Syst. Biol. 64(1):e26–e41, 2015

DOI:10.1093/sysbio/syu053
Advance Access publication August 7, 2014

Phylogenetics and the Human Microbiome

FREDERICK A. MATSEN IV*

Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 91802, USA;
**Correspondence to be sent to: Program in Computational Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 91802, USA;*
E-mail: matsen@fhcrc.org.

Received 21 October 2013; reviews returned 28 January 2014; accepted 24 July 2014
Associate Editor: Olivier Gascuel

Abstract.—The human microbiome is the ensemble of genes in the 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 and genetic analysis of 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 article, 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. [human microbiome; human microbiota; metagenome; microbial ecology; phylogenetic methods; 16S]

The parameter regime and focus of human-associated microbial 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. Microbes are now primarily identified by their molecular sequences because such molecular identification 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 delimitation. 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 could be used in a wider setting and may deserve broader consideration by the phylogenetics community.

The human-associated microbial assemblage 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. This collection of microbes living inside and on our surfaces is called the *human microbiota*, and the ensemble of genetic information in those microbes is called the *human microbiome*, although usage of these terms varies (Boon et al. 2013).

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 microbiota, 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 Microbiota

The human microbiota 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 microbiota have remarkable metabolic potential, being 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 microbiota, 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. Traditionally, our microbiota have been transmitted from mother to infant in the birth canal and by breastfeeding (reviewed in Funkhouser and Bordenstein 2013). In this section, I will briefly review what is known about the human microbiota and its effect on our health.

The human microbiota form an ecosystem. It is dynamic in terms of taxonomic representation but apparently constant in terms of function (Consortium 2012). There is a "core" microbiota which is shared between all humans (Turnbaugh et al. 2008). The human microbiota 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 interindividual versus intraindividual variation (Consortium 2012).

Our actions can shift the composition of our microbiota. Changes in diet can very quickly shift its composition, but there is also a strong correlation between long-term diet and microbiota (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 microbiota interact on many levels with host phenotype (reviewed in Cho and Blaser 2012). The gut microbiota, 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 gut microbiota 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 microbiota leading to obesity has been established for mice: an obese phenotype can be transferred from mouse to mouse by gut microbial transplantation (Turnbaugh et al. 2006), the pregnant human gut microbiota 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 microbiota research, and other domains have been investigated to a lesser extent. Changes in archaeal and fungal populations have been shown to covary with bacterial residents (Hoffmann et al. 2013) and have a nonuniform distribution across the human skin (Findley 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 article, we will primarily be describing the human microbiota 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 microbiota 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 these data are available on a dedicated website (http://www.hmpdacc.org/). The MetaHIT study (Qin et al. 2010) also generated lots of data but it is not available to outside researchers.

Inferring Microbial Community Composition 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 detection of almost all bacteria and archaea regardless of whether they can be cultured.

In microbial ecology, the census of bacteria in a given environment using marker gene amplification and sequencing are generally called "marker gene surveys." This terminology is equivalent to the "barcoding" terminology more commonly used for eukaryotic surveys using 18S or the fungal internal transcribed spacer (ITS). Such surveys would ideally return a census of all the microbes in a sample along with their abundances.

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 trade-off 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, to compare more similar samples deeper sequencing is required. In addition to 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 fifteen copies and there can be diversity within the copies (Klappenbach et al. 2001). This can distort inferences concerning actual microbe abundances based on read copy number. Recent work by Kembel et al. (2012) implements the independent contrasts method (Felsenstein 1985) 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., McNabb et al. 2004; Case et al. 2007); 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 (OTUs) as described below.

Metagenomes

As described above, "metagenome" means that DNA 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 the organism from which it came. Because metagenomes do not proceed through an amplification step, they do 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 subset metagenomic data to 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 microorganisms. This is proven to be a useful strategy, and several groups have built databases of core gene families as well as provided programs and/or web tools to phylogenetically analyze metagenomes subset to those core genes (Von Mering et al. 2007a; Wu and Eisen 2008a; Stark et al. 2010; Kembel et al. 2011; Darling et al. 2014). However, because of the variability of gene repertoire in microbes, this core gene set may be relatively small: even the largest collection of genes in these databases only recruits around 1% of a metagenome. At least some portion of the rest of the other approximately 99% of the metagenome can be taxonomically classified using one of the methods described below.

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

Whole Genomes

In addition to conventional genome sequencing for microbial 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 (http://www.fda.gov/Food/FoodScienceResearch/WholeGenomeSequencingProgramWGS/). These data may become more common for unculturable organisms as single-cell sequencing methods improve (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 (Emerson et al. 2012; Howe et al. 2012; Iverson et al. 2012; Pell 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.

Before proceeding, we note that phylogenetics for microbes differs in some important respects from, say, mammalian phylogenetics. Where phylogenetic research on mammals is converging on one or two possible basic structures for their evolution, much of the deep history of microbes remains obscure. Inference of this history is complicated by the fact that any tree-like signal in the deep evolutionary relationships between bacteria is restricted to a small set of so-called "core" genes (Bapteste et al. 2009; Leigh et al. 2011; Lang et al. 2013).

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 similar to bacteria in their gross morphology, 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 (Holt et al. 1984), although the actual revision process appears opaque. The GreenGenes 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 reassign sequence taxonomic labels so that they are concordant with a given rooted phylogenetic tree in a way that allows polyphyletic taxonomic 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 or EMBL sequence databases.

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 counts of DNA substrings of a given length, which are often called *k*-mer classifiers (Wang et al. 2007; Rosen et al. 2008). For those studies and others, the conventional metrics of classification performance such as precision and recall are applied to data simulated from known and taxonomically classified genomes. 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; as such it is somewhat phylogenetic in that it uses the structure of the taxonomic tree. Munch et al. (2008a, 2008b) infer taxonomic assignment by automatically retrieving sequences equipped with taxonomic information and building a tree on them along with an unknown sequence. Srinivasan et al. (2012) find that phylogenetic methods to do taxonomic classification can outperform composition-based techniques at least for certain taxonomic groups. 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. Other metagenomic classification techniques have been reviewed by Mande et al. (2012); interesting recent progress has been made by Lanzén et al. (2012), Koslicki et al. (2013), and Dröge et al. (2014). Treangen et al. (2013) report shorter run times and much higher accuracy (again, for taxonomic classification using reads simulated from full genomes) for such metagenomic classification when reads are assembled before they are classified.

The role of 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 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.

