11 Bioinformatics for Studying Environmental Microorganisms

Adriana M. Fróes and Bas E. Dutilh

CONTENTS

11.1 INTRODUCTION

The term "bioinformatics" was first introduced in 1970 by Ben Hesper and Paulien Hogeweg to indicate the study of information processing in biotic systems (Hesper and Hogeweg 1970). Since the late 1980s, bioinformatics has mostly been used to refer to computational analysis of biological data (Hogeweg 2011), and has brought a revolution to biology. Bioinformatics is an interdisciplinary field that combines computer science, statistics, mathematics, and biology. Current-day biological research is highly data driven, generated by microarray gene expression measurements, high-throughput sequencing technologies, and proteomic mass spectrometry, to name a few examples. Bioinformaticists develop critical computational tools for organizing, visualizing, storing, and analyzing the data obtained (Kesh 2004). As these large-scale data generation methods become more accessible to researchers around the world, the biological questions that may be addressed become more advanced, and the amount and complexity of biological data increase. For example, the speed and throughput of DNA sequencing platforms have increased faster than the computational power required to analyze the resulting data (Carlson 2003), making research projects increasingly dependent on innovative bioinformatics solutions. This is known as the "bioinformatics bottleneck." Bioinformatics has become an intrinsic component in molecular biological research, in which it plays an important role in the large-scale analysis of sequence data. These sequences may represent proteins (amino acids), or genes and genomes (DNA), and are stored in large databases, maintained

by, among others, the European Molecular Biology Laboratory (EMBL), the National Center for Biotechnology Information (NCBI, GenBank), and the DNA Data Bank of Japan (DDBJ). The Basic Local Alignment Search Tool (BLAST) was introduced in 1990 as a fast computational algorithm for searching similar sequences within these databases (Altschul et al. 1990), and it is the most widely used bioinformatic tool today.

Microorganisms (microbes) comprise the majority of the biological diversity on the planet and consist of all life forms that are so small that they cannot be seen by the naked eye, including bacteria, archaea, eukaryotes, and viruses. Natural microbial communities are complex biological systems that contain a wide variety of species. Together, the combined genomic potential of these species is known as their metagenome, and depending on the environment, it encodes a remarkable functional diversity that allows the community to respond to environmental changes. The traditional method to tap this reservoir is by culturing individual microorganisms and characterizing them one by one. However, the large majority of microorganisms has not yet been cultured and cannot be grown using standard laboratory culturing techniques. High-throughput sequencing technologies have provided a new way of analyzing metagenomes, by sequencing DNA directly isolated from an environmental sample. This has led to the very data-intensive research field of metagenomics, which provides a powerful tool to bypass the limitation of culture-dependent analyses.

Metagenomics employs untargeted sequencing of environmental DNA to analyze mixed communities of microbes. Here, untargeted means that random regions of the microbial genomes in the community are sampled, also known as shotgun metagenomics (a shotgun fires a random hail of pellets). This is in contrast with targeted approaches, such as environmental amplicon sequencing, which first amplify a specific gene or genetic region from the metagenome and use it as a marker to analyze the diversity of the sampled microbes. The small subunit of ribosomal RNA (SSU rRNA) is the most commonly analyzed marker gene in environmental amplicon sequencing studies (see later), but other genes have been used as well. The SSU rRNA gene is universally conserved and consists of 16S rRNA in Bacteria and Archaea, and 18S rRNA in Eukaryotes. Viruses do not contain a universally conserved marker gene, so the most promising way to analyze the viruses in a microbial community is by using shotgun metagenomics. In 2002, the first shotgun metagenomics study analyzed uncultured marine viral communities (Breitbart et al. 2002).

The two classical questions in metagenomics that are asked about the sampled microorganisms are as follows: "Who is there?" and "What they are doing?" (Handelsman 2004). Because shotgun metagenomes provide a uniquely rich profile of the genetic content of these communities, they have the potential to answer both these questions, and many more. Traditionally, the most straightforward approach to answering the classical questions in metagenomics involves aligning the metagenomic sequencing reads to a reference database of known, annotated sequences. For example, these known sequences may be derived from sequencing cultured strains or by performing cloning experiments. The first question can then be answered by looking at the taxonomic annotations of the hits in the reference database, whereas the second question can be answered by assessing their functional annotations. The advantage of this approach is that the genes that are present in the metagenome can immediately be placed in the context of existing knowledge about these taxa or functions, including all the related scientific literature. However, the disadvantage is that not all sequences can be aligned to the reference database: in some cases, up to 99% of the sequences in a shotgun metagenome cannot be annotated, especially if the sample was taken from an understudied environment (Mokili et al. 2012). These unmappable sequences, known as "unknowns," reflect the enormous natural diversity of microorganisms that we are only beginning to unveil with metagenomics, and highlight how much remains to be discovered in biology (Dutilh 2014).

One of the earliest shotgun metagenomics studies analyzed the microbes in four sites within the Sargasso Sea, a well-studied region of the Atlantic Ocean close to Bermuda (Venter et al. 2004). DNA was isolated from microbes with a size range from 0.1 to 3.0 μm, generating almost 2 million sequencing reads, and over 1.6 billion nucleotides of data. Based on sequence relatedness and unique SSU rRNA gene counts, the analysis suggested that these DNA fragments were derived from

at least 1800 species, including 148 previously unknown bacteria. Bioinformatic analysis of the data identified spatial variation in species richness and relative abundance among the four sampled sites. Moreover, the analysis identified over 1.2 million potential unique protein-coding genes. This is an astonishing number, especially if we consider that at the time, only 140,000 proteins were available in the reference protein sequence database. Among these 1.2 million potential new proteins, at least 782 new rhodopsin-like photoreceptors were identified, confirming the importance of this type of phototrophy in the open sea (Venter et al. 2004). In this chapter, we will introduce some of the bioinformatics behind metagenomic analyses.

11.2 DATABASES

A biological database is a collection of information, or data from a biological system, stored in a computer-readable format. As mentioned previously, biology is increasingly becoming a technology-driven science, and databases have become indispensable to store not only primary data but also the results of (bioinformatic) experiments generated by different research projects around the world. Examples that are relevant to analyzing microbial communities include databases of (1) nucleotide sequences, such as next-generation sequencing reads, genes, or genomes (GenBank—http://www.ncbi.nlm.nih.gov/genbank, DDBJ—www.ddbj.nig.ac.jp, MicrobesOnline [Dehal et al. 2009]—www.microbesonline.org); (2) amino acid sequences of proteins (UniProtKB/Swiss-Prot: reviewed, manually annotated; UniProtKB/TrEMBL: unreviewed, automatically annotated); (3) protein structures (Protein DataBank [Berman et al. 2003]—www.wwpdb.org); (4) information about interactions between proteins (Search Tool for the Retrieval of Interacting of Genes/Proteins [von Mering et al. 2003]—string-db.org); (5) metabolic modules or groups of proteins that work together in a metabolic pathway (the Kyoto Encyclopedia of Genes and Genomes [KEGG] [Kanehisa 2002]—www.genome.jp/kegg); (6) protein families (Pfam [Finn et al. 2014]—pfam.xfam.org); (7) scientific literature (PubMed—www.ncbi.nlm.nih.gov/pubmed); and so on. Moreover, some specialized databases focus on particular organisms or groups of organisms (e.g., *Escherichia coli* or viruses), particular types of nucleic acids or proteins (transfer RNA [tRNA], or gene regulatory regions).

11.2.1 Database Accessibility

The information in a database is stored in a computer-readable format, so that it can be accessed and retrieved automatically. To facilitate this access, each data element in the database has a unique identifier or accession number that can also be used to refer to it in scientific publications. Obviously, the identifier for a certain data element should be conserved, so that the same element can be retrieved from the database at any later point in time. For manual access, easy searching options should be available to search the database and access the stored data quickly. Moreover, automated access options to the database can be made available through an application programming interface, so that external computer programs and bioinformatic tools can easily perform high-throughput searches or queries. Finally, it is often very practical for bioinformaticists to be able to download the full database content at once, so that they can perform experiments with the data on their own computer. Because the information present in a database might change, such downloads are really snapshots of the data at a certain point in time. Thus, when accessing data in any database, it is always important to record the date and version of the database, and mention this in communications with other scientists, for example, when writing scientific articles.

Every entry in the database is annotated with relevant information, and it is important to have a consistent annotation of the entries in the database, with as little missing information as possible. The data in databases may be collected by conducting (high-throughput) experiments, by browsing through (scientific) literature, or by collecting or interpreting existing datasets. Often, databases obtain their information by automated methods such as robotized experiments or large-scale bioinformatic mining of new or existing data. It may be very important to scientists to know how a

specific piece of information was obtained, because different ways of obtaining information may be more or less reliable. Therefore, a good database is transparent about the background and history of every entry. For example, if certain information was manually curated, the annotation could tell the user who performed the quality checks. Or, if the information was automatically obtained, the annotation should contain details of the tool or program that was used for this. Finally, an important feature of science is that it is continually moving forward as new knowledge becomes available. Therefore, to avoid getting outdated, a good database is regularly updated as new information becomes available. However, it should always be possible to retrieve previous entries, or even previous versions of the entire database, so that older experiments can be repeated and validated.

