CHAPTER 1

MICROBIOME ANALYSIS: STATE-OF-THE-ART AND FUTURE TRENDS

MITCH FERNANDEZ[1], VANESSA AGUIAR-PULIDO[1], JUAN RIVEROS[1], WENRUI HUANG[1], JONATHAN SEGAL[2], ERLIANG ZENG[3], MICHAEL CAMPOS[4], KALAI MATHEE[2], GIRI NARASIMHAN[1]

[1]Bioinformatics Research Group (BioRG), School of Computing and Information Sciences, Florida International University, Miami, FL 33199, USA

[2]Herbert Wertheim College of Medicine, Florida International University, Miami, FL 33199, USA

[3]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA

[4]Miller School of Medicine, University of Miami, Miami, FL 33136, USA

1.1 Introduction

Microbes form complex heterogeneous interacting communities, whether in the environment or in specific niches within humans and other host organisms [13]. *Metagenomics* approaches have been used to study the composition and dynamics of such microbial communities. Distinct communities of bacteria are present at different sites of the human body, and changes in their structure have strong implications for human health. The Human Microbiome Project (HMP) focuses on the study of microbial communities that inhabit the healthy human body [50, 61]. It is known that 90% of the cells in the human body are bacterial cells and that bacterial communities play such critical roles as aiding in the digestion of food, synthesizing essential vitamins, and assisting the immune system in fending off pathogenic invaders. Human microbiome studies have revealed that diseases and disorders are strongly correlated

In I. I. Măndoiu and A. Zelikovsky (Eds.), Computational Methods for Next Generation Sequencing Data Analysis.

with changes in microbial community profiles [60, 43, 52]. These studies have also demonstrated that microbial community structure in five niches of the human body (gut, mouth, airways, urogenital, and skin) are quite distinct, and appear to transcend gender, age, and ethnicity [32].

It is possible to employ classical microbiological methods for the analysis of microbiomes, by which one could culture individual taxa and then use specific protocols developed to identify them. However, classical approaches will always be limited by (a) the ability to culture – it is broadly believed that most bacterial taxa are not readily culturable by traditional culturing techniques (although more extensive cultivation approaches have yielded considerably higher percentages of cultured taxa, as shown by the work of M. Surette and colleagues [55]); (b) time it takes to culture – it could take several weeks for a culture to grow; and (c) the potentially staggering number and diversity of microbial taxa present in the sample.

Molecular approaches to microbiome studies involve extracting microbial DNA from a sample followed by a process of determining the profile of the microbial community present. A reasonable approach is that of PCR amplifications of marker genes, such as the gene for 16S rRNA, followed by one of a variety of approaches to investigate length heteoregeneity of variable regions of the microbial genomes [16]. Here the limitations are the result of the fact that length heterogeneity of specific regions of certain marker genes in the bacterial world is considerably smaller than the number of taxa.

Exploiting sequence heterogeneity is clearly a more informative approach than using length heterogeneity. More recent methods involve the use of next generation sequencing. The DNA is first extracted and a portion of the marker gene is amplified with appropriate PCR primers. Then, next generation sequencing techniques generate a large number of reads from the amplicons. Bioinformatics tools are finally used to claissify the reads by the different taxa from which they have arisen.

While this will be discussed in detail later, the limitations of these methods are the following: (a) the extent of our databases and the sequences cataloged in them limits the bacterial taxa that can be successfully identified; (b) next generation sequencing techniques produce short reads that may or may not be from distinguishable regions of the individual taxon; and (c) the software classifier used to classify individual reads may produce inaccurate results, in part because the independence assumptions inherent in existing classifiers may not be biologically valid.

Microbiome studies have largely been applied on phylogenetic marker genes, most often the gene for 16S rRNA (for example, see [9, 28, 5, 26, 38, 21]), but others include 23S rRNA, *recA*, DNA *gyrA*, and many more [57, 39, 70, 37]. There is no consensus on the ideal marker gene, and the popularity of 16S rRNA is simply due to its use in standard reference manuals [37]. However, recent studies have taken a different approach by employing whole genome sequencing [32, 48, 10]. The advantage of marker genes is the availability of large databases [11, 51, 15] to aid in classification of reads. The advantage of the whole genome approach is that it becomes possible to obtain functional profiles of the microbial community by investigating the functional annotations of the genes to which the reads map. We discuss later why this is potentially of greater interest than the marker gene approach.

Regardless of the method of identifying bacterial taxa or the functional elements present, the next step is to leverage the information to understand the difference between samples. Such measures include $\alpha-$ and $\beta-$diversity indices, which have been shown to be useful in distinguishing between samples [18, 45]. Specific indices for calculating these types of diversity include the Shannon diversity index, popular for its dealing with the uncertainty in predicting unobserved species [54]; the Jaccard index, which measures similarity between samples by the proportion of shared clusters or species between them [33]; and the popular inverse-Simpson index, which estimates diversity by considering the probability that a randomly selected individual from a sample has already been observed [56]. It should be noted that there are situations where well-established diversity indices are not sufficient to differentiate between groups of samples. In recent studies, it has been observed that the diversity indices for the airways microbiome of smoking and non-smoking subjects are not distinguishable, even though these studies have readily identified individual bacterial genera that tend to favor the airways of smokers over non-smokers or former smokers [19, 47].

More advanced techniques to understand and differentiate microbial community profiles have been investigated. It is clear that metagenomic studies have to go further and dig deeper to uncover interesting features of microbial communities. The next set of promising investigations have to focus on understanding the structure of microbial communities and their interactions in any environmental niche. One of the greatest challenges in understanding human health is uncovering the large number of complex interactions that occur within the microbial community, and between the community and the human host. There is a great need to interpret the results in a way that is useful to both research scientists and clinicians. Studying the structure of microbial communities will shed light on the nature of bacterial "social networks" and their consequences.

1.2 The Metagenomics Analysis Pipeline

No single approach is sufficient to fully characterize a source environment due to the complexity of the microbial communities in metagenomic samples. Thus, any metagenomics analysis pipeline will involve a series of sequential steps. This begins with data pre-processing to filter reads by quality and length, remove contaminants, remove chimeric sequences generated during PCR amplification, and prepare data for subsequent analysis. A classification and clustering step is used to map each read to a sequence in a reference database and group reads by taxonomy or sequence similarity. This will be discussed in more detail later. A single-sample analysis step employs standard measures, such as making estimates of the richness and diversity of taxa in each sample; when whole genome sequencing is used, the analysis involves studying the functional pathways and protein families that are present and/or overrepresented. Finally, multiple-sample comparisons are used to identify patterns in related groups of samples. Every step in the pipeline has the potential of introducing errors and biases which will be carried forward to the rest of the analyses [22].

It is essential that parameters be selected with care, understanding the limitations of the data and the biases introduced during collection, amplification, and sequencing.

The marker gene approach is unable to directly provide information about the functional elements present in a sample due to its limited ability to resolve taxonomic identity. The gene for 16S rRNA is present in every bacterial genome. This gene contains a mixture of highly conserved and hypervariable regions, the former making it an easy target for amplification and the latter for mapping reads to taxa [66]. However, it is generally not possible to use the 16S rRNA gene to distinguish between strains or species, especially with short reads, so identity is normally limited to the genus level or above [25]. This can pose a problem for our purposes, since members of the same genera can behave very differently. For example, *Campylobacter hominis* is considered a member of the normal flora of the gut, whereas *Campylobacter jejuni* is known to be pathogenic [69]. Furthermore, closely related bacterial species and strains are often competitors for the same environmental niches [30]. It is desirable to differentiate between any species and strains present in order to better understand the dynamics of the communities being studied.

