Dissertation

submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruperto Carola University Heidelberg, Germany
for the degree of

Doctor of Natural Sciences

Presented by:

Renato Alves

Lic. Molecular and Cellular Biology

born in: Covilhã, Portugal

Oral examination on 22nd November, 2019

Integrating metatranscriptomes and metagenomes for deconvolution of composition and expression in human gut and artificial communities

Referees:

Dr. Athanasios Typas

Prof. Dr. Robert Russell

Acknowledgements

During the development of the work that culminated in the creation of this PhD thesis, I was privileged to work with and meet many colleagues, nurture friendships with nerdy (scientific) discussions, and be part of a warm and welcoming community of interesting and interested folks. To you all an enormous thank you!

First and foremost, I would like to thank my supervisor Peer Bork for guiding my quest towards the light and offering me the opportunity to work in such challenging topics. Peer has always been supportive of my curiosity and *geekiness* allowing me to explore different topics and technologies with independence. At the same time, his patience and persistence always ensured projects were constantly being pushed further, even when everything seemed to be reaching a standstill. I am deeply thankful to him, for being an inspiring mentor, for the easy and hard lessons, and overall support that helped me shaping my scientific self into what it is today.

I am also grateful to my TAC committee members, Georg Zeller, Nassos Typas and Rob Russell, for invaluable feedback and an always critical view of the work. Their input was vital to shape this work over the years into its final form.

From all the people I had the pleasure of working with I would like to thank: Paul Costea for his sharp and often clashing sense of scientific correctness and integrity, Luis Coelho for his mentorship and expertise, both technical and biological, that resulted in several published works presented in this thesis, Sebastian Schmidt for his guidance and intuition that lead to some of the directions and work herein discussed and Jaime Huerta-Cepas, for his expertise and input with regards to the use of *eggNOG* and for great times as part of Google Summer of Code.

I would also like to acknowledge Lisa Maier, Ece Kartal, Sonja Blasche, Tobias Wenzel, Marja Driessen, Rajna Hercog and Anja Telzerow for their collaboration and contribution that culminated in most of the work presented in the third chapter. Without their input and knowledge this chapter would simply not exist.

A special *thank you* is reserved for Thea Van Rossum, Michael Kuhn and Vladimir Benes, without whom this work would likely not be possible, but most of all, for being three of the most

supportive, friendly and caring colleagues I had the pleasure of working with.

For a daily dosage of good mood and to always create a genuinely fun environment to work in, I cannot go by without thanking my office mates Lucas, Anna, Ece, Oleks and Pamela. Your humor and friendship will far outlive this work **:)**.

Last but not least, I would like to thank my parents, my partner and our sprout for keeping up with me, shaping me into who I am and always pushing me higher and further, even at the cost of personal sacrifices.

A special mention is reserved to a rather unique grandma cat that withstood all the squeezing and petting required to produce this monograph. It is fair to say that no cats were harmed during the production of this work.

Table of Contents

Summary

Over the last 15 years the human microbiome has received increasing attention. During this time, many studies have contributed to shed light on the complex network of interactions both between the microorganisms and their host, and within microbial communities themselves.

While traditionally aiming at assessing composition, recent studies have broadened this scope to multi-dimensional aspects, using multi-omics approaches. By integrating information about genomes, transcripts, proteins and metabolites, a holistic understanding of the microbiome is now within reach. However progressive, these studies generally suffer from a lack of closure, as interpretation and integration of this data is all but straightforward.

In the particular case of metatranscriptomes, species abundance and gene expression are coupled into a single readout. Consequently, normalization of this data is a crucial but poorly understood and unresolved problem. Here I present different approaches to normalise metatranscriptomes and highlight procedural concerns when obtaining this type of data.

Results show that better normalization strategies are necessary when integrating multi-omics data and that controlled pilot experiments are required for a better understanding of the intricate dynamics and interactions between members of these communities. This work further exposes concerns about the interpretation of functional aspects of microbial populations, primarily driven by the many uncontrolled sources of variation herein discussed.

As these new data types become more widespread, methods will certainly evolve towards better standardization and controlled procedures. This will help the microbiome field to move beyond its descriptive state into one able to provide a more detailed and mechanistic understanding.

Zusammenfassung

In den letzten 15 Jahren hat das menschliche Mikrobiom zunehmend Aufmerksamkeit erhalten. In dieser Zeit haben viele Studien dazu beigetragen, das komplexe Netzwerk der Wechselwirkungen zwischen den Mikroorganismen und ihrem Wirt, sowie innerhalb der mikrobiellen Gemeinschaften selbst zu beleuchten.

Existierende Studien beschäftigten sich vor allem mit der Analyse der bakteriellen Zusammensetzung. Eine neue Entwicklung ist, auch mehrdimensionale Aspekte zu betrachten, vor allem mit sogenannten "Multi-Omics"-Studien. Durch die Integration von Informationen über Genome, Transkripte, Proteine und Metaboliten ist nun ein ganzheitliches Verständnis des Mikrobioms in greifbarer Nähe. Trotz aller Fortschritte mangelt es im Allgemeinen an greifbaren Schlussfolgerungen die alle Stränge miteinander verbinden, da die Interpretation und Integration dieser Daten alles andere als einfach ist.

Im speziellen Fall von Metatranskriptomen werden Spezies-Abundanz und Genexpression in einer einzigen Messung erfasst. Folglich ist die Normalisierung dieser Daten ein entscheidendes, aber wenig verstandenes und ungelöstes Problem. Hier stelle ich verschiedene Ansätze zur Normalisierung von Metatranskriptomen vor und erörtere die praktischen Probleme bei der Gewinnung dieser Art von Daten.

Die Ergebnisse zeigen, dass bei der Integration von Multi-Omics-Daten bessere Normalisierungsstrategien erforderlich sind und dass kontrollierte Vorstudien erforderlich sind, um die komplexe Dynamik und Interaktion zwischen Mitgliedern dieser Gemeinschaften besser zu verstehen. Diese Arbeit zeigt weiterhin die momentanen Grenzen hinsichtlich der Interpretation funktionaler Aspekte mikrobieller Populationen auf. Diese sind hauptsächlich durch die vielen unkontrollierten Variationsquellen bedingt.

Mit zunehmender Verbreitung dieser neuen Datentypen werden sich die Methoden zu einer besseren Standardisierung und zu kontrollierten Verfahren entwickeln. Dies wird der Mikrobiom-Forschung helfen, über qualitative Studien hinaus zu quantitativen Ansätzen zu gelangen, die ein detailliertes und mechanistisches Verständnis liefern können.

1

Introduction

1.1 Microbiome studies

With recent technological advances in DNA and RNA sequencing technologies scientific disciplines such as microbial ecology and in particular experienced a burst of information availability. In little over a decade the repertoire of known prokaryotic species and the knowledge of their interactions and importance has expanded several fold. In the context of human health, the findings that these microorganisms have a significant impact on well-being and can be causative or predictive of some diseases (*68*, *129*, *133*) has lead to an increased interest in their study, the study of the microbiome.

A large part of this scientific advance has been made possible thanks to techniques such as metagenome and metatranscriptome shotgun sequencing, targeting DNA and RNA, respectively. These techniques allow peeking into the biological potential and activity of entire communities by providing a snapshot of the DNA and RNA content at a given point in time. However, the dynamic nature of these environments poses interesting challenges both logistically (sampling, preservation, …) and experimentally (design, extraction, sequencing, analysis, …). In the case of the human gut, periodic cycles linked to behavioral patterns such as daily or weekly routines (*28*), as well as, diet, geographical, demographic and epidemiological changes (*131*, *136*) define a multidimensional problem whose implications and relevance is still

an active topic of research. Together with technical aspects, the interpretation of these results is very challenging and has lead to conflicting interpretations (*1*, *39*).

Comparative studies using metagenomes and metatranscriptomes while having great potential are also very challenging from a technical standpoint with elaborate protocols and time sensitive steps. Latest studies follow a *best effort* approach where some but not all recommended practices are followed. This includes some form of preservation and conservation of the biological material after collection, simultaneous extraction of DNA and RNA and depletion of ribosomal RNA (rRNA) and transfer RNA (tRNA). One should also emphasise that these practices, although recommended, also introduce bias of different nature. Due to the availability of several options and commercial kits, the field has yet to converge on universal and standard protocols that address most of these issues. As such, while these bias are somewhat uniform within project, they pose significant challenges when integrating several projects.

1.2 Moving beyond function potential with metatranscriptomics

While working towards the understanding of the complex interactions in the microbiome, one must not only focus on the functional potential but also in its realization. Many studies (*9*, *11*, *19*, *83*, *132*) evaluate and discuss functional aspects of the microbiome. Yet, by using only metagenomic data, their interpretation is but an assessment of the possible functions that the microbiome can perform, not of those actually being performed. In order to assess the actual functional activity one must move beyond metagenomes DNA towards readouts that better reflect what these communities are doing, such as metatranscriptomics RNA, metaproteomics (proteins) and metabolomics.

While moving towards functional analysis, both proteins and metabolites are highly variable and available methods are still plagued with obstacles in assigning experimental readouts to their respective sources. Consequently, the use of metatranscriptomics provides the best compromise. As protocols not unlike those used for metagenomics can be used to sequence transcriptomes, this readout is reachable and approachable. However, the highly dynamic nature of RNA poses technical challenges that transpire into the final result.

In the specific case of metatranscriptomics, composition and expression are intertwined in the final output (fig. 1.1). As the use of metatranscriptomic for human studies is still in its early days (*37*, *43*, *70*), current approaches overlook many of the technical aspects underlying this data. In hope that biological signal outweighs noise, such studies focus primarily on the end-goal, relying on several assumptions along the way.

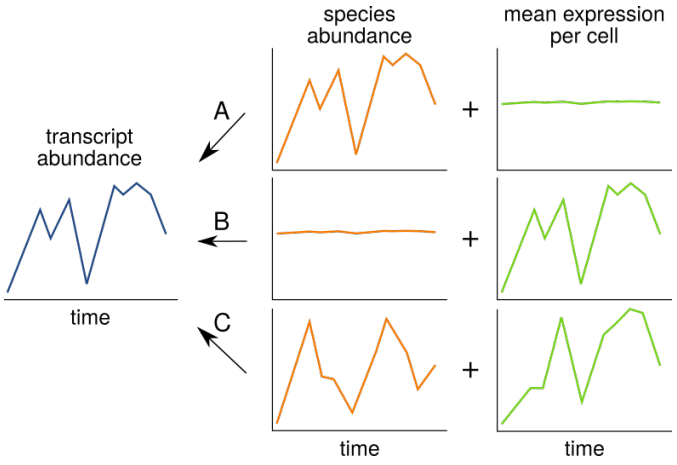

Figure # 1.1: ***Convolution of abundance and expression signals*** *- illustrating scenarios with different biological behavior and same transcriptomic readout.*
A*-A species that has not changed its transcriptional programme despite changing its representation in the community,*
B*-A stably prevalent species that displays a different transcriptional programme over the specified time frame*
C*-The most common scenario, where expression and abundance are convolved over time.*
Neither of the three scenarios is distinguishable without the ability to deconvolute abundance from expression.

One example is seen in *Franzosa et al* (*36*) and the *HUMANn2* software. Here, an older software *MetaPhlAn2* (*118*) is used to taxonomically profile metagenomic samples and subsequently used as reference abundance to normalise associated metatranscriptomes. This approach is in principle valid, but relies on the assumption that the estimates produced by the profiler are accurate and concordant with the species composition of the sample. Were this assumption to fail, the functional responses seen in the metatranscriptome would be normalized to an incorrect level, resulting in an erroneous interpretation of the results.

One additional confounder in metatranscriptomic studies is that of rRNA. Bacterial RNA is known to be composed of 95-97% rRNA and tRNA (*97*). Such high percentage would result in a high proportion of sequenced reads belonging to rRNA reads while shadowing remaining transcripts. Consequently, metatranscriptomic protocols often require the removal of rRNA and tRNA transcripts by different means. In the case of eukaryotes, mRNAs contain a poly-A sequence amenable to hybridization. This principle is exploited by kits that use poly-T rich columns to bind poly-A sequences allowing rRNA and tRNA, that lack such sequence, to be washed through. Bound molecules can afterwards be released yielding almost pure mRNA sequences. On bacteria however, RNA does not possess poly-A and as such this approach cannot be used. Instead solutions often involve binding to specific regions of rRNA and tRNA, either with a similar principle to the poly-A or as part of an enzymatic reaction that degrades bound regions. With this approach, rRNA and tRNA can be brought down to as little as 1% or even less. Yet, designing universal sequences targeting a wide range of species is challenging, leading to species-specific biases that can be misinterpreted in the context of relative abundances (*119*).

In light of these challenges, efforts towards a better understanding of the implications of different

normalization strategies and the impact of technical noise in the interpretation of metatranscriptomes, is in order.

1.3 Microbiome software

Over the years, the expression *academic software* has been used as synonym of poor quality. The main contribution to this reputation is the fact that scientific projects are fast moving and short lived and software is often released to the public with little care for its future. Thanks to initiatives such as the *European Academic Software Award* (*10*), the adoption of a *code required* policy by most publishing institutions, increasing awareness towards reproducibility and standardization in science (*91*) and large community efforts to improve the scientific bioinformatics ecosystem (*3*, *41*), this reputation has changed to more positive notes.

The complex and dynamic nature of the microbiome field is also reflected in its software and available methods. A plethora of popular and well maintained tools exist (*13*, *21*, *36*, *75*, *82*, *101*, *110*, *118*) allowing the design of studies that require the analysis of large or very large amounts of data (*2*, *43*, *44*, *70*, *100*, *102*, *109*, *120*). However their use often involves building custom multi-step pipelines and execution with non default parameters. While perfectly natural this approach adds to the challenge of standardization by making projects isolated in their analysis and difficult to compare without significant effort. Similarly the use of different experimental and computational protocols and technologies increases the likelyhood of effects of unknown or uncontrolled origin (*24*). Although some of these effects can be mitigated batch correction techniques, their use often comes at the cost of loss of signal (*26*, *38*). Be that as it may, efforts towards standardization, transparency and reproducibility of computational approaches are crucial not just as best practice but also to ensure a sustainable scientific endeavour.

1.4 Artificial communities

The so called microbiome field is often perceived to be at odds with traditional microbiology. Both of these disciplines try to understand microbial communities. While microbiome studies often use a top-down approach, collecting large amounts of data, and trying to find patterns in the data that are explanatory or associate with different conditions, traditional microbiology focuses on small scale experiments, controlled environments and mechanistic understanding of the underlying biology of these species.

In recent years, the microbiome field is recognizing that the top-down approach has limitations and that larger amounts of data are unlikely to change this trend. Consequently, these disciplines are converging, primarily driven by an attempt to move microbiome studies beyond associative and into a more mechanistic direction (*111*). Similarly, traditional microbiology is scaling up

by using high-throughput technologies such as microfluidics (*121*), or introducing automation through robot assisted experimental setups.

These efforts typically aim at creating *in vitro*, and controllable environments, mimicking the conditions experienced *in vivo*, where multiple species are grown together. This kind of setup allows mechanistic studies by analyzing the interactions between the different members of the community, and evaluate their evolution under different conditions and perturbations. Some of these efforts are now coming to public and are already providing interesting biological insight into these communities and both synergistic and antagonistic behaviors seen when organisms are brought together (*42*, *55*, *72*).

The use of these systems also brings advantages in terms of protocol validation and optimization. One often mentioned concern in microbiome studies is the poor use of standards and controls. This criticism is often coupled with recommendations on approaches that could address these issues. Such is the case of spike-ins (*99*), used to control efficiency of DNA/RNA extraction and sequencing bias. Despite reasonable, these recommendations have been received with resistance and are yet to been implemented and integrated into common practice. An often used argument is that metagenomics and metatranscriptomics protocols are already quite complex, both in terms of sampling and processing, and introducing additional steps will only increase the sources of bias, a point of view that while valid is arguably flawed. Consequently, the adoption of these and other best practices has given microbiome studies a reputation of being lax in terms of scientific rigor.

In all, the study of artificial communities while losing some of the properties of *in vivo* systems, allows an entirely different level of control over many source of noise. In time, these communities will certainly provide the much needed mechanistic insight that will help push the microbiome field forward.

1.5 Objectives

Following from the need to develop better approaches to analyse metatranscriptomic data, several objectives are outlined:

- Define a set of standard approaches for metagenomic studies
- Identify a robust strategy to normalize and analyse metatranscriptomic data
- Understand the impact and validity of these strategies under different conditions and perturbations

Each of these will be explored in a different chapter of this thesis.

2

Method development in microbiome analysis

2.1 Introduction

A large number of microbiome studies use, as the basis of their analysis, DNA or RNA sequencing data. Be it 16S rRNA amplicon, shotgun metagenomics or shotgun metatranscriptomics, these datasets are typically large in both size and numbers. As a consequence, the use of computer assisted methods for their analysis is inevitable.

In this chapter I describe two softwares and one resource created to provide support for such analysis. The first software is a taxonomic profiler which provides estimates of the abundance of species in metagenomic samples. This kind of tool is often a starting point of many metagenomic studies. The second is a generalist nucleotide sequence processing, pipelining, analysis tool and language. This tool provides an all around modular framework for quality control, mapping, assembling, annotating, profiling and feature counting of sequence data. Last but not least, I describe a global gene catalog resource that can be used both to discover unknown and novel genes in different habitats and also as reference to future metagenomic and metatranscriptomic studies.

2.2 Consistency between taxonomic profilers

As many, if not most, microbiome analysis require an assessment of the species composition and abundance of any given sample, tools such as *mOTUs* (*110*) and *MetaPhlAn2* (*118*) are frequently a critical point in these studies, either as standalone or integrated into larger pipelines such as *HUMANn2* (*36*). These tools try to estimate abundance by using universal (*mOTUs*) or clade-specific (*MetaPhlAn2*) marker genes. Other tools rely on assigning individual reads to genomic or taxonomic databases (*48*, *127*, *128*). However, as closely related genomes tend to have regions of high similarity, reads mapping to such regions have to either be distributed across all equally scoring hits or be masked and ignored entirely. Both of these approaches introduce normalization and technical bias leading to skewed abundance estimates. This problem can be partially overcome by using marker genes which are by design unique and taxon- or clade- specific. This is the case for both *mOTUs* and *MetaPhlAn2*. Additionally, if marker genes can be universally defined, that is, encompassing the entirety of the tree of life (here only the prokaryotic branch), estimating abundance from species lacking a complete reference genome becomes a possibility. This is the case for *mOTUs* and in contrast to *MetaPhlAn2*, which requires reference genomes in order to construct its database of clade-specific markers.

Using simulated test datasets generated by the Critical Assessment of Metagenome Interpretation (CAMI) group (*103*), the performance of both tools was assessed. These datasets include three groups with variable complexity reflecting different habitats, environments and sequencing depths (see tbl. 6.1).

While comparing output of both tools (see fig. 2.1), two outstanding issues were noticeable. First, the results show reasonable performance at genus level but poor at species level, with *MetaPhlAn2* scoring slightly better. Second, the biggest challenge, and a definite source of loss of signal, is the mapping of the outputs of both tools to a common and comparable taxonomic reference. This step only slightly affects the profiling obtained by *mOTUs* tool, since NCBI taxonomy identifiers are included in the output, However, the mapping step represents a bigger problem when using *MetaPhlAn2*-derived profiles. In fact, not only *MetaPhlAn2* doesn't provide any NCBI taxonomy identifiers, it further complicates this integration by using species labels with masked non-alphanumeric characters. This masking approach results in over 60,000 labels, that could not be mapped back to the NCBI taxonomy. As a result, this method leads to an inflated rate of false positives, and, consequently, a lower precision rate. Moreover, the fact that these tools were released at different points in time implies that their underlying taxonomic reference is different, requiring additional translation to common ground and comparable identifiers. Similar hurdles were found by the authors of the CAMI study that, although using a slightly different strategy, achieved an identical result (*103*).

