Review article

An overview of Metagenomics: The current era of bio-analysis

Aravinthkumar A[1]*, Savita Jandaik [1], Harender Raj Gautam [2] and J.Sheela [3]

[1] Dr. YS Parmar University of Horticulture and Forestry, Nauni-Solan, India
[2] Dept. of Plant Pathology, Dr. YS Parmar University of Horticulture and Forestry, Nauni-Solan, India
[3] Dept. of Plant Pathology, Tamil Nadu Agricultural University, Coimbatore, India
*** Correspondence author**: Aravinthkumar A: aravinth98a@gmail.com

Abstract: Microbes are ubiquitous in nature, and exist as the true elixir of mankind. The estimated number of microbial cells on Earth hovers around to be 10^{30}. The conventional microbial techniques enable us to grow microbes in the laboratory and study them under *in vitro* conditions. But those techniques are not powerful enough to culture diverse microbes at once, as they are time-consuming and contamination prone. Omics sciences have become core scientific tools to characterize microbiomes and study the natural communities and discover new microbes and their genes from the environment at a real fast time. Metagenomics seems to be the ideal culture-independent technique for unravelling the biodiversity of samples in addition to its clinical and diagnostic approaches. Metagenomics is the study of genomes recovered from the environmental sample using the advanced bioinformatics tools and genetic technologies This Omics approach is a vast field of growing interest among the scientific community and the need for efficient cultivation strategies has led to many rapid methodological and technological advances. The most difficult issue for the user is definitively not to get lost among all possible choices. Genome-resolved metagenomics, an amazing hypothesis generator, has revolutionized our ability to understand the uncultured microbes and catalysed unprecedented discoveries that have impacted the multiple fields from biogeochemistry to evolutionary biology. Thus, in this review, we have quantified the development of metagenomic application in examining the microbial ecology and we have also discussed some of the important informatics tools deployed in the metagenomic studies.

Keywords: Metagenomics, Microbiome, High-throughput sequence, Culture-independent, Meta barcodes.

Citation: Aravinthkumar et al. 2022. An overview of Metagenomics: The current era of bio-analysis. Octa J. Biosci. Vol. 10 (1):13-26

Received: 03/02/2022
Revised: 2/6/2022
Accepted: 25/6/2022
Published:30/6/2022

1. Introduction

Microbiology has experienced a transformation during the past two decades that has altered the view of micro-organisms and how to study them. The study of microbes that have been spread throughout the planet encounters many challenges through the discovery of novel unknown organisms and by establishing their interactions with the environment (Pavlovic et al., 2020). The advent of computational methods to collect, process and extract valuable biological information from complex microbial communities made way for the new frontier of science, Metagenomics (Alves et al., 2018). The metagenomics stream of science is new to us because it is not traditionally used in genetics and it has evolved so recently. The study is an interdisciplinary science that unites chemistry and biology. This evolving science integrates a computational approach that transits classical microbiology to modern omics science. Now, the availability of high throughput sequencing technologies tackles the chaos of DNA sequencing err. Also, the metagenomic approach bypasses the requirement of obtaining pure culture for sequencing and holds the promise of

revealing the genomes of the majority of microorganisms (Pérez-Cobas et al., 2020). Through these automated technologies, the test that takes days to perform can be performed in a few hours.

The word Metagenomics was first coined in 1998 by Jo Handlesman in her paper, "Molecular biological access to the chemistry of unknown soil microbes". The word Metagenomics has its root words from Greek, 'Meta- Transcendent' means sampling of organisms together, and 'Genomics- genome' means genetic complement *i.e.*, DNA sequences of the microbiome(Handelsman, 2005). Metagenomics, the branch of Omics sciences, is defined as the microbial genetic material analysis that is recovered from the environmental samples. The genomic analysis of microorganisms by direct DNA extraction and cloning it to form an assemblage of micro-organisms. The approach involves culture-independent analysis of the collective genome of organisms. The metagenomics studies are now commonly used in microbial ecology studies about the microbial communities in more detail. The informatics studies make it possible to mine huge datasets that govern microbial ecosystems(Kennedy et al., 2010). It also refers to Eco-genomics or Community genomics or Environmental genomics (Lorenzi et al., 2011). In simple assumption, it is considered as the study of the microbiome. Microbiome in the sense, microbiota which is the totality of microbes *viz.*, Bacteria, Fungi, Viruses, Protozoa, Algae, and their theatre of activity in a well-defined distinct bio-physio-chemical properties. The study discourses the microbial structural elements especially their proteins, lipids, polysaccharides, nucleic acids, and internal or external interaction of structural elements like primary and secondary metabolite production involving signaling, toxins production, and organic molecules. Coupled with next-generation sequencing (NGS) technologies, metagenomics even quick spots plant diseases to taxonomic ranks without the necessity of erstwhile information about the host or pathogen. It provides an extensive and accurate assessment of the abundance of phylogenetic and functional diversity of microbes in a niche (Dash and Das, 2018).