It is not straightforward to define a clear notion of optimality for OTU inference. Although in phylogenetics we would like to compare reconstructions to an object that is generally not knowable—the "true" historical phylogenetic tree—OTU inference has a variety of desirable outcomes, only some of which are knowable. Wang et al. (2013) divide performance measures into *external* measures, which compare an inferred clustering to some defined outcome, and *internal* measures, which give overall descriptive statistics on an inferred clustering. External measures applied to observed or simulated data include obtaining the same number of OTUs as taxonomic groups at some level (e.g., Edgar 2013), scoring deviation from a decomposition of a phylogenetic tree (e.g., Navlakha et al. 2010), or a given set of taxonomic classifications (e.g., Cai and Sun 2011). Internal measures evaluate the clustering in various ways without reference to a "true" clustering, with the general idea that within-cluster distances should be small compared to between-cluster distances.

There are many OTU inference methods with various speeds and strategies. Some methods proceed through a list of sequences and progressively add each either to an existing cluster or start a new cluster with that sequence, such as CD-HIT (Li and Godzik 2006) and USEARCH (Edgar 2010), whereas UPARSE (Edgar 2013) uses a similar strategy while also attempting to correct for sequencing error. White et al. (2010) show that different ways of doing this genre of heuristic clustering can result in very different results. Cai and Sun (2011) perform highly efficient clustering using a pseudometric-based partition tree, which can be thought of as a hierarchical clustering tree with a fixed set of internal node heights. Navlakha et al. (2010) take a semi-supervised approach in that input includes sequences along with a subset of sequences that are equipped with taxonomic classifications; the algorithm then groups all sequences (many in novel clusters) into clusters that have similar properties as the example taxonomic clusters. Wang et al. (2013) optimize a criterion of cluster modularity. Hao et al. (2011) use a Gaussian mixture model formulation for clustering to avoid fixed cutoff values, and Cheng et al. (2012) use a two-step process, first with a Dirichlet multinomial mixture on 3-mer profiles, and then a minimum description length criterion. Zhang et al. (2013) have developed a phylogenetic means to do species delimitation that scales to a relatively large number of sequencing reads and so can be used as an OTU clustering method.

The centrality of the OTU concept can be seen by the fact that the by-sample table of OTU observations (i.e., the matrix of counts with rows representing samples and columns representing OTUs) 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 nonoverlapping 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 the work of Sharpton et al. (2011) and Zhang et al. (2013), 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. For example, the most commonly used index is the Simpson index (Simpson 1949), which is simply the sum of the squared frequencies of the OTUs. The drawback of applying such a diversity metric to a collection of OTUs is that the large-scale structure of the diversity is lost, such that two closely related OTUs contribute as much to the diversity measure as two distant ones. Phylogenetic diversity (PD) metrics, which do take this overall diversity structure into account, are also used. However, whereas just about every one of the hundreds or thousands of 16S surveys applies an OTU-based alpha diversity estimate, only a few involve PD.

PD measures use the structure and branch lengths of a phylogenetic tree to quantify the diversity of a sample (Fig. 1). The (unweighted) PD of a set of taxa S in an unrooted phylogenetic tree is simply the total length of the branches that sit between taxa in S (Faith 1992). It quantifies the "amount of evolution" contained in the evolutionary history of those taxa. 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 nonphylogenetic diversity measures such as Simpson (1949) and Shannon

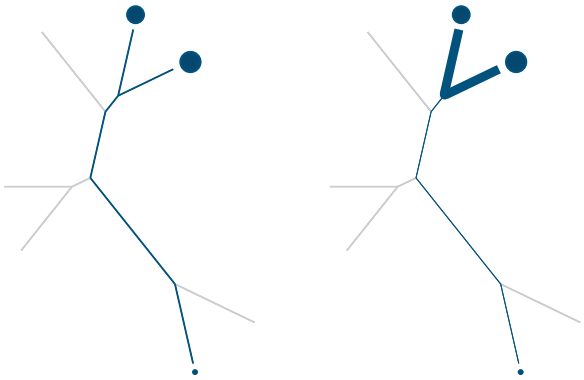

FIGURE 1. Unweighted phylogenetic diversity (PD, left) and an abundance-weighted PD measure (right), where taxa present in a sample are shown as circles and abundances are shown as the size of the circles. Unweighted PD takes the total length of branches sitting between tree tips represented in a sample. 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 total area of the edges. One such abundance-weighted measure simply takes the absolute value of the difference of the total read abundance on one side compared with the other.

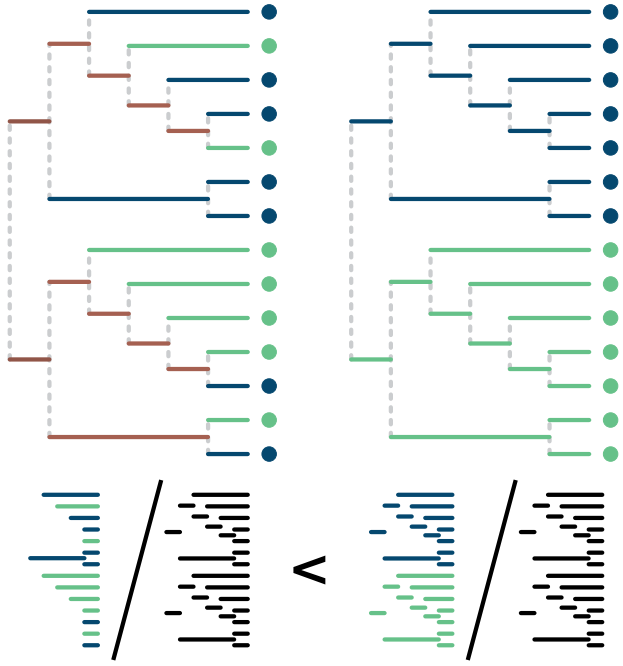

FIGURE 2. The UniFrac divergence measure (figure adapted from Lozupone and Knight 2005). Assume that the sequence data to build the phylogenetic tree derives from two samples: the light-shaded sample and the dark-shaded sample (green and blue in the online version). When the samples are interspersed across the tree (left tree), they have a smaller *fraction* of branch length that sits ancestral to clades that are *uniquely* composed of one sample or another, compared with when they are separate (right tree). The bottom pictorial equation shows the ratios of interest for UniFrac: the branch length unique to one sample divided by the total branch length. The ratio is smaller when the samples are interspersed (left) than they are when separate (right tree).

(1948) are commonly used in human microbiome studies, abundance weighted PD measures are not. 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 with abundant taxa. Such measures do exist (Rao 1982; Barker 2002; Allen et al. 2009; Chao et al. 2010; Vellend et al. 2010). Abundance-weighted measures commonly weight edges in proportion to abundance, but one can also construct "partially abundance weighted" measures by weighting edges by abundances transformed by a sublinear function. McCoy and Matsen (2013) have recently shown that such partially abundance weighted diversity measures do a good job of distinguishing between dysbiotic and "normal" states of the human microbiota; 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 subsampling (see discussion of rarefaction below).