The limited resolution of marker gene methods does not mean that reads cannot be intelligently assigned to distinct groups, but rather that those groups cannot be easily mapped to specific taxa. A commonly used approach is to first cluster reads based on sequence similarity into *operational taxonomic units* (OTUs), and then assign the best available taxonomic identity to each OTU. This typically results in several OTUs mapping to the same taxon, with each OTU being roughly analogous to a strain or species at or above a specified sequence similarity. A commonly used threshold for the 16S rRNA marker gene is 97% similarity. It is sometimes useful to approximate higher taxonomic levels, for example, to conduct an analysis at the phylum level. This can be accomplished by adjusting the similarity threshold. Labeling each OTU with reference to its taxon is a convenience, but should be done with caution. There is experimental evidence that even the best 16S classifiers exhibit some taxon-specific biases [23]. The use of OTUs may help minimize the effects of these biases.

There are other technical limitations with marker gene approaches. The choice of primers used for PCR can have an appreciable effect on amplification efficiency and resulting coverage [63]. Primers are required for the process of amplification, and are used by DNA polymerases to identify start sites for copying the complementary strand as well as to serve as a scaffold on which to build the new strand. Another advantage of primers is that they can be designed to amplify the specific marker genes of our choosing by matching the start of the genomic sequences that interest us. The process of evolution will inevitably produce minor differences in DNA sequences as species diverge, even in the highly conserved regions of marker genes. Ideally designed primers must be able to bind to a maximum number of different target species while capturing genomic regions that maximize their distinguishability. The most commonly used primers were developed many years ago when 16S rRNA databases were still relatively sparse [64, 62, 39]. More recently designed degenerate primers, which contain certain "wildcard" bases, have been shown to be a vast improvement over these and can capture substantially more taxa [34]. However, one can never know what has been missed by this approach. Another confounding problem with

marker genes is the issue of copy number. All bacteria usually have multiple copies of 16S rRNA genes. Differences in copy number might give an inflated estimate of the relative abundance of certain taxa [14]. More troubling is the problem of noise, the presence of large numbers of reads that form very low membership OTUs, resulting in thousands of OTUs with as little as a single read. It can be difficult to determine if these OTUs should properly be merged with some other larger OTU (by adjusting the cutoff for similarity), if they represent amplification or sequencing artifacts, or if they are indicative of truly low abundance OTUs present in the sample [44]. All of these limitations beg for intelligent ways of filtering and normalizing the data, and a standard for dealing with sequencing efficiency, copy number, and noise has yet to emerge. For now, it is important to recognize that every study is based on a blurry temporal snapshot of a microbial community potentially filled with large, gaping holes. This is obviously a challenge for making inferences about interactions between bacteria when trying to describe the microbial social network, but not an insurmountable one.

A large number of open-source analytical tools for metagenomics have emerged in recent years. These include MG-RAST, MOTHUR, QIIME, CloVR, VAMPS, and others [46, 53, 8, 4, 3, 31]. These can automatically step through a basic analytical pipeline for metagenomics. However, this is just a starting point for determining how these billions of cells are interacting and the implications for the health of the host. Are there groups of bacteria that tend to work together? Do they appear in healthy as well as diseased tissue? If these groups are present, do they actively compete with other groups for nutrition in the environmental niche, and just how fierce is this competition? What are the implications of the presence of specifc groups on the health status of the human host? Can we classify certain bacteria as implicitly pathogenic based on their group membership? Can we infer causal relationships between these groups and disease states or the prognosis of the subject? These more interesting and challenging questions require reflection and are not always answered using cookie-cutter tools.

1.3 Data Limitations and Sources of Errors

As stated earlier, every step in the metagenomics pipeline introduces new errors and biases which will be carried forward to the rest of the analyses. Sources of error in the analyses are listed below.

1. Contamination is a widespread problem, especially during sample collection and DNA extraction. However, propagation of contamination issues can be mitigated by incorporating appropriate controls to subtract out in the downstream analysis.

2. Partial or complete inhibition of amplification from different compounds or chemicals present in the sample, despite efforts to remove them in DNA extraction.

3. PCR amplification of some regions of bacterial genomes may not work efficiently due to issues such as inappropriate PCR amplicon size (outside the range of ideal amplicon size), low copy number, unfavorable nucleotide content (AT-rich or repetitive sequences), faulty universal primers, improperly optimized PCR conditions (temperature, reagent concentrations or cycling parameters), or stochastic effects related to DNA concentration.

4. Multiple sources of error exist in the sequencing phase, resulting in many "noisy" reads. One way to eliminate these reads is to set a cutoff or threshold for including OTUs in the analysis. However, an incorrectly chosen threshold can be another source of error.

5. Clustering and classification of reads is a major source of error. If clustering of reads is used, then it uses evidence from a collection of reads in order to decide on a classification, thus reducing the reliance on the classification of individual reads. On the other hand, if the classification is incorrect, it could misclassify entire clusters.

6. When dealing with whole genome sequencing data, often more than one cluster is mapped to the same taxon, which is typically inferred to be caused by the presence of subcategories of that taxon (e.g., species or strains of the same bacterial group). Depending on the region used for the classification, this inference may be incorrect. It is possible that members of a different bacterial group acquire critical genes by *horizontal gene transfer* and end up being misclassified.

7. Normalization is necessary, but can be fraught with errors because of the assumptions made. The choice of normalization method remains controversial [49, 12].

8. Data imputations, if needed, can be another source of error.

9. It is essential that parameters be selected with care and awareness of known limitations in how data was collected, amplified, and sequenced.

10. Any multi-step inference process can be erroneous because of potential compounding of errors.

11. There are a multitude of external factors which may influence the structure and activity of microbial communities but which are not considered by either amplicon or whole genome sequencing. These factors might be highly dynamic, and can include temporary influxes of invading taxa from environmental exposure, the presence of non-microbial biological entities, changes in the availability of nutrients or the build-up of toxins due to host behavior or cyclical patterns (i.e., inconsistent sleep patterns, dietary changes, hormonal fluctuations, etc.), and other unknowns which may profoundly affect any analysis.

1.3.1 Designing degenerate primers for Microbiome work

Metagenomics studies require an efficient PCR amplification of the DNA. This is currently achieved using "universal" forward and reverse primers that amplify appropriate regions of DNA from all targeted mircobes. Jaric et al. [34] argue that currently used universal primers for the 16S rRNA region were designed many years ago, are not as efficient, and fail to bind to recently cataloged species. Their analysis shows that 22 of the most widely used primer pairs are far from optimal in the sense that they do not abide by primer design rules and fail to produce amplicons in a "virtual PCR" experiment, because they fail to hybridize to a large number of 16S sequences in the current database.

Jaric et al. [34] focus on the optimal design of degenerate primer pairs so that (a) the number of 16S sequences "virtually" amplified by the primer pair is maximized, and (b) the number of 16S sequences that produce distinguishable amplicons is maximized. Jaric et al. proposed an automated general method of designing PCR primer pairs that abide by primer design rules and uses current sequence databases as input. Since the method is automated, primers can be designed for targeted microbial species or updated as species are added or deleted from the database.