The interest in "Who they are?" and "What they are doing?" and the realization of the microbial world changed our view on biological diversity. At this age, visualization of microbes and quantifying them to categorize through a microscope had its limits (Bharti and Grimm, 2021). Microbes with similar phenotypic characters in community assemblage can't be classified without knowing the specific metabolites. To address this crisis of microbial physiology and taxonomy, the "Pure culture technique" was developed by Robert Koch and due to the discovery of the Petri plate, conical flask, autoclave, and basic laboratory techniques in his age, it is regarded as the Golden age of Microbiology (Forbes et al., 2017). It paved the way for an exponential increase in knowledge of microbial diversity. As years went on it became clear that culturing techniques were not kept on its immense diversity of microbes.

Fredrick Sanger's chain-terminating dideoxynucleotides sequencing and Maxam-Gilbert's chemical cleavage methods ushered the study of microbial ecology and its diversity in 1977. Staley et al. deemed the discrepancy diversity of organisms that can be cultured versus non-culturable organisms by giving "The Great Plate Count Anomaly" theory of 1985. However, the representative culture organisms of different niches could not be considered for establishing phylogenetic relations (Stewart, 2012). Carl Woese in lathe te 9s, addressed that in the extreme, interspecies gene exchanges could be so rampant and broadspread that a bacterium would not have its history on its own; it would be an evolutionary chimera, each with its history. Carl Woese and George E. Fox were the two people who pioneered the usage of 16s rRNA sequencing technology to study the

phylogenetic relationships in organisms in 1977. These studies expanded our knowledge in better understanding culture-independent microbial diversity in a community.

In 1985, Carl Woese's experimental advance of rRNA gene providing chronometers radically changed the way of visualization of the microbial world. In the same year, Pace et al. developed direct cloning of environmental samples without culturing them prior. But the first successful metagenomic library was created in 1991 (Lee et al., 2011). Early studies relied on direct RNA sequencing or reverse transcription-generated DNA copies. The technical breakthrough through the development of PCR primers and the disgorgement of novel microorganisms for the study seamlessly ended the limitations of sequencing technologies. However, it does not provide information on enzymatic abilities and has a major workload. Even today we are battling this chaos and now it has a new term "The Uncultured Majority" (Xu and Zhao, 2018). These anomalies had been exploited by Stein et al. (1996) by pushing the field forward by metagenomic sequencing of Hawaiian ocean water (but yet the name has not been coined). They employed the selective media concept to search for specific metabolic phenotypes in biosynthesis gene clusters (BGCs) in Araceae. This approach of natural sample-driven isolation has been considered as meta profiling which uses the 16s amplicon sequencing technique (access one genome at a time) (Trindade et al., 2015). Then Jo Handelsman recognized the use of the entire environmental sample's DNA (e-DNA) to discover novel BGC loci. Jo Handlesman and co-workers (1998) transferred the potential genomic fragments i.e., e-DNA obtained from soil samples into Fosmid vectors and expressed them in *E. coli*, they screened the specific phenotype of interest i.e., the collective genome of the soil. Her groundwork laid the foundation to analyze the functional and taxonomic sequences from a collective genome of a sample and she termed this new branch of Omic sciences "Metagenomics".

The collective genome obtained raised the concern of assessing the immense amount of data from the environmental samples. To resolve this chaos, numerous data analysis innovations in comparative metagenomics from clustering orthologs to gene catalogs have evolved (Dai and I, 2016). These innovations made the technique possible to assemble and extract groups of metagenomic contigs that represent a collective genome from a population of similar microbes. Tyson et al. unprecedentedly used shotgun metagenomic sequencing as Stein et al. for DNA extraction in 1996 and they accessed the functional repertoire of challenging contigs in 2004. But the key strategies they obtained to construct the genome are, adjusting the assembly to not penalize non-uniform read depth and allowing reading mapping of 95% identity. The work by Tyson et al. made a ground-breaking transition of the metagenomic approach primarily from a real-time phase to genome centric phase. This random community genomic approach, the metagenomic approach led to the development of next-generation technologies to sequence total DNA from the sample and the unprecedented discoveries of novel organisms through High Throughput Sequencing (HTS) led to the advancement of knowledge on the diversity in nature. Advances in high-throughput sequencing (HTS) has fostered rapid developments in the field of microbiome research and massive microbiome datasets are now being generated with the help of long-reads. Long-read sequencing technologies are overcoming early limitations in accuracy and throughput, broadening their application domains in genomics. In recent years, even high-throughput sequencing technologies have enabled us to identify even many novel species, new viruses, and enzymatic pathways in various crops. HTS techniques, particularly metabarcoding, are useful in the surveillance of soilborne, seed-borne, and airborne pathogens, as well as for identifying new pathogens and determining the origin of outbreaks. These sequencing studies can also give insights into the

