From DEPARTMENT OF CELL AND MOLECULAR BIOLOGY
Karolinska Institutet, Stockholm, Sweden

TAILORING BIOINFORMATICS STRATEGIES FOR THE CHARACTERIZATION OF THE HUMAN MICROBIOME IN HEALTH AND DISEASE

Mauricio Barrientos Somarribas

Karolinska
Institutet

Stockholm 2019

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet.
Printed by Eprint AB 2019
© Mauricio Barrientos Somarribas, 2019
ISBN 978-91-7831-508-6

TAILORING BIOINFORMATICS STRATEGIES FOR THE CHARACTERIZATION OF THE HUMAN MICROBIOME IN HEALTH AND DISEASE
THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Mauricio Barrientos Somarribas

Principal Supervisor:
Prof. Björn Andersson
Karolinska Institutet
Department of Cell and Molecular Biology

Co-supervisor(s):
Tobias Allander, Ph.D
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

Stefanie Präst-Nielsen, Ph.D
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology
Center for Translational Microbiome Research

Opponent:
Associate Professor Leo Lahti
University of Turku
Department of Mathematics and Statistics

Examination Board:
Associate Professor Tomas Bergström
Swedish University of Agricultural Sciences
Department of Animal Breeding and Genetics

Associate Professor Anders Andersson
Kungliga Tekniska Högskolan
Department of Gene Technology
Science for Life Laboratory

Professor Martin Rottenberg
Karolinska Institutet
Department of Microbiology, Tumor and Cell Biology

To my mother, and everyone who helped me get here

ABSTRACT

The human microbiome is a very active area of research due to its potential to explain health and disease. Advances in high throughput DNA sequencing in the last decade have catalyzed the growth of microbiome research; DNA sequencing allows for a cost-effective method to characterize entire microbial communities directly, including unculturable microbes which were previously difficult to study. 16S rRNA sequencing and shotgun metagenomics, coupled with bioinformatics methods have powered the characterization of the human microbiome in different parts of the body. This has led to the discovery of novel links between the microbiome and diseases such as allergies, cancer, and autoimmune diseases.

This thesis focuses on the application of both 16S rRNA sequencing and shotgun metagenomics for the characterization of the human microbiome and its relationship with health and disease. We established two methodologies to address these questions. The first methodology is a bench-to-bioinformatics pipeline to discover putative viral pathogens involved in disease using shotgun metagenomics technology. In **paper I**, we apply the proposed pipeline to explore the hypothesis of viral infection as a putative cause of childhood Acute Lymphoblastic Leukemia. In **paper II**, we propose a complementary method to the pipeline to improve the detection of unknown viruses, especially those with little or no homology to currently known viruses. We applied this method on a collection of viral-enriched libraries which resulted in the characterization of a new viral-like genome.

The second methodology was developed to explore and generate hypothesis from a human skin microbiome dataset of Psoriasis and Atopic Dermatitis patients. The results of the analysis are presented in Paper III and Paper IV. **Paper III** is a pure data-driven exploration of the dataset to discover different aspects on how the microbiome is linked to both diseases. **Paper IV** follows up from the results of paper III but focuses on characterizing the skin site microbiome variability in Atopic Dermatitis.

LIST OF SCIENTIFIC PAPERS

I. Bogdanovic G, Pou C, **Barrientos-Somarribas M**, Bjerkner A, Honkaniemi E, Allander T, et al. **Virome characterisation from Guthrie cards in children who later developed acute lymphoblastic leukaemia**. Br J Cancer. 2016 Oct 11;115(8):1008–14.

II. **Barrientos-Somarribas M***, Messina DN*, Pou C, Lysholm F, Bjerkner A, Allander T, et al. **Discovering viral genomes in human metagenomic data by predicting unknown protein families**. Sci Rep. 2018;

III. Fyhrquist, N., Muirhead, G., Prast-Nielsen, S., Jeanmougin, M., Olah, P., Skoog, T., Jules-Clement, G., Feld, M., **Barrientos-Somarribas, M.**, Pennino, D., Suomela, S., Tessas, I., Lybeck, E., Baran, A.M., Darban, H., Gangwar, R.S., Gerstel, U., Jahn, K., Karisola, P., Yan, L., Hansmann, B., Katayama, S., Meller, S., Bylesjö, M., Hupé, P., Levi-Schaffer, F., Greco, D., Ranki, A., Schröder, J.M., Barker, J., Kere, J., Tsoka, S., Lauerma, A., Soumelis, V., Nestle, F.O., Homey, B., Andersson, B., Alenius, H. **Microbe-host interplay in atopic dermatitis and psoriasis.** Unpublished manuscript. 2019.

IV. **Barrientos-Somarribas M*,** Ottman N*, MAARS Consortium, Andersson B, Alenius H. **Microbial and transcriptional differences elucidate atopic dermatitis heterogeneity across skin sites.** Unpublished manuscript. 2019

OTHER PUBLICATIONS

V. Pou C, **Barrientos-Somarribas M**, Marin-Juan S, Bogdanovic G, Bjerkner A, Allander T, et al. **Virome definition in cerebrospinal fluid of patients with neurological complications after hematopoietic stem cell transplantation**. J Clin Virol. 2018 Nov;108:112–20.

CONTENTS

LIST OF ABBREVIATIONS

16S	16S subunit rRNA gene
AD	Atopic Dermatitis / Eczema
ALL	Acute Lymphoblastic Leukemia
ASV	Amplicon Sequence Variant
CSF	Cerebrospinal Fluid
DNA	Deoxyribonucleic Acid
EM	Expectation Maximization
ESV	Exact Sequence Varians
GLM	Generalized Linear Model
HMP	Human Microbiome Project
ITS	Internal Transcribed Spacer
LCA	Lowest Common Ancestor
MAE	Mean Absolute Error
MAP	Maximum *a posteriori*
MDS	Multidimensional Scaling
MSA	Multiple Sequence Alignment
OTU	Operational Taxonomic Unit
PacBio	Pacific Biosciences
PASI	Psoriasis Area Severity Index
PCA	Principal Component Analysis
PCoA	Principal Coordinate Analysis
PSO	Psoriasis
RDP	Ribosomal Database Project
RNA	Ribonucleic Acid

SCORAD	SCORing Atopic Dermatitis
SISPA	Sequence-Independent, Single Primer Amplification
SNV	Single Nucleotide Variation
ssDNA / ssRNA	Single stranded DNA or RNA
UMAP	Uniform Manifold Approximation and Projection
VLP	Viral-like particle

1 INTRODUCTION

1.1 The human microbiome

Nearly four decades ago, the bacterium *H. pylori* was isolated from the stomach, debunking the common belief that the stomach is a sterile environment. We have now realized that the human body is host to a vast range of bacteria, fungi, viruses and other eukaryotes that interact with our organism. The existence of microbial communities in places such as the gut, the oral cavities or the skin have been well documented for many years now, but recent studies continue to find microbes residing in unexpected places (Dickson and Huffnagle 2015). For example, the lung and the eye have been recently found to contain microbes (Hilty et al. 2010; O'Dwyer, Dickson, and Moore 2016; Huffnagle, Dickson, and Lukacs 2017; St. Leger et al. 2017; Shin et al. 2016; Cavuoto et al. 2018). Other studies also suggest that microbes live in the placenta (Stout et al. 2013), and evidence from immunosuppressed individuals suggests the existence of viral communities in the blood (L. Li et al. 2013; Popgeorgiev et al. 2013).

We refer to the collection of microbial communities colonizing the different sites in the human body as 'the human microbiome[1]' (E. A. Grice and Segre 2012). These communities are dynamical entities (Gonze et al. 2018; Faust et al. 2015); the composition of any microbial community in the human body will depend on the physiological conditions (e.g. temperature, pH, oxygen), resource availability, host-microbe interactions (Gilbert et al. 2018; Virgin 2014) and the interactions within the community (Fredricks 2001; J. Xu 2006). Considering the wide variability of environments in the human body and across individuals, it is not unexpected that the variability of the human microbiome is huge, among individuals and between sites. A recent study even suggests microbiota is so unique that it could identify individuals (Franzosa et al. 2015), although host genetics appear not to be a strong determinant of the microbiome (Rothschild et al. 2018).

The microbiome plays an important role in maintaining human homeostasis, contributing to metabolism(LeBlanc et al. 2013; Metges 2000; Flint et al. 2012), training of the immune system and modulation of the immune response (E. A. Grice and Segre 2012; Ursell et al. 2012; Naik et al. 2012). Consequently, alterations to the healthy resident microbial

[1] In this thesis, the term 'microbiome' will be used as defined by Ledeberg & McCray(2001): "the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space and have been all but ignored as determinants of health and disease". However, it is noted that in the literature the term 'microbiome' can also be defined as 'the collection of genomes from the aforementioned organisms'.

communities can have a considerable impact on health. These imbalanced states are termed 'dysbiosis' (Petersen and Round 2014) and they have been linked to a wide range of diseases such as inflammatory bowel disease (Frank et al. 2007; Norman et al. 2015), asthma (Hilty et al. 2010), atopic dermatitis (Oh et al. 2013; Kong, Oh, Deming, Conlan, Grice, Beatson, Nomicos, Polley, Komarow, Murray, et al. 2012), and depression (Foster and McVey Neufeld 2013).

Understanding the role of the microbiome in disease holds the potential for developing new diagnostic, therapeutic or preventive tools. Examples of this include the fecal transplants for the treatment of *C. difficile* infection (Aas, Gessert, and Bakken 2003) and the development of probiotics (Khalesi et al. 2019). However, further research is required to elucidate the complexity of the microbiome and to translate these insights gained from the research into clinical practice.

1.2 An ecological framework to study the microbiome

To tackle the complexity of the microbiome and its effect on health and disease, the field has borrowed a conceptual framework from ecology. Biological complexity is partitioned into stratified levels of organization: organism, population, community, ecosystem, biome, and biosphere. However, for the study of the human microbiome, we will only consider the population, community and ecosystem levels (J. Xu 2006).

At the population level, each species is studied independently from the community. The goal is to understand a microbial species in isolation and characterize its functions (e.g. metabolism, replication strategies, virulence, cell biology), patterns of evolution and variation(J. Xu 2006). Studying the behavior of individual microbial populations forms the basis to understand how a species will interact with the other members of the community and with the human body in different environments.

At the community level, the aim is to investigate the composition, behavior and the spatio-temporal dynamics of microbial populations that share a common environment (Konopka 2009). Challenges at this level include determining the species and strain composition and profiling the functional and metabolic content of the entire community and the contribution of each population (Kuczynski et al. 2011).

Another key aim when examining microbial communities is to determine the type and mechanisms behind microbe-microbe interactions and their spatial distribution (Mah and O'Toole 2001). Microbial interactions can be classified into 5 broad categories: 1) mutualistic, when both organisms benefit from the interaction 2) commensal, when one organism benefits from the interaction and the other one remains unaffected, 3) competitive, when one organism will kill or inhibit the growth of another that requires the same resources; 4) parasitic, when one organism benefits while harming its host, and 5) predatory, when one organism kills another one as part of its life cycle(Lang and Benbow 2013).

At the ecosystem level, we examine how both individual populations and the entire community interact with the environment: the human body. The focus of study at this level include understanding how communities adapt to changes in the environmental conditions (e.g. nutrients, oxygen, pH) and describing the mechanisms and effects of the crosstalk between the microbes, the immune system and surrounding cells. Understanding human health and disease can only happen when the ecosystem is considered (Gilbert et al. 2018).

This introduction describes different techniques to extract population and community level information from microbial communities using DNA sequencing technologies and how to associate them with health and disease. The following sections (1.3–1.5) describe how sequencing can be used for characterizing the microbiome and two protocols to achieve this. Section 1.6 describes statistical techniques and algorithms to summarize population and community-level information and find ecosystem-level insights through associations with human information using clinical metadata or other omics datasets. Finally, section 1.7 outlines two challenges in human microbiome research that will be addressed by the work in this thesis.

1.3 Characterizing the microbiome using DNA sequencing: a conceptual overview

Microbiome studies depend on our ability to accurately profile microbial communities. One way of achieving this goal is to study their metagenome: the collection of genomes from all microbes in the community. The information from the metagenome enables the reconstruction of the microbial species profile from all kingdoms, including bacteria, viruses, archaea and other microscopic eukaryotes. Furthermore, because we are sampling complete genomes, we can infer the functional potential of communities and individual populations. Another advantage of metagenomics (i.e. the study of metagenomes), is that it allows the study of communities without isolating and culturing the organisms (K. Chen and Pachter 2005).

The concept of an ideal human metagenomics pipeline is presented in Figure 1.3.1. The data generation follows a single-cell-like protocol, that can be broken down into three steps: sample processing, nucleic acid extraction and sequencing. The sample processing step consists of isolating all microbial cells and viral particles from any type of human sample (e.g. a skin biopsy, stool, or blood). An ideal sample processing step discards human cells and other molecules since these comprise a sizeable fraction of the original sample but contain no microbiome information. Purified microbes are then subjected to nucleic acid extraction, which isolates the genomic content from each microbe and labels them with a unique barcode that identifies genomic fragments from the same microbial cell or virus. Then, the barcoded genomic fragments are sequenced, generating error-free digital representations of the nucleotide composition of each genome.

The resulting sequenced genomes can be used to extract biological information from the microbial community. Typically, the information we are interested to extract from the metagenomes falls into three categories (Knight et al. 2018):

- Taxonomical abundance profile: A list of species present in the sample coupled with a measure of abundance for each species
- Strain-level population information: For microbial populations of interest, inferred strain variation based on genotypes (SNVs) or gene content.
- Functional profile: The coding potential of the entire community, as well as the coding content of the different populations. This include gene, gene family and pathway abundances

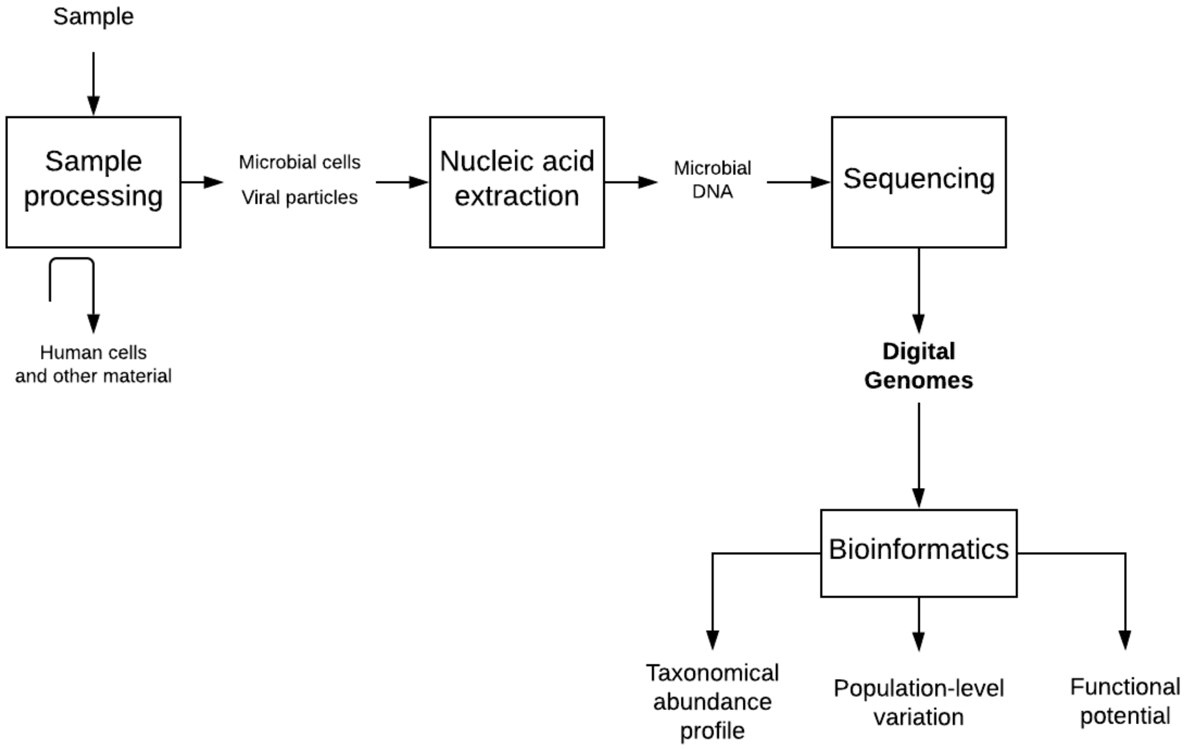

Figure 1.3.1 - An ideal metagenomics assay

The biological information can be used to extract insights depending on the question under study. For example, we can use the taxonomical profile to determine whether certain populations expand or contract in relation to a clinical condition. We can also examine population-level variability within the different microbial populations to determine whether strains with different functional profiles are associated with a phenotype of interest. Finally, we can also examine the functional potential of the metagenome, both at the global level to understand what the community is able to synthesize and react to as a whole, and at a per-clade level to predict mechanisms though which key populations partake in specific processes.

Unfortunately, the state of technology is far away from this proposed scenario. Sections 1.4 and 1.5 describe the two main protocols for human microbiome profiling using DNA sequencing technologies: shotgun metagenomics and amplicon sequencing respectively. It is worth noting that other techniques exist, such as metatranscriptomics, but they will not be covered in the scope of this document.

1.4 Shotgun Metagenomics

Shotgun metagenomics is the real-life protocol to obtain metagenomes from a sample. The first pioneering metagenomics studies surveyed the microbial diversity of the oceans (Rusch et al. 2007) and it has since been used to characterize soil (Howe et al. 2014), wastewater (Munck et al. 2015) and even the New York City metro (Afshinnekoo et al. 2015). The human microbiome field has also caught up, with the Human Microbiome Project utilizing shotgun metagenomics for the latter part of the project (Wylie et al. 2012). Now it is a common choice for the development of large studies such as MetaHIT consortium (Qin et al. 2010a).

Figure 1.4.1 outlines a typical shotgun metagenomics pipeline. The following subsections will describe the steps of the process. Section 1.4.1 describes the sequencing data generation steps. Then Sections 1.4.2 and 1.4.3 describe two alternative but complementary ways of inferring biological information from the sequences: read-based analysis and metagenomic assembly.

1.4.1 Data generation

The sample processing steps from the shotgun metagenomics pipeline differ considerably from those proposed in the conceptual pipeline due to technological limitations. Compared to the conceptual pipeline, there is no purifying step that isolates and sorts the microbial cells and viral particles, so the nucleic acid extraction step is performed on the entire sample. The isolated DNA is then fragmented, amplified and adapters are ligated for sequencing. This has important implications for the downstream data processing such as:

- The nucleic acid extraction efficiency for each type of microbe will differ depending on the method, so the extracted genomic DNA concentrations will be biased with respect to the microbial abundances (Kong et al. 2017).
- If the sample contains human cells, the extracted DNA will be a mixture of human and microbial DNA. This can impact the sensitivity of the assay if the human-to-microbial DNA ratio is very high since human fragments will waste part of the sequencing capacity (Pereira-Marques et al. 2019).
- All DNA is extracted in bulk, meaning it is not possible to track which DNA fragments come from a particular microbial cell or viral particle. This has to be inferred computationally downstream.

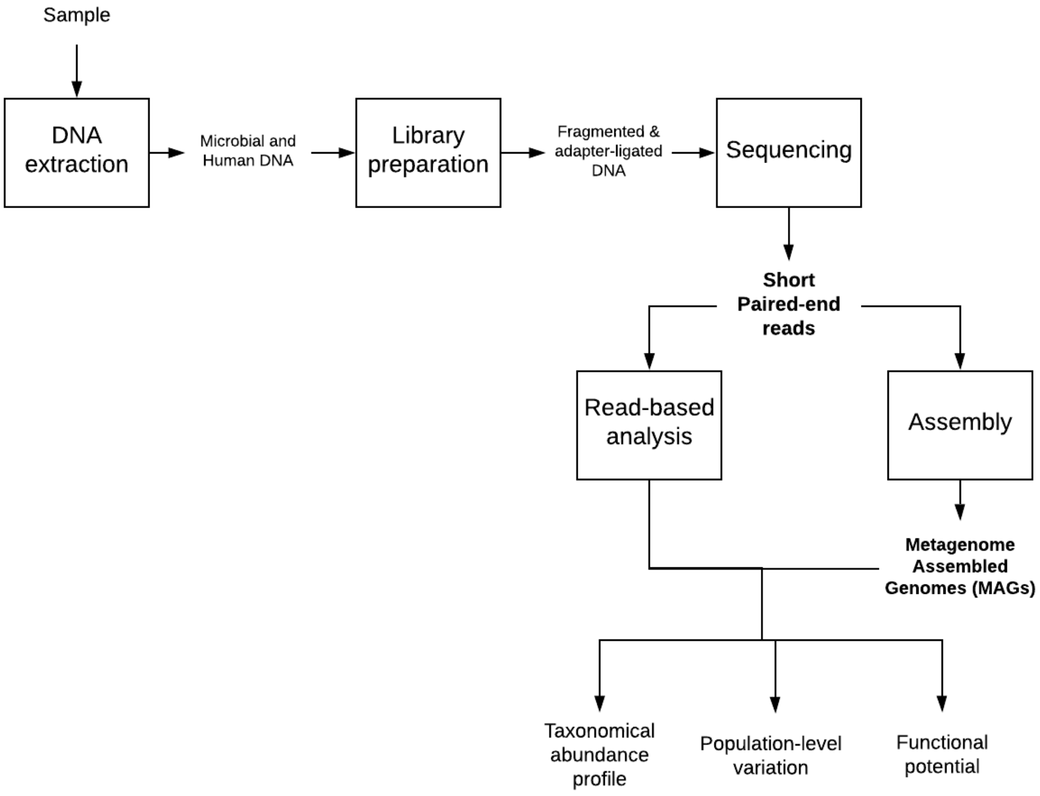

Figure 1.4.1 - The shotgun metagenomics pipeline

- The presence of laboratory reagent contamination, especially on low biomass samples (Salter, Cox, Turek, Calus, Cookson, Pop, et al. 2014; Mollerup et al. 2016). The use of negative controls can help reduce the impact of this problem during analysis.

Additionally, current sequencing technology also imposes further constraints on the data:

- Most current metagenomics studies nowadays are performed using short-read sequencing technologies. Even though the output is not full genome sequences, short-reads can be used to infer information about the community either directly (see Section 1.4.2) or by reconstructing the genomic fragments from which they originate (see Section 1.4.3) (Knight et al. 2018).
- DNA-sequencing is error-prone. Although the error rates of short-read sequencers are relatively low, existing errors convolute genome reconstruction and other analysis (Schirmer et al. 2016).
- Sequencing instruments have a limited capacity below the total DNA molecule number in a typical microbiome sample. This implies that only relative abundances from can be obtained from sequencing data since the total sequence counts are constrained by the instrument capacity (Gloor et al. 2017).

- High complexity communities require deep sequencing to survey the low abundance members (Rodriguez-R and Konstantinidis 2014; Cleary et al. 2015)

The resulting sequencing dataset will consist of millions of short sequences. Most metagenomic sequencing is performed with Illumina technology due to economic reasons, although other alternatives such as Ion Torrent or long-read technologies like Pacific Biosciences and Oxford Nanopore exist. This means the output sequences will be paired-ended (a fragment is sequenced from 5' to 3' and also from 3' to 5') and read length will range from 100 to 300 bp depending on the sequencing instrument and kit.

1.4.2 Read-based microbiome reconstruction

One way of reconstructing a microbial community from shotgun metagenomics data is to leverage the information about existing microbial genomes in public databases to infer what is contained in the metagenome without explicitly reconstructing the genomes. These approaches are commonly known as 'read-based' or 'reference-based'. Generally speaking, read-based analyses yield good results when the microbial diversity of the samples is relatively well known (Knight et al. 2018), but the final results can be heavily influenced by the choice of database (Shaiber and Eren 2019; Breitwieser et al. 2019).

Read-based tools can recover different types of microbial information from the reads. The following subsections describe different types of tools categorized by the microbial information they recover.

1.4.2.1 Taxonomic profile inference

Taxonomic profile inference tools aim to identify all taxa present in the sample and estimate the abundance of each taxon. These tools can be broadly classified into three categories depending on the strategy they use:

- Mapping-based with a strategy to deal with multimapped reads
- Mapping-based with Lowest Common Ancestry
- Marker gene-based

The first approach is analogous to the transcriptome quantification problem. The problem can be framed as estimating the abundance of a set of reference genomes in a given dataset. The general concept consists of mapping the reads to the reference genomes and analyzing the mapping results to estimate the abundance of each species in the set. One of the main challenges with the abundance estimation step is to select a model to incorporate the information from multi-mapped reads (reads that are assigned to more than one species) to obtain accurate abundance measurements. This approach is very reliable if the genomes of the community members are well characterized. However, if unknown species are present in the samples, reads from these unknown species can be misclassified and skew the estimated profiles.

Tools like Pathoscope (Francis et al. 2013) originally implemented this concept, analyzing the output of traditional mapping tools such as bowtie or bwa and using the Expectation-Maximization (EM) algorithm to estimate the abundances from all mapped reads. Newer tools optimize the mapping step by determining if the read is compatible with the reference, instead of calculating the full alignment. Examples of this include Centrifuge (Kim et al. 2016), which uses an FM-index based genome index for assignment and then implements a Cufflinks-like EM approach for multimapping read assignment. Pseudo-mapping tools for RNA-seq such as Salmon (Patro et al. 2017) and Kallisto (N. L. Bray et al. 2016) incorporate their own models to disambiguate multi-compatible reads in the quantification, as well as bias estimation. However, due to the large size of the microbial genome databases compared to a human or mouse transcriptome, RNA-seq tools must be adapted for metagenomics use, e.g FastViromeExplorer (Tithi et al. 2018), or metakallisto (Schaeffer et al. 2017). Minhash-based strategies for metagenome profiling could also be classified in this category (C. T. Brown and Irber 2016; Ondov et al. 2016).

The second approach originates from addressing one of the main causes of read multimapping in microbial datasets: homologous regions. **The lowest common ancestor (LCA) strategy** states that if a sequence originates from a conserved region in different bacteria, the appropriate classification for the sequence is the taxonomical level where the region is conserved. For example, if a read maps to a conserved region of the 16S rRNA gene, it should be classified as ‘Kingdom: Bacteria’, whereas a read from a gene conserved only in Staphylococcal species should be classified as ‘Genus: Staphylococcus’. In practice, LCA approaches analyze all the database matches for each read and use the taxonomic lineage information to determine what is the lowest node in the taxonomical tree that spans all matches.

The first LCA tools like MEGAN (Huson et al. 2007) analyze blast search results against nucleotide or protein databases such as NCBI nt or nr, using the NCBI taxonomy to place reads at the appropriate taxonomic level. However, blast searches are computationally expensive, and scaling up to millions of sequences becomes prohibitive. Kraken (Wood and

Salzberg 2014) pioneered an efficient k-mer based algorithm to implement the LCA. Kraken's algorithm depends on building a database of informative k-mers from a set of reference genomes, where each k-mer is annotated with the LCA of the reference genomes where it was observed. Then, the database is used to classify reads by annotating each k-mer of the read and performing an LCA of the k-mer hits to determine the final annotation.

LCA-based methods have both advantages and disadvantages. In principle, these methods are more robust than other mapping approaches when classifying sequences from genomes that are not present in the database. However, in practice, the accuracy of most tools suffers when considering classifications below order (Vollmers, Wiegand, and Kaster 2017; Lindgreen, Adair, and Gardner 2015). Also, many LCA based tools do not include an abundance estimation model for the different clades (Schaeffer et al. 2017), so they need to be complemented with tools like Bracken for accurate abundance estimation (Lu et al. 2017). Finally, modern LCA approaches such as Kraken2 and Clark-S (Ounit and Lonardi 2016) are extremely fast classifiers, but the database construction can be very resource-consuming.

The last strategy for classification is to use marker genes for profiling. There are two types of marker genes: universal genes, and 'clade-specific' marker genes. Universal marker genes, such as the 16S rRNA gene, can be used for taxonomical classification if the gene is conserved in the species of interest and contains sufficient variability to distinguish clades. A tool that implements this approach is mOTU2 (Milanese et al. 2019), and it uses a set of 40 marker genes for profiling.

On the other hand, "clade-specific" marker genes are present in members of a clade but absent or too divergent in any other clades. This approach was pioneered by MetaPhlAn (Segata et al. 2012) and more currently MetaPhlAn2 (Truong et al. 2015). Using the marker database, clade abundance profiles can be estimated from read-mappings to the marker gene database, since reads are in general expected to map 'uniquely' to markers. For example, the Metaphlan2 gene database was built by first identifying core genes from all clades at all taxonomic levels. Then, the suitability of each candidate marker gene was evaluated by establishing a sequence identity threshold to check if the sequence is divergent enough from all other potential markers.

Marker-gene based strategies address many limitations from other strategies. Marker gene databases are considerably smaller than a full genome database, and it streamlines the abundance estimation by removing the problem of multimapped reads. With appropriate markers, the approach can be highly specific but sacrificing sensitivity (Vollmers, Wiegand, and Kaster 2017), since only a fraction of the metagenomic data will be used for the classification and the profile will be tied to the marker selection.

