BRIEFINGS IN BIOINFORMATICS. VOL 13. NO 4. 420–429
Advance Access published on 4 February 2012

doi:10.1093/bib/bbr080

Measuring the microbiome: perspectives on advances in DNA-based techniques for exploring microbial life

James A. Foster, John Bunge, Jack A. Gilbert and Jason H. Moore

Submitted: 23rd September 2011; Received (in revised form): 16th December 2011

Abstract

This article reviews recent advances in 'microbiome studies': molecular, statistical and graphical techniques to explore and quantify how microbial organisms affect our environments and ourselves given recent increases in sequencing technology. Microbiome studies are moving beyond mere inventories of specific ecosystems to quantifications of community diversity and descriptions of their ecological function. We review the last 24 months of progress in this sort of research, and anticipate where the next 2 years will take us. We hope that bioinformaticians will find this a helpful springboard for new collaborations with microbiologists.

Keywords: *microbial ecology; biodiversity; metagenomics; next generation sequencing; microbiome; visual analytics*

INTRODUCTION

We live in a microbial world, with microscopic organisms filling discrete ecosystems in such environments as soil, lakes and oceans, the human gut or skin, and even computer keyboards. Though microbiota include bacteria, archea, viruses and microscopic eukaria, we will consider only bacterial examples in this article. Bacteria comprise most of the Earth's biomass and richness [1]. They dominate ecological functions such as carbon cycling, greenhouse gas emission and oxygen production. Ninety per cent of the cells in a human body are bacterial, as are 99% of the gene transcripts [2]. However, most of the microbial world has been inaccessible to us, a kind of biological 'dark matter', since we do not know how to culture over 97% of all bacteria, and since older cultivation-independent microbial survey techniques such as TRFLP (Terminal Restriction Fragment Length Polymorphism), ARISA (Automated Intergenic Spacer Analysis) and gradient gel electophoresis have significant limitations. 'Next Generation' sequencing technologies have enabled, for the first time, high-throughput microbial sampling [3].

Current microbiome studies extract DNA from a microbiome sample, quantify how many representatives of distinct populations (species, ecological functions or other properties of interest) were observed in the sample, and then estimate a model of the original community. Ambitious projects are underway to catalog microbial life for the entire Earth, the ocean and the human body [4–6]. Surveys of transcriptomes and entire genomes have revealed more than half of all known protein sequences. Existing methods for estimating richness and community structure from observed samples are becoming

Corresponding author. James A. Foster, Department of Biological Sciences, Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844-3051, USA. Tel: +1 208 885 7062; Fax: +1 208 885 7905; E-mail: foster@uidaho.edu

Dr James A. Foster is a Professor in the Department of Biological Sciences and the Institute for Bioinformatics and Evolutionary STudies (IBEST) at the University of Idaho.

Dr John Bunge is Associate Professor in the Department of Statistical Science at Cornell University.

Dr Jack A. Gilbert is an Environmental Microbiologist at Argonne National Laboratory, and Assistant Professor in the Department of Ecology and Evolution at University of Chicago.

Dr Jason H. Moore is the Third Century Professor of Genetics and Community and Family Medicine and Director of the Institute for Quantitative Biomedical Sciences at Dartmouth College.

more refined, improving model estimation, confidence quantification and comparative methods [7–9]. Finally, interactive, visual techniques are emerging with which to explore these complicated data sets prior to formal analysis.

The new sequencing technologies have idiosyncratic strengths and weaknesses, which are not fully understood, and are beyond the scope of this review [10]. Currently, most researchers use the Roche 454 GS-FLX or Illumina GAIIx/HiSeq2000 sequencing platforms. The Roche 454 GS-FLX Titanium can now generate in excess of 1 million reads per run, which takes 23 h, with read lengths up to 1000 bp (average ~500 bp); the average run generates 750 Mbp of sequencing data. The Illumina HiSeq2000 platform can now generate ~4 billion paired-end reads per run (with two flow cells of 1 billion fragments each), which takes 10 days, with (usually) 150 bp paired-end reads to create an ~250-bp product; the average run generates 1 Tbp of sequencing data. Of course, there is wide variation between individual labs for these statistics. Emerging technologies, such as single molecule sequencing and smaller single lab devices are not widely used yet, and Sanger sequencing of large-insert libraries is still significant [11].

Recent bioinformatics advances have significantly improved sequencing and assembly errors detection and correction. Several packages provide pipelines to bring these new algorithms into the lab [12, 13]. Bioinformaticists continue to improve algorithms for detecting specific types of error, such as chimeric sequences [14] and precise but inaccurate reads [15, 16].

In this review, we survey recent advances in genome-based analytical techniques to measure the diversity of complete microbial communities. There are, of course, many other ways for analytical scientists to advance microbiome studies, which we do not review here, such as new quality control methods, large-scale data curation, knowledge mining and novel data-analytic techniques such as metaproteomics and advanced mass spectrometry. So, for working purposes here we consider a 'microbiome' to be a well-defined patch of an ecosystem, such as all bacteria in a prescribed sector of the ocean or all bacteria from a specific body part of several humans. We use microbial ecology terminology rather than statistical conventions, so that a 'population' is a collection of all organisms of a given species, a 'community' is a collection of 'populations' that share a specific ecosystem, and a 'sample' or 'specimen' is a physical extract from a given microbiome. Finally, we limit references for the most part to recent publications that serve as jumping off points for further exploration, rather than a complete literature survey.

