BRIEFINGS IN BIOINFORMATICS. VOL 13. NO 6. 669–681
Advance Access published on 8 September 2012
doi:10.1093/bib/bbs054

Classification of metagenomic sequences: methods and challenges

Sharmila S. Mande, Monzoorul Haque Mohammed and Tarini Shankar Ghosh

Submitted: Ist March 2012; Received (in revised form): 24th July 2012

Abstract

Characterizing the taxonomic diversity of microbial communities is one of the primary objectives of metagenomic studies. Taxonomic analysis of microbial communities, a process referred to as binning, is challenging for the following reasons. Primarily, query sequences originating from the genomes of most microbes in an environmental sample lack taxonomically related sequences in existing reference databases. This absence of a taxonomic context makes binning a very challenging task. Limitations of current sequencing platforms, with respect to short read lengths and sequencing errors/artifacts, are also key factors that determine the overall binning efficiency. Furthermore, the sheer volume of metagenomic datasets also demands highly efficient algorithms that can operate within reasonable requirements of compute power. This review discusses the premise, methodologies, advantages, limitations and challenges of various methods available for binning of metagenomic datasets obtained using the shotgun sequencing approach. Various parameters as well as strategies used for evaluating binning efficiency are then reviewed.

Keywords: *binning algorithms; metagenomics; taxonomic classification; lowest common ancestor; oligo-nucleotide composition; taxonomic diversity*

INTRODUCTION

A majority of microbes residing in diverse environments cannot be cultured in the laboratory [1]. Consequently, traditional genomics-based approaches, requiring prior cloning and culturing of individual microbes, cannot be used to study entire microbial communities residing in any given environment. The advent of the 'metagenomics' approach has enabled researchers to circumvent this limitation by facilitating direct extraction, sequencing and analysis of specific phylogenetic marker genes or the entire genomic content of microbial communities. These microbial communities (microbiomes) can display a wide degree of spatial/temporal variations in their taxonomic composition [2]. Consequently, a key preliminary step in metagenomic analysis is to decipher the microbial community structure of the given environment by categorizing various microbes residing therein and quantifying their diversity in terms of species richness/abundance. In the context of microbial communities, the term 'species' refers to a fundamental and distinct rank of taxonomic hierarchy. Organisms are grouped at the rank of 'species' primarily on the basis of their overall genotypic and/or morphological similarity. However, the criteria adopted by researchers for grouping individuals into the same species are currently not universal and are generally observed to be context dependent [3]. Species richness, a frequently employed diversity metric, refers to the number of distinct species (within a given unit area) inhabiting a particular biological community, habitat, or ecosystem type [4]. In contrast, species abundance incorporates calculations with respect to species evenness and/or dominance,

Corresponding author. Sharmila S. Mande. TCS Innovation Labs, Tata Consultancy Services Ltd., 54-B Hadapsar Industrial Estate, Pune 411013, Maharashtra, India. Tel: +91 020 6608 6432. Fax: 020 6608 6399. E-mail: sharmila@atc.tcs.com, sharmila.mande@tcs.com

Sharmila S. Mande heads the Bio-Sciences Division, TCS Innovation Labs, India. Her research interests include metagenomics, comparative genomics, algorithm development, mathematical modeling of biological systems and structural biology.

Monzoorul Haque Mohammed works as a scientist in Bio-Sciences Division, TCS Innovation Labs, India. His research interest includes development of algorithms for analyzing metagenomic datasets and compression of biological data.

Tarini Shankar Ghosh is a scientist in Bio-Sciences Division, TCS Innovation Labs, India. His research interests include algorithm development, metagenome analysis and understanding mechanisms of bacterial pathogenesis.

i.e. the pattern of the relative abundances of species in a given environment. Values of relative abundance in turn indicate the quantitative pattern of rarity and commonness among species in a sample or a community [4]. Obtaining such insights into the microbial diversity helps in identifying and associating specific organisms or taxonomic groups (and the genes/proteins they encompass) with various phenotypic/functional traits characterizing a given environment.

Two approaches are generally adopted for characterizing taxonomic diversity of metagenomes. In the first approach, referred to as the shotgun sequencing based approach, genomic fragments originating from genomes of organisms constituting a microbiome are extracted and sequenced [2]. The sequencing step typically generates millions of sequences. These sequences (genomic fragments), also referred to as 'reads', can be considered to represent the compositional properties of their source genomes. Analyzing these reads can thus provide insights into the composition of various microbes constituting a microbiome. The second approach focuses on the isolation, extraction and sequencing of amplicons corresponding to entire (or specific portions of) phylogenetic marker genes (e.g. 16S rRNA, rpoB, etc., in the case of prokaryotic organisms) or specific genomic regions such as the Internal Transcribed Spacer (ITS) regions (for fungal species) [5, 6]. Various structural properties of these genes and genomic regions enable their use as 'species-specific taxonomic barcodes' that can be employed for obtaining quick estimates of taxonomic diversity [6].

In both approaches (described above), the subsequent step of analyzing the sequenced data in order to get the taxonomic diversity profile of an environmental sample is referred to as 'binning'. Binning, a process conceptually similar/analogous to established machine learning techniques, involves classifying and/or clustering reads into specific bins. Given the two approaches of characterizing taxonomic diversity, binning methods can be classified into two groups, namely shotgun sequencing based and amplicon based. Based on their methodologies and final objectives, binning methods can be further categorized as 'taxonomy dependent' and 'taxonomy independent'. Methods belonging to the former category follow 'supervised learning procedures', wherein, individual reads are taxonomically classified by comparing them to sequences/models (of known phylogenetic origin) present in reference databases. Reads classified under similar taxonomic categories are finally grouped into bins. However, assignments of individual reads by taxonomy-dependent methods are subject to obtaining sufficient levels of similarity, between reads and sequences/models in reference databases. In a typical metagenomics scenario, a majority of reads originate from genomes of hitherto unknown organisms. In other words, sequences belonging to the source genomes of these reads are absent in existing reference databases. Such reads, lacking a 'genomic reference', typically fail to exceed the predetermined similarity threshold criteria, and consequently cannot be mapped to the known 'taxonomic reference' tree. Existing taxonomy-dependent binning methods generally categorize such reads as unassigned. Therefore, the overall objective (and applicability) of taxonomy-dependent methods is to obtain estimates of the profile/abundance of 'known' taxonomic groups in a given environmental sample. In contrast, taxonomy-independent methods simply group/bin reads in a given dataset based on their mutual similarity and do not involve a database comparison step. The methodology followed by taxonomy-independent methods is therefore similar to 'unsupervised' machine learning procedures.

