Metagenomic Tools for Taxonomic and Functional Annotation

Chapter · November 2021

DOI: 10.1201/9781003042570-4

7 authors, including:

Yordanis Pérez Llano
Universidad Nacional Autónoma de México
15 PUBLICATIONS **114** CITATIONS
SEE PROFILE

Ramón Batista García
Universidad Autónoma del Estado de Morelos
68 PUBLICATIONS **565** CITATIONS
SEE PROFILE

María del Rayo Sánchez
Universidad Autónoma del Estado de Morelos
52 PUBLICATIONS **902** CITATIONS
SEE PROFILE

Jorge Luis Folch-Mallol
Universidad Autónoma del Estado de Morelos
116 PUBLICATIONS **1,861** CITATIONS
SEE PROFILE

Section II

Metagenomics Tools to Access Microbial Diversity

2

Metagenomic Tools for Taxonomic and Functional Annotation

Yordanis Pérez-Llano, Ramon Alberto Batista-García, María del Rayo Sánchez-Carbente, Jorge Luis Folch-Mallol
Universidad Autónoma del Estado de Morelos, México
Sara Cuadros-Orellana
Universidad Católica del Maule, Chile
Alma Delia Nicolás-Morales, Natividad Castro-Alarcón,
Universidad Autónoma de Guerrero, México

CONTENTS

2.1 Introduction 21
2.2 General Methods for Taxonomic Annotation of Sequencing Data 22
2.3 Benchmarking Studies of Metagenomic Annotation Software 26
2.4 Tools for Taxonomic Annotation 26
2.5 Databases for Taxonomic Annotation 29
2.6 Microbiome Functional Inference from Community Structure 31
2.7 Functional Annotation of Shotgun Metagenome Data 32
2.8 Main Databases Used for Functional Annotation 34
2.9 Visualization of Metagenomic Data Annotation 34
2.10 Challenges and Future Perspectives 36

2.1 Introduction

The rise of next-generation sequencing (NGS) technologies and their application to describe microbial communities have reshaped our understanding of biology. Metagenomics allowed the uncovering of a microscopic universe that is not only part of our environment but also a key player in ecosystem interactions that were until recently elusive. Microbes are involved in processes ranging from worldwide geochemical carbon cycling to changing human physiology and behavior. Metagenomic analysis of environments such as deep-sea or extreme ecosystems has significantly expanded and restructured the tree of life, rewriting our current knowledge about the evolution of life on Earth (Parks et al. 2019). Though it is a powerful technology, it is not free of pitfalls and limitations that arise both from the technical processing of samples and the computational analysis of the large amount of data it generates.

Traditionally, metagenomics comprises a group of analytical and microbiological techniques that allow the identification of the microorganisms present in an environment and their relative abundance (Fosso et al. 2018). Although functional screening of environmental DNA libraries has also been considered as a metagenomic approach, currently the term is preferentially applied to the analysis of datasets obtained from next-generation sequencing platforms. There are two major technological approaches to obtaining metagenomic data: amplicon-based sequencing (or metabarcoding) and whole-genome shotgun (WGS) sequencing. In the former, the DNA extracted from the biological sample is used as a

DOI: 10.1201/9781003042570-4

template for the amplification of marker genes. The amplicons are then sequenced by NGS technology, allowing, for example, identification of the individual members of a microbial community by phylogeny-based approaches. Shotgun metagenomics, on the contrary, relies on the untargeted sequencing of all available genome fragments after DNA extraction from the biological sample. A pivotal procedure in understanding the type of data generated by both technologies is sequence annotation. In the field of 'omics' bioinformatics, annotation is the process of finding biologically relevant features to genomic elements, and it consists of gene prediction and taxonomy or function assignment.

Other steps, such as sequence quality control and assembly, usually precede the annotation of genes in genomes and metagenomes. Analyzing microbial community datasets follows the same general principles of genome analysis. For instance, sequence assembly (often described with the analogy of putting together the pieces of a puzzle) is a general process of concatenation of overlapping fragmented sequences. For some time, it was intuitive to think that taxonomic and functional assignment profited from having the full gene or even several contiguous genes (as in contigs) rather than just short reads. However, most current methods are assembly indifferent and instead rely on the direct annotation of raw reads.

In this chapter, we intend to cover recent advances in the annotation of metagenomic sequencing data. The aim is not to present an exhaustive description of all available tools, as this is a very dynamic field. We also do not intend to make tool recommendations, as the best tool to use probably depends on the sample and data characteristics. Instead, this chapter aims to review some of the general strategies and methods used for metagenome taxonomic and functional annotation and their strengths and limitations and refer our readers to recently published benchmarking studies for the selection of tools that are appropriate for their research purpose.

2.2 General Methods for Taxonomic Annotation of Sequencing Data

Taxonomic annotation methods essentially aim to solve a classification problem in which the raw or processed sequencing data must be categorized using a reference taxonomy. This generally implies that the elements of the query set (sample data) must be matched to labeled elements on a reference database. Although it might seem simple, it becomes a cumbersome task because the reference databases are generally incomplete and contain sparse or even skewed data (Balvočiute and Huson 2017). Another particularity of this classification problem is that the biological taxonomy system follows a hierarchical structure by which the organisms are grouped into different ranks or taxa. Hence, the classification can be performed at different depths on the hierarchy, and in doing so, the classification will be more accurate for high-ranking taxa (e.g., kingdom or phylum) than for low-ranking ones (e.g., genus or species). To further complicate the scenario, the discovery of new microorganisms and the generation of new phylogenomic data impose changes in the reference taxonomy. This often leads to specimens being renamed to accommodate the new taxonomic system. The available databases should be updated every time this renaming occurs, but in practice, this is generally not the case. Therefore, achieving a highly confident annotation depends on the target taxonomy rank and relies on the matching algorithm and the reference database quality.

Figure 2.1 shows the common annotation pipelines that are followed for amplicon-based and shotgun sequencing data from metagenomic samples. The raw metagenomic data consist of many relatively short DNA fragment sequences, called reads. Whether these reads come from a metabarcoding or a WGS experiment, the annotation can be performed on the raw sequence reads or assembled contigs. The assembly step is often resource intensive, especially in terms of memory usage, though modern assemblers have somewhat reduced this limitation. Assembly tools use various algorithms and heuristics that may produce different results and thus directly affect taxonomy annotation. Probably due to these issues, assembly is not always performed, and alternatively, the taxonomic annotation is obtained directly from the raw reads, especially if the aim is to characterize the taxonomic composition of the microbiota (Tamames, Cobo-Simón, and Puente-Sánchez 2019). For some other purposes, such as functional annotation or genome reconstruction, assembly could be a required step.

Another common stage in data processing is the clustering of elements of the same apparent genomic origin. In the case of amplicon-based sequencing, the goal is to identify either the reads derived from the

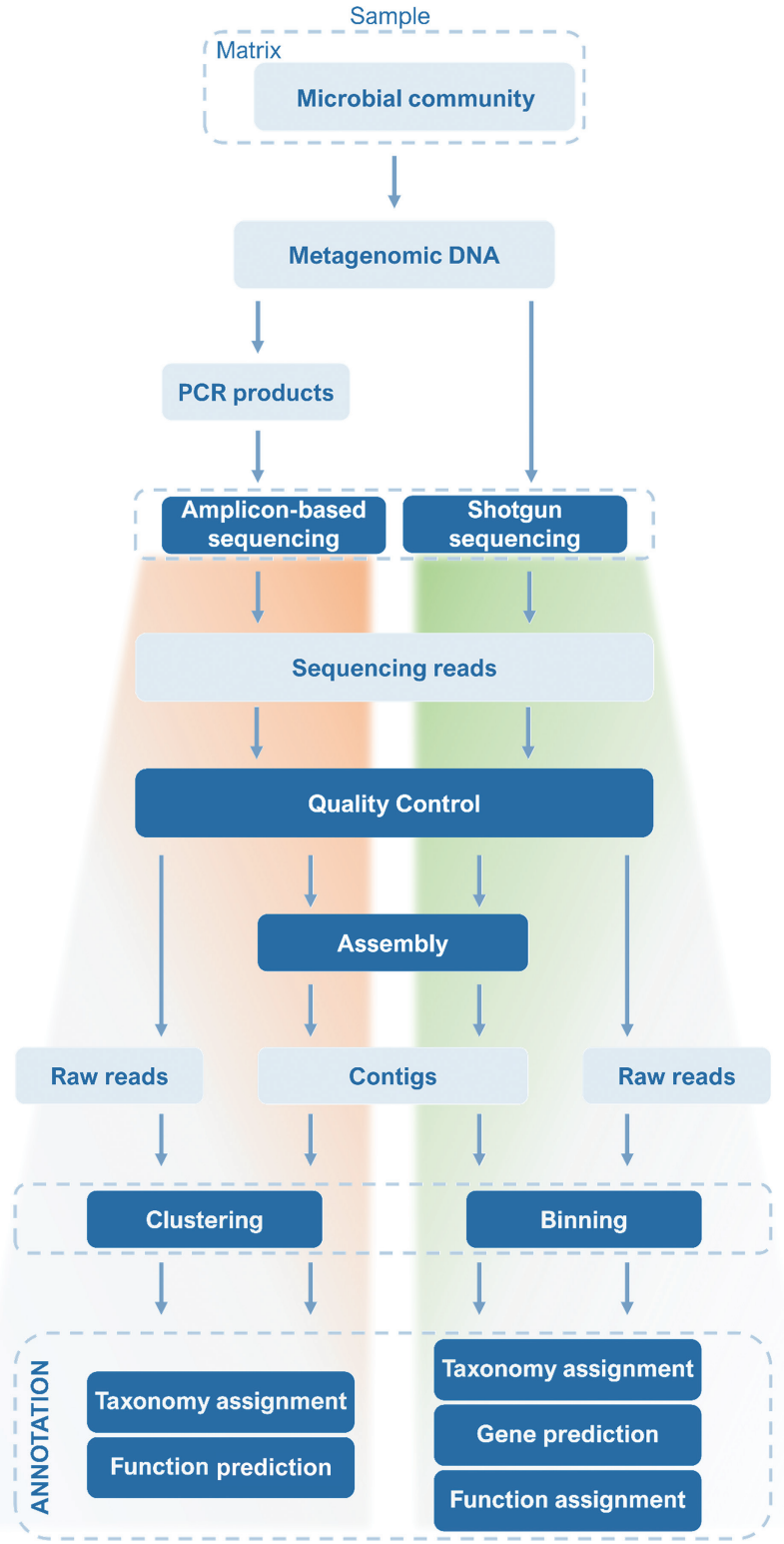

FIGURE 2.1 General workflows of metagenome taxonomic annotation of amplicon-based sequencing and WGS metagenomics.

same amplicon or from different amplicons. This identification is hampered by sequencing errors, which can be misinterpreted as biological variation. To overcome this sequencing platform limitation, assembled contigs are clustered if they have up to 3% sequence divergence (calculated for 16S rRNA variable region fragments), aiming to cancel the effect of errors (Edgar 2018b). These clusters of sequences are known as operational taxonomic units (OTUs) and are represented by only one sequence in the ensuing annotation and post-processing steps. The 'operational' part in the term emphasizes the notion that these clusters are not an accurate representation of the biological variation in the sample. Current evidence coming from assembled genomes has shown that the initial 97% threshold does not adequately capture species variation, as this must be above the 99% identity threshold for the entire 16S rRNA gene to achieve accurate species classification (Edgar 2018b). More recent methods (e.g., DADA2, Callahan et al. 2016; Deblur, Amir et al. 2017) take into consideration the lower error rates in current sequencing platforms and perform a clustering (sometimes referred to as denoising) that aims to differentiate technical variation from biological variation. The clusters generated by these methods, known as amplicon sequence variants (ASVs), tend to have less sequence divergence than OTUs.

The equivalent clustering step for WGS data is known as binning, that is, separating reads or contigs into bins that should ideally contain all the elements derived from a single genome. Instead, in practice, the bins usually contain sequences from closely related strains or species within a community. Binning might serve various purposes, such as genome reconstruction, functional classification of microbial entities, and microbial abundance estimation. Binning methods are divided into the ones that are taxonomy independent, that is, that use alignment and compositional metrics to gather all elements with shared characteristics, and taxonomy dependent, which first perform read taxonomy assignment and then cluster sequences based on taxonomic proximity or identity (Breitwieser, Lu, and Salzberg 2018).