The designed primers were shown (in "virtual PCR" epxeriments) to achieve the goals mentioned above. They take the work one step further to design sets of primer pairs that extend the number of distinguishable amplicons for 16S sequences in the database. Wet lab experiments lend further credence to their claims on the effectiveness of the designed primer pairs.

1.4 Diversity and Richness Measures

Though the early steps in an analytical pipeline produce a noisy snapshot, there are a number of preliminary measures that can be employed to gain some insight into the environment being studied, including making estimates about the richness and diversity of the community [17]. *Richness* simply indicates the number of different OTUs present in a sample given that it is not possible to make an exact count. It assumes that not every OTU present has been directly observed. *Diversity*, like entropy, measures the evenness of the distribution of the abundance of the OTUs. If each OTU has a similar number of members, then the diversity of the community is relatively high, whereas if a few OTUs make up the bulk of all the cells present, then diversity is considered very low.

This, of course, tells you very little about how the members of these communities are interacting. Richness and diversity measures are most useful when compared between different communities, such as the communities housed by different human subjects at the same body site, or between different body sites in the same individual. This is where the coarsest patterns may begin to emerge. Consider samples collected from two groups of subjects, one of healthy controls and the other suffering from some illness. You observe that the number of distinct OTUs present in both groups is about the same, but that diversity is much higher in your healthy subjects. This

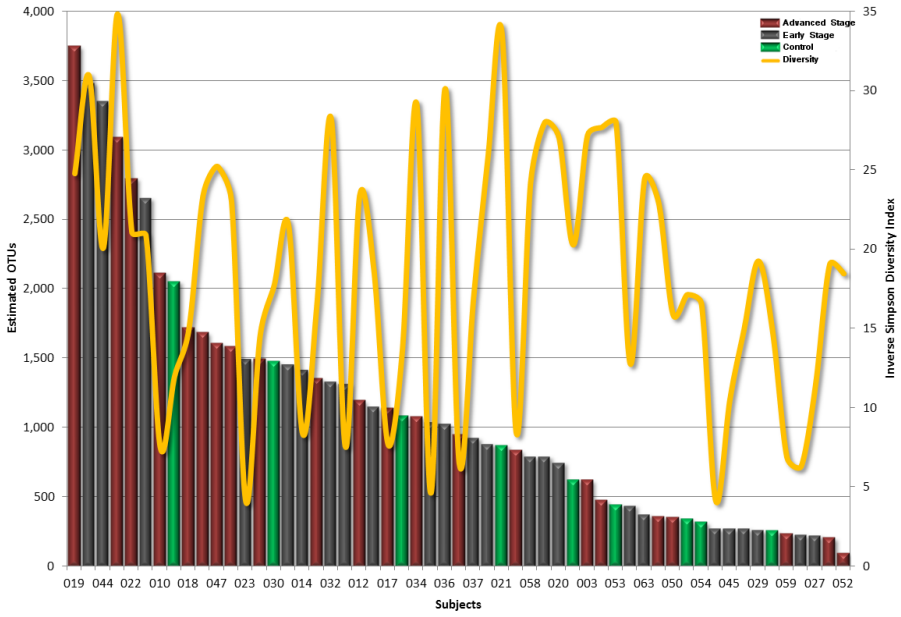

Figure 1.1 Plot of the standard measures of richness and diversity for a collection of samples. Each bar represents the richness (i.e., estimated number of OTUs present in the sample), and the gold line indicates the diversity of the sample. Bars are color-coded by clinical category and are arranged in descending order. However, no apparent pattern is discernible between the categories that are being compared, either in richness or in diversity (unpublished data).

suggests that a small number of OTUs in your diseased subjects have found an environment that favors them, and they are using up the bulk of the resources that are available at the expense of the other OTUs present. This would be a nice, simple pattern and you could claim to have found a biomarker for this disease. Unfortunately, in most studies to date, patterns like this have failed to emerge. There is simply too much variability between humans, and it is often found that richness and diversity tell you almost nothing about an individual's health status. (See Fig. 1.1, where richness and diversity say little that is useful about the three groups being compared.) What has been observed is that an individual's diversity levels tend to be consistent across body sites [32]. In other words, someone who has high diversity at body site A is likely to also have high diversity at body site B. In addition, these relative diversity levels seem to remain constant over time. It is not known if diversity levels change when an individual changes status, say from healthy to diseased. Although these broad standard measures merit further tracking, they are probably mainly useful as indicators of the complexity of the community being studied. It is generally more interesting to know which particular OTUs are present and how their abundance levels are changing at different stages of disease progression. This is the true starting point for constructing a network of interactions.

1.5 Correlations and Association Rules

It is important to know which OTUs tend to occur together and which do not. More precisely, it is interesting to identify instances when high relative abundance levels for one OTU tend to coincide with high abundance levels for another OTU. Alternatively, it is also interesting to identify when high relative abundance levels for one OTU coincide with low abundance levels for another OTU. In addition, there could be other factors that might be considered along with OTU abundance levels. For biomedical studies, demographics of the subject from which samples were taken, or other known environmental conditions (such as acidity or aerobicity of the niche) can be included.

Finding these types of correlations may be the first step in deducing interactions between OTUs in a community. If two OTUs are positively correlated, this could suggest that they are not competing for the same resources or niche in the community, and could even indicate some kind of dependency between them. For example, the waste products of one OTU may be modifying the pH levels in the environment, making it more agreeable to a second OTU. Likewise, if the change in pH is toxic to another OTU, we would observe a negative correlation between them. However, there are many alternative interpretations for correlations in the abundance of OTUs. Consider a situation in which three OTUs are competing for the same limited resource. If one OTU is especially efficient at utilizing this resource, its numbers would likely increase while the abundance of its competitors would decline. So there should be a negative correlation between this exceptionally competitive OTU and its two rivals. The abundance of the rivals, however, would tend to change concordantly, assuming they are equally fit. The presence and success of the efficient OTU would cause the number of members of the two less efficient OTUs to decline concordantly, resulting in a positive correlation between them, despite the fact that they are competitors for the same limited resource. Although it is not usually straightforward to understand the reason for the existence of a correlation, establishing the connections can help build a picture of what is going on in a community. We will explore ways of visualizing a network of correlations shortly.

Of greater interest than just pairs of correlated OTUs, is to determine sets of OTUs that tend to co-occur, and whether the presence of one OTU is preditive of the presence of another for a subset of the subjects. For this we can turn to statistical data mining tools for generating association rules, such as the Apriori algorithm for basket data [1]. This works by first identifying specific sets (combinations) of bacterial taxa or other items that are frequently observed together in a large number of subjects, and then generating a rule for each item based on the other items in the set. The validity of the rule can be quantified for the given data by computing quantities such as *support*, *confidence*, and *lift* [41]. Apriori can be used to essentially cluster OTUs by how often they occur together, and suggest which OTUs we should be seeing based on "guilt-by-association". Although association rules are of interest, Apriori can also be used as a classifier to predict disease states. If certain item sets are frequently found in a group of diseased individuals but are absent from otherwise healthy individuals, the item sets can act as markers for the disease.