The present review first summarizes the premise, methodologies, advantages and limitations of existing binning methods. We also discuss various aspects with respect to (i) strategies employed for evaluating binning methods, (ii) parameters used for evaluating binning efficiency and (iii) existing challenges. This review lays specific emphasis on binning methodologies designed for analyzing metagenomic datasets obtained using the shotgun sequencing approach. Since two recent reviews [7,8] provide a detailed description and performance evaluation of various methods available for binning 16S datasets (obtained using the amplicon-based sequencing approach), these methods have not been covered in this review.

BINNING ALGORITHMS FOR DATASETS OBTAINED USING SHOTGUN SEQUENCING

Analyses of datasets obtained using shotgun sequencing involve characterizing the taxonomic and functional diversity of a given environment by analyzing DNA fragments originating from the genomes of resident microbes. Existing binning methods for such datasets (summarized in Figure 1) can be classified into two categories, namely taxonomy dependent and taxonomy independent.

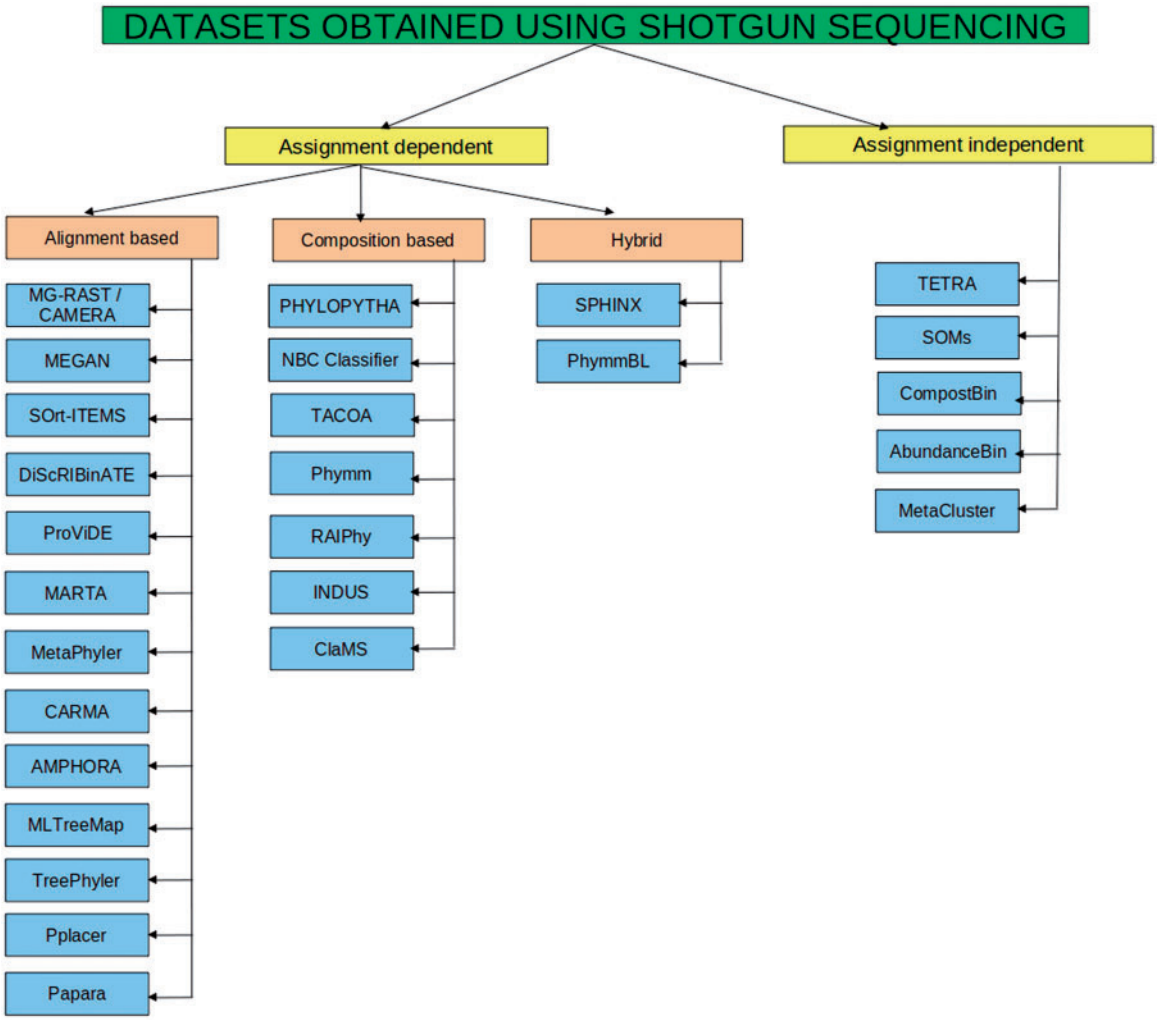


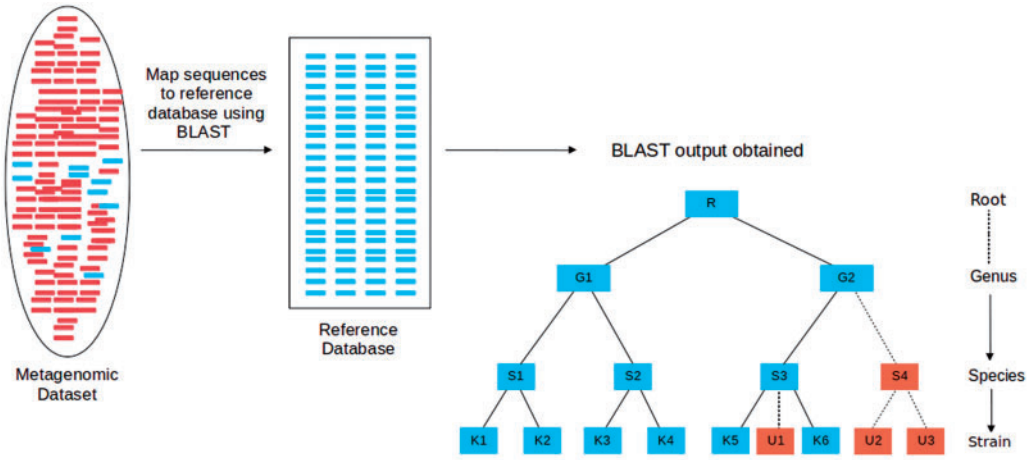
Figure 1: A schematic representation of various categories of algorithms available for binning metagenomic datasets obtained using shotgun sequencing.