As initially discussed, taxonomic assignment or classification requires two general steps: 1) a matching step that implies that the reads, contigs, k-mers, or any other structure generated from these must be matched to elements on a reference database and 2) the label assignment method based on the matching results. For some methods, discerning between the two processes might be challenging, sometimes because of the complexity of the methods or because both processes are not delineated in the method description provided by the authors. In any case, classifying methods according to their internal logic or functioning does not aim to generate superfluous classes but instead help us determine if a group of methods is the best performer or suffers from a specific type of errors/characteristics (e.g., high false positive rate, low recall, high resource consumption).

Although different classifications are used in the literature (Escobar-Zepeda et al. 2018; Breitwieser, Lu, and Salzberg 2018; Ye et al. 2019; Hleap et al. 2020), here we consider that the matching algorithms are based on three different strategies: sequence alignment, pseudoalignment, and composition. Figure 2.2 shows the programs that use each of these strategies during matching.

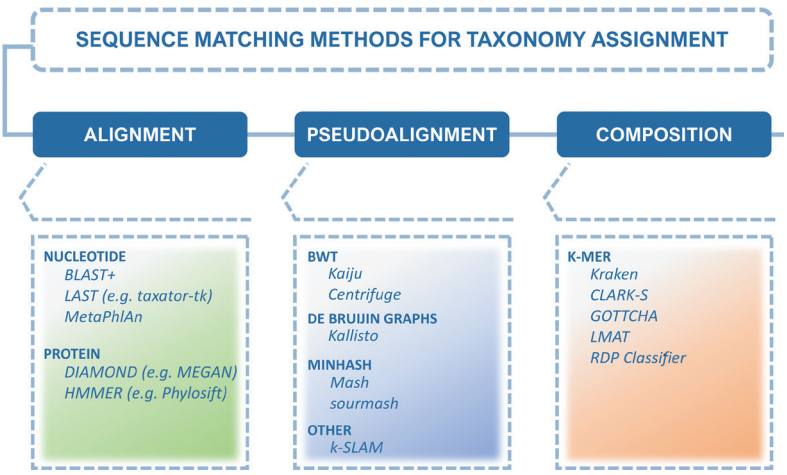

FIGURE 2.2 Types of sequence-matching algorithms used by different tools for metagenome taxonomy assignment.

Alignment-based matching algorithms perform local or global alignment of reads to a reference database. The alignment can be performed using nucleotide sequences against nucleotide databases or translated sequences against protein databases. BLAST is among the best performers in terms of accuracy and is the best-known alignment-based algorithm. The major drawback of these methods is that the alignment is resource intensive, so all the tools that rely on them are burdened with high memory and time consumption.

Once the alignment is performed, taxonomy assignment occurs by several methods. In the case of BLAST (or MegaBLAST, the most-used version in metagenomics), assignment is achieved by the best scoring hit of the sequence search. Other methods like MEGAN, which also uses BLAST, perform the assignment by a method called lowest common ancestor (LCA) (Huson et al. 2011). The LCA taxonomy assignment method is currently used by several tools, independent of what matching algorithm they use. The idea behind this method is to assign the most confident label based on annotations placed on distinct levels of the taxonomical hierarchy. In the case of MEGAN, the alignment of sequences is performed against various databases, and the taxonomic label is that of the lower-ranked taxa that unambiguously accommodates the annotations obtained from all databases (Huson et al. 2011). A third algorithm (or metric) to assign taxonomic labels is the average nucleotide identity (ANI), a measure of overall genome relatedness that is inspired in the DNA-to-DNA hybridization technique. As ANI is more related to complete genomes, its application is focused on the taxonomic annotation of metagenome-assembled genomes (MAGs) or metagenomic bins. A recent implementation of this algorithm is the tool OrthoANI, in which the ANI is calculated from the BLAST alignment of orthologous gene sequences from different genomes (Lee et al. 2016). The currently accepted boundary that delineates species is around 95–96% ANI. This method performs well for species classification where closely related species are available in the database, whereas it underperforms when comparing genomes from different genera (Lee et al. 2016; Yoon et al. 2017; Jain et al. 2018).

The pseudo-alignment term was used by the authors of kallisto to describe their method of matching reads to genomes based on k-mers represented in de Bruijn graphs (Yi et al. 2018; Bray et al. 2016). Here we chose the same term in a different context to include all the methods that perform taxonomic classification based on an alignment conducted in a space other than the sequence space of the original reads or their direct translation into peptides. In this group of methods, the reads or contigs are generally transformed into a different space using some form of convolution process, achieving a new way to represent the sequences. These methods aim to take advantage of the properties of the new space to optimize the matching step by one parameter (e.g., speed, RAM requirement, database storage, etc.). The most-used method for sequence transformation is the Burrow-Wheelers transform (BWT), a compression system that speeds up the matching to a reference database and at the same time reduces the database storage space. Some tools that use BWT in their algorithm are Kaiju (Menzel, Ng, and Krogh 2016) and Centrifuge (D. Kim et al. 2016).

Sequence composition methods are generally based on k-mer composition, although some methods might use k-mers to perform alignment or pseudoalignment as a primary matching algorithm, such as in the case of kallisto. In k-mer composition algorithms, the reads or contigs are split into small fragments of defined length (called k-mers) that are mapped by alignment to a reference database. The matching is performed by assessing the k-mer composition of the query sequence against the defined k-mer composition in the database. In some tools, this matching is not performed explicitly, as the taxonomy assignment occurs by using this composition information instead. For example, tools like the RDP classifier use probabilistic methods (in this case, naïve Bayes classifier) to determine the probability that a query sequence belongs to a set of known sequences within a genus (Wang et al. 2007; Bacci et al. 2015).

The current database growth due to metagenome-derived genomes, particularly in the case of RefSeq, has led to an expansion in species and genera that has reshaped taxonomical structures (Nasko et al. 2018). One direct repercussion of this growth is the reduction in the species classification accuracy of k-mer based methods, especially when several closely related genomes are found in the database (Nasko et al. 2018). Another confounding factor for species classification is the poorly defined species boundaries in taxa where high horizontal gene transfers occur (Nasko et al. 2018). This implies that, in the foreseeable future, k-mer–based methods will be increasingly prone to false positives for the classification of

lower-ranked taxa (e.g., species or strain), so other alternative methods (even if more resource intensive) should be adopted.

2.3 Benchmarking Studies of Metagenomic Annotation Software

The pursuit of optimal methods for metagenomic analysis has resulted in an explosion of tools, all oriented to accomplish the same set of tasks but pledging to outperform their predecessors by some metric (Marx 2020). New tools are generally evaluated within a specific context or with a particular data type or application in mind. Assessing whether these tools can be applied broadly and how they compare to alternative software is generally carried out in so-called benchmarking studies. When conducted by relatively neutral authors (e.g., they are not involved in the development of tools evaluated in the study) and using appropriate test data and metrics, these studies are of vital importance to guide software users in the selection of the most accurate and/or fast tools for a specific task.

Referring to benchmarking studies for pipeline selection should be a common practice for software users and bioinformaticians, and we highly encourage it. A detailed discussion of the many reasons used by researchers to select bioinformatic tools can be found in (Gardner et al. 2017). Among these reasons, we can find the notions that:

1. Recently published software should be better or have improvements over older software.
2. Highly cited tools are widely accepted by the community (including potential reviewers of your work) and therefore more desirable.
3. The reputation of the authors or the journal is a guarantee of the quality of the tool.
4. Software tools often trade accuracy for speed, and therefore slower software should be more accurate.

Unfortunately, researchers might select tools only because they are user friendly or extensively documented. By correlating software accuracy with speed, age (i.e., to test the effect of recency), citation number (to test the wide use of the tools), and commonly accepted merit indexes such as journal impact or author reputation (H-index), Gardner et al. (2017) demonstrated that these metrics are not reliable predictors of accuracy. Basing research decisions on these widely accepted preconceptions could affect the reliability of our results and should therefore be avoided.

Systematic and standardized guidelines for benchmarking omics tools have recently been proposed (Mangul et al. 2019). Ensuring a scientifically rigorous comparison of different tools requires a gold standard dataset that serves as ground truth and a general set of metrics that can score the performance of any given method. Thus, as gold standard tools are not yet available for many bioinformatics applications, resorting to benchmarking studies is highly recommended for the selection of appropriate tools that fit our experimental setup and data.

Several benchmark studies assess the taxonomic annotation accuracy and recall of bioinformatics tools or pipelines (Bazinet and Cummings 2012; Peabody et al. 2015; Lindgreen, Adair, and Gardner 2016; Siegwald et al. 2017; McIntyre et al. 2017; Sczyrba et al. 2017; Almeida et al. 2018; Escobar-Zepeda et al. 2018; Gardner et al. 2019; Ye et al. 2019; Velsko et al. 2018). Some of these also evaluate the effect of databases on classification accuracy and recall (Escobar-Zepeda et al. 2018; Velsko et al. 2018). There are even benchmarking studies oriented to evaluate tools for specific tasks or datatypes, such as in rumen microbiome analysis (López-García et al. 2018) or ancient microbiome samples (Velsko et al. 2018). Some of these studies will be covered further in this chapter.

2.4 Tools for Taxonomic Annotation

The study of the microbial diversity of different natural environments can be carried out by targeting specific genes that can inform us about the taxonomic composition of the ecosystem. The 16S

ribosomal RNA (rRNA) gene is commonly used for diversity analysis of Bacteria and Archaea communities, as it is present in all members of both domains and contains enough variability (up to nine hypervariable regions) to make it a very reliable marker for genus and species identification (Yang, Wang, and Qian 2016). For eukaryotes, the internal transcribed spacer (ITS), the 18S rRNA, and 28S rRNA genes are used.

Although marker-based metagenomic sequencing is still a standard procedure for the taxonomical description of microbial communities, various studies have shown that this technique has biases that can be alleviated by WGS sequencing (Khachatryan et al. 2020). Although WGS experiments are considerably more expensive and require computationally intensive analyses, higher accuracy in terms of identified taxa and abundance estimation is achieved by this methodology (Khachatryan et al. 2020). An affordable yet more accurate alternative is to perform experiments with little sequencing effort (approximately 0.5 million sequences for gut microbiomes), which are known as shallow sequencing, and that have been shown to achieve similar diversity estimation to ultradeep WGS at a price comparable to marker-based sequencing experiments (Hillmann et al. 2018). Shallow- or moderate-depth shotgun sequencing may be used by researchers to obtain species-level taxonomic and functional data at approximately the same cost as amplicon sequencing.

With the increasing size of metagenomic projects, biological databases are also growing exponentially in size. Under this scenario, the computational cost of sequence alignment needs to be considered. The development of faster tools, such as UBLAST and USEARCH (Edgar 2010), LAST (Kiełbasa et al. 2011), RAPSearch2 (Zhao, Tang, and Ye 2012), and DIAMOND (Buchfink, Xie, and Huson 2015), to cite a few, represents a new trend in bioinformatics.

The USEARCH algorithm is mostly used for high-identity searches. Conversely, UBLAST is used to search for more divergent protein or translated nucleotide sequences. LAST improves seed-and-extend heuristic methods by using an adaptive seed. RAPSearch (reduced alphabet based protein similarity search) uses an optimized suffix array data structure to accelerate the identification of alignment seeds, together with multi-thread modes (available in RAPSearch2 only) to further speed the process.

Among these tools, DIAMOND is probably the fastest one. It aligns short reads to the complete National Center for Biotechnology Information (NCBI)'s nr protein database. For this alignment, both the database and the query sequences are indexed, making the matching process more computationally feasible (Buchfink, Xie, and Huson 2015). Like BLASTX, DIAMOND is an 'all mapper' that attempts to exhaustively determine all significant alignments for the query (Buchfink, Xie, and Huson 2015). Nevertheless, it outperforms BLASTX by an impressive 20,000 times on short reads while maintaining a similar degree of sensitivity.

For the alignment of long reads, the most-used method is LAST, which allows a quick and sensitive comparison of sequences with an arbitrarily non-uniform composition. The LAST aligner is tolerant to single-base insertions and deletions, therefore outperforming non-gapped aligners (Kiełbasa et al. 2011).

