CHAPTER EIGHTEEN

Computational Methods for High-Throughput Comparative Analyses of Natural Microbial Communities

Sarah P. Preheim[*], Allison R. Perrotta[†], Jonathan Friedman[‡,§], Chris Smilie[§], Ilana Brito[*], Mark B. Smith[¶], Eric Alm[*,1]

[*]Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[†]Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[‡]Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[§]Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[¶]Microbiology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[1]Corresponding author: e-mail address: ejalm@mit.edu

Contents

Abstract

One of the most widely employed methods in metagenomics is the amplification and sequencing of the highly conserved ribosomal RNA (rRNA) genes from organisms in complex microbial communities. rRNA surveys, typically using the 16S rRNA gene for prokaryotic identification, provide information about the total diversity and taxonomic affiliation of organisms present in a sample. Greatly enhanced by high-throughput sequencing, these surveys have uncovered the remarkable diversity of uncultured organisms and revealed unappreciated ecological roles ranging from nutrient cycling

Methods in Enzymology, Volume 531
ISSN 0076-6879
http://dx.doi.org/10.1016/B978-0-12-407863-5.00018-6

to human health. This chapter outlines the best practices for comparative analyses of microbial community surveys. We explain how to transform raw data into meaningful units for further analysis and discuss how to calculate sample diversity and community distance metrics. Finally, we outline how to find associations of species with specific metadata and true correlations between species from compositional data. We focus on data generated by next-generation sequencing platforms, using the Illumina platform as a test case, because of its widespread use especially among researchers just entering the field.

1. INTRODUCTION

Prokaryotic cells make up the majority of biomass on the planet (Whitman, Coleman, & Wiebe, 1998) and influence every ecosystem from the world's oceans to the human gut. Metagenomics approaches, including amplicon-based techniques targeting the conserved small subunit ribosomal RNA genes (commonly 16S rRNA in prokaryotes), facilitate research of the structure, function, and stability of microbial communities (Amann, Ludwig, & Schleifer, 1995; Pace, 1997; Woese, Kandler, & Wheelis, 1990).

Although 16S rRNA surveys of microbial communities are widely used to characterize the composition and diversity of microorganisms present in a sample, there are many problems associated with transforming 16S rRNA sequences into proxies for species, given the ambiguity of bacterial species. Additionally, although there is no consensus for what constitutes a microbial species, the operational taxonomic unit (OTU) is a widely used construct of clustered sequence data that approximates "species" in subsequent analysis steps. OTUs are typically formed by grouping together sequences that are within a defined genetic cut-off, even though such grouping does not accurately reflect current opinion about microbial species (Gevers et al., 2005), typically overestimates sample diversity (Huse, Welch, Morrison, & Sogin, 2010), and inappropriately assumes that diverse organisms evolve at similar rates. Single base changes and chimeras can create artificial diversity during both PCR and sequencing (Qiu et al., 2001; Zhou et al., 2011). Furthermore, prior to sequencing, the preparation of samples incorporates many known biases due to differential access to microbial DNA for amplification (Forney, Zhou, & Brown, 2004). Steps should be taken to reduce these errors and biases whenever possible—from library preparation to computational processing and analysis.

Despite these limitations, 16S rRNA surveys have emerged as the most popular approach to study microbial communities, largely due to the speed,

convenience, and low cost of this analysis as a result of next-generation sequencing technologies. These technologies allow researchers to compare hundreds of community profiles marked with molecular barcodes in a single sequencing run. As such, rRNA surveys have been used to complete comprehensive body site sampling in healthy adults through the Human Microbiome project (Huttenhower, Gevers, Knight, Abubucker, Badger, Chinwalla et al., 2012) and to sequence hundreds of thousands of environmental samples from across the world through the Earth Microbiome project (http://www.earthmicrobiome.org/), as well as countless other projects characterizing microbial community variation in space or time.

This chapter outlines the bioinformatics approaches used to compare microbial communities using high-throughput sequencing technologies, focusing on data generated by the Illumina sequencing platform (San Francisco, CA). We use an example of a 16S rRNA survey generated using Illumina sequencing, although many of the principles are relevant to other platforms. We focus on the following topics in the comparative analyses of microbial communities:

- Background
- Sequence processing
- Comparative analyses

2. SEQUENCING TERMINOLOGY

Identifying microorganisms present in a natural community using a sequencing-based 16S rRNA survey begins with the construction of a library. A library is a collection of DNA fragments that represents the sequence diversity in a sample. These fragments are enriched from the rest of the community genomic DNA by PCR using primers which match the microbial population or gene of interest. For Illumina sequencing, the complete molecular construct contains the amplified genomic DNA, sequences that identify the sample it originated from (i.e., index or barcode sequences) and sequences required by the platform to adhere library fragments to the solid matrix and provide a priming site for the sequencing reaction. Adding a barcode sequence to the molecular construct identifying which sample the library originated from allows for hundreds of libraries to be sequenced in the same reaction, commonly called multiplexing. Illumina offers paired-end sequencing, which provides a sequence for the forward and reverse strands of a template, enabling resequencing of very short template

sequences for improved accuracy, or longer effective reads with enhanced positional information for longer template sequences.

