Briefings in Bioinformatics, 22(1), 2021, 88–95

doi: 10.1093/bib/bbaa121
Advance Access Publication Date: 24 June 2020
Method Review

Comparison of microbiome samples: methods and computational challenges

Matteo Comin, Barbara Di Camillo and Cinzia Pizzi and Fabio Vandin

Corresponding author: Fabio Vandin, Department of Information Engineering, University of Padova, Padova 35131, Italy. Tel: +39-0498277946;
Fax: +39-0498277799; E-mail: fabio.vandin@unipd.it
All authors contributed equally to the work.

Abstract

The study of microbial communities crucially relies on the comparison of metagenomic next-generation sequencing data sets, for which several methods have been designed in recent years. Here, we review three key challenges in the comparison of such data sets: species identification and quantification, the efficient computation of distances between metagenomic samples and the identification of metagenomic features associated with a phenotype such as disease status. We present current solutions for such challenges, considering both reference-based methods relying on a database of reference genomes and reference-free methods working directly on all sequencing reads from the samples.

Key words: microbiome; metagenomics; next-generation sequencing.

Introduction

The study of the microbiome, the collective genetic material from the microorganisms found in a given environment, has seen a rapid expansion in recent years due to the exceptional advances in sequencing technologies [1]. These advances have enabled a tremendous expansion in our ability to interrogate microbial genomes, moving from the study of microbes that are amenable to cultivation to the interrogation of all microbes in a sample.

Sequencing-based approaches for microbiome research consist mostly of high-throughput sequencing of marker genes, usually regions of one rRNA gene (e.g. 16S for bacteria), and the ‘shotgun’ sequencing of microbial DNA, without selection of any specific gene [2]. While both approaches are sometimes referred to as metagenomics, the term ‘metataxonomics’ is usually preferred for marker gene sequencing, since it does not allow to investigate full-genome information. Marker gene sequencing is appealing in terms of costs and the possibility, assuming the same marker gene region is targeted, to compare species abundances across samples (even for lowly abundant species), but it does not provide genomic information for all microbiome members, since only the sequenced gene is available. On the other hand, shotgun sequencing can, potentially, provide (partial) information for all microbiome members.

The analysis of the microbiome is elucidating the crucial role of microbial communities in the environment and in human health [3, 4]. A typical metagenomic study comprises several steps, ranging from the collection and sequencing of samples

Matteo Comin is an associate professor of computer engineering at the University of Padova. His research interests focus on the area of algorithms for computational biology.
Barbara Di Camillo is an associate professor of computer engineering at the University of Padova. Her research focuses on the area of machine learning and modeling applied to bioinformatics.
Cinzia Pizzi is an associate professor of computer engineering at the University of Padova. Her research focuses on algorithms and data structures for genome analysis, applied to phylogenomics and metagenomics.
Fabio Vandin is a professor of computer engineering at the University of Padova. His research focuses on computational methods for genomics, including the analysis of cancer genomes and metagenomics.
Department of Information Engineering, University of Padova, Padova, 35131, Italy

Submitted: 22 December 2019; **Received (in revised form):** 9 May 2020

to the computational analysis of the results and to an eventual validation [5]. While some applications, for example the surveillance of food supply, mostly require to analyze single samples in isolation, one step that is common to most studies is the comparison of metagenomic samples. For example, in ecology, metagenomic samples are compared to identify similarities and differences between microbial habitats, while in clinical research, metagenomic samples are compared to identify metagenomic features distinguishing groups with different characteristics (e.g. cases versus controls). The challenges arising in the comparison of metagenomic samples and currently available computational solutions are the focus of this review. For a discussion of other challenges arising in the computational analysis of metagenomes, we refer to other recent reviews (e.g. [5–8]), not focused on the comparison of metagenomic samples. In brief, [5] assesses the common problems in the several steps of the design and analysis of shotgun sequencing experiments, including critical steps outside the data analysis (e.g. the collection, processing and sequencing of the samples and the validation of results). Breitwieser *et al.* [6] focus on methods and databases used for the tasks of read classification and metagenomic assembly. Chiu and Miller [7] present the challenges of implementing next-generation sequencing of metagenomes in the clinical laboratory. In addition, benchmarks of some of the components of a metagenomic data analysis pipeline have been recently published. For example, [8] presents a benchmark of 20 metagenomic classifiers using synthetic and real data, while the CAMI initiative [9] has produced a comprehensive assessment of methods for metagenomic species detection and other tools in metagenomics sample analysis.

There are two major classes of methods for metagenome analysis: reference-based methods, which map the reads obtained from sequencing the microbiome against a database of reference genomes (for example, to detect and quantify the species in the sample), and reference-free methods, which process the reads without relying on an external set of references. By comparing the sequencing reads to a set of reference genomes, reference-based methods allow the identification and quantification of known species or, more generally, of operational taxonomic unit (OTU) as well as the identification of biological functional pathways. However, reference-based methods suffer from the lack of comprehensiveness of the genome catalogs they rely on, which is due to most microbes being difficult to culture. On the other hand, reference-free methods do not suffer from the biases in the reference sequence resources but do not directly provide information on known species or insights at the functional level, even if such information can be obtained after an inferential step to identify putative taxonomy or putative functional features (e.g. biological functional pathways can be identified relying on computational methods for the inference of putative genes and other functional features [10]).