Kraken uses the k-mer composition and LCA methods to perform taxonomy classification. Initially, a reference database is constructed by finding unique k-mers within genomic sequences and assigning a taxid (taxonomy label) based on the lowest common ancestor that shares that k-mer.

A meta-analysis of classifier benchmarking studies revealed that, when performed without strict guidelines, these studies can lead to contradictory results (Gardner et al. 2019). Part of the findings of this meta-analysis are shown in Figure 2.3. The first issue raised by the authors is that many of the tools available for taxonomic classification have not been independently evaluated (see Figure 2.3a). The reasons for this are many, but the authors underscore as more likely causes that the unevaluated tools are the most recently published, are no longer available/functional, or provide results that are not suited for comparison (Gardner et al. 2019). One other cause could be the popularity of the tools, but among the unevaluated set are some highly cited tools (e.g., UPARSE, PathSeq, Phylosift), which can be identified by the black labels in Figure 2.3b. In any case, this hidden group may contain promising tools that did not make it to the spotlight.

A more concerning observation in (Gardner et al. 2019) was that the number of citations per year does not correlate with the tool accuracy estimates, with less accurate tools being highly cited (see Figure 2.3c). To normalize comparison across different studies, these authors transformed the reported F-measures (a measure of accuracy) into a Z-score metric that eliminated the methodological and data

differences. According to this metric, less accurate tools score negatively, while a positive score is given to those tools that consistently have a higher F-measure. As can be observed in Figure 2.3c, the classifiers in tools such as QIIME and mothur are highly used but are significantly less accurate than other tools such as Kraken, NBC, CLARK, or OneCodex (Gardner et al. 2019).

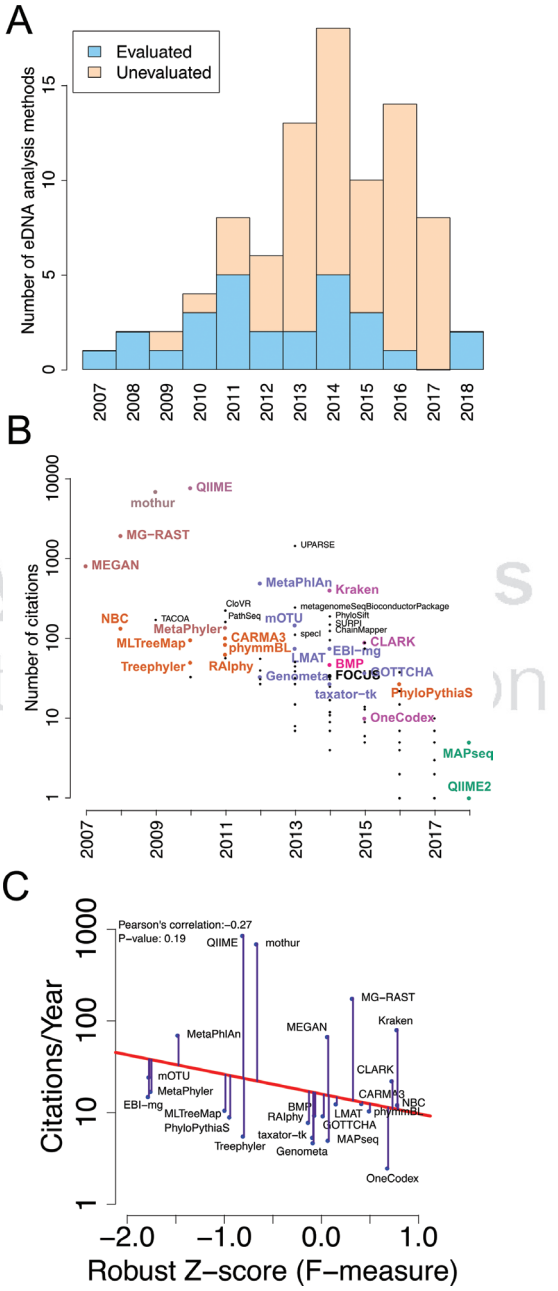

FIGURE 2.3 Analysis of software tool benchmarking studies, publication date, and citations as part of a meta-analysis of software accuracy conducted by Gardner et al. (2019). The results show (a) many of the available tools have not been independently evaluated as part of a benchmark study; (b) the number of citations for each software tool versus the year it was published, where tools that have been evaluated are colored and labeled, while highly cited tools that have not been independently evaluated are labeled in black, and (c) the number of citations per year does not correlate with the tool accuracy estimates, with less accurate tools being highly cited. (Modified from Gardner et al.[2019] with permission).

A recent comprehensive benchmarking of metagenomic sequence classifiers was conducted to assess the species-level classification precision and recall of 20 different tools using a uniform database (Ye et al. 2019). This study considers tools that were not previously evaluated, such as PathSeq, MMseqs2, taxMaps, and mOTUs2. The authors of this study make a distinction of methods according to the type of sequence they handle during the matching algorithm. Hence, they assess DNA-to-DNA classifiers such as Kraken, k-SLAM, MegaBLAST, metaOthello, CLARK, GOTTCHA, taxMap, prophyle, PathSeq, Centrifuge, and Karp. DIAMOND, Kaiju, and MMseqs2 were evaluated as DNA-to-protein classifiers, and marker-based classifiers were MetaPhlAn2 and mOTUs2. All the tools were executed with default parameters as would be used by the average user of these packages. The precision of the abundance estimation of each of the methods was also assessed. Bracken, one of the tested tools, is an add-on to Kraken that allows more accurate abundance estimation, which is originally not performed by Kraken. According to their results, and in agreement with previous findings, the marker-based methods significantly underperformed compared to the shotgun-based methods. From the rest of the tested methods, Centrifuge, MegaBLAST, PathSeq, prophyle, DIAMOND, Kaiju, and MMseqs2 performed poorly in one or several metrics compared to Kraken, Kraken2, KrakenUniq, Bracken, CLARK, CLARK-S, k-SLAM, and taxMaps. From the latter methods, given that precision, recall, and abundance estimation are fairly similar, the selection for the optimal tool relies on the computing performance (i.e., CPU and memory requirements, execution time). For this reason, the combination of Kraken2/Bracken seems to be the most attractive among the tested classifiers, with execution time around 1 minute and memory consumption of ~36 Gb for Kraken2 and less than 1 Gb for Bracken (Ye et al. 2019).

As new tools or improvements on existing tools are likely to emerge in the future and no standardize pipeline applicable to most case scenarios can be envisioned, benchmarking studies will remain necessary. The Live Evaluation of Computational Methods for the Metagenome Investigation (LEMMI) tool has deployed a platform for the automatic benchmarking of taxonomic classification software (Seppey, Manni, and Zdobnov 2020). Results obtained from the assessment of the more popular tools can be obtained from the site. But, more importantly, users can set up evaluations of new algorithms with general or custom datasets. This platform will streamline future benchmarking studies and serves as an initiative to standardize research within the metagenomic field.

2.5 Databases for Taxonomic Annotation

Taxonomic assignment, as discussed earlier, depends on the availability of a reference database containing sequences labeled with known taxonomic information. As barcoding genes such as the 16S rRNA gene and ITS are well established in microbiome profiling, several databases collect taxonomic information based on these markers. Widely used RNA databases include GreenGenes (DeSantis et al. 2006), Ribosomal Database Project/RDP (Cole et al. 2014), SILVA (Quast et al. 2013), UNITE (Kõljalg et al. 2005), the NCBI Taxonomy database (Schoch et al. 2020), and the Genome Taxonomy Database (GTDB) (Parks et al. 2018).

The GreenGenes database stores only 16S (rRNA small subunit, SSU) sequences recovered from different sources (other databases, mainly NCBI). With this information, the taxonomy of Bacteria and Archaea was automatically constructed (DeSantis et al. 2006). The database was constructed with ~90,000 sequences in 2006, and its last release was in August 2013. Of the databases covered here, GreenGenes has the smallest number of taxonomic nodes and is the least supported. On the other hand, unlike the rest of the databases, all nodes in the GreenGene database phylogeny are assigned to a defined taxonomic rank (i.e., they have a corresponding domain, phylum, class, order, family, genus or species) (Balvočiute and Huson 2017). It is also a small database in terms of memory requirements and therefore easy to load and query.

The RDP database provides 16S (SSU) sequences from Bacteria and Archaea and 28S (rRNA large subunit, LSU) from Fungi. Its last documented release (September 30, 2016) contained ~3,356,000 aligned and annotated 16S rRNA sequences and ~125,000 28S rRNA sequences. The taxonomy assignment in this database is not phylogeny based but instead relies on a naïve Bayes classifier trained on a smaller subset of the database. RDP developed the RDPipeline program, an online complimentary

service designed to perform several common processing steps for taxonomy-dependent analysis (using the RDP classifier) and for taxonomy-independent analysis (using hierarchical clustering) of large datasets (Cole et al. 2005; Bacci et al. 2015).

SILVA is a comprehensive, up-to-date quality-controlled database of rRNA gene sequences from Bacteria, Archaea, and Eukaryota (Glöckner et al. 2017; Woloszynek et al. 2018). This database is updated in a timely manner. Its last release (version 138, December 16, 2019) contains ~9,500,000 SSU sequences, and from these, ~510,000 are represented in the phylogenetic guide tree. Although the taxonomy is based on Bergey's Taxonomic Outlines and the List of Prokaryotic Names with Standing in Nomenclature (LPSN), its taxonomy classification process is phylogeny based, as it uses guide trees to resolve inconsistencies in nomenclature (Quast et al. 2013; Yilmaz et al. 2014). The taxonomy assignment is manually curated, which is the main difference from the previously mentioned databases. More recently, this database adopted the GTDB taxonomy (Parks et al. 2018), which implied a major rearrangement of the taxonomy, mainly in Bacteria and Archaea. Like RDP, the SILVA database has a web service (SILVAngs) that provides a fully automated analysis of rRNA gene amplicon sequencing data.

A comparison searching for inconsistencies in the taxonomies of these three databases estimated that the annotation error rate in the RDP database is ~10%, while for GreenGenes and SILVA, this value is around 17% (Edgar 2018a). As the author of this study points out, it is striking that the RDP database, in which taxonomy is not assigned explicitly from phylogeny, can be more accurate than the phylogeny-based and/or manually curated alternatives. This again reinforces the notion that 'more is not necessarily better' and that selecting any of these databases for a study should be taken under careful consideration.

These databases all contain marker-based taxonomic associations, which have been useful in the investigation of metagenomes. However, as discussed earlier, using exclusively amplicon sequence and OTU-based methods for taxonomic enumeration can substantially underestimate species diversity. To overcome this limitation, other taxonomies are constructed on more general sequence information. Such are the cases of the Microbial Genome Atlas (MiGA) (Rodriguez-R et al. 2018), the Genome Taxonomy Database (Parks et al. 2018, 2020), and the National Center for Biotechnology Information Taxonomy (Balvočiute and Huson 2017).

The GTDB database taxonomy was constructed using a phylogenomic alignment of 120 concatenated gene markers from more than 150,000 bacterial genomes, followed by the resolution of polyphyletic groups and taxonomy rank normalization by relative evolutionary divergence (Parks et al. 2018, 2019, 2020). One of its most relevant features is the phylogenetical consistency in the sense that it is highly congruent with the relative evolutionary divergence among species and, at the same time, eliminates phyletic conflicts. Also, the taxonomy required the reclassification of ~58% of genomes in the database, as well as the definition of new phyla (Parks et al. 2018). This database also includes several genomes of uncultured species that have been assembled from shotgun metagenomic data and are therefore unnamed. The total amount of unnamed species in the database was estimated around 40%, and more recently, the authors utilized the ANI and alignment fraction (AF) metrics to determine a species-level cluster of genomes that implied a major reordering and renaming at the genus and species rank level (Parks et al. 2019, 2020; Rinke et al. 2020). In our opinion, the GTDB seems to be the most accurate (i.e., closest to the true evolutionary tree) among the currently available taxonomic structures, but an independent formal evaluation of this statement has not been provided so far.

Finally, the largest sequence-associated taxonomy structure used for metagenomic annotation is the National Center for Biotechnology Information Taxonomy (Balvočiute and Huson 2017). There are larger taxonomy structures (e.g., the Open Tree of life Taxonomy; Hinchliff et al. 2015; Rees and Cranston 2017), but these have no sequence information available and are therefore not relevant for NGS sequence analysis.