3. SEQUENCE PROCESSING

Methodological artifacts must be minimized through quality filtering, while computational analyses require sequences to be organized into appropriate groups, and biological interpretation is facilitated by assignment of a meaningful taxonomic label to each group. Errors at any of these steps can lead to inappropriate interpretation. There are many different methods for analyzing and grouping sequence data. We will outline our approach, highlighting alternatives to commonly applied techniques when appropriate. Both the mothur (Schloss et al., 2009) and QIIME (Caporaso et al., 2010) software packages provide a comprehensive suite of tools for 16S rRNA analysis, including fastq quality filtering, OTU assignment, and classification for comparative analyses. The standard protocols associated with each package include many of the steps outlined below to process 16S rRNA sequence data.

We will use an example dataset throughout this chapter for illustration purposes. These libraries were prepared using a two-step PCR approach with primers targeting rRNA sequence positions 515 and 786 of the *Escherichia coli* genome (region V4) similar to a previously published method (Knight et al., 2011). The raw data and library construction protocol can be downloaded from the distribution-based clustering (DBC) Web site (https://github.com/spacocha/Distribution-based-clustering/blob/master/MIE_dir/). Libraries were sequenced on an Illumina MiSeq instrument (Illumina, San Diego, CA) with 2×251 paired end reads with an 8 base indexing run. Our samples were multiplexed with additional samples, including a defined mock community. 1.1 million reads were obtained from 20 freshwater samples and approximately 230,000 from the mock community.

Below, we will outline steps necessary to process the sequencing data into manageable units for further analysis, although our most recent protocol for sequence processing and OTU calling with DBC can be found at https://github.com/spacocha/Distribution-based-clustering/. The basic outline includes:

- Quality filtering, dereplication, and trimming of sequences
- Creating OTUs
- Removing artifacts and assigning informative labels to OTUs

3.1. Quality filtering, dereplication, and trimming of sequences

The overall quality of the run was visualized from the fastq file with the program FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (Fig. 18.1). Across any sequencing platform, the average quality decreases with the length of DNA that is sequenced. One way to increase the quality of bases toward the end of a read and increase the read length is to use overlapping paired-end sequencing, commonly applied with Illumina. The observation of the same base in the reverse orientation can be used to reduce the error rate. Paired-end sequences can be joined with programs such as SHE-RA (Rodrigue et al., 2010), PandaSeq (Masella, Bartram, Truszkowski, Brown, & Neufeld, 2012), or within mothur using make.contigs. Alternatively, nonoverlapping, paired-end sequences have been concatenated and analyzed together (Werner, Zhou, Caporaso, Knight, & Angenent, 2012). Otherwise, two separate analyses can be done independently on the 5′ and 3′ reads. Although the information content is typically increased with overlapping reads, our protocol is standardized to use information from each end independently.

Raw data are first quality filtered, dereplicated, and trimmed. Improvement in diversity estimates can be gained after quality filtering (Bokulich et al., 2013) although overfiltering can lead to a loss of information. We use the mock community to determine the optimal amount of quality filtering. Our mock community is created from purified, linearized plasmids containing 16S rRNA sequences from a clone library. Thus, the concentration and sequence of each template is known, which can be used to evaluate quality-filtering performance. We use QIIME's split_libraries_fastq.py to demultiplex and quality filter the sequences, although other programs can be used to accomplish a similar filtering. [Note: we used a custom perl script (fastq2Qiime_barcode.pl) to modify the output of the Illumina data to correspond with the format required by split_libraries_fastq.py. This may not be necessary in all cases, depending on the form of the data from the sequencer and the position of the barcode sequence.]

Overfiltering can skew the abundance ratios in 16S-rRNA amplicon data and underfiltering allows many more errors into the analysis. We have identified two parameters in QIIME's split_libraries_fastq.py that increase the quality of the resulting data: (1) the maximum number of bad bases allowed before trimming (–max_bad_run_length) and (2) the threshold quality score (–last_bad_quality_char).

