contains computationally predicted proteins as well as proteins with valid names in Uniprot. These proteins were then scanned for the presence of Pfams. However, the other host specific datasets were available in different forms. For some species, the protein annotations were available while for the others only contigs or raw reads. In the end, all the genomes were processed from the required stage to produce the Pfam profiles in desirable formats. All genomes and metagenomes were processed using the same tools and parameters as described above (3.3.1) to retain the consistency.

3.4 Making Pfam binary vector

The main MiMiC script needs Pfam binary vector of microbial species as reference genome dataset and the metagenome to predict the corresponding minimal microbial consortium. This is needed as two individual tab delimited files where the rows are the Pfams and the columns are the species genomes/metagenomes. The values are represented as 0 and 1 for the absence and presence of the Pfam in each genome/metagenome. These files can be provided in required format by the user or it can be generated by MiMiC's preprocessing scripts. The hmmscan, a sub-package of HMMER that was used to compare a protein sequence against a protein family profile database. The HMM Pfam profile database had to be built prior to mapping protein sequence using hmmpress. I built a HMM profile database for Pfam version A-17 to be scanned in the genomes/metagenomes in this study. ***The hmmscan output was parsed using a custom perl script (provided within MiMiC) to select the first best hit for each Pfam. The presence and absence of Pfams were represented as 1 and 0 respectively to generate a binary Pfam vector. This binary vector of Pfam was considered as the functional profile of the metagenome or individual microbial species***.

3.5 Reducing redundancy in NCBI reference genome database

I collected the bacterial genomes available in NCBI Refseq genome database and other publicly available databases. I made three host specific (PiBAC, miBC, Human gut) and one NCBI RefSeq reference database. To reduce the redundancy, I decided to take the assemblies marked as reference or representative genome per species. However, all species did not have a representative or reference genome and in that case I used a genome completeness quality score from checkM tool (Parks et al. 2015). An assembly is defined as a reference genome to the species if the assembly has high sequence quality, backed with experimental data and has extensive proteome support from Uniprot (Rolf Apweiler et al. 2004; U. Consortium 2018). A representative genome was decided based on the strain, first complete genome or highest assembly quality and/or clade analysis where the genome with the maximum number of proteins in the cluster was selected. checkM (Parks et al. 2015) tool provides the assessment of quality of the genome based on ubiquitous and single copy genes within a phylogenetic lineage. Moreover, the scores like completeness and contamination fraction provided by checkM helps identify the best assembly out of the pool. The assembly was selected in the priority order of it being a reference genome, representative genome or for the species lacking either of the genome, based on its checkM score. If a species had several reference genomes then all those assemblies were retained. I took all *Escherichia Coli* genomes and clustered them based on the Pfam binary vector using the jaccard index (Jaccard 1901). I noticed that all 4 reference genomes fall into different clusters. Considering that the multiple reference genomes might be useful to maintain the diversity within the species. I calculated the completeness score of the species with redundant assemblies and selected the assembly with the best score.

The resulting non-redundant reference datasets were then subjected to either protein prediction or directly Pfam scanning and binary vector generation. I named the Pfam binary vector database of non-redundant NCBI RefSeq as Pfam Signature Database (hereafter referred as PfamSigDB). PfamSigDB represents one Pfam binary vector per species. The scripts used for downloading, annotating, parsing of the output, vector generation were collectively made available within MiMiC repository.

3.6 Scoring species

Scoring the species from a reference dataset for a given metagenome is the central part of the MiMiC approach.

3.6.1 Calculating MiMiC-score

MiMiC-score is the actual numerical value based on which the species is selected as a member of the minimal microbial consortium. ***I calculate MiMiC-score considering the maximum Pfam matches and least Pfam mismatches of the reference species to the metagenome in comparison in each cycle (iteration) of the selection. However, I explored several other ways of calculating MiMiC-score.*** These scoring methods were compared with the MiMiC method (x/x+y) (Figure 3) using PiBAC as a reference dataset (n=111) (Wylensek et al. 2020) on pig gut metagenomes (Xiao et al. 2016). Under the notion that more prevalent Pfams represent more common functionalities while less prevalent represent unique functions of the species, I used the rareness index of the Pfam as weight to adjust the species score in weighted scoring methods.

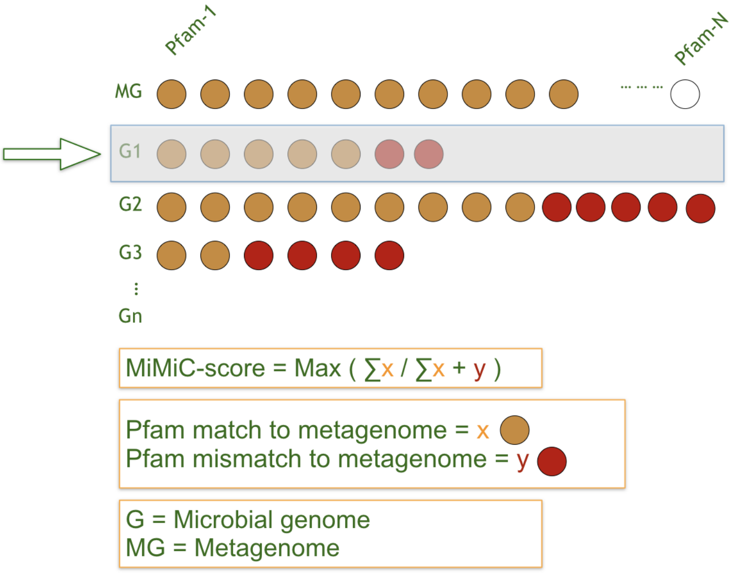
Pfam-1
Pfam-N
MG
G1
G2
G3
⋮
Gn
MiMiC-score = Max (∑x / ∑x + y)
Pfam match to metagenome = x ●
Pfam mismatch to metagenome = y ●
G = Microbial genome
MG = Metagenome

Figure 3: MiMiC scoring system. The circles represent the Pfam presence in the metagenome (MG) and reference genomes (G1, G2, G3 … Gn). The grey transparent box shows which genome is being matched to the metagenome in the current iteration.

3.6.1.1 Unweighted score method

Unweighted MiMiC-score is the score given to each microbial species based on the number of Pfam matches and mismatches of the metagenome to the species genome. Each Pfam was given the same importance in the calculation, no matter if it was a core or rather unique function (Figure 3). This was done iteratively. Iterations were the defined number of species to be selected from the reference genome dataset. In the first cycle, Pfams in the species vector were compared with Pfams in the metagenome. Pfams common in genome and metagenome were called as match (x) and the Pfams present only in the reference species but not present in the metagenome were considered as mismatch (y). Matches (x) were meant to be closer to the metagenome while mismatches showed that function wise how different a bacterium was to the metagenome. To reduce the difference and bring bacteria which had more function closer to the metagenome, I always divided the number of matches (x) by the number of matches plus mismatches (x/x+y). The species having the highest score is picked in the first iteration. Before going to the second iteration, the functions (Pfams) in the metagenome covered by the first species and in the reference genome dataset is replaced by zero. This reduced the chance of redundant functions covered by different species and picked the next species based on more novel function match to metagenome.

3.6.1.2 Weighted score method

The weighted score was assigned to the Pfam based on its uniqueness or rareness in one species compared to all other species within the reference dataset. I calculated the prevalence of protein families amongst the available genomes in a reference genome database. Pfams which were present in most of the genomes could be responsible for housekeeping functions. Such Pfams could be given the lowest score. Pfams, shared rarely or less common among the genomes, could be more specific to certain functions in the species.

The ubiquitous index was calculated for each Pfam based on its presence in the reference genome dataset. This ubiquitous index was used to give weightage to the score of each genome calculated based on the Pfam present in the genome. This added weightage to the selection of the genome that retained more unique functionality. Ultimately, the weighted scoring system might lead to a minimal consortium that covers a wider range of functionality. I calculated the percentage of the absence of Pfams across the reference genome dataset by dividing the number of absences by the total number of genomes in the reference dataset. Then the obtained score was added to the present score (1) of each metagenome. This way the Pfam gets the higher score with more number of absent counts in the reference dataset.

The weighted scoring system is explained with an example below with a dummy table representing the reference genome dataset (Table 2,3). In the following example the species genome is represented with the letter “G” and Pfam with the latter of “P”. Suppose genome G1 and genome G9, have the equal number of Pfam counts but different Pfams correspond to the metagenomes. From genome G1, protein family P1, P2, P4, and genome G9 protein family P1, P3 and P4 matched to the metagenome. If we observe the binary based match score will be the same for both the genomes(Table 2). If we count the number of Pfam matches without assigning any weight to the protein family based on their uniqueness, both the genomes get an equal score 3. This way we may miss the selection of the genome which carries a protein family with more unique functions in the reference dataset.

G1 = P1 (1) + P2 (1) + P4 (1) = 3 -> unweighted MiMiC-score
G9 = P1 (1) + P3 (1) + P4 (1) = 3 -> unweighted MiMiC-score

However, in the case of weighted Pfam score genome G1 gets a higher score as it has a Pfam p2 which has more score because of its absence in more genomes compared to any other Pfam (Table 3). The uniqueness of the Pfam p2 makes it a more weighted protein family in the database.

G1 = P1 (1.1) + P2 (1.9) + P4 (1.7) = 4.7 -> weighted MiMiC-score
G9 = P1 (1.1) + P3 (1.5) + P4 (1.7) = 4.3 -> weighted MiMiC-score

Table 2: Example of unweighted Pfam profile

	G1	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**	Present-Count	Absent-Count	Absent-Score	Absent-Score
P1	**1**	1	1	1	1	1	1	1	**1**	o	9	1	1/10	0.1
P2	**1**	0	0	0	0	0	0	0	**0**	0	1	9	9/10	0.9
P3	**0**	0	0	0	0	1	1	1	**1**	1	5	0	5/10	0.5
P4	**1**	1	0	0	0	0	0	0	**1**	0	3	7	7/10	0.7

Table 3: Example of weighted Pfam profile

	G1	**G2**	**G3**	**G4**	**G5**	**G6**	**G7**	**G8**	**G9**	**G10**	Present-Count	Absent-Count	Absent-Score	Absent-Score
P1	**1.1**	1.1	1.1	1.1	1.1	1.1	1.1	1.1	**1.1**	0	9	1	1/10	0.1
P2	**1.9**	0	0	0	0	0	0	0	0	0	1	9	9/10	0.9
P3	**0**	0	0	0	0	1.5	1.5	1.5	**1.5**	1.5	5	5	5/10	0.5
P4	**1.7**	1.7	0	0	0	0	0	0	**1.7**	0	3	7	7/10	0.7

3.6.2 Scoring equations

I made different combinations of match and mismatch for weighted and unweighted scores to evaluate if the proposed criteria (x/x+y) was better than any other equations or concept of weighted score (Table 4). All equations were compared in terms of functional coverage (number of matches in metagenome) and the number of mismatches in the metagenome (Table 5,6). Following are the equations which were applied on pig gut metagenomes with the reference dataset PiBAC.

Table 4: Description of the variables used in calculating MiMiC-score

Variables for score calculation	**Description**
x	Common number of Pfams between microbial species and corresponding metagenomes.
y	Number of Pfam which are present in microbial species but not in metagenomes.
w	Weighted score given to a protein family in the reference genome database. { PfamPresent + (Pfam absent count across the reference genome dataset / total number of genome in the reference database) }
xw	Weighted match (sum of weighted Pfam matched to metagenome)
yw	Weighted mismatch (sum of weighted Pfam which did not match to metagenome)

3.6.2.1 Unweighted scoring equation

Following is the description of the scoring equation applied by two different concepts with corresponding titles given to compare each version of the equation.

Table 5: Description of the variables used in the unweighted MiMiC scoring method

Scoring equation	**Equation description**	**Equation title**
max(x)	Maximum number of match to metagenome	MiMiC-A
max{x/(x+y)}	***Number of match to metagenome / (Number of match to metagenome+number of mismatch to metagenome)***	***MiMiC-B***
max(x/y)	Number of match to metagenome/ Number of mismatch to metagenome	MiMiC-C

3.6.2.2 Weighted scoring equation

Following are the equations for the weighted score method.

Table 6: Description of the variables used in the weighted MiMiC scoring method

Scoring equation	**Equation description**	**Equation title**
max(xw/yw)	{ weighted sum of matches to metagenome / (unweighted sum of matches to metagenome + weighted sum of mismatches to metagenome) }	MiMiC-Wa
max{ xw / (xw+yw) }	{weighted sum of matches to metagenome / (weighted sum of matches to metagenome + weighted sum of mismatches)}	MiMiC-Wb
max{ xw / (x+y) }	{ weighted sum of matches / (sum of matches to metagenome + unweighted sum of mismatches to metagenome) }	MiMiC-Wc
max{ xw / (x/yw) }	{ weighted sum of matches to metagenome / (unweighted sum of matches + weighted sum of mismatches) }	MiMiC-Wd

Each weighted equation has 2 more versions of increasing the weighted score and converting all the values as 10^w and 100^w to increase the exponential of the values.

3.7 Descriptive statistics generated by MiMiC

A number of calculations were performed while calculating the minimal microbial consortium (Table 7). The whole descriptive statistics reported in the MiMiC output file for each metagenome is detailed in the table below.

Table 7: Explanation of the output generated by MiMiC

Title of the column in output table	**Description**
MetaGenome	The metagenomic ID for which the consortium was predicted
Bacteria	The RefSeq genome assembly accession or the identification number of reference genome datasets for special cases.
Percentage	***Percentage of Pfam matches of metagenome to the selected species in each iteration in a cumulative manner.*** ***Percentage= (Pfam match to metagenome/total number of Pfam in metagenome)*100***

Title of the column in output table	**Description**
NovelMatch	***These are the number of Pfam matches to the meta genome by each species per iteration which were not covered by the previous one.***
ScorePerIteration	The MiMiC-score based on which the species was selected.
PfamInGenomePerItration	Total remaining number of Pfam in species Pfam vector after removing the matched Pfam by previous genome. These Pfam were used to find new novel matches which were not covered in the previous iteration.
exclusiveMismatchPerIteration	***Number of Pfam that did not match the metagenome in the vector of "PfamInGenomePerItration".***
RelativeMatchPerIteration	This is reported to observe the relative number Pfam match compared to the number of mismatch in each iteration. ("PfamInGenomePerItration"/"exclusiveMismatchPerIteration") per iteration.
BactTotalPfam	The number of Pfam present in selected species originally before removing the Pfam in each iteration that match the metageome.
AbsoluteMatch	The number of Pfam that matched the original Pfam profile of microbial species without removing the Pfam in each iteration that matched the metagenome. The common Pfam numbers between BactTotalPfam and Pfam in metagenomes.
AbsoluteMismatch	The number of Pfam mismatches compared to the original Pfam profile of species to the metagenome without removing Pfam at each iteration. The Pfam in BactTotalPfam not mapped to the original profile of metagenome.
CumNovelMatch	***Cumulative values of "NovelMatch" added by each species.***
AbsoluteRelativeMatch	AbsoluteRelativeMatch is the ratio (AbsoluteMatch / AbsoluteMismatch) of AbsoluteMatch to AbsoluteMismatch for each species per iteration.
CumAbsoluteRelativeMatch	These are the cumulative added value of AbsoluteRelativeMatch.
CumExclusiveMismatchPerIteration	These are the cumulative added value of "ExclusiveMismatchPerIteration".
Rank	***Rank of species picked in minimal consortium.***

3.8 Threshold for minimal microbial consortium

It is not trivial to make a decision on the number of species to be selected within minimal consortia. I explored different criteria of suggesting the size of a minimal microbial consortium. It can be set as a hard cutoff by the user or depending on the size of the reference dataset it can be run for exhaustive search of Pfam and then apply a knee point on the function coverage covered by the selected species. The exhaustive search of Pfams in the entire reference database or a selected dataset could be a computationally expensive task. To automate the process of selecting the number of species I compared two knee point methods using R package *inflection (Christopoulos 2016)* to find the elbow point on the cumulative function coverage curve. ***This knee point acts as a threshold of selecting the number of species for each metagenome***. As a bench mark of the two methods of knee point we applied both the knee point on the cutoff of iteration 50 and 100. ***In our study, I applied iteration cutoff 50 with the knee point method uik for the reference NCBI refseq database. With the reference datasets PiBAC and miBC I applied exhaustive search of Pfam in the reference dataset and then allowed the knee point uik based cutoff to select minimal microbial consortia.***

3.9 Taxonomic filter in the reference dataset

There is a lack of solid information and background about the source or host of the bacterial species in NCBI biosample database. Taxonomic based filters could be an alternative to filter the most closely related species to the ecosystem from the MiMiC reference database. Taxonomic profiles of all genomes from NCBI RefSeq are provided to select the species more related to metagenomes. This subset can be used further to generate a minimal microbial consortium for the corresponding metagenomes.

3.10 Comparison of functional profile of different ecosystem

Comparison of functional profiles of different ecosystems were examined based on Pfam binary vectors. Distance matrices were generated using the jaccard index (Jaccard 1901). Multidimensional scaling (MDS) (Mead 1992) plot was performed using cmdscale (Balakrishnan et al. 2014) on the distance matrix generated by jaccard index (Jaccard 1901).

3.11 Prevalence based analysis

Post analysis included the prevalence of selected species across the group and number of Pfams in each category. This provides an overview of how similar a minimal microbial consortium is between two metagenomes. A cumulative graph is generated for each metagenome to visualize the number of Pfams covered in the metagenome by MiMiC picked species for each metagenome. The prevalence was calculated by counting the number of presence of a species in minimal consortia divided by the total number of metagenomes in specific groups and multiplied by 100. This provide prevalence as a percentage of the metagenomes in which the species are present.

4 Results

The result part provides detailed insight into the curated reference genome database, MiMiC method validation, selection criteria of species, Pfam based classification and prediction of minimal microbial consortia for different ecosystems.

4.1 Building a non-redundant NCBI-based genome reference database

4.1.1 Overview of microbial genomes in the RefSeq genome database

As genome assemblies did not have information of taxonomic classification in the NCBI RefSeq database, I collected all bacterial genome assemblies until February 2019 and merged them with taxonomic information available within the NCBI taxonomic database using “taxid” as key identification number. The taxonomic details were created in a separate file obtained from "lineages-2019-02-20.csv" available at (https://gitlab.com/zyxue/ncbitax2lin-lineages) (Mahmoudabadi and Phillips 2018). Five main kingdoms exist in the NCBI taxonomy: Bacteria, Archaea, Eukaryota, Viruses, and Viroids, including 489,796 bacterial, 13,142 archaeal, and 195,721 viral taxonomies to the aforementioned date. An assembly summary updated till 28th May 2019 and containing 156,800 assemblies was downloaded from NCBI RefSeq (ftp://ftp.ncbi.nlm.nih.gov/genomes/RefSeq/bacteria/assembly_sromummary.txt). There were 45,567 strains available in the database. Assemblies were given a reference “species_taxid”. Each strain had a unique “taxid”. Organism names and taxid had the same number of entities. I found that 7 deprecated assemblies in the RefSeq database and those had to be removed. The RefSeq genome assemblies update regularly resulting in addition and deprecation of the assemblies. The assembly_summary.txt file provides the metadata information about the genomes and their assemblies; however, it lacks certain features like taxonomic classification, number of genes, number of proteins, genome size and GC% (guanine and cytosine). These features were then merged from the genomic reports taken from another source (ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS).

In NCBI the taxid is not species specific, it is assigned to a new strain level entry to the taxonomic database. In some cases “species_taxid” and “taxid” are same for the bacterium or archaeal species. In the assembly_summary.txt file of RefSeq, there were 44,853 taxid out of 45,567 that had taxonomy lineage information in the lineage file. Which meant 714 taxid did not have lineage information. In the end, 156,062 assemblies were found to have proper taxonomic information merged using assembly_summary.txt (May 2019) and taxonomy lineage (February 2019) which was used as a reference genome database in this work. The “assembly_accession” is a unique identity given to each genome assembly in the RefSeq genome database. There is an exponential growth after the year of 2005-06 in the number of available genome assemblies of bacterial species in NCBI RefSeq bacterial genome database (Figure 4A). ***I observed that the number of species is less than the number of genome assemblies added to the database, which causes redundancy of multiple genome***

assemblies per species (Figure 4B). I looked at the redundancy level of the species and found that 0.44% species (n=100 reference assemblies) had 61.5% (n=96,526) of total number of assemblies (n=156,062). The Figure 4C represents the species having most redundant genome assemblies.

As taxid is not unique to the species, so the redundancy had to be reduced at the level of species_taxid**.** There were 44,853 taxid and 22,436 species_taxids in the updated version. There were 19,978 species that had the similar species_taxid and taxid. There were 2,478 speicies_taxid that had different taxid. There were 18,631 non-redundant species out of 22,436 species. There were 3,805 species that had 137,431 redundant genome assemblies which made it clear that there is a huge redundancy at the assembly level which needs to be simplified into a least redundant version. There were 3,805 species (species_taxid) that had 26,222 taxids indicating several strains of a species.

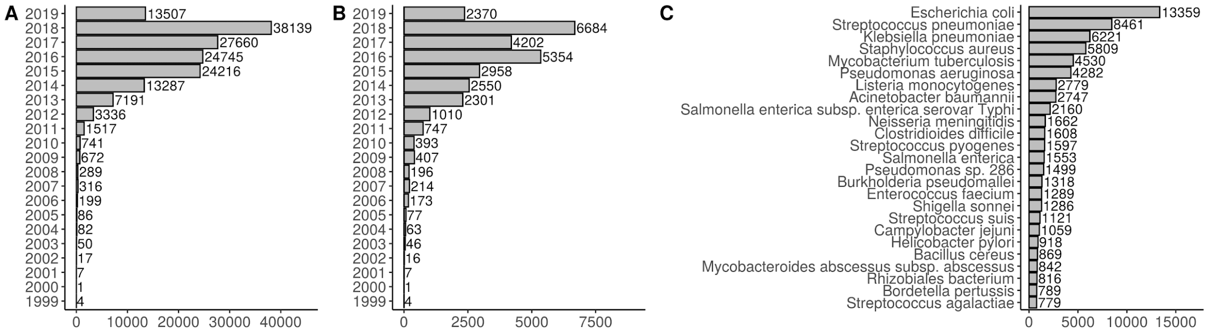

Figure 4: Overview of bacterial genomes deposited in the NCBI reference sequence (RefSeq) database. (A) Number of genome assemblies per year. **(B)** Number of species present per year. **(C)** Top-25 most redundant bacterial species as an indication of redundancy.

In general, genome assemblies are categorized in four different types described as complete, scaffold, contig and chromosome. These assembly resolution levels play a major role in the outcome of the genomic and functional analysis. I showed the number of assemblies resolved at different levels (Figure 5A) and explained the definition of each category below. A complete genome assembly includes all chromosomes as gapless sequences including plasmids and organelles. A complete genome assembly does not contain unplaced and unlocalized scaffolds. A chromosome level can represent one or more chromosomes from respective species. This level of assembly could be a complete sequence of chromosomes without gaps or contigs/scaffolds of chromosomes with gaps. A scaffold is the sequence where contig sequences were connected across the gaps. Scaffolds are unplaced and can be unlocalized. Contigs are just assembled reads into larger reads. They were not assembled further into the scaffold.

Most assemblies were submitted as contigs and scaffolds in the RefSeq. The number of complete genomes have increased significantly in the last two years. Among the redundant and non-redundant species, there were further two categories of assemblies which were classified as representative and reference genome (Figure 5B). I observed that even the non-redundant species were classified as representative genomes. There were only 9 non-redundant reference species. ***The reference genome category is classified from the complete genome only. However, representative genome assemblies are classified from all types of assemblies including contigs, scaffolds and chromosomes***. Considering only the species with complete genome was limiting the number and the diversity of the species available in the genome database to be used to predict the minimal microbial consortia. Based on the observation of representative genomes, ***I decided to include the genome assembly that was the best representative of the species irrespective of the level of assemblies.*** Hence, I considered the reference or the representative genome assembly for the redundant species. Figure 5C represents the reduced redundancy of all bacterial species based on both reference and representative genome assemblies.