In this article, first we discuss studies based on 16S rRNA amplicons. Next, we review analyses of metagenomic and metatranscriptomic data from shotgun sequencing of multiple genomes or genome transcripts. We then consider advances and limitations in statistical techniques for diversity estimation. Then we discuss visual analytics, hypothesis generation by visually exploring these very large sequence data sets. Finally, we speculate on how microbiome studies may change in the next 2 years.

16S RRNA AMPLICON ANALYSIS

Hypervariable regions of individual, highly conserved genes, such as the small ribosomal subunit in noneukaryotes, have served as proxies for species since Woese and Fox [17–19] first used them to demonstrate that archea were a separate kingdom. With new sequencing technologies based on the polymerase chain reaction (PCR) it became possible to sample all the 16S rRNA genes in a specimen without having to isolate and cultivate organisms in order to amplify DNA separately. By tagging specimens with molecular barcodes, labs can multiplex several treatments and controls into a single sequencing run, making it possible to survey and compare different specimens with very few sequencing jobs, dramatically shrinking the time between sample preparation and data analysis and the sequencing costs.

The 16S rRNA gene remains a good but far from ideal molecular marker for microbial diversity, and there is no obvious alternative. 16S rRNA genes from hundreds of thousands of organisms have been fully sequenced and classified [13, 20]. As with all databases, ribosomal databases are growing larger and better, so analysis relying on them can only improve. The secondary structure of the 16S rRNA molecule is well characterized, at least for reference strains, which makes it possible to perform fast, secondary structure driven alignments [21, 22]. However, as with any single gene, the diversity of the 16S rRNA gene does not always reflect phylogenetic relationships or metabolic potentials that are known from other sources [23]. Current studies rarely resolve sequences below the family level (even for known strains) due to limited database depth, though the algorithms themselves are capable of finder resolution. Consequently, results are often reported at the order or even phyla level, even though different species or even strains are likely to have very different roles in microbiomes. Database sequences are surely biased samples of reality, since they assume at least that their targets are amenable to existing sequencing and annotation methodologies. They have been further biased by a historical fixation on potential pathogens and environmental contaminants. However, the 16S rRNA gene is likely to remain the most reliable and broadly applicable marker for some time.

To date, only small 16S rRNA gene fragments, rather than entire genes or genomes, have been amenable to sequencing. Primers exist for hypervariable regions known as V1 through V9, of widely varying lengths and phylogenetic resolution [24]. Different regions, and combinations of regions, have different strengths and weaknesses [25, 26]. Historically, human microbiome surveys typically sample from regions near V3, while environmental surveys often sample from regions near V6, though evidence indicates that V2 and V4 are less error prone and most project in the NIH Human Microbiome Project use the V3–V5 region [27]. As sequencing technologies and protocols improve, projects are sequencing longer regions, such as V3–V5 (from the beginning of V3 to the end of V5) or V6–V9. Eventually, it may become routine to use the entire 16S gene, multiple marker genes, or even entire genomes.

There are two types of algorithms for inferring microbiome diversity and structure from 'clean' sequences, and both have improved greatly in the last 2 years.

Clustering methods group sequences by similarity, computing statistics from the number and size of clusters. Clustering methods are sensitive to how one measures similarity and what similarity threshold one uses [25, 28]. Older distance clustering methods begin by comparing all pairs of sequences, producing massive distance matrices. Newer algorithms compute clusters on the fly, requiring far less computer memory. Clusters are often called Operational Taxonomic Units (OTUs), a term borrowed from systematics, though the basis for clustering does not always reflect organismal phylogeny or functional diversity. Recent studies have shown that, in general, average neighbor clustering (usually at a 97% similarity threshold) following single linkage clustering (usually at a 98% similarity threshold) works better for estimating community diversity than alternatives [16]. Very few algorithms exist that rigorously fit statistical models to sequence data in order to estimate microbiome structure (see below).

Classification methods, on the other hand, weight their analysis with metadata such as estimates of phylogenetic or functional relationships. Increasingly sophisticated algorithms, including Bayesian inference, match experimental sequences to those in existing databases [13, 20, 29], which are continually updated [13, 29]. Classification methods, including phylogeny-informed analyses [30, 31], help with research projects where it is important to know more than the diversity of a microbiome; for instance the number of organisms likely to be related to potential pathogens or the likely functional capacity of a community. UniFrac algorithms estimate between-population (so called 'beta') diversity, informed by estimated phylogenetic divergence between samples [32]. These techniques will improve over time with rapidly improving databases and phylogenetic estimation algorithms. However, they are limited by the very small number of sequenced organisms relative to what exists in nature, by the computational complexity of current phylogenetic estimation algorithms, and by the problematic nature of the species concept for bacteria. Moreover, many organisms in the databases are still unclassified, having been recalcitrant to current taxonomic methods [25, 33].

