OXFORD UNIVERSITY PRESS Systematic Biology

Phylogenetics and the human microbiome

Journal:	*Systematic Biology*
Manuscript ID:	USYB-2013-214
Manuscript Type:	Regular Manuscript
Date Submitted by the Author:	21-Oct-2013
Complete List of Authors:	Matsen, Frederick; FHCRC, Computational Biology
Keywords:	human microbiome, microbial ecology, phylogenetic methods, 16S, metagenome

Version dated: October 21, 2013

RH: PHYLOGENETICS AND THE HUMAN MICROBIOME

Phylogenetics and the human microbiome.

FREDERICK A. MATSEN IV[1]

[1] *Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, WA, 91802, USA*

Corresponding author: Frederick A Matsen, Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, WA, 91802, USA; E-mail: matsen@fhcrc.org.

ABSTRACT

The human microbiome is the collection of microbes that live inside and on the surface of humans. Because microbial sequencing information is now much easier to come by than phenotypic information, there has been an explosion of sequencing information from microbiome samples. Much of the analytical work for these sequences involves phylogenetics, at least indirectly, but methodology has developed in a somewhat different direction than for other applications of phylogenetics. In this paper I review the field and its methods from the perspective of a phylogeneticist, as well as describing current challenges for phylogenetics coming from this type of work.

(Keywords: human microbiome; microbial ecology; phylogenetic methods; 16S; metagenome)

INTRODUCTION

The parameter regime and focus of microbiome research sits outside of the traditional setting for phylogenetics methods development and application; why should our community be interested in what microbial ecologists and medical researchers have done? The answer is simple: this system is data- and question-rich. It absolutely requires molecular methods, because microbes are now primarily identified by their molecular sequences, which is much more straightforward to do in high throughput than morphological or phenotypic characterization. Indeed, microbial ecology has recently become for the most part the study of the relative abundances of various sequences derived from the environment, even if the framework for understanding between-microbe relationships includes metabolic information and other information not derived directly from sampled molecular sequences.

Although there is something of a divide between phylogeny as practiced as part of microbial ecology on one hand and that for multicellular organisms on the other, there are many parallels between the two enterprises. Both communities struggle with issues of sequence alignment, large-scale tree reconstruction, and species definition. However, approaches differ between the microbial ecology community and that of eukaryotic phylogenetics, in part because the scope of the former contains an almost unlimited diversity of organisms, leading to additional problems above the usual. The species concept is even more problematic for microbes than for multicellular organisms, and hence there is also considerable discussion concerning how to group them into species-like units. Organizing microbes into a sensible taxonomy is a serious challenge, especially in the absence of obvious morphological features.

Because of this high level of diversity and challenges with species definitions, microbial ecology researchers have developed their own explicitly phylogenetic techniques for comparing samples rather than comparing on the level of species

abundances. Although there is some overlap with previous literature, these techniques and could be used in a wider setting and may deserve broader consideration by the phylogenetics community.

The human microbiome is specifically interesting because questions of microbial genomics, translated into questions of function, have important consequences for human health. Additionally, due to more than a century of hospital laboratory work, our knowledge about human-associated microbes is relatively rich. It is also common to manipulate the human microbiome or animal models thereof via intervention studies and germ-free animals, which is easier to do than for many other microbial communities.

In this review I will describe phylogenetics-related research happening in microbial ecology and contrast approaches between microbial researchers and what I think of as the typical *Systematic Biology* audience. Despite an obvious oversimplification, I will use *eukaryotic phylogenetics* to indicate what I think of as the mainstream of SB readership, and *microbial phylogenetics* to denote the other. I realize that there is substantial overlap— for instance the microbial community is very interested in unicellular fungi, and additionally many in the SB community do work on microbes— but this terminology will be useful for concreteness. There is of course also substantial overlap in methodology, however as we will see there are significant differences in approach and the two areas have developed somewhat in parallel. I will first briefly review the recent literature on the human microbiome, then describe novel ways in which human microbiome researchers have used trees. I will finish with opportunities for the *Systematic Biology* audience to contribute to this field. I have made an explicit effort to make a neutral comparison between two directions rather than criticize the approximate methods common in microbial phylogenetics, indeed, microbial phylogenetics requires algorithms and ideas that work in parameter regimes an order of magnitude larger than typical for eukaryotic phylogenetics.

THE HUMAN MICROBIOME

The human microbiome is the collection of microbial organisms that live inside of and on the surface of humans. These organisms are populous: it has been estimated that there are ten times as many bacteria associated with each individual than there are human cells of that individual. The microbiome has remarkable metabolic potential, with an ensemble of genes estimated to be about 150 times larger than the human collection of genes (Qin et al. 2010). Much of our metabolic interaction with the outside world is mediated by our microbiome, as it has important roles in immune system development, nutrition, and drug metabolism (Kau et al. 2011; Maurice et al. 2013); our food and drug intake in turn impacts the diversity of microbes present. In this section I will briefly review what is known about the human microbiome and its effect on our health.

The human microbiome is an ecosystem. It is dynamic in terms of taxonomic representation but apparently constant in terms of function (Consortium 2012). There is a "core" microbiome which is shared between all humans (Turnbaugh et al. 2008). The human microbiome is spatially organized, as can be seen on skin (Grice et al. 2009), with substantial variation in human body habitats across space and time (Costello et al. 2009). There is a substantial range of inter-individual versus intra-individual variation (Consortium 2012).

Our actions can shift the composition of our microbiome. Changes in diet can very quickly shift its composition, but there is also a strong correlation between long-term diet and microbiome (Li et al. 2009; Wu et al. 2011). Antibiotics fundamentally disturb microbial communities, resulting in an effect that lasts for years (Jernberg et al. 2007; Dethlefsen et al. 2008; Jakobsson et al. 2010; Dethlefsen and Relman 2011).

The microbiome interacts on many levels with host phenotype (reviewed in Cho and Blaser 2012). The gut microbiome in particular correlates with health of individuals from the elderly in industrialized nations (Claesson et al. 2012) to children with acute metabolic dysfunction in rural Africa (Smith et al. 2013). Considerable attention has also been given to the interaction between the gut microbiome and obesity, although the story is not yet clear. An intervention study has

established human gut microbes associated with obesity (Ley et al. 2006). A causal role for the microbiome leading to obesity is established for mice: an obese phenotype can be transferred from mouse to mouse by gut microbiome transplantation (Turnbaugh et al. 2006), the pregnant human gut microbiome leads to obesity in mice (Koren et al. 2012), and probiotics can lead to a lean phenotype and healthy eating behavior (Poutahidis et al. 2013). However, these promising leads have not yet been confirmed causally or in population studies of humans (Zhao 2013). For example, a study of obesity in the old-order Amish did not find any correlation between obesity and particular gut communities (Zupancic et al. 2012).

Bacteria have been the primary focus of human microbiome research, and other domains have been investigated though to a lesser extent. Changes in archaeal and fungal populations have been shown to covary with bacterial residents (Hoffmann et al. 2013). Viral populations have been observed to be highly dynamic and variable across individuals (Reyes et al. 2010; Minot et al. 2011, 2013). We will focus on bacteria here.

In this paper we will be primarily be describing the human microbiome from a community-level phylogenetic perspective rather than from the fine-scale perspective of immune-mediated interactions between host and microbe (reviewed in Hooper et al. 2012). Our understanding of the true effect of the microbiome will eventually come from such a molecular-level understanding, although until we can characterize all of the molecular interactions between microbes and the human body, a broad perspective will continue to be important.

INVESTIGATING THE HUMAN MICROBIOME VIA SEQUENCING

It is now possible to assay microbial communities in high throughput using sequencing. One way is to amplify a specific gene in the genome for sequencing using polymerase chain reaction (PCR). Scientists typically pick a "marker" gene in that case that is meant to recapitulate the "overall" evolutionary history of the microbes. Another way is to randomly shear input DNA and/or RNA and then perform sequencing directly. We will consistently refer to the former as a *survey*

and the second a *metagenome*, although these words have not always been consistently used in the literature.

The Human Microbiome Project (Methé et al. 2012) generated lots of survey, metagenome, and whole-genome sequencing data and this data is available on a dedicated website[1]. The MetaHIT study (Qin et al. 2010) also generated lots of data but it is not available to outside researchers.

Microbial community estimation using marker gene surveys

Our modern knowledge of the microbial world is in a large part derived from the methods of Carl Woese and colleagues who pioneered the use of marker genes as a way to distinguish between microbial lineages (Fox et al. 1977). Their work, and the scientists who followed them, focused on the 16S ribosomal gene (henceforth simply "16S") as a genetic marker. This gene was chosen because it has regions of high and low diversity, which enable resolution on a variety of evolutionary time scales. Regions of low diversity in 16S also enabled the development of the first "universal" 16S PCR primers (Lane et al. 1985) which enabled surveys of almost all bacteria regardless of whether they can be cultured.

Where Woese and colleagues labored over digestion and gel electrophoresis to infer sequences, modern researchers have the luxury of high throughput sequencing. This can be done with a high level of multiplexing, making an explicit tradeoff between depth of sequencing for each specimen and the number of specimens able to be put on the sequencer at the same time. This has led to extensive parallelization, most recently by sequencing dozens of samples at a time on the Illumina instrument (Degnan and Ochman 2011; Caporaso et al. 2012). This brings up the question of how many sequences are needed to characterize the microbial diversity of a given environment. To distinguish between two rather different samples, relatively few sequences per sample are required (Kuczynski et al. 2010) however for more subtle information deeper sequencing is required. In addition to

[1] http://www.hmpdacc.org/

sequencing samples across individuals, this parallelization has also enabled sampling through time (e.g. Caporaso et al. 2011).