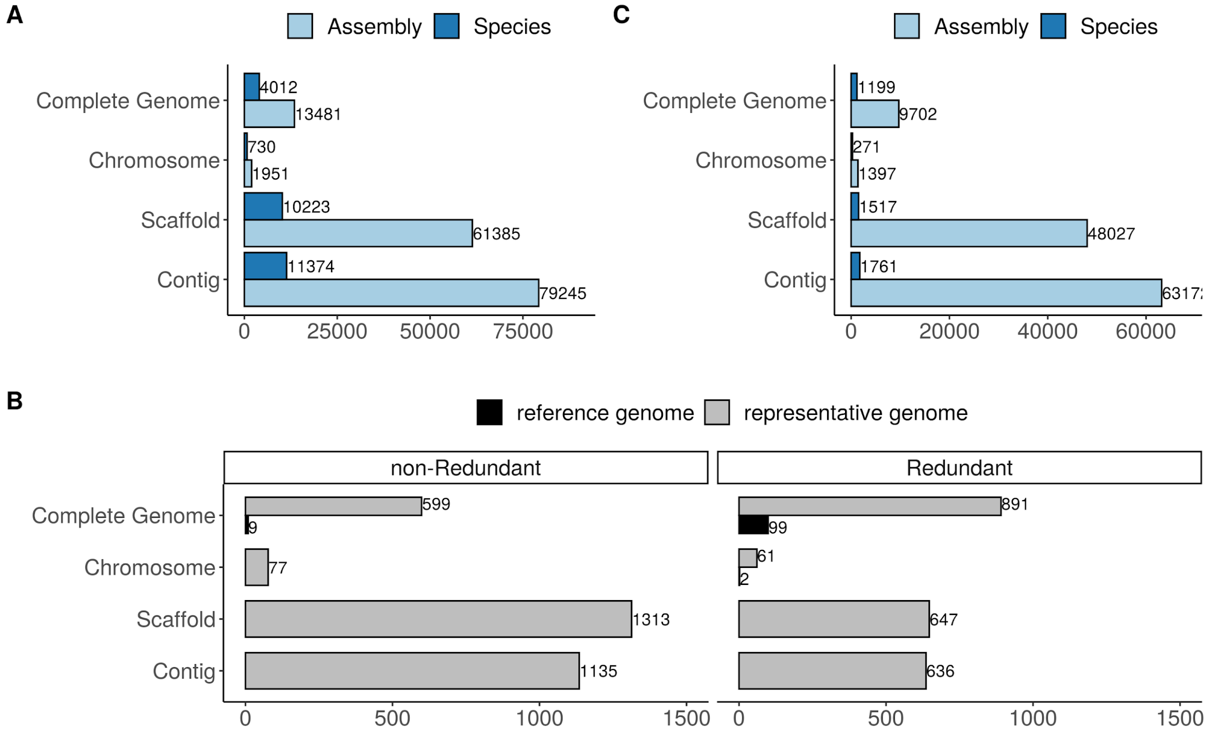

Figure 5: Descriptive statistics of RefSeq bacterial genome database and reducing redundancy. (A) Total number of genome assemblies (light blue) and corresponding species (dark blue) per assembly level. **(B)** The number of reference and representative assemblies categorized into non-redundant and redundant species at each assembly level. **(C)** Number of assemblies (light blue) and unique species (dark blue) reduced to only reference or representative genome assemblies.

4.1.2 Reducing redundancy

4.1.2.1 Based on reference and representative genome

It was important to reduce the redundancy for computational aspects whilst maintaining the diversity of the database by keeping reference or representative genome assemblies. In the redundant assemblies, there were only 100 species (represented by 111 strains), whose reference genome was available in the database. ***These 100 species represented 96,526 genome assemblies in the database*** *(Table 8)*. It was observed that comparatively pathogens had more redundant species (Figure 4C). It might be because disease specific bacteria are studied widely compared to other species. There were 2,306 representative genome assemblies corresponding to 2,270 strains and 2,166 species. ***There were 2,166 species with representative genomes covering 29,389 genome assemblies***. If a species had more than one reference genome, redundancy was allowed at that level. Same criteria was applied for the representative genome assemblies. There were 15 species that had both reference and representative genomes covering 3,617 assemblies. In total 2,252 species had either reference or representative genome

4.1.2.2 Based on the best completeness score

After reducing the redundancy based on reference and representative assemblies, there were 1,554 species which had 15,133 redundant assemblies (Table 8). These species did not have a representative or the reference genome assembly. ***In the second step, checkM scores*** *(Parks et al. 2015)* ***were calculated for this set of assemblies. The assembly with the maximum completeness was taken into account.*** There were 20,883 species which made a non-redundant final set of assemblies, either one assembly per genome or marked as reference or representative genome. Following is the table of the total number of species included in the reference genome for the MiMiC analysis. If a species had both types of genomes as representative or reference genome, I included both the genome assemblies in the final database. Summing up all the information I ended up with 22,418 species with the assembly number of 22,584.

Table 8: Redundant assemblies and selection criteria

Number of redundant species	**Number of non-redundant species**	**Criteria to remove redundancy**
96,526	100	Reference genome
29,389	2,166	Representative genome
15,133	1,554	Completeness (highest CheckM-score)
	18,631	Already non-redundant

4.1.3 Pfam binary vector of the refined NCBI bacterial genome database

A non-redundant Pfam binary vector was prepared for this set of bacterial genomes in the form of presence and absence of Pfam in the genome and named it as BactPfamSigDB (Bacterial Pfam Signature Database). BactPfamSigDB represents the functional profile of the overall isolated bacterial community present in the NCBI RefSeq database. This database saves a lot of time for processing the genomes and serves as a further guideline to understand, curate, and add the new bacterial species in the future development of the approach.

4.1.4 Pfam binary vectors of archaeal species

Annotated files (translated_cd.faa.gz) of archaeal species were downloaded from NCBI RefSeq genome database (ftp://ftp.ncbi.nlm.nih.gov/genomes/RefSeq/archaea/) updated until February 2019 (n=977) and the Pfam binary vector were generated for each annotated assembly. All archaeal assembly accessions were curated and mapped to "assembly_summary.txt" file from NCBI RefSeq. This way I merged the assembly summary information to the assembly accession. The taxonomic details were created in a separate file obtained from "lineages-2019-02-20.csv" available at (https://gitlab.com/zyxue/ncbitax2lin-lineages) (Mahmoudabadi and Phillips 2018). There were multiple assemblies of archaeal species in NCBI RefSeq, the same as in the case of Bacteria. The assembly with maximum check-M completeness score or labeled as "representative/reference", were included in the final database. Applying above criteria, a non-redundant Pfam binary vector database with the taxonomic information in a separate file linked with assembly_accession was created for the archaeal genome dataset. There are 581 species present in the archaeal Pfam binay vector dataset named and referred as "ArchPfamSigDB" (Archaeal Pfam Signature database) in this study.

4.1.5 Bacterial and Archaeal Pfam binary vector datasets were merged together

I merged the bacterial and archaeal Pfam binary vector database comprising 22,999 species. While looking at the Pfam distribution and the genome size of these species, I saw that species having fewer Pfams also had relatively smaller genome sizes. Some of these species appeared to be outliers falling under the lower quartile in Pfam based distribution (Figure 6). These species (n=295) were removed from the Pfam vector database leaving 22,627 species in the final version of the MiMiC reference database.

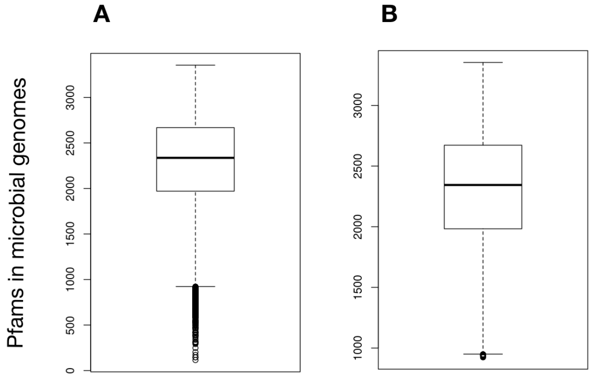

Figure 6: Pfam distribution across single microbial species in the entire genome database (PfamSigDB). Number of Pfams (y-axis) in each bacterial/archaeal genome. PfamSigDB **(A)** before and **(B)** after removing the outliers (species falling under the lower quartile).

4.1.6 Taxonomic distribution in final Pfam vector database

In the combined database of bacteria and archaea (n=22,806 assemblies), there were 54 Phyla, 95 classes, 213 orders, 481 families, 2,617 genera, and 22,627 species taxid (Figure 7A). Bacteria covered 48 Phyla, 82 classes, 189 orders, 439 families, 2,484 genera, 22,138 species taxid. In the case of archaea, there were only 5 Phyla, 12 classes, 23 orders, 41 families, 132 genera, and 488 species taxid in the database (Figure 7B).

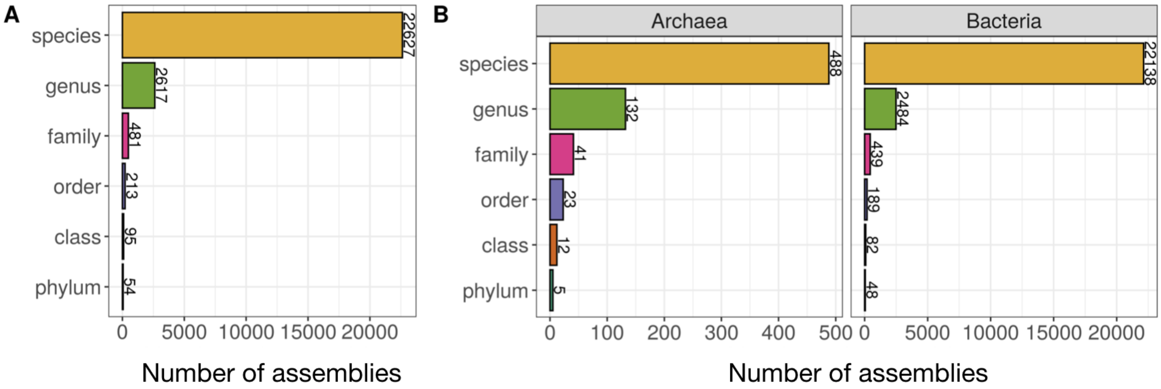

Figure 7: Taxonomic distribution of the Pfam reference database. (A) Total number of phyla, classes, orders, families, genera and species-level assemblies in the database. **(B)** The same data split into Archaea and Bacteria.

4.1.7 Number of Pfam vectors in MiMiC reference genome datasets

Following table represents the number of genome vectors in each category of the Pfam reference genome database. I categorized the datasets into six categories, the Pig Intestinal Bacterial collection (PiBAC), Mouse Intestinal Bacterial Collection (miBC), Human Intestinal Bacterial Collection (Human gut), NCBI RefSeq Bacterial Collection, NCBI RefSeq Archaeal Collection, and the combined NCBI Archaeal and Bacterial Collection (Table 9).

Table 9: Number of genome assemblies in PfamSigDB and the host-specific database

Reference dataset	**Number of species**
PiBAC (Wylensek et al. 2020)	111
miBC (Lagkouvardos, Pukall, et al. 2016)	104
Human gut (Zou et al. 2019)	803
NCBI RefSeq (Bacteria)	22,138
NCBI RefSeq (Archaea)	488
NCBI RefSeq (Bacteria +Archaea)* after removing outliers	22,627

4.2 Impact of genomic features on MiMiC outcome

I observed that the Pfam distribution varied substantially between species. Genomes with fewer Pfams tend to be of smaller size. I calculated correlations between Pfam distribution and the genome size as well as other factors like number of genes, proteins and even GC percentage. The distribution of Pfam was symmetric whereas the genome size, number of genes and proteins showed a long tail on the right and the GC percentage had bimodal presentation. The genome size showed a significant correlation of 0.98 (pearson correlation, p-value=0.001) with number of proteins and Pfams (Figure 8). GC percentage did not show high correlation to Pfams or proteins. ***Hence, the genome size of a microorganism can affect the MiMiC outcome and should be taken into account to avoid bias.***

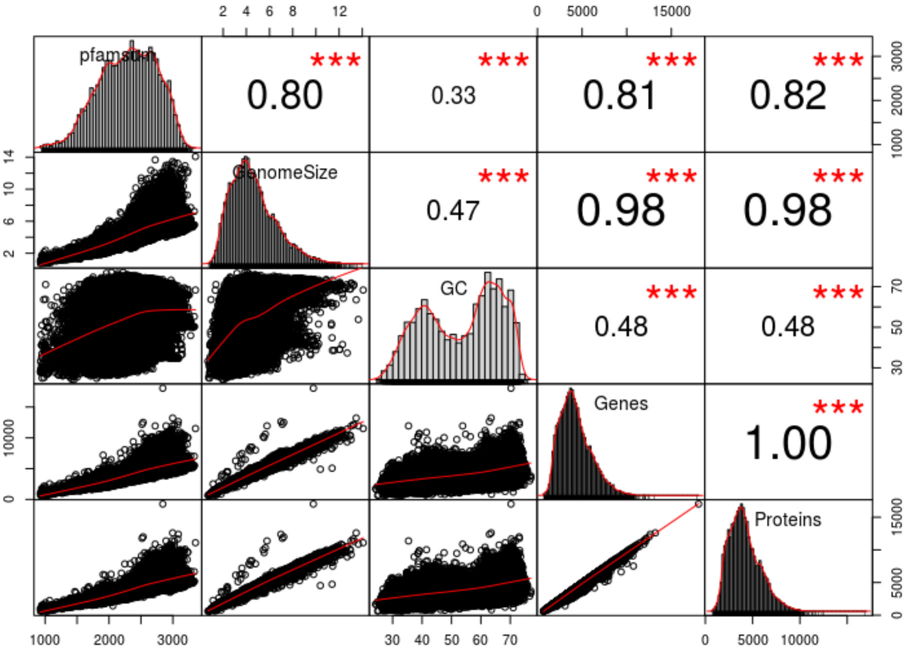

Figure 8: Pairwise correlation analysis for genome size, number of genes, proteins and Pfam. Correlation values are shown as numbers above the diagonal where the size of the font corresponds to the value of correlation coefficient. The significance p-value of correlation test is shown as red stars (*** p < 0.0001). The distribution of each factor is shown on the diagonal as histograms. The scatter plots below the diagonal suggest a linear correlation between the number of genes and proteins to the genome size.

4.3 Functional profile of different ecosystems can be distinguished based on Pfam profile

The essential foundation of MiMiC is that the Pfam binary vectors can be used to distinguish the functional profile of two different ecosystems. To establish that, I used Pfam as a binary unit to measure the functional profile of metagenomes and microbial species. Mostly, the number of

Pfams across the metagenomes within an ecosystem showed minor variation except for the environmental ecosystems soil and ocean (Figure 9A,B). The number of Pfams varied from ca. 2000 to 7000 for soil metagenomes and ca. 3000 to 7500 for ocean while for the other ecosystems this variation remained under ca. 2000 Pfams (Figure 9B).

I compared Pfam profiles of microbiota from different environments and body sites of different organisms. Pfam binary vectors of metagenomes from different ecosystems showed significantly different clusters in multidimensional scaling plot (MDS) (Figure 9C,D). The distance was calculated using the Jaccard index. I observed that environmental metagenomes clustered separate from organismal body sites (mouse, pig, human). Additionally, the gut microbiota of different organisms could also be significantly distinguished from one another using Pfam profiles (Figure 9C,D).

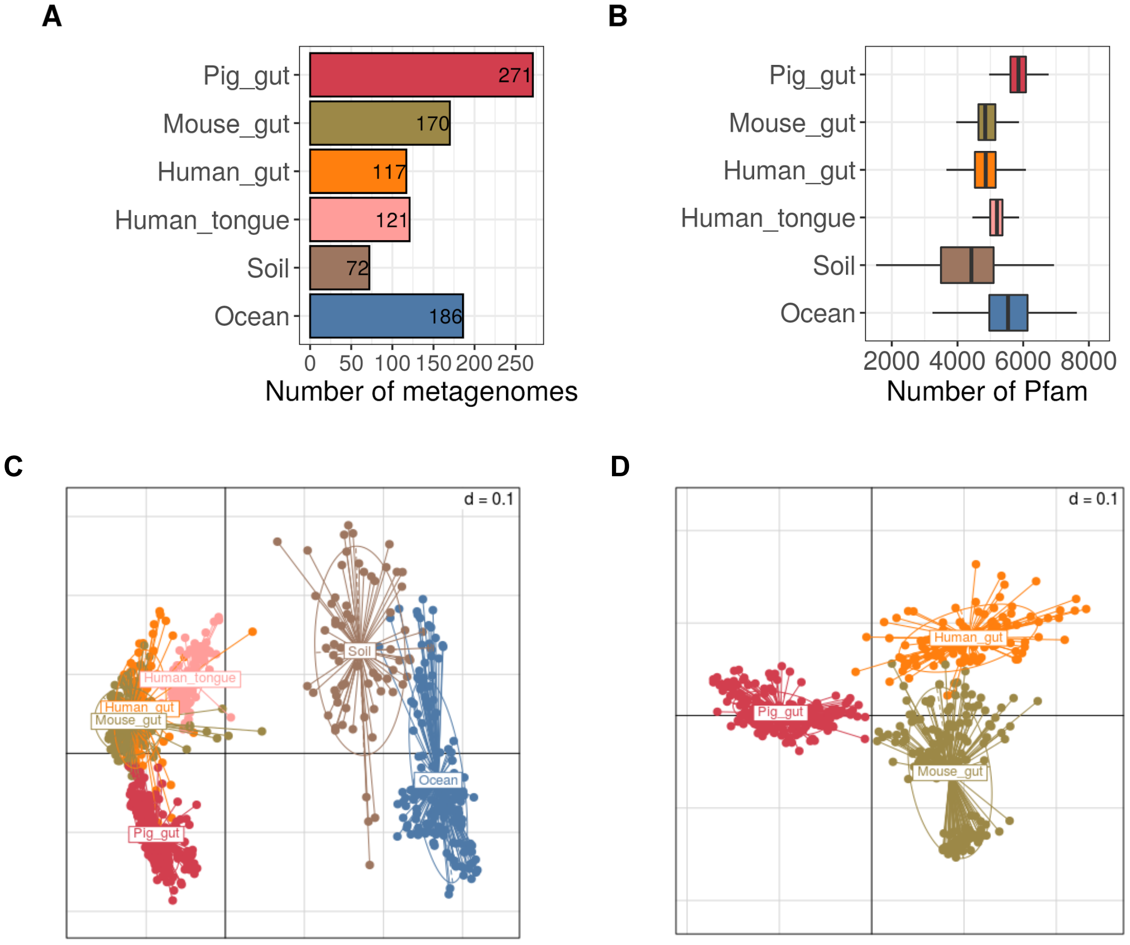

Figure 9: Pfam-based classification of different ecosystems. (A) Number of metagenomes in each category. **(B)** Number of Pfam per metagenome in each category. The Multidimensional scaling (MDS) plots in **(C)** all categories and **(D)** only gut samples based on the Pfam profiles of metagenomes using the jaccard index distance method (p-value 0.001). Each dot corresponds to a metagenome. The value of “d” in the MDS plots is the mesh of the grid in each plot.

4.4 Devising a method for scoring species in MiMiC

I established that Pfam is a valid unit to distinguish ecosystem specific communities. The next question is to score the species to generate minimal consortium to represent the functionality of the ecosystem. ***In this process, MATCHES represent Pfams present in both microbial genomes and the input metagenome whilst MISMATCHES are the Pfams included in any given microbial genome but not in the metagenome to be matched.*** To this end, I used score=matches/(matches+mismatches) on pig gut metagenomes (n = 284) from pig gut gene catalog (Xiao et al. 2016) with PiBAC (Pig Intestinal Culture Collection, n=111 species) (https://www.dsmz.de/pibac) as reference genome dataset. I also explored several other scoring methods to verify my approach (table 10) explained in method section 3.6.2. I created a weighted Pfam dataset where the binary Pfam matrix was scored according to the rareness of the Pfams across the reference dataset (method 3.6.2.2).

Table 10: Different versions of MiMiC

Number	Method Name	Equation
1	MiMiC A	x
2	**MiMiC B**	**x/(x+y)**
3	MiMiC C	x/y
4	MiMiC Wa	xw/yw
5	MiMiC Wb	xw/(xw+yw)
6	MiMiC Wc	xw/(x+y)
7	MiMiC Wd	xw/(x+yw)
8	MiMiC Wa_10	10^(xw/yw)
9	MiMiC Wb_10	10^(xw/(xw+yw))
10	MiMiC Wc_10	10^(xw/(x+y))
11	MiMiC Wd_10	10^(xw/(x+yw))
12	MiMiC Wa_100	100^(xw/yw)
13	MiMiC Wb_100	100^(xw/(xw+yw))
14	MiMiC Wc_100	100^(xw/(x+y))
15	MiMiC Wd_100	100^(xw/(x+yw))

I compared the median of Pfam matches (function coverage) in minimal consortia and median of mismatches in minimal consortia of 284 metagenomes for each MiMiC version. MiMiC version-A covered the maximum function (Figure 10A). We see three patterns in figure 10A, upper part is MiMiC_A, middle overlapping part contained MiMiC versions B, Wa, Wa_10, Wa_100, Wb, Wb_10, Wb_100, Wd, Wd_10 and Wd_100, the third and the lowest contained

Wc and Wc_100. However, the number of mismatches were also highest in MiMiC-A (Figure 10B). The middle part in the Figure B contained method Wc and Wc_100 whilst the rest of the methods appeared to be very close to each other or overlapped. I decided to pick a method which covers maximum function with the minimum mismatch level. Accounting for these criteria, MiMiC-B, MiMiC-C, MiMiC-Wa, and MiMiC-Wb performed equally well (Figure 10C,D). The weighted method MiMiC-Wa and MiMiC-Wb did not make any significant difference compared to the unweighted method MiMiC-B and MiMiC-C. To reduce the unnecessary computational time, I decided to use unweighted methods. ***With the assumption that adding match to mismatch in denominator can help to reduce the genome size effect, I decided to use the method MiMiC-B (matches/matches+mismatches) for further development in this study***.

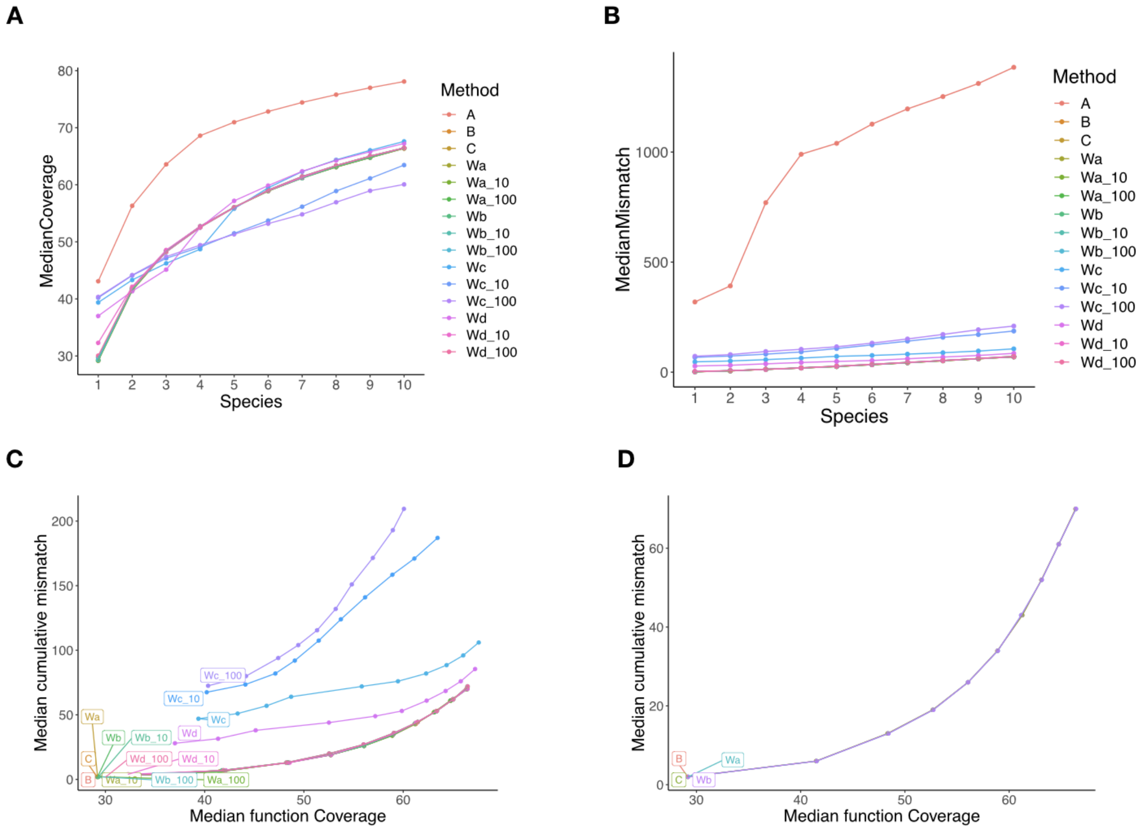

Figure 10: Comparison of different MiMiC methods. (A) Percentage of median Pfam coverage of Pig gut metagenomes by first 10 species selected from the PiBAC bacterial genome collection. MiMiC method A is on the top. The middle part MiMiC versions B, Wa, Wa_10, Wa_100, Wb, Wb_10, Wb_100, Wd, Wd_10 and Wd_100 are either very close to each other or overlapped while lowest part Wc and Wc_100 versions showed different trends **(B)** The number of mismatches were highest in MiMiC-A. The middle part in the Figure B contained method Wc and Wc_100 whilst rest of the methods appeared to be very close to each other or overlapped **(C)** Comparison of median cumulative Pfam coverage (x-axis) and median mismatches (y-axis) of all metagenomes by all MiMiC methods. **(D)** The least plotted line of Figure C plotted separately in Figure D showing MiMiC methods B, C, Wa and Wb overlapped.