Taxonomy-dependent methods

A majority of methods available for binning datasets obtained using shotgun sequencing belong to the taxonomy-dependent category. In these methods, the extent of 'similarity' of reads with sequences (in reference databases) or pre-computed models (built using sequences in reference databases) drives the assignment process. Reads failing to exceed pre-determined similarity thresholds are categorized as 'unassigned'. Based on the strategy used for comparing reads with sequences/pre-computed models, taxonomy-dependent methods can be sub-classified into alignment-based, composition-based and hybrid methods.

Alignment-based methods

A majority of these methods work by aligning reads to sequences or Hidden Markov Models (HMMs) corresponding to known taxonomic groups. Alignment-based methods typically employ algorithms like BLAST [9], BLAT [10], or read-mapping methods like BWA [11], BOWTIE [12] to first align individual reads to nucleotide/protein sequences belonging to known and characterized genomes. Collections of such reference sequences are present in major public repositories such as NCBI (ftp://ftp.ncbi.nih.gov/blast/db/), PFAM (http://pfam.sanger.ac.uk/), UniProt (http://www.uniprot.org/), EMBL (http://www.ebi.ac.uk/embl/), NCBI Genbank (http://www.ncbi.nlm.nih.gov/genbank/), NCBI Refseq (http://www.ncbi.nlm.nih.gov/RefSeq/), DDBJ (http://www.ddbj.nig.ac.jp/) and Ensembl (http://www.ensembl.org/). Reads are finally assigned to different taxonomic groups by analyzing the quality of their alignments with various hit sequences. This approach in its



Reads originate from	Significant BLAST Hits	Assingment Strategies	
		Best BLAST Hit Approach	LCA
K1	K1, K2, K3	K1 (√)	G1 (√)
U1	K5, K6	K5 (X)	S3 (√)
U2 and U3	K5, K6	K5 (X)	S3 (X)

Figure 2: Accuracy of taxonomic assignments using the best BLAST hit and the LCA approach in two different database scenarios. **Scenario I:** Read originating from the genome of known strains (KI), sequences of which are present in the reference database. In most cases, the best hit for this read will correspond to KI. A few significant hits may also be obtained with related organisms K2 and K3. Assignments using both approaches (best BLAST approach and the LCA approach) are generally observed to be correct (indicated using a tick sign). **Scenario 2:** Reads originating from genomes of new/unknown strains (UI, U2 and U3), sequences of which are absent in the reference database. Adopting the best blast approach in this scenario results in wrong assignments (indicated by a X sign). The success of the LCA approach is observed to be dependent on the extent of representation of organisms related to the source organism of the reads.

simplest form is adopted by the MG-RAST server [13] and the CAMERA pipeline [14], wherein reads are assigned to taxa of the organisms corresponding to their respective best BLAST hits. However, a primary limitation of BLAST-based approaches is the requirement of huge compute power for aligning millions of reads against huge number of sequences constituting reference databases. Moreover, a large proportion of reads in datasets obtained using the shotgun sequencing approach typically originate from hitherto unknown taxa belonging to either an entirely new species or genus or family or order or class or even a new phylum. It is incorrect to assign such reads to the organism corresponding to the best BLAST hit (Figure 2). To address this, the MG-RAST server also provides a Lowest Common Ancestor (LCA)-based option to infer taxonomic affiliation. In the LCA approach, a read is assigned to the lowest common taxonomic ancestor of the organisms corresponding to the set of significant hits (Figure 2). The LCA approach also forms the premise of the popular standalone binning software MEGAN [15].

For reads originating from hitherto unknown genomes, taxonomic affiliations using the LCA approach, although obtained at higher taxonomic levels, are expected to be more accurate as compared to that obtained using the best BLAST hit approach (Figure 2). However, the most critical step of this work-flow pertains to identifying the set of 'significant' hits which can be provided as inputs to the LCA procedure. MEGAN utilizes bit-score (of individual hits) as the sole parameter for judging significance. However, studies have indicated that this single-parameter approach adversely affects the specificity/accuracy of taxonomic assignments in different scenarios [16, 17]. Approaches like SOrt-ITEMS [16], DiScRIBinATE [17], ProViDE [18],

MetaPhyler [19] and MARTA [20] have addressed this limitation by utilizing, apart from bit-scores, pre-computed thresholds of other alignment parameters like the numbers/percentages of identities, positives and gaps to judge the quality of alignments. For each read, an appropriate level of taxonomic assignment is identified based on the observed alignment quality. The final assignment of a given read is made to a taxon that lies at or above this identified level. Though, the overall premise appears somewhat similar, these five methods differ with respect to the reference databases against which the similarity searches are performed. While SOrt-ITEMS, DiScRIBinATE employ the nr database, similarity searches of MARTA are done either against the nt database or against custom collections of genome sequences. In contrast, MetaPhyler performs a similarity search of reads against a customized database that contains sequences belonging to 31 phylogenetic marker gene families. MetaPhyler, in its work-flow, employs alignment parameter thresholds that are separately pre-computed for individual gene families. The idea is to capture variation patterns specific for each gene family, rather than using universal thresholds. It is worthwhile to note that ProViDE [18], a method customized for binning reads in viral metagenomic datasets, uses pre-computed thresholds of alignment parameters that are empirically determined by specifically analyzing the patterns of sequence divergence within the viral kingdom.

CARMA [21] and AMPHORA [22] are two well known methods which adopt HMM-based binning approaches. CARMA first compares reads (using BLASTx) against protein sequences in the PFAM database. Subsequently, it generates a phylogenetic tree by comparing each read to the HMM(s) of the protein families having significant hit(s). The final taxonomy is inferred based on the placement of the read in the constructed phylogenetic tree. In contrast, AMPHORA first compares reads against HMMs pre-built using sequences belonging to 31 phylogenetic marker gene families. Subsequently, a phylogenetic tree is constructed between the read and the sequences belonging to the best scoring HMM. The final taxonomic assignments are then obtained in a manner similar to CARMA. However, AMPHORA incorporates an additional bootstrapping step to improve the confidence of the final assignments. Other methods that utilize HMMs or reference trees in their assignment process are MLTreeMap [23], Treephyler [24], pplacer [25] and papara [26]. These methods additionally utilize either Maximum-Likelihood estimates or Bayesian-based strategies to compute confidence scores. While MLTreeMap compares query sequences against HMMs built using protein sequences of 40 marker gene families, Treephyler employs the PFAM database in its work-flow. In contrast, pplacer and papara provide a generalized algorithmic frame-work which can be utilized for placing reads onto the best scoring insertion edge on a user-specified reference phylogenetic tree. The pplacer method is also employed as a core component in the assignment work-flow of AMPHORA.