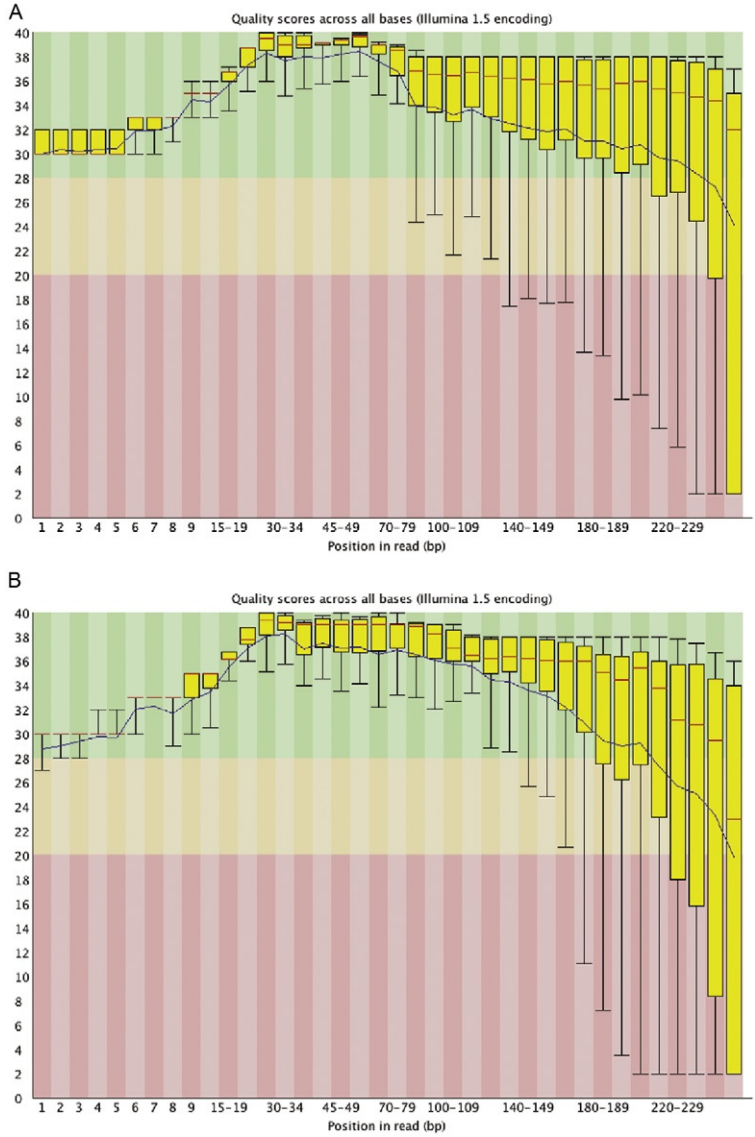

Figure 18.1 The per base sequence quality as determined by FastQC for (A) the 5′ (forward) read and (B) 3′ (reverse) end of a paired-end 250 × 250 MiSeq (Illumina) sequencing run. Raw base qualities are highest from approximately 20 (after the primer) to 70 bps or so.