4.5 Determining the number of species to be proposed in minimal consortia

Based on the comparative analysis with other versions of MiMiC, I decided to use the method MiMiC-B. However, there was still an open question to determine the number of species to be selected to propose a minimal microbial consortium. One way to do this was to define a hard cutoff (e.g. 10 or 12 species) or decide based on a cumulative function coverage curve individually for every metagenome. However, hard cutoff could bias the comparative studies where one condition potentially might have more diverse microbiota than the other. On the other hand, deciding for every sample individually could also be very exhaustive. To address this question, I carefully benchmarked various approaches and proposed an automated species selection criteria in MiMiC. To this end, I first used the MiMiC-B method on all metagenomes in this study using PfamSigDB (Pfam Signature Database, species = 22,627) as a reference database and analyzed the number of species covering the functions in each category at the iteration cutoffs 50 and 100. ***I observed that the cumulative functional coverage of metagenomes by MiMiC-picked species turned into plateau and did not add any distinguishable number of Pfams after approximately 50 iterations in most of the cases*** (Figure 11).

The number of species found to be variable within and across the category ecosystems with the same iteration cutoffs. For instance, the lowest number of species were picked in soil metagenomes and highest number of species were picked in ocean and pig metagenomes. This suggested that the hard cutoffs would not work even for the metagenomes within an ecosystem.

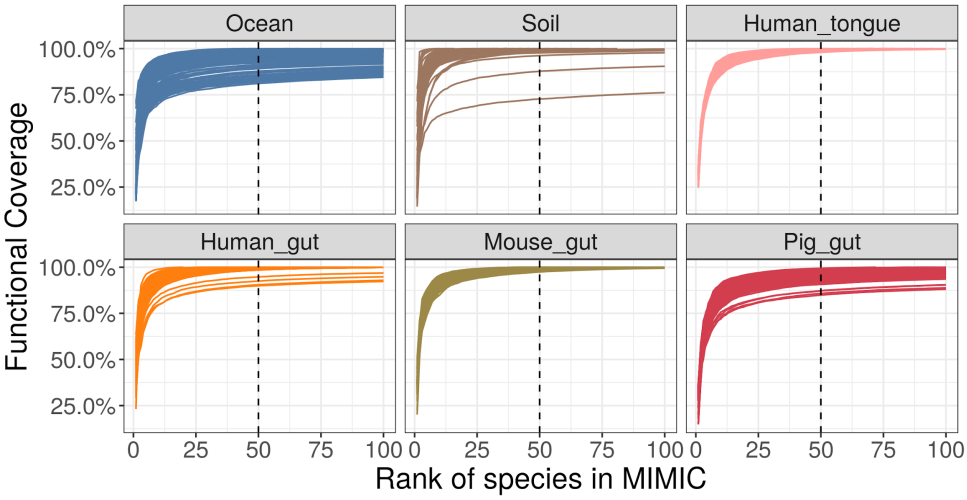

Figure 11: Function coverage of metagenomes by MiMiC derived species. The cumulative sum of the percentage of Pfams covered over the number of species. Each line is an individual metagenome. The y-axis is the percentage of Pfam covered in metagenome and x-axis is the rank of species given by MiMiC.

Next, I implemented two different knee point methods to determine the elbow point on the cumulative function coverage curve. ***With the notion that the species covering most numbers of Pfam cover more diverse functions of the metagenome, the elbow point on the curve would suggest a non-significant addition of functions.*** I looked into the function coverage at different iteration cutoffs and knee point methods to determine the number of species to be included in the MiMiC proposed minimal microbial consortia. I applied two knee point methods uik and d2uik implemented in R package inflection (Christopoulos 2016) to decide the number of species in minimal consortia. At the same time, to benchmark the iteration number I set two cutoffs as 50 and 100. I observed that the method d2uik was very stringent compared to method uik. Method d2uik calculated the knee point at a very low number of species in both the iteration cutoff 50 and 100 (Figure 12). The average knee point range was 4 to 6 species at iteration 100 and 4 to 5 at iteration 50 in all the categories after applying method d2uik (Figure 12). In case of method uik the average keepoint remained 7 to 12 species at iteration 100 and 6 to 10 at iteration 50 in most categories (Figure 12). Furthermore, I observed that both the methods did not make any significant difference in terms of adding more numbers of species at 100 iterations compared to 50.

At the iteration 50 and knee point uik it was observed that the maximum number of species (n=13) were picked in pig gut metagenomes (Xiao et al. 2016) and the least number of species were picked in soil samples (n=2). ***Metagenomes within the category and among the category had different numbers of species picked. This suggested the heterogeneity of the individuals in human, pig, mice and in environment metagenomes as well.***

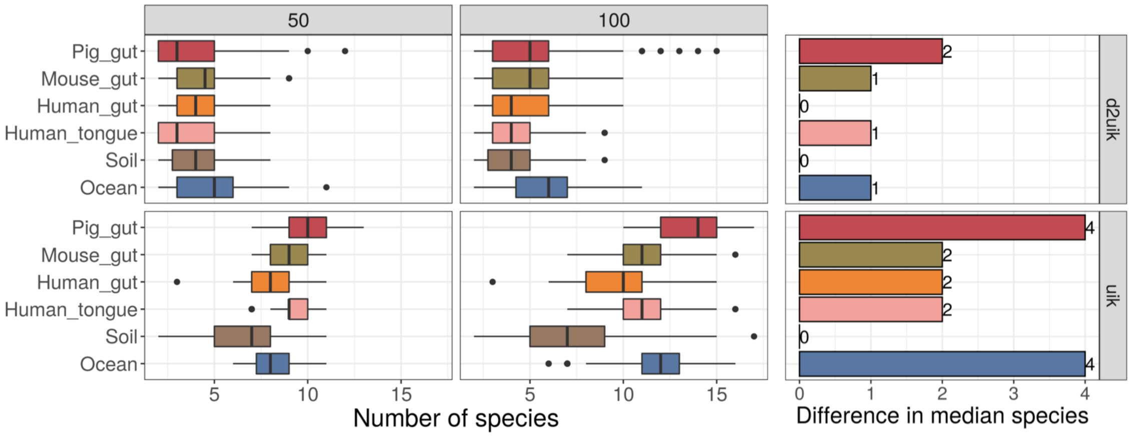

Figure 12: Number of species derived by MiMiC using different knee point methods. Numbers of species were calculated per metagenome based on different knee point methods (uik - bottom panel, d2uik - top panel) and iterations (50 - left panel and 100 - right panel). Box plots represent the number of species and bar plots represent the median number difference between iteration cutoff 50 and 100.

In addition to deciding the number of species it was important to make sure to give the best possible function coverage. I compared the function coverage by the number of proposed species at different iterations and applied knee point methods. The increasing number of iterations did not make a considerable difference in both the cases of knee point methods. The median function coverage difference between iteration 50 to 100 remains 0.48% to 5% at both the knee point methods. In most cases (mouse, human, soil), it remained under 2.8% (Figure 13). Addition of novel Pfam by new species after 10 species declined extremely in most cases (Figure 11). ***MiMiC selected species at knee point uik and iteration number 50 were able to cover 85% to 95% function coverage in metagenomes including all categories (Figure 13). Which remained lower (75% to 85%) in method d2uik at both the range of iteration 50 and 100.***

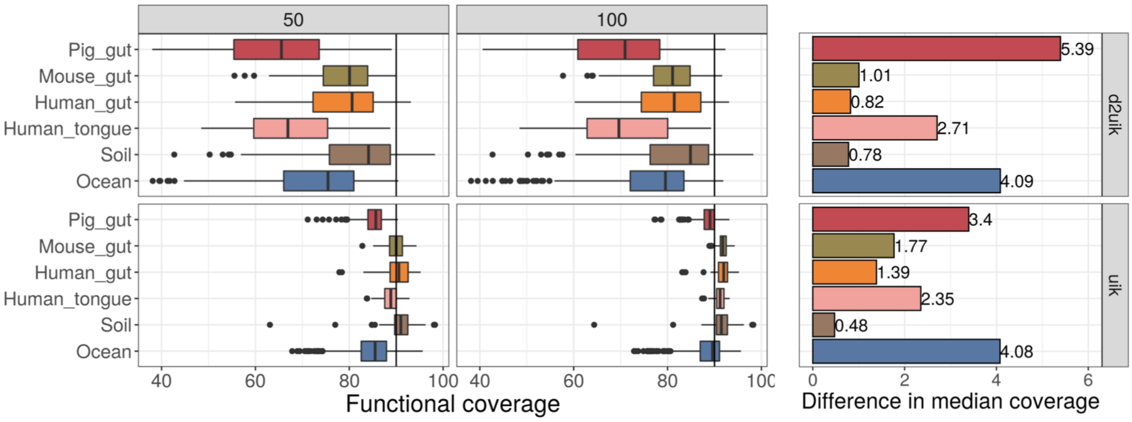

Figure 13: Comparison of function coverage of metagenome using different knee point methods. Percentage of metagenomic function covered by MiMiC species in each category based on different knee point methods (uik - bottom panel, d2uik - top panel) and iterations (50 - left panel and 100 - right panel). Box plots represent the function coverage and bar plot represents the median function coverage difference between iteration cutoff 50 and 100. The vertical black line in the boxplots is drawn at 90 to easily see the relative difference amongst different methods and iteration cutoffs.

Based on these results and the principle of considerably minimum number of species and maximum number of function coverage, I decided to have a soft cutoff of the knee point method uik instead of d2uik. ***Keeping a constant number (50) of iterations across all the metagenomes while applying the knee point methods makes the approach to select a sound automated number of the minimum number of species for each metagenome***. It is recommended to use the method uik for knee point determination. However, both the options were made available to the user to explore the data with multiple choices.

4.7 Evaluation and validation of MiMiC method-B on mock community

Pfam profile of metagenome from MBARC (Singer et al. 2016) (Mock Bacteria Archaea Community) mock community was used to evaluate MiMiC version-B. MBARC contains 23 bacterial species and 3 archaeal species. MiMiC version-B was run with the exhaustive iterations option until all the functions of the metagenome were covered by the reference species. PfamSigDB (n=22,627) consisting of both bacteria and archaea was used as a reference to generate the MiMiC proposed minimal consortium (Figure 14)**.**

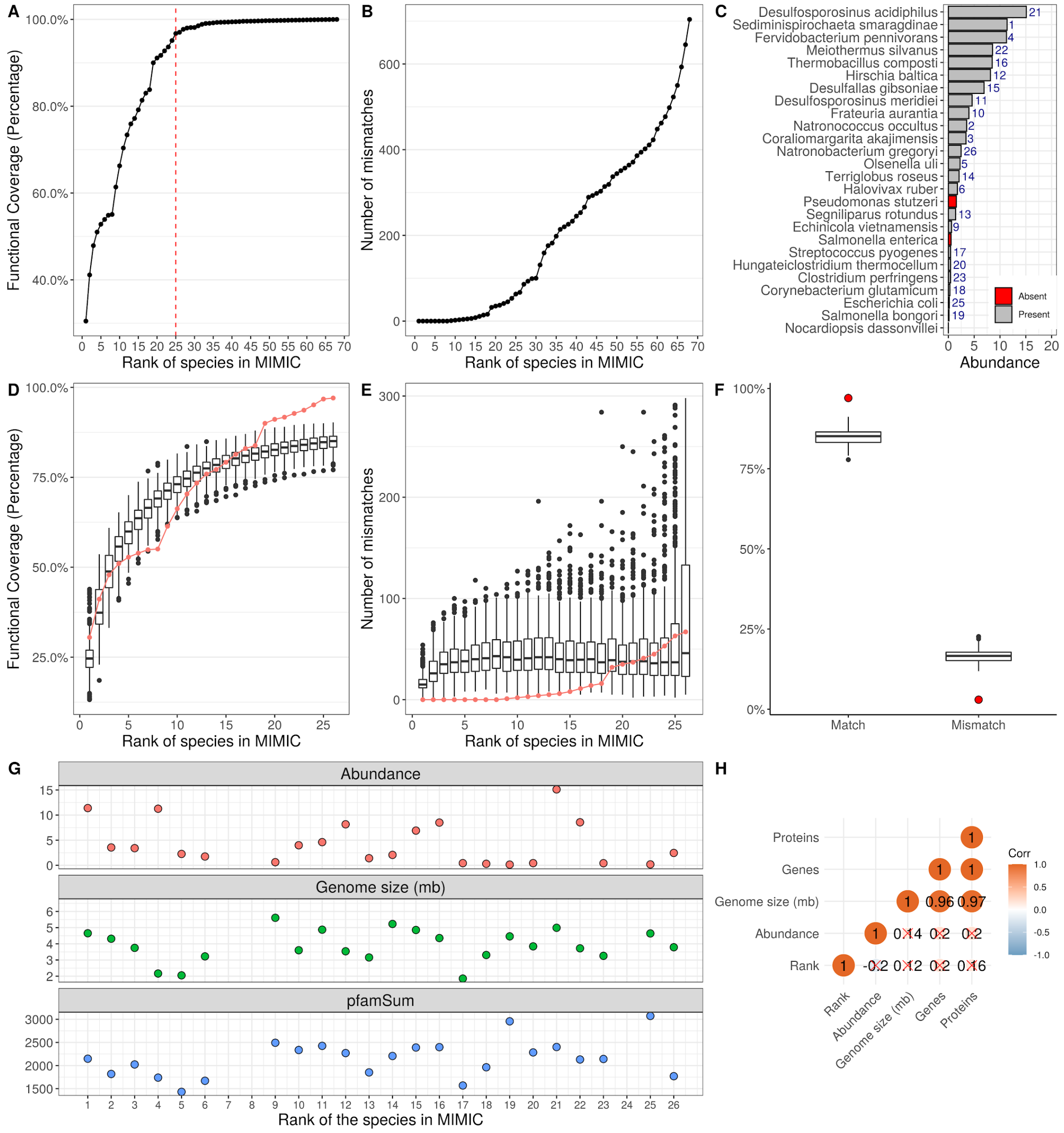

Figure 14: Evaluation of MiMiC method-B using MBARC mock community. (A) The cumulative function coverage of species in the MBARC mock community. The red dotted line corresponds to the knee point deduced by MiMiC. **(B)** Similar to (A) where the y-axis shows the number of Pfam mismatches between a species genome and the metagenome. **(C)** The taxonomic relative abundance of each mock community member in the metagenome. Grey color represents the presence and red shows the absence of the species in the MiMiC derived community. **(D)** function coverage and **(E)** mismatches in 500 randomly selected sets of 26 species where the red line corresponds to the scores of MiMiC derived species. **(F)** overall percentage of Pfam coverage (Match) and mismatches of metagenomes by MiMiC selected species were compared with the random sets. Red dots represent the Pfam coverage and mismatch by MiMiC picked species. **(G)** The genomic characteristics of the MiMiC derived species arranged by rank on the x-axis and the corresponding feature on the y-axis. **(E)** Pearson's Correlation calculation among all the genomic features and species rank. The color gradient from dark to light represents higher to lower correlation coefficient and the size of the dot corresponds to p-value where the numbers with "X" marked are non-significant correlation tests.

The MBARC metagenomic Pfams were covered 100% by 68 species (Figure 14A). The number of mismatches are reported in Figure 14B. The knee point cutoff gave 25 species and 23 of these 25 species were found to be the source species from MBARC mock community. These 23 species included 20 bacterial and all 3 archaeal species of the community (Figure 14C). Three species, *Nocardiopsis dassonvillei DSM 43111*, *Pseudomonas stutzeri RCH2*, and *Salmonella enterica subsp. arizonae serovar RSK2980* were not present in the MiMiC outcome. *Nocardiopsis dassonvillei DSM 43111* had zero percent (0.00%) relative abundance in taxonomic profile of MBARC metagenome so it was expected to be absent from the minimal consortium. This explained that overall MiMiC only missed 2 bacterial species of MBARC in the consortium generated by MiMiC. I took the first 26 microbial species from MiMiC derived consortium and compared the union of their Pfam profile with the MBARC metagenome. I found that these 26 species covered 97.03% of the metagenomic Pfam profile. Only 2.97% Pfams were not covered by MiMiC generated consortium. ***This explained that even considering the large number of species and their diversity, MiMiC was still able to pick a taxonomically relevant and functionally sound profile to the corresponding metagenome.***

To evaluate further, I compared the match and mismatch of these 26 species with the 500 randomly selected sets of 26 species from the reference genome database (Figure 14D,14E). ***In total, MiMiC derived species covered 97.03% function coverage in metagenome while random sets covered 87.5% on average*** *(Figure 14F).* First 2 and last 11 species contributed more matches compared to the median matches of random sets at the same position***.*** I observed that there were no mismatches until the first 8 species and thereafter only 1 mismatch per species upto the 14th species. Number of mismatches gradually increased as the number of matches increased compared to random sets in MiMiC picked species (Figure 14D,E). I generated the union Pfam profile of MiMiC generated consortium and compared it with the metagenome Pfam profile of MBARC. The median Pfam coverage of random 500 sets was 87.5% and the median mismatch was 16.63%. The highest Pfam coverage by 500 random sets was 91.9% and the lowest Pfam coverage was 77%. The highest mismatch was 22% and the lowest mismatch was 11.84%. The Pfam coverage by MiMiC picked species was higher than the median and the highest covering random set. Similarly, the mismatch percentage was lower in MiMiC picked species (2.97%) compared to the median (16.63%) and the lowest mismatch percentage (11.84%) of the random set (Figure 14F).

I carried out a correlation analysis with the factors such as genome-size, relative abundance of the species in the community, and number of Pfam (Figure 14G,H). **Unsurprisingly, *genome-size is significantly correlated with the number of genes and proteins with the Pearson's correlation coefficient 0.96 and 0.97 respectively. However, the relative abundance, genome size, number of genes, and proteins were not correlated with the rank of species in MiMiC outcome (Figure H). This suggested that the MiMiC derived minimal microbial consortium was not affected by such factors*. *With this evaluation by the mock community, it was established that MiMiC provides the selection of species with best function coverage and least mismatch compared to random sets. Moreover, the MiMiC picked species were taxonomically relevant (92%) to the metagenomic profile*.**

4.8 MiMiC performed better than random sets across different ecosystems

To assess the MiMiC performance on native metagenomes, I compared function coverage per category by MiMiC picked species with median function coverage of the same number of randomly picked species (Figure 15).

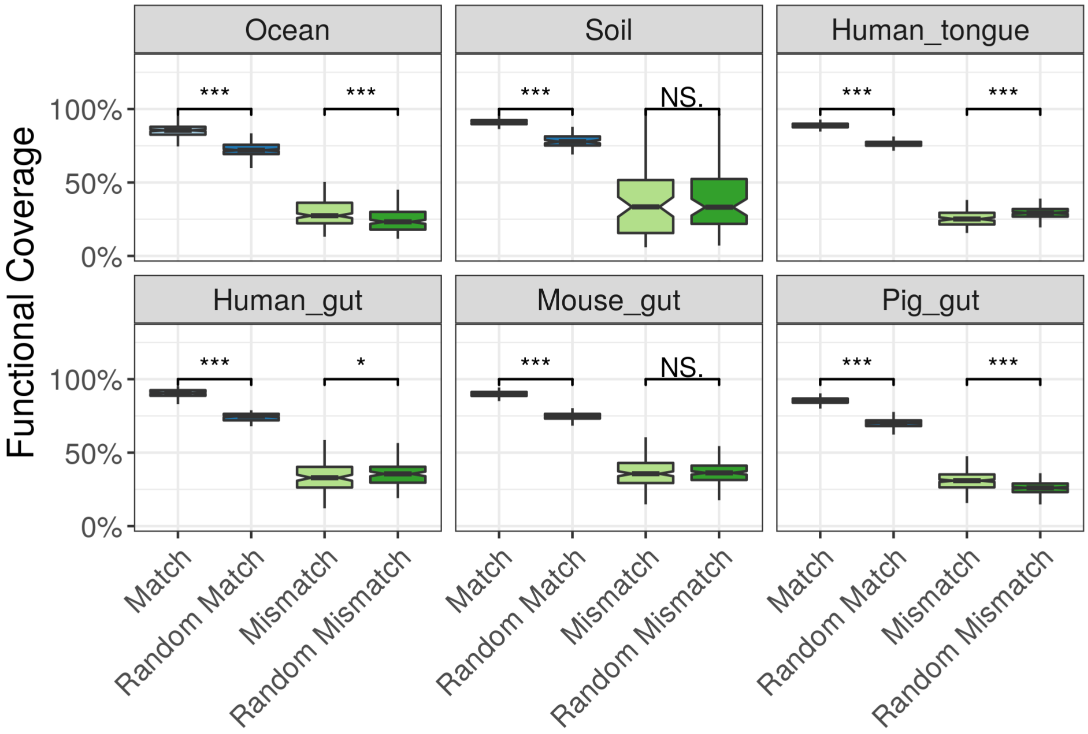

Figure 15: Comparison of Pfam coverage and mismatches to the random sets. MiMiC matches and mismatches to the metagenomes are shown with the light green color and median match and mismatch of random sets are shown in dark green color. The p-value significance of wilcoxon rank sum test is represented by the number of stars where NS is the non-significant difference between the comparisons.

I generated 100 random sets without replacement for each metagenome. For each set, the number of Pfam matches and number of Pfam mismatches to the respective metagenome were calculated and later converted into the percentage of function coverage ((Pfam match to the metagenome/total number of Pfam in the metagenome)*100)) and percentage of mismatch ((Pfam that did not match to metagenome/total number of Pfam in the metagenome)*100). Medians of function coverage percentage and median of mismatch percentage of the random sets to the metagenome were compared with those of MiMiC picked species. Such comparison provided a ground to evaluate if MiMiC derived minimal consortia had the actual potential to be functionally close to the native ecosystem.

A Wilcoxon rank sum test was applied to test the difference between the MiMiC picked species and random sets using "Wilcoxon.test" function in R. Pfam matches were significantly higher than the random sets at the p-value of 0.001 in all metagenomes. The difference in the mismatches between MiMiC picked species and random sets was not significant except pig gut and ocean samples. ***These results confirmed that the microbial community proposed by MiMiC was significantly closer to the native ecosystem than by chance***.

4.9 Pfam profile of MiMiC derived minimal microbial consortia could distinguish the ecosystems

The original Pfam profile of metagenomes from different environments showed distinct clusters (Figure 9C,D). This explained the heterogeneity of functional profiles within and across the ecosystems. Considering this I expected that a combined Pfam profile of MiMiC derived minimal microbial consortium from each metagenome should show a similar pattern. I generated a Pfam union of MiMiC-selected species for each metagenome and this was considered as a mimicked function profile of that metagenome. Using this profile I generated a multidimensional scaling plot (MDS plot) using the jaccard index as a distance calculating method of different groups. ***Ocean and soil made a separate cluster from the consortia from human, mice and pig*** (Figure 16A). However, there was an overlap of the gut metagenomes groups. Even then the pig gut metagenomes were more likely separated from the mouse and human gut metagenomes. This observation reassured that the MiMiC proposed minimal consortia captures the functionality of the ecosystem.

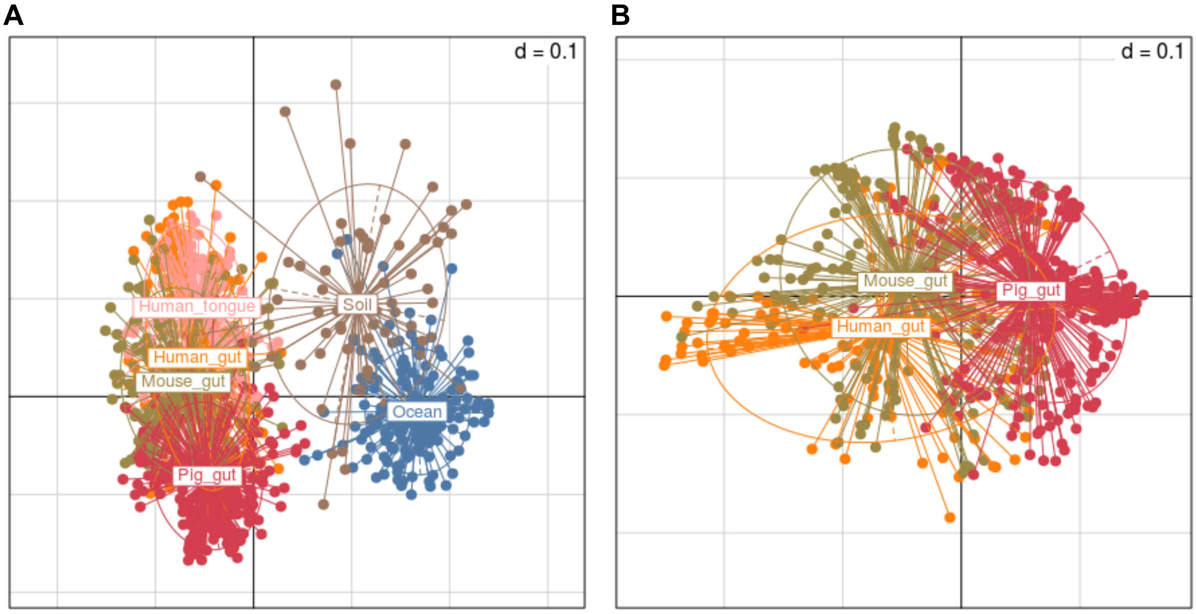

Figure 16: Pfam profile of minimal consortia of metagenomes selected by MiMiC. The MDS plots of **(A)** all categories and **(B)** only gut samples based on Pfam union profiles of minimal microbial consortia generated by MiMiC. The colors represent the sample category and each dot is a metagenome.