METAGENOMICS/ METATRANSCRIPTOMICS

Researchers use metagenomic and metatranscriptomic sequencing to explore the functional and expressed potentials of microbial communities. Most studies have performed extensive sequencing of bacterial communities [34]. But viral [35] and eukaryotic [36] communities have also been studied. Indeed, recent metagenomic data analysis is being used to expand the breadth of perceived phylogenetic space [37].

The difficulty of assembling and annotating the data, due to short read lengths, has been the primary challenge to analyzing high-throughput metagenomic/metatranscriptomic data [38]. Assembly is important for the reconstruction of genes and operons for functional assignment and improved annotation of taxonomy [39], but also for re-assembly of whole genomes from metagenomic DNA [40]. Independently of assembly problems, functional annotation is a difficult problem, compounded by the sheer quantity of sequence data. Consequently, automated annotation has become routine, with little or no manual assessment of accuracy [41]. One of the most appropriate ways of defining the accuracy of assembly and annotation of metagenomic data are to use *in silico* simulated data from fragmented genomes [42] or actual fragmented genomic DNA from known organisms [43].

Nonetheless, comparative metagenomics remains one of the most powerful ways to explore gene distribution across different ecosystems [44]. Several tools and technologies exist for comparing functional community dynamics across different metagenomic data sets [45]. Current techniques are limited by difficulties in contextualizing sequencing data with environmental metadata from the target ecosystem [46]. However, techniques are being developed to improve these analyses, once environmental metadata about the niche space in which the community was structured becomes available [47].

It is possible to model complex community dynamics in relation to the chemical and physical dynamics of the ecosystem, even without exhaustive sequence and environmental data. For example, tools exist to derive the abundance of gene/transcript fragments annotated to known enzyme activities from metagenomic and metatranscriptomic data [48]. In addition, bioclimatic models are being developed to extrapolate the responses of bacterial community structure to environmental change, and how this will affect relative changes in the consumption or production of metabolites in an ecosystem [49].

STATISTICS FOR DIVERSITY ESTIMATION

The statistical challenges for microbiome studies are to estimate population richness and diversity, model community structure, quantify uncertainty and compare estimates rigorously [50]. This is true whether the analysis is based on clustering or classification-based methodologies. We divide the relevant procedures into two groups: (i) methods that treat the observed sample as the community and (ii) methods that account for the existence of unobserved (unsampled) organisms or taxa in the community. The former group is represented by procedures such as UniFrac [32]. These methods are extremely useful and informative and are well-documented and implemented in current software (e.g. mothur, QIIME) so we do not address them here. The latter group consists of quantitative, inferential statistical procedures, that is, methods that estimate true but unknown numerical measures of diversity, such as the total taxonomic richness of a community, both observed and unobserved). These methods are described mainly in the theoretical statistical literature, which bioinformatics specialists are less likely to read. So we focus on them in this expository article.

Most current techniques begin with frequency count data, which groups observations into bins and report the number of members of each bin. There are two main approaches to richness estimation from such count data. The *classical* or *frequentist* approach is better represented both in the literature and in available software. Coverage-based nonparametric estimators like Chao and ACE are popular, being simple to compute, and are available in bioinformatics packages such as mothur and QIIME [12, 51]. But they are known to underestimate the true diversity in high-diversity situations, and to behave erratically when outliers are present [50]. Recently, more stable but computationally intensive parametric mixture models have been introduced. Both types of estimate are available in a single package, CatchAll [7]. Further, CatchAll computes several different estimates and returns a ranked comparison of the 'best' analyses for a given data set.

The *Bayesian* approach, in contrast, begins with a prior probability distribution that represents what is known or believed about the diversity before collecting any data. Using Bayes' Theorem, this approach then derives a posterior distribution using the observed data, which yields the final estimate of diversity along with error terms and confidence intervals. There are two ways to define the prior. In 'objective' or 'non-informative' Bayesian analysis one minimizes the amount of information in the prior so that it influences the end result as little as possible; while in *subjective* or *informative* Bayesian analysis the prior expresses the experimenter's beliefs about the diversity, or weights the results according to known factors that are unrelated to the observed data. Both have been studied in the diversity estimation literature, but the objective Bayesian approach is more widely accepted [52, 53]. Indeed it promises to be statistically and computationally stable and flexible, and may well be a strong competitor to the frequentist methods. But at present there is no simple and generally accessible Bayesian diversity estimation software, so we have less applied experience than with the classical approach.