11.2.2 FASTA Format

Sequence data files can be retrieved from databases in various formats. The most commonly used is the FASTA format, which was initially designed for the FASTA sequence similarity search program (Lipman and Pearson 1985), but is now very widely used. Key to the success of the FASTA file format is that it is very basic, providing only the DNA or protein sequences and their identifiers, with a minimum of annotation information. Each sequence in a FASTA file begins with a one line description, followed by one or more lines containing the sequence itself. The description line is distinguished from the lines containing the sequence by a greater-than symbol ">" at the beginning of the line, immediately followed by a unique identifier of the sequence (no space in between). In the example in Figure 11.1, "gi|667481592|gb|KM000119.1|" is the unique identifier of the displayed sequence, consisting of a "GenInfo identifier" (gi: 667481592) and a database identifier (here GenBank, gb: KM000119.1). Other database acronyms can also appear in the header, including "embl," "dbj," and "sp." In some cases, the sequence identifier is followed by a description, in the exemplified case, a description of the species (uncultured phage crAssphage) and clone (clone TS4.2-F23R23-35). However, because there is a space after the identifier, these optional elements of the header line are not part of the unique sequence identifier and may be left out. FASTA files that contain metagenomic sequences often include comments on the header line that describe the metadata, such as the sampling location and other characteristics. In the lines containing the sequence itself, spaces and new-line characters may be added to make the FASTA file more readable to the human eye, but these space characters have no meaning for the biological sequence. As a result, a single sequence in a FASTA file can cover several lines, as illustrated in Figure 11.1, and allow the length of the sequence to be quickly observed (485 nucleotides). Some FASTA files may also contain the entire sequence on a single line with no space characters; this is also known as the FASTA-wide format.

11.2.3 GenBank Format

GenBank is a sequence database at the NCBI (http://www.ncbi.nlm.nih.gov/genbank/) that has been existed since 1985 (Burks et al. 1985, Benson et al. 2014). It supports biological and bibliographic annotations, and is accessible through Entrez (Global Query Cross-Database Search System), a central search engine on the website of the NCBI. Currently, it contains more than 654 billion nucleotides in sequences from over 280,000 formally described species (Benson et al. 2014). The GenBank format

```
>gi|667481592|gb|KM000119.1| Uncultured phage crAssphage clone TS4.2-F23R23-35
GAGATATTAT AATGCTATTG CTTATCTTGA AGATGAAAAT ATTATTAAAA GAACTAATAT TAGAAATATT
TATGTTGTTA ATCCTATTTA CATATTTAGA GGTGATGTTA ACAAACTTAT TAATATTATT AGTGAAGCTA
AATTAATAAA AACTTTTGAT GATAAAGATA GACTTATAGT TGATAAATTT ATTTTATTTA AAAATGATAC
TGATAAAGGT ATTGTAATTG CGAATAAAGA TTTGTATGCA ACTGAAGTTA TAGATATTGG TGAGGACTGA
GTTAAATGCG ATGATAAAAA TGATGATAAT TATAAAGATG AGGGCGAAGA TAATGGAAAT AATGAAGATG
AAGGTAATAA AGGTAACCAT AATGAAAATG AAATGAAAAT GAAATGAAAA GTTATAATGA GAGTTATACC
GAAAATAATA GCAATGGGAA TAATAATGAA CACGAACGCA ACGGGTATAT AGATAATGGA CAGAA
```

FIGURE 11.1 Example of a sequence in FASTA format.

is a rich format for storing both sequences and annotations, and allows detailed annotations for each sequence to be included (Figure 11.2). An entry in the GenBank format starts with a line containing the word "LOCUS," that is followed by a number of annotation lines. The sequence itself is listed at the bottom, and the entry is closed with two forward-slashes "//." GenBank file consists of three main sections: the header section, the features, and the sequence. Within these sections, there is a set order in which the annotation attributes are listed (note that not all annotation attributes have to be included for each sequence). The header section contains attributes such as keywords, accession and version numbers, source organisms and taxonomic classification, comments, and references to journal articles related to the sequence. The features section contains elements encoded on the sequence, such as genes, or regulatory elements. A line containing the word "ORIGIN" marks the start of the sequence section, and this section includes numbers indicating the position in the sequence (Figure 11.2).

```
LOCUS       KM000119                 485 bp    DNA     linear   ENV 28-JUL-2014
DEFINITION  Uncultured phage crAssphage clone TS4.2-F23R23-35 orf00102
            (orf00102) gene, partial cds.
ACCESSION   KM000119
VERSION     KM000119.1  GI:667481592
KEYWORDS    ENV.
SOURCE      uncultured phage crAssphage
  ORGANISM  uncultured phage crAssphage
            Viruses; unclassified phages; environmental samples.
REFERENCE   1  (bases 1 to 485)
  AUTHORS   Dutilh, B.E.
  TITLE     A highly abundant bacteriophage discovered in the unknown sequences
            of human faecal metagenomes
  JOURNAL   Nat Commun (2014)
FEATURES             Location/Qualifiers
     source          1..485
                     /organism="uncultured phage crAssphage"
                     /mol_type="genomic DNA"
                     /isolation_source="feces"
                     /host="Homo sapiens"
                     /db_xref="taxon:1211417"
                     /clone="TS4.2-F23R23-35"
                     /environmental_sample
     gene            <1..280
                     /gene="orf00102"
     CDS             <1..280
                     /gene="orf00102"
                     /codon_start=2
                     /transl_table=11
                     /product="orf00102"
                     /protein_id="AIG55372.1"
                     /db_xref="GI:667481593"
                     /translation="RYYNAIAYLEDENIIKRTNIRNIYVVNPIYIFRGDVNKLINIIS
                     EAKLIKTFDDKDRLIVDKFILFKNDTDKGIVIANKDLYATEVIDIGED"
ORIGIN
        1 gagatattat aatgctattg cttatcttga agatgaaaat attattaaaa gaactaatat
       61 tagaaatatt tatgttgtta atcctattta catatttaga ggtgatgtta acaaacttat
      121 taatattatt agtgaagcta aattaataaa aacttttgat gataaagata gacttatagt
      181 tgataaattt attttattta aaaatgatac tgataaaggt attgtaattg cgaataaaga
      241 tttgtatgca actgaagtta tagatattgg tgaggactga gttaaatgcg atgataaaaa
      301 tgatgataat tataaagatg agggcgaaga taatggaaat aatgaagatg aaggtaataa
      361 aggtaaccat aatgaaaatg aaatgaaaat gaaatgaaaa gttataatga gagttatacc
      421 gaaaataata gcaatgggaa taataatgaa cacgaacgca acgggtatat agataatgga
      481 cagaa
//
```

FIGURE 11.2 Example of a sequence in GenBank format.

11.2.4 General Databases

Taxonomic and functional profiling of environmental sequencing samples depends on the availability of a good reference database. As mentioned previously, many general databases for nucleotide and protein sequences are publicly available, such as DDBJ, EMBL-European Bioinformatics Institute (EBI), GenBank, the KEGG (Kanehisa 2002), and SEED (Overbeek et al. 2005). These databases generally contain validated sequences that have been studied in reasonable detail, and alignment of metagenomic sequencing reads to the database sequences allows the annotations to be transferred from the database to the newly sequenced reads. The wide variety of taxonomic groups and functional genes present in these general databases allows a wide variety of metagenomic sequencing reads to be taxonomically and functionally annotated.

11.2.5 Specialized Enzyme Category Databases

Some databases specialize in one category of genes, such as specific functions that may be particularly important for analyzing environmental microbes. For example, a database of enzymes involved in carbohydrate metabolism is the Carbohydrate-Active enZymes database (Lombard et al. 2014), or databases of antibiotic resistance genes include the Antibiotic Resistance Database (Liu and Pop 2009) and the Comprehensive Antibiotic Resistance Database (McArthur et al. 2013). These databases are very useful when you are searching for specific functions in the environmental sequences, because they contain all known sequences for these functions that have been described in the literature. Thus, by aligning metagenomic reads to the sequences in a specialized enzyme category database, homologs of these sequences can be identified and their annotations transferred. However, for most of the genes in natural microbial communities, we have no idea of their biological function. Using specialized enzyme category databases can only tell you which known functions are found, and this might be an underestimate of the true functional content of, for example, the carbohydrate metabolism genes, or the antibiotic resistance genes that are really present. Conversely, screening an environment for functions by using specialized enzyme category databases might also lead to overestimations. This can happen if the sequences in the metagenome look like (parts of) genes in the database but have a different function, for example, if the genes contain a conserved region that is present in different genes, or a region with low complexity (a repetitive sequence) that might be present in unrelated genes. If this happens, a metagenomic DNA sequence may be interpreted as having the enzymatic function you are interested in, although it is actually not functionally related to the genes in the enzyme category database. One way to avoid such biases is to validate the subset of reads that gave a hit to the specialized enzyme category database by querying them to a more comprehensive, general database, such as GenBank. If the read had a spurious hit in the specialized enzyme category database, this will be revealed if the second search yields a better alignment score to an unrelated sequence in the general database. Because this validation search is only performed with the fraction of the original dataset that produced a hit against the smaller, specialized enzyme category database, the computational burden is greatly reduced. This approach allows the use of more sensitive similarity search algorithms such as tblastx to annotate metagenomic reads (Dutilh et al. 2013).

