

Metagenomics: the boon for microbial world knowledge and current challenges

J.K. Choudhari¹, J. Choubey², M.K. Verma^{1,3}, T. Chatterjee² and B.P. Sahariah¹

¹Chhattisgarh Swami Vivekanand Technical University, Bhilai, India, ²Raipur Institute of Technology, Raipur, India, ³National Institute of Technology Raipur, Raipur, India

10.1 Introduction: an overview of metagenomics

The tiny alive microorganisms, that is, bacteria, fungi, viruses present in the biosphere are invisible with bare eyes performing enormous numbers of positive as well as negative tasks that help in proper functioning of ecosystems, balancing environmental equilibrium, and health issues in other species in the world. Microorganisms can reside in diverse habitats of soil, water, and fossil remaining as the alimentary tract system of organisms. Several microbes are able to survive in the extreme environmental condition of pH/temperature or so while implementing their functions. Microbes are predominantly involved in ecosystem succession, photosynthesis, nutrient biogeochemical cycle, food chain, symbiosis, research and development, pollution abatement, industrial activities, agricultural activities, metabolism in humans or other organisms, disease-causing/cure in other living beings, and many more. Microbes possess great genetic diversity, chemical property, and utility. Tremendous significant information regarding the identification, structure, function, habitat, niche, and other properties of microbes for various purposes is documented from laboratory studies and provided in public databases. The documentation of the vast microbial world in terms of diversity, taxonomy, functions, and structure is still under continuous study. Till date, a huge portion of world microbes are left unidentified and/or not explored in an ecosystem. Therefore it is beyond one's imagination about the incredibility, contribution, and significance of the yet to be identified, noncultured complex groups of microorganisms. Researchers are continuously involved for better understanding of the microbial world with sufficient accurate information extraction but always encounter various limitations and challenges.

Metagenome designates the whole genetic material of many individual organisms present in an environmental sample. The term metagenomics was first coined in the year 1998 (Handelsman et al., 1998). Metagenomic approaches to study various usage and key points of the metagenomes are received directly from the environmental samples that have been difficult to culture in the laboratory. Metagenomics investigations encompass various datasets, computational and biological technologies for analysis, and information extraction from datasets of environmental samples. Giovannoni, Britschgi, Moyer, and Field (1990) phylogenetically analyzed clone libraries enriched with 16S RNA of Sargasso Sea picoplankton derived by polymerase chain reaction (PCR) in the year 1990. In the year 1991 Schmidt and collaborators generated another library of sequencing information about DNA (deoxyribonucleic acid) from the marine picoplankton community using 16S rRNA gene cloning and sequencing.

Fig. 10.1 presents the chronology of the metagenomics study developments noted with major milestones. DNA of a species is the signature in a sample with a metagenome records for the identification of their significant configuration, structure, and function. Therefore microbes from the domains of bacteria, viruses, or fungi can be efficiently assessed by isolating their individual DNA. The DNA isolation precedes cloning of the same into genome library of cultured species and thereafter screening of the resultant clones to find out new elements. These steps are highly erudite with the requirement of proficient efforts, determination, precision, selection of correct tools, accurate inputs, methods, and analysis process to avoid misinterpretation of the input data. The 16S sequencing and Shotgun sequencing are the two basic metagenome approaches in the current scenario. Marker genes, such as 16S rRNA, 18S rRNA, and internal transcribed

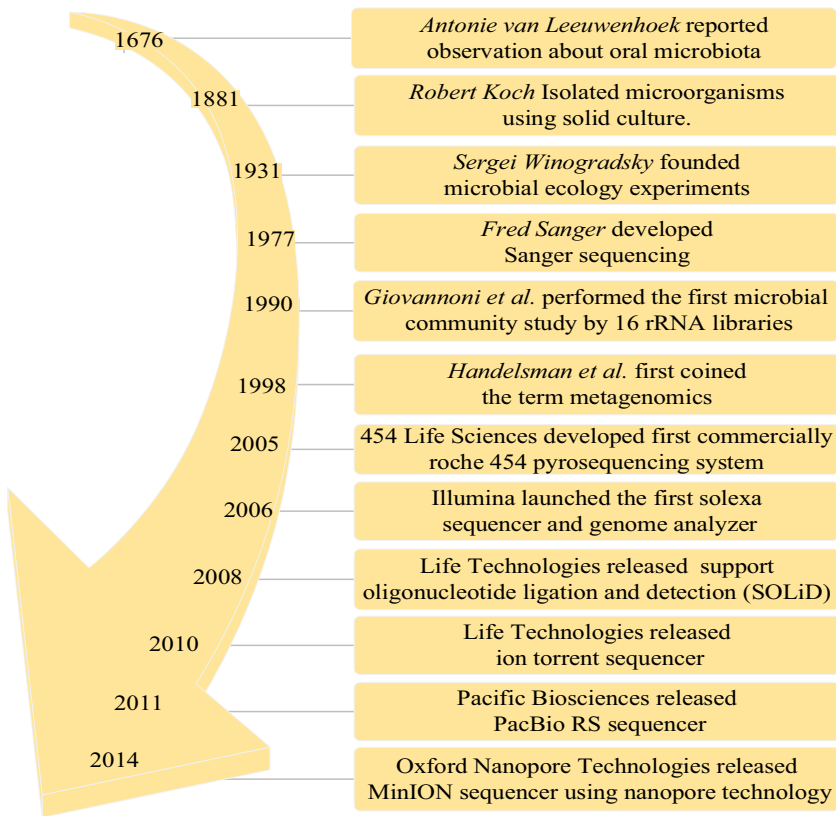


FIGURE 10.1 Chronology of the principal breakthroughs in Metagenomics.

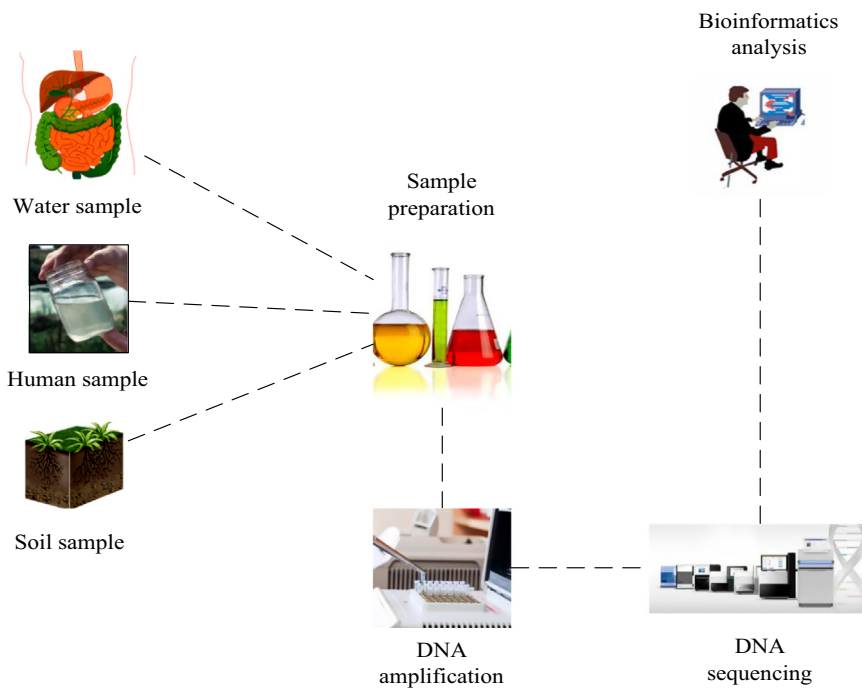


FIGURE 10.2 Metagenomics workflow for sequencing and analysis of the sample.

spacers, along with their relevant amplicon sequencing methods are considered in 16S-targeted sequencing. On the other hand, shotgun sequencing metagenomics encompasses information of both taxonomic and functional attributes. A schematic representation of workflow for shotgun sequencing metagenomics analysis using next-generation sequencing (NGS) techniques is given in Fig. 10.2.