Recently, statistical methods have been developed that adjust estimates according to patterns in or assumptions about the frequency count data. For example, the successive ratios of frequencies (the number of doubletons divided by the number of singletons, tripletons divided by doubletons, etc.) have known statistical properties, which led to a new estimation method (available in CatchAll) [54]. Another example incorporates suspected unreliability of low frequency counts into diversity estimates. Recent analyses of artificially constructed communities with known diversity and structure indicate that existing methods may systematically lead to inflated low frequency counts. Strategies to address such biases include: (i) using a Bayesian prior weighted toward lower diversity values; (ii) reporting lower bounds rather than direct estimates for the total diversity; (iii) statistically separating the projected population into low and high-diversity components and deleting or downweighting the latter and (iv) by pooling low frequency counts up to some cutoff (say, the singletons and doubletons) and re-estimating the total diversity from these left-censored data [55]. All of these strategies are statistically feasible, although not all have been implemented in software [CatchAll includes (ii) and (iii)], and this remains an area of current research.

The next logical step is to move from estimating the diversity of a single community ('population' in the statistical sense) to comparing diversity levels across two or more communities. Given reliable richness estimates for individual communities, it is straightforward to make statistical comparisons of richness between microbiomes. It is considerably more challenging to quantify how much population structure is shared between two or more communities. One common metric for two communities is the Jaccard index, which is the ratio of the number of shared populations to the total number of populations observed. Other between-community diversity metrics include Sørensen, Bray-Curtis and Morisita-Horn [51]. However, these formulae are often used to compare observed samples rather than estimated communities, leading to statistically indefensible practices such as discarding data to 'normalize' samples to the same size. What is lacking is between-community diversity metrics that account for both observed and unobserved populations. This appears to be a challenging statistical problem. Chao *et al.* [8] provided a nonparametric estimator of the true, community-level Jaccard and Sørensen indices. But, few other solutions have been proposed [56].

Finally, microbiome studies need to model or predict richness and diversity using covariate data, such as observable biological, chemical, or other environmental variables. If the response or dependent variable is simply the (estimated) richness then standard statistical modeling techniques such as regression are appropriate. But, modeling diversity and structure, rather than just richness, as a function of the predictors, requires techniques such as canonical correspondence analysis [9].

All these analyses should be based on estimates of unobserved structure, rather than exclusively observed data, since substantial unobserved diversity is typical of microbial ecology studies.

VISUALIZING THE RESULTS

Microbiome data are inherently high dimensional and complex. Suppose the goal of a project is to relate bacterial community structure at a particular body site to clinical observations. A typical data set might include a list of hundreds of bacterial species that are hierarchically organized into different groups, including genera, families, orders, classes and phyla. This is further complicated by information about genes and pathways that are present in each of the bacterial species and how these relate to clinical endpoints. The genomic information of the host, such as demographic data, patient specifics and lifestyle data may also be important. The ultimate challenge is to put these many different layers of information together in a statistical or machine learning analysis to identify clinically useful patterns.

Given this level of data complexity, it is important for the researcher to have tools with which to visualize and explore data. Visual interaction allows the researcher to critically explore the measurements themselves for quality control, for discovering patterns that lead to new hypotheses, and for interpreting results. Also, it is often desirable to communicate results visually to other scientists and clinicians. However, it is challenging to choose the right visualization technique for the right type of data or information, given that there are so many information visualization methods [57].

Several different information visualization methods have been useful for the analysis of microbiome data. For example, heat maps, introduced >50 years ago [58], have become popular and useful for visualizing population structure in large microbial communities and for clusters of expression patterns in genomics [59]. A heat map consists of a 2D grid or matrix of colored squares where each square represents an observation of a variable and the color of the square is proportional to the value of that observation. It is common to order the squares along the two axes with additional categorical data such as bacterial phyla and tissue type. For example, a recent study by Wu *et al.* [60] explored the relationship between long-term dietary patterns and gut microbial enterotypes. This study used Spearman correlations to estimate the association between different nutrients and bacterial genera in 98 healthy volunteers. It summarized the results with a heat map, where each column represented different taxa, each row represented a different nutrient and the color of each square represented the magnitude of the correlation, with darker red representing stronger positive correlations and darker blue representing stronger negative correlations. Wu *et al.* also performed a hierarchical cluster analysis to organize the results into visual patterns that were easier to interpret. For example, the authors found that fat-related nutrients tended to be more similar in the correlations across taxa than other nutrient groups. In addition to heat maps, the authors also used principal components analysis (PCA) to identify linear combinations of gut microbial taxa associated with long-term diet. They used 2D and 3D scatterplots to identify clusters of patients defined by the first two or three principal components. This type of multivariate analysis is inherently visual and can prove to be a very useful information visualization tool for microbiome analysis.

Some recent projects move beyond visualization into visual analytics, which closely integrates computational analysis and visualization and human–computer interaction [61]. This is distinct from information visualization, which focuses on methods such as heat maps for showing high-dimensional research results, and scientific visualization, which focuses on the mathematics and physics of visualizing complex objects. What distinguishes visual analytics is the integration of data analysis with visualization methods so that data analysis can be launched directly from the visualization, and the visualization adjusted in response to the data analysis. Computer hardware such as the Microsoft Surface Computer or the Apple iPad enable and democratize visual analytics. All of this combined with a 3D visualization screen or display wall provides a modern visual analytics discovery environment that immerses users in their data and research results.