Despite the high throughput and low cost of modern sequencing, inherent challenges remain for applications of marker gene sequencing to take a census of microbes. Most fundamentally, various microbes have different DNA extraction efficiencies, even with stringent protocols, meaning that a collection of sequences need not be representative of the communities from which they were derived (Morgan et al. 2010). Current high throughput sequencing technology is limited to a length that is shorter than most genes, which limits the resolution of the analyses. "Primer bias," or differing amplification levels of various sequences based on their affinity for the primers (Suzuki and Giovannoni 1996; Polz and Cavanaugh 1998), is a challenge and has led to the standardization of primers (Methé et al. 2012). Worse, multiplex PCR is known to create chimeric (i.e. spurious recombinant) sequences via partial PCR products (Hugenholtz and Huber 2003; Ashelford et al. 2005; Haas et al. 2011; Schloss et al. 2011). Correspondingly, chimera checking software has been developed (including Ashelford et al. 2006; Edgar et al. 2011). Also, 16S can be present in up to 15 copies and there can be diversity within the copies (Klappenbach et al. 2001). Recent work by Kembel et al. (2012) implements the independent contrasts (Felsenstein 1985) method to correct for copy number, which has been helpful despite a moderate evolutionary signal in copy number variation (Klappenbach et al. 2000). Some groups have reported advantages to using alternate single-copy genes as markers for characterization of microbial communities (e.g. Case et al. 2007; McNabb et al. 2004), however 16S remains the dominant locus used by a large margin. A final cause of noise is next-generation sequencing error: this is certainly a problem for both surveys and metagenomes, but is becoming less of a problem as technology improves. I will not address it specifically except in the inference of operational taxonomic units as described below.

Metagenomes

As described above "metagenome" means that data is sheared randomly across the genome rather than amplified from a specific location, and thus the genetic region of a read is unknown in addition to what organism it was derived from. Because metagenomes do not proceed through an amplification step, it does not have the same PCR primer biases as a marker gene survey, however extraction efficiency concerns remain and multiplex sequencing is known to have biases of its own.

It is possible to use metagenomic data as an expanded set of marker genes. That is, one can use 16S reads that appear in the metagenome as well as reads from other "core" genes that are expected to follow the same evolutionary path and are present in a large proportion of micro-organisms. This is proven to be a useful strategy (Von Mering et al. 2007a; Wu and Eisen 2008a; Stark et al. 2010; Kembel et al. 2011), however, because of the diversity of gene repertoire in microbes, this core gene set may be relatively small. Indeed, even the largest collection of genes in these databases only recruits around 1 percent of a metagenome. At least some portion of the rest of the data can be taxonomically classified (methods reviewed by Mande et al. 2012); Treangen et al. (2013) report speedups and much higher accuracy when reads are assembled before they are classified.

Metagenomic data is often used to infer information about metabolism rather than phylogenetic nature (Greenblum et al. 2012; Abubucker et al. 2012). Discussing these methods is beyond the scope of this paper, as is "metatranscriptomics": the sequencing of mRNA in bulk.

Whole genomes

Whole-genome sequencing from culture is currently being used for microbial outbreak tracking (Köser et al. 2012; Snitkin et al. 2012). The Food and Drug Administration maintains GenomeTrakr, an openly accessible database of whole genomes sampled from the environment and grown in culture[2]. This data may become common for unculturable organisms as single-cell sequencing methods improve

[2] http://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/

(reviewed in Kalisky and Quake 2011). The assembly of complete genomes from metagenomes, once limited to samples with a very small number of organisms (Baker et al. 2010), is now becoming feasible for more diverse populations with improved sequencing technology and computational approaches (Howe et al. 2012; Pell et al. 2012; Iverson et al. 2012; Emerson et al. 2012; Podell et al. 2013).

TREE-THINKING IN HUMAN MICROBIOME RESEARCH

In this section I consider the ways in which phylogenetic methodology has impacted human microbiome research. What may be most interesting for the *Systematic Biology* audience is the way in which phylogenetic trees are being used to actively revise taxonomy as well as being used as a structure on which to perform sample comparison.

Phylogenetics and taxonomy

Phylogenetic inference has had a substantial impact on microbial ecology research by changing our view of the taxonomic relationships between microorganisms. The clearest such example is the discovery that archaea, although morphologically similar to bacteria, form their own separate lineage (Woese and Fox 1977).

Several groups are continually revising taxonomy using the results of phylogenetic tree inference. These attempts are less ambitious than the PhyloCode project (to develop a taxonomic scheme expressed directly in terms of a phylogeny; see Forey 2001), and simply work to revise the hierarchical structure of the taxonomy while for the most part leaving taxonomic names fixed. Bergey's Manual of Systematic Bacteriology has officially adopted 16S as the basis for their taxonomy, although the actual revision process appears opaque (Kreig et al. 1984). The Green-Genes group (DeSantis et al. 2006a) has been very active in updating their taxonomy according to 16S, first with their GRUNT tool (Dalevi et al. 2007) and more recently with their tax2tree tool (McDonald et al. 2011). Tax2tree uses a heuristic algorithm to optimize the F-measure of precision and recall for its taxonomic assignments. Interestingly, tax2tree allows for polyphyletic taxonomic groups as

allowance for either phylogenetic error, lack of resolution of the 16S gene or taxonomic groups with no evolutionary basis; the method does not attempt to signal the cause of such polyphyletic groups. Matsen and Gallagher (2012) developed algorithms to quantify discordance between phylogeny and taxonomy based on a coloring problem previously described in the computer science literature (Moran and Snir 2008). Although it is wonderful that several groups are actively working on taxonomic revision, it can be frustrating to have multiple different taxonomies with no easy way to translate between them or to the taxonomic names provided in the NCBI sequence database.

An obvious application of phylogenetics is to perform taxonomic classification, as the taxonomy is at least in part defined by phylogeny. However, comparisons of taxonomic classification programs (Liu et al. 2008; Bazinet and Cummings 2012) have indicated that current implementations of phylogenetic methods do not perform as well as simple classifiers based on k-mer composition (Wang et al. 2007; Rosen et al. 2008). Srinivasan et al. (2012) find that phylogenetic methods to do taxonomic classification can outperform composition-based techniques at least for certain taxonomic groups. Some authors report that a combination of composition-based and homology-based classifiers work best (Brady and Salzberg 2009; Parks et al. 2011). The MEGAN program (Huson et al. 2007, 2011) BLASTs an unknown sequence onto a database of sequences with taxonomic labels and assigns the sequence the lowest (i.e. narrowest) taxonomic group shared by all of the high-quality hits. Munch et al. (2008a,b) infer taxonomic assignment by automatically retrieving sequences equipped with taxonomic information and building a tree on them along with an unknown sequence. Segata et al. (2012) propose a clever approach to inferring organisms present in a metagenomic sample by compiling a database of clade-specific genes, then classifying a given read as being from the only clade that has the corresponding gene. They show that this has good sensitivity and specificity, however, this method can only be used to identify the presence of organisms whose genome has been sequenced.

The role of Operational Taxonomic Units (OTUs)

Although there continues to be a lively debate on if there is a meaningful concept of species for microbes (Bapteste et al. 2009; Caro-Quintero and Konstantinidis 2012), a substantial part of human microbiome research has replaced any traditional species concept with the notion of Operational Taxonomic Units (OTUs). An OTU is a proxy species concept that is typically defined with a fixed divergence cutoff, most commonly at 97% sequence identity, such that each OTU is a cluster of sequences that are closer to each other than that cutoff. It is common for trees to be built on sequence representatives from these OTUs, and the abundance of an OTU to be given by the number of sequences that sit within that cluster. I briefly describe the mini-industry of OTU clustering techniques to contrast with the phylogenetic literature on species delimitation (Pons et al. 2006; Yang and Rannala 2010). I will use the term *OTU inference* despite the fact that there is no clear definition of the OTU concept.

There are many OTU inference methods with various speeds and strategies. Of the fixed-cutoff methods, the traditional choice has been CD-HIT (Li and Godzik 2006), which seems to have been supplanted by clustering features of USEARCH (Edgar 2010) and perhaps now its descendant UPARSE (Edgar 2013). These methods are very fast and are heuristic in that they are described as an algorithm rather than as global optimization of some notion of goodness of clustering. White et al. (2010) show that different ways of doing this heuristic clustering can result in very different results. Navlakha et al. (2009) develop methods that try to come up with sequence groupings that more closely reflect the type of species divisions found in taxonomies. Hao et al. (2011) use a Gaussian mixture model formulation for clustering to avoid fixed cutoff values.

The centrality of the OTU concept can be seen by the fact that the contingency table of OTU observations is considered to be the fundamental data type for 16S studies (McDonald et al. 2012), or that methods have been devised to find OTUs from non-overlapping sequences (Sharpton et al. 2011). A significant amount of effort has been made to distinguish sequencing error in environmental samples from true rare variants; much of this work has played out in the OTU inference

literature (Quince et al. 2009, 2011; Bragg et al. 2012; Edgar 2013) as such errors are especially problematic there. With the exception of Sharpton et al. (2011), OTU inference is not considered to be a phylogenetic problem but rather something to be performed before phylogenetic inference begins.

Diversity estimates using phylogenetics

Because 16S surveys are inherently complex and noisy data, summary statistics are often used; summaries of the diversity of a single sample are often called *alpha diversity*. For the most part, this literature adapts methods from the classical ecological literature by substituting OTUs for taxonomic groups. However, phylogenetic diversity metrics are also used and here we will focus on their applications and methods.

Despite the enthusiasm with which microbial ecologists have accepted between-community comparison using phylogenetics (next section), phylogenetic alpha diversity seems under-developed. Phylogenetic diversity (PD) measures use the structure and branch lengths of a phylogenetic tree to measure the diversity of a sample (Fig. 1). The original (unweighted) PD of a set of taxa is simply the total of the lengths of the branches connecting taxa in the set Faith (1992). It quantifies the "amount of evolution" contained in the evolutionary history of those taxa. Whereas just about every 16S survey investigation involves an OTU-based alpha diversity estimate, only a few involve PD: unweighted PD has been applied to some 16S survey data (Lozupone and Knight 2007; Costello et al. 2009) and to metagenomic reads in a set of marker genes (Kembel et al. 2011).

Although abundance weighted non-phylogenetic diversity measures such as Simpson (1949) and Shannon (1948) are among the most common, abundance weighted phylogenetic diversity measures are not used in human microbiome studies. Abundance-weighted measures take a sum of branch lengths weighted by abundance, such that branches that connect abundant taxa get a higher weight than ones that do not (Fig. 1). Thus rare taxa and artifactual sequences are down-weighted compared to abundant taxa. Such measures do exist (Rao 1982; Barker

2002; Allen et al. 2009; Chao et al. 2010; Vellend et al. 2011). McCoy and Matsen (2013) have recently shown that partially abundance weighted diversity measures do a good job of distinguishing between dysbiotic and "normal" states of the human microbiome; in particular that they do a better job than the commonly-used OTU-based measures. Nipperess and Matsen (2013) have also determined formulas for the expectation and variance of PD under random sub-sampling, which is often applied to enable comparison between samples of different sequencing depths.