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).

To calculate the unweighted UniFrac ("unique fraction") divergence between two samples of reads, one builds a global tree based on both samples and then calculates the fraction of the total tree length found in only one (i.e., the fraction of the tree unique to one) of the two 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 specific case of the earthmovers distance, and that the commonly used randomization procedure for significance estimation has a central limit theorem approximation. The earthmovers distance (Monge 1781; Villani 2003) between two probability distributions in this case can be defined using a physical analogy as the minimum amount of "work," defined as mass times distance, required to move probability mass in one distribution 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

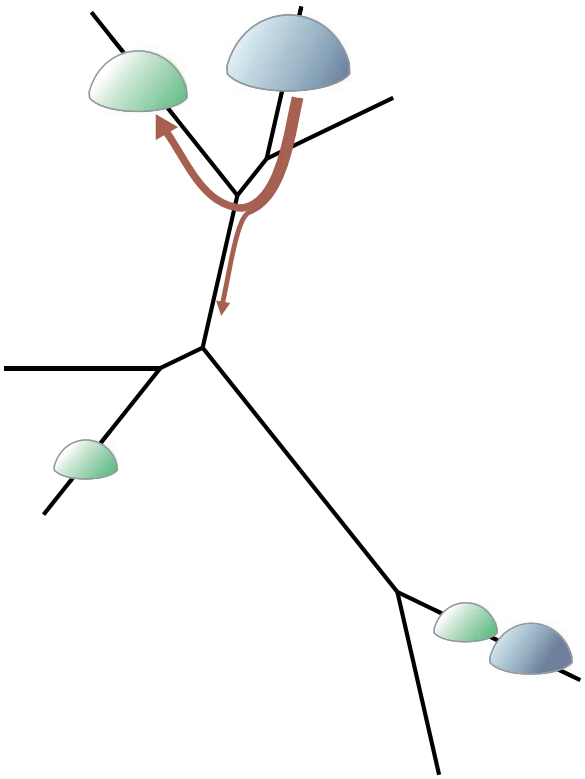

FIGURE 3. Part of a minimal mass movement to calculate the earth-mover's distance between two probability distributions on a phylogenetic tree. For this, each probability distribution is considered as a configuration of dirt piles (round bumps in the figure) on the tree, and the distance between two such dirt pile configurations is defined to be the minimum amount of physical "work" required to move the dirt in one configuration to the other.

than either unweighted or weighted UniFrac. There are clear connections between UniFrac and PD explored by Faith et al. (2009), who point out related measures in Faith's earlier work.

The most common way to use a distance matrix obtained from applying UniFrac to all pairs of samples 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 several 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 the 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, for which they use common distance measures in phylogenetics such as the subtree prune–regraft metric.

Phylogeny and Function

16S distance is frequently used as a proxy for a functional comparison between human microbiome samples. Indeed, researchers using UniFrac do not always think of their comparisons as being in terms of a single gene, but rather in terms of an abstracted measure of community function. 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. They have recovered more novel protein families using these phylogeny-based approaches than they would have using methods organized by selecting microbes to sequence based on their taxonomic labels. 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.

Horizontal Gene Transfer

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 sets are 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 concerning 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 microbiota of Japanese such that individuals with this resulting microbiota 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) found that in a mouse model, horizontal transfer between pathogenic bacteria is blocked by commensal bacteria except for periods of gut inflammation. Horizontal transfer of genes is inferred in these studies by finding highly similar subsequences in otherwise less related organisms.

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 sets are 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 (Morgan et al. 2013; Romiguier et al. 2013). 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 2009; Nawrocki et al. 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. Researchers do most phylogenetic inferences as part of a pipeline such as mothur (Schloss et al. 2009) which has incorporated the clear-cut code (Sheneman et al. 2006), or QIIME (Caporaso et al. 2010b), which wraps clear-cut, FastTree, and RAxML (Stamatakis 2006).

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 long been used as a means to build a phylogenetic tree sequentially (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 (Von Mering et al. 2007b; Monier et al. 2008; Wu and Eisen 2008b; Matsen et al. 2010; Stark 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. Brown and Truszkowski (2013) use locality-sensitive hashing to obtain placement more than two orders of magnitude faster than the *pplacer* program of Matsen et al. (2010).

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, whereas the 1,262,986 sequence GreenGenes 13_5 alignment 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 sequence 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 large phylogenies 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, with the idea that such an algorithm would retain the efficiency characteristics of placement algorithms while allowing the reference tree to change. For example, sequence data sets are 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. Izquierdo-Carrasco et al. (2014) have taken a step in this direction by developing an informatic framework that updates alignments and builds larger trees using previous smaller trees as starting points.

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 raises 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? Parks and Beiko (2012) have expanded the range of choices by defining community comparison metrics on phylogenetic split systems. An alternative would be to use collections of reconciled gene trees in the presence of gene deletion, transfer, and loss (e.g., Szöllősi et al. 2013a, 2013b).

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 the work by 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 might one 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.

Although developing community assembly models forms an important project for microbial ecology generally and human-associated microbial ecology in particular, phylogeny-aware methods could be further developed. One way to model microbial community assembly is to apply Hubbell's neutral theory (Costello et al. 2012; Fierer 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, where some initial work has been done by Lambert and Steel (2013) in another setting.

Diversity preservation is of interest for microbiota researchers like it is for eukaryotic organisms, but has not received the formalization and algorithmic treatment surrounding PD for larger organisms (Hartmann and Steel 2006; Pardi and Goldman 2007). Martin Blaser, in particular, has argued that changes in our microbiota 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 microbiota diversity (Blaser 2011). Because a child's initial microbiota is transmitted from the mother (reviewed in Funkhouser and Bordenstein 2013), there is a somewhat equivalent notion of microbiota extinction when the chain is interrupted via cesarean section and infant formula. In order to characterize extant diversity, Yatsunenko et al. (2012) have explicitly contrasted microbiota development in urban, forest-dwelling, and rural populations, whereas Tito et al. (2012) have endeavored to characterize the microbiota from ancient feces. How might phylogenetic methods be used in these preservation efforts?

There are indications of coevolution between microbiota and their hosts. Ochman et al. (2010) found identical tree topologies for primate and microbiome evolution. For the microbiota, they used maximum parsimony such that each column represented 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

microbiota 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?