For example, Ravel *et al.* [62] used movies to explore and display the temporal variation in the vaginal microbiome of 396 women from different racial groups, and work is underway to incorporate temporal and patient metadata. The use of movies allows users to interact with the visualization in a way that is not possible with static images. As another example, one can extend the traditional heat map by integrating and rendering additional information along the *z*-axis [63]. This additional visual dimension enhances the visual discovery process. In this study, the authors implemented the 3D heat map using a commercial 3D video game engine called Unity 3D. (The authors chose the Unity3D development tool because it uses Mono, the open-source, cross-platform. NET implementation, so as to not be limited to code libraries supplied by the vendor.) Unity makes graphic-user interface (GUI) code easy to write, enabling rapid prototyping, and the workflow for incorporating assets from other tools such as Maya and Photoshop is straightforward. An additional advantage is that Unity can use Direct3D on Windows machines, which allows users to employ off-the-shelf drivers to view 3D heat maps in stereo on suitable equipment. OpenGL would require explicit coding to see the view from each eye to produce stereo. The ability to easily see 3D heat maps in stereo is important given the widespread and emerging availability of 3D televisions and computer monitors, and leveraging game development systems for data analysis engages powerful market forces to enhance scientific analyses.

Another important benefit of using video game engines for visual analytics is that they make it possible to interact with the 3D visualization as you would in a video game. Animation, sound and point and click interaction with the data on the screen enable the user to experience their data in creative ways. The end result is an open-source software package that combines human–computer interaction and visualization in a 3D heat map in a way that is not possible with common analysis tools such as Microsoft Excel or R. Figure 1 illustrates the GUI for the 3D heat map software package. Also, illustrated is one view of the microbiome data from

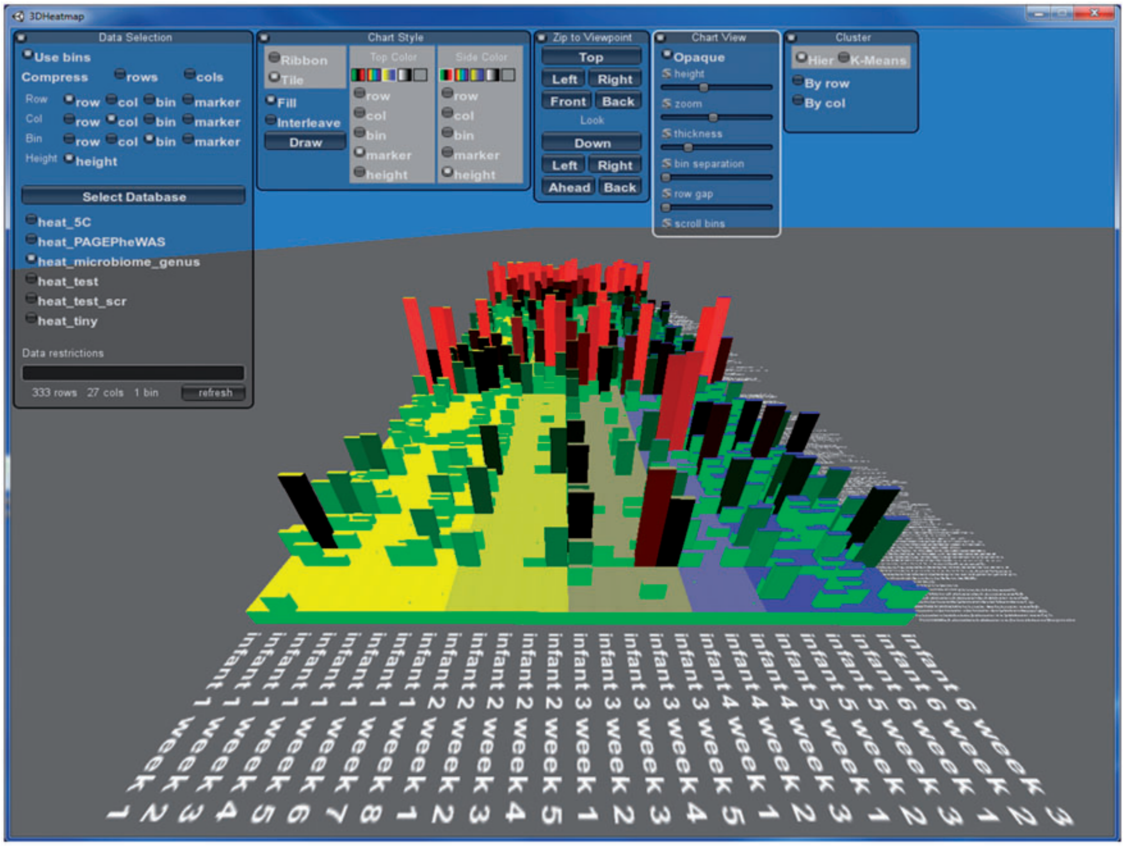

Figure 1: Screenshot of the 3D heat map application showing menus for data selection, chart style, viewpoint, chart view and cluster analysis. Each menu can be minimized or hidden. Illustrated are human microbiome data. Each row is a microbe with the name shown on the *y*-axis. Each column is a different subject and time point. The *z*-axis represents the relative abundance of the microbes. The 3D heat map makes it possible to add additional layers of information in the fourth and fifth dimensions, using colors (see online documents).

