An introduction to the analysis of shotgun metagenomic data

Thomas J. Sharpton*

Department of Microbiology and Department of Statistics, Oregon State University, Corvallis, OR, USA

Environmental DNA sequencing has revealed the expansive biodiversity of microorganisms and clarified the relationship between host-associated microbial communities and host phenotype. Shotgun metagenomic DNA sequencing is a relatively new and powerful environmental sequencing approach that provides insight into community biodiversity and function. But, the analysis of metagenomic sequences is complicated due to the complex structure of the data. Fortunately, new tools and data resources have been developed to circumvent these complexities and allow researchers to determine which microbes are present in the community and what they might be doing. This review describes the analytical strategies and specific tools that can be applied to metagenomic data and the considerations and caveats associated with their use. Specifically, it documents how metagenomes can be analyzed to quantify community structure and diversity, assemble novel genomes, identify new taxa and genes, and determine which metabolic pathways are encoded in the community. It also discusses several methods that can be used compare metagenomes to identify taxa and functions that differentiate communities.

Keywords: metagenome, bioinformatics, microbiota, microbiome, microbial diversity, host–microbe interactions, review

INTRODUCTION

Microorganisms are essentially everywhere in nature. Diverse communities of microbes thrive in environments ranging from the human gut (Walter and Ley, 2011), to the rhizosphere (Philippot et al., 2013), to conventionally inhospitable habitats such as acid mine runoff (Simmons et al., 2008) and geothermal hot springs (Sharp et al., 2014). Studies of cultured microbes reveal that they are critical components of these environments and provide essential ecosystem services (Arrigo, 2005; van der Heijden et al., 2008). Microbes that associate with a macroscopic host organism are no exception, and, in the subsequent discussion, are referred to as microbiota (note other definitions exist, e.g., Aminov, 2011). Microbiota can interact with their host to influence physiology and contribute to health, growth, or fitness (van der Heijden et al., 2008; Dimkpa et al., 2009; Hooper et al., 2012). For example, studies of model rhizosphere microbiota have taught us that they can impact plant growth (Kennedy et al., 2007), stress response (Redman et al., 2002; Yang et al., 2009), and pathogenic defense (Cook et al., 1995; Raupach and Kloepper, 1998). A comprehensive understanding of a macroscopic organism's physiology requires investigation of its microbiota. Unfortunately, most microbes are notoriously difficult to culture in the laboratory.

Advances in DNA sequencing and biocomputing enable exploration of the genetic diversity of the uncultured component of host-associated microbial communities. Amplicon sequencing, for example, is the most widely used method for characterizing the diversity of microbiota. Here, a community is sampled (e.g., water, soil, tissue biopsy) and DNA is extracted from all cells in the sample. A taxonomically informative genomic marker that is common to virtually all organisms of interest is then targeted and amplified by PCR. The resultant amplicons are sequenced and bioinformatically characterized to determine which microbes are present in the sample and at what relative abundance. In the case of bacteria and archaea, amplicon sequencing studies usually target the small-subunit ribosomal RNA (16S) locus, which is both a taxonomically and phylogenetically informative marker (Pace et al., 1986; Hugenholtz and Pace, 1996). Amplicon sequencing of the 16S locus revealed an a tremendous amount of microbial diversity on Earth (Pace, 1997; Rappé and Giovannoni, 2003; Lozupone and Knight, 2007) and has been used to characterize the biodiversity of microbes from a great range of environments including the human gut (Human Microbiome Project Consortium, 2012a; Yatsunenko et al., 2012), *Arabidopsis thaliana* roots (Lundberg et al., 2012), ocean thermal vents (McCliment et al., 2006), hot springs (Bowen De León et al., 2013), and Antarctic volcano mineral soils (Soo et al., 2009). Comparing 16S sequence profiles across samples clarifies how microbial diversity associates with and scales across environmental conditions. In the case of microbiota, such observations have generated insight into host–microbe interactions and yielded hypotheses about microbiota-based disease mechanisms (Turnbaugh et al., 2009; Muegge et al., 2011; Bulgarelli et al., 2012; Smith et al., 2013). Follow-up microbiota-manipulation studies often confirm these hypotheses (Smith et al., 2013; David et al., 2014). Experimental design plays an important role in these analyses, as the most promising hypotheses tend to derive from comparisons of microbiota associated with cohorts of hosts of distinct genotypes or treatment conditions. (Kuczynski et al., 2011) provide a thorough review of how 16S amplicon sequencing can be used to study microbiota.

While powerful, amplicon sequencing is not without limitation. First, it may fail to resolve a substantial fraction of the diversity in a community given various biases associated with PCR (Hong et al., 2009; Sharpton et al., 2011; Logares et al., 2013). Second, amplicon sequencing can produce widely varying estimates of diversity (Jumpstart Consortium Human Microbiome Project Data Generation Working Group, 2012). For example, different genomic loci have differential power at resolving taxa (Liu et al., 2008; Schloss, 2010; Jumpstart Consortium Human Microbiome Project Data Generation Working Group, 2012). In addition, sequencing error and incorrectly assembled amplicons (i.e., chimeras), can produce artificial sequences that are often difficult to identify (Wylie et al., 2012). Third, amplicon sequencing typically only provides insight into the taxonomic composition of the microbial community. It is impossible to directly resolve the biological functions associated with these taxa using this approach. In some cases, phylogenetic reconstruction can be used to infer those biological functions that are encoded in a genome containing a particular 16S sequence (Langille et al., 2013). But, the accuracy with which these methods estimate the true functional diversity of a community is tied to how well the genomic diversity of the community is represented by the genomes available in sequence databases. Finally, amplicon sequencing is limited to the analysis of taxa for which taxonomically informative genetic markers are known and can be amplified. Novel or highly diverged microbes, especially viruses, are difficult to study using this approach. Additionally, because the 16S locus can be transferred between distantly related taxa (i.e., horizontal gene transfer), analysis of 16S sequences can result in overestimations of the community diversity (Acinas et al., 2004).

Shotgun metagenomic sequencing is an alternative approach to the study of uncultured microbiota that avoids these limitations. Here, DNA is again extracted from all cells in a community. But, instead of targeting a specific genomic locus for amplification, all DNA is subsequently sheared into tiny fragments that are independently sequenced. This results in DNA sequences (i.e., reads) that align to various genomic locations for the myriad genomes present in the sample, including non-microbes. Some of these reads will be sampled from taxonomically informative genomic loci (e.g., 16S), and others will be sampled from coding sequences that provide insight into the biological functions encoded in the genome. As a result, metagenomic data provides the opportunity to simultaneously explore two aspects of a microbial community: *who is there* and *what are they capable of doing?*

Despite these benefits, metagenomic sequence data presents several challenges. First, metagenomic data is relatively complex and large, complicating its informatic analysis. For example, it can be difficult to determine the genome from which a read was derived. Additionally, most communities are so diverse that most genomes are not completely represented by reads. As a result, two reads from the same gene may not overlap and are thus impossible to directly compare through sequence alignment (Schloss and Handelsman, 2008; Sharpton et al., 2011). When reads do overlap, it is not always evident if they are from the same or distinct genomes, which can challenge sequence assembly (Mavromatis et al., 2007; Mende et al., 2012). Also, metagenomic analysis

tends to require a large volume of data to identify meaningful results because of the vast amount of genomic information being sampled. This requirement can pose computational problems. Fortunately, informatic software development is rapidly advancing and improving the ease and efficiency of metagenomic analysis. Second, metagenomes may contain unwanted host DNA, especially in the case of microbiota. In some situations, host DNA can so overwhelm community DNA that intricate molecular methods must be applied to selectively enrich microbial DNA prior to sequencing. Molecular and bioinformatic methods to filter host DNA from metagenomes either prior or subsequent to sequencing of the data are in development (Woyke et al., 2006; Chew and Holmes, 2009; Delmotte et al., 2009; Schmieder and Edwards, 2011b; Garcia-Garcerà et al., 2013). Third, while contamination is a challenge general to environmental sequencing studies (Degnan and Ochman, 2012), the identification and removal of metagenomic sequence contaminants is especially problematic (Kunin et al., 2008). For example, it can be difficult to determine which reads were generated from a detected contaminant's genome. A metagenomic contaminant can mislead analyses of community genetic diversity if the contaminant's genome is enriched for genes that are uncommon in the community, especially when the contaminant is highly abundant or has a large genome. Fortunately, software tools that identify and filter contaminant sequences in a metagenome exist (Schmieder and Edwards, 2011a). Finally, metagenomes tend to be relatively expensive to generate compared to amplicon sequences, especially in complex communities or when host DNA greatly outnumbers microbial DNA. Ongoing advances in DNA sequencing technology are improving the affordability of metagenomic sequencing.

These challenges have limited the application of metagenomic investigation. But, thanks to the aforementioned research advances, this analytical strategy has become more tractable for most laboratories. In recent years, metagenomic sequencing has been used to identify new viruses (Yozwiak et al., 2012), characterize the genomic diversity and function of uncultured bacteria (Wrighton et al., 2012), reveal novel and ecologically important proteins (Godzik, 2011), and identify taxa and metabolic pathways that differentiate gut microbiota associated with healthy and diseased humans (Morgan et al., 2012). The analysis of metagenomes has also been used to characterize plant microbiota, especially those associated with roots and leaves [as reviewed in Bulgarelli et al. (2013), Vorholt (2012)]. For example, metagenomic analysis has been used to identify physiological traits that differentiate rice leaf- and root-associating communities (Knief et al., 2012), characterize root endophytes of rice (Sessitsch et al., 2012), and quantify the physiological differences between microbiota associating with clover, soybean, and *Arabidopsis* leaves (Delmotte et al., 2009). The study of plant metagenomes can be difficult given that plants can have large genomes, which can overwhelm the genomic representation of the microbial community in the metagenome. Advances in laboratory procedures that physically separate microbiota from plant tissue (e.g., Jiao et al., 2006; Delmotte et al., 2009) will continue to improve the efficacy of metagenomic investigations in plant systems.

What follows in this review is a discussion of how metagenomic sequencing can be used to explore the taxonomic and functional diversity of microbial communities (**Figure 1**). It will briefly introduce analytical tools to this end, though it will not provide an exhaustive listing of all such tools. This review will assume that the reader is able to generate shotgun metagenomes, quality control the sequence data, and bioinformatically filter host DNA when relevant. The Human Microbiome Project standard operating protocols (http://www.hmpdacc.org/tools_protocols/tools_protocols.php) provide a thorough guide on how to conduct these steps (Human Microbiome Project Consortium, 2012a). Additional information on metagenomic analysis can be found in other reviews (e.g., Kunin et al., 2008; Kuczynski et al., 2011; Thomas et al., 2012; Davenport and Tümmler, 2013).