The raw information about microbes is predominantly extracted employing NGS techniques for sequencing millions of small fragments of DNA in parallel followed by mapping the individual reads. Advantages of NGS to name a few are: its proficiency for sequencing entire genomes can be limited to a specific area of interest or a few individual genes within a short duration. This sequencing output highly influences genomic research with a great beneficial impact. Two broad NGS strategies namely, (1) amplicon sequencing, predominantly of the 16S rRNA gene as a phylogenetic marker based on PCR to generate amplicons and (2) shotgun sequencing based on direct sequencing of DNA samples without PCR to capture the complete breadth of DNA within a sample are followed.

In the amplicon sequencing (16S rRNA), PCR is used to identify species correctly and to analyze the significance of genetic diversity in a specific genomic region. The amplicon sequencing analysis results in complete quality census data of microbial communities. Shotgun metagenome sequencing being new and commanding environmental sequencing approach involves taxonomic profiling and functional analysis, thus providing advanced information of a community niche. This chapter describes advanced metagenomics approaches, tools used in metagenomics, and possible challenges and addresses the issues considering a dataset from mangrove soils as case study.

10.2 Resources in metagenomics

Databases containing various sequence datasets, numerous tools, and servers required for processing the datasets are major metagenomics resources. The employment of NGS techniques for metagenomics generates information in the form of sequence datasets. A few major databases considered for microbial studies of taxonomic binning of 16S reads are Greengenes, SILVA, Greengenes normalized, National Center for Biotechnology Information (NCBI), and Ribosomal Database Project (RDP). A database performs as a reference platform along with taxonomic information while determining the phylogenetic structure of the metagenome sequence reads of an unknown sample. GreenGenes is a comprehensive 16S reference database and taxonomy in database based on a de novo phylogeny that provides standard operational taxonomic unit sets. It can be browse at <http://greengenes.lbl.gov/>. Another rich database SILVA (<https://www.arb-silva.de/>) has manually curated taxonomic rank assignments and predominantly possesses taxonomic information about all three main domains namely, bacteria, archaea, and eukarya. The datasets here are predominantly constructed on phylogenesis of small subunit rRNAs namely, 16S and 18S for prokaryotes and eukaryotes, respectively. RDP (<http://rdp.cme.msu.edu/>) is also based on 16S rRNA sequences available from the International Nucleotide Sequence Database Collaboration and represents bacteria, archaea, and fungi (eukarya). The NCBI (<https://www.ncbi.nlm.nih.gov/>) is a physically curated database created from up-to-date efficient literature comprising names of all creatures related with submissions to the NCBI. These databases are further analyzed following relevant tools as mentioned in Tables 10.1–10.4 according to the workflow for final extraction of information from the database.

10.3 Challenges in metagenomics

Metagenomics being a culture-independent technology is capable to provide enormous informative resources about diverse microbes from their natural habitat encouraging researchers to solve the query to estimate the individual species, habitat, structural, and functional niche. A metagenomic project typically consists of basic steps, such as the sequencing of DNA or RNA using NGS (such as Illumina and Roche 454), followed by the utilization of specialized bioinformatics methods and tools in the form of computational and biological technologies. The processing of data includes assembling the sequences, identification of genetic materials, discarding unwanted data, physiological (functional metabolic comparative) analysis, extraction of important information, and interpretation. Minor errors in sample preparation or selection of computational technology may give rise to erroneous results and wrong interpretation. Also, each step has to be properly taken care of in presence of massive metagenomic data.

The major challenges encountered are in terms of storage of the huge data generated, selection of the correct and appropriate tool, and proficiency in checking of the data quality. The availability of tools in the form of coding instead of graphical user interface often raises trouble in data analysis. Variation of the results for the same data source is very likely with the utilization of different tools. Storage scarcity for the analyzed data is often encountered by various researchers. Species abundance and high diversity are prone to encounter these challenges with the generation of the gigantic size of NGS data. Researchers continuously come across the challenges while connecting the dots between sequencing reads as well as the origin of population and functions encoding. The arrangement of millions of reads from short to large reference genome is a staggering task in itself. In addition, other conditions, such as the presence of minor nonconformities [e.g., short nuclear polymorphism (SNP) and indels], sequencing errors with the millions of reads, and unavoidable repeats in the reference genome, should be considered to avoid erroneous output. These attributes often

make the process much more complex and always need to be correctly rectified. The assembly-based methods help in reducing the data size as it overlaps reads into contiguous sequences (contigs) and become computationally intensive but challenges arise as sequencing errors or genomic repeats.

Plentiful numbers of tools exist to perform metagenomics analysis that demands profound training for the selection and application of the appropriate tool. Other a few common challenges encountered by researchers that hamper successful metagenome analysis and interpretation of information are sequencing errors and genomic duplication during experiment conditions; comparisons of multiple datasets; and requirement of intensive computation and cumbersome downstream analysis including statistical analysis. With every day there is the development of new bioinformatics tools and technology; however, in face of the huge and diverse microbial world, it becomes limited in either way for the quest of never-ending curiosity and knowledge of researchers. To have a clear understanding of the challenges faced by researchers and ways to overcome the same, the workflow of metagenome data analysis is discussed in the following sections.

10.4 The workflow in metagenome analysis

For metagenome data analysis, numerous strategies are in existence possessing a few basic common steps, such as data processing, quality check, assembling of contigs, and pass the data for taxonomic classification (binning). These individual steps are configured with multiple substeps to achieve complete data analysis and result interpretation.

10.5 Dataset acquire and processing

Most of the sequencers provide sequence files in a standard format, which is commonly known as the FASTQ format. FASTQ format is very similar to the FASTA format commonly used by bioinformaticians where the "Q" stands for quality. It suggests that a FASTQ file contains not only the sequence but also the quality of the sequence bases along with the sequence. These are text files that contain the sequence information including the base qualities. As a sequencer can read the sequence of millions of fragments of DNA, the files are also expected to be huge, and in many cases, one would not be able to open the file in a regular text editor. But there are commands (`more/head/tail`) that enable the user to see a particular portion of such files.

There are two ways available for attaining the FASTQ and Short Read Archive (SRA) files. Firstly, if the researcher outsources the sequencing, they can get the designated files directly from sequence provider. Another way to obtain the FASTQ and SRA files from public repositories that host next-generation sequence data such as NCBI Short Read or SRA archive, European nucleotide archive and Japan's DNA data bank. NCBI Short Read or SRA archive is one of the popular public repositories hosting sequence data. The European nucleotide archive and Japan's DNA data bank are also to be named that host short raw data for reading new generation sequencers.

10.6 Quality control analysis

Quality control (QC) analysis is generally an initial step for metagenome data analysis and a set of computational tools are applied to identify rich/poor-quality sequences and thereafter discard the low-quality sequences and impurities. This helps in retaining only suitable sequences and reduces the size of the data significantly. QC encounters limitations in speed and impurity screening as metagenomic data basically possess sequences of multiple genomes, the majority of which is initially unknown. A few computational steps for QC analysis are mentioned below.

10.6.1 Base quality score

Base quality score (BQS) is the confidence value of a base based on Phred score that designates series of quality scores for entire considered bases at their respective location in a fastQ file. A BQS beyond Q30 is considered as good whereas any score above Q20 is generally accepted.

10.6.2 Sequence quality score

Sequence quality score indicates quality score distribution of entire considered sequences provided with provision to filter out as well as retain the low-quality values and good section, respectively, of a report. It allows researchers to examine the quality score distribution across all sequences.

10.6.3 Overrepresented sequences

The overrepresented sequences refer to those sequences, which are above an expected frequency in the data and generally of high biological significance, or an indication of adapter contamination used in sequencing.

10.6.4 Per base N content

The undetected base with sufficient confidence is labeled as “N” and a plot is generated representing the percentage of N base calls across its position.

10.6.5 Duplicate sequences

The number of reads/particular part of the genome that is encoded more than once is termed as a duplicate sequence. Enrichment bias, such as PCR over-amplification and repeated sequencing of a particular portion, is generally responsible for duplication causing a significant increase in the data size. The function “de-duplicate” is used to correct duplicate sequences. It is recommended that while comparing the raw and de-duplicated versions of the library, there is only a single time occurrence of the vast majority of reads within the dataset.