Several limitations hamper the use of the Apriori algorithm. For one, it requires discrete values as inputs, and the data resulting from the metagenomics pipeline after normalization consists of relative abundance values for each OTU, which are continuous values. Numerous discretization methods are available to categorize abundance levels. One could discretize the values as "present" or "absent", or as "high", "medium", or "low", or into even finer categories, but these tend to flatten the data, and finding appropriate cutoffs for each category can be fairly arbitrary. There has been some effort toward methods that find frequent itemsets and association rules without discretization, but relying on correlations [6]. In addition, the number of item sets and association rules generated is normally very high, with many of the item sets showing interesections and similarities. Effective and compact visualizations of rules can be difficult (see Figure 1.2), although graph-based visualizations have great promise (see Section 1.7). Furthermore, determining statistical significance remains an open problem [20]. Still, if properly wielded, the use of Apriori can lead to the discovery of interesting associations and interactions between bacterial taxa and help identify indicators of critical changes in human health.

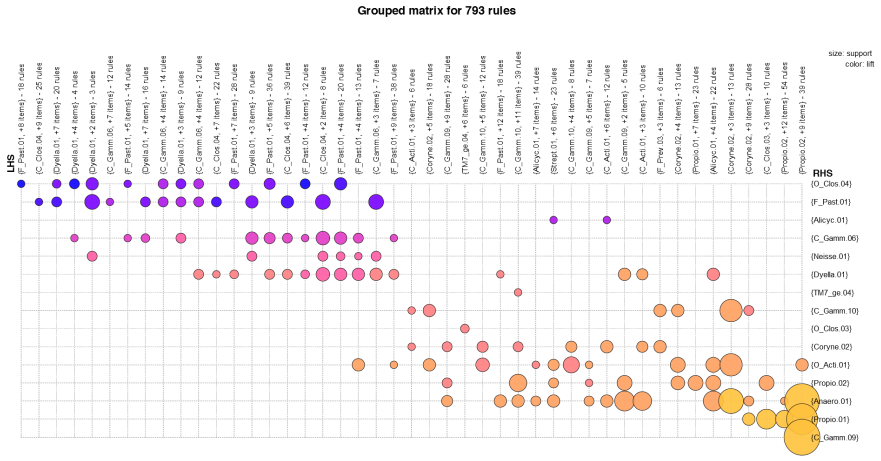

Figure 1.2 Example of a bubble plot for visualizing association rules. Columns consist of overlapping antecedents (item sets) leading to the consequents (the rules) in the rows. Bubble size indicates support and color the strength of the interest measure.

1.6 Microbial Functional Profiles

A recent study has revealed that the gene content of the bacterial community is more constant than the phylogenetic content [7]. Therefore, there is a compelling impetus to move beyond species composition-centric studies of microbial communities and towards functional composition analysis. As discussed earlier, a key preliminary step

in metagenomic analysis is to decipher the microbial community structure of a given niche by categorizing the microbes residing therein and understanding their diversity. In the context of microbial communities, the term "species" refers to a fundamental and distinct rank of taxonomic hierarchy. There are limits to what taxonomic profiling can tell us about a community. Although interesting, community membership does little to describe the role being played by the taxa present. It is therefore much more important to investigate the functional profile of microbial communities. The *functional profile* of a sample can be defined as the set of expressed genes and active metabolic pathways in a community.

Consider a situation in which two species have very similar characteristics and functions, both of them being very efficient at synthesizing vitamin B12, for example. These species would be competitors in the same niche and their presence likely to be mutually exclusive. Imagine doing a comparison of two communities, one in which species A is present in high abundance, and the other in which species B is present. Based on the taxonomic profile, it would seem that these two communities are different, but in fact since the species play the same role, all other things being equal, the communities are actually similar. There is evidence to support situations like the one described. The Human Microbiome Project has observed that although taxonomic membership is highly variable between human subjects at particular body sites, the functional profiles at those sites vary little [32].

Differences in functional profiles of the microbiota can be predictive of the health of a host. Knights et al. [35] reviewed supervised classification methods that exploit this fact. It is therefore of paramount importance to understand the functional profiles of microbial communities in various sites and their dynamics. Can the dynamics of the profile shed light on the health of the host and will it also enlighten us on the etiology, progression, and prognosis?

As previously discussed, two approaches have been adopted for characterizing the diversity of metagenomes. The first approach focuses on the sequencing of phylogenetic marker genes, such as the gene for 16S rRNA [9, 28, 5, 26, 38, 21]. The second approach is based on whole genome sequencing. Whole genome sequencing is superior to 16S rRNA approaches primarily because it is able to directly characterize the functional profile of the community. Typically, sequences are compared to public databases and genes predicted by homology. Taxonomic identification can be helpful, but is not essential. The major limitation here is that functional annotations are still far from complete, and accurately ascribing functions to newly predicted gene sequences will take many years. Ontology databases will not catch up to the available data anytime soon.

As mentioned earlier, the most popular approaches continue to be those based on 16S rRNA [9, 28, 5, 26, 38, 21]. The functional analysis of 16S rRNA metagenomic data focuses on what "species" are present. In cases where the roles of the bacterial species are well understood, taxonomic identity can be an indicator of the more active functions in the community. Methods also exist for imputing functional profiles based on 16S data. These involve reconstruction of the ancestral state of taxa to predict the presence of gene families in descendants [40]. Although these methods have been tested with good results, problems remain. The 16S approaches cannot resolve

taxonomy below the genus level, and it is known that different species or strains in the same genus can behave very differently. The specific changes leading to these behaviors cannot be predicted from ancestry. Furthermore, taxonomic identification is limited by the completeness of marker gene databases. The number of sequences that cannot even be classified down to the genus level can be very high in some studies [24]. Additionally, when analyzing whole genome sequencing data, the problem of horizontal gene transfer (HGT) needs to be considered. There are indications that the HGT rate can be as high as 8% in some bacterial communities [58]. Although there is merit to attempts to create functional profiles based on taxonomy, these methods may be most useful for identifying environments which would strongly benefit from whole genome sequencing.

Zhang et al. [71, 72, 73, 74] have developed a novel approach to analyze the whole genome data of a microbiome. Their strategy is to compute a matrix, where the rows correspond to features such as genes, gene families, functional groups, or even pathways, and the columns correspond to OTUs. Thus, instead of mapping OTUs to bacterial taxa, the matrix maps them to *functional entities* that enable us to ascribe functional characteristics. The matrix can be seen as a collection of columns (OTUs with a functional profile), or as a collection of rows (COGs with measures of their contributions to each OTU). Clusters of Orthologous Groups (COGs) was used as functional entities to create an edge-weighted network of COGs where the edge weights are the correlations between the row vectors. It is possible to construct a single COG correlation network for each sample or set of samples. If dense subgraphs of this network are identified as clusters, then such a cluster corresponds to COGs that have a similar abundance profile across all OTUs. The authors extrapolated that these COGs have similar functions and are involved in the same pathways. Using GO annotations of these COGs, functional enrichment analysis was pursued. Given two networks, one for each of two samples or sets of samples, differential analysis can highlight the functional differences between them. Another possibility is to investigate which network motifs are conserved or changed between the samples. Differential modules can be mapped to KEGG pathways using iPath, an interactive pathway explorer [67]. An example of such a mapped module is shown in Figure 1.3. As can be seen, the pathways of genes in this module include functional categories such as energy metabolism; glycan biosynthesis and metabolism; metabolism of terpenoids and polyketides; and metabolism of cofactors and vitamins. An added advantage of using functional entities is that it helps reduce the number of dimensions nearly tenfold. OrthoMCL [42], a BLAST-driven method, can also be used to identify putative COGs[1].