11.2.6 SSU rRNA: A Universal Taxonomic Marker Gene for Cellular Organisms

Taxonomic marker genes have distinct sequence characteristics that enable reliable taxonomic classification. The SSU rRNA gene is universally shared by all cellular life forms, consisting of a 16S rRNA gene for bacteria and archaea, and 18S rRNA for eukaryotes. It is the most often used taxonomic marker gene and can be used to discriminate different species, or build phylogenies to identify and distinguish taxonomic groups (Woese and Fox 1977). Several SSU databases are dedicated to this taxonomic marker gene, including the Ribosomal Database Project (RDP) (Cole et al. 2014), ARB-SILVA (Quast et al. 2013), and GreenGenes (DeSantis et al. 2006). Since the

early 1990s (Lane et al. 1985, Schmidt et al. 1991), many environmental DNA sequencing projects have focused on identifying naturally occurring SSU rRNA genes. These studies used polymerase chain reaction (PCR) primers targeting conserved regions in the gene, to amplify the variable regions lying in between. Although these experiments have revealed the extraordinary evolutionary diversity of environmental microorganisms, it may be difficult to draw firm conclusions about the functional potential of a microbial community based on these SSU rRNA surveys. It is becoming increasingly clear that microbial genomes are often highly plastic, and the functional content of a genome may vary widely between strains, even if the SSU rRNA gene is (almost) identical. Thus, drawing conclusions about the functional properties of a microbial community based on a profile of SSU rRNA sequences is questionable, although this is still common practice in environmental microbiology research, and recently a bioinformatic tool has been introduced that makes this kind of analysis possible at a large scale (Langille et al. 2013). Besides the disparity between the function and the taxonomy of organisms based on their SSU rRNA sequence, several other biases are worth mentioning. First, depending on the variable regions that are amplified, different conclusions may be reached about the genetic diversity, richness, and diversity of a microbial community (Schloss 2010). Second, not all microbes contain the same copy number of the SSU rRNA gene within their genome (Klappenbach et al. 2001), so counting the abundance of this gene in a dataset may lead to overestimation of the microbes with many copies relative to those with only a single copy. Third, there is variability in PCR amplification efficiency of the gene between species (Engelbrektson et al. 2010). Notwithstanding these biases, SSU rRNA remains by far the most popular taxonomic marker gene for analyzing the species composition of environmental communities. Based on the length ratio of SSU rRNA genes (~1500 nucleotides) and bacterial genomes (~3 million nucleotides), an average metagenomic dataset consisting of short sequencing reads is expected to contain one in every 2000 reads aligning to SSU rRNA, and this number is often surprisingly accurate. Thus, SSU rRNA reads in shotgun metagenomes can be used to analyze which species are present in a random sample of short nucleotide sequences from the environment.

11.2.7 Signature Genes

Although SSU rRNA is the most commonly used taxonomic marker gene, information about the taxonomic composition of a microbial community may also be derived from other genes. For example, for fungi, the internal transcribed spacer region is the marker of choice for the exploration of fungal diversity in environmental samples (Kõljalg et al. 2013), and Cyanobacteria contain unique signature genes involved in photosynthesis (Martin et al. 2003). Signature genes occur only within a specific taxonomic clade, and such specific genes occur at all levels in the tree of life (Figure 11.3), from species and genera to phyla and kingdoms (Dutilh et al. 2008b). Thus, the presence of a signature gene in a metagenome indicates that an organism belonging to that clade is present in the sampled environment, and this can be used to taxonomically profile the sampled microbial community. A tool to create such a taxonomic profile is Signature (Dutilh et al. 2008a), a web server for taxonomic characterization of sequencing samples using signature genes based on gene content. Another approach is MetaPhlAn (Segata et al. 2012), which predefined over 400,000 taxonomic marker genes for the taxonomic characterization of metagenomic samples. These genes form a relatively small database that can be used to align the metagenomic reads and create a taxonomic profile of the community.

11.3 SIMILARITY SEARCH ALGORITHMS

The two classical questions in metagenomics are commonly answered by aligning the metagenomic sequences to a reference database of annotated sequences and transferring the taxonomic or functional annotation from the closest database hit to the metagenomic sequence. Keys in this approach are both the similarity search algorithm and the reference database that are used. We have discussed reference databases earlier. For similarity search algorithms, there is generally a trade-off

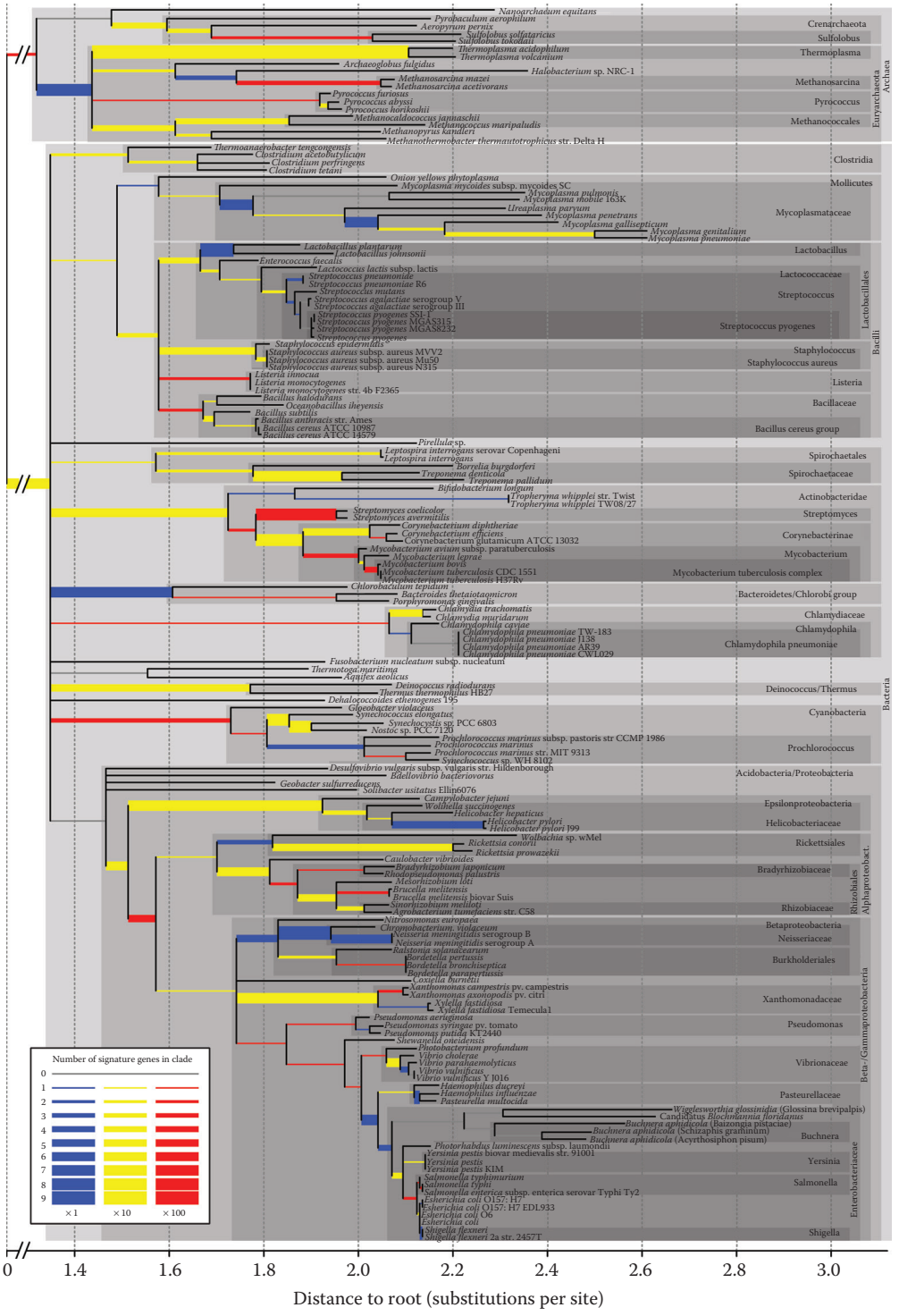

FIGURE 11.3 Number of signature genes in different branches of the tree of life. Branch widths and colors indicate the number of signature genes found for each node (see legend). (Adapted from Dutilh, B.E. et al., *Mol. Biol. Evol.*, 25, 1659–1667, 2008b.)

between sensitivity and speed. For years, the BLAST has offered the perfect balance between these two qualities (Altschul et al. 1990). Briefly, BLAST works in two steps. First, it finds short "seeds" (k-mers or subsequences of length k; for example, $k = 11$ for nucleotides or $k = 3$ for amino acids) that are matched between the query sequence and the database (Figure 11.4). This is a very fast step because such a database can be indexed and stored in the rapidly accessible memory (RAM) of a computer. Second, these k-mer seeds are extended in both directions to maximize the local region where the sequences can be homologously aligned. There are five main BLAST programs that allow different types of query and database sequences to be compared:

- "blastn" compares a nucleotide query sequence to a nucleotide database. Thus, it can answer questions such as the following: "From what gene family is my metagenomic sequencing read derived?"
- "blastp" compares a protein query sequence to a protein database. Thus, it can answer questions such as the following: "What other proteins are related to my query protein?"
- "blastx" compares a nucleotide query sequence to a protein database, by first translating the DNA query to protein in six frames. Because protein sequences are often more conserved in evolution than nucleotide sequences, blastx is more sensitive than blastn if you have a nucleotide query.
- "tblastn" compares a protein query sequence to a nucleotide database, by first translating the DNA sequences in the database to protein in six frames.
- "tblastx" compares a nucleotide query sequence to a nucleotide database, but does so at the level of protein sequences by translating both the query and database sequences to protein in six frames, and comparing all possible combinations. Thus, tblastx is the most computationally intensive BLAST algorithm, but also more sensitive than blastn at identifying evolutionarily distant homologs, if both the query and the database contain nucleotide sequences.