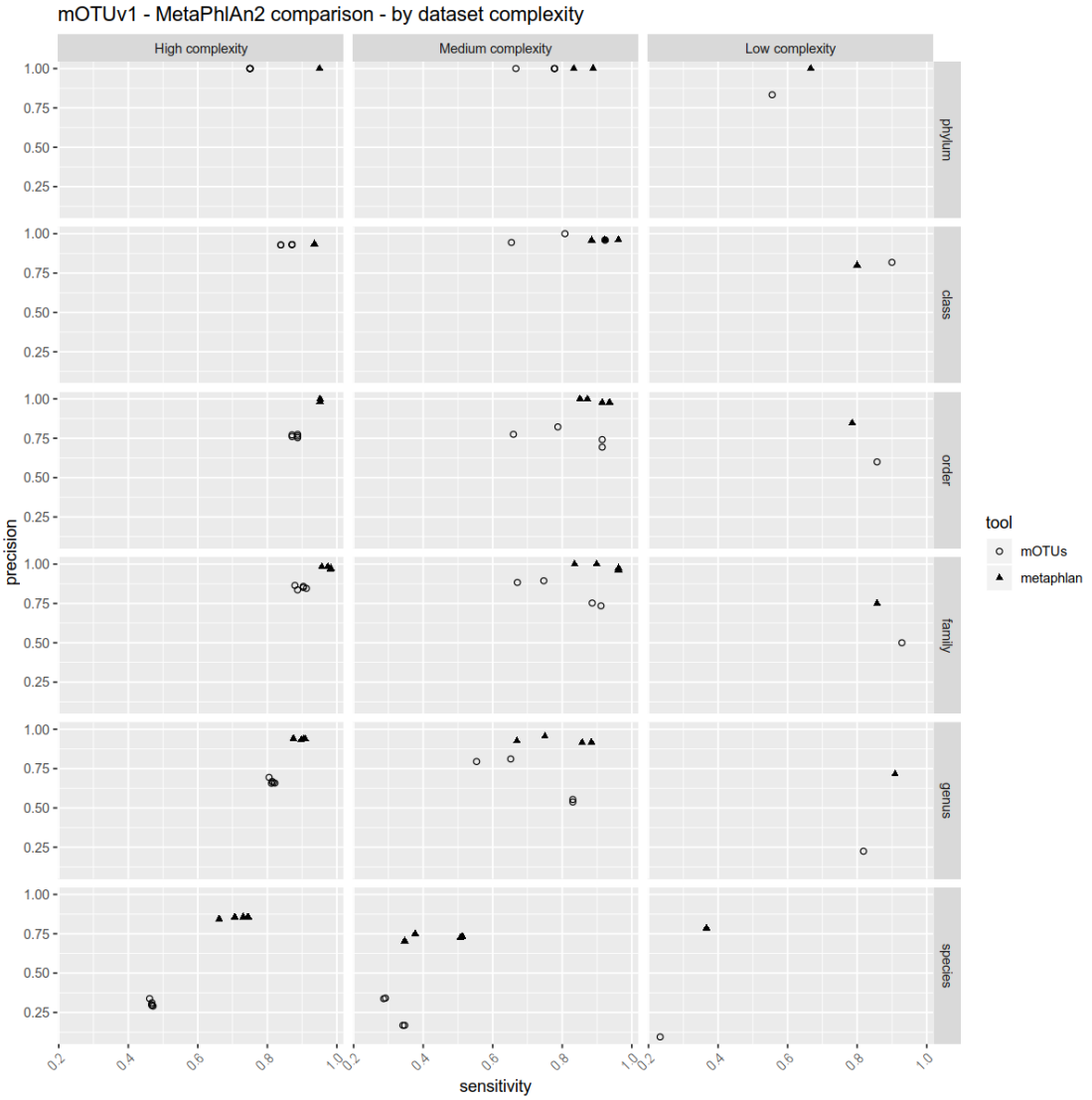

Figure # 2.1: ***Evaluation of mOTUs-v1 profiler and MetaPhlAn v2.5 on CAMI test datasets*** *- medium and high complexity datasets approximate a real sample and as such represent a more realistic outcome. MetaPhlan2 shows overall better performance*

These and many other methods have been since benchmarked, with larger datasets, as part of the CAMI effort (*103*), which aims at providing a broad and unbiased assessment of different metagenomic analysis tools.

After this initial assessment, and as part of the work towards an improved version of the *mOTUs* profiler, this analysis was re-done using the newer version of the profiler, *mOTUs2*. This analysis has been also included in the *mOTUs2* publication (*82*). In the paper, a significant improvement is also visible, specially when compared to the original *mOTUs1* (see fig. 2.2). where a significant improvement is visible compared to its previous version (see fig. 2.2). This improvement spans throughout the entire range of taxonomic levels we investigated (panels **h-o**), as well as in genomes assembled from metagenomic samples (panels **a-g**). For the latter, the improvement is particularly outstanding, compared not only to *mOTUs* but also to alternative publicly available profiling tools. One relevant exception is visible in (panel **n**), where species-level profiling remains challenging, especially in terms of recall, for all the tools considered, due to the high-complexity nature of the dataset we investigated. This scenario reflects a simulation strategy that penalizes most profiling tools by including circular elements (see tbl. 6.1) that are particularly challenging to profile.

As of the writing of this work, *mOTUs v2.5* has been made public. This version expands the underlying database with close to 60,000 new reference genomes, effectively duplicating the range of species that can be profiled with this tool. *MetaPhlAn2*'s database, although large, has not been updated since its initial release. It will be interesting to see how the latest versions of these tools perform, something which is likely to happen soon since a second iteration of the CAMI challenge is, at the time of writing, underway.

Despite their partial disagreement and their sub-optimal ability to profile simulated samples at species level, *mOTUs* and *MetaPhlAn2* remain among the most widely used and best tools for this task. The increasing availability of reference genomes will help improving the internal databases of these tools and consequently their profiling performance.

2.3 Reproducible microbiome analysis with NGLess

The topic of reproducibility in science has since long been a concern and recurs when high impact science faces concerns of irreproducibility (*53*, *96*), in some cases by the same authors of the initial study (*69*). Reproducibility in metagenomic studies is particularly challenging, both experimentally and computationally. Experimental reproducibility can be improved by the use of standard protocols and procedures (*74*, *115*, *116*). On the other hand, computational reproducibility requires a set of good practices, such as the use of versioning for resources and software, logging for pipeline steps and settings (*40*), just to name a few (well illustrated in fig. 2.3).

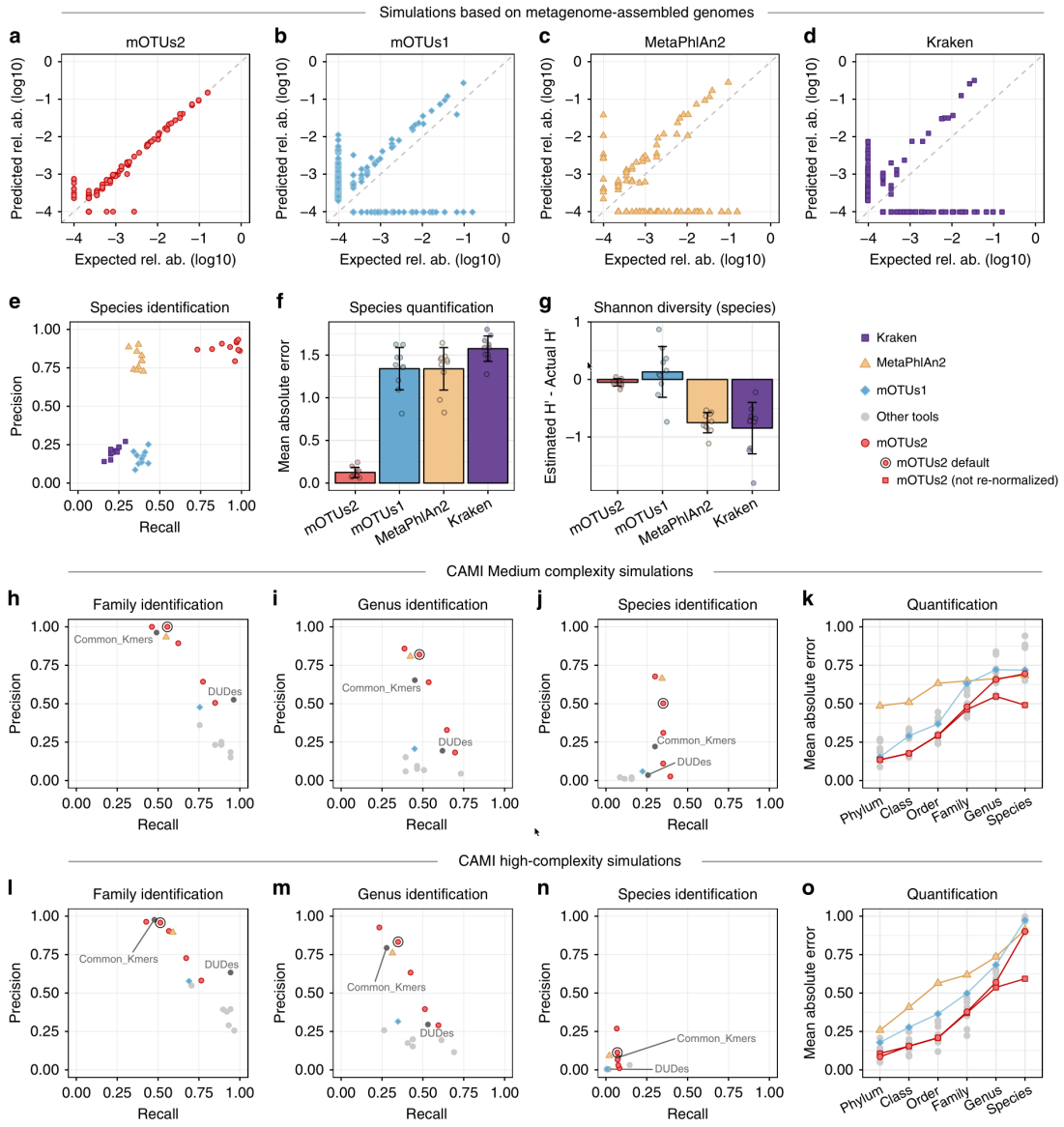

Figure # 2.2: ***Evaluation of mOTUs profiling against MetaPhlAnv2 and other tools*** *- in Milanese* et al. *(82)*

Conciseness, readability and extensive and detailed documentation also remarkably improve the reusability of computational pipelines and the reproducibility of their results. *NGLess* and its *ng-meta-profiler*s (*21*) (see also lst. A.1) have been conceived and developed keeping in mind these aspects, in order to maximise reproducibility, accessibility and readability.

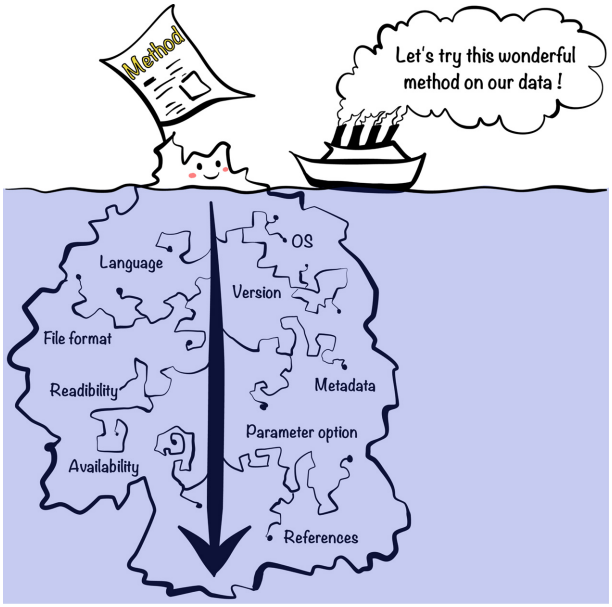

Figure # 2.3: ***Hidden reproducibility challenges in computational analysis*** *- being able to use a published method often involves dealing with issues not considered by the authors such as system incompatibilities, a big list of checks - in Kim* et al. *(56)*

NGLess is both a language and a framework. Software and resources are versioned and either built-in or downloaded on first use. Parameters and pipeline steps are explicitly defined as part of the *NGLess* language, itself also versioned. Moreover, computations are automatically considered outdated if either the recipe or any of its parameters is modified, thereby ensuring a consistent and clean execution of the task, avoiding unwanted mixture of results from different configurations.

NGLess has borrowed inspiration from existing software, in particular MOCAT2 (*60*), which used a convenient folder organization. Similarly, the first version of *mOTUs* (*110*) was reimplemented in *NGLess* in a more convenient, bug ridden and reusable format. Additionally, *NGLess* was built with modularity in mind and as such provides interfaces to extend the vocabulary of actions, allowing users to use tools beyond those provided as built-in. This is the case of `ngless-contrib` (*138*), a collection of external modules open to community contribution and one of my contributions to this project.

As part of the process to build *NGLess* into a robust tool, thousands of samples were processed (see sec. 2.4), hundreds of bug reports, bug fixes and new features were implemented, including the addition of algorithms to better support the analysis of metatranscriptomic data. This

involved not just a conspicuous software development effort over several years but also constant testing, improving documentation and occasionally providing training to users. Direct contact with users and feedback through surveys allowed us to reassess design decisions, improve overall usability and readability. On average 70% of users found the *NGLess* language easy to interpret on first contact, reaching almost 100% after training.

In all, the *batteries included* approach of *NGLess*, the emphasis on reproducibility and the effort in standardizing common metagenomic analysis through *ng-meta-profilers*, make this framework a convenient and reliable platform for microbiome studies and more generically, processing of Next Generation Sequencing (NGS) sequencing data. This work is publicly available at `http://ngless.embl.de` and has been published in *Coelho et al. 2019* (*21*).

2.4 Capturing genetic novelty beyond reference genomes

High quality genome assemblies are the most reliable reference for metagenomic analysis. Well studied species and strains often have detailed genetic annotations both in terms of the definition of genes and its functional capabilities and behaviors. Model bacterial organisms such as *Escherichia coli*, *Bacillus subtilis* and *Mycoplasma genitalium* are a small example in an ever growing list of high quality references. Reaching such high depth of understanding of a species often requires culturing in isolation. While culturing is in itself challenging (*84*, *108*), large efforts are ongoing (*35*, *137*). Once successful, these efforts will open up a new range of possibilities. The availability of isolated species will allow the design of perturbation studies in controlled environment. Similarly, pure cultures will warrant high quality DNA, free of contaminations, that will provide high quality reference genomes.

One of the great findings enabled by metagenomic studies is the fact that microscopic life is much richer and more diverse than initially thought. Many new species have been found through 16S rRNA profiling and metagenomic shotgun sequencing. Occasionally, entirely new taxonomic clades have been discovered, such as our proposed *Borkfalki ceftriaxensis*, representative of a new *Comantemales* order (*44*). In the same direction, recent studies managed to almost double the current repository of reference genomes by assembling and binning a large amount of metagenomes in order to obtain Metagenome Assembled Genomes (MAGs) (*4*), a technique that can yield low-to-high quality reference genomes (*85*, *89*, *126*). Several strategies have been used to generate MAGs, some such as *MetaBat2*, packaged as a reusable pipeline (*54*). These approaches usually consist of two main steps. First, contigs are assembled from individual samples or co-assembled from many samples and scanned for marker genes. Second, by assessing the abundance of these contigs and marker genes across samples it is possible to bin them based on co-abundance under the assumption that contigs originating from the same species have approximately the same abundance. A MAGs is therefore the result of binning co-assembled contigs based on their co-

abundance across several samples. While these techniques are highly prone to the generation of chimeric sequences by fusion of unrelated species, their usefulness is undeniable and continued efforts will most certainly improve the quality of the resulting genomes and minimize the creation of artifacts.

As microbiome studies often focus on understanding the dynamic nature of these communities, another useful approach is to center the analysis on genes. With the increase of metatranscriptomic studies, genes became the primary target of most analysis, be it as part of differentially expression analysis or simply as a proxy for functional inference. A logical step in this direction are gene catalogues (*23*, *67*, *92*, *109*, *130*) and MetaGenomic Species (MGS) (*87*). The first, as the name implies, consists of catalogues of genes compiled from assembled metagenomes and/or metatranscriptomes. The second, and akin to MAGs, consists of binning genes based on their co-abundance across samples. While gene catalogues and MGSs can primarily serve as reference to future studies, they are also a good way to discover novel genes potentially originating from unknown species and in some cases having no homology to existing databases.

Given the focus of this work on metatranscriptomic analysis I had the opportunity to collaborate in the creation of the first version of the Global Microbial Gene Catalogue (GMGC). This catalogue was derived from 12,743 publicly available high quality metagenomes spanning different habitats and body sites tbl. 2.1. A total of 2,007,736,046 Open Reading Frames (ORFs) were predicted and clustered together with 312,020,843 genes from 84,029 high quality genomes in the ProGenomes database (*78*) for a total of 302,655,267 gene clusters (*22*). The analysis and generation of this catalogue required over 10 million Central Processing Unit (CPU) hours distributed over the period of a year on a cluster with over 3000 CPU cores at European Molecular Biology Laboratory (EMBL). My contributions to this work involved an initial benchmark, co-designing the first steps of the pipeline (which included processing the 12,743 metagenomes into the catalogue of 300 million gene clusters), re-processing the initial samples against the gene clusters and annotating the catalogue. In addition to driving a large part of the computation effort required for this project, I was also responsible for a strategic decision that allowed reducing the computational time required from 20 million CPU hours (estimated) to little over 6 million CPU hours, effectively reducing the project time frame from 9 to 3 months.

Table 2.1: Distribution of samples used to build GMGC(v1)

Habitat / Body site	Category	Number of samples
human_gut	Human	7059
human_oral	Human	1593
human_skin	Human	1139
human_nose	Human	228

Habitat / Body site	Category	Number of samples
human_vagina	Human	173
pig_gut	Animal	295
mouse_gut	Animal	230
dog_gut	Animal	129
cat_gut	Animal	124
soil	Soil	312
marine	Water	130
freshwater	Water	104
wastewater	Water	22
built-environment	Buildings	1205
Total		12743

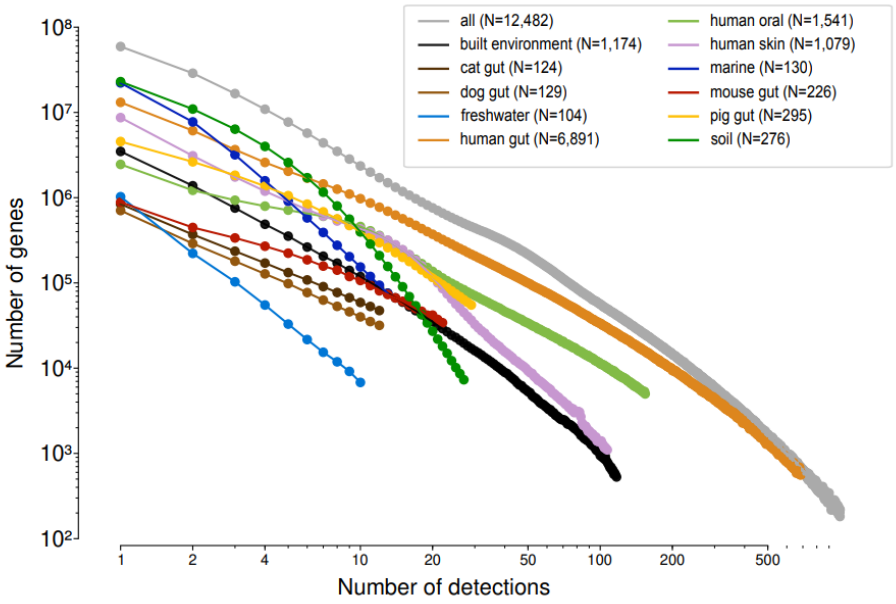

Figure # 2.4: ***Identifying novel and rare genes*** *- number of genes and their prevalence across samples in different habitats - in Coelho* et al. *(in review)* (22)

As part of the many outcomes of this work, of special mention is the somewhat surprising size of the catalogue, which reflects a richer breadth and wider coverage than initially expected, and an increase in the number of novel and rare genes (see fig. 2.4). This work expands previous efforts (*23*, *67*, *92*, *109*, *130*) with over 100 million genes, largely due to the increased number of habitats and samples included. Thanks to this effort we now have a resource that is able to capture a significantly larger portion of novelty in a wider range habitats. Yet, similar future efforts will likely be necessary, since we still didn't capture a large fraction of the biodiversity present in highly diverse and less studied habitats such as *marine* and *soil* environments (see fig. 2.5).

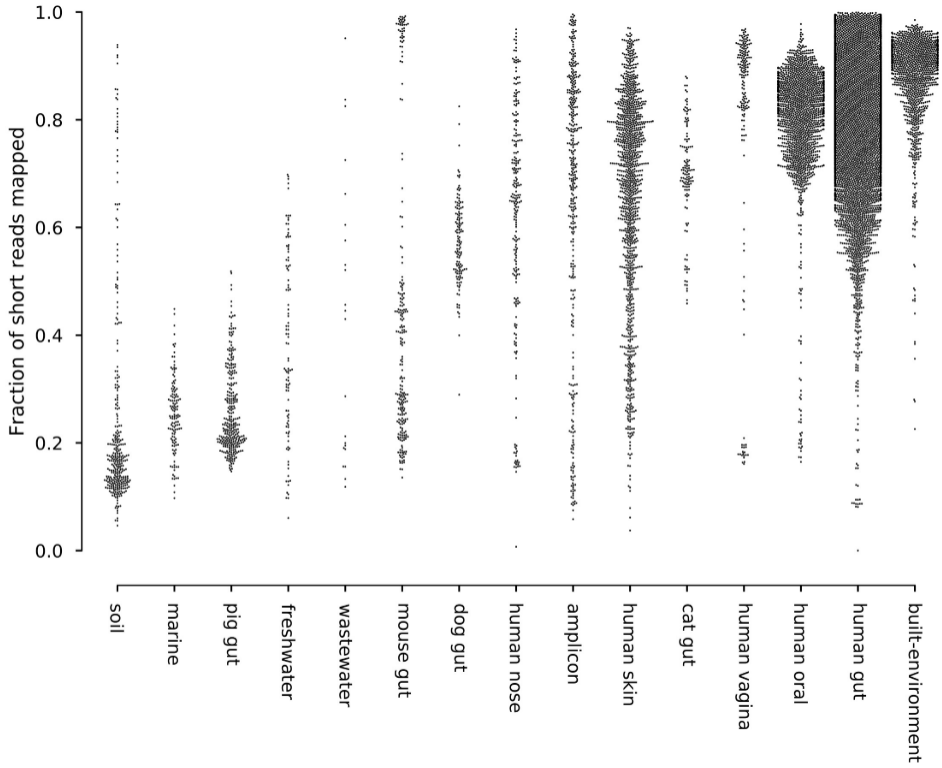

Figure # 2.5: ***Mapping rates against GMGC-v1*** *- habitats with higher biodiversity are still poorly represented in the catalogue, reflecting also lower number of samples, while well studied habitats such as human gut are well represented - in Coelho* et al. *(in review) (22)*

2.5 Discussion

As previously shown, (see fig. 2.2) *mOTUs* (particularly *mOTUs2*) and *MetaPhlan2* provide the best performance to estimate abundance from metagenomes and a starting ground for metatranscriptomic analysis. In the following chapter (sec. 3) I will revisit these tools and their application for profiling both metagenomes and, more importantly, metatranscriptomes.
Together with *NGLess*, these tools provide the foundation for the integration of metagenomes and metatranscriptomes.

As following work required the analysis of thousands of samples, having a robust, performant and streamlined solution is invaluable. In large computational data processing efforts it is inevitable that unexpected system failures happen due to the many components involved, both at software and hardware levels. Being able to perform the bulk of the data processing in a computationally efficient way and with the reassurance of data consistency and reproducibility is an invaluable advantage.