1.4.2.2 Strain-level population inference

Strain-level population analysis aims to recover strain variation from specific populations in metagenomics samples. Determining strain profiles requires fine-grained analysis, so strain-calling tools typically work on a per-species basis. The available tools use two main strategies to infer strain information: analyze gene content or extract single nucleotide variation(SNV). Ideally, these tools should be able to deconvolute the mixture of strains, but in practice, most of them focus on characterizing the most abundant strain or computing some form of a representative strain profile.

Tools that focus on gene content for strain profiling, such as PanPhlAn (Scholz et al. 2016), are based on the concept of a species pangenome: the complete set of genes present in any strain of a species. If the pangenome of a species is well known, it is possible to determine whether different samples contain different strains by examining the gene presence/absence profiles. The main advantage of this approach is that pangenome profiles enable direct functional interpretations: the presence or absence of genes with known function can be directly associated with any phenotype of interest.

On the other hand, SNV-based strain profilers aim to identify informative SNVs that can be used to distinguish between the strain content of the samples. Some tools like StrainEST (Albanese and Donati 2017) use a precomputed SNVs database collected from reference genomes, while other tools like metaSNV (Costea et al. 2017) and StrainPhlAn (Truong et al. 2017) call the SNVs directly from the sequencing dataset by mapping to a reference database (metaSNV maps to genomes, while StrainPhlAn uses the MetaPhlAn2 marker gene database).

Strain-profiling tools are sensitive to sequencing coverage requirements. For example, PanPhlAn requires a 'nearly uniform coverage' across the median number of genes in a strain to detect a strain in a sample, and imposes further coverage requirements for a gene to be called as 'present' (Scholz et al. 2016). This makes them applicable for high and medium abundance microbes, but not so effective for low abundance bacteria.

1.4.2.3 Functional profile inference

Functional analysis is concerned with the prediction of the coding potential present in the metagenome. In metagenome analysis, we are interested in identifying the functional potential of the metagenome as a whole, as well as understanding the coding potential of individual taxa. Read-based functional profilers mainly focus on characterizing the global functional potential by quantifying gene families and pathways from the reads. Associating genes with taxa using reads can be very challenging since the information in short sequences is often insufficient to distinguish between homologs.

Read-based functional analysis depends on one or more databases of gene and protein sequences to annotate the reads. Database searches can be performed at the nucleotide level or using translated searches using blastx or similar tools like Diamond (Buchfink, Xie, and Huson 2015) or RapSearch2 (Zhao, Tang, and Ye 2012). Once the mappings have been calculated, gene family level abundances are calculated by aggregating mapping to sequences in the databases using gene family annotations. Commonly used gene family annotations come from databases such as COG (Galperin et al. 2015), EggNOG (Huerta-Cepas et al. 2019) or KEGG (Kanehisa et al. 2016). Alternatively, it is also possible to map the reads to protein families directly with the HMMER suite (Mistry et al. 2013b) using databases such as Pfam (Finn et al. 2014) or TIGRfam (Haft, Selengut, and White 2003). Finally, pathway coverage and abundance can be calculated from the gene family abundances, using databases such as MetaCyc (Caspi et al. 2018) or KEGG.

In contrast with taxonomical annotation tools, there are fewer tools for read-based functional profile inference. HUMAnN2 (Abubucker et al. 2012), ShotMAP (Nayfach et al. 2015) and Fun4me (Sharifi and Ye 2017) are some examples of the available command line- based pipelines for functional characterization. Some online services exist as well, such as MG-RAST (Keegan, Glass, and Meyer 2016), IMG/M (I.-M. A. Chen et al. 2017) and EBI metagenomics (Hunter et al. 2014).

1.4.3 Assembly-based microbiome reconstruction

An alternative approach to read-based microbiome profiling methods consists of reconstructing the genomes present in the samples using sequence assembly. Metagenomics assembly can be considered a special case of the genome assembly problem, with the added complexity that different genomes are mixed in the same sample and the uneven fragment coverage due to differences in microbial abundances.

Different metagenomics assemblers have been published in recent years, most of them based on de Bruijn graphs. Common cited tools include Megahit (D. Li et al. 2015), SPAdes (Bankevich et al. 2012) or MetaSPAdes (Nurk et al. 2016), IDBA-UD (Peng et al. 2012) and Ray Meta (Boisvert et al. 2012). Although the choice of assembly tool is not trivial, the Megahit assembler is often recommended as a starting point based on assembly quality, resource usage and speed compared to other assemblers (Sczyrba et al. 2017; Ayling, Clark, and Leggett 2019).

Assembly strategies will differ depending on the number of samples, the sequencing depth, and community complexity (Ghurye, Cepeda-Espinoza, and Pop 2016). Ideally, samples should be reconstructed individually to maximize the capture of sample-specific variation. However, to improve the recovery of low-abundance microbes, samples can also be co-assembled together at the expense of disregarding individual sample variation. Assembly quality can be assessed by checking statistics such as the contig length distribution, checking contig read coverage, and measuring the percentage of read incorporation of the assembly. Reconstructed sequences are typically called ‘contigs’.

The next step after assembly is binning: the aim is to group (or ‘bin’) contigs that belong to the same genome. Available *de novo* binning tools rely on two main strategies for grouping contigs: sequence composition or coverage. Sequence composition-based binning is derived from the observation that the tetramer composition of fragments of the same genome will be similar. In contrast, coverage-based binning is based on the assumption that the coverage of contigs from the same genome will co-vary among different samples. State-of-the-art binning tools such as CONCOCT (Alneberg et al. 2014), Metabat2 (Kang et al. 2015) and MaxBin 2.0 (Wu, Simmons, and Singer 2016) use a combination of both strategies. Recently, a meta-binner called Das-Tool (Sieber et al. 2018) was released, to combines the results from different binners. Alternatively, genome bins can be inferred by taxonomically annotating each contig and grouping contigs with the same taxonomical annotation (Sczyrba et al. 2017).

Bins can be evaluated using different tools to determine the quality of the recovered genomes. Tools like checkM (Parks et al. 2014) or metaQUAST (Mikheenko, Saveliev, and Gurevich 2016) can be used for this purpose. Common metrics include contamination, genome completeness, and purity. Bins can be manually refined with the help of annotation and tools like Anvi'o (Eren et al. 2015). Finally, bins above certain quality criteria can be labeled taxonomically using tools like Kraken or MEGAN(Wood and Salzberg 2014; Huson et al. 2007), phylogenetically using PhyloPhlAn (Segata et al. 2013) and run through gene prediction and annotation pipelines like MetaGeneMark or Prokka. (Rho, Tang, and Ye 2010; Zhu, Lomsadze, and Borodovsky 2010; Seemann 2014).

Finally, the new genomes and their annotation can be used to infer the taxonomical abundance profiles of each sample and perform functional analysis both at global and per-genome level, using read-based tools or ad-hoc strategies similar to the ones implemented in read-based analysis tools.

1.5 Amplicon Sequencing: A cost-effective alternative

An alternative for microbiome profiling is marker gene or amplicon sequencing, also known as metabarcoding. Unlike shotgun metagenomics, the approach targets conserved genes in the genomes of microbes which can be used to infer the taxonomic and phylogenetic structure of the microbial community in a sample. In practice, most human microbiome studies use the 16S ribosomal RNA subunit (16S rRNA) gene to characterize bacteria and some types of archaea, since the gene is well conserved and with sufficient variation to distinguish between subclades (Olsen et al. 1986; Hugenholtz 2002). In this thesis, I will refer to amplicon sequencing to mean 16S rRNA gene amplicon sequencing unless noted otherwise. However, the same strategy can be extended to characterize archaeal (Gantner et al. 2011; Chaban and Hill 2012) and fungal populations (Lindahl et al. 2013) using different conserved genes.

The amplicon sequencing approach was instrumental for the first large scale analysis of human bacterial communities (E. a. Grice et al. 2009; Costello et al. 2009; Ravel et al. 2011) and for the establishment and execution of the Human Microbiome Project(HMP), a systematic effort to map the diversity of the microbiota in the body (The Human Microbiome Project Consortium 2012).

A typical 16S amplicon sequencing pipeline is described in Figure 1.5.1. The following subsections describe the data generation and community profiling processes.

1.5.1 Data generation

The first step for amplicon sequencing is DNA extraction, similar to shotgun metagenomics. In the library preparation step, a PCR is used to enrich for the desired conserved regions and sequencing adapters are ligated to the fragments. Then, the amplicons are forwarded for sequencing.

The 16S rRNA gene is around 1500bp in bacteria and contains 9 hypervariable regions that can be targeted for phylogenetic analysis. Due to the constraints of short-read sequencing length, a selection of hypervariable regions is required. This selection must be performed carefully, since the resolution of a variable regions to distinguish members of certain clades vary between variable regions and this impacts directly the downstream interpretation of results (Graspeuntner et al. 2018; Yang, Wang, and Qian 2016; Teng et al. 2018).

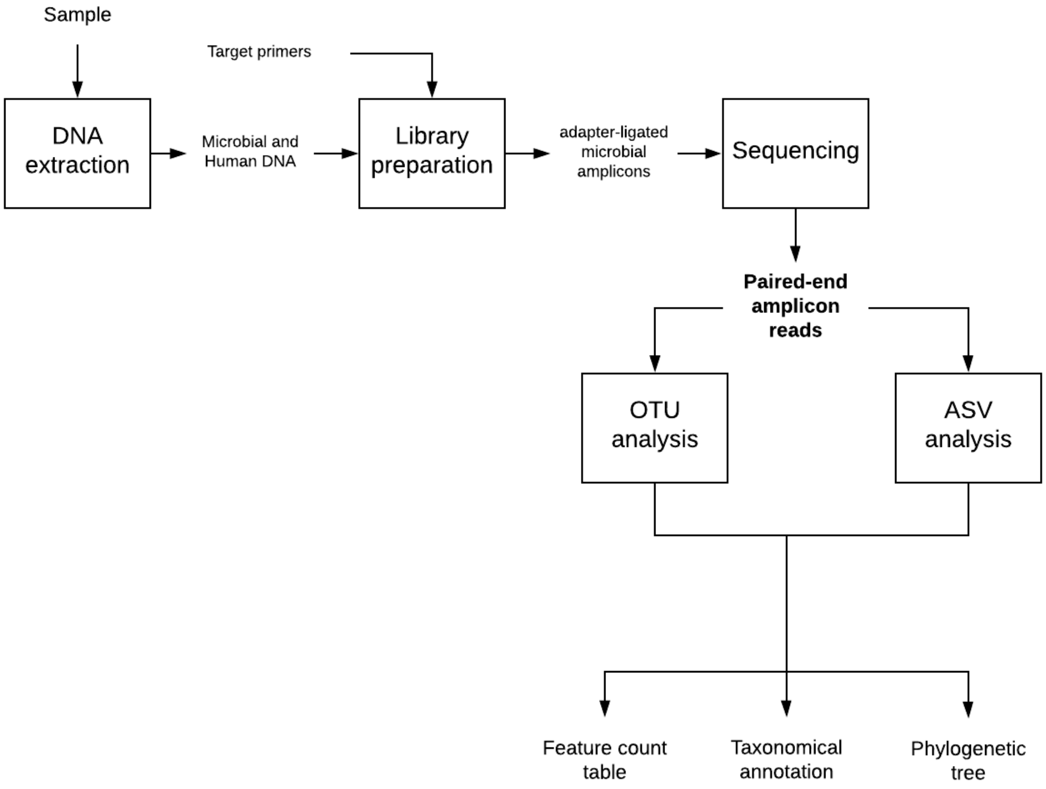

Figure 1.5.1 - 16S rRNA amplicon sequencing pipeline

16S rRNA-based microbiome profiling is a cost-effective method for microbiome studies; it avoids many of the pitfalls from shotgun metagenomics while enabling taxonomic profiling. The maturity of the technology has even led to its consideration for clinical diagnostics (Almonacid et al. 2016). For example, human DNA does not waste any sequencing since the 16S gene is not conserved in humans. Additionally, sequencing depth requirements are much lower since we are not attempting to reconstruct genomes, and thus lower abundance species are easier to detect. Bioinformatics analysis of 16S rRNA sequences is also more standardized compared to shotgun metagenomics (Hillmann et al. 2018).

The main drawback of the 16S rRNA amplicon sequencing approach is that it only yields taxonomical information, and species and subspecies (strain) resolution can be difficult to attain (Knight et al. 2018). Another issue with this approach is the dependence on PCR. It has been shown that the choice of primers and PCR cycles will bias the compositional estimation, as some primers will favor the amplification of certain clades (Eloe-Fadrosh et al. 2016). Additionally, PCR amplification carries the risk of chimera formation, which must be addressed during the analysis steps (Haas et al. 2011). Finally, while the approach can be extended for archaea and fungi, it cannot be extended for universal virome inspection due to the low level of gene conservation among viruses (Kuczynski et al. 2011).

1.5.2 OTU-based profiling

One way of describing the taxonomical profile of a microbiome sample based on amplicon data is to create Operational Taxonomic Units or OTUs. OTUs are clusters of amplicon sequences above a certain similarity threshold and represent the lowest-level "unit" of analysis.

Amplicon sequencing data processing starts with the quality control and filtering of the reads, removing low-quality bases and any adapters, primers or linkers used during the library preparation process. In the case of Illumina sequences, the quality trimming must take care that the paired-ends still overlap sufficiently so that they can be merged at a later stage.

The next step is to identify and correct errors in the 16S sequences. Denoising and chimera removal algorithms typically rely on error models and take advantage of unique sequence counts and base quality scores to predict which sequences are likely to contain errors. Then, sequences can be either corrected or discarded if they are PCR artifacts (Quince et al. 2011; R. C. Edgar et al. 2011).

Denoised sequences are subsequently clustered based on similarity to form the OTUs. An identity threshold should be selected; a 97% similarity threshold is commonly used in the literature (Konstantinidis KT 2005), but other thresholds have been suggested (R. C. Edgar 2018; Yarza et al. 2014; Nguyen et al. 2016). Subsequently, a representative sequence is selected for each OTU for downstream analyses, typically the cluster 'medoid'.

OTU clustering can be performed in three ways: open-reference, close- reference and de novo (Caporaso et al. 2010). In closed-reference OTU picking, the clustering is performed against a set of sequences (known as seeds) from a database like Greengenes (DeSantis et al. 2006) or Silva (Yilmaz et al. 2014). Sequences that do not cluster over the sequence identity threshold are discarded. Alternatively, de novo OTU picking compares each sequence in the dataset to each other and clusters them using the set identity threshold. De novo OTU picking is computationally expensive since each unique sequence must be compared to every other. An intermediate solution is to perform open-reference OTU picking, where sequences are subject to closed-reference OTU picking, and 'discarded' sequences are then subject to de novo OTU picking.

Once the OTUs have been formed, three main artifacts are produced to enable downstream analysis (McMurdie and Holmes 2013):

- a feature count matrix (OTUs vs samples)
- a phylogenetic tree
- taxonomical annotation of the OTUs

The **feature count matrix** is constructed by counting the reads belonging to each OTU in each of the samples.

The **phylogenetic tree** is built from performing a multiple sequence alignment from the representative sequences of each OTU, and then using the alignment to create a rooted phylogenetic tree.

Finally, sequences can be **taxonomically annotated** by comparing them to any database of sequences with taxonomic annotation. Typically, taxonomical classification is performed using a Naïve Bayes Classifier.

Clustering sequences into OTUs has several advantages for downstream processing and analysis. It helps eliminate left-over errors from the denoising process, as the sequences will most likely end up in the same OTU. Additionally, since the 16S gene is multi-copy, clustering can collapse paralogs into the same OTU, simplifying interpretation of the results. It also reduces the computational load of inferring phylogeny, since only the representative sequences will be considered (R. Edgar 2019).

However, the biological interpretation of OTUs is not straightforward because 1) OTU formation is sensitive to the choice of clustering algorithm and parameters (Mahé et al. 2014; W. Chen et al. 2013) 2) OTUs are dataset specific, which means they cannot be easily compared between datasets (Callahan, McMurdie, and Holmes 2017). 3) clustering can mask relevant biological variation from species and strains (Tikhonov, Leach, and Wingreen 2015) and 4) OTUs within a single dataset will correspond to different taxonomical levels, as there is no standardized definition of a bacterial species. 5) The taxonomical annotation can vary depending on the choice of database (Park and Won 2018).

1.5.3 ASV-based profiling

In recent years, the analysis of 16S datasets has shifted from the use of OTUs to the use of Exact Sequence Variants (ESVs) or Amplicon Sequence Variants (ASVs). ASVs give the maximal possible taxonomical resolution, allowing some ASVs to reach species or subspecies resolution. Additionally, it makes it easier to compare between datasets, since the sequences are directly comparable to each other (Callahan, McMurdie, and Holmes 2017).

The process of inferring ASVs from a dataset is almost identical to creating OTUs, except for the sequence clustering step. This step is substituted by a more complex denoising step that attempts to recover all ‘true’ sequence variants in the sample. The resulting ASVs can be annotated taxonomically, used to construct a feature count table (ASVs vs samples) and subject to phylogenetic analysis with the same methods as OTUs.

Three denoising pipelines are the most commonly used in the literature to produce ASVs: DADA2 (Callahan et al. 2016), Deblur (Amir et al. 2017) and UNOISE (R. C. Edgar 2016). All denoisers are Illumina-specific, although dada2 authors suggest the method can be applied to 454 sequences as well. The DADA2 denoising strategy relies on learning parametric error profiles from the sequencing data, and then employ a divisive partitioning algorithm to infer the true sequence variants. Deblur uses a per-sample approach, using Hamming pairwise distances calculated from multiple sequence alignments coupled with a parametric error profile specified by the user to identify noisy sequences. Deblur uses the UCHIME (R. C. Edgar et al. 2011) algorithm implemented in VSEARCH (Rognes et al. 2016) for chimera removal. The UNOISE3 algorithm is based on predicting ‘zero-radius OTUs’ or zOTUs. The general idea is to perform one-pass clustering such that the centroids of the cluster are inferred to be the ‘true sequences. For this, preset values for two parameters for the clustering have been optimized for different datasets.

A recent benchmark suggests the results from the three pipelines are mostly comparable, although the total number of predicted ASVs varies with the dataset. From a computational resource point of view, UNOISE3 is the fastest of the alternatives, but dada2 might be better at recovering low-abundance organisms(Nearing et al. 2018).

1.6 Tools and techniques to summarize and derive biological insights from the microbiome

1.6.1 Data distribution

The biological information obtained from both shotgun metagenomics and amplicon sequencing datasets after processing can be generally summarized as a set of abundance or binary matrixes of samples vs. features, with associated metadata for both features and samples.

In amplicon sequencing, the features will be OTUs or ASVs, and the feature metadata consists of the taxonomical annotation and phylogenetic tree relating the OTUs/ASVs. In the case of shotgun metagenomics, it is possible to infer a more varied set of features such as taxa, genes, gene families, pathways, or strains. The associated metadata varies depending on the type of data: e.g. a full taxonomical lineage for taxa, or some kind of categorical annotation for genes and pathways (e.g. Gene Ontology).

Abundance matrixes, as the name implies, store an abundance estimate of each feature in the samples. The estimates are inferred from some form of sequence counts (e.g. number of reads mapped). In the case of amplicon sequences, the raw abundance matrix consists of sequence counts, whereas in shotgun, the abundances can sometimes be already normalized or transformed by the tool that generated it.

Microbiome data often resembles transcriptome data, where the results are summarized in a gene or transcript count matrix. However, there are a few key differences. Microbiome data tends to be sparser, or zero-inflated (Jonsson et al. 2018a; L. Xu et al. 2015; Kaul et al. 2017): not all features are expected to be present in any sample, and different samples can have feature subsets with little overlap in content. Also, microbiome features are related along ‘natural hierarchies’ (e.g. taxonomy, phylogeny). This implies that features can be aggregated at different levels to create different features for downstream analysis.

1.6.1.1 Data normalization

Due to the relative abundance nature of the data, the most common normalization is performed to adjust the abundances to the sequencing depth of the sample. Some shotgun tools will output relative abundances (e.g. as percentages) directly. For shotgun data, abundance estimates can be also adjusted for microbial genome size or gene length.

1.6.1.2 Compositional data analysis

Quantification estimates from sequencing data, despite appearing as sequence counts, should be regarded as relative abundances (Gloor et al. 2017), since the total number of counts is constrained by the instrument's capacity. In other words, the abundances within a sample behave like percentages, where the features are dependent on each other. This has important implications for data analysis, because standard statistical techniques do not consider the dependencies between the features. This problem can be overridden if an estimate of total microbial abundance is available to scale the relative abundances into absolute abundance, but this is usually not the case.

The key property of relative abundances is that they store information about the ratios between the features. Ratios, unlike the relative abundances, are amenable to standard statistical techniques. Thus, different Log-Ratio data transformations have been proposed, each with different mathematical properties and interpretations. There are three main types: Additive Log Ratio (ALR), Centered Log Ratio (CLR), and Isoform Log Ratio (ILR) transformations(Quinn et al. 2018).

- ALR: Every feature is divided by a feature of reference and then log-transformed.
- CLR: Every feature is divided by the geometric mean of all the features.
- ILR: New features are created based on the concept of balances, which are ratios defined from a sequential binary partition of the features (e.g. a feature dendrogram).

The main issue with log-ratio transformations is the problem of features with zeroes (Silverman et al. 2018). Typically, zero imputation techniques are applied to deal with this problem (Quinn et al. 2018).

1.6.2 Global analysis – Introduction

To analyze microbiomes at the community level, we require methods to summarize characteristics of the entire community such as the microbe distribution or find ways to compare directly entire microbial profiles. Alpha and beta diversity measures accomplish this. Common packages to perform these analyses include vegan (Oksanen et al. 2019), microbiome (Lahti, Shetty, and et al 2017), phyloseq (McMurdie and Holmes 2013) and Qiime (Caporaso et al. 2010).

1.6.3 Global analysis – Alpha diversity

Alpha diversity is a measure describing the ‘local’ species diversity of a site. Although the use of ‘local’ can vary between ecologists, in human microbiome research one estimate of an alpha diversity index is calculated per sample. The distribution of the index is then analyzed according to the relevant sample groups (e.g. disease condition, anatomical site). Alpha diversity indexes typically consider the number of species (i.e. richness) and the distribution of species abundances (i.e. evenness) (Knight et al. 2018).

Species richness can be estimated directly from the number of observed species in a sample. However, due to limited sequencing depth, we expect the number of observed species to be an underestimate of the real richness value (Hughes et al. 2001). Therefore, some indexes like Chao1 and ACE aim to estimate species richness as the “effective number of species”. For example, Chao1 estimates the number of missing species based on the number of species with low counts (such as singletons and doubletons) (Chao 1984). ACE type estimators rely on the inverse of the singleton count (Hughes et al. 2001). Rarefaction curves can also be used to estimate richness, but it has been suggested to be suboptimal (McMurdie and Holmes 2014).

As mentioned previously, alpha diversity measures can also account for evenness or species abundance. In general, more ‘even’ samples (samples in which every species has similar numbers) are more diverse than samples where few species have high numbers and the rest are low abundant. The most commonly used index is the Shannon-Wiener index, based on Shannon’s entropy which measures the information entropy (or uncertainty) in a sample (Shannon 1948). Intuitively, the Shannon entropy measures the difficulty of guessing what species would come out if a random microbe was drawn from a community. A similar measure is the Simpson diversity index, which is formulated as the probability that a random sample of 2 species will result in both individuals being of the same species (Simpson 1949).

Alpha diversity indexes are useful proxies to identify differences or changes in the distribution of microbes between groups of samples. The measures can help identify total changes in the number of species in the samples, or differences in microbial abundances.

1.6.4 Global Analysis - Beta diversity

Beta diversity was formally defined as the ratio between local (alpha) diversity and total diversity (of the ecosystem under study). However, in human microbiome research, beta diversity indexes are used in practice as pairwise distance or dissimilarity measures to explore the variability within and between sample groups (e.g. clinical condition, sample type). Similar to alpha diversity indexes, different beta diversity measures make different assumptions and therefore have different biological interpretations. Some examples of commonly used beta diversity indexes include:

- **Jaccard index:** a metric that considers only the binary presence/absence of each taxon. It is calculated as the ratio of common species between both samples divided by the total number of unique species.
- **Bray-Curtis dissimilarity:** a semi-metric (does not necessarily satisfy triangle inequality) which considers not only the presence but also the abundance of the common taxa between the samples (J. R. Bray and Curtis 1957)
- **Yue-Clayton** is a similarity measure that considers both the abundance of common species and also the abundances of species unique to one of the samples (Yue and Clayton 2005).
- **Aitchison distance**, a compositional data analysis concept defined the Euclidean distance of Centered or Isometric Log Ratio transformed vectors can also be used as a beta diversity measure (Gloor et al. 2017).

The aforementioned indexes assume each taxon is equally distinct from each other. However, beta diversity indexes can also consider the phylogenetic distance between community members in the calculation. Indexes based on phylogenetic distances include:

- **UniFrac**: Considers only presence/absence of taxa, similar to the Jaccard index (Lozupone and Knight 2005).
- **Weighted and Generalized UniFrac**: Considers phylogenetic distance but also considers taxon abundance (J. Chen et al. 2012).

The beta diversity index should be carefully selected depending on the question under study. Indexes like the Jaccard or unweighted UniFrac weight low and high abundance microbes equally, whereas weighted UniFrac or Bray Curtis will highlight communities where the dominant microbes are similar. Similarly, the choice between phylogenetic-based and 'standard' indexes depends on what functions are the focus of the study. In the case of the human microbiome, closely related microbes can have very different roles in disease (e.g. Staphylococci in AD), so a phylogenetic-based measure could obscure this difference. However, when studying processes that are more likely to be conserved between related species, phylogenetic-based distance measure will identify better groups with similar core functions.

1.6.4.1 Statistical testing

Distances calculated from beta diversity indexes can be used to compare groups using a statistical test called Permutational Multivariable Analysis of Variance or PERMANOVA(Tang, Chen, and Alekseyenko 2016). The PERMANOVA tests for the null hypothesis that the centroids and dispersion of groups are equal. Other alternatives include the Analysis of Similarities (ANOSIM) test, which uses ranks to test whether similarities between two groups are different from the similarities within the groups, and the Mantel test (Anderson and Walsh 2013).

1.6.4.2 Ordination

Ordination is the process of projecting samples into a low-dimensional space for visualization. The most well-known ordination method is Principal Component Analysis (PCA), but the method is not appropriate for compositional data. It is however possible to apply PCA to CLR- or ILR-transformed data (Section 1.6.1.2).

In microbiome studies, distance-based ordination methods are quite common. Methods like Metric Multidimensional Scaling or (m)MDS can be applied to beta diversity metrics (i.e. not appropriate for Bray Curtis). Other methods like Non-metric Multidimensional Scaling, and more recently t-SNE (Van Der Maaten and Hinton 2008) and UMAP (McInnes et al. 2018) can also be used for ordination, but should be used with caution since they can introduce distortion(Cooley et al. 2019).

1.6.4.3 Clustering

Beta diversity indexes can also be used for cluster analysis: automatically find sample groups based on the features. Different clustering algorithms exist, with different requirements and complexity. Examples include k-means and its variants, mixture-model based clustering and density-based clustering such as DBSCAN. Another clustering technique commonly used in the microbiome research is hierarchical agglomerative clustering (HAC), which is used in conjunction with heatmap visualizations to relate features to hierarchical groupings.

Briefly, hierarchical agglomerative clustering is an iterative clustering algorithm that begins by placing each sample in an individual cluster. Then, at each step, the two clusters with the closest distance are merged into a new cluster. This is repeated until all clusters are merged into a single one, and this process generates a binary-tree structure which can be easily visualized with a dendrogram. An important consideration for HAC based clustering is the choice of cluster-to-cluster distances a.k.a linkage: alternatives include single-linkage, complete-linkage, weighted or unweighted average linkage clustering (UPGMA or WPGMA) among others.

1.6.5 Feature-level analysis

Another way to approach microbiome data is to examine the association between individual features (microbes or genes) and phenotypes or sample groups (like disease condition). Typically, the problem is framed as a comparison between groups (i.e. an association with a categorical variable), so it is also known as differential abundance analysis. For continuous measurements, it is possible to extend some models for continuous variables; alternatively, the discretization of values is an option.

Many different approaches and models have been used in the literature for differential abundance testing. The following subsections summarize the most common choices.

1.6.5.1 Non-parametric or normalization-based approaches

Many studies perform differential abundance analysis using non-parametric statistical tests such as Kruskal-Wallis or Wilcoxon test, as they make no data distribution assumptions. Another method is to arcsin-transform the relative abundance data and use multivariate linear regression for testing. This approach is implemented in tools such as MaAsLin (Morgan et al. 2012).

1.6.5.2 Count-based models - GLMMs

The RNA-seq differential expression problem is analogous to the microbiome differential abundance problem, so many authors have chosen to apply state-of-the-art tools of the transcriptomics field such as DESeq2 (Love, Huber, and Anders 2014) and edgeR (Robinson, McCarthy, and Smyth 2010) for microbiome differential abundance analysis. A recent benchmark suggests these methods perform well for microbiome data (Weiss et al. 2017a).