Many parameters in the BLAST algorithm can be adjusted, and it is worthwhile to investigate the optimal settings for your experiment (Kerfeld and Scott 2011). Because the alignment extension step in the BLAST algorithm is rather computationally expensive, it is often the rate-limiting step in metagenomic analyses where tens of thousands to millions of sequences are compared to the database. Thus, metagenome annotation is often performed with faster similarity search algorithms, which comes at a cost of sensitivity and leaves evolutionarily distant homology between the query and database sequences undetected ("unknowns"). For example, we will see below that the metagenomic analysis pipeline of Metagenomics Rapid Annotation using Subsystem Technology (MG-RAST) employs the BLAST-like alignment tool (BLAT) algorithm (Kent 2002) to facilitate faster similarity searches. Another fast algorithm is Real-Time Metagenomics (RTMG) (Edwards et al. 2012) that

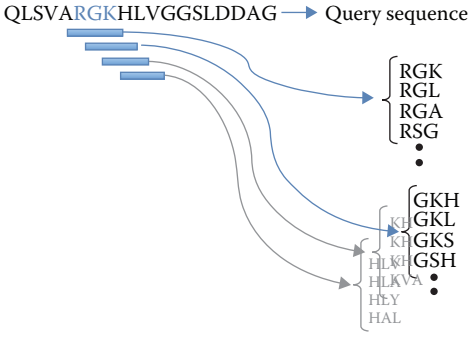

FIGURE 11.4 Example illustrating the first phase of BLAST, where a k-mer (here: $k = 3$) or "three-letter word" from the query sequence is compared to a list that contains the indexes of all instances of that word existing in the database.

only stores and searches exact *k*-mers ($k = 7$–12 amino acids), allowing no evolutionary divergence in the sequence. By bypassing the slow alignment extension step, RTMG is much less sensitive, but also orders of magnitude faster than BLAST. Finally, Bowtie is a very fast algorithm for aligning large numbers of nucleotide query sequences to a database that indexes the reference database with a Burrows–Wheeler index that limits the amount of computational memory required by the program (Langmead and Salzberg 2012). Bowtie2 can also extend the alignment to the full length of the read and produce gapped alignments like BLAST. With the huge amounts of sequencing data generated by the modern high-throughput sequencing machines, the development of efficient, innovative tools to analyze this data has become one of the major bioinformatic goals in metagenome analysis.

11.4 METAGENOMIC ANALYSIS

Metagenomics analyzes the combined genetic information of whole communities of microorganisms. Metagenomic research projects often generate enormous amounts of sequence data, especially in the present era of high-throughput DNA sequencing technologies. It is critical that these datasets are published in openly accessible databases, to enable scientific transparency, to allow the scientific community to reproduce experiments and analyses, and to be able to use the data for any other unforeseen purposes, such as answering new research questions in the future. When a scientific article is published, many journals require the associated datasets to be made available through publicly available data repositories. For metagenomics, there are several options available, including MG-RAST (Meyer et al. 2008), EBI Metagenomics (Hunter et al. 2014), Integrated Microbial Genomes with Microbial samples (IMG/M, Markowitz et al. 2014), and the environmental nucleotide section of the GenBank BioProject Database. These databases store several types of datasets, including shotgun metagenomes and environmental amplicon sequencing data. Two databases that focus on viral shotgun metagenomics are MetaVir (Roux et al. 2014) and the Viral Informatics Resource for Metagenome Exploration (VIRoME, Wommack et al. 2012). All the databases require the uploaded information to be annotated with standard metadata fields, such as the Minimum Information about a (Meta)Genome Sequence (MIGS/MIMS) metadata (Field et al. 2011). This minimum annotation information includes the sample name, the location and date that the sample was taken, the sequencing method, and details of the sampled environment, and can be supplemented with other specific information that makes it easier for future researchers to work with the datasets. The Genomic Standards Consortium (http://gensc.org/) (Field et al. 2008) is an international group working to standardize the description of genomes and metagenomes, and the exchange of genomic data and metadata.

11.4.1 Profiling

After annotating metagenomic reads by aligning them to a reference database, the annotations can be summarized to provide a taxonomic or functional profile of the metagenome. This can be done with programs such as MEGAN (Huson et al. 2007) or CARMA (Gerlach and Stoye 2011). To account for ambiguities in the sample, such as novel species that may only be distantly related to the organisms in the database, the annotations for each read are generalized by identifying the last common ancestor of the organisms identified among the hits for that read. The generalized annotations of all identified reads in the metagenome are then summarized by displaying them on a taxonomic tree or functional diagram (Huson and Mitra 2012). Figure 11.5 shows the results of two shotgun metagenomes from hospital sewage (unpublished data) that were taxonomically profiled based on the NCBI taxonomy (Figure 11.5a), and functionally profiled based on SEED subsystem classification (Figure 11.5b) (Overbeek et al. 2005). To perform a functional analysis, MEGAN analyzes the BLAST results and reports the number of reads assigned to each functional role. MEGAN also includes a module called "KEGGviewer" that allows the functional profile of the metagenome to be viewed in the context of pathways (Mitra et al. 2011). Figure 11.5b shows an example of the glycolysis/gluconeogenesis pathway, in which boxes indicate enzymes that are colored to indicate the number of reads assigned.