4.10 MiMiC-selected species recapitulated differences between the different environment systems

The microbial community proposed by MiMiC varied across the metagenomes of the same ecosystem. ***Certain species were found to be more prevalent within and across the ecosystems while some were mutually exclusive.*** I combined the MiMiC species profile of individual metagenomes in a group and applied the prevalence approach. I selected the top 10 most prevalent species per category and compared them to find the most common and unique species among ecosystems (Figure 17). *Anaerotruncus sp. 1XD22-93* species was prevalent in more than 50% metagenomes from the mouse and human gut, 40% in the human tongue, and 25% in pig gut. The same species had less than 10% prevalence in environment samples (soil and ocean). *Anaerotruncus sp. 1XD22-93* was isolated from a murine cecal sample (PRJNA486904). *Anaerotruncus* genus had been reported in the human tongue and human intestine as well (B. Sun et al. 2017; Lau et al. 2006). The soil samples had the least prevalent species whereas gut samples had the most prevalent and common species between different organisms. *Candidatus Kinetoplastibacterium galatii, Thermobifida fusca, Hydrotalea flava* were the most prevalent species and only picked in soil metagenomes*.* *Thermobifida fusca* is an aerobic, thermophilic soil bacterium known to be a degrader of plant cell wall (Lykidis et al. 2007). *Prevotella sp. Archaeoglobus profundus, alteromonas sp. 190, Candidatus Pelagibacter sp. RS39* were picked only in ocean metagenomic samples. *Nostoc linckia, Lewinella nigricans, Mesorhizobium sp. F7* are few more species that were most prevalent in the ocean and least in gut minimal consortia. *Campylobacter sp. 10_1_50, Pasteurella caecimuris, Rodentibacter myodis, Neisseria sp. HMSC07C12, Campylobacter sp.AAUH-44UCsig-a, Streptococcus sp. BCA20,* were most prevalent and only present in human tongue specific minimal consortia. *Parabacteroides sp. AM44-16, Bacteroides thetaiotaomicron,Clostridium cavendishii,* were only present in human minimal microbial consortia. Prevotella sp. P5-126 was present only in pig gut minimal microbial consortia and prevalent about 50% metagenomes. Prevotella sp. P5-126 is an anaerobic bacterium. It was isolated from the large intestine of pig (PRJNA396820). *Bacterium D16-50, Streptococcus sp. FDAARGOS_521, Geosporobacter ferrireducens,* were more prevalent in pig gut consortia. *Lachnospiraceae bacterium 28-4, Lactobacillus johnsonii, Bilophila wadsworthia* were prevalent only in mouse gut metagenomes. *Lachnospiraceae bacterium 28-4 species is anaerobic* and isolated from mice (PRJNA175982). There was a pattern of high prevalent species that were more specific to certain groups than less prevalent species.

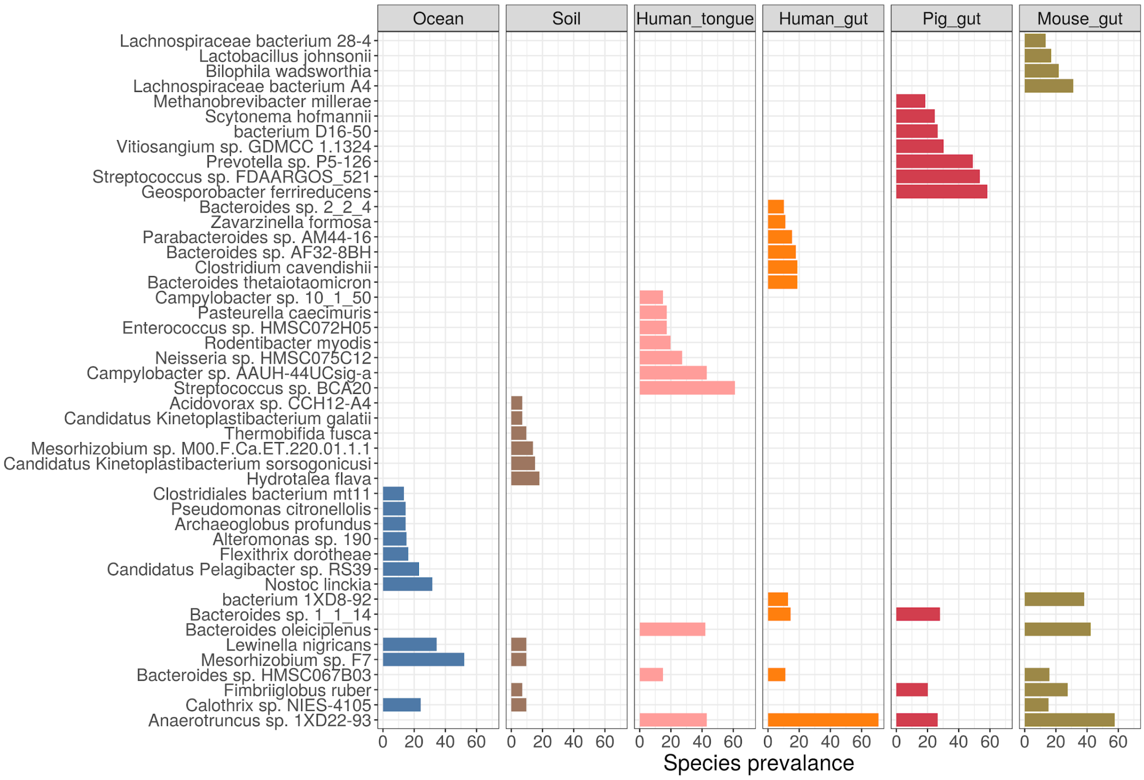

Figure 17: Top 10 MiMiC derived species prevalence across ecosystems. The prevalence of the top 10 most prevalent species derived by MiMiC in each category across ecosystems.

4.11 Predicting Minimal microbial consortia for Pig gut metagenomes

4.11.1 Pfam profiles of pig gut metagenomes could be distinguished based on their geographical condition

Pigs have been used as model organisms for microbial infectious diseases related studies in humans (Meurens et al. 2012). I used a collection of 111 bacterial species cultured from pig intestine (PiBAC) to MiMiC the profile of pig gut metagenome. I used 284 metagenomes from the study of pig gut gene catalog (Xiao et al. 2016). I processed these metagenomes from raw reads to protein prediction and generated Pfam binary vector profiles.

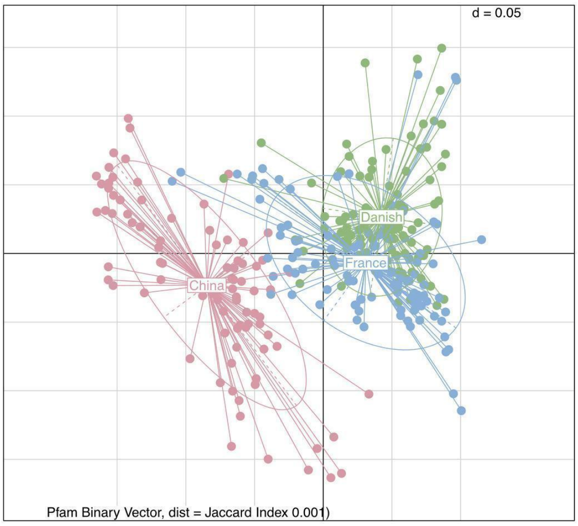

Figure 18: MDS plot of pig gut metagenomes based on full functional profile. The MDS plot of pig gut metagenomes generated using the full Pfam function profiles. The colors represent the country source of the metagenomes.

These metagenomes formed distinct clusters based on their source country suggesting country specificity in their gut microbiota (Figure 18). The samples from France and Denmark showed higher similarity compared to those from China. All metagenomes were able to be classified only country wise and not based on facility, gender or diet. I planned our research to predict a minimal microbial consortium and analyse the geographical (country) effect on pig gut microbial functions.

4.11.2 Function coverage of pig gut metagenomes by PiBAC species

I looked into the function coverage (Pfam) of metagenomes of pig gut using PiBAC (Wylensek et al. 2020) dataset as reference. A union Pfam vector was created of 111 PiBAC bacterial species to be used to match with the Pfam profiles of all the pig gut metagenomes. The Pfam coverage percentage of each metagenome was calculated by dividing the total number of Pfam matches by the total number of Pfams in the metagenome. ***On average more than 90% of Pfams were covered by 111 bacterial species in the pig metagenome*** (Figure 19A). The Pfam coverage was relatively higher in France compared to China and Danish metagenomes (Figure 19B).

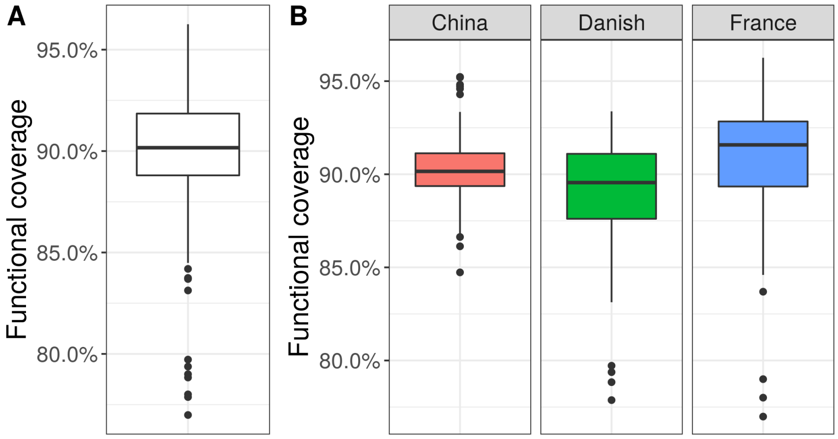

Figure 19: Pfam coverage of pig gut metagenomes by the PiBAC reference genome datasets for **(A)** all metagenomes and **(B)** per country.

4.11.3 Applying MiMiC to predict the minimal microbial consortia

A minimal microbial consortium was generated for each metagenome using MiMiC. I let the bacteria to be picked until no Pfam left in the reference genome database which matches to the metagenome. The species up to the knee point (uik) were considered as the minimal microbial consortia for the respective metagenomes. On average 20 species were selected from PiBAC based on knee points for each metagenome.

Surprisingly, 75 of the 111 species were picked in any of the minimal microbial consortium for all metagenomes (Figure 20A). Some species were more prevalent in specific countries than in others. The prevalence of the MiMiC picked species were calculated across all three countries (n=284) and within the country group (Figure 20B**)**. I analysed the results setting up different prevalence cutoffs. There were only 17 species that had a prevalence more than 50% across all the metagenomes.

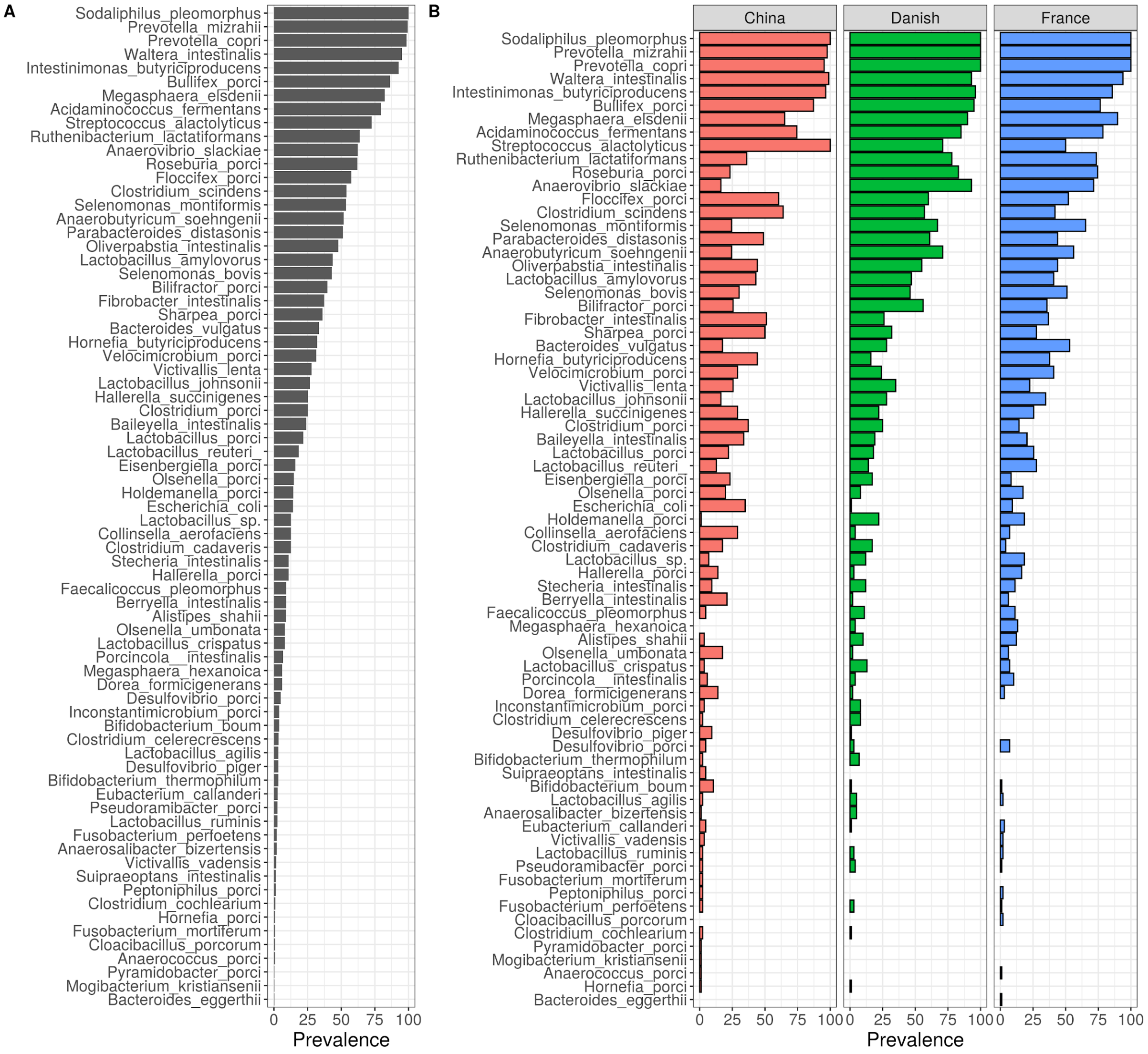

Figure 20: Prevalence analysis of MiMiC derived species. (A) The prevalence of each species across all the metagenomes and **(B)** per source country.

Some species were found to be more prevalent in a specific country than the whole cohort. I included these species by selecting the species present at least in 50% of the metagenomes of any of the countries. Following this criteria I added 8 more species in our selection. These 23 species represent the group of species specific to the country and also most common across the countries. Species *Sodaliphilus pleomorphus*, *Prevotella mizrahii, Prevotella copri, Waltera intestinalis, Intestinimonas butyriciproducens, Bullifex porci*, occured in more than 75% of metagenomes in each categories. Species *Ruthenibacterium lactatiformans, Roseburia porci, Anaerovibrio slackiae* had less than 40% prevalence in China while more than 75% prevalence in France and Danish metagenomes. *Fibrobacter intestinalis* and *Sharpea porci* were more prevalent in China compared to the France and Danish metagenomes.

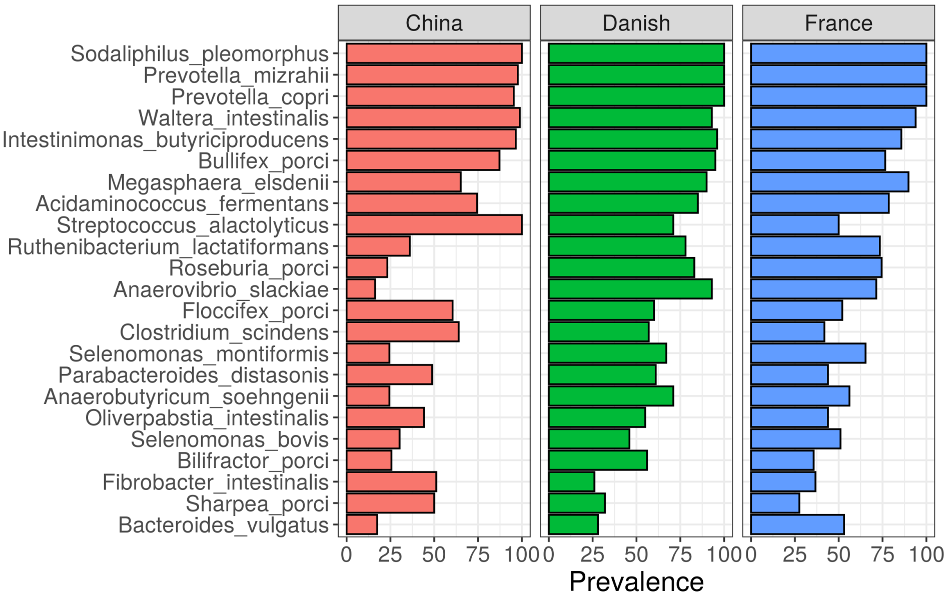

Figure 21: The prevalence of MiMiC picked species based on 50% cutoff in at least one country.

4.11.4 Function coverage by MiMiC-selected 23 species in pig gut metagenomes

These 23 species provided a generalized function based minimum microbial consortium for pig gut metagenomes based on PiBAC culture collection. I calculated the function coverage of pig gut metagenomes by these 23 species and compared it with the randomly selected species of the same sample size. The function coverage by these 23 species was not uniform for all the metagenomes (Figure 21B). On average these species covered 75% of the metagenomic functions. The most functions were covered in France metagenomes.

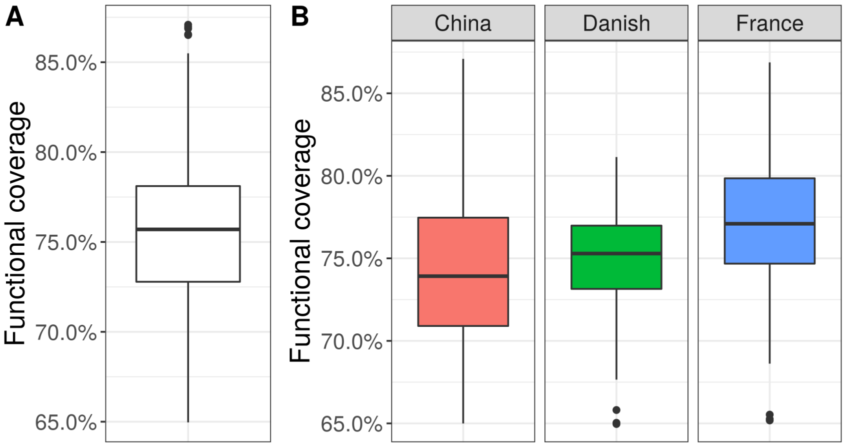

Figure 22: The function coverage of pig gut metagenomes by 23 MiMiC picked species. (A) in the whole metagenomic cohort **(B)** per source country.

Additionally, I compared the function coverage by 23 species with the function coverage of 500 random sets from the PiBAC collection for each metagenome. This provided an overview of how good these 23 species were compared to random sets. Function coverage and mismatches were calculated for each individual metagenome with each reference set of 23 species of 500 random sets from PiBAC. Medians of function coverage and mismatches were calculated for each metagenome of 500 random sets of 23 species for each metagenome. Further median of the median function coverage and median of median mismatches of randomly picked sets were compared with the MiMiC-selected 23 species.

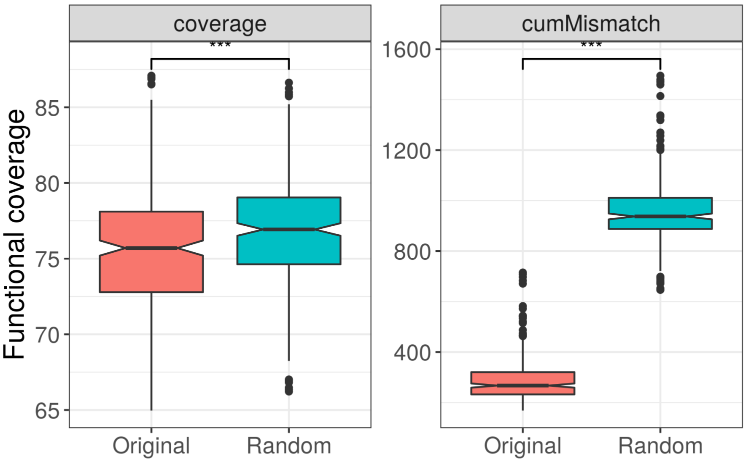

Figure 23: Comparison of MiMiC derived species with random sets. Function coverage (left) and mismatches (right) comparison between the MiMiC derived 23 species with randomly selected 23 species. The p-value significance of wilcoxon rank sum test is indicated by stars.

MiMiC-selected 23 species covered a similar number of functions compared to a median of 500 random sets of 23 species in all metagenomes (Figure 23). ***The functional coverage of random sets of 23 species was significantly higher compared to original functional coverage, the cumulative mismatches were also significantly increasing. However, the increase in the coverage is only 1.2% whereas in case of mismatches, I see an increment of more than 100% in random sets.*** MiMiC picked 23 species that might contribute or cover the same amount of function as in random sets; however, the number of mismatches were always lower in MiMiC Picked species (23) than in the random sets.

4.12 Predicting minimal microbial consortia for mice gut metagenomes

4.12.1 Pfam profiles of mice gut metagenomes can be distinguished based on their gut location and disease susceptibility

I applied MiMiC on another model organism mouse. I applied MiMiC on a case control condition like the effect of DSS-colitis susceptibility in mice gut microbiota. For this purpose I took mice gut metagenomes from colitis susceptible and non-colitis susceptible groups. Following the DSS treatment, it was found that not all mice developed colitis. I compared the Pfam profile of metagenomes and found that the conditions formed two distinct clusters (Figure 24B). Furthermore, the Pfam profile of metagenomes from different gut locations (ileum, colon and cecum) also formed separate clusters (Figure 24A). It was expected as the small intestine is supposed to harbour more bacterial species compared to ileum and cecum. However, this finding suggested that there was a possibility to have a distinct consortium representing the colitis susceptibility in different gut locations.

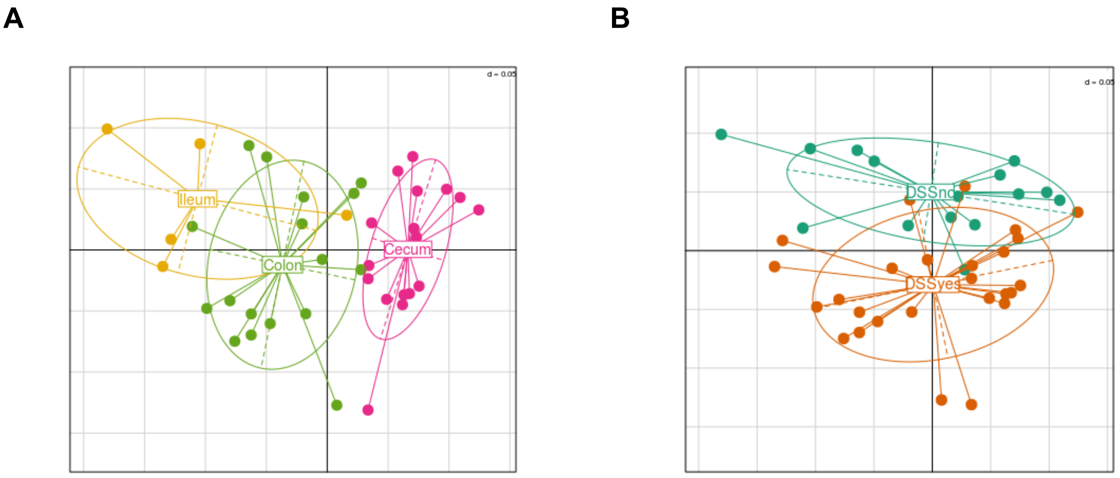

Figure 24: Pfam profile based classification of mouse gut metagenomes. MDS plot based on Pfam profile of metagenomes from **(A)** mice gut locations (colon, cecum, and ileum) and **(B)** DSS-colitis (DSSyes) and non-colitis (DSSno) susceptible mice.