In the remaining of this review, we focus on three major computational challenges arising in the comparison of metagenomic samples and provide an overview of the currently available solutions. In particular, we first discuss the problem of species detection and quantification and consider both reference-based and reference-free approaches. We then consider the computation of distances between metagenomic samples. Finally, we discuss the comparison of metagenomic samples in the context of metagenome-based disease status classification. While the 1st two challenges arise mostly in the analysis of shotgun sequencing data, the 3rd one is relevant for both the analysis of marker gene sequencing data and shotgun sequencing data.

Computational challenges and methods

Species detection and quantification

Microbial communities can be analyzed and compared through the detection and quantification of the species they contain. The detection and quantification of species in a sample can be carried out using a set of reference genomes, e.g. bacteria and viruses, or without them (reference free). In the first case, also known as taxonomic classification or taxonomic binning, the input sequences are clustered into bins corresponding to their taxonomic ID. These reference-based taxonomic classification methods are useful for the identification of organisms with close relatives in the reference database. When no close relative of a species is in the reference database, reference-free binning of the reads may be a useful 1st step in the analysis.

Reference-based species detection

The reference-based methods, a.k.a. taxonomic binning, can be broadly divided into three categories: (1) alignment-based methods; (2) marker-based methods, where certain specific marker sequences are used to identify the species; and (3) sequence-composition-based methods, which are based on the nucleotide composition (e.g. k-mers usage). Traditionally, the 1st strategy was to use BLAST [11] to align each read with all sequences in GenBank. Later, faster methods have been deployed for this task; popular examples are MegaBlast [12] and Megan [13]. However, as the reference databases and the size of sequencing data sets have grown, alignment has become computationally infeasible, leading to the development of metagenomics classifiers that provide much faster results.

Marker-based methods use clade-specific marker genes as a taxonomic reference, so that the identification of one of these genes can be used as evidence that a given taxa is present. This allows faster assignment because the database of marker genes is far smaller than a database of the full genomes for all species. Popular examples of marker gene methods are MetaPhlAn [14], which uses Bowtie2 as fast and sensitive read aligner, and Phylosift [15], which is based on the aligner LAST. These algorithms do not classify the input reads directly; instead, they provide the microbial composition, expressed in terms of relative abundance for all taxa that they recognize in the sample.

The fastest and most promising approaches belong to the composition-based category [16]. The composition-based methods exploit the full potential of sequencing, as opposed to marker gene methods, where most of the reads in a sample do not receive a classification because they are not mapped to a marker gene. The basic principles of composition-based methods can be summarized as follows: each genome of reference organisms is represented by its k-mers and the associated taxonomic label of the organisms, then the reads are searched and classified throughout this k-mers database. For example, Kraken [17] constructs a data structure that is an augmented taxonomic tree in which a list of significant k-mers is associated to each node, leafs and internal nodes. Given a node on this taxonomic tree, its list of k-mers is the representative for the taxonomic label, and it will be used for the classification of metagenomic reads. In the classification step, each read is decomposed into its k-mers and these k-mers are searched in the tree, then the read is classified by searching the highest-weighted path in the taxonomic tree. Clark [18] uses a similar approach, building databases of species- or genus-level specific k-mers and discarding any k-mer mapping to higher levels. The precision of these methods is as good as MegaBlast

[12]; nevertheless, the processing speed is much higher [16]. Several other composition-based methods have been proposed in recent years. In Girotto *et al.* [19], the number of unassigned reads is decreased through reads overlap detection and species imputation. Centrifuge and Kraken 2 [20, 21] try to reduce the size of the *k*-mer database with the use of FM-index and minimizers, respectively. The sensitivity can be improved by filtering uninformative *k*-mers [22] or by using spaced seeds instead of *k*-mers [23].

Reference-free species quantification

Reference-based methods are based on a database of reference genomes and the associated taxonomic labels. This information is usually indexed in a *k*-mers database for fast queries. Although taxonomic read classification is very efficient, the construction of *k*-mers databases usually is very demanding, requiring large amounts of RAM and disk space. Another drawback is the fact that most bacteria found in environmental samples are unknown and cannot be cultured and separated in the laboratory [24]. As a consequence, the genomes of most microbes in an environmental sample lack a taxonomically related sequences in existing reference databases.