[1]Note that in this work, COGs refer to orthologous gene groups obtained directly from results of the computational tool OrthoMCL [42] on genes from the genomes of interest. It should be noted that these COGs are different from the NCBI COG database [59], which contains clusters from 66 Unicellular organisms.

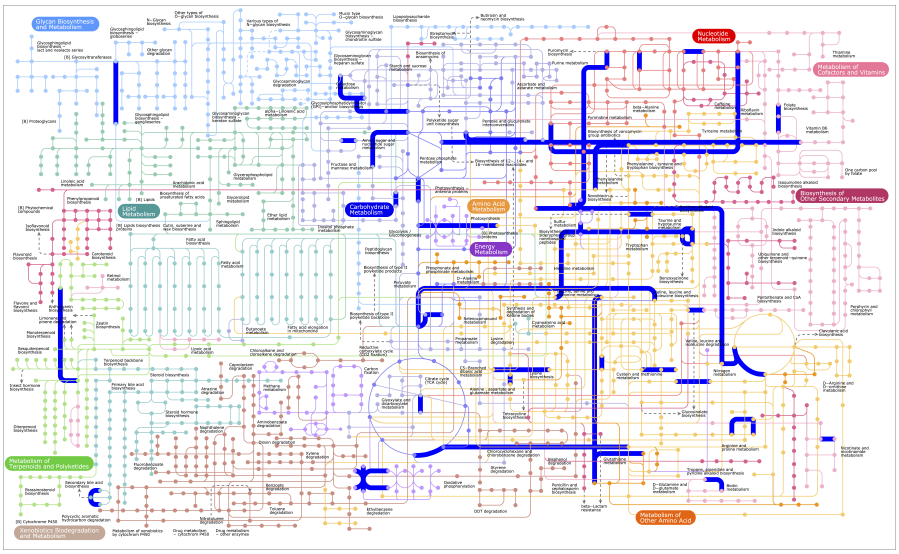

Figure 1.3 Metabolic map of a module identified from a gene network. Nodes symbolize compounds, and lines connecting nodes are enzymes. All enzymes (lines) corresponding to a single KEGG map have the same color. All enzymes (lines) corresponding to a single module are highlighted and colored with module color.

1.7 Microbial Social Interactions and Visualizations

Bacterial communities are intricate collections of bacterial species that each provide functions which contribute to the stability of the community. In most natural environments, microbes do not live in isolation but form a complex ecological interaction web. Recent research shows that complex social behaviors are commonly observed not only in animals but also in bacterial species [2, 65]. Those social behaviors involve complex systems of cooperation, communication, and synchronization. Thus, the communities are dynamic consortia of microbial species populations.

Because of the complexity of interactions and the sheer size of the communities being studied, efficient ways of analyzing and visually summarizing our analysis are needed. The use of network diagrams are appropriate for this purpose. These consist of nodes to represent each OTU, with edges connecting those nodes that have some kind of relationship. Nodes may have one, multiple, or no edges between them. An example of a complex network diagram is shown below (Figure 1.4).

Here we see some of the more abundant OTUs found in the lungs of active smokers (unpublished data). Each OTU is labeled with the best taxon to which it maps. In the case of multiple OTUs mapping to the same taxon, an arbitrary number is appended to the label to distinguish one from the other. Edges between the nodes indicate correlations between the populations of the OTUs present. The size of the node is adjusted to indicate the size of the population. Correlations are a measure of

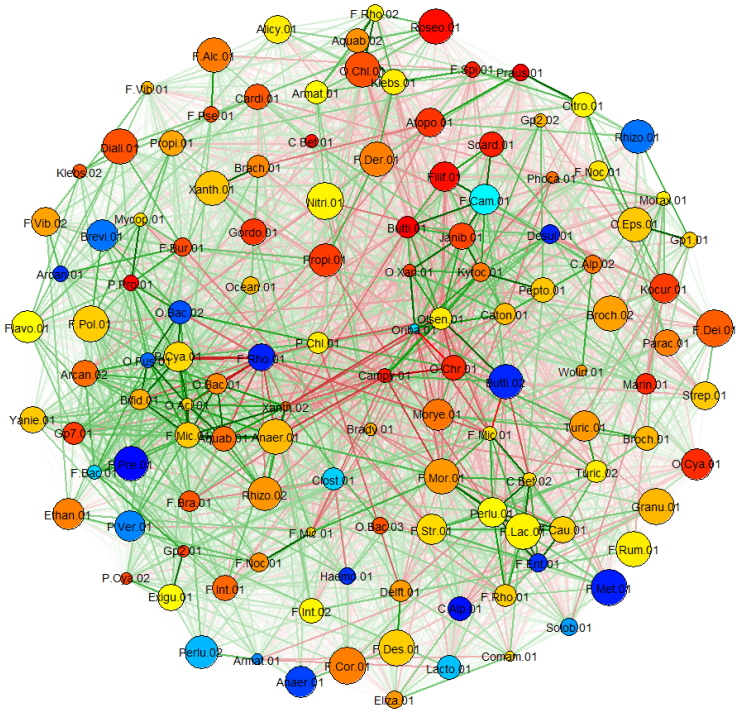

Figure 1.4 Basic network diagram where nodes represent OTUs and edges represent co-occurrence in subjects. Edge color indicates positive (green) or negative (red) correlations. Node size is adjusted to reflect relative abundance using a log scale. A force-directed layout using the Fruchterman-Reingold algorithm is used. The position of each node is dependent on the strength of its interactions with all other nodes in the system. A heatscale has been used to assign a color to each node based on differential abundance between two groups of subjects. The greater the significance in the difference, the hotter (redder) the color of the node.

co-occurrence of the OTUs in the subjects being considered. A green edge is used for a positive correlation, a red edge for a negative correlation. The thickness of the edge increases with the strength of the correlation. The size of each node is adjusted to express the relative abundance of the OTU. The nodes are colored according to a heat scale. The less significant the difference in abundance between two groups, the cooler (bluer) the color. The more significant the difference, the hotter (redder) the color. Finally, each OTU is positioned in space according to the Fruchterman-Reingold algorithm [27], which works as follows: Imagine that a spring exists be-

tween every pair of nodes; the strength of the springs varies depending on the size of the node and the strength of the correlation between them. Initially, each node is placed at an arbitrary position in space, and the overall energy of the system due to the pull of the springs is calculated. Two strongly (positively or negatively) correlated nodes will tend to attract each other, but there may be many other interactions acting to pull them apart. Springs that are "stretched" store energy. The positions of the nodes are iteratively readjusted according to these combinations of forces, with the overall goal of minimizing the total energy of the system. This iterative process stops when the total energy of the system remains unchanged, i.e., a local minimum is reached.

The diagram above contains a large amount of information in a very compact way. However, the density of diagrams like these requires careful study to properly interpret them. Some relationships can be easily picked out visually, while others are much less clear. As should be expected, two strongly positively (or negatively) correlated OTUs tend to be located in close proximity. However, the converse is not true. Two OTUs located in close proximity may have no detectable correlation. There is typically no clear delineation between groups of positively and negatively correlated OTUs.

The network diagrams shown above indicates relationships similar to those found in social networks [29]. Cooperation and competition between bacteria has been well studied in the field of bacterial ecology. It is therefore natural to ask whether these network diagrams reveal interactions between bacterial taxa. In this context, it makes sense to ask if there are "clusters" in the network graphs and if these clusters are different for different groups of subjects.