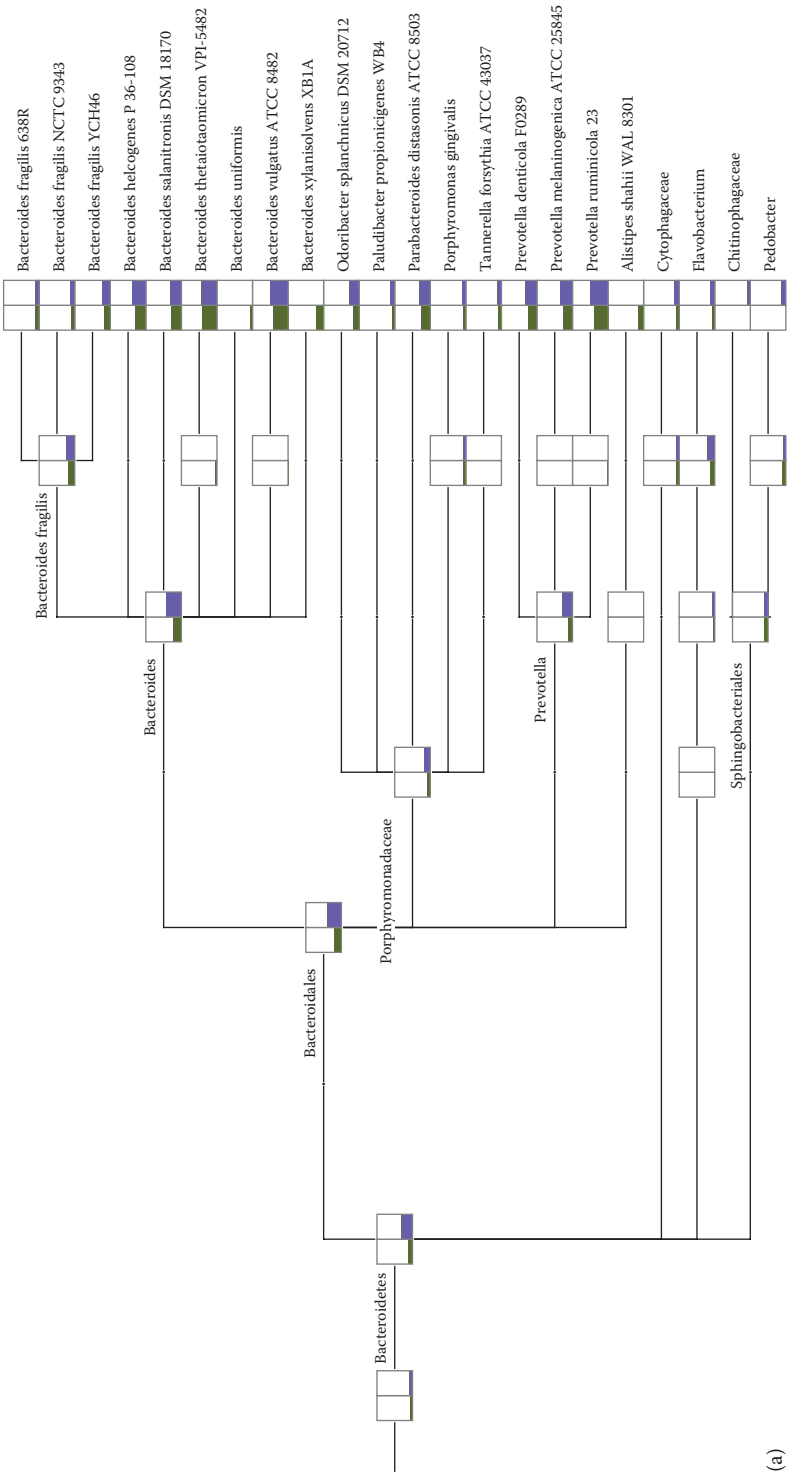

FIGURE 11.5 Analysis of two microbial metagenomes from hospital sewage (unpublished data) with MEGAN (Huson, D.H. et al., *Genome. Res*., 17, 377–386, 2007). (a) Part of the taxonomic overview (Bacteroidetes branch). Boxes indicate taxonomic groups and are colored according to the number of reads in each metagenome. (*Continued*)

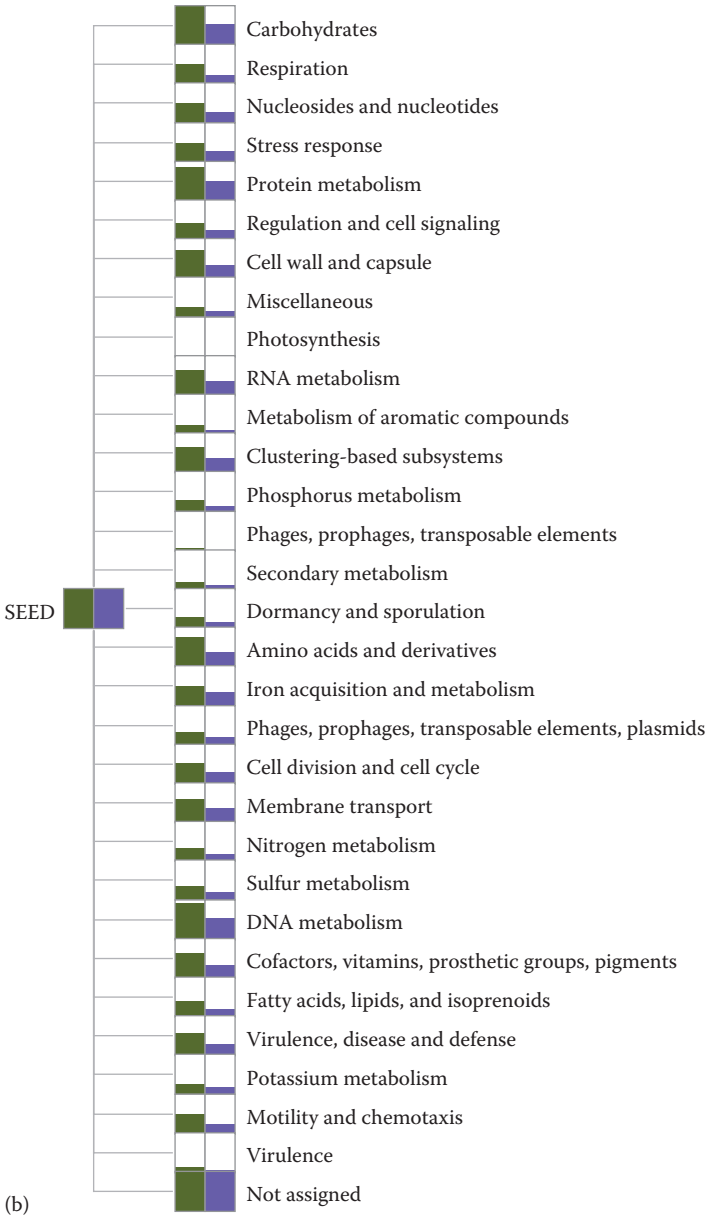

FIGURE 11.5 (Continued) Analysis of two microbial metagenomes from hospital sewage (unpublished data) with MEGAN (Huson, D.H. et al., *Genome. Res*., 17, 377–386, 2007). (b) An overview of the functions identified in the metagenomes represented as SEED level-1 subsystems. Boxes indicate functional categories and are colored according to the number of reads in each metagenome. (Overbeek, R. et al., *Nucl. Acids. Res*., 33 (17), 5691–5702, 2005; Huson, D.H. and Mitra, S., *Evol. Genomics*.: *Stat. Comput. Meth*., 856, 415–429, 2012.) (*Continued*)

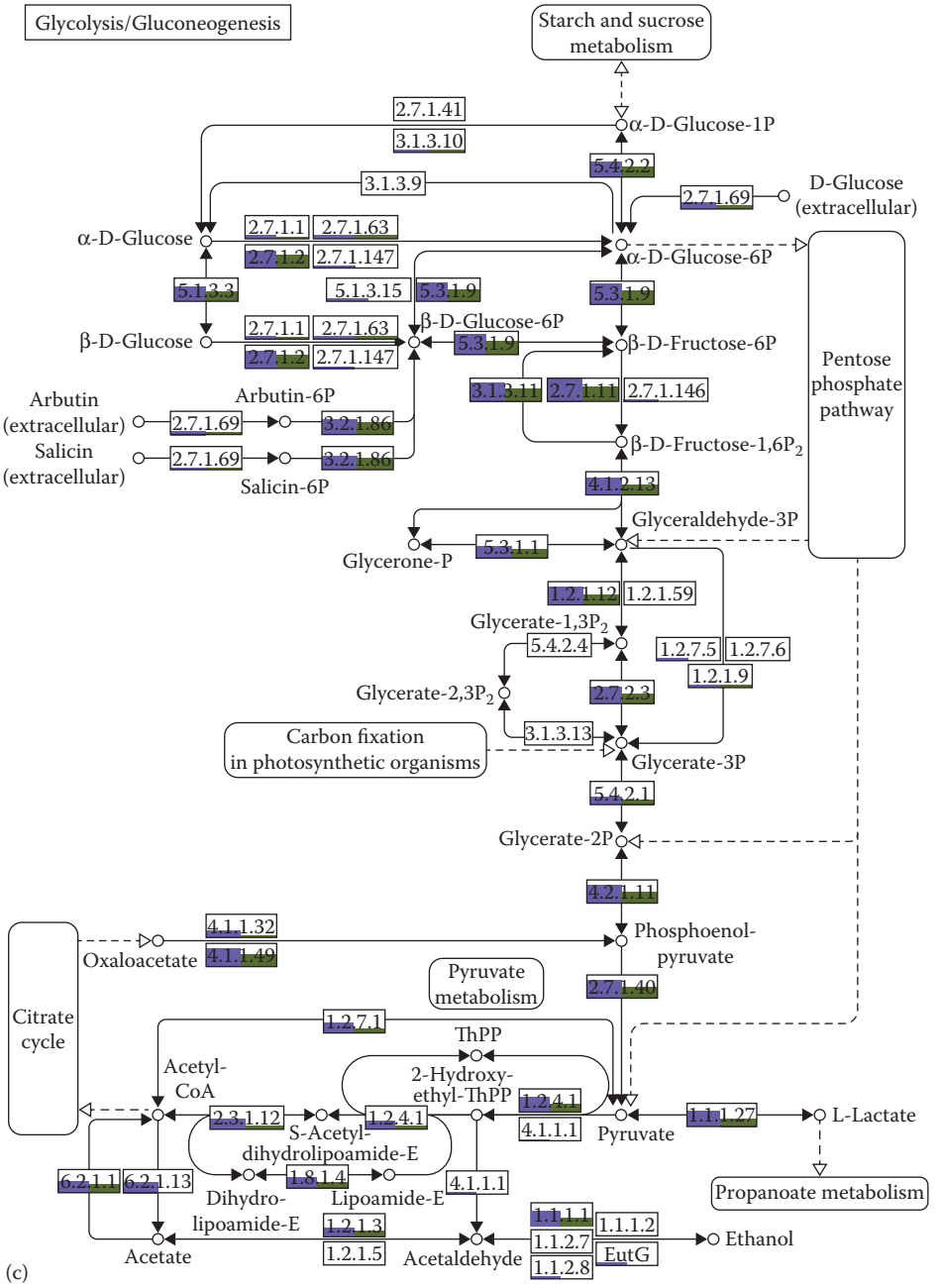

FIGURE 11.5 (Continued) Analysis of two microbial metagenomes from hospital sewage (unpublished data) with MEGAN (Huson, D.H. et al., *Genome. Res.*, 17, 377–386, 2007). (c) Detailed view of the functional genes related to the of the glycolysis/gluconeogenesis pathway, visualized using the KEGGviewer. Boxes indicate enzymes and are colored according to the number of reads in each metagenome.