10.6.6 Read length distribution

Read length distribution (RLD) denotes the relative amounts of the diverse size of sequence fragments. Generally, sequencing produces uniform length reads, but activities, such as adapter removal and trimming, can influence the length distribution. This information is beneficial to discard the reads with low RLD than a particular threshold length.

10.6.7 Per base sequence content

Per base sequence content generates a parallel graph for each base ideally that are similar at each position. Any deviation indicates the presence of overrepresented sequences, PCR duplication, or other issues from systemic sequencing problems.

10.6.8 Guanine–cytosine content

A normal distribution of guanine–cytosine (GC) content is desirable and any deviation from the normal suggests library contamination or some bias in sequencing or library preparation.

10.6.9 Overrepresented k-mers

The k-mers present in the reads and their distribution along the read length is expressed in a distribution graph presenting the most frequently occurring k-mers and their sequence, the count, and the most observed position of each k-mer. Those k-mers, which do not have even coverage, could be possible sources of bias in the library.

10.6.10 Quality analysis and improving software tools

Various types of sequencing pieces comprising low-quality reads and contaminating reads can be present in the raw metagenomic data. Defective sequencing gadget property or flawed sample preparation are at most to be blamed for low sequencing quality reads that can significantly influence downstream analysis. Quality filtering highly benefits toward the detection of accurate information from the microbial sample through metagenomic sequencing. Impurities originated from impure samples and/or erroneous sample preparation include genomes from microbes along with other species (e.g., high eukaryotic species) in the metagenomic samples. The quality of raw reads obtained from sequencing experiments is initially evaluated before aligning them. Based on the evaluation, one should discard, trim, or modify the reads that are devoid of the defined specification before alignment. Without filtering and trimming low-quality reads, the results obtained might lead to wrong biological interpretation. A few tools for QC analysis are enlisted in [Table 10.1](#).

TABLE 10.1 Software tools used for quality control analysis and metagenomics data improvement.

Tool	Platform	Function	References
FastQC	Windows/Linux/Mac	Sequencing quality control	Andrews (2010)
FastQ Screen	Linux	Sequencing quality control, read mapping	Wingett and Andrews (2018)
BBtools	Linux/Mac/Windows	Sequencing quality control and trimming	Bushnell (2018)
MultiQC	Linux/Mac/Windows	Sequencing quality control, validation	Ewels, Magnusson, Lundin, and Källér (2016)
FASTX-Toolkit	Linux/OpenSolaris/FreeBSD/Mac	Read preprocessing, sequence editing; sequencing quality analysis	Gordon and Hannon (2010)
PRINSEQ	Web version	Read preprocessing; sequence trimming; removing sequence artifact	Cantu, Sadural, and Edwards (2019)
TagDust	Unix/Linux	Exclude known contaminants and filter out low complexity sequences based on operating a library of hidden Markov models HMM	Lassmann, Hayashizaki, and Daub (2009)
SolexaQA	UNIX	Solexaqa computes sequence quality statistic of Illumina system and also support on Torrent and 454 data	Cox, Peterson, and Biggs (2010)
TagCleaner	Web version	Identify and discard tagged sequences from genomic or metagenomic datasets	Schmieder, Lim, Rohwer, and Edwards (2010)
CANGS	Unix/Linux/Mac	Screens poor quality sequences, cleans PCR primers, filters singletons, and identifies barcodes	Pandey, Nolte, and Schlötterer (2010)
QC-Chain	Linux	Quality, trimming, and filtering	Zhou, Su, Wang, Xu, and Ning (2013)
PIQA	Linux	Check quality scores and identification of defective tiles from the data	Martínez-Alcántara et al. (2009)
Trimmomatic	Unix/Linux/Mac	Sequence trimming	Bolger, Lohse, and Usadel (2014)
Cutadapt	Mac/Windows	Sequence trimming	Martin (2011)

10.7 Genome assembly tools in metagenomics

Following quality control check and improvement derived short sequences of DNA or reads are added as an input to the computational process to decipher and compare with genome library. This is a complex task due to uncertainties in the reconstruction and duplication caused by genomic repeats, nonuniform representation of genomes in a sample, and genomic variant between the sequences of closely related organisms. A few tools used for genome assembly are mentioned with their applications in [Table 10.2](#).

10.8 Binning tools in metagenomics

Metagenomics binning implies clustering of the assembled sequences into a similar hierarchy, for example, species, genus, and even higher categories employing binning tools. Reference-based binning and reference-free binning are two principal categories of metagenomics binning approaches. Reference-based binning methods align sequences to databases of reference genomes and identify the taxonomic group identical to the input sequence. Reference-free binning methods make use of sequence information, without any prior knowledge and group sequences into unlabeled bins (Sedlar, Kupkova, & Provaznik, 2017; Song, Ren, & Sun, 2019; Yue et al., 2020). This binning provides light in terms of the diversity of species present in the metagenome. Researchers can identify the extreme conditions considering the sampling conditions about the survivability of the species. A few tools used for binning are mentioned in [Table 10.3](#).

TABLE 10.2 Tools for genome assembly.

Tool	Platform	Function	References
Megahit	Linux	Assembles metagenomics following succinct de Bruijn graph principle	Li, Liu, Luo, Sadakane, and Lam (2015)
SPAdes	Linux/Mac	Assembles genomes	Bankevich et al. (2012)
MetaSPAdes	Linux/Mac	Versatile metagenomic assembler	Nurk, Meleshko, Korobeynikov, and Pevzner (2017)
Ray Meta	Linux	De novo assembly of metagenomes using distributed computing to enable parallel assemblies of multiple genomes	Boisvert, Raymond, Godzaridis, Laviolette, and Corbeil (2012)
MetaVelvet-SL	Linux	Single-genome assembler	Sato and Sakakibara (2015)
IDBA-UD	Linux	A de novo assembler recognized for single-cell and metagenomic sequencing data characterized with highly uneven depth	Peng, Leung, Yiu, and Chin (2012)
MetAMOS	Linux	Assembles modular metagenomic provided with analysis steps for metagenomic data	Treangen et al. (2013)
MOCAT2	Linux	Assembles metagenomic sequence followed by gene prediction with reference to novel features for taxonomic	Kultima et al. (2016)
MetaQuast	Linux/Mac	Evaluates and compares metagenome assemblies with reference to nearby alignments	Mikheenko, Saveliev, and Gurevich (2016)
BUSCO	Linux	Evaluate the comprehensiveness of genome assembly	Sepepey, Manni, and Zdobnov (2019)
CheckM	Linux	Compute quality of genomes retrieved from isolates, single cell, or metagenomes	Parks, Imelfort, Skennerton, Hugenholtz, and Tyson (2015)

TABLE 10.3 Binning tools.

Tool	Platform	Function	References
AbundanceBin	Linux/Mac	Abundance-based binning for estimations of species abundances and characterizing a microbial community	Wu and Ye (2011)
MaxBin	Linux	Binning based on an expectation–maximization algorithm to find out the taxonomy	Wu, Simmons, and Singer (2016)
MetaWatt	Linux/Mac	Primarily depends on multivariate statistics of tetranucleotide frequencies associated with the application of incorporated Markov models and taxonomic assessment of binning quality	Strous, Kraft, Bisdorf, and Tegetmeyer (2012)
SolidBin	Linux	Binning tool based on using sequence feature similarity for taxonomy assignments of some contigs	Wang, Wang, Lu, Sun, and Zhu (2019)
MetaBCC-LR	Linux/Mac	Encompasses binning with the inclusion of long reads composition	Wickramarachchi, Mallawaarachchi, Rajan, and Lin (2020)
BMC3C	Linux	Codon practice, sequence configuration, and read coverage are encompassed while binning metagenomic contigs	Yu, Jiang, Wang, Zhang, and Luo (2018)
COCACOLA	Linux/Unix	Sequence composition, read coverage, coalignment, and paired-end read linkage are incorporated during binning metagenomic contigs	Lu, Chen, Fuhrman, and Sun (2017)
GraphBin	Linux	Categorize and improve binning of metagenomic contigs employing assembly graphs	Mallawaarachchi, Wickramarachchi, and Lin (2020)
GroopM	Linux	A dedicated tool specialized with visual interactive metagenomic bin editing properties is available	Imelfort et al. (2014)
MetaBAT	Linux	Abundance and tetranucleotide frequency incorporate and categorize metagenome binning	Kang, Froula, Egan, and Wang (2015)

10.8.1 Statistical analysis

For complete identification of the richness of the taxonomic group present in the genome, many tools are used for statistical analyses. The Primer-E package allows a series of diverse statistical tools together with the generation of multi-dimensional scaling plots, analysis of similarities, and recognition of the species or functions with special attributes to identify variance between two samples (SIMPER). Metastats, a web-based tool offers high confidence discriminatory functions between the replicated metagenome datasets. Kristiansson, Hugenholtz, and Dalevi (2009) reported that ShotgunFunctionalizeR package is efficient for commuting a number of statistical procedures to assess functional variances for comparison within the samples, both for individual genes and for entire pathways following the popular R statistical package.