Finally, while the use of the GMGC for further metatranscriptomic analysis could not be pursued due to time constraints, a publication (*22*) is, at the time of writing, under review. Regardless, this resource will certainly be of value for future metagenomic and metatranscriptomic studies as it provides the widest survey of prokaryotic genes to date.

3

Integrating metatranscriptomes and metagenomes

3.1 Introduction

As previously introduced, the readout of a metatranscriptomics experiment is the result of a combination of two sources of variation: abundance and expression (fig. 1.1). When paired metagenomes and metatranscriptomes are available, current strategies to normalize metatranscriptomes (*36*), simply subtract metagenome derived abundances from metatranscriptome counts. This approach assumes that the metagenome reflects the real species abundance and that the same baseline applies to the paired metatranscriptome.

Since the discrepancy between metagenomes and metatranscriptomes could be due to biological effects, one could argue the assumption still holds. However, given the fact that RNA is quickly degraded, is very dynamically regulated and RNA experiments typically involve protocols with many steps, the most likely source of discrepancy is of technical rather than biological nature. Following this, I postulated the possibility of deconvoluting species abundance and gene expression from metatranscriptomes alone, allowing normalization independent from metagenomes. If successful, one would be able to rely solely on metatranscriptomes, and perform both taxonomic and functional assessments of a community. Additionally, being able to rely entirely and

solely on metatranscriptomes would also have a significant economic impact as the need for metagenomes would be drastically reduced. This principle is not entirely novel in the field of RNA-seq, but its application to metatranscriptomes is limited. One example of this approach is seen in *Klingenberg et al.* (*57*). In this work, *scaling factors* are calculated for every species in order remove the effect of species abundance, and before performing differential gene expression analysis. Said *scaling factors* are extrapolated from gene counts and are roughly equivalent to an average across all genes. While suitable for differential gene expression, this approach is not reliable for species abundance estimation. Factors such as genome size, length of genes, paralogy and orthology are not considered, leading to different kinds of bias when estimating species abundance with these methods.

On a different note, abundance is estimated based on sequencing coverage of a genome or, as is the case of the tools considered, marker genes. This approach assumes that different species have an approximately equivalent ratio of DNA to cell number. This assumption has been challenged in the past (*90*) and recently revisited (*107*) but due to technical limitations, relying on coverage remains the best approximation currently available.

Being able to deconvolute species abundance from expression would warrant a more reliable solution to the problem of normalization. In this chapter I will explore existing approaches and present an alternative method to normalize metatranscriptomic data.

3.2 Metagenomes, metatranscriptomes and reference genomes

In order to pursue the goals of this chapter, publicly available paired metatranscriptomes and metagenomes, that is, extracted simultaneously and from the same biological sample, were obtained.

All publicly available datasets considered in this work consist of human stool samples (sec. 6.2), originating from projects researching diseases potentially related with the human gut. As a well studied environment, the majority of most prevalent species has had its genome sequenced and reference genomes are publicly available. Building on existing knowledge, the present work uses representative reference genomes obtained from the curated proGenomes database (*78*) and further reduced to species found to be present in the human body (sec. 6.2.2).

3.2.1 Fragmentation in high quality genomes

Although the reference genomes obtained from proGenomes are curated, many still display a high degree of fragmentation and are either of scaffold or contig quality.

As an initial analysis I assessed the distribution of genome quality. Results are illustrated in

(fig. 3.1).

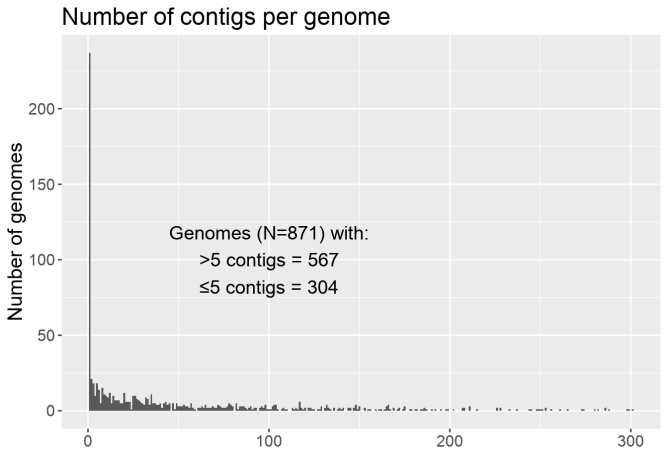

Figure # 3.1: ***Contig distribution in proGenomes reference genomes***

Surprisingly, despite considered high quality, most representative genomes still display a high level of fragmentation. Of the 871 genomes considered, 35% are composed of more than 5 contigs and more than 20 have 200 or more fragments.

The high level of fragmentation complicates the analysis of these species. The lack of complete reference genomes affects the ability to map reads to poorly assembled regions, increasing the fraction of unmapped reads. Additionally, as assemblies are imperfect, higher levels of fragmentation are likely to reflect in increased rates of chimeric contigs due to low coverage at the time of assembly.

After this result, manual inspection revealed that the majority of the affected species are typically reported as having low prevalence and abundance in human gut samples. In light of these results, the exclusion of fragmented genomes was considered but deemed unnecessary.

As genomes are a foundational part of this work, future iterations will certainly benefit from the availability of higher quality and complete references.

3.2.2 Public datasets and taxonomic variation

Public datasets from six public projects (tbl. 6.2) were found to meet the requirements for this work (sec. 6.2). In order to assess the distribution of samples in terms of richness and composition, taxonomic abundances were estimated with *mOTUs2* and displayed in a Principal Coordinates Analysis (PCoA) plot (fig. 3.2) using Bray-Curtis dissimilarity (*14*). In the plot we can see clustering most `HPFS-MLVS` samples together with `T1D_LCSB`, `InternalGT` and `Franzosa2014`, and a few outlier samples. A significant spread is also seen in *Axis 2* affecting some samples from `HPFS-MLVS` and `T1D_LCSB`. A more concerning effect is visible on `HMP2-IBD` and `HMP2-IBDMDB`

samples. These not only display the largest internal dissimilarity, they also sub-cluster by sampling/sequencing institution. Data from `HMP2-IBDMDB` was collected in three different locations, of which at least two sub-clusters are clearly distinguishable in fig. 3.2.

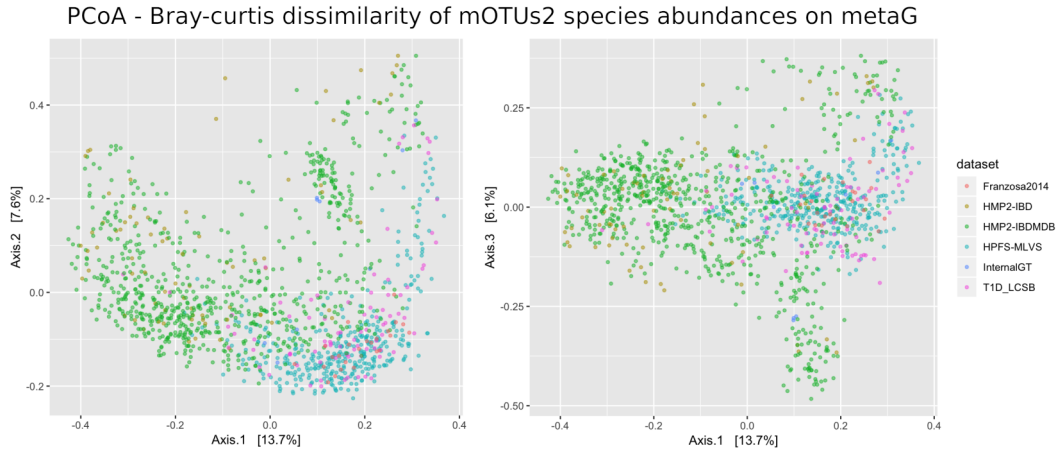

Figure # 3.2: ***PCoA plot of Bray-Curtis dissimilarity on metagenome derived mOTUs2 abundances***

3.3 Taxonomic profilers in metatranscriptomic data

As presented in the previous chapter (sec. 2.2), taxonomic profiling is a frequent and often critical step in metagenomic analysis. Here I evaluate the use of metagenomic profilers on metatranscriptomes and their agreement with profiles generated from the paired metagenomes.

Similarly to what was seen in the previous section (sec. 3.2.2), projects that displayed increased dissimilarity also show accentuated disagreement between paired metagenomes and metatranscriptomes fig. 3.3. Such is the case of `HMP2-IBD` and a contrast to `HMP2-IBDMDB`. Surprisingly, samples from `HPFS-MLVS` also show accentuated lack of correlation. This discrepancy can be partially explained by metadata inconsistencies, identified after several interactions with the authors of this work. Regardless, the fact that a considerable number of samples, from this and other projects, display low correlation or even anti-correlation, reflects poorly on the pairing of this data. Likely, this lack of agreement is a symptom of difficulties in handling the preparation of RNA samples.

On a different note, two additional aspects are visible in fig. 3.3 and highlighted in fig. A.1.
First, profiles generated with *mOTUs2* show consistently better performance than those with *MetaPhlAn2*. This aspect is not entirely unexpected given than *MetaPhlAn2* uses a larger number of marker genes than *mOTUs2*, and consequently, is more likely to be affected by fluctuations in expression.
Second, despite reasonable consistency and performance in some datasets, correlation scores rarely reach `0.8` and show an overall median score of `0.574` for *mOTUs2* and `0.403` for

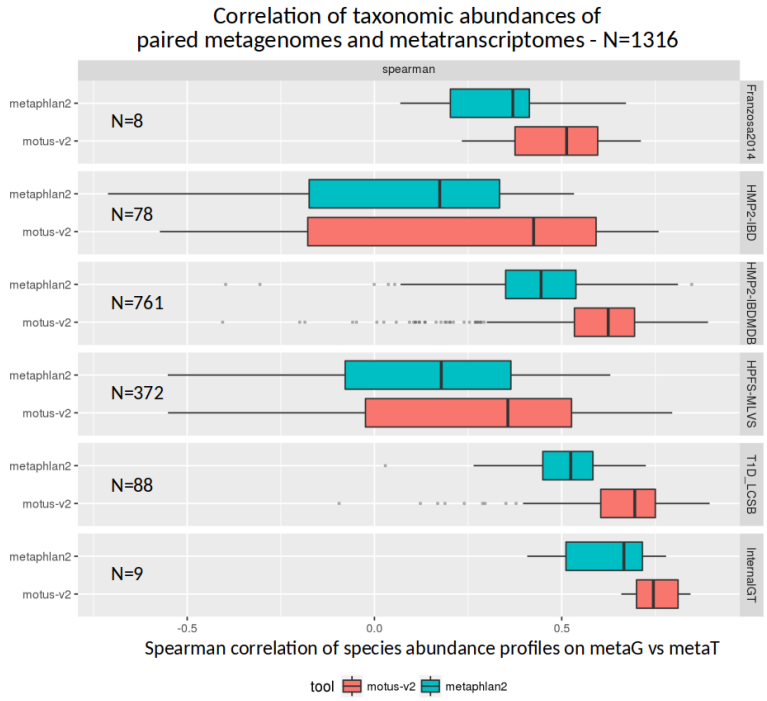

Figure # 3.3: ***Correlation of taxonomic profiles on paired metagenomes and metatranscriptomes*** *- spearman correlation of taxonomic abundances profiled with* mOTUs2 *and* MetaPhlAn2 *on paired metagenomes and metatranscriptomes*

MetaPhlAn2 (tbl. A.2).

In summary, we see that while these tools were designed to profile metagenomes, their performance on metatranscriptomes is subpar.

Both taxonomic profiling tools considered (*mOTUs* and *MetaPhlAn2*) rely on marker genes designed to profile metagenomes. Given the suboptimal performance above, I considered that a different set of marker genes, selected specifically to profile metatranscriptomes, could outperform existing tools.

As such, throughout the remaining of this chapter, I will explore alternatives aiming at improving existing methods while addressing the problem of normalization.

3.4 Counting strategies and mapping thresholds

An initial step in the approach consisted of identifying if the 97% threshold for species delineation previously defined on metagenomes (*79*), is equally valid for metatranscriptomes.

The result is shown in fig. 3.4 together with unique mapping rates at different identity thresholds. Given this result, the 97% identity threshold was deemed equally appropriate.

In metagenomic studies, the problem of reads mapping non-uniquely to different regions of the

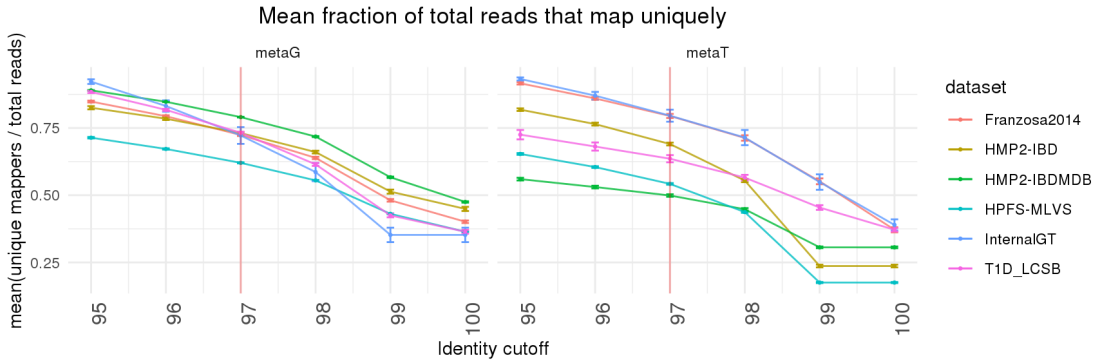

Figure # 3.4: ***Validation of mapping identity threshold*** *- shown is the fraction of reads mapping uniquely at different identity thresholds. The red vertical line highlights 97% identity corresponding to the species threshold as defined in* Mende et al. 2013

same or different reference genomes is common. This issue is aggravated by the high sequence similarity shared between many of the species and strains present in reference databases. A frequent strategy to avoid this problem is to reduce databases to have only representative genomes and few or no related strains. For this reason, `NGLess` implements several different strategies to count such reads, ranging from counting all hits equally to distributing non-unique reads based on the proportion of uniquely mapped reads. In this work, and in order to avoid introducing noise by any such strategies, only uniquely mapped reads were considered. This decision is supported by the relatively high mapping rates seen (fig. 3.4) even when discarding non-unique reads.

In summary, the threshold of 97% sequence identity and the sole use of uniquely mapped reads, were considered adequate.

3.5 Marker gene selection

3.5.1 eggNOG orthology

Much like *mOTUs* and in contrast to *MetaPhlAn2*, universally defined marker genes were considered better targets for profiling metatranscriptomes. Following the same approach used by *mOTUs*, genes were selected based on sequence conservation across species using orthology annotations derived from `eggNOG` (*46*). *mOTUs* relies on universal and single-copy marker genes (*110*). In `eggNOG` this corresponds to the Non-supervised Orthologous Group (NOG) level (`eggNOG 4.5`) and *mOTUs* genes are a subset of Cluster of Orthologous Groups (COGs) (*114*), originally defined at the highest level of taxonomy (*superkingdom*).

For the approach used in this work, the criteria of *single-copy* was dropped as it was considered too restrictive, yielding the same set of markers used by *mOTUs*. Instead, all orthologous groups defined at `eggNOG`'s, `bactNOG` level were considered.

In order to identify and annotate NOGs in the reference genomes considered (sec. 3.2), `eggnog-mapper` (*45*) was used together with its database constructed from `eggNOG v4.5`. These annotations were in turn used to create an annotation file consisting of genes NOGs and their corresponding annotated regions for every genome.

In turn, once quality controlled and mapped against the reference genomes, samples were profiled using said annotation file as described in sec. 6.2.5.

Finally, candidate marker genes were subsequently selected from these gene profiles, using a combination of supervised and unsupervised approaches described in the following sections.

3.5.2 Supervised strategies

Using an empirical understanding of what could define an ideal set of marker genes, two strategies were outlined (sec. 6.2.6.1).

1. Display small and stable dynamic range of expression
2. Display good correlation with species abundance

The top and bottom ranking genes for each approach are shown in fig. 3.5.

The results from the two approaches, while intuitive by design, reveal interesting properties about the selected candidate genes.

Upon closer inspection, strategy **1.** aiming at identifying genes with a small dynamic range of expression, showed inconsistent results across datasets. This resulted in a different rank order and therefore, a different set of genes for each project, which is undesirable. Interestingly, only one of gene families included in the *mOTUs* set appeared in the top 20 of this approach (fig. 3.5 - panel **A**).

Similarly, strategy **2.**, while stable across datasets, benefited gene families with large numbers of paralogs, such as *COG0642 - Histidine kinase* and *COG0534 - Mate efflux family protein*. This property interestingly reveals that large families, while covering a larger portion of the genome of each species, recruit a larger number of reads. Consequently, this drives the correlation score to higher values by introducing outliers with higher count numbers. Gene families with high numbers of paralogs, while better at recapitulating the metagenomic abundance, are also very variable across species, something, once again, undesirable. In contrast to the first strategy, none of the 10 gene families included in the *mOTUs* set ranked in either the top or bottom 20 sets (fig. 3.5 - panel **B**).

In order to have a better understanding of the performance of these strategies, in the following section, both selections will be revisited and assessed in context with unsupervised approaches.

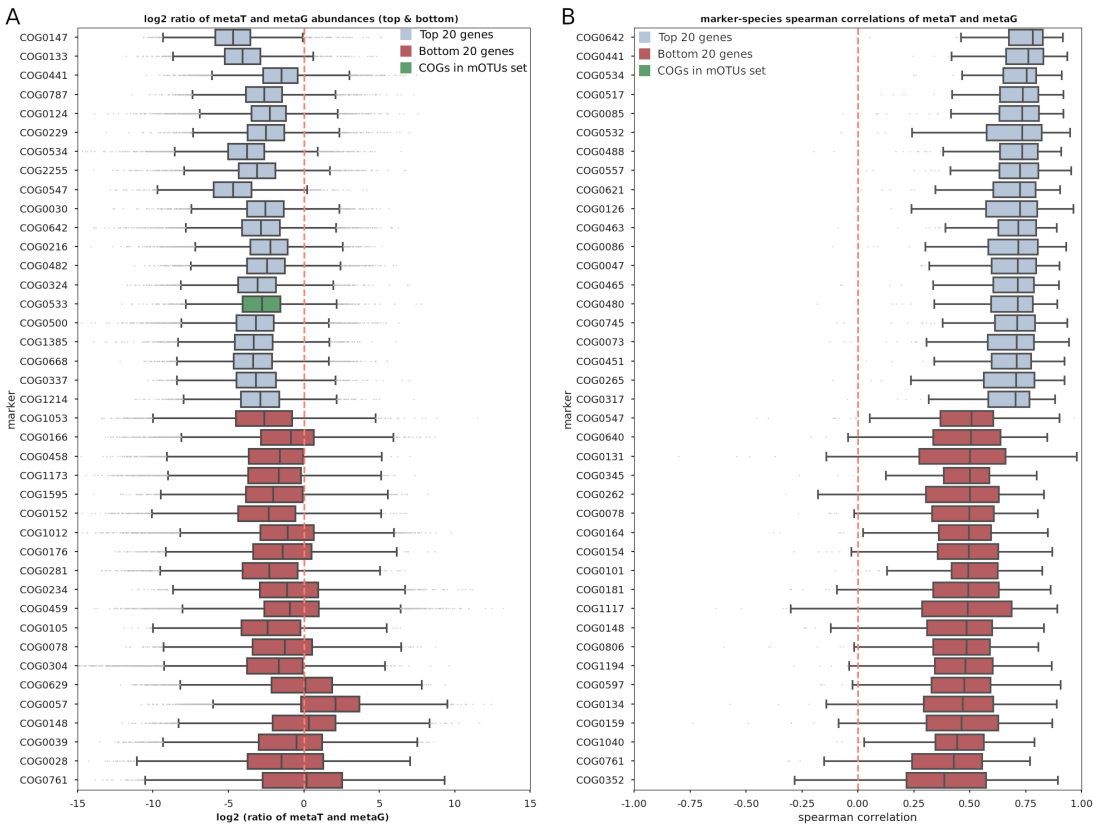

Figure # 3.5: ***Ranked genes in supervised strategies*** *- shown are top and bottom 20 ranking genes in both strategies considered.*
A-Ranked dynamic expression range of conserved marker genes - *plotted is mean log2 of gene abundance derived from metatranscriptomes an normalized by species abundance estimated on metagenomes. The red dashed line represents a mean expression identical to species abundance. Highlighted in green are genes also used by the* mOTU *tool. Genes are ranked by standard deviation.*
B-Ranked spearman correlation of gene and species abundance - *plotted is correlation of gene abundances derived from metatranscriptomes and species abundance estimated on metagenomes*

3.5.3 Unsupervised strategies

Upon realisation that the supervised strategies underperformed, a set of neutral approaches were considered. Using machine-learning and regression models, additional sets of candidate genes were selected and evaluated in context with the previous strategies (see sec. 6.2.6.2).