11.4.2 Example of a Pipeline for Metagenomic Analysis: MG-RAST

Several databases offer a metagenomic analysis pipeline, and although, for example, the MG-RAST (Meyer et al. 2008), EBI Metagenomics (Hunter et al. 2014), and IMG/M (Markowitz et al. 2014) annotation pipelines report similar types of taxonomical and/or functional annotations, there are

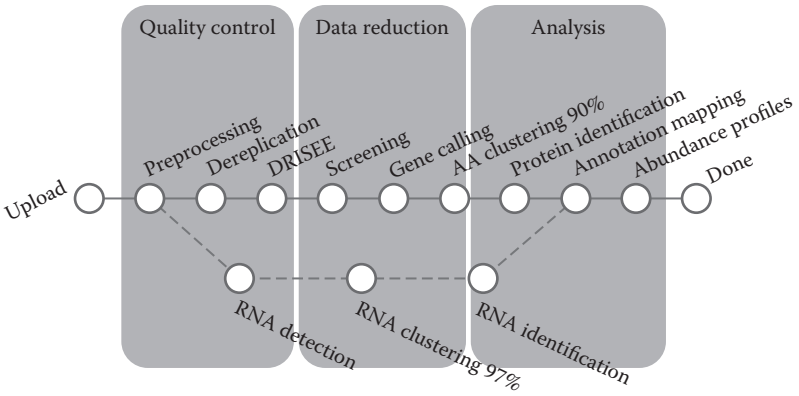

FIGURE 11.6 MG-RAST metagenomic analysis pipeline version 3. (Meyer, F et al., *BMC Bioinformatics*, 9, 386, 2008.)

subtle differences with respect to the algorithmic steps and bioinformatic programs used. Here, we will illustrate one example of a metagenomic analysis pipeline, MG-RAST (Meyer et al. 2008), one of the first online services to study the functional and taxonomic composition of metagenomes (http://metagenomics.anl.gov; see Figure 11.6). However, we want to emphasize that all are valid approaches to analyze metagenomic data.

After creating an account, users can upload their metagenomic sequences in formats, including FASTA, FASTQ, and SFF. Metadata annotations should be uploaded in a separate file. The user can choose several pipeline options such as minimum nucleotide quality and sequence length cutoffs, which MG-RAST will apply while performing quality control. Unreliable sequences and parts of sequences are removed, and not analyzed further. A dereplication step removes reads that are exactly identical; these reads are most likely artifacts of the sequencing process, and would bias the results. Moreover, sequences that are nearly identical to regions on the genomes of model organisms such as cow, fly, human, and mouse may not be derived from the microbial community, but consist of contamination, for example, in host-associated metagenomes, and these are removed as well. For the remaining sequences, genetic regions that represent protein-coding genes are identified and translated into amino acids. Amino acid sequences that are >90% identical are clustered, and the longest sequence is retained for further analysis. This clustering step can vastly reduce the size of the dataset, and thus the computational burden in the following steps. Next, the representative longest sequence of each cluster of proteins is included in a similarity search that identifies similar proteins in databases of known, annotated sequences. This similarity search is performed with the BLAT (Kent 2002), which is faster than BLAST (see later). As a result, users are supplied with a profile that indicates which taxonomic groups and protein functions are present in their metagenomic dataset, allowing them to answer the following classical questions in metagenomics: "Who is there?" and "What they are doing?" (Handelsman 2004). MG-RAST supports analysis of datasets containing bacterial, archaeal, and viral sequences, but currently not eukaryotic ones.

Databases such as MG-RAST, IMG/M, and EBI Metagenomics not only provide a metagenome analysis pipeline, but also function as repositories where researchers can store and download metagenomic datasets. There are currently several thousand metagenomes available, and this number is expected to grow rapidly in the coming years, providing an invaluable resource to researchers around the world. Uploaded datasets do not have to be made public immediately upon submission, and users may choose to keep their data private. However, users are urged to release the datasets as soon as the associated manuscript is reviewed and published.

11.4.3 Alignment-Free Metagenome Profiling

Annotation of metagenomic sequences by aligning them to a reference database is still rather computationally expensive, and in some cases, computer power may be a limiting factor in high-throughput metagenomic research. Recently, two very fast tools have been developed for taxonomic profiling of metagenomes, without alignment or assembly of the individual sequencing reads, FOCUS (Silva et al. 2014) and Taxy (Meinicke et al. 2011). FOCUS and Taxy count the *k*-mers (short "words" in the DNA, for example, length $k = 7–12$; see Figure 11.4) that are present in the whole metagenome and represent them as a vector that profiles the entire community at once. Then, it uses a database of vectors, each representing the *k*-mer profile in one of the reference genomes, to calculate a distribution of these vectors, whose combined *k*-mer profile best approximates the observed *k*-mer profile of the whole metagenome. Because FOCUS and Taxy do not need to align each individual sequence to a reference database, this results in a very fast and scalable approach to predict the taxonomic profile of a metagenome. These tools only provide a taxonomic annotation of the metagenome, that is, the most likely distribution by which the genomes in the reference database are present in the metagenome. Functional annotation is currently not provided (Silva et al. 2014, Meinicke et al. 2011).

11.4.4 Operational Taxonomic Units

As mentioned previously, the SSU rRNA gene sequences of many organisms were studied and have been collected in databases, including RDP (Cole et al. 2014), ARB-SILVA (Quast et al. 2013), and GreenGenes (DeSantis et al. 2006). Moreover, SSU rRNA environmental amplicon surveys have studied the presence of this gene in a wide variety of environments, allowing detailed descriptions of the species in these environments. To do this, the reads sequenced from environmentally amplified SSU rRNA are first grouped into operational taxonomic units (OTUs) or clusters of sequences that are highly identical (Blaxter et al. 2005). Literally, the term OTU means "a taxonomic group within the context of the present study," but in practice, the taxonomic group could be a strain, species, or genus, or a group with undetermined evolutionary relationships that shares a given set of characters. It is up to the scientist to specify and justify the definition of OTUs in the context of their particular study. In the analysis of SSU rRNA amplicons, OTUs are typically defined as clusters of sequences with 97% identity (or 3% difference), which is motivated by the observation that this corresponds approximately to the level of sequence identity observed between different species (Sokal and Sneath 1963, Konstantinidis and Tiedje 2005). Because OTUs are defined *de novo* by clustering the sequences based on their identity, OTU analysis does not depend on a reference database, and does not require the sequences to be known in advance. Thus, this data-first approach allows new species to be discovered, provided that their SSU rRNA gene sequences match the PCR primers used during amplification. Although the number of OTUs can be overestimated as a result of sequencing errors (Kunin et al. 2010), this approach has revolutionized our view of the diversity in the microbial world (Sogin et al. 2006).

There are some online tools for 16S rRNA sequence analysis such as Microbial Community Analysis (MiCA) (Shyu et al. 2007) and the RDP pipeline (Cole et al. 2009), which require the data to be uploaded through a user-friendly web interface. However, online analyses often have limitations for the number of sequences that can be analyzed, or the number of runs or permutations that can be performed. Quantitative Insights into Microbial Ecology (QIIME) (Caporaso et al. 2010) and Mothur (Schloss et al. 2009) are two frequently used, open-source bioinformatic tools for microbial ecology community studies that allow amplicon datasets to be analyzed. Both of them can be downloaded and run locally.

Analysis of microbial communities often includes rarefaction curves (Figure 11.7). A rarefaction curve is a way to assess the richness in species or gene families in a sample, that is, the alpha-diversity of microbes in the sampled environment. Rarefaction curves plot the number of new species or gene families that are observed as a function of the number of sequences sampled from the dataset. They are created by randomly subsampling *n* sequences from the total dataset of *N* sequences and measuring the

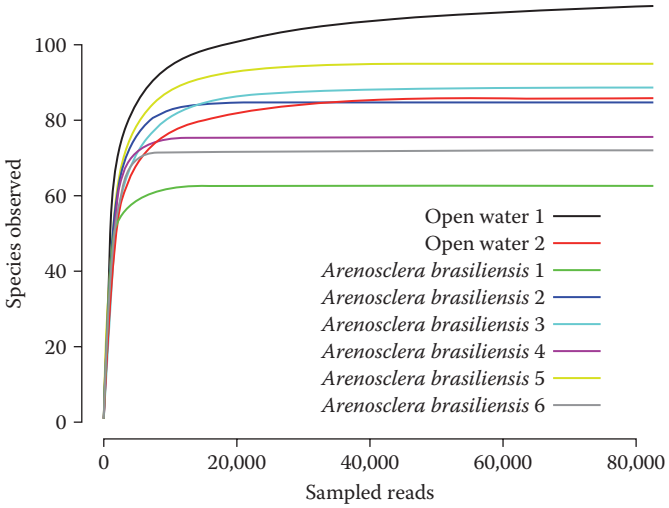

FIGURE 11.7 Rarefaction curves of eight metagenomes from the marine sponge *Arenosclera brasiliensis* and open water. The observed number of species is displayed as a function of increased sampling effort. (Trindade-Silva, A.E. et al., *PLoS One*, 7, e39905, 2012.)

number of species or gene families found in each subsample ($n = 1, 2, ..., N$ sequences). Each subsample of size n is taken several times, and the number of observed species or gene families averaged. Thus, rarefaction curves show how the number of species or gene families is expected to grow as the sampling effort increases (Gotelli and Colwell 2001). Curves generally grow rapidly at first, because initially each observed sequence is novel. Then, depending on the diversity of the microbes in the sample, the curves reach a plateau, which only slowly increases as the rarest species are found (Figure 11.7). The example in Figure 11.7 shows that the microbial diversity of the open water metagenomes keeps increasing slightly with an increasing number of reads, indicating the presence of a long tail of rare species. The six metagenomes associated with the marine sponge *Arenosclera brasiliensis* reach saturation at approximately 60–100 species, indicating that the metagenomic sequencing experiment covered most of the biodiversity in this environment at around 20,000–40,000 sampled reads (Trindade-Silva et al. 2012).