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 poses difficult problems. The randomization of group membership commonly used in combination with UniFrac to determine significance does not have appropriate properties in the regime of incomplete sampling with nonindependent observations, which is certainly the correct regime for surveys and metagenomes. Such nonindependence can lead to incorrect rejection of the null hypothesis. Imagine, for example, that we have a random process as follows. Each sample from the process takes a random subset of "observations" from the leaves and then throws down some number of reads for each observation in that subset, with the number of reads having a mean significantly greater than 1. If the number of leaves is large compared with the number of sample observations, then two draws will always appear significantly different even though they are from the same underlying process. In trying to remedy such false-positive identification of differences, it becomes clear that even basic definitions pose a challenge: the question of whether two communities are the "same" and "different" probably needs to be approached from the perspective of ecosystem modeling.

For both alpha and beta diversity measurement, 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 subsampling to the number of reads in the lowest abundance sample (Schloss et al. 2009; Caporaso et al. 2010b). In addition to throwing away data, this normalization implicitly assumes a model whereby reads are sequenced independently of one another. This 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 mixture model for read counts, whereas McMurdie and Holmes (2014) claim better performance using a Gamma-Poisson mixture.

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 microbiota? At least for the next several years we can expect the research questions described above to continue to unfold. Future research projects will continue to bring deeper sequencing on more samples. The uBiome (http://ubiome.com/) and American Gut (http://americangut.org/) projects promise to bring gut microbiome sequencing to the average citizen for a low price. Comparative studies will continue to investigate what shapes and is shaped by the microbiota. 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 genetic sequences 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 will present themselves, although the specifics will change. For example, routine 16S sequencing is likely to be replaced

soon by Matrix-Assisted Laser Desorption Ionization–Time of Flight (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 (possibly including phylogenetics) 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), and this genome-scale data require evolutionary analysis methods to interpret it. 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 research on the human microbiome and will continue to do so.

Funding

This work was supported by National Institutes of health grant [R01-HG005966-01] and National Science Foundation grant [1223057].

Acknowledgements

I thank Olivier Gascuel for the opportunity to present on this subject during the 2013 MCEB 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 article. The article was greatly improved by thoughtful peer review from Frank (Andy) Anderson, Olivier Gascuel, Alexis Stamatakis, Frdric Delsuc, and an anonymous referee.

References

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

Aldous D. J., Krikun M. A., Popovic L. 2011. Five statistical questions about the tree of life. Syst. Biol. 60:318–328.

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

Anders S., Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol. 11:R106.

Ashelford K. E., Chuzhanova N. A., Fry J. C., Jones A. J., Weightman A. J. 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Appl. Environ. Microbiol. 71:7724–7736.

Ashelford K. E., Chuzhanova N. A., Fry J. C., Jones A. J., Weightman A. J. 2006. New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. Appl. Environ. Microbiol. 72:5734–5741.

Baker B. J., Comolli L. R., Dick G. J., Hauser L. J., Hyatt D., Dill B. D., Land M. L., VerBerkmoes N. C., Hettich R. L., Banfield J. F. 2010. Enigmatic, ultrasmall, uncultivated Archaea. Proc. Nat. Acad. Sci. 107:8806–8811.

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

Barker, G. 2002. Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation. Biol. J. Linnean Soc. 76:165–194.

Bazinet A., Cummings M. 2012. A comparative evaluation of sequence classification programs. BMC Bioinformatics 13:92.

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

Berger S., Krompass D., Stamatakis A. 2011. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. Syst. Biol. 60:291–302.

Blaser M. 2011. Antibiotic overuse: stop the killing of beneficial bacteria. Nature 476:393–394.

Boon E., Meehan C. J., Whidden C., Wong D. H.-J., Langille M. G. I., Beiko R. G. 2013. Interactions in the microbiome: communities of organisms and communities of genes. FEMS Microbiol. Rev. 38:90–118.

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

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

Brown, D. G., Truszkowski J. 2013. LSHPlace: fast phylogenetic placement using locality-sensitive hashing. In: Luay Nakhleh, Noah Rosenberg, and Tandy Warnow, editors. 18th Pacific Symposium on Biocomputing. Singapore: World Scientific. p. 310–319.

Cai Y., Sun Y. 2011. ESPRIT-Tree: hierarchical clustering analysis of millions of 16S rRNA pyrosequences in quasilinear computational time. Nucleic Acids Res. 39:e95.

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

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

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

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

Caro-Quintero A., Konstantinidis K. T. 2012. Bacterial species may exist, metagenomics reveal. Environ. Microbiol. 14:347–355.

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

Chao A., Chiu C., Jost L. 2010. Phylogenetic diversity measures based on Hill numbers. Philos. Trans. R. Soc. B Biol. Sci. 365:3599–3609.

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

Chen T., Yu W., Izard J., Baranova O., Lakshmanan A., Dewhirst F. 2010. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database doi:10.1093/database/baq013.

Cheng L., Walker A. W., Corander J. 2012. Bayesian estimation of bacterial community composition from 454 sequencing data. Nucleic Acids Res. 40:5240–5249.

Cho I., Blaser M. J. 2012. The human microbiome: at the interface of health and disease. Nat. Rev. Genet. 13:260–270.

Claesson M. J., Jeffery I. B., Conde S., Power S. E., O'Connor E. M., Cusack S., Harris H. M. B., Coakley M., Lakshminarayanan B., O'Sullivan O., Fitzgerald G. F., Deane J., O'Connor M., Harnedy N., O'Connor K., O'Mahony D., van Sinderen D., Wallace M., Brennan L., Stanton C., Marchesi J. R., Fitzgerald A. P., Shanahan F., Hill C., Ross R. P., O'Toole P. W. 2012. Gut microbiota composition correlates with diet and health in the elderly. Nature 488:178–184.

Clark A. E., Kaleta E. J., Arora A., Wolk D. M. 2013. Matrix-assisted laser desorption ionization–time of flight mass spectrometry: a fundamental shift in the routine practice of clinical microbiology. Clin. Microbiol. Rev. 26:547–603.

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

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

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

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

Darling A. E., Jospin G., Lowe E., Matsen 4th, F. A., Bik H. M., Eisen J. A. 2014. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ 2:e243.

Degnan P. H., Ochman H. 2011. Illumina-based analysis of microbial community diversity. ISME J. 6:183–194.

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

Delsuc F., Metcalf J. L., Wegener Parfrey L., Song S. J., González A., Knight R. 2013. Convergence of gut microbiomes in myrmecophagous mammals. Mol. Ecol. 23:1301–1317.

DeSantis T., Hugenholtz P., Larsen N., Rojas M., Brodie E., Keller K., Huber T., Dalevi D., Hu P., Andersen G. 2006a. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72:5069.