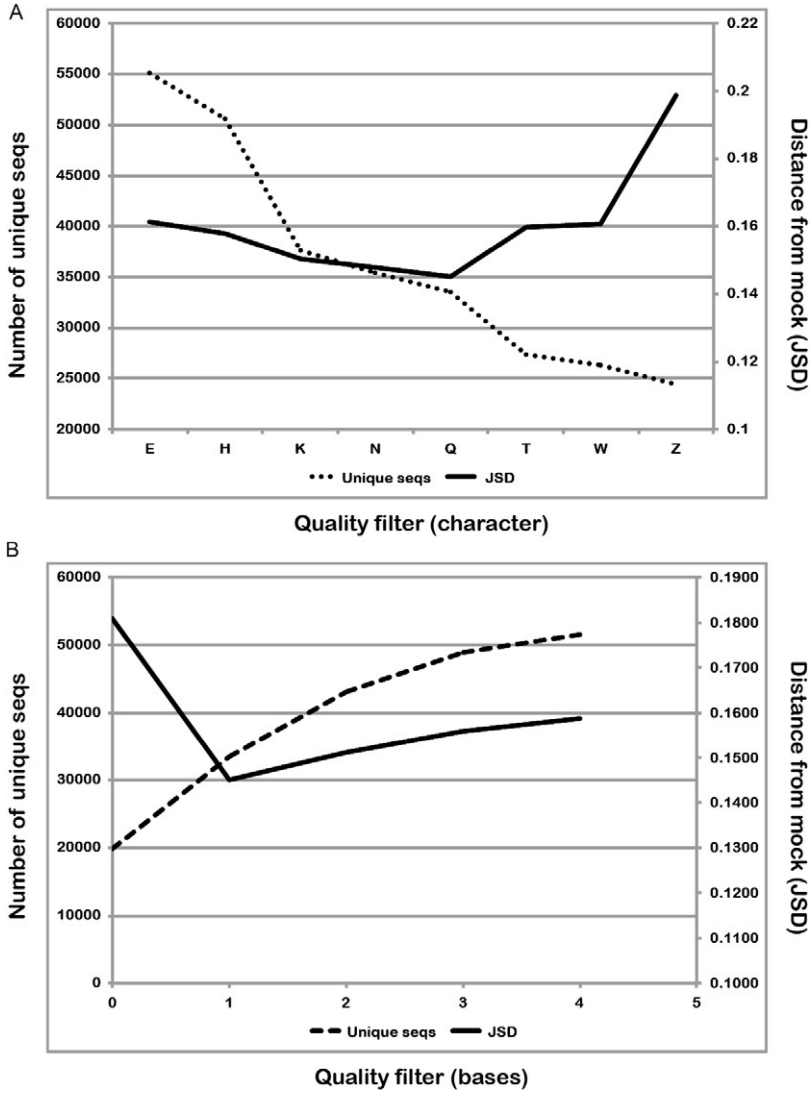

Figure 18.2 Intermediate quality filtering results in a dataset that is most similar to the input (defined, mock) community. (A) The solid line shows the distance of the resulting data from the input sequences after quality filtering using various quality thresholds (right axis). The dotted line shows the total number of unique sequences remaining after every filtering step (left axis). X-axis is the ASCII characters (Illumina's version 1.5 encoding) used in the split_library_fastq.py program of QIIME (–last_bad_char), representing different filtering stringencies ranging from E to Z, corresponding to Phred scores of 5 to 26 and probabilities of error from 0.3165 to 0.0025, respectively. (B) The solid line shows the distance of the resulting data from the input sequencing after allowing a specified number of bases to fall below the quality threshold before truncating the sequence (right axis). The dashed line shows the total number of unique sequences remaining after each step (left axis). X-axis is the number of consecutive bases with quality scores below the quality threshold that are allowed before truncating the sequence in the split_libraries_fastq.py program of QIIME (–max_bad_run_length). JSD, Jensen Shannon divergence.

We vary these two parameters and compare the resulting sequences to the known, mock community to identify the optimal filtering criteria. We found that with a quality threshold of Phred 17 (–last_bad_char Q; error rate ~0.02) and truncating the sequence when two bases fall below the quality filter (–max_bad_run_length 2) results in a mock community that has the smallest distance from our input (Fig. 18.2).

3.2. Creating OTUs

Grouping sequences into OTUs is important, because all downstream analyses are dependent on them, but there is little consensus on how this should be done. Some methods group sequences into clusters with sequences from a well-curated database as the "seed" (e.g., QIIME's "closed-reference" clustering). While this method is quick and convenient, it is not recommended because it can discard novel sequences in the dataset, even when they are abundant. This can be overcome with the "open-reference" approach (http://qiime.org/tutorials/open_reference_illumina_processing.html), where novel sequences are instead retained. Other programs cluster sequences *de novo*, forming OTUs based on their relation to other sequences in the dataset [e.g., average-linkage clustering in mothur (Schloss et al., 2009), heuristics such as USEARCH (http://www.drive5.com/usearch/) and ESPRIT (Sun et al., 2009)].

We favor an approach called DBC. DBC is an alternative method of forming OTUs *de novo* that is accurate and discriminating. This approach is different from other clustering methods because it uses the additional information contained in the distribution of sequences across libraries. Using the distribution and genetic information, DBC can reduce much of the false diversity created by sequencing error for a mock community. It can also provide the power to identify differentially distributed sequences that would otherwise be clustered together because of their sequence similarity. A manuscript describing the details of DBC will appear in the near future (Preheim, Perrotta, Martin-Platero, & Gupta, submitted for publication) and the most up-to-date, detailed protocol for running this algorithm can be found at the Web site (https://github.com/spacocha/Distribution-based-clustering).

DBC works to group genetically similar sequences that have a similar distribution across samples. The sequence by library matrix is used along with the pairwise genetic distance file to inform the clustering. The DBC algorithm can accept both aligned and unaligned phylogenetic distances, using the lowest of both to improve OTU calling. We used mothur align.seqs to

create an alignment to a reference dataset (http://www.mothur.org/wiki/Silva_reference_alignment), generated from the mothur formatted Silva alignment and trimmed to the positions of our amplicon. Typically, we generate a distance matrix using FastTree—make_matix (Price, Dehal, & Arkin, 2009), although other distance programs can be used. The distribution similarity is evaluated for two sequences using the chi-squared test and the Jensen–Shannon divergence.

DBC is typically run in parallel for datasets of any significant size. This requires that the data are preclustered to a very low identity (~90% identity clusters), and each cluster is analyzed independently to form the final OTUs. Thus, the DBC algorithm evaluates whether to divide sequences found within each 90% cluster into additional OTUs based on the distribution and genetic information provided. OTUs called within the 90% clusters independently are merged into a final cluster list.

3.3. Artifact removal and classification of OTUs

Artifacts can inflate the total number of OTUs and diversity estimates. DBC can reduce the impact of sequencing errors, but non-rRNA sequences and chimeras may still be present which will inflate total diversity. We follow the recommendations on the mothur Web site (Schloss et al., 2009; http://www.mothur.org/wiki/MiSeq_SOP) and use align.seqs command with the modified reference alignment (Schloss, 2009). We discard sequences that do not start and end at the same position in the alignment using mothur: screen.seqs command with start and end options. Chimeras create false diversity, but are not easily distinguished from true rRNA sequences. They should be identified with specialized software such as UCHIME (Edgar, Haas, Clemente, Quince, & Knight, 2011) and removed from further analysis. In the example dataset, 6.2% of the resulting OTUs were chimeric, using the forward read only (35% from the overlapped reads).