11.5 UNKNOWNS AND THE SIZE OF BIOLOGICAL SEQUENCE SPACE

In the previous examples, interpretation of metagenomic sequences relied on individual sequencing reads that were annotated by sequence similarity searches to a general or specialized database. This approach assumes that the environmental DNA sequences are derived from similar organisms as those present in the database and perform a similar genetic function. Although this is a valid paradigm, our knowledge of the global biological sequence space is very limited (Dutilh 2014). Sequence space is defined as the multidimensional space of all possible nucleotide (or protein) sequences (Smith 1970). Sequence space contains n dimensions; one dimension per residue that can take one of four (or 20, for proteins) states, with a total volume of $\Sigma 4^n$ sequences when summed over all possible sequence lengths n. It remains an open question how large the current biological sequence space is, that is, the fraction occupied by currently living organisms. Not only are the gene annotations in currently available (meta)genome sequences incomplete, but also recent advances in metagenomics have shown that most genes have never even been seen before. This unknown fraction of biological sequence space is sometimes referred to as "biological dark matter" (Youle et al. 2012, Rinke et al. 2013), and can reach up to 99% of the metagenomic reads, depending on the sampled environment, the type of DNA and sequencing technology, the homology search algorithm, and the reference database (Mokili et al. 2012). The unknowns limit our ability to annotate metagenomes, and consequently, they preclude a full understanding of the studied microbial ecosystem. Efforts

are under way to improve the coverage of biological sequence space by sequencing new reference genomes (Wu et al. 2009, Rinke et al. 2013). Similarly, laboratories around the world are validating the functions of new genes. Although these efforts continuously improve the reference databases, the technology to link the small body of experimentally generated evidence to large "families" of similar proteins is still in flux (Gilbert et al. 2011). For example, inconsistencies in the accuracy of genome annotation have been the subject of heated discussions since the beginning of the genomic era (Brenner 1999). Still, the so-called 70% hurdle holds for genome annotation, as functions of only ~50%–70% of the genes in any given genome can be predicted with reasonable confidence. The remaining genes either (1) are homologous to genes of unknown function, and are typically referred to as "conserved hypothetical" genes or (2) do not have any known homologs termed "hypothetical" or "noncharacterized" or "unknown" because it is unclear whether they encode actual proteins (Sivashankari and Shanmughavel 2006). Similarly, the majority of sequences in a metagenomic dataset cannot be annotated, because they do not have any homologs in the database.

11.6 METAGENOMIC ASSEMBLY

As mentioned previously, metagenomes are often characterized by a high fraction of unknown sequences. This not only depends on the species that are present in the sampled environment but also on the sampling protocol. For example, metagenomes derived from virus-like particles often contain many more unknowns than bacterial metagenomes (Mokili et al. 2012), reflecting the fact that viruses are an understudied group of organisms. Thus, it may be expected that expanding reference databases will be the most promising approach toward resolving the unknowns (Dutilh 2014). For example, the first shotgun metagenomes that were sequenced from the human gut revealed a high interindividual diversity and many unknown genes (Gill et al. 2006). Later, an improved reference database of sequences specific for the human gut microbiome decreased the percentage of unknown sequences from ~85% to ~20% (Qin et al. 2010). This database, known as the "human gut microbial gene catalog," was created by analyzing the metagenomic sequencing reads derived from the gut microbiomes of 124 individuals. Thus, the sequences in the new database better represented the microbial community in the human gut than the previous reference database, which contained previously sequenced genomes derived from other sources.

The gut catalog was created by assembling the sequencing reads from 124 shotgun metagenomes into longer sequences called "contigs," and identifying genes within these contigs (Qin et al. 2010). In a way, this is similar to the identification of OTUs for the analysis of environmental amplicon datasets (discussed earlier), because it is a data-first approach that does not require the sequences to be known in advance. Metagenome assembly can be done in two different ways. First, reads may be aligned to a closely related reference genome. This can identify close relatives of the reference within the sample (e.g., different strains of the same species), especially if the reads are iteratively aligned and assembled (Dutilh et al. 2009, 2010). However, reference-based assembly precludes the identification of completely novel genomes. This can be resolved by *de novo* assembly, a process where all reads are compared to each other, and overlapping sequences combined into longer contigs without a reference sequence. *De novo* assembly of a metagenome is more challenging than for a single genome, because the sequences are derived from a large number of genomes, and depending on the species richness and diversity, as well as the number of sequencing reads obtained, some or all these genomes may not be completely sampled. Moreover, genetic heterogeneity or microdiversity between strains leads to similar, but not identical sequences that need to be assembled into consensus sequences. Finally, one can imagine the situation in which the sequences from different species are assembled into one contig, leading to chimeras. Chimeras are sequences that are composed of parts from different genomes and do not exist in the original sample. Chimeras occur if there are highly identical regions between two genomes that are not recognizably different. Thus, it may be expected that chimeras are more frequent between closely related species and have less impact between very divergent species. Because chimeras frequently arise during PCR amplification, they are expected to be more abundant in environmental

amplicon sequencing than in shotgun metagenomics, and they can be detected using bioinformatic tools (Edgar et al. 2011). Finally, simultaneous cross-assembly (crAss) of reads from different metagenomes has been used to identify similarity between metagenomes (Dutilh et al. 2012), where the assembled cross-contig sequences directly represent the shared entities between the metagenomes. Because it is a data-first approach, crAss can identify sequences that are shared between specific samples *de novo*. Next, these potentially unknown sequences can then be analyzed using traditional approaches, such as sequence similarity searches, but also the laboratory-based molecular biology toolbox.

11.7 CONCLUSION

Our knowledge about the microbial world is still very limited. Metagenomics provides a powerful tool to explore biological sequence space, by unbiased sequencing of genetic material isolated directly from the environment. Metagenomic experiments provide large amounts of data, and analyzing these datasets requires efficient bioinformatic tools and fast computers. Thus, bioinformatics is a critical factor in studies of environmental microorganisms. Bioinformaticists can address biological questions by integrating information from very different sources and identifying patterns, which is not achievable through experimentation alone. By analyzing large metagenomic datasets, bioinformaticists are often the first to make interesting observations, and they allow biological hypotheses to be falsified or validated based on their analysis of the data. Biology, and especially the study of environmental microorganisms, is a highly data-driven science, and bioinformatics is the only possibility to unlock this data and discover novel science.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW et al. 1990. Basic local alignment search tool. *J Mol Biol* 215: 403–410.

Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ et al. 2014. Genbank. *Nucleic Acids Res* 42: D32–D37.

Berman HM, Henrick K, Nakamura H. 2003. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10 (12): 980.

Blaxter M, Mann J, Chapman T, Thomas F et al. 2005. Defining operational taxonomic units using DNA barcode data. *Philos Trans R Soc Lond B Biol Sci* 360: 1935–1943.

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F. 2002. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* 99: 14250–14255.

Brenner SE. 1999. Errors in genome annotation. *Trends Genet* 15 (4): 132–133.

Burks C, Fickett JW, Goad WB, Kanehisa M et al. 1985. The Genbank nucleic acid sequence database. *Comput Appl Biosci* 1: 225–233.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K et al. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 462: 1056–1060.

Carlson R. 2003. The pace and proliferation of biological technologies. *Biosecur Bioterror* 1: 203–214.

Cole JR, Wang Q, Cardenas E, Fish J et al. 2009. The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucl Acids Res* 37: D141–D145.

Cole JR, Wang Q, Fish JA, Chai B et al. 2014. Ribosomal Database Project: Data and tools for high throughput rRNA analysis. *Nucl Acids Res* 42: D633–D642.

Dehal PS, Price MN, Bates JT, Baumohl JK et al. 2009. MicrobesOnline: An integrated portal for comparative and functional genomics. *Nucl Acids Res* 38: 1–5.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M et al. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* 72: 5069–5072.

Dutilh BE. 2014. Metagenomic ventures into outer sequence space. *Bacteriophage* 4 (4): e979664.

Dutilh BE, Backus L, van Hijum SA, Tjalsma H. 2013. Screening metatranscriptomes for toxin genes as functional drivers of human colorectal cancer. *Best Pract Res Clin Gastroenterol* 27: 85–99.