We informally define a *club* as a cluster of bacterial taxa such that every member of the group has a stronger correlation with the rest of the group than it does with members not in the group. The concept of clustering has been well studied in the statistics and computing literature and has been defined and computed in a multitude of ways [68]. Differernt clustering methods have used different distance and similarity measures or modeled the data using statistical distributions. Others have defined different objective functions using density and tightness measures, variance ratios, weakest links, relative margins, and other statistical measures, resulting in a wide range of definitions of clusters and algorithms for clustering.

Clubs with positive correlations between each other are likely to be indications of "cooperation" between the members of the group, while negative correlations are likely to be evidence of "competition". The former group of interacting OTUs may represent taxa that complement each other in a given environment and their composition could indicate a core group of functions needed to thrive.

A substantially more interesting structure to be found in the network diagrams turns out to be "competing groups" of bacterial taxa. We informally define *rival clubs* to be a pair of clubs such that every member of one club has a negative correlation with some or all of the members of the "rival" group. As with the definition of a club, this concept can be formalized and computed in a variety of different ways. Rival clubs are likely to indicate "competing" groups requiring the same scarce resources

in a niche. In the example shown in Figure 1.5, Club-A, Club-B, and Club-C form a trio of rival clubs (unpublished data).

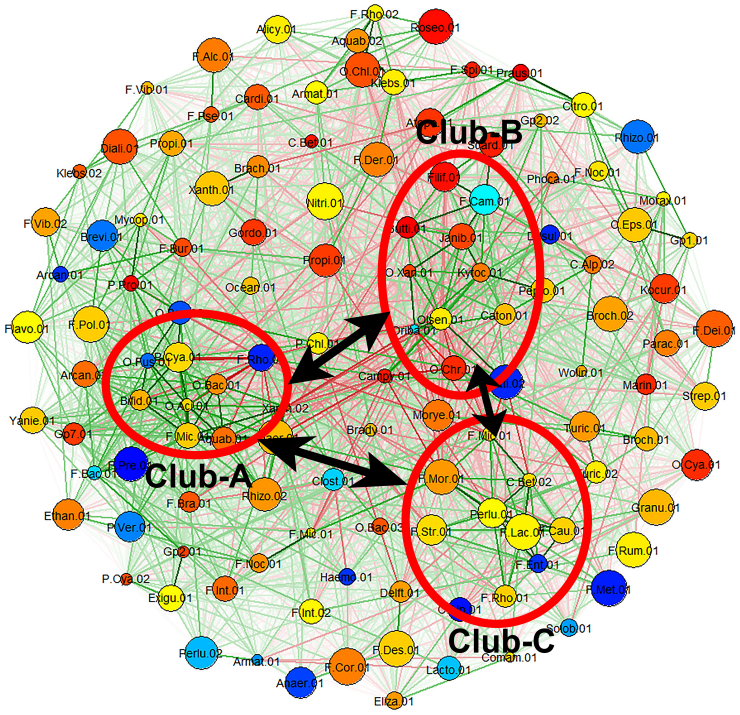

Figure 1.5 Once clubs have been identified, rival clubs are charcterized by many negative edges between them.

Interpreting a cluster of this kind raises many unanswered questions. What can be inferred about a cluster of OTUs? Do these OTUs compete for different nutrients in the environment? Are they critical to each other? Does the presence of one help to recruit others to the same environment? What is the effect of the elimination of one of the members of a cluster? Can anything be said about their pathogenicity? Are they all pathogenic or all non-pathogenic? Or can a mix occur? What functional genes do their genomes represent? Do they collectively represent a useful functional profile that helps them to thrive? Are there differences in gene expression? Do they collectively represent a useful gene expression profile? How do they communicate, and what chemicals and metabolites are present in the niche? What can be said about

the level of horizontal gene transfer within the members of a cluster? Are there quorum sensing genes or other communication genes strongly expressed within a cluster? Is a cluster meaningful in posting a response to environmental stimulus or stress, such as an antibiotic?

There are some limitations to these network diagrams herein. We are clearly extrapolating interactions based on information on co-occurrence. While positive interactions do suggest co-occurrence, the converse need not be true.

1.8 Bayesian Inferences

Although the correlation networks described above can produce meaningful insights, there are other network structures that can be learned from data which can be helpful in answering interesting questions about the communities being studied. The use of probabilistic graphical models (PGM) for analyzing microbial communities is still uncommon, but they can serve as useful complements by helping to infer dependencies between OTUs and to possibly identify causal reasons for disease states. A PGM is a directed acyclic graph in which each node represents a variable of interest and each directed arc indicates a conditional dependence relationship between nodes [36]. A probability distribution is associated with each node, and through the application of Bayes rule, predictions can be made about the presence of a node given information about other connected nodes. As a simple example, consider the network represented in Figure 1.6.

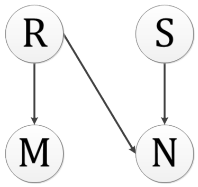

Figure 1.6 The likelihood of event M ("lawn is wet"), depends only on the probability of event R ("it rained"). In contrast, the likelihood of event N ("neighbor's lawn is wet") depends on event R and the probability of the event S ("neighbor turned on the sprinkler".

Here we see that the probability of the lawn being wet depends only on the probability of it having rained. In contrast, the probability of the neighbor's lawn being wet depends on the probability of rain as well as the probability of the neighbor turning on the sprinkler system. Given the condition that it has rained, the probability of the lawn being wet changes and can be easily calculated. However, given that the neighbor turned on the sprinkler system gives no additional information about whether or not the lawn is wet. In essence, to determine the state of the lawn, the relevant node is only R since the state of M does not depend on any of the other nodes in the network. In a large network, a structure like this dramatically reduces the number of calculations required for determining the probability of the state of a node of interest.

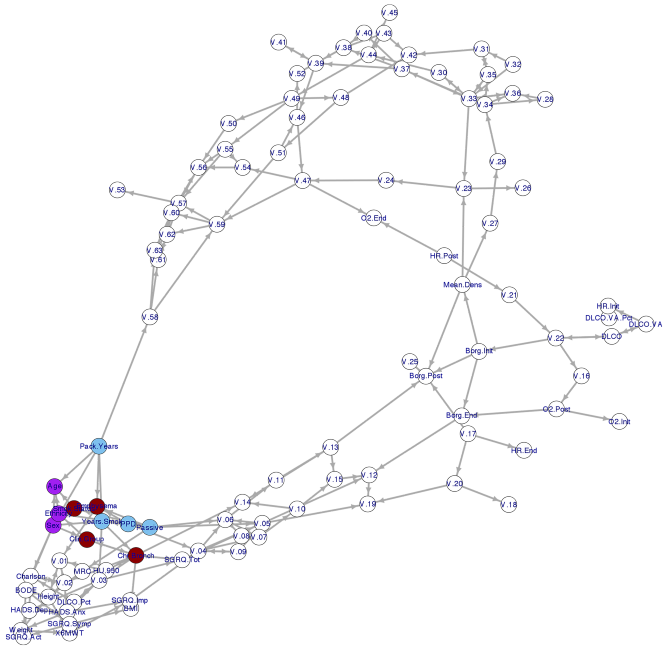

Figure 1.7 A complex PGM

Figure 1.7 gives an example of a more complex PGM constructed from metagenomic data (unpublished data). In this case, most nodes represent OTUs, but also included are nodes which represent disease states, host demographics, and other aspects of host phenotype. Once a model is constructed, it can be used to predict the likelihood of a subject developing the disease when any combination of dependent nodes is observed. According to Bayesian theories, certain substructures in these networks can also be used to infer causation.