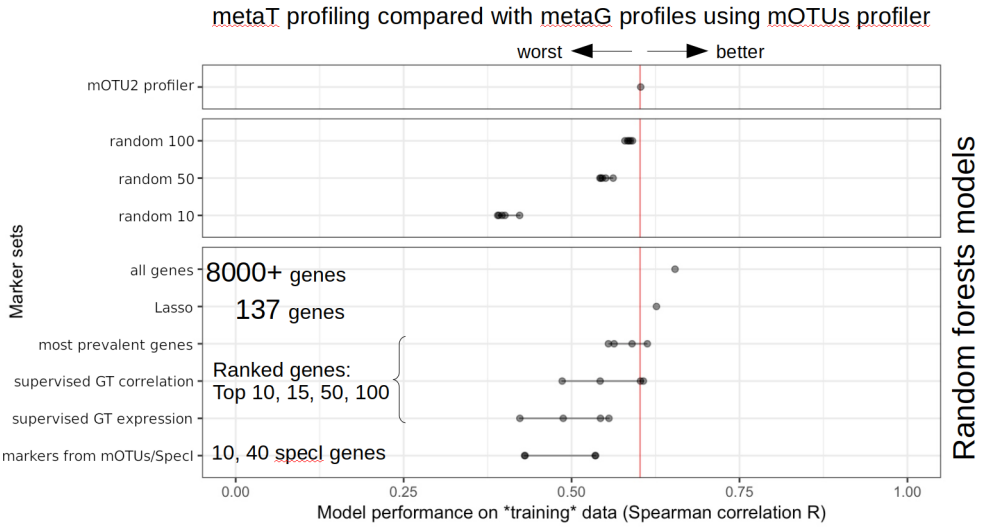

Figure # 3.6: ***Assessment of supervised and unsupervised strategies*** *- shown are models trained on all available data and assessed on the same data (training data). The red line highlights the performance of the mOTUs2 profiler*

In fig. 3.6 we see the summary of all strategies considered. As expected, random sets of markers always performed worse than the *mOTUs profiler* but, selecting a larger number of markers clearly shows an increase in overall performance. In fact, making use of all available genes results in a theoretical maximum performance of this approach. However one should keep in mind that these results are assessed on the training data and as such including all available genes is quite likely to result in over-fitting of the models to the data.

Most promisingly, the LASSO approach displays the overall best performance with a reduced set of marker genes.

Similarly, selecting the most prevalent, that is, the genes most frequently found across all samples, also displays good performance. Once again, this strategy follows the same direction as the supervised approach **2.** in sec. 3.5.2, by prioritizing large families of genes which are more likely to be detected in a wider ranger of species.

Complementing what has been presented in the previous chapter (sec. 3.5.2) we now see that, both supervised approaches display poor performance, with exception of the *correlation approach* (approach 2. in sec. 3.5.2) that the *mOTUs* profiler when the top 50 or 100 genes are used. This result is somewhat expected since, in this particular strategy, both the selection and assessment share the same underlying principle.

Finally, and to some degree surprisingly, the selection of 10 *mOTUs* genes and 40 *specI* underperformed. This reflects a loss of signal due to the fact that *eggNOG v4.5* annotations are used.

The results shown in this figure while assessed on training data already display a concerning pattern.

Unfortunately upon evaluation of these models on test data, by using a cross-validation approach across and within projects, all but one model (*all genes*), underperformed.

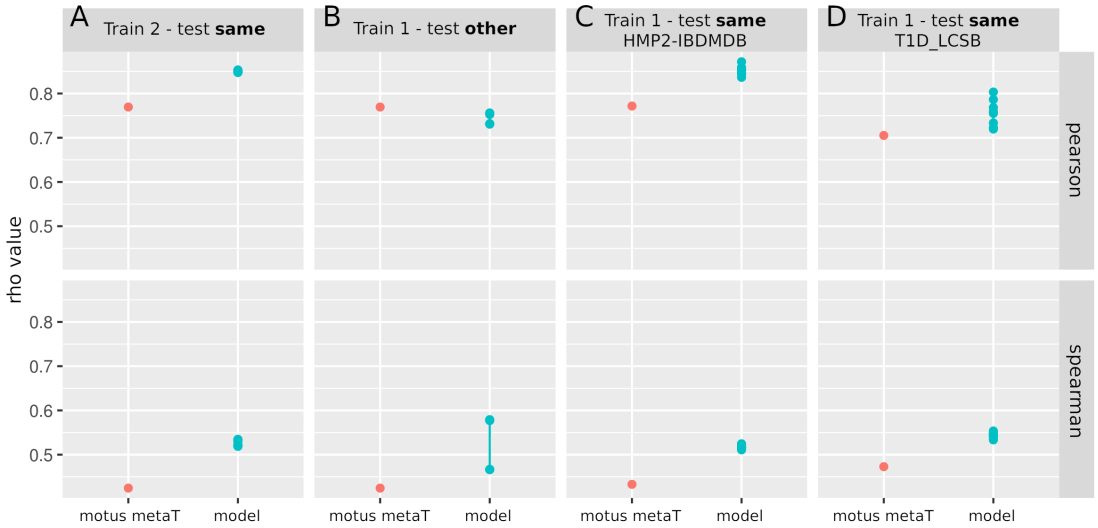

Figure # 3.7: ***Model performance on the two largest datasets*** *- approx. 1000 genes were used to build the models shown,*
A*-model was training on both datasets and evaluated on the same data*
B*-model was trained in one dataset and assessed on the other*
C-D*-models were trained in one dataset and assessed on the same data*

At this point, it was suggested that the approach might be suffering from batch effects or other sources of non-homogeneous variation across all considered datasets. To confirm if this was the case, the largest datasets (`HMP2-IBDMDB` and `T1D_LCSB`), that also displayed a reasonable agreement in sec. 3.3, were used to train and cross-validate a larger model using approximately 1000 candidate genes.

The results shown in fig. 3.7, highlight the underlying problem. While the models show an improvement over the results by the *mOTU profiler*, they generalise poorly to other datasets.

The lack of generalization of the models is a symptom of two likely causes. First, the nature of the datasets in terms of the disease being studied (`HMP2-IBDMDB`-Inflammatory Bowel Disease; `T1D_LCSB`-Type-1 diabetes) and the geographical location of sampling (`HMP2-IBDMDB`-United States of America, North America; `T1D_LCSB`-Luxembourg, Europe). Second, not enough data, something which is known to be a problem with most machine-learning approaches when dealing with hard or noisy data.

If considering the above together with sources of technical noise derived from experimental handling, the problem becomes extremely challenging or even intractable.

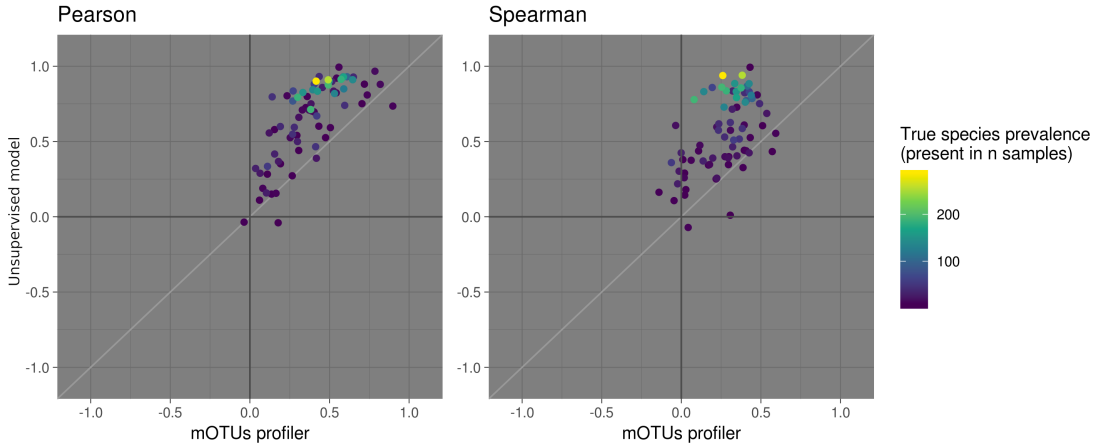

Figure # 3.8: ***Model performance on high prevalence species*** *- shown are 73 species detected in both datasets (`HMP2-IBDMDB, T1D_LCSB`)*

Once again and reassessing the main source of noise, a decision was made to discard low abundance species. As these are too variable across samples, a high rate of absence is likely to confuse and penalize most genes negatively when training the models. By focusing on highly abundant (`>1%`) and prevalent species (73 species detected in both datasets), we managed to improve upon the previous result considerably (see fig. 3.8). In addition to this improvement, the result also shows that prevalence is a clear driver of performance. Prevalent species are consistently better predicted using our model when compared to results from the *mOTU profiler*.

Lastly, this result also supports the argument that more data is required, not just in terms of absolute number of samples, but, more importantly, samples in which the same species consistently detected. In other words, the fact that many species are lowly abundant in many samples introduces uncertainty. This is primarily due to a sampling effect which ultimately affects the final gene counts,

To complement this approach, a wide range of machine-learning algorithms, available in the `mlr` package, were also evaluated. Additionally, samples were partitioned into controls and cases using available metadata, in order to assess the impact of these variables in variation and the performance of the models. No significant improvements over the random forest approach described above were seen. Deep neuronal networks were also considered but were found to mostly over-fit the data, something which, once again, supports the argument in favor of the need of more data. Finally, a marker gene selection strategy using a genetic algorithm was also tested but the result, once again, did not yield significant improvements.

In all, the results presented in this section revealed that the approach of estimating species abundance using only metatranscriptomes is viable but requires larger datasets in order to build gen-

eralizable models. On the other hand, in light of this work it is possible that for sufficiently large projects, the number metagenomes required for adequate abundance normalization is reduced, having a desirable economic impact.

3.5.4 Selected orthologous groups

After the findings in sec. 3.5.2 regarding the properties of the selected marker genes, I decided to investigate these same properties in the final selection of genes from the previous chapter.

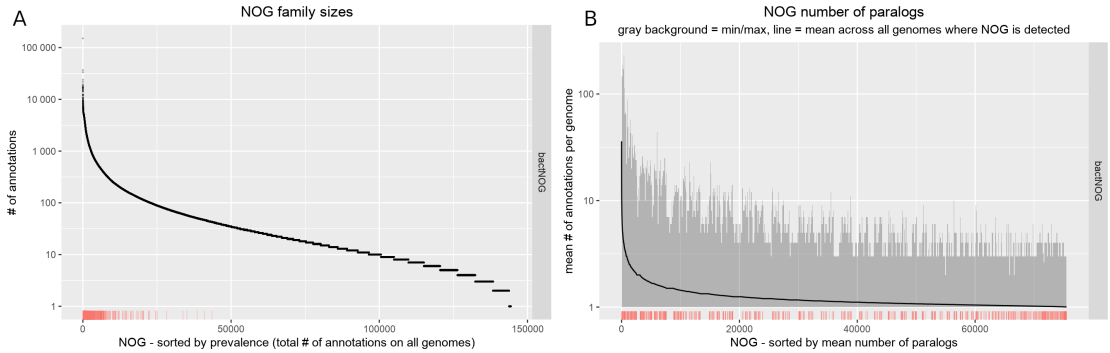

Figure # 3.9: ***Properties of marker genes (NOGs) selected through unsupervised approaches*** *- red vertical lines represent marker genes selected through the final unsupervised strategy (~1000 genes)*
A-Number of orthologous group occurrences in reference genomes - *shown are the total number of annotations of each NOG in the reference genomes considered.*
B-Number of paralogs in orthologous groups - *shown are the mean number of annotations for the same orthologous group (paralogs) in each reference genome.*

Shown in fig. 3.9 - panel **A** are the sizes of gene families considered in the study sorted by the number of times they are identified in the reference genomes. Here we see that most selected genes originate from prevalent gene families, that is, families with members present in most species considered. Surprisingly, in fig. 3.9 - panel **B**, the final selection of genes covers a wide range of families with variable paralogy sizes. This result is somewhat unexpected and contrasts with the selection of genes from supervised strategies (sec. 3.5.2), where all selected genes corresponded to families with high number of paralogs.

3.6 Discussion

Given the results shown in this chapter, we see that the problem of normalization of metatranscriptomes and in particular deconvoluting species abundance and gene expression is a difficult challenge. The many strategies outlined revealed that this approach is viable but larger samples sizes are required to properly calibrate adequate and generalizable models.

As shown, predicting lowly abundant or lowly prevalent species is specially challenging due to lack of data. Similarly, when focusing on high prevalence species the problem becomes considerably easier. Regardless, even when enough data is available, predictions are still subpar,

reflecting the inherent variation in this data. This aspect became clear when comparing projects from different backgrounds (disease and geography). Models were able to capture project specific traits manifesting their underlying biology.

Another aspect which was initially considered but not revisited is that of the quality of the reference genomes. The effect of genome fragmentation on the final estimates was not assessed but will certainly correlate with poor performance and abundance estimates.

A key aspect and weakness in this work is the lack of a *true* species abundance or, at the very least, a second source of such information. This limitation presents itself in two forms. On one hand, the entire approach relies on good pairing of metagenomes and metatranscriptomes. Specifically, that underlying technical and biological variation between the two outputs will be shadowed by the real species abundance. On the other, the species abundance used as reference is determined by what was considered the best metagenomic profiler available. As consequence, the inherent inaccuracies of the profiler will persist as part of the trained models unless better abundance estimates are possible.

In what concerns technical and biological variation, one should also consider the sampling procedure and how it affects the sample. For instance, exclusive anaerobic species are likely to see larger deviations than facultative species, due to exposure to oxygen during sample collection. As seen in some of the projects, the agreement of metatranscriptomes and metagenomes is often poor. One could argue that the difference is due to biological causes but it is unlikely that such high variation is seen if samples are collected with minimal atmospheric exposure and rapidly preserved in liquid nitrogen. Despite considered best practices in the microbiome field, these steps are rarely followed due to practical reasons. Sampling with such rigor requires a controlled environment, not always available, and a procedure that minimizes the risks to the donor, often an ethical concern. In all, while ideal, technical variation will remain an obstacle and, unless controlled for, an unknown source of variation.

With regards to alternative sources for *real* abundances one could consider two options. First, alternative software based on different approaches. This would be viable if the bacterial species present in the community were consistently the same across samples, something which is known not to be the case. Second, alternative readouts from the same community. While this aligns with the motivations behind multi-omics studies, not many publicly available projects have followed this direction. The few examples of such practice often focus on obtaining readouts from different sources (genome, transcripts, proteins, metabolites), rather than multiple from the same source.

In the following chapter I will present additional work in this direction following a multi-omics approach as an attempt to control for some of the limitations mentioned above.

4

Small scale community dynamics

4.1 Introduction

Most microbiome studies focus on analysing complex communities with many different species, often poorly characterized. This complexity is also reflected in the interpretation and analysis of the data by creating confounding effects due to unexpected or uncontrolled interactions.

Many studies have come to the conclusion that in order cope with this complexity and to obtain mechanistic understanding, experimental designs targeting concrete questions are required. Such designs, while still far from simple, control for several aspects that are difficult to address with *in vivo* models. For instance, while mice are an often studied model organism, they naturally display coprophagic behavior that poses challenges to the interpretation of their gut dynamics. This is specially true when trying to establish parallelisms with human or other non-coprophagic organisms. As such, using *in vitro* systems allows, not only to have tightly controlled environments, it enables designs targeting specific questions often inaccessible *in vivo* due to ethical or practical reasons.

I this chapter I will present a pilot experimental *in vitro* setup using artificial communities designed to mimic the human gut environment. Although primarily focusing on the challenge of normalizing metatranscriptomes, this experimental setup also addresses practical aspects affecting most microbiome studies. Additionally, in order to avoid developing methods and

approaches that only apply to artificial scenarios, and to allow a broader understanding of transcriptomic variation in different conditions, the experimental setup was designed to include perturbations by addition of human targeting drugs. Given their design targets, these drugs are not expected to have an effect on the community. However, recent studies (72) have shown that this is not the case, revealing interesting dynamics and unforeseen mechanisms of action for different human drugs. Specifically, together with genetic and phenotypic characteristics of host, the microbiome is likely to play a role in the plasticity of the response to a drug treatment.

By pursuing such experimental setup in a controlled environment, sampling and handling biases are minimized, creating the ideal conditions to obtain metatranscriptomic and metagenomic readouts with a much desired, high quality and agreement.

4.2 Designing a controlled experiment

As seen in the previous chapter, the need for better control over experimental variables, and an additional source of abundance estimation, is required to properly calibrate and assess the approach previously outlined.

For this purpose a pilot experimental setup was designed using bacterial species commonly found in the human gut and for which medium-high quality genomes were available. These species were grown together to allow for interactions and to reach a stable state, at which point they were perturbed by the addition of a drug (details in sec. 6.3).

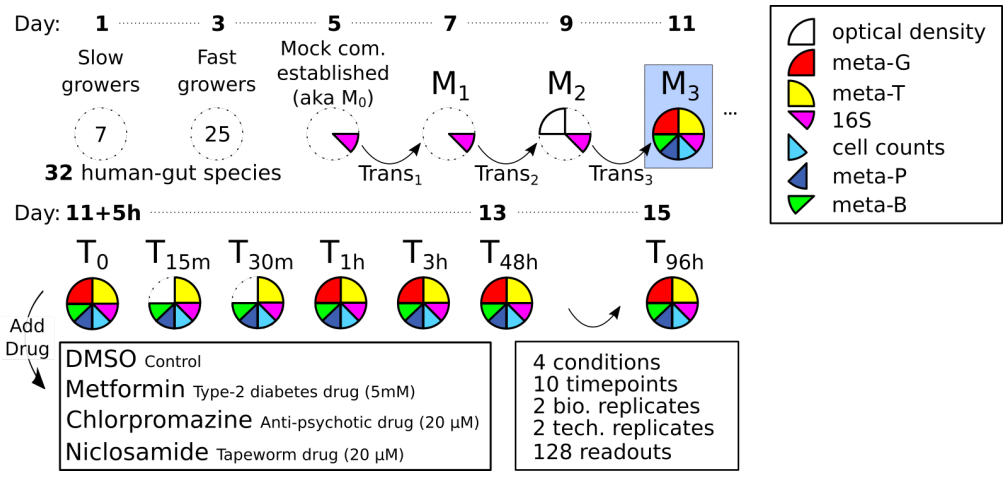

Figure # 4.1: ***Mock experiment design, community stabilization, drug perturbations, timepoints and data types collected***

The experimental setup, conditions, timepoints and datatypes collected is illustrated in fig. 4.1 and described in detail in the methods (sec. 6.3).

The inclusion of perturbations but use of drugs, allows not just studying how these affect the

community, but also increasing variability across samples to cover a wider ranger of potential expression. Similarly, having multiple close timepoints after exposure, allows inspecting how the community is reacting and rearranging its expression to cope with the perturbation.

While the experiment was designed with a multi-omics approach in mind, this work focuses only on 16S rRNA readouts, metagenomes and metatranscriptomes.

During the execution of the design, several obstacles of technical and practical nature were encountered.

First, *RiboZero*, the kit used to deplete rRNA and tRNA in all projects included in this study, and mentioned in several standard metagenomics protocols, was found to have been discontinued as of 2018 and no alternative from the same company was available (August 2019 (*139*)). With an unclear justification, the absence of this kit will, yet again, introduce changes in standard protocols that will likely complicate integration attempts of existing and future projects.

Second, to our surprise we realized that gene prediction and annotation was not a solved problem, at least for some of the species considered in this study. As having accurate gene predictions is a critical starting point for metatranscriptomic analysis, this point was further investigated. Results are discussed below.

Lastly, while considered essential for the goals of this work and included in the design, it was not possible to obtain reliable measurements for cell counts.

4.3 Ribosomal RNA depletion challenges

With the discontinuation of *RiboZero* an alternative kit had to be considered. As available kits are typically designed to target a subset of species, we were approached by New England Biolabs (NEB) which offered the possibility of using a new kit under development, targetting a wider ranger of species. Due to its development state and thanks to knowing the exact species in the community, the efficiency of depletion was assessed for every species and sample.

In fig. 4.2 we can see an overview of the rRNA depletion efficiency. As expected the depletion was not homogeneous across species. This primarily reflects the mechanism of action of this kit, that uses specific DNA probes to target rRNAs and degrade them, by use of an enzyme that cleaves DNA/RNA hybrids.

Notably, as the basis of this kit was constructed from *Escherichia coli*, depletion of this species was almost 100% efficient. In contrast, species like *Bacteroides vulgatus*, *Bifidobacterium adolescentis*, *Clostridium perfringens* and *Dorea formicigenerans* had considerably worse depletion efficiency (fig. 4.2 and also, fig. A.3, fig. A.4, fig. A.5).

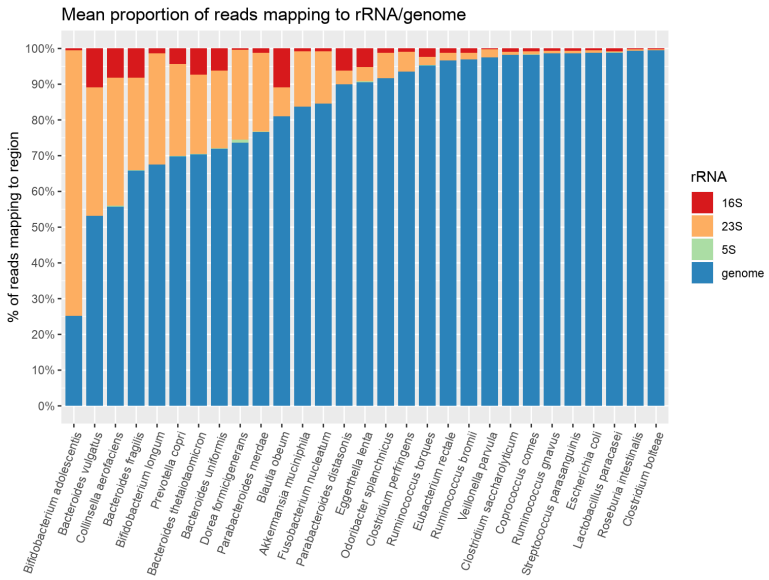

Figure # 4.2: ***Overall rRNA depletion efficiency***

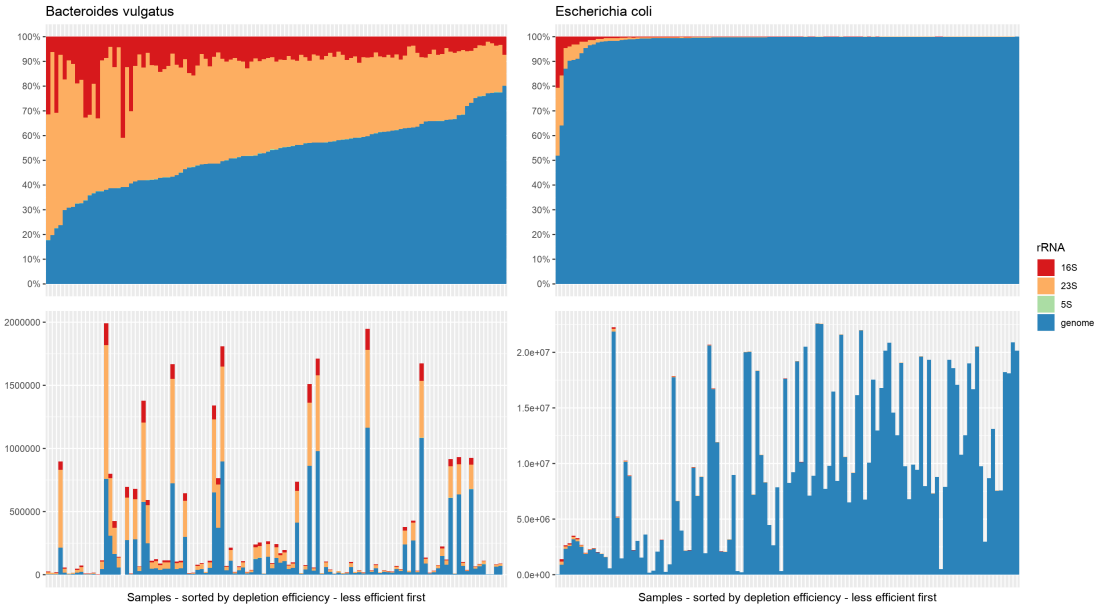

Figure # 4.3: ***B.vulgatus and E.coli rRNA analysis after depletion*** *- data shown for all samples*

Finally, and also surprisingly, the efficiency of depletion across samples is quite variable fig. 4.3, something which may be caused by chemical interactions with reagents from extraction, or more likely, variable RNA qualities and degradation (see fig. 4.3).