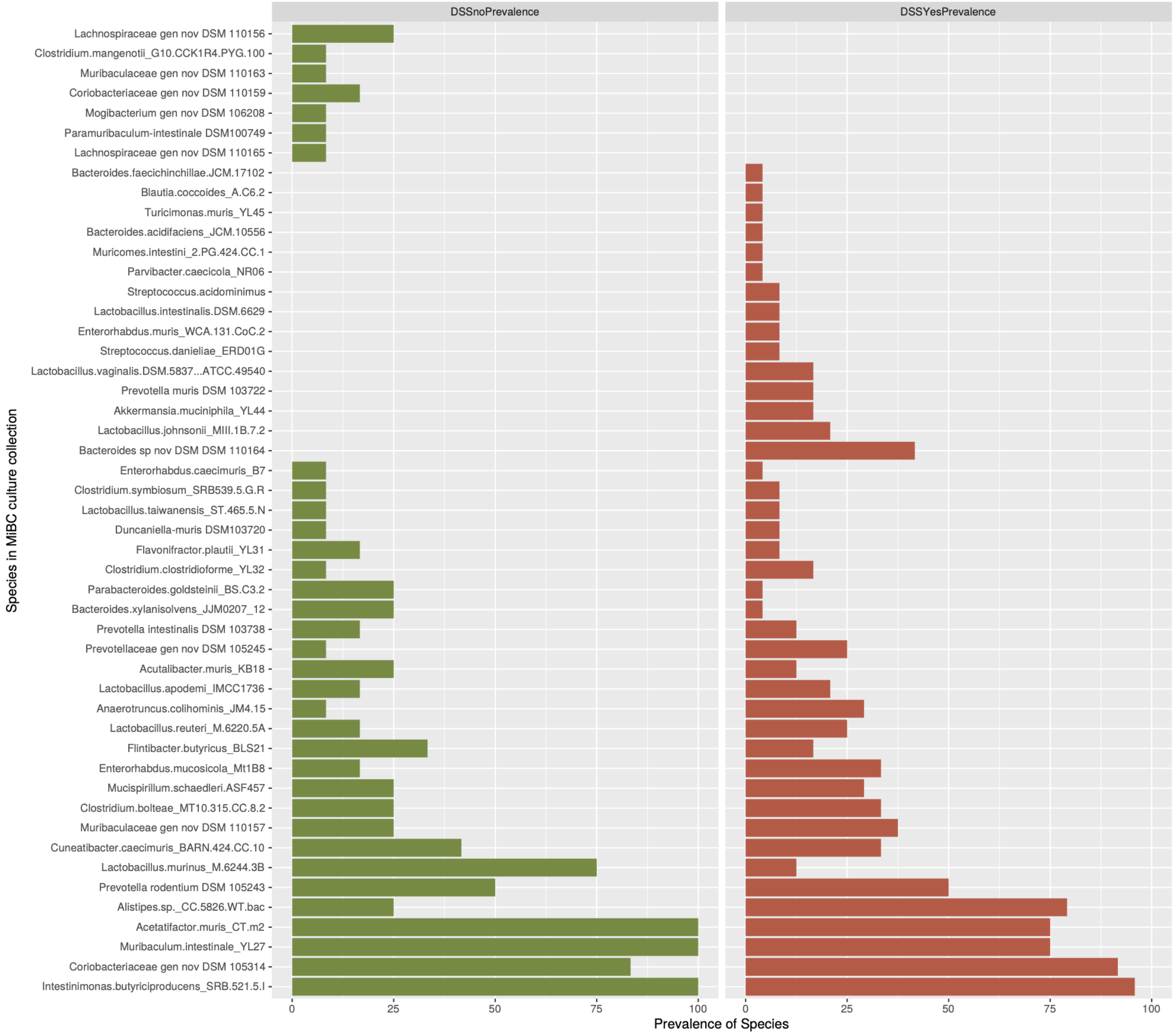

Figure 25: The prevalence of MiMiC picked species in non-colitis susceptible (DSSnoPrevalence) and DSS-colitis susceptible (DSSyesPrevalence) mice.

4.12.2 Applying MiMiC to predict the minimal microbial consortium

I applied MiMiC on colon and cecum metagenomes to observe the difference between DSS-colitis susceptible and non-susceptible Pfam profile. I took 36 metagenomic samples (Colon = 18, Cecum = 18) for disease specific MiMiC study where 24 samples were from DSS-colitis susceptible and 12 from non-susceptible mice. Minimal microbial consortia were generated for each metagenome using MiMiC. As a reference dataset I used 104 bacterial species from in-house mouse intestinal microbial collection. I used the iteration cutoff 50 and knee point method *uik* for the selection of minimal consortia. The Pfam profiles varied across individuals so it was not expected to be identical within condition or group of mice. However, some bacterial species were found to be specific to the condition while some were commonly present across all the mice (Figure 25). MiMiC found 49 out of 104 species from the reference dataset across all the metagenomes with the iteration cutoff 50 and knee point method uik.

After generating an individual minimal microbial consortium for each metagenome, I calculated the prevalence of each species between the group of DSS-colitis and non-colitis susceptible metagenomes.

With the prevalence analysis I found 15 species exclusive to DSS-colitis susceptible while 7 for non-susceptible mice and 28 present across all (Figure 25). In some cases, despite the presence of species in both the categories there was a significant difference in the prevalence between two groups. For example, *Lactobacillus murinus* was more prevalent (75%) in non-colitis susceptible than the DSS-colitis susceptible (<20%) mice metagenomes. *Lactobacillus murinus* had shown to mediate the anti-inflammaging effects in calorie restricted mice (Pan et al. 2018). This could be one reason why *"Lactobacillus murinus*" was picked in non-colitis (DSSno) specific minimal microbial consortia.

Acetatifactor.muris_CT.m2, Muribaculum.intestinale_YL27, Coriobacteriaceae gen nov DSM 105314, intestinimonas.butyriciproducens_SRB.521.5.1, were some species present in 75% metagenomes of both the non-colitis and DSS-colitis susceptible mice. *Lachnospiraceae.gen nov DSM 110156, Clostridium.mangenotii_G10.CCK1R4.PYG.100, Muribaculaceae.gen nov DSM 110163, Coriobacteriaceae gen nov DSM110159, Mogibacterium gen nov DSM 106208, Paramuribaculum-intestinale DSM100749, NM72.1.8_Lachnospiraceae gen nov DSM 110165*, are 7 species that were prevalent only in non-colitis susceptible mice. *Bacteroides sp nov DSM110164, Lactobacillus.johnsonii_MIII.1B.7.2, Akkermansia.muciniphilia_YL44, Prevotella.muris DSM 103722, Lactobacillus.vaginalis.DSM.5837...ATCC.49540* are 5 species that were prevalent in more than 20% colitis susceptible mice gut metagenomes. Several members of the Prevotella genus had been associated with colitis disease (Iljazovic et al. 2021; Peaper et al. 2011).The genus *Bacteroides* had been shown to be increased in DSS induced colitis group mice (Kanwal et al. 2020). ***These results suggested that MiMiC was able to derive species that were functionally common across different conditions as well as the exclusive species related to certain conditions.***

4.13 MiMiC as a bioinformatic pipeline

We established that MiMiC could predict the functionally representative community for an ecosystem with most related matches from a reference datasets. It allowed to include the community member with shared function by matching Pfam profile with metagenome and unique function by excluding the function shared by other community members. I incorporated all concepts discussed above and integrated them into a bioinformatic pipeline to provide a quick solution to find a functionally representative minimal microbial community from a given metagenomic data. As an input, MiMiC needs a functional profile of an ecosystem in the form of a binary vector. While MiMiC is not a metagenomic processing pipeline, for the ease of usage, I incorporated most commonly used metagenomic data processing tools into a simplified pipeline and provided separate scripts to preprocess the metagenomic reads, assign them to proteins, predict ORFs and scan the Pfams (Figure 26A). The right panel of Figure 26 describes the microbial species included in the building block of the reference datasets. I included three host specific reference genome datasets from human, mice and pig described in the method section. These datasets provide an opportunity to predict a host specific minimal microbial consortium while species from NCBI refseq genome database provide a wider range.

The middle part of the Figure 26 is the central and core part of MiMiC which takes input from both the other panels as Pfam binary vectors to provide the best selection of species mimicking the most functions of metagenome from the left panel. The preprocessing steps to handle metagenome data are wrapped up as shell scripts while the Pfam parsing is provided as perl script. The core part of MiMiC to generate minimal microbial consortium is written in R. The collection of assorted scripts to process individual steps provide flexibility to the user to start from any step. The final MiMiC output provides summary statistics including species accession, Pfam matches and mismatches for each iteration, cumulative percentage of Pfam coverage in metagenome, number of novel Pfams added at every iteration and the MiMiC score used to determine the species rank. Moreover, the user also has a flexibility to select the number of species to be printed in the final outcome as a defined number or automated selection by knee point method. MiMiC will summarise all the observed species until the user defined cutoff and the rank is given in the descending order of MiMiC-score.

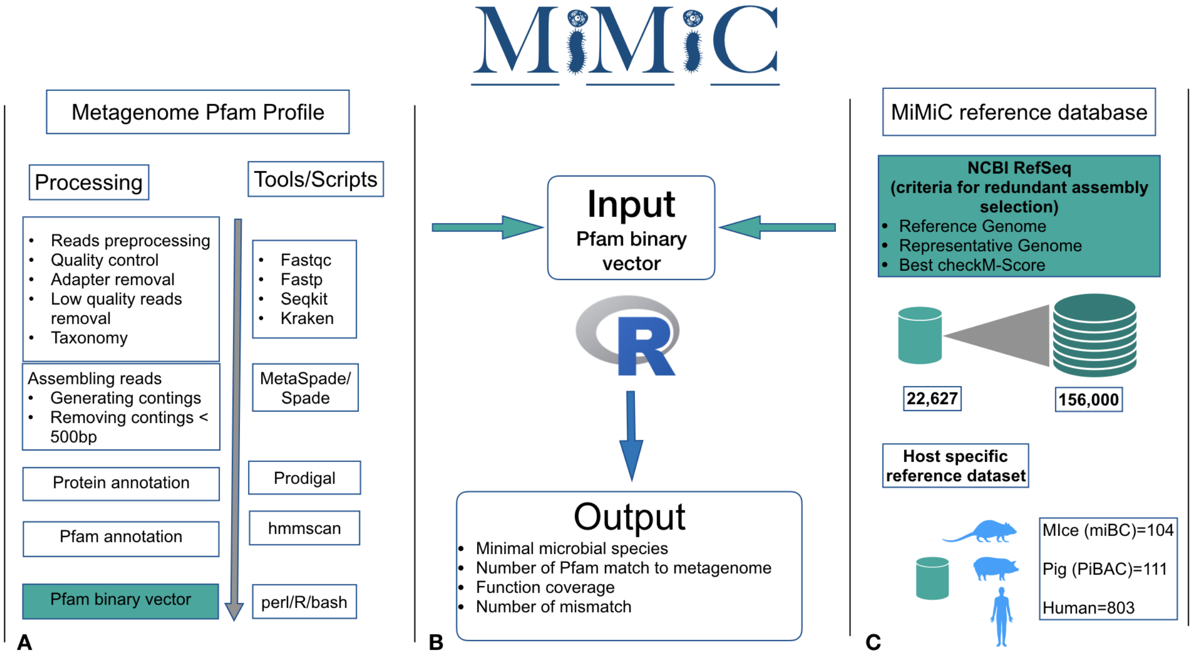

Figure 26: MiMiC infrastructure as a bioinformatic pipeline. The middle panel (B) shows the core part of MiMiC that takes input as Pfam binary vectors of the metagenome (left A) and reference genome database (right C) and generates a minimal microbial consortium.

5 Discussion

Building a model, mimicking an ecology of an environment or human/mice gut microbiota is a big challenge in the field of microbiological science. Several models have been defined for the similar purposes like SIHUMI, oligoMM, ASF, and MAMC (Minimal Active Microbial Consortia)(Puentes-Téllez and Falcao Salles 2018). These consortia were designed for certain purposes and can not be used for personalisation of minimal microbial consortium for individual’s need. Here, we propose a computational method to generate an individualized function based minimal microbial consortium of the metagenome in question.

Often not all the species in an environment have their isolates available and effectively no genomic reference of their functions. However, borrowing the information from functionally similar bacterial species could help getting better insights into the functional profiles of a certain metagenome. These species could be taken as representative and functional proxy of the bacteria whose genomic information is missing or not able to cultivate. Our function based minimal microbial consortia opens a new door to the scientific community to generate personalized, experiment specific, environment specific, and cultivable consortia.

As a matter of fact, it is impossible to devise an experiment directly on the host or huge environmental ecosystem. Therefore, there is always a need for a defined and representative synthetic model system which could mimic the environment or host specific microbiota. Most minimal microbial consortia are designed based on the taxonomic abundance, fermentation and their ability to be cultured. Here, I looked into a different possibility of selecting species from a pool of cultivable species based on their function capability instead of their taxonomic abundance in the ecosystem.

Our primary results support the assumption that the function profile of different ecosystems using the Pfam binary vector profile can be classified into the distinct cluster of respective ecosystems. These results provided a positive indication towards our approach that Pfam can be used to distinguish the metagenome from different ecosystems and can be used for further implementation to generate a representative minimal microbial consortium.

The MiMiC method serves a purpose of identifying a minimal microbial community for individual and more personalized needs. However, one needs to be aware of technical artefacts arising from human handling of the samples while proposing a microbiota of an ecosystem. For example, genetically similar mice from different vendors tend to have a different microbial composition (Ericsson et al. 2015). If an experiment is conducted, the response by the microbial community will vary in different facilities. It is practical to define a minimal microbial consortium which mimics the functions of a specific environment rather than proposing a single microbial consortium for different experiments and facilities.

I did detailed comparative analysis of different approaches to rank the representative species and selected the best approach for further implementation of methods to generate minimal consortia for different ecosystems. The proposed scoring system ranks the species based on the maximum coverage of the functions of the ecosystem while having the least irrelevant Pfam mismatch.

A huge part of the microbial world is still uncovered despite the rapid evolution in the high throughput sequencing technologies (Filée et al. 2005; Lloyd et al. 2018). MiMiC as a bioinformatic method, combining the functional profile of cultivable isolated bacterial genome will provide an initial solution to the current challenge of selecting bacterium for minimal microbial consortia from a pool of cultivable species. We initiated the approach to predict the minimal microbial consortia for two model organisms mouse and pig. A cultivable culture collection of pig (Wylensek et al. 2020) and mice (Lagkouvardos, Pukall, et al. 2016) provide a wide range of cultivable species from pig and mice intestine. We demonstrate that these species cover a significant proportion of metagenomic functions from their host environment. Pfam annotated profile of bacterial species for pig, mice and human (Zou et al. 2019) intestine provides a platform to predict host specific culturable minimal microbial consortia. Upon the success of the applicability of this approach on a host specific reference, we extended our reference genome database to be more inclusive. We included all bacterial and archaeal species available in NCBI RefSeq genome database and processed them to the minimum redundancy. The reduced redundancy provides better precision and lower computational processing cost. Our reference genome database consists of the best possible assembly quality genomes for each species. In order to assess the impact of biases due to genome and annotation quality of the reference genome database, we conducted the correlation study of these factors with the Pfam annotations. Unsurprisingly, our results show that the genome size of the microbial species correlates with the number of genes, proteins and Pfams. Given that the Pfam matches relies on sequence similarity, bigger and better resolved genomes could be prioritised in the minimal consortium if we consider only the number of Pfam matches between the genome and the metagenome. This brings another question of deconvoluting the real representative genome from the genome size bias. In a refined and final version of MiMiC we considered the number of Pfam mismatches between the microbial species and metagenome to penalize the calculated MiMiC-score to minimize the bias towards the bigger genome size. Using this scoring system, first we analysed the MBARC mock community where the ground truth of the species in the community is known. We could show that our approach allows the unbiased selection of the species in the rank order of their abundance level in the community.

The MiMiC derived consortium of MBARC mock community represents 97% of functional profile and 92% taxonomic relevance despite the size of reference database (n=22,627). MiMiC was able to pick up all three archaeal species which provides an assurance and certainty of the approach. Furthermore, giving higher priority to the species with the novel function in each iteration increases the chances of more shared and exclusive function between the species. The automated species selection by knee point derivation also provides sound results with respect to the number of species picked in consortium. We also showed that our approach can not only successfully distinguish hosts but also different environmental ecosystems. In this work, I investigated minimal microbial consortia from human, pig and mouse gut, different body parts of humans and also ocean and soil metagenomes. This way demonstrated the wide applicability of MiMiC on various ecosystems.

Considering that the MiMiC is a reference genome based approach, it is inevitable that the proposed representative minimal microbial consortium would depend on the selection and resolution of the reference dataset and that of the metagenome in question. To make the selection more specific to the host or relative environment we depend on the taxonomic lineage and the source from where the bacterium was isolated. These areas are still under refinement in NCBI and can influence the outcome of the MiMiC. I have created three host specific databases for human, pig and mouse which can be used to generate the minimal microbial consortia for the respective host. However, as a useful source for the future development we provided the taxonomic lineage for all the microbial species used for Pfam reference databases. To increase the precision of the outcome, a filtered reference dataset can be created using the taxonomic lineage of the Pfam reference database provided. This way it allows the MiMiC to predict the minimal microbial consortia from the taxonomically related host specific dataset.

After scoring the species, the next question is to select the minimal number of representative species from the given community. Consequently, I evaluated the number of iterations (species) to be selected with the maximal functional coverage. Generally, we found that the first 50 species provide the considerable representation of functional coverage and good ground to select the species visually based on their coverage information. Nonetheless, an automated approach to derive a defined minimal microbial consortium provides a more sensitive and individualistic approach for each metagenome. To this end, I compared two methods uik and d2uik from the R package ‘inflection’ to find a knee point on cumulative function coverage of metagenomes by species in minimal microbial consortia. The results indicate that the function d2uik gives a more robust knee point and is rarely dependent on the iteration number. However, it is more stringent with respect to the number of species picked due to detection of knee points at very low rank of species. On the other end, the method uik depends highly on the shape of the cumulative function coverage curve and effectively on the number of iterations the knee point is applied on. Our results indicate that when no less than 50 iterations are included, the knee point method uik derives substantial number species without discernible loss of information. However, both the options of knee point and the iteration cutoffs are provided for the analysis and a user can apply different options and explore the effects of different choices. The processing time depends on the number of species in the reference genome database.

As MiMiC output we provide a transparent results table containing the values for different steps used during the process at each iteration. We provide the reference genome accession for each species, cumulative function of metagenome covered by microbial species at each iteration, novel function added to the consortium by species, total number of Pfam present in microbial species and the metagenome both, Pfam used of species in each iteration, and absolute number of match and mismatch for each species to the metagenome. These outcomes provide a transparent view of the results and can be used for further exploration of the consortia and ecosystem.

For an affirmation of the MiMiC method we compared the MiMiC generated results with the same number of randomly selected species sets to make sure that consortium derived by

MiMiC always bring the community members covering significantly more functions than the average functions covered by random sets. MiMiC generated consortia always performed better with respect to relative number of Pfam match to metagenome than mismatch regardless of the choice of reference dataset or the ecosystem from where the metagenome was derived.

Despite the fact that shotgun sequencing of microbial DNA provides a deep insight of microbial ecology and functions of microbial species present in it, there are several primary factors like different sequencing technology, sequence length, sequencing depth, which influence the coverage of microbial DNA in respective samples (Gweon et al. 2019). This could be one reason which may lead to dissimilar results of the microbiota from the same environment. Among the 937 metagenomes analysed, we observed that the individual metagenomes from the same ecosystem do not have a similar microbial consortia profile.

The accuracy of the whole approach of scanning the Pfams which the backbone of MiMiC ultimately depends on how accurately the sequenced reads can be assembled and the coding sequences can be predicted. A technical bias can cause a big difference even in technical replicates from the same source of sample collection of microbial DNA. For example irrespective of low abundance of *Ruminococcus bromii* compared to any other highly abundant species of amylolytic bacteria in human colon (Ze et al. 2012), it has superior capability to degrade particulate resistant starch and perform a role of keystone species (Ze et al. 2013). A Microbial sample with low depth of sequencing or with more sequencing error can miss to build a genome reconstruct or identification of a species through taxonomic or binning tools compared to the sample with higher depth and coverage of sequencing. Having said this, it is highly recommended to revisit the goal of the study before sequencing a metagenome and carefully evaluate the sequencing quality and depth.

Apart from using MiMiC to derive a minimal consortium per metagenome, one can also analyse the prevalence of the species from MiMiC consortia across several metagenomes from the same ecosystem for wider applicability. Our results indicate that the prevalence of the species per ecosystem can be used to identify the most and least prevalent species in the ecosystem to propose a set of conclusive and a generalized minimal microbial consortium for that ecosystem. We demonstrated this approach using reference datasets, pig intestinal bacterial collection (PiBAC), mice intestinal bacterial collection (miBC) and NCBI RefSeq genome database to generate the minimal microbial consortia for model organism pig, mice and in a extended study to the human and environmental metagenomes like soil and ocean as well. Having said this, one can also observe that the soil samples represent higher variability amongst the metagenomes taken from various sites compared to the metagenomes of living organisms. The most prevalent species also show the presence in maximum 20% of the metagenomes of soil while in any of the organismal metagenome samples the prevalence goes over 60%. This indicates that the organismal ecosystems are comparatively more similar than the environmental ecosystems.

We demonstrated using MiMiC that 18 species (MiBAC-1) which were selected based on the pre taxonomic abundance, cover the most function in mouse gut metagenome compared to both ASF and SIHUMI (Lagkouvardos, Pukall, et al. 2016). However, numbers of species were not equal in all the consortia and significantly higher numbers of species in MiBAC-1 could propose better functional coverage in the mouse gut metagenome.

We found that pig metagenomes profile from different geographical areas can be distinguished in distinct clusters. In our study, we showed that pig metagenomes from different countries China, France and Denmark made distinct clusters based on their Pfam profile (Wylensek et al. 2020). We used PiBAC bacterial collection to identify the minimal microbial consortia for the metagenomic dataset from study of pig gut gene catalog (Xiao et al. 2016). The study demonstrates that MiMiC derived 23 species cover the maximum function than the randomly picked set of 23 species. At the same time MiMiC gives an initial ground to consider the species to be selected in a minimal microbial consortium. We found that there are 36 species which were not picked by any of the metagenomes. This brings the benefits that MiMiC can sort out the functionally least important species from the consortia. Besides, the comparison of 23 most prevalent species with random sets gives assurance to MiMiC and the prevalence approach to predict a generalized set of microbial communities which cover the considerable range of function in all the metagenomes from different geographical and breeding backgrounds.

Moreover, MiMiC is able to provide the species which are more prevalent in disease susceptible mice compared to non susceptible mice. Using miBC we found that some species were more prevalent in the disease susceptible mice and less prevalent in non susceptible mice. This suggests that MiMiC is able pick the species which share the similar function in the different conditions/environment. At the same time it is able to derive the species which are more susceptible or responsible for the disease or disease-free environment. At the same time we demonstrated that the mice gut metagenomes can be classified based on the Pfam profile from different parts of the mouse gut like colon, ileum, and cecum. Same fact is applied to function annotation. Still there is a need for further exploration of these species and their function. We hypothesize that common species are responsible for general function and unique species are the one which are responsible for causing colitis.

Microbial Gene catalogs from different organisms provide a wonderful opportunity to build a functional profile based on the genes and sequencing data available (Almeida et al. 2021; J. Li et al. 2014; Xiao et al. 2016, 2015). On the other hand the prokaryotic genomes from different culture collections and databases make it possible to use the functional profile of individual species (Wylensek et al. 2020; Lagkouvardos, Pukall, et al. 2016; Zou et al. 2019). Gene catalogues can be translated into functional profiles as Pfam binary vectors using preprocessing MiMiC pipeline. Furthermore, MiMiC can be used to provide a minimal set of cultivable microbial species for human, mice and pig gut microbiota based on gene cataloge translated profiles. Predicted consortia could further be used in germ free mice to mimic a gut environment from respective hosts. At the same time projects like MetaHit (Metagenomics of the Human Intestinal Tract) cover a wide population and diversity of human gut microbiota including healthy and patient data. MetaHit provides access to microbial profiles from Inflammatory bowel disease (IBD) Ulcerative colitis (UC) and Crohn's disease (CD). MiMiC can be applied to provide a further insight of function based minimal microbial communities specific to disease or healthy human microbiota.

Behzad D. Karkaria *et al. 2021* proposed the automated design of the preselected species using bacteriocin as anti-microbial peptides to inhibit the similar strain to be outnumbered then others in the same environment (Karkaria, Fedorec, and Barnes 2021). This method integrated the quorum sensing, chemostat with bayes approach to monitor the growth of the species in certain environments. However, it did not provide the ground of how to initiate selecting species from an ecosystem at first hand. Our approach of MiMiC provides a unique platform to the user to select the species from an environment taking advantage of the metagenomic sequencing.
MiMiC is implemented as a method using R, perl and shell scripts. The core method of generating Pfam binary vectors and calculating MiMiC-score are written in R. The Pfam annotation of metagenomes or genomes are parsed using perl script. A Pfam binary vector of the reference genome and the metagenome are expected as inputs to the main script of MiMiC. For the ease of usage, I extended the repository to include scripts to generate the metagenomic assembly, taxonomic profile, coding sequence prediction and Pfam annotation. The Pfam binary vectors of the host specific datasets for pig (n=111 species), mice (n=104 species) and human (n=803 species) are provided to be used directly for the host specific applications. Since all these species are cultivable, the proposed minimal microbial consortia using these reference datasets provide a cultivable set of microbiome communities. I extended the reference genome repository to the RefSeq genome database and provided a Pfam binay vector of 22,627 species. The user can use their own customised reference genome datasets to predict the functionally best representative species to the metagenome. Benchmarking of the iteration number cutoff, knee point methods, and different versions of MiMiC provides a reference guide to the user to set up and use different cutoff and choice of methods. For example, a minimal microbial consortium can be generated by a defined number of community members or it can be decided based on the knee point choice. ***All in all, MiMiC is the approach first proposed to generate a representative cultivable set of microbiome communities from metagenome(s). MiMiC provides flexible, transparent and open source scripts amenable for easy adaptation to the user's need for every feature.***

6 Conclusion and perspective

In this study, I laid the foundation for important building blocks of a bioinformatic pipeline to select species for automated generation of function-based minimal microbial consortium. I built a non-redundant prokaryotic reference genome database in terms of Pfam profile, developed and optimised a method to generate minimal microbial community to represent an ecosystem, validated this approach on a mock community and proposed minimal consortia for various ecosystems. I showed that Pfam is a suitable unit to define the function profile of metagenomes and microbial genomes. The proposed Pfam profiles of several host specific datasets provide an opportunity to generate minimal microbial consortia of cultured species for further mechanistic studies. The wide range of processed species-level genomes from NCBI RefSeq provide an opportunity to explore multiple microbial communities and generate an individualized consortium. Minimal microbial consortia predicted for pig and mice intestinal microbiota can provide initial ground to select the strain for further study or designing synthetic microbial community to mimic the pig and mice intestinal gut microbiota in germ free models. The evaluation of MiMiC using a mock community (MBARC) provides a sound validation that MiMiC derived minimal consortia is not only functionally representative, but also taxonomically.