Reference-free methods, a.k.a. genome binning, do not require to know all the genomes in the sample; instead, they try to divide the reads into groups so that reads from the same species are clustered together. Reference-free classification tools, also known as binning tools, are based on the observation that the *k*-mer distributions of the DNA fragments from the same genome are more similar than those from different genomes. Thus, without using any reference genome, one can determine if two fragments are from genomes of similar species based on their *k*-mer distributions. The major problem when processing metagenomic data is the fact that the proportion of species in a sample, a.k.a. abundance rate, can vary greatly. Most of the tools can only handle species with even abundance ratios, and their binning performances degrade significantly in real situations when the abundance ratios of the species are different. For example, BiMeta [25] and MetaCluster [26] try to group the reads into many small clusters so that reads from minority species (with low abundance ratios) could exist as isolated clusters. Both these methods use as means of comparison the Euclidean distance between the vectors of *k*-mers counts on the clusters groups. AbundanceBin [27] works well for very different abundance ratios, but problems arise when some species have similar abundance ratios. In Girotto *et al.* [28], reads are clustered based on a self-standardized statistic, derived from alignment-free statistics, that is not dominated by the noise in the individual sequences and that can compare groups of reads with different abundance ratios.

Another important step of metagenome analysis is the reconstruction of new genomes through assembly. Sample metagenomes can be assembled into contigs, and contig binning serves as the key step toward the detection of new species, taxonomic profiling and downstream functional analysis. Grouping contigs into bins of putative species is one of the hurdles faced when analyzing metagenomic data. Typically, a few issues are encountered including struggling to differentiate related microorganisms, repetitive sequence regions within or across genomes, sequencing errors and strain-level variation within the same species, decreasing accuracy for contigs below a size threshold or excluding low-coverage and low-abundance organisms [29, 30]. Several techniques have been developed for contig binning, where studies extract features from contigs to infer bins based on sequence composition [31, 32], abundance [33] or hybrids of both sequence composition and abundance [29, 34–37]. Some hybrid binning tools, such as CONCOCT [29], MaxBin2.0 [34], GroopM [35] and MetaCon [37], are designed to bin contigs based on multiple related metagenomic samples. Among these methods, GroopM [35] is advantageous in its visualized and interactive pipeline. On one hand, it is flexible, allowing users to merge and split bins; on the other hand, in the absence of expert intervention, the automatic binning results of GroopM are not as satisfactory as CONCOCT [29]. CONCOCT [29] makes use of the Gaussian mixture model to cluster contigs into bins. MetaBAT2 [36] calculates integrated distance for pairwise contigs and then clusters contigs by an iterative graph partitioning procedure. MaxBin [34] compares the distributions of distances between and within the same genomes. In Qian and Comin [37], metagenomic contigs are clustered based on probabilistic *k*-mers statistics, contigs coverage and length. Therefore, these approaches can be applied to bin contigs from incomplete or uncultivated genomes.

The CAMI initiative [9] has developed a comprehensive assessment for metagenomics species detection and others challenges in metagenomics sample analysis. The authors of [9] concluded that most reference-based read classification methods are able to reconstruct taxon bins of acceptable quality down to the family rank. Overall, all tools are more precise when reconstructing genomes than for species or genus bins, indicating that the decreased performance for low ranks is due partly to limitations of the reference taxonomy. As for the reference-free methods, most genome binners performed well when no closely related strains are present.

Computing distances between metagenomic samples

After the detection and quantification of the species in each sample using reference-based methods, microbial communities can be analyzed and compared using ecological measures, such as species diversity, richness and uniformity. However, reference-based methods suffer from the biases of the reference databases they rely on. A different approach is provided by *de novo* comparative metagenomics, which is the comparison among metagenomic samples based entirely on their reads content. *De novo* comparative metagenomics enables new insights, which are not restricted to the availability and completeness of *ad hoc* databases.

The main issue with the *de novo* comparison between two (or more) metagenomics samples is that, since they often includes millions of reads, an all-against-all comparison of their content becomes impractical from the computational point of view. To overcome this problem, several methods have been proposed in literature to make metagenomic samples comparison computationally affordable, while still effective for the identification of biological diversity. The computational efficiency is especially important for large metagenomic projects (e.g. the Human Microbiome Project [38]) where a large number of samples are sequenced and compared. Many of these methods consider the Jaccard distance [39–43], which measures shared content between samples. Extension to a variety of more powerful environmental distances has been implemented more recently [42–46]. Vector distance between *k*-mers abundance descriptors has also been considered [45].

A common factor to all these methods is that the computation of such distances is based on *k*-mers presence and/or abundance. In fact, while the Jaccard similarity measure and many ecological distances have been originally defined in terms of species presence and/or abundance, alternative *k*-mer-based definitions have been proposed [39, 42], supported by recent studies [47] showing that *k*-mer-based distances are well correlated to taxonomic ones.

In order to achieve computational scalability for similarity measures between all pairs of samples in a metagenome project, several techniques have been developed. In the remaining of this section, we focus on the three main techniques that are currently employed, in different ways, by tools for metagenomic samples comparison: probabilistic data structures [39, 40], parallel/distributed computation [42, 45] and dimensionality reduction through sampling [41, 43, 44, 46].