functional potential of the micro-organisms identified from the diversified communities. In additionthe, Metagenomics technique is revolutionizing the field of microbiology by analyzing the population of many unculturable and unknown microbes which has excited researchers in many disciplines that could benefit from the study of microbiomes, including those in ecology, environmental sciences, and life sciences. A better understanding of these evolutionary drivers of microbial interaction, including the identification of origin will be crucial for rationally developing the treatments for the microbiomes.

1.1. Need for Metagenomic technique

Metagenomics analyzed the complete genome of an organism easier in a real quick time by high throughput sequencing of base pairs or nucleotides. In contrast, it involves sampling the genome sequences of a community of organisms inhabiting a common environment (Fanning et al., 2017). It provides a relatively unbiased view of community structure i.e., species richness and distribution, and also the functional (metabolic) potential of the microbial community. The main advantage of metagenomic is that any amenable environment sample provided can be used for analysis (Chikere et al., 2019). As every sequence read is derived from a different individual from a different community, it provides broad insights into the sympatric populations. The whole or near-complete genome of dominant species can be reconstructed using random sequencing(Kunin et al., 2008). It strictly speaks on the genome-centric approach, which is becoming a basic lab technique to understand the ecology and evolution of microbial ecosystems. Before the advancement of metagenomic science, we were completely oblivious to what we didn't know (unknown unknowns). Even with the latest computational resources or blueprints, we can't read many of the instructions (known unknowns) (Chatzivassiliou, 2021). This science advances the research remedies in the areas of medicine, agriculture, energy production, and bioremediation (Prasad et al., 2021).

2. Experimental design

The first and foremost part of a scientific study or an experiment is the design (Garlapati et al., 2019). The experiment should answer the biological query that has been raised. It should fit the work's objective. Many contrasting results will be termed when various approaches or studies are carried out. So, one should choose the best approach for analyzing the needed one. This Omics science also deals with two kinds of approaches towards microbiota under study based on the target of sequencing (Pavlovic et al., 2020).

Though the approach of both target and untargeted methods differs, the workflow is similar for both amplicon and whole-genome sequencing.

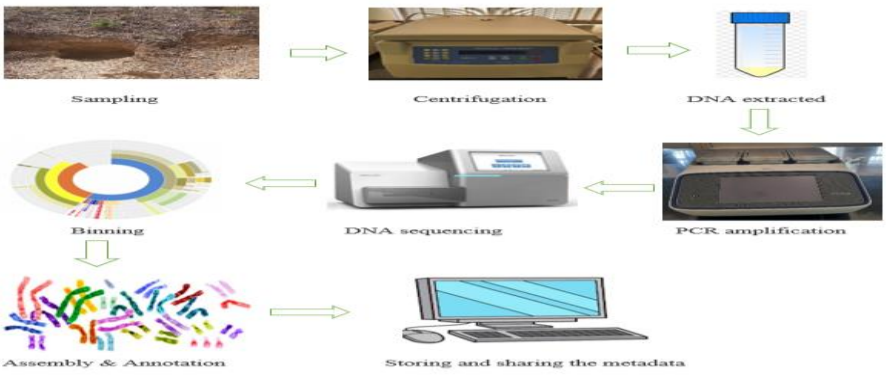

Figure 1. Overview of the Metagenomics study

Table 1: Target method vs. Untargeted method (Csabel & Tinahones, 2019)

Characteristics	Target method	Untargeted method
Sequencing of	Specific microbial amplicon 16s rRNA method	Whole-genome shotgun method for microbial community
Applies to	Only one microbial gene	Random genomic fragments
Performs	Phylogenetic and genera level taxonomic profiling. Functional pathways can also be exploited with the advent of bio-informatic software	Function Profiling, Genomic assembly, Species-level taxonomic profiling, Gene prediction, and metabolic pathways
Operational Taxonomic Units (OTUs)	Directly matches reference taxa that are available priorly	Assembling contigs and scaffolds, and annotating datasets for 95% identity
Resolution	Low level of dataset resolution	High level of dataset resolution
Requirement	Highly skilled laborers are needed to run the analysis	This automated pipeline needs less skilled labors
Cost of analysis	Lower compared to the shotgun method	Higher
The approach can be intensified by	Surveying a large number of individuals	Sequencing subset of samples
Pros	Great depth, precise	Sequence 100%, higher resolution, assess function and identify novel features
Cons	Unequal amplification, Lesser resolution, Limited OUT's, impossible to access functional potential	Not as much depth, Sequences contaminants, complex dataset