Transcript abundance data is commonly modeled as a negative binomial distribution of counts, but as discussed in section 1.6.1, microbial data tends to be sparser than transcriptome data. To address the excess zeroes that might not be captured by the overdispersion parameter of the negative binomial model, different models have been proposed for analysis: zero-inflated gaussian(ZIG) (Paulson et al. 2013), zero-inflated Poisson (ZIP) (Jonsson et al. 2018b), zero-inflated negative binomial (ZINB) (J. Chen et al. 2018) and the use of hurdle models (a two part model for 0's and a normal model for values) (L. Xu et al. 2015).

1.6.5.3 Compositional Data-based testing

The compositional data analysis field has developed methods for differential abundance testing as well. For example, the Analysis of Composition of Microbiomes (ANCOM) method (Mandal et al. 2015), derives a pseudo F statistic that can be used to determine statistical significance per feature. ANCOM makes two strong assumptions to simplify the problem with compositionality: 1) at least two of the tested features are not differentially expressed and 2) features are not all differentially expressed by the same amount. Another method for differential abundance analysis is ALDEx2, which performs classical statistical tests on CLR or ALR transformed data (Fernandes et al. 2013). In this case, the differential abundance result should be interpreted with respect to the chosen reference (e.g. geometric mean, or a linear combination of features). Finally, a very recent paper from the Knight lab describes a method based on the ranking of differentials (i.e. ratios of relative features between two groups) (Morton et al. 2019).

1.6.5.4 Machine learning

A less traditional alternative to identify features associated with metadata variables or phenotypes is to use machine learning models. One possibility is to use models from which a measure of variable importance for classification or regression can be calculated, such as decision trees or random forests. Another possibility is to combine machine learning techniques with classical statistical testing. For example, MaAsLin uses gradient boosting machines for feature selection but performs GLM based testing (Morgan et al. 2012).

1.6.6 Microbe-microbe interactions

Another task in the analysis of microbiome datasets is to infer relationships between microbial populations (or genes) in a community. From a data analysis perspective, it is possible to capture patterns of co-variation and use these to hypothesize about the relationships.

1.6.6.1 Network analysis

Microbial co-abundance networks are built with the same methods applied to create transcriptional co-expression networks. To create a network, the analyst can define one or more criteria to determine if two microbes share an edge. Common criteria include setting a minimum similarity threshold between the microbe abundance profiles. Examples of similarity measures include the Pearson or Spearman correlation, Euclidean distance, or the mutual information criteria. The compositional data analysis field has contributed the SparCC (Friedman and Alm 2012) and proportionality (Lovell et al. 2015) similarity measures as well.

Once the network has been constructed, different network analysis techniques can be applied depending on the question. For example, calculating network statistics, such as node degree and centrality, are helpful to identify microbes with a strong influence in the community. Additionally, community discovery algorithms can help identify subsets of microbes that depend on each other. Network visualizations can be constructed using force-directed layouts.

1.6.6.2 Frequent Pattern mining

Frequent Pattern Mining algorithms are designed to search efficiently for sets of items that co-occur together often in a set of transactions. For microbiome data, by defining items as microbes, and the transactions as samples, we can adapt these algorithms to identify all the sets of microbes that occur together ‘frequently’ in samples. In contrast with network analysis, pattern mining only considers the presence (and not the abundance) of the items (i.e. microbes). Additionally, these algorithms are not ideal to identify microbes with antagonistic relationships, since this can imply the microbes will often not co-occur.

Commonly used algorithms for pattern mining include a priori, ECLAT and FP-growth (Chee et al. 2018).

1.6.7 Data Integration

The questions relating the human microbiome to health and disease are becoming increasingly more complex. To answer these, the analysis requires the integration of different datasets, both microbial and human.

Data integration is a complex subject, and it typically requires large sample sizes for statistical power. In the case of microbiome studies, integration methods are developed ad hoc for the data and questions at hand. For example, Morgan et al. 2015 implemented a strategy to associate human transcriptome data to microbial data in inflammatory bowel disease. Another study, based on the LifeLinesDeep cohort, examined the relationship between host genetics and the microbiome integrating 16S data, shotgun metagenomics, genomics and other phenotype variables like fasting glucose levels (Rothschild et al. 2018).

However, there are also efforts to develop general tools to aid with general data integration. A promising recent development is Hierarchical All-against-All significance testing or HALLA (huttenhower.sph.harvard.edu/halla), which attempts to reduce the multiple testing burden by performing statistical testing based on a hierarchical aggregation of features. Other examples include tools like MVDA (Serra et al. 2015) which employs a multiview clustering approach, and the framework proposed by Pedersen et al. 2018.

1.7 Challenges in human microbiome characterization

The methods discussed in the previous sections have enabled the characterization of the diversity of the human microbiome in general, especially in the gut. A PubMed search on July 2019 revealed more than 15000 hits for gut microbiome studies compared to less than 10 000 hits for oral, skin and lung microbiome together. In the gut, diverse aspects of the microbiome have been explored, like the characterization of novel bacterial (Wylie et al. 2012) and viral (Dutilh et al. 2014) taxa, temporal dynamics (Caporaso et al. 2011) and bacterial genomic variation (Schloissnig et al. 2013). A human gut gene catalog has also been established (J. Li et al. 2014b) by combining data from other gut microbiome studies (Qin et al. 2010b, 2012).

However, many studies have also explored other body sites, like the respiratory tract (Dickson et al. 2016; O'Dwyer, Dickson, and Moore 2016; Huffnagle, Dickson, and Lukacs 2017) and the biogeographical and temporal dynamics of skin (Oh et al. 2014, 2016). The latest report from the Human Microbiome Project (HMP) focused on sampling the nose, the mouth and the vagina(Lloyd-Price et al. 2017). Data from 48 different body sites is available in the HMP data portal (https://portal.hmpdacc.org/search)

Other microbiome studies have focused on characterizing the human virome in different conditions. For example, (Minot et al. 2013) described the variability and temporal dynamics of the human virome in the gut, and Norman et al. 2015 examined the role of the virome in inflammatory bowel disease. Other studies have explored the virome composition in the lower respiratory tract (Lysholm et al. 2012) skin (Hannigan et al. 2015) and the oropharynx(Yolken et al. 2014).

Shotgun metagenomics and 16s rRNA sequencing have been instrumental in associating the microbiome to a wide range of diseases as well. Examples include obesity, diabetes and inflammatory bowel disease in gut (Gevers et al. 2014; Frank et al. 2007; Hartstra et al. 2015), urinary tract infections (Stapleton 2016), atopic dermatitis and psoriasis in skin (Kobayashi et al. 2015; Gong et al. 2006; Waldman et al. 2001; Alekseyenko et al. 2013) and neurological disorders like depression and anxiety(Foster and McVey Neufeld 2013; Zheng et al. 2016). The microbiome has also been linked to cancer (Neto, Whitaker, and Pei 2016; Guma et al. 2016).

However, there are still many open questions and challenges regarding the role of the human microbiome in health and disease. The following subsections summarize current challenges in two areas of human microbiome research that are addressed in this thesis: virome exploration and the association between skin microbiome and disease.

1.7.1 Virome characterization

From the different components of the human microbiome, the virome is the most understudied (Virgin 2014). This fact is not surprising if we examine the numbers of viral genomes available in public databases. To July 2017, the NCBI genome database contains 205 692 entries for prokaryotic genomes compared to only 32 212 viral genome entries. Moreover, it has been estimated that we have only explored about 1% of the global virome at the sequence level (Mokili, Rohwer, and Dutilh 2012).

We have yet to decode most of the variety of the human virome. One effort in this direction is viral discovery projects, which aim to discover novel viruses by studying diseases with unknown etiology. Clinically relevant viruses such as the Human Bocavirus (Allander et al. 2005), old-world arenavirus (Palacios et al. 2008), and the Merkel cell polyomavirus (Spurgeon and Lambert 2013) have been discovered with these approaches. However, careful result validation is paramount in these kinds of studies, since they can often result in false positives due to contamination (Naccache et al. 2013).

Viral discovery projects hold great potential, since many complex diseases of unknown etiology are hypothesized to be caused or mediated by infections (Kraszewska-Głomba, Matkowska-Kocjan, and Szenborn 2015; Virtanen and Jacobson 2012; Pou et al. 2018). One example of such diseases is childhood acute lymphoblastic leukemia or ALL. There is evidence of pre-leukemic clones *in utero* developing into ALL later in life (Wiemels et al. 2002; Maia et al. 2004). DNA viral infections have been hypothesized to play a role in generating the aberrant clones that give rise to ALL, but attempts to identify a candidate virus using traditional methods have yielded no results (Gustafsson and Carstensen 1999; Isa et al. 2004).

Another challenge in the study of human viruses is to understand the link of many 'asymptomatic' viruses to health and disease. Most individuals are estimated to carry at least 10 asymptomatic viral infections during their lifetime (Virgin, Wherry, and Ahmed 2009). One hypothesis is that these viruses play an important role in immunomodulation. For example, it has been observed that GBV-C, an asymptomatic chronic infection, confers some protection in the presence of HIV infection (Lanteri et al. 2015). Characterizing and studying the interactions between asymptomatic viruses, such as the giant marseille-like virus (Popgeorgiev et al. 2013), anelloviruses and human endogenous retroviruses (L. Li et al. 2013), with our immune system might help us elucidate mechanisms of immune system development, maturation and their potential involvement in immune pathologies such as allergies or autoimmunity.

Also, the human body is colonized by bacteriophages (Dutilh et al. 2014; Minot et al. 2013; Hannigan et al. 2015) which could have a direct impact on human health (Barr et al. 2013; Kernbauer, Ding, and Cadwell 2014; Hsu et al. 2019) , and disease (Tetz and Tetz 2016). Interestingly, the human prokaryotic virome appears to be stable across time. Oh et al. 2016 observed that while the human skin virome is quite variable, the phageome is more stable. This is consistent with observations in saliva (Pride et al. 2012) and gut (Minot et al. 2013), which found an 80% conservation of viral diversity during a 2.5 year follow up period. Although it has been suggested that this stability is intrinsically linked to the stability of the bacterial microbiome, the specific mechanisms behind phage modulation of microbiota are still mostly unknown.

Shotgun metagenomics projects should theoretically facilitate the study of the entire viral fraction of any sample, but practical limitations hamper its effectiveness. For instance, viral fragments are often orders of magnitude lower in abundance compared to host and bacterial fragments after library preparation, rendering them hard to detect. To address this shortcoming, many virome studies opt for viral particle enrichment. The most common enrichment procedures are based on filtration, centrifugation and homogenization. These protocols are effective at reducing the host and bacterial load, but they can also bias the viral composition (Conceição-Neto et al. 2015).

Virome characterization also presents specific challenges at the data analysis level which are usually not accounted for by more general microbiome analysis tools (Rose et al. 2016). The main challenge is the recovery and characterization of unknown viruses with no homology to known species (Hurwitz, U'Ren, and Youens-Clark 2016). Other challenges include low coverage and high within-species and within-population genomic heterogeneity, which complicate the accurate reconstruction and annotation of viral fragments. For this reason, researchers have developed specialized methods, such as assemblers that address the variability of viral populations (Hunt et al. 2015), and methods for viral haplotype reconstruction (Prosperi and Salemi 2012; Giallonardo et al. 2014). A comprehensive review of these challenges has already been addressed in Rose et al. 2016.

1.7.2 The Skin microbiome

The skin microbiome plays an important role in maintaining homeostasis. For example, the microbiota is involved in processes like perspiration (Natsch, Schmid, and Flachsmann 2004), immune system modulation (Lai et al. 2010) and in the maintenance of barrier function of the skin, where the microbial communities exert defense against pathogens through colonization resistance mechanisms (Otto 2010).

One of the most salient characteristics of the skin microbiome is that it is not a global homogeneous community, but rather a heterogeneous collection of communities in balance with their microenvironment (E. a. Grice et al. 2009; Costello et al. 2009). These microenvironments are divided into three categories: dry, sebaceous and moist (E. a. Grice et al. 2009), and are characterized by different physiological conditions such as humidity, temperature, pH, nutrients, and antimicrobial peptide composition (Schommer and Gallo 2013). The microenvironments, in conjunction with the immune system, help determine the composition of the microbial community residing in any given site. In fact, body sites have been found to be very different in terms of bacterial species richness, diversity, stability and species composition (Oh et al. 2014, 2016) Moreover, body site diversity is so striking that diversity between body sites within the same individual is greater than diversity in the same body site between individuals (Oh et al. 2014) , but stable over time (Oh et al. 2016).

Designing and executing a good skin microbiome study is challenging, despite being easily accessible for sampling. The selection of the appropriate sampling strategy is paramount. Besides skin site variation, both the sampling method and level of skin under study have been shown to influence the final characterization (E. A. Grice et al. 2008; Nakatsuji et al. 2013). In addition, many studies also control for external factors that could influence the microbiota and therefore confound the analysis. Examples of these include antibiotic use (E. a. Grice et al. 2009; Costello et al. 2009), topical medications and emollients (Kong, Oh, Deming, Conlan, Grice, Beatson, Nomicos, Polley, Komarow, Mullikin, et al. 2012) and washing or hygiene habits(Oh et al. 2014). Another factor to consider when designing a study is that skin microbiome is low-biomass, so there is a high risk of contamination (Kong et al. 2017).

The skin microbiome is very rich, hosting not only a wide range of bacteria but also viruses, fungi and other eukaryotes (Oh et al. 2014, 2016; Lacey, Ní Raghallaigh, and Powell 2011). From the bacterial side, most studies suggest there are 4 dominating bacterial phyla on healthy skin: Actinobacteria, Firmicutes, Proteobacteria and Bacteroidetes. It has also been observed that the Propionibacterium, Staphylococcus and Corynebacterium genus are common in most skin sites(E. a. Grice et al. 2009; Costello et al. 2009).

Fungal and viral species have also been observed in the skin. The most common fungus is *Malassezia spp.* (Oh et al. 2014, 2013), but other fungi such as *Candida spp*.(Waldman et al. 2001) and dermatophytes (Findley et al. 2013) are commonly associated with skin. Other eukaryotes such as Demodex mites (Lacey, Ní Raghallaigh, and Powell 2011) have also been associated with the skin. Human viruses such as papillomavirus and polyomavirus are frequently detected in skin (Foulongne et al. 2012; Hannigan et al. 2015; Schowalter et al. 2010; Oh et al. 2016), although Poxviridae and Circoviridae have also been observed (Oh et al. 2016; Foulongne et al. 2012). Polyomaviruses are of particular scientific interest, since two previously unknown strains are shed chronically from healthy skin, but are yet to be linked with any disease, while another member of the family has been associated with cancer (Schowalter et al. 2010). Phages like Staphylococcus and Streptococcus phages were found ubiquitously in Oh et al. 2016, but their relevance is still unclear.

The skin microbiome has long been considered to play a role in many skin pathologies. With the microbiome revolution, we have begun to elucidate its effect in diseases such atopic dermatitis (Kong, Oh, Deming, Conlan, Grice, Beatson, Nomicos, Polley, Komarow, Mullikin, et al. 2012), psoriasis (Fry et al. 2013) and other diseases like acne and rosacea (Fredricks 2001). In this thesis, we focused on atopic dermatitis and psoriasis, since these diseases can be used as models to study the allergy and autoimmunity.

1.7.2.1 Atopic dermatitis

Atopic dermatitis (AD), is a chronic allergic disease associated with genetic defects in the skin barrier function. This is supported by evidence that 10 to 30% of patients suffer from mutations in the *FILAGGRIN* gene (S. J. Brown and Mclean 2012). The skin microbiome is suspected to influence the development and severity of AD (Williams and Gallo 2015). For instance, Kong et al. 2012 observed a link between disease severity and lower bacterial diversity, and a corresponding increased diversity post-treatment, supporting the hypothesis.

Most microbiome studies in AD have focused on *Staphylococcus aureus*, which colonizes 90% of patients (Rudikoff and Lebwohl 1998). *S. aureus* is believed to be the main driver of inflammation and it is closely linked to the severity of the flares (Gong et al. 2006;

Pascolini et al. 2011). Proposed mechanisms are *S. aureus* virulence factors such as the alpha- and delta- hemolysins (Brauweiler, Goleva, and Leung 2014; Nakamura et al. 2013). Interestingly, Kobayashi et al. 2015 demonstrated a dysbiosis-mediated mechanism for *S. aureus* colonization that results in inflammation, highlighting the role of other local microbial players.

Other suspected players in AD are *Staphylococcus epidermidis* and *Malassezia spp*. *S. epidermidis* is a skin commensal involved in immune modulation (Lai et al. 2009; Naik et al. 2012, 2015). In the context of AD, *S. epidermidis* has been shown to inhibit *S. aureus* growth and biofilm formation, through the production of antimicrobial peptides (Sugimoto et al. 2013; Cogen et al. 2010) and immune modulation (Lai et al. 2010). However, its role in the disease is still unclear, as Kong et al. 2012 observed a corresponding increase in *S. epidermidis* during flares. Whether *S. epidermidis* has a commensal relation with *S. aureus* or grows as a counter-response to the increase of S aureus is an open question. The involvement of *Malassezia spp.* in AD is also unclear, as there is conflicting evidence in favor and against its implication (Jo, Kennedy, and Kong 2016).

Another open question in AD microbiome research is the role of skin site with respect to the skin microbiome. Skin site is known to be an important determinant of the healthy skin microbiome (Oh et al. 2014; E. a. Grice et al. 2009), but skin variability in AD is not well studied. A recent study by Baurecht et al. 2018, suggest that there is a decrease in diversity between skin sites compared to healthy skin.

1.7.2.2 Psoriasis

Psoriasis is an autoimmune disease characterized by keratinocyte hyperproliferation and immune cell infiltration. There are 5 main types of psoriasis based on the clinical presentation, and while genetic predisposition factors have been identified, the causes remain elusive (Parisi et al. 2012).

The role of the microbiome in psoriasis is less understood compared to AD, but some associations have been identified. For example, psoriasis is generally characterized by a lower bacterial diversity on the skin (Fahlén et al. 2012; Gao et al. 2008; Alekseyenko et al. 2013) but increased fungal diversity compared to healthy skin (Takemoto et al. 2015). Statnikov et al. 2013 proposed that the skin microbiome could be used as an accurate marker for psoriasis diagnosis. Several studies also linked the guttate form of psoriasis to Group A Streptococcal infection (Leung et al. 1995). The role of fungal microbiota in psoriasis is of special interest, as lesions can resemble the ones caused by some fungal infections. Several studies focusing on Malassezia spp. and Candida spp. exist, but no definite association has been proven (L. C. Paulino et al. 2006; Luciana C. Paulino, Tseng, and Blaser 2008; Waldman et al. 2001). A recent shotgun metagenomics study suggests that key insights about the influence of the microbiome in psoriasis can be obtained by studying strain-level differences in the microbiota (Tett et al. 2017).

The skin microbiome field is growing rapidly, and the increasing evidence of its involvement in diseases such as atopic dermatitis and psoriasis call for increased efforts to characterize the microbiome profiles in-depth, which in turn might result in better diagnostics and treatment for these maladies.

2 AIMS

The overall aim of this thesis is to explore ways to improve the state-of-the-art methods used to investigate the human microbiome and its role in health and disease. The specific aims of the constituent studies are as follows:

- To propose a bench-to-bioinformatics viral discovery pipeline to explore hypotheses related to unknown etiological agents causing disease.
- To utilize the viral discovery pipeline to explore as a way to explore viral-mediated pathogenesis hypothesis of childhood Acute Lymphoblastic Leukemia (Paper I)
- To propose a method for the discovery of novel viruses in virome datasets (Paper II)
- To establish a methodology for the exploration and hypothesis generation of human skin microbiome datasets
- To apply our exploration methodology on a human skin microbiome dataset from atopic dermatitis and psoriasis patients (Papers III & IV)

3 RESULTS AND DISCUSSION

3.1 Papers I & II: Viral discovery

One strategy to explore the human microbiome and its role in health and disease is to investigate the role of viruses in diseases with cases of unknown etiology. Studies that use shotgun metagenomics for this purpose are commonly categorized as Viral Discovery projects. A typical viral discovery study starts with a disease of interest, where a fraction of the cases is of unknown etiology and ideally some indication of a viral infection playing a role in pathogenesis exists. A cohort of patients suffering from the disease are subsequently identified and relevant clinical specimens are collected. The choice of clinical specimens must be carefully selected based on the working hypothesis of the study. Specimens can then be pooled and then run through a shotgun sequencing-based pipeline for virome characterization, which results in a list of candidate sequences for validation. Viral candidates resulting from the analysis can then be confirmed in the samples via PCR or other methods, and based on these results, follow-up studies for genome reconstruction (in the case of a novel virus) and targeted association studies with the new pathogen can be designed.

The core element of any Viral Discovery project is the shotgun metagenomics-based pipeline for virome characterization. For this thesis, we propose a bench-to-bioinformatics pipeline, which is summarized in Figure 3.1.1. The data generation takes place in four steps: **a)** The pipeline takes as input samples or sample pools and performs a viral particle enrichment. In our implementation, the viral enrichment procedure consists of a filtration step, using filters of 0.45 and 0.22 nanometers, or a low-speed (1000g) centrifugation step for 5 minutes at room temperature to pellet bacterial and human cells. Tests on the enrichment procedure showed that low-speed centrifugation was similarly effective than filters at enriching for small viruses, and less prone to discard larger viral particles (such as Cytomegalovirus). **b)** After the viral enrichment, we aliquot the sample and run separate DNA and RNA extractions to ensure the capture of both DNA and RNA viruses. **c)** Unbiased amplification and library preparation are performed to account for the low nucleic acid concentration and to ligate adapters for sequencing. In our implementation, we use sequence-independent single primer amplification (SISPA), a common choice for virus studies (Reyes and Kim 1991; Chrzastek et al. 2017) **d)** Finally, the resulting libraries are submitted for sequencing. We use Illumina sequencers, typically MiSeq due to its longer read length (2x300bp) or NovaSeq or HiSeq if higher coverage is desired.

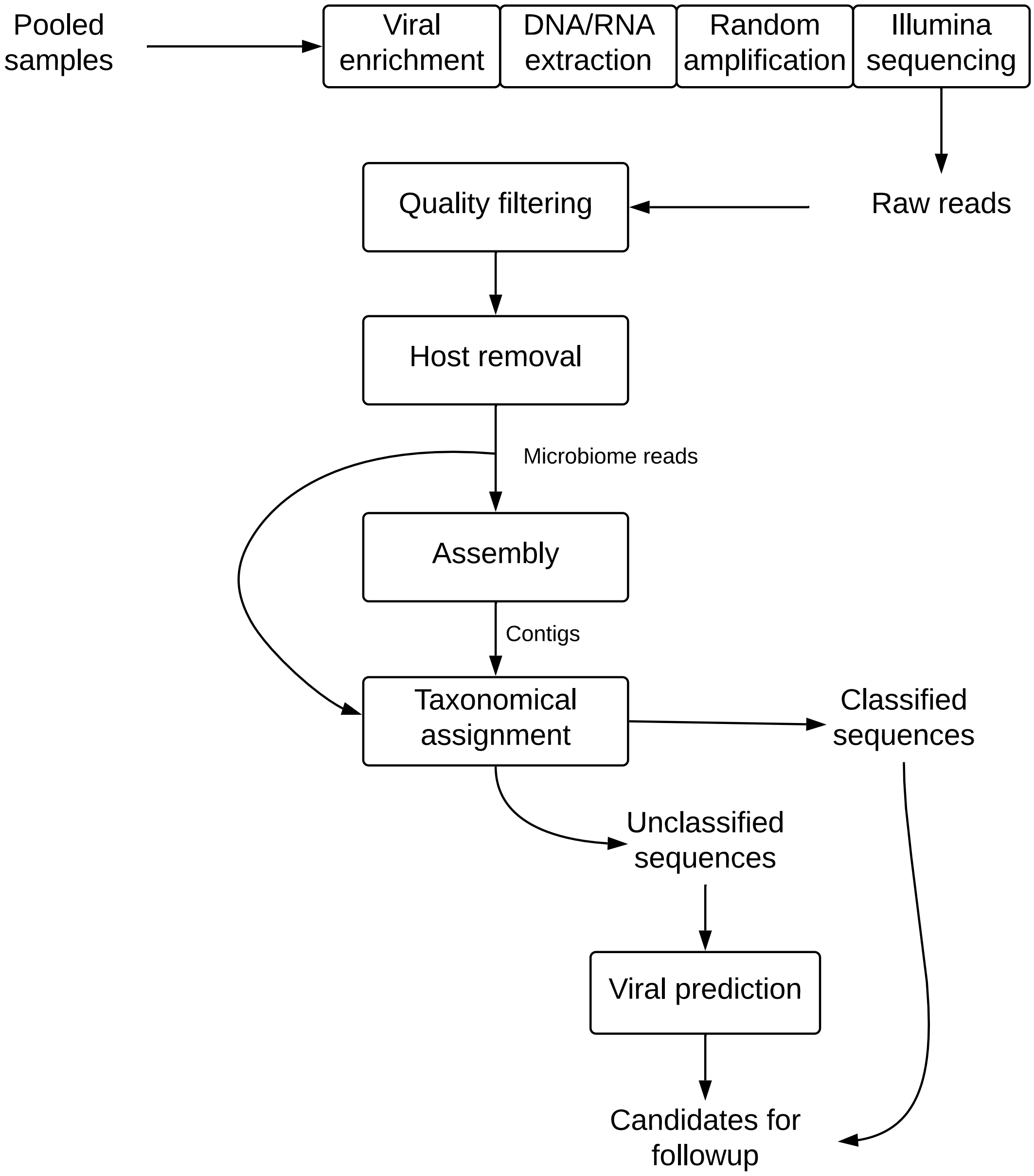

Figure 3.1.1 Bench-to-Bioinformatics viral discovery pipeline

The sequencing data is analyzed through our bioinformatics pipeline, which can be split into two ‘conceptual’ modules: preprocessing and discovery. The preprocessing module starts by removing adapters, both from Illumina and SISPA amplification and trimming low-quality bases, which typically occur at the end of Illumina reads. Then, the pipeline performs a filtering step to remove human sequences by mapping the quality-filtered reads against the human genome reference. This step yields a set of reads we label as “microbial reads” and an estimate of microbial-to-human ratio, which varies greatly with the type of clinical specimen, the condition (disease vs. healthy) and the viral enrichment procedure. The human filtering step reduces the downstream processing load and removes the identifiable human sequences to ensure patient anonymity.

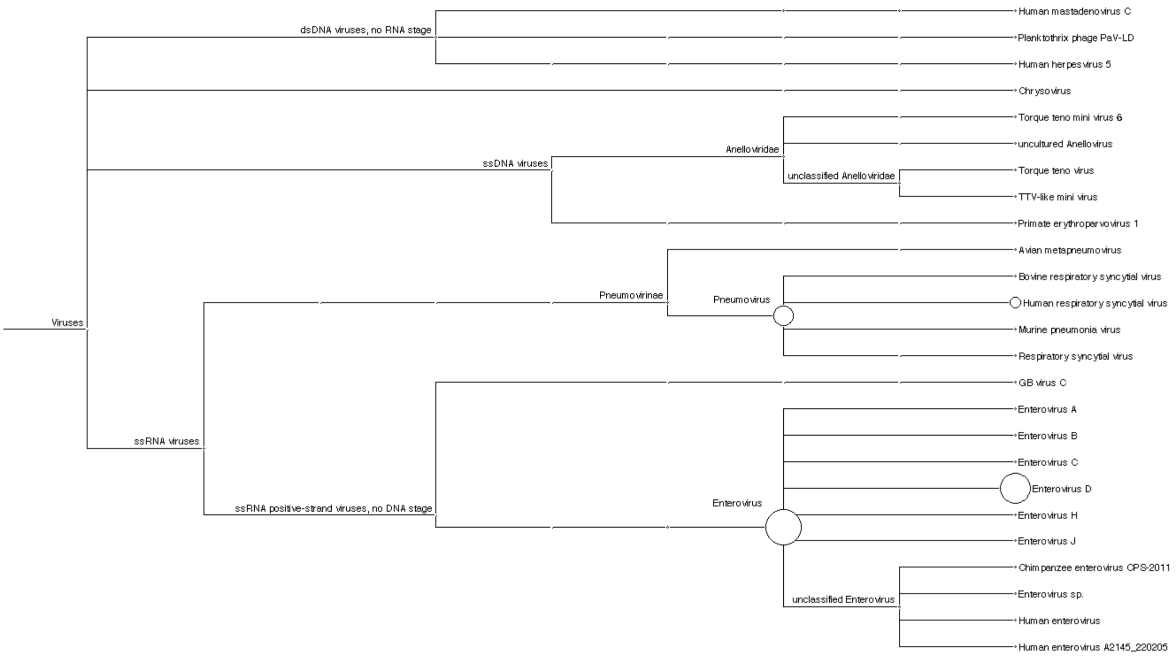

Figure 3.1.2 – MEGAN output of the validation of the viral discovery pipeline with a mock sample

The discovery module takes the filtered microbial reads and processes them using two different strategies: 1) reads are analyzed directly with read-based taxonomical inference tools, or 2) reads are through metagenomics assembly pipelines, and the resulting contigs annotated downstream. Earlier iterations of the pipeline (used in papers I & II) relied heavily on blast-like searches for annotation and LCA tools such as MEGAN for taxonomical annotation. However, method developments for efficient taxonomical profiling and taxonomical sequence annotation have become available, and these have been incorporated into the latest versions of the pipeline: https://github.com/maubarsom/VD_pipeline

We validated the suitability of our pipeline for viral discovery by assembling a mock sample and processing it through our method. The mock was created by combining four clinical specimens positive for DNA and RNA pathogens: parvovirus (plasma, ssDNA), adenovirus (nasopharyngeal aspirate, dsDNA) and cytomegalovirus (CMV or HHV-5, serum, dsDNA), respiratory syncytial virus type A (RSV, tracheal secretion, ssRNA) and enterovirus (nasopharyngeal aspirate, ssRNA). The sample containing CMV(HHV-5) was merged in a lower proportion to assess the sensitivity of the integrative approach to detect low copy viruses. The pipeline was able to recover sequences from four out of the five viruses including CMV which was merged at a lower proportion (Figure 3.1.2).