4.4 Identifying genes in metagenomic samples

As part of the preparation of the reference genomes for metatranscriptomics analysis, an unusual annotation pattern was noticed. Annotations originating from *RefSeq* where quite variable across species, both in terms of number of predicted genes and the ratio of coding to non-coding bases. This motivated the use of an alternative method that would provide consistent annotations across genomes.

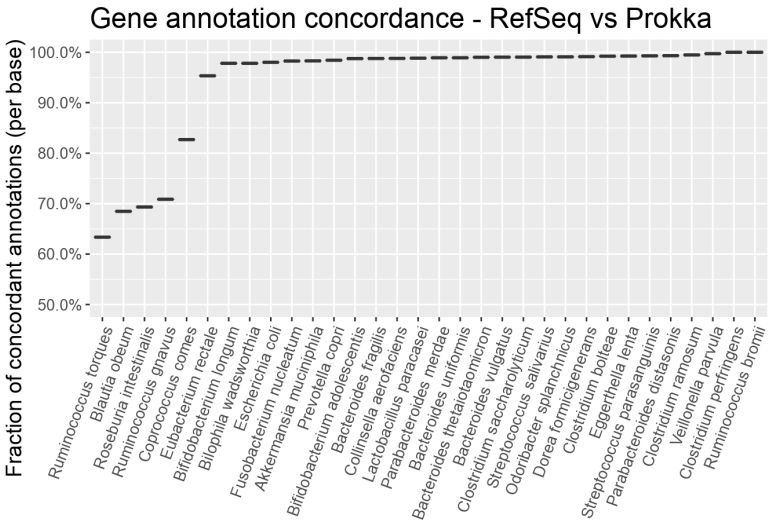

Figure # 4.4: ***Agreement between gene predictions*** *- gene predictions from Prokka and RefSeq were compared by counting the number of bases that were predicted as coding by both approaches. Species are sorted by ascending order of agreement*

Following this finding, *Prodigal* (*50*) was used to predict genes in all 32 genomes (see sec. 6.3.9). Results are displayed in fig. 4.8, where we can see that a few genomes showed severe disagreement between the two methods.

In order to avoid skewing results on these genomes, *Prodigal* gene predictions were used throughout the remaining of this analysis.

Despite consistent use of gene predictions, once reads were mapped against the genomes, it was found that a significant proportion originated from anti-sense transcripts (see fig. A.2). This phenomenon is known in bacteria and speculated to serve regulatory purposes by an RNA silencing mechanism. However, the trigger for anti-sense expression and how this process is regulated is not well understood. No correlation was seen between variation of anti-sense expression and

the drug treatments, suggesting the absence of a regulatory response induced by the drugs.

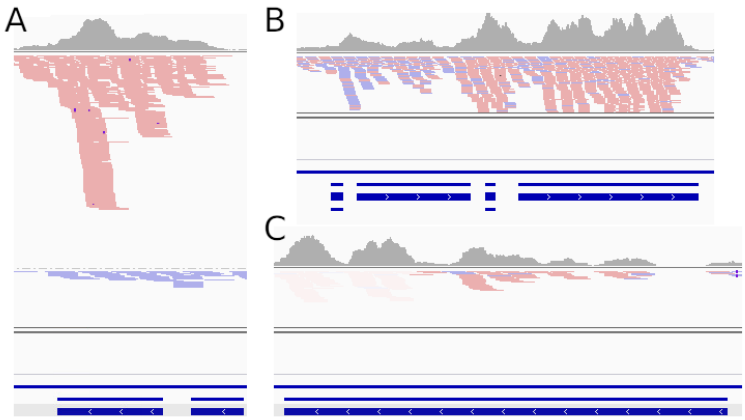

Figure # 4.5: ***Overlap between gene predictions and mapped reads*** *- reads mapped against reference genomes do not always reflect predicted gene annotation. Red and light blue lines represent reads mapping to the forward and reverse strand, respectively. Thick blue lines define predicted genes. The profiles in gray show read coverage.*

Finally, the overlap between reads and predicted ORFs is also frequently inconsistent. While this may hint at poor gene predictions, the appearance of peaks and troughs (fig. 4.5) in coverage profiles, is known to be primarily caused by sequencing bias and related to thermodynamic aspects of RNA, including the formation of hairpins and other secondary structure elements.

4.5 Species abundance through different approaches

In continuation of the work presented in the previous chapter, abundance estimations were performed using three of the readouts from the experiment.

Results are shown in fig. 4.6 for the two biological replicates, where we see that both biological runs are mostly consistent across samples and conditions, displaying only significant variation in the last time-points. In the figure a dominance of blue and green colors is visible corresponding to *E.coli* and *C.perfringens*, respectively. These two species overtook the community in the initial time-points, possibly due to better efficiency in replication or pathways that make use of nutrients readily available in the medium. Similarly, *V. parvula* only bloomed at later time-points reflecting its known secondary metabolizer role (*32*, *106*).

In global terms, community mostly recovered its initial state in the latest time-points (see fig. 4.7). One exception is the *chlorpromazine* treatment, that displayed the largest response to the drug, and where *Bacteroides*, *Parabacteroides*, *Bifidobacterium* and *Fusobacterium* were almost entirely removed from the community. Additionally, and proportionally to the rest of the community, the species *A. muciniphila* and *P. merdae*, were also significantly reduced in the *niclosamide* treatment. Last but not least, the *metformin* treatment had no noticeable effect on the community. This is consistent with the results found by *Maier et al.* (*72*) and, to some degree, a contrast to *Forslund*

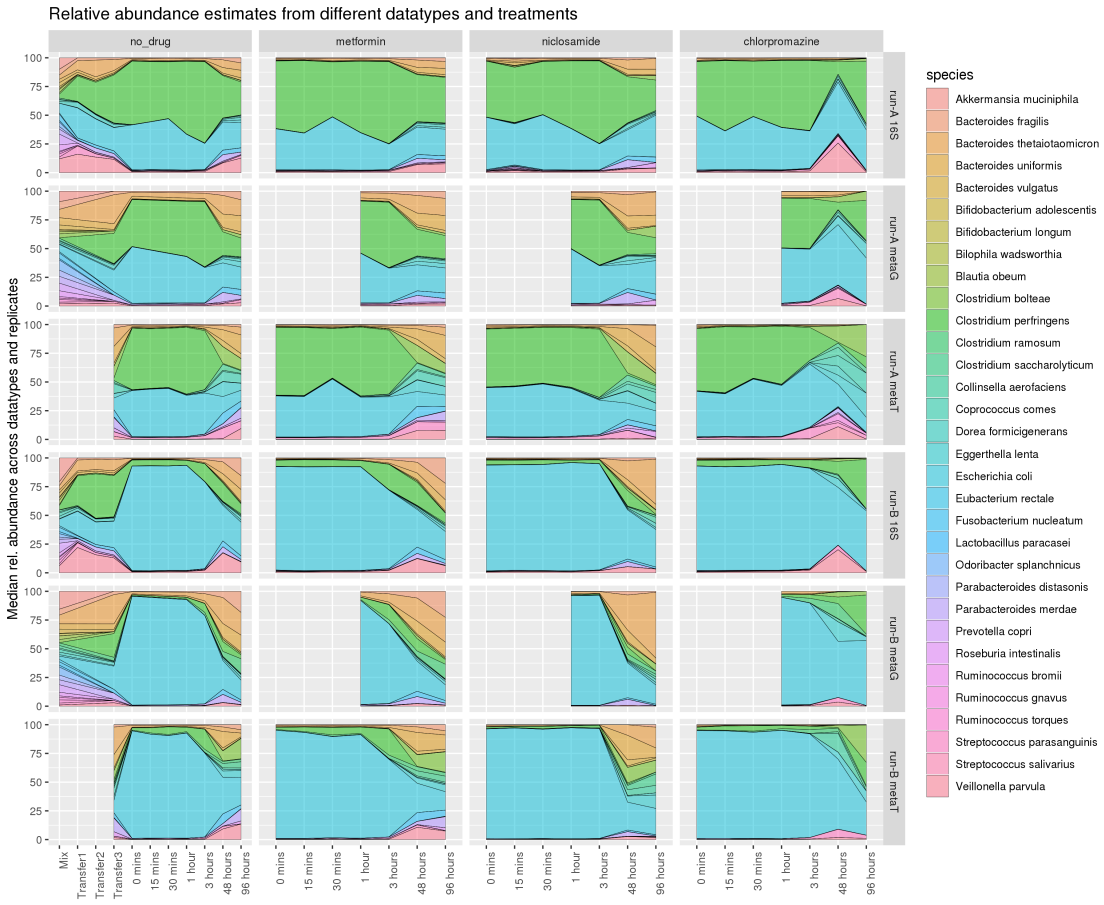

Figure # 4.6: ***Relative abundance estimated using different readouts*** *- shown are two biological replicates (runA and runB) and abundance estimates under in control (no_drug) and different drug treatments (metforming, niclosamide and chlorpromazine)*
16S*-abundance estimated by mapping reads to a database of 16S rRNA regions matching the strains in the study*
metaG*-abundance estimated by mapping reads to complete genomes*
metaT*-abundance estimated by mapping reads to complete genomes after excluding rRNA regions and reads*

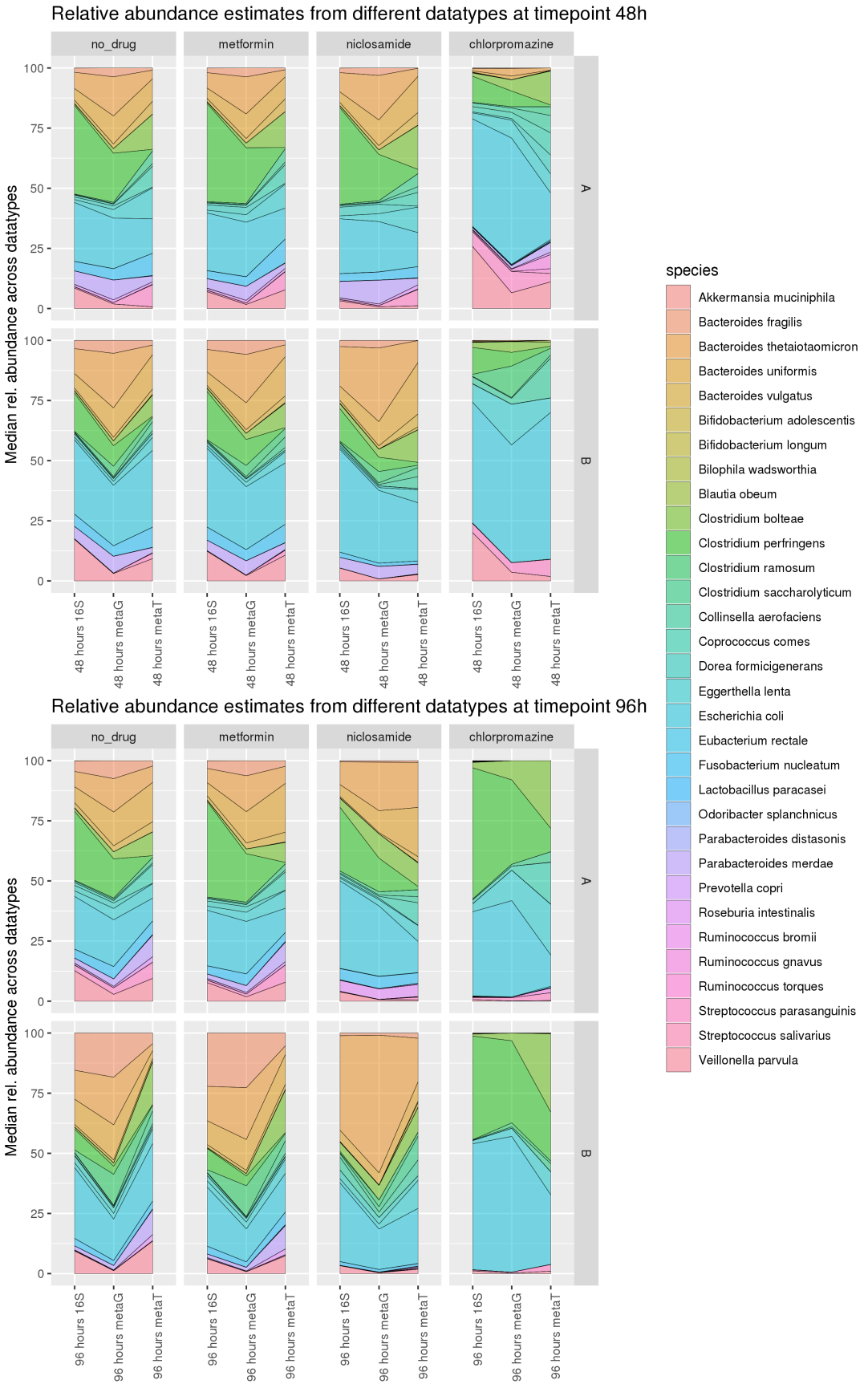

Figure # 4.7: ***Fraction of metatranscriptomic reads mapped to entire genome*** *- reads were mapped against the genome of each species and normalized by its size. Ribosomal regions were excluded from the analysis*

et al. (*34*) where an effect was measurable. This result likely reflects an effect driven by the host rather than the community. *Metformin* is a drug that regulates sugar levels in the blood. Consequently, reduction of the availability of sugar is a more likely driver of the changes described in the work by *Forslund et al.* (*34*). On top of this, due to the mode of action of this drug, the effect is unlikely to be reproducible in artificial communities where no host-mediated sugar regulation occurs.

An interesting result, clearly visible in fig. 4.7, is the disagreement between estimates from metagenomes (metaG) and metatranscriptomes (metaT). Noticeably, species that have a low representation in the community (by *16S* and *metaG*) have high expression levels (*metaT*), or vice-versa. Such are the case of *C.bolteae* that is low abundant in early time-points and comparatively, highly active in the final time-point, and *C.perfringens* that although highly abundant, shows almost no expression at later time-points, reflecting what seems to be a stationary or dormant phase.

Finally, the different relative abundance estimates produced in this section will be used in the following to further validate the deconvolution of metatranscriptomic composition and gene expression.

4.6 Revisiting abundance estimation from metatranscriptomes

Revisiting the work in the previous chapter, now applied to the metatranscriptomes from the artificial communities, we can now make use of the additional readouts as alternative reference abundance estimates.

The results are shown in fig. 4.8, where predictions with the same 10 marker genes used by the *mOTU* profiler, perfectly recapitulate the abundance derived from metagenomes using the *mOTUs* profiler. In contrast, when using the 40 *specI* markers (*79*), an increase in dispersal of estimates becomes visible, something which is equally true for models that include a larger number of genes.

These results, intuitively, suggest that the output provided by the mOTUs profiler on metatranscriptomes reflects the expression of the 10 marker genes and that these perfectly recapitulate the species abundance. Such striking result is unsettling when considering results obtained from publicly available datasets. While having the exact genomes for the species included in the study certainly would help improving the overall approach, this by itself is not enough justification.

Reasoning on the possible causes for such an excellent result, a few arguments can be presented. On one hand, the species considered for the artificial community are common and prevalent gut members with available reference genomes. As such, enough data is available allowing for the

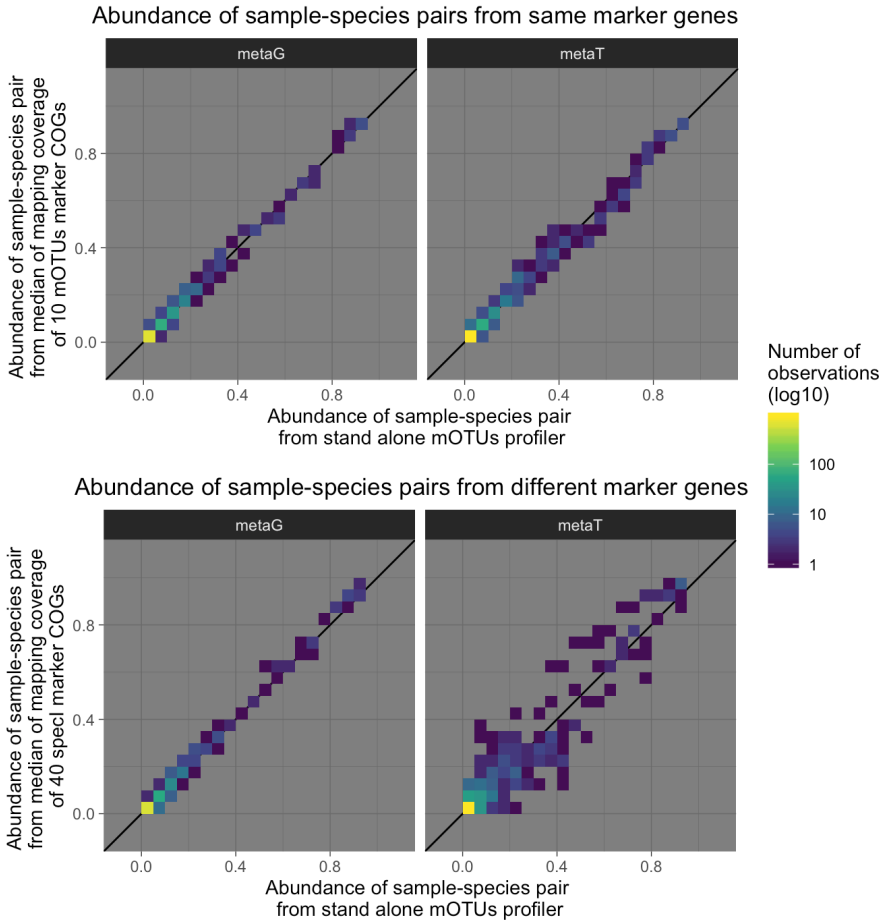

Figure # 4.8: ***Performance of model built using the 10 mOTUs and 40 specI markers*** *- shown are abundance estimates on both metagenomes (metaG) and metatranscriptomes (metaT) plotted against reference abundances generated with mOTUs-profiler on metagenomes*

mOTUs approach to include better representations of the sequences of these species and consequently perform reliably. More importantly, the fact that this artificial community was grown and sampled in a controlled environment, following all best-practices in terms of preservation and extraction of DNA and RNA, reinforces the argument that most discrepancy seen is due to technical noise, introduced during or after sampling.

Whereas the outcome of this experiment is strikingly positive, a few aspect will benefit from further investigation. Primarily, assessing if species beyond those included in this artificial community display equally good agreement when profiled with mOTUs when sampled in a controlled environment. And subsequently, if the variation seen in public data is somehow revealing of a biological state of the gut environment. In other words, that the gut, and particularly the colon, is in constant state of change, not allowing for metagenomic and metatranscriptomic readouts to agree with the same level of exactitude seen in this artificial community.

4.7 Discussion

In line with what was discussed in the previous chapter, the results outlined here are supportive of poor technical handling of the stool samples. While this is less of a concern for metagenomes, metatranscriptomes are very sensitive due to their dynamic nature, consequently leading to discrepancies.

In light of these results, it is fair to argue that, the approaches discussed before were effectively capturing primarily technical noise. This is further supported by the fact that the models while being trained failed to generalize across datasets, something which is to be expected if assuming that experimental bias is a major contributor to the variation in the data.

In summary, I am inclined to support the 10 marker genes used by the mOTUs profiler as the best candidates to achieve the objective proposed for this work, and to defend that better experimental practices are required when targeting metatranscriptomics.

Other issues were presented in this chapter. Gene prediction and its overlap with sequenced reads was assessed, revealing a reasonable effect introduced by the sequencing platform. This is a widely known problem affecting short-read sequencing platforms and often disregarded. Newer long-read sequencing platforms are known to not be as dramatically affected by this problem and may prove to be a more reliable option.

The effect of the drug perturbations was evaluated both in terms of regulation of anti-sense transcription and overall composition of samples. A few species were found to be affected by the drugs included in this study. In particular, *chlorpromazine* showed a dramatic impact in the community, with several *genera* being greatly reduced. While the modes of action of these drugs

on bacteria are not known, the fact that *chlorpromazine* is an antipsychotic drug poses interesting questions with regards to the relation between the microbiome and mental health.