Probabilistic data structures

Compareads [39] and its evolution COMMET (COmpare Multiple METagenomes) [40] are among the 1st approaches proposed in literature to compare metagenomic samples. They both aim at the identification of shared contents, in terms of reads, between two (or more) samples, which allows the computation of an approximation of the Jaccard distance. In Compareads and COMMET, two reads are considered similar if they share a number of *k*-mers above a given threshold. To achieve a small memory footprint, Compareads uses a probabilistic data structure, based on a modified Bloom filter, which returns an over-estimation of the number of similar reads shared by the two reads data sets. If the volume of *k*-mers to consider is above a given threshold, then the computation is split in chunks and the union of the results of each computation is taken. Compareads creates large intermediate files and does redundant computation, which makes it not feasible for comparative analysis in very large metagenomic projects.

COMMET is an evolution of Compareads where each metagenomic sample is indexed only once, and such index is then used to compute the intersection with all the remaining samples. Moreover, intersections are stored as bitvectors, reducing the storage space of two orders of magnitude with respect to Compareads. Once the number of shared reads is estimated, it can be used to compute the Jaccard distance between samples.

Parallel and distributed computation

Another direction explored by comparative metagenomics is that of exploiting the additive properties of some ecological distances, which are often used to measure metagenomic samples similarity, by exploiting parallelism and/or distributed computation.

Ecological indices, such as Bray-Curtis, are originally defined over the number of species that can be found in the samples. However, in Simka [42] such distances are computed in terms of shared *k*-mers, based on the observation that *k*-mer based distances have a high correlation with taxonomic-based distances [47]. Simka achieves scalability by exploiting a parallel *k*-mer counting strategy on several samples at once. It then combines the results in a cumulative distance matrix, without storing large intermediate files.

Libra [45] performs an all-against-all comparison of metagenomes based on *k*-mer content. It employs the cosine similarity to compare samples using sequence composition and abundance, taking into account for sequencing depth. It also implements ecological indices such as Bray–Curtis and Jensen–Shannon. Libra relies on the Hadoop platform for scalability, which provides fault tolerance and simplifies the implementation in a distributed environment. It uses the *k*-mer histogram for load balancing, an inverted index to avoid storing large vectors for each sample, and it performs an aggregate distance matrix computation with a sweep line algorithm.

Dimensionality reduction

Dimensionality reduction can be used to obtain a reduced feature vector description of a metagenomic sample. Pairwise similarity between vectors representing different samples in a metagenomic experiment can then be computed and used to build a similarity matrix that describes the similarity among all pairs of samples.

Mash [41] uses the MinHash technique (local sensitive hashing) to subsample the *k*-mers from the reads of the sample. The set of sequences are reduced into small sketches, by mapping unique *k*-mers into hashes. It then computes distances among such sketches, considerably reducing the time required for distance computation. Among the several applications proposed in [41], Mash has been tested also for metagenomic samples comparison, proving to be considerably faster than several other methods. However, the loss of *k*-mer frequency information in the computation of the similarity matrix impacts on the resolution of large-scale comparative analysis. In addition, to obtain a sketch, Mash requires the extraction of all *k*-mers in the sample, which is computationally expensive.

In MetaFast [44], *de novo* partial assembly of the reads of a sample into pseudocontigs is performed first. Then, these pseudocontigs are merged into a single De Bruijn graph. The subset of *k*-mers to use as features is obtained from the segmentation of such graph into components: the relative abundance of the component is used to compute Bray–Curtis dissimilarity measure.

SAKEIMA (sampling algorithm for *k*-mers approximation) [43] is a sampling approach for approximate frequent *k*-mers counting. Rigorous provable bounds on the approximation with respect to the real counts are given by a characterization through the Vapnik-Chervonenkis (VC) dimension, a core concept from statistical learning. In Pellegrina *et al.* [43], several abundance-based ecological indices (e.g. Bray–Curtis) among metagenomic samples were considered, showing that SAKEIMA is able to compute a very close approximation of the distances, by using only a small fraction of the *k*-mers for the computation. A similar good approximation was shown also for the computation of a presence-based distance such as the widely used Jaccard distance.

HULK (histosketching using little *k*-mers) [46] uses the *k*-mer spectrum (normalized vector of *k*-mer frequencies) to represent microbiome diversity. To avoid computing and storing the *k*-mer spectrum, the histosketch data structure [48] is used instead. Such data structure maintains a set of fixed size sketches to approximate the *k*-mer spectrum taking the data from an input stream. Since the data structure is both updatable and similarity preserving, it is possible to update both the content of the data structure and approximate the similarity to other spectra. HULK includes the computation of (weighted) Jaccard, Eucledian and some ecological indices.