3.1.1 Paper I: Inspecting the roles of viral infections in the pathogenesis of Acute Lymphoblastic Leukemia

Paper 1 is an application of our pipeline to examine the possible role of a viral infection in the pathogenesis of childhood acute lymphoblastic leukemia (ALL). In this study, we inspected the virome of neonatal bloodspots (NBS) taken from newborns who later developed ALL. The virome of ALL children was then compared to the virome of healthy children. Viruses from the normal human flora were identified, and sequences belonging to pathogenic viruses such as human herpesvirus 6 and parvovirus B19 were detected.

To validate the results, we designed primers specific to the detected viruses and tested the sequenced pools and the individual clinical specimens by PCR. The presence of HHV-6 was confirmed in 2 patients and 3 controls, while 1 patient and 0 controls were positive for parvovirus B19. These findings, despite being potentially interesting candidates, do not suggest the existence of an association between the observed viruses and the disease.

Paper I illustrates many of experimental design and implementation challenges that arise in viral discovery projects. The goal of the study is to find evidence of a viral infection in children that developed ALL, to serve as a basis for subsequent studies to determine possible causality and elucidate mechanisms. This study chose to inspect neonatal bloodspots, since they are sampled close to the date of birth and being convenient for research purposes. Neonatal bloodspots could in principle contain traces of the infection, but depending on the time of the event, evidence of the causal viral infection is likely to be cleared. Better sampling strategies such as periodically collecting maternal blood or amniotic fluid over the course of a pregnancy in a prospective study could be envisioned. However, without other evidence, a project of this scale would be hard to justify.

Another challenge, which is common to most viral discovery projects, is the choice of pooling strategy and sequencing depth. Appropriate choices for these variables will depend greatly on the disease under study, the number of patients and the type of samples. Factors to consider in the decision include the 1) expected prevalence of the virus in the population of interest (from which we will obtain our samples), 2) the virus load and especially in relation to human and bacterial load in the clinical specimens of interest, and 3) whether the pathogen is completely unknown or at least a close relative to known ones. Since many of these are impossible to know or estimate beforehand, these choices commonly made based on a combination of previous experience and economical constraints.

From the bioinformatics perspective, our pipeline depends heavily on database searches to characterize the virome population. However, this approach is insufficient to detect viruses that are unrelated or very distantly related to known viruses; database searches would yield either no matches or misclassifications. Considering most of the viral sequence biodiversity is unknown(Virgin 2014), we need to complement our strategy with other approaches to detect uncharacterized viral sequences.

3.1.2 Paper II: ORFan protein prediction in viral metagenomics datasets

The aim of Paper II was to prototype a novel bioinformatics strategy to detect novel protein families in metagenomics datasets and apply the method to explore the set of unannotated sequences from viral-enriched human metagenomics libraries. The motivation behind the project is that the prediction of novel or 'orfan' proteins in virome datasets can be used as anchor points for the discovery of novel viruses.

An outline of the method is available in Figure 3.1.3. The method starts with a set of unannotated sequences from metagenomics samples and clusters them based on sequence identity. The aim of the clustering step is to group sequences that potentially belong to the same gene family. Then, sequences with identity > 95% within each cluster are merged into a consensus sequence, since they likely represent gene fragments from the same species. Subsequently, the reduced clusters are subject to codon-aware multiple sequence alignment. The resulting MSAs (which belong to a potential gene family) are trimmed and split (manually) into smaller subalignments to improve the quality of the alignments. Finally, the curated alignments are evaluated using RNAcode for coding potential, and predictions are ranked using the RNAcode score, alignment length and sequence complexity.

We applied the method to an unannotated set of sequences from a collection of viral discovery libraries generated from serum, nasopharyngeal and throat swabs, feces, and cerebrospinal fluid (CSF). This resulted in a set of 32 high-quality candidate 'orfan' proteins. Homology searches for protein annotation revealed that most gene families had been recovered in other metagenomics datasets such as the MetaHIT human gut microbiome gene catalog(J. Li et al. 2014a) but only 6 were functionally annotated. Additionally, the method also predicted gene families that were not present in any of the inspected metagenomic datasets, and from one of these, it was possible to reconstruct a 5.7kb circular genome with some evidence of phage origin.

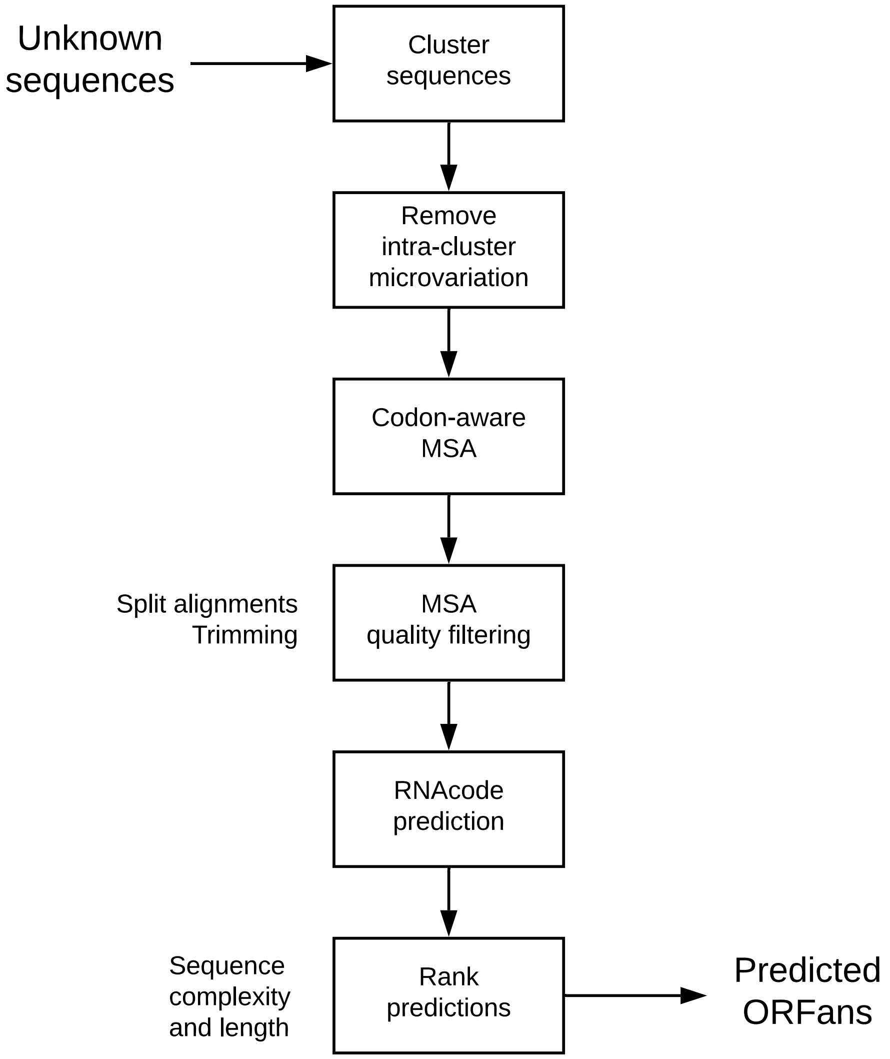

Figure 3.1.3 - ORFan protein prediction method summary

Our results suggest that the method is effective to recover ORFan proteins from viral metagenomics datasets. However, the implementation of the method requires further work to become widely applicable in practice. For example, the main clustering step with MCL together with the single-linkage subclustering on each cluster to reduce redundancy were very time-consuming. Reimplementing the method with state-of-the-art methods for clustering such as VSEARCH (Rognes et al. 2016) or MMSeqs2 (Steinegger and Söding 2017) could improve the speed and throughput of the method without compromising cluster quality. The other main limitation lies in the curation of high-quality multiple sequence alignments. The choice of a codon-aware alignment tool such as MACSE was key, but without automatized methods for alignment curation, the scalability of the method is severely limited. The development of methods in this area would improve the scalability of this method greatly.

From a viral discovery perspective, these results also suggest that predicting novel ORFan can be effectively used to recover unknown viral genomes. The recovery of the bacteriophage-like genome in paper II was based on one of the 32 predicted ORFan proteins. The presence of the fragment in the fecal sample pool was confirmed by PCR and subsequently tested in the individual fecal specimens. From these, only one out of 10 individuals tested positive for the fragment. Interestingly, this observation provides further evidence about the efficacy of our viral discovery pipeline, despite possible limitations imposed by the pooling and sequencing depth.

The genome reconstruction follow-up was performed using an inverse PCR, using primers designed from the new protein family sequences. This resulted in an amplicon of around 7 kbps, suggesting the genomic fragment was circular. We then sequenced and assembled the fragment using a combination of Illumina and Sanger sequencing, with the help of the IVA assembler(Hunt et al. 2015). Annotation of the circular fragment was performed using ab initio gene prediction and homology searches with HMMER3 (Mistry et al. 2013a). Phylogenetic placement was based on one ORF with Pfam (Finn et al. 2014) annotation. An alternative for genome recovery in other viral discovery projects could involve pooling positive-testing samples and performing deep shotgun sequencing together with viral enrichment strategies.

The role of bacteriophages in human health is not well understood, but studies in recent years have revived the interest in the field. For example, Hsu et al. 2019 recently demonstrated the impact of phage predation on the microbiome and its impact on the metabolome in a mouse model, suggesting that the phages could actively play a similar role in humans. Additionally, phage therapy is regaining traction as an alternative to antibiotics, with recent studies suggesting it can fend off bacterial infections (Hu et al. 2018; Schooley et al. 2017; Chan et al. 2018). Thus, the characterization of phages could have important repercussions in the fight against antibiotic-resistant bacteria in the near future (Kortright et al. 2019).

3.1.3 Conclusions

In conclusion, paper I and II showcase the development of effective methods to execute viral discovery projects. However, continuous research is needed to improve the quality of virome characterization, both from the data generation and bioinformatics angles. For example, the development of efficient methods for unbiased viral enrichment would drastically impact the quality of virome characterization and novel virus discovery. From the bioinformatics side, novel strategies for the characterization and annotation of unknown viruses are required. Tools based on machine learning techniques like VirFinder (Ren et al. 2017) and Metavir2 (Roux et al. 2014) are pushing the boundaries for novel virome characterization, but there is plenty of room for development.

3.2 Papers III & IV: Associating the human skin microbiome to disease

Another way to study the influence of the microbiome in health and disease is to sample and characterize human microbial communities of interest and study the community variation in relation to the phenotype of interest (e.g. a comparison between communities from healthy individuals and patients suffering from a disease). In this thesis, we studied a cohort of Atopic Dermatitis (AD), Psoriasis (PSO) and healthy controls to understand the role of the skin microbiome in allergy and autoimmunity, using AD and PSO as model diseases. One lesional skin sample and one non-lesional skin sample from matching skin locations were collected from each patient and every sample was processed with 16S rRNA sequencing, shotgun metagenomics for microbiome data and microarrays for human transcriptome profiling. Extensive clinical data from each study participant was also collected for analysis.

The results of the analysis of the cohort are presented in papers III and IV. Paper III describes the results from an open-ended exploratory analysis to examine the role of the skin microbiome in AD and psoriasis versus healthy skin. In this study, lesional samples from both diseases were compared to the samples from healthy individuals. Paper IV focused on understanding the microbial and transcriptional variability in atopic dermatitis across skin sites. This study was motivated by the observation that few studies had analyzed diseased skin-site variability as thoroughly as it has been studied in healthy skin (Segre 2014, 2016). In paper IV, we analyzed both lesional and non-lesional sample from Atopic Dermatitis patients and the healthy controls samples.

The bacterial taxonomical profile for both papers was obtained from the V3-V4 16S rRNA 454 pyrosequencing data using the QIIME pipeline. Open OTU picking was performed against the Greengenes database (13.8) with a similarity threshold of 99.3%. The choice of threshold was selected to optimize distinguishing *S. aureus* from other Staphylococcus species. The 16S rRNA processing pipeline choice is now outdated by 2019 standards, but the choices were state-of-the-art at the time of data generation, processing and analysis. In any case, most ASV-based methods are not compatible with the data since they are specific for Illumina data. Shotgun metagenomics data was not used for taxonomical profiling due to issues with data quality that made the taxonomical profiles unreliable (data not shown).

To analyze this dataset, we established a general methodology for microbiome analytics, focused on exploration and hypothesis generation. The methodology follows a top-down approach based on the ecological levels of complexity from Section 1.2. The methodology can be summarized as follows:

- Community-level analysis
    - Identify community-level variation associated with the condition under study
    - Identify clades driving the observed community-level variation
    - Deduce microbial interactions between the clades that drive variation
    - Study associations with other disease-related variables (disease subtypes, risk factors, severity)
- Population-level analysis
    - Explore strain-level variation on key microbial populations
- Mechanisms:
    - Use functional data and data integration to elucidate mechanisms

The following sections will describe each of the six steps of the methodology and exemplify how these were applied to explore biological questions in both paper III and IV.

3.2.1 Community-level analysis

The first step in the analysis of the skin microbiome dataset was to identify community-level associations with the variables of interest. Since communities are typically high-dimensional vectors of values, finding global associations requires statistical techniques to summarize some characteristics of the community. In microbiome analysis, alpha and beta diversity indexes are commonly used for this purpose.

In paper III, we were interested in understanding the variability in the microbial communities among the two diseases and healthy skin. For this, we compared the Shannon index between AD, psoriasis, and healthy skin and observed decreased that AD samples had a decreased alpha diversity with respect to the controls and psoriasis samples. Beta diversity ordinations based on the Bray-Curtis dissimilarity show a segregation between the AD and healthy populations, and similarly between PSO and healthy. In paper IV, the same diversity measures allowed us to identify varying ‘microbial’ dynamics between the thigh and the upper back in AD samples. Consistent with other studies, we observed decreased inter-site variability in AD with respect to healthy using beta diversity (Baurecht et al. 2018).

The choice of alpha and beta diversity measures should be tuned to the phenomenon under study. In the context of the human skin microbiome, indexes like the Shannon or Simpson diversity are useful summaries to identify dysbiosis, since they can reflect changes in the composition (both in number and abundance) of the microbiome. However, alpha diversity measures interpreted in isolation can be misleading. In paper IV, the Shannon diversity index from the upper back samples does not change significantly between the groups (Figure 3.2.1), but further investigation with beta diversity measures and taxonomical barplots reveal that the healthy upper back community has a very different bacterial composition compared to AD upper back (Figure 3.2.2).

Identify clades driving global variation

Once patterns in global community variation have been identified, a natural follow-up is to identify which microbes drive these patterns. In paper III, we employed two different strategies to identify OTUs driving the differences between the diseases. First, we performed a differential abundance analysis, comparing the arcsin transformed abundance between the three diseases while accounting for possible confounding factors with a GLM. The second strategy consisted of training Random Forest classifiers to discriminate between the conditions (healthy vs PSO, healthy vs AD, and AD vs PSO), and using the variable importance Z-score to identify the OTUs that most reliably distinguished between the healthy and diseased groups. Both analyses yielded coherent results, with differentially abundant microbes having high Z-scores. Important microbes in AD include *S. aureus, S. epidermidis, Burkholderia sp.*, *Staphylococcus sp.* and in psoriasis, *C. simulans, C. Kroppenstedtii and Lactobacillus spp.*

In paper IV, we performed differential abundance analysis between healthy controls, lesional and non-lesional for each skin site (thigh and upper back). For this, we applied a compositionally-aware method called ANCOM (Mandal et al. 2015). The analysis identified *S. aureus* and a *Staphylococcus spp.* as being differential abundant between the groups in the thigh. Two other OTUs were identified as differentially abundant in upper back, but these are known kit contaminants (Glassing et al. 2016; Salter, Cox, Turek, Calus, Cookson, Moffatt, et al. 2014). A possible explanation for this is that upper back controls are lower-biomass with respect to AD upper back and therefore more likely to be contaminated during the sample processing. The choice of ANCOM was due to its effective control of the false discovery rate (FDR) compared to other methods (Weiss et al. 2017b).

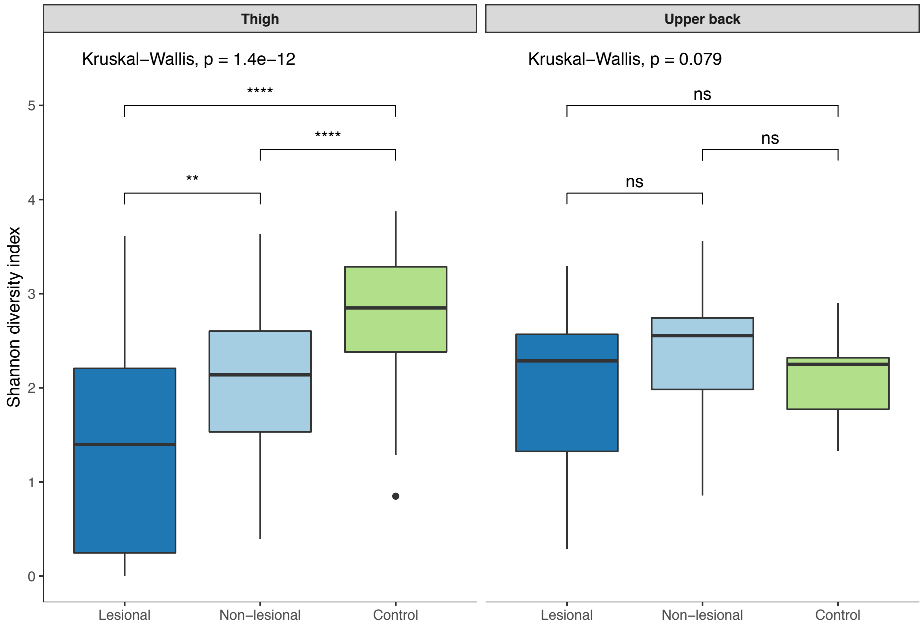

Figure 3.2.1 - Shannon diversity index of AD and healthy samples between skin sites (thigh and upper back)

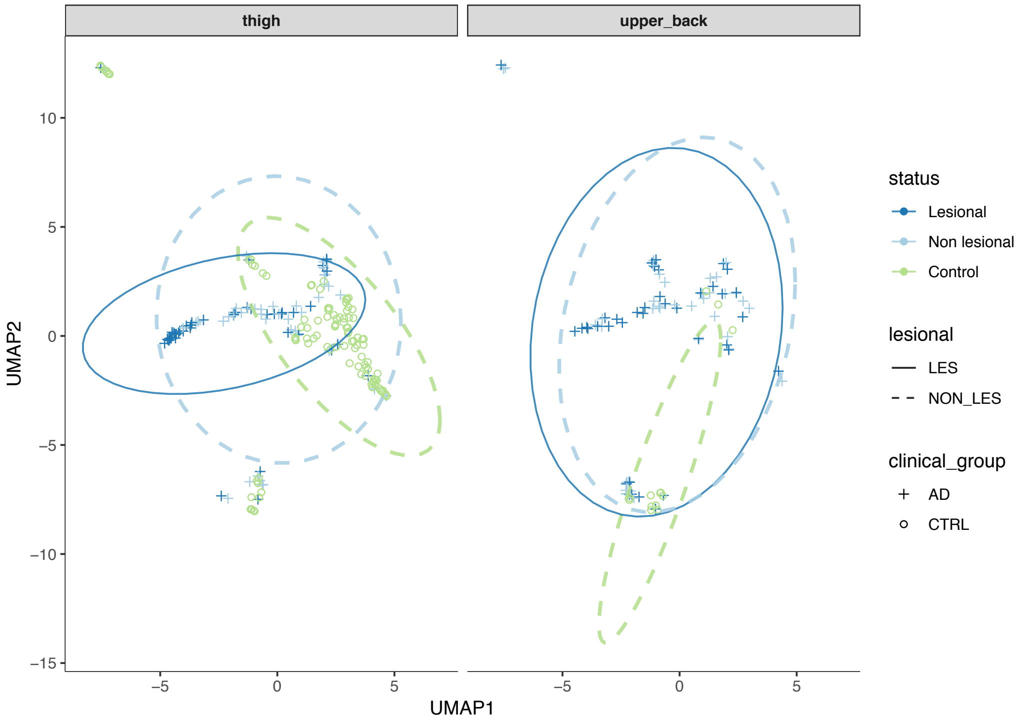

Figure 3.2.2 - Beta diversity plot of AD and healthy samples based on the Bray Curtis dissimilarity using UMAP (Paper IV)

Microbe-microbe interactions

Once key microbial players have been identified, we can use the abundance data to infer to understand dependencies between the microbial key players and the rest of the community. In paper III, we constructed two microbial co-abundance networks using SparCC (Friedman and Alm 2012) using all OTUs present in more than 5% of the samples, one for AD and one for psoriasis. The resulting AD co-abundance network resulted in a single connected component. Salient features of the network include that *S. aureus* correlations with most bacteria were negative, and the existence of a clique between *S. epidermidis*, two *Corynebacterium sp.* OTUs and *Finegoldia sp.*, with links to *P. acnes* (now *C. acnes*). In psoriasis, the co-abundance network yielded a set of five independent connected components. The network suggests that *C. simulans*, which is a strong determinant for psoriasis is associated with two OTUs: *Streptococcus spp.* and *P. anaerobius.* Another notable feature of the network is that the node with the highest degree is an OTU labeled as *C. kroppenstedtii*.

Association with disease-related variables

Another way of understanding how the community is involved in the disease is to examine the association with measures of disease variability. Examples include disease severity, risk factors, comorbidities, or the presence of specific symptoms. In paper III, we examined the association between microbial features and disease severity for both AD and PSO. We employed a forward selection strategy to train regression models containing 5 to 50 microbial features (in increments of 5) to predict SCORAD (AD severity index) or PASI (PSO severity index) scores. Microbial features were ranked according to their individual correlation with the severity score. The best model was achieved using 35 OTUs to predict SCORAD with a Mean Absolute Error (MAE) of 9.84.

In paper IV, we re-examined the microbial association with severity with respect to skin site. In this analysis, we focused only on *S. aureus* association to severity, based on our results from paper III and previous studies (Kong, Oh, Deming, Conlan, Grice, Beatson, Nomicos, Polley, Komarow, Murray, et al. 2012). Interestingly, the association between *S. aureus* and severity could only be observed in the thigh but not in the upper back.

3.2.2 Population-level analysis

Strain analysis can help provide insights about the results derived from the community-level analysis. For example, we hypothesized that the differences in the association between S. aureus and severity observed in paper IV could be partially explained by the presence of different S. aureus strains. To investigate this, we profiled *S. aureus* strains using StrainEST (Albanese and Donati 2017) from the available shotgun metagenomics data. The tool outputs relative abundance strain profiles for *S. aureus* for each sample based on the available genomes from NCBI. Figure 3.2.3 summarizes the number of detected *S. aureus* strains per sample. These results suggest that most samples have more than one *S. aureus* strain co-colonizing the skin, both in lesions and non-lesions.

The *S. aureus* strain data allowed us to examine different questions. Firstly, we inspected the *S. aureus* strain profiles at the level of each individual: are the strains that colonize nonlesional skin the same that are present in lesional skin? Analyzing subjects with both a lesional and a nonlesional sample, we observed that the most dominant strain matches in 19 out of 21 patients sampled in the thigh and 7 out of 7 patients sampled in the upper back. However, the Bray-Curtis dissimilarity of the profiles is not very high for most samples (Figure 3.2.4). Another observation is that thigh lesional samples tend to have more detected strains than nonlesional (Figure 3.2.5).

We can generate different hypotheses from these observations. One possibility is that the dominant *S. aureus* strain present in unaffected skin, plays a role in exacerbating severity when external factors (e.g. inflammation) together with other microbiome changes (e.g. loss of ‘protective’ species) trigger its virulence. Another possible hypothesis is that the thigh lesional ‘microenvironment’ is more favorable than the upper back for *S. aureus* strains colonization. Another hypothesis could be that the thigh lesional microenvironment is more favorable for the recruitment of virulent *S. aureus* strains.