The core part of MiMiC exists as a flexible wrap up of R scripts which can easily be integrated as a shiny app. Such interface will enable users from diverse backgrounds to autonomously explore and interpret their own results. For instance, a user can select taxonomically relevant reference datasets and analyse the proposed microbial communities tailored to their needs. There are also other databases providing refined genomic and metadata information which can be integrated in the current MiMiC reference database (Kyrpides 1999; Mukherjee et al. 2021). If a user wants to use a pathogen free reference database, this can also be implemented by flagging the pathogens in the reference database obtaining information from available soure like https://www.ncbi.nlm.nih.gov/pathogens/organisms/ (McNeil et al. 2007). Apart from the proposed automated selection of species, the shiny app could provide flexibility for the end user to choose a custom number of species by looking at the cumulative graph interactively. Currently, MiMiC uses protein families (Pfam) as a classifier to select functionally closer species to the native ecosystems, however, in future other means of functional annotations can be implemented such as EggNOG, COG, GEMs and KEGG. These annotation methods can be used for special needs such as metabolic pathway based classification, orthologous group and metabolic models. Currently, MiMiC scoring system relies on metagenomic function coverage. Nevertheless, the scoring can be customised to suit the annotation method used.

Projects like the integrative Human Microbiome Project (iHMP), MetaHit have opened up exciting opportunities to better our understanding of human microbiota. Such comprehensive information will enable researchers to understand the complexity of the microbiota via proposed minimal consortia with MiMiC.

7 List of figures

8 List of tables

9 Abbreviations

PiBAC	Pig intestinal bacterial collection
miBC	Mouse intestinal bacterial collection
SIHUMI	Simplified human intestinal microbiota
ASF	Altered Schaedler Flora
iHMP	Integrative human microbiome project
NCBI	National center for Biotechnology information
RefSeq	Reference Sequence
MBARC-26	Mock Bacteria ARchaea Community
PCR	Polymerase chain reaction
rRNA	Ribosomal ribonucleic acid

RNA	Ribonucleic acid
DNA	Deoxyribonucleic acid
NGS	Next generation sequencing
MAGs	metagenome-assembled genomes
CDS	Conserved domain sequence
KEGG	Kyoto Encyclopedia of Genes and Genomes
GO	Gene ontology
COGs	Cluster of orthologous group
EggNOG	Evolutionary genealogy of genes Non supervised orthologous group
Pfam	Protein family
MDS	Multidimensional scale
HMM	Hidden Markov Model
IBD	Inflammatory bowel disease
FMT	Fecal Microbiota transplant
FBA	Flux balance analysis
GSM	Genome scale metabolic reconstruction
BactPfamSigDB	Bacterial Pfam signature database

10 References

1. Aguiar-Pulido, Vanessa, Wenrui Huang, Victoria Suarez-Ulloa, Trevor Cickovski, Kalai Mathee, and Giri Narasimhan. 2016. “Metagenomics, Metatranscriptomics, and Metabolomics Approaches for Microbiome Analysis.” *Evolutionary Bioinformatics Online* 12 (Suppl 1): 5–16. https://doi.org/10.4137/EBO.S36436.

2. Alexander, A. Deloris, Roger P. Orcutt, Janell C. Henry, Joseph Baker, Anika C. Bissahoyo, and David W. Threadgill. 2006. “Quantitative PCR Assays for Mouse Enteric Flora Reveal Strain-Dependent Differences in Composition That Are Influenced by the Microenvironment.” *Mammalian Genome: Official Journal of the International Mammalian Genome Society* 17 (11): 1093–1104. https://idp.springer.com/authorize/casa?redirect_uri=https://link.springer.com/article/10.1007/s00335-006-0063-1&casa_token=H31YviaYUiQAAAAA:Ne03WGXQ4RAbPL4xP655TJIOaRnJCurwIDF2Cpox2PXd7I28ByiyVi5q_fIQZfEeIkisdevhyd_42yL8Pg.

3. Almeida, Alexandre, Stephen Nayfach, Miguel Boland, Francesco Strozzi, Martin Beracochea, Zhou Jason Shi, Katherine S. Pollard, et al. 2021. “A Unified Catalog of 204,938 Reference Genomes from the Human Gut Microbiome.” *Nature Biotechnology* 39 (1): 105–14. https://doi.org/10.1038/s41587-020-0603-3.

4. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

5. Apweiler, R., T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, et al. 2001. “The InterPro Database, an Integrated Documentation Resource for Protein Families, Domains and Functional Sites.” *Nucleic Acids Research* 29 (1): 37–40. https://doi.org/10.1093/nar/29.1.37.

6. Apweiler, Rolf, Amos Bairoch, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, et al. 2004. “UniProt: The Universal Protein Knowledgebase.” *Nucleic Acids Research* 32 (Database issue): D115–19. https://doi.org/10.1093/nar/gkh131.

7. Armanhi, Jaderson Silveira Leite, Rafael Soares Correa de Souza, Natália de Brito Damasceno, Laura M. de Araújo, Juan Imperial, and Paulo Arruda. 2017. “A Community-Based Culture Collection for Targeting Novel Plant Growth-Promoting Bacteria from the Sugarcane Microbiome.” *Frontiers in Plant Science* 8: 2191. https://doi.org/10.3389/fpls.2017.02191.

8. Attwood, T. K., P. Bradley, D. R. Flower, A. Gaulton, N. Maudling, A. L. Mitchell, G. Moulton, et al. 2003. “PRINTS and Its Automatic Supplement, prePRINTS.” *Nucleic Acids Research* 31 (1): 400–402. https://doi.org/10.1093/nar/gkg030.

9. Babaei, Parizad, Saeed Shoaie, Boyang Ji, and Jens Nielsen. 2018. “Challenges in

Modeling the Human Gut Microbiome.” *Nature Biotechnology* 36 (8): 682–86. https://doi.org/10.1038/nbt.4213.

10. Bairoch, A. 2000. “The ENZYME Database in 2000.” *Nucleic Acids Research* 28 (1): 304–5. https://doi.org/10.1093/nar/28.1.304.

11. Balakrishnan, N., Theodore Colton, Brian Everitt, Walter Piegorsch, Fabrizio Ruggeri, and Jozef L. Teugels, eds. 2014. “Principal Coordinates Analysis.” In *Wiley StatsRef: Statistics Reference Online*, 27:297. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat05670.

12. Baldini, Federico, Almut Heinken, Laurent Heirendt, Stefania Magnusdottir, Ronan M. T. Fleming, and Ines Thiele. 2019. “The Microbiome Modeling Toolbox: From Microbial Interactions to Personalized Microbial Communities.” *Bioinformatics* 35 (13): 2332–34. https://doi.org/10.1093/bioinformatics/bty941.

13. Bankevich, Anton, Sergey Nurk, Dmitry Antipov, Alexey A. Gurevich, Mikhail Dvorkin, Alexander S. Kulikov, Valery M. Lesin, et al. 2012. “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing.” *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 19 (5): 455–77. https://doi.org/10.1089/cmb.2012.0021.

14. Bassols, Anna, Cristina Costa, P. David Eckersall, Jesús Osada, Josefa Sabrià, and Joan Tibau. 2014. “The Pig as an Animal Model for Human Pathologies: A Proteomics Perspective.” *Proteomics. Clinical Applications* 8 (9-10): 715–31. https://doi.org/10.1002/prca.201300099.

15. Bateman, Alex, Ewan Birney, Lorenzo Cerruti, Richard Durbin, Laurence Etwiller, Sean R. Eddy, Sam Griffiths-Jones, Kevin L. Howe, Mhairi Marshall, and Erik L. L. Sonnhammer. 2002. “The Pfam Protein Families Database.” *Nucleic Acids Research* 30 (1): 276–80. https://doi.org/10.1093/nar/30.1.276.

16. Becker, Natalie, Julia Kunath, Gunnar Loh, and Michael Blaut. 2011. “Human Intestinal Microbiota: Characterization of a Simplified and Stable Gnotobiotic Rat Model.” *Gut Microbes* 2 (1): 25–33. https://doi.org/10.4161/gmic.2.1.14651.

17. Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. 2012. “GenBank.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gks1195.

18. Bilen, Melhem, Jean-Charles Dufour, Jean-Christophe Lagier, Fréderic Cadoret, Ziad Daoud, Grégory Dubourg, and Didier Raoult. 2018. “The Contribution of Culturomics to the Repertoire of Isolated Human Bacterial and Archaeal Species.” *Microbiome* 6 (1): 94. https://doi.org/10.1186/s40168-018-0485-5.

19. Bokulich, Nicholas A., Jai Ram Rideout, William G. Mercurio, Benjamin Wolfe, Corinne F. Maurice, Rachel J. Dutton, Peter J. Turnbaugh, Rob Knight, and J. Gregory

Caporaso. n.d. “Mockrobiota: A Public Resource for Microbiome Bioinformatics Benchmarking.” https://doi.org/10.7287/peerj.preprints.2065.

20. Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. 2019. “Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2.” *Nature Biotechnology* 37 (8): 852–57. https://doi.org/10.1038/s41587-019-0209-9.

21. Bonnet, R. 2002. “Differences in rDNA Libraries of Faecal Bacteria Derived from 10- and 25-Cycle PCRs.” *INTERNATIONAL JOURNAL OF SYSTEMATIC AND EVOLUTIONARY MICROBIOLOGY*. https://doi.org/10.1099/ijs.0.01755-0.

22. Brenner, Katie, Lingchong You, and Frances H. Arnold. 2008. “Engineering Microbial Consortia: A New Frontier in Synthetic Biology.” *Trends in Biotechnology* 26 (9): 483–89. https://doi.org/10.1016/j.tibtech.2008.05.004.

23. Britschgi, T. B., and S. J. Giovannoni. 1991. “Phylogenetic Analysis of a Natural Marine Bacterioplankton Population by rRNA Gene Cloning and Sequencing.” *Applied and Environmental Microbiology* 57 (6): 1707–13. https://www.ncbi.nlm.nih.gov/pubmed/1714704.

24. Brugiroux, Sandrine, Markus Beutler, Carina Pfann, Debora Garzetti, Hans-Joachim Ruscheweyh, Diana Ring, Manuel Diehl, et al. 2016. “Genome-Guided Design of a Defined Mouse Microbiota That Confers Colonization Resistance against Salmonella Enterica Serovar Typhimurium.” *Nature Microbiology* 2 (November): 16215. https://doi.org/10.1038/nmicrobiol.2016.215.

25. Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. “Fast and Sensitive Protein Alignment Using DIAMOND.” *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

26. Buie, Timothy. 2015. “Potential Etiologic Factors of Microbiome Disruption in Autism.” *Clinical Therapeutics* 37 (5): 976–83. https://doi.org/10.1016/j.clinthera.2015.04.001.

27. Bushnell, Brian. 2014. “BBMap: A Fast, Accurate, Splice-Aware Aligner.” Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States). https://www.osti.gov/biblio/1241166.

28. Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. 2016. “DADA2: High-Resolution Sample Inference from Illumina Amplicon Data.” *Nature Methods* 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.

29. Canard, B., and R. S. Sarfati. 1994. “DNA Polymerase Fluorescent Substrates with Reversible 3’-Tags.” *Gene* 148 (1): 1–6. https://doi.org/10.1016/0378-1119(94)90226-7.

30. Carlos, Camila, Huan Fan, and Cameron R. Currie. 2018. “Substrate Shift Reveals Roles for Members of Bacterial Consortia in Degradation of Plant Cell Wall Polymers.”

Frontiers in Microbiology 9 (March): 364. https://doi.org/10.3389/fmicb.2018.00364.

31. Carradec, Quentin, Eric Pelletier, Corinne Da Silva, Adriana Alberti, Yoann Seeleuthner, Romain Blanc-Mathieu, Gipsi Lima-Mendez, et al. 2018. “A Global Ocean Atlas of Eukaryotic Genes.” *Nature Communications* 9 (1): 373. https://doi.org/10.1038/s41467-017-02342-1.

32. Caspi, Ron, Richard Billington, Ingrid M. Keseler, Anamika Kothari, Markus Krummenacker, Peter E. Midford, Wai Kit Ong, Suzanne Paley, Pallavi Subhraveti, and Peter D. Karp. 2020. “The MetaCyc Database of Metabolic Pathways and Enzymes - a 2019 Update.” *Nucleic Acids Research* 48 (D1): D445–53. https://doi.org/10.1093/nar/gkz862.

33. Castillo, Sandra, Dorothee Barth, Mikko Arvas, Tiina M. Pakula, Esa Pitkänen, Peter Blomberg, Tuulikki Seppanen-Laakso, et al. 2016. “Whole-Genome Metabolic Model of Trichoderma Reesei Built by Comparative Reconstruction.” *Biotechnology for Biofuels* 9 (November): 252. https://doi.org/10.1186/s13068-016-0665-0.

34. Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. “Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor.” *Bioinformatics* 34 (17): i884–90. https://doi.org/10.1093/bioinformatics/bty560.

35. Christopoulos, Demetris. 2016. “Introducing Unit Invariant Knee (UIK) As an Objective Choice for Elbow Point in Multivariate Data Analysis Techniques.” https://doi.org/10.2139/ssrn.3043076.

36. Clarridge, Jill E., 3rd. 2004. “Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases.” *Clinical Microbiology Reviews* 17 (4): 840–62, table of contents. https://doi.org/10.1128/CMR.17.4.840-862.2004.

37. Coenye, Tom, and Peter Vandamme. 2003. “Intragenomic Heterogeneity between Multiple 16S Ribosomal RNA Operons in Sequenced Bacterial Genomes.” *FEMS Microbiology Letters* 228 (1): 45–49. https://doi.org/10.1016/S0378-1097(03)00717-1.

38. Coico, Richard. 2005. “Gram Staining.” *Current Protocols in Microbiology* Appendix 3 (October): Appendix 3C. https://doi.org/10.1002/9780471729259.mca03cs00.

39. Cole, James R., Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. 2014. “Ribosomal Database Project: Data and Tools for High Throughput rRNA Analysis.” *Nucleic Acids Research* 42 (Database issue): D633–42. https://doi.org/10.1093/nar/gkt1244.

40. Consortium, The Gene Ontology, and The Gene Ontology Consortium. 2008. “The Gene Ontology Project in 2008.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkm883.

41. Consortium, Uniprot. 2018. “UniProt: The Universal Protein Knowledgebase.” *Nucleic Acids Research*. https://www.ncbi.nlm.nih.gov/pmc/articles/pmc5861450/.

42. Coskun, Mehmet. 2014. “Intestinal Epithelium in Inflammatory Bowel Disease.” *Frontiers of Medicine* 1 (August): 24. https://doi.org/10.3389/fmed.2014.00024.

43. Costea, Paul I., Georg Zeller, Shinichi Sunagawa, Eric Pelletier, Adriana Alberti, Florence Levenez, Melanie Tramontano, et al. 2017. “Towards Standards for Human Fecal Sample Processing in Metagenomic Studies.” *Nature Biotechnology* 35 (11): 1069–76. https://doi.org/10.1038/nbt.3960.

44. Coupland, Paul, Tamir Chandra, Mike Quail, Wolf Reik, and Harold Swerdlow. 2012. “Direct Sequencing of Small Genomes on the Pacific Biosciences RS without Library Preparation.” *BioTechniques* 53 (6): 365–72. https://doi.org/10.2144/000113962.

45. Cummings, J. H. 1981. “Short Chain Fatty Acids in the Human Colon.” *Gut* 22 (9): 763–79. https://doi.org/10.1136/gut.22.9.763.

46. Delcher, Arthur L., Kirsten A. Bratke, Edwin C. Powers, and Steven L. Salzberg. 2007. “Identifying Bacterial Genes and Endosymbiont DNA with Glimmer.” *Bioinformatics* 23 (6): 673–79. https://doi.org/10.1093/bioinformatics/btm009.

47. Dewhirst, F. E., C. C. Chien, B. J. Paster, R. L. Ericson, R. P. Orcutt, D. B. Schauer, and J. G. Fox. 1999. “Phylogeny of the Defined Murine Microbiota: Altered Schaedler Flora.” *Applied and Environmental Microbiology* 65 (8): 3287–92. https://www.ncbi.nlm.nih.gov/pubmed/10427008.

48. DiBaise, John K., Husen Zhang, Michael D. Crowell, Rosa Krajmalnik-Brown, G. Anton Decker, and Bruce E. Rittmann. 2008. “Gut Microbiota and Its Possible Relationship with Obesity.” *Mayo Clinic Proceedings. Mayo Clinic* 83 (4): 460–69. https://doi.org/10.4065/83.4.460.

49. Dilthey, Alexander T., Chirag Jain, Sergey Koren, and Adam M. Phillippy. 2019. “Strain-Level Metagenomic Assignment and Compositional Estimation for Long Reads with MetaMaps.” *Nature Communications*. https://doi.org/10.1038/s41467-019-10934-2.

50. Dinan, Timothy G., and John F. Cryan. 2015. “The Impact of Gut Microbiota on Brain and Behaviour: Implications for Psychiatry.” *Current Opinion in Clinical Nutrition and Metabolic Care* 18 (6): 552–58. https://doi.org/10.1097/MCO.0000000000000221.

51. Durbin, Richard, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. https://play.google.com/store/books/details?id=HUUhAwAAQBAJ.

52. Eddy, Sean R. 2004. “What Is a Hidden Markov Model?” *Nature Biotechnology*. https://doi.org/10.1038/nbt1004-1315.

53. Edgar, Robert C. 2013. “UPARSE: Highly Accurate OTU Sequences from Microbial

Amplicon Reads." *Nature Methods* 10 (10): 996–98. https://doi.org/10.1038/nmeth.2604.

54. El-Gebali, Sara, Jaina Mistry, Alex Bateman, Sean R. Eddy, Aurélien Luciani, Simon C. Potter, Matloob Qureshi, et al. 2019. "The Pfam Protein Families Database in 2019." *Nucleic Acids Research* 47 (D1): D427–32. https://doi.org/10.1093/nar/gky995.

55. Ericsson, Aaron C., J. Wade Davis, William Spollen, Nathan Bivens, Scott Givan, Catherine E. Hagan, Mark McIntosh, and Craig L. Franklin. 2015. "Effects of Vendor and Genetic Background on the Composition of the Fecal Microbiota of Inbred Mice." *PloS One* 10 (2): e0116704. https://doi.org/10.1371/journal.pone.0116704.

56. Escobar-Zepeda, Alejandra, Arturo Vera-Ponce de León, and Alejandro Sanchez-Flores. 2015. "The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics." *Frontiers in Genetics* 6 (December): 348. https://doi.org/10.3389/fgene.2015.00348.

57. Ewing, Brent, Ladeana Hillier, Michael C. Wendl, and Phil Green. 1998. "Base-Calling of Automated Sequencer Traces UsingPhred. I. Accuracy Assessment." *Genome Research*. https://doi.org/10.1101/gr.8.3.175.

58. Fang, Gang, Nitin Bhardwaj, Rebecca Robilotto, and Mark B. Gerstein. 2010. "Getting Started in Gene Orthology and Functional Analysis." *PLoS Computational Biology* 6 (3): e1000703. https://doi.org/10.1371/journal.pcbi.1000703.

59. Filée, Jonathan, Françoise Tétart, Curtis A. Suttle, and H. M. Krisch. 2005. "Marine T4-Type Bacteriophages, a Ubiquitous Component of the Dark Matter of the Biosphere." *Proceedings of the National Academy of Sciences of the United States of America* 102 (35): 12471–76. https://doi.org/10.1073/pnas.0503404102.

60. Filee, J., F. Tetart, C. A. Suttle, and H. M. Krisch. 2005. "Marine T4-Type Bacteriophages, a Ubiquitous Component of the Dark Matter of the Biosphere." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0503404102.

61. Finn, Robert D., Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, et al. 2014. "Pfam: The Protein Families Database." *Nucleic Acids Research* 42 (Database issue): D222–30. https://doi.org/10.1093/nar/gkt1223.

62. Finn, Robert D., Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, et al. 2016. "The Pfam Protein Families Database: Towards a More Sustainable Future." *Nucleic Acids Research* 44 (D1): D279–85. https://doi.org/10.1093/nar/gkv1344.

63. Finn, Robert D., Jaina Mistry, Benjamin Schuster-Böckler, Sam Griffiths-Jones, Volker Hollich, Timo Lassmann, Simon Moxon, et al. 2006. "Pfam: Clans, Web Tools and

Services.” *Nucleic Acids Research* 34 (Database issue): D247–51. https://doi.org/10.1093/nar/gkj149.

64. Finn, Robert D., John Tate, Jaina Mistry, Penny C. Coggill, Stephen John Sammut, Hans-Rudolf Hotz, Goran Ceric, et al. 2008. “The Pfam Protein Families Database.” *Nucleic Acids Research* 36 (Database issue): D281–88. https://doi.org/10.1093/nar/gkm960.

65. Fisher, Matthew. 2001. “Lehninger Principles of Biochemistry, 3rd Edition; By David L. Nelson and Michael M. Cox.” *The Chemical Educator*. https://doi.org/10.1007/s00897000455a.

66. Fitch, W. M. 1970. “Distinguishing Homologous from Analogous Proteins.” *Systematic Zoology* 19 (2): 99–113. https://www.ncbi.nlm.nih.gov/pubmed/5449325.

67. Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. 1995. “Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd.” *Science* 269 (5223): 496–512. https://doi.org/10.1126/science.7542800.

68. Francke, Christof, Roland J. Siezen, and Bas Teusink. 2005. “Reconstructing the Metabolic Network of a Bacterium from Its Genome.” *Trends in Microbiology*. https://doi.org/10.1016/j.tim.2005.09.001.

69. Frank, Daniel N., Charles E. Robertson, Christina M. Hamm, Zegbeh Kpadeh, Tianyi Zhang, Hongyan Chen, Wei Zhu, et al. 2011. “Disease Phenotype and Genotype Are Associated with Shifts in Intestinal-Associated Microbiota in Inflammatory Bowel Diseases.” *Inflammatory Bowel Diseases* 17 (1): 179–84. https://doi.org/10.1002/ibd.21339.

70. Franzosa, Eric A., Lauren J. McIver, Gholamali Rahnavard, Luke R. Thompson, Melanie Schirmer, George Weingart, Karen Schwarzberg Lipson, et al. 2018. “Species-Level Functional Profiling of Metagenomes and Metatranscriptomes.” *Nature Methods* 15 (11): 962–68. https://doi.org/10.1038/s41592-018-0176-y.

71. Fritz, Adrian, Peter Hofmann, Stephan Majda, Eik Dahms, Johannes Dröge, Jessika Fiedler, Till R. Lesker, et al. 2019. “CAMISIM: Simulating Metagenomes and Microbial Communities.” *Microbiome* 7 (1): 17. https://doi.org/10.1186/s40168-019-0633-6.

72. Galperin, Michael Y., Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin. 2015. “Expanded Microbial Genome Coverage and Improved Protein Family Annotation in the COG Database.” *Nucleic Acids Research* 43 (Database issue): D261–69. https://doi.org/10.1093/nar/gku1223.

73. Galperin, Michael Y., Yuri I. Wolf, Kira S. Makarova, Roberto Vera Alvarez, David Landsman, and Eugene V. Koonin. 2020. “COG Database Update: Focus on Microbial Diversity, Model Organisms, and Widespread Pathogens.” *Nucleic Acids Research*,

November. https://doi.org/10.1093/nar/gkaa1018.

74. Garibyan, L., and N. Avashia. 2013. “Research Techniques Made Simple: Polymerase Chain Reaction (PCR).” *The Journal of Investigative Dermatology*. https://www.ncbi.nlm.nih.gov/pmc/articles/pmc4102308/.

75. Gest, Howard. 2004. “The Discovery of Microorganisms by Robert Hooke and Antoni Van Leeuwenhoek, Fellows of the Royal Society.” *Notes and Records of the Royal Society of London* 58 (2): 187–201. https://doi.org/10.1098/rsnr.2004.0055.

76. Gray, M. W., D. Sankoff, and R. J. Cedergren. 1984. “On the Evolutionary Descent of Organisms and Organelles: A Global Phylogeny Based on a Highly Conserved Structural Core in Small Subunit Ribosomal RNA.” *Nucleic Acids Research* 12 (14): 5837–52. https://doi.org/10.1093/nar/12.14.5837.

77. Grüning, Björn, The Bioconda Team, Ryan Dale, Andreas Sjödin, Brad A. Chapman, Jillian Rowe, Christopher H. Tomkins-Tinch, Renan Valieris, and Johannes Köster. 2018. “Bioconda: Sustainable and Comprehensive Software Distribution for the Life Sciences.” *Nature Methods*. https://doi.org/10.1038/s41592-018-0046-7.