Metagenome-based disease status classification

After the appropriate preprocessing of the sequencing reads, metagenomics data can be summed up so to give information about the microbiota, i.e. the specific microorganisms that are found within a specific environment, or the microbiome, i.e. the collection of genomes from all the microorganisms found in a particular environment. Data matrices are organized to report all bacterial taxa present in samples. Taxa are organized with respect to a specific taxonomic order, for example genus or species or a more generic OTUs. As previously seen, shotgun metagenomics give access to additional data related to the microbiome, i.e. the collection of genomes from all the microorganisms in the environment, which can be mapped to biological functional pathways [49, 50]. The analysis of this type of data is extremely useful for identifying bacterial species that can act as biomarkers for the samples in different classes of interest. The scope of application ranges from the clinic, for example in the identification of markers for early diagnosis of colorectal cancer [51] or the identification of healthy donors for fecal transplantation [52], to the food industry, for example in the control of food safety and fermentative processes [53].

Beside statistical hypothesis testing, the use of artificial intelligence and machine learning is the method of choice when the number of samples is sufficiently high, to be able to classify samples and identify bacterial biomarkers. Moreover, the use of multivariate approaches might highlight the role of ecological interactions and niche in the sample class specification [54]. However, in order to obtain correct and robust results, it is necessary to take into account some characteristics of the data when applying machine learning approaches [55].

First, as a result of sequencing, there is a large difference, even of several orders of magnitude, in terms of the number of total species between different samples. Normalization is usually applied to NGS data to eliminate this bias between samples and make species abundances comparable. Most of the approaches in metagenomics are based on total sum scaling (divide raw counts by the total number of reads in the sample), quantile-based normalization, and normalization techniques borrowed from RNA sequencing such as edgeR [56], DESeq2 [57] and scran [58]. edgeR and DESeq2 take into account the bias introduced by the most abundant and the rarest species in the sample; in addition, methods like scran [58] also takes into account the sparsity of the data set, as in single-cell RNA sequencing. More recently, *ad hoc* normalization techniques have been implemented for microbiota data, explicitly addressing data sparsity, such as GMPR [59] (based on geometric mean of pairwise ratios) and Wrench [60] (based on an empirical Bayes normalization approach).

Indeed, data deriving from metagenomics experiments are typically very sparse due to the double sampling carried out at the steps of (i) sample collection and (ii) sequencing. This leads to data matrices with a very high percentage of zeros. The problem is that some of these zeros represent actual zero abundance of a given species in the sample, while others represent missing values. Therefore, it becomes essential to distinguish ones from the others even in light of the fact that many machine learning methods are not able to automatically manage missing values. Nevertheless, at the present, most of the metagenomics studies do not explicitly impute data before the analysis.

Finally, the measured counts are not proportional to the abundance of the species because they add up to a total (the sequencing depth of the sample). Therefore, they represent a relative abundance related to a probability of sampling. The data of this type are called compositional and are not in Euclidean space but are constrained by the simplex; therefore, standard methods such as correlation or Euclidean distance are not applicable to unpreprocessed data [61–63]. A typical solution is to use the transformations introduced by Aitchison [61] and in later works [64]. The simplest transformation is the pairwise log ratio, in which an OTU is chosen as reference and each of the other OTU counts in a sample is expressed as the log base 2 ratio with respect to the reference. As an alternative, the centered log ratio, implemented as the log ratio between OTU counts and the geometric mean of all the counts in a sample, and the isometric log ratio, defined by Egozcue *et al*. [64], can be used.

Once the preprocessing has been performed, the biological interest focuses on the identification of the biomarker species and on the ability to classify a new sample in the correct class. The 1st aim of these studies is to identify bacterial taxa that are characteristic of a sample compared to others. There are a number of methods that perform univariate statistical tests to identify bacterial biomarkers. From metastats [65], which has been a pioneering work in this field, to different approaches that couple statistical tests on differential abundant taxa with tests on biological consistency and effect size like LefSe [66], or with phylogenetic information like MEGAN [67]. More recent approaches adjust for potential confounding factors and covariates such as biotmle [68] and, in addition, account for the compositional nature and the sparsity of the data sets, such as in the case of ANCOM [62] and ALDEx2 [69]. Many classical supervised learning algorithm can be applied to this problem, from ridge regression, lasso and elastic net to support vector machines, neural networks, random forests and gradient boosting. Instead of reviewing the different supervised learning methodologies or the hundreds of application on microbiota and microbiome studies, we focus here on some open challenges related to the analysis of this kind of data. The identification of robust lists of bacterial biomarkers related to a class is an important step to identify microbial patterns associated with disease status and to develop methods for disease identification and prediction. However, very often, different studies come to different conclusions in terms of species or pathways of interest identified as marker of a class, with an overlap that is often lower than 50%. This limited overlap between different studies is attributable to a number of factors such as (1) the high technical and biological variability of the data, (2) the size of the data set (few subjects with respect to the number of analyzed variables)and (3) the heterogeneity of experimental protocols and computational pipelines used in the analysis.