Assigning a taxonomic label to each OTU created during clustering can help to interpret downstream analyses, especially when comparing against the published literature. Overlapped, paired-end sequences facilitate taxonomy assignment because their greater length increases the power of taxonomic classification. When overlapping is not possible and short reads must be used, taxonomic classification can be optimized for the specific design and read length (Mizrahi-Man, Davenport, & Gilad, 2013). The Ribosomal Database Project (RDP) naive Bayesian Classifier (Wang, Garrity, Tiedje, & Cole, 2007) is one of the most popular approaches, although others are available, such as Greengenes (DeSantis et al., 2006). The RDP classifier typically performs well

on assigning short sequences at the family level (Claesson et al., 2010). However, when applying to short sequences, it is best to refine the training set that the RDP classifier uses for classification to include only the sequenced region as other regions may be characterized by quite different compositions (Werner et al., 2012). Concatenated, nonoverlapping sequences can be classified with Rtax (Soergel, Dey, Knight, & Brenner, 2012), which can be implemented through QIIME.

We typically match OTU representatives with Sanger-sequenced environmental clones from the same sample to increase the phylogenetic information associated with the most abundant organisms. This may be particularly useful in environments that are not well studied (i.e., the bottom of a lake), as opposed to more commonly analyzed samples (i.e., human microbiome). Although this type of analysis is not high-throughput, it does enhance the information gained from either the Sanger-sequenced or the Illumina-sequenced libraries alone and can be a powerful additional tool for the analysis of the most abundant sequences.

After completing these steps, the representative sequences, an OTU by library matrix and associated phylogenetic information from either the matching Sanger clones or the RDP classification can be used in downstream analysis.

4. COMMUNITY ANALYSIS

Because there are many options available to compare microbial communities, it is critical to match the tools utilized with the specific biological questions being pursued. Data generated from most high-throughput sequencing platforms (e.g., Illumina or 454) will be similar after OTU calling and classification, so many of the tools and reviews of community analyses in relation to pyrosequencing data can be applied to Illumina datasets as well. QIIME community analysis (http://qiime.org/tutorials/tutorial.html) can be used to guide the user through common types of analyses. Therefore, the goal of this section is to identify caveats of a few commonly applied analyses and provide an implementation of appropriate alternatives in areas such as:

- Diversity estimates
- Community distance metrics
- Associations of OTUs with metadata
- Correlation analysis

4.1. Diversity estimates

Estimating the total diversity (often called alpha diversity) within a community is a common first step to summarize community composition. True diversity reflects both the number of species present and the evenness of their distribution. However, because molecular surveys only capture a fraction of all species actually present in a community, all metrics reflect only an estimate of true diversity and will be sensitive to the sampling depth used. This complicates comparisons across samples with different sampling depths as more deeply sequenced samples may spuriously appear to be more diverse, independently of their actual composition. There are several metrics used to measure diversity, including the Shannon and Simpson indices (Hill, 1973), which measure both richness and evenness, and the Chao-1 estimator, which estimates species richness from the rarefaction of observed sequences (Chao, 1984; Chao, Colwell, Lin, & Gotelli, 2009). Shannon and Simpson indices are strongly preferred due to their reduced sensitivity to sampling depth (Haegeman et al., 2013).

To demonstrate the potential for artifacts of sequencing depth to confound meaningful analysis of diversity, we compared the sensitivity of richness and Shannon and Simpson diversity to sampling depth (Fig. 18.3). Using the OTU versus library matrix generated from the DBC algorithm above, we chose a library with 80,047 reads to compute the Chao-1 richness and Shannon and Simpson diversity metrics in PySurvey (sample_diversity with indices=['richness'], methods=['Chao1'] and indices=['shannon', 'simpson']). We then take random subsamples of the full dataset to simulate sequencing experiments with fewer reads, ranging from 1000 reads up to the full set 80,047 reads at 500 read intervals. We compute estimates of the three metrics in 100 iterations of random samples at each sequencing depth. While the Shannon and Simpson estimates of effective species quickly stabilize around 101 and 26, respectively, the species richness estimated by Chao-1 continues to rise with sequencing depth. Thus, while indices based on both species richness and abundance are relatively insensitive to sequencing depth, indices such as the Chao-1 estimator that measure only species richness remain highly sensitive to sequencing depth, even at high coverage. Although the Simpson index is less dependent on sequencing depth, it is also less sensitive to rare OTUs. For this reason, Shannon is the preferred metric for rRNA community analysis—achieving a balance of sensitivity to rare OTUs without undue dependency on sequencing depth.

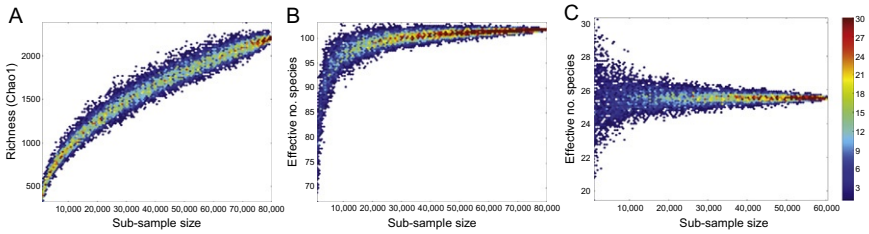

Figure 18.3 Richness is a function of sequencing depth, whereas Shannon and Simpson diversity indices stabilize with sequencing depth. (A) Chao-1 estimator (Chao1) of richness for the same library subsampled to different depths. (B) Effective number of species calculated with the Shannon diversity index (SDI) for the same library subsampled to different depths. Because SDI is computed on a log scale, we transform the SDI to the effective number of species present, which is on linear scale, by computing e^{SDI}. (C) Effective number of species calculated as the inverse of the Simpson diversity index (1/Simpsons diversity index) for the same library subsampled to different depths.