10.9 Data storage and sharing

Data storage and sharing are unavoidable to genome researchers. However, for the huge metagenomic information and results with time, it requires an advanced society and association for storage, communication of metadata, and federal services. Genome Sciences Centre is a platform for large-scale high-throughput genomics data, helps researchers in the processing of complex biological samples and provide various sequencing and bioinformatics services. The minimum information about a genome sequence (MIxS) currently provides a series of standard metadata languages with minimum information about any (x) sequence checklists of the considered genome. The minimum information about a metagenome sequence and the minimum information about a MARKer sequence possess standard arrangements for processing and storing environmental and experimental data. Few tools are enlisted in Table 10.4 for the reader's reference.

10.10 Metagenomics analysis: a case study

Metagenomic sequencing assesses microbial diversity and detects microorganisms provided new diagnostic tools for microbiologists. The metagenomics methodology consists of three principal activities namely, (1) wet laboratory methodology; (2) sequencing; and (3) data analysis. The integration of these three parts is of crucial importance for the fruitful findings and sample interpretation. The technologies are specially selected to distinguish microbial identity without prior knowledge of what a sample contains to open new doors in disciplines, such as microbial ecology. Gratified to bioinformatic technologies, now it is not only identifying existing microbial species but also is thrust into the functional roles, such as metabolic activities of the microorganisms. This section provides insight into the metagenomics research style for the identification of amazing microbes considering a publicly available dataset.

10.11 Material, methodology, and outcome

10.11.1 Metagenome dataset

The work "Microbial diversity in mangrove soils" is considered, and the SRR171305 dataset is derived from the NCBI SRA (<https://www.ncbi.nlm.nih.gov/sra>) as an example to describe the ways and tools' utilization to neutralize the challenges raised during the analysis of a dataset. If a dataset is not available in FASTQ format, the same can be transformed into FASTQ format using the fastq-dump program of the SRA toolkit.

10.11.2 Sequencing quality analysis

The quality of SRA data is assessed using the FastQC tool as it is an important aspect of analyzing metagenomes. A series of analysis modules, such as BQS, SCS, overrepresentative sequence, per base "N", duplicate sequence, RND, per base sequence content, GC content, and Kmer content, is applied to the raw data and HTML report is achieved with a summary of the module for rapid assessment of the outcome. Based on the quality control report generated, it is possible to decide on the pretreatment steps, which must be carried out before the actual alignment step. The preprocessing procedure includes clipping the adaptor and removal of low-quality bases. This can be determined by using the FastQC summary report and then matching the format details displayed in the reads having the overall low quality to filter out and trimming of reads to remove low-quality portions of the sequence while retaining the high-quality portion. The data set SRR171305 contains 254,583 sequences in total 100,157,019 base pairs (bp) with an average length of 393 bp. Among the sequences tested, 26,500 sequences (10.41%) are identified as unable to pass the QC pipeline and duplication identified 25,840 sequences as artificial duplicate reads.

TABLE 10.4 Web server for metagenome analysis, data storage, and sharing.

Tool	Function	References
MG-RAST	Phylogenetic and functional analyses of metagenomes and data storage	Keegan, Glass, and Meyer (2016)
IMG/M	Comparative analysis of publicly available genome, data storage, management, and analysis system	Chen et al. (2016)
METAREP	Comparative metagenomics analysis as well as taxonomic and functional classifications, gene ontology, NCBI taxonomy, and pathway analysis	Goll et al. (2010)
CoMet	Comparative functional profiling of metagenomes as well as ORF to recognize as well as successively assign protein sequences	Herath, Tang, Tandon, Ackland, and Halgamuge (2017)
METAGENassist	Comparative metagenomic analysis using bacterial census data from the different biological host and different environment site for taxonomic or automatically generated phenotypic labeling	Arndt et al. (2012)
MetaABC	Intermingling binning tools with methods, such as artifacts discard, unassigned reads analysis, and sampling biases control	Su et al. (2011)
MyTaxa	Categorize all genes present and identify the degree of novelty in an unknown sequence to taxonomical strata and the occurrence of horizontal gene transmission	Luo, Rodriguez-r, and Konstantinidis (2014)
metaMicrobes Online	Proportional and functional genome analysis	Chivian, Dehal, Keller, and Arkin (2012)
EBI Metagenomics	Taxonomical, functional, and proportional metagenomics investigations	Hunter et al. (2014)
CAMERA	Offers a unique arrangement for placing, localizing, investigating, picturing, and allotment of information on microbial biology	Chen, Sun, Li, and Wooley (2011)
METAVIR	To annotate viral metagenomic sequences	Roux, Tournayre, Mahul, Debrosas, and Enault (2014)
VIROME	Designed to enable users to have the flexibility for retrieving reads, open reading frames, projected peptides sequences, and search outputs for considered secondary analyses	Wommack et al. (2012)
QIIME	Microbial taxonomy and function analysis	Kuczynski et al. (2012)
Mothur	Taxonomy and function analysis	Schloss et al. (2009)
PICRUSt	Functional analysis designed for indicator gene research and full genomes	Langille et al. (2013)

Fig. 10.3A and B represents the distribution of sequence lengths of base pairs for this metagenome before and after processing. Each position denotes sequences number within a length bp range based on upload and post-QC sequencing. The distribution of sequence lengths and GC content (%) increased after processing as the sequence artifacts, and low-quality sequence lengths get discarded considering Phred score and sequences length 15 and 5, respectively, keeping only a high-quality base with Phred score more than 15 where sequence length is higher than 5.

Following quality trimming, assembling the contig with reference to the genome is performed using MG-RAST server considering parameters namely, assembled (if input sequence contains assembled data), dereplication (removing the replicate sequence) screening, (removing host-specific species), and dynamic trimming (discarding poor sequencing). These parameters are essential for efficient biological interpretation in terms of taxonomic and functional properties. An accurate and effective k-mer-based algorithm is generally applied for the alignment of the contig, which offers due consideration to all the issues of the arrangement of millions of reads from the short large reference genome, presence of minor nonconformities, such as SNPs/indels, sequencing errors, and unavoidable repeats in the reference genome. The accuracy of alignment impacts efficient identification of microbial community diversity, whereas erroneous alignment of reads leads to errors and wrong interpretation. Moreover, the algorithm's speed and memory footprint are also important parameters for fast analysis.