Finally, while a valuable multi-omics dataset was produced as the result of the experiment herein described, many directions remain unexplored. The analysis of this data will continue beyond the work presented here and will certainly provide interesting biological insights about this artificial community and its response to the different drug perturbations.

5

Conclusion

In this work, I presented an overview of methods and their application to metatranscriptomics studies.

In light of the objectives defined for this project, the endeavour was mostly successful. A survey of available tools was performed and standard approaches to analyse metagenomic data were outlined. Two products (*NGLess* and *GMGC*) were possible thanks to the many contributions during this period. Similarly, while its robustness is debatable, a strategy to normalize and analyse metatranscriptomic data was presented and its limitations discussed. Last but not least, these strategies were evaluated in an artificial bacterial community, revealing surprising results that strongly suggest a significant contribution from technical noise to the variation seen in the data.

In short, the main messages to be taken are that metatranscriptomic data is noisy and careful analysis is required. Several limitations of current approaches were highlighted and both technical and methodological concerns were expressed. As such, functional studies making use of this type of data should be interpreted with caution, particularly in what concerns technical handling, experimental controls and best-practices. More importantly, when reasoning over metatranscriptomic results from a functional point of view, one should be careful not to interpret functional alterations due to experimental handling as biologically relevant for the condition being studied.

During the realization of this work, many analysis were performed and many more were considered. In its present form, it is fair to say that, time was its biggest obstacle and that, although reaching the proposed objectives, a definite and final answer to this topic has not been reached. Several analysis could have benefited from additional thoroughness and rigor but, regardless, this thesis stands as a proof-of-concept and paves the way to future efforts in the same or similar directions.

6

Materials and Methods

6.1 Comparison of taxonomic profilers

6.1.1 Assessment of mOTUs1 and MetaPhlan2

The two tools were initially assessed using simulated metagenomes (see tbl. 6.1) from the first iteration of CAMI. Metagenomes, a reference taxonomy database and the expected result after profiling (*gold standard*) were downloaded from the official CAMI website (https://data.cami-challenge.org).

Table 6.1: Characteristics of the CAMI test datasets

Complexity	Base-pairs	Genomes	Circular elements	Number of Samples
Low	15Gbp	40	20	1
Medium	40Gpb	132	100	4
High	75Gbp	596	478	5

All simulated metagenomes were profiled with both *mOTUs1* and *MetaPhlAn2*. Results were converted to comply with the CAMI format and assesed against the *gold standard* using the provided scripts. As the CAMI format requires the presence of National Center for Biotechnology

Information (NCBI) taxonomic identifiers and their entire lineage, the output of each tool was remapped to the same reference taxonomy included as part of the CAMI challenge. Since no NCBI taxonomy identifers are present in the output of *MetaPhlAn2*, a name matching strategy was used. This strategy is error prone and aggravated by the fact that *MetaPhlAn2* masks certain characters from the species names. The performance of both tools is possibly affected by the lossy nature of this identifier translation.

6.1.2 GMGC creation

The procedure to create the GMGC is described in *Coelho et al.* (*22*).

In brief, metagenomes and metatranscriptomes were preprocessed and quality controlled using `NGLess` (*21*) and individually assembled into contigs using `MEGAHIT` (*62*). ORFs were predicted from contigs using `MetaGeneMark` (*135*) and clustered into groups of genes using a custom graph approach (*22*). Samples were then mapped to the clustered genes using `minimap2` (*64*) and rarity of genes was assessed using coverage as a proxy. Genes were 6-frame translated to proteins using `fna2faa` (*6*) and functionally annotated with `eggnog-mapper` (*45*). Taxonomic assignments for each gene were obtained by using a *Last Common Ancestor* approach (*22*) on sequence similarity results obtained by using `diamond` (*15*) against `UniRef90` (*112*) and `UniProt+TrEMBL` (*140*).

6.2 Integration of metagenomes and metatranscriptomes

In order to identify comparable public human gut datasets containing metagenomes and metatranscriptomes, a literature survey was performed and projects meeting the following requirements were considered:

- Samples preserved and conserved with chemical agent (e.g. RNALater) and/or temperature (-20°C/-80°C)
- Simultaneous DNA/RNA extraction (paired extraction)
- rRNA and tRNA depletion (e.g. RiboZero)

A total of 5 projects were found to meet these requirements in addition to a few samples generated in the Bork group (see tbl. 6.2).

Table 6.2: Project identifiers and publications used in the study

Download source	Alias	Samples	Publication
PRJNA188481	Franzosa_2014	8	(*37*)
PRJNA389280	HMP2-IBD	78	(*100*)
PRJNA398089	HMP2-IBDMDB	761	(*70*)

Download source	Alias	Samples	Publication
PRJNA354235	HPFS-MLVS	372	(*77*)
PRJNA289586	T1D_LCSB	88	(*43*)
Generated in-house	InternalGT	9	

6.2.1 Data acquisition

Public data was downloaded from European Nucleotide Archive (ENA) between February and June 2018. The project `HMP2-IBDMDB` was originally downloaded from `ibdmdb.org`, but most of its data has since been made available on ENA under the accession number `PRJNA398089`.

6.2.2 Reference genomes

Bacteria and Archaea reference genomes were obtained from proGenomes `v1.0` (*78*), containing 5510 genomes. In order to further improve the quality of the genomes used, contigs smaller than 100 base-pairs (bp) were excluded and only species associated with human body sites were considered. Body site associations were curated from Patric database metadata (*124*). In total, 871 genomes were used for subsequent analysis.

6.2.3 Orthology and functional annotations

Orthology and functional annotations for all genomes in proGenomes `v1.0` were generated using an early development version of eggNOG-mapper (*45*) using NOGs as defined in eggNOG v4.5 (*46*).

6.2.4 Sequence processing using NGLess

Quality control of raw reads was performed using NGLess (*21*) and the script lst. 6.1. After this step, reads shorter than 45bp were discarded. Reads that passed quality control were subsequently mapped against a human reference database at a 90% identity threshold. Reads mapping more than 45 contiguous bases at this identity cutoff were discarded. The human reference contains a primary assembly of the human genome (release `GRCh38.p10`), cDNA sequences according to Ensembl gene predictions and sequences corresponding to 45S ribosomal clusters, all of which downloaded from the Ensembl project repository (*134*).

Once quality controlled, reads were mapped against the filtered database of reference prokaryotic genomes using the NGLess script lst. 6.2 and the default BWA-MEM (*65*) mapper together with the option to report all hits (`mode_all=True`).

Listing 6.1 NGLess code used to quality control raw reads and remove human contamination

```
ngless "0.8"
import "parallel" version "0.6"
import "mocat" version "0.0"

samples = readlines(ARGV[2])
sample = lock1(samples)
input = load_mocat_sample(ARGV[1] + '/' + sample)

input = preprocess(input, keep_singles=True) using |read|:
    read = substrim(read, min_quality=25)
    if len(read) < 45:
        discard
mapped = map(input,
             fafile='<references>/Homo_sapiens.GRCh38.p10.cdna+dna+45S.fna')

mapped = select(mapped) using |mr|:
    mr = mr.filter(min_match_size=45, min_identity_pc=90, action={unmatch})
    if mr.flag({mapped}):
        discard

write(as_reads(mapped),
      ofile='outputs/' + sample + '/' + sample + '.filtered.fq.gz')
collect(qcstats({fastq}),
        ofile='outputs/preprocessing_fqstats.txt',
        current=sample, allneeded=samples)
```

The choice of a 97% mapping identity threshold (*79*) was validated in parallel by mapping reads at a minimum identity threshold of 95% and subsequently filtering in increments of 1% up to 100% identity (see fig. 3.4).

6.2.5 Counting reads overlapping regions of interest

NGLess was once again used together with BAM files produced in the previous steps and General Feature Format (GFF) files containing eggNOG orthology annotations to generate gene count tables from metatranscriptomes.

Several variations of counts were produced, including different strategies to handle reads mapping to multiple regions (*multiple-mappers*), normalization based on size of the genome, features being counted as well as eggNOG annotations defined at different taxonomic resolution levels.

6.2.6 Selection of candidate genes for normalization

6.2.6.1 Supervised approaches

Two strategies were used to rank and select genes based on empirical knowledge. Ideal candidate genes for normalization were considered to meet a set of criteria *a priori*. The assumptions and formulation used are:

Listing 6.2 NGLess code used to map against reference genomes

```
ngless "0.8"
import "parallel" version "0.6"
import "mocat" version "0.0"
import "samtools" version "0.0"

samples = readlines(ARGV[3])
sample = lock1([ARGV[1]])
input = load_mocat_sample(ARGV[2] + '/' + sample)

mapped = map(input,
             fafile="<references>/v11-v2-rep-v2UL-contigs-min100-human_gut.fna",
             mode_all=True)
mapped = select(mapped) using |mr|:
    mr = mr.filter(min_match_size=45, min_identity_pc=97, action={drop})

write(mapped, ofile='outputs/' + sample + '.human_gut-v11UL.bam')

sorted = samtools_sort(mapped, by={name})
write(sorted,
      ofile='outputs/namesorted/' + sample + '.human_gut-v11UL.namesorted.bam')
```

Listing 6.3 NGLess code used to count reads overlapping regions of interest

```
ngless "0.8"
import "parallel" version "0.6"

samples = readlines("all.samplefile")
sample = lock1([ARGV[1]])
FEATURE = ARGV[2]  # one of "bactNOG", "arNOG"
input = samfile('outputs/namesorted/' + sample + '.human_gut-v11UL.namesorted.bam')

counts = count(input, features=['eggnog45'], subfeatures=[FEATURE],
               multiple={unique_only},
               gff_file="<references>/v11-v2-rep-v2UL-(…)_annotations-v2.gff")
collect(counts, ofile="outputs/feature_counts.tsv",
        current=ARGV[3], allneeded=samples)
```

1. Display small and stable dynamic range of expression - $log2(metaT/metaG)$
2. Display good correlation with species abundance - $spearman(metaT, metaG)$

where $metaG$ represents the abundance of each species as estimated by the *mOTUs2* profiler (*82*) on the paired metagenome and $metaT$, the expression of each gene (on the paired meta-transcriptome) belonging to the same group of orthologs (*bactNOG* level as of `eggNOG v4.5`). Genes were sorted according to these measures and their performance assessed by comparing the mean gene abundance across all candidates.

Additionally, as the *mOTUs* approach uses 10 marker genes, a subset of an initial larger set of 40 universal markers (*specI* markers) (*79*), both the 10 and 40 marker list was re-evaluated, using the same counting strategy outlined above.

6.2.6.2 Unsupervised approaches - Machine learning models

Following the supervised approaches described in sec. 6.2.6.1 several unsupervised approaches were considered with different degrees of stringency. Genes with less than 100 non-zero observations were excluded and only species with average abundance above 0.1% were considered.

A LASSO regression was used to select a small and representative list of candidates. Random sets of genes with cardinality 10, 50, 100 and 144 (same number as LASSO approach) were also selected and evaluated together with the previous approaches.

Random forest models were trained on the selected sets of genes in order to tune the coefficients for each marker. Initial models were 6-fold cross-validated across project and repeated 4 times.

The final evaluation was performed on only the two largest datasets (`HMP-IBDMDB` and `T1D_LSCB` - see tbl. 6.2), the same that displayed better metagenome - metatranscriptome aggreements (see fig. 3.3). Random forest models were 2-fold cross-validated across and within project and the process repeated 4 times.

This analysis was implemented using the `mlr` package (*12*) and the `R` language (*95*). Plots were generated from derived models and results using `ggplot2` part of `tidyverse` (*30*, *125*).

6.3 Artificial gut communities

6.3.1 Species selection and reference genomes

The selection of species used in this study represents a subset of those most abundant and prevalent in the human gut. In total, 32 species (see tbl. 6.3) were selected based on two prior in-house studies (*117*) and availability of reference genomes in public databases. Reference genomes, proteomes as well as gene and functional annotations were downloaded from RefSeq on March 2019 (release 92).

Table 6.3: Species/strains considered in the study

Identifier	Species/Sub-species	Strains
NT5001	Bacteroides vulgatus	DSM 1447, ATCC 8482
NT5002	Bacteroides uniformis	DSM 6597, ATCC 8492
NT5003	Bacteroides fragilis	DSM 2151, ATCC 25285
NT5004	Bacteroides thetaiotaomicron	DSM 2079, ATCC 29148
NT5006	Clostridium ramosum	DSM 1402, ATCC 25582
NT5009	Eubacterium rectale	DSM 17629
NT5011	Roseburia intestinalis	DSM 14610

Identifier	Species/Sub-species	Strains
NT5017	Veillonella parvula	DSM 2008, ATCC 10790
NT5019	Prevotella copri	DSM 18205
NT5021	Akkermansia muciniphila	DSM 22959, ATCC BAA-835
NT5022	Bifidobacterium adolescentis	DSM 20083, ATCC 15703
NT5024	Eggerthella lenta	DSM 2243, ATCC 25559
NT5025	Fusobacterium nucleatum nucleatum	DSM 15643, ATCC 25586
NT5026	Clostridium bolteae	DSM 15670, ATCC BAA-613
NT5028	Bifidobacterium longum longum	DSM 20219, ATCC 15707
NT5032	Clostridium perfringens	DSM 11782
NT5036	Bilophila wadsworthia	ATCC 49260
NT5037	Clostridium saccharolyticum	DSM 2544, ATCC 35040
NT5038	Streptococcus salivarius	DSM 20560, ATCC 7073
NT5042	Lactobacillus paracasei	ATCC SD5275
NT5045	Ruminococcus bromii	ATCC 27255
NT5046	Ruminococcus gnavus	ATCC 29149
NT5047	Ruminococcus torques	ATCC 27756
NT5048	Coprococcus comes	ATCC 27758
NT5069	Blautia obeum	DSM 25238, ATCC 29174
NT5071	Parabacteroides merdae	DSM 19495, ATCC 43184
NT5072	Streptococcus parasanguinis	DSM 6778, ATCC 15912
NT5073	Collinsella aerofaciens	DSM 3979, ATCC 25986
NT5074	Parabacteroides distasonis	DSM 20701, ATCC 8503
NT5076	Dorea formicigenerans	DSM 3992, ATCC 27755
NT5078	Escherichia coli	ED1a
NT5081	Odoribacter splanchnicus	DSM 20712, ATCC 29572

A database of 16S rRNA regions was constructed, for amplicon analysis, by manually querying the SILVA rRNA database (*93*) and extracting the nearest strain representative sequence. Complementarily Prokka (v1.14.0) (*104*) was used to predict the location of rRNA regions in the 32 genomes. The generated GFF file containing predicted coordinates was used for subsequent analysis.

6.3.2 Selection of drugs

Three human targeted drugs were selected based on previous results. In addition to its primary target, all drugs included in the study have been reported to have antimicrobial effects (*25*, *34*,

72, *113*). Drugs used in this study are listed in tbl. 6.4.

Table 6.4: Drugs used to perturb the mock community

Drug	Human disease target
Chlorpromazine	Schizophrenia
Metformin	Type-2 Diabetes (T2D)
Niclosamide	Tapeworm infections

6.3.3 Pre-inoculation and stabilization of mixed culture

Species were pre-inoculated in isolation on liquid medium from pure stocks and incubated at 37 °C under anaerobic condition for a period of 3 or 5 days, depending on the growth rate of each species (see tbl. A.1). Optical density (OD) of all monocultures was measured after this period and used to calculate ratios that would warrant a mixed culture with equal proportions of each species under the assumption that all species have equal OD properties. The mix of species was then inoculated into mGAM liquid medium.

In order to allow species to reach a stable state (stabilization stage), the culture was grown for 48 hours before being transferred into fresh medium, in duplicate. This process was repeated 3 times for a total of 144 hours at which stage the experimental perturbation was started. OD measurements were performed in the penultimate transfer in order to determine the start of exponential phase and the ideal time to start the drug exposure.

6.3.4 Medium and drug preparation

mGAM medium was prepared according to manufacturer's instructions. Chlorpromazine and Niclosamide were added from DMSO stock solution. Metformin was added as powder directly into the medium at 10x concentration. The medium was then filtered sterile and diluted further to 1x concentration. Final concentrations of each drug in the medium were Metformin - 5 mM, Chlorpromazine - 20 µM, Niclosamide - 20 µM, chosen based on previous work (*72*).

6.3.5 Start of experiment and sample collection

Following the stabilization phase OD of the culture was monitored in order to select the optimal start of the experiment. The experiment was started when the community reached logarithmic phase (OD ~ 2-3) at which point they were inoculated to each of the four conditions (control + 3 drugs) in duplicate. The cultures were sampled at fixed time intervals (see fig. 4.1). The first sample, corresponding to time-point 0, was obtained immediately after the addition of the culture to the medium with drug. Samples were also collected at time-points: 15 minutes, 30m, 1hour, 3h,

48h and 96h. At 48h the culture was transferred to fresh medium and collected prior to transfer. After centrifugation and removal of medium, samples were kept as pellet at -80 °C until extraction. Collected samples were then processed to obtain multiple readouts including, rRNA 16S amplicon sequencing, metagenomic and metatranscriptomic shotgun sequencing, and extra cellular metabolomics (secretome) and metaproteomics via mass spectrometry. Mass spectroscopy readouts were processed by external collaborators and have not been used in this study.

6.3.6 DNA extraction and library preparation

Genomic DNA and total RNA were extracted from flash-frozen samples using Qiagen Allprep Powerfecal DNA/RNA kit (ID: 80244) following the included protocol with minor changes for an additional 700µl phenol-chloroform step after lysing samples via TissueLyser II. DNA was measured by Invitrogen Qubit fluorometric quantitation using Qubit™ dsDNA HS Assay Kit (ID: Q32854) and was stored at −20°C. RNA was measured via Bioanalyzer (Aligent) with Pico and Nano chips depending on the sample concentration and stored at -80°C for further analysis.

Extracted DNA was split into two aliquots for ribosomal 16S amplicon sequencing and metagenomics shotgun sequencing. DNA samples were amplified using primers targeting the V4 region of the 16S rRNA gene with the following primer sequences: F515 5'-GTGCCAGCMGCCGCGGTAA-3'and R806 5'-GGACTACHVGGGTWTCTAAT-3' (*18*). PCR was performed according to the manufacturer's instructions of the KAPA HiFi HotStart PCR Kits (Kapabiosystems) using barcoded primers with minor modifications after a two-step PCR protocol (NEXTflex™ 16S V4 Amplicon-Seq Kit, Bioo Scientific, Austin, Texas, USA). PCR products were removed with SPRIselect magnetic beads (0.8x ratio) after pooling.

RNA samples were depleted for rRNA and tRNA using the NEB experimental kit *NEBNext Bacteria rRNA Depletion Kit (2018)* with additional experimental probes for *Eggerthella lenta* and subsequently libraries were prepared using the *NEBNext Ultra II Directional RNA Library Prep Kit*, multiplexed into three pools and single-end sequenced on an Illumina NextSeq500 platform with a High Output 75 cycles flow cell.

All samples were sequenced at Genomics Core Facility (GeneCore) at EMBL on Illumina© (San Diego, USA) platforms with configurations listed in tbl. 6.5.

Table 6.5: Platforms and settings used to sequence the different mock data types

Data type	Platform	Sequencing	Target depth
16S amplicon	Illumina© MiSeq	250bp paired-end	50 Mbp
metagenomic shotgut	Illumina© HiSeq 4000	150bp paired-end	1.5 Gbp
metatranscriptomic shotgut	Illumina© NextSeq 500	84bp single-end	2.5 Gbp

6.3.7 Sequence processing using NGLess

Amplicon and Metagenomic reads were pre-processed with NGLess using the same approach described in sec. 6.2.4. BAM files were kept for subsequent analysis. Metatranscriptomic reads were processed using the same approach but the quality trimming algorithm `smoothtrim()` was used instead due to a high rate of single-base quality drop affecting the Illumina NextSeq platform.

Amplicon, metagenomes and metatranscriptomes were mapped against the reference database of 32 species (see sec. 6.3.1) using the same approach described in sec. 6.2.4.

Amplicon reads were additionally mapped against the database of 16S regions extracted from SILVA (see sec. 6.3.1) using MAPseq (`v1.2.4`) (*75*). Paired reads were mapped independently and assignments were only considered upon agreement.

6.3.8 Metatranscriptome rRNA data analysis

Two approaches were used to quantify the fraction and distribution of rRNA present in all samples after depletion. In order to assess the proportion of reads originating from every rRNA subunit (`5S`, `16S` and `23S`) of every species, NGLess was used to count the number of mapped reads intersecting the rRNA regions predicted by Prokka (see sec. 6.3.1). Additionally SortMeRna `v3.0.3` (*59*) was used to estimate total rRNA subunit contributions.

6.3.9 Gene annotation concordance

Gene annotations were obtained from RefSeq together with reference genomes (see sec. 6.3.1) and compared with *Prokka* (v1.14.0) (*104*) predictions. Agreement between the two annotations was assessed by comparing the number of predicted genes and total coding bases on both DNA strands.

6.3.10 Abundance estimation using metagenomes and metatranscriptomes

Abundance estimates were produced by counting the number of reads mapping to each genome included in the study. For metagenome derived estimates, total counts were normalized by the size of the genome (number of base-pairs).

For metatranscriptome derived estimates, additional steps were required. rRNA reads were removed using `SortMeRNA` with default parameters. Gene predictions by *Prokka*/*Prodigal* were used to calculate the total number of coding bases per genome, after exclusion of rRNA regions. Finally, total read counts were normalized by the number of coding bases on each genome.