Community comparison using phylogenetics

The level of similarity between samples or groups of samples is called *beta diversity*. As with alpha diversity, it is not uncommon to use classical measures (e.g. Jaccard 1908) applied to OTU counts, however phylogenetics-based methods are the most popular. They are generally variants of the "UniFrac" phylogenetic dissimilarity measure (as described and named by Lozupone and Knight 2005). Kuczynski et al. (2010) claim that the UniFrac framework is superior to other methods for community comparison via real data and simulations (for a contrary viewpoint using simulations see Schloss 2008).

The name UniFrac is a contraction of "Unique Fraction," which refers to the fact that the original UniFrac definition compares the fraction of edges that connect only tips from one sample via the shortest path rather than edges that connect between samples (Fig. 2; Lozupone and Knight 2005). Weighted UniFrac is an abundance weighted version (Lozupone et al. 2007). These dissimilarity measures have hundreds of citations. Evans and Matsen (2012) showed that weighted UniFrac is in fact a special case of the earth-movers distance, is special in that it can be calculated in linear time, and that the commonly-used randomization procedure to attach a p-value to an observed distance has a central limit theorem approximation as a Gaussian process. The earth-movers distance in this case can be defined as the minimum amount of work required to move piles of dirt in one

configuration to another along the tree; in this case the size of the dirt piles is proportional to the number of reads mapping to that location in the tree (Fig. 3). Chen et al. (2012) have shown that a partially abundance-weighted variant of UniFrac may have greater power to resolve community differences than either unweighted or weighted UniFrac.

The most common way to use a between-sample pairwise distance matrix found from an application of UniFrac is to apply an ordination method such as principal coordinates analysis. Indeed, the separation of two communities in a principal components plot is often used as *prima facie* evidence of a difference between them (e.g. Lozupone and Knight 2007; Costello et al. 2009; Yatsunenko et al. 2012), while the lack of such a difference is interpreted as showing that the communities are not different overall.

There have been efforts to augment these ordination visualizations with additional information giving more structure to the visualizations. Biplots display variables (in the microbial case summarized by taxonomic labels) as points along with points representing samples (e.g. Hewitt et al. 2013; Lozupone et al. 2013). Purdom (2008) describes how generalized principal component eigenvectors can be interpreted via weightings on the leaves of a phylogenetic tree. Matsen and Evans (2013) have developed a variant of principal components analysis that explicitly labels the axes with weightings on phylogenetic trees that indicate their influence.

In another vein, La Rosa et al. (2012b) consider the induced taxonomic tree of a sample as a statistical object and, using a framework where a sampling probability is defined in terms of a distance between such induced trees, define and investigate maximum likelihood estimation of and likelihood ratio tests for these trees. They focus on distances between trees induced by matrix metrics on the corresponding adjacency matrices. A similar framework was used by Steel and Rodrigo (2008) to construct maximum likelihood supertrees.

Phylogeny and function

16S distance is frequently used as a proxy for a functional comparison between human microbiome samples. Those accustomed to microbial genetics may think this surprising, because the genetic repertoire of microbes is commonly acquired horizontally as well as vertically, and horizontal transmission leaves no trace in 16S ancestry.

However, Zaneveld et al. (2010) have shown that organisms that are more distant in terms of 16S are also more divergent in terms of gene repertoire. Such observations surround a fit nonlinear curve, and the extent to which they lay on the curve appears to be phylum-dependent. This "proxy" approach has recently been taken to its logical conclusion by Langille et al. (2013), who develop methods to infer functional characteristics from a 16S sample using discrete trait evolution models on 16S gene trees by either parsimony (Kluge and Farris 1969) or likelihood (Pagel 1994) methods via the ape package (Paradis et al. 2004).

Similar logic has been applied to prioritize microbes for sequencing. Wu et al. (2009) have derived a "phylogeny-driven genomic encyclopaedia of Bacteria and Archaea" by selecting organisms for sequencing that are divergent from sequenced organisms. By doing so they have recovered more novel protein families than they would have using methods organized by taxonomy. In a similar effort for the human microbiome (Fodor et al. 2012), phylogenetic results were not shown although the authors state that phylogenetic methods did give similar results to their analysis.

Genome-scale inquiries using phylogenetics

With some notable exceptions, mainstream applications of phylogenetics to a collection of human-associated microbes have typically been with the idea of finding "the" tree of such a collection rather than explicitly exploring divergence between various gene trees. As described above, whole-genome data is typically used to directly infer functional information rather than information concerning ancestry. The continuing debate concerning whether a microbial tree of life is a useful concept (Bapteste et al. 2009; Caro-Quintero and Konstantinidis 2012) does not seem

to have dampened human microbiome researchers' enthusiasm for using a single such tree.

Nevertheless, the work that has been done to infer horizontal gene transfer in the human microbiome has revealed interesting results. Hehemann et al. (2010) found that a seaweed gene has been transferred into a bacterium in the gut microbiome of Japanese such that individuals with this resulting microbiome are better able to digest the algae in their diet. Following on this work, Smillie et al. (2011) found that the human microbiome is in fact a common location for gene transfer. Stecher et al. (2012) find that in a mouse model, horizontal transfer between pathogenic bacteria is blocked by commensal bacteria except for periods of gut inflammation.

PHYLOGENETIC INFERENCE AS PRACTICED BY HUMAN MICROBIOME RESEARCHERS

Alignment and tree inference

In general, human microbiome researchers are interested in quickly doing phylogenetic inference on large data sets, and are less interested in clade-level accuracy or measures of uncertainty. This is defended by saying that for applications such as UniFrac, the tree is used as a framework to structure the data, and there is a certain amount of flexibility in that framework that will give the same results. Furthermore, given that the underlying data is typically 16S alone we can expect some topological inaccuracy in reconstructing the "tree of cells" even with the best methods. Additionally, as specified below, these data sets can be very large. There does not seem to be contentious discussion of specific features of the inferred trees equivalent to, say, the current discussion around the rooting of the placental mammal tree. Given this perspective, it is not surprising that Bayesian phylogenetic methods and methods that incorporate alignment uncertainty are absent.

Alignment methods are primarily focused on developing automated methods to extend a relatively small hand-curated "seed alignment" with additional sequences; several tools have been created with exactly this application for 16S in mind (DeSantis et al. 2006b; Caporaso et al. 2010a; Pruesse et al. 2012). The community also uses profile hidden Markov models (Eddy 1998) and CM models (Nawrocki et al. 2009; Nawrocki 2009) to achieve the same result.

The large data sets associated with human microbiome analysis require highly efficient algorithms for *de novo* tree inference. Historically this has meant relaxed neighbor joining (Evans et al. 2006), but more recently FastTree 2 (Price et al. 2010) has emerged as the *de facto* standard. People do most phylogenetic inferences as part of a pipeline such as mothur (Schloss et al. 2009) which has ported in the clearcut program (Sheneman et al. 2006), or QIIME (Caporaso et al. 2010b), which wraps clearcut and FastTree.

The scale of the data has motivated strategies other than complete phylogenetic inference, such as the insertion of sequences into an existing phylogenetic tree. Although such insertion has a long history as a means to sequentially build a phylogenetic tree (Kluge and Farris 1969), the first software with insertion specifically as a goal was the parsimony insertion tool in the ARB program by Ludwig et al. (2004). ARB is commonly used to reconstruct a full tree by direct insertion.

There are also other methods with the less ambitious goal of mapping sequences of unknown origin into a so-called fixed *reference tree*, sometimes with uncertainty estimates. These programs (Wu and Eisen 2008b; Monier et al. 2008; Von Mering et al. 2007b; Stark et al. 2010; Matsen et al. 2010; Berger et al. 2011) have various speeds and features. This work has also spurred development of specialized alignment tools for this mapping process. Berger and Stamatakis (2011) focus on the problem of inferring the optimal alignment and insertion of sequences into a tree. Mirarab et al. (2012) use data set partitioning to improve alignments on subsets of taxa specifically for this application.

Considerable effort goes to the creation of large curated alignments and phylogenetic trees on 16S. There are two major projects to do so: one is the SILVA

database (Pruesse et al. 2007; Quast et al. 2013), and the other is the GreenGenes database (DeSantis et al. 2006a; McDonald et al. 2011). Because of the high rate of insertion and deletion of nucleotides in 16S, these alignments have a high percentage of gap. Taking the length of 16S to be 1543 nucleotides, the 479,726 sequence SILVA reference alignment version 115 is over 96% gap, while the 1,262,986 sequence GreenGenes 13_5 alignment has is almost 80% gap. The SILVA-associated 'all-species living tree' project (Yarza et al. 2008) started with a tree inferred by maximum likelihood and has been continually updated by inserting sequences via parsimony. The GreenGenes tree is updated by running FastTree from scratch for every release. There appears to be a commonly held belief that FastTree in particular works well even with such gappy alignments (e.g. Sharpton et al. 2011).

In addition to these 16S-based resources, the MicrobesOnline resource (Dehal et al. 2010) offers a very nice interactive tree-based genome browser. On a much smaller scale, there are microbiome body-site specific reference sets (Chen et al. 2010; Griffen et al. 2011; Srinivasan et al. 2012)

PHYLOGENETIC CHALLENGES AND OPPORTUNITIES IN HUMAN MICROBIOME RESEARCH

Many phylogenetic challenges remain in human microbiome research. Some of them are familiar, such as how to build phylogenies on the scale needed here on data that has many insertions and deletions. I review some others here.

One clear challenge is to fill the gap between on one hand complete *de novo* tree inference versus sequence insertion or placement that leaves the "reference tree" fixed. These two extremes are each useful in different parameter regimes, but something between the two would be helpful as well. For example, sequence data is continually being added to large databases, motivating methods that could continually update trees with this new sequence data while allowing the previous tree to change according to this new information.

In this review I have devoted considerable space to the ways in which microbial ecologists have used the 16S tree as a proxy structure for the complete evolutionary history of their favorite organisms. They have even shown that 16S distance recapitulates gene content divergence and used this correlation to predict gene functions. It is well known, however, that any single tree will not give a complete representation of the evolutionary history of a collection of microbes.