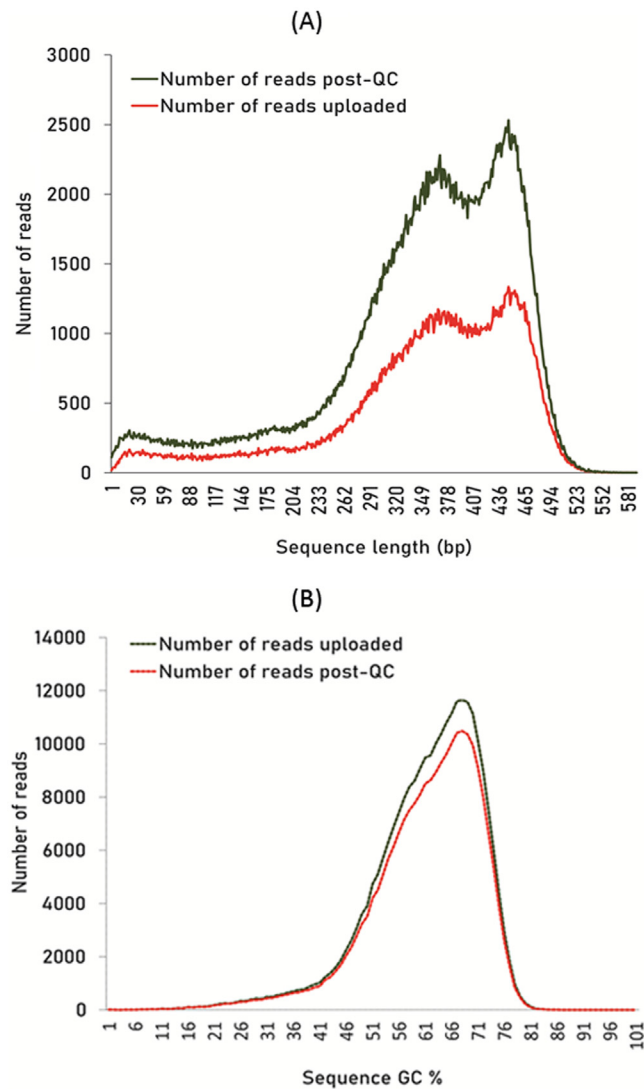


FIGURE 10.3 Sequence distribution of metagenome dataset: (A) sequence length histogram and (B) sequence guanine–cytosine distribution.

10.11.3 Hits distribution of metagenome from the database sources

Hits distribution of metagenome analysis provides a chart that shows the number of annotated reads hits in the various databases listed namely, protein databases, protein databases containing functional hierarchy information, and ribosomal RNA databases. Fig. 10.4 indicates that the considered mangrove soil metagenome dataset possess highly hit matching with RefSeq, TrEMBL, GeneBank, IMG, and PATRIC databases.

10.11.4 Hits distribution of functional group

The hits distribution of functional group analysis for the processed sequences from the considered database is given in Fig. 10.5A–C to illustrate the distribution of functional categories, such as Kyoto Encyclopedia of Genes and Genomes orthologous groups (KOs), subsystems, clusters of orthologous groups (COGs), and the evolutionary genealogy of genes: nonsupervised orthologous groups (eggNOGs). Each share of the pie charts denotes the fraction (%) of reads with identified protein functions with reference to considered marked source category. For the considered database, the top five hits distribution of functional categories (metabolism, genetic information processing, environmental information processing, cellular processes, and human diseases) followed by 24 subcategories are identified in KO where maximum sequence matching is recorded for genetic information processing (16.87%) (Fig. 10.5A). This included

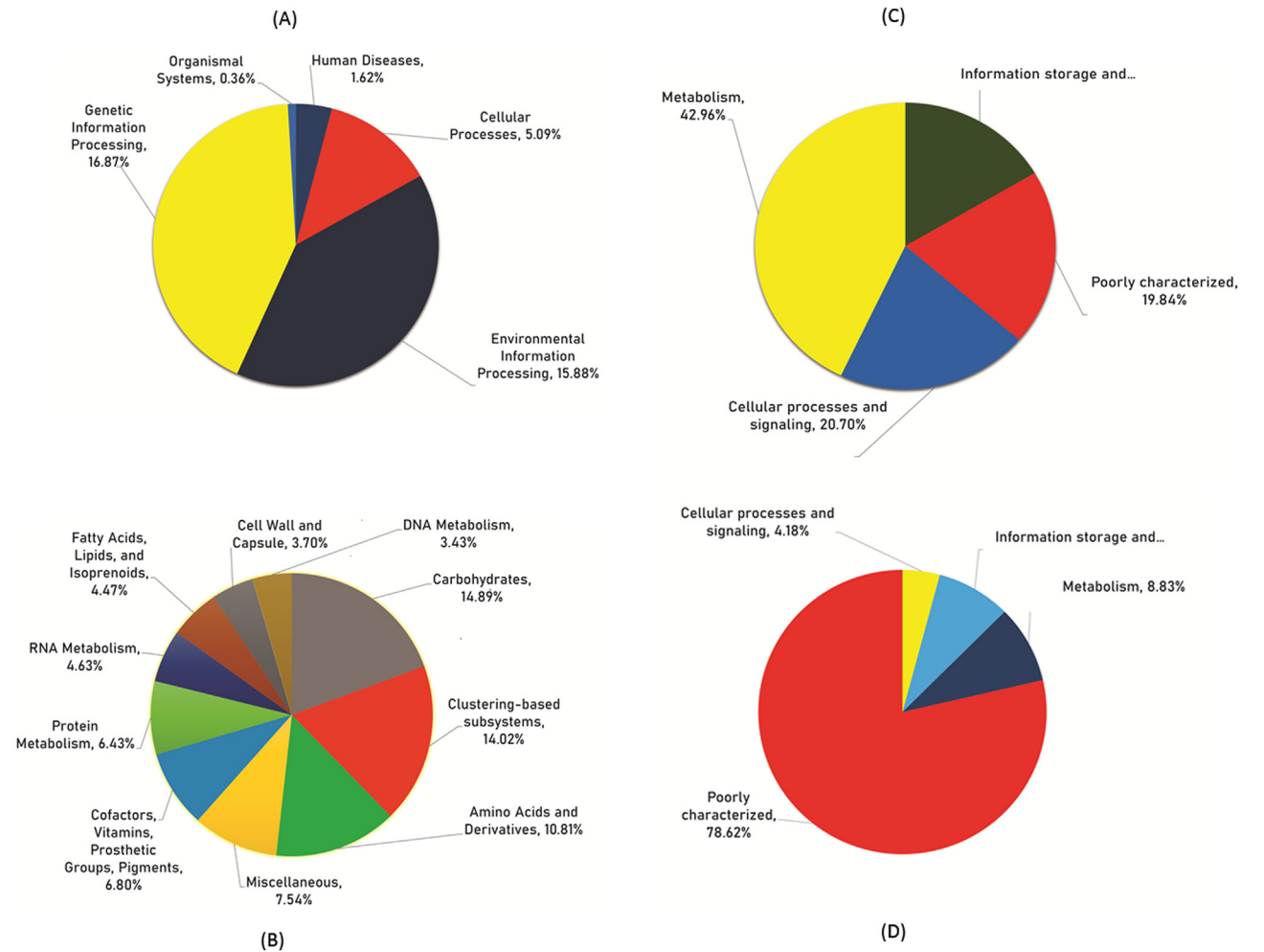
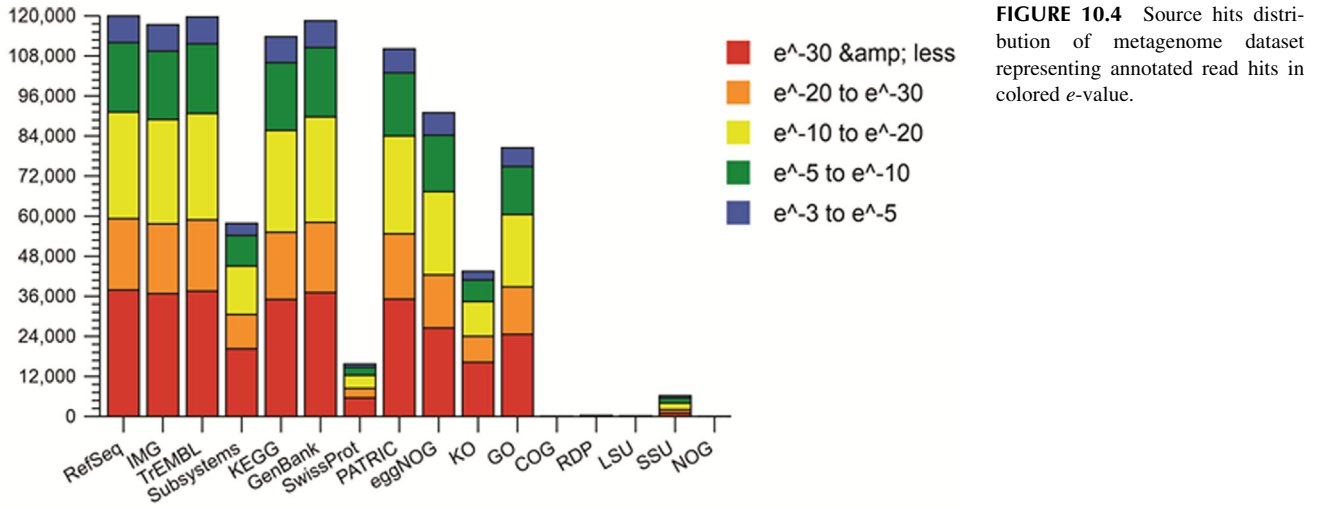