78. Gweon, H. Soon, Liam P. Shaw, Jeremy Swann, Nicola De Maio, Manal AbuOun, Rene Niehus, Alasdair T. M. Hubbard, et al. 2019. “The Impact of Sequencing Depth on the Inferred Taxonomic Composition and AMR Gene Content of Metagenomic Samples.” *Environmental Microbiome* 14 (1): 7. https://doi.org/10.1186/s40793-019-0347-1.

79. Haft, Daniel H., Michael DiCuccio, Azat Badretdin, Vyacheslav Brover, Vyacheslav Chetvernin, Kathleen O’Neill, Wenjun Li, et al. 2018. “RefSeq: An Update on Prokaryotic Genome Annotation and Curation.” *Nucleic Acids Research* 46 (D1): D851–60. https://doi.org/10.1093/nar/gkx1068.

80. Handelsman, Jo. 2004. “Metagenomics: Application of Genomics to Uncultured Microorganisms.” *MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS* 68 (4): 669–85. https://doi.org/10.1128/MBR.68.4.669-685.2004.

81. Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. 1998. “Molecular Biological Access to the Chemistry of Unknown Soil Microbes: A New Frontier for Natural Products.” *Chemistry & Biology* 5 (10): R245–49. https://www.ncbi.nlm.nih.gov/pubmed/9818143.

82. Henry, Christopher S., Matthew DeJongh, Aaron A. Best, Paul M. Frybarger, Ben Linsay, and Rick L. Stevens. 2010. “High-Throughput Generation, Optimization and Analysis of Genome-Scale Metabolic Models.” *Nature Biotechnology* 28 (9): 977–82. https://doi.org/10.1038/nbt.1672.

83. Herrera Paredes, Sur, Tianxiang Gao, Theresa F. Law, Omri M. Finkel, Tatiana Mucyn, Paulo José Pereira Lima Teixeira, Isaí Salas González, et al. 2018. “Design of Synthetic Bacterial Communities for Predictable Plant Phenotypes.” *PLoS Biology* 16 (2):

e2003962. https://doi.org/10.1371/journal.pbio.2003962.

84. Highlander, Sarah. 2013. “Mock Community Analysis.” *Encyclopedia of Metagenomics. New York, Springer New York*, 1–7.

85. Highlander, Sarah. 2015. “Mock Community Analysis.” In *Encyclopedia of Metagenomics: Genes, Genomes and Metagenomes: Basics, Methods, Databases and Tools*, edited by Karen E. Nelson, 497–503. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4899-7478-5_54.

86. Hubbell, Stephen P. 2001. *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Princeton University Press. https://play.google.com/store/books/details?id=EIQpFBu84NoC.

87. Huerta-Cepas, Jaime, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, et al. 2016. “eggNOG 4.5: A Hierarchical Orthology Framework with Improved Functional Annotations for Eukaryotic, Prokaryotic and Viral Sequences.” *Nucleic Acids Research* 44 (D1): D286–93. https://doi.org/10.1093/nar/gkv1248.

88. Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K. Forslund, Helen Cook, Daniel R. Mende, et al. 2019. “eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses.” *Nucleic Acids Research* 47 (D1): D309–14. https://doi.org/10.1093/nar/gky1085.

89. Hugenholtz, Philip, Brett M. Goebel, and Norman R. Pace. 1998. “Impact of Culture-Independent Studies on the Emerging Phylogenetic View of Bacterial Diversity.” *Journal of Bacteriology*. https://doi.org/10.1128/jb.180.24.6793-6793.1998.

90. Human Microbiome Project Consortium. 2012. “Structure, Function and Diversity of the Healthy Human Microbiome.” *Nature* 486 (7402): 207–14. https://doi.org/10.1038/nature11234.

91. Hyatt, Doug, Gwo-Liang Chen, Philip F. Locascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics* 11 (March): 119. https://doi.org/10.1186/1471-2105-11-119.

92. Iljazovic, Aida, Urmi Roy, Eric J. C. Gálvez, Till R. Lesker, Bei Zhao, Achim Gronow, Lena Amend, et al. 2021. “Perturbation of the Gut Microbiome by Prevotella Spp. Enhances Host Susceptibility to Mucosal Inflammation.” *Mucosal Immunology* 14 (1): 113–24. https://doi.org/10.1038/s41385-020-0296-4.

93. Jaccard, P. 1901. “Étude Comparative de La Distribution Florale Dans Une Portion Des Alpes et Des Jura.” *Bulletin de La SociÃ©tÃ© Vaudoise Des Sciences Naturelles*. https://ci.nii.ac.jp/naid/10019961020/.

94. Jensen, Lars Juhl, Philippe Julien, Michael Kuhn, Christian von Mering, Jean Muller, Tobias Doerks, and Peer Bork. 2008. “eggNOG: Automated Construction and Annotation of Orthologous Groups of Genes.” *Nucleic Acids Research* 36 (Database issue): D250–54. https://doi.org/10.1093/nar/gkm796.

95. Jeske, Lisa, Sandra Placzek, Ida Schomburg, Antje Chang, and Dietmar Schomburg. 2019. “BRENDA in 2019: A European ELIXIR Core Data Resource.” *Nucleic Acids Research* 47 (D1): D542–49. https://doi.org/10.1093/nar/gky1048.

96. Johns, Nathan I., Tomasz Blazejewski, Antonio Lc Gomes, and Harris H. Wang. 2016. “Principles for Designing Synthetic Microbial Communities.” *Current Opinion in Microbiology* 31 (June): 146–53. https://doi.org/10.1016/j.mib.2016.03.010.

97. Johnston, Eric R., Luis M. Rodriguez-R, Chengwei Luo, Mengting M. Yuan, Liyou Wu, Zhili He, Edward A. G. Schuur, et al. 2016. “Metagenomics Reveals Pervasive Bacterial Populations and Reduced Community Diversity across the Alaska Tundra Ecosystem.” *Frontiers in Microbiology* 7 (April): 579. https://doi.org/10.3389/fmicb.2016.00579.

98. Kanehisa, M. 1997. “A Database for Post-Genome Analysis.” *Trends in Genetics: TIG* 13 (9): 375–76. https://doi.org/10.1016/s0168-9525(97)01223-7.

99. Kanehisa, M., and S. Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28 (1): 27–30. https://doi.org/10.1093/nar/28.1.27.

100. Kanehisa, Minoru, Susumu Goto, Masahiro Hattori, Kiyoko F. Aoki-Kinoshita, Masumi Itoh, Shuichi Kawashima, Toshiaki Katayama, Michihiro Araki, and Mika Hirakawa. 2006. “From Genomics to Chemical Genomics: New Developments in KEGG.” *Nucleic Acids Research* 34 (Database issue): D354–57. https://doi.org/10.1093/nar/gkj102.

101. Kang, Dongwan D., Jeff Froula, Rob Egan, and Zhong Wang. 2015. “MetaBAT, an Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities.” *PeerJ* 3 (August): e1165. https://doi.org/10.7717/peerj.1165.

102. Kanwal, Sadia, Thomson Patrick Joseph, Shams Aliya, Siyuan Song, Muhammad Zubair Saleem, Muhammad Azhar Nisar, Yue Wang, Abdo Meyiah, Yufang Ma, and Yi Xin. 2020. “Attenuation of DSS Induced Colitis by Dictyophora Indusiata Polysaccharide (DIP) via Modulation of Gut Microbiota and Inflammatory Related Signaling Pathways.” *Journal of Functional Foods*. https://doi.org/10.1016/j.jff.2019.103641.

103. Karkaria, Behzad D., Alex J. H. Fedorec, and Chris P. Barnes. 2021. “Automated Design of Synthetic Microbial Communities.” *Nature Communications* 12 (1): 672. https://doi.org/10.1038/s41467-020-20756-2.

104. Karp, Peter D., Suzanne M. Paley, Markus Krummenacker, Mario Latendresse, Joseph M. Dale, Thomas J. Lee, Pallavi Kaipa, et al. 2010. “Pathway Tools Version 13.0: Integrated Software for Pathway/genome Informatics and Systems Biology.”

Briefings in Bioinformatics 11 (1): 40–79. https://doi.org/10.1093/bib/bbp043.

105. King, Zachary A., Justin Lu, Andreas Dräger, Philip Miller, Stephen Federowicz, Joshua A. Lerman, Ali Ebrahim, Bernhard O. Palsson, and Nathan E. Lewis. 2016. “BiGG Models: A Platform for Integrating, Standardizing and Sharing Genome-Scale Models.” *Nucleic Acids Research* 44 (D1): D515–22. https://doi.org/10.1093/nar/gkv1049.

106. Klindworth, Anna, Elmar Pruesse, Timmy Schweer, Jörg Peplies, Christian Quast, Matthias Horn, and Frank Oliver Glöckner. 2013. “Evaluation of General 16S Ribosomal RNA Gene PCR Primers for Classical and next-Generation Sequencing-Based Diversity Studies.” *Nucleic Acids Research* 41 (1): e1. https://doi.org/10.1093/nar/gks808.

107. Kristensen, David M., Yuri I. Wolf, and Eugene V. Koonin. 2017. “ATGC Database and ATGC-COGs: An Updated Resource for Micro- and Macro-Evolutionary Studies of Prokaryotic Genomes and Protein Family Annotation.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw934.

108. Kull, Kalevi. 2010. “Ecosystems Are Made of Semiosic Bonds: Consortia, Umwelten, Biophony and Ecological Codes.” *Biosemiotics* 3 (3): 347–57. https://doi.org/10.1007/s12304-010-9081-1.

109. Kultima, Jens Roat, Luis Pedro Coelho, Kristoffer Forslund, Jaime Huerta-Cepas, Simone S. Li, Marja Driessen, Anita Yvonne Voigt, Georg Zeller, Shinichi Sunagawa, and Peer Bork. 2016. “MOCAT2: A Metagenomic Assembly, Annotation and Profiling Framework.” *Bioinformatics* 32 (16): 2520–23. https://doi.org/10.1093/bioinformatics/btw183.

110. Kyrpides, N. C. 1999. “Genomes OnLine Database (GOLD 1.0): A Monitor of Complete and Ongoing Genome Projects World-Wide.” *Bioinformatics* 15 (9): 773–74. https://doi.org/10.1093/bioinformatics/15.9.773.

111. Lagkouvardos, Ilias, Sandra Fischer, Neeraj Kumar, and Thomas Clavel. 2017. “Rhea: A Transparent and Modular R Pipeline for Microbial Profiling Based on 16S rRNA Gene Amplicons.” *PeerJ*. https://doi.org/10.7717/peerj.2836.

112. Lagkouvardos, Ilias, Divya Joseph, Martin Kapfhammer, Sabahattin Giritli, Matthias Horn, Dirk Haller, and Thomas Clavel. 2016. “IMNGS: A Comprehensive Open Resource of Processed 16S rRNA Microbial Profiles for Ecology and Diversity Studies.” *Scientific Reports* 6 (September): 33721. https://doi.org/10.1038/srep33721.

113. Lagkouvardos, Ilias, Rüdiger Pukall, Birte Abt, Bärbel U. Foesel, Jan P. Meier-Kolthoff, Neeraj Kumar, Anne Bresciani, et al. 2016. “The Mouse Intestinal Bacterial Collection (miBC) Provides Host-Specific Insight into Cultured Diversity and Functional Potential of the Gut Microbiota.” *Nature Microbiology* 1 (10): 16131. https://doi.org/10.1038/nmicrobiol.2016.131.

114. Land, Miriam, Loren Hauser, Se-Ran Jun, Intawat Nookaew, Michael R. Leuze, Tae-Hyuk Ahn, Tatiana Karpinets, et al. 2015. “Insights from 20 Years of Bacterial Genome Sequencing.” *Functional & Integrative Genomics* 15 (2): 141–61. https://doi.org/10.1007/s10142-015-0433-4.

115. Larsbrink, Johan, Theresa E. Rogers, Glyn R. Hemsworth, Lauren S. McKee, Alexandra S. Tauzin, Oliver Spadiut, Stefan Klinter, et al. 2014. “A Discrete Genetic Locus Confers Xyloglucan Metabolism in Select Human Gut Bacteroidetes.” *Nature* 506 (7489): 498–502. https://doi.org/10.1038/nature12907.

116. Lau, S. K. P., P. C. Y. Woo, G. K. S. Woo, A. M. Y. Fung, A. H. Y. Ngan, Y. Song, C. Liu, P. Summanen, S. M. Finegold, and K. Yuen. 2006. “Bacteraemia Caused by Anaerotruncus Colihominis and Emended Description of the Species.” *Journal of Clinical Pathology* 59 (7): 748–52. https://doi.org/10.1136/jcp.2005.031773.

117. Lee Lerner, K., and Brenda Wilmoth Lerner. 2003. *World of Microbiology and Immunology: A-L*. Gale. https://play.google.com/store/books/details?id=CtluVhm06DwC.

118. Lenormant, Charles. 1861. “Mémoire Sur Les Représentations Qui Avaient Lieu Dans Les Mystères d’Éleusis.” *Mémoires de l’Institut National de France*. https://doi.org/10.3406/minf.1861.1426.

119. Letunic, Ivica, Tobias Doerks, and Peer Bork. 2009. “SMART 6: Recent Updates and New Developments.” *Nucleic Acids Research* 37 (Database issue): D229–32. https://doi.org/10.1093/nar/gkn808.

120. Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. “MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph.” *Bioinformatics* 31 (10): 1674–76. https://doi.org/10.1093/bioinformatics/btv033.

121. Liebl, Wolfgang. 2011. “Metagenomics.” *Encyclopedia of Geobiology*. https://doi.org/10.1007/978-1-4020-9212-1_133.

122. Li, Junhua, Huijue Jia, Xianghang Cai, Huanzi Zhong, Qiang Feng, Shinichi Sunagawa, Manimozhiyan Arumugam, et al. 2014. “An Integrated Catalog of Reference Genes in the Human Gut Microbiome.” *Nature Biotechnology* 32 (8): 834–41. https://doi.org/10.1038/nbt.2942.

123. Lloyd, Karen G., Andrew D. Steen, Joshua Ladau, Junqi Yin, and Lonnie Crosby. 2018. “Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes.” *mSystems* 3 (5). https://doi.org/10.1128/mSystems.00055-18.

124. Lloyd-Price, Jason, Cesar Arze, Ashwin N. Ananthakrishnan, Melanie Schirmer, Julian Avila-Pacheco, Tiffany W. Poon, Elizabeth Andrews, et al. 2019. “Multi-Omics of the Gut Microbial Ecosystem in Inflammatory Bowel Diseases.” *Nature* 569 (7758):

655–62. https://doi.org/10.1038/s41586-019-1237-9.

125. Luckey, T. D. 1972. “Introduction to Intestinal Microecology.” *The American Journal of Clinical Nutrition* 25 (12): 1292–94. https://doi.org/10.1093/ajcn/25.12.1292.

126. Lukashin, A. 1998. “GeneMark.hmm: New Solutions for Gene Finding.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/26.4.1107.

127. Lykidis, Athanasios, Konstantinos Mavromatis, Natalia Ivanova, Iain Anderson, Miriam Land, Genevieve DiBartolo, Michele Martinez, et al. 2007. “Genome Sequence and Analysis of the Soil Cellulolytic Actinomycete Thermobifida Fusca YX.” *Journal of Bacteriology* 189 (6): 2477–86. https://doi.org/10.1128/JB.01899-06.

128. Macfarlane, S., G. T. Macfarlane, and J. H. Cummings. 2006. “Review Article: Prebiotics in the Gastrointestinal Tract.” *Alimentary Pharmacology & Therapeutics* 24 (5): 701–14. https://doi.org/10.1111/j.1365-2036.2006.03042.x.

129. Mahmoudabadi, Gita, and Rob Phillips. 2018. “A Comprehensive and Quantitative Exploration of Thousands of Viral Genomes.” *eLife* 7 (April). https://doi.org/10.7554/eLife.31955.

130. Makarova, Kira, Yuri Wolf, and Eugene Koonin. 2015. “Archaeal Clusters of Orthologous Genes (arCOGs): An Update and Application for Analysis of Shared Features between Thermococcales, Methanococcales, and Methanobacteriales.” *Life*. https://doi.org/10.3390/life5010818.

131. Maloy, Kevin J., and Fiona Powrie. 2011. “Intestinal Homeostasis and Its Breakdown in Inflammatory Bowel Disease.” *Nature* 474 (7351): 298–306. https://doi.org/10.1038/nature10208.

132. Marchesi, Julian R., and Jacques Ravel. 2015. “The Vocabulary of Microbiome Research: A Proposal.” *Microbiome* 3 (July): 31. https://doi.org/10.1186/s40168-015-0094-5.

133. Marchler-Bauer, Aron, Chanjuan Zheng, Farideh Chitsaz, Myra K. Derbyshire, Lewis Y. Geer, Renata C. Geer, Noreen R. Gonzales, et al. 2013. “CDD: Conserved Domains and Protein Three-Dimensional Structure.” *Nucleic Acids Research* 41 (Database issue): D348–52. https://doi.org/10.1093/nar/gks1243.

134. Mayer, Emeran A., Rob Knight, Sarkis K. Mazmanian, John F. Cryan, and Kirsten Tillisch. 2014. “Gut Microbes and the Brain: Paradigm Shift in Neuroscience.” *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 34 (46): 15490–96. https://doi.org/10.1523/JNEUROSCI.3299-14.2014.

135. McCarty, Nicholas S., and Rodrigo Ledesma-Amaro. 2019. “Synthetic Biology Tools to Engineer Microbial Communities for Biotechnology.” *Trends in Biotechnology* 37 (2): 181–97. https://doi.org/10.1016/j.tibtech.2018.11.002.

136. McGill, Brian J., Rampal S. Etienne, John S. Gray, David Alonso, Marti J. Anderson, Habtamu Kassa Benecha, Maria Dornelas, et al. 2007. “Species Abundance Distributions: Moving beyond Single Prediction Theories to Integration within an Ecological Framework.” *Ecology Letters*. https://doi.org/10.1111/j.1461-0248.2007.01094.x.

137. McNeil, L. K., C. Reich, R. K. Aziz, D. Bartels, M. Cohoon, T. Disz, R. A. Edwards, et al. 2007. “The National Microbial Pathogen Database Resource (NMPDR): A Genomics Platform Based on Subsystem Annotation.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkl947.

138. Mead, Al. 1992. “Review of the Development of Multidimensional Scaling Methods.” *Journal of the Royal Statistical Society: Series D (The Statistician)* 41 (1): 27–39. https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2348634.

139. Meier, Matthew J., E. Suzanne Paterson, and Iain B. Lambert. 2016. “Use of Substrate-Induced Gene Expression in Metagenomic Analysis of an Aromatic Hydrocarbon-Contaminated Soil.” *Applied and Environmental Microbiology* 82 (3): 897–909. https://doi.org/10.1128/AEM.03306-15.

140. Meurens, F., A. Summerfield, H. Nauwynck, and L. Saif. 2012. “The Pig: A Model for Human Infectious Diseases.” *Trends in*. https://www.sciencedirect.com/science/article/pii/S0966842X11001958.

141. Meyer, F., D. Paarmann, M. D’Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, et al. 2008. “The Metagenomics RAST Server – a Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes.” *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-9-386.

142. Mitchell, Alex L., Maxim Scheremetjew, Hubert Denise, Simon Potter, Aleksandra Tarkowska, Matloob Qureshi, Gustavo A. Salazar, et al. 2018. “EBI Metagenomics in 2017: Enriching the Analysis of Microbial Communities, from Sequence Reads to Assemblies.” *Nucleic Acids Research* 46 (D1): D726–35. https://doi.org/10.1093/nar/gkx967.

143. Moayyedi, Paul, Yuhong Yuan, Harith Baharith, and Alexander C. Ford. 2017. “Faecal Microbiota Transplantation for Clostridium Difficile-Associated Diarrhoea: A Systematic Review of Randomised Controlled Trials.” *The Medical Journal of Australia* 207 (4): 166–72. https://onlinelibrary.wiley.com/doi/abs/10.5694/mja17.00295.

144. Morowitz, Michael J., Erica M. Carlisle, and John C. Alverdy. 2011. “Contributions of Intestinal Bacteria to Nutrition and Metabolism in the Critically Ill.” *The Surgical Clinics of North America* 91 (4): 771–85, viii. https://doi.org/10.1016/j.suc.2011.05.001.

145. Mukherjee, Supratim, Dimitri Stamatis, Jon Bertsch, Galina Ovchinnikova, Jagadish Chandrabose Sundaramurthi, Janey Lee, Mahathi Kandimalla, I-Min A. Chen, Nikos C. Kyrpides, and T. B. K. Reddy. 2021. “Genomes OnLine Database (GOLD) v.8: Overview

and Updates." *Nucleic Acids Research* 49 (D1): D723–33. https://doi.org/10.1093/nar/gkaa983.

146. Mukhopadhya, Indrani, Richard Hansen, Emad M. El-Omar, and Georgina L. Hold. 2012. "IBD—what Role Do Proteobacteria Play?" *Nature Reviews. Gastroenterology & Hepatology* 9 (February): 219. https://doi.org/10.1038/nrgastro.2012.14.

147. Muller, J., D. Szklarczyk, P. Julien, I. Letunic, A. Roth, M. Kuhn, S. Powell, et al. 2010. "eggNOG v2.0: Extending the Evolutionary Genealogy of Genes with Enhanced Non-Supervised Orthologous Groups, Species and Functional Annotations." *Nucleic Acids Research* 38 (Database issue): D190–95. https://doi.org/10.1093/nar/gkp951.

148. Mullis, Kary B., Henry A. Erlich, Norman Arnheim, Glenn T. Horn, Randall K. Saiki, and Stephen J. Scharf. 1987. Process for amplifying, detecting, and/or-cloning nucleic acid sequences. USPTO 4683195. *US Patent*, filed February 7, 1986, and issued July 28, 1987. https://patentimages.storage.googleapis.com/ec/14/bf/0a414f77b2d203/US4683195.pdf.

149. Mullis, K. B. 1990. "The Unusual Origin of the Polymerase Chain Reaction." *Scientific American* 262 (4): 56–61, 64–65. https://doi.org/10.1038/scientificamerican0490-56.

150. Nemoto, Hideyuki, Keiko Kataoka, Hideki Ishikawa, Kazue Ikata, Hideki Arimochi, Teruaki Iwasaki, Yoshinari Ohnishi, Tomomi Kuwahara, and Koji Yasutomo. 2012. "Reduced Diversity and Imbalance of Fecal Microbiota in Patients with Ulcerative Colitis." *Digestive Diseases and Sciences* 57 (11): 2955–64. https://doi.org/10.1007/s10620-012-2236-y.

151. Niu, Ben, Joseph Nathaniel Paulson, Xiaoqi Zheng, and Roberto Kolter. 2017. "Simplified and Representative Bacterial Community of Maize Roots." *Proceedings of the National Academy of Sciences of the United States of America* 114 (12): E2450–59. https://doi.org/10.1073/pnas.1616148114.

152. Nood, Els van, Anne Vrieze, Max Nieuwdorp, Susana Fuentes, Erwin G. Zoetendal, Willem M. de Vos, Caroline E. Visser, et al. 2013. "Duodenal Infusion of Donor Feces for Recurrent Clostridium Difficile." *The New England Journal of Medicine* 368 (5): 407–15. https://doi.org/10.1056/NEJMoa1205037.

153. Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. "metaSPAdes: A New Versatile Metagenomic Assembler." *Genome Research*. https://doi.org/10.1101/gr.213959.116.

154. Nyrén, P., B. Pettersson, and M. Uhlén. 1993. "Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay." *Analytical Biochemistry* 208 (1): 171–75. https://doi.org/10.1006/abio.1993.1024.

155. Pan, Fengwei, Liying Zhang, Min Li, Yingxin Hu, Benhua Zeng, Huijuan Yuan, Liping Zhao, and Chenhong Zhang. 2018. “Predominant Gut Lactobacillus Murinus Strain Mediates Anti-Inflammaging Effects in Calorie-Restricted Mice.” *Microbiome* 6 (1): 54. https://doi.org/10.1186/s40168-018-0440-5.

156. Park, John Chulhoon, and Sin-Hyeog Im. 2020. “Of Men in Mice: The Development and Application of a Humanized Gnotobiotic Mouse Model for Microbiome Therapeutics.” *Experimental & Molecular Medicine* 52 (9): 1383–96. https://doi.org/10.1038/s12276-020-0473-2.

157. Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2018. “A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life.” *Nature Biotechnology* 36 (10): 996–1004. https://doi.org/10.1038/nbt.4229.

158. Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. “CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes.” *Genome Research* 25 (7): 1043–55. https://doi.org/10.1101/gr.186072.114.

159. Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 2019. “Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.” *Cell* 176 (3): 649–62.e20. https://doi.org/10.1016/j.cell.2019.01.001.

160. Pasteur, Louis, Chamberland, and Roux. 2002. “Summary Report of the Experiments Conducted at Pouilly-Le-Fort, near Melun, on the Anthrax Vaccination, 1881.” *The Yale Journal of Biology and Medicine* 75 (1): 59–62. https://www.ncbi.nlm.nih.gov/pubmed/12074483.

161. Peaper, D. R., J. Bertin, S. C. Eisenbarth, and J. I. Gordon. 2011. “NLRP6 Inflammasome Is a Regulator of Colonic Microbial Ecology and Risk for Colitis.” *Cell*.