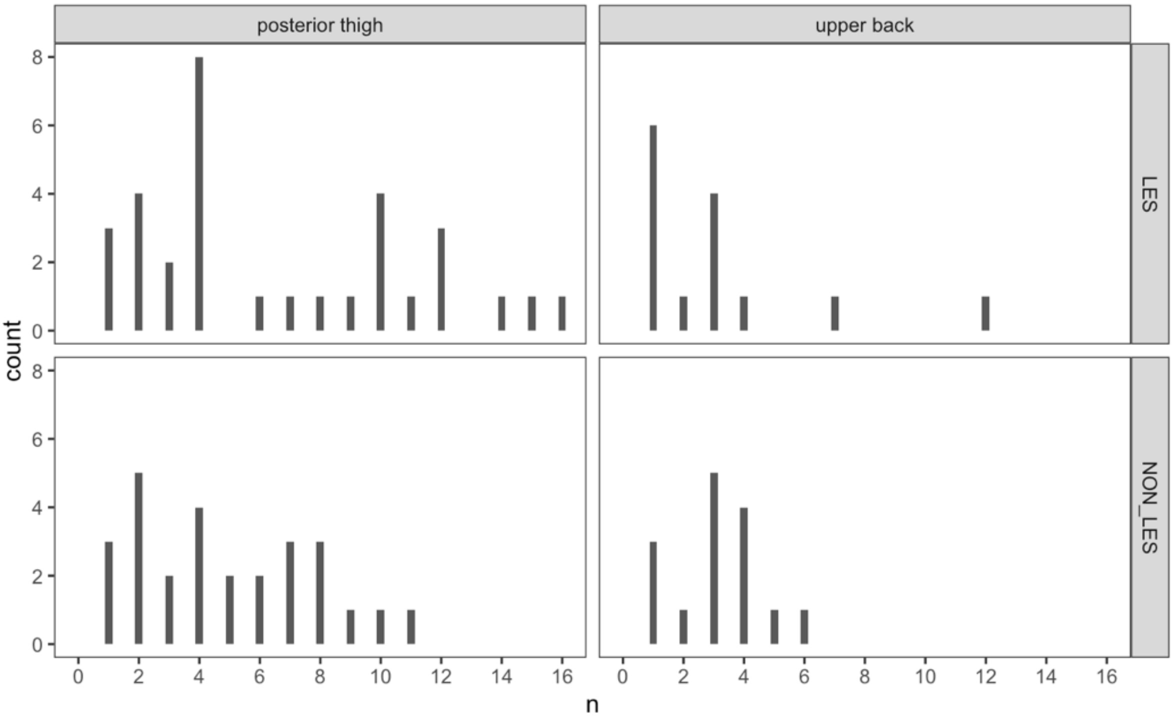

Figure 3.2.3 Distribution of number of strains per sample

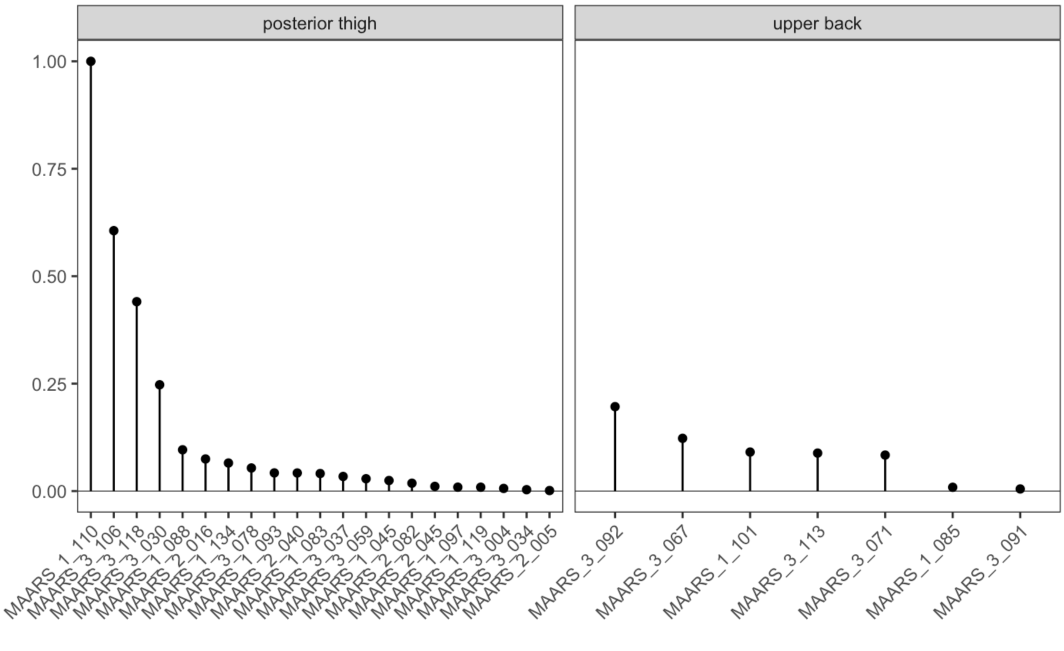

Figure 3.2.4 Bray-Curtis dissimilarity between lesional and nonlesional samples per individual

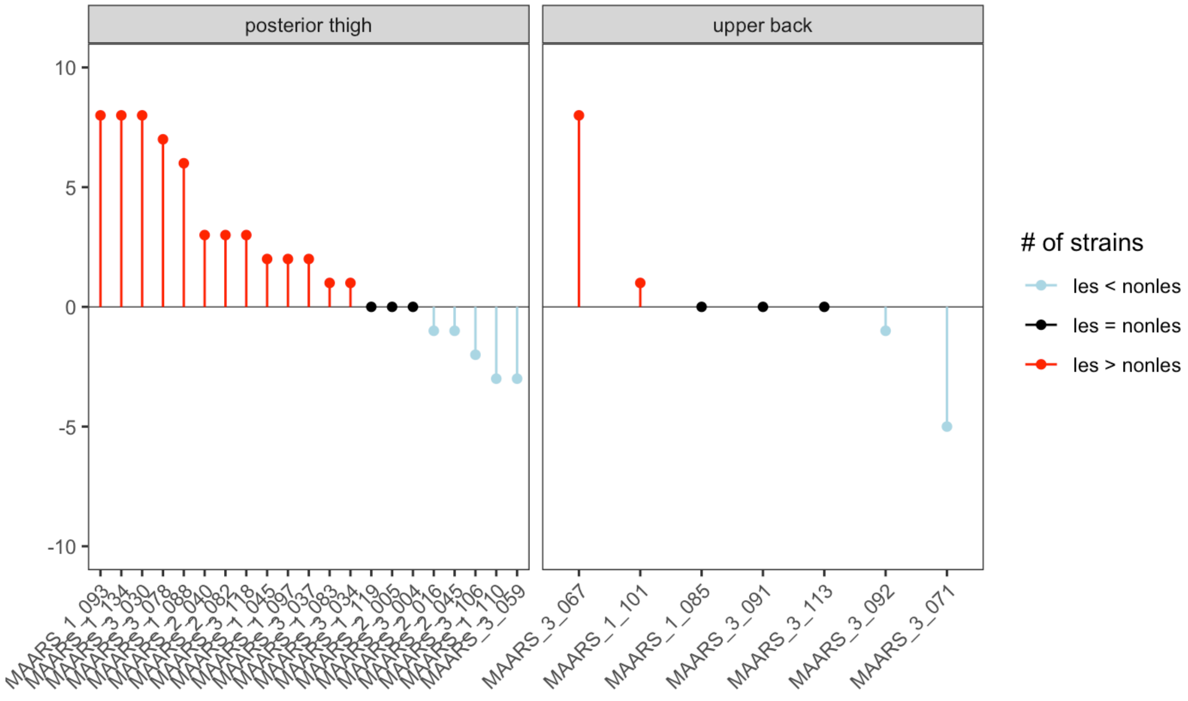

Figure 3.2.5 Difference in number of strains between lesional and nonlesional samples per individual

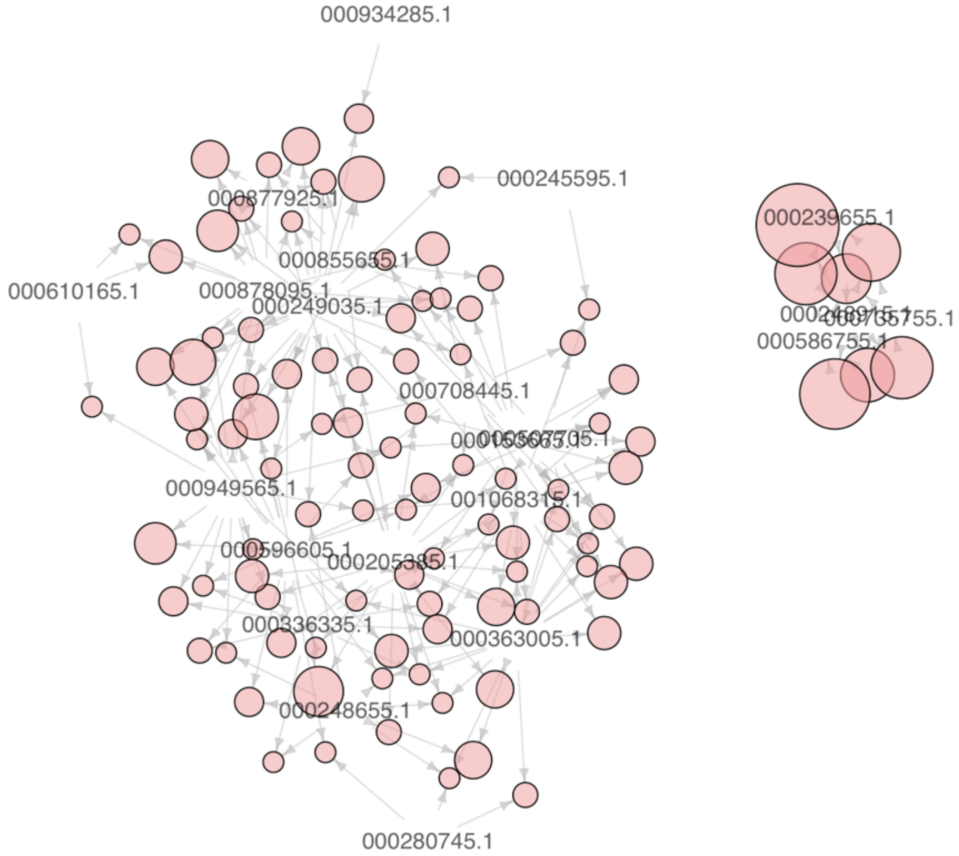

Figure 3.2.6 Graph-based visualization of closed frequent itemsets of S. aureus strains

We also examined the co-occurrence of S. aureus strains. To accomplish this, we applied the ECLAT algorithm to find closed frequent itemsets with a minimum support of 6 (arbitrary choice). This resulted in a list of strain-sets represented in a graph-based layout in Figure 3.2.6. The most notable feature of these results was a set of four strains which co-occur very frequently, but only with each other. In particular, one of the strains (000248915.1) appears to be central to the other three, since it appears together with them 21,18 and 16 samples, which are the closed itemsets with the highest support in the data.

We further investigated which strains are associated with AD severity in our samples. We used a simple Bayesian model to estimate the posterior likelihood of a strain to occur in a patient with severe AD as opposed to a patient with moderate AD. Assuming a probability *p* of a strain colonizing a severe patient (with probability *1-p* of colonizing a moderate patient), the likelihood of observing a number of severe patients carrying the strain from a set of *N* patients is distributed as a Bernoulli(*N,p*). We set a Beta(2,2) prior for the parameter *p,* to enforce the belief that by default, strains do not show a preference between severe and moderate patients. The posteriors for the 12 strains with the highest MAP estimates for *p* are shown in Figure 3.2.7. Interestingly, the group of 4 highly co-occurring strains (000239655.1, 000248915.1, 000586755.1 and 000735755.1) are all among the strains with the highest MAP estimates for probability of occurring in severe patients. We asked also if these severity-associated strains showed any distinct preference for any of the sites. Figure 3.2.8 reveals that most of these strains appear in both sites.

The strain analysis helped refine some hypotheses concerning the site-specific differences in severity, but some open questions remain. For example, we identified a set of strains that are associated to severe disease, but these are present in both skin sites. A possible follow-up study could focus on S. aureus strain variability across skin sites. Another caveat about these results is the reliability of the strain inference algorithm and the database. StrainEST uses NCBI genomes as ‘strains’, so it is possible that the algorithm misclassifies a SNP profile that is not present in the database as combinations of close strains. However, the fact that the most dominant strain matches across lesional and non-lesional samples of the same individual suggests the algorithm works correctly to some degree. Further work is required to validate and characterize these strains.

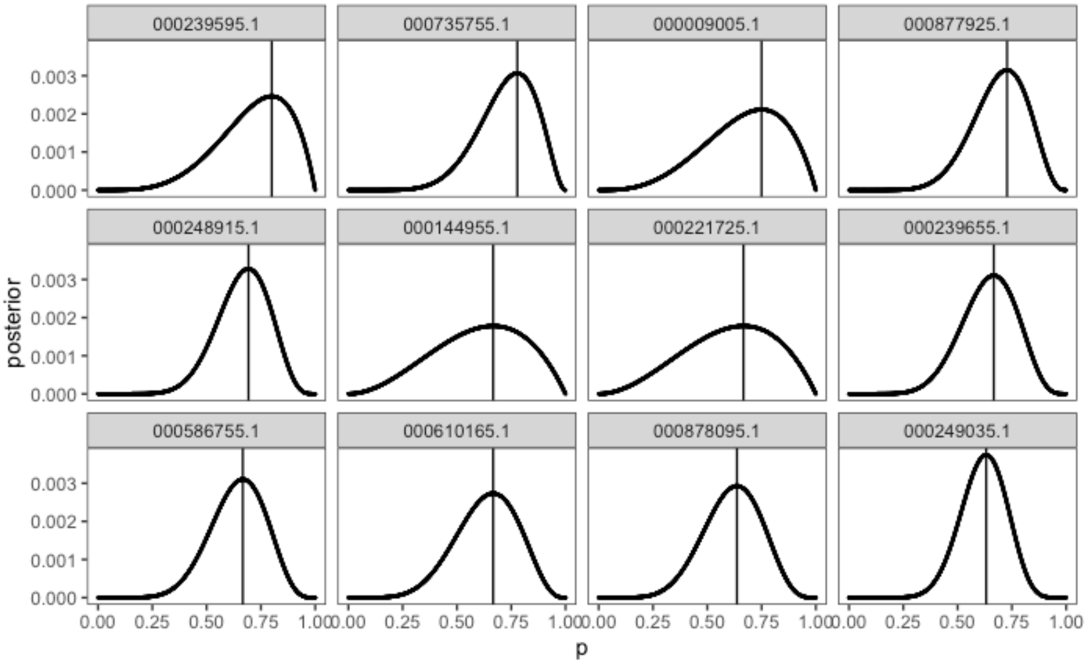

Figure 3.2.7 Posterior distribution of the probability of a given strain appearing in a severe sample (vs. a moderate sample). The figure shows the 12 strains with the highest MAP estimate.

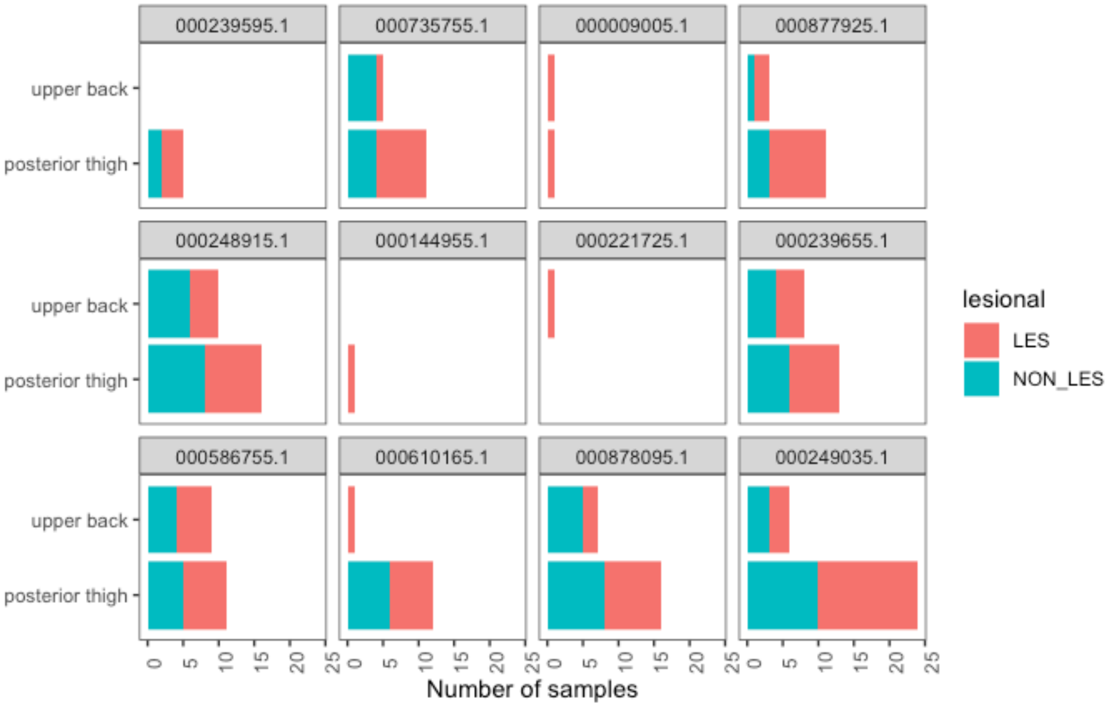

Figure 3.2.8 Number of samples where each of the severe-associated strains was detected

3.2.3 Data integration and functional analysis

The final aim of microbiome studies is to find possible mechanisms that drive the microbial and phenotype variation identified in previous analyses. Shotgun metagenomics data allows the identification of gene families and pathways present in the microbial community. Mechanisms can be also inferred by integrating human-level data, such as direct measurements of physiological conditions (oxygen, pH), molecules (lipid content, antibacterial peptides), or omics datasets. In the case of papers III and IV, we used both shotgun metagenomics and the skin transcriptome data for this purpose.

In AD, we wanted to understand the effects of *S. aureus* colonization on the human skin. For this, we divided the lesional samples based on their S. aureus abundance profile into two categories: samples with 'high' (> 87% relative abundance) or 'low' abundance of the OTU. Based on this partition we explored 1) which microbial functions were associated with the groups, and 2) the effect of *S. aureus* dominance on the skin transcriptome.

For (1), we used PiCRUST (Langille et al. 2013) to perform an initial prediction of functions that are differentially represented between the *S. aureus* high and low groups. The analysis resulted in gene families from different KEGG pathways like "Bacterial toxins", the "Two-component system" and "Glycolysis / Gluconeogenesis" enriched in the 'high' samples. PiCRUST predictions were then validated using a subset of shotgun metagenomics samples. Gene family abundance prediction was obtained using HUMAnN2, and statistical analysis confirmed the enrichment of the genes in the predicted pathways. Notable examples include the alpha and delta-toxin *S.aureus* genes.

PiCRUST is a useful hypothesis generation for microbial mechanisms from 16S data. However, results should be interpreted with care, since the taxonomical resolution of 16S is not sufficient to make inferences about strain-level functional variation. Shotgun metagenomics data or targeted experiments are good alternatives to validate the predictions test the predictions made by PiCRUST. However, validation approaches must be planned carefully: merging annotations between PiCRUST and HUMAnN2 was problematic since they employ different databases: KEGG and UniRef with COG annotations respectively. We chose to match the annotations based on gene name but recognize this is far from ideal.

To explore (2), the association between S. aureus and the skin transcriptome, we performed a differential expression analysis between *S. aureus* high and low groups, which resulted in a set of 256 DEGs (referred in the manuscript as "*S. aureus* signature"). Annotation of the DEGs using Ingenuity Pathway Analysis observed an enrichment of different categories, such as Keratinocyte differentiation and different components of the kynurenine pathway. Based on these results, we hypothesized that the host could employ a tryptophan 'starvation' strategy to control *S. aureus* colonization. Follow-up experiments based on this hypothesis revealed that AD *S. aureus* strains have low dependence on tryptophan, so the hypothesis was discarded. We performed a similar analysis in paper IV but focusing on the AD thigh samples. Results like 'keratinocyte differentiation' were also present, but a surprising association with 'Circadian Rhythm signaling' were also predicted. Further research is required to determine the significance of this finding.

To elucidate mechanisms in psoriasis, we selected a different strategy since there were no bacterial OTU candidates with a strong association with the disease as *S. aureus* in AD. Therefore, we decided to test the associations between data-defined gene groups and the top differentially abundant OTUs using MaAsLin. Gene groups were inferred from the results of module detection on the Psoriasis-versus-healthy DEG co-expression network. We performed association tests of 12 gene modules against the 25 most differentially abundant OTUs, accounting for body site, age, gender and institution. This resulted in 6 associations after multiple testing correction, three of which related an OTU annotated as *Corynebacterium spp.* to genes enriched for cell cycle, and inflammatory pathways. This strategy is similar to the one used by (Morgan et al. 2015).

3.2.4 Conclusions

The work on papers III and IV illustrate the use of our established methodology for the exploration of a human skin microbiome dataset. This allowed us to draw new insights, generate and test hypothesis, and confirm previous observations from other studies about the relationship between the host and the skin microbiome in psoriasis and atopic dermatitis. The methodology proved to be flexible enough to allow the use of a broad range of statistical techniques and algorithms, while providing enough structure to facilitate reasoning about the results and connecting insights drawn from different analyses.

Many choices in technology, data processing and analysis for this skin microbiome project do not reflect the current state of the art. The project was conceived and executed when 454 pyrosequencing was the standard for 16S rRNA sequencing (circa 2011). Nevertheless, the cohort is still (to our knowledge) one of the largest studies of the skin microbiome in both diseases. Also, solid study design choices, e.g. the use of negative controls, were implemented in the study when they were not standard practice. Nevertheless, the results of the analysis are congruent with the literature.

A modern redesign of the study would be developed using shotgun metagenomics. However, ensuring good data quality and sufficient sequencing depth to characterize skin communities is still a challenge due to the low biomass of the samples and the high proportion of human DNA. In the case of designing a similar amplicon-based study, sequencing the V1-V3 regions would be a better choice than V3-V4, as the V1-V3 region has a better resolution for Staphylococcus species (Meisel et al. 2016).

However, the analysis strategies and most of the methods we used are still widely used in current microbiome papers, and many modern analysis techniques like machine learning and compositional-based data analysis were applied in this work. The observations and generated hypothesis will lead to interesting biological insights with the appropriate follow up studies. Future work will include further mining of the shotgun metagenomics dataset for non-bacterial microbes, and bench work to validate some of the observations.

An important limitation of this study was the choice of anatomical locations. During the patient sampling phase, a second AD skin site (upper back) was included in the design. The measure enabled the study in paper IV, since the number of AD samples from each site was balanced (~40 per group). However, the collection of healthy controls for the upper back site was overlooked, so only few healthy controls are available for the upper back. This implies that the interpretation of results where samples are compared to upper back controls must account for this sample size unbalance.

In conclusion, the results of paper III and paper IV will serve as a basis for the design and development of future studies to further understand the role of skin microbiome in atopic dermatitis and psoriasis.

4 CONCLUDING REMARKS AND FUTURE PERSPECTIVES

The human microbiome field has been growing at an accelerated pace for the past decade due to the developments in high throughput sequencing and bioinformatics methods. This has resulted in numerous studies looking for possible associations between the human microbiome and every disease. However, the lack of standardized practices has hampered reproducibility and the ability to compare between studies due to the relative novelty of the field.

The field has started to reach an inflection point: the cumulative knowledge has finally started to push towards a new stage where we are demanding more stringency about experimental design, data generation and analysis methodologies. Examples of these trends include big efforts like the CAMI challenge (Vollmers, Wiegand, and Kaster 2017) and reviews for conducting high quality microbiome research (Knight et al. 2018; Kong et al. 2017). As human microbiome research evolves into a more rigorous state, we will be able to distinguish true associations between the microbiome with disease and better characterize the mechanisms behind these associations, hopefully leading to translational research that impacts human health.

The work in this thesis has sought to contribute to the state of human microbiome research by developing methods and executing analysis following the best practices in the field. We proposed a bench to bioinformatics pipeline for executing viral discovery projects, together with a method prototype for unknown virus discovery. The methods helped explore the role of the virome in acute lymphoblastic leukemia and also enabled the discovery of a new bacteriophage genome. Additionally, we analyzed one of the largest skin microbiome datasets regarding atopic dermatitis and psoriasis. This resulted in many insights and testable hypotheses that can potentially form the basis of further studies to understand the relationship between the microbiome and these diseases.

From the technology development perspective, the microbiome field will surely be impacted by developments in single-molecule long-read sequencing. Currently, Pacific Biosciences and Oxford Nanopore have already revolutionized the way we approach genome sequencing projects(Koren et al. 2013), helped monitor the spread of viruses in an outbreak(Quick et al. 2016), and coupled with tools like What's In My Pot(WIMP) enable real-time profiling of any metagenomics sample(Juul et al. 2015). Although throughput, error rates and costs are currently limiting, it is only a matter of time before they become routinely employed for metagenomics projects. In parallel, developments in single-cell technologies will provide an alternative method to perform metagenomics studies that address many of the limitations we have with current techniques (Rinke et al. 2013). New bioinformatics methods will surely accompany the fast-paced developments in technology. The progress in methodology and technology coupled with improved study designs should allow the field to solve many open questions. Hopefully, these answers will impact human health during the coming decades.

5 ACKNOWLEDGEMENTS

First of all, I would like to thank my main supervisor **Björn Andersson** for giving me the opportunity to do my PhD in the lab and for all the scientific and academic discussions. To **Hamid** for all the life advice and the chats and all the data (maybe you went a bit too crazy on the data!). To **all present and former lab members**, for the nice environment and for being very quiet in the office specially during the last two years. **Talavera**: maje, estoy acá no se si gracias a vos o por tu culpa (lol). Gracias por todo voj, los taquitos, el kebab, las habladas, las birras, las obscenidades, los modelos probabilísticos, la vida y todo lo demás que es básicamente innumerable, me vale muy pero muy v*** si no te parece, vos te leés la tesis y me decís que te gustó. Al **Pou**, merci per tot tíooo, si a alguien le debo literalmente el tener los requisitos para estarme graduando es a vos <3. Como dijo un Poueta alguna vez: "no existe más muerte que la de nuestros sueños". Also, I would like to thank my **co-supervisors Tobias** and **Stefanie** for all the scientific discussions.

I would also like to thank all my collaborators, with whom I have had the pleasure to work and to learn from during these years. In particular, I would like to thank **Filipe**, **Helen** and **Alan** in KCL for all the great work and support in the MAARS project. I really learned a lot from you and had a lot of fun hanging out too! Also, the Alenius lab, the Ranki lab and the MAARS consortium, and in particular **Noora**, **Hanna**, **Elina** and **Annamari** for all the great work with the skin projects! A **Andre** gracias por toda la paciencia con el proyecto de EB y suerte con tu tésis que ya casi estás! To **Anna** for being the best student one could ever wish for! To the **metagenomics JC: Luisa, Fredrik, Johannes, Yue, Paulo, Steffi, Justine, John** for giving me a platform to discuss microbiome research with such a smart bunch of people, for the grilling (I hope) and all the fika!

To all the friends I found during these years in these Nordic lands, I would like to thank you for keeping me grounded, for all your support, fun times and for helping me maintain a little bit of sanity. It is crazy the amount of people I have met during all these years; too many people have come and gone, and a piece of my heart has certainly gone with each of you. Despite my best efforts, I'll for sure miss many of you here, but your mark in my life will linger on :).

Chicki, thanks for being my best friend during these years I've gotten to know you. Even despite how different we are, you were always there for me, and I'll be forever thankful for this even if I might not have fully appreciated it at the time. Thanks for always helping me see things from a different perspective <3. **Tati**, mi panchi poooorc favorita, crec que non hi ha paraules per explicar-te l'important que has estat (i encara ets) per mi durant tots aquets anys; con todo y todo me alegro de aún tenerte en mi vida. Gracias por abrir mi mente, por todo el apoyo y cariño, las vaciladas, los viajes y todo lo que me has enseñado. **Sandrita**, basicamente empezamos el doctorado casi juntos, y casi que sos PI y yo aqui intentando terminar jaja. Gracias por la energía, la actitud y por siempre encontrar tiempo para vernos (aunque sea 3 horas después de las acordadas jajaja). **Vale**, grazie per tutte le serate in giro, la musica, i concerti, tutto il funk, per prendere tutto sciallo, e per le storie di tutti i casini e belinate in questa vita. Anche per il bellisimo viaggio a Londra di break di tesi :D (sei di fianco a me in aereo mentre scrivo questo). **Noora**, my first Nordic friend after sooo many years, it was lots of fun (also a bit intense hehe) to work together. Thanks for all the fun, the beers and brunches, talks about life and specially paper IV in this thesis! **Paula**, thanks for all the chats and the support, and the awesome movie and singing evenings too. **Konrad**, probably my most ADHD friend ever :D thanks for everything, the advice, the geekness, the help the music, the jams, everything. Come back to Stockholm soon, so many future projects to do! **Carlos Mata**, por todo el chingue, las birritas, los taquitos y las pachucadas de los barrios del sur jaja.

Sarah, thanks for being the best awkward friend one might ask for (lololol) For all the never-ending talks about everything, the swears, the posh English, the football (you're s***), all the intense married-couple-like discussions and never giving up on me. **Yildiz**, thanks for being a voice of wisdom and for helping me realize there are other people with a similar kind of crazy as mine. Thanks for all our discussions both deep and light-hearted, for the all the interesting articles, books, ALL the help editing this document, fikas, board games and everything else. **Rafa**, thanks for all the games, music, concerts, random talks and for meeting up even if meant for you to come out of the cave into the light! Also, thanks specially for screwing up all our Hanabi games we didn't finish, and for telling me the wrong dress code for your wedding ahueahueahue. **Vero**, thanks for the support, for all the pizzas, for giving me a second home for Christmas and all the awesome lazy days.

To my swedish girls **Frida** and **Amanda**: thank you for helping me reconnect with this complicated piece of frozen land I found myself living in for longer than expected. **Fri**(dita hehe), thank you for becoming a very unlikely friend in a crazy time: the writing of this document. For always telling me things straight yet being one of the most caring and warm-hearted persons I've met here. Also, for the badminton matches, teaching me a bit of football vision, and ... I'll get my nutmeg revenge sooner or later! **Amanda,** thanks for being as random, happy, sweet and awesome as you are. Also, for the awesome basketball skills, for getting me to do random front-flips in Rålis and constantly inviting me to swim despite my incapability.

To all my **friends from KI,** thank you for making this 'interesting' place a bit easier to withstand. I would like to thank specially: **Alberto**: gracias por todas las fiestas y la buena vibra que siempre está con vos. **Shahul** (paki!), thanks for feeding me and the fun, the fooseball, the deep talks, and the advice (even if unsolicited sometimes). **Daniela**: thanks for being my big sister, all the advice the fikas and the cakes! To **Irene**, thank you for always being there, for staying in touch, for the help with the thesis editing, and for your constant stream of energy and positivity (hehe <3). **Joanne**, for being a true hardliner, for the awesome karaoke and fun times. **Alča**, thanks for having such an open heart and infinite energy, for the best laser tag session ever and your czechness. To **Alek** for the awesome music making. A **Ale Fernandez**, gracias por tenerme en tu lab de intern hace millones de años, y por las habladas y consejos a lo largo de estos años. To **Alba,** infinite kudos and thousand thanks for the amazing thesis cover!

To all my friends from the CMB times: **Gonçalo , Simona, Katrin, Giuseppe, Mehdi, Marion, Laure, Eric, Ashwini, Philine, Lisa, Laura, Riccardo** ... thanks for all the fun, pubs, parties and fika's :D. To Olov Andersson's group, **Dodo, Christos, Jeremy, Nicki** thanks for adopting me for lunch when I first got left with no group hehe.

To the **9D mitochondria gang**: **Sarah**, **Roby, Mara, Jelena, Miriam, Rodolfo, Laura, Diana, David(s)** thanks for being my foster lunch group once we moved to biomedicum and for all the fun outside work too (and filling up our concert!). To the **people in 9C**, thanks for taking about 3 months to start talking to the guy alone in the office (thus forcing me to socialize with other people in the building hehe) , but also for being very nice to me (and sharing your fikas) after you finally got to it :D , specially, N**uno, Vero, Ju, Graciela, Leo, Kathy , Cassandra, Cajsa. Cagla and Pontus** (to be fair, you probably were the first to talk to me after a week). **Amir, Juancito, Pierre, Hassan and Christian**, for all the fooseball matches and shitty goals! I have the feeling I will still be waiting for the fooseball tournament by the time I'm done. And of course, so many cool people around the rest of KI: **Theresa, Jorge Correia, Nuria, Jennine, Magali, Emilio, Sindhu, Polina, Jill , Anders, Paulo, Milana, Alvaro & Rafa, Maria Azorin, Paola P., Tanya, Roberta , Walter, Elena, Edu, Shady, Sharesta**, **Tina, Varsha, Florian** and everyone else my chicken thesis-fried brain has forgotten to mention (sorry), for all the lunches, dinners, parties, bbqs, events, beers and everything else.