Composition-based methods

Methods in this category utilize compositional properties like GC percentage, codon usage and oligonucleotide usage patterns for first comparing reads to sequences or models present in reference databases. Final taxonomic assignments are based on the extent of compositional similarity in relative and/or absolute terms. Being alignment free, these methods are faster and require lesser compute power as compared to alignment-based methods. However, in order to generate a robust compositional signal, having taxonomic discrimination capability, the methods require query sequences of sufficient length.

Composition-based methodologies differ with respect to the way they represent, quantify and compare compositional properties. Most methods involve an initial training step during which one or more compositional properties of known genomes are used for building 'genome-specific' reference models or classifiers. For instance, Phylopythia [27] and the NBC classifier [28] build genome or clade-specific classifiers using Support Vector Machines (SVMs) and Naive-Bayesian approaches, respectively, in order to capture and represent oligonucleotide usage patterns observed in known taxonomic clades. In contrast, TACOA [29] first builds genome-specific models by analyzing tetra and penta-nucleotide usage patterns. A kernelized-Nearest Neighbor (k-NN) approach is subsequently employed to decipher the taxonomic assignments of individual reads. Another method, namely Phymm [30] represents oligonucleotide usage patterns of reference genomes as Interpolated Markov Models (IMMs). Reads are scored against these models and a Bayesian approach is subsequently employed for drawing taxonomic inferences. Markovian properties are also used by another method, namely

ClaMS [31]. This method generates signatures/models of training sequences using de Bruijn graphs and Markovian chain properties. During the classification phase, a similar procedure is used for generating and comparing signatures of query reads with pre-computed signatures of training sequences. RAIphy [32], a recently developed semi-supervised method, bases its classification on a index (referred to as the Relative Abundance Index) that indicates the over/under-abundance patterns of k-mers in sequences belonging to various known taxonomic clades. This index is subsequently used as a measure to associate a given taxon to a query sequence.

All methods described above assume that a single compositional model comprehensively represents the oligonucleotide usage patterns of a genome. However, certain genomes are known to be characterized by distinct regions of heterogeneity as compared to the rest of the genome [33]. The assumption of a 'one genome—one composition model' is thus not appropriate for such scenarios. The recently published INDUS algorithm [34] discards this assumption and represents each genome in the form of multiple vectors. Each vector captures the pattern of tetranucleotide frequencies of individual (non-overlapping) 1-Kb segments generated by dicing the respective genome. During the assignment process, INDUS utilizes the compositional distance between the query read and the 'closest' identified set of reference segments for determining an appropriate taxonomic level of assignment for the query. The final assignment is made to a consensus taxon that corresponds to the closest reference segments at or above the identified taxonomic level.

Hybrid methods

Binning methods under this category incorporate a combination of alignment and composition-based strategies for taxonomic classification. For instance, SPHINX [35] algorithm adopts a two-phase binning approach. The first phase compares the composition of a given read (represented as a tetra-nucleotide frequency vector) with those of reference sequences (in a pre-clustered format). The objective of this phase is to quickly identify a subset of clusters of reference sequences that are closest in composition to the given read. In the second phase, the taxonomic classification of the query read is inferred by first aligning the query read to reference sequences in the closest cluster and then employing a similarity-based approach like SOrt-ITEMS. While the first phase aids in reducing the search-space (and consequently binning time), the second phase ensures the accuracy/specificity of assignment. PhymmBL [30] is another hybrid method that combines the composition-based methodology of Phymm (described previously) along with an alignment-based step (BLAST) to improve the confidence of taxonomic assignments.

Taxonomy-independent methods

Methods under this category include TETRA [36], variants of SOMs [37, 38], CompostBin [39], AbundanceBin [40] and MetaCluster [41]. Among these, the simplest methodology is adopted by TETRA. For a given sequence dataset, TETRA computes the pairwise correlations between tetra-nucleotide usage patterns of all reads. This information is used for segregating reads (expected to originate from related taxonomic clades) into distinct bins. Self Organizing Maps (SOMs) are neural network-based approaches which involve clustering of multidimensional data (e.g. tetra-nucleotide frequencies). The results of this clustering are then represented on a two dimensional map. The usage of 4-mer frequencies, by both these methods, is based on previous observations that 4-mers have optimal taxonomic discrimination capability as opposed to other k-mer frequencies [42]. In contrast, the CompostBin method involves computing frequencies of k-mers of various lengths and subsequently adopting a weighted PCA-based strategy to reduce the dimensionality of compositional space. However, both TETRA and CompostBin require sequences of sufficient length for optimal binning performance. Furthermore, in scenarios where the sample contains multiple species with highly varying levels of abundance, methods like TETRA tend to create multiple clusters (bins) for reads originating from the highly abundant species. This limitation is addressed by the recently published AbundanceBin method [40] which models the number of reads originating from different species using separate Poisson distributions. The objective of AbundanceBin is to form bins containing reads originating from species having similar abundance levels. Although AbundanceBin works efficiently with samples having highly varying abundance levels, its binning efficiency is observed to decrease with simulated samples having an even distribution of species. However, given that environmental samples with an even species distribution are highly unlikely to occur, this limitation (poor

efficiency in samples with an even species distribution) is not an issue in practice. Nevertheless, another method, namely MetaCluster [41], attempts to address such hypothetical 'even distribution' scenarios by adopting a two-phase strategy. In the first phase, reads are segregated into taxonomically homogeneous clusters of similar sizes. However, since this phase is likely to result in distributing reads of species with high abundances into several clusters, the second phase of MetaCluster involves merging of different clusters (expected to contain fragments from the same species) by generating probabilistic models that are based on the l-mer distributions of the fragments constituting each cluster.

SELECTION OF BINNING METHODS

Taxonomic classification of metagenomic data is primarily performed with the objective of cataloging/classifying various microbial groups inhabiting a given environment. Subsequent analyses involve comparing the obtained taxonomic profile with those of related environments to identify spatial and/or temporal variations in the microbial community structure and characterizing the identified variations in taxonomic/functional terms. The extent to which these subsequent analyses are successful is intricately dependent on the resolution and the accuracy of the obtained taxonomic profiles. Accurate results obtained with high resolution (i.e. at specific taxonomic levels) aid in identifying the subtle differences between metagenomes.