The NCBI database is a collection of resources that contains nucleotides and protein sequence entries from different experimental origins, and its manually curated taxonomy covers all the sequences in the database (Wheeler et al. 2008; Schoch et al. 2020). The taxonomy assembles taxa naming information contained in 23 external resources, and it is updated weekly. The number of taxa in the NCBI taxonomy is currently over 460,000, which roughly represents a quarter of the described species so far, although it also contains sequence information on about 1.34 million species without formal names that are commonly regarded as 'dark taxa' (Schoch et al. 2020).

A comparison of the taxonomic structures of GreenGenes, RDP, SILVA, NCBI, and the Open Tree of Life Taxonomy (OTT) revealed that the 16S rRNA databases map well onto the larger databases (NCBI and OTT), but these do not map well onto the smaller taxonomies (Balvočiute and Huson 2017). This indicates that NCBI and OTT taxonomies are more explanatory than the 16S rRNA databases, although SILVA, as mentioned earlier, recently adopted the GDBT taxonomy, which could have significantly improved the content of this database. These authors also stated that, while larger, the OTT is not significantly more diverse than the NCBI taxonomy (Balvočiute and Huson 2017). To the best of our knowledge, more recent updates of independent database benchmarking including the latest genome-centered databases have not been published.

2.6 Microbiome Functional Inference from Community Structure

Ecological inferences from marker-based sequencing experiments are needed to fully understand the ecological niches of microbial communities in any environment. These inferences rely on verified metagenomic and metatranscriptomic profiles in several habitats. They allow correlating the taxonomic community structure with different functions in any ecosystem. The microbiome function is difficult to interpret when complete microbial profiling—bacteria, fungi, algae, protozoa—is obtained. Then robust algorithms to predict the ecological functions in nature are needed, and further efforts should be made to get bioinformatics tools to generate new insights related to the role of microorganisms in different ecosystems.

Microbial communities have been routinely studied using different molecular markers to provide powerful taxonomic interpretations. For example, as previously mentioned, 16S and 28S ribosomal RNA genes have been extensively used to assess bacterial and fungal biodiversity, respectively. However, these molecular markers are not useful in establishing the ecological roles of microbes. In this sense, metabolic and ecological functions should be predicted by bioinformatic algorithms using associations between the taxonomic profiles and experimentally demonstrated metabolic capabilities in well-studied species. Thus, methods based on ancestral state reconstruction should be developed, but many challenges are frequently related to these methods. For example, these strategies assign functional annotations by extrapolation of marker gene sequences with those known species. In this scenario, a database with a huge amount of complete microbial genomes properly annotated is necessary. That is an important bottleneck because a few thousand complete microbial genomes are currently available, and the annotation is deficient in many cases. For these reasons, interpretation of the microbial ecological functionality could be limited, and experimental validation should be performed to demonstrate that functional predictions are accurate and realistic.

SINAPS is a method that predicts microbial function from marker gene sequences (Edgar 2017). This algorithm was successfully validated to predict functions related to energy metabolism, Gram-positive staining, presence of flagella, 16S copy number, and number of V4 primer mismatches from 16S V4 ribosomal sequences. The validation of this method demonstrated that a large number of functions were correctly assigned.

Several bioinformatics tools have been developed to infer functional roles of taxa within a community: PICRUSt (Langille et al. 2013), Tax4Fun (Aßhauer et al. 2015), Piphillin (Iwai et al. 2016), Faprotax (Louca, Parfrey, and Doebeli 2016), and PAPRICA (Bowman and Ducklow 2015). Tax4Fun2 (improving on its predecessor Tax4Fun) is a tool for the prediction of functional profiles and redundancy of a metagenomic sequencing based on 16S ribosomal markers (Wemheuer et al. 2020). The accuracy and robustness of this tool were notably enhanced by the incorporation of habitat-specific genomic information.

These methods can predict functional capabilities from prokaryotic ribosomal sequences but not from metagenomic shotgun sequencing. These algorithms have been used to interpret ecological niches of microbial communities inhabiting soil, marine seawater, microbial mats, and so on (Wemheuer et al. 2020). However, the robustness of all these methods is based on the genomes available in public databases, and these only represent a very limited fraction of the functional diversity in nature. Thus, the accuracy and reliability of these methods could be unsatisfactory. To solve this inconvenience, new tools

for specific habitats such as the rumen (Wilkinson et al. 2018) or marine ecosystems (Louca, Parfrey, and Doebeli 2016) have been developed.

A major question in environmental microbiology reflects the need to know if microbial communities contain redundant functional members. This is an urgent question, and its answer could probably allow understanding of how microbial communities provide functional stability to ecosystems. For example, this has importance in polluted habitats where environmental changes could produce ecological successions in microbial communities with unknown impacts on biodegradative functionalities. To address this need, Tax4Fun2 now offers a robust algorithm based on a functional redundancy index that reflects the proportion of species with the capabilities to perform a particular metabolic function and their phylogenetic relationship with others (Wemheuer et al. 2020). Although the authors stated that this tool is also available for fungi, to date, it is only validated for 16S rRNA gene data. Thus, further research is needed to generate new and robust tools useful to predict functions from eukaryotic gene data (18S or 28S rRNA).

PICRUSt2 is another predictive tool to provide functional inferences based on marker gene sequencing (Douglas et al. 2019). This method allows the analysis of eukaryotic communities and is compatible with any OTU-based algorithm. This method was successfully used to identify functional signatures in bacterial communities from humans with inflammatory diseases (Douglas et al. 2019).

Although there is some progress in the development of diverse tools to generate functional microbial inferences, there are important challenges associated with predicting metabolic networks from marker-based sequencing. This aids in designing robust methodologies with a positive impact on bioremediation, for example (Faust 2019). The generation of functional networks could allow optimizing the fitness of microbial consortia to enhance the biodegradation of certain pollutants, such as atrazine (Xu et al. 2019).

2.7 Functional Annotation of Shotgun Metagenome Data

The tools used to annotate metagenomic data often consist of new developments over previously existing tools designed to annotate isolated genomes. In both cases, the first step in the functional annotation is gene prediction. Once the candidate open reading frames (ORFs) are found, they can be linked to biological information based on current knowledge. This means that the functional annotation of the same dataset can be improved over time as our knowledge of biological systems increases.

Accurate gene prediction is a fundamental step in most metagenomics pipelines. Methods for gene prediction are classified as extrinsic (e.g., homology search) or intrinsic (e.g., sequence composition analysis).

Probably the most reliable way to predict a gene is to find a close homolog from another organism, and this is precisely what homology-based methods do: They perform pairwise alignments between metagenomic reads and a given database of known proteins. However, the main drawback of methods based exclusively on homology evidence is that they can only annotate previously known genes.

So, it soon became clear that computational methods that score the coding region using intrinsic sequence features are required for those genes lacking a significant homology to known genes. Intrinsic features can include signal sensors such as start/stop codons and promoters, as well as content sensors, such as patterns of codon usage, k-mer frequency profiles, or any other statistically inferable feature. These *ab initio* approaches increase the possibility of detecting novel genes, as they use linguistic or pattern recognition algorithms to detect specific sequence motifs or global statistical patterns that can consistently help in the process of gene finding. However, these methods are suitable for annotating assembled contigs rather than short reads, as they require a large number of genes for model training.

Methods based on intrinsic evidence include MetaGeneMark (Zhu, Lomsadze, and Borodovsky 2010), FragGeneScan (Rho, Tang, and Ye 2010), Glimmer-MG (Kelley et al. 2012), and MetaProdigal (Hyatt et al. 2012). MetaGeneMark, for instance, is a gene-prediction tool based on GeneMark-HMM. It uses a heuristic method to compute the parameters as functions of intrinsic features of individual sequences, which makes it efficient in predicting genes in metagenomic datasets, as such features (e.g., G+C content, codons, and oligomer frequencies) will probably vary widely among reads.

FragGeneScan is designed to find complete and fragmented genes in short reads (Rho, Tang, and Ye 2010). It combines codon usage information, sequencing error models, and start/stop codon patterns in a

hidden Markov model (HMM) to find the most likely path of hidden states from a given input sequence. It accepts as inputs both short reads or assembled contigs, and it represents a suitable tool for gene prediction in incomplete metagenome assemblies.

Other popular tools are Glimmer-MG (Kelley et al. 2012; Salzberg et al. 1998) and MetaProdigal (Hyatt et al. 2012). The former incorporates classification and clustering of sequences prior to gene prediction using the Glimmer framework and uses a probabilistic model for prediction of gene length and start/stop codon presence in the case of truncated genes that are typical of shotgun metagenome sequencing. The latter is a metagenomic version of the gene prediction program Prodigal (Hyatt et al. 2010), which provides enhanced translation initiation site identification, the ability to identify sequences that use alternate genetic codes, and confidence values for each gene prediction.

Some pipelines use a combination of evidence-based (extrinsic) and *ab initio* (intrinsic) methods. This is the case of an in-house pipeline developed by researchers from the Max Planck Institute for Marine Microbiology, named meta-ORF-finder or mORFind (unpublished), which uses a combination of Orpheus (Frishman et al. 1998), CRITICA (Badger and Olsen 1999), and the previously mentioned Glimmer framework. Both CRITICA and Orpheus are BLAST-based tools that aim to identify coding regions in genomes invoking comparative analysis. CRITICA first considers the observed amino acid identity for the translated aligned sequences once their percentage nucleotide identity is known; if it is higher than expected, this is taken as evidence for coding. Next, it incorporates information of the relative hexanucleotide frequencies in coding frames versus other contexts, and this feature makes it less dependent on the accuracy of sequence annotation in databases and thus well suited for the analysis of novel genomes. Orpheus uses a very similar approach, in which the similarity-derived seed ORFs have their coding potential parameters calculated and scored. Those features, combined with those included in Glimmer, make mORFind a versatile tool suitable for metagenome gene prediction.

Identifying eukaryotic protein-coding genes in metagenomes is more challenging than identifying prokaryotic ones due to the exon-intron architecture of eukaryotic genes. A tool specially designed to meet this challenge is MetaEuk (Levy Karin, Mirdita, and Söding 2020). This toolkit allows high-throughput reference-based discovery and annotation of protein-coding genes in eukaryotic metagenomic contigs. Instead of doing a spliced alignment, which would be computationally costly, it takes as input a set of assembled contigs and scans each contig in all six reading frames to extract putative protein fragments between stop codons in each frame. Then it uses a sensitive and efficient method called MMseqs2 to perform an iterative search through any target database.

GeneMark-ES is an *ab initio* algorithm iterative unsupervised training to identify protein-coding genes in eukaryotic genomes. Augustus is one of the most accurate tools for eukaryotic gene prediction and is based on a generalized hidden Markov model (Stanke and Waack 2003). A web interface (WebAugustus) allows users to train their own gene structures or upload a training gene structure file or genome file and then perform eukaryotic gene predictions (Hoff and Stanke 2013).

Predicting protein function is the next step in metagenome functional annotation. It can be based on different sources of information, such as sequence similarity (mapping to databases), phylogenetic profiles, protein–protein interactions, and protein complexes.

A common approach is to perform a translated BLAST search to determine the annotation that will be assigned to a read based on its alignment scores, and this criterion can vary according to the protocol. For instance, the final annotation can consider only optimal alignments, including suboptimal alignments, or even use the average of multiple high-scoring hits. However, it is important to consider that the efficiency of the alignment methods is influenced by sequencing errors, read length, the phylogenetic coverage of the reference database, and the differences in annotation accuracy across the clades.

Also, the alignment of large datasets can be a computationally intensive and limiting step in the analysis of metagenomes. Thus, some tools were developed to achieve faster and accurate functional annotation. One example is Woods, an orthology-based functional classifier that uses a combination of machine learning (random forest) and similarity-based classification (RAPsearch2).

Methods based on protein interactions include interolog mapping. Interologs are interacting pairs of proteins that have homologs with conserved interaction in another organism. In interolog mapping, a known interolog in an organism is extended to a second organism assuming that the homologous proteins in different organisms maintain their interaction properties. A tool that uses the interolog concept is

STRING. It uses hierarchically arranged orthologous group relations, as defined in eggNOG, to transfer associations between organisms (prokaryotes and eukaryotes) where applicable.