The use of force-directed layouts in constructing the graphical representations provides similar insights as in the correlation networks above. However, learning the structure of the network can be much more computationally challenging. Even a modestly-sized network constructed using Markov chain Monte Carlo sampling will normally require parallelization on a high performance computer cluster. The joint probability distribution may include prior knowledge about relationships or can be learned on the fly. Although the use of priors will not affect the structure of the network, having sound probabilities is essential for any inferences being made.

1.9 Conclusion

To describe the structure of a microbial community, we have considered all of the elements present and the connections between them, which constitutes a network of relationships. Too few microbiome studies attempt to analyze their subject from this perspective. Those that do, tend to focus on taxa, or sometimes the functional elements present, without much regard to other possible nodes. To be sure, a great deal of information can be inferred about these networks by measuring changes in the abundance of either taxa or genes in different environments. These noisy snapshots can clue us in on many of the interactions that are occurring. Network diagrams are powerful tools for examining the complex relationships in microbial communities. A large number of connections can be studied at once, and they allow for the discovery of unexpected associations between OTUs, both cooperative and competitive. Yet this is still an incomplete picture. There are too many elements that are not captured by most current methods, and in many cases it is likely that we lack sufficient information to correctly deduce the direction of individual interactions or the topology of the networks being studied. More effort needs to be focused on ways of discovering the structure of microbial networks.

Studying these communities in humans is particularly challenging. We all have different experiences, eat different foods, live in different places, participate in different social networks. All of these differences contribute to the high variability in the infrastructure our bodies provide to our bacterial residents. Yet some markers may exist that can distinguish between the microbial communities present at different body sites. Our intuition tells us that there should be a difference between the community residing in healthy tissue versus diseased tissue. Something must be different about the network of bacteria in the lungs of someone with emphysema as compared to someone with no respiratory problems at all. Is there a signature that we have yet to discover? If there is validity to the concept of an enterotype, a classification of individuals based on the bacteria present in their gut, is there also a "*respirotype*"? A "*dermatype*"? If so, it is likely not sufficient to simply define these types based on absence or presence of specific bacteria. What role does host genotype play? Age? Life history? It is the interactions between functional elements, taxon-specific characteristics, host tissue properties, the presence of metabolites, toxins, sensing molecules, and exogenous influences that will produce distinct signatures between healthy and diseased environments. Discovery of these highly heterogeneous subnetworks will be challenging, but their value for rapidly assessing state of health, predicting disease progression, and developing intervention strategies cannot be understated.

REFERENCES

1. R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

2. E. E. Allen and J. F. Banfield. Community genomics in microbial ecology and evolution. *Nature Reviews Microbiology*, 3(6):489–498, June 2005.
3. S. V. Angiuoli, M. Matalka, A. Gussman, K. Galens, M. Vangala, D. R. Riley, C. Arze, J. R. White, O. White and W. F. Fricke. CloVR: a virtual machine for automated and portable sequence analysis from the desktop using cloud computing. *BMC bioinformatics*, 12(1):356, 2011.
4. M. Arumugam, E. D. Harrington, K. U. Foerstner, J. Raes and P. Bork. Smashcommunity: a metagenomic annotation and analysis tool. *Bioinformatics*, 26(23):2977–2978, 2010.
5. A. Barberan, S. T. Bates, E. O. Casamayor and N. Fierer. Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J*, 6:343–351, September 2011.
6. S. Brin, R. Motwani and C. Silverstein. Beyond market baskets: Generalizing association rules to correlations. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, SIGMOD '97, pages 265–276, New York, NY, USA, 1997. ACM.
7. C. Burke, P. Steinberg, D. Rusch, S. Kjelleberg and T. Thomas. Bacterial community assembly based on functional genes rather than species. *Proceedings of the National Academy of Sciences*, 108(34):14288–14293, 2011.
8. J Gregory Caporaso et al. QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336, 2010.
9. S. Chaffron, H. Rehrauer, J. Pernthaler and C. von Mering. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, 20(7):947–959, July 2010.
10. P. S. Chain et al. Genomics. Genome project standards in a new era of sequencing. *Science*, 326(5950):236–237, Oct 2009.
11. J. R. Cole, Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske and J. M. Tiedje. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.*, 42(Database issue):D633–642, Jan 2014.
12. P. I. Costea, G. Zeller, S. Sunagawa and P. Bork. A fair comparison. *Nat. Methods*, 11(4):359, Apr 2014.
13. E. K. Costello, C. L. Lauber, M. Hamady, N. Fierer, J. I. Gordon and R. Knight. Bacterial community variation in human body habitats across space and time. *Science*, 326(5960):1694–1697, Dec 2009.
14. L. D. Crosby and C. S. Criddle. Understanding bias in microbial community analysis techniques due to rrn operon copy number heterogeneity. *BioTechniques*, 34(4):790–794, Apr 2003.
15. T. Z. DeSantis, P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu and G. L. Andersen. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, 72(7):5069–5072, Jul 2006.
16. M. S. Doud, M. Light, G. Gonzalez, G. Narasimhan and K. Mathee. Combination of 16S rRNA variable regions provides a detailed analysis of bacterial community dynamics in the lungs of cystic fibrosis patients. *Hum. Genomics*, 4(3):147–169, Feb 2010.

17. M. Doud, E. Zeng, L. Schneper, G. Narasimhan and K. Mathee. Approaches to analyse dynamic microbial communities such as those seen in cystic fibrosis lung. *Human Genomics*, 3(3):246–256, 2009.

18. P. B. Eckburg, E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen, M. Sargent, S. R. Gill, K. E. Nelson and D. A. Relman. Diversity of the human intestinal microbial flora. *Science*, 308(5728):1635–1638, Jun 2005.

19. J. R. Erb-Downward et al. Analysis of the lung microbiome in the "healthy" smoker and in COPD. *PLoS ONE*, 6(2):e16384, 2011.

20. K. Faust and J. Raes. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.*, 10(8):538–550, Aug 2012.

21. K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes and C. Huttenhower. Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Comput Biol*, 8(7):e1002606+, July 2012.

22. K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes and C. Huttenhower. Microbial co-occurrence relationships in the human microbiome. *PLoS Comput. Biol.*, 8(7):e1002606, 2012.

23. M. Fernandez. An analytical workflow for metagenomic data and its application to the study of chronic obstructive pulmonary disease. Florida International University Undergraduate Honors Thesis, 2013. Undergraduate Thesis.

24. M. Fernandez and G. Narasimhan. Unpublished results.

25. G. E. Fox, J. D. Wisotzkey, P. Jurtshuk and G. E. Fox. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int. J. Syst. Bacteriol.*, 42(1):166–170, Jan 1992.

26. S. Freilich, A. Kreimer, I. Meilijson, U. Gophna, R. Sharan and E. Ruppin. The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Research*, 38(12):3857–3868, 2010.

27. T. M. J. Fruchterman and E. M. Reingold. Graph drawing by forcedirected placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

28. A. Gonzalez and R. Knight. Advancing analytical algorithms and pipelines for billions of microbial sequences. *Current Opinion in Biotechnology*, 23(1):64–71, February 2012.