Depending on the context/setting of the metagenomic study, results of binning are used for addressing/answering various questions. In clinical settings, results of binning help in the identification of key microbial groups responsible for the onset and progression of various diseases and/or physiological disorders. For example, a comparison of the taxonomic profiles (obtained using SPHINX) of gut metagenomes sequenced from healthy and malnourished children have implicated bacterial species belonging to Campylobacterales lineage to be associated with malnourishment [43]. In yet another recent study, comparative analysis of taxonomic profiles (obtained using PhymmBL and MEGAN) of oral metagenomes [44] has indicated that dental cavities house a complex community of microbes belonging to diverse bacterial lineages and are not specifically dominated by *Streptococcus mutans* (earlier considered as the prime causative agent of dental caries). Interestingly, results of binning, in this study, have therefore helped in negating the role of a microbial species which was earlier thought to be associated with a specific disease (in this case, dental caries). In the context of ecological studies, taxonomic diversity estimates (obtained through binning) provide crucial insights with respect to the spatial and temporal variations within microbial communities residing in diverse environments. For instance, the taxonomic characterization of the terephthalate (TA) wastewater metagenome (using Phylopythia) has helped in identifying specific microbial species that play a key role not only in the degradation of TA but also in maintaining the stability of this unique microbial community [45].

In principle, any taxonomy-dependent binning method can be employed to classify sequences constituting metagenomic datasets. However, the length of the metagenomic sequences (which is dependent on the sequencing platform) is generally observed to be a key factor that drives the selection of the binning method. Sequences of relatively longer lengths are amenable for both alignment-based as well as composition-based binning methods. However, the latter category of methods are preferable due to their faster execution speeds and low memory requirements. For example, in the TA wastewater metagenome study mentioned above [45], a composition-based method (Phylopythia) was employed, given that input sequences were of sufficient length. In contrast, for lower length sequences (with weak compositional signals), it is preferable to adopt alignment-based or hybrid binning methods. For instance, the lower length of sequences (200–400 bp) in both the malnourished gut metagenomic study [43] and the oral metagenomes [44] necessitated the adoption of hybrid (SPHINX and PhymmBL) and alignment-based (MEGAN) binning methods. On the other hand, for ultra-short sequences, a pre-assembly step becomes a necessity prior to performing taxonomic classification. For example, in a comparative study on human gut metagenomes [46], metagenomic sequences of length of ~75 bp (obtained using the Illumina sequencing platform) were first assembled into contigs and subsequently classified using MEGAN. The extent of coverage achieved using the present generation of sequencing technologies is yet another important factor that affects binning results. Results obtained with metagenomes sequenced at a higher coverage

are expected to capture the rare taxa resident in a particular environment.

Taxonomy-independent methods are more relevant in cases of metagenomes where the proportion of taxonomically classifiable sequences using taxonomy-dependent methods is very low. The results obtained using taxonomy-independent methods may also aid in downstream processes like assembly. A progressive assembly of sequences constituting individual clusters (expected to be more or less taxonomically homogeneous) is likely to reduce the time/memory requirements of the downstream assembly process.

VALIDATION STRATEGIES

Taxonomy-dependent methods

In order to validate taxonomy-dependent methodologies, validation should be performed using simulated metagenomic datasets and databases (Figure 3). Reads in these datasets should simulate the lengths as well as the error models associated with various sequencing technologies. For measuring scalability, multiple datasets of varying sizes should be used. Three simulated datasets currently being employed as a 'gold-standard' for evaluating the performance of various metagenomics analysis algorithms (including binning) are the 'Fidelity of Analysis of Metagenomic Samples' (FAMeS) datasets [47]. These datasets of varying taxonomic complexity contain approximately 100 000 reads having lengths ranging between 650 and 1000 bp. These reads, sampled from 112 real-world genome sequencing projects data, contain typical sequencing errors associated with Sanger sequencing technology. However, evaluating binning efficiency using only these datasets is not comprehensive, given that reads in these datasets do not represent lengths and error models corresponding to the present generation of sequencing technologies. Given this, softwares like MetaSim [48] and ART [49] have been developed to simulate reads generated using the latter technologies. A comparison of the pattern of taxonomic assignments, obtained for simulated datasets generated using these read-simulators (Supplementary Material 1), indicates that end-users

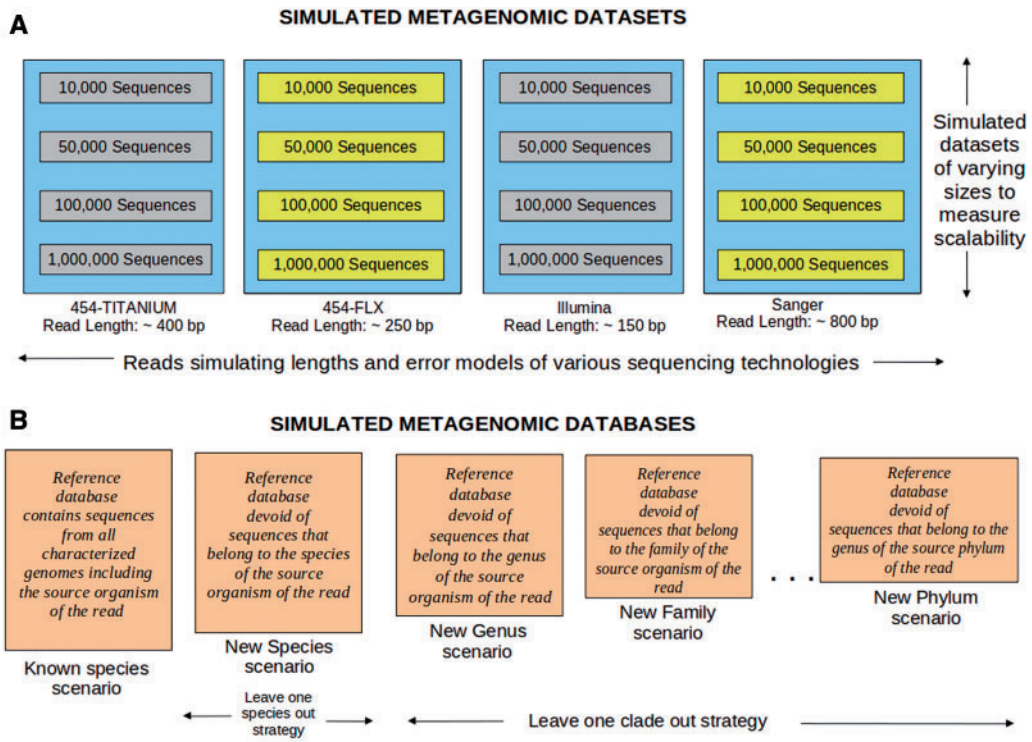


Figure 3: Suggested design of simulated (**A**) metagenomic datasets and (**B**) databases for validation of assignment-dependent binning algorithms. Simulated datasets should ideally mimic read lengths and error models associated with various sequencing technologies. Datasets containing varying number of input sequences should be considered to assess scalability. A range of databases simulating a 'leave one species out' or 'leave one clade out' scenario should be constructed.

can employ either of these softwares for generating evaluation datasets.