Moore *et al.* [63]. The software allows you to load data from an SQLite database, select color schemes, select visualization settings and even perform a cluster analysis as a way to organize the results. Here, each row represents a different microbe. The height and side color (green to red) of the bars represent the relative abundance of each microbe while the columns represent different patients (colored yellow to blue) at different time points in chronological order. A 3D mouse or keyboard controls and a standard mouse make it possible to interactively explore the data. A central challenge for adapting these kinds of visualization tools for microbiome data will be the integration of phylogenetic information.

WHERE THINGS ARE GOING

Sequencing technologies will continue to improve in both accuracy and throughput, and bench top sequencers will become standard equipment in individual labs. Amplicon techniques will rely more on whole gene samples, perhaps from multiple genes, removing the bias associated with selecting fixed fragments of a particular gene. This will increase the need for tools that deduce phylogenies from gene genealogies. Complete 16S rRNA gene sequences will remain the standard for microbial systematics for some time. However, we anticipate that amplicon analysis will become a quick screening technique, preliminary to more detailed metagenomic studies, rather than the final stage in ecological analysis.

The ideal data set for genomic-based microbial studies of any given ecosystem, including those associated with animals, including humans, is a complete genome for every organism at a given time in the ecosystem. When combined with temporal observations, it might be possible to completely characterize the genetic diversity of the system by sequencing the dominant organisms as the system changes. When the temporally situated, approximated genomes of the dominant members are sequenced, it may be possible to generate comprehensive models of microbial metabolism and interactions and to design experiments that manipulate the system by adding or removing specific populations. The most obvious route toward such comprehensive data sets is single genome isolation and sequencing [64]. This technology is currently performed by isolating single microbial cells and sequencing them directly. It is used to identify the functional potential of organisms and to design economically feasible, rather than exhaustive, shotgun metagenomics studies. Naturally, it will be difficult to sample very low-abundance organisms, or to sample deeply enough to detect minor genomic variations. Limited coverage is a technological challenge, which is likely to be overcome by new technologies. But sequencing depth, may be endemic to microbiome studies if small genomic variations are discovered that significantly alter community functions.

But the ultimate objective of microbiome studies is to build complete, predictive models of how microbiomes interact and respond to stimuli such as climate change, agricultural practices and disease [65]. Parameterizing such complex models will continue to require metatranscriptomic and other 'omic' studies of the expressed capability of community members [33, 66, 67]. Using techniques such as autonomous collection and preservation of microbial communities for metatranscriptomic analysis combined with quantitative characterization of transcription in metatranscriptomic data, we may start to see a revolution in our ability to quantify functional capability [68, 69].

Statistical improvements will occur in parametric model estimation, error and uncertainty bounds, and in comparing diversity statistics, especially in terms of comparison of communities. These improvements are likely to include refined techniques for censoring unreliable data, without first characterizing where the noise comes from. We also anticipate that software tools will become more available for sophisticated analyses, but that interpreting results will still require statistical expertise.

Information visualization and visual analytics will become standard parts of microbiome research workflows. Integration into statistical computing software such as R is already underway, so that analyses can be launched directly from visualization applications. The ability to launch statistical analyses directly from the visualization environment opens the door to making discoveries that are inspired by visual cues, rather than preconceived hypotheses that are dependent on existing knowledge.

Key points

- Next-generation sequencing technologies have made it possible to collect thorough samples of nucleotide sequence data from given microbial ecosystems.
- Comprehensive samples of particular genes or gene fragments, including amplicons, processed with recently developed software can lead to accurate characterizations of microbial community richness and diversity.
- Comprehensive samples of all genomic or transcriptomic data in a given microbial ecosystem, metagenomics, with recently developed software, can lead to accurate characterizations of microbial community function.
- Statistical methods for estimating the structure of microbial populations is moving beyond mere quantification, making it possible to begin developing predictive models.
- Visual analytics, which combines interactive, visual exploration of gene-based and metagenomic data sets with statistical analysis software, is developing rapidly.

Acknowledgements

We also thank the organizers of the Pacific Symposium on Biocomputing for their special session on 'Microbiome Studies'.

FUNDING

NIH COBRE (grant P20RR16448 to J.A.F., partial); NIH INBRE (grant P20RR016454 to J.A.F., partial); NSF STC 'BEACON Center for the Study of Evolution in Action' NSF STC (DBI-0939454 to J.A.F., partial); National Science Foundation (grant DEB-08-16638 to J.B., partial); U.S. Department of Energy (Contract DE-AC02-06CH11357 to J.A.G., partial) and NIH (grants LM009012, LM010098 and AI59694 to J.H.M., partial). This research was conducted using the resources of the Cornell Center for Advanced Computing, which receives funding from Cornell University, the National Science Foundation, and other leading public agencies, foundations, and corporations.