The apparent success of 16S-tree-based comparisons begs the question of if a more complete representation of the evolutionary history of the microbes would yield better comparisons. This suggests a practical perspective on the theoretical issue of the tree of life: what is the representation of the genetic ancestry of a set of microbes that allows us to best perform proxy whole-genome comparison? This representation could be simple. For example, one of the results of Zaneveld et al. (2010) is that 16S correlates better with gene repertoire in some taxonomic groups than others. If we were to equip the 16S tree with some measure of the strength of that correlation, would that allow for more precise comparison? If we allow an arbitrary "hidden" object, what such object would perform best? For example, collections of reconciled gene trees in the presence of gene deletion, transfer, and loss (see Szöllősi et al. 2013a,b, for interesting recent results) could be used.

It appears that neutral models involving phylogenetics could be more fully developed. Methods explicitly invoking trait evolution are notably absent, with the recent exception of (Langille et al. 2013). The results of this simple method are reasonable, but would a collection of gene trees reconciled with a species tree allow for better prediction? Perhaps improved methods, say involving whole-genome evolutionary modeling or models of metabolic network evolution, could shed light on the problem. Here again the "tree of life" problem can be formulated in a practical light: what representation allows for the best prediction of features of underlying genomes? How to formulate a useful notion of independent contrasts (Felsenstein 1985) on such an object? It is quite possible that inference using a more complete representation would not be able to overcome the inherent noise of the

data, but further exploration seems warranted as simple methods give reasonable results.

An important project for microbial ecology is to model community assembly, and perhaps further phylogeny-aware methods could be used. One way to model community assembly is to apply Hubbell's neutral theory (Fierer et al. 2012; Costello et al. 2012). O'Dwyer et al. (2012) model community assembly with an explicitly phylogenetic perspective, and include some comparison of models to data. Continued work in this direction seems warranted, given the way in which phylogenetic tree shape statistics have had a significant impact on macroevolutionary modeling (Mooers and Heard 1997; Aldous et al. 2011). In this case various (alpha and beta) diversity statistics would play the role of tree shape statistics by reducing a distribution on the tips of a tree down to a real number. Another challenge is to bring together macroevolutionary modeling with species abundance modeling, but some initial steps have been made in another setting (Lambert and Steel 2013).

Diversity preservation is of interest for microbiome researchers like it is for eukaryotic organisms, but has not received the formalization and algorithmic treatment surrounding phylogenetic diversity for larger organisms (Hartmann and Steel 2006; Pardi and Goldman 2007). Martin Blaser in particular has argued that changes in our microbiomes are leading to an increase in autoimmune disease and certain types of cancer (reviewed in Cho and Blaser 2012) and has made passionate appeals to preserve microbiome diversity (Blaser 2011). Because a child's initial microbiome is transmitted from the mother (reviewed in Funkhouser and Bordenstein 2013), there is a somewhat equivalent notion of microbiome extinction when the chain is interrupted. Consequently, Yatsunenko et al. (2012) have explicitly contrasted microbiome development in urban, forest-dwelling, and rural populations, while Tito et al. (2012) have endeavored to characterize the microbiome from ancient feces. How might phylogenetic methods be used in these preservation efforts?

Perhaps in part because of the importance of mother-to-child transmission, there are indications of coevolution between microbiomes and their hosts. Ochman et al.

(2010) found identical tree topologies for primate and microbiome evolution. For the microbiome, they used maximum parsimony such that each column was a microbe and each such entry took discrete states according to how much of that microbe was present. Although parsimony gave an interesting answer here, the presence of such coevolution raises the question of what sort of forward-time models are appropriate for microbiome change? Would methods using these models do better than parsimony or commonly-applied phenetic methods applied to the distances described above? Some studies (e.g. Phillips et al. 2012; Delsuc et al. 2013) see a combination of historical and dietary influences. How can such forces be compared in this setting?

The approach of considering a collection of genes and their metabolic network as a meta-organism has yielded some interesting results (Borenstein et al. 2008; Greenblum et al. 2012). A clear limitation to this approach is that cellular boundaries are ignored: populations are not a freely diffusing soup. Could these approaches be improved by using phylogenetic methods to reconstruct the compartmentalization of genomes and processes into cells?

As described above, Morgan et al. (2010) showed that various microbes have different DNA extraction efficiencies, meaning that the representation of marker gene sequences is not representative of the actual communities. Furthermore, there was no clear taxonomic signal in their observations of the variability of extraction efficiency, which seems to preclude a correction strategy based on "species-tree" phylogenetic modeling. However, presumably *something* about their genome is determining extraction efficiency; it would be interesting and useful to search for the genetic determinants. As described above, abundances are commonly used as part of community comparison, thus a better quantification of error in those observations of abundance would be a great help.

In a similar vein, assessing the significance level of an observed difference between communities is not straightforward and poses difficult problems. The randomization of group membership commonly used in combination with UniFrac to determine significance does not have appropriate properties when in the regime

of incomplete sampling with non-independent observations, which is certainly the correct regime for surveys and metagenomes. Such non-independence can lead to incorrect rejection of the null hypothesis. Imagine, for example, that we have a random observation process on the tree equipped with some collection of "base observations" at the leaves. Each sample from the process takes a random subset of those base observations and then throws down some number of reads for each observation in that set, the number of which has mean significantly greater than 1. If the set of base observations is large compared to the number of sample observations, then two draws will always appear significantly different even though they are from the same underlying process. This would be a false positive. Even basic definitions pose a challenge here: the question of whether two communities are the "same" and "different" probably needs to be approached from the perspective of ecosystem modeling.

In general, read count normalization has not received nearly the attention that it has in other applications of high throughput sequencing (such as RNA-Seq, e.g. Anders and Huber 2010; Robinson et al. 2010). One type of normalization handles differential depths of sequencing across samples. The presently used approach is *rarefaction*, which means uniform sub-sampling, typically without replacement to the number of reads in the lowest abundance sample (Caporaso et al. 2010b; Schloss et al. 2009). In addition to throwing away data, this normalization implicitly assumes a model whereby reads are sequenced independently of one another, which is not the case. An alternative is provided by O'Dwyer et al. (2012), who provide a "UniFrac score normalization curve" based on a sampling model of community assembly. This is a good start, but more work should be done exploring results under deviation from that model.

Another type of normalization seeks to infer the true abundances from noisy observations of the various taxonomic groups or OTUs. Holmes et al. (2012) and La Rosa et al. (2012a) use models where read counts are modeled as overdispersed samples of the true abundance and provide methods for statistical testing. Paulson et al. (2013) estimate true abundances using a zero-inflated Gaussian model for

read counts. This work could be extended to a phylogenetic context by making use of the relationship between OTUs, and modeling the way in which the abundance of one OTU may increase the abundance of a related OTU because of sequencing error or a change of condition that changes the abundance of both.

Finally, the conventional wisdom that UniFrac analysis is robust to tree reconstruction methodology begs further exploration. Would it be possible to infer an equivalence class of phylogenetic trees, where two trees are deemed equivalent if they induce the same principal coordinates projection given the same underlying presence/absence or count data? Given that a tree is an integral part of a UniFrac analysis, it would be interesting to be able to infer the features of a tree that determine the primary trends in a projection.

Discussion

What can we expect next at the intersection of phylogenetics and the human microbiome? At least for the next several years we can expect more related work. Future will continue to bring deeper sequencing on more samples. The uBiome project[3] promises to bring gut microbiome sequencing to the average citizen for a low price (Costandi 2013). Comparative studies will continue to investigate what shapes and is shaped by the microbiome. However, some of the initial excitement may have died down, as neither the Human Microbiome Project nor the MetaHIT project were extended.

There are limitations to what we can learn using genetics because more intricate processes such as gene regulation may be at play, limiting what sequence-level phylogenetics can do. Future work may move from general ecological models to models that include specific interactions between microbes and the host (reviewed in Hooper et al. 2012).

Opportunities for clinical applications may present themselves, but only some of these will be sequencing based. For example, routine 16S sequencing is likely to be replaced soon by Matrix-Assisted Laser Desorption Ionization–Time of Flight

[3] http://ubiome.com/

(MALDI-TOF) mass spectrometry for assignment of a single microbe grown in culture to a database entry (Clark et al. 2013). However, for diagnoses that are on the level of microbial communities, sequencing and consequent analysis methods will still be required (reviewed in Rogers et al. 2013). Inexpensive whole-genome sequencing will certainly have a profound impact on clinical practice and epidemiological studies (Didelot et al. 2012). For all of these measures, it will be important to have rigorous means of quantifying uncertainty for robust diagnostic applications.

Human microbiome research has experienced a frenetic rate of expansion over the past decade, and sometimes the hype has outmatched the science. However, our microbes are here to stay and so is research on them. Thus we can look forward to the field of human microbiome analysis settling down to a comfortable and mature middle age as an interesting intersection between ecology and medicine. Phylogenetics has already contributed significantly to human microbiome research and will continue to do so.

ACKNOWLEDGEMENTS

I thank Olivier Gascuel for the opportunity to present on this subject during the 2013 LIRMM Mathematical and Computational Evolutionary Biology workshop and for organizing a corresponding special section of *Systematic Biology*. I am grateful to Aaron Darling, David Fredricks, Noah Hoffman, Steven Kembel, Connor McCoy, Martin Morgan, and Sujatha Srinivasan for interesting discussions that informed this review, thank Bastien Boussau, Noah Hoffman, Christopher Small and Björn Winckler for providing feedback on the manuscript, and thank the NIH (R01 HG005966-01) and NSF (1223057) for financial support.

FIGURE LEGENDS

Figure 1: Unweighted phylogenetic diversity (PD, left) and an abundance-weighted PD measure (right), where taxa present in a sample are shown as blue circles and abundances are shown as the size of the circles. Unweighted PD takes the sum of the branch lengths connecting samples. Abundance-weighted measures take a weighted sum of branch lengths where weight is determined in some way by the abundance of the taxa on either side of the branch: if we give edges width according to their weight, the abundance-weighted measure can be thought of as the sum of the areas of the edges.

Figure 2: The UniFrac divergence measure (figure adapted from Lozupone and Knight 2005). When samples are interspersed across the tree (left tree), they have a smaller *fraction* of branch lengths that can be *uniquely* assigned to one sample or another compared to when they are separate (right tree).

Figure 3: Part of a minimal mass movement to calculate the earth-mover's distance on a phylogenetic tree.