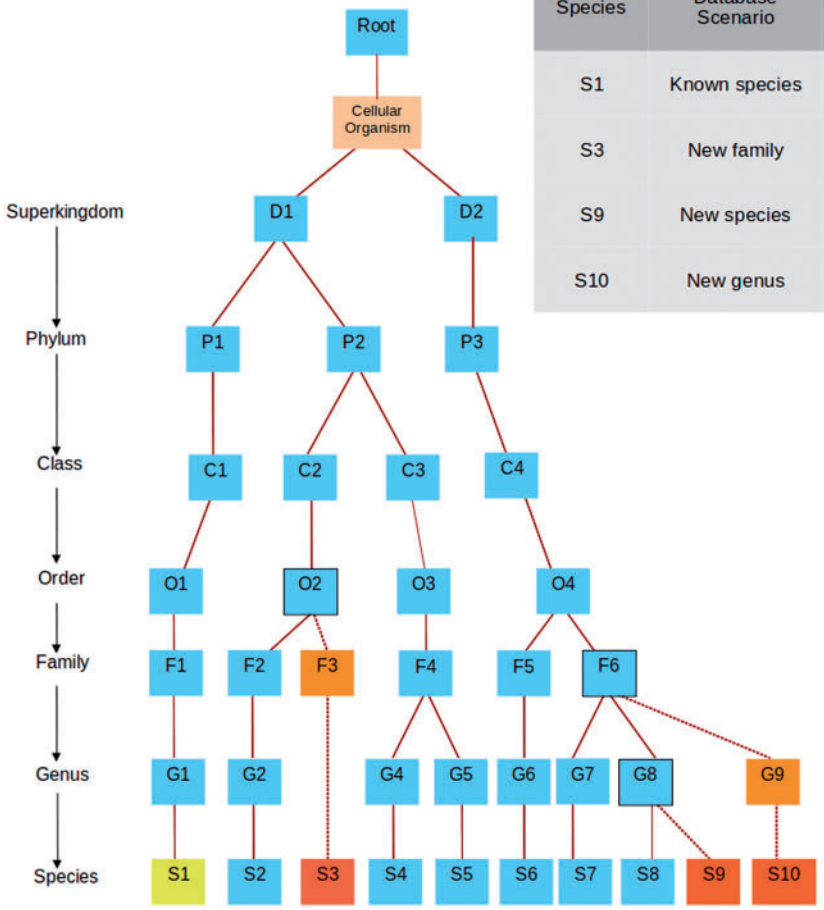
The design of the reference database is another important aspect to be considered during evaluation of taxonomy-dependent methods. Simulated databases should mimic real-world metagenomic scenarios wherein they are devoid of sequences that are taxonomically related to the input reads at various taxonomic levels. This simulation is typically done by adopting a 'leave one (species) out' strategy, wherein sequences or models belonging to only a single species are removed from reference databases and validation is performed using reads from this species [15, 21]. However, in typical metagenomic scenarios, query reads may originate from entirely new taxonomic clades, not necessarily diverged at the species level. Consequently, to encompass such scenarios, it is preferable to adopt 'leave one clade out' strategies, wherein sequences belonging to an entire clade (genus, family, order, class, phylum, etc.) are removed from the reference database (Figure 3). Such a validation strategy has been adopted in methodologies like SOrt-ITEMS [16] and DiScRIBinATE [17].

Binning efficiency of taxonomy-dependent methods is typically quantified in terms of four parameters, viz. accuracy, specificity, execution time and requirements of compute power. Assignment of a read is considered to be accurate if it is assigned to 'any' taxon that lies in the taxonomic lineage of the source organism of the read. On the other hand, assignment specificity is defined in terms of the taxonomic level (strain, species, genus, family, order, class, phylum, superkingdom) at which the read is assigned. Assignments at strain level are considered to be the most specific in scenarios where the reads originate from a known strain, sequences of which are present in the reference database. However, in most metagenomic scenarios, the taxon corresponding to the source organism of the read is absent in the reference database. In such scenarios, an ideal binning method (with high specificity) is expected to identify and assign such a read to an appropriately higher level taxon that represents the point of evolutionary divergence between the source organism (of the read) and various taxa in the known reference taxonomic tree (Figure 4). Various taxonomy-dependent methods currently use different measures for quantifying accuracy and specificity. Methods like MEGAN and SOrt-ITEMS compute the percentage of correctly assigned reads at different taxonomic levels, and use this information as a measure to quantify accuracy and specificity. On the other hand, methods like CARMA and TACOA evaluate these parameters in terms of true and false positive rates.

Requirements of compute power and the overall processing time are also significant parameters that need to be considered while evaluating binning efficiency. It has been observed that there exists a trade-off between the accuracy and specificity of a method and the requirements of time/compute power. For instance, though composition-based methods have been shown to score over alignment-based methods in terms of execution time and compute power, the relatively lower accuracy and specificity of these methods as compared to alignment-based methods and their limited applicability with metagenomic datasets containing short reads, still remain a point of concern. It is encouraging to note that the recently reported hybrid binning methods [30, 35] utilize the principles of both alignment and composition-based approaches in order to capitalize on the relative advantages of both. Comparative evaluation of binning efficiency (with respect to the trade-off between accuracy, specificity and execution time) of methods belonging to all three categories (alignment-based, composition-based and hybrid) have already been performed and discussed in earlier studies [34, 35]. A summary of these results is provided in Supplementary Material 2.

Taxonomy-independent methods

In contrast to taxonomy-dependent methods, performance evaluation of taxonomy-independent methods is typically done in the following manner. Simulated datasets containing a mixture of reads originating from multiple species are first binned using these methods. Binning efficiency is then evaluated using parameters such as taxonomic homogeneity of the resulting bins and the number as well as the size of bins generated. Ideally, an efficient method is expected to form '*n*' number of taxonomically homogenous bins where '*n*' is the number of species constituting the validation dataset. However, in cases where in multiple homogeneous bins arise due to segregation of reads from the same species, additional evaluation parameters, such as normalized mutual information (NMI) and F-score need to be employed. A comprehensive description of these parameters is provided in an earlier review by



Species	Database Scenario	Nodes representing 'correct' assignments (in decreasing order of binning 'specificity')
S1	Known species	S1, G1, F1, O1, C1, P1, D1, Cellular Organisms, Root
S3	New family	O2, C2, P2, D1, Cellular Organisms, Root
S9	New species	G8, F6, O4, C4, P3, D2, Cellular Organisms, Root
S10	New genus	F6, O4, C4, P3, D2, Cellular Organisms, Root

Figure 4: Schematic representation of a taxonomic tree and an associated table indicating the accuracy and specificity of assignments under various database scenarios. In this figure, species SI represents a known species whose sequences are present in the reference database. Species S3, SI0 represent hitherto unknown species belonging to a new family (F3) and new genus (G9), respectively. Species S9 represents a hitherto unknown species belonging to known genus (G8). Dotted lines indicate novel lineages. The inset table indicates the accuracy and specificity of assignments for reads in various database scenarios.