FIGURE 10.5 Functional hits distribution: (A) Kyoto Encyclopedia of Genes and Genomes orthologous groups, (B) subsystems, (C) clusters of orthologous groups, and (D) evolutionary genealogy of genes: nonsupervised orthologous groups.

transcription, translation, folding sorting and degradation, replication, and repair. Environmental information processing matched 15.88%, where principal functions are membrane transport, signal transduction, and signaling molecules and interaction. The subsystems functional database designates the assembly of associated functional roles within a system, for example, a metabolic pathway. For the current study, nine significant subsystem functional roles are identified, where carbohydrates metabolism and clustering based subsystem exhibited higher hits (Fig. 10.5B).

The COGs' functional database provides information about orthologous functions of a genome considering their protein orthologs. COG analysis with the mangrove dataset specified the top five highly matched functional categories in descending order as metabolism, cellular processes, information storage, and processing, and poorly characterized (Fig. 10.5C). The eggNOG database provides biological information similar to COGs but also incorporates nonsupervised orthologous groups originated from various organisms. While analyzing eggNOG, the mangrove dataset expressed maximum functional hit for the category of poorly characterized followed by metabolism, information storage, processing, and cellular processes (Fig. 10.5D). The analysis of all these functional attribute databases can be performed simultaneously in the MG-RAST server within 24–48 hours.

10.11.5 Taxonomic hits distribution

Taxonomic hits distribution is investigated using a contigLCA algorithm for the dataset to find the unique consensual taxonomic entity for all characteristics of each individual sequence in terms of taxonomical level (domain, phylum, class, family, and genus). The results are generated from direct input and associated with detailed statistics. In the taxonomic classification of a biological system, the domain ranks the highest position above the kingdom level. Taxonomic analysis or binning suggests that the dataset derived from the mangrove sample is dominated mainly by the bacteria domain (92%) and eukaryota (7%) as can be seen in Fig. 10.6A.

Fig. 10.6B presents the taxonomic information at the phylum level as Actinobacteria (35.20%) and Proteobacteria (31.86%) occupy maximum coverages for the mangrove soil. Taxonomic (phylum) distribution of top sequence hits. Actinobacteria is a large phylum present in abundance at both aquatic and terrestrial ecosystems and characterized by high GC content. Proteobacteria is recognized as a possible microbial signature of disease related to metabolism. Recent studies correlated the influence of proteobacteria with a few respiratory disease conditions influencing lungs (asthma, chronic obstructive pulmonary disease, inflammation, etc.). Taxonomic class distribution of the sample suggested Actinobacteria, a class of bacteria in the phylum Actinobacteria occupies a huge part (67.92%) followed by Alphaproteobacteria (35.30%) and other few classes in the mangrove soil (Fig. 10.6C).

The taxonomic distribution identified 10 rich orders in the considered dataset dominated by order *Actinomycetales* (37.24%) from class *Actinobacteria*. This is followed by order *Planctomycetales* (13.8%, from class *Planctomycetacia*) and *Solibacterales* (6.73%, from class *Acidobacteria*) and others as given in Fig. 10.6D. Fig. 10.6E presents the taxonomic distribution of the considered mangrove soil dataset at the rank of family. The rich family belongs to *Solibacteraceae* (9.99%), *Bradyrhizobiales* (8.69%), *Conexibacteriales* (7.22%), and *Streptomycetales* (7.04%). Fig. 10.6F denotes taxonomic distribution of considered mangrove soil dataset at the rank of Genus showing *Candidatus solibacter* (10.09%), *Conexibacter* (7.29%), and *Streptomyces* (6.47%) to be rich over other available genera. Thus the analysis helps in identifying the dominant taxonomic groups along with the functional attributes in the dataset for further information extraction as well as interpretation.

10.11.6 Rarefaction curve

The rarefaction curve (Fig. 10.7) computed from species abundance, denotes species richness showing a total number of distinct species from a complete data set. Quick rise with a high individual number at first followed by level off toward an asymptote as fewer new species per unit of individuals collected is a major characteristic of this curve. The steep slope on the left is an indicator of a large fraction of the species diversity that remains to be discovered. The flatter curve to the right suggests that a reasonable number of individuals are sampled, where even more intensive sampling yields only a few additional species.

10.11.7 Alpha diversity

Alpha diversity (α -diversity) expresses the diversity of organisms in a sample in a single number, which is estimated from the distribution of species level annotation of the sample using MG-RAST. For the current study, the range of α -diversity values is 413 species. Annotated species richness is the number of distinct species annotations in the

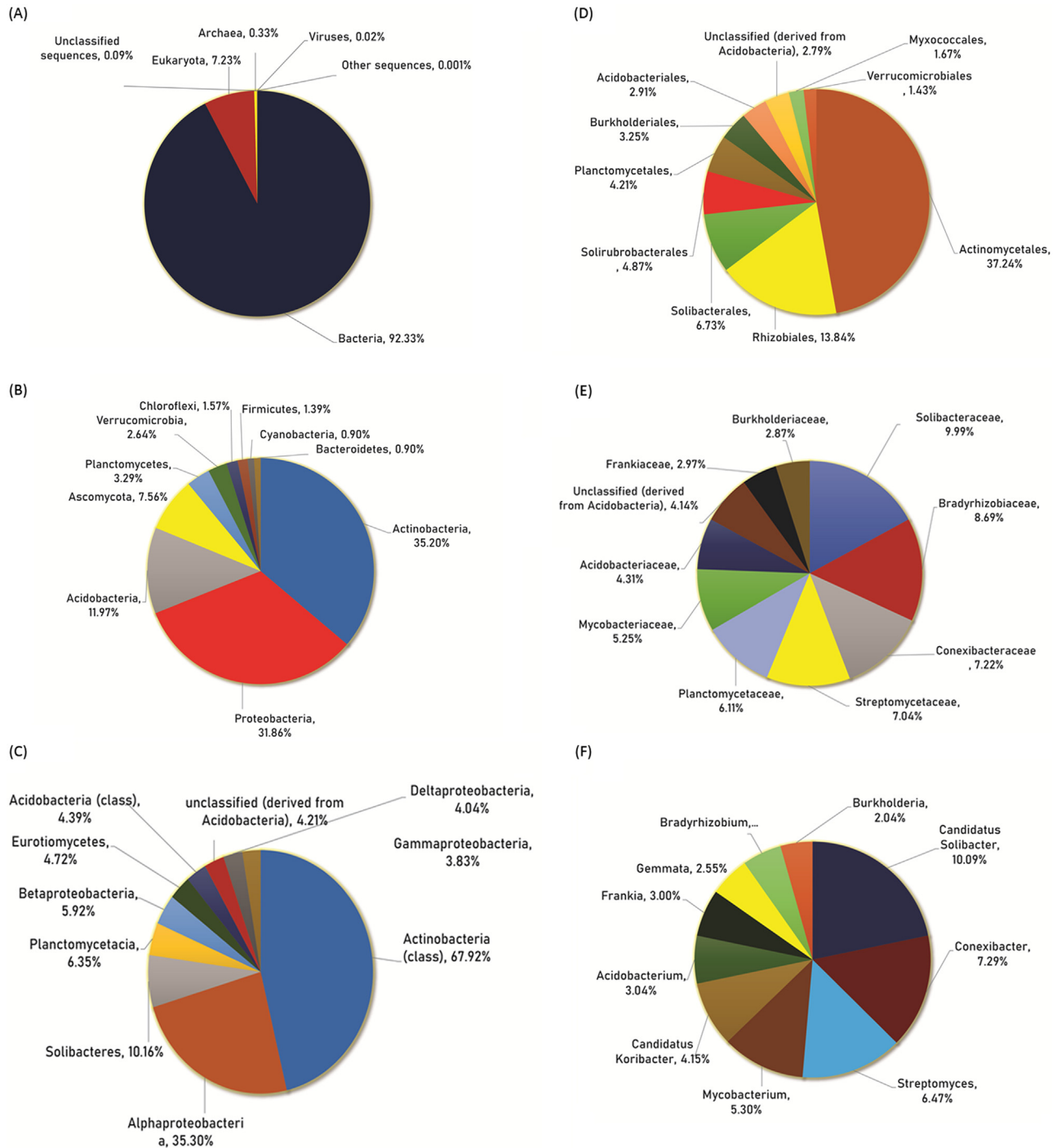


FIGURE 10.6 Taxonomic hits distribution: (A) domain, (B) phylum, (C) class, (D) order, (E) family, and (F) genus.

combined MG-RAST data set. Contrary, Shannon diversity is an abundance-weighted average of the logarithm of the relative abundances of annotated species.