List of publications

Published

1. A. Milanese, D. R. Mende, L. Paoli, G. Salazar, H.-J. J. Ruscheweyh, M. Cuenca, P. Hingamp, **R. Alves** *et al.*, Microbial abundance, activity and population genomic profiling with mOTUs2, *Nature Communications* **10, 1014** (2019). (*82*)

2. L. P. Coelho, **R. Alves** *et al.*, NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language, *Microbiome* **7, 84** (2019). (*21*)

3. T. S. B. Schmidt, M. R. Hayward, L. P. Coelho, S. S. Li, P. I. Costea, A. Y. Voigt, J. Wirbel, O. M. Maistrenko, **R. Alves** *et al.*, Extensive transmission of microbes along the gastrointestinal tract, *eLife* **8** (2019), doi:10.7554/eLife.42693. (*102*)

4. F. Hildebrand, L. Moitinho-Silva, S. Blasche, M. T. Jahn, T. I. Gossmann, J. Huerta-Cepas, R. Hercog, M. Luetge, M. Bahram, A. Pryszlak, **R. Alves** *et al.*, Antibiotics-induced monodominance of a novel gut bacterial order, *Gut*, gutjnl–2018–317715 (2019). (*44*)

In review

1. L. P. Coelho, **R. Alves** *et al.*, Towards the biogeography of prokaryotic genes, submitted to *Nature* (2019). (*22*)

2. S. K. Forslund, R. Chakaroun, J. Aron-Wisnewsky, T. Nielsen, T. S. B. Schmidt, L. Moitinho-Silva, S. Adriouch, **R. Alves** *et al.*, Medication signatures dominate microbiome and metabolome alter- ations in cardiometabolic disease, submitted to *Nature* (2019). (*33*)

Citable works

1. **Alves, R.**, Coelho, L. P., Huerta-Cepas, J., & Bork, P. (2019, January). Simulated metagenomes with quality and abundance distributions derived from real samples. https://doi.org/10.5281/zenodo.2560288

2. Huerta-Cepas, J., Coelho, L. P., & **Alves, R.** (2018, June). eggNOG Mapper annotations of Mouse, Dog and Pig gut gene catalogs. https://doi.org/10.5281/zenodo.1299267

3. **Alves, R.**, Coelho, L. P. (2018). fna2faa - a fast DNA/RNA 6 frames aminoacid translator IUPAC ambiguity aware - https://github.com/unode/fna2faa

4. **Alves, R.**, (2018). jug-schedule - a jug subcommand for job submission to DRMAA compatible clusters - https://gitlab.com/unode/jug_schedule

Abbreviations and glossary

amplicon a segment of DNA or RNA that is both the source and/or product of amplification or replication. 7

assembly the computational process by which overlapping reads are merged into contigs. 21

BAM binary version of a SAM file. 56

bp base-pairs. 49

CAMI Critical Assessment of Metagenome Interpretation. 8, 10, 47, 48

COG Cluster of Orthologous Group. 24

contig a set of overlapping DNA sequence fragments used to construct a physical map of a chromosome. 13, 20, 21, 48, 59, 60

CPU Central Processing Unit. 14

eggNOG evolutionary genealogy of genes: Non-supervised Orthologous Groups. 50

EMBL European Molecular Biology Laboratory. 14, 55

ENA European Nucleotide Archive. 49

Ensembl An online platform that provides integrated genome, gene, variation, gene regulation and comparative genomics data. 49

GeneCore Genomics Core Facility. 55

GFF General Feature Format. 50, 53

GMGC Global Microbial Gene Catalogue. 14

MAG Metagenome Assembled Genome. 13, 14

marker gene genes that due to their conserved sequence or properties are of special interest to a specific application. 8, 13

metagenome DNA sequencing from microbial communities. 2, 48, 52

metatranscriptome RNA sequencing from microbial communities. 2, 48, 52

MGS MetaGenomic Species. 14

NCBI National Center for Biotechnology Information. 47, 48, 60

NEB New England Biolabs. 35, 55

NGLess a domain-specific language for NGS data processing. 49, 50

NGS Next Generation Sequencing. 13, 60

NOG Non-supervised Orthologous Group. 24, 25, 49

ORF Open Reading Frame. 14, 38, 48

PCoA Principal Coordinates Analysis. 21

RefSeq NCBI's curated database of annotated genomic, transcript, and protein sequence records. 52, 56

rRNA ribosomal RNA. 2, 3, 7, 13, 35, 48, 53, 55, 56

SAM tab-delimited text file that contains sequence alignment data. 59

scaffold the result of merging contigs when their order is known but they do not overlap - placeholder sequences are often used to fill gaps. 20

tRNA transfer RNA. 2, 3, 35, 48, 55

List of figures, tables and listings

List of Figures

List of Tables

List of Listings

Bibliography

1. K. Aagaard, J. Ma *et al.*, The placenta harbors a unique microbiome, *Science Translational Medicine* **6** (2014), doi:10.1126/scitranslmed.3008599.

2. G. S. Abu-Ali, R. S. Mehta *et al.*, Metatranscriptome of human faecal microbial communities in a cohort of adult men, *Nature Microbiology* (2018), doi:10.1038/s41564-017-0084-4.

3. E. Afgan, D. Baker *et al.*, The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update, *Nucleic Acids Research* **46**, W537–W544 (2018).

4. M. Albertsen, P. Hugenholtz *et al.*, Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes, *Nature Biotechnology* **31**, 533–538 (2013).

5. R. Alves, jug-schedule - a jug subcommand for job submission to DRMAA compatible clusters, (available at https://gitlab.com/unode/jug).

6. R. Alves, P. Coelho, fna2faa - a fast DNA/RNA 6 frames aminoacid translator IUPAC ambiguity aware, (available at https://github.com/unode/fna2faa).

7. S. Anders, P. T. Pyl, W. Huber, HTSeq–a Python framework to work with high-throughput sequencing data, *Bioinformatics* **31**, 166–169 (2015).

8. S. Andrews, FastQC: a quality control tool for high throughput sequence data, (2010) (available at http://www.bioinformatics.babraham.ac.uk/projects/fastqc).

9. C. R. Armour, S. Nayfach, K. S. Pollard, T. J. Sharpton, A Metagenomic Meta-analysis Reveals Functional Signatures of Health and Disease in the Human Gut Microbiome, *mSystems* **4**, 1–15 (2019).

10. P. Baumgartner, S. Payr, in *Proceedings of ed-media 97–world conference on educational multimedia and hypermedia*, (Citeseer, 1997).

11. J. R. Bedarf, F. Hildebrand *et al.*, Functional implications of microbial and viral gut

metagenome changes in early stage L-DOPA-naïve Parkinson's disease patients., *Genome medicine* **9**, 39 (2017).

12. B. Bischl, M. Lang *et al.*, mlr: Machine Learning in R, *Journal of Machine Learning Research* **17**, 1–5 (2016).

13. E. Bolyen, J. R. Rideout *et al.*, Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2, *Nature Biotechnology* **37**, 852–857 (2019).

14. J. R. Bray, J. T. Curtis, An Ordination of the Upland Forest Communities of Southern Wisconsin, *Ecological Monographs* **27**, 325–349 (1957).

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

16. B. Bushnell, BBMap, (available at sourceforge.net/projects/bbmap/).

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

18. J. G. Caporaso, C. L. Lauber *et al.*, Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample, *Proceedings of the National Academy of Sciences of the United States of America* **108**, 4516–4522 (2011).

19. R. Carr, E. Borenstein, Comparative analysis of functional metagenomic annotation and the mappability of short reads., *PLoS ONE* **9**, e105776 (2014).

20. L. P. Coelho, Jug: Software for Parallel Reproducible Computation in Python, *Journal of Open Research Software* **5** (2017), doi:10.5334/jors.161.

21. L. P. Coelho, R. Alves *et al.*, NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language, *Microbiome* **7**, 84 (2019).

22. L. P. Coelho, R. Alves *et al.*, Towards the biogeography of prokaryotic genes, *in review in Nature* (2019).

23. L. P. Coelho, J. R. Kultima *et al.*, Similarity of the dog and human gut microbiomes in gene content and response to diet, *Microbiome* **6**, 72 (2018).

24. P. I. Costea, G. Zeller *et al.*, Towards standards for human fecal sample processing in metagenomic studies, *Nature Biotechnology* **35**, 1069–1076 (2017).

25. H. D. M. Coutinho, J. G. M. Costa, E. O. Lima, V. S. Falcão-Silva, J. P. Siqueira-Júnior, Enhancement of the antibiotic activity against a multiresistant Escherichia coli by Mentha arvensis

L. and chlorpromazine., *Chemotherapy* **54**, 328–30 (2008).

26. Z. Dai, S. H. Wong, J. Yu, Y. Wei, Batch effects correction for microbiome data with Dirichlet-multinomial regression, *Bioinformatics* **35**, 807–814 (2019).

27. A. C. E. Darling, Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements, *Genome Research* **14**, 1394–1403 (2004).

28. L. A. David, C. F. Maurice *et al.*, Diet rapidly and reproducibly alters the human gut microbiome., *Nature* **505**, 559–63 (2014).

29. D. E. Deatherage, J. E. Barrick, in (2014), pp. 165–188.

30. T. developers, *tidyverse: Easily Install and Load the 'Tidyverse'* (2017; https://cran.r-project.org/package=tidyverse).

31. S. R. Eddy, W. R. Pearson, Ed. Accelerated Profile HMM Searches, *PLoS Computational Biology* **7**, e1002195 (2011).

32. A. Edlund, Y. Yang *et al.*, Uncovering complex microbiome activities via metatranscriptomics during 24 hours of oral biofilm assembly and maturation, *Microbiome* **6**, 217 (2018).

33. K. Forslund, R. Chakaroun *et al.*, Medication signatures dominate microbiome and metabolome alterations in cardiometabolic disease, *in review in Nature* (2019).

34. K. Forslund, F. Hildebrand *et al.*, Disentangling type 2 diabetes and metformin treatment signatures in the human gut microbiota, *Nature* **528**, 262–266 (2015).

35. S. C. Forster, N. Kumar *et al.*, A human gut bacterial genome and culture collection for improved metagenomic analyses, *Nature Biotechnology* **37**, 186–192 (2019).

36. E. A. Franzosa, L. J. McIver *et al.*, Species-level functional profiling of metagenomes and metatranscriptomes, *Nature Methods* **15**, 962–968 (2018).

37. E. A. Franzosa, X. C. Morgan *et al.*, Relating the metatranscriptome and metagenome of the human gut, *Proceedings of the National Academy of Sciences* **111**, E2329–E2338 (2014).

38. S. M. Gibbons, C. Duvallet, E. J. Alm, Correcting for batch effects in case-control microbiome studies, *PLoS Computational Biology* **14**, 1–17 (2018).

39. M. C. de Goffau, S. Lager *et al.*, Human placenta has no microbiome but can contain potential pathogens, *Nature* (2019), doi:10.1038/s41586-019-1451-5.

40. B. Grüning, J. Chilton *et al.*, Practical Computational Reproducibility in the Life Sciences, *Cell Systems* **6**, 631–635 (2018).

41. B. Grüning, R. Dale *et al.*, Bioconda: sustainable and comprehensive software distribution for the life sciences, *Nature Methods* **15**, 475–476 (2018).

42. S. A. Hardwick, W. Y. Chen *et al.*, Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis, *Nature Communications* **9**, 1–10 (2018).

43. A. Heintz-Buschart, P. May *et al.*, Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes., *Nature microbiology* **2**, 16180 (2016).

44. F. Hildebrand, L. Moitinho-Silva *et al.*, Antibiotics-induced monodominance of a novel gut bacterial order, *Gut*, gutjnl–2018–317715 (2019).

45. J. Huerta-Cepas, K. Forslund *et al.*, Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper, *Molecular Biology and Evolution* **34**, 2115–2122 (2017).

46. J. Huerta-Cepas, D. Szklarczyk *et al.*, eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences, *Nucleic Acids Research* **44**, D286–D293 (2016).

47. J. D. Hunter, Matplotlib: A 2D graphics environment, *Computing in Science & Engineering* **9**, 90–95 (2007).

48. D. H. Huson, N. Weber, *Microbial community analysis using MEGAN* (Elsevier Inc., ed. 1, 2013; http://dx.doi.org/10.1016/B978-0-12-407863-5.00021-6), pp. 465–485.

49. C. Huttenhower, R. Knight *et al.*, Advancing the microbiome research community, *Cell* **159**, 227–230 (2014).

50. D. Hyatt, G.-L. Chen *et al.*, Prodigal: prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics* **11**, 119 (2010).

51. C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries, *Nature Communications* **9**, 1–8 (2018).

52. E. Jones, T. Oliphant, P. Peterson, Others, SciPy: Open source scientific tools for Python (available at http://www.scipy.org/).

53. J. Kaiser, Plan to replicate 50 high-impact cancer papers shrinks to just 18, *Science* (2018), doi:10.1126/science.aau9619.

54. D. D. Kang, F. Li *et al.*, MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies, *PeerJ* **7**, e7359 (2019).

55. D. Kevin, S. Vet, K. Faust, F. Moens, Integrated culturing , modeling and transcriptomics uncovers complex interactions and emergent behavior in a synthetic gut community, (2018).

56. Y. M. Kim, J. B. Poline, G. Dumas, Experimenting with reproducibility: A case study of robustness in bioinformatics, *GigaScience* **7**, 1–8 (2018).

57. H. Klingenberg, P. Meinicke, How to normalize metatranscriptomic count data for differential expression analysis, *PeerJ* **5**, e3859 (2017).

58. T. Kluyver, B. Ragan-Kelley *et al.*, F. Loizides, B. Schmidt, Eds. Jupyter Notebooks – a publishing format for reproducible computational workflows, 87–90 (2016).

59. E. Kopylova, L. Noé, H. Touzet, SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data, *Bioinformatics* **28**, 3211–3217 (2012).

60. J. R. Kultima, L. P. Coelho *et al.*, MOCAT2: a metagenomic assembly, annotation and profiling framework, *Bioinformatics*, btw183 (2016).

61. G. M. Kurtzer, V. Sochat, M. W. Bauer, A. Gursoy, Ed. Singularity: Scientific containers for mobility of compute, *PLOS ONE* **12**, e0177459 (2017).

62. D. Li, C.-M. M. Liu, R. Luo, K. Sadakane, T.-W. W. Lam, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, *Bioinformatics* **31**, 1674–1676 (2015).

63. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data, *Bioinformatics* **27**, 2987–2993 (2011).

64. H. Li, I. Birol, Ed. Minimap2: pairwise alignment for nucleotide sequences, *Bioinformatics* **34**, 3094–3100 (2018).

65. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, **00**, 1–3 (2013).

66. H. Li, Seqtk - a fast and lightweight tool for processing FASTA or FASTQ sequences, (available at https://github.com/lh3/seqtk).

67. J. Li, H. Jia *et al.*, An integrated catalog of reference genes in the human gut microbiome, *Nature Biotechnology* **32**, 834–841 (2014).

68. S. S. Li, A. Zhu *et al.*, Durable coexistence of donor and recipient strains after fecal microbiota transplantation, *Science* **352**, 586–589 (2016).

69. G. J. Lithgow, M. Driscoll, P. Phillips, A long journey to reproducible results, *Nature* **548**, 387–388 (2017).

70. J. Lloyd-Price, C. Arze *et al.*, Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases, *Nature* **569**, 655–662 (2019).

71. R. Lorenz, S. H. Bernhart *et al.*, ViennaRNA Package 2.0, *Algorithms for Molecular Biology* **6**, 26 (2011).

72. L. Maier, M. Pruteanu *et al.*, Extensive impact of non-antibiotic drugs on human gut bacteria, *Nature* **555**, 623–628 (2018).

73. G. Marçais, C. Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, *Bioinformatics* **27**, 764–770 (2011).

74. C. E. Mason, E. Afshinnekoo, S. Tighe, S. Wu, S. Levy, International standards for genomes, transcriptomes, and metagenomes, *Journal of Biomolecular Techniques* **28**, 8–18 (2017).

75. J. F. Matias Rodrigues, T. S. B. Schmidt, J. Tackmann, C. Von Mering, MAPseq: Highly efficient k-mer search with confidence estimates, for rRNA sequence analysis, *Bioinformatics* **33**, 3808–3810 (2017).

76. W. McKinney, in *Proceedings of the 9th python in science conference*, S. van der Walt, J. Millman, Eds. (2010), pp. 51–56.

77. R. S. Mehta, G. S. Abu-Ali *et al.*, Stability of the human faecal microbiome in a cohort of adult men, *Nature Microbiology* (2018), doi:10.1038/s41564-017-0096-0.

78. D. R. Mende, I. Letunic *et al.*, proGenomes: a resource for consistent functional and taxonomic annotations of prokaryotic genomes, *Nucleic Acids Research*, gkw989 (2016).

79. D. R. Mende, S. Sunagawa, G. Zeller, P. Bork, Accurate and universal delineation of prokaryotic species, *Nature Methods* **10**, 881–884 (2013).

80. D. R. Mende, A. S. Waller *et al.*, Assessment of metagenomic assembly using simulated next generation sequencing data, *PLoS ONE* **7** (2012), doi:10.1371/journal.pone.0031386.

81. D. Merkel, Docker: Lightweight Linux Containers for Consistent Development and Deployment, *Linux J.* **2014** (2014) (available at http://dl.acm.org/citation.cfm?id=2600239.2600241).

82. A. Milanese, D. R. Mende *et al.*, Microbial abundance, activity and population genomic profiling with mOTUs2, *Nature Communications* **10**, 1014 (2019).

83. A. Moya, M. Ferrer, Functional Redundancy-Induced Stability of Gut Microbiota Subjected

to Disturbance, *Trends in Microbiology* **24**, 402–413 (2016).

84. T. Narihiro, Y. Kamagata, Cultivating yet-to-be cultivated microbes: The challenge continues, *Microbes and Environments* **28**, 163–165 (2013).

85. S. Nayfach, Z. J. Shi, R. Seshadri, K. S. Pollard, N. C. Kyrpides, New insights from uncultivated genomes of the global human gut microbiome, *Nature* **568**, 505–510 (2019).

86. S. Neph, M. S. Kuehn *et al.*, BEDOPS: high-performance genomic feature operations, *Bioinformatics* **28**, 1919–1920 (2012).

87. H. B. Nielsen, M. Almeida *et al.*, Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes, *Nature Biotechnology* **32**, 822–828 (2014).

88. T. Oliphant, {NumPy}: A guide to {NumPy} (available at http://www.numpy.org/).

89. E. Pasolli, F. Asnicar *et al.*, Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle, *Cell* **176**, 649–662.e20 (2019).

90. V. Pecoraro, K. Zerulla, C. Lange, J. Soppa, S. Aziz, Ed. Quantification of Ploidy in Proteobacteria Revealed the Existence of Monoploid, (Mero-)Oligoploid and Polyploid Species, *PLoS ONE* **6**, e16392 (2011).

91. C. Poussin, N. Sierro *et al.*, Interrogating the microbiome: experimental and computational considerations in support of study reproducibility, *Drug Discovery Today* **23**, 1644–1657 (2018).

92. J. Qin, R. Li *et al.*, A human gut microbial gene catalogue established by metagenomic sequencing, *Nature* **464**, 59–65 (2010).

93. C. Quast, E. Pruesse *et al.*, The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools, *Nucleic Acids Research* **41**, 590–596 (2013).

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

95. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2018; https://www.r-project.org/).

96. J. Repass, R. Project, C. Biology, Elife-25801-V1,, 1–12 (2018).

97. C. Rosenow, Prokaryotic RNA preparation methods useful for high density array analysis: comparison of two approaches, *Nucleic Acids Research* **29**, 112e–112 (2001).

98. RStudio Team, *RStudio: Integrated Development Environment for R* (RStudio, Inc., Boston, MA, 2015; http://www.rstudio.com/).

99. B. M. Satinsky, S. M. Gifford, B. C. Crump, M. A. Moran, *Use of internal standards for quantitative metatranscriptome and metagenome analysis* (Elsevier Inc., ed. 1, 2013; http://dx.doi.org/10.1016/B978-0-12-407863-5.00012-5), pp. 237–250.

100. M. Schirmer, E. A. Franzosa *et al.*, Dynamics of metatranscription in the inflammatory bowel disease gut microbiome, *Nature Microbiology* **2**, 17004 (2018).

101. P. D. Schloss, S. L. Westcott *et al.*, Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities, *Applied and Environmental Microbiology* **75**, 7537–7541 (2009).

102. T. S. Schmidt, M. R. Hayward *et al.*, Extensive transmission of microbes along the gastrointestinal tract, *eLife* **8** (2019), doi:10.7554/eLife.42693.

103. A. Sczyrba, P. Hofmann *et al.*, Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software, *Nature Methods* **14**, 1063–1071 (2017).

104. T. Seemann, Prokka: Rapid prokaryotic genome annotation, *Bioinformatics* **30**, 2068–2069 (2014).

105. T. Seemann, Barrnap - BAsic Rapid Ribosomal RNA Predictor, (available at https://github.com/tseemann/barrnap).

106. N. Segata, S. K. Haake *et al.*, Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples., *Genome biology* **13**, R42 (2012).

107. J. Soppa, Polyploidy and community structure, *Nature Publishing Group* **2**, 1–2 (2017).

108. E. J. Stewart, Growing unculturable bacteria, *Journal of Bacteriology* **194**, 4151–4160 (2012).

109. S. Sunagawa, L. P. Coelho *et al.*, Structure and function of the global ocean microbiome, *Science* **348**, 1261359 (2015).

110. S. Sunagawa, D. R. Mende *et al.*, Metagenomic species profiling using universal phylogenetic marker genes, *Nature Methods* **10**, 1196–1199 (2013).

111. N. K. Surana, D. L. Kasper, Moving beyond microbiome-wide associations to causal microbe identification, *Nature* **552**, 244–247 (2017).

112. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches, *Bioinformatics* **31**, 926–932

(2015).

113. J. Tam, T. Hamza *et al.*, Host-targeted niclosamide inhibits C. difficile virulence and prevents disease in mice without disrupting the gut microbiota, *Nature Communications* **9**, 1–11 (2018).

114. R. L. Tatusov, The COG database: a tool for genome-scale analysis of protein functions and evolution, *Nucleic Acids Research* **28**, 33–36 (2000).