*

References

S. Abubucker, N. Segata, J. Goll, A. M. Schubert, J. Izard, B. L. Cantarel, B. Rodriguez-Mueller, J. Zucker, M. Thiagarajan, B. Henrissat, O. White, S. T. Kelley, B. Meth, P. D. Schloss, D. Gevers, M. Mitreva, and C. Huttenhower. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLOS Computational Biology*, 8(6):e1002358, 2012.

D. J. Aldous, M. A. Krikun, and L. Popovic. Five statistical questions about the tree of life. *Systematic Biology*, 60(3):318–328, 2011.

B. Allen, M. Kon, and Y. Bar-Yam. A new phylogenetic diversity measure generalizing the Shannon index and its application to phyllostomid bats. *The American Naturalist*, 174(2):236–243, 2009.

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

K. E. Ashelford, N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Applied and Environmental Microbiology*, 71(12):7724–7736, 2005.

K. E. Ashelford, N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman. New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. *Applied and Environmental Microbiology*, 72(9):5734–5741, 2006.

B. J. Baker, L. R. Comolli, G. J. Dick, L. J. Hauser, D. Hyatt, B. D. Dill, M. L. Land, N. C. VerBerkmoes, R. L. Hettich, and J. F. Banfield. Enigmatic, ultrasmall, uncultivated Archaea. *Proceedings of the National Academy of Sciences*, 107(19):8806–8811, 2010.

E. Bapteste, M. A. O'Malley, R. G. Beiko, M. Ereshefsky, J. P. Gogarten, L. Franklin-Hall, F.-J. Lapointe, J. Dupré, T. Dagan, Y. Boucher, and W. Martin. Prokaryotic evolution and the tree of life are two different things. *Biol Direct*, 4(1):34, 2009.

G. Barker. Phylogenetic diversity: A quantitative framework for measurement of priority and achievement in biodiversity conservation. *Biological Journal of the*

Linnean Society, 76(2):165–194, 2002.

A. Bazinet and M. Cummings. A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, 13(1):92, 2012.

S. Berger and A. Stamatakis. Aligning short reads to reference alignments and trees. *Bioinformatics*, 27(15):2068–2075, 2011.

S. Berger, D. Krompass, and A. Stamatakis. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biology*, 60(3):291–302, 2011.

M. Blaser. Antibiotic overuse: Stop the killing of beneficial bacteria. *Nature*, 476 (7361):393–394, 2011.

E. Borenstein, M. Kupiec, M. W. Feldman, and E. Ruppin. Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proceedings of the National Academy of Sciences*, 105(38):14482–14487, 2008.

A. Brady and S. L. Salzberg. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nature methods*, 6(9):673–676, 2009.

L. Bragg, G. Stone, M. Imelfort, P. Hugenholtz, and G. W. Tyson. Fast, accurate error-correction of amplicon pyrosequences using Acacia. *Nature Methods*, 9(5): 425–426, 2012.

J. G. Caporaso, K. Bittinger, F. D. Bushman, T. Z. DeSantis, G. L. Andersen, and R. Knight. PyNAST: A flexible tool for aligning sequences to a template alignment. *Bioinformatics*, 26(2):266–267, 2010a.

J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E. Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, and R. Knight. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336, 2010b.

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

J. G. Caporaso, C. L. Lauber, W. A. Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S. M. Owens, J. Betley, L. Fraser, M. Bauer, N. Gormley, J. A. Gilbert, G. Smith, and R. Knight. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal*, 6(8):1621–1624, 2012.

A. Caro-Quintero and K. T. Konstantinidis. Bacterial species may exist, metagenomics reveal. *Environmental Microbiology*, 14(2):347–355, 2012.

R. J. Case, Y. Boucher, I. Dahllof, C. Holmstrom, W. F. Doolittle, and S. Kjelleberg. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl. Environ. Microbiol.*, 73(1):278–288, Jan 2007.

A. Chao, C. Chiu, and L. Jost. Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365 (1558):3599–3609, 2010.

J. Chen, K. Bittinger, E. S. Charlson, C. Hoffmann, J. Lewis, G. D. Wu, R. G. Collman, F. D. Bushman, and H. Li. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*, 28 (16):2106–2113, 2012.

T. Chen, W. Yu, J. Izard, O. Baranova, A. Lakshmanan, and F. Dewhirst. The Human Oral Microbiome Database: A web accessible resource for investigating oral microbe taxonomic and genomic information. *Database: the Journal of Biological Databases and Curation*, 2010, 2010.

I. Cho and M. J. Blaser. The human microbiome: At the interface of health and disease. *Nature Reviews Genetics*, 13(4):260–270, 2012.

M. J. Claesson, I. B. Jeffery, S. Conde, S. E. Power, E. M. O'Connor, S. Cusack, H. M. B. Harris, M. Coakley, B. Lakshminarayanan, O. O'Sullivan, G. F. Fitzgerald, J. Deane, M. O'Connor, N. Harnedy, K. O'Connor, D. O'Mahony, D. van Sinderen, M. Wallace, L. Brennan, C. Stanton, J. R. Marchesi, A. P. Fitzgerald, F. Shanahan, C. Hill, R. P. Ross, and P. W. O'Toole. Gut microbiota composition

correlates with diet and health in the elderly. *Nature*, 488(7410):178–184, 2012.

A. E. Clark, E. J. Kaleta, A. Arora, and D. M. Wolk. Matrix-assisted laser desorption ionization–time of flight mass spectrometry: A fundamental shift in the routine practice of clinical microbiology. *Clinical Microbiology Reviews*, 26(3):547–603, 2013.

H. M. P. Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486:207–214, 2012.

M. Costandi. Citizen microbiome. *Nature Biotechnology*, 31(2):90–90, 2013.

E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon, and R. Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–1697, 2009.

E. K. Costello, K. Stagaman, L. Dethlefsen, B. J. Bohannan, and D. A. Relman. The application of ecological theory toward an understanding of the human microbiome. *Science*, 336(6086):1255–1262, 2012.

D. Dalevi, T. DeSantis, J. Fredslund, G. Andersen, V. Markowitz, and P. Hugenholtz. Automated group assignment in large phylogenetic trees using GRUNT: GRouping, Ungrouping, Naming Tool. *BMC Bioinformatics*, 8(1):402, 2007.

P. H. Degnan and H. Ochman. Illumina-based analysis of microbial community diversity. *The ISME Journal*, 6(1):183–194, 2011.

P. S. Dehal, M. P. Joachimiak, M. N. Price, J. T. Bates, J. K. Baumohl, D. Chivian, G. D. Friedland, K. H. Huang, K. Keller, P. S. Novichkov, I. L. Dubchak, E. J. Alm, and A. P. Arkin. MicrobesOnline: An integrated portal for comparative and functional genomics. *Nucleic Acids Research*, 38(suppl 1):D396–D400, 2010.

F. Delsuc, J. L. Metcalf, L. Wegener Parfrey, S. J. Song, A. González, and R. Knight. Convergence of gut microbiomes in myrmecophagous mammals. *advance access, Molecular Ecology*, 2013.

T. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. Andersen. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 72(7):5069, 2006a.

T. DeSantis, P. Hugenholtz, K. Keller, E. Brodie, N. Larsen, Y. Piceno, R. Phan, and G. L. Andersen. NAST: A multiple sequence alignment server for comparative analysis of 16S rRNA genes. *Nucleic Acids Research*, 34(suppl 2):W394–W399, 2006b.

L. Dethlefsen and D. A. Relman. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4554–4561, 2011.

L. Dethlefsen, S. Huse, M. L. Sogin, and D. A. Relman. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. *PLOS Biology*, 6(11):e280, 2008.

X. Didelot, R. Bowden, D. J. Wilson, T. E. Peto, and D. W. Crook. Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics*, 13 (9):601–612, 2012.

S. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

R. C. Edgar. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19):2460–2461, Oct 2010.

R. C. Edgar. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *advance access, Nature Methods*, 2013.

R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200, 2011.

J. B. Emerson, B. C. Thomas, K. Andrade, E. E. Allen, K. B. Heidelberg, and J. F. Banfield. Metagenomic assembly reveals dynamic viral populations in hypersaline systems. *Applied and Environmental Microbiology*, 2012.

J. Evans, L. Sheneman, and J. Foster. Relaxed neighbor joining: A fast distance-based phylogenetic tree construction method. *Journal of Molecular Evolution*, 62 (6):785–792, 2006.

S. N. Evans and F. A. Matsen. The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples. *J. Royal Stat. Soc. (B)*, 74(3):569–592, 2012.

D. Faith. Conservation evaluation and phylogenetic diversity. *Biological Conservation*, 61(1):1–10, 1992.

J. Felsenstein. Phylogenies and the comparative method. *American Naturalist*, pages 1–15, 1985.

N. Fierer, S. Ferrenberg, G. E. Flores, A. González, J. Kueneman, T. Legg, R. C. Lynch, D. McDonald, J. R. Mihaljevic, S. P. O'Neill, M. E. Rhodes, S. J. Song, and W. A. Walters. From animalcules to an ecosystem: Application of ecological concepts to the human microbiome. *Annual Review of Ecology, Evolution, and Systematics*, 43:137–155, 2012.

A. A. Fodor, T. Z. DeSantis, K. M. Wylie, J. H. Badger, Y. Ye, T. Hepburn, P. Hu, E. Sodergren, K. Liolios, H. Huot-Creasy, B. W. Birren, and A. M. Earl. The most wanted taxa from the human microbiome for whole genome sequencing. *PLOS ONE*, 7(7):e41294, 2012.

P. Forey. The PhyloCode: Description and commentary. *Bulletin of Zoological Nomenclature*, 58:81–96, 2001.

G. E. Fox, K. R. Pechman, and C. R. Woese. Comparative cataloging of 16S ribosomal ribonucleic acid: Molecular approach to procaryotic systematics. *International Journal of Systematic Bacteriology*, 27(1):44–57, 1977.

L. J. Funkhouser and S. R. Bordenstein. Mom knows best: The universality of maternal microbial transmission. *PLOS Biology*, 11(8):e1001631, 2013.

S. Greenblum, P. J. Turnbaugh, and E. Borenstein. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences*, 109(2):594–599, 2012.