Function prediction based on multilayer protein networks (FP-MPN) is a method that integrates protein–protein interaction (PPI) networks, protein domain content, and protein complex subunit information to predict protein function. This method assumes that diverse types of connections between groups of proteins reflect distinct roles and importance.

2.8 Main Databases Used for Functional Annotation

The main resources for functional annotation are currently the National Center for Biotechnology Information databases. The nr database is a protein database that contains non-identical sequences from GenBank CDS translations, Protein Data Bank (PDB), Swiss-Prot, Protein Information Resource (PIR), and Protein Research Foundation (PRF). RefSeq is an open-access, curated, and non-redundant database of publicly available genomes, transcripts, and protein products.

The UniProt database is a large collection of protein sequences and annotations from all domains of life. It contains more than 120 million protein sequences, of which the majority are derived from the translated genome and MAG sequence information deposited in ENA/GenBank/DDBJ databases (Bateman 2019). Over one-half of those proteins have annotations obtained from the literature by expert curators, while the remaining entries are automatically annotated using information from several databases, mainly InterPro.

The KEGG database allows estimating metabolic pathways in a metagenome. KEGG integrates information from 15 other databases by a computational database construction algorithm (Kanehisa and Goto 2000). The genomic information category, which is based on the KO (KEGG Orthologues) database, contains genomes and genes derived from different databases (RefSeq, Genbank, and NCBI Taxonomy), giving them original KEGG annotations. KEGG mapping can be performed with the KEGG Mapper tool, together with the KOALA tools (BlastKOALA and GhostKOALA), which allow for an automatic assignment of KO (KEGG orthology) identifiers used in the mapping (Kanehisa and Goto 2000; Kanehisa et al. 2016; Kanehisa and Sato 2020). PFAM is a collection of curated protein families, each represented by multiple sequence alignments generated using hidden Markov models (Finn et al. 2014). eggNOG (Evolutionary Genealogy of Genes: Non-supervised Orthologous Groups) is a database that provides orthologous gene mappings for Bacteria, Archaea, and Eukaryotes (Powell et al. 2012; Huerta-Cepas et al. 2016). Specialized databases include dbCAN for the carbohydrate-active enzyme (CAZYmes) (Yin et al. 2012); MEROPS for proteolytic enzymes (Rawlings, Barrett, and Finn 2016), their substrates, and inhibitors; and the Lipase Engineering Database (Fischer and Pleiss 2003).

2.9 Visualization of Metagenomic Data Annotation

A challenging task that researchers usually face when analyzing big data such as NGS sequencing results is representing the data to appropriately communicate the findings of the studies. Ideally, the representations should be as simple as possible, they should not represent individual entities or behaviors of the data but rather aggregated groups or tendencies and colors of different entities, all the details of the figure should be clearly distinguishable, and the resulting figures should fit in the printing area of most paper formats. The most common representations used in the field can be observed in Figure 2.4.

In metagenome analysis, stacked bar plots are the most used and abused representation of taxonomic annotation and abundance. This representation can be obtained from almost all analysis pipelines and is very useful to see major sample differences in abundance or composition. However, stacked bar plots are inadequate to see differences in minor taxa, can be hard to read (especially for color-blind people), and are not suited to represent diversity in complex communities. Several other diagrams have been used to represent diversity and abundance, each with its pros and cons.

Stack bar plots and other representations for visualizing metagenomic data omit or distort quantitative hierarchical relationships and cannot display secondary variables. Krona, Centrifuge, and other tools

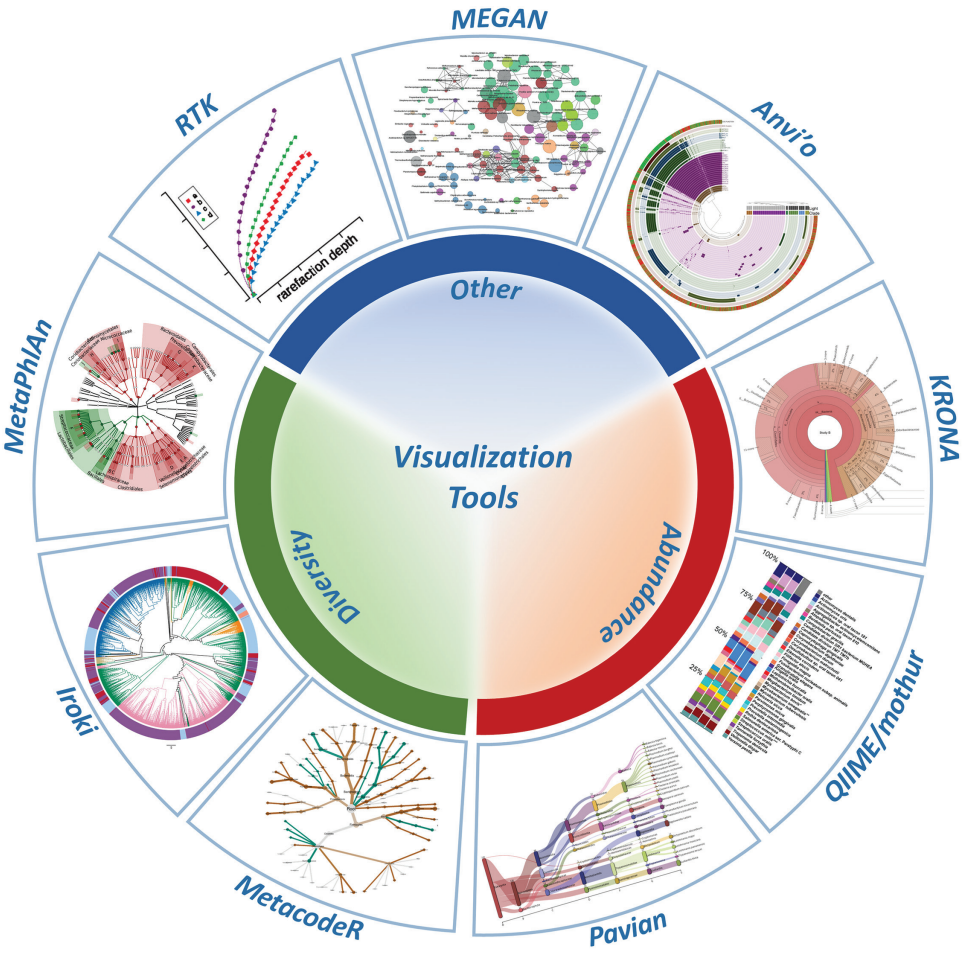

FIGURE 2.4 Graphs and types of visualizations produced by different metagenome analysis tools aiming to represent taxonomical information. Visualizations can accommodate metagenome diversity, abundance, or other relevant experimental information. Diversity can be explicitly represented as cladograms or trees in which each branch is an OTU or ASV or can be formally analyzed using rarefaction curves. Other visualizations such as stacked bar plots, pie charts, or Sankey diagrams can also incorporate abundance information. Other representations can depict the presence, co-occurrence, or differential abundance of taxa in several samples. See the main text for a detailed explanation.

enable the interactive visualization of complex metagenomic data using multi-layer pie charts, which depict the abundance of the most commons microbes of a sample. Krona distinguishes itself among others, as it has a lightweight implementation and is easily integrated into existing portals such as MG-RAST, METAREP, and Galaxy (Ondov, Bergman, and Phillippy 2011). The major disadvantages of this representation are that it can only accommodate information from one sample or experimental group (i.e., it is not suited to compare different samples), and the information that is interactively displayed (embedded in the graph) is lost when printed.

Sankey flow diagrams are equivalent to pie charts in terms of displayed information and interpretation but are more visually appealing. This representation also considers quantitative taxonomic hierarchy information, with taxonomic levels distributed in the horizontal axis. Pavian (Breitwieser and Salzberg 2020) and BioSankey (Platzer et al. 2018) are among the tools that can produce this type of diagram and, similarly, can embed information on the graph to allow the interactive representation of additional data such as sample comparisons. This representation has the same disadvantages as pie charts.

Taxonomical information has traditionally been represented using cladograms and phylogenetic trees. This simple representation is suited to display all the taxonomic entities in a sample or group at a given rank. Annotated trees and cladograms are widely used in the metagenomic field, as they can display diversity information more precisely than the previously mentioned representations. Tunable tools that allow annotation of phylogenetic trees include iTOL (Letunic and Bork 2007, 2019), EvolView (Zhang et al. 2012; Subramanian et al. 2019), and Iroki (Moore et al. 2020), among others. Most of them are not specially designed to handle the output of metagenome annotation tools. However, iTOL recently incorporated an option to read the tree files output by QIIME (Letunic and Bork 2019), while Iroki can deal with trees from QIIME, SILVAngs classifier, and other tools (Moore et al. 2020).

Phylogeny trees and cladograms can also be represented by many specialized metagenomic pipelines. That is the case of tools like MetaPhlAn (Segata et al. 2012; Truong et al. 2015), Anvi'o (Eren et al. 2015), and MetacodeR (Foster, Sharpton, and Grünwald 2017). MetaPhAn produces cladograms that can be annotated by shading the branches; modifying the size and color of nodes; and separating the different ranks of the taxonomy to achieve a more comprehensive understanding of diversity, abundance, and even sample differences (Segata et al. 2012; Truong et al. 2015). MetacodeR performs similar tree representations but also allows producing layouts of various trees to compare groups of samples and other segmentations of the study design (Foster, Sharpton, and Grünwald 2017). Anvi'o produces densely annotated and interactive trees to assist researchers in the human-guided binning of metagenomes (Eren et al. 2015). These trees can be exported in publication-ready format and might serve to display several metadata types associated with samples and experimental groups, but a high number of annotations can interfere with the readability of the final diagram and obstruct the diversity representation as the tree is shrunk.

For sample comparison, sample rarefaction analysis can help determine a sequencing depth at which all samples have the same amount of 'sequencing effort.' This is a widespread normalization strategy to avoid sequencing depth bias when estimating abundances. Rarefaction curves are also informative about the amount of ecosystem diversity that was effectively captured by the sequencing experiment and therefore are very useful during the experiment design stage. Although many tools can produce rarefaction analyses, a recently published tool called RTK (Saary et al. 2017) seems to be among the best current solutions to efficiently calculate rarefaction curves and normalize a high number of samples.

STAMP (STatistical Analysis of Metagenomic Profiles) is a graphical tool focused on sample comparison with a relevant and rigorous statistical treatment of biological effects (Parks and Beiko 2010; Parks et al. 2014). This tool provides pairwise or multiple comparisons of annotated metagenomic profiles with the added feature of reporting effect size and confidence intervals associated with an annotation feature (gene, pathway, gene ontology, or enzyme class). Effect size statistics calculated using STAMP can be complementary and sometimes pivotal in assessing the biological relevance of *p*-values in hypothesis testing during sample comparison.

Species co-occurrence networks can give information that is complementary to differential abundance analysis. In this case, the nodes of the network represent identified OTUs or taxa (generally species), and the connections represent a correlation coefficient that is obtained from the matrix of sample species abundance. This co-occurrence network can be obtained from several pipelines, including MEGAN (Huson et al. 2016; Bağcı et al. 2019). The direct implication of this analysis is identifying hub species that can shape or influence community structure. As discussed earlier, this could represent a valuable tool to design intervention strategies to influence soil communities, gut microbiota, water treatment biodigester microbes, and so on.

Deciding on the most appropriate graphical representation for a dataset or result is pivotal for efficient communication of experimental findings. Several pipelines can produce the same types of representation with different aesthetics and varying levels of difficulty. It is a task that should be considered with the same caution as the decision on tools for scientific computation.

2.10 Challenges and Future Perspectives

From its beginnings in the early 2000s, metagenomics has become a popular approach to evaluate the taxonomic and functional composition of microbiomes. The reduction in sequencing costs, increased computational power, and the surge of bioinformatics tools have enabled many laboratories to study

microbial communities in diverse environments. As noted throughout the chapter, there are still several challenges that researchers in the field are facing.

The most pressing of subjects is arriving at a consensus on annotation pipelines. Without a unified way of processing and reporting data, much of the findings will not be reusable for validation by peers or under future hypotheses, making it harder to achieve knowledge integration. There should be a culture among bioinformaticians and biologists working with computational resources to ensure that their use of tools is guided by truly scientific reasons and not by trends or ease of use. In most cases, the highly cited tools are already evaluated independently by a benchmarking study, which should provide enough support for deciding on a pipeline. If that is not the case, several platforms (some discussed in this chapter) are available to guide researchers through performing a benchmarking study. The time spent on this task would not be fruitless, as it will reflect on the reliability of the results obtained from our methodology.