As regards the technical variability, a possible approach to tackle it is to preprocess the data in order to correct for batch effect and systematic sources of variability, as explained above. On the other hand, biological variability can be taken into account by appropriate experimental design, matching samples in different classes on covariates, since it is known that the microbiota is extremely sensitive to environmental covariates. For example, in *Homo sapiens*, there is a marked sensitivity to age and nutrition. Furthermore, it is essential to address the studies with an adequate number of samples available for each class [70] and, at least in some cases, with sufficient sequencing depth and coverage of diverse genome regions to detect different species. A paradigmatic case is represented by *Prevotella copri*, a common human gut microbe that has been both positively and negatively associated with health. A recent analysis [71] revealed that *Prevotella* is probably composed of four distinct species with high variability between individuals and lower alpha diversity in western population with respect to the rest of the world. In this specific case, 16S sequencing might be insufficient to distinguish the different bacterial clades, thus resulting in apparently incoherent results of different studies.

As regards the 2nd point, there are two stability issues arising in metagenomics classification. First, since training data are often scarce, predictive models obtained from different data sets can be extremely different. A possible solution to this issue could be borrowed from other fields, in which classification methods are often used with a bootstrap Monte Carlo resampling scheme. This strategy has proved effective countermeasure against effects of unwanted selection bias and tends to stabilize the lists of selected biomarkers [72]. However, since the number of variables is generally very high and these variables interact with each other, i.e. they are often correlated or co-regulated, they can be combined in many different ways to give many possible sets of features that are equally good in terms of classification accuracy. To address this issue, additional information available on the relationships between species should be used to improve the stability of the classifiers. The basic idea of this strategy was implemented in gene expression classification [73] and could be potentially useful in metagenomics studies to take into account the complex species relationships, instead of considering them as independent features.

Finally, as part of the study of the microbiome and microbiota, there is in general a lack of assessment of the preprocessing and analysis pipelines, compared to other areas of application of the NGS, such as RNA sequencing. Given the multiple steps of the analysis, each of which can introduce bias and compromise the biological conclusions and the reproducibility of the study, it is essential to carefully document the analysis pipeline and use reliable and validated methods for analysis [74]. This last step requires, on the one hand, the possibility of using reference data sets, real or simulated, which act as a golden standard and, on the other, resort to collaborative and internationally coordinated efforts involving the entire scientific community.

Conclusion

The study of microbiomal communities has grown tremendously in recent years, thanks to the advances in shotgun sequencing that allow one to interrogate the whole microbiome in a sample. While different studies require different computational analyses, a common step in several applications is the comparison of metagenomes, either directly, by computing the distance between pairs of microbiome samples, or indirectly, by identifying features that relate to a phenotype of interest. Here, we reviewed some of the computationally challenges and the tools that have been designed to tackle such challenges. One challenge is the identification and quantification of all species in a microbial sample, for which both reference-based and reference-free methods have been designed. A 2nd challenge is the computation of metagenomic distances, which poses severe computational issues that have been tackled by several *de novo* approaches. A 3rd challenge is the identification of features that distinguish two or more classes of samples that we discussed in the context of disease status classification from metagenomes. Many of the problems that arise in metagenomic analysis may be solved in the next few years by methodological or technological advances (e.g. long-read sequencing technologies such as Pacific Biosciences and Oxford Nanopore Technologies), but the awareness of the issues arising in the comparison of metagenomic samples is a requirement to avoid critical errors in the analysis of large and complex metagenomic data sets.

Key Points

- The comparison of microbiome samples is crucial in several microbial studies, requiring advanced computational tools for critical steps. These tools are either reference based, comparing reads in a data set with reference genomes in a database, or reference free, which consider all reads in the data set.
- The identification and quantification of the species in the sample is a critical step for techniques based on the similarity between the number or the proportion of species in samples.
- *De novo* methods for computing distances between metagenomic samples have to overcome severe computational issues posed by the large sizes of next-generation sequencing data sets but are not subject to the biases in the reference sequence resources.
- Applications include the identification of similarities and differences between microbial communities in different habitats and the prediction of disease status from metagenomic information.

Acknowledgments

We thank Stefano Campanaro for providing feedback on the manuscript.

Funding

University of Padova (grants SID2017 and STARS: Algorithms for Inferential Data Mining).

References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016; **17**(6): 333–351.
2. Scholz M, Ward DV, Pasolli E, *et al*. Strain-level microbial epidemiology and population genomics from shotgun metagenomics. *Nat Methods* 2016; **13**(5): 435.
3. Turnbaugh PJ, Ley RE, Hamady M, *et al*. The human microbiome project. *Nature* 2007; **449**(7164): 804–10.
4. Integrative HMP iHMP Research Network Consortium. The integrative human microbiome project *Nature* 2019; **569**: 641–8.
5. Quince C, Walker AW, Simpson JT, *et al*. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol* 2017; **35**(9): 833.
6. Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* 2017; **20**(4): 1125–36.
7. Chiu CY, Miller SA. Clinical metagenomics. *Nat Rev Genet* 2019; **20**:341–55.
8. Simon HY, Siddle KJ, Park DJ, *et al*. Benchmarking metagenomics tools for taxonomic classification. *Cell* 2019; **178**(4): 779–94.
9. Sczyrba A, Hofmann P, Belmann P, *et al*. Critical assessment of metagenome interpretation-a benchmark of metagenomics software. *Nat Methods* 2017; **14**:1063–71.
10. Huerta-Cepas J, Szklarczyk D, Forslund K, *et al*. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res* 2016; **44**(D1): D286–293.
11. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990; **215**(3): 403–10.