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

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

Dethlefsen L., Relman D. A. 2011. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. Proc. Nat. Acad. Sci. 108:4554–4561.

Didelot X., Bowden R., Wilson D. J., Peto T. E., Crook D. W. 2012. Transforming clinical microbiology with bacterial genome sequencing. Nat. Rev. Genet. 13:601–612.

Dröge J., Gregor I., McHardy A. C. 2014. Taxator-tk: fast and precise taxonomic assignment of metagenomes by approximating evolutionary neighborhoods. http://arxiv.org/abs/1404.1029.

Eddy S. 1998. Profile hidden Markov models. Bioinformatics 14:755–763.

Edgar R. C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

Edgar R. C. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat. Methods 10:996–998.

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

Emerson J. B., Thomas B. C., Andrade K., Allen E. E., Heidelberg K. B., Banfield J. F. 2012. Metagenomic assembly reveals dynamic viral populations in hypersaline systems. Appl. Environ. Microbiol. 78:6309–6320.

Evans J., Sheneman L., Foster J. 2006. Relaxed neighbor joining: a fast distance-based phylogenetic tree construction method. J. Mol. Evol. 62:785–792.

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

Faith D. 1992. Conservation evaluation and phylogenetic diversity. Biol. Conser. 61:1–10.

Faith D. P., Lozupone C. A., Nipperess D., Knight R. 2009. The cladistic basis for the phylogenetic diversity measure links evolutionary features to environmental gradients. Int. J. Mol. Sci. 10:4723–4741.

Felsenstein J. 1985. Phylogenies and the comparative method. Am. Nat. 125:1–15.

Fierer N., Ferrenberg S., Flores G. E., González A., Kueneman J., Legg T., Lynch R. C., McDonald D., Mihaljevic J. R., O'Neill S. P., Rhodes M. E., Song S. J., Walters W. A. 2012. From animalcules to an ecosystem: application of ecological concepts to the human microbiome. Ann. Rev. Ecol. Evol. Syst. 43:137–155.

Findley K., Oh J., Yang J., Conlan S., Deming C., Meyer J. A., Schoenfeld D., Nomicos E., Park M., NIH Intramural Sequencing Center Comparative Sequencing Program, Kong H. H., Segre J. A. 2013. Topographic diversity of fungal and bacterial communities in human skin. Nature 498:367–370.

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

Forey P. 2001. The PhyloCode: description and commentary. Bull. Zool. Nomencl. 58:81–96.

Fox G. E., Pechman K. R., Woese C. R. 1977. Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. Int. J. Syst. Bacteriol. 27:44–57.

Funkhouser L. J., Bordenstein S. R. 2013. Mom knows best: the universality of maternal microbial transmission. PloS Biol. 11:e1001631.

Greenblum S., Turnbaugh P. J., Borenstein E. 2012. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc. Nat. Acad. Sci. 109:594–599.

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

Griffen A., Beall C., Firestone N., Gross E., DiFranco J., Hardman J., Vriesendorp B., Faust R., Janies D., Leys E. 2011. CORE: a phylogenetically-curated 16S rDNA database of the core oral microbiome. PloS ONE 6:e19051.

Holt, J. G., Krieg N. R., Sneath P. 1984. Bergey's manual of systematic bacteriology. Baltimore: Williams and Wilkins.