There is also a global demand from the scientific community to improve on how researchers produce, interpret, and publish data to ensure that the information complies with four criteria: It should be findable, accessible, interoperable, and reusable (Wilkinson et al. 2016). The FAIR guidelines (common jargon) aim to provide a framework for researchers, mainly those working on big data, to produce data with enough value to be used by others and enough structure to be handled by machines.

For scientists working on metagenomics, this implies that, as a general recommendation, the raw data of every experiment should be available along with the processed results. Moreover, these data should be accompanied by structured and detailed metadata that describes the experimental conditions, treatments, subject details, locations, sample processing, and any other relevant information that could be influencing the results.

Given the heterogeneity of methods, each with its own set of deficiencies, the used pipeline should also be available and preferably the used code be deposited in a reputable DOI-issuing repository so that others can access and cite it (Wilson et al. 2017). This becomes increasingly easy as researchers working in scientific computing adopt good practices when performing computational experiments. A detailed compendium of good practices for scientific computing can be found in (Wilson et al. 2017), and these should be considered as important as good pipetting practices for wet-lab experiments.

Another challenge in the field is homogenizing the nomenclature of techniques, processes, data types, and computational tasks. In the case of bioinformatics data and tasks, which are somehow less ambiguous, there is a set of ontologies (EDAM ontologies) to describe the more relevant and used operations, formats, and types (Ison et al. 2013). The proper use of these terms, although technically complex, is more normalized than the use of terms directly related to biological experiments. For example, terms such as metagenome, microbiome, metaprofiling, and metabarcoding are used interchangeably in literature to refer to the same technique. An effort to generate a common vocabulary was initially published by Marchesi and Ravel (2015). Unfortunately, some of the terms and definitions used in that text, such as metataxonomics, have not achieved widespread acceptance in the field. In the case of metataxonomics, in our opinion, the term does not properly justify its intended uses (e.g., marker-based and WGS sequencing for the sole purpose of taxonomical description of communities). On the other hand, their definition of the microbiome (i.e., covering both biotic and abiotic factors of an environment) rules out the most widespread uses of this term, which by force of habit only refer to the biotic component of an ecosystem. We encourage our readers to review the cited resources on nomenclature and actively commit to incorporating proper language in the description of scientific results. Adopting these or other future nomenclatures is a challenge that the community should face responsibly, as normalizing language will improve brevity, clarity, precision, and ultimately communication among specialists.

REFERENCES

Aßhauer, Kathrin P., Bernd Wemheuer, Rolf Daniel, and Peter Meinicke. 2015. "Tax4Fun: Predicting Functional Profiles from Metagenomic 16S RRNA Data." *Bioinformatics* 31 (17): 2882–84. https://doi.org/10.1093/bioinformatics/btv287.

Almeida, Alexandre, Alex L. Mitchell, Aleksandra Tarkowska, and Robert D. Finn. 2018. "Benchmarking Taxonomic Assignments Based on 16S RRNA Gene Profiling of the Microbiota from Commonly Sampled Environments." *GigaScience* 7 (5): 1–10. https://doi.org/10.1093/gigascience/giy054.

Amir, Amnon, Daniel McDonald, Jose A. Navas-Molina, Evguenia Kopylova, James T. Morton, Zhenjiang Zech Xu, Eric P. Kightley, et al. 2017. "Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns." Edited by Jack A. Gilbert. *MSystems* 2 (2): 1–7. https://doi.org/10.1128/mSystems.00191-16.

Bacci, Giovanni, Alessia Bani, Marco Bazzicalupo, Maria Teresa Ceccherini, Marco Galardini, Paolo Nannipieri, Giacomo Pietramellara, and Alessio Mengoni. 2015. "Evaluation of the Performances of Ribosomal Database Project (RDP) Classifier for Taxonomic Assignment of 16S RRNA Metabarcoding Sequences Generated from Illumina-Solexa NGS." *Journal of Genomics* 3: 36–39. https://doi.org/10.7150/jgen.9204.

Badger, Jonathan H., and Gary J. Olsen. 1999. "CRITICA: Coding Region Identification Tool Invoking Comparative Analysis." *Molecular Biology and Evolution* 16 (4): 512–24. https://doi.org/10.1093/oxfordjournals.molbev.a026133.

Bağcı, Caner, Sina Beier, Anna Górska, and Daniel H. Huson. 2019. "Introduction to the Analysis of Environmental Sequences: Metagenomics with MEGAN." In *Methods in Molecular Biology*, 1910: 591–604. Humana Press Inc. https://doi.org/10.1007/978-1-4939-9074-0_19.

Balvočiute, Monika, and Daniel H. Huson. 2017. "SILVA, RDP, Greengenes, NCBI and OTT—How Do These Taxonomies Compare?" *BMC Genomics* 18 (Suppl 2): 1–8. https://doi.org/10.1186/s12864-017-3501-4.

Bateman, Alex. 2019. "UniProt: A Worldwide Hub of Protein Knowledge." *Nucleic Acids Research* 47 (D1): D506–15. https://doi.org/10.1093/nar/gky1049.

Bazinet, Adam L., and Michael P. Cummings. 2012. "A Comparative Evaluation of Sequence Classification Programs." *BMC Bioinformatics* 13 (1). https://doi.org/10.1186/1471-2105-13-92.

Bowman, Jeff S., and Hugh W. Ducklow. 2015. "Microbial Communities Can Be Described by Metabolic Structure: A General Framework and Application to a Seasonally Variable, Depth-Stratified Microbial Community from the Coastal West Antarctic Peninsula." *PLoS One* 10 (8): 1–18. https://doi.org/10.1371/journal.pone.0135868.

Bray, Nicolas L., Harold Pimentel, Páll Melsted, and Lior Pachter. 2016. "Near-Optimal Probabilistic RNA-Seq Quantification." *Nature Biotechnology* 34 (5): 525–27. https://doi.org/10.1038/nbt.3519.

Breitwieser, Florian P., Jennifer Lu, and Steven L. Salzberg. 2018. "A Review of Methods and Databases for Metagenomic Classification and Assembly." *Briefings in Bioinformatics* 20 (4): 1125–39. https://doi.org/10.1093/bib/bbx120.

Breitwieser, Florian P., and Steven L. Salzberg. 2020. "Pavian: Interactive Analysis of Metagenomics Data for Microbiome Studies and Pathogen Identification." *Bioinformatics* 36 (4): 1303–4. https://doi.org/10.1093/bioinformatics/btz715.

Buchfink, Benjamin, Chao Xie, and Daniel H. Huson. 2015. "Fast and Sensitive Protein Alignment Using DIAMOND." *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

Callahan, Benjamin J., Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.

Cole, J. R., B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje. 2005. "The Ribosomal Database Project (RDP-II): Sequences and Tools for High-Throughput RRNA Analysis." *Nucleic Acids Research* 33 (DATABASE ISS.): 294–6. https://doi.org/10.1093/nar/gki038.

Cole, J. R., Qiong Wang, Jordan A. Fish, Benli Chai, Donna M. McGarrell, Yanni Sun, C. Titus Brown, Andrea Porras-Alfaro, Cheryl R. Kuske, and James M. Tiedje. 2014. "Ribosomal Database Project: Data and Tools for High Throughput RRNA Analysis." *Nucleic Acids Research* 42 (D1): 633–42. https://doi.org/10.1093/nar/gkt1244.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. 2006. "Greengenes, a Chimera-Checked 16S RRNA Gene Database and Workbench Compatible with ARB." *Applied and Environmental Microbiology* 72 (7): 5069–72. https://doi.org/10.1128/AEM.03006-05.

Douglas, Gavin M., Vincent J. Maffei, Jesse Zaneveld, Svetlana N. Yurgel, James R. Brown, Christopher M. Taylor, Curtis Huttenhower, and Morgan G. I. Langille. 2019. "PICRUSt2: An Improved and Extensible Approach for Metagenome Inference." *BioRxiv*. https://doi.org/10.1101/672295.

Edgar, Robert C. 2010. "Search and Clustering Orders of Magnitude Faster Than BLAST." *Bioinformatics* 26 (19): 2460–1. https://doi.org/10.1093/bioinformatics/btq461.

Edgar, Robert C. 2017. "SINAPS: Prediction of microbial traits from marker gene sequences." *BioRxiv*, no. Moran 2015. https://doi.org/10.1101/124156.

Edgar, Robert C. 2018a. "Taxonomy Annotation and Guide Tree Errors in 16S RRNA Databases." *PeerJ* 2018 (6). https://doi.org/10.7717/peerj.5030.

Edgar, Robert C. 2018b. "Updating the 97% Identity Threshold for 16S Ribosomal RNA OTUs." *Bioinformatics* 34 (14): 2371–75. https://doi.org/10.1093/bioinformatics/bty113.

Eren, A. Murat, Ozcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. 2015. "Anvi'o: An Advanced Analysis and Visualization Platform for 'omics Data." *PeerJ* 2015 (10): 1–29. https://doi.org/10.7717/peerj.1319.

Escobar-Zepeda, Alejandra, Elizabeth Ernestina Godoy-Lozano, Luciana Raggi, Lorenzo Segovia, Enrique Merino, Rosa María Gutiérrez-Rios, Katy Juarez, Alexei F. Licea-Navarro, Liliana Pardo-Lopez, and Alejandro Sanchez-Flores. 2018. "Analysis of Sequencing Strategies and Tools for Taxonomic Annotation: Defining Standards for Progressive Metagenomics." *Scientific Reports* 8 (1): 1–13. https://doi.org/10.1038/s41598-018-30515-5.

Faust, Karoline. 2019. "Microbial Consortium Design Benefits from Metabolic Modeling." *Trends in Biotechnology* 37 (2): 123–5. https://doi.org/10.1016/j.tibtech.2018.11.004.

Finn, Robert D., Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, et al. 2014. "Pfam: The Protein Families Database." *Nucleic Acids Research* 42 (D1): 222–30. https://doi.org/10.1093/nar/gkt1223.

Fischer, Markus, and Jürgen Pleiss. 2003. "The Lipase Engineering Database: A Navigation and Analysis Tool for Protein Families." *Nucleic Acids Research* 31 (1): 319–21. https://doi.org/10.1093/nar/gkg015.

Fosso, Bruno, Graziano Pesole, Francesc Rosselló, and Gabrie Valiente. 2018. "Unbiased Taxonomic Annotation of Metagenomic Samples." *Journal of Computational Biology* 25 (3): 348–60. https://doi.org/10.1089/cmb.2017.0144.

Foster, Zachary S. L., Thomas J. Sharpton, and Niklaus J. Grünwald. 2017. "Metacoder: An R Package for Visualization and Manipulation of Community Taxonomic Diversity Data." *PLoS Computational Biology* 13 (2): 1–15. https://doi.org/10.1371/journal.pcbi.1005404.

Frishman, Dmitrij, Andrey Mironov, Hans Werner Mewes, and Mikhail Gelfand. 1998. "Combining Diverse Evidence for Gene Recognition in Completely Sequenced Bacterial Genomes." *Nucleic Acids Research* 26 (12): 2941–7. https://doi.org/10.1093/nar/26.12.2941.

Gardner, Paul P., Renee J. Watson, Xochitl C. Morgan, Jenny L. Draper, Robert D. Finn, Sergio E. Morales, and Matthew B. Stott. 2019. "Identifying Accurate Metagenome and Amplicon Software via a Meta-Analysis of Sequence to Taxonomy Benchmarking Studies." *PeerJ* 2019 (1): 1–19. https://doi.org/10.7717/peerj.6160.

Gardner, Paul P., James Paterson, Fatemeh Ashari-Ghomi, Sinan Umu, Stephanie McGimpsey, and Aleksandra Pawlik. 2017. "A Meta-Analysis of Bioinformatics Software Benchmarks Reveals That Publication-Bias Unduly Influences Software Accuracy." *BioRxiv*. https://doi.org/10.1101/092205.