2.1. Sampling and DNA extraction

Accuracy of the result is the foremost need of any sample investigation and successful evaluation. So, to meet it, one should collect proper species enriched samples from the environment (Amarasinghe et al., 2020). Thereby then one can carry out the proper processing techniques as per the need of study before analysis. The processing is nothing other than the environmental DNA extraction by either of the two methods viz., Direct or Indirect. In the indirect method, DNA is separated from cell debris by direct cell wall lysis in the environmental sample. But this technique end by representing the mixture of DNA with less sheered quality of cells. The presence of high humic acid content also blocks some enzymatic reactions. In the indirect method, the cells are separated from the sample and lysed by enzymes or by chemical or mechanical disruptions including thermal or free-shaw shocks, ultrasonication, and bead homogenization (Ahmadi et al., 2018).

The certain limitations for DNA extraction include (Shuikan et al., 2020): 1) DNA extracted should be representative of all cells present in the sample 2) Sufficient amounts of high-quality nucleic acids must be obtained for subsequent library production and sequencing 3) Sample fractionation should be checked and ensured to enrich the target material 4) Physical separation and isolation of cells plays a vital role in DNA yield 5) Lysis of cells has a quantifiable bias in terms of microbial diversity, DNA yield, fragment sequence length 6) Some type of sample yield very small amounts of DNA but in library production for most sequencing technologies require high amounts of DNA (ng or µg), and hence amplification of starting material might be required. Multiple displacement amplification (MDA) using random hexamers and phage phi29 polymerase is one option employed to increase DNA yields, this method has been widely used in single-cell genomics and to a certain extent in metagenomics.

2.1.1.Metagenomic Library

The extracted DNA of the representative organisms can be investigated by integrating it into the host vector for further analysis. To construct a genomic library, the predominant key feature like copy number, insert capacity, choice of host, and screening procedure should be taken into consideration. Bacterial artificial chromosomes (BAC - 200 KBS insert capacity) and Fosmids (35-45 KBS insert capacity) are the vectors for amplifying large fragments. For short fragments Plasmids (15kbs insert capacity) are the most frequently used vectors for constructing a metagenomic library. Currently, *Escherichia coli* is used as the ideal host for the library construction at utmost importance for the effective cloning and expression of insert genes.

2.1.2.High Throughput Sequencing platforms

- ***First-generation sequencing platform***

The pioneering work of Fiers et al. (1976) that enabled the entire genome sequenced belonged to a small ssRNA virus, bacteriophage. The first DNA-based genome of PhiX 174 was sequenced by Sanger (Sanger et al., 1977). It is considered first-generation sequencing or Sanger sequencing technology. The first organism, *Haemophilus influenza* whose entire genome of 1.8 billion base pairs was sequenced in 1995 (Fleischmann et al., 1995). However, the heterogeneity of mixed microbial communities, since the extracted DNA has to be cloned and integrated into a bacterial vector, considerably limits the metagenomic studies to low-diversity microbiomes. Due to this obligation of the cloning step, only the culturable organisms in the laboratory were suitable for the sequencing. It created the challenges of cost limitation, low throughput, and the requirement of a large quantity of starting material (Köser et al., 2012). A highly diverse microbiome may produce an oversimplified picture and not be adequately characterized using cloning, and sequencing of a multitude of samples, making Sanger sequencing challenging, to say the least, for metagenomic studies. These limitations of the pre-HTS platform led to the development of Next Generation Sequencing technologies (Heather & Chain, 2016).

- ***Second-generation sequencing platform***

The first High-throughput sequencing (HTS) technique is Roche 454. Roche 454 was originally released in 2005 but later acquired by Roche in 2007 (Lee et al., 2011). The technique involves, capturing the molecule in beads which are further loaded in wells of picotiter plate for amplification using PCR emulsion. Finally, the molecules are sequenced using Pyrosequencing. It uses optical detection to identify the proximate base of nucleotides and uses polymerase to drive sequences. GS FLX+ Titanium is the last manufactured sequencer of this type. With concomitant termination of support, Roche discontinued this platform in 2013.

Solexa released a sequencer in 2005 (now Illumina sequencer). It is based on sequencing by