References

1. Whitman WB, Coleman DC, Wiebe WJ. Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* 1998;**95**(12): 6578–83.
2. Sears CL. A dynamic partnership: celebrating our gut flora. *Anaerobe* 2005;**11**(5):247–51.

3. Rodrigue S, Materna AC, Timberlake SC, *et al.* Unlocking short read sequencing for metagenomics. *PLoS ONE* 2010; **5**(7):e11840.
4. Gilbert JA, Meyer F, Jansson J, *et al.* The Earth Microbiome Project: Meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6 2010. *Stand Genomic Sci* 2010;**3**(3):249–53.
5. Bowler C, Karl D. Microbial oceanography in a sea of opportunity. *Nature* 2009;**459**:180–4.
6. Brüls T, Weissenbach J. The human metagenome: our other genome? *Hum Mol Genet* 2011;**20**:142–8.
7. Bunge J. Estimating the number of species with CatchAll. *Pac Symp Biocomput* 2011;121–30.
8. Chao A, Chazdon RL, Colwell RK, Shen T-J. Abundance-based similarity indices and their estimation when there are unseen species in samples. *Biometrics* 2006; **62**(2):361–71.
9. Legendre P, Legendre L. *Numerical Ecology*, Vol. 24, 3rd edn (Developments in Environmental Modelling). Amsterdam: Elsevier, 2012.
10. Suzuki S, Ono N, Furusawa C, *et al.* Comparison of sequence reads obtained from three next-generation sequencing platforms. *PLoS ONE* 2011;**6**(5):e19534.
11. Brochier-Armanet C, Deschamps P, López-García P, *et al.* Complete-fosmid and fosmid-end sequences reveal frequent horizontal gene transfers in marine uncultured planktonic archaea. *ISME J* 2011;**5**(8):1291–302.
12. Schloss PD, Westcott SL, Ryabin T, *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009; **75**(23):7537–41.
13. Cole JR, Chai B, Farris RJ, *et al.* The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res* 2007;**35**(Database issue):D169–72.
14. Edgar RC, Haas BJ, Clemente JC, *et al.* UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 2011;**27**(16):2194–200.
15. Reeder J, Knight R. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat Methods* 2010;**7**(9):668–9.
16. Huse SM, Welch DM, Morrison HG, *et al.* Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ Microbiol* 2010;**12**(7):1889–98.
17. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci USA* 1977;**74**(11):5088–90.
18. Woese CR. A new biology for a new century. *Microbiol Mol Biol Rev* 2004;**68**(2):173–86.
19. Schmidt TM. The maturing of microbial ecology. *Int Microbiol* 2006;**9**(3):217–23.
20. Pruesse E, Quast C, Knittel K, *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 2007;**35**(21):7188–96.
21. Nawrocki E, Kolbe D, Eddy S. Infernal 1.0: inference of RNA alignments. *Bioinformatics* 2009;**25**(10):1335–7.
22. Birin H, Gal-Or Z, Elias I, Tuller T. Inferring horizontal transfers in the presence of rearrangements by the minimum evolution criterion. *Bioinformatics* 2008;**24**(6):826–32.
23. Kuczynski J, Liu Z, Lozupone C, *et al.* Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat Methods* 2010; **7**(10):813–9.
24. Kim M, Morrison M, Yu Z. Evaluation of different partial 16S rRNA gene sequence regions for phylogenetic analysis of microbiomes. *J Microb Meth* 2011;**84**(1):81–7.
25. Schloss P. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol* 2010; **6**(7):e1000844.
26. Wommack K, Bhavsar J, Ravel J. Metagenomics: read length matters. *Appl Environ Microbiol* 2008;**74**(5):1453–63.
27. Liu Z, DeSantis TZ, Andersen GL, Knight R. Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res* 2008;**36**:e120.
28. Sun Y, Cai Y, Huse SM, *et al.* A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief Bioinformatics* 2011; **13**(1):107–31.
29. DeSantis TZ, Hugenholtz P, Larsen N, *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 2006;**72**(7): 5069–72.
30. Knights D, Costello E. Supervised classification of human microbiota. *FEMS Microbiol* 2011;**35**(2):343–59.
31. Cadotte MW, Jonathan Davies T, Regetz J, *et al.* Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. *Ecol Lett* 2010;**13**(1):96–105.
32. Lozupone C, Lladser M, Knights D, *et al.* UniFrac: an effective distance metric for microbial community comparison. *ISME J* 2010;**5**(2):169–72.
33. Huse SM, Huber JA, Morrison HG, *et al.* Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* 2007;**8**(7):R143.
34. Gilbert JA, Field D, Swift P, *et al.* The taxonomic and functional diversity of microbes at a temperate coastal site: a "multi-omic" study of seasonal and diel temporal variation. *PLoS ONE* 2010;**5**(11):e15545.
35. Kristensen DM, Mushegian AR, Dolja VV, *et al.* New dimensions of the virus world discovered through metagenomics. *Trends Microbiol* 2010;**18**(1):11–9.
36. Cuvelier ML, Allen AE, Monier A, *et al.* Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc Natl Acad Sci USA* 2010; **107**(33):14679–84.
37. Wu D, Wu M, Halpern A, *et al.* Stalking the fourth domain in metagenomic data: searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees. *PLoS ONE* 2011;**6**(3):e18011.
38. Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol* 2010;**6**(2):e1000667.
39. Warren RL, Holt RA. Targeted assembly of short sequence reads. *PLoS ONE* 2011;**6**(5):e19816.
40. Chitsaz H, Yee-Greenbaum JL, Tesler G, *et al.* Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nat Biotechnol* 2011;**29**(10):915–21.
41. Temperton B, Gilbert JA, Quinn JP, *et al.* Novel analysis of oceanic surface water metagenomes suggests importance of