Glöckner, Frank Oliver, Pelin Yilmaz, Christian Quast, Jan Gerken, Alan Beccati, Andreea Ciuprina, Gerrit Bruns, et al. 2017. "25 Years of Serving the Community with Ribosomal RNA Gene Reference Databases and Tools." *Journal of Biotechnology* 261 (February): 169–76. https://doi.org/10.1016/j.jbiotec.2017.06.1198.

Hillmann, Benjamin, Gabriel A. Al-Ghalith, Robin R. Shields-Cutler, Qiyun Zhu, Daryl M. Gohl, Kenneth B. Beckman, Rob Knight, and Dan Knights. 2018. "Evaluating the Information Content of Shallow Shotgun Metagenomics." Edited by John F. Rawls. *MSystems* 3 (6): e00069–18. https://doi.org/10.1128/mSystems.00069-18.

Hinchliff, Cody E., Stephen A. Smith, James F. Allman, J. Gordon Burleigh, Ruchi Chaudhary, Lyndon M. Coghill, Keith A. Crandall, et al. 2015. "Synthesis of Phylogeny and Taxonomy into a Comprehensive Tree of Life." *Proceedings of the National Academy of Sciences of the United States of America* 112 (41): 12764–9. https://doi.org/10.1073/pnas.1423041112.

Hleap, Jose S., Joanne E. Littlefair, Dirk Steinke, Paul D. N. Hebert, and Melania E. Cristescu. 2020. "Assessment of Current Taxonomic Assignment Strategies for Metabarcoding Eukaryotes." *BioRxiv*, 40. https://doi.org/10.1101/2020.07.21.214270.

Hoff, Katharina J., and Mario Stanke. 2013. "WebAUGUSTUS-a Web Service for Training AUGUSTUS and Predicting Genes in Eukaryotes." *Nucleic Acids Research* 41 (Web Server issue): 123–8. https://doi.org/10.1093/nar/gkt418.

Huerta-Cepas, Jaime, Damian Szklarczyk, Kristoffer Forslund, Helen Cook, Davide Heller, Mathias C. Walter, Thomas Rattei, et al. 2016. "EGGNOG 4.5: A Hierarchical Orthology Framework with Improved Functional Annotations for Eukaryotic, Prokaryotic and Viral Sequences." *Nucleic Acids Research* 44 (D1): D286–93. https://doi.org/10.1093/nar/gkv1248.

Huson, D. H., Sina Beier, Isabell Flade, Anna Górska, Mohamed El-Hadidi, Suparna Mitra, Hans-Joachim Joachim Ruscheweyh, and Rewati Tappu. 2016. "MEGAN Community Edition—Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data." Edited by Timothée Poisot. *PLOS Computational Biology* 12 (6): e1004957. https://doi.org/10.1371/journal.pcbi.1004957.

Huson, D. H., Suparna Mitra, Hj Ruscheweyh, N. Weber, and S. C. Schuster. 2011. "Integrative Analysis of Environmental Sequences Using MEGAN4." *Genome Research* 21 (9): 1552–60. https://doi.org/10.1101/gr.120618.111.

Hyatt, Doug, Gwo Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. "Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification." *BMC Bioinformatics* 11. https://doi.org/10.1186/1471-2105-11-119.

Hyatt, Doug, Philip F. Locascio, Loren J. Hauser, and Edward C. Uberbacher. 2012. "Gene and Translation Initiation Site Prediction in Metagenomic Sequences." *Bioinformatics* 28 (17): 2223–30. https://doi.org/10.1093/bioinformatics/bts429.

Ison, Jon, Matúš Kalaš, Inge Jonassen, Dan Bolser, Mahmut Uludag, Hamish McWilliam, James Malone, Rodrigo Lopez, Steve Pettifer, and Peter Rice. 2013. "EDAM: An Ontology of Bioinformatics Operations, Types of Data and Identifiers, Topics and Formats." *Bioinformatics* 29 (10): 1325–32. https://doi.org/10.1093/bioinformatics/btt113.

Iwai, Shoko, Thomas Weinmaier, Brian L. Schmidt, Donna G. Albertson, Neil J. Poloso, Karim Dabbagh, and Todd Z. DeSantis. 2016. "Piphillin: Improved Prediction of Metagenomic Content by Direct Inference from Human Microbiomes." *PLoS One* 11 (11): 1–18. https://doi.org/10.1371/journal.pone.0166104.

Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. "High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries." *Nature Communications* 9 (1): 1–8. https://doi.org/10.1038/s41467-018-07641-9.

Kanehisa, Minoru, and Susumu Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research*. Oxford University Press. https://doi.org/10.1093/nar/28.1.27.

Kanehisa, Minoru, and Yoko Sato. 2020. "KEGG Mapper for Inferring Cellular Functions from Protein Sequences." *Protein Science* 29 (1): 28–35. https://doi.org/10.1002/pro.3711.

Kanehisa, Minoru, Yoko Sato, Kanae Morishima, Kanehisa M. Sato, Y. Morishima, K. Minoru Kanehisa, Yoko Sato, and Kanae Morishima. 2016. "BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences." *Journal of Molecular Biology* 428 (4): 726–31. https://doi.org/10.1016/j.jmb.2015.11.006.

Kelley, David R., Bo Liu, Arthur L. Delcher, Mihai Pop, and Steven L. Salzberg. 2012. "Gene Prediction with Glimmer for Metagenomic Sequences Augmented by Classification and Clustering." *Nucleic Acids Research* 40 (1): 1–12. https://doi.org/10.1093/nar/gkr1067.

Khachatryan, Lusine, Rick H. de Leeuw, Margriet E. M. Kraakman, Nikos Pappas, Marije te Raa, Hailiang Mei, Peter de Knijff, and Jeroen F. J. Laros. 2020. "Taxonomic Classification and Abundance Estimation Using 16S and WGS—A Comparison Using Controlled Reference Samples." *Forensic Science International: Genetics* 46 (December 2019): 102257. https://doi.org/10.1016/j.fsigen.2020.102257.

Kiełbasa, Szymon M., Raymond Wan, Kengo Sato, Paul Horton, and Martin C. Frith. 2011. "Adaptive Seeds Tame Genomic Sequence Comparison." *Genome Research* 21 (3): 487–93. https://doi.org/10.1101/gr.113985.110.

Kim, Daehwan, Li Song, Florian P. Breitwieser, and Steven L. Salzberg. 2016. "Centrifuge: Rapid and Sensitive Classification of Metagenomic Sequences." *Genome Research* 26 (12): 1721–9. https://doi.org/10.1101/gr.210641.116.

Kõljalg, Urmas, Karl Henrik Larsson, Kessy Abarenkov, R. Henrik Nilsson, Ian J. Alexander, Ursula Eberhardt, Susanne Erland, et al. 2005. "UNITE: A Database Providing Web-Based Methods for the Molecular Identification of Ectomycorrhizal Fungi." *New Phytologist* 166 (3): 1063–68. https://doi.org/10.1111/j.1469-8137.2005.01376.x.

Langille, Morgan G. I., Jesse Zaneveld, J. Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A. Reyes, Jose C. Clemente, et al. 2013. "Predictive Functional Profiling of Microbial Communities Using 16S RRNA Marker Gene Sequences." *Nature Biotechnology* 31 (9): 814–21. https://doi.org/10.1038/nbt.2676.

Lee, Imchang, Yeong Ouk Kim, Sang Cheol Park, and Jongsik Chun. 2016. "OrthoANI: An Improved Algorithm and Software for Calculating Average Nucleotide Identity." *International Journal of Systematic and Evolutionary Microbiology* 66 (2): 1100–3. https://doi.org/10.1099/ijsem.0.000760.

Letunic, Ivica, and Peer Bork. 2007. "Interactive Tree of Life (ITOL): An Online Tool for Phylogenetic Tree Display and Annotation." *Bioinformatics* 23 (1): 127–8. https://doi.org/10.1093/bioinformatics/btl529.

Letunic, Ivica, and Peer Bork. 2019. "Interactive Tree of Life (ITOL) v4: Recent Updates and New Developments." *Nucleic Acids Research* 47 (W1): 256–9. https://doi.org/10.1093/nar/gkz239.

Levy Karin, Eli, Milot Mirdita, and Johannes Söding. 2020. "MetaEuk-Sensitive, High-Throughput Gene Discovery, and Annotation for Large-Scale Eukaryotic Metagenomics." *Microbiome* 8 (1): 1–15. https://doi.org/10.1186/s40168-020-00808-x.

Lindgreen, Stinus, Karen L. Adair, and Paul P. Gardner. 2016. "An Evaluation of the Accuracy and Speed of Metagenome Analysis Tools." *Scientific Reports* 6: 1–14. https://doi.org/10.1038/srep19233.

López-García, Adrian, Carolina Pineda-Quiroga, Raquel Atxaerandio, Adrian Pérez, Itziar Hernández, Aser García-Rodríguez, and Oscar González-Recio. 2018. "Comparison of Mothur and QIIME for the Analysis of Rumen Microbiota Composition Based on 16S RRNA Amplicon Sequences." *Frontiers in Microbiology* 9 (DEC): 1–11. https://doi.org/10.3389/fmicb.2018.03010.

Louca, S., L. W. Parfrey, and M. Doebeli. 2016. "Decoupling Function and Taxonomy in the Global Ocean Microbiome." *Science* 353 (6305): 1272–7. https://doi.org/10.1126/science.aaf4507.

Mangul, Serghei, Lana S. Martin, Brian L. Hill, Angela Ka Mei Lam, Margaret G. Distler, Alex Zelikovsky, Eleazar Eskin, and Jonathan Flint. 2019. "Systematic Benchmarking of Omics Computational Tools." *Nature Communications* 10 (1): 1–11. https://doi.org/10.1038/s41467-019-09406-4.

Marchesi, Julian R., and Jacques Ravel. 2015. "The Vocabulary of Microbiome Research: A Proposal." *Microbiome* 3 (1): 1–3. https://doi.org/10.1186/s40168-015-0094-5.

Marx, Vivien. 2020. "When Computational Pipelines Go 'Clank'." *Nature Methods* 17 (7): 659–62. https://doi.org/10.1038/s41592-020-0886-9.

McIntyre, Alexa B. R., Rachid Ounit, Ebrahim Afshinnekoo, Robert J. Prill, Elizabeth Hénaff, Noah Alexander, Samuel S. Minot, et al. 2017. "Comprehensive Benchmarking and Ensemble Approaches for Metagenomic Classifiers." *Genome Biology* 18 (1): 1–19. https://doi.org/10.1186/s13059-017-1299-7.

Menzel, Peter, Kim Lee Ng, and Anders Krogh. 2016. "Fast and Sensitive Taxonomic Classification for Metagenomics with Kaiju." *Nature Communications* 7. https://doi.org/10.1038/ncomms11257.

Moore, Ryan M., Amelia O. Harrison, Sean M. McAllister, Shawn W. Polson, and K. Eric Wommack. 2020. "Iroki: Automatic Customization and Visualization of Phylogenetic Trees." *PeerJ* 8: e8584. https://doi.org/10.7717/peerj.8584.

Nasko, Daniel J., Sergey Koren, Adam M. Phillippy, and Todd J. Treangen. 2018. "RefSeq Database Growth Influences the Accuracy of K-Mer-Based Lowest Common Ancestor Species Identification." *Genome Biology* 19 (1): 165. https://doi.org/10.1186/s13059-018-1554-6.

Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. 2011. "Interactive Metagenomic Visualization in a Web Browser." *BMC Bioinformatics* 12 (September). https://doi.org/10.1186/1471-2105-12-385.

Parks, Donovan H., and Robert G. Beiko. 2010. "Identifying Biologically Relevant Differences Between Metagenomic Communities." *Bioinformatics* 26 (6): 715–21. https://doi.org/10.1093/bioinformatics/btq041.

Parks, Donovan H., Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J. Mussig, and Philip Hugenholtz. 2019. "Selection of Representative Genomes for 24,706 Bacterial and Archaeal Species Clusters Provide a Complete Genome-Based Taxonomy." *BioRxiv*, 1–25. https://doi.org/10.1101/771964.

Parks, Donovan H., Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J. Mussig, and Philip Hugenholtz. 2020. "A Complete Domain-to-Species Taxonomy for Bacteria and Archaea." *Nature Biotechnology* (April). https://doi.org/10.1038/s41587-020-0501-8.