Sun *et al.* [8]. To ensure comprehensive evaluation, the methods should ideally be tested with simulated datasets of varying taxonomic complexity. For instance, CompostBin [39] first computes the error rate of each generated bin in terms of the number of misclassified DNA sequences present in the bin. Subsequently, the error rate of the method on the given dataset is computed as the average of the error rates obtained for each of the individual bins. A similar evaluation strategy has been used by MetaCluster [41].

CHALLENGES

In spite of the availability of several approaches for binning metagenomic sequence datasets, there are several challenges which still remain to be addressed. These challenges are discussed below.

Pre-processing stage

Notwithstanding the availability of efficient binning methods, estimating the taxonomic diversity of any microbial community critically depends on the efficacy of the initial experimental steps like sample collection, preparation, DNA extraction and sequencing. Limitations in DNA extraction/sequencing protocols can severely bias the representation of different species in the extracted DNA sample, consequently leading to erroneous estimates of taxonomic diversity [50]. Furthermore, host-associated metagenomes are frequently known to contain a significant proportion of contaminating sequences

originating from the host genome. Ideally, such datasets should be 'de-contaminated' using available methods like Eu-Detect [51] or DeConseq [52] prior to binning.

Optimization of algorithms

An important consideration that still remains to be addressed by most of the binning methods pertains to the representation bias of different taxonomic groups in existing reference databases. Due to priorities of scientific research, reference databases are observed to be biased with sequences from organisms/clades having pathogenic or industrial implications. For instance, >60% of prokaryotic sequences in the nr database belong to phylum Proteobacteria. In contrast, phyla like Chlorobi and Fusobacteria have <1% representation. This uneven representation biases the scoring/classification step of binning methods (especially composition-based approaches) towards highly represented taxonomic groups, thereby adversely impacting binning accuracy. Currently, a few (existing) binning methods (INDUS being an example) incorporate suitable normalization procedures in their work-flow to address this issue.

Both taxonomy-dependent and taxonomy-independent methods have their own drawbacks. Existing taxonomy-dependent methods fail to classify a large fraction of reads (originating from hitherto unknown organisms). This in turn affects abundance and diversity estimates in unknown ways. On the other hand, although taxonomy-independent methods cluster all reads, the taxonomic affiliations of these reads cannot be identified. Development of methods that can capitalize on the advantages of taxonomy dependent as well as taxonomy-independent methods still remains an open challenge. Furthermore, the compute requirements of binning methods (especially the alignment-based methods) are still observed to be high. This limits their usage to research labs having sufficient computational resources/infrastructure. Though hybrid methods such as PhymmBL [30] and SPHINX [35] significantly reduce the overall compute requirements and binning time, their overall accuracy is yet to match the levels attained by pure alignment-based methods. In addition, the efficiency of binning methods is generally observed to be relatively low with ultra-short reads (length <50 bp) generated using technologies like SOLiD (http://solid.appliedbiosystems.com). Without a pre-assembly step, it is still challenging to classify such reads. However, given that current generation of sequencing technologies are gradually moving towards generating reads of relatively longer length, the above aspect may not be a challenge in the near future. In addition, a majority of binning methods are optimized for binning reads originating from prokaryotic genomes. Methods specifically designed for binning reads originating from pico-eukaryotic and fungal genomes are currently unavailable.

Post-processing stage

Some of the binning algorithms (e.g. Phymm, PhymmBL, NBC classifier, etc.) are observed to classify all reads in a dataset (irrespective of their origin from known or hitherto unknown organisms) at the level of strain/species. Although these methods possess reasonably high levels of classification efficiency, the absence of a correlation between the assignment score and the taxonomic level of divergence makes it difficult for end-users to properly interpret these results. Furthermore, it is observed that binning estimates obtained using most of the currently available methods are generally not normalized with respect to varying sizes of the source genomes. This aspect still remains to be addressed.

CONCLUSIONS

With rapid advances in sequencing technologies and the increased focus on personalized genomic solutions, the field of metagenomics research is currently witnessing exponential growth. This has necessitated the development of computational tools that enable efficient and accurate analysis of metagenomic datasets. The problem of binning, i.e. taxonomic characterization of metagenomes, is currently being addressed by several research groups. Typical challenges related to binning, due to incomplete databases, insufficient read lengths and high sequencing error rates, are expected to ease with improvements in sequencing technologies. Furthermore, the recent emergence of single cell sequencing technologies [53, 54], which attempt to perform an experimental pre-segregation of microbial cells prior to sequencing, are expected to further ease the computational challenges associated with binning. In summary, technological and computational advances seen in the recent past provide a positive outlook with respect to using the power of the metagenomics approach to obtain greater insights about the vast

majority of hitherto unknown microbes present in various environments.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Accurate estimates of taxonomic diversity, obtained using binning methods, provide valuable insights regarding the structure and dynamics of microbial communities residing in any given environment.
- Need to comprehensively evaluate efficiency of binning algorithms not only in terms of accuracy and specificity, but also with respect to time and compute requirements.
- Alignment-based binning procedures have relatively higher binning accuracy and specificity than composition and hybrid approaches. However, it is difficult to employ them in resource poor settings.
- Advancements in sequencing technologies are expected to improve binning efficiency in the near future.
- Need for development of gold-standard simulated datasets as well as databases and uniform standards for evaluating binning methods.