Dutilh BE, He Y, Hekkelman ML, Huynen MA. 2008a. Signature, a web server for taxonomic characterization of sequence samples using signature genes. *Nucl Acids Res* 36: W470–W474.

Dutilh BE, Huynen MA, Gloerich J, Strous M. 2010. Iterative read mapping and assembly allows the use of a more distant reference in metagenome assembly. In Frans J. De Bruijn (ed.), *Handbook of Molecular Microbial Ecology I: Metagenomics and Complementary Approaches*. Hoboken, NJ: Wiley-Blackwell.

Dutilh BE, Huynen MA, Strous M. 2009. Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. *Bioinformatics* 25: 2878–2881.
Dutilh BE, Schmieder R, Nulton J, Felts B et al. 2012. Reference-independent comparative metagenomics using cross-assembly: crAss. *Bioinformatics* 28: 3225–3231.
Dutilh BE, Snel B, Ettema TJ, Huynen MA. 2008b. Signature genes as a phylogenomic tool. *Mol Biol Evol* 25: 1659–1667.
Edgar RC, Haas BJ, Clemente JC, Quince C et al. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27: 2194–2200.
Edwards RA, Olson R, Disz T, Pusch GD et al. 2012. Real time metagenomics: Using k-mers to annotate metagenomes. *Bioinformatics* 28: 3316–3317.
Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N et al. 2010. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISME J* 4: 642–647.
Field D, Amaral-Zettler L, Cochrane G, Cole JR et al. 2011. The genomic standards consortium. *PLoS Biol* 9: e1001088.
Field D, Garrity G, Gray T, Morrison N et al. 2008. The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* 26: 541–547.
Finn RD, Bateman A, Clements J, Coggill P et al. 2014. The Pfam protein families database. *Nucl Acids Res* 42: D222–D230.
Gerlach W, Stoye J. 2011. Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res* 39: e91.
Gilbert JA, O'Dor R, King N, Vogel T. 2011. The importance of metagenomic surveys to microbial ecology: or why Darwin would have been a metagenomic scientist. *Microb Inform Exp* 1: 5.
Gill SR, Pop M, Deboy RT, Eckburg PB et al. 2006. Metagenomic analysis of the human distal gut microbiome. *Science* 312: 1355–1359.
Gotelli NJ, Colwell RK. 2001. Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecol Lett* 4 (4): 379–391.
Handelsman J. 2004. Metagenomics: Application of genomics to uncultured micro-organisms. *Microbiol Mol Biol Rev* 68: 669–685.
Hesper B, Hogeweg P. 1970. Bioinformatica: een werkconcept. *Kameleon* 1 (6): 28–29.
Hogeweg P. 2011. The roots of bioinformatics in theoretical biology. *PLoS Comp Biol* 7 (3): e1002021.
Hunter S, Corbett M, Denise H, Fraser M et al. 2014. EBI metagenomics—A new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Res* 42: D600–D606.
Huson DH, Auch AF, Qi J, Schuster SC. 2007. MEGAN analysis of metagenomic data. *Genome Res* 17: 377–386.
Huson DH, Mitra S. 2012. Introduction to the analysis of environmental sequences: Metagenomics with MEGAN. *Evol Genomics: Stat Comput Meth* 856: 415–429.
Kanehisa M. 2002. The KEGG database. *Novartis Found Symp* 247: 91–101; discussion 101–103: 119–128, 244–252.
Kent WJ. 2002. BLAT—The BLAST-like alignment tool. *Genome Re*s 12 (4): 656–664.
Kerfeld CA, Scott KM. 2011. Using BLAST to teach "E-value-tionary" concepts. *PLoS Biol* 9: e1001014.
Kesh, S. 2004. Critical Issues in bioinformatics and computing. *Perspect Health Inf Manag* 1: 9.
Klappenbach JA, Saxman PR, Cole JR, Schmidt TM. 2001. rrndb: The Ribosomal RNA Operon Copy Number Database. *Nucl Acids Res* 29: 181–184.
Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L et al. 2013. Towards a unified paradigm for sequence-based identification of fungi. *Mol Ecol* 22: 5271–5277.
Konstantinidis KT, Tiedje JM. 2005. Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* 187: 6258–6264.
Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. 2010. Wrinkles in the rare biosphere: Pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ Microbiol* 12: 118–123.
Lane DJ, Pace B, Olsen GJ, Stahl DA et al. 1985. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc Natl Acad Sci USA* 82 (20): 6955–6959.
Langille MG, Zaneveld J, Caporaso JG, McDonald D et al. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 31: 814–821.
Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9: 357–359.
Lipman DJ, Pearson WR. 1985. Rapid and sensitive protein similarity searches. *Science* 227 (4693): 1435–1441.
Liu B, Pop M. 2009. ARDB—Antibiotic Resistance Genes Database. *Nucl Acids Res* 37: D443–D447.
Lombard V, Golaconda Ramulu H, Drula E et al. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucl Acids Res* 42: D490–D495.

Markowitz VM, Chen IM, Chu K, Szeto E et al. 2014. IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucl Acids Res* 42: D568–D573.
Martin KA, Siefert JL, Yerrapragada S, Lu Y et al. 2003. Cyanobacterial signature genes. *Photosynth Res* 75: 211–221.
McArthur AG, Waglechner N, Nizam F, Yan A et al. 2013. The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother* 57: 3348–3357.
Meinicke P, Aßhauer KP, Lingner T. 2011. Mixture models for analysis of the taxonomic composition of metagenomes. *Bioinformatics* 27: 1618–1624.
Meyer F, Paarmann D, D'Souza M, Glass EM et al. 2008. The metagenomic RAST server—A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9: 386.
Mitra S, Rupek P, Richter DC, Urich T et al. 2011. Functional analysis of metagenomes and metatranscriptomes using SEED and KEGG. *BMC Bioinfo* 12 (Suppl. 1): S21.
Mokili JL, Rohwer F, Dutilh BE. 2012. Metagenomics and future perspectives in virus discovery. *Curr Opin Virol* 2: 63–77.
Overbeek R, Begley T, Butler RM, Choudhuri JV et al. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucl Acids Res* 33 (17): 5691–5702.
Qin J, Li R, Raes J, Arumugam M et al. 2010. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464: 59–65.
Quast C, Pruesse E, Yilmaz P, Gerken J et al. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucl Acids Res* 41: D590–D596.
Rinke C, Schwientek P, Sczyrba A, Ivanova NN et al. 2013. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499: 431–437.
Roux S, Tournayre J, Mahul A, Debroas D et al. 2014. Metavir 2: New tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinform* 15: 76.
Schloss PD. 2010. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput Biol* 6: e1000844.
Schloss PD, Westcott SL, Ryabin T, Hall JR et al. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75: 7537–7541.
Schmidt TM, DeLong EF, Pace NR. 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J Bacteriol* 173: 4371–4378.
Segata N, Waldron L, Ballarini A, Narasimhan V et al. 2012. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 9: 811–814.
Shyu C, Soule T, Bent SJ, Foster JA et al. 2007. MiCA: A web-based tool for the analysis of microbial communities based on terminal-restriction fragment length polymorphisms of 16S and 18S rRNA genes. *Microb Ecol* 53 (4): 562–570.
Silva GGZ, Cuevas DA, Dutilh BE, Edwards RA. 2014. FOCUS: An alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ* 2: e425.
Sivashankari S, Shanmughavel P. 2006. Functional annotation of hypothetical proteins—A review. *Bioinformation* 1 (8): 335–338.
Smith JM. 1970. Natural selection and the concept of a protein space. *Nature* 225: 563–564.
Sogin ML, Morrison HG, Huber JA, Welch DM et al. 2006. Microbial diversity in the deep sea and the underexplored "rare Biosphere." *Proc Natl Acad Sci USA* 103 (32): 12115–12120.
Sokal PHA, Sneath RR. 1963. *Principles of Numerical Taxonomy*. San Francisco, CA: W. H. Freeman.
Trindade-Silva AE, Rua C, Silva GGZ, Dutilh BE et al. 2012. Taxonomic and functional microbial signatures of the endemic marine sponge *Arenosclera brasiliensis*. *PLoS One* 7: e39905.
Venter JCK, Remington JF, Heidelberg AL, Halpern D et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304 (5667):66–74.
von Mering C, Huynen M, Jaeggi D, Schmidt S et al. 2003. STRING: A database of predicted functional associations between proteins. *Nucl Acids Res* 31 (1): 258–261.
Woese CR, Fox GE. 1977. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci USA* 74: 5088–5090.
Wommack KE, Bhavsar J, Polson SW, Chen J et al. 2012. VIROME, a standard operating procedure for analysis of viral metagenome sequences. *Stand Genomic Sci* 6: 427–439.
Wu D, Hugenholtz P, Mavromatis K, Pukall R et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* 462: 1056–1060.
Youle M, Haynes M, Rohwer F. 2012. Scratching the surface of biology's dark matter. In Witzany G (ed.), *Viruses: Essential Agents of Life*. pp. 61–81. The Netherlands: Springer.