polyphosphate metabolism in oligotrophic environments. *PLoS ONE* 2011;**6**(1):e16499.

42. Pignatelli M, Moya A. Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PLoS ONE* 2011;**6**(5):e19984.
43. Morgan JL, Darling AE, Eisen JA. Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS ONE* 2010;**5**(4):e10209.
44. Biddle J, White J, Teske A. Metagenomics of the subsurface Brazos-Trinity Basin (IODP site 1320): comparison with other sediment and pyrosequenced metagenomes. *ISME J* 2011;**5**(6):1038–47.
45. Mitra S, Rupek P, Richter DC, *et al*. Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinformatics* 2011;**12(Suppl 1)**:S21.
46. Yilmaz P, Kottmann R, Field D, *et al*. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol* 2011;**29**(5):415–20.
47. Weber M, Teeling H, Huang S, *et al*. TaxSOM: application of self-organizing maps to link biodiversity and functional data in environmental metagenomics. *ISME J* 2011;**5**(5): 918–28.
48. Meyer F, Paarmann D, D'Souza M, *et al*. The Metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 2008;**9**:386.
49. Larsen PE, Collart F, Field D, *et al*. Predicted Relative Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine metagenomic dataset. *Microbial Inform Exp* 2011;**1**:4.
50. Bunge J. Statistical estimation of uncultivated microbial diversity. In: *Uncultivated Microorganisms*, Vol. 10. New York, NY: Springer Verlag, 2009, 160–78.
51. Caporaso JG, Kuczynski J, Stombaugh J, *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010;**7**(5):335–6.
52. Barger K. Objective Bayesian estimation for the number of species. *Bayesian Analysis* 2010;**5**(4):765–86.
53. Quince C, Curtis TP, Sloan WT. The rational exploration of microbial diversity. *ISME J* 2008;**2**(10):997–1006.
54. Rocchetti I, Bunge J. Population size estimation based upon ratios of recapture probabilities. *J Appl Stat* 2011;**5**(2B): 1512–33.
55. Bunge J, Böhning D, Allen H, *et al*. Estimating population diversity with unreliable low frequency counts. In: B, editor. Kohala, Vol. 17 HI: Pacific Symposium in Biocomputing, 2012, 203–12.
56. Engen S, Grøtan V. Estimating similarity of communities: a parametric approach to spatio-temporal analysis of species diversity. *Ecography* 2010;**34**:220–31.
57. Heer J, Bostock M, Ogievetsky V. A tour through the visualization zoo. *Commun ACM* 2010;**53**(6):59–67.
58. Sneath PH. The application of computers to taxonomy. *J Gen Microbiol* 1957;**17**(1):201–26.
59. Eisen MB, Spellman PT, Brown PO, *et al*. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998;**95**(25):14863–8.
60. Wu GD, Chen J, Hoffmann C, *et al*. Linking long-term dietary patterns with gut microbial enterotypes. *Science* 2011;**334**(6052):105–8.
61. Zhang Q, Segall R, Cao M. (eds.). Visual analytics and interactive technologies: data, text, and web mining applications. *Information Science Reference*. 2011. Hershey, PA.
62. Ravel J, Gajer P, Abdo Z, *et al*. Vaginal microbiome of reproductive-age women. *Proc Natl Acad Sci USA* 2011; **108(Suppl 1)**:4680–7.
63. Moore J, Lari R. Human microbiome visualization using 3D technology. *Pac Symp Biocomput* 2011;**15**:154–64.
64. Woyke T, Tighe D, Mavromatis K, *et al*. One bacterial cell, one complete genome. *PLoS ONE* 2010;**5**(4):e10314.
65. Denef VJ, Mueller RS, Banfield JF. AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *ISME J* 2010;**4**(5):599–610.
66. Gilbert JA, Steele JA, Caporaso JG, *et al*. Defining seasonal marine microbial community dynamics. *ISME J* 2011;**18**: 1–11.
67. Gosalbes MJ, Durbán A, Pignatelli M, *et al*. Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS ONE* 2011;**6**(3):e17447.
68. Ottesen EA, Marin R, Preston CM, *et al*. Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. *ISME J* 2011;**5**(12):1881–95.
69. Field D, Amaral-Zettler L, Cochrane G, *et al*. The Genomic Standards Consortium. *PLoS Biol* 2011;**9**(6):e1001088.