115. The Integrative HMP (iHMP) Research Network Consortium, The Integrative Human Microbiome Project, *Nature* **569**, 641–648 (2019).

116. The NIH HMP Working Group, J. Peterson *et al.*, The NIH Human Microbiome Project, *Genome Research* **19**, 2317–2323 (2009).

117. M. Tramontano, S. Andrejev *et al.*, Nutritional preferences of human gut bacteria reveal their metabolic idiosyncrasies, *Nature microbiology* **3**, 514–522 (2018).

118. D. T. Truong, E. A. Franzosa *et al.*, MetaPhlAn2 for enhanced metagenomic taxonomic profiling., *Nature methods* **12**, 902–3 (2015).

119. D. Vandeputte, G. Kathagen *et al.*, Quantitative microbiome profiling links gut community variation to microbial load, *Nature* **551**, 507–511 (2017).

120. T. Vatanen, E. A. Franzosa *et al.*, The human gut microbiome in early-onset type 1 diabetes from the TEDDY study, *Nature* **562**, 589–594 (2018).

121. M. Villa, R. J. Bloom *et al.*, High-throughput isolation and culture of human gut bacteria with droplet microfluidics, *bioRxiv*, 630822 (2019).

122. M. Waskom, O. Botvinnik *et al.*, mwaskom/seaborn: v0.8.1 (September 2017), (2017), doi:10.5281/ZENODO.883859.

123. A. R. Wattam, D. Abraham *et al.*, PATRIC, the bacterial bioinformatics database and analysis resource, *Nucleic Acids Research* **42**, 581–591 (2014).

124. A. R. Wattam, J. J. Davis *et al.*, Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center, *Nucleic Acids Research* **45**, D535–D542 (2017).

125. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag New York, 2016; https://ggplot2.tidyverse.org).

126. L. G. E. Wilkins, C. L. Ettinger, G. Jospin, J. A. Eisen, Metagenome-assembled genomes provide new insight into the microbial diversity of two thermal pools in Kamchatka, Russia, *Scientific Reports* **9**, 1–15 (2019).

127. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2, *bioRxiv*, 1–16 (2019).

128. D. E. Wood, S. L. Salzberg, Kraken: ultrafast metagenomic sequence classification using exact alignments, *Genome Biology* **15**, R46 (2014).

129. G. D. Wu, J. Chen *et al.*, Linking long-term dietary patterns with gut microbial enterotypes., *Science (New York, N.Y.)* **334**, 105–8 (2011).

130. L. Xiao, Q. Feng *et al.*, A catalog of the mouse gut metagenome, *Nature Biotechnology* **33**, 1103–1108 (2015).

131. T. Yatsunenko, F. E. Rey *et al.*, Human gut microbiome viewed across age and geography, *Nature* **486**, 222–227 (2012).

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

133. G. Zeller, J. Tap *et al.*, Potential of fecal microbiota for early-stage detection of colorectal cancer., *Molecular systems biology* **10**, 766 (2014).

134. D. R. Zerbino, P. Achuthan *et al.*, Ensembl 2018, *Nucleic Acids Research* **46**, D754–D761 (2018).

135. W. Zhu, A. Lomsadze, M. Borodovsky, Ab initio gene identification in metagenomic sequences, *Nucleic Acids Research* **38**, e132–e132 (2010).

136. N. Zmora, J. Suez, E. Elinav, You are what you eat: diet, health and the gut microbiota, *Nature Reviews Gastroenterology and Hepatology* **16**, 35–56 (2019).

137. Y. Zou, W. Xue *et al.*, 1,520 Reference Genomes From Cultivated Human Gut Bacteria Enable Functional Microbiome Analyses, *Nature Biotechnology* **37**, 179–185 (2019).

138. ngless-toolkit/ngless-contrib: A collection of NGLess modules open to community contributions (available at https://github.com/ngless-toolkit/ngless-contrib).

139. Illumina discontinued rRNA depletion product list (available at https://www.illumina.com/products/selection-tools/rrna-depletion-selection-guide.html).

140. UniProt: a worldwide hub of protein knowledge, *Nucleic Acids Research* **47**, D506–D515 (2019).

A

Appendix - Images, tables and code

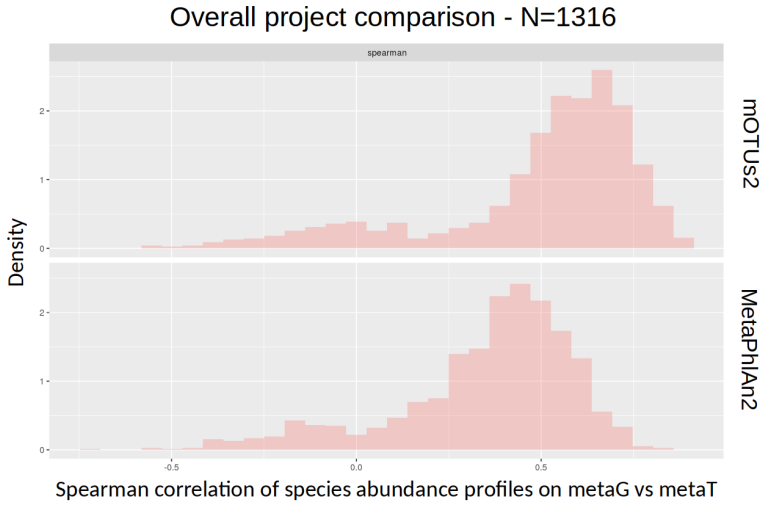

Figure # A.1: ***Overall correlation of taxonomic profiles on paired metagenomes and metatranscriptomes*** *- spearman correlation of taxonomic abundances profiled with* mOTUs2 *and* MetaPhlAn2

Table A.1: Characteristics of the species included in the study and medium used during individual growth

Identifier	Genera	O2 tolerance	Gram	Growth	Medium
NT5001	Bacteroides	anaerobic	negative	normal	mGAM
NT5002	Bacteroides	anaerobic	negative	normal	mGAM
NT5003	Bacteroides	anaerobic	negative	normal	mGAM

Identifier	Genera	O2 tolerance	Gram	Growth	Medium
NT5004	Bacteroides	anaerobic	negative	normal	mGAM
NT5006	Clostridium	anaerobic	positive	normal	mGAM
NT5009	Eubacterium	anaerobic	positive	slow	mGAM
NT5011	Roseburia	anaerobic	positive	normal	mGAM
NT5017	Veilonela	anaerobic	negative	normal	Todd-Hewitt+0.6% sodium lactate
NT5019	Prevotella	anaerobic	negative	slow	mGAM
NT5021	Akkermansia	anaerobic	negative	slow	mGAM
NT5022	Bifidobacterium	anaerobic	positive	normal	mGAM
NT5024	Eggerthella	anaerobic	positive	slow	mGAM
NT5025	Fusobacterium	anaerobic	negative	slow	mGAM
NT5026	Clostridium	anaerobic	positive	normal	mGAM
NT5028	Bifidobacterium	anaerobic	positive	normal	mGAM
NT5032	Clostridium	anaerobic	positive	normal	mGAM
NT5036	Bilophila	anaerobic	negative	normal	60 mM sodium formate, 10 mM taurine
NT5037	Clostridium	anaerobic	positive	normal	mGAM
NT5038	Streptococcus	fac. aerobic	positive	normal	mGAM
NT5042	Lactobacillus	fac. aerobic	positive	normal	mGAM
NT5045	Ruminococcus	anaerobic	positive	slow	mGAM
NT5046	Ruminococcus	anaerobic	positive	normal	mGAM
NT5047	Ruminococcus	anaerobic	positive	slow	mGAM
NT5048	Coprococcus	anaerobic	positive	normal	mGAM
NT5069	Blautia	anaerobic	positive	normal	mGAM
NT5071	Parabacteroides	anaerobic	positive	normal	mGAM
NT5072	Streptococcus	anaerobic	positive	normal	mGAM
NT5073	Collinsella	anaerobic	positive	normal	mGAM
NT5074	Parabacteroides	anaerobic	negative	normal	mGAM
NT5076	Dorea	anaerobic	positive	normal	mGAM
NT5078	Escherichia	anaerobic	negative	normal	mGAM
NT5081	Odoribacter	anaerobic	negative	normal	mGAM

Table A.2: Spearman correlation of taxonomic profiles from paired metagenomes and metatranscriptomes

Dataset	Tool	`mean(spearman)`	`median(spearman)`
Franzosa2014	*mOTUs2*	0.483	0.513
	MetaPhlAn2	0.338	0.369
HMP2-IBD	*mOTUs2*	0.221	0.425

Dataset	Tool	mean(spearman)	median(spearman)
	MetaPhlAn2	0.076	0.174
HMP2-IBDMDB	*mOTUs2*	0.602	0.624
	MetaPhlAn2	0.440	0.445
HPFS-MLVS	*mOTUs2*	0.261	0.356
	MetaPhlAn2	0.149	0.178
InternalGT	*mOTUs2*	0.746	0.745
	MetaPhlAn2	0.618	0.666
T1D_LCSB	*mOTUs2*	0.656	0.695
	MetaPhlAn2	0.511	0.524
Overall	*mOTUs2*	0.496	0.574
	MetaPhlAn2	0.351	0.403

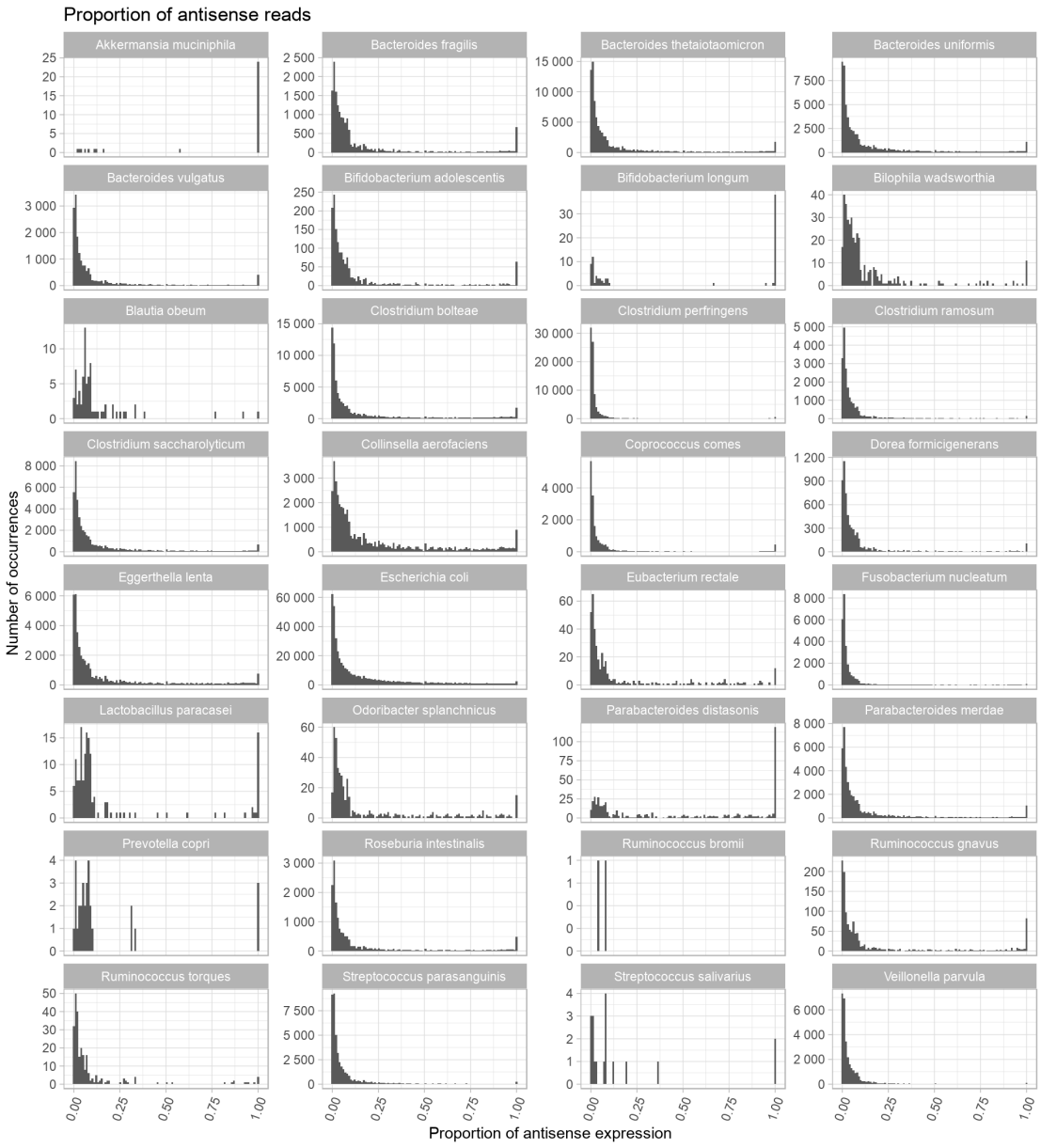

Figure # A.2: ***Proportion of antisense reads per gene*** *- shown is the ratio of antisense and sense mapped reads across all genes with at least 10 reads and proportion* $> 10^{-7}$*. Sense was defined using ORF predictions from Prokka/Prodigal*

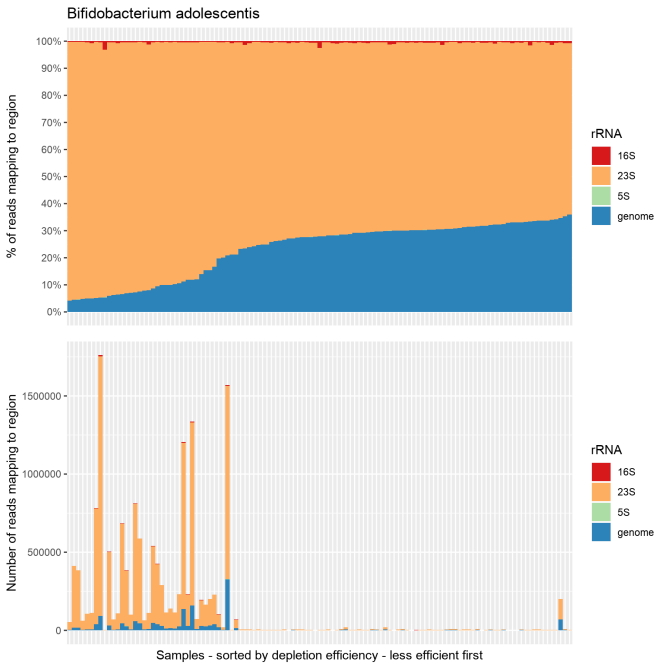

Figure # A.3: ***B.adolescentis rRNA analysis after depletion***

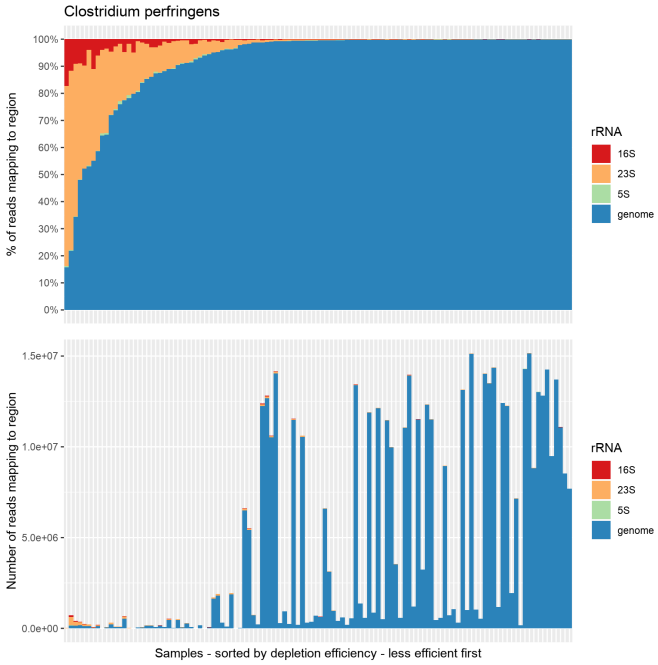

Figure # A.4: ***C.perfringens rRNA analysis after depletion***

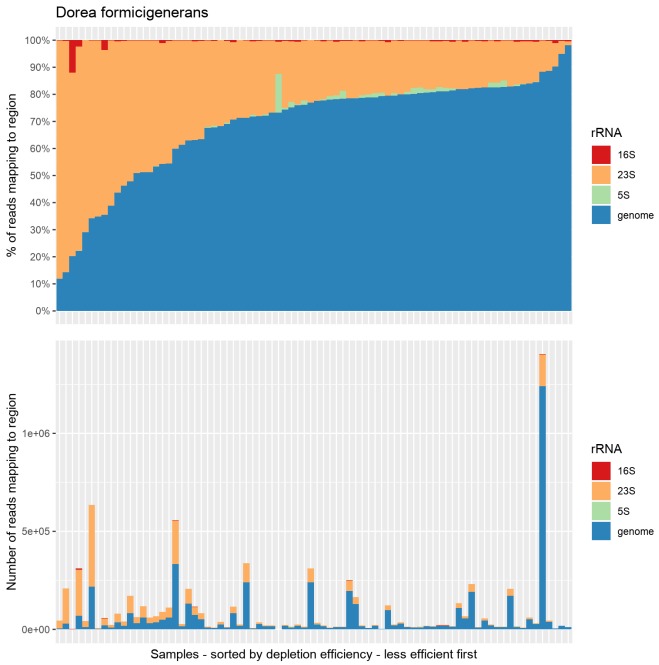

Figure # A.5: ***D.formicigenerans rRNA analysis after depletion***

Listing A.1 ng-meta-profiler - human-gut-profiler

```
#!/usr/bin/env ngless
ngless "0.9"
import "mocat" version "0.0"
import "specI" version "0.1"
import "motus" version "0.1"
import "igc" version "0.9"

input = load_mocat_sample(ARGV[1])
RESULTS = ARGV[2]

qc_reads = preprocess(input, keep_singles=False) using |read|:
    read = substrim(read, min_quality=25)
    if len(read) < 45:
        discard

human_mapped = map(qc_reads, reference='hg19')

non_human = select(human_mapped) using |mr|:
    mr = mr.filter(min_match_size=45, min_identity_pc=90, action={unmatch})
    if mr.flag({mapped}):
        discard

non_human_reads = as_reads(non_human)

igc_mapped = map(non_human_reads, reference='igc', mode_all=True)
igc_mapped_post = select(igc_mapped) using |mr|:
    mr = mr.filter(min_match_size=45, min_identity_pc=95, action={drop})
    if not mr.flag({mapped}):
        discard

igc_counts = count(igc_mapped_post, features=['OGs'],
                   multiple={dist1}, normalization={scaled})
write(igc_counts, ofile=RESULTS </> 'eggNOG.traditional.counts.txt',
      auto_comments=[{hash}, {script}])

mapped_refmg = map(non_human_reads, reference='refmg')
mapped_refmg = select(mapped_refmg) using |mr|:
    mr = mr.filter(min_match_size=45, min_identity_pc=97, action={drop})
    if not mr.flag({mapped}):
        discard

write(count(mapped_refmg, features=['specI_cluster']),
      ofile=RESULTS </> 'specI.raw.counts.txt')

write(count(mapped_refmg, features=['specI_cluster'], normalization={scaled}),
      ofile=RESULTS </> 'specI.scaled.counts.txt')

specI_reads = as_reads(mapped_refmg)

motus_mapped = map(specI_reads, reference='motus', mode_all=True)
motus_raw_gene_counts = count(motus_mapped, features=['gene'], multiple={dist1})

write(motus(motus_raw_gene_counts), ofile=RESULTS </> 'motus.counts.txt')
```

B

Appendix - Software

The following software was used throughout the development of this work.

- `barrnap` v0.9 (*105*)
- `bbmap` v38.26 (*16*)
- `bedops` v2.4.35 (*86*)
- `bedtools` v2.27.1 (*94*)
- `breseq` v0.33.2 (*29*)
- `BWA-MEM` (*65*)
- `cmessi` v1.2.1 (*80*)
- `datamash` v1.2
- `diamond` v0.9.34-38 (*15*)
- `docker` (*81*)
- `eggnog-mapper` (*45*)
- `fastANI` v1.1 (*51*)
- `FastQC` v0.11.5 (*8*)
- `fastx-toolkit` v0.0.14
- `fetchMG` v1.0 (*60*)
- `fna2faa` v0.1.1 (*6*)
- `hmmer` v3.1b2 (*31*)

- HUMAnN2 (*36*)
- jellyfish v2.2.10 (*73*)
- jug-schedule (*5*)
- Jupyter (*58*)
- Linux operating system
- MapSeq v1.2.3 (*75*)
- Mauve (*27*)
- MEGAHIT (*62*)
- MetaGeneMark v3.38 (*135*)
- MetaPhlAn v2.5.0-2.7.0 (*118*)
- minimap2 v2.14 (*64*)
- MOCAT2 (*60*)
- mOTUs v1-v2.0 (*82*, *110*)
- ncbiBLAST+ v2.8.1 (*17*)
- NGLess v0.5-1.0.0 (*22*)
- Pandoc
- prodigal v2.6.3 (*50*)
- prokka v1.1.0 (*104*)
- Python v2.7, v3.5-3.6
- Python packages:
  - HTSeq (*7*)
  - jug 1.6.4-1.6.7 (*20*)
  - matplotlib (*47*)
  - numpy (*88*)
  - pandas (*76*)
  - SciPy (*52*)
  - seaborn (*122*)
- R language v3.4, v3.5 (*95*)
- R packages:
  - mlr (*12*)
  - patchwork
  - rmarkdown (*12*)
  - tidyverse (*30*, *125*)
- RStudio (*98*)
- samtools (*63*)
- seqtk v1.3 (*66*)
- singularity v3.0 (*61*)

- `SortMeRNA` v3.0.3 (*59*)
- `syncthing` v0.14.51
- `ViennaRNA` v2.4.13 (*71*)
- `vim`
- `Visual Studio Code`
- `xsv` v0.12.2