The use of the MG-RAST server for metagenomic analysis of the mangrove soil dataset requires 24–30 hours and provides detailed results with statistical information in a single platform. The principal functional attributes recognized from functional analysis of the dataset are genetic information processing and metabolism. The taxonomic profile analysis suggests that the sample is rich in the domain bacteria, where maximum abundant phylum, class, order, family, and genus are *Actinobacteria*, *Actinobacteria*, *Actinomycetales*, *Solibacteraceae*, and *C. solibacter*, respectively.

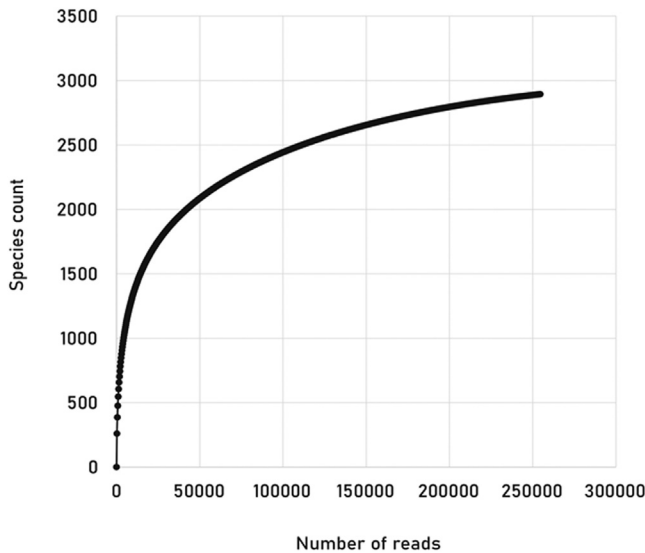


FIGURE 10.7 Rarefaction curve for the computation of species richness.

10.12 Advantages of metagenomics study

Microbial populations in nature are continuously exposed to changes in habitat and environmental conditions. There is the possibility of several physiological/functional and biochemical properties' transformations through acclimation to adapt to the new conditions. Laboratory study of microbes is feasible to deal with a selective and small number of populations compared to the existing population in the biosphere. Metagenomic studies deal with these natural microbes considered for sequencing directly from their existing environment/natural sample without any corruption to provide detailed information of their existence and properties. Identification of new species from various sites is prone with the library enlisted species compared with similar sequence categorize if identical species is not listed beforehand. Metagenomics provides the advantage to study acquiring knowledge about microbes that is difficult to culture in the laboratory directly from their true habitat.

The information received is from the natural terrain of the microbes. These enlighten the complex interactions in the ecosystems, microbial diversity, and genetic information and identify new elements in the microbes in terms of gene/protein/enzymes, etc., together with various working biochemical pathways. The identification of new elements gets relevant utilizations in pollution treatment, production of antibiotics, or other novel products that are hugely used in environmental/medical/industrial research. The successful metagenomic analysis is indeed a boon to know the insight of the microbes and extract more significant information for the welfare of humans and the environment. Therefore the metagenomics approach stands highly to encompass and discover knowledge about the advantages as well as proficiencies of microbes for its various applications.

10.13 Limitations and future perspective

Metagenomics study arrangements with sample and data analysis from specie's natural niche and therefore provide huge information identifying about species diversity, their function, and interaction. However, it is prone to loss of information about species whose extracts are lost during sample preparation or at other states. Reference sequence in the laboratory is necessary for the complete notation of the species. Effective screening methods and threshold conditions for the analytical activities are still under investigation along with other bioinformatics tools. Also, the genomes that are recorded with incompetent expression may be discarded which may lead to loss of sample identification. Along with it, the researchers are required to experience satisfactory knowledge about metagenomes as well as computational tools/language for correct interpretation of the coded results preceded by the correct selection of methodology.

10.14 Conclusion

Metagenomics is a broad discipline facilitating the taxonomic and physiological information of species collected from their true habitat. Metagenomics successfully encompasses diverse environment samples and its species metagenome

information with proficiency due to advanced techniques and resources. Microbes are providing their service to earth in terms of structural and functional balance in the ecosystem, medical and health status, and industrial and economic activity with significant attributions. Its utmost importance that microbes are subjected to detailed study for identification, physiological and functional aspects at a taxonomic level for future reference. Metagenomics is worthy for providing huge information gathered with advanced technology and everyday increasing database about existed species efficiency or newly identified species with the help of computational biology and reference library database. However, metagenomics is also accompanied by various challenges to be dealt with in terms of data selection, selection of tools for processing and information extraction, and interpretation of the correct result. Thorough studies of literature and practice of computational biology help to solve these issues to great extent. Researchers are continuously involved in deploying updated bioinformatics tools, protocols, and databases; many of which are publicly available. Therefore expertise and careful selection of resources according to the target in chronology to process the metagenomics data and extract/interpret the valuable information are necessary.

Conflict of interest

The authors declare no conflict of interests.

References

- Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.
- Arndt, D., Xia, J., Liu, Y., Zhou, Y., Guo, A. C., Cruz, J. A., ... Wishart, D. S. (2012). METAGENassist: A comprehensive web server for comparative metagenomics. *Nucleic Acids Research*, *40*, W88–W95.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*, 455–477. Available from <https://doi.org/10.1089/cmb.2012.0021>.
- Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., & Corbeil, J. (2012). Ray Meta: Scalable de novo metagenome assembly and profiling. *Genome Biology*, *13*, R122. Available from <https://doi.org/10.1186/gb-2012-13-12-r122>.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*, 2114–2120.
- Bushnell, B. (2018). *BBTools: A suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data*. Joint Genome Institute.
- Cantu, V. A., Sadural, J., & Edwards, R. (2019). PRINSEQ++, a multi-threaded tool for fast and efficient quality control and preprocessing of sequencing datasets. *PeerJ*. Available from <https://doi.org/10.7287/peerj.preprints.27553v1>.
- Chen, I.-M. A., Markowitz, V. M., Chu, K., Palaniappan, K., Szeto, E., Pillay, M., ... Huntemann, M. (2016). IMG/M: Integrated genome and metagenome comparative data analysis system. *Nucleic Acids Research*, gkw929.
- Chen, J., Sun, S., Li, W., & Wooley, J. C. (2011). A community cyberinfrastructure resource for metagenomics research: CAMERA 2.0. *Metagenomics: Current Innovations and Future Trends*, 119.
- Chivian, D., Dehal, P. S., Keller, K., & Arkin, A. P. (2012). MetaMicrobesOnline: Phylogenomic analysis of microbial communities. *Nucleic Acids Research*, *41*, D648–D654.
- Cox, M. P., Peterson, D. A., & Biggs, P. J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics*, *11*, 1–6.
- Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics (Oxford, England)*, *32*, 3047–3048.
- Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., & Field, K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, *345*, 60–63.
- Goll, J., Rusch, D. B., Tanenbaum, D. M., Thiagarajan, M., Li, K., Methé, B. A., & Yooseph, S. (2010). METAREP: JCVI metagenomics reports—An open source tool for high-performance comparative metagenomics. *Bioinformatics (Oxford, England)*, *26*, 2631–2632.
- Gordon, A., & Hannon, G. J. (2010). *Fastx-toolkit. FASTQ/A short-reads preprocessing tools* (unpublished) (p. 5). http://hannonlab.cshl.edu/fastx_toolkit.
- Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology*, *5*(10), R245–R249.
- Herath, D., Tang, S.-L., Tandon, K., Ackland, D., & Halgamuge, S. K. (2017). CoMet: A workflow using contig coverage and composition for binning a metagenomic sample with high precision. *BMC Bioinformatics*, *18*, 571.
- Hunter, S., Corbett, M., Denise, H., Fraser, M., Gonzalez-Beltran, A., Hunter, C., ... Maguire, E. (2014). EBI metagenomics—A new resource for the analysis and archiving of metagenomic data. *Nucleic Acids Research*, *42*, D600–D606.
- Imelfort, M., Parks, D., Woodcroft, B. J., Dennis, P., Hugenholtz, P., & Tyson, G. W. (2014). GroopM: An automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, *2*, e603.

- Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, 3, e1165.
- Keegan, K. P., Glass, E. M., & Meyer, F. (2016). MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods in Molecular Biology (Clifton, N.J.)*, 1399, 207–233. Available from https://doi.org/10.1007/978-1-4939-3369-3_13.
- Kristiansson, E., Hugenholtz, P., & Dalevi, D. (2009). ShotgunFunctionalizeR: An R-package for functional comparison of metagenomes. *Bioinformatics (Oxford, England)*, 25, 2737–2738. Available from <https://doi.org/10.1093/bioinformatics/btp508>.
- Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., & Knight, R. (2012). Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Current Protocols in Microbiology*, 27, 1E–5E.
- Kultima, J. R., Coelho, L. P., Forslund, K., Huerta-Cepas, J., Li, S. S., Driessen, M., . . . Bork, P. (2016). MOCAT2: A metagenomic assembly, annotation and profiling framework. *Bioinformatics (Oxford, England)*, 32, 2520–2523.
- Langille, M. G., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., . . . Knight, R. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31, 814–821.
- Lassmann, T., Hayashizaki, Y., & Daub, C. O. (2009). TagDust—A program to eliminate artifacts from next generation sequencing data. *Bioinformatics (Oxford, England)*, 25, 2839–2840.
- Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics (Oxford, England)*, 31, 1674–1676. Available from <https://doi.org/10.1093/bioinformatics/btv033>.
- Lu, Y. Y., Chen, T., Fuhrman, J. A., & Sun, F. (2017). COCACOLA: Binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge. *Bioinformatics (Oxford, England)*, 33, 791–798.
- Luo, C., Rodríguez-r, L. M., & Konstantinidis, K. T. (2014). MyTaxa: An advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Research*, 42, e73.
- Mallawaarachchi, V., Wickramarachchi, A., & Lin, Y. (2020). GraphBin: Refined binning of metagenomic contigs using assembly graphs. *Bioinformatics (Oxford, England)*, 36, 3307–3313.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetJournal*, 17, 10–12.
- Martínez-Alcántara, A., Ballesteros, E., Feng, C., Rojas, M., Koshinsky, H., Fofanov, V. Y., . . . Fofanov, Y. (2009). PIQA: Pipeline for Illumina G1 genome analyzer data quality assessment. *Bioinformatics (Oxford, England)*, 25, 2438–2439.
- Mikheenko, A., Saveliev, V., & Gurevich, A. (2016). MetaQUAST: Evaluation of metagenome assemblies. *Bioinformatics (Oxford, England)*, 32, 1088–1090.
- Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: A new versatile metagenomic assembler. *Genome Research*, 27, 824–834.
- Pandey, R. V., Nolte, V., & Schlötterer, C. (2010). CANGS: A user-friendly utility for processing and analyzing 454 GS-FLX data in biodiversity studies. *BMC Research Notes*, 3, 3.
- Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25, 1043–1055.
- Peng, Y., Leung, H. C., Yiu, S.-M., & Chin, F. Y. (2012). IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics (Oxford, England)*, 28, 1420–1428.
- Roux, S., Tournayre, J., Mahul, A., Debroas, D., & Enault, F. (2014). Metavir 2: New tools for viral metagenome comparison and assembled virome analysis. *BMC Bioinformatics*, 15, 1–12.
- Sato, K., & Sakakibara, Y. (2015). MetaVelvet-SL: An extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. *DNA Research*, 22, 69–77.
- Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., . . . Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75, 7537–7541. Available from <https://doi.org/10.1128/AEM.01541-09>.
- Schmieder, R., Lim, Y. W., Rohwer, F., & Edwards, R. (2010). TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics*, 11, 341.
- Sedlar, K., Kupkova, K., & Provaznik, I. (2017). Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Computational and Structural Biotechnology Journal*, 15, 48–55. Available from <https://doi.org/10.1016/j.csbj.2016.11.005>.
- Seppy, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing genome assembly and annotation completeness. *Gene prediction* (pp. 227–245). Springer.
- Song, K., Ren, J., & Sun, F. (2019). Reads binning improves alignment-free metagenome comparison. *Frontiers in Genetics*, 10. Available from <https://doi.org/10.3389/fgene.2019.01156>.
- Strous, M., Kraft, B., Bisdorf, R., & Tegetmeyer, H. (2012). The binning of metagenomic contigs for microbial physiology of mixed cultures. *Frontiers in Microbiology*, 3, 410.
- Su, C.-H., Hsu, M.-T., Wang, T.-Y., Chiang, S., Cheng, J.-H., Weng, F. C., . . . Tsai, H.-K. (2011). MetaABC—An integrated metagenomics platform for data adjustment, binning and clustering. *Bioinformatics (Oxford, England)*, 27, 2298–2299.
- Treangen, T. J., Koren, S., Sommer, D. D., Liu, B., Astrovskaia, I., Ondov, B., . . . Pop, M. (2013). MetAMOS: A modular and open source metagenomic assembly and analysis pipeline. *Genome Biology*, 14, R2.

- Wang, Z., Wang, Z., Lu, Y. Y., Sun, F., & Zhu, S. (2019). SolidBin: Improving metagenome binning with semi-supervised normalized cut. *Bioinformatics (Oxford, England)*, *35*, 4229–4238.
- Wickramarachchi, A., Mallawaarachchi, V., Rajan, V., & Lin, Y. (2020). MetaBCC-LR: Meta genomics binning by coverage and composition for long reads. *Bioinformatics (Oxford, England)*, *36*, i3–i11.
- Wingett, S. W., & Andrews, S. (2018). FastQ Screen: A tool for multi-genome mapping and quality control. *F1000Research*, *7*.
- Wommack, K. E., Bhavsar, J., Polson, S. W., Chen, J., Dumas, M., Srinivasiah, S., . . . Nasko, D. J. (2012). VIROME: A standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences*, *6*, 421–433.
- Wu, Y.-W., & Ye, Y. (2011). A novel abundance-based algorithm for binning metagenomic sequences using l-tuples. *Journal of Computational Biology*, *18*, 523–534. Available from <https://doi.org/10.1089/cmb.2010.0245>.
- Wu, Y.-W., Simmons, B. A., & Singer, S. W. (2016). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics (Oxford, England)*, *32*, 605–607.
- Yu, G., Jiang, Y., Wang, J., Zhang, H., & Luo, H. (2018). BMC3C: Binning metagenomic contigs using codon usage, sequence composition and read coverage. *Bioinformatics (Oxford, England)*, *34*, 4172–4179.
- Yue, Y., Huang, H., Qi, Z., Dou, H.-M., Liu, X.-Y., Han, T.-F., . . . Tu, J. (2020). Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinformatics*, *21*, 334. Available from <https://doi.org/10.1186/s12859-020-03667-3>.
- Zhou, Q., Su, X., Wang, A., Xu, J., & Ning, K. (2013). QC-Chain: Fast and holistic quality control method for next-generation sequencing data. *PLoS One*, *8*, e60234. Available from <https://doi.org/10.1371/journal.pone.0060234>.