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

Hao X., Jiang R., Chen T. 2011. Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics 27:611–618.
Hartmann K., Steel M. 2006. Maximizing phylogenetic diversity in biodiversity conservation: greedy solutions to the Noah's Ark problem. Syst. Biol. 55:644–651.
Hehemann J.-H., Correc G., Barbeyron T., Helbert W., Czjzek M., Michel G. 2010. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature 464:908–912.
Hewitt K. M., Mannino F. L., Gonzalez A., Chase J. H., Caporaso J. G., Knight R., Kelley S. T. 2013. Bacterial diversity in two neonatal intensive care units (NICUs). PloS ONE 8:e54703.
Hoffmann C., Dollive S., Grunberg S., Chen J., Li H., Wu G. D., Lewis J. D., Bushman F. D. 2013. Archaea and Fungi of the human gut microbiome: correlations with diet and bacterial residents. PloS ONE 8:e66019.
Holmes I., Harris K., Quince C. 2012. Dirichlet multinomial mixtures: generative models for microbial metagenomics. PloS ONE 7:e30126.
Hooper L. V., Littman D. R., Macpherson A. J. 2012. Interactions between the microbiota and the immune system. Science 336:1268–1273.
Howe A. C., Jansson J., Malfatti S. A., Tringe S. G., Tiedje J. M., Brown C. T. 2012. Assembling large, complex environmental metagenomes. arXiv preprint arXiv:1212.2832.
Hugenholtz P., Huber T. 2003. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. Int. J. Syst. Evol. Microbiol. 53:289–293.
Huson D. H., Auch A. F., Qi J., Schuster S. C. 2007. Megan analysis of metagenomic data. Genome Res. 17:377–386.
Huson D. H., Mitra S., Ruscheweyh H.-J., Weber N., Schuster S. C. 2011. Integrative analysis of environmental sequences using MEGAN4. Genome Res. 21:1552–1560.
Iverson V., Morris R. M., Frazar C. D., Berthiaume C. T., Morales R. L., Armbrust E. V. 2012. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. Science 335:587–590.
Izquierdo-Carrasco F., Cazes J., Smith S. A., Stamatakis A. 2014 PUmPER: phylogenies updated perpetually. Bioinformatics 30:1476–1477.
Jaccard P. 1908. Nouvelles recherches sur la distribution florale. Bull. Soc. Vaudoise Sci. Nat. 44:223–270.
Jakobsson H. E., Jernberg C., Andersson A. F., Sjölund-Karlsson M., Jansson J. K., Engstrand L. 2010. Short-term antibiotic treatment has differing long-term impacts on the human throat and gut microbiome. PloS ONE 5:e9836.
Jernberg C., Löfmark S., Edlund C., Jansson J. K. 2007. Long-term ecological impacts of antibiotic administration on the human intestinal microbiota. The ISME J. 1:56–66.
Kalisky T., Quake S. R. 2011. Single-cell genomics. Nat. Methods 8:311–314.
Kau A. L., Ahern P. P., Griffin N. W., Goodman A. L., Gordon J. I. 2011. Human nutrition, the gut microbiome and the immune system. Nature 474:327–336.
Kembel S. W., Eisen J. A., Pollard K. S., Green J. L. 2011. The phylogenetic diversity of metagenomes. PloS ONE 6:e23214.
Kembel S. W., Wu M., Eisen J. A., Green J. L. 2012. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. PloS Comput. Biol. 8:e1002743.
Klappenbach J. A., Dunbar J. M., Schmidt T. M. 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl. Environ. Microbiol. 66:1328–1333.
Klappenbach J. A., Saxman P. R., Cole J. R., Schmidt T. M. 2001. rrndb: the ribosomal RNA operon copy number database. Nucleic Acids Res. 29:181–184.
Kluge A. G., Farris J. S. 1969. Quantitative phyletics and the evolution of anurans. Syst. Biol. 18:1–32.
Koren O., Goodrich J. K., Cullender T. C., Spor A., Laitinen K., Kling Bäckhed H., Gonzalez A., Werner J. J., Angenent L. T., Knight R., Bäckhed F., Isolauri E., Salminen S., Ley R. E. 2012. Host remodeling of the gut microbiome and metabolic changes during pregnancy. Cell 150:470–480.
Köser C. U., Holden M. T., Ellington M. J., Cartwright E. J., Brown N. M., Ogilvy-Stuart A. L., Hsu L. Y., Chewapreecha C., Croucher N. J., Harris S. R., Sanders M., Enright M. C., Dougan G., Bentley S. D., Parkhill J., Fraser L. J., Betley J. R., Schulz-Trieglaff O. B., Smith G. P., Peacock S. J. 2012. Rapid whole-genome sequencing for investigation of a neonatal MRSA outbreak. N. Engl. J. Med. 366:2267–2275.
Koslicki D., Foucart S., Rosen G. 2013. Quikr: a method for rapid reconstruction of bacterial communities via compressive sensing. Bioinformatics 29:2096–2102.
Kuczynski J., Liu Z., Lozupone C., McDonald D., Fierer N., Knight R. 2010. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nat. Methods 7:813–819.
La Rosa P. S., Brooks J. P., Deych E., Boone E. L., Edwards D. J., Wang Q., Sodergren E., Weinstock G., Shannon W. D. 2012a. Hypothesis testing and power calculations for taxonomic-based human microbiome data. PloS ONE 7:e52078.
La Rosa P. S., Shands B., Deych E., Zhou Y., Sodergren E., Weinstock G., Shannon W. D. 2012b. Statistical object data analysis of taxonomic trees from human microbiome data. PloS ONE 7:e48996.
Lambert A., Steel M. 2013. Predicting the loss of phylogenetic diversity under non-stationary diversification models. arXiv preprint arXiv:1306.2710.
Lane D. J., Pace B., Olsen G. J., Stahl D. A., Sogin M. L., Pace N. R. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc. Nat. Acad. Sci. 82:6955–6959.
Lang J. M., Darling A. E., Eisen J. A. 2013. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. PloS ONE 8:e62510.
Langille M. G., Zaneveld J., Caporaso J. G., McDonald D., Knights D., Reyes J. A., Clemente J. C., Burkepile D. E., Thurber R. L. V., Knight R., Beiko R. G., Huttenhower C. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat. Biotechnol. 31:814–821.
Lanzén A., Jørgensen S. L., Huson D. H., Gorfer M., Grindhaug S. H., Jonassen I., Øvreås L., Urich T. 2012. CREST classification resources for environmental sequence tags. PloS ONE 7:e49334.
Leigh J. W., Schliep K., Lopez P., Bapteste E. 2011. Let them fall where they may: congruence analysis in massive phylogenetically messy data sets. Mol. Biol. Evol. 28:2773–2785.
Ley R., Turnbaugh P., Klein S., Gordon J. 2006. Microbial ecology: human gut microbes associated with obesity. Nature 444: 1022–1023.
Li W., Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659.
Li F., Hullar M. A., Schwarz Y., Lampe J. W. 2009. Human gut bacterial communities are altered by addition of cruciferous vegetables to a controlled fruit-and vegetable-free diet. J. Nutrition 139:1685–1691.
Liu Z., DeSantis T. Z., Andersen G. L., Knight R. 2008. Accurate taxonomy assignments from 16s rrna sequences produced by highly parallel pyrosequencers. Nucleic Acids Res. 36:e120–e120.
Lozupone C., Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. 71:8228.
Lozupone C. A., Knight R. 2007. Global patterns in bacterial diversity. Proc. Nat. Acad. Sci. 104:11436–11440.
Lozupone C. A., Hamady M., Kelley S. T., Knight R. 2007. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl. Environ. Microbiol. 73:1576–85.
Lozupone C. A., Stombaugh J., Gonzalez A., Ackermann G., Wendel D., Vázquez-Baeza Y., Jansson J. K., Gordon J. I., Knight R. 2013. Meta-analyses of studies of the human microbiota. Genome Res. 23:1704–1714.
Ludwig W., Strunk O., Westram R., Richter L., Meier H., Yadhukumar Buchner A., Lai T., Steppi S., Jobb G., Frster, W., Brettske I., Gerber S., Ginhart A. W., Gross O., Grumann S., Hermann S., Jost R., Knig A., Liss T., Lmann R., May M., Nonhoff B., Reichel B., Strehlow R., Stamatakis A., Stuckmann N., Vilbig A., Lenke M., Ludwig T., Bode A., Schleifer K. 2004. ARB: a software environment for sequence data. Nucleic Acids Res. 32:1363.
Mande S. S., Mohammed M. H., Ghosh T. S. 2012. Classification of metagenomic sequences: methods and challenges. Brief. Bioinformatics 13:669–681.

Matsen F. A., Evans S. N. 2013. Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. PloS ONE 8:e56859.

Matsen F. A., Gallagher A. 2012. Reconciling taxonomy and phylogenetic inference: formalism and algorithms for describing discord and inferring taxonomic roots. Algorithms Mol. Biol. 7:8.

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

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

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

McDonald D., Clemente J. C., Kuczynski J., Rideout J. R., Stombaugh J., Wendel D., Wilke A., Huse S., Hufnagle J., Meyer F., Knight R., Caporaso J. G. 2012. The biological observation matrix (BIOM) format or: how i learned to stop worrying and love the ome-ome. GigaScience 1:7.

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

McMurdie P. J., Holmes S. 2014. Waste not, want not: why rarefying microbiome data is inadmissible. PloS Comput. Biol. 10:e1003531.

McNabb A., Eisler D., Adie K., Amos M., Rodrigues M., Stephens G., Black W. A., Isaac-Renton J. 2004. 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. 42:3000–3011.