To everyone in **KEFF and in the SSIF basketball**, for all the years of fun and friendly-spirited football and basketball, for accepting (and helping me partially recover from) my addiction to ballhogging. In particular, thanks to **Xico, David, Rob, André, Andreia, Moutinho, Frida, Emma & Teresa** for the fun in and out of the football pitch, BBQs, badminton, beers, randomness. **Fede**, maejtro!, por darme todos los pases , la magia en el terreno de juego , las recomendaciones musicales y consejos de vida. A la KEFF comunidad latina **(Edu, Juan, Moises, Joao)** por llevarme al colombiano a comer bandejita paisa calidad . **Andreia** for introducing me to badminton and paddel tennis (I'll kick your ass soon enough!) A **Eloy** por todos las discusiones de baloncesto y vida post-partido. **Amanda Ainali**, for helping me with the swedish and the beers (the basketball too)! To the bands I've played along the years, **All covers (L-G, Daniel and crew, in loving memory of Annika and Anders)**, as well as everyone in **Divans orkester (Lars, Jerry 1 & 2, Siv, and Åke)** for the music, and for taking me in to play with you, as different a member as I might have been. To present and former members of **Bossa Academica: Konrad, Markel, Ana, Abraham, Marcelo and Jorge**, for teaching me so much about music, for helping me take my musicianship to another level, for all the gigs and rehearsals, for not hating me everytime I use 'beautiful cadenzas' and go too 'jazzy', or swing around in the rhythm randomly in any given song(hehe).

To Shady & Asa, for being the best corridor mates. **Mahmoud and Anders**, for being my concert-going friends, without which most of my concerts would have been very boring and keep the metal alive \m/. To the people in **Magiska: Sara, Keo, Lorenzone, Ste, Teo**, thanks for making me feel at home there. To the **CR crew: Andre, Luis Carlos, Carlos, Yani, Marce y Antea** (tica honoraria): PV mooops, gracias por darme un pedazo de tiquicia por acá, las birritas y el bullying por no saber cuales son las Julieta jajaja.

To my friends abroad: **Gabo y Ale**, por ser mis senpai con esto del doctorado, no estaría aca si no fuera por uds de fijo. A Gabo también por los comentarios de la tesis! A **Oscar,** por muy conocedoramente convencerme de no ir a gringoland y venir a Europa. Posiblemente una de las mejores decisiones que he tomado. **To Priyata**, for staying in my life throughout all these years, for all the visits, for all the talks, and for being my lifeline abroad. **Marcia,** obrigado! such a random friendship that happened to go through the years as nothing, and for meeting in real life quite regularly. Looking forward to meeting you around soon, it's been a while now :) **Mateja**, hvala, for being one of my first friends in Sweden, and having me over in Croatia as if it hadn't been 5 years since we last met <3. **Sephora**, for being my favorite nurse, for teaching me French and all the visits and talks. **To Mo(nica)** for all the postcards! To my friends from the year in Finland: **Tin, Monica, Selam, Dish, Amjad, Julia, Chiara**: for staying around in one way or another <3. **To Maddie**, for coming to visit and giving me a place to crash in Marseille and taking me around!

A la **gente de tiquicia**, gracias porque a pesar de los años aún han seguido estado conmigo de un modo u otro a lo largo de estos 8 años desde que jalé. Gracias por sacar un rato para verme cuando bajo (diría Leo, "por qué no subir?"). En especial a la **Flaca**, por ser una constante en mi vida <3, a **Peto y Leo** por las habladas profundas, las estupideces y la música calidad. A **Jose** por las smash mejengas (pero no por usar las pestes), a **Andrés** por ordenar las pizzas cada año que vuelvo, a **Ali** por ser la única de "miau" en venir a visitarme, a la **China** por caerte a estocolmo, y siempre llegar tarde pero igual venir a verme jaja, a **Katha** por sacar el rato de tu ocupada vida para verme y llevarme a comer por ahi. A **July** por ser un conocedor, a **Luiiish** por ser mi compañero de doctorado a la distancia y a **Naty** por llevarme a pasear con los peludos y ser una peleona. Al resto de mis amigos: **Las Anitas (Moreno y Chevez), Chebas, Rebe, Carito Padilla, Luisk, Desiree, Iosi, Chan** por tenerme presente siempre con uds de un modo u otro. **A Cindy y Rodrigo**, por cuidar a mi madre y llevarme a pasear cuando ando por allá!

Last but not least, **a mis padres**, gracias por todas las oportunidades que me dieron en esta vida sin las cuales obviamente no estaría aquí. Gracias por enseñarme como vivir, y por apoyarme en todas las decisiones que he tomado, aunque estas me hayan llevado a vivir a 10 000 km de uds. **Ma**, gracias por estar siempre ahi cuando te necesito y por la paz de todos estos años aunque no esté por allá. **Dad**, gracias por siempre estar pendiente, y los paseos a la playa para agarrar sol que aca hace falta. Y al resto de mi familia por siempre recibirme con los brazos abiertos cuando voy a visitar.

6 REFERENCES

Aas, Johannes, Charles E Gessert, and Johan S Bakken. 2003. “Recurrent Clostridium Difficile Colitis: Case Series Involving 18 Patients Treated with Donor Stool Administered via a Nasogastric Tube.” *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America* 36 (5): 580–85. https://doi.org/10.1086/367657.

Abubucker, Sahar, Nicola Segata, Johannes Goll, Alyxandria M. Schubert, Jacques Izard, Brandi L. Cantarel, Beltran Rodriguez-Mueller, et al. 2012. “Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome.” *PLoS Computational Biology* 8 (6). https://doi.org/10.1371/journal.pcbi.1002358.

Afshinnekoo, Ebrahim, Cem Meydan, Shanin Chowdhury, Dyala Jaroudi, Collin Boyer, Nick Bernstein, Julia M. Maritz, et al. 2015. “Geospatial Resolution of Human and Bacterial Diversity with City-Scale Metagenomics.” *Cell Systems* 1 (1): 97-97.e3. https://doi.org/10.1016/j.cels.2015.07.006.

Albanese, Davide, and Claudio Donati. 2017. “Strain Profiling and Epidemiology of Bacterial Species from Metagenomic Sequencing.” *Nature Communications* 8 (1): 2260. https://doi.org/10.1038/s41467-017-02209-5.

Alekseyenko, Alexander V, Guillermo I Perez-Perez, Aieska De Souza, Bruce Strober, Zhan Gao, Monika Bihan, Kelvin Li, Barbara A Methé, and Martin J Blaser. 2013. “Community Differentiation of the Cutaneous Microbiota in Psoriasis.” *Microbiome* 1 (1): 31. https://doi.org/10.1186/2049-2618-1-31.

Allander, Tobias, Martti T Tammi, Margareta Eriksson, Annelie Bjerkner, Annika Tiveljung-Lindell, and Björn Andersson. 2005. “Cloning of a Human Parvovirus by Molecular Screening of Respiratory Tract Samples.” Proceedings of the National Academy of Sciences of the United States of America 102 (36). https://doi.org/10.1073/pnas.0504666102.

Almonacid, Daniel E, Laurens Kraal, Francisco J Ossandon, Yelena V Budovskaya, Juan Pablo Cardenas, Jessica Richman, and Zachary S Apte. 2016. “16S RRNA Gene Sequencing as a Clinical Diagnostic Aid for Gastrointestinal-Related Conditions.” *BioRxiv*.

Alneberg, Johannes, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson, and Christopher Quince. 2014. "Binning Metagenomic Contigs by Coverage and Composition." *Nature Methods* 11 (11): 1144–46. https://doi.org/10.1038/nmeth.3103.

Amir, Amnon, Daniel McDonald, Jose A. Navas-Molina, Evguenia Kopylova, James T. Morton, Zhenjiang Zech Xu, Eric P. Kightley, et al. 2017. "Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns." Edited by Jack A. Gilbert. *MSystems* 2 (2). https://doi.org/10.1128/mSystems.00191-16.

Anderson, Marti J., and Daniel C. I. Walsh. 2013. "PERMANOVA, ANOSIM, and the Mantel Test in the Face of Heterogeneous Dispersions: What Null Hypothesis Are You Testing?" *Ecological Monographs* 83 (4): 557–74. https://doi.org/10.1890/12-2010.1.

Ayling, Martin, Matthew D Clark, and Richard M Leggett. 2019. "New Approaches for Metagenome Assembly with Short Reads." *Briefings in Bioinformatics*, February. https://doi.org/10.1093/bib/bbz020.

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

Barr, Jeremy J, Rita Auro, Mike Furlan, Katrine L Whiteson, Marcella L Erb, Joe Pogliano, Aleksandr Stotland, et al. 2013. "Bacteriophage Adhering to Mucus Provide a Non-Host-Derived Immunity." *Proceedings of the National Academy of Sciences of the United States of America* 110 (26): 10771–76. https://doi.org/10.1073/pnas.1305923110.

Baurecht, Hansjörg, Malte C. Rühlemann, Elke Rodríguez, Frederieke Thielking, Inken Harder, Anna-Sophie Erkens, Dora Stölzl, et al. 2018. "Epidermal Lipid Composition, Barrier Integrity, and Eczematous Inflammation Are Associated with Skin Microbiome Configuration." *The Journal of Allergy and Clinical Immunology* 141 (5): 1668-1676.e16. https://doi.org/10.1016/j.jaci.2018.01.019.

Boisvert, Sébastien, Frédéric Raymond, Élénie Godzaridis, François Laviolette, and Jacques Corbeil. 2012. "Ray Meta: Scalable de Novo Metagenome Assembly and

Profiling." *Genome Biology* 13 (12): R122. https://doi.org/10.1186/gb-2012-13-12-r122.

Brauweiler, Anne M, Elena Goleva, and Donald Y M Leung. 2014. "Th2 Cytokines Increase Staphylococcus Aureus Alpha Toxin-Induced Keratinocyte Death through the Signal Transducer and Activator of Transcription 6 (STAT6)." *The Journal of Investigative Dermatology* 134 (8): 2114–21. https://doi.org/10.1038/jid.2014.43.

Bray, J. Roger, and J. T. Curtis. 1957. "An Ordination of the Upland Forest Communities of Southern Wisconsin." *Ecological Monographs* 27 (4): 325–49. https://doi.org/10.2307/1942268.

Bray, Nicolas L, Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Near-Optimal Probabilistic RNA-Seq Quantification." *Nature Biotechnology* 34 (5): 525–27. https://doi.org/10.1038/nbt.3519.

Breitwieser, Florian P, Mihaela Pertea, Aleksey Zimin, and Steven L Salzberg. 2019. "Human Contamination in Bacterial Genomes Has Created Thousands of Spurious Proteins." *Genome Research* 29 (6): gr.245373.118. https://doi.org/10.1101/GR.245373.118.

Brown, C Titus, and Luiz Irber. 2016. "Sourmash: A Library for MinHash Sketching of DNA." *The Journal of Open Source Software*. https://doi.org/10.21105/joss.00027.

Brown, Sara J, and W H Irwin Mclean. 2012. "One Remarkable Molecule: Filaggrin." *J Invest Dermatol* 132 (3 Pt 2): 751–62. https://doi.org/10.1038/jid.2011.393.One.

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

Callahan, Benjamin J, Paul J McMurdie, and Susan P Holmes. 2017. "Exact Sequence Variants Should Replace Operational Taxonomic Units in Marker-Gene Data Analysis." *The ISME Journal* 11 (12): 2639–43. https://doi.org/10.1038/ismej.2017.119.

Callahan, Benjamin J, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13 (7): 581–83.

https://doi.org/10.1038/nmeth.3869.

Caporaso, J. Gregory, Christian L. Lauber, Elizabeth K. Costello, Donna Berg-Lyons, Antonio Gonzalez, Jesse Stombaugh, Dan Knights, et al. 2011. “Moving Pictures of the Human Microbiome.” *Genome Biology* 12 (5): R50. https://doi.org/10.1186/gb-2011-12-5-r50.

Caporaso, J Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, et al. 2010. “QIIME Allows Analysis of High-Throughput Community Sequencing Data.” *Nature Methods*. https://doi.org/10.1038/nmeth.f.303.

Caspi, Ron, Richard Billington, Carol A Fulcher, Ingrid M Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, et al. 2018. “The MetaCyc Database of Metabolic Pathways and Enzymes.” *Nucleic Acids Research* 46 (D1): D633–39. https://doi.org/10.1093/nar/gkx935.

Cavuoto, Kara M, Santanu Banerjee, Darlene Miller, and Anat Galor. 2018. “Composition and Comparison of the Ocular Surface Microbiome in Infants and Older Children.” *Translational Vision Science & Technology* 7 (6): 16. https://doi.org/10.1167/tvst.7.6.16.

Chaban, Bonnie, and Janet E Hill. 2012. “A ‘universal’ Type II Chaperonin PCR Detection System for the Investigation of Archaea in Complex Microbial Communities.” *The ISME Journal* 6 (2): 430–39. https://doi.org/10.1038/ismej.2011.96.

Chan, Benjamin K, Paul E Turner, Samuel Kim, Hamid R Mojibian, John A Elefteriades, and Deepak Narayan. 2018. “Phage Treatment of an Aortic Graft Infected with Pseudomonas Aeruginosa.” *Evolution, Medicine, and Public Health* 2018 (1): 60–66. https://doi.org/10.1093/emph/eoy005.

Chao, Anne. 1984. “Nonparametric Estimation of the Number of Classes in a Population.” *Scandinavian Journal of Statistics*. WileyBoard of the Foundation of the Scandinavian Journal of Statistics. https://doi.org/10.2307/4615964.

Chee, Chin-Hoong, Jafreezal Jaafar, Izzatdin Abdul Aziz, Mohd Hilmi Hasan, and William Yeoh. 2018. “Algorithms for Frequent Itemset Mining: A Literature Review.” *Artificial Intelligence Review*, March, 1–19. https://doi.org/10.1007/s10462-018-9629-

z.

Chen, I-Min A., Victor M. Markowitz, Ken Chu, Krishna Palaniappan, Ernest Szeto, Manoj Pillay, Anna Ratner, et al. 2017. "IMG/M: Integrated Genome and Metagenome Comparative Data Analysis System." *Nucleic Acids Research* 45 (D1): D507–16. https://doi.org/10.1093/nar/gkw929.

Chen, Jun, Kyle Bittinger, Emily S Charlson, Christian Hoffmann, James Lewis, Gary D Wu, Ronald G Collman, Frederic D Bushman, and Hongzhe Li. 2012. "Associating Microbiome Composition with Environmental Covariates Using Generalized UniFrac Distances." *Bioinformatics (Oxford, England)* 28 (16): 2106–13. https://doi.org/10.1093/bioinformatics/bts342.

Chen, Jun, Emily King, Rebecca Deek, Zhi Wei, Yue Yu, Diane Grill, Karla Ballman, and Oliver Stegle. 2018. "An Omnibus Test for Differential Distribution Analysis of Microbiome Sequencing Data." Edited by Oliver Stegle. *Bioinformatics* 34 (4): 643–51. https://doi.org/10.1093/bioinformatics/btx650.

Chen, Kevin, and Lior Pachter. 2005. "Bioinformatics for Whole-Genome Shotgun Sequencing of Microbial Communities." *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.0010024.

Chen, Wei, Clarence K. Zhang, Yongmei Cheng, Shaowu Zhang, and Hongyu Zhao. 2013. "A Comparison of Methods for Clustering 16S RRNA Sequences into OTUs." Edited by Maurizio Casiraghi. *PLoS ONE* 8 (8): e70837. https://doi.org/10.1371/journal.pone.0070837.

Chrzastek, Klaudia, Dong-hun Lee, Diane Smith, Poonam Sharma, David L. Suarez, Mary Pantin-Jackwood, and Darrell R. Kapczynski. 2017. "Use of Sequence-Independent, Single-Primer-Amplification (SISPA) for Rapid Detection, Identification, and Characterization of Avian RNA Viruses." *Virology* 509 (September): 159–66. https://doi.org/10.1016/j.virol.2017.06.019.

Cleary, Brian, Ilana Lauren Brito, Katherine Huang, Dirk Gevers, Terrance Shea, Sarah Young, and Eric J Alm. 2015. "Detection of Low-Abundance Bacterial Strains in Metagenomic Datasets by Eigengenome Partitioning." *Nature Biotechnology* 33 (10): 1053–60. https://doi.org/10.1038/nbt.3329.

Cogen, Anna L, Kenshi Yamasaki, Jun Muto, Katheryn M Sanchez, Laura Crotty Alexander, Jackelyn Tanios, Yuping Lai, Judy E Kim, Victor Nizet, and Richard L Gallo. 2010. "Staphylococcus Epidermidis Antimicrobial Delta-Toxin (Phenol-Soluble Modulin-Gamma) Cooperates with Host Antimicrobial Peptides to Kill Group A Streptococcus." *PloS One* 5 (1): e8557. https://doi.org/10.1371/journal.pone.0008557.

Conceição-Neto, Nádia, Mark Zeller, Hanne Lefrère, Pieter De Bruyn, Leen Beller, Ward Deboutte, Claude Kwe Yinda, et al. 2015. "Modular Approach to Customise Sample Preparation Procedures for Viral Metagenomics: A Reproducible Protocol for Virome Analysis." *Scientific Reports* 5 (January): 16532. https://doi.org/10.1038/srep16532.

Cooley, Shamus M., Timothy Hamilton, Eric J. Deeds, and J. Christian J. Ray. 2019. "A Novel Metric Reveals Previously Unrecognized Distortion in Dimensionality Reduction of ScRNA-Seq Data." *BioRxiv*, July, 689851. https://doi.org/10.1101/689851.

Costea, Paul Igor, Robin Munch, Luis Pedro Coelho, Lucas Paoli, Shinichi Sunagawa, and Peer Bork. 2017. "MetaSNV: A Tool for Metagenomic Strain Level Analysis." *PLoS ONE* 12 (7). https://doi.org/10.1371/journal.pone.0182392.

Costello, Elizabeth K, Christian L Lauber, Micah Hamady, Noah Fierer, Jeffrey I Gordon, and Rob Knight. 2009. "Bacterial Community Variation in Human Body Habitats across Space and Time." *Science* 326 (5960): 1694–97. https://doi.org/10.1126/science.1177486.Bacterial.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. "Greengenes, a Chimera-Checked 16S RRNA Gene Database and Workbench Compatible with ARB." *Applied and Environmental Microbiology* 72 (7): 5069–72. https://doi.org/10.1128/AEM.03006-05.

Dickson, Robert P., John R. Erb-Downward, Fernando J. Martinez, and Gary B. Huffnagle. 2016. "The Microbiome and the Respiratory Tract." *Annual Review of Physiology* 78 (1): 481–504. https://doi.org/10.1146/annurev-physiol-021115-105238.

Dickson, Robert P, and Gary B Huffnagle. 2015. "The Lung Microbiome: New Principles for Respiratory Bacteriology in Health and Disease." *PLoS Pathogens* 11 (7): e1004923. https://doi.org/10.1371/journal.ppat.1004923.

Dutilh, Bas E., Noriko Cassman, Katelyn McNair, Savannah E. Sanchez, Genivaldo G. Z. Silva, Lance Boling, Jeremy J. Barr, et al. 2014. “A Highly Abundant Bacteriophage Discovered in the Unknown Sequences of Human Faecal Metagenomes.” *Nature Communications* 5 (July). https://doi.org/10.1038/ncomms5498.

Edgar, Robert. 2019. “Defining and Interpreting OTUs.” 2019.

Edgar, Robert C. 2016. “UNOISE2: Improved Error-Correction for Illumina 16S and ITS Amplicon Sequencing.” *BioRxiv*, October, 081257. https://doi.org/10.1101/081257.

Edgar, Robert C., Brian J. Haas, Jose C. Clemente, Christopher Quince, and Rob Knight. 2011. “UCHIME Improves Sensitivity and Speed of Chimera Detection.” *Bioinformatics* 27 (16): 2194–2200. https://doi.org/10.1093/bioinformatics/btr381.

Edgar, Robert C. 2018. “Updating the 97% Identity Threshold for 16S Ribosomal RNA OTUs.” Edited by Alfonso Valencia. *Bioinformatics (Oxford, England)* 34 (14): 2371–75. https://doi.org/10.1093/bioinformatics/bty113.

Eloe-Fadrosh, Emiley A., Natalia N. Ivanova, Tanja Woyke, and Nikos C. Kyrpides. 2016. “Metagenomics Uncovers Gaps in Amplicon-Based Detection of Microbial Diversity.” *Nature Microbiology* 1 (4): 15032. https://doi.org/10.1038/nmicrobiol.2015.32.

Eren, A. Murat, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. “Anvi’o: An Advanced Analysis and Visualization Platform for ‘omics Data.” *PeerJ* 3 (October): e1319. https://doi.org/10.7717/peerj.1319.

Fahlén, Annika, Lars Engstrand, Barbara S. Baker, Anne Powles, and Lionel Fry. 2012. “Comparison of Bacterial Microbiota in Skin Biopsies from Normal and Psoriatic Skin.” *Archives of Dermatological Research* 304 (1): 15–22. https://doi.org/10.1007/s00403-011-1189-x.

Faust, Karoline, Leo Lahti, Didier Gonze, Willem M de Vos, and Jeroen Raes. 2015. “Metagenomics Meets Time Series Analysis: Unraveling Microbial Community Dynamics.” *Current Opinion in Microbiology* 25 (June): 56–66. https://doi.org/10.1016/j.mib.2015.04.004.

Fernandes, Andrew D., Jean M. Macklaim, Thomas G. Linn, Gregor Reid, and Gregory B. Gloor. 2013. “ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed

Population RNA-Seq." Edited by John Parkinson. *PLoS ONE* 8 (7): e67019. https://doi.org/10.1371/journal.pone.0067019.

Findley, Keisha, Julia Oh, Joy Yang, Sean Conlan, Clayton Deming, Jennifer A Meyer, Deborah Schoenfeld, et al. 2013. "Topographic Diversity of Fungal and Bacterial Communities in Human Skin." *Nature* 498 (7454): 367–70. https://doi.org/10.1038/nature12171.

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

Flint, Harry J., Karen P. Scott, Sylvia H. Duncan, Petra Louis, and Evelyne Forano. 2012. "Microbial Degradation of Complex Carbohydrates in the Gut." *Gut Microbes* 3 (4): 289–306. https://doi.org/10.4161/gmic.19897.

Foster, Jane A, and Karen-Anne McVey Neufeld. 2013. "Gut-Brain Axis: How the Microbiome Influences Anxiety and Depression." *Trends in Neurosciences* 36 (5): 305–12. https://doi.org/10.1016/j.tins.2013.01.005.

Foulongne, Vincent, Virginie Sauvage, Charles Hebert, Olivier Dereure, Justine Cheval, Meriadeg Ar Gouilh, Kevin Pariente, et al. 2012. "Human Skin Microbiota: High Diversity of DNA Viruses Identified on the Human Skin by High Throughput Sequencing." Edited by Amanda Ewart Toland. *PLoS ONE* 7 (6): e38499. https://doi.org/10.1371/journal.pone.0038499.

Francis, O. E., M. Bendall, S. Manimaran, C. Hong, N. L. Clement, E. Castro-Nallar, Q. Snell, et al. 2013. "Pathoscope: Species Identification and Strain Attribution with Unassembled Sequencing Data." *Genome Research* 23 (10): 1721–29. https://doi.org/10.1101/gr.150151.112.

Frank, D. N., A. L. St. Amand, R. A. Feldman, E. C. Boedeker, N. Harpaz, and N. R. Pace. 2007. "Molecular-Phylogenetic Characterization of Microbial Community Imbalances in Human Inflammatory Bowel Diseases." *Proceedings of the National Academy of Sciences* 104 (34): 13780–85. https://doi.org/10.1073/pnas.0706625104.

Franzosa, Eric A, Katherine Huang, James F Meadow, Dirk Gevers, Katherine P Lemon, Brendan J M Bohannan, and Curtis Huttenhower. 2015. "Identifying Personal

Microbiomes Using Metagenomic Codes." *Proceedings of the National Academy of Sciences* 112 (22): E2930–38. https://doi.org/10.1073/pnas.1423854112.

Fredricks, David N. 2001. "Microbial Ecology of Human Skin in Health and Disease." *Journal of Investigative Dermatology Symposium Proceedings* 6 (3): 167–69. https://doi.org/10.1046/j.0022-202x.2001.00039.x.

Friedman, Jonathan, and Eric J. Alm. 2012. "Inferring Correlation Networks from Genomic Survey Data." Edited by Christian von Mering. *PLoS Computational Biology* 8 (9): e1002687. https://doi.org/10.1371/journal.pcbi.1002687.

Fry, L., B. S. Baker, A. V. Powles, A. Fahlen, and L. Engstrand. 2013. "Is Chronic Plaque Psoriasis Triggered by Microbiota in the Skin?" *British Journal of Dermatology*. https://doi.org/10.1111/bjd.12322.

Galperin, Michael Y., Kira S. Makarova, Yuri I. Wolf, and Eugene V. Koonin. 2015. "Expanded Microbial Genome Coverage and Improved Protein Family Annotation in the COG Database." *Nucleic Acids Research* 43 (D1): D261–69. https://doi.org/10.1093/nar/gku1223.

Gantner, Stephan, Anders F. Andersson, Laura Alonso-Sáez, and Stefan Bertilsson. 2011. "Novel Primers for 16S RRNA-Based Archaeal Community Analyses in Environmental Samples." *Journal of Microbiological Methods* 84 (1): 12–18. https://doi.org/10.1016/j.mimet.2010.10.001.

Gao, Zhan, Chi-hong Tseng, Bruce E. Strober, Zhiheng Pei, and Martin J. Blaser. 2008. "Substantial Alterations of the Cutaneous Bacterial Biota in Psoriatic Lesions." Edited by Niyaz Ahmed. *PLoS ONE* 3 (7): e2719. https://doi.org/10.1371/journal.pone.0002719.

Gevers, Dirk, Subra Kugathasan, Lee A. Denson, Yoshiki Vázquez-Baeza, Will Van Treuren, Boyu Ren, Emma Schwager, et al. 2014. "The Treatment-Naive Microbiome in New-Onset Crohn's Disease." *Cell Host & Microbe* 15 (3): 382–92. https://doi.org/10.1016/j.chom.2014.02.005.

Ghurye, Jay S., Victoria Cepeda-Espinoza, and Mihai Pop. 2016. "Metagenomic Assembly: Overview, Challenges and Applications." *The Yale Journal of Biology and Medicine* 89 (3): 353. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5045144/.

Giallonardo, Francesca Di, Armin Töpfer, Melanie Rey, Sandhya Prabhakaran, Yannick Duport, Christine Leemann, Stefan Schmutz, et al. 2014. "Full-Length Haplotype Reconstruction to Infer the Structure of Heterogeneous Virus Populations." *Nucleic Acids Research* 42 (14): e115. https://doi.org/10.1093/nar/gku537.

Gilbert, Jack A., Martin J. Blaser, J. Gregory Caporaso, Janet K. Jansson, Susan V. Lynch, and Rob Knight. 2018. "Current Understanding of the Human Microbiome." *Nature Medicine* 24 (4): 392–400. https://doi.org/10.1038/nm.4517.

Glassing, Angela, Scot E. Dowd, Susan Galandiuk, Brian Davis, and Rodrick J. Chiodini. 2016. "Inherent Bacterial DNA Contamination of Extraction and Sequencing Reagents May Affect Interpretation of Microbiota in Low Bacterial Biomass Samples." *Gut Pathogens* 8 (1): 24. https://doi.org/10.1186/s13099-016-0103-7.

Gloor, Gregory B., Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. 2017. "Microbiome Datasets Are Compositional: And This Is Not Optional." *Frontiers in Microbiology* 8 (November): 2224. https://doi.org/10.3389/fmicb.2017.02224.

Gong, J.Q., L. Lin, T. Lin, F. Hao, F.Q. Zeng, Z.G. Bi, D. Yi, and B. Zhao. 2006. "Skin Colonization by Staphylococcus Aureus in Patients with Eczema and Atopic Dermatitis and Relevant Combined Topical Therapy: A Double-Blind Multicentre Randomized Controlled Trial." *British Journal of Dermatology* 155 (4): 680–87. https://doi.org/10.1111/j.1365-2133.2006.07410.x.

Gonze, Didier, Katharine Z Coyte, Leo Lahti, and Karoline Faust. 2018. "Microbial Communities as Dynamical Systems." *Current Opinion in Microbiology* 44 (August): 41–49. https://doi.org/10.1016/j.mib.2018.07.004.

Graspeuntner, Simon, Nathalie Loeper, Sven Künzel, John F. Baines, and Jan Rupp. 2018. "Selection of Validated Hypervariable Regions Is Crucial in 16S-Based Microbiota Studies of the Female Genital Tract." *Scientific Reports* 8 (1): 9678. https://doi.org/10.1038/s41598-018-27757-8.

Grice, Elizabeth a., Heidi H. Kong, Sean Conlan, Clayton B. Deming, Joie Davis, Alice C. Young, Nisc Comparative Sequencing Program, et al. 2009. "Topographical and Temporal Diversity of the Human Skin." *Science* 324 (5931): 1190–92. https://doi.org/10.1126/science.1171700.Topographical.