E. A. Grice, H. H. Kong, S. Conlan, C. B. Deming, J. Davis, A. C. Young, G. G. Bouffard, R. W. Blakesley, P. R. Murray, E. D. Green, M. L. Turner, and J. A. Segre. Topographical and temporal diversity of the human skin microbiome. *Science*, 324(5931):1190–1192, 2009.

A. Griffen, C. Beall, N. Firestone, E. Gross, J. DiFranco, J. Hardman, B. Vriesendorp, R. Faust, D. Janies, and E. Leys. CORE: A phylogenetically-curated 16S rDNA

database of the core oral microbiome. *PLOS ONE*, 6(4):e19051, 2011.

B. J. Haas, D. Gevers, A. M. Earl, M. Feldgarden, D. V. Ward, G. Giannoukos, D. Ciulla, D. Tabbaa, S. K. Highlander, E. Sodergren, B. Methé, T. Z. DeSantis, T. H. M. Consortium, J. F. Petrosino, R. Knight, and B. W. Birren. Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, 21(3):494–504, 2011.

X. Hao, R. Jiang, and T. Chen. Clustering 16S rRNA for OTU prediction: A method of unsupervised Bayesian clustering. *Bioinformatics*, 27(5):611–618, 2011.

K. Hartmann and M. Steel. Maximizing phylogenetic diversity in biodiversity conservation: Greedy solutions to the Noah's Ark problem. *Systematic Biology*, 55(4):644–651, 2006.

J.-H. Hehemann, G. Correc, T. Barbeyron, W. Helbert, M. Czjzek, and G. Michel. Transfer of carbohydrate-active enzymes from marine bacteria to japanese gut microbiota. *Nature*, 464(7290):908–912, 2010.

K. M. Hewitt, F. L. Mannino, A. Gonzalez, J. H. Chase, J. G. Caporaso, R. Knight, and S. T. Kelley. Bacterial diversity in two neonatal intensive care units (NICUs). *PLOS ONE*, 8(1):e54703, 2013.

C. Hoffmann, S. Dollive, S. Grunberg, J. Chen, H. Li, G. D. Wu, J. D. Lewis, and F. D. Bushman. Archaea and Fungi of the human gut microbiome: Correlations with diet and bacterial residents. *PLOS ONE*, 8(6):e66019, 2013.

I. Holmes, K. Harris, and C. Quince. Dirichlet multinomial mixtures: Generative models for microbial metagenomics. *PLOS ONE*, 7(2):e30126, 2012.

L. V. Hooper, D. R. Littman, and A. J. Macpherson. Interactions between the microbiota and the immune system. *Science*, 336(6086):1268–1273, 2012.

A. C. Howe, J. Jansson, S. A. Malfatti, S. G. Tringe, J. M. Tiedje, and C. T. Brown. Assembling large, complex environmental metagenomes. *arXiv preprint arXiv:1212.2832*, 2012.

P. Hugenholtz and T. Huber. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *International Journal of Systematic and Evolutionary Microbiology*, 53(1):289–293, 2003.

D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster. Megan analysis of metagenomic data. *Genome research*, 17(3):377–386, 2007.

D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster. Integrative analysis of environmental sequences using MEGAN4. *Genome Research*, 21(9): 1552–1560, 2011.

V. Iverson, R. M. Morris, C. D. Frazar, C. T. Berthiaume, R. L. Morales, and E. V. Armbrust. Untangling genomes from metagenomes: Revealing an uncultured class of marine Euryarchaeota. *Science*, 335(6068):587–590, 2012.

P. Jaccard. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaudoise Sci. Nat.*, 44:223–270, 1908.

H. E. Jakobsson, C. Jernberg, A. F. Andersson, M. Sjölund-Karlsson, J. K. Jansson, and L. Engstrand. Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. *PLOS ONE*, 5(3):e9836, 2010.

C. Jernberg, S. Löfmark, C. Edlund, and J. K. Jansson. Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. *The ISME Journal*, 1(1):56–66, 2007.

T. Kalisky and S. R. Quake. Single-cell genomics. *Nature Methods*, 8(4):311–314, 2011.

A. L. Kau, P. P. Ahern, N. W. Griffin, A. L. Goodman, and J. I. Gordon. Human nutrition, the gut microbiome and the immune system. *Nature*, 474(7351):327–336, 2011.

S. W. Kembel, J. A. Eisen, K. S. Pollard, and J. L. Green. The phylogenetic diversity of metagenomes. *PLOS ONE*, 6(8):e23214, 2011.

S. W. Kembel, M. Wu, J. A. Eisen, and J. L. Green. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLOS Computational Biology*, 8(10):e1002743, 2012.

J. A. Klappenbach, J. M. Dunbar, and T. M. Schmidt. rRNA operon copy number reflects ecological strategies of bacteria. *Applied and Environmental Microbiology*, 66(4):1328–1333, 2000.

J. A. Klappenbach, P. R. Saxman, J. R. Cole, and T. M. Schmidt. rrndb: The ribosomal RNA operon copy number database. *Nucleic Acids Research*, 29(1):181–184, 2001.

A. G. Kluge and J. S. Farris. Quantitative phyletics and the evolution of anurans. *Systematic Biology*, 18(1):1–32, 1969.

O. Koren, J. K. Goodrich, T. C. Cullender, A. Spor, K. Laitinen, H. Kling Bäckhed, A. Gonzalez, J. J. Werner, L. T. Angenent, R. Knight, F. Bäckhed, E. Isolauri, S. Salminen, and R. E. Ley. Host remodeling of the gut microbiome and metabolic changes during pregnancy. *Cell*, 150(3):470–480, 2012.

C. U. Köser, M. T. Holden, M. J. Ellington, E. J. Cartwright, N. M. Brown, A. L. Ogilvy-Stuart, L. Y. Hsu, C. Chewapreecha, N. J. Croucher, S. R. Harris, M. Sanders, M. C. Enright, G. Dougan, S. D. Bentley, J. Parkhill, L. J. Fraser, J. R. Betley, O. B. Schulz-Trieglaff, G. P. Smith, and S. J. Peacock. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. *New England Journal of Medicine*, 366(24):2267–2275, 2012.

N. Kreig, J. Holt, R. Murray, D. Breener, M. Bryant, J. Moulder, N. Pfennig, P. Sneath, and J. Staley. *Bergey's Manual of Systematic Bacteriology.* Williams & Wilkins, 1984.

J. Kuczynski, Z. Liu, C. Lozupone, D. McDonald, N. Fierer, and R. Knight. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nature Methods*, 7:813–819, 2010.

P. S. La Rosa, J. P. Brooks, E. Deych, E. L. Boone, D. J. Edwards, Q. Wang, E. Sodergren, G. Weinstock, and W. D. Shannon. Hypothesis testing and power calculations for taxonomic-based human microbiome data. *PLOS ONE*, 7(12):e52078, 2012a.

P. S. La Rosa, B. Shands, E. Deych, Y. Zhou, E. Sodergren, G. Weinstock, and W. D. Shannon. Statistical object data analysis of taxonomic trees from human microbiome data. *PLOS ONE*, 7(11):e48996, 2012b.

A. Lambert and M. Steel. Predicting the loss of phylogenetic diversity under non-stationary diversification models. *arXiv preprint arXiv:1306.2710*, 2013.

D. J. Lane, B. Pace, G. J. Olsen, D. A. Stahl, M. L. Sogin, and N. R. Pace. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences*, 82(20):6955–6959, 1985.

M. G. Langille, J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. A. Reyes, J. C. Clemente, D. E. Burkepile, R. L. V. Thurber, R. Knight, R. G. Beiko, and C. Huttenhower. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31(9):814–821, 2013.

R. Ley, P. Turnbaugh, S. Klein, and J. Gordon. Microbial ecology: Human gut microbes associated with obesity. *Nature*, 444(7122):1022–1023, 2006.

F. Li, M. A. Hullar, Y. Schwarz, and J. W. Lampe. Human gut bacterial communities are altered by addition of cruciferous vegetables to a controlled fruit-and vegetable-free diet. *The Journal of Nutrition*, 139(9):1685–1691, 2009.

W. Li and A. Godzik. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, Jul 2006.

Z. Liu, T. Z. DeSantis, G. L. Andersen, and R. Knight. Accurate taxonomy assignments from 16s rrna sequences produced by highly parallel pyrosequencers. *Nucleic acids research*, 36(18):e120–e120, 2008.

C. Lozupone and R. Knight. UniFrac: A new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.*, 71(12):8228, 2005. ISSN 0099-2240. doi: 10.1128/AEM.71.12.8228.

C. A. Lozupone, M. Hamady, S. T. Kelley, and R. Knight. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, 73(5):1576–85, 2007. ISSN 0099-2240.

C. Lozupone, J. Stombaugh, A. Gonzalez, G. Ackermann, D. Wendel, Y. Vazquez-Baeza, J. K. Jansson, J. I. Gordon, and R. Knight. Meta-analyses of studies of the human microbiota. *advance access, Genome Research*, 2013.

C. A. Lozupone and R. Knight. Global patterns in bacterial diversity. *Proceedings of the National Academy of Sciences*, 104(27):11436–11440, 2007.

W. Ludwig, O. Strunk, R. Westram, L. Richter, H. Meier, Yadhukumar, A. Buchner, T. Lai, S. Steppi, G. Jobb, W. Frster, I. Brettske, S. Gerber, A. W. Ginhart, O. Gross, S. Grumann, S. Hermann, R. Jost, A. Knig, T. Liss, R. Lmann, M. May, B. Nonhoff, B. Reichel, R. Strehlow, A. Stamatakis, N. Stuckmann, A. Vilbig, M. Lenke, T. Ludwig, A. Bode, and K. Schleifer. ARB: A software environment for sequence data. *Nucleic Acids Res*, 32(4):1363, 2004.

S. S. Mande, M. H. Mohammed, and T. S. Ghosh. Classification of metagenomic sequences: Methods and challenges. *Briefings in Bioinformatics*, 13(6):669–681, 2012.

F. A. Matsen and A. Gallagher. Reconciling taxonomy and phylogenetic inference: Formalism and algorithms for describing discord and inferring taxonomic roots. *Algorithms for Molecular Biology*, 7:8, 2012. doi: 10.1186/1748-7188-7-8.

F. Matsen, R. Kodner, and E. Armbrust. pplacer: Linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11(1):538, 2010.

F. A. Matsen and S. N. Evans. Edge principal components and squash clustering: Using the special structure of phylogenetic placement data for sample comparison. *PLOS ONE*, 8(3):e56859, 2013.