References

1. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 1995;**59**(1):143–69.
2. Tyson GW, Chapman J, Hugenholtz P, *et al*. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 2004; **428**(6978):37–43.
3. Doolittle WF, Zhaxybayeva O. On the origin of prokaryotic species. *Genome Res* 2009;**19**(5):744–56.
4. Colwell RK. Biodiversity: concepts, patterns, and measurement. In: Levin A, (ed). *The Princeton Guide to Ecology*. Princeton, NJ, USA: Princeton University Press, 2009; 257–63.
5. Clarridge JE III. Impact of 16S rRNA gene sequence analysis for identification of bacteria on clinical microbiology and infectious diseases. *Clin Microbiol Rev* 2004;**17**:840–62.
6. Santamaria M, Fosso B, Consiglio A, *et al*. Reference databases for taxonomic assignment in metagenomics. *Brief Bioinform* 2012. [Epub ahead of print].
7. Ribeca P, Valiente G. Computational challenges of sequence classification in microbiomic data. *Brief Bioinform* 2011;**12**(6):614–25.
8. Sun Y, Cai Y, Huse SM, *et al*. A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief Bioinform* 2011;**13**(1): 107–21.
9. Altschul S, Gish W, Miller W. Basic local alignment search tool. *J Mol Biol* 1990;**215**(3): 403–10.
10. Kent WJ. BLAT - the BLAST-like alignment tool. *Genome Res* 2002;**12**(4):656–64.
11. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 2010;**26**(5): 589–95.
12. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**:R25.
13. 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.
14. Seshadri R, Kravitz SA, Smarr L, *et al*. CAMERA - a community resource for metagenomics. *PLoS Biol* 2007;**5**(3): e75.
15. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res* 2007;**17**(3):377–86.
16. Monzoorul Haque M, Ghosh TS, Komanduri D, *et al*. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* 2009;**25**(14):1722–30.
17. Ghosh TS, Monzoorul Haque M, Mande SS. DiScRIBinATE: a rapid method for accurate taxonomic classification of metagenomic sequences. *BMC Bioinformatics* 2010;**11(Suppl 7)**:S14.
18. Ghosh TS, Mohammed MH, Komanduri D, *et al*. ProViDE: a software tool for accurate estimation of viral diversity in metagenomic samples. *Bioinformation* 2011;**6**(2): 91–4.
19. Liu B, Gibbons T, Ghodsi M, *et al*. MetaPhyler: taxonomic profiling for metagenomic sequences. *Proc IEEE Bioinform Biomed* 2010;95–100.
20. Horton M, Bodenhausen N, Bergelson J. MARTA: a suite of java-based tools for assigning taxonomic status to DNA sequences. *Bioinformatics* 2010;**26**(4):568–9.
21. Krause L, Diaz NN, Goesmann A, *et al*. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res* 2008;**36**(7):2230–9.
22. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* 2008; **9**(10):R151.
23. Stark M, Berger SA, Stamatakis A, *et al*. MLTreeMap–accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics* 2010;**11**:461.
24. Schreiber F, Gumrich P, Daniel R, *et al*. Treephyler: fast taxonomic profiling of metagenomes. *Bioinformatics*. 2010; **26**(7):960–1.
25. Matsen AFA, 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.
26. Berger SA, Stamatakis A. Aligning short reads to reference alignments and trees. *Bioinformatics* 2011;**27**(15):2068–75.
27. McHardy AC, Martín HG, Tsirigos A, *et al*. Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* 2007;**4**(1):63–72.
28. Rosen GL, Reichenberger ER, Rosenfeld AM. NBC: the naive bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics* 2010;**27**(1): 127–9.
29. 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.

30. Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated markov models. *Nat Methods* 2009;**6**(9):673–6.
31. Pati A, Heath LS, Kyrpides NC, *et al*. ClaMS: a classifier for metagenomic sequences. *Stand Genomic Sci* 2011;**5**(2): 248–53.
32. Nalbantoglu OU, Way SF, Hinrichs SH, *et al*. RAIphy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles. *BMC Bioinformatics* 2011;**12**:41.
33. Cole ST, Brosch R, Parkhill J, *et al*. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* 1998;**393**:537–44.
34. Mohammed MH, Ghosh TS, Reddy RM, *et al*. INDUS - a composition-based approach for rapid and accurate taxonomic classification of metagenomic sequences. *BMC Genomics* 2011;**12(Suppl 3)**:S4.
35. Mohammed MH, Ghosh TS, Singh NK, *et al*. SPHINX - an algorithm for taxonomic binning of metagenomic sequences. *Bioinformatics* 2011;**27**(1):22–30.
36. 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.
37. Chan CK, Hsu AL, Halgamuge SK, *et al*. Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics* 2008;**9**:215.
38. Ultsch A, Moerchen F. ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. Technical Report, Vol. 46. Germany: Department of Mathematics and Computer Science, University of Marburg, 2005.
39. Chatterji S, Yamazaki I, Bai Z, *et al*. CompostBin: a DNA composition-based algorithm for binning environmental shotgun reads. *Res in Comp Mol Biol (LNCS)* 2008;**4955**: 17–28.
40. Wu YW, Ye Y. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J Comput Biol* 2011;**18**(3):523–34.
41. Leung HCM, Yiu SM, Yang B, *et al*. A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. *Bioinformatics* 2011;**27**(11): 1489–95.
42. Pride DT, Meinersmann RJ, Wassenaar TM, *et al*. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* 2003;**13**:145–58.
43. Gupta SS, Mohammed MH, Ghosh TS, *et al*. Metagenome of the gut of a malnourished child. *Gut Pathog* 2011;**3**:7.
44. Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, *et al*. The oral metagenome in health and disease. *ISME J* 2012;**6**(1): 46–56.
45. Lykidis A, Chen CL, Tringe SG, *et al*. Multiple syntrophic interactions in a terephthalate-degrading methanogenic consortium. *ISME J* 2011;**5**(1):122–30.
46. Qin J, Li R, Raes J, *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 2010;**464**(7285):59–65.
47. Mavromatis K, Ivanova N, Barry K, *et al*. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* 2007;**4**(6):495–500.
48. Richter DC, Ott F, Auch AF, *et al*. MetaSim—a sequencing simulator for genomics and metagenomics. *PLoS One* 2008; **3**(10):e3373.
49. Huang W, Li L, Myers JR, *et al*. ART: a next-generation sequencing read simulator. *Bioinformatics* 2011;**28**(4):593–4.
50. Morgan JL, Darling AE, Eisen JA. Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One* 2010;**5**(4):e10209.
51. Mohammed MH, Chadaram S, Komanduri D, *et al*. Eu-Detect: an algorithm for detecting eukaryotic sequences in metagenomic data sets. *J Biosci* 2011;**36**(4):709–17.
52. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLoS One* 2011;**6**(3):e17288.
53. Raghunathan A, Ferguson HR, Bornarth CJ, *et al*. Genomic DNA amplification from a single bacterium. *Appl Environ Microbiol* 2005;**71**:3342–7.
54. Stepanauskas R, Sieracki ME. Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc Natl Acad Sci USA* 2007;**104**:9052–7.