Grice, Elizabeth A., Heidi H. Kong, Gabriel Renaud, Alice C. Young, Gerard G. Bouffard, Robert W. Blakesley, Tyra G. Wolfsberg, Maria L. Turner, and Julia A. Segre. 2008. “A Diversity Profile of the Human Skin Microbiota.” *Genome Research* 18 (7): 1043–50. https://doi.org/10.1101/gr.075549.107.

Grice, Elizabeth A, and Julia A Segre. 2012. “The Human Microbiome: Our Second Genome.” *Annual Review of Genomics and Human Genetics* 13: 151–70. https://doi.org/10.1146/annurev-genom-090711-163814.

Guma, Sergei, Remegio Maglantay, Ryan Lau, Rosemary Wieczorek, Jonathan Melamed, Fang-Ming Deng, Ming Zhou, et al. 2016. “Papillary Urothelial Carcinoma with Squamous Differentiation in Association with Human Papilloma Virus: Case Report and Literature Review.” *American Journal of Clinical and Experimental Urology* 4 (1): 12–16. http://www.ncbi.nlm.nih.gov/pubmed/27069958.

Gustafsson, B, and J Carstensen. 1999. “Evidence of Space-Time Clustering of Childhood Acute Lymphoblastic Leukaemia in Sweden.” *British Journal of Cancer* 79 (3–4): 655–57. https://doi.org/10.1038/sj.bjc.6690103.

Haas, Brian J, Dirk Gevers, Ashlee M Earl, Mike Feldgarden, Doyle V Ward, Georgia Giannoukos, Dawn Ciulla, et al. 2011. “Chimeric 16S RRNA Sequence Formation and Detection in Sanger and 454-Pyrosequenced PCR Amplicons.” *Genome Research* 21 (3): 494–504. https://doi.org/10.1101/gr.112730.110.

Haft, Daniel H, Jeremy D Selengut, and Owen White. 2003. “The TIGRFAMs Database of Protein Families.” *Nucleic Acids Research* 31 (1): 371–73. https://doi.org/10.1093/nar/gkg128.

Hannigan, Geoffrey D., Jacquelyn S. Meisel, Amanda S. Tyldsley, Qi Zheng, Brendan P. Hodkinson, Adam J. Sanmiguel, Samuel Minot, Frederic D. Bushman, and Elizabeth A. Grice. 2015. “The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity, Genetic Enrichment, and Dynamic Associations with the Host Microbiome.” *MBio* 6 (5). https://doi.org/10.1128/mBio.01578-15.

Hartstra, Annick V., Kristien E.C. Bouter, Fredrik Bäckhed, and Max Nieuwdorp. 2015. “Insights Into the Role of the Microbiome in Obesity and Type 2 Diabetes.” *Diabetes Care* 38 (1): 159–65. https://doi.org/10.2337/dc14-0769.

Hillmann, Benjamin, Gabriel A. Al-Ghalith, Robin R. Shields-Cutler, Qiyun Zhu, Daryl M. Gohl, Kenneth B. Beckman, Rob Knight, and Dan Knights. 2018. "Evaluating the Information Content of Shallow Shotgun Metagenomics." Edited by John F. Rawls. *MSystems* 3 (6): e00069-18. https://doi.org/10.1128/mSystems.00069-18.

Hilty, Markus, Conor Burke, Helder Pedro, Paul Cardenas, Andy Bush, Cara Bossley, Jane Davies, et al. 2010. "Disordered Microbial Communities in Asthmatic Airways." Edited by Olivier Neyrolles. *PLoS ONE* 5 (1): e8578. https://doi.org/10.1371/journal.pone.0008578.

Howe, Adina Chuang, Janet K Jansson, Stephanie A Malfatti, Susannah G Tringe, James M Tiedje, and C Titus Brown. 2014. "Tackling Soil Diversity with the Assembly of Large, Complex Metagenomes." *Proceedings of the National Academy of Sciences of the United States of America* 111 (13): 4904–9. https://doi.org/10.1073/pnas.1402564111.

Hsu, Bryan B., Travis E. Gibson, Vladimir Yeliseyev, Qing Liu, Lorena Lyon, Lynn Bry, Pamela A. Silver, and Georg K. Gerber. 2019. "Dynamic Modulation of the Gut Microbiota and Metabolome by Bacteriophages in a Mouse Model." *Cell Host & Microbe* 25 (6): 803-814.e5. https://doi.org/10.1016/j.chom.2019.05.001.

Hu, Yue O O, Luisa W Hugerth, Carina Bengtsson, Arlisa Alisjahbana, Maike Seifert, Anaga Kamal, Åsa Sjöling, et al. 2018. "Bacteriophages Synergize with the Gut Microbial Community To Combat Salmonella." Edited by Katrine L. Whiteson. *MSystems* 3 (5). https://doi.org/10.1128/mSystems.00119-18.

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

Huffnagle, G B, R P Dickson, and N W Lukacs. 2017. "The Respiratory Tract Microbiome and Lung Inflammation: A Two-Way Street." *Mucosal Immunology* 10 (2): 299–306. https://doi.org/10.1038/mi.2016.108.

Hugenholtz, Philip. 2002. "Exploring Prokaryotic Diversity in the Genomic Era." *Genome Biology* 3 (2): REVIEWS0003. http://www.ncbi.nlm.nih.gov/pubmed/11864374.

Hughes, J B, J J Hellmann, T H Ricketts, and B J Bohannan. 2001. “Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity.” *Applied and Environmental Microbiology* 67 (10): 4399–4406. https://doi.org/10.1128/aem.67.10.4399-4406.2001.

Hunt, Martin, Astrid Gall, Swee Hoe Ong, Jacqui Brener, Bridget Ferns, Philip Goulder, Eleni Nastouli, Jacqueline A Keane, Paul Kellam, and Thomas D Otto. 2015. “IVA: Accurate de Novo Assembly of RNA Virus Genomes.” *Bioinformatics (Oxford, England)* 31 (14): 2374–76. https://doi.org/10.1093/bioinformatics/btv120.

Hunter, Sarah, Matthew Corbett, Hubert Denise, Matthew Fraser, Alejandra Gonzalez-Beltran, Christopher Hunter, Philip Jones, et al. 2014. “EBI Metagenomics—a New Resource for the Analysis and Archiving of Metagenomic Data.” *Nucleic Acids Research* 42 (D1): D600–606. https://doi.org/10.1093/nar/gkt961.

Hurwitz, Bonnie L., Jana M. U’Ren, and Ken Youens-Clark. 2016. “Computational Prospecting the Great Viral Unknown.” Edited by Andrew Millard. *FEMS Microbiology Letters* 363 (10): fnw077. https://doi.org/10.1093/femsle/fnw077.

Huson, Daniel H, Alexander F Auch, Ji Qi, and Stephan C Schuster. 2007. “MEGAN Analysis of Metagenomic Data.” *Genome Research* 17 (3): 377–86. https://doi.org/10.1101/gr.5969107.

Isa, Adiba, Peter Priftakis, Kristina Broliden, and Britt Gustafsson. 2004. “Human Parvovirus B19 DNA Is Not Detected in Guthrie Cards from Children Who Have Developed Acute Lymphoblastic Leukemia.” *Pediatric Blood & Cancer* 42 (4): 357–60. https://doi.org/10.1002/pbc.20001.

Jo, Jay-Hyun, Elizabeth A. Kennedy, and Heidi H. Kong. 2016. “Topographical and Physiological Differences of the Skin Mycobiome in Health and Disease.” *Virulence*, October, 00–00. https://doi.org/10.1080/21505594.2016.1249093.

Jonsson, Viktor, Tobias Österlund, Olle Nerman, and Erik Kristiansson. 2018a. “Modelling of Zero-Inflation Improves Inference of Metagenomic Gene Count Data.” *Statistical Methods in Medical Research*, November, 096228021881135. https://doi.org/10.1177/0962280218811354.

Juul, Sissel, Fernando Izquierdo, Adam Hurst, Xiaoguang Dai, Amber Wright, Eugene

Kulesha, Roger Pettett, and Daniel J. Turner. 2015. “What’s in My Pot? Real-Time Species Identification on the MinION^{TM}.” *BioRxiv*, November, 030742. https://doi.org/10.1101/030742.

Kanehisa, Minoru, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2016. “KEGG as a Reference Resource for Gene and Protein Annotation.” *Nucleic Acids Research* 44 (D1): D457–62. https://doi.org/10.1093/nar/gkv1070.

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

Kaul, Abhishek, Siddhartha Mandal, Ori Davidov, and Shyamal D Peddada. 2017. “Analysis of Microbiome Data in the Presence of Excess Zeros.” *Frontiers in Microbiology* 8: 2114. https://doi.org/10.3389/fmicb.2017.02114.

Keegan, Kevin P., Elizabeth M. Glass, and Folker Meyer. 2016. “MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function.” In *Methods in Molecular Biology (Clifton, N.J.)*, 1399:207–33. https://doi.org/10.1007/978-1-4939-3369-3_13.

Kernbauer, Elisabeth, Yi Ding, and Ken Cadwell. 2014. “An Enteric Virus Can Replace the Beneficial Function of Commensal Bacteria.” *Nature* 516 (7529): 94–98. https://doi.org/10.1038/nature13960.

Khalesi, Saman, Nick Bellissimo, Corneel Vandelanotte, Susan Williams, Dragana Stanley, and Christopher Irwin. 2019. “A Review of Probiotic Supplementation in Healthy Adults: Helpful or Hype?” *European Journal of Clinical Nutrition* 73 (1): 24–37. https://doi.org/10.1038/s41430-018-0135-9.

Kim, Daehwan, Li Song, Florian P Breitwieser, and Steven L Salzberg. 2016. “Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences.” *Genome Research* 26 (12): 1721–29. https://doi.org/10.1101/gr.210641.116.

Knight, Rob, Alison Vrbanac, Bryn C. Taylor, Alexander Aksenov, Chris Callewaert, Justine Debelius, Antonio Gonzalez, et al. 2018. “Best Practices for Analysing Microbiomes.” *Nature Reviews Microbiology* 16 (7): 410–22. https://doi.org/10.1038/s41579-018-0029-9.

Kobayashi, Tetsuro, Martin Glatz, Keisuke Horiuchi, Hiroshi Kawasaki, Haruhiko Akiyama, Daniel H. Kaplan, Heidi H. Kong, Masayuki Amagai, and Keisuke Nagao. 2015. "Dysbiosis and Staphyloccus Aureus Colonization Drives Inflammation in Atopic Dermatitis." *Immunity* 42 (4): 756–66. https://doi.org/10.1016/j.immuni.2015.03.014.

Kong, Heidi H., Julia Oh, Clay Deming, Sean Conlan, Elizabeth A. Grice, Melony A. Beatson, Effie Nomicos, Eric C. Polley, Hirsh D. Komarow, Jim Mullikin, et al. 2012. "Temporal Shifts in the Skin Microbiome Associated with Disease Flares and Treatment in Children with Atopic Dermatitis." *Genome Research* 22 (5): 850–59. https://doi.org/10.1101/gr.131029.111.

Kong, Heidi H., Julia Oh, Clay Deming, Sean Conlan, Elizabeth A. Grice, Melony A. Beatson, Effie Nomicos, Eric C. Polley, Hirsh D. Komarow, Patrick R. Murray, et al. 2012. "Temporal Shifts in the Skin Microbiome Associated with Disease Flares and Treatment in Children with Atopic Dermatitis." *Genome Research* 22 (5): 850–59. https://doi.org/10.1101/gr.131029.111.

Kong, Heidi H, Björn Andersson, Thomas Clavel, John E Common, Scott A Jackson, Nathan D Olson, Julia A Segre, and Claudia Traidl-Hoffmann. 2017. "Performing Skin Microbiome Research: A Method to the Madness." *The Journal of Investigative Dermatology* 137 (3): 561–68. https://doi.org/10.1016/j.jid.2016.10.033.

Konopka, Allan. 2009. "What Is Microbial Community Ecology?" *The ISME Journal* 3 (11): 1223–30. https://doi.org/10.1038/ismej.2009.88.

Koren, Sergey, Gregory P Harhay, Timothy P L Smith, James L Bono, Dayna M Harhay, Scott D Mcvey, Diana Radune, Nicholas H Bergman, and Adam M Phillippy. 2013. "Reducing Assembly Complexity of Microbial Genomes with Single-Molecule Sequencing." *Genome Biology* 14 (9): R101. https://doi.org/10.1186/gb-2013-14-9-r101.

Kortright, Kaitlyn E, Benjamin K Chan, Jonathan L Koff, and Paul E Turner. 2019. "Phage Therapy: A Renewed Approach to Combat Antibiotic-Resistant Bacteria." *Cell Host & Microbe* 25 (2): 219–32. https://doi.org/10.1016/j.chom.2019.01.014.

Kraszewska-Głomba, Barbara, Agnieszka Matkowska-Kocjan, and Leszek Szenborn. 2015. "The Pathogenesis of Periodic Fever, Aphthous Stomatitis, Pharyngitis, and Cervical

Adenitis Syndrome: A Review of Current Research." *Mediators of Inflammation* 2015: 563876. https://doi.org/10.1155/2015/563876.

Kuczynski, Justin, Christian L. Lauber, William A. Walters, Laura Wegener Parfrey, José C. Clemente, Dirk Gevers, and Rob Knight. 2011. "Experimental and Analytical Tools for Studying the Human Microbiome." *Nature Reviews Genetics* 13 (1): 47–58. https://doi.org/10.1038/nrg3129.

Lacey, Noreen, Síona Ní Raghallaigh, and Frank C Powell. 2011. "Demodex Mites--Commensals, Parasites or Mutualistic Organisms?" *Dermatology* 222 (2): 128–30. https://doi.org/10.1159/000323009.

Lahti, Leo, Sudarshan Shetty, and et al. 2017. "Tools for Microbiome Analysis in R." http://microbiome.github.com/microbiome.

Lai, Yuping, Anna L. Cogen, Katherine A. Radek, Hyun Jeong Park, Daniel T. MacLeod, Anke Leichtle, Allen F. Ryan, Anna Di Nardo, and Richard L. Gallo. 2010. "Activation of TLR2 by a Small Molecule Produced by Staphylococcus Epidermidis Increases Antimicrobial Defense against Bacterial Skin Infections." *Journal of Investigative Dermatology* 130 (9): 2211–21. https://doi.org/10.1038/jid.2010.123.

Lai, Yuping, Anna Di Nardo, Teruaki Nakatsuji, Anke Leichtle, Yan Yang, Anna L Cogen, Zi-Rong Wu, et al. 2009. "Commensal Bacteria Regulate Toll-like Receptor 3–Dependent Inflammation after Skin Injury." *Nature Medicine* 15 (12): 1377–82. https://doi.org/10.1038/nm.2062.

Lang, Jennifer M., and M Eric Benbow. 2013. "Species Interactions and Competition." *Nature Education Knowledge* 4 (8). https://www.nature.com/scitable/knowledge/library/species-interactions-and-competition-102131429.

Langille, Morgan G I, Jesse Zaneveld, J Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A Reyes, Jose C Clemente, et al. 2013. "Predictive Functional Profiling of Microbial Communities Using 16S RRNA Marker Gene Sequences." *Nature Biotechnology* 31 (9): 814–21. https://doi.org/10.1038/nbt.2676.

Lanteri, M. C., F. Vahidnia, S. Tan, J. T. Stapleton, P. J. Norris, J. Heitman, X. Deng, et al. 2015. "Downregulation of Cytokines and Chemokines by GB Virus C After

Transmission Via Blood Transfusion in HIV-Positive Blood Recipients." *Journal of Infectious Diseases* 211 (10): 1585–96. https://doi.org/10.1093/infdis/jiu660.

LeBlanc, Jean Guy, Christian Milani, Graciela Savoy de Giori, Fernando Sesma, Douwe van Sinderen, and Marco Ventura. 2013. "Bacteria as Vitamin Suppliers to Their Host: A Gut Microbiota Perspective." *Current Opinion in Biotechnology* 24 (2): 160–68. https://doi.org/10.1016/J.COPBIO.2012.08.005.

Leger, Anthony J. St., Jigar V. Desai, Rebecca A. Drummond, Abirami Kugadas, Fatimah Almaghrabi, Phyllis Silver, Kumarkrishna Raychaudhuri, et al. 2017. "An Ocular Commensal Protects against Corneal Infection by Driving an Interleukin-17 Response from Mucosal Γδ T Cells." *Immunity* 47 (1): 148-158.e5. https://doi.org/10.1016/j.immuni.2017.06.014.

Leung, D. Y M, J. B. Travers, R. Giorno, D. A. Norris, R. Skinner, J. Aelion, L. V. Kazemi, et al. 1995. "Evidence for a Streptococcal Superantigen-Driven Process in Acute Guttate Psoriasis." *Journal of Clinical Investigation* 96 (5): 2106–12. https://doi.org/10.1172/JCI118263.

Li, D., C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. 2015. "MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph." *Bioinformatics* 31 (10): 1674–76. https://doi.org/10.1093/bioinformatics/btv033.

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

Li, Linlin, Xutao Deng, Piyada Linsuwanon, David Bangsberg, Mwebesa Bosco Bwana, Peter Hunt, Jeffrey N Martin, Steven G Deeks, and Eric Delwart. 2013. "AIDS Alters the Commensal Plasma Virome." *Journal of Virology* 87 (19): 10912–15. https://doi.org/10.1128/JVI.01839-13.

Lindahl, B D, R H Nilsson, L Tedersoo, K Abarenkov, T Carlsen, R Kjoller, U Koljalg, et al. 2013. "Fungal Community Analysis by High-Throughput Sequencing of Amplified Markers - a User's Guide." *New Phytologist* 199 (1): 288–99. https://doi.org/10.1111/nph.12243.

Lindgreen, Stinus, Karen L Adair, and Paul Gardner. 2015. “An Evaluation of the Accuracy and Speed of Metagenome Analysis Tools.” *BioRxiv*. Cold Spring Harbor Labs Journals. https://doi.org/10.1101/017830.

Lloyd-Price, Jason, Anup Mahurkar, Gholamali Rahnavard, Jonathan Crabtree, Joshua Orvis, A. Brantley Hall, Arthur Brady, et al. 2017. “Strains, Functions and Dynamics in the Expanded Human Microbiome Project.” *Nature* 550 (7674): 61–66. https://doi.org/10.1038/nature23889.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550. https://doi.org/10.1186/s13059-014-0550-8.

Lovell, David, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. 2015. “Proportionality: A Valid Alternative to Correlation for Relative Data.” Edited by Roland L. Dunbrack Jr. *PLOS Computational Biology* 11 (3): e1004075. https://doi.org/10.1371/journal.pcbi.1004075.

Lozupone, C., and R. Knight. 2005. “UniFrac: A New Phylogenetic Method for Comparing Microbial Communities.” *Applied and Environmental Microbiology* 71 (12): 8228–35. https://doi.org/10.1128/AEM.71.12.8228-8235.2005.

Lu, Jennifer, Florian P. Breitwieser, Peter Thielen, and Steven L. Salzberg. 2017. “Bracken: Estimating Species Abundance in Metagenomics Data.” *PeerJ Computer Science* 3 (January): e104. https://doi.org/10.7717/peerj-cs.104.

Lysholm, Fredrik, Anna Wetterbom, Cecilia Lindau, Hamid Darban, Annelie Bjerkner, Kristina Fahlander, A. Michael Lindberg, Bengt Persson, Tobias Allander, and Björn Andersson. 2012. “Characterization of the Viral Microbiome in Patients with Severe Lower Respiratory Tract Infections, Using Metagenomic Sequencing.” Edited by Sarah K. Highlander. *PloS One* 7 (2): e30875. https://doi.org/10.1371/journal.pone.0030875.

Maaten, Laurens Van Der, and Geoffrey Hinton. 2008. “Visualizing Data Using T-SNE.” *Journal of Machine Learning Research*.

Mah, Thien-Fah C, and George A O’Toole. 2001. “Mechanisms of Biofilm Resistance to Antimicrobial Agents.” *Trends in Microbiology* 9 (1): 34–39.

https://doi.org/10.1016/S0966-842X(00)01913-2.

Mahé, Frédéric, Torbjørn Rognes, Christopher Quince, Colomban de Vargas, and Micah Dunthorn. 2014. “Swarm: Robust and Fast Clustering Method for Amplicon-Based Studies.” *PeerJ* 2 (September): e593. https://doi.org/10.7717/peerj.593.

Maia, Ana Teresa, Roxane Tussiwand, Giovanni Cazzaniga, Paolo Rebulla, Susan Colman, Andrea Biondi, and Mel Greaves. 2004. “Identification of Preleukemic Precursors of Hyperdiploid Acute Lymphoblastic Leukemia in Cord Blood.” *Genes, Chromosomes and Cancer* 40 (1): 38–43. https://doi.org/10.1002/gcc.20010.

Mandal, Siddhartha, Will Van Treuren, Richard A White, Merete Eggesbø, Rob Knight, and Shyamal D Peddada. 2015. “Analysis of Composition of Microbiomes: A Novel Method for Studying Microbial Composition.” *Microbial Ecology in Health & Disease* 26 (May): 27663. https://doi.org/10.3402/mehd.v26.27663.

McInnes, Leland, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. “UMAP: Uniform Manifold Approximation and Projection.” *Journal of Open Source Software* 3 (29): 861. https://doi.org/10.21105/joss.00861.

McMurdie, Paul J., and Susan Holmes. 2013. “Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data.” Edited by Michael Watson. *PLoS ONE* 8 (4): e61217. https://doi.org/10.1371/journal.pone.0061217.

———. 2014. “Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible.” Edited by Alice Carolyn McHardy. *PLoS Computational Biology* 10 (4): e1003531. https://doi.org/10.1371/journal.pcbi.1003531.

Meisel, Jacquelyn S., Geoffrey D. Hannigan, Amanda S. Tyldsley, Adam J. SanMiguel, Brendan P. Hodkinson, Qi Zheng, and Elizabeth A. Grice. 2016. “Skin Microbiome Surveys Are Strongly Influenced by Experimental Design.” *Journal of Investigative Dermatology* 136 (5). https://doi.org/10.1016/j.jid.2016.01.016.

Metges, Cornelia C. 2000. “Contribution of Microbial Amino Acids to Amino Acid Homeostasis of the Host.” *The Journal of Nutrition* 130 (7): 1857S-1864S. https://doi.org/10.1093/jn/130.7.1857S.

Mikheenko, Alla, Vladislav Saveliev, and Alexey Gurevich. 2016. “MetaQUAST: Evaluation of Metagenome Assemblies.” *Bioinformatics (Oxford, England)* 32 (7):

1088–90. https://doi.org/10.1093/bioinformatics/btv697.

Milanese, Alessio, Daniel R Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Pascal Hingamp, et al. 2019. “Microbial Abundance, Activity and Population Genomic Profiling with MOTUs2.” *Nature Communications* 10 (1): 1014. https://doi.org/10.1038/s41467-019-08844-4.

Minot, Samuel, Alexandra Bryson, Christel Chehoud, Gary D Wu, James D Lewis, and Frederic D Bushman. 2013. “Rapid Evolution of the Human Gut Virome.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (30): 12450–55. https://doi.org/10.1073/pnas.1300833110.

Mistry, Jaina, Robert D. Finn, Sean R. Eddy, Alex Bateman, and Marco Punta. 2013a. “Challenges in Homology Search: HMMER3 and Convergent Evolution of Coiled-Coil Regions.” *Nucleic Acids Research* 41 (12). https://doi.org/10.1093/nar/gkt263.

Mokili, John L, Forest Rohwer, and Bas E Dutilh. 2012. “Metagenomics and Future Perspectives in Virus Discovery.” *Current Opinion in Virology* 2 (1): 63–77. https://doi.org/10.1016/j.coviro.2011.12.004.

Mollerup, Sarah, Jens Friis-Nielsen, Lasse Vinner, Thomas Arn Hansen, Stine Raith Richter, Helena Fridholm, Jose Alejandro Romero Herrera, et al. 2016. “Propionibacterium Acnes - Disease Causing Agent or Common Contaminant? Detection in Diverse Patient Samples by next Generation Sequencing.” *Journal of Clinical Microbiology*, January, JCM.02723-15-. https://doi.org/10.1128/JCM.02723-15.

Morgan, Xochitl C, Boyko Kabakchiev, Levi Waldron, Andrea D Tyler, Timothy L Tickle, Raquel Milgrom, Joanne M Stempak, et al. 2015. “Associations between Host Gene Expression, the Mucosal Microbiome, and Clinical Outcome in the Pelvic Pouch of Patients with Inflammatory Bowel Disease.” *Genome Biology* 16 (1): 67. https://doi.org/10.1186/s13059-015-0637-x.

Morgan, Xochitl C, Timothy L Tickle, Harry Sokol, Dirk Gevers, Kathryn L Devaney, Doyle V Ward, Joshua A Reyes, et al. 2012. “Dysfunction of the Intestinal Microbiome in Inflammatory Bowel Disease and Treatment.” *Genome Biology* 13 (9): R79. https://doi.org/10.1186/gb-2012-13-9-r79.

Morton, James T., Clarisse Marotz, Alex Washburne, Justin Silverman, Livia S. Zaramela, Anna Edlund, Karsten Zengler, and Rob Knight. 2019. "Establishing Microbial Composition Measurement Standards with Reference Frames." *Nature Communications* 10 (1): 2719. https://doi.org/10.1038/s41467-019-10656-5.

Munck, Christian, Mads Albertsen, Amar Telke, Mostafa Ellabaan, Per Halkjær Nielsen, and Morten O A Sommer. 2015. "Limited Dissemination of the Wastewater Treatment Plant Core Resistome." *Nature Communications* 6: 8452. https://doi.org/10.1038/ncomms9452.

Naccache, Samia N, Alexander L Greninger, Deanna Lee, Lark L Coffey, Tung Phan, Annie Rein-Weston, Andrew Aronsohn, John Hackett, Eric L Delwart, and Charles Y Chiu. 2013. "The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns." *Journal of Virology* 87: 11966–77. https://doi.org/10.1128/JVI.02323-13.

Naik, Shruti, Nicolas Bouladoux, Jonathan L. Linehan, Seong-Ji Han, Oliver J. Harrison, Christoph Wilhelm, Sean Conlan, et al. 2015. "Commensal–Dendritic-Cell Interaction Specifies a Unique Protective Skin Immune Signature." *Nature* 520 (7545): 104–8. https://doi.org/10.1038/nature14052.

Naik, Shruti, Nicolas Bouladoux, Christoph Wilhelm, Michael J Molloy, Rosalba Salcedo, Wolfgang Kastenmuller, Clayton Deming, et al. 2012. "Compartmentalized Control of Skin Immunity by Resident Commensals." *Science (New York, N.Y.)* 337 (6098): 1115–19. https://doi.org/10.1126/science.1225152.

Nakamura, Yuumi, Jon Oscherwitz, Kemp B. Cease, Susana M. Chan, Raul Muñoz-Planillo, Mizuho Hasegawa, Amer E. Villaruz, et al. 2013. "Staphylococcus δ-Toxin Induces Allergic Skin Disease by Activating Mast Cells." *Nature* 503 (7476): 397–401. https://doi.org/10.1038/nature12655.

Nakatsuji, Teruaki, Hsin-I Chiang, Shangi B Jiang, Harish Nagarajan, Karsten Zengler, and Richard L Gallo. 2013. "The Microbiome Extends to Subepidermal Compartments of Normal Skin." *Nature Communications* 4: 1431. https://doi.org/10.1038/ncomms2441.

Natsch, Andreas, Joachim Schmid, and Felix Flachsmann. 2004. "Identification of Odoriferous Sulfanylalkanols in Human Axilla Secretions and Their Formation through Cleavage of Cysteine Precursors by a C-S Lyase Isolated from Axilla

Bacteria.” *Chemistry and Biodiversity* 1 (7): 1058–72. https://doi.org/10.1002/cbdv.200490079.

Nayfach, Stephen, Patrick H Bradley, Stacia K Wyman, Timothy J Laurent, Alex Williams, Jonathan A Eisen, Katherine S Pollard, and Thomas J Sharpton. 2015. “Automated and Accurate Estimation of Gene Family Abundance from Shotgun Metagenomes.” *PLoS Computational Biology* 11 (11): e1004573. https://doi.org/10.1371/journal.pcbi.1004573.

Nearing, Jacob T, Gavin M Douglas, André M Comeau, and Morgan G.I. Langille. 2018. “Denoising the Denoisers: An Independent Evaluation of Microbiome Sequence Error-Correction Approaches.” *PeerJ* 6 (August): e5364. https://doi.org/10.7717/peerj.5364.

Neto, Antonio Galvao, April Whitaker, and Zhiheng Pei. 2016. “Microbiome and Potential Targets for Chemoprevention of Esophageal Adenocarcinoma.” *Seminars in Oncology* 43 (1): 86–96. https://doi.org/10.1053/j.seminoncol.2015.09.005.

Nguyen, Nam-Phuong, Tandy Warnow, Mihai Pop, and Bryan White. 2016. “A Perspective on 16S RRNA Operational Taxonomic Unit Clustering Using Sequence Similarity.” *Npj Biofilms and Microbiomes* 2 (1): 16004. https://doi.org/10.1038/npjbiofilms.2016.4.