4.2. Community distance metrics

Many applications call for estimating the distance between two distinct communities, a task that is impacted by sampling noise and "compositional" effects, which refer to biases that arise from the fact that only relative abundances rather than absolute values are measured (i.e., relative frequencies for all species must sum to 100%). Measures of distance between communities can be categorized based on the properties of the communities they account for: incidence (presence/absence), abundance (absolute or relative), and/or phylogenetic relatedness. Generally, incidence-based distances, where all components contribute equally, are especially challenging to estimate reliably (Chao, Chazdon, Colwell, & Shen, 2005), whereas measures which give larger weights to more abundant components are more readily estimated (Bent & Forney, 2008). This is analogous to the alpha diversity example in Fig. 18.3, where species richness is much more challenging to estimate than the Shannon entropy or the Simpson diversity.

To overcome the sampling noise-associated 16S rRNA survey data, we propose adopting the square root of the Jensen–Shannon divergence $JSD^{1/2}$, defined as:

$$d_{JS}^2\left(\underline{x},\underline{y}\right) = \frac{1}{2}\sum_{i=1}^{D}\left[x_i \log_2 \frac{x_i}{M_i} + y_i \log_2 \frac{y_i}{M_i}\right],$$

$$M_i = \frac{x_i + y_i}{2}$$

JSD has a clear information theoretic interpretation and is commonly used to measure differences between distributions. In this context, it gives the logarithm base 2 of the effective number of distinct communities (communities that have no species in common) and is the beta component of the Shannon entropy (Jost, 2006). The square root is of JSD is used, as it transforms JSD into a proper metric (Endres & Schindelin, 2003), enhancing interpretability. JSD accounts for the problem of compositional effects, as it is a measure of divergence between distributions, and maintains accurate and robust performance on datasets that confound Euclidean distances.

JSD was used to compare the samples from a stratified lake. Counts were normalized [fracs_f=ps.normalize(counts)] and the $JSD^{1/2}$ was used to compare communities from different extraction techniques and different depths in a stratified lake [D=ps.dist_mat (fracs_f , axis='rows', metric='JSsqrt')] and plotted with PySurvey [ps.plot_dist_heatmap (D, plot_rlabels=True, rlabel_width=0.06)]. Two major zones were identified, as the longest branches in the tree, dividing the upper (1–7 m) and lower (9–19 m) samples. Additionally, four total subzones were identified (sub1, 1–5 m; sub2, 7 m; sub3, 9–11 m; sub4, 13–19 m). Although the different extraction techniques (MoBio vs. Qiagen) tended to cluster with other samples from the same extraction technique within subzones sub1 and sub4, samples otherwise clustered by depth (Fig. 18.4).

4.3. Associations of OTUs with metadata

Multivariate statistical techniques have been developed specifically to handle high-dimensional data, including Principal Coordinates Analysis. These techniques work by first reducing the dimensionality of the data by identifying principal "axes" of differentiation among communities, which represent distinct combinations of bacterial species. These tools are particularly effective when the variable of interest (e.g., pH, disease state) is associated with major changes in community structure, but are less effective at detecting subtle variations in community structure. Furthermore, they have trouble pinpointing the specific bacterial species that drive these associations. Recently, statistical learning techniques have been employed to detect associations between bacterial species and environmental metadata. Statistical learning has many advantages over multivariate statistical approaches, including the ability to detect both major and minor variations in microbial community structure, the ability to detect nonlinear associations involving combinations of bacterial species, and the ability to pinpoint the specific bacterial species that underlie these associations.