C. F. Maurice, H. J. Haiser, and P. J. Turnbaugh. Xenobiotics shape the physiology and gene expression of the active human gut microbiome. *Cell*, 152(1):39–50, 2013.

C. McCoy and F. Matsen. Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. *PeerJ*, 9:e157, 2013. doi: 10.7717/peerj.157.

D. McDonald, M. N. Price, J. Goodrich, E. P. Nawrocki, T. Z. DeSantis, A. Probst, G. L. Andersen, R. Knight, and P. Hugenholtz. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, 6(3):610–618, 2011.

D. McDonald, J. C. Clemente, J. Kuczynski, J. R. Rideout, J. Stombaugh, D. Wendel, A. Wilke, S. Huse, J. Hufnagle, F. Meyer, R. Knight, and J. G. Caporaso. The

biological observation matrix (BIOM) format or: How i learned to stop worrying and love the ome-ome. *GigaScience*, 1(1):7, 2012.

A. McNabb, D. Eisler, K. Adie, M. Amos, M. Rodrigues, G. Stephens, W. A. Black, and J. Isaac-Renton. Assessment of partial sequencing of the 65-kilodalton heat shock protein gene (hsp65) for routine identification of Mycobacterium species isolated from clinical sources. *J. Clin. Microbiol.*, 42(7):3000–3011, Jul 2004.

B. A. Methé, K. E. Nelson, M. Pop, H. H. Creasy, M. G. Giglio, C. Huttenhower, D. Gevers, J. F. Petrosino, S. Abubucker, J. H. Badger, A. T. Chinwalla, A. M. Earl, M. G. FitzGerald, R. S. Fulton, K. Hallsworth-Pepin, E. A. Lobos, R. Madupu, V. Magrini, J. C. Martin, M. Mitreva, D. M. Muzny, E. J. Sodergren, J. Versalovic, A. M. Wollam, K. C. Worley, J. R. Wortman, S. K. Young, Q. Zeng, K. M. Aagaard, O. O. Abolude, E. Allen-Vercoe, E. J. Alm, L. Alvarado, G. L. Andersen, S. Anderson, E. Appelbaum, H. M. Arachchi, G. Armitage, C. A. Arze, T. Ayvaz, C. C. Baker, L. Begg, T. Belachew, V. Bhonagiri, M. Bihan, M. J. Blaser, T. Bloom, V. R. Bonazzi, P. Brooks, G. A. Buck, C. J. Buhay, D. A. Busam, J. L. Campbell, S. R. Canon, B. L. Cantarel, P. S. Chain, I.-M. A. Chen, L. Chen, S. Chhibba, K. Chu, D. M. Ciulla, J. C. Clemente, S. W. Clifton, S. Conlan, J. Crabtree, M. A. Cutting, N. J. Davidovics, C. C. Davis, T. Z. DeSantis, C. Deal, K. D. Delehaunty, F. E. Dewhirst, E. Deych, Y. Ding, D. J. Dooling, S. P. Dugan, W. Michael Dunne, A. Scott Durkin, R. C. Edgar, R. L. Erlich, C. N. Farmer, R. M. Farrell, K. Faust, M. Feldgarden, V. M. Felix, S. Fisher, A. A. Fodor, L. Forney, L. Foster, V. Di Francesco, J. Friedman, D. C. Friedrich, C. C. Fronick, L. L. Fulton, H. Gao, N. Garcia, G. Giannoukos, C. Giblin, M. Y. Giovanni, J. M. Goldberg, J. Goll, A. Gonzalez, A. Griggs, S. Gujja, B. J. Haas, H. A. Hamilton, E. L. Harris, T. A. Hepburn, B. Herter, D. E. Hoffmann, M. E. Holder, C. Howarth, K. H. Huang, S. M. Huse, J. Izard, J. K. Jansson, H. Jiang, C. Jordan, V. Joshi, J. A. Katancik, W. A. Keitel, S. T. Kelley, C. Kells, S. Kinder-Haake, N. B. King, R. Knight, D. Knights, H. H. Kong, O. Koren, S. Koren, K. C. Kota, C. L. Kovar, N. C. Kyrpides, P. S. La Rosa, S. L. Lee, K. P. Lemon, N. Lennon, C. M. Lewis, L. Lewis, R. E. Ley, K. Li, K. Liolios, B. Liu, Y. Liu, C.-C. Lo, C. A. Lozupone,

R. Dwayne Lunsford, T. Madden, A. A. Mahurkar, P. J. Mannon, E. R. Mardis, V. M. Markowitz, K. Mavrommatis, J. M. McCorrison, D. McDonald, J. McEwen, A. L. McGuire, P. McInnes, T. Mehta, K. A. Mihindukulasuriya, J. R. Miller, P. J. Minx, I. Newsham, C. Nusbaum, M. OLaughlin, J. Orvis, I. Pagani, K. Palaniappan, S. M. Patel, M. Pearson, J. Peterson, M. Podar, C. Pohl, K. S. Pollard, M. E. Priest, L. M. Proctor, X. Qin, J. Raes, J. Ravel, J. G. Reid, M. Rho, R. Rhodes, K. P. Riehle, M. C. Rivera, B. Rodriguez-Mueller, Y.-H. Rogers, M. C. Ross, C. Russ, R. K. Sanka, P. Sankar, J. Fah Sathirapongsasuti, J. A. Schloss, P. D. Schloss, T. M. Schmidt, M. Scholz, L. Schriml, A. M. Schubert, N. Segata, J. A. Segre, W. D. Shannon, R. R. Sharp, T. J. Sharpton, N. Shenoy, N. U. Sheth, G. A. Simone, I. Singh, C. S. Smillie, J. D. Sobel, D. D. Sommer, P. Spicer, G. G. Sutton, S. M. Sykes, D. G. Tabbaa, M. Thiagarajan, C. M. Tomlinson, M. Torralba, T. J. Treangen, R. M. Truty, T. A. Vishnivetskaya, J. Walker, L. Wang, Z. Wang, D. V. Ward, W. Warren, M. A. Watson, C. Wellington, K. A. Wetterstrand, J. R. White, K. Wilczek-Boney, Y. Qing Wu, K. M. Wylie, T. Wylie, C. Yandava, L. Ye, Y. Ye, S. Yooseph, B. P. Youmans, L. Zhang, Y. Zhou, Y. Zhu, L. Zoloth, J. D. Zucker, B. W. Birren, R. A. Gibbs, S. K. Highlander, G. M. Weinstock, R. K. Wilson, and O. White. A framework for human microbiome research. *Nature*, 486(7402): 215–221, 2012.

S. Minot, R. Sinha, J. Chen, H. Li, S. A. Keilbaugh, G. D. Wu, J. D. Lewis, and F. D. Bushman. The human gut virome: Inter-individual variation and dynamic response to diet. *Genome Research*, 21(10):1616–1625, 2011.

S. Minot, A. Bryson, C. Chehoud, G. D. Wu, J. D. Lewis, and F. D. Bushman. Rapid evolution of the human gut virome. *Proceedings of the National Academy of Sciences*, 110(30):12450–12455, 2013.

S. Mirarab, N. Nguyen, and T. Warnow. SEPP: SATé-enabled phylogenetic placement. *Accepted to the Pacific Symposium on Biocomputing*, 2012. http://www.cs.utexas.edu/ tandy/warnow-psb2012.pdf.

A. Monier, J. Claverie, and H. Ogata. Taxonomic distribution of large DNA viruses in the sea. *Genome Biol*, 9(7):R106, 2008.

A. O. Mooers and S. B. Heard. Inferring evolutionary process from phylogenetic tree shape. *Q. Rev. Biol.*, pages 31–54, 1997.

S. Moran and S. Snir. Convex recolorings of strings and trees: Definitions, hardness results and algorithms. *Journal of Computer and System Sciences*, 74(5):850–869, 2008.

J. L. Morgan, A. E. Darling, and J. A. Eisen. Metagenomic sequencing of an in vitro-simulated microbial community. *PLOS ONE*, 5(4):e10209, 2010.

K. Munch, W. Boomsma, J. P. Huelsenbeck, E. Willerslev, and R. Nielsen. Statistical assignment of DNA sequences using Bayesian phylogenetics. *Systematic Biology*, 57(5):750–757, 2008a.

K. Munch, W. Boomsma, E. Willerslev, and R. Nielsen. Fast phylogenetic DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363 (1512):3997–4002, 2008b.

S. Navlakha, J. White, N. Nagarajan, M. Pop, and C. Kingsford. Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information. In *Research in Computational Molecular Biology*, pages 400–417. Springer, 2009.

E. Nawrocki, D. Kolbe, and S. Eddy. Infernal 1.0: Inference of RNA alignments. *Bioinformatics*, 25(10):1335–1337, 2009.

E. P. Nawrocki. *Structural RNA homology search and alignment using covariance models*. PhD thesis, Washington University, 2009. Advisor: Sean R Eddy.

D. Nipperess and F. Matsen. The mean and variance of phylogenetic diversity under rarefaction. *Methods in Ecology and Evolution*, 2013. doi: 10.1111/2041-210X.12042.

H. Ochman, M. Worobey, C.-H. Kuo, J.-B. N. Ndjango, M. Peeters, B. H. Hahn, and P. Hugenholtz. Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLOS Biology*, 8(11):e1000546, 2010.

J. P. O'Dwyer, S. W. Kembel, and J. L. Green. Phylogenetic diversity theory sheds light on the structure of microbial communities. *PLOS Computational Biology*, 8 (12):e1002832, 2012.

M. Pagel. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 255(1342):37–45, 1994.

E. Paradis, J. Claude, and K. Strimmer. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20(2):289–290, 2004.

F. Pardi and N. Goldman. Resource-aware taxon selection for maximizing phylogenetic diversity. *Systematic Biology*, 56(3):431–444, 2007.

D. Parks, N. MacDonald, and R. Beiko. Classifying short genomic fragments from novel lineages using composition and homology. *BMC Bioinformatics*, 12(1):328, 2011.

J. N. Paulson, O. C. Stine, H. C. Bravo, and M. Pop. Differential abundance analysis for microbial marker-gene surveys. *advance access,Nature methods*, 2013. http://dx.doi.org/10.1038/nmeth.2658.