Norman, Jason M, Scott A Handley, Megan T Baldridge, Lindsay Droit, Catherine Y Liu, Brian C Keller, Amal Kambal, et al. 2015. “Disease-Specific Alterations in the Enteric Virome in Inflammatory Bowel Disease.” *Cell* 160 (3): 447–60. https://doi.org/10.1016/j.cell.2015.01.002.

Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel Pevzner. 2016. “MetaSPAdes: A New Versatile de Novo Metagenomics Assembler,” April. http://arxiv.org/abs/1604.03071.

O’Dwyer, David N., Robert P. Dickson, and Bethany B. Moore. 2016. “The Lung Microbiome, Immunity, and the Pathogenesis of Chronic Lung Disease.” *The Journal of Immunology* 196 (12): 4839–47. https://doi.org/10.4049/jimmunol.1600279.

Oh, Julia, Allyson L. Byrd, Clay Deming, Sean Conlan, Heidi H. Kong, and Julia A. Segre. 2014. “Biogeography and Individuality Shape Function in the Human Skin Metagenome.” *Nature* 514 (7520): 59–64. https://doi.org/10.1038/nature13786.

Oh, Julia, Allyson L. Byrd, Morgan Park, Heidi H. Kong, and Julia A. Segre. 2016. "Temporal Stability of the Human Skin Microbiome." *Cell* 165 (4): 854–66. https://doi.org/10.1016/j.cell.2016.04.008.

Oh, Julia, Alexandra F. Freeman, Morgan Park, Robert Sokolic, Fabio Candotti, Steven M. Holland, Julia A. Segre, and Heidi H. Kong. 2013. "The Altered Landscape of the Human Skin Microbiome in Patients with Primary Immunodeficiencies." *Genome Research* 23 (12): 2103–14. https://doi.org/10.1101/gr.159467.113.

Oksanen, Jari, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, et al. 2019. "The Vegan Package."

Olsen, G J, D J Lane, S J Giovannoni, N R Pace, and D A Stahl. 1986. "Microbial Ecology and Evolution: A Ribosomal RNA Approach." *Annual Review of Microbiology* 40 (1): 337–65. https://doi.org/10.1146/annurev.mi.40.100186.002005.

Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. "Mash: Fast Genome and Metagenome Distance Estimation Using MinHash." *Genome Biology* 17 (1): 132. https://doi.org/10.1186/s13059-016-0997-x.

Otto, Michael. 2010. "Staphylococcus Colonization of the Skin and Antimicrobial Peptides." *Expert Review of Dermatology* 5 (2): 183–95. https://doi.org/10.1586/edm.10.6.

Ounit, Rachid, and Stefano Lonardi. 2016. "Higher Classification Sensitivity of Short Metagenomic Reads with CLARK-S-S." *Bioinformatics* 32 (24): 3823–25. https://doi.org/10.1093/bioinformatics/btw542.

Palacios, Gustavo, Julian Druce, Lei Du, Thomas Tran, Chris Birch, Thomas Briese, Sean Conlan, et al. 2008. "A New Arenavirus in a Cluster of Fatal Transplant-Associated Diseases." *The New England Journal of Medicine* 358 (10): 991–98. https://doi.org/10.1056/NEJMoa073785.

Parisi, Rosa, Deborah P M Symmons, Christopher E M Griffiths, and Darren M Ashcroft. 2012. "Global Epidemiology of Psoriasis: A Systematic Review of Incidence and Prevalence." *Journal of Investigative Dermatology* 133 (2): 377–85. https://doi.org/10.1038/jid.2012.339.

Park, Sang-Cheol, and Sungho Won. 2018. “Evaluation of 16S RRNA Databases for Taxonomic Assignments Using Mock Community.” *Genomics & Informatics* 16 (4): e24. https://doi.org/10.5808/GI.2018.16.4.e24.

Parks, Donovan H, Michael Imelfort, Connor T Skennerton, Philip Hugenholtz, and Gene W Tyson. 2014. “CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes.” *Genome Research*, October. https://doi.org/10.7287/peerj.preprints.554v1.

Pascolini, Chiara, Jolinda Sinagra, Simone Pecetta, Valentina Bordignon, Alessandra De Santis, Laura Cilli, Viviana Cafiso, et al. 2011. “Molecular and Immunological Characterization of Staphylococcus Aureus in Pediatric Atopic Dermatitis: Implications for Prophylaxis and Clinical Management.” *Clinical & Developmental Immunology* 2011: 718708. https://doi.org/10.1155/2011/718708.

Patro, Rob, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. 2017. “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.” *Nature Methods* 14 (4): 417–19. https://doi.org/10.1038/nmeth.4197.

Paulino, L. C., C.-H. Tseng, B. E. Strober, and M. J. Blaser. 2006. “Molecular Analysis of Fungal Microbiota in Samples from Healthy Human Skin and Psoriatic Lesions.” *Journal of Clinical Microbiology* 44 (8): 2933–41. https://doi.org/10.1128/JCM.00785-06.

Paulino, Luciana C., Chi-Hong Tseng, and Martin J. Blaser. 2008. “Analysis of Malassezia Microbiota in Healthy Superficial Human Skin and in Psoriatic Lesions by Multiplex Real-Time PCR.” *FEMS Yeast Research* 8 (3): 460–71. https://doi.org/10.1111/j.1567-1364.2008.00359.x.

Paulson, Joseph N, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. 2013. “Differential Abundance Analysis for Microbial Marker-Gene Surveys.” *Nature Methods* 10 (12): 1200–1202. https://doi.org/10.1038/nmeth.2658.

Pedersen, Helle Krogh, Sofia K. Forslund, Valborg Gudmundsdottir, Anders Østergaard Petersen, Falk Hildebrand, Tuulia Hyötyläinen, Trine Nielsen, et al. 2018. “A Computational Framework to Integrate High-Throughput ‘-Omics’ Datasets for the Identification of Potential Mechanistic Links.” *Nature Protocols* 13 (12): 2781–2800. https://doi.org/10.1038/s41596-018-0064-z.

Peng, Y., H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin. 2012. "IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth." *Bioinformatics* 28 (11): 1420–28. https://doi.org/10.1093/bioinformatics/bts174.

Pereira-Marques, Joana, Anne Hout, Rui M. Ferreira, Michiel Weber, Ines Pinto-Ribeiro, Leen-Jan van Doorn, Cornelis Willem Knetsch, and Ceu Figueiredo. 2019. "Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis." *Frontiers in Microbiology* 10 (June): 1277. https://doi.org/10.3389/fmicb.2019.01277.

Petersen, Charisse, and June L. Round. 2014. "Defining Dysbiosis and Its Influence on Host Immunity and Disease." *Cellular Microbiology* 16 (7): 1024–33. https://doi.org/10.1111/cmi.12308.

Popgeorgiev, Nikolay, Mickaël Boyer, Laura Fancello, Sonia Monteil, Catherine Robert, Romain Rivet, Claude Nappez, et al. 2013. "Marseillevirus-like Virus Recovered from Blood Donated by Asymptomatic Humans." *The Journal of Infectious Diseases* 208 (7): 1042–50. https://doi.org/10.1093/infdis/jit292.

Pou, C., M. Barrientos-Somarribas, S. Marin-Juan, G. Bogdanovic, A. Bjerkner, T. Allander, B. Gustafsson, and B. Andersson. 2018. "Virome Definition in Cerebrospinal Fluid of Patients with Neurological Complications after Hematopoietic Stem Cell Transplantation." *Journal of Clinical Virology* 108 (November): 112–20. https://doi.org/10.1016/j.jcv.2018.09.014.

Pride, David T, Julia Salzman, Matthew Haynes, Forest Rohwer, Clara Davis-Long, Richard A White, Peter Loomer, Gary C Armitage, and David A Relman. 2012. "Evidence of a Robust Resident Bacteriophage Population Revealed through Analysis of the Human Salivary Virome." *The ISME Journal* 6 (5): 915–26. https://doi.org/10.1038/ismej.2011.169.

Prosperi, Mattia C F, and Marco Salemi. 2012. "QuRe: Software for Viral Quasispecies Reconstruction from next-Generation Sequencing Data." *Bioinformatics* 28 (1): 132–33. https://doi.org/10.1093/bioinformatics/btr627.

Qin, Junjie, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, et al. 2010a. "A Human Gut

Microbial Gene Catalogue Established by Metagenomic Sequencing." *Nature* 464 (7285): 59–65. https://doi.org/10.1038/nature08821.

———. 2010b. "A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing." *Nature* 464 (7285): 59–65. https://doi.org/10.1038/nature08821.

Qin, Junjie, Yingrui Li, Zhiming Cai, Shenghui Songgang Li, Jianfeng Zhu, Fan Zhang, Suisha Liang, et al. 2012. "A Metagenome-Wide Association Study of Gut Microbiota in Type 2 Diabetes." *Nature* 490 (7418): 55–60. https://doi.org/10.1038/nature11450.

Quick, Joshua, Nicholas J. Loman, Sophie Duraffour, Jared T. Simpson, Ettore Severi, Lauren Cowley, Joseph Akoi Bore, et al. 2016. "Real-Time, Portable Genome Sequencing for Ebola Surveillance." *Nature* 530 (7589): 228–32. https://doi.org/10.1038/nature16996.

Quince, Christopher, Anders Lanzen, Russell J Davenport, and Peter J Turnbaugh. 2011. "Removing Noise From Pyrosequenced Amplicons." *BMC Bioinformatics* 12 (1): 38. https://doi.org/10.1186/1471-2105-12-38.

Quinn, Thomas P., Ionas Erb, Greg Gloor, Cedric Notredame, Mark F. Richardson, and Tamsyn M. Crowley. 2018. "A Field Guide for the Compositional Analysis of Any-Omics Data." *BioRxiv*, December, 484766. https://doi.org/10.1101/484766.

Ravel, J., P. Gajer, Z. Abdo, G. M. Schneider, S. S. K. Koenig, S. L. McCulle, S. Karlebach, et al. 2011. "Vaginal Microbiome of Reproductive-Age Women." *Proceedings of the National Academy of Sciences* 108 (Supplement_1): 4680–87. https://doi.org/10.1073/pnas.1002611107.

Ren, Jie, Nathan A. Ahlgren, Yang Young Lu, Jed A. Fuhrman, and Fengzhu Sun. 2017. "VirFinder: A Novel k-Mer Based Tool for Identifying Viral Sequences from Assembled Metagenomic Data." *Microbiome* 5 (1): 69. https://doi.org/10.1186/s40168-017-0283-5.

Reyes, G R, and J P Kim. 1991. "Sequence-Independent, Single-Primer Amplification (SISPA) of Complex DNA Populations." *Molecular and Cellular Probes* 5 (6): 473–81. http://www.ncbi.nlm.nih.gov/pubmed/1664049.

Rho, Mina, Haixu Tang, and Yuzhen Ye. 2010. "FragGeneScan: Predicting Genes in Short and Error-Prone Reads." *Nucleic Acids Research* 38 (20).

https://doi.org/10.1093/nar/gkq747.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40. https://doi.org/10.1093/bioinformatics/btp616.

Rodriguez-R, Luis M, and Konstantinos T Konstantinidis. 2014. "Estimating Coverage in Metagenomic Data Sets and Why It Matters." *The ISME Journal* 8 (11): 2349–51. https://doi.org/10.1038/ismej.2014.76.

Rognes, Torbjørn, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. 2016. "VSEARCH: A Versatile Open Source Tool for Metagenomics." *PeerJ* 4 (October): e2584. https://doi.org/10.7717/peerj.2584.

Rose, Rebecca, Bede Constantinides, Avraam Tapinos, David L Robertson, and Mattia Prosperi. 2016. "Challenges in the Analysis of Viral Metagenomes." *Virus Evolution* 2 (2): 1–11. https://doi.org/10.1093/ve/vew022.

Rothschild, Daphna, Omer Weissbrod, Elad Barkan, Alexander Kurilshikov, Tal Korem, David Zeevi, Paul I. Costea, et al. 2018. "Environment Dominates over Host Genetics in Shaping Human Gut Microbiota." *Nature* 555 (7695): 210–15. https://doi.org/10.1038/nature25973.

Roux, Simon, Jeremy Tournayre, Antoine Mahul, Didier Debroas, and François Enault. 2014. "Metavir 2: New Tools for Viral Metagenome Comparison and Assembled Virome Analysis." *BMC Bioinformatics* 15 (1): 76. https://doi.org/10.1186/1471-2105-15-76.

Rudikoff, Donald, and Mark Lebwohl. 1998. "Atopic Dermatitis." *The Lancet* 351 (9117): 1715–21. https://doi.org/10.1016/S0140-6736(97)12082-7.

Rusch, Douglas B., Aaron L. Halpern, Granger Sutton, Karla B. Heidelberg, Shannon Williamson, Shibu Yooseph, Dongying Wu, et al. 2007. "The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific." *PLoS Biology* 5 (3): 0398–0431. https://doi.org/10.1371/journal.pbio.0050077.

Salter, Susannah J, Michael J Cox, Elena M Turek, Szymon T Calus, William O Cookson, Miriam F Moffatt, Paul Turner, et al. 2014. "Reagent and Laboratory Contamination Can Critically Impact Sequence-Based Microbiome Analyses." *BMC Biology* 12 (1):

87. https://doi.org/10.1186/s12915-014-0087-z.

Salter, Susannah J, Michael J Cox, Elena M Turek, Szymon T Calus, William O Cookson, Miriam F Moffatt, Paul Turner, Julian Parkhill, Nicholas J Loman, and Alan W Walker. 2014. "Reagent and Laboratory Contamination Can Critically Impact Sequence-Based Microbiome Analyses." *BMC Biology* 12 (1): 87. https://doi.org/10.1186/s12915-014-0087-z.

Schaeffer, L, H Pimentel, N Bray, P Melsted, and L Pachter. 2017. "Pseudoalignment for Metagenomic Read Assignment." Edited by Bonnie Berger. *Bioinformatics* 33 (14): 2082–88. https://doi.org/10.1093/bioinformatics/btx106.

Schirmer, Melanie, Rosalinda D'Amore, Umer Z. Ijaz, Neil Hall, and Christopher Quince. 2016. "Illumina Error Profiles: Resolving Fine-Scale Variation in Metagenomic Sequencing Data." *BMC Bioinformatics* 17 (1): 125. https://doi.org/10.1186/s12859-016-0976-y.

Schloissnig, Siegfried, Manimozhiyan Arumugam, Shinichi Sunagawa, Makedonka Mitreva, Julien Tap, Ana Zhu, Alison Waller, et al. 2013. "Genomic Variation Landscape of the Human Gut Microbiome." *Nature* 493 (7430): 45–50. https://doi.org/10.1038/nature11711.

Scholz, Matthias, Doyle V Ward, Edoardo Pasolli, Thomas Tolio, Moreno Zolfo, Francesco Asnicar, Duy Tin Truong, Adrian Tett, Ardythe L Morrow, and Nicola Segata. 2016. "Strain-Level Microbial Epidemiology and Population Genomics from Shotgun Metagenomics." *Nature Methods* 13 (5): 435–38. https://doi.org/10.1038/nmeth.3802.

Schommer, Nina N., and Richard L. Gallo. 2013. "Structure and Function of the Human Skin Microbiome." *Trends in Microbiology*. https://doi.org/10.1016/j.tim.2013.10.001.

Schooley, Robert T, Biswajit Biswas, Jason J Gill, Adriana Hernandez-Morales, Jacob Lancaster, Lauren Lessor, Jeremy J Barr, et al. 2017. "Development and Use of Personalized Bacteriophage-Based Therapeutic Cocktails To Treat a Patient with a Disseminated Resistant Acinetobacter Baumannii Infection." *Antimicrobial Agents and Chemotherapy* 61 (10). https://doi.org/10.1128/AAC.00954-17.

Schowalter, Rachel M, Diana V Pastrana, Katherine A Pumphrey, Adam L Moyer, and Christopher B Buck. 2010. "Merkel Cell Polyomavirus and Two Previously Unknown

Polyomaviruses Are Chronically Shed from Human Skin." *Cell Host & Microbe* 7 (6): 509–15. https://doi.org/10.1016/j.chom.2010.05.006.

Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, et al. 2017. "Critical Assessment of Metagenome Interpretation—a Benchmark of Metagenomics Software." *Nature Methods* 14 (11): 1063–71. https://doi.org/10.1038/nmeth.4458.

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

Segata, Nicola, Daniela Börnigen, Xochitl C Morgan, and Curtis Huttenhower. 2013. "PhyloPhlAn Is a New Method for Improved Phylogenetic and Taxonomic Placement of Microbes." *Nature Communications* 4 (1): 2304. https://doi.org/10.1038/ncomms3304.

Segata, Nicola, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. 2012. "Metagenomic Microbial Community Profiling Using Unique Clade-Specific Marker Genes." *Nature Methods* 9 (8): 811–14. https://doi.org/10.1038/nmeth.2066.

Serra, Angela, Michele Fratello, Vittorio Fortino, Giancarlo Raiconi, Roberto Tagliaferri, and Dario Greco. 2015. "MVDA: A Multi-View Genomic Data Integration Methodology." *BMC Bioinformatics* 16 (1): 261. https://doi.org/10.1186/s12859-015-0680-3.

Shaiber, Alon, and A Murat Eren. 2019. "Composite Metagenome-Assembled Genomes Reduce the Quality of Public Genome Repositories." *MBio* 10 (3): e00725-19. https://doi.org/10.1128/mBio.00725-19.

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

Sharifi, Fatemeh, and Yuzhen Ye. 2017. "From Gene Annotation to Function Prediction for Metagenomics." In *Methods in Molecular Biology (Clifton, N.J.)*, 1611:27–34. https://doi.org/10.1007/978-1-4939-7015-5_3.

Shin, Hakdong, Kenneth Price, Luong Albert, Jack Dodick, Lisa Park, and Maria Gloria Dominguez-Bello. 2016. "Changes in the Eye Microbiota Associated with Contact

Lens Wearing." *MBio* 7 (2): e00198. https://doi.org/10.1128/mBio.00198-16.

Sieber, Christian M. K., Alexander J. Probst, Allison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, and Jillian F. Banfield. 2018. "Recovery of Genomes from Metagenomes via a Dereplication, Aggregation and Scoring Strategy." *Nature Microbiology* 3 (7): 836–43. https://doi.org/10.1038/s41564-018-0171-1.

Silverman, Justin D., Kimberly Roche, Sayan Mukherjee, and Lawrence A. David. 2018. "Naught All Zeros in Sequence Count Data Are the Same." *BioRxiv*, November, 477794. https://doi.org/10.1101/477794.

Simpson, E. H. 1949. "Measurement of Diversity." *Nature* 163 (4148): 688–688. https://doi.org/10.1038/163688a0.

Spurgeon, Megan E, and Paul F Lambert. 2013. "Merkel Cell Polyomavirus: A Newly Discovered Human Virus with Oncogenic Potential." *Virology* 435 (1): 118–30. https://doi.org/10.1016/j.virol.2012.09.029.

Stapleton, Ann E. 2016. "The Vaginal Microbiota and Urinary Tract Infection." *Microbiology Spectrum* 4 (6). https://doi.org/10.1128/microbiolspec.UTI-0025-2016.

Statnikov, A, A V Alekseyenko, Z Li, M Henaff, G I Perez-Perez, M J Blaser, and C F Aliferis. 2013. "Microbiomic Signatures of Psoriasis: Feasibility and Methodology Comparison." *Sci Rep* 3 (September): 2620. https://doi.org/srep02620 [pii]\r10.1038/srep02620.

Steinegger, Martin, and Johannes Söding. 2017. "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets." *Nature Biotechnology* 35 (11): 1026–28. https://doi.org/10.1038/nbt.3988.

Stout, Molly J., Bridget Conlon, Michele Landeau, Iris Lee, Carolyn Bower, Qiuhong Zhao, Kimberly A. Roehl, et al. 2013. "Identification of Intracellular Bacteria in the Basal Plate of the Human Placenta in Term and Preterm Gestations." *American Journal of Obstetrics and Gynecology* 208 (3): 226.e1-226.e7. https://doi.org/10.1016/j.ajog.2013.01.018.

Sugimoto, Shinya, Takeo Iwamoto, Koji Takada, Ken-Ichi Okuda, Akiko Tajima, Tadayuki Iwase, and Yoshimitsu Mizunoe. 2013. "Staphylococcus Epidermidis Esp Degrades Specific Proteins Associated with Staphylococcus Aureus Biofilm Formation and

Host-Pathogen Interaction." *Journal of Bacteriology* 195 (8): 1645–55. https://doi.org/10.1128/JB.01672-12.

Takemoto, Akemi, Otomi Cho, Yuka Morohoshi, Takashi Sugita, and Masahiko Muto. 2015. "Molecular Characterization of the Skin Fungal Microbiome in Patients with Psoriasis." *The Journal of Dermatology* 42 (2): 166–70. https://doi.org/10.1111/1346-8138.12739.

Tang, Zheng-Zheng, Guanhua Chen, and Alexander V. Alekseyenko. 2016. "PERMANOVA-S: Association Test for Microbial Community Composition That Accommodates Confounders and Multiple Distances." *Bioinformatics* 32 (17): 2618–25. https://doi.org/10.1093/bioinformatics/btw311.

Teng, Fei, Sree Sankar Darveekaran Nair, Pengfei Zhu, Shanshan Li, Shi Huang, Xiaolan Li, Jian Xu, and Fang Yang. 2018. "Impact of DNA Extraction Method and Targeted 16S-RRNA Hypervariable Region on Oral Microbiota Profiling." *Scientific Reports* 8 (1): 16321. https://doi.org/10.1038/s41598-018-34294-x.

Tett, Adrian, Edoardo Pasolli, Stefania Farina, Duy Tin Truong, Francesco Asnicar, Moreno Zolfo, Francesco Beghini, et al. 2017. "Unexplored Diversity and Strain-Level Structure of the Skin Microbiome Associated with Psoriasis." *Npj Biofilms and Microbiomes* 3 (1): 14. https://doi.org/10.1038/s41522-017-0022-5.

Tetz, George, and Victor Tetz. 2016. "Bacteriophage Infections of Microbiota Can Lead to Leaky Gut in an Experimental Rodent Model." *Gut Pathogens*, 1–4. https://doi.org/10.1186/s13099-016-0109-1.

Tikhonov, Mikhail, Robert W Leach, and Ned S Wingreen. 2015. "Interpreting 16S Metagenomic Data without Clustering to Achieve Sub-OTU Resolution." *The ISME Journal* 9 (1): 68–80. https://doi.org/10.1038/ismej.2014.117.

Tithi, Saima Sultana, Frank O. Aylward, Roderick V. Jensen, and Liqing Zhang. 2018. "FastViromeExplorer: A Pipeline for Virus and Phage Identification and Abundance Profiling in Metagenomics Data." *PeerJ* 6 (January): e4227. https://doi.org/10.7717/peerj.4227.

Truong, Duy Tin, Eric a Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015.

"MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling." *Nature Methods* 12 (10): 902–3. https://doi.org/10.1038/nmeth.3589.

Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. 2017. "Microbial Strain-Level Population Structure and Genetic Diversity from Metagenomes." *Genome Research* 27 (4): 626–38. https://doi.org/10.1101/gr.216242.116.

Ursell, Luke K, Jessica L Metcalf, Laura Wegener Parfrey, and Rob Knight. 2012. "Defining the Human Microbiome." *Nutrition Reviews* 70 Suppl 1 (Suppl 1): S38-44. https://doi.org/10.1111/j.1753-4887.2012.00493.x.

Virgin, Herbert W. 2014. "The Virome in Mammalian Physiology and Disease." *Cell*. https://doi.org/10.1016/j.cell.2014.02.032.

Virgin, Herbert W, E John Wherry, and Rafi Ahmed. 2009. "Redefining Chronic Viral Infection." *Cell* 138 (1): 30–50. https://doi.org/10.1016/j.cell.2009.06.036.

Virtanen, Jussi Oskari, and Steve Jacobson. 2012. "Viruses and Multiple Sclerosis." *CNS & Neurological Disorders Drug Targets* 11 (5): 528–44. http://www.ncbi.nlm.nih.gov/pubmed/22583435.

Vollmers, John, Sandra Wiegand, and Anne Kristin Kaster. 2017. "Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters!" *PLoS ONE*. https://doi.org/10.1371/journal.pone.0169662.

Waldman, A., A. Gilhar, L. Duek, and Israela Berdicevsky. 2001. "Incidence of Candida in Psoriasis - A Study on the Fungal Flora of Psoriatic Patients." *Mycoses* 44 (3–4): 77–81. https://doi.org/10.1046/j.1439-0507.2001.00608.x.

Weiss, Sophie, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, et al. 2017a. "Normalization and Microbial Differential Abundance Strategies Depend upon Data Characteristics." *Microbiome* 5 (1): 27. https://doi.org/10.1186/s40168-017-0237-y.

Wiemels, J. L., B. C. Leonard, Y. Wang, M. R. Segal, S. P. Hunger, M. T. Smith, V. Crouse, X. Ma, P. A. Buffler, and S. R. Pine. 2002. "Site-Specific Translocation and Evidence of Postnatal Origin of the t(1;19) E2A-PBX1 Fusion in Childhood Acute Lymphoblastic Leukemia." *Proceedings of the National Academy of Sciences* 99 (23):

15101–6. https://doi.org/10.1073/pnas.222481199.

Williams, Michael R., and Richard L. Gallo. 2015. “The Role of the Skin Microbiome in Atopic Dermatitis.” *Current Allergy and Asthma Reports* 15 (11): 65. https://doi.org/10.1007/s11882-015-0567-4.

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

Wu, Yu-Wei, Blake A. Simmons, and Steven W. Singer. 2016. “MaxBin 2.0: An Automated Binning Algorithm to Recover Genomes from Multiple Metagenomic Datasets.” *Bioinformatics* 32 (4): 605–7. https://doi.org/10.1093/bioinformatics/btv638.

Wylie, Kristine M., Rebecca M. Truty, Thomas J. TJ Sharpton, Kathie A. Mihindukulasuriya, Yanjiao Zhou, Hongyu Gao, Erica Sodergren, George M. Weinstock, and Katherine S. Pollard. 2012. “Novel Bacterial Taxa in the Human Microbiome.” Edited by Sarah K. Highlander. *PLoS ONE* 7 (6): e35294. https://doi.org/10.1371/journal.pone.0035294.

Xu, Jianping. 2006. “Microbial Ecology in the Age of Genomics and Metagenomics: Concepts, Tools, and Recent Advances.” *Molecular Ecology* 15 (7): 1713–31. https://doi.org/10.1111/j.1365-294X.2006.02882.x.

Xu, Lizhen, Andrew D. Paterson, Williams Turpin, and Wei Xu. 2015. “Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data.” Edited by Yinglin Xia. *PLOS ONE* 10 (7): e0129606. https://doi.org/10.1371/journal.pone.0129606.

Yang, Bo, Yong Wang, and Pei-Yuan Qian. 2016. “Sensitivity and Correlation of Hypervariable Regions in 16S RRNA Genes in Phylogenetic Analysis.” *BMC Bioinformatics* 17 (1): 135. https://doi.org/10.1186/s12859-016-0992-y.

Yarza, Pablo, Pelin Yilmaz, Elmar Pruesse, Frank Oliver Glöckner, Wolfgang Ludwig, Karl-Heinz Schleifer, William B. Whitman, Jean Euzéby, Rudolf Amann, and Ramon Rosselló-Móra. 2014. “Uniting the Classification of Cultured and Uncultured Bacteria and Archaea Using 16S RRNA Gene Sequences.” *Nature Reviews Microbiology* 12

(9): 635–45. https://doi.org/10.1038/nrmicro3330.

Yilmaz, Pelin, Laura Wegener Parfrey, Pablo Yarza, Jan Gerken, Elmar Pruesse, Christian Quast, Timmy Schweer, Jörg Peplies, Wolfgang Ludwig, and Frank Oliver Glöckner. 2014. "The SILVA and 'All-Species Living Tree Project (LTP)' Taxonomic Frameworks." *Nucleic Acids Research* 42 (D1). https://doi.org/10.1093/nar/gkt1209.

Yolken, R. H., L. Jones-Brando, D. D. Dunigan, G. Kannan, F. Dickerson, E. Severance, S. Sabunciyan, et al. 2014. "Chlorovirus ATCV-1 Is Part of the Human Oropharyngeal Virome and Is Associated with Changes in Cognitive Functions in Humans and Mice." *Proceedings of the National Academy of Sciences* 111 (45): 16106–11. https://doi.org/10.1073/pnas.1418895111.

Yue, Jack C., and Murray K. Clayton. 2005. "A Similarity Measure Based on Species Proportions." *Communications in Statistics - Theory and Methods* 34 (11): 2123–31. https://doi.org/10.1080/STA-200066418.

Zhao, Yongan, Haixu Tang, and Yuzhen Ye. 2012. "RAPSearch2: A Fast and Memory-Efficient Protein Similarity Search Tool for next-Generation Sequencing Data." *Bioinformatics* 28 (1): 125–26. https://doi.org/10.1093/bioinformatics/btr595.

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

Zhu, Wenhan, Alexandre Lomsadze, and Mark Borodovsky. 2010. "Ab Initio Gene Identification in Metagenomic Sequences." *Nucleic Acids Research* 38 (12). https://doi.org/10.1093/nar/gkq275.