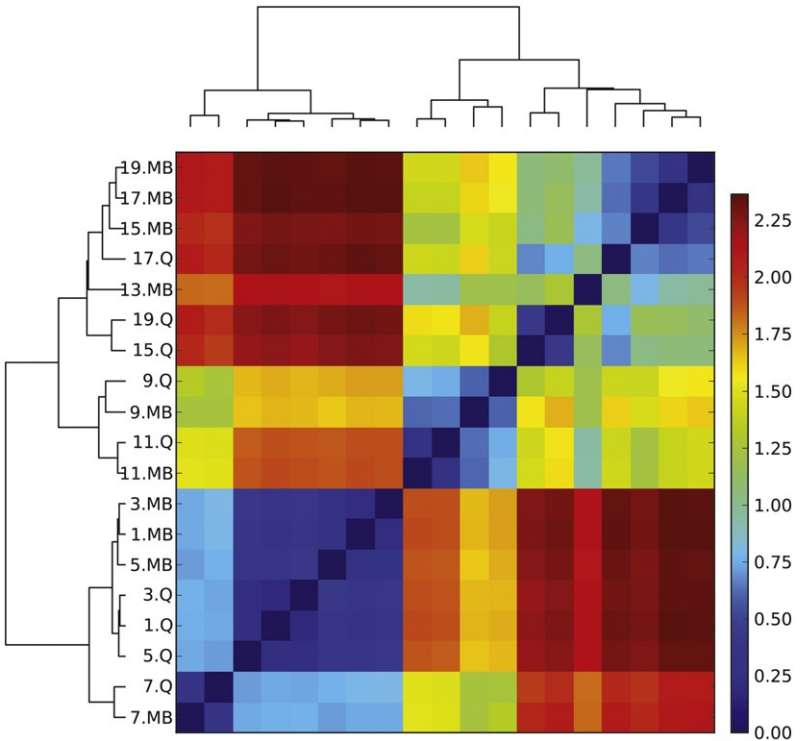

Figure 18.4 A heatmap depicting different biogeochemical zones of a stratified lake. The bacterial community for each depth was compared to all other libraries using the square root of the Jensen–Shannon divergence ($JSD^{1/2}$) and plotted as a heatmap. Bacterial communities cluster into two major zones (upper, oxic and lower, anoxic zone) as shown by the largest branch on the dendrogram (i.e., tree) and four subzones. Libraries are labeled with the depth (in meters) and the extraction technique. MB, MoBio Power Water DNA extraction kit; Q, Qiagen Blood and Tissue Kit. For example, 1.MB refers to 1 m depth in the water extracted with the MoBio kit and 19.Q refers to 19 m depth with the Qiagen kit. The dendrogram above and to the left of the heatmap are identical.

One statistical learning technique that is well suited to microbial community analysis is Random Forest classification, which is implemented in the SLIME software package. For example, SLIME was recently used to discriminate patients with inflammatory bowel disease from healthy controls and to identify the specific bacterial taxa underlying this association (Papa et al., 2012). SLIME can detect associations using categorical or quantitative metadata and runs permutation tests to assess statistical significance. It is best used for relatively large datasets where the total number of samples exceeds 50. The input for this analysis is a table of OTU counts and corresponding metadata. It will proceed to test the significance of every column of metadata, reporting the *P*-value of each association, the most important bacterial taxa, and advanced metrics such as the area under the curve and other classification results.

4.4. Correlation analysis

Microorganisms interact with each other in various ways, through competition for shared resources, antagonistic mechanisms (e.g., antibiotic production), or in mutual or symbiotic associations. Correlation analysis can identify possibly interacting OTUs by identifying groups that change across a large sample in a similar way. However, correlations from low-diversity samples are often artifacts related to the compositional nature of sampling and sequencing (Friedman & Alm, 2012). For example, variation in a single, dominant OTU causes all other OTUs to have a spurious negative correlation to this dominant strain, although their absolute abundance may not have changed. As a result, standard correlation metrics (Pearson, Spearman, etc.) should not be used to compare OTUs. Instead, SparCC (https://bitbucket.org/yonatanf/sparcc) can be used to infer correlations from compositional data with high accuracy, even in low-diversity samples, and can be thought of as a replacement for Pearson correlations. Nonetheless, correlations do not imply causality and can arise when two OTUs respond to a third unmeasured environmental factor. Thus, additional experiments are needed to verify results from SparCC or other similar methods.

We used SparCC to identify correlated OTUs using the OTU by library matrix (created as stated earlier) after filtering out OTUs with fewer than 100 counts (-c 100) across at least two libraries (-s 2) using QIIME's filter_otu_table.py (removing the first line of the resulting table). We then ran SparCC on the data, using 50 bootstraps and a "one-sided" test. Interestingly, of the two pairs with correlations of 0.7, one is classified as *Saprospiraceae* (ID0000005M), which have been found attached

to filamentous bacteria (Xia, Kong, Thomsen, & Nielsen, 2008). The other bacteria could not be classified (ID0000388M). This analysis provides the underlying information to identify important interactions in the environment.

5. SUMMARY

In this chapter, we provide an overview of the methods used to interpret high-throughput microbial community analysis, with a focus on data from the widely used Illumina sequencing platform. We explain how to transform Illumina sequence data into meaningful units for further analysis, including a specialized OTU calling method (DBC). We discuss why Shannon diversity is a more appropriate measure of diversity than richness and use the $JSD^{1/2}$ as a metric for distance between microbial communities. Finally, we use Random Forest classification to identify associations between specific OTUs and metadata and use SparCC to identify true correlations from proportional composition data. Armed with these tools and the increasing power of high-throughput sequencing, we expect that researchers will continue to discover exciting new connections between bacterial communities and their diverse environments.

REFERENCES

Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic identification and in-situ detection of individual microbial-cells without cultivation. *Microbiological Reviews*, *59*(1), 143–169.

Bent, S. J., & Forney, L. J. (2008). The tragedy of the uncommon: Understanding limitations in the analysis of microbial diversity. *ISME Journal*, *2*(7), 689–695.

Bokulich, N. A., Subramanian, S., Faith, J. J., Gevers, D., Gordon, J. I., Knight, R., et al. (2013). Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing. *Nature Methods*, *10*(1), 57–59.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5), 335–336.