J. Pell, A. Hintze, R. Canino-Koning, A. Howe, J. M. Tiedje, and C. T. Brown. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proceedings of the National Academy of Sciences*, 109(33):13272–13277, 2012.

C. Phillips, G. Phelan, S. Dowd, M. MCdonough, A. Ferguson, J. Delton Hanson, L. Siles, N. Ordóñez-garza, M. San Francisco, and R. Baker. Microbiome analysis among bats describes influences of host phylogeny, life history, physiology and geography. *Molecular Ecology*, 21:26172627, 2012.

S. Podell, J. A. Ugalde, P. Narasingarao, J. F. Banfield, K. B. Heidelberg, and E. E. Allen. Assembly-driven community genomics of a hypersaline microbial ecosystem. *PLOS ONE*, 8(4):e61692, 2013.

M. F. Polz and C. M. Cavanaugh. Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, 64(10):3724–3730, 1998.

J. Pons, T. G. Barraclough, J. Gomez-Zurita, A. Cardoso, D. P. Duran, S. Hazell, S. Kamoun, W. D. Sumlin, and A. P. Vogler. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology*, 55(4):595–609, 2006.

T. Poutahidis, M. Kleinewietfeld, C. Smillie, T. Levkovich, A. Perrotta, S. Bhela, B. J. Varian, Y. M. Ibrahim, J. R. Lakritz, S. M. Kearney, A. Chatzigiagkos, D. A. Hafler, E. J. Alm, and S. E. Erdman. Microbial reprogramming inhibits Western diet-associated obesity. *PLOS ONE*, 8(7):e68596, 2013.

M. Price, P. Dehal, and A. Arkin. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLOS ONE*, 5(3):e9490, 2010.

E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner. SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35 (21):7188–7196, 2007.

E. Pruesse, J. Peplies, and F. O. Glöckner. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, 28(14):1823–1829, 2012.

E. Purdom. Analyzing data with graphs: Metagenomic data and the phylogenetic tree. *UC Berkeley Statistics Technical Reports*, 766:1–22, 2008.

J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Dore, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S. D. Ehrlich, and J. Wang. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.

C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, 2013.

C. Quince, A. Lanzén, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read, and W. T. Sloan. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature methods*, 6(9):639–641, 2009.

C. Quince, A. Lanzen, R. J. Davenport, and P. J. Turnbaugh. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, 12(1):38, 2011.

C. Rao. Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology*, 21(1):24–43, 1982.

A. Reyes, M. Haynes, N. Hanson, F. E. Angly, A. C. Heath, F. Rohwer, and J. I. Gordon. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, 466(7304):334–338, 2010.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

G. B. Rogers, L. R. Hoffman, M. P. Carroll, and K. D. Bruce. Interpreting infective microbiota: The importance of an ecological perspective. *Trends in Microbiology*, 21(6):271–276, 2013. ISSN 0966-842X. doi: http://dx.doi.org/10.1016/j.tim.2013.03.004.

G. Rosen, E. Garbarine, D. Caseiro, R. Polikar, and B. Sokhansanj. Metagenome fragment classification using n-mer frequency profiles. *Advances in Bioinformatics*, 2008, 2008.

P. D. Schloss. Evaluating different approaches that test whether microbial communities have the same structure. *The ISME Journal*, 2(3):265–275, 2008.

P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75(23):7537–7541, 2009.

P. D. Schloss, D. Gevers, and S. L. Westcott. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLOS ONE*, 6(12):

e27310, 2011.

N. Segata, L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson, and C. Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, 2012.

C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(1):379–423, 1948. ISSN 15591662. doi: 10.1145/584091.584093.

T. J. Sharpton, S. J. Riesenfeld, S. W. Kembel, J. Ladau, J. P. O'Dwyer, J. L. Green, J. A. Eisen, and K. S. Pollard. PhylOTU: A high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLOS Computational Biology*, 7(1):e1001061, 2011.

L. Sheneman, J. Evans, and J. A. Foster. Clearcut: A fast implementation of relaxed neighbor joining. *Bioinformatics*, 22(22):2823–2824, 2006.

E. Simpson. Measurement of diversity. *Nature*, 163(4148):688, 1949.

C. S. Smillie, M. B. Smith, J. Friedman, O. X. Cordero, L. A. David, and E. J. Alm. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*, 480(7376):241–244, 2011.

M. I. Smith, T. Yatsunenko, M. J. Manary, I. Trehan, R. Mkakosya, J. Cheng, A. L. Kau, S. S. Rich, P. Concannon, J. C. Mychaleckyj, J. Liu, E. Houpt, J. V. Li, E. Holmes, J. Nicholson, D. Knights, L. K. Ursell, R. Knight, and J. I. Gordon. Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science*, 339(6119):548–554, 2013.

E. S. Snitkin, A. M. Zelazny, P. J. Thomas, F. Stock, D. K. Henderson, T. N. Palmore, and J. A. Segre. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. *Science Translational Medicine*, 4 (148):148ra116–148ra116, 2012.

S. Srinivasan, N. G. Hoffman, M. T. Morgan, F. A. Matsen, T. L. Fiedler, R. W. Hall, F. J. Ross, C. O. McCoy, R. Bumgarner, J. M. Marrazzo, and D. N. Fredricks. Bacterial communities in women with bacterial vaginosis: High resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. *PLOS ONE*, 7(6):e37818, 2012.

M. Stark, S. Berger, A. Stamatakis, and C. von Mering. Mltreemap-accurate maximum likelihood placement of environmental dna sequences into taxonomic and functional reference phylogenies. *BMC genomics*, 11(1):461, 2010.

B. Stecher, R. Denzler, L. Maier, F. Bernet, M. J. Sanders, D. J. Pickard, M. Barthel, A. M. Westendorf, K. A. Krogfelt, A. W. Walker, M. Ackermann, U. Dobrindt, N. R. Thomson, and W.-D. Hardt. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. *Proceedings of the National Academy of Sciences*, 109(4):1269–1274, 2012.

M. Steel and A. Rodrigo. Maximum likelihood supertrees. *Systematic Biology*, 57 (2):243–250, 2008.

M. T. Suzuki and S. J. Giovannoni. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology*, 62(2):625–630, 1996.

G. J. Szöllősi, W. Rosikiewicz, B. Boussau, E. Tannier, and V. Daubin. Efficient exploration of the space of reconciled gene trees. *arXiv preprint arXiv:1306.2167*, 2013a.

G. J. Szöllősi, E. Tannier, N. Lartillot, and V. Daubin. Lateral gene transfer from the dead. *Systematic Biology*, 62(3):386–397, 2013b.

R. Y. Tito, D. Knights, J. Metcalf, A. J. Obregon-Tito, L. Cleeland, F. Najar, B. Roe, K. Reinhard, K. Sobolik, S. Belknap, M. Foster, P. Spicer, R. Knight, and C. M. Lewis, Jr. Insights from characterizing extinct human gut microbiomes. *PLOS ONE*, 7(12):e51146, 2012.

T. J. Treangen, S. Koren, D. D. Sommer, B. Liu, I. Astrovskaya, B. Ondov, A. E. Darling, A. M. Phillippy, and M. Pop. MetAMOS: A modular and open source metagenomic assembly and analysis pipeline. *Genome Biology*, 14(1):R2, 2013.

P. J. Turnbaugh, R. E. Ley, M. A. Mahowald, V. Magrini, E. R. Mardis, and J. I. Gordon. An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature*, 444(7122):1027–131, 2006.

P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C.

Heath, R. Knight, and J. I. Gordon. A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–484, 2008.

M. Vellend, W. Cornwell, K. Magnuson-Ford, and A. Mooers. Measuring phylogenetic biodiversity. In A. Magurran and B. McGill, editors, *Biological Diversity: Frontiers in Measurement and Assessment*, pages 194–207. Oxford University Press, 2011.

C. Von Mering, P. Hugenholtz, J. Raes, S. Tringe, T. Doerks, L. Jensen, N. Ward, and P. Bork. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, 315(5815):1126–1130, 2007a.

C. Von Mering, P. Hugenholtz, J. Raes, S. Tringe, T. Doerks, L. Jensen, N. Ward, and P. Bork. Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science*, 315(5815):1126, 2007b.

Q. Wang, G. Garrity, J. Tiedje, and J. Cole. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–5267, 2007.

J. White, S. Navlakha, N. Nagarajan, M. Ghodsi, C. Kingsford, and M. Pop. Alignment and clustering of phylogenetic markers-implications for microbial diversity studies. *BMC Bioinformatics*, 11(1):152, 2010.

C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977.

D. Wu, P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. Dhaeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J.-F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H.-P. Klenk, and J. A. Eisen. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, 462(7276):1056–1060, 2009.

G. D. Wu, J. Chen, C. Hoffmann, K. Bittinger, Y.-Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight, R. Sinha, E. Gilroy, K. Gupta, R. Baldassano, L. Nessel, H. Li, F. D. Bushman, and J. D. Lewis. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.

M. Wu and J. A. Eisen. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.*, 9(10):R151, 2008a.

M. Wu and J. Eisen. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol*, 9(10):1–11, 2008b.

Z. Yang and B. Rannala. Bayesian species delimitation using multilocus sequence data. *Proceedings of the National Academy of Sciences*, 107(20):9264–9269, 2010.

P. Yarza, M. Richter, J. Peplies, J. Euzeby, R. Amann, K.-H. Schleifer, W. Ludwig, F. O. Glöckner, and R. Rosselló-Móra. The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. *Systematic and Applied Microbiology*, 31(4):241–250, 2008.

T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, and J. I. Knight, Rob andGordon. Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227, 2012.

J. R. Zaneveld, C. Lozupone, J. I. Gordon, and R. Knight. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Research*, 38(12):3869–3879, 2010.

L. Zhao. The gut microbiota and obesity: From correlation to causality. *Nature Reviews Microbiology*, 11(9):639–647, 2013.

M. L. Zupancic, B. L. Cantarel, Z. Liu, E. F. Drabek, K. A. Ryan, S. Cirimotich, C. Jones, R. Knight, W. A. Walters, D. Knights, E. F. Mongodin, R. B. Horenstein, B. D. Mitchell, N. Steinle, S. Snitker, A. R. Shuldiner, and C. M. Fraser. Analysis of the gut microbiota in the old order Amish and its relation to the metabolic syndrome. *PLOS ONE*, 7(8):e43052, 2012.