Methé B. A., Nelson K. E., Pop M., Creasy H. H., Giglio M. G., Huttenhower C., Gevers D., Petrosino J. F., Abubucker S., Badger J. H., Chinwalla A. T., Earl A. M., FitzGerald M. G., Fulton R. S., Hallsworth-Pepin K., Lobos E. A., Madupu R., Magrini V., Martin J. C., Mitreva M., Muzny D. M., Sodergren E. J., Versalovic J., Wollam A. M., Worley K. C., Wortman J. R., Young S. K., Zeng Q., Aagaard K. M., Abolude O. O., Allen-Vercoe E., Alm E. J., Alvarado L., Andersen G. L., Anderson S., Appelbaum E., Arachchi H. M., Armitage G., Arze C. A., Ayvaz T., Baker C. C., Begg L., Belachew T., Bhonagiri V., Bihan M., Blaser M. J., Bloom T., Bonazzi V. R., Brooks P., Buck G. A., Buhay C. J., Busam D. A., Campbell J. L., Canon S. R., Cantarel B. L., Chain P. S., Chen I.-M. A., Chen L., Chhibba S., Chu K., Ciulla D. M., Clemente J. C., Clifton S. W., Conlan S., Crabtree J., Cutting M. A., Davidovics N. J., Davis C. C., DeSantis T. Z., Deal C., Delehaunty K. D., Dewhirst F. E., Deych E., Ding Y., Dooling D. J., Dugan S. P., Dunne W. M., Durkin S. A., Edgar R. C., Erlich R. L., Farmer C. N., Farrell R. M., Faust K., Feldgarden M., Felix V. M., Fisher S., Fodor A. A., Forney L., Foster L., Di Francesco V, Friedman J, Friedrich D. C., Fronick C. C., Fulton L. L., Gao H., Garcia, N., Giannoukos G., Giblin C., Giovanni M. Y., Goldberg J. M., Goll J., Gonzalez A., Griggs A., Gujja, S., Haas B. J., Hamilton H. A., Harris, E. L., Hepburn T. A., Herter, B., Hoffmann, D. E., Holder, M. E., Howarth C., Huang K. H., Huse S. M., Izard J., Jansson J. K., Jiang H., Jordan C., Joshi V., Katancik J. A., Keitel W. A., Kelley S. T., Kells C., Kinder-Haake S., King N. B., Knight R., Knights D., Kong H. H., Koren O., Koren S., Kota K. C., Kovar C. L., Kyrpides N. C., La Rosa P. S., Lee S. L., Lemon K. P., Lennon N., Lewis C. M., Lewis L., Ley R. E., Li K., Liolios K., Liu B., Liu Y., Lo, C.-C., Lozupone C. A., Lunsford R. D., Madden T., Mahurkar A. A., Mannon P. J., Mardis, E. R., Markowitz V. M., Mavrommatis K., McCorrison J. M., McDonald, D., McEwen J., McGuire A. L., McInnes P., Mehta T., Mihindukulasuriya K. A., Miller J. R., Minx, P. J., Newsham I., Nusbaum C., OLaughlin M., Orvis J., Pagani I., Palaniappan K., Patel S. M., Pearson M., Peterson J., Podar M., Pohl C., Pollard, K. S., Priest, M. E., Proctor L. M., Qin X., Raes J., Ravel J., Reid J. G., Rho M., Rhodes R., Riehle K. P., Rivera M. C., Rodriguez-Mueller B., Rogers Y.-H., Ross M. C., Russ C., Sanka R. K., Sankar P., Fah Sathirapongsasuti J., Schloss J. A., Schloss P. D., Schmidt T. M., Scholz M., Schriml L., Schubert A. M., Segata N., Segre J. A., Shannon W. D., Sharp R. R., Sharpton T. J., Shenoy N., Sheth N. U., Simone G. A., Singh I., Smillie C. S., Sobel J. D., Sommer D. D., Spicer P., Sutton G. G., Sykes S. M., Tabbaa D. G., Thiagarajan M., Tomlinson C. M., Torralba M., Treangen T. J., Truty R. M., Vishnivetskaya T. A., Walker J., Wang L., Wang Z., Ward D. V., Warren W., Watson M. A., Wellington C., Wetterstrand K. A., White J. R., Wilczek-Boney K., Qing Wu Y., Wylie K. M., Wylie T., Yandava C., Ye L., Ye Y., Yooseph S., Youmans B. P., Zhang L., Zhou Y., Zhu Y., Zoloth L., Zucker J. D., Birren B. W., Gibbs R. A., Highlander S. K., Weinstock G. M., Wilson R. K., White O. 2012. A framework for human microbiome research. Nature 486:215–221.

Minot S., Bryson A., Chehoud C., Wu G. D., Lewis J. D., Bushman F. D. 2013. Rapid evolution of the human gut virome. Proc. Nat. Acad. Sci. 110:12450–12455.

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

Mirarab S., Nguyen N., Warnow T. 2012. SEPP: SATé-enabled phylogenetic placement. In: James A. Foster, Jason Moore, Jack Gilbert, John Bunge, editors. Pacific Symposium on Biocomputing. World Scientific. p. 247–258. doi:10.1142/9789814366496_0024.

Monge G. 1781. Mémoire sur la théorie des déblais et des remblais. Histoire de l'Académie Royale des Sciences de Paris, Paris: l'Imprimerie Royale.

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

Mooers A. O., Heard S. B. 1997. Inferring evolutionary process from phylogenetic tree shape. Q. Rev. Biol. 72:31–54.

Moran S., Snir S. 2008. Convex recolorings of strings and trees: definitions, hardness results and algorithms. J. Computer Syst. Sci. 74:850–869.

Morgan J. L., Darling A. E., Eisen J. A. 2010. Metagenomic sequencing of an in vitro-simulated microbial community. PloS ONE 5:e10209.

Morgan C. C., Foster P. G., Webb A. E., Pisani D., McInerney J. O., OConnell M. J. 2013. Heterogeneous models place the root of the placental mammal phylogeny. Mol. Biol. Evol. 30:2145–2156.

Munch K., Boomsma W., Huelsenbeck J. P., Willerslev E., Nielsen R. 2008a. Statistical assignment of DNA sequences using Bayesian phylogenetics. Syst. Biol. 57:750–757.

Munch K., Boomsma W., Willerslev E., Nielsen R. 2008b. Fast phylogenetic DNA barcoding. Philos. Trans. R. Soc. B Biol. Sci. 363:3997–4002.

Navlakha S., White J., Nagarajan N., Pop M., Kingsford C. 2010. Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information. J. Comput. Biol. 17:503–516.

Nawrocki E. P. 2009. Structural RNA homology search and alignment using covariance models. [Ph.D. thesis]. Washington University. Advisor: Sean R Eddy.

Nawrocki E., Kolbe D., Eddy S. 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics 25:1335–1337.