29. C. Haythornthwaite. Social network analysis: An approach and technique for the study of information exchange. *Library & Information Science Research*, 18(4):323–42, 1996.

30. M. E. Hibbing, C. Fuqua, M. R. Parsek and S. B. Peterson. Bacterial competition: surviving and thriving in the microbial jungle. *Nat. Rev. Microbiol.*, 8(1):15–25, Jan 2010.

31. S. M. Huse, D. B. M. Welch, A. Voorhis, A. Shipunova, H. G. Morrison, A M. Eren and M. L Sogin. Vamps: a website for visualization and analysis of microbial population structures. *BMC bioinformatics*, 15(1):41, 2014.

32. C. Huttenhower et al. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, Jun 2012.

33. P. Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, February 1912.

34. M. Jaric, J. Segal, E. Silva-Herzog, L. Schneper, K. Mathee and G. Narasimhan. Better primer design for metagenomics applications by increasing taxonomic distinguishability. *BMC Proc*, 7(Suppl 7):S4, Dec 2013.
35. D. Knights, E. K. Costello and R. Knight. Supervised classification of human microbiota. *Microbiology Reviews*, 35(2):343–359, Oct 2010.
36. D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.
37. Konstantinos T Konstantinidis and James M Tiedje. Towards a genome-based taxonomy for prokaryotes. *Journal of bacteriology*, 187(18):6258–6264, 2005.
38. J. Kuczynski, Z. Liu, C. Lozupone, D. McDonald, N. Fierer and R. Knight. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Meth*, 7(10):813–819, October 2010.
39. D. J. Lane. 16S/23S rRNA sequencing. *Nucleic Acid Techniques in Bacterial Systematics*, pages 125–175, 1991.
40. M. G. Langille et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.*, 31(9):814–821, Sep 2013.
41. P. Lenca, P. Meyer, B. Vaillant and S. Lallich. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, 184(2):610 – 626, 2008.
42. L. Li, C. J. Stoeckert Jr. and D. S. Roos. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13:2178–2189, 2003.
43. J. M. Marrazzo, D. H. Martin, D. H. Watts, J. Schulte, J. D. Sobel, S. L. Hillier, C. Deal and D. N. Fredricks. Bacterial vaginosis: identifying research gaps proceedings of a workshop sponsored by DHHS/NIH/NIAID. *Sex Transm Dis*, 37(12):732–744, Dec 2010.
44. A. C. McHardy and I. Rigoutsos. What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr. Opin. Microbiol.*, 10(5):499–503, Oct 2007.
45. P. McKenna, C. Hoffmann, N. Minkah, P. P. Aye, A. Lackner, Z. Liu, C. A. Lozupone, M. Hamady, R. Knight and F. D. Bushman. The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. *PLoS Pathog.*, 4(2):e20, Feb 2008.
46. F. Meyer et al. The metagenomics rast server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1):386, 2008.
47. A. Morris et al. Comparison of the respiratory microbiome in healthy nonsmokers and smokers. *Am. J. Respir. Crit. Care Med.*, 187(10):1067–1075, May 2013.
48. K. E. Nelson et al. A catalog of reference genomes from the human microbiome. *Science*, 328(5981):994–999, May 2010.
49. J. N. Paulson, O. C. Stine, H. C. Bravo and M. Pop. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods*, 10(12):1200–1202, Dec 2013.
50. J. Peterson et al. The NIH Human Microbiome Project. *Genome Res.*, 19(12):2317–2323, Dec 2009.
51. E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies and F. O. Glockner. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, 35(21):7188–7196, 2007.

52. J. Qin et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, Mar 2010.

53. P. D. Schloss et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541, 2009.

54. C. E. Shannon and C. E. Shannon. The mathematical theory of communication. 1963. *MD Comput*, 14(4):306–317, 1997.

55. C. D. Sibley, M. E. Grinwis, T. R. Field, C. S. Eshaghurshan, M. M. Faria, S. E. Dowd, M. D. Parkins, H. R. Rabin and M. G. Surette. Culture enriched molecular profiling of the cystic fibrosis airway microbiome. *PLoS ONE*, 6(7):e22702, 2011.

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

57. SJ Sogin, ML Sogin and CR Woese. Phylogenetic measurement in procaryotes by primary structural characterization. *Journal of molecular evolution*, 1(2):173–184, 1972.

58. J. Tamames and A. Moya. Estimating the extent of horizontal gene transfer in metagenomic sequences. *BMC Genomics*, 9:136, 2008.

59. R. L. Tatusov, E. V. Koonin and D. J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, October 1997.

60. P. J. Turnbaugh and J. I. Gordon. The core gut microbiome, energy balance and obesity. *J. Physiol. (Lond.)*, 587(Pt 17):4153–4158, Sep 2009.

61. P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight and J. I. Gordon. The human microbiome project. *Nature*, 449(7164):804–810, Oct 2007.

62. S. Turner, K. M. Pryer, V. P. W. Miao and J. D. Palmer. Investigating Deep Phylogenetic Relationships among Cyanobacteria and Plastids by Small Subunit rRNA Sequence Analysis1. *Journal of Eukaryotic Microbiology*, 46(4):327–338, 1999.

63. X. Wei, D. N. Kuhn and G. Narasimhan. Degenerate primer design via clustering. In *CSB*, pages 75–83. IEEE Computer Society, 2003.

64. W. G. Weisburg, S. M. Barns, D. A. Pelletier and D. J. Lane. 16s ribosomal dna amplification for phylogenetic study. *Journal of Bacteriology*, 173(2):697–703, 1991.

65. S. A. West, S. P. Diggle, A. Buckling, A. Gardner and A. S¿ Griffin. The social lives of microbes. *Annual Review of Ecology, Evolution, and Systematics*, 38(1):53–77, 2007.

66. C. R. Woese, R. Gutell, R. Gupta and H. F. Noller. Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids. *Microbiol. Rev.*, 47(4):621–669, Dec 1983.

67. T. Yamada, I. Letunic, S. Okuda, M. Kanehisa and P. Bork. iPath2.0: interactive pathway explorer. *Nucleic Acids Res*, 39(Web Server issue):W412–5, 2011.

68. M. J. Zaki and W. Meira Jr. Fundamentals of data mining algorithms, 2011.

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

70. D. R. Zeigler. Gene sequences useful for predicting relatedness of whole genomes in bacteria. *International journal of systematic and evolutionary microbiology*, 53(6):1893–1900, 2003.

71. W. Zhang, S. J. Emrich and E. Zeng. A two-stage machine learning approach for pathway analysis. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2010,*, pages 274–279, 2010.

72. W. Zhang, E. Zeng, D. Liu, S. Jones and S. J. Emrich. A machine learning framework for trait based genomics. In *Proceedings of IEEE 2nd International Conference on Computational Advances in Bio and Medical Sciences, ICCABS 2012,*, pages 1–6, 2012.

73. W. Zhang, E. Zeng, D. Liu, S. Jones and S. J. Emrich. Mapping genomic features to functional traits through microbial whole genome sequences. *To appear in the International Journal of Bioinformatics Research and Applications,*, 2012.

74. W. Zhang, E. Zeng, S.J. E., J. Livermore, D. Liu and S.E. Jones. Predicting bacterial functional traits from whole genome sequences using random forest. In *Computational Advances in Bio and Medical Sciences (ICCABS), 2013 IEEE 3rd International Conference on*, pages 1–2, June 2013.