162. Peisl, B. Y. Loulou, Emma L. Schymanski, and Paul Wilmes. 2017. “Dark Matter in Host-Microbiome Metabolomics: Tackling the unknowns–A Review.” *Analytica Chimica Acta*, December. https://doi.org/10.1016/j.aca.2017.12.034.

163. Pitkänen, Esa, Paula Jouhten, Jian Hou, Muhammad Fahad Syed, Peter Blomberg, Jana Kludas, Merja Oja, et al. 2014. “Comparative Genome-Scale Reconstruction of Gapless Metabolic Networks for Present and Ancestral Species.” *PLoS Computational Biology* 10 (2): e1003465. https://doi.org/10.1371/journal.pcbi.1003465.

164. Plaza Oñate, Florian, Emmanuelle Le Chatelier, Mathieu Almeida, Alessandra C. L. Cervino, Franck Gauthier, Frédéric Magoulès, S. Dusko Ehrlich, and Matthieu Pichaud. 2019. “MSPminer: Abundance-Based Reconstitution of Microbial Pan-Genomes from Shotgun Metagenomic Data.” *Bioinformatics* 35 (9): 1544–52.

https://doi.org/10.1093/bioinformatics/bty830.

165. Powell, Sean, Kristoffer Forslund, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Jaime Huerta-Cepas, Toni Gabaldón, et al. 2014. “eggNOG v4.0: Nested Orthology Inference across 3686 Organisms.” *Nucleic Acids Research* 42 (Database issue): D231–39. https://doi.org/10.1093/nar/gkt1253.

166. Powell, Sean, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Michael Kuhn, Jean Muller, Roland Arnold, et al. 2012. “eggNOG v3.0: Orthologous Groups Covering 1133 Organisms at 41 Different Taxonomic Ranges.” *Nucleic Acids Research* 40 (Database issue): D284–89. https://doi.org/10.1093/nar/gkr1060.

167. Pruesse, Elmar, Christian Quast, Katrin Knittel, Bernhard M. Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. 2007. “SILVA: A Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB.” *Nucleic Acids Research* 35 (21): 7188–96. https://doi.org/10.1093/nar/gkm864.

168. Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. 2005. “NCBI Reference Sequence (RefSeq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins.” *Nucleic Acids Research* 33 (Database issue): D501–4. https://doi.org/10.1093/nar/gki025.

169. Puentes-Téllez, Pilar Eliana, and Joana Falcao Salles. 2018. “Construction of Effective Minimal Active Microbial Consortia for Lignocellulose Degradation.” *Microbial Ecology* 76 (2): 419–29. https://doi.org/10.1007/s00248-017-1141-5.

170. Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, and Nicola Segata. 2017. “Shotgun Metagenomics, from Sampling to Analysis.” *Nature Biotechnology* 35 (9): 833–44. https://doi.org/10.1038/nbt.3935.

171. Ramírez-Pérez, Oscar, Vania Cruz-Ramón, Paulina Chinchilla-López, and Nahum Méndez-Sánchez. 2017. “The Role of the Gut Microbiota in Bile Acid Metabolism.” *Annals of Hepatology* 16 Suppl 1 (November): S21–26. https://doi.org/10.5604/01.3001.0010.5672.

172. Reinke, Johannes. 1872. “Ueber Die Anatomischen Verhältnisse Einiger Arten von Gunnera L.” *Nachrichten von Der Königl. Gesellschaft Der Wissenschaften Und Der Georg-Augusts-Universität Zu Göttingen* 1872: 100–108. https://eudml.org/doc/179542.

173. Rosenthal, Nadia, and Steve Brown. 2007. “The Mouse Ascending: Perspectives for Human-Disease Models.” *Nature Cell Biology* 9 (9): 993–99. https://doi.org/10.1038/ncb437.

174. Roy, Urmi, Eric J. C. Gálvez, Aida Iljazovic, Till Robin Lesker, Adrian J. Błażejewski, Marina C. Pils, Ulrike Heise, Samuel Huber, Richard A. Flavell, and Till Strowig. 2017. “Distinct Microbial Communities Trigger Colitis Development upon Intestinal Barrier

Damage via Innate or Adaptive Immune Cells." *Cell Reports* 21 (4): 994–1008. https://doi.org/10.1016/j.celrep.2017.09.097.

175. Sammut, Stephen John, Robert D. Finn, and Alex Bateman. 2008. "Pfam 10 Years on: 10,000 Families and Still Growing." *Briefings in Bioinformatics* 9 (3): 210–19. https://doi.org/10.1093/bib/bbn010.

176. Sanger, F., and A. R. Coulson. 1989. "A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase." *Molecular Biology*. https://doi.org/10.1016/b978-0-12-131200-8.50040-x.

177. Sanger, F., S. Nicklen, and A. R. Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74 (12): 5463–67. https://doi.org/10.1073/pnas.74.12.5463.

178. Sarkar, Amar, Soili M. Lehto, Siobhán Harty, Timothy G. Dinan, John F. Cryan, and Philip W. J. Burnet. 2016. "Psychobiotics and the Manipulation of Bacteria–Gut–Brain Signals." *Trends in Neurosciences* 39 (11): 763–81. https://doi.org/10.1016/j.tins.2016.09.002.

179. Schink, Bernhard, and Alfons J. M. Stams. 2013. "Syntrophism Among Prokaryotes." *The Prokaryotes*. https://doi.org/10.1007/978-3-642-30123-0_59.

180. Schloss, Patrick D., Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, et al. 2009. "Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities." *Applied and Environmental Microbiology* 75 (23): 7537–41. https://doi.org/10.1128/AEM.01541-09.

181. Schluenzen, F., A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, et al. 2000. "Structure of Functionally Activated Small Ribosomal Subunit at 3.3 Angstroms Resolution." *Cell* 102 (5): 615–23. https://doi.org/10.1016/s0092-8674(00)00084-2.

182. Schou Larsen, Thomas, and Anders Krogh. 2003. "EasyGene--a Prokaryotic Gene Finder That Ranks ORFs by Statistical Significance." *BMC Bioinformatics* 4 (1): 1–15.

183. Scordis, P., D. R. Flower, and T. K. Attwood. 1999. "FingerPRINTScan: Intelligent Searching of the PRINTS Motif Database." *Bioinformatics* 15 (10): 799–806. https://doi.org/10.1093/bioinformatics/15.10.799.

184. Seemann, Torsten. 2014. "Prokka: Rapid Prokaryotic Genome Annotation." *Bioinformatics* 30 (14): 2068–69. https://doi.org/10.1093/bioinformatics/btu153.

185. Sender, Ron, Shai Fuchs, and Ron Milo. 2016. "Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans." *Cell* 164 (3): 337–40. https://doi.org/10.1016/j.cell.2016.01.013.

186. Sen, Partho, and Matej Orešič. 2019. “Metabolic Modeling of Human Gut Microbiota on a Genome Scale: An Overview.” *Metabolites* 9 (2). https://doi.org/10.3390/metabo9020022.

187. Shannon, C. E. 1948. “A Mathematical Theory of Communication.” *The Bell System Technical Journal* 27 (3): 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

188. Sharon, Donald, Hagen Tilgner, Fabian Grubert, and Michael Snyder. 2013. “A Single-Molecule Long-Read Survey of the Human Transcriptome.” *Nature Biotechnology* 31 (11): 1009–14. https://doi.org/10.1038/nbt.2705.

189. Sha, Sumei, Bin Xu, Xin Wang, Yongguo Zhang, Honghong Wang, Xiangyun Kong, Hongwu Zhu, and Kaichun Wu. 2013. “The Biodiversity and Composition of the Dominant Fecal Microbiota in Patients with Inflammatory Bowel Disease.” *Diagnostic Microbiology and Infectious Disease* 75 (3): 245–51. https://doi.org/10.1016/j.diagmicrobio.2012.11.022.

190. Shen, Wei, Shuai Le, Yan Li, and Fuquan Hu. 2016. “SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation.” *PloS One* 11 (10): e0163962. https://doi.org/10.1371/journal.pone.0163962.

191. Sigrist, Christian J. A., Lorenzo Cerutti, Edouard de Castro, Petra S. Langendijk-Genevaux, Virginie Bulliard, Amos Bairoch, and Nicolas Hulo. 2010. “PROSITE, a Protein Domain Database for Functional Characterization and Annotation.” *Nucleic Acids Research* 38 (Database issue): D161–66. https://doi.org/10.1093/nar/gkp885.

192. Singer, Esther, Bill Andreopoulos, Robert M. Bowers, Janey Lee, Shweta Deshpande, Jennifer Chiniquy, Doina Ciobanu, et al. 2016. “Next Generation Sequencing Data of a Defined Microbial Mock Community.” *Scientific Data* 3 (September): 160081. https://doi.org/10.1038/sdata.2016.81.

193. Smith, Kendall A. 2012. “Louis Pasteur, the Father of Immunology?” *Frontiers in Immunology*. https://doi.org/10.3389/fimmu.2012.00068.

194. Song, Hao, Ming-Zhu Ding, Xiao-Qiang Jia, Qian Ma, and Ying-Jin Yuan. 2014. “Synthetic Microbial Consortia: From Systematic Analysis to Construction and Applications.” *Chemical Society Reviews* 43 (20): 6954–81. https://doi.org/10.1039/c4cs00114a.

195. Sonnhammer, E. L., S. R. Eddy, E. Birney, A. Bateman, and R. Durbin. 1998. “Pfam: Multiple Sequence Alignments and HMM-Profiles of Protein Domains.” *Nucleic Acids Research* 26 (1): 320–22. https://doi.org/10.1093/nar/26.1.320.

196. Sonnhammer, E. L., S. R. Eddy, and R. Durbin. 1997. “Pfam: A Comprehensive Database of Protein Domain Families Based on Seed Alignments.” *Proteins* 28 (3): 405–20. https://www.ncbi.nlm.nih.gov/pubmed/9223186.

197. Spellerberg, I. F., and P. J. Fedor. 2003. “A Tribute to Claude Shannon (1916–2001) and a Plea for More Rigorous Use of Species Richness, Species Diversity and the ‘Shannon–Wiener’Index.” *Global Ecology and Biogeography: A Journal of Macroecology*. https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1466-822X.2003.00015.x.

198. Stolyar, Sergey, Steve Van Dien, Kristina Linnea Hillesland, Nicolas Pinel, Thomas J. Lie, John A. Leigh, and David A. Stahl. 2007. “Metabolic Modeling of a Mutualistic Microbial Community.” *Molecular Systems Biology* 3 (March): 92. https://doi.org/10.1038/msb4100131.

199. Sudo, Nobuyuki, Yoichi Chida, Yuji Aiba, Junko Sonoda, Naomi Oyama, Xiao-Nian Yu, Chiharu Kubo, and Yasuhiro Koga. 2004. “Postnatal Microbial Colonization Programs the Hypothalamic-Pituitary-Adrenal System for Stress Response in Mice.” *The Journal of Physiology*. https://doi.org/10.1113/jphysiol.2004.063388.

200. Sun, Beili, Dongrui Zhou, Jing Tu, and Zuhong Lu. 2017. “Evaluation of the Bacterial Diversity in the Human Tongue Coating Based on Genus-Specific Primers for 16S rRNA Sequencing.” *BioMed Research International* 2017 (August): 8184160. https://doi.org/10.1155/2017/8184160.

201. Sun, Xue, Lei Song, Wenjuan Yang, Lili Zhang, Meng Liu, Xiaoshuang Li, Geng Tian, and Weiwei Wang. 2020. “Nanopore Sequencing and Its Clinical Applications.” *Methods in Molecular Biology* 2204: 13–32. https://doi.org/10.1007/978-1-0716-0904-0_2.

202. Tatusova, Tatiana, Michael DiCuccio, Azat Badretdin, Vyacheslav Chetvernin, Eric P. Nawrocki, Leonid Zaslavsky, Alexandre Lomsadze, Kim D. Pruitt, Mark Borodovsky, and James Ostell. 2016. “NCBI Prokaryotic Genome Annotation Pipeline.” *Nucleic Acids Research* 44 (14): 6614–24. https://doi.org/10.1093/nar/gkw569.

203. Tatusov, R. L., and N. D. Fedorova. 2003. “The COG Database: An Updated Version Includes Eukaryotes.” *Biomedical Chromatography: BMC*. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-4-41.

204. Tatusov, R. L., E. V. Koonin, and D. J. Lipman. 1997. “A Genomic Perspective on Protein Families- COG Main Paper.” *Science* 278 (5338): 631–37. https://www.ncbi.nlm.nih.gov/pubmed/9381173.

205. “The Integrative Human Microbiome Project: Dynamic Analysis of Microbiome-Host Omics Profiles during Periods of Human Health and Disease.” 2014. *Cell Host & Microbe* 16 (3): 276–89. https://doi.org/10.1016/j.chom.2014.08.014.

206. Thiele, Ines, and Bernhard Ø. Palsson. 2010. “A Protocol for Generating a High-Quality Genome-Scale Metabolic Reconstruction.” *Nature Protocols* 5 (1): 93–121. https://doi.org/10.1038/nprot.2009.203.

207. Tierney, Braden T., Zhen Yang, Jacob M. Luber, Marc Beaudin, Marsha C. Wibowo, Christina Baek, Eleanor Mehlenbacher, Chirag J. Patel, and Aleksandar D. Kostic. 2019. “The Landscape of Genetic Content in the Gut and Oral Human Microbiome.” *Cell Host & Microbe* 26 (2): 283–95.e8. https://doi.org/10.1016/j.chom.2019.07.008.

208. Tringe, Susannah Green, Christian von Mering, Arthur Kobayashi, Asaf A. Salamov, Kevin Chen, Hwai W. Chang, Mircea Podar, et al. 2005. “Comparative Metagenomics of Microbial Communities.” *Science* 308 (5721): 554–57. https://doi.org/10.1126/science.1107851.

209. Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. “MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling.” *Nature Methods*. https://doi.org/10.1038/nmeth.3589.

210. “Ueber Die Isolirte Färbung Der Schizomyceten in Schnitt- Und Trockenpräparaten von Dr. Gram, Kopenhagen. — Fortschritte Der Medicin 1884 No. 6. Ref. Dr. Becker.” 1884. *Deutsche Medizinische Wochenschrift* 10 (15): 234–35. https://doi.org/10.1055/s-0029-1209285.

211. “UniProt: A Hub for Protein Information.” 2015. *Nucleic Acids Research* 43 (D1): D204–12. https://doi.org/10.1093/nar/gku989.

212. Van Goethem, Marc W., Rian Pierneef, Oliver K. I. Bezuidt, Yves Van De Peer, Don A. Cowan, and Thulani P. Makhalanyane. 2018. “A Reservoir of ‘Historical’ Antibiotic Resistance Genes in Remote Pristine Antarctic Soils.” *Microbiome* 6 (1): 40. https://doi.org/10.1186/s40168-018-0424-5.

213. Wang, Baohong, Mingfei Yao, Longxian Lv, Zongxin Ling, and Lanjuan Li. n.d. “The Human Microbiota in Health and Disease | Elsevier Enhanced Reader.” Accessed August 18, 2019. https://doi.org/10.1016/J.ENG.2017.01.008.

214. Wang, Yan, and Lloyd H. Kasper. 2014. “The Role of Microbiome in Central Nervous System Disorders.” *Brain, Behavior, and Immunity* 38 (May): 1–12. https://doi.org/10.1016/j.bbi.2013.12.015.

215. Whittaker, R. H. 1960. “Vegetation of the Siskiyou Mountains, Oregon and California.” *Ecological Monographs* 30 (3): 279–338. https://doi.org/10.2307/1943563.

216. Whittaker, R. H. 1972. “EVOLUTION AND MEASUREMENT OF SPECIES DIVERSITY.” *Taxon* 21 (2-3): 213–51. https://doi.org/10.2307/1218190.

217. Woese, C. R., and G. E. Fox. 1977. “Phylogenetic Structure of the Prokaryotic Domain: The Primary Kingdoms.” *Proceedings of the National Academy of Sciences of the United States of America* 74 (11): 5088–90. https://doi.org/10.1073/pnas.74.11.5088.

218. Woese, C. R., O. Kandler, and M. L. Wheelis. 1990. “Towards a Natural System of

Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya." *Proceedings of the National Academy of Sciences of the United States of America* 87 (12): 4576–79. https://doi.org/10.1073/pnas.87.12.4576.

219. Wood, Derrick E., Jennifer Lu, and Ben Langmead. 2019. "Improved Metagenomic Analysis with Kraken 2." *Genome Biology* 20 (1): 257. https://doi.org/10.1186/s13059-019-1891-0.

220. Wood, Derrick E., and Steven L. Salzberg. 2014. "Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments." *Genome Biology* 15 (3): R46. https://doi.org/10.1186/gb-2014-15-3-r46.

221. Wooley, John C., Adam Godzik, and Iddo Friedberg. 2010. "A Primer on Metagenomics." *PLoS Computational Biology* 6 (2): e1000667. https://doi.org/10.1371/journal.pcbi.1000667.

222. Wylensek, David, Thomas C. A. Hitch, Thomas Riedel, Afrizal Afrizal, Neeraj Kumar, Esther Wortmann, Tianzhe Liu, et al. 2020. "A Collection of Bacterial Isolates from the Pig Intestine Reveals Functional and Taxonomic Diversity." *Nature Communications* 11 (1): 6389. https://doi.org/10.1038/s41467-020-19929-w.

223. Xiao, Liang, Jordi Estellé, Pia Kiilerich, Yuliaxis Ramayo-Caldas, Zhongkui Xia, Qiang Feng, Suisha Liang, et al. 2016. "A Reference Gene Catalogue of the Pig Gut Microbiome." *Nature Microbiology* 1 (September): 16161. https://doi.org/10.1038/nmicrobiol.2016.161.

224. Xiao, Liang, Qiang Feng, Suisha Liang, Si Brask Sonne, Zhongkui Xia, Xinmin Qiu, Xiaoping Li, et al. 2015. "A Catalog of the Mouse Gut Metagenome." *Nature Biotechnology* 33 (10): 1103–8. https://doi.org/10.1038/nbt.3353.

225. Ye, Simon H., Katherine J. Siddle, Daniel J. Park, and Pardis C. Sabeti. 2019. "Benchmarking Metagenomics Tools for Taxonomic Classification." *Cell* 178 (4): 779–94. https://doi.org/10.1016/j.cell.2019.07.010.

226. Yoon, Seok-Hwan, Sung-Min Ha, Soonjae Kwon, Jeongmin Lim, Yeseul Kim, Hyungseok Seo, and Jongsik Chun. 2017. "Introducing EzBioCloud: A Taxonomically United Database of 16S rRNA Gene Sequences and Whole-Genome Assemblies." *International Journal of Systematic and Evolutionary Microbiology* 67 (5): 1613–17. https://doi.org/10.1099/ijsem.0.001755.

227. Ze, Xiaolei, Sylvia H. Duncan, Petra Louis, and Harry J. Flint. 2012. "Ruminococcus Bromii Is a Keystone Species for the Degradation of Resistant Starch in the Human Colon." *The ISME Journal* 6 (8): 1535–43. https://doi.org/10.1038/ismej.2012.4.

228. Ze, Xiaolei, Fanny Le Mougen, Sylvia H. Duncan, Petra Louis, and Harry J. Flint. 2013. "Some Are More Equal than Others: The Role of 'Keystone' Species in the Degradation of Recalcitrant Substrates." *Gut Microbes* 4 (3): 236–40.

https://doi.org/10.4161/gmic.23998.

229. Zhalnina, Kateryna, Karsten Zengler, Dianne Newman, and Trent R. Northen. 2018. “Need for Laboratory Ecosystems To Unravel the Structures and Functions of Soil Microbial Communities Mediated by Chemistry.” *mBio* 9 (4). https://doi.org/10.1128/mBio.01175-18.

230. Zhang, Jingying, Yong-Xin Liu, Na Zhang, Bin Hu, Tao Jin, Haoran Xu, Yuan Qin, et al. 2019. “NRT1.1B Is Associated with Root Microbiota Composition and Nitrogen Use in Field-Grown Rice.” *Nature Biotechnology*. https://doi.org/10.1038/s41587-019-0104-4.

231. Zheng, P., B. Zeng, C. Zhou, M. Liu, Z. Fang, X. Xu, L. Zeng, et al. 2016. “Gut Microbiome Remodeling Induces Depressive-like Behaviors through a Pathway Mediated by the Host’s Metabolism.” *Molecular Psychiatry* 21 (6): 786–96. https://doi.org/10.1038/mp.2016.44.

232. Zhu, Huaiqiu, Gang-Qing Hu, Yi-Fan Yang, Jin Wang, and Zhen-Su She. 2007. “MED: A New Non-Supervised Gene Prediction Algorithm for Bacterial and Archaeal Genomes.” *BMC Bioinformatics* 8 (March): 97. https://doi.org/10.1186/1471-2105-8-97.

233. Zou, Yuanqiang, Wenbin Xue, Guangwen Luo, Ziqing Deng, Panpan Qin, Ruijin Guo, Haipeng Sun, et al. 2019. “1,520 Reference Genomes from Cultivated Human Gut Bacteria Enable Functional Microbiome Analyses.” *Nature Biotechnology* 37 (2): 179–85. https://doi.org/10.1038/s41587-018-0008-8.

11 Acknowledgement

First and foremost, I would like to express my gratitude to my supervisors Prof Dr. Dirk Haller and Prof. Dr. Thomas Clavel for their insightful feedback and guidance throughout this research work. I am thankful to Prof. Dirk Haller for providing me an opportunity as doctoral student at his group and allowed me to work in a fascinating world of microbiome research. I am grateful to have Prof. Thomas Clavel as my mentor who helped me to reshape my research into this dissertation with constructive feedback and support. I would like to give a special thanks to Dr. Ilias Lagvakurdos for his initial support to understand microbial ecology via computational methods.

I would like to acknowledge Prof. Dirk Haller's group and entire team at the Core Facility Microbiome (ZIEL) for their pleasant company and working environment throughout my stay there. I extend my special thanks to Brita Sturm (Haller lab) for helping me with all official work throughout my PhD. I am thankful to Sevana for helping me to get through administrative work. I would like to thank all members of the Clavel group in Aachen (Afrizal, David Wylensek, Alina Viehof, Theresa Streidl, Atscharah Panyot, Esther Wortmann, Susan Jennings, Ntana Kousetzi, Dr. Thomas Hitch and Dr. Nicole Treichel) for their supportive working environment and special thanks to Maike Schleibach and Soheila Razavi for her caring and supportive attitude.

I would like to acknowledge Prof. Paturu Kondaiah (IISc bangalore) and Prof. Dr. Wolfgang Enard (LMU Munich) to encourage and motivate me to choose my career path in science. I would like to thank the TUM graduate school team and special thanks to Dr. Daniela Röder for continuous support to understand and fulfill the official requirement of the PhD.

This journey would have not been so beautiful without many individuals and friends I met along the way. I would like to thank all my friends who helped, encouraged and were always there for me especially Sabrina. I am very grateful to my parents and a word of thanks will not be enough to complement their support and sacrifices for shaping my career. My heartfelt thanks to my wife Swati for her constant support, love and always being there as a best friend and wonderful companion. Thank you my little daughter Ananya for always cheering and inspiring me during the writing phase.

12 Publications

1. **Neeraj Kumar**, Thomas Hitch, Ilias Lagvakurdos, Dirk Haller, Thomas Clavel. MiMiC: A bioinformatic approach for generation of synthetic communities from metagenomes. *Microbial biotechnology.* 2021; (submitted: MICROBIO 2021-136-RA)
2. Wylensek D, Hitch TCA, Riedel T, Afrizal A, **Kumar N**, Wortmann E, et al. A collection of bacterial isolates from the pig intestine reveals functional and taxonomic diversity. *Nat Commun.* 2020;11: 6389. doi:10.1038/s41467-020-19929-w
3. Streidl T, **Kumar N**, Suarez LN, Rohn S, Clavel T. Extibacter. Bergey's Manual of Systematics of Archaea and Bacteria. *Wiley*. 2019; pp. 1–7. doi:10.1002/9781118960608.gbm01749
4. Lagkouvardos I, Fischer S, **Kumar N**, Clavel T. Rhea: a transparent and modular R pipeline for microbial profiling based on 16S rRNA gene amplicons. *PeerJ.* 2017;5: e2836. doi:10.7717/peerj.2836
5. Lagkouvardos I, Pukall R, Abt B, Foesel BU, Meier-Kolthoff JP, **Kumar N**, et al. The Mouse Intestinal Bacterial Collection (miBC) provides host-specific insight into cultured diversity and functional potential of the gut microbiota. *Nat Microbiol.* 2016;1: 16131. doi:10.1038/nmicrobiol.2016.131

13 Presentations

Talk: Modeling and managing of microbial communities.

- 3rd International Metaproteome Symposium, Helmholtz Center for Environmental Research, Leipzig, Germany. 05/12/2018.
- BMGC (Biomedical Graduate school Aachen), 15/02/2019

Poster: MiMiC-A computational method for automated generation of minimal microbial consortia.

- Microbiome, Cold Spring Harbor Laboratory (CSHL), Newyork, United States. 19/06/2019
- BMGC (Biomedical Graduate school Aachen), 02/12/2017