Nipperess D. A., Matsen 4th, F. A. 2013. The mean and variance of phylogenetic diversity under rarefaction. Methods Ecol. Evol. 4:566–572.

Ochman H., Worobey M., Kuo C.-H., Ndjango J.-B. N., Peeters M., Hahn B. H., Hugenholtz P. 2010. Evolutionary relationships of wild hominids recapitulated by gut microbial communities. PloS Biol. 8:e1000546.

O'Dwyer J. P., Kembel S. W., Green J. L. 2012. Phylogenetic diversity theory sheds light on the structure of microbial communities. PloS Comput. Biol. 8:e1002832.

Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proc. R. Soc. Lond. Ser. B Biol. Sci. 255:37–45.

Paradis E., Claude J., Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Pardi F., Goldman N. 2007. Resource-aware taxon selection for maximizing phylogenetic diversity. Syst. Biol. 56:431–444.

Parks D. H., Beiko R. G. 2012. Measuring community similarity with phylogenetic networks. Mol. Biol. Evol. 29:3947–3958.

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

Paulson J. N., Stine O. C., Bravo H. C., Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. Nat. Methods 10:1200–1202.

Pell J., Hintze A., Canino-Koning R., Howe A., Tiedje J. M., Brown C. T. 2012. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. Proc. Nat. Acad. Sci. 109:13272–13277.

Phillips C., Phelan G., Dowd S., MCdonough M., Ferguson A., Delton Hanson J., Siles L., Ordóñez-garza N., San Francisco M., Baker R. 2012. Microbiome analysis among bats describes influences of host phylogeny, life history, physiology and geography. Mol. Ecol. 21:2617–2627.

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

Polz M. F., Cavanaugh C. M. 1998. Bias in template-to-product ratios in multitemplate PCR. Appl. Environ. Microbiol. 64:3724–3730.

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

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

Price M., Dehal P., Arkin A. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS ONE 5:e9490.

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

Pruesse E., Peplies J., Glöckner F. O. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28:1823–1829.

Purdom E. 2008. Analyzing data with graphs: metagenomic data and the phylogenetic tree. UC Berkeley Stat. Tech. Rep. 766:1–22.

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

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

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

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

Rao C. 1982. Diversity and dissimilarity coefficients: a unified approach. Theor. Popul. Biol. 21:24–43.

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

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

Rogers G. B., Hoffman L. R., Carroll M. P., Bruce K. D. 2013. Interpreting infective microbiota: the importance of an ecological perspective. Trends Microbiol. 21:271–276.

Romiguier J., Ranwez V., Delsuc F., Galtier N., Douzery E. J. 2013. Less is more in mammalian phylogenomics: at-rich genes minimize tree conflicts and unravel the root of placental mammals. Mol. Biol. Evol. 30:2134–2144.

Rosen G., Garbarine E., Caseiro D., Polikar R., Sokhansanj B. 2008. Metagenome fragment classification using N-mer frequency profiles. Adv. Bioinformatics 2008:205969.

Schloss P. D. 2008. Evaluating different approaches that test whether microbial communities have the same structure. ISME J. 2:265–275.

Schloss P. D., Gevers D., Westcott S. L. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PloS ONE 6:e27310.

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

Segata N., Waldron L., Ballarini A., Narasimhan V., Jousson O., Huttenhower C. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat. Methods 9:811–814.

Shannon C. 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27:379–423.

Sharpton T. J., Riesenfeld S. J., Kembel S. W., Ladau J., O'Dwyer J. P., Green J. L., Eisen J. A., Pollard K. S. 2011. PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. PloS Comput. Biol. 7:e1001061.

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

Simpson E. 1949. Measurement of diversity. Nature 163:688.

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

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

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

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

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Stark M., Berger S., Stamatakis A., von Mering C. 2010. MLTreeMap-accurate maximum likelihood placement of environmental dna sequences into taxonomic and functional reference phylogenies. BMC Genomics 11:461.

Stecher B., Denzler R., Maier L., Bernet F., Sanders M. J., Pickard D. J., Barthel M., Westendorf A. M., Krogfelt K. A., Walker A. W., Ackermann M., Dobrindt U., Thomson N. R., Hardt W.-D. 2012. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. Proc. Nat. Acad. Sci. 109:1269–1274.

Steel M., Rodrigo A. 2008. Maximum likelihood supertrees. Syst. Biol. 57:243–250.

Suzuki M. T., Giovannoni S. J. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl. Environ. Microbiol. 62:625–630.

Szöllősi G. J., Rosikiewicz W., Boussau B., Tannier E., Daubin V. 2013a. Efficient exploration of the space of reconciled gene trees. Syst. Biol. 62:901–912.

Szöllősi G. J., Tannier E., Lartillot N., Daubin V. 2013b. Lateral gene transfer from the dead. Syst. Biol. 62:386–397.

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

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

Turnbaugh P. J., Ley R. E., Mahowald M. A., Magrini V., Mardis E. R., Gordon J. I. 2006. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature 444:1027–1131.

Turnbaugh P. J., Hamady M., Yatsunenko T., Cantarel B. L., Duncan A., Ley R. E., Sogin M. L., Jones W. J., Roe B. A., Affourtit J. P., Egholm M., Henrissat B., Heath A. C., Knight R., Gordon J. I. 2008. A core gut microbiome in obese and lean twins. Nature 457:480–484.

Vellend M., Cornwell W. K., Magnuson-Ford K., Mooers A. 2010. In: Magurran A. E., McGill B. J., editors. Biological Diversity: Frontiers in Measurement and Assessment. Oxford: Oxford University Press.

Villani C. 2003. Topics in Optimal Transportation. Providence: American Mathematical Society.

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

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

Wang Q., Garrity G., Tiedje J., Cole J. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73:5261–5267.

Wang X., Yao J., Sun Y., Mai V. 2013. M-pick, a modularity-based method for OTU picking of 16S rRNA sequences. BMC Bioinformatics 14:43.

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

Woese C. R., Fox G. E. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc. Nat. Acad. Sci. 74:5088–5090.

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

Wu M., Eisen J. 2008a. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 9:1–11.

Wu M., Eisen J. A. 2008b. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 9:R151.

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

Yang Z., Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. Proc. Nat. Acad. Sci. 107:9264–9269.

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

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

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

Zhang J., Kapli P., Pavlidis P., Stamatakis A. 2013. A general species delimitation method with applications to phylogenetic placements. Bioinformatics 29:2869–2876.

Zhao L. 2013. The gut microbiota and obesity: From correlation to causality. Nat. Rev. Microbiol. 11:639–647.

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