WHO IS THERE? ASSESSING TAXONOMIC DIVERSITY

One of the primary ways by which a microbial community can be characterized is the quantification of its taxonomic diversity. This involves determining which microbes are present in a community (i.e., richness) and at what abundance. Taxonomic diversity serves as a way of profiling a community and can be used to ascertain the similarity of two or more communities (e.g., communities with more shared taxa are more similar). Additionally, taxonomic diversity may provide some insight into the biological function of the community when it contains members of functionally described taxa (e.g., the presence of Cyanobacteria suggests

that the community is photosynthetic). In the case of metagenomics, taxonomic diversity is typically quantified by either (1) analyzing taxonomically informative marker genes, (2) grouping sequences into defined taxonomic groups (i.e., binning), or (3) assembling sequences into distinct genomes. These approaches are not mutually exclusive and may be synergistic (**Figure 2**). For example, in some situations, it may be appropriate to bin sequences into taxonomic groups and then subject each group's sequences to assembly, while other cases may warrant conducting an initial assembly and then subjecting the assembled sequences to binning.

MARKER GENE ANALYSIS

Marker gene analysis is one of the most straightforward and computationally efficient ways of quantifying a metagenome's taxonomic diversity. This procedure involves comparing metagenomic reads to a database of taxonomically informative gene families (i.e., marker genes), identifying those reads that are marker gene homologs, and using sequence or phylogenetic similarity to the marker gene database sequences to taxonomically annotate each metagenomic homolog. The most frequently used marker genes include rRNA genes or protein coding genes that tend to be single copy and common to microbial genomes. Because this approach involves comparing metagenomic reads to a relatively small database for the purpose of a similarity search (e.g., not all gene families are taxonomically informative), marker gene analysis can be a relatively rapid way to estimate the diversity
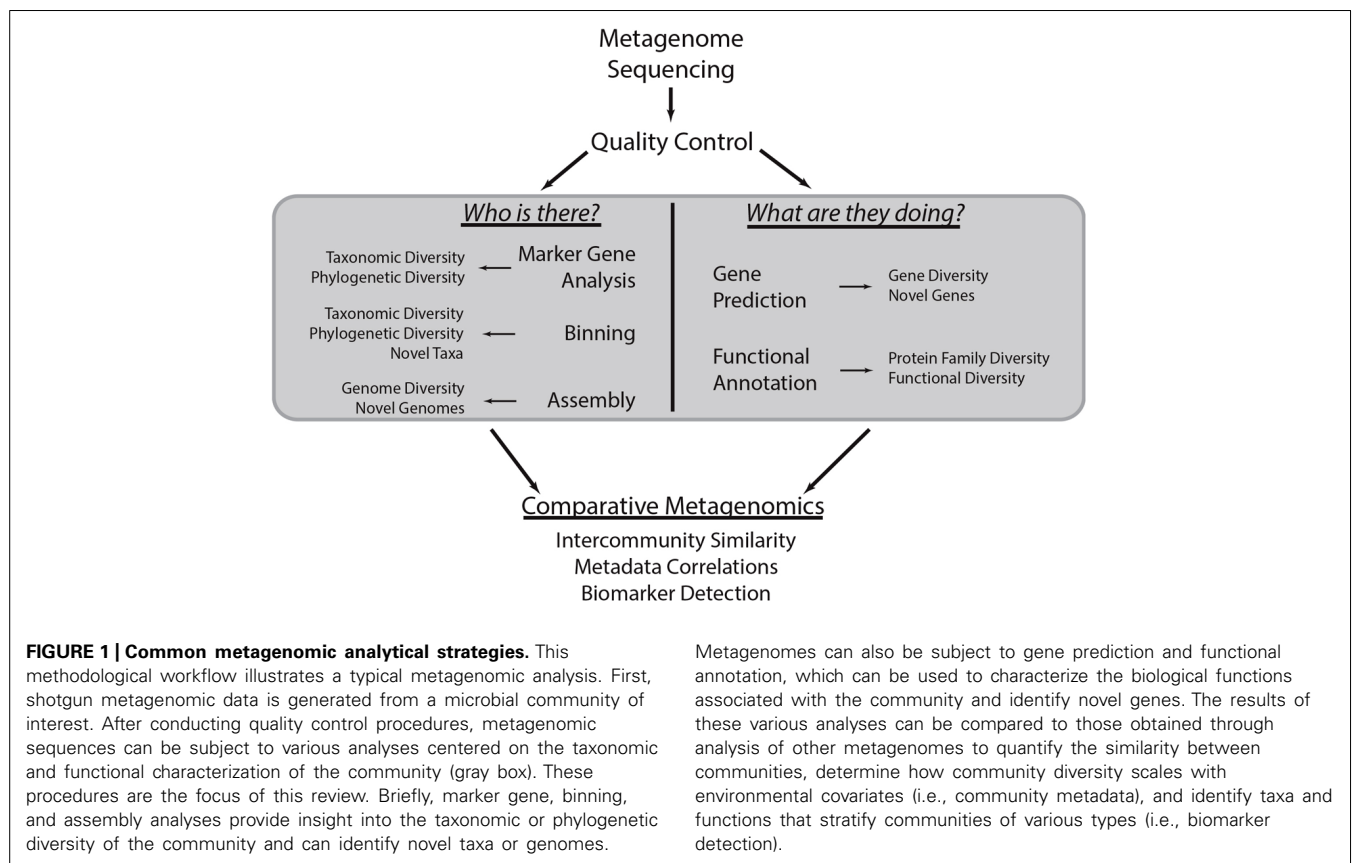

FIGURE 1 | Common metagenomic analytical strategies. This methodological workflow illustrates a typical metagenomic analysis. First, shotgun metagenomic data is generated from a microbial community of interest. After conducting quality control procedures, metagenomic sequences can be subject to various analyses centered on the taxonomic and functional characterization of the community (gray box). These procedures are the focus of this review. Briefly, marker gene, binning, and assembly analyses provide insight into the taxonomic or phylogenetic diversity of the community and can identify novel taxa or genomes.

Metagenomes can also be subject to gene prediction and functional annotation, which can be used to characterize the biological functions associated with the community and identify novel genes. The results of these various analyses can be compared to those obtained through analysis of other metagenomes to quantify the similarity between communities, determine how community diversity scales with environmental covariates (i.e., community metadata), and identify taxa and functions that stratify communities of various types (i.e., biomarker detection).
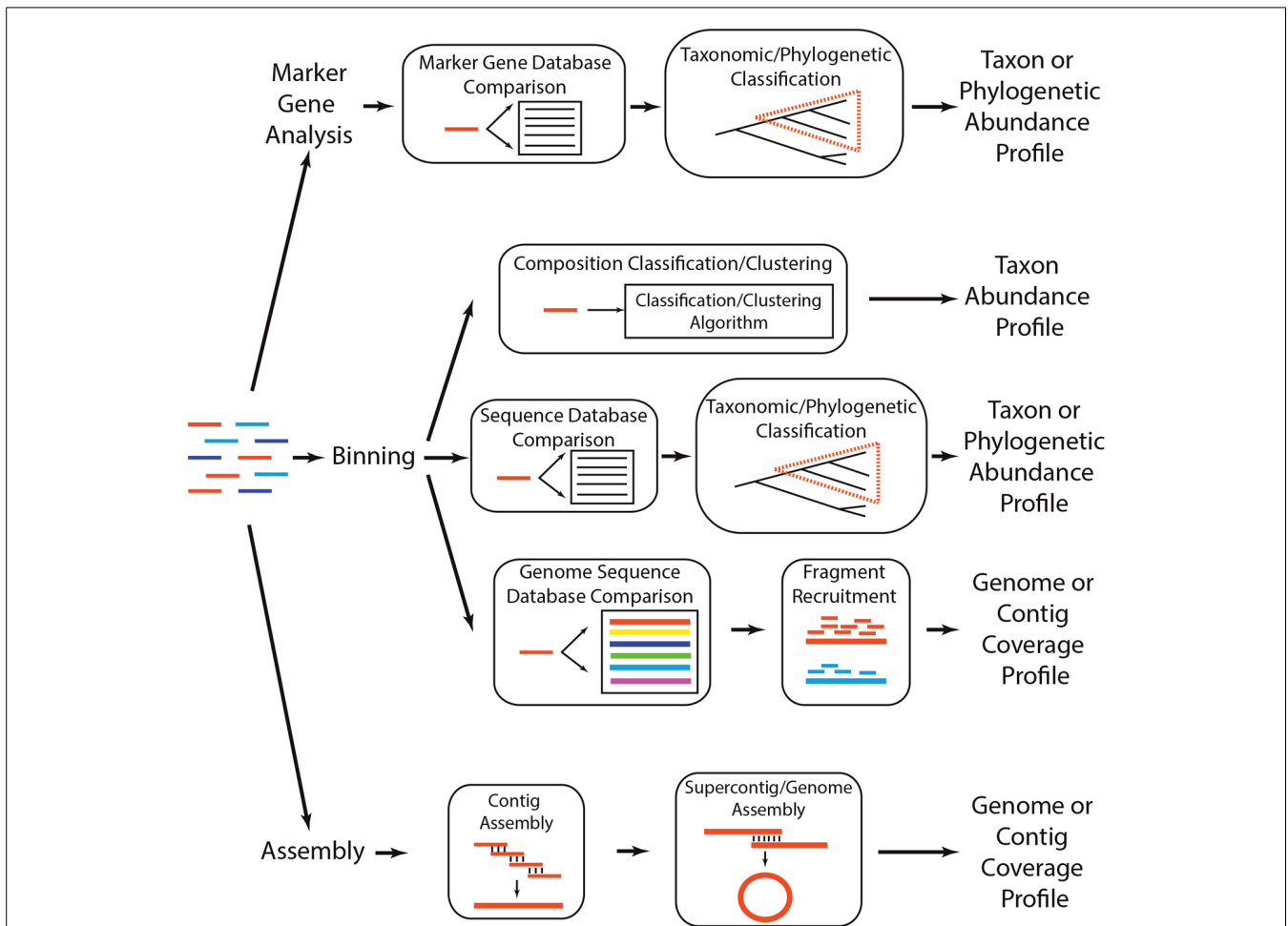
FIGURE 2 | Analytical strategies to determine which taxa are present in a metagenome. A metagenome (colored lines, left) can be subject to three general analytical strategies that ultimately produce a profile of the taxa, phylogenetic lineages, or genomes present in the community. *Marker gene analyses* involve comparing each read to a reference database of taxonomically or phylogenetically informative sequences (i.e., marker genes), using a classification algorithm to determine if the read is a homolog of a marker gene, and annotating classified reads based on their similarity across marker gene sequences. There are several methods for *binning* metagenomes, including (1) compositional binning, which uses sequence composition to classify or cluster metagenomic reads into taxonomic groups, (2) similarity binning, which classifies a read into a taxonomic or phylogenetic group based on its similarity to previously identified genes or proteins, and (3) fragment recruitment, wherein reads are aligned to nearly identical genome sequences to produce metagenomic coverage estimates of the genome. Finally, sequences can be subject to *assembly*, wherein reads that share nearly identical sequence at their ends are merged to create contigs, which can subsequently be assembled into supercontigs or complete genomes.

of a metagenome. Additionally, focusing on single-copy gene families may provide more accurate estimates of taxonomic abundance than methods that consider families known to widely vary in copy number across genomes (e.g., similarity based binning, below; Liu et al., 2011). This general strategy may be applied to assembled or unassembled reads, though some specific methods may only be applicable to one of these two data types.

There are two general methods by which marker genes are used to taxonomically annotate metagenomes. The first relies on sequence similarity between the read and the marker genes. For example, MetaPhyler uses the results of a pairwise sequence search between metagenomic reads and a database of marker genes as well as a series of custom classifiers that are considerate of family (e.g., rate of evolution) and read (e.g., sequence length) properties

to determine the taxonomy of the metagenomic sequence (Liu et al., 2011). MetaPhlAn also relies on sequence similarity to taxonomically characterize metagenomic marker gene homologs. It uses an extensive database of phylogenetic clade-specific markers (i.e., families that are single copy and generally only common to a monophyletic group of taxa) to assign metagenomic sequences to specific taxonomic groups (Segata et al., 2012). The second approach uses phylogenetic information, which may take longer to calculate, but may also provide greater accuracy. For example, AMPHORA (Wu and Eisen, 2008; Wu and Scott, 2012), uses hidden Markov models (HMMs) to identify metagenomic homologs of phylogenetically informative, single copy protein-coding genes that are common to sequenced genomes from either bacteria or archaea. It then assembles a marker gene phylogeny that includes metagenomic homologs, which are annotated based

on their relative location in the tree (i.e., phylotyping). PhyloSift (Darling et al., 2014) is similar, but uses an expanded marker database, including an extensive viral gene family database, and edge PCA (Matsen and Evans, 2013) to identify specific lineages in a marker gene's phylogenetic tree that differ between communities. PhylOTU uses a phylogenetic tree to relate non-overlapping metagenomic 16S homologs, which are subsequently clustered into taxonomic groups based on phylogenetic distance (Sharpton et al., 2011).

There are several important caveats to be aware of when using marker genes to analyze metagenomes. First, this strategy operates under the assumption that the relatively small fraction of the metagenome that is homologous to marker genes represents an accurate sampling of the entire taxonomic distribution of the community. While researchers take great effort to identify marker genes that are uniformly present across clades of genomes, the genome sequences available to researchers during marker gene identification may not adequately reflect the diversity of genomes present in the community under investigation. Second, marker gene analysis is not appropriate for taxa that do not contain the markers being explored. Thanks to recent efforts to identify phylogenetic clade-specific marker genes (Segata et al., 2012; Wu et al., 2013; Darling et al., 2014) and expand the phylogenetic diversity represented in genome sequence databases (Wu et al., 2009), this may be a diminishing problem. Third, the accuracy of annotation is based on properties of the marker family and likely varies across markers. Accuracy is also a function of how well the reference database reflects the community under investigation. Expanding the phylogenetic diversity of available genomes sequences can mitigate these limitations.

BINNING

A related strategy, known as binning, attempts to assign every metagenomic sequence to a taxonomic group. Generally, each sequence is either (1) classified into a taxonomic group (e.g., OTU, genus, family) through comparison to some referential data or (2) clustered into groups of sequences that represent taxonomic groups based on shared characteristics (e.g., GC content). Binning plays an important role in the analysis of metagenomes. First, depending on the method used, binning may provide insight into the presence of novel genomes that are difficult to otherwise identify. Second, it provides insight into the distinct numbers and types of taxa in the community. While many approaches provide a coarse resolution of taxonomy, some are capable of indicating strain-level variation, though usually at the expense of fewer binned reads. Third, binning provides a way of reducing the complexity of the data, such that post-binning analyses (e.g., assembly) can be performed independently on each set of binned reads rather than on the entire population of data. Binning may be conducted on assembled or unassembled data, though most algorithms report that binning accuracy improves as sequence lengths increase. Binning algorithms generally come in one of three flavors: sequence composition, sequence similarity, and fragment recruitment.

Sequence compositional binning uses metagenome sequence characteristics (e.g., tetramer frequency) to cluster or classify sequences into taxonomic groups. These methods generally do not require the alignment of reads to a reference sequence database and, as a result, can process large metagenomes relatively rapidly. Some of these methods instead analyze whole genome sequences ahead of time to train classifiers that stratify sequences into taxonomic groups. For example, PhyloPithia and PhylopithiaS (McHardy et al., 2007; Patil et al., 2011) use support vector machines, which analyze training sequences associated with various phylogenetic groups to build oligonucleotide frequency models that determine whether a new sequence (e.g., a metagenomic read) is a member of the group. A related tool, Phymm (Brady and Salzberg, 2009; Brady and Salzberg, 2011), uses interpolated Markov models (Salzberg et al., 1998), which combine prediction probabilities derived from a variety of training sequence oligonucleotide lengths, and, optionally, blast search results, to classify metagenomics reads into phylogenetic lineages. Other methods use sequence characteristics to cluster metagenomic reads into distinct groups without querying a reference database, and thus may identify previously unknown organisms. For example, emergent self-organizing maps (ESOMs) can be used to cluster assembled metagenomic reads based on tetranucleotide frequency and, optionally, contig coverage and abundance distributions (Dick et al., 2009). While taxonomic annotations are not identified directly from this approach, it has proven useful for partitioning contigs into groups that can be subsequently assembled into nearly complete genomes representing uncharacterized organisms (Wrighton et al., 2012). Two-tiered clustering (Saeed et al., 2012) is a related approach that first bins sequences into coarse groups based on GC content and the oligonucleotide frequency-derived error gradient (Saeed and Halgamuge, 2009), which assesses the variance in oligonucleotide frequency across the length of a read, and then subdivides these initial clusters based on tetranucleotide frequency. There are many additional compositional binning algorithms – including NBC, a naïve bayes classifier that has been shown to annotate more sequences than some sequence-similarity based procedures (Rosen et al., 2008) – and listing them all is beyond the scope of this review. While compositional binning algorithms have proven useful for the analysis of metagenomes, they generally operate under the assumption that the sequence characteristics being interrogated tend to be phylogenetically informative. Variation in the taxonomic bias of these sequence characteristics may result in inaccurate assignments for a fraction of the data. Also, the accuracy of these methods, especially the classifiers, is tied to the selection of genomes used to train the classification algorithm.

Metagenomic reads can also be binned based on their sequence similarity to a database of taxonomically annotated sequences. Compared to compositional binning tools, these methods tend to require greater computational resources as every read is usually aligned to a large volume of sequences. In addition, these methods, like classification-based compositional binning algorithms, are not necessarily ideal for the identification of novel genomes, though they may be used to identify phylogenetic nodes that contain putatively novel lineages. That said, similarity based methods may provide higher annotation accuracy and resolution compared to compositional binning. One of the most widely used tools is MEGAN, which is a sequence similarity approach

that uses blast to compare metagenomic reads to a database of sequences that are annotated with NCBI taxonomy (Huson et al., 2011). It then infers the taxonomy of the sequence by placing the read on the node in the NCBI taxonomy tree that corresponds to the last common ancestor of all the taxa that contain a homolog of the read. MG-RAST is also widely used; it uses phylogenomic reconstruction of database sequences to which a read is similar to infer the read's taxonomy (Meyer et al., 2008). CARMA uses reciprocal best hits between database sequences and metagenomic reads and models gene-family specific rates of evolution to infer the appropriate taxonomic rank of each metagenomic read (Gerlach and Stoye, 2011). Note that while MetaPhlAn and PhyloSift were introduced in the marker gene analysis section, one may consider these types of methods as binning algorithms, especially as the database of marker genes expands.

A related approach, called fragment recruitment, identifies reads that exhibit nearly identical alignments to genome sequences (i.e., read mapping) and partitions reads based on the genome to which they map. This approach was used to taxonomically characterize reads in the global ocean survey Rusch et al. (2007) and Qin et al. (2010) used it to estimate the abundance of gut microbiota in healthy and inflammatory bowel diseased individuals. MOSAIK (Lee et al., 2013) was used by Schloissnig et al. (2013) to map reads to microbial genomes for the purpose of characterizing strain-level variation in the human microbiome. Fragment recruitment can also be used at the level of genes to quantify the abundance of metabolic pathways (Desai et al., 2013; see Gene Prediction). There are currently few tools that will handle both the mapping of reads to a database of genomes and the calculation of genome abundance. Genometa (Davenport et al., 2012) is one such tool and provides a graphical user interface. (Martin et al., 2012) evaluated the performance of several commonly used short read mapping algorithms (e.g., SOAP, BWA, CLC) in fragment recruitment using RefCov, which analyzes the output files produced by these algorithms and calculates recruitment statistics such as coverage depth and breadth. These methods are not necessarily ideal for communities that contain genomes outside of the scope of genome sequences in reference databases and are not useful for the analysis of novel taxa.

There are several general caveats associated with binning. First, there is usually a trade-off between the number of reads that are binned and the taxonomic specificity of the annotations assigned to each bin. Additionally, while binning provides a way of annotating a substantial fraction of the metagenome, there may be large variance in the accuracy and specificity of the annotations across reads. Second, convergent evolutionary characteristics, including horizontal gene transfer, may diminish the accuracy of binning, especially for composition-based approaches and for the study of those taxa that may not be well-represented by the training data. Finally, in the case of novel organisms, it is often difficult to validate an algorithm's predictions. Multiple independent predictions of the organism's existence (e.g., different algorithms, different communities) can provide additional support, but subsequent experimental verification may be necessary.

ASSEMBLY

Assembly merges collinear metagenomic reads from the same genome into a single contiguous sequence (i.e., contig) and is useful for generating longer sequences, which can simplify bioinformatic analysis relative to unassembled short metagenomic reads. In some instances, complete or nearly complete genomes can be assembled, which provides insight into the genomic composition of uncultured organisms found in a community (Iverson et al., 2012; Wrighton et al., 2012; Ruby et al., 2013). If used to quantify taxonomic abundance, one must be careful to track contig coverage (i.e., the number of assembled reads that align to the average base in the contig), as contigs are subsequently treated as a single sequence in most downstream analyses, and analytical tools may thus not accurately quantify the abundance of the taxon as it is represented in the raw data. The major challenge associated with assembly is the generation of chimeras, wherein sequences from two distinct genomes are spuriously assembled into a contig due to shared sequence similarity. Chimeras are more likely to be generated in relatively complex communities (Luo et al., 2012), so researchers often bin reads and independently assemble each bin to mitigate the risk of generating chimeras.

While there are many algorithms for assembling nucleic acid sequences, relatively few have been designed to deal with the specific informatic challenges associated with metagenomes. Many tools build upon the traditional de Brujin graph approach to genome assembly [thoroughly reviewed in Compeau et al. (2011)], wherein a network (i.e., graph) models the contiguous sequence overlap between all subsequences of a specified length (i.e., k-$mers$) in a read as well as the corresponding k-$mers$ in all other reads that are linked through overlapping sequence identity. For example, tools like MetaVelvet (Namiki et al., 2012) and Meta-IBDA (Peng et al., 2011) generate a de Brujin graph from the entire metagenome and use properties of the graph or sequence data to identify sub-graphs that represent genome-specific assemblies. Genovo constructs a probabilistic model of assembly and outputs the set of contigs with the highest likelihood (Laserson et al., 2011). Because of the complexity associated with *de novo* metagenome assembly, several recent tools implement data reduction or efficiency procedures to reduce the amount of memory or time needed to complete assembly. These tools may be the only options for those labs without sophisticated computing environment (e.g., big memory machines). For example, diginorm (Brown et al., 2012) filters redundant reads by normalizing the distribution of k-$mers$ in a metagenome. Another tool, khmer, stores the nodes of a de Brujin graph in a memory-efficient structure (McDonald and Brown, 2013). PRICE implements a series of data reduction procedures to minimize the complexity associated with generating an initial set of contigs and then uses paired-end information associated with reads to merge contigs (Ruby et al., 2013). Ray Meta (Boisvert et al., 2012) uses a distributed computing environment (e.g., a cloud or computer cluster) to disburse the computationally expensive task of assembly across multiple computers, which improves the rate at which massive sequence libraries can be assembled. MetAMOS is a modular workflow that executes a variety of assembly algorithms and conducts taxonomic and

functional annotation on the resulting contigs (Treangen et al., 2013).

There are several considerations associated with assembling metagenomic sequences. First, assembly tends to be limited to the most abundant taxa in the community. Without extensive sequencing, it may be difficult to assemble genomes of rare microbiota. Second, assembly may produce *in silico* chimeras, so it should be used cautiously and with consideration. Repetitive regions within a genome are also notoriously difficult to assemble; analysis of repeat copy number variation from assemblies should be carefully evaluated. Combining long-read (e.g., Pacific Bioscience) and short-read (e.g., Illumina) sequences in the same assembly may limit these errors, though there are currently few tools that can combine these types of data (Deshpande et al., 2013). Third, assembly can be computationally intensive, especially in its requirements for RAM. In addition to the efficiency tools mentioned above, binning sequences prior to assembly can be a good way to cut down on the computational complexity.

WHAT ARE THEY CAPABLE OF DOING? INFERRING BIOLOGICAL FUNCTION

Metagenomes provide insight into a community's physiology by clarifying the collective functions that are encoded in the genomes of the organisms that make up the community. The functional diversity of a community can be quantified by annotating metagenomic sequences with functions (**Figure 3**). This usually involves identifying metagenomic reads that contain protein coding sequences and comparing the coding sequence to a database of genes, proteins, protein families, or metabolic pathways for which some functional information is known. The function of the coding sequence is inferred based on its similarity to sequences in the database. Doing this for all metagenomic sequences produces a profile that describes the number of distinct types of functions and their relative abundance in the metagenome. This profile can be used to compare metagenomes to identify those communities that are metabolically similar (Human Microbiome Project Consortium, 2012b), ascertain how various treatments influence the functional composition of the community (Looft et al., 2012), and reveal those functions that associate with specific environmental or host-physiological variables (i.e., biomarkers) and may be useful for environmental or host diagnosis (Morgan et al., 2012). Metagenomes may also reveal the presence of novel genes (Nacke et al., 2012) or provide insight into the ecological conditions associated with those genes for which the function is currently unknown (Buttigieg et al., 2013). In general, metagenome functional annotation involves two non-mutually exclusive steps: gene prediction and gene annotation.
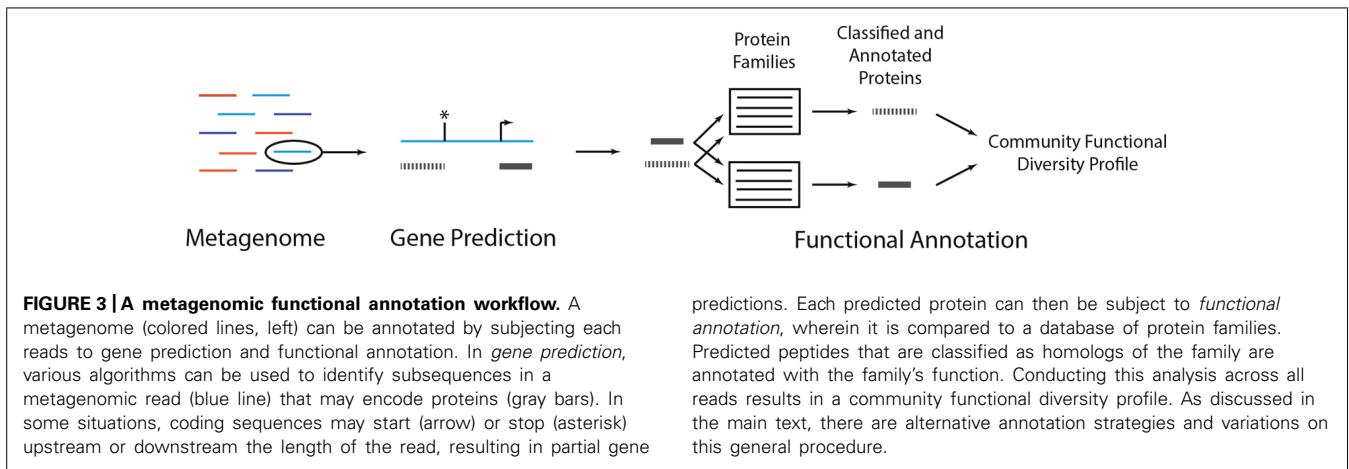
GENE PREDICTION

Gene prediction determines which metagenomic reads contain coding sequences. Once identified, coding sequences can be functionally annotated. Gene prediction can be conducted on assembled or unassembled metagenomic sequences. For assembled metagenomes with full-length coding sequences, gene prediction is akin to the framework used during the analysis of whole genome

sequences, with the caveat that some prediction algorithms require species-specific parameters that may not always be appropriate when the contigs have been sampled from diverse or novel lineages. For example, many gene prediction algorithms are typically trained using sequence features of closely related organisms. An extensive review on gene prediction in assembled genomes is covered by Yandell and Ence (2012), Richardson and Watson (2013). For unassembled or poorly assembled metagenomes, the problem is more challenging and involves predicting partial coding sequences, in the case that a gene starts upstream or stops downstream of the length of the read. There tend to be three ways by which genes are predicted in metagenomes: (1) gene fragment recruitment, (2) protein family classification, and (3) *de novo* gene prediction. Note that because of the considerable diversity of genomes in nature compared to those in sequence databases (Wu et al., 2009), not all predicted genes will exhibit homology to known sequences. Some of these predictions may be spurious, while others will represent novel or highly diverged proteins. Thus, gene prediction is not just an important step in functional annotation, but is critical to the identification of novel genes.

One of the most straightforward ways of identifying coding sequences in a metagenome is to use fragment recruitment (see Binning) to map metagenomic reads or contigs to a database of gene sequences. Metagenomic sequences that are identical or nearly identical to a full-length gene sequence are considered representative subsequences of the gene. In the case that the gene has a functional annotation, this method of gene prediction can also simultaneously provide a functional annotation for the recruited metagenomic sequences (Desai et al., 2013). This procedure has been used to quantify the genetic diversity of marine communities (Rusch et al., 2007) and gut microbiota (Qin et al., 2010; Human Microbiome Project Consortium, 2012a), and is generally useful for cataloging the specific genes present in the metagenome. This is generally a high-throughput gene prediction procedure because it tends to rely on read mapping algorithms that rapidly assess whether a genomic fragment is nearly identical to a database sequence. However, this comes at the expense of being able to identify diverse homologs of a known gene. As a result, it may not be the most appropriate gene prediction procedure for metagenomes derived from communities with genomes that are underrepresented in sequence databases, especially if the identification of novel or highly divergent genes is desired.

A related approach involves translating each metagenomic read into all six possible protein coding frames and comparing each of the resulting peptides to a database of protein sequences by sequence alignment. The alignments can then be analyzed to identify those metagenomic sequences that encode translated peptides that exhibit homology to proteins in the database. This can be conducted by using sequence translation tools like transeq (Rice et al., 2000) to translate reads prior to conducting protein sequence alignment using blastp or fast blast algorithms like USEARCH (Edgar, 2010), RAPsearch (Zhao et al., 2012), or lastp (Kiełbasa et al., 2011). Alternatively, alignment algorithms that translate nucleic acid sequences on the fly, like blastx (Altschul et al., 1997), USEARCH with the ublast option, or lastx (Kiełbasa

FIGURE 3 | A metagenomic functional annotation workflow. A metagenome (colored lines, left) can be annotated by subjecting each reads to gene prediction and functional annotation. In *gene prediction*, various algorithms can be used to identify subsequences in a metagenomic read (blue line) that may encode proteins (gray bars). In some situations, coding sequences may start (arrow) or stop (asterisk) upstream or downstream the length of the read, resulting in partial gene predictions. Each predicted protein can then be subject to *functional annotation*, wherein it is compared to a database of protein families. Predicted peptides that are classified as homologs of the family are annotated with the family's function. Conducting this analysis across all reads results in a community functional diversity profile. As discussed in the main text, there are alternative annotation strategies and variations on this general procedure.

et al., 2011) can be used. This gene prediction procedure is most frequently used in concert with functional annotation, wherein the annotation of the protein sequence to which the translated metagenomic read is homologous is used to infer the read's annotation (discussed in more depth below). Since this method relies on comparing metagenomic sequences to a reference database of known sequences, it is not useful for identifying novel types of proteins. But, it can reveal diverged homologs of known proteins.

De novo gene prediction, on the other hand, can potentially identify novel genes. Here, gene prediction models, which are trained by evaluating various properties of microbial genes (e.g., length, codon usage, GC bias), are used to assess whether a metagenomic read or contig contains a gene and does not rely on sequence similarity to a reference database to do so. As a result, these methods can identify genes in the metagenome that share common properties with other microbial genes but that may be highly diverged from any gene that has been discovered to date. There are several tools that can be used for *de novo* gene prediction, including MetaGene (Noguchi et al., 2006), Glimmer-MG (Kelley et al., 2012), MetaGeneMark (Zhu et al., 2010), FragGeneScan (Rho et al., 2010), Orphelia (Hoff et al., 2009), and MetaGun (Liu et al., 2013). In Trimble et al. (2012), many of these methods were compared using statistical simulations. The authors found that their performance varied as a function of read properties (e.g., length and sequencing error rate), with different methods producing optimal accuracies at different property thresholds, which suggests that researchers need to be careful about selecting the appropriate algorithm for their data. As in genome annotation, Yok and Rosen (2011) found that gene prediction in metagenomes improves when multiple methods are applied to the same data (e.g., a consensus approach). While these methods can require a fair bit of time and resources to predict genes, they tend to be more discriminating than 6-frame translation and may save time when functionally annotating sequences as fewer pairwise sequence comparisons may be necessary (Trimble et al., 2012). In the case that the predicted gene is novel relative to database sequences, it can be difficult to determine if the gene is real or a spurious prediction. Identifying homologs of the gene in other communities may be one way of reinforcing *de novo* predictions.

FUNCTIONAL ANNOTATION BY PROTEIN FAMILY CLASSIFICATION

Once coding sequences in a metagenome are predicted, they can be subject to functional annotation. The most common way this is accomplished is by classifying the predicted metagenomic proteins into protein families. A protein family is a group of evolutionarily related protein sequences, or subsequences in the case of protein domain families (e.g., Pfam; Finn et al., 2014). They are usually characterized by comparing full-length protein sequences that have been identified through genome sequencing projects. Because the proteins in a family share a common ancestor, they are thought to encode similar biological functions. If a metagenomic sequence is determined to be a homolog of this family (i.e., it is classified as being a member of the family), then it is inferred that the sequence encodes the family's function. Classification of an assembled or unassembled metagenomic protein sequence into a protein family usually requires comparing the metagenomic protein to either a database of protein sequences, each of which is designated as being a member of a family, or comparison of the sequence to a probabilistic model that describes the diversity of proteins in the family (e.g., HMMs). Once the metagenomic sequence has been compared to all proteins or all models, it can either be classified into (1) a single family (e.g., the family with the best hit), (2) a series of families (e.g., all families that exhibit a significant classification score), or (3) no family, which suggests that the protein may be novel, highly diverged, or spurious. There are exceptions to this annotation framework, such as the gene recruitment procedure mentioned in the Gene Prediction section, though they are less commonly used.

There are many databases that can be used to functionally annotate metagenomic proteins. They generally come in two varieties: sequence databases and HMM databases. Comparing metagenomic reads to a database of sequences tends to be relatively fast and may produce more specific hits for reads that are closely related to sequences in the database, whereas comparing metagenomic reads to a database of HMMs tends to identify more distantly related and diverged members of a family, though their precision for very short sequences is not well explored. Commonly used sequence databases include the SEED annotation system, which is employed by MG-RAST and links

specific family level functions into higher-order functional sub-systems (Overbeek et al., 2014). KEGG orthology groups have proven to be particularly useful as they conveniently map to KEGG metabolic pathway modules (Kanehisa et al., 2014). MetaCyc is similar in that the families are mapped to highly curated and well-described metabolic pathways, though their reliance on functional precision comes at the expense of database sequence diversity (Caspi et al., 2014). EggNOG is a database of non-supervised orthologs groups of proteins that tends to be frequently updated so as to include a relatively large amount of sequence diversity (Powell et al., 2014). The use of HMM databases in metage-nomic analyses tends to be limited to Pfam, which uses HMMs to model protein domains (Finn et al., 2014). Recent years have seen the generation of databases of HMMs of full-length and phylogenetically diverse protein families. This includes Phylofacts (Afrasiabi et al., 2013) and the SiftingFamilies database (Sharp-ton et al., 2012), which, like EggNOG, tend to be frequently updated.

Protein family classification of metagenomic reads tends to require substantial computing resources because all metagenomic peptides are compared to all protein sequences or models in the database. Fortunately, each comparison is independent, so computing clusters and multi-core servers can distribute the com-putational load in parallel to improve throughput. There are several web servers that interface with distributed computing clusters to conduct gene prediction, the database search, family classification and annotation. These include MG-RAST (Meyer et al., 2008), CAMERA (Sun et al., 2011), and IMG/M (Markowitz et al., 2014). These tools tend to be relatively easy to use, though they do place some constraints on the analysis (e.g., the protein family database). As an added benefit, these resources provide pub-lic access to many metagenomes and comparative metagenomic tools. There are also standalone workflows that researchers can install on their own systems, such as RAAMCAP (Li, 2009), Smash-Community (Arumugam et al., 2010) and MetAMOS (Treangen et al., 2013), which often provide more analytical flexibility. The Human Microbiome Project data was annotated using HUMAnN (Abubucker et al., 2012), which maps metagenomic reads to KEGG pathways to produce pathway coverage and abundance profiles. There are also post-processing tools that analyze protein family classification results produced independently by the researcher. For example, ShotgunFunctionalizeR (Kristiansson et al., 2009) is an R package that enables comparative metagenomic analy-ses including the identification of families and pathways that are overrepresented in particular samples or that correlate with partic-ular sample properties (e.g., environmental conditions). Similarly, LefSe (Segata et al., 2012) conducts robust statistical tests to iden-tify those taxa, genes, or pathways that stratify two or more metagenomes.

While protein family classification of metagenomic reads is a useful way of inferring community function, it is imperfect. First, the functional diversity encoded in the metagenome may only approximate the community's functional activity. The pres-ence of a gene does not mean that it is expressed at the time of sampling. That said, comparative metagenomic and metatran-scriptomic analyses indicate that differences between communities at the transcriptional level are often mirrored at the genomic

level, suggesting that metagenomes may provide a meaningful proxy for activity (Mason et al., 2012). Additionally, the detec-tion of enriched functions in a metagenome suggests that they are important to some aspect of the dynamic interaction between the community and its environment or host. Analysis of meta-transcriptomic and metaproteomic data can provide additional insight into which pathways are actively expressed in the commu-nity, though they provide lower coverage of the functions found in the community (Gilbert et al., 2010; Simon and Daniel, 2011; Mason et al., 2012). Second, most databases contain families that have no known functional annotation. Metagenomic reads that are determined to be homologs of such families will not be ascribed a function. These families can still be informative, as they can pro-vide support for metagenomic coding sequence predictions and may be useful diagnostics. Third, the protein family database used to annotate the sequences may be subject to phylogenetic biases, such that certain communities are disproportionately more accu-rately or more thoroughly annotated than others (Wu et al., 2009). Each database also uses different approaches for identifying fami-lies and functionally annotating them. The result is that different databases may annotate different proportions of the metagenome and may produce different functional profiles that describe the community. Fourth, this method presumes that function is rela-tively evolutionarily static. Evolutionarily plastic functions erode the specificity with which function can be inferred. Finally, there may be more proteins and functions in nature than those that have been described by current sequence databases (Wu et al., 2009; Godzik, 2011). Novel strategies for functionally annotating metagenomes and improvements in the way predicted metage-nomic proteins are integrated into protein family databases are needed.

CONCLUSION

Researchers interested in analyzing metagenomes to characterize microbial community diversity and function now have a litany of tools and data resources at their disposal. Many of the tools discussed here were developed for researchers comfortable inter-facing with a command-line environment. This is understandable given the complexity of metagenomic data and the computational requirements traditionally associated with its analysis. But, many researchers interested in metagenomic analysis may not have expe-rience working with this type of software or access to the necessary computational resources. Fortunately, there are many web-based tools that centralize metagenome data management and analysis and provide researchers with the means to annotate and compare metagenomes through an easy-to-use interface (**Table 1**). These tools will not necessarily conduct all analytical strategies and fre-quently do not provide the flexibility and customization of their command-line counterparts.

Knowing which analyses to conduct and which tools to apply remain confusing questions for many scientists. The answer depends largely on several variables, including the hypothesis and goals, the experimental design, and the known properties of the community. For example, a researcher that is interested in identi-fying well-curated metabolic pathways that are overrepresented in a community may elect to use a database optimized for pathway curation, like KEGG or MetaCyc, to annotate their metagenomes.

Table 1 | Web-based metagenomic analysis resources.

Resource	Methods	Citation	Web link
AmphoraNet	Marker gene analysis: phylogeny	Kerepesi et al. (2014)	http://pitgroup.org/amphoranet/
CAMERA	Various: taxonomic and functional annotation, comparative analyses	Sun et al. (2011)	http://camera.calit2.net/
Comet	Functional annotation, comparative analyses	Lingner et al. (2011)	http://comet.gobics.de/
LEfSe (Galaxy)	Comparative analyses	Segata et al. (2011)	http://huttenhower.sph.harvard.edu/galaxy/
IMG/M	Various: taxonomic and functional annotation, comparative analyses	Markowitz et al. (2014)	https://img.jgi.doe.gov/m/
MG-RAST	Various: taxonomic and functional annotation, comparative analyses	Meyer et al. (2008)	http://metagenomics.anl.gov/
MALINA	Various: taxonomic and functional annotation, comparative analyses	Tyakht et al. (2012)	http://malina.metagenome.ru/
METAGENassist	Various: taxonomic annotation, comparative analyses	Arndt et al. (2012)	http://www.metagenassist.ca/
MetaPhlAn (Galaxy)	Marker gene analysis: similarity	Segata et al. (2012)	http://huttenhower.sph.harvard.edu/galaxy/
NBC	Binning: compositional classification	Rosen et al. (2011)	http://nbc.ece.drexel.edu
Orphelia	Gene prediction	Hoff et al. (2009)	http://orphelia.gobics.de/
Phylopithia webserver	Binning: compositional classification	Patil et al. (2012)	http://phylopythias.cs.uni-duesseldorf.de/
Real time metagenomics	Functional annotation	Edwards et al. (2012)	http://edwards.sdsu.edu/rtmg/
WebCARMA	Binning: sequence similarity	Gerlach et al. (2009)	http://webcarma.cebitec.uni-bielefeld.de/
WebMGA	Various: taxonomic and functional annotation	Wu et al. (2011)	http://weizhong-lab.ucsd.edu/metagenomic-analysis/

Conversely, researchers interested in counting the total distinct types of proteins may want to use a database that optimizes for phylogenetic diversity. If the community is known to contain phylogenetically diverged lineages relative to genome sequence databases, then it may be better to use taxonomic annotation techniques that are more tolerant of sequence divergence than fragment recruitment methods. If the main objective is to characterize the genome of a relatively abundant organism in the community, then metagenomic assembly may be warranted. Consideration of the assumptions and limitations of the analytical strategies and tools is critical as the improper approach may fail to detect a meaningful signature or, worse, identify a spurious result.

There are many areas where metagenomic analysis can be improved. First, the precision, thoroughness, and throughput of the analytical strategies reviewed here can be increased. Additional analytical methods (e.g., non-coding RNA detection; Weinberg et al., 2009) are also needed. Second, many of the tools that are currently available would benefit from expansion of the diversity of genome sequence databases, which are frequently queried as referential information during metagenomic analysis. Third, infrastructural developments associated with managing and serving sequence data are needed. Given the plummeting costs of DNA sequencing, it is realistic for researchers to generate massive metagenomes across a large number of samples. The rapid growth in the size of data complicates its storage, organization, and distribution. Fourth, improved statistical methodology is needed, especially for metagenomes generated from complex communities where data for any given taxon or protein may be sparse. Statistical methodology can also improve the identification of biomarkers from comparative studies where a large number of covariates (e.g., environmental or host physiological parameters) are collected for each sample. Finally, additional experimental systems that provide opportunities to manipulate communities, especially microbiota, are needed. Because the results identified through the comparison of metagenomes are typically associative, most studies only produce hypotheses about how communities interact with their environment. Modulating the community composition (e.g., antibiotic administration, gnotobiotic hosts, probiotic supplementation, community transplantation, mono-association of specific taxa) and evaluating the effect on the environment or host provides a direct test of these hypotheses. By coupling metagenomics with this experimental framework across a diverse array of systems, insight into the general rules and properties of community-environment interaction can be gleaned.

Finally, I thank the Gordon and Betty Moore Foundation for generously sponsoring this work (grant #3300) and two anonymous reviewers for their scientific and editorial suggestions.

REFERENCES

Abubucker, S., Segata, N., Goll, J., Schubert, A. M., Izard, J., Cantarel, B. L., et al. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.* 8:e1002358. doi: 10.1371/journal.pcbi.1002358

Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., and Polz, M. F. (2004). Divergence and redundancy of 16S rRNA sequences in genomes with multiple Rrn operons. *J. Bacteriol.* 186, 2629–2635. doi: 10.1128/JB.186.9.2629-2635.2004

Afrasiabi, C., Samad, B., Dineen, D., Meacham, C., and Sjölander. K. (2013). The PhyloFacts FAT-CAT web server: ortholog identification and function prediction using fast approximate tree classification. *Nucleic Acids Res.* 41, W242–W248. doi: 10.1093/nar/gkt399

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Aminov, R. I. (2011). Horizontal gene exchange in environmental microbiota. *Front. Microbiol.* 2:158. doi: 10.3389/fmicb.2011.00158

Arndt, D., Xia, J., Liu, Y., Zhou, Y., Guo, A. C., Cruz, J. A., et al. (2012). META-GENassist: a comprehensive web server for comparative metagenomics. *Nucleic Acids Res.* 40, W88–W95. doi: 10.1093/nar/gks497

Arrigo, K. R. (2005). Marine microorganisms and global nutrient cycles. *Nature* 437, 349–355. doi: 10.1038/nature04159

Arumugam, M., Harrington, E. D., Foerstner, K. U., Raes, J., and Bork, P. (2010). SmashCommunity: a metagenomic annotation and analysis tool. *Bioinformatics* 26, 2977–2978. doi: 10.1093/bioinformatics/btq536

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., and Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol.* 13:R122. doi: 10.1186/gb-2012-13-12-r122

Bowen De León, K., Gerlach, R., Peyton, B. M., and Fields, M. W. (2013). Archaeal and bacterial communities in three alkaline hot springs in Heart Lake Geyser Basin, Yellowstone National Park. *Front. Microbiol.* 4:330. doi: 10.3389/fmicb.2013.00330

Brady, A., and Salzberg, S. (2011). PhymmBL expanded: confidence scores, custom databases, parallelization and more. *Nat. Methods* 8:367. doi: 10.1038/nmeth0511-367

Brady, A., and Salzberg, S. L. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat. Methods* 6, 673–676. doi: 10.1038/nmeth.1358

Brown, C. T., Howe, A., Zhang, Q., Pyrkosz, A. B., and Brom, T. H. (2012). A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. Genomics, arXiv:1203.4802.

Bulgarelli, D., Rott, M., Schlaeppi, K., Ver Loren van Themaat, E., Ahmadinejad, N., Assenza, F., et al. (2012). Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. *Nature* 488, 91–95. doi: 10.1038/nature11336

Bulgarelli, D., Schlaeppi, K., Spaepen, S., Ver Loren van Themaat, E., and Schulze-Lefert, P. (2013). Structure and functions of the bacterial microbiota of plants. *Annu. Rev. Plant Biol.* 64, 807–838. doi: 10.1146/annurev-arplant-050312-120106

Buttigieg, P. L., Hankeln, W., Kostadinov, I., Kottmann, R., Yilmaz, P., Duhaime, M. B., et al. (2013). Ecogenomic perspectives on domains of unknown function: correlation-based exploration of marine metagenomes. *PLoS ONE* 8:e50869. doi: 10.1371/journal.pone.0050869

Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., et al. (2014). The MetaCyc database of metabolic pathways and enzymes and the Bio-Cyc collection of pathway/genome databases. *Nucleic Acids Res.* 42, D459–D471. doi: 10.1093/nar/gkt1103

Chew, Y. V., and Holmes, A. J. (2009). Suppression subtractive hybridisation allows selective sampling of metagenomic subsets of interest. *J. Microbiol. Methods* 78, 136–143. doi: 10.1016/j.mimet.2009.05.003

Compeau, P. E., Pevzner, P. A., and Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* 29, 987–991. doi: 10.1038/nbt.2023

Cook, R. J., Thomashow, L. S., Weller, D. M., Fujimoto, D., Mazzola, M., Bangera, G., et al. (1995). Molecular mechanisms of defense by rhizobacteria against root disease. *Proc. Natl. Acad. Sci. U.S.A.* 92, 4197–4201. doi: 10.1073/pnas.92.10.4197

Darling, A. E., Jospin, G., Lowe, E., Matsen, F. A. IV, Bik, H. M., and Eisen, J. A. (2014). PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2:e243. doi: 10.7717/peerj.243

Davenport, C. F., Neugebauer, J., Beckmann, N., Friedrich, B., Kameri, B., Kokott, S., et al. (2012). Genometa – a fast and accurate classifier for short metagenomic shotgun reads. *PLoS ONE* 7:e41224. doi: 10.1371/journal.pone.0041224

Davenport, C. F., and Tümmler, B. (2013). Advances in computational analysis of metagenome sequences. *Environ. Microbiol.* 15, 1–5. doi: 10.1111/j.1462-2920.2012.02843.x

David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., et al. (2014). Diet rapidly and reproducibly alters the human gut microbiome. *Nature* 505, 559–563. doi: 10.1038/nature12820

Degnan, P. H., and Ochman, H. (2012). Illumina-based analysis of microbial community diversity. *ISME J.* 6, 183–194. doi: 10.1038/ismej.2011.74

Delmotte, N., Knief, C., Chaffron, S., Innerebner, G., Roschitzki, B., Schlapbach, R., et al. (2009). Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 106, 16428–16433. doi: 10.1073/pnas.0905240106

Deshpande, V., Fung, E. D. K., Pham, S., and Bafna, V. (2013). Cerulean: A Hybrid Assembly Using High Throughput Short and Long Reads. Quantitative Methods Genomics, arXiv:1307.7933.

Deshpande, V., Fung, E. D. K., Pham, S., and Bafna, V. (2013). Cerulean: a hybrid assembly using high throughput short and long reads. *Quant. Methods Genomics.* doi: 10.1093/bioinformatics/bts721

Dick, G. J., Andersson, A. F., Baker, B. J., Simmons, S. L., Thomas, B. C., Yelton, A. P., et al. (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol.* 10:R85. doi: 10.1186/gb-2009-10-8-r85

Dimkpa, C., Weinand, T., and Asch, F. (2009). Plant-rhizobacteria interactions alleviate abiotic stress conditions. *Plant Cell Environ.* 32, 1682–1694. doi: 10.1111/j.1365-3040.2009.02028.x

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460–2461. doi: 10.1093/bioinformatics/btq461

Edwards, R. A., Olson, R., Disz, T., Pusch, G. D., Vonstein, V., Stevens, R., et al. (2012). Real time metagenomics: using K-mers to annotate metagenomes. *Bioinformatics* 28, 3316–3317. doi: 10.1093/bioinformatics/bts599

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi: 10.1093/nar/gkt1223

Garcia-Garcerà, M., Garcia-Etxebarria, K., Coscollà, M., Latorre, A., and Calafell, F. (2013). A new method for extracting skin microbes allows metagenomic analysis of whole-deep skin. *PLoS ONE* 8:e74914. doi: 10.1371/journal.pone.0074914

Gerlach, W., Jünemann, S., Tille, F., Goesmann, A., and Stoye, J. (2009). WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics* 10:430. doi: 10.1186/1471-2105-10-430

Gerlach, W., and Stoye, J. (2011). Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res.* 39:e91. doi: 10.1093/nar/gkr225

Gilbert, J. A., Field, D., Swift, P., Thomas, S., Cummings, D., Temperton, B., et al. (2010). The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. *PLoS ONE* 5:e15545. doi: 10.1371/journal.pone.0015545

Godzik, A. (2011). Metagenomics and the protein universe. *Curr. Opin. Struct. Biol.* 21, 398–403. doi: 10.1016/j.sbi.2011.03.010

Hoff, K. J., Lingner, T., Meinicke, P., and Tech, M. (2009). Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res.* 37, W101–W105. doi: 10.1093/nar/gkp327

Hong, S., Bunge, J., Leslin, C., Jeon, S., and Epstein, S. S. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J.* 3, 1365–1373. doi: 10.1038/ismej.2009.89

Hooper, L. V., Littman, D. R., and Macpherson, A. J. (2012). Interactions between the microbiota and the immune system. *Science* 336, 1268–1273. doi: 10.1126/science.1223490

Hugenholtz, P., and Pace, N. R. (1996). Identifying microbial diversity in the natural environment: a molecular phylogenetic approach. *Trends Biotechnol.* 14, 190–197. doi: 10.1016/0167-7799(96)10025-1

Human Microbiome Project Consortium. (2012a). Structure, function and diversity of the healthy human microbiome. *Nature* 486, 207–214. doi: 10.1038/nature11234

Human Microbiome Project Consortium. (2012b). A framework for human microbiome research. *Nature* 486, 215–221. doi: 10.1038/nature11209

Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N., and Schuster, S. C. (2011). Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 21, 1552–1560. doi: 10.1101/gr.120618.111

Iverson, V., Morris, R. M., Frazar, C. D., Berthiaume, C. T., Morales, R. L., and Armbrust, E. V. (2012). Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* 335, 587–590. doi: 10.1126/science.1212665

Jiao, J.-Y., Wang, H.-X., Zeng, Y., and Shen, Y.-M. (2006). Enrichment for microbes living in association with plant tissues. *J. Appl. Microbiol.* 100, 830–837. doi: 10.1111/j.1365-2672.2006.02830.x

Jumpstart Consortium Human Microbiome Project Data Generation Working Group. (2012). Evaluation of 16S rDNA-based community profiling for human microbiome research. *PLoS ONE* 7:e39315. doi: 10.1371/journal.pone.0039315

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205. doi: 10.1093/nar/gkt1076

Kelley, D. R., Liu, B., Delcher, A. L., Pop, M., and Salzberg, S. L. (2012). Gene prediction with glimmer for metagenomic sequences augmented by classification and clustering. *Nucleic Acids Res.* 40:e9. doi: 10.1093/nar/gkr1067

Kennedy, P. G., Hortal, S., Bergemann, S. E., and Bruns, T. D. (2007). Competitive interactions among three ectomycorrhizal fungi and their relation to host plant performance. *J. Ecol.* 95, 1338–1345. doi: 10.1111/j.1365-2745.2007.01306.x

Kerepesi, C., Bánky, D., and Grolmusz, V. (2014). AmphoraNet: the webserver implementation of the AMPHORA2 metagenomic workflow suite. *Gene* 533, 538–540. doi: 10.1016/j.gene.2013.10.015

Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., and Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Res.* 21, 487–493. doi: 10.1101/gr.113985.110

Knief, C., Delmotte, N., Chaffron, S., Stark, M., Innerebner, G., Wassmann, R., et al. (2012). Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. *ISME J.* 6, 1378–1390. doi: 10.1038/ismej.2011.192

Kristiansson, E., Hugenholtz, P., and Dalevi, D. (2009). ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* 25, 2737–2738. doi: 10.1093/bioinformatics/btp508

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., et al. (2011). Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.* 13, 47–58. doi: 10.1038/nrg3129

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., and Hugenholtz, P. (2008). A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.* 72, 557–578. doi: 10.1128/MMBR.00009-08

Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Laserson, J., Jojic, V., and Koller, D. (2011). Genovo: de novo assembly for metagenomes. *J. Comput. Biol.* 18, 429–443.

Lee, W. P., Stromberg, M., Ward, A., Stewart, C., Garrison, E., and Marth, G. T. (2013). MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Read Mapping. Genomics Quantitative Methods, arXiv:1309.1149.

Li, W. (2009). Analysis and comparison of very large metagenomes with fast clustering and functional annotation. *BMC Bioinformatics* 10:359. doi: 10.1186/1471-2105-10-359

Lingner, T., Asshauer, K. P., Schreiber, F., and Meinicke, P. (2011). CoMet–a web server for comparative functional profiling of metagenomes. *Nucleic Acids Res.* 39, W518–W523. doi: 10.1093/nar/gkr388

Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., and Pop, M. (2011). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics* 12(Suppl. 2):S4. doi: 10.1186/1471-2164-12-S2-S4

Liu, Y., Guo, J., Hu, G., and Zhu, H. (2013). Gene prediction in metagenomic fragments based on the SVM algorithm. *BMC Bioinformatics* 14(Suppl. 5):S12. doi: 10.1186/1471-2105-14-S5-S12

Liu, Z., DeSantis, T. Z., Andersen, G. L., and Knight, R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res.* 36:e120. doi: 10.1093/nar/gkn491

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmento, H., et al. (2013). Metagenomic 16S rDNA illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* doi: 10.1111/1462-2920.12250 [Epub ahead of print].

Looft, T., Johnson, T. A., Allen, H. K., Bayles, D. O., Alt, D. P., Stedtfeld, R. D., et al. (2012). In-feed antibiotic effects on the swine intestinal microbiome. *Proc. Natl. Acad. Sci. U.S.A.* 109, 1691–1696. doi: 10.1073/pnas.1120238109

Lozupone, C. A., and Knight, R. (2007). Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. U.S.A.* 104, 11436–11440. doi: 10.1073/pnas.0611525104

Lundberg, D. S., Lebeis, S. L., Paredes, S. H., Yourstone, S., Gehring, J., Malfatti, S., et al. (2012). Defining the core *Arabidopsis thaliana* root microbiome. *Nature* 488, 86–90. doi: 10.1038/nature11237

Luo, C., Tsementzi, D., Kyrpides, N. C., and Konstantinidis, K. T. (2012). Individual genome assembly from complex community short-read metagenomic datasets. *ISME J.* 6, 898–901. doi: 10.1038/ismej.2011.147

Markowitz, V. M., Chen, I.-M., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., et al. (2014). IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Res.* 42, D568–D573. doi: 10.1093/nar/gkt919

Martin, J., Sykes, S., Young, S., Kota, K., Sanka, R., Sheth, N., et al. (2012). Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. *PLoS ONE* 7:e36427. doi: 10.1371/journal.pone.0036427

Mason, O. U., Hazen, T. C., Borglin, S., Chain, P. S., Dubinsky, E. A., Fortney, J. L., et al. (2012). Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME J.* 6, 1715–1727. doi: 10.1038/ismej.2012.59

Matsen, F. A. IV, and Evans, S. N. (2013). Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. *PLoS ONE* 8:e56859. doi: 10.1371/journal.pone.0056859

Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A. C., et al. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat. Methods* 4, 495–500. doi: 10.1038/nmeth1043

McCliment, E. A., Voglesonger, K. M., O'Day, P. A., Dunn, E. E., Holloway, J. R., and Cary, S. C. (2006). Colonization of nascent, deep-sea hydrothermal vents by a novel Archaeal and Nanoarchaeal assemblage. *Environ. Microbiol.* 8, 114–125. doi: 10.1111/j.1462-2920.2005.00874.x

McDonald, E., and Brown, C. T. (2013). Khmer: Working with Big Data in Bioinformatics. Computational Engineering, Finance, and Science, arXiv:1303.2223.

McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., and Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods* 4, 63–72. doi: 10.1038/nmeth976

Mende, D. R., Waller, A. S., Sunagawa, S., Järvelin, A. I., Chan, M. M., Arumugam, M., et al. (2012). Assessment of metagenomic assembly using simulated next generation sequencing data. *PLoS ONE* 7:e31386. doi: 10.1371/journal.pone.0031386

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST Server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386

Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., et al. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol.* 13:R79. doi: 10.1186/gb-2012-13-9-r79

Muegge, B. D., Kuczynski, J., Knights, D., Clemente, J. C., González, A., Fontana, L., et al. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* 332, 970–974. doi: 10.1126/science.1198719

Nacke, H., Engelhaupt, M., Brady, S., Fischer, C., Tautzt, J., and Daniel, R. (2012). Identification and characterization of novel cellulolytic and hemicellulolytic genes

and enzymes derived from german grassland soil metagenomes. *Biotechnol. Lett.* 34, 663–675. doi: 10.1007/s10529-011-0830-2

Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.* 40:e155. doi: 10.1093/nar/gks678

Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res.* 34, 5623–5630. doi: 10.1093/nar/gkl723

Overbeek, R., Olson, R., Pusch, G. D., Olsen, G. J., Davis, J. J., Disz, T., et al. (2014). The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res.* 42, D206–D214. doi: 10.1093/nar/gkt1226

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science* 276, 734–740. doi: 10.1126/science.276.5313.734

Pace, N. R., Stahl, D. A., Lane, D. J., and Olsen, G. J. (1986). The analysis of natural microbial populations by ribosomal RNA sequences. *Adv. Microb. Ecol.* 9, 1–55. doi: 10.1007/978-1-4757-0611-6

Patil, K. R., Haider, P., Pope, P. B., Turnbaugh, P. J., Morrison, M., Scheffer, T., et al. (2011). Taxonomic metagenome sequence assignment with structured output models. *Nat. Methods* 8, 191–192. doi: 10.1038/nmeth0311-191

Patil, K. R., Roune, L., and McHardy, A. C. (2012). The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. *PLoS ONE* 7:e38581. doi: 10.1371/journal.pone.0038581

Peng, Y., Leung, H. C., Yiu, S. M., and Chin, F. Y. (2011). Meta-IDBA: a de novo assembler for metagenomic data. *Bioinformatics* 27, i94–i101. doi: 10.1093/bioinformatics/btr216

Philippot, L., Raaijmakers, J. M., Lemanceau, P., and van der Putten, W. H. (2013). Going back to the roots: the microbial ecology of the rhizosphere. *Nat. Rev. Microbiol.* 11, 789–799. doi: 10.1038/nrmicro3109

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., et al. (2014). eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* 42, D231–D239. doi: 10.1093/nar/gkt1253

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464, 59–65. doi: 10.1038/nature08821

Rappé, M. S., and Giovannoni, S. J. (2003). The uncultured microbial majority. *Annu. Rev. Microbiol.* 57, 369–394. doi: 10.1146/annurev.micro.57.030502.090759

Raupach, G. S., and Kloepper, J. W. (1998). Mixtures of plant growth-promoting rhizobacteria enhance biological control of multiple cucumber pathogens. *Phytopathology* 88, 1158–1164. doi: 10.1094/PHYTO.1998.88.11.1158

Redman, R. S., Sheehan, K. B., Stout, R. G., Rodriguez, R. J., and Henson, J. M. (2002). Thermotolerance generated by plant/fungal symbiosis. *Science* 298:1581. doi: 10.1126/science.1078055

Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 38:e191. doi: 10.1093/nar/gkq747

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the european molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/S0168-9525(00)02024-2

Richardson, E. J., and Watson, M. (2013). The automatic annotation of bacterial genomes. *Brief. Bioinform.* 14, 1–12. doi: 10.1093/bib/bbs007

Rosen, G., Garbarine, E., Caseiro, D., Polikar, R., and Sokhansanj, B. (2008). Metagenome fragment classification using N-mer frequency profiles. *Adv. Bioinformatics* 2008:205969. doi: 10.1155/2008/205969

Rosen, G. L., Reichenberger, E. R., and Rosenfeld, A. M. (2011). NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* 27, 127–129. doi: 10.1093/bioinformatics/btq619

Ruby, J. G., Bellare, P., and Derisi, J. L. (2013). PRICE: software for the targeted assembly of components of (Meta) genomic sequence data. *G3* (*Bethesda*) 3, 865–880. doi: 10.1534/g3.113.005967

Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., et al. (2007). The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol.* 5:e77. doi: 10.1371/journal.pbio.0050077

Saeed, I., and Halgamuge, S. K. (2009). The oligonucleotide frequency derived error gradient and its application to the binning of metagenome fragments. *BMC Genomics* 10(Suppl. 3):S10. doi: 10.1186/1471-2164-10-S3-S10

Saeed, I., Tang, S. L., and Halgamuge, S. K. (2012). Unsupervised discovery of microbial population structure within metagenomes using nucleotide base composition. *Nucleic Acids Res.* 40:e34. doi: 10.1093/nar/gkr1204

Salzberg, S. L., Delcher, A. L., Kasif, S., and White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* 26, 544–548. doi: 10.1093/nar/26.2.544

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., et al. (2013). Genomic variation landscape of the human gut microbiome. *Nature* 493, 45–50. doi: 10.1038/nature11711

Schloss, P. D. (2010). The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput. Biol.* 6:e1000844. doi: 10.1371/journal.pcbi.1000844

Schloss, P. D., and Handelsman, J. (2008). A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics* 9:34. doi: 10.1186/1471-2105-9-34

Schmieder, R., and Edwards, R. (2011a). Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS ONE* 6:e17288. doi: 10.1371/journal.pone.0017288

Schmieder, R., and Edwards, R. (2011b). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27, 863–864. doi: 10.1093/bioinformatics/btr026

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12:R60. doi: 10.1186/gb-2011-12-6-r60

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066

Sessitsch, A., Hardoim, P., Döring, J., Weilharter, A., Krause, A., Woyke, T., et al. (2012). Functional characteristics of an endophyte community colonizing rice roots as revealed by metagenomic analysis. *Mol. Plant Microbe Interact.* 25, 28–36. doi: 10.1094/MPMI-08-11-0204

Sharp, C. E., Brady, A. L., Sharp, G. H., Grasby, S. E., Stott, M. B., and Dunfield, P. F. (2014). Humboldt's spa: microbial diversity is controlled by temperature in geothermal environments. *ISME J.* doi: 10.1038/ismej.2013.237 [Epub ahead of print].

Sharpton, T. J., Jospin, G., Wu, D., Langille, M. G., Pollard, K. S., and Eisen, J. A. (2012). Sifting through genomes with iterative-sequence clustering produces a large, phylogenetically diverse protein-family resource. *BMC Bioinformatics* 13:264. doi: 10.1186/1471-2105-13-264

Sharpton, T. J., Riesenfeld, S. J., Kembel, S. W., Ladau, J., O'Dwyer, J. P., Green, J. L., et al. (2011). PhylOTU: a high-throughput procedure quantifies microbial community diversity and resolves novel taxa from metagenomic data. *PLoS Comput. Biol.* 7:e1001061. doi: 10.1371/journal.pcbi.1001061

Simmons, S. L., Dibartolo, G., Denef, V. J., Goltsman, D. S., Thelen, M. P., and Banfield, J. F. (2008). Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biol.* 6:e177. doi: 10.1371/journal.pbio.0060177

Simon, C., and Daniel, R. (2011). Metagenomic analyses: past and future trends. *Appl. Environ. Microbiol.* 77, 1153–1161. doi: 10.1128/AEM.02345-10

Smith, M. I., Yatsunenko, T., Manary, M. J., Trehan, I., Mkakosya, R., Cheng, J., et al. (2013). Gut microbiomes of Malawian twin pairs discordant for kwashiorkor. *Science* 339, 548–554. doi: 10.1126/science.1229000

Soo, R. M., Wood, S. A., Grzymski, J. J., McDonald, I. R., and Cary, S. C. (2009). Microbial biodiversity of thermophilic communities in hot mineral soils of Tramway Ridge, Mount Erebus, Antarctica. *Environ. Microbiol.* 11, 715–728. doi: 10.1111/j.1462-2920.2009.01859.x

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., et al. (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and analysis: the CAMERA resource. *Nucleic Acids Res.* 39, D546–D551. doi: 10.1093/nar/gkq1102

Thomas, T., Gilbert, J., and Meyer, F. (2012). Metagenomics – a guide from sampling to data analysis. *Microb. Inform. Exp.* 2:3. doi: 10.1186/2042-5783-2-3

Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaya, I., Ondov, B., et al. (2013). MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol.* 14:R2. doi: 10.1186/gb-2013-14-1-r2

Trimble, W. L., Keegan, K. P., D'Souza, M., Wilke, A., Wilkening, J., Gilbert, J., et al. (2012). Short-read reading-frame predictors are not created equal: sequence

error causes loss of signal. *BMC Bioinformatics* 13:183. doi: 10.1186/1471-2105-13-183

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., et al. (2009). A core gut microbiome in obese and lean twins. *Nature* 457, 480–484. doi: 10.1038/nature07540

Tyakht, A. V., Popenko, A. S., Belenikin, M. S., Altukhov, I. A., Pavlenko, A. V., Kostryukova, E. S., et al. (2012). MALINA: a web service for visual analytics of human gut microbiota whole-genome metagenomic reads. *Source Code Biol. Med.* 7:13. doi: 10.1186/1751-0473-7-13

van der Heijden, M. G., Bardgett, R. D., and van Straalen, N. M. (2008). The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecol. Lett.* 11, 296–310. doi: 10.1111/j.1461-0248.2007.01139.x

Vorholt, J. A. (2012). Microbial life in the phyllosphere. *Nat. Rev. Microbiol.* 10, 828–840. doi: 10.1038/nrmicro2910

Walter, J., and Ley, R. (2011). The human gut microbiome: ecology and recent evolutionary changes. *Annu. Rev. Microbiol.* 65, 411–429. doi: 10.1146/annurev-micro-090110-102830

Weinberg, Z., Perreault, J., Meyer, M. M., and Breaker, R. R. (2009). Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. *Nature* 462, 656–659. doi: 10.1038/nature08586

Woyke, T., Teeling, H., Ivanova, N. N., Huntemann, M., Richter, M., Gloeckner, F. O., et al. (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* 443, 950–955. doi: 10.1038/nature05192

Wrighton, K. C., Thomas, B. C., Sharon, I., Miller, C. S., Castelle, C. J., VerBerkmoes, N. C., et al. (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 337, 1661–1665. doi: 10.1126/science.1224041

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., et al. (2009). A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature* 462, 1056–1060. doi: 10.1038/nature08656

Wu, D., Jospin, G., and Eisen, J. A. (2013). Systematic identification of gene families for use as 'markers' for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. *PLoS ONE* 8:e77033. doi: 10.1371/journal.pone.0077033

Wu, M., and Eisen, J. A. (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* 9:R151. doi: 10.1186/gb-2008-9-10-r151

Wu, M., and Scott, A. J. (2012). Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* 28, 1033–1034. doi: 10.1093/bioinformatics/bts079

Wu, S., Zhu, Z., Fu, L., Niu, B., and Li, W. (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics* 12:444. doi: 10.1186/1471-2164-12-444

Wylie, K. M., Truty, R. M., Sharpton, T. J., Mihindukulasuriya, K. A., Zhou, Y., Gao, H., et al. (2012). Novel bacterial taxa in the human microbiome. *PLoS ONE* 7:e35294. doi: 10.1371/journal.pone.0035294

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13, 329–342. doi: 10.1038/nrg3174

Yang, J., Kloepper, J. W., and Ryu, C. M. (2009). Rhizosphere bacteria help plants tolerate abiotic stress. *Trends Plant Sci.* 14, 1–4. doi: 10.1016/j.tplants.2008.10.004

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., et al. (2012). Human gut microbiome viewed across age and geography. *Nature* 486, 222–227. doi: 10.1038/nature11053

Yok, N. G., and Rosen, G. L. (2011). Combining gene prediction methods to improve metagenomic gene annotation. *BMC Bioinformatics* 12:20. doi: 10.1186/1471-2105-12-20

Yozwiak, N. L., Skewes-Cox, P., Stenglein, M. D., Balmaseda, A., Harris, E., and DeRisi, J. L. (2012). Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl. Trop. Dis.* 6:e1485. doi: 10.1371/journal.pntd.0001485

Zhao, Y., Tang, H., and Ye, Y. (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* 28, 125–126. doi: 10.1093/bioinformatics/btr595

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res.* 38:e132. doi: 10.1093/nar/gkq275