Parks, Donovan H., Maria Chuvochina, David W. Waite, Christian Rinke, Adam Skarshewski, Pierre Alain Chaumeil, and Philip Hugenholtz. 2018. "A Standardized Bacterial Taxonomy Based on Genome Phylogeny Substantially Revises the Tree of Life." *Nature Biotechnology* 36 (10): 996. https://doi.org/10.1038/nbt.4229.

Parks, Donovan H., Gene W. Tyson, Philip Hugenholtz, and Robert G. Beiko. 2014. "STAMP: Statistical Analysis of Taxonomic and Functional Profiles." *Bioinformatics* 30 (21): 3123–4. https://doi.org/10.1093/bioinformatics/btu494.

Peabody, Michael A., Thea Van Rossum, Raymond Lo, and Fiona S. L. Brinkman. 2015. "Evaluation of Shotgun Metagenomics Sequence Classification Methods Using in Silico and In Vitro Simulated Communities." *BMC Bioinformatics* 16 (1). https://doi.org/10.1186/s12859-015-0788-5.

Platzer, Alexander, Julia Polzin, Klaus Rembart, Ping Penny Han, Denise Rauer, and Thomas Nussbaumer. 2018. "BioSankey: Visualization of Microbial Communities over Time." *Journal of Integrative Bioinformatics* 15 (4): 1–7. https://doi.org/10.1515/jib-2017-0063.

Powell, Sean, Damian Szklarczyk, Kalliopi Trachana, Alexander Roth, Michael Kuhn, Jean Muller, Roland Arnold, et al. 2012. "EggNOG v3.0: Orthologous Groups Covering 1133 Organisms at 41 Different Taxonomic Ranges." *Nucleic Acids Research* 40 (D1): 284–9. https://doi.org/10.1093/nar/gkr1060.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. "The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools." *Nucleic Acids Research* 41 (D1): 590–6. https://doi.org/10.1093/nar/gks1219.

Rawlings, Neil D., Alan J. Barrett, and Robert Finn. 2016. "Twenty Years of the MEROPS Database of Proteolytic Enzymes, Their Substrates and Inhibitors." *Nucleic Acids Research* 44 (D1): D343–50. https://doi.org/10.1093/nar/gkv1118.

Rees, Jonathan A., and Karen Cranston. 2017. "Automated Assembly of a Reference Taxonomy for Phylogenetic Data Synthesis." *Biodiversity Data Journal* 5 (1). https://doi.org/10.3897/BDJ.5.e12581.

Rho, Mina, Haixu Tang, and Yuzhen Ye. 2010. "FragGeneScan: Predicting Genes in Short and Error-Prone Reads." *Nucleic Acids Research* 38 (20): e191–e191. https://doi.org/10.1093/nar/gkq747.

Rinke, Christian, Maria Chuvochina, Aaron Mussig, Pierre-Alain Chaumeil, David Waite, William Whitman, Donovan Parks, and Philip Hugenholtz. 2020. "A Rank-Normalized Archaeal Taxonomy Based on Genome Phylogeny Resolves Widespread Incomplete and Uneven Classifications." 1–24. https://doi.org/10.1101/2020.03.01.972265.

Rodriguez-R, Luis M., Santosh Gunturu, William T. Harvey, Ramon Rosselló-Mora, James M. Tiedje, James R. Cole, and Konstantinos T. Konstantinidis. 2018. "The Microbial Genomes Atlas (MiGA) Webserver: Taxonomic and Gene Diversity Analysis of Archaea and Bacteria at the Whole Genome Level." *Nucleic Acids Research* 46 (W1): W282–8. https://doi.org/10.1093/nar/gky467.

Saary, Paul, Kristoffer Forslund, Peer Bork, and Falk Hildebrand. 2017. "RTK: Efficient Rarefaction Analysis of Large Datasets." *Bioinformatics (Oxford, England)* 33 (16): 2594–5. https://doi.org/10.1093/bioinformatics/btx206.

Salzberg, Steven L., Arthur L. Deicher, Simon Kasif, and Owen White. 1998. "Microbial Gene Identification Using Interpolated Markov Models." *Nucleic Acids Research* 26 (2): 544–8. https://doi.org/10.1093/nar/26.2.544.

Schoch, Conrad L., Stacy Ciufo, Mikhail Domrachev, Carol L. Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, et al. 2020. "NCBI Taxonomy: A Comprehensive Update on Curation, Resources and Tools." *Database : The Journal of Biological Databases and Curation* 2020 (2): 1–21. https://doi.org/10.1093/database/baaa062.

Sczyrba, Alexander, Peter Hofmann, Peter Belmann, David Koslicki, Stefan Janssen, Johannes Dröge, Ivan Gregor, et al. 2017. "Critical Assessment of Metagenome Interpretation—A Benchmark of Metagenomics Software." *Nature Methods* 14 (11): 1063–71. https://doi.org/10.1038/nmeth.4458.

Segata, Nicola, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. 2012. "Metagenomic Microbial Community Profiling Using Unique Clade-Specific Marker Genes." *Nature Methods* 9 (8): 811–14. https://doi.org/10.1038/nmeth.2066.

Seppey, Mathieu, Mosè Manni, and Evgeny M. Zdobnov. 2020. "LEMMI: A Continuous Benchmarking Platform for Metagenomics Classifiers." *Genome Research*, 1208–16. https://doi.org/10.1101/gr.260398.119.

Siegwald, Léa, Hélène Touzet, Yves Lemoine, David Hot, Christophe Audebert, and Ségolène Caboche. 2017. "Assessment of Common and Emerging Bioinformatics Pipelines for Targeted Metagenomics." *PLoS One* 12 (1): 1–26. https://doi.org/10.1371/journal.pone.0169563.

Stanke, Mario, and Stephan Waack. 2003. "Gene Prediction with a Hidden Markov Model and a New Intron Submodel." *Bioinformatics* 19 (SUPPL.2): 215–25. https://doi.org/10.1093/bioinformatics/btg1080.

Subramanian, Balakrishnan, Shenghan Gao, Martin J. Lercher, Songnian Hu, and Wei Hua Chen. 2019. "Evolview v3: A Webserver for Visualization, Annotation, and Management of Phylogenetic Trees." *Nucleic Acids Research* 47 (W1): W270–5. https://doi.org/10.1093/nar/gkz357.

Tamames, Javier, Marta Cobo-Simón, and Fernando Puente-Sánchez. 2019. "Assessing the Performance of Different Approaches for Functional and Taxonomic Annotation of Metagenomes." *BMC Genomics* 20 (1): 1–16. https://doi.org/10.1186/s12864-019-6289-6.

Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. "MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling." *Nature Methods* 12 (10): 902–3. https://doi.org/10.1038/nmeth.3589.

Velsko, Irina M., Laurent A. F. Frantz, Alexander Herbig, Greger Larson, and Christina Warinner. 2018. "Selection of Appropriate Metagenome Taxonomic Classifiers for Ancient Microbiome Research." *MSystems* 3 (4): 1–41. https://doi.org/10.1128/msystems.00080-18.

Wang, Qiong, George M. Garrity, James M. Tiedje, and James R. Cole. 2007. "Naïve Bayesian Classifier for Rapid Assignment of RRNA Sequences into the New Bacterial Taxonomy." *Applied and Environmental Microbiology* 73 (16): 5261–7. https://doi.org/10.1128/AEM.00062-07.

Wemheuer, Franziska, Jessica A. Taylor, Rolf Daniel, Emma Johnston, Peter Meinicke, Torsten Thomas, and Bernd Wemheuer. 2020. "Tax4Fun2: Prediction of Habitat-Specific Functional Profiles and Functional Redundancy Based on 16S RRNA Gene Sequences." *Environmental Microbiome* 15 (1): 11. https://doi.org/10.1186/s40793-020-00358-7.

Wheeler, David L., Tanya Barrett, Dennis A. Benson, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, et al. 2008. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 36 (SUPPL.1): D13–21. https://doi.org/10.1093/nar/gkm1000.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "Comment: The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3: 1–9. https://doi.org/10.1038/sdata.2016.18.

Wilkinson, Toby J., Sharon A. Huws, Joan E. Edwards, Alison H. Kingston-Smith, Karen Siu-Ting, Martin Hughes, Francesco Rubino, Maximillian Friedersdorff, and Christopher J. Creevey. 2018. "CowPI: A Rumen Microbiome Focussed Version of the PICRUSt Functional Inference Software." *Frontiers in Microbiology* 9 (May): 1–10. https://doi.org/10.3389/fmicb.2018.01095.

Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal. 2017. "Good Enough Practices in Scientific Computing." Edited by Francis Ouellette. *PLOS Computational Biology* 13 (6): e1005510. https://doi.org/10.1371/journal.pcbi.1005510.

Woloszynek, Stephen, Zhengqiao Zhao, Gregory Ditzler, Jacob R. Price, Erin R. Reichenberger, Yemin Lan, Jian Chen, et al. 2018. "Analysis Methods for Shotgun Metagenomics." In *Theoretical and Applied Aspects OfSystems Biology*, edited by F. Alves Barbosa da Silva, N. Carels, and F. Paes Silva Junior, 71–112. https://doi.org/10.1007/978-3-319-74974-7_5.

Xu, Xihui, Raphy Zarecki, Shlomit Medina, Shany Ofaim, Xiaowei Liu, Chen Chen, Shunli Hu, et al. 2019. "Modeling Microbial Communities from Atrazine Contaminated Soils Promotes the Development of Biostimulation Solutions." *ISME Journal* 13 (2): 494–508. https://doi.org/10.1038/s41396-018-0288-5.

Yang, Bo, Yong Wang, and Pei Yuan Qian. 2016. "Sensitivity and Correlation of Hypervariable Regions in 16S RRNA Genes in Phylogenetic Analysis." *BMC Bioinformatics* 17 (1): 1–8. https://doi.org/10.1186/s12859-016-0992-y.

Ye, Simon H., Katherine J. Siddle, Daniel J. Park, and Pardis C. Sabeti. 2019. "Benchmarking Metagenomics Tools for Taxonomic Classification." *Cell* 178 (4): 779–94. https://doi.org/10.1016/j.cell.2019.07.010.

Yi, Lynn, Harold Pimentel, Nicolas L. Bray, and Lior Pachter. 2018. "Gene-Level Differential Analysis at Transcript-Level Resolution." *Genome Biology* 19 (1): 1–11. https://doi.org/10.1186/s13059-018-1419-z.

Yilmaz, Pelin, Laura Wegener Parfrey, Pablo Yarza, Jan Gerken, Elmar Pruesse, Christian Quast, Timmy Schweer, Jörg Peplies, Wolfgang Ludwig, and Frank Oliver Glöckner. 2014. "The SILVA and 'All-Species Living Tree Project (LTP)' Taxonomic Frameworks." *Nucleic Acids Research* 42 (D1): 643–48. https://doi.org/10.1093/nar/gkt1209.

Yin, Yanbin, Xizeng Mao, Jincai Yang, Xin Chen, Fenglou Mao, and Ying Xu. 2012. "DbCAN: A Web Resource for Automated Carbohydrate-Active Enzyme Annotation." *Nucleic Acids Research* 40 (W1): 445–51. https://doi.org/10.1093/nar/gks479.

Yoon, Seok Hwan, Sung min Ha, Jeongmin Lim, Soonjae Kwon, and Jongsik Chun. 2017. "A Large-Scale Evaluation of Algorithms to Calculate Average Nucleotide Identity." *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology* 110 (10): 1281–6. https://doi.org/10.1007/s10482-017-0844-4.

Zhang, Huangkai, Shenghan Gao, Martin J. Lercher, Songnian Hu, and Wei Hua Chen. 2012. "EvolView, an Online Tool for Visualizing, Annotating and Managing Phylogenetic Trees." *Nucleic Acids Research* 40 (W1): 569–72. https://doi.org/10.1093/nar/gks576.

Zhao, Yongan, Haixu Tang, and Yuzhen Ye. 2012. "RAPSearch2: A Fast and Memory-Efficient Protein Similarity Search Tool for Next-Generation Sequencing Data." *Bioinformatics* 28 (1): 125–6. https://doi.org/10.1093/bioinformatics/btr595.

Zhu, Wenhan, Alexandre Lomsadze, and Mark Borodovsky. 2010. "Ab Initio Gene Identification in Metagenomic Sequences." *Nucleic Acids Research* 38 (12): 1–15. https://doi.org/10.1093/nar/gkq275.