12. Zhang Z, Schwartz S, Wagner L, *et al*. A greedy algorithm for aligning dna sequences. *J Comput Biol* 2004; **7**(1–2): 203–14.
13. Huson DH, Auch AF, Qi J, *et al*. Megan analysis of metagenomic data. *Genome Res* 2007; **17**: 377–86.
14. Segata N, Waldron L, Ballarini A, *et al*. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 2012; **9**.
15. Darling AE, Jospin G, Lowe E, *et al*. Phylosift: phylogenetic analysis of genomes and metagenomes. *PeerJ* 2014; **2**(e243): 1–28.
16. Lindgreen S, Adair KL, Gardner P. An evaluation of the accuracy and speed of metagenome analysis tools. *Scientific reports* 2016; **6**(19233): 1–14.
17. Wood D, Salzberg S. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014; **15**: 1–12.
18. Ounit R, Wanamaker S, Close TJ, *et al*. Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics* 2015; **16**(1): 1–13.
19. Girotto S, Comin M, Pizzi C. Higher recall in metagenomic sequence classification exploiting overlapping reads. *BMC Genomics* 2017; **18**(10): 917.
20. Kim D, Song L, Breitwieser F, *et al*. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* 2016; **26**: gr.210641.116.
21. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* 2019; **20**: 257, 1–13.
22. Qian J, Marchiori D, Comin M. Fast and sensitive classification of short metagenomic reads with skraken. In: Peixoto N, Silveira M, Ali HH, *et al*. (eds). *Biomedical Engineering Systems and Technologies*. Cham: Springer International Publishing, 2018, 212–226.
23. Binda K, Sykulski M, Kucherov G. Spaced seeds improve k-mer-based metagenomic classification. *Bioinformatics* 2015; **31**(22): 3584.
24. Eisen JA. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. *PLoS Biol* 2007; **5**: 384–88.
25. Van Vinh L, Van Lang T, Binh LT, *et al*. A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads. *Algorithm Mol Biol* 2015; **10**(1): 1–12.
26. Wang Y, Leung HC, Yiu SM, *et al*. Metacluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. *Bioinformatics* 2012; **28**: i356–i62.
27. Wu YW, Ye Y. A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *J Comput Biol* 2011; **18**: 523–34.
28. Girotto S, Pizzi C, Comin M. Metaprob: accurate metagenomic reads binning based on probabilistic sequence signatures. *Bioinformatics* 2016; **32**(17): i567–75.
29. Alneberg J, Bjarnason BS, de Bruijn I, *et al*. Binning metagenomic contigs by coverage and composition. *Nat Methods* 2014; **11**:1144–6.
30. Bowers RM, Clum A, Tice H, *et al*. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics* 2015; **16**(1): 856.
31. Kislyuk A, Bhatnagar S, Dushoff J, *et al*. Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics* 2009; **10**: 1–16.
32. Kelley DR, Salzberg SL. Clustering metagenomic sequences with interpolated markov models. *BMC Bioinformatics* 2010; **11**: 1–12.
33. 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): 1489–1495.
34. Wu Y-W, Simmons BA, Singer SW. Maxbin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 2016; **32**(4): 605–7.
35. Imelfort M, Parks p, Woodcroft BJ, *et al*. Groopm: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* 2014; **2**: 1–16.
36. Kang DD, Li F, Kirton E, *et al*. Metabat 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 2019; **7**: e7359.
37. Qian J, Comin M. Metacon: unsupervised clustering of metagenomic contigs with probabilistic k-mers statistics and coverage. *BMC Bioinformatics* **20**(9): 367.
38. Huttenhower C, Gevers D, Knight R, *et al*. Structure, function and diversity of the healthy human microbiome. *Nature* 2012; **486**(7402): 207.
39. Maillet N, Lemaitre C, Chikhi R, *et al*. Compareads: comparing huge metagenomic experiments. *BMC Bioinformatics* 2012; **13**(S10): 1–10.
40. Maillet N, Collet G, Vannier T, *et al*. Commet: Comparing and combining multiple metagenomic datasets. In: *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2014, pp. 94–98
41. Ondov BD, Treangen TJ, Melsted P, *et al*. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biol* 2016; **17**(1): 132.
42. Benoit G, Peterlongo P, Mariadassou M, *et al*. Multiple comparative metagenomics using multiset k-mer counting. *PeerJ Comput Sci* 2016; **2**:e94.
43. Pellegrina L, Pizzi C, Vandin F. Fast approximation of frequent k-mers and applications to metagenomics. *J Comput Biol* 2020; **27**(4): 534–49.
44. Ulyantsev VI, Kazakov SV, Dubinkina VB, *et al*. MetaFast: fast reference-free graph-based comparison of shotgun metagenomic data. *Bioinformatics* 2016; **32**(18): 2760–2767.
45. Choi I, Ponsero AJ, Bomhoff M, *et al*. Libra: scalable k-mer-based tool for massive all-vs-all metagenome comparisons. *GigaScience* 2018; **8**(2): giy165.
46. Rowe WPM, Carrieri AP, Alcon-Giner C, *et al*. Streaming histogram sketching for rapid microbiome analytics. *Microbiome* 2019; **7**(1): 40.
47. Dubinkina VB, Ischenko DS, Ulyantsev VI, *et al*. Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. *BMC Bioinformatics* 2016; **17**(1): 38.
48. Yang D, Li B, Rettig L, *et al*. Histosketch: fast similarity-preserving sketching of streaming histograms with concept drift. In: *2017 IEEE International Conference on Data Mining (ICDM)*, 2017, 545–554.
49. Franzosa EA, McIver LJ, Rahnavard G, *et al*. Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* 2018; **15**(11): 962.
50. McIver LJ, Abu-Ali G, Franzosa EA, *et al*. Biobakery: a meta'omic analysis environment. *Bioinformatics* 2017; **34**(7): 1235–7.
51. Dai Z, Coker OO, Nakatsu G, *et al*. Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* 2018; **6**(1): 70.
52. Duvallet C, Zellmer C, Panchal P, *et al*. Framework for rational donor selection in fecal microbiota transplant clinical trials. *PloS One* 2019; **14**(10): e0222881.