Chao, A. (1984). Nonparametric-estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, *11*(4), 265–270.

Chao, A., Chazdon, R. L., Colwell, R. K., & Shen, T. J. (2005). A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, *8*(2), 148–159.

Chao, A., Colwell, R. K., Lin, C. W., & Gotelli, N. J. (2009). Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*, *90*(4), 1125–1133.

Claesson, M. J., Wang, Q. O., O'Sullivan, O., Greene-Diniz, R., Cole, J. R., Ross, R. P., et al. (2010). Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Research*, *38*(22), e200.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, *72*(7), 5069–5072.

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

Endres, D. M., & Schindelin, J. E. (2003). A new metric for probability distributions. *IEEE Transactions on Information Theory*, *49*(7), 1858–1860.

Forney, L. J., Zhou, X., & Brown, C. J. (2004). Molecular microbial ecology: Land of the one-eyed king. [Review]. *Current Opinion in Microbiology*, *7*(3), 210–220.

Friedman, J., & Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *Plos Computational Biology*, *8*(9), e1002687.

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., et al. (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology*, *3*(9), 733–739.

Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J., & Weitz, J. S. (2013). Robust estimation of microbial diversity in theory and in practice. *The ISME Journal*, *7*, 1092–1101.

Hill, M. O. (1973). Diversity and evenness—Unifying notation and its consequences. *Ecology*, *54*(2), 427–432.

Huse, S. M., Welch, D. M., Morrison, H. G., & Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology*, *12*(7), 1889–1898.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., et al. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207–214.

Jost, L. (2006). Entropy and diversity. [Editorial Material]. *Oikos*, *113*(2), 363–375.

Knight, R., Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., et al. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, 4516–4522.

Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G., & Neufeld, J. D. (2012). PANDAseq: PAired-eND Assembler for Illumina sequences. *BMC Bioinformatics*, *13*, 31.

Mizrahi-Man, O., Davenport, E. R., & Gilad, Y. (2013). Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: Evaluation of effective study designs. *PLoS One*, *8*(1), e53608.

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, *276*(5313), 734–740.

Papa, E., Docktor, M., Smillie, C., Weber, S., Preheim, S. P., Gevers, D., et al. (2012). Non-invasive mapping of the gastrointestinal microbiota identifies children with inflammatory bowel disease. *PLoS One*, *7*(6), e39242.

Preheim, S. P., Perrotta, A. R., Martin-Platero, A. M., Gupta, A., & Alm, E. J. Distribution-based clustering: Using ecology to refine the operational taxonomic unit (Accepted, *Applied and Environmental Microbiology*).

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, *26*(7), 1641–1650.

Qiu, X. Y., Wu, L. Y., Huang, H. S., McDonel, P. E., Palumbo, A. V., Tiedje, J. M., et al. (2001). Evaluation of PCR-generated chimeras: Mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Applied and Environmental Microbiology*, *67*(2), 880–887.

Rodrigue, S., Materna, A. C., Timberlake, S. C., Blackburn, M. C., Malmstrom, R. R., Alm, E. J., et al. (2010). Unlocking short read sequencing for metagenomics. *PLoS One*, *5*(7), e11840.

Schloss, P. D. (2009). A high-throughput DNA sequence aligner for microbial ecology studies. *PLoS One, 4*(12).

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541.

Soergel, D. A. W., Dey, N., Knight, R., & Brenner, S. E. (2012). Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences. *ISME Journal*, *6*(7), 1440–1444.

Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M. L., McKendree, W., et al. (2009). ESPRIT: Estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic Acids Research*, *37*(10), e76.

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

Werner, J. J., Koren, O., Hugenholtz, P., DeSantis, T. Z., Walters, W. A., Caporaso, J. G., et al. (2012). Impact of training sets on classification of high-throughput bacterial 16s rRNA gene surveys. *ISME Journal*, *6*(1), 94–103.

Werner, J. J., Zhou, D., Caporaso, J. G., Knight, R., & Angenent, L. T. (2012). Comparison of illumina paired-end and single-direction sequencing for microbial 16S rRNA gene amplicon surveys. *ISME Journal*, *6*(7), 1273–1276.

Whitman, W. B., Coleman, D. C., & Wiebe, W. J. (1998). Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(12), 6578–6583.

Woese, C. R., Kandler, O., & Wheelis, M. L. (1990). Towards a natural system of organisms—Proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(12), 4576–4579.

Xia, Y., Kong, Y. H., Thomsen, T. R., & Nielsen, P. H. (2008). Identification and ecophysiological characterization of epiphytic protein-hydrolyzing Saprospiraceae ("Candidatus epiflobacter" spp.) in activated sludge. *Applied and Environmental Microbiology*, *74*(7), 2229–2238.

Zhou, H. W., Li, D. F., Tam, N. F. Y., Jiang, X. T., Zhang, H., Sheng, H. F., et al. (2011). BIPES, a cost-effective high-throughput method for assessing microbial diversity. *ISME Journal*, *5*(4), 741–749.