BRIEFINGS IN BIOINFORMATICS. VOL 13. NO 6. 646–655
Advance Access published on 31 July 2012
doi:10.1093/bib/bbs031

Taxonomic binning of metagenome samples generated by next-generation sequencing technologies

Johannes Dröge and Alice C. McHardy

Submitted: 15th March 2012; Received (in revised form): 17th May 2012

Abstract

Metagenome research uses random shotgun sequencing of microbial community DNA to study the genetic sequences of its members without cultivation. This development has been strongly supported by improvements in sequencing technologies, which have rendered sequencing cheaper than before. As a consequence, downstream computational analysis of metagenome sequence samples is now faced with large amounts of complex data. One of the essential steps in metagenome analysis is reconstruction of draft genomes for populations of a community or of draft 'pan-genomes' for higher level clades. 'Taxonomic binning' corresponds to the process of assigning a taxonomic identifier to sequence fragments, based on information such as sequence similarity, sequence composition or read coverage. This is used for draft genome reconstruction, if sequencing coverage is insufficient for reconstruction based on assembly information alone. Subsequent functional and metabolic annotation of draft genomes allows a genome-level analysis of novel uncultured microbial species and even inference of their cultivation requirements.

Keywords: *metagenomics; taxonomic binning; next-generation sequencing*

INTRODUCTION

The application of genome sequencing technologies to the study of an entire community of microbial organisms, as opposed to a clonal culture of an individual isolate strain, is known as metagenomics [1, 2]. Such analysis allows one to determine genome sequence information for a vast portion of the microbial world for which cultivation conditions are unknown or difficult to reproduce under laboratory conditions [3, 4]. Even the first metagenome studies, investigating the Sargasso Sea [5] and Minnesota farm soil [6], were able to demonstrate the enormous potential of the microbial world to serve as a treasure trove of genes with novel functionalities, as these studies resulted in the discovery of many thousands of new gene sequences that were only remotely similar to genes of known function. They also revealed the unexpected complexity of microbial communities in terms of the number of taxa contained therein. Since then, much research has explored microbial ecosystems, soil, aquatic and host associated, in more detail [7–11] and has revealed a great wealth of novel genetic information from microbial species that are only distantly related to well-studied model organisms.

Both amplicon sequencing and random shotgun sequencing of microbial communities are sometimes referred to as metagenomics. Amplicon sequencing, or environmental tag sequencing, is used to determine the taxonomic composition and phylogenetic structure of a microbial community. In amplicon sequencing, informative marker regions of the

Corresponding author. A. C. McHardy, Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Institute for Computer Science, Universitätsstraße 1, 40225 Düsseldorf, Germany. Tel: +49-211-81-10591; Fax: +49-211-81-13464; E-mail: alice.mchardy@uni-duesseldorf.de; Max Planck Research Group for Computational Genomics & Epidemiology, Max Planck Institute for Informatics, Campus E1 4, 66123 Saarbrücken, Germany.

Johannes Dröge is a PhD student in the department of Algorithmic Bioinformatics at Heinrich Heine University Düsseldorf.

Alice Carolyn McHardy obtained her PhD in bioinformatics in 2004 and holds the chair of Algorithmic Bioinformatics at Heinrich Heine University Düsseldorf.

genomes from a microbial community are amplified by polymerase chain reaction, and used as a proxy to determine which phylotypes or operational taxonomic units (OTUs) are present in a microbial community, and their relative abundance. Commonly used markers regions are the ribosomal genes [12] and the internal transcribed spacer region [13], which is positioned between ribosomal genes. In terms of numbers and the evolutionary closeness of the distinct species present, microbial community profiles can be correlated across environments and communities, linked to environmental parameters. They can be indicative of the presence of genes that are relevant for particular metabolic functionalities [14], given that the respective genes are already known. However, the gene inventory and the encoded functionality of most microbial species are largely unknown and may also vary considerably between strains.

Shotgun sequencing can be used to study the genetic information of microbial communities by sequencing DNA that has been extracted and randomly sheared into smaller fragments. Even though subject to different technology-dependent biases, this procedure allows functional and process-level characterization of microbial communities as a whole and the reconstruction of draft genome sequences for individual community members.

NEXT-GENERATION SEQUENCING TECHNOLOGIES

DNA sequencing technologies have rapidly advanced over the last 5 years and these developments have substantially shaped the way metagenome research is performed. Post-Sanger sequencing technologies are commonly referred to as next-generation sequencing (NGS) [15, 16]. In comparison with Sanger sequencing, NGS methods can sequence DNA more quickly and at lower cost through massive parallelization. This is generally achieved by amplification and fixation of millions of individual template molecules or their enzyme counterparts on a solid phase prior to sequencing. Although Sanger sequencing results in read lengths of ~800 bp, the commercially available NGS technologies (Table 1) currently generate reads of ~50–75 bp (Applied Biosciences/Life Technologies – SOLiD), 75–150 bp (Solexa/Illumina – Sequencing by Synthesis), 100–200 bp (IonTorrent/Life Technologies – Semiconductor Chip Sequencing) and 550–1000 bp (454/Roche – Pyrosequencing). The upcoming generation [17, 18] of sequencers using single molecule sequencing produces read lengths of more than 1 kb (PacBio, SMRT, 15–20% assumed error rate [17]) and of 5–10 kb (Oxford Nanopore technology, 5% assumed error rate). Besides different read lengths and amounts of sequence data produced, each technology has a characteristic profile of sequencing errors, resulting from the technology-specific preparation and detection procedures. The choice of an appropriate sequencing technology depends on the scientific questions asked. For instance, while an 80 bp read is sufficient to cover a hypervariable region in the 16S gene [12] for analysis of microbial community composition, *de novo* recovery of draft microbial genome sequences by taxonomic binning from a complex organismal mixture requires substantially longer reads or higher sequencing depth and sequencing of short paired reads [9, 11, 19, 20].

BIOINFORMATIC ANALYSIS OF METAGENOME SAMPLES

NGS produces large volumes of sequence data (Table 1). Currently, a single run of an Illumina HiSeq machine generates up to 600 Gb per run (www.illumina.com), which is of the order of 10^4 times the amount of data produced in a similar timeframe by a Sanger sequencing chemistry based sequencer (Table 1). This, in turn, results in drastically increased runtimes for all the bioinformatics procedures applied in metagenomics [21], such as assembly of sequence fragments, taxonomic binning, prediction of protein encoding genes, as well as functional and process-level gene annotation. Together, taxonomic binning and assembly allow draft genome reconstructions for community members for which sequencing has recovered substantial amounts of sequence. Assembly corresponds to the computational process of placing individual reads into longer pieces of contiguous sequences, known as contigs, based on sequence overlaps and paired read information. Taxonomic binning sorts the contigs of a metagenome sample into 'bins' that represent the populations or higher-level clades of community members. Though both tasks are performed independently and evaluate different types of information, the problem of metagenome sequence assembly is closely related to taxonomic binning, as both allow the reconstruction of draft

Table 1: Throughput and read lengths of different sequencing technologies

Manufacturer and technology	Length (bp)	Throughput*	Normalized throughput** (Mb/h)	Throughput scale***	Time per run
Solexa/Illumina Sequencing by Synthesis	100	300 Gb/8.5 days	1500	10^4	8.5 days
	–150	– 600 Gb/11 days	–2300		–11 days
Life Technologies/Applied Biosystems SOLiD	50	7 Gb/day	300	10^3–10^4	2 days
	–75	–20 Gb/day	–800		–7 days
Life Technologies/Ion Torrent	100	10 Mb/2 h	5	10^1–10^3	2 h
	–200	–1 Gb/2 h	–500		
Roche/454 Pyrosequencing	550	450 Mb/10 h	30	10^2	10 h
	–1000	–700 Mb/23 h	–45		–23 h
Life Technologies Capillary Sanger sequencing	600	690 kb/day	0.029	10^0	~7 h [15]
	–900	–2100 kb/day	–0.088		

*Numbers are based on vendor information: Illumina Inc. (www.illumina.com), Life Technologies (www.lifetechnologies.com), Roche/454 (www.454.com). **Normalized throughput is scaled to a 1-h period and rounded. ***The throughput scale is compared with Life Technologies 3730 Sanger chemistry-based sequencer and shows the ratio of throughput values in terms of order of magnitude. Because lack of information on sequencing statistics or commercial availability, Pacific Biosciences (www.pacificbiosciences.com), Oxford Nanopore Technologies (www.nanoporetech.com) and Helicos Biosciences (www.helicosbio.com) are excluded.

genome sequences. The terms 'taxonomic' and 'phylogenetic' binning are both used in the literature, as modern taxonomies such as the NCBI taxonomy [22] or the ribosomal gene based RDP-II [23], GreenGenes [24] and ARB-SILVA [25] taxonomies are built upon phylogenetic principles. Even though it is less consistent, taxonomic binning software for shotgun metagenomics most frequently relies on the NCBI taxonomy, probably due to its widespread use in annotation of public sequence data.

Similar to the assembly of individual isolated genomes [26], assembly in metagenomics aims to recover long contiguous pieces of sequence from the sequence collection of reads that represent parts of the genomes of individual community members. Massively increased amounts of data, varying organism abundances within a sampled community, differing complexities in terms of the overall number of organisms contained and the presence of multiple closely related organisms all challenge the sequence assemblers that were originally designed for isolated genomes. To address these challenges, methods designed for assembly of microbial community NGS data [27–30] are being developed. Paired-end or mate-pair protocols, which add distance information between two individual reads, can greatly aid in the assembly process. Assembly information such as the ordering of contigs within a scaffold can also be used to check binning quality, and binning has been used to refine assembly in a feedback process. In recent studies, the joint analysis of assembly information and sequence composition allowed the reconstruction of several partial genomes by taxonomic binning [19, 20]. Thus, a closer integration of the two approaches appears promising for draft genome reconstruction from NGS metagenome data.

Following assembly and binning, further bioinformatic analyses include the prediction of genes, as well as functional annotation and reconstruction of potential pathways. For these steps, dedicated web servers exist, such as MG-RAST [31], IMG/M [32] and CAMERA [33]. Analysis of the gene content of individual bins allows inference of the functional and metabolic capabilities of individual community members, and allows a metagenome sample to be studied in its entirety. If read lengths or sequencing depth are insufficient for assembly, the functional analysis of a metagenome sample is restricted to what can be inferred without partial genome reconstructions for individual community members.

BINNING STRATEGIES

The term binning was originally coined for the problem of separating the sequence fragments of a metagenome according to the microbial populations they originate from [7, 34]. The definition has been extended to include bins that represent all fragments that originate from a common higher level clade, in cases where resolution down to individual populations is not possible. For placement of sequence fragments into taxonomic bins, attributes which are

indicative of the taxonomic origin of a fragment are evaluated. Different types of information can be used for this purpose: (a) local sequence similarity to sequences of known taxa (used in similarity-based taxonomic assignment), (b) similarity in sequence composition to sequences of a given taxon (used in composition-based taxonomic assignment) or to other sequences in the sample (used in composition-based clustering) or (c) similarity in read coverage and linkage information from assembly for contigs within a metagenome sample. The underlying rationale of using read coverage is that similar coverage of two contigs in the sample indicates similar abundance and therefore potentially the same underlying source population in the community.

How accurately fragments can be assigned to taxonomic bins depends on several factors. The first is fragment length. Shorter, noisier fragments cannot be assigned as accurately as longer fragments of 2 kb or more [35]. In particular, assignment of individual reads or of fragments less than 1 kb in length poses significant challenges. Reported assignment accuracies for 100 bp fragments to a clade at the genus level are 60% under somewhat idealized conditions, with only reference data from the same species being removed. This, however, means that 40% of fragments are misassigned [36]. Furthermore, accuracy drops to less than 30% if the reference data are depleted of sequences from the same genus, meaning 70% of 100 bp fragments are misassigned at the family level.

Another influential factor for binning accuracy is the community's complexity in terms of the number of distinct phylotypes it comprises. Metagenome sequencing of complex communities, such as those found in soil [11], results in lower sequencing coverage of most populations and therefore shorter contigs in assembly. This amounts to many short fragments, or even predominantly unassembled samples, which have to be separated into a multitude of taxonomic bins. The larger the number of bins, the harder the problem becomes, as the chances of randomly assigning a fragment correctly decrease with increasing numbers of bins. Finally, for taxonomic assignment, the availability of reference data from taxa that are closely related to the microbes of the sequenced community is important for accurate assignment. Similarity-based assignment of metagenome shotgun sequence data requires homologous reference sequences from related taxa to be available for a fragment to be assigned; ideally, entire sequenced genomes should be available. The sequencing of many isolate genomes of the human microbiome in the Human Microbiome Project has immensely helped similarity-based taxonomic assignment of human gut metagenome samples [43, 44]. A 'shallow' (i.e. to high-ranking clades only) taxonomic assignment of a sample based on sequence similarities indicates the presence of many taxa that are only distantly related to isolated sequenced genomes. If no sequenced genomes from related taxa are available, composition-based assignment can be used for higher resolution taxonomic binning. Clustering of metagenome fragments based on sequence composition does not require reference sequences and comparably small amounts of non-homologous reference sequences are required for composition-based taxonomic classification. Table 2 lists available web-based applications for phylotyping and taxonomic binning of metagenome samples.

Taxonomic binning based on sequence similarities

Similarity-based taxonomic assignment utilizes the local similarity of a query sequence to sequences of known taxonomic origin. Taxonomic identifiers are commonly assigned either by identifying the lowest common ancestor (LCA) from the taxonomy for the taxa of the most similar sequences found [35] or by using phylogenetic placement methods. Phylogenetic placement methods, such as pplacer [45], EPA/RaxML [46] and SEPP [47] place the query sequence within a fixed reference tree. The taxonomic label assigned then corresponds to the LCA of the taxa associated with the first ancestral node's children. Both methods are related to 'nearest neighbor' classification. In both cases, there has to be a search phase in which such similarities are identified. Typically, local similarities to sequence database entries are searched for with alignment programs such as BLAST [48]. Searches for gene family or protein domain motifs in the query sequence can be performed with a reference collection of profile Hidden Markov Models (HMMs). HMMER 3.0, released in 2010, has a 100-fold increase in speed compared with prior versions, with runtimes being competitive to blastp [49]. Screening a large metagenome sample with a collection of profile HMMs for marker genes is computationally much less demanding than a full search for similar regions in large sequence collections [49]. This is because the number of entries to be searched against is typically

Table 2: Overview of existing web applications for taxonomic assignment and phylotyping of metagenome sequence samples

Name	Phylotyping	Taxonomic assignment	Functional annotation	Techniques and web link
CAMERA [33] (v.2)	✓	—	✓	Reverse Psi-BLAST (*http:*//camera.calit2.net)
MetaABC [37]	—	✓	—	BLAST, PhymmBL, MEGAN, Sort-ITEMS (*http:*//bits2.iis.sinica.edu.tw/MetaABC/)
MG-RAST [31] (v.3.1.2)	✓	—	✓	BLAST/BLAT (*http:*//metagenomics.anl.gov)
MLTreeMap [38] (v.2.06.1)	✓	—	✓	BLAST, HMMER, RaxML (*http:*//mltreemap.org)
NBC [39] (v.1.1 CLI)	—	✓	—	Naïve Bayesian Classifier (*http:*//nbc.ece.drexel.edu)
PhyloPythia [40], PhyloPythiaS [35]	—	✓	—	(Structured) SVM (*http:*//cbcsrv.watson.ibm.com/phylopythia.html; *http:*//binning.bioinf.mpi-inf.mpg.de)
TaxSOM [41]	—	✓	—	Self-Organizing Maps (*http:*//soma.arb-silva.de)
WebCARMA [42] (v.3.0)	✓	✓	✓	BLAST, HMM search versus Pfam (*http:*//webcarma.cebitec.uni-bielefeld.de)

Phylotyping methods assign only a subset of contigs based on taxonomic marker genes.

several orders of magnitude lower. HMMs are popular in combination with phylogenetic placement approaches, as the required multiple alignment of a query sequence to the homologs can be directly deduced from the state path of the sequence through the HMM and the multiple alignment used in its construction. However, known marker genes or protein families from reference collections such as PFAM only cover a small part of the genes found across diverse environments. Therefore, most HMM-based approaches [38, 50, 51] may be seen as phylotypers of metagenome samples, rather than binning methods, as they indicate the taxonomic composition of the sample based on placement of a fraction of the fragments, rather than assigning the entire sample.

Searching for similar sequences in large sequence collections results in a higher fragment coverage with hits than when profile HMMs are used. Analysis of a metagenome sequence sample therefore comes with high computational costs, beyond what a typical desktop computer is capable of. When using a similarity search, one is therefore confronted with the question of which reference sequences to compare with. The choice depends on the available time and computational resources. Databases that are often searched are NCBI RefSeq, a non-redundant nucleotide and protein collection for medical, functional and diversity studies; NCBI whole genomes; NCBI nt, a large nucleotide collection; and NCBI nr, a large non-redundant protein collection [22]. Software such as MEGAN [52] allows the output of BLAST to be interpreted for the taxonomic and functional characterization of metagenome samples based on sequence similarity. If sequenced genomes of related species to the sampled taxa exist, recruitment analysis has been used [43]. Here, each read is compared with a set of genome sequences and 'recruited' to the most similar genome, allowing the identification of reads of the prevalent species that are closely related to a sequenced reference collection, if performed with stringent alignment cut-offs [53].

Case study 1

Recruitment analysis. In [54], Illumina and Roche/454 sequencing were jointly used to generate 860 Mb of non-human sequence data from a microbial community of human dental plaque. All obtained reads were aligned against 50 available reference genomes for human oral microbes from the Human Microbiome Project using Mummer, resulting in recruitment of 4% of all reads with more than 97% sequence identity to one of the reference genomes. This indicates that most of the sampled microbes originate from species that are too distantly related to the sequenced reference collection for similarity-based recruitment.

Taxonomic binning based on sequence composition

The composition-based approach to taxonomic binning is to utilize the taxonomic signal contained in fragment-wide GC content, codon usage or the use of short oligomers (kmers), typically 4–6 bp long. The observation that such properties tend to vary more across the genomes of different species than within a given one gave rise to the term genome signatures [54, 55]. Such signatures can also be inferred for higher-level clades, allowing their use

for taxonomic fragment assignment across various ranks [40].

Taxonomic binning based on sequence composition can be performed with supervised or unsupervised methods. The choice of which to use depends on the availability of suitable reference data. Unsupervised methods group fragments with similar composition profiles into clusters, corresponding to individual taxonomic bins. Inference of the taxonomic label for a bin can be performed based on taxonomic assignment of marker genes found in the fragments of a bin. To infer the clustering of fragments, existing methods use, for example, a graph-cut algorithm or variations of a self-organizing map algorithm [56, 57]. A sample can also be binned with supervised methods, which assign fragments to clades using a model trained with available reference sequences. Supervised methods tend to have higher accuracy than unsupervised methods for taxonomic assignment and are more easily applied to complex microbial mixtures with skewed organism abundances. However, they require sufficient amounts of reference sequences to be identified for the sample populations or higher-level clades which are to be included in the model. In practice, therefore, each approach has its own appeal and both are being applied. Methods used for supervised classification are, for example, (structural) Support Vector Machines (SVMs) [40], the naive Bayes classifier [39], a k-nearest neighbor classifier [58] and Interpolated Markov Models [36]. As composition-based signatures are a global attribute of sequences, no entire reference genomes are required, but only sufficient amounts of sequences for inference of a composition-based signature. For SVM-based classification, this has been found to be ~100 kb per clade [35]. Reference sequences can be identified among publicly available genomes or by taxonomic assignment of conserved marker-genes of the sample contigs, which allows the respective contigs to be used as training material. If necessary, fosmids carrying marker genes can be sequenced to generate training material for interesting sample populations or higher level clades [10, 59, 60].

Case study 2

Taxonomic binning by composition-based taxonomic assignment. In [59], a microbial gut community from the Australian Tammar wallaby was studied by Sanger and 454 sequencing of metagenome plasmid and fosmid libraries. This microbial community is involved in the breakdown of plant biomass consumed by the host animal. Using 16S rRNA analysis, 236 distinct phylotypes were observed. Of the 16S rRNA sequences, 9% originated from a novel species, Wallaby group 1 (WG-1), in the family of Succinivibrionaceae. PhyloPythia, a composition-based taxonomic classifier, was used to train a model including the WG-1 and other relevant clades for species present in the community. Composition-based taxonomic assignment of the metagenome sample recovered a 2 Mb draft genome for WG-1. Metabolic reconstruction based on the draft genome allowed the cultivation requirements for WG-1 to be deduced, leading to isolation, characterization and a draft genome sequence for the previously unknown species. It also resulted in the finding that WG-1 contributes to the low-methane emission phenotype of plant biomass degradation in the Tammar wallaby. The draft genome sequences from the isolate culture showed 98.9% sequence identity to the WG-1 metagenome bin, and 90% of shared reads and assemblies, indicating accurate reconstruction of the draft genome from the metagenome sample by composition-based taxonomic binning.

Hybrid methods

Several methods combine different types of information to improve predictive accuracy [19, 20, 36, 52]. For instance, read coverage is combined with an analysis of kmer frequencies in clustering of fragments [19, 34]. Searches for similar sequences and analysis of linkage information from an assembly are also combined with composition-based taxonomic assignment, if the computational burden can be borne. This has particular advantages for short fragment analysis. Kmer signatures for fragments below 1 kb in length, particularly those of individual reads, are noisy, even more so than taxonomic conservation of sequence similarities [35].

Case study 3

Taxonomic binning based on clustering by sequence composition and read coverage. In one of the most in-depth metagenome studies of a particular environment undertaken so far, 286 Gb of paired-end Illumina sequence reads were generated from a sample of the plant-fiber adherent microbiome from a cow rumen [19]. Rarefaction analysis of 16S rRNA indicated the presence of ~1000 distinct OTUs. Clustering of assembled contigs by agglomerative hierarchical

clustering, based on tetramer frequencies and read coverage, resulted in the formation of 466 taxonomic bins. Fifteen of these were estimated to represent largely complete genomes (between 60% and 92%), based on their association with fully sequenced genomes from their respective clades. This estimate was based on the presence of a minimal set of core genes found in all sequenced genomes from the respective phylogenetic order.

Case study 4

Taxonomic binning based on assembly information and sequence composition [20]. SOLID sequencing of two marine samples generated 58.5 Gb of mate-paired reads of 50 bps in length. The number of phylotypes observed with16S rRNA analysis was not specified in detail; however, family-level taxonomic groups were observed with abundances of less than 10%. From the metagenome data, 300 Mb of contigs were assembled. Scaffolds—linked sets of contigs assumed to originate from one genome—were generated by splitting the assembly graph, which links contigs based on mate-pair information, according to mate-pair linkage scores, read coverage and tetranucleotide usage. Scaffold clustering by tetranucleotide usage generated 14 partial genome reconstructions from the two samples, for populations ranging in abundance from 4 to 10% each in one of the samples. Reassembly of 11 mate-pair connected scaffolds that are binned together based on similar tetranucleotide statistics and manual gap closure allowed the recovery of a closed circular 2 Mb genome from an uncultured group, the marine group II Euryarchaeota.

ADVANTAGES AND DISADVANTAGES OF DIFFERENT BINNING APPROACHES

Which binning methodology to use depends on multiple factors, such as the complexity of the analyzed microbial community, available reference sequences and computing resources. For taxonomic assignment of arbitrary sequence fragments to a particular species based on sequence similarity, completely sequenced reference genomes of closely related taxa are ideally required, which are often not available. If no reference data exist for the species of the metagenome sample, homology-based taxonomic assignment to higher level clades is more accurate than composition-based taxonomic assignment for short fragments of 1 kb or less [35]. This length corresponds to individual reads with most sequencing technologies. The assignment of individual reads in general is, however, notably less accurate than assignment of longer fragments.

The runtime of sequence similarity searches increases proportional to the product of the metagenome sample size (number and length of contigs) and the size of the reference sequence collection. This makes it a computationally very demanding task for NGS data sets. The required computing resources are not available in many experimental laboratories. If researchers are willing to submit their data to external facilities, data can be processed by web servers such as MG-RAST, IMG-M or CAMERA, which offer their computational resources to the community.

The choice of whether to cluster or classify based on sequence composition depends on availability of some reference data to train a composition-based classifier. Classification is likely to be more accurate than clustering in taxonomic assignment. However, if no reference data is available, clustering will allow resolution of taxonomic bins which otherwise would go undetected. If multiple types of information are included into the binning process, like it is done in hybrid approaches, this is likely to increase the overall amount and accuracy of assignments. Composition-based taxonomic assignment requires less reference sequences than homology-based assignment. This is because sequence composition is a globally conserved property, while sequence similarity depends on local sequence conservation between a query and target. Training times of a composition-based taxonomic classifier depend on the method used, but it requires typically considerably less time than searching a reference sequence collection. Once a composition-based model for taxonomic classification has been trained, execution times for classification again typically scale linearly with the metagenome sample size and are independent of a reference sequence collection. For composition-based clustering, no training phase is needed. The runtime of clustering typically scales at least quadratically with the sample size, as it often involves pairwise comparisons.

FUTURE DIRECTIONS

The recent developments in sequencing technologies have considerably pushed the boundaries in terms of

what can be learned from metagenome sequence samples. The high sequencing depth of microbial communities, in combination with the application of sophisticated algorithms, has allowed the retrieval of near-complete draft genomes from the metagenomes of many microbial communities, including highly complex ones, such as those found in soil [11]. However, the size and heterogeneity of the different data types produced by the various novel techniques have created new challenges, which remain to be addressed. A prominent one is how to further reduce the computational requirements of searching for local similarities between giga- and even terabase-sized sequence samples and equivalently large reference sequence collections. Second, it remains to be explored how taxonomic assignment accuracy can be further improved for the vast majority of microbial community members that are only distantly related to sequenced isolate genomes. Because of the value of available sequences from related taxa for the taxonomic binning of a particular sample, efforts such as GEBA might help in this regard [61]. The GEBA project aims to construct a 'Genomic Encyclopedia for Bacteria and Archaea' by strategic sequencing of microbial genomes from all major and minor taxonomic groups. As the cost of sequencing has decreased, partial genome reconstruction by single-cell genome sequencing is an attractive option for obtaining reference sequences for taxonomic binning and draft genome reconstruction [62] from metagenomes. Here, an individual cell from a microbial population within a community is isolated using techniques such as optical tweezers, fluorescence-assisted cell sorting and others, and is then lysed and its genome sequence amplified with multiple displacement amplification prior to random shotgun sequencing.

Advances in single-molecule sequencing technologies now allow longer reads to be generated than what was possible using traditional Sanger sequencing. Even though this promises to resolve several issues associated with short read analysis, such as high error rates in binning, assembly and functional annotation, the larger sequencing error of some of these technologies, currently estimated to be ~15%, presents a different substantial hurdle. Therefore, assessing technology-specific errors and developing technology-specific denoising procedures, such as have been developed for 454 amplicon data [63], will be prerequisite to leveraging the value of these techniques for metagenome research.

An interesting research direction is to investigate whether composition-based binning is applicable for the analysis of samples with both microbial and viral content. Composition-based taxonomic binning has been successfully applied for the analysis of viral metagenome samples; however, bacteriophage codon usage to some extent reflects properties of the host [64, 65]. Therefore, classification accuracy and level of taxonomic resolution attainable for viral taxa will have to be investigated in more detail.

Key Points

- NGS technologies generate massive amounts of sequencing data allowing the in-depth analysis of microbial communities.
- Taxonomic binning has allowed draft genomes of microbial species from many environments to be reconstructed, and the cultivation requirements of a novel uncultured species to be deduced.
- To further advance draft genome reconstruction from metagenome samples, the existing techniques could be further refined by integrating multiple sources of information and by appropriately denoising the data under consideration to remove technology-specific sequencing errors.

Acknowledgements
The authors thank Chris Quince and Alex Scyrba for providing comments.

FUNDING
A.C.M and J.D. gratefully acknowledge funding by the German Max Planck society and Heinrich-Heine University Düsseldorf.

References

1. Simon C, Daniel R. Metagenomic analyses: past and future trends. *Appl Environ Microbiol* 2011;**77**:1153–61.
2. Kunin V, Copeland A, Lapidus A, *et al*. A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev* 2008;**72**: 557–78.
3. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 1995;**59**:143–69.
4. Hugenholtz P. Exploring prokaryotic diversity in the genomic era. *Genome Biol* 2002;**3**:1–8.
5. Venter JC, Remington K, Heidelberg JF, *et al*. Environmental genome shotgun sequencing of the Sargasso Sea. *Science (New York, NY)* 2004;**304**:66–74.
6. Tringe SG, von Mering C, Kobayashi A, *et al*. Comparative metagenomics of microbial communities. *Science (New York, NY)* 2005;**308**:554–7.
7. Woyke T, Teeling H, Ivanova NN, *et al*. Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* 2006;**443**:950–5.

8. Suen G, Scott JJ, Aylward FO, *et al*. An insect herbivore microbiome with high plant biomass-degrading capacity. *PLoS Genet* 2010;**6**:e1001129.
9. Turnbaugh PJ, Quince C, Faith JJ, *et al*. Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proc Natl Acad Sci U S A* 2010;**107**:7503–8.
10. Warnecke F, Luginbühl P, Ivanova N, *et al*. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* 2007;**450**:560–5.
11. Mackelprang R, Waldrop MP, DeAngelis KM, *et al*. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 2011;**480**:368–71.
12. Huse SM, Dethlefsen L, Huber JA, *et al*. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet* 2008;**4**:e1000255.
13. Jeewon R, Hyde KD. Detection and diversity of fungi from environmental samples: traditional versus molecular approaches. *Adv Tech Soil Microbiol* 2007;**11**:1–15.
14. Fuhrman JA. Microbial community structure and its functional implications. *Nature* 2009;**459**:193–9.
15. Mardis ER. The impact of next-generation sequencing technology on genetics. *Trends Genet: TIG* 2008;**24**:133–41.
16. Metzker ML. Sequencing technologies—the next generation. *Nature Rev Genet* 2009;**11**:31–46.
17. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet* 2010;**19**: R227–40.
18. Thompson JF, Milos PM. The properties and applications of single-molecule DNA sequencing. *Genome Biol* 2011;**12**:217.
19. Hess M, Sczyrba A, Egan R, *et al*. Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science (New York, NY)* 2011;**331**:463–7.
20. Iverson V, Morris RM, Frazar CD, *et al*. Untangling genomes from metagenomes: revealing an uncultured class of marine Euryarchaeota. *Science* 2012;**335**:587–90.
21. Wilkening J, Wilke A, Desai N, Meyer DF. Using clouds for metagenomics: A case study. *IEEE Cluster 2009*. New Orleans, LA, USA: IEEE;1–6.
22. Sayers EW, Barrett T, Benson D, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2009;**37**:D5–15.
23. Cole JR, Wang Q, Cardenas E, *et al*. The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* 2009;**37**:D141–5.
24. 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**: 5069–72.
25. 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**:7188–96.
26. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics* 2010;**95**:315–27.
27. Pell J, Hintze A, Canino-Koning R, *et al*. Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. Arxiv Preprint arXiv: 1112.4193 I:1–11.
28. Koren S, Treangen TJ, Pop M. Bambus 2: scaffolding metagenomes. *Bioinformatics* 2011;**27**:2964–71.
29. Peng Y, Leung HCM, Yiu SM, *et al*. Meta-IDBA: a de novo assembler for metagenomic data. *Bioinformatics (Oxford, England)* 2011;**27**:i94–i101.
30. Laserson J, Jojic V, Koller D. Genovo: de novo assembly for metagenomes. *J Comput Biol* 2011;**18**:429–43.
31. 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.
32. Markowitz VM, Chen I-M, Chu K, *et al*. IMG/M: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res* 2012;**40**(Database issue): D123–9.
33. Sun S, Chen J, Li W, *et al*. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. *Nucleic Acids Res* 2011;**39**:D546–51.
34. Tyson GW, Chapman J, Hugenholtz P, *et al*. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 2004;**428**: 37–43.
35. Patil KR, Haider P, Pope PB, *et al*. Taxonomic metagenome sequence assignment with structured output models. *Nat Methods* 2011;**8**:191–2.
36. Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* 2009;**6**:673–6.
37. Su C-H, Hsu M-T, Wang T-Y, *et al*. MetaABC—an integrated metagenomics platform for data adjustment, binning and clustering. *Bioinformatics (Oxford, England)* 2011;**27**: 2298–9.
38. Stark M, Berger S, Stamatakis A, *et al*. MLTreeMap—accurate maximum likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics* 2010;**11**:461.
39. Rosen GL, Reichenberger ER, Rosenfeld AM. NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics (Oxford, England)* 2011;**27**:127–9.
40. McHardy AC, Martín HG, Tsirigos A, *et al*. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* 2007;**4**:63–72.
41. Weber M, Teeling H, Huang S, *et al*. Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. *ISME J* 2011;**5**: 918–28.
42. Gerlach W, Jünemann S, Tille F, *et al*. WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. *BMC Bioinformatics* 2009; **10**:430.
43. Qin J, Li R, Raes J, *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;**464**:59–65.
44. Nelson KE, Weinstock GM, Highlander SK, *et al*. A catalog of reference genomes from the human microbiome. *Science (New York, NY)* 2010;**328**:994–9.
45. Matsen FA, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* 2010;**11**:538.
46. Berger SA, Krompass D, Stamatakis A. Performance, accuracy, and web server for evolutionary placement of short

sequence reads under maximum likelihood. *Systematic Biol* 2011;**60**:291–302.

47. Mirarab S, Nguyen N, Warnow T. SEPP: SATé-Enabled Phylogenetic Placement. *Pacific Symp Biocomput* 2012; 247–58.
48. Camacho C, Coulouris G, Avagyan V, *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.
49. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011;**39(Suppl 2)**:W29–37.
50. Gerlach W, Stoye J. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res* 2011;1–11.
51. Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics (Oxford, England)* 2012;1–2.
52. Huson DH, Mitra S, Ruscheweyh H-J, *et al*. Integrative analysis of environmental sequences using MEGAN4. *Genome Res* 2011;**21**:1552–60.
53. Xie G, Chain PSG, Lo C-C, *et al*. Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. *Mol Oral Microbiol* 2010;**25**: 391–405.
54. Karlin S, Burge C. Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* 1995;**11**:283–90.
55. Deschavanne PJ, Giron A, Vilain J, *et al*. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* 1999;**16**: 1391–9.
56. Chatterji S, Yamazaki I, Bai Z, Eisen J, Vingron M, Wong L. CompostBin: a DNA composition-based algorithm for binning environmental shotgun reads. *Research in Computational Molecular Biology*, Vol. 4955. Berlin/ Heidelberg: Springer, 2008, 17–28.
57. Teeling H, Waldmann J, Lombardot T, *et al*. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics* 2004;**5**:163.
58. Diaz NN, Krause L, Goesmann A, *et al*. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics* 2009;**10**:56.
59. Pope PB, Smith W, Denman SE, *et al*. Isolation of Succinivibrionaceae implicated in low methane emissions from Tammar wallabies. *Science (New York, NY)* 2011;**333**: 646–8.
60. Pope PB, Denman SE, Jones M, *et al*. Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase profiles different from other herbivores. *Proc Natl Acad Sci U S A* 2010;**107**:14793–8.
61. Wu D, Hugenholtz P, Mavromatis K, *et al*. A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature* 2009;**462**:1056–60.
62. Woyke T, Tighe D, Mavromatis K, *et al*. One bacterial cell, one complete genome. *PloS One* 2010;**5**:e10314.
63. Quince C, Lanzén A, Curtis TP, *et al*. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* 2009;**6**:639–41.
64. Pride DT, Schoenfeld T. Genome signature analysis of thermal virus metagenomes reveals archaea and thermophilic signatures. *BMC Genomics* 2008;**9**:420.
65. Lucks JB, Nelson DR, Kudla GR, *et al*. Genome landscapes and bacteriophage codon usage. *PLoS Comput Biol* 2008;**4**: e1000001.