53. Alkema W, Boekhorst J, Wels M, *et al*. Microbial bioinformatics for food safety and production. *Brief Bioinform* 2015; **17**(2): 283–92.
54. Zhou Y-H, Gallins P. A review and tutorial of machine learning methods for microbiome host trait prediction. *Front Genet* 2019; **10**:579.
55. Weiss S, Xu ZZ, Peddada S, *et al*. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 2017; **5**(1): 27.
56. Robinson MD, McCarthy DJ, Smyth GK. Edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; **26**(1): 139–40.
57. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol* 2014; **15**(12): 550.
58. Bacher R, Chu L-F, Leng N, *et al*. Scnorm: robust normalization of single-cell rna-seq data. *Nat Methods* 2017; **14**(6): 584.
59. Chen L, Reeve J, Zhang L, *et al*. Gmpr: a robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* 2018; **6**:e4600.
60. Kumar MS, Slud EV, Okrah K, *et al*. Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics* 2018; **19**(1): 799.
61. Aitchison J. The statistical analysis of compositional data. *J R Stat Soc B Methodol* 1982; **44**(2): 139–60.
62. Mandal S, Van Treuren W, White RA, *et al*. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* 2015; **26**(1): 27663.
63. Patuzzi I, Baruzzo G, Losasso C, *et al*. Metasparsim: a 16s rrna gene sequencing count data simulator. *BMC Bioinformatics* 2019: 1–13.
64. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, *et al*. Isometric logratio transformations for compositional data analysis. *Math Geol* 2003; **35**(3): 279–300.
65. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput Biol* 2009; **5**(4): 1–11.
66. Segata N, Izard J, Waldron L, *et al*. Metagenomic biomarker discovery and explanation. *Genome Biol* 2011; **12**(6): R60.
67. Mitra S, Gilbert JA, Field D, *et al*. Comparison of multiple metagenomes using phylogenetic networks based on ecological indices. *ISME J* 2010; **4**(10): 1236–42.
68. Hejazi NS, van der Laan MJ, Hubbard AE. A generalization of moderated statistics to data adaptive semiparametric estimation in high-dimensional biology. preprint arXiv:1710.05451, 2017.
69. Fernandes AD, Macklaim JM, Linn TG, *et al*. Anova-like differential expression (aldex) analysis for mixed population rna-seq. *PLoS One* 2013; **8**(7): 1–15.
70. Thomas AM, Manghi P, Asnicar F, *et al*. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nat Med* 2019; **25**(4): 667.
71. Tett A, Huang KD, Asnicar F, *et al*. The prevotella copri complex comprises four distinct clades underrepresented in westernized populations. *Cell Host Microbe* 2019; **26**(5): 666–79.
72. Di Camillo B, Sanavia T, Martini M, *et al*. Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment. *PLoS One* 2012; **7**(3): e32200.
73. Sanavia T, Aiolli F, Martino GDS, *et al*. Improving biomarker list stability by integration of biological knowledge in the learning process. *BMC Bioinformatics* 2012; **13**(4): S22.
74. Poussin C, Sierro N, Boué S, *et al*. Interrogating the microbiome: experimental and computational considerations in support of study reproducibility. *Drug Discov Today* 2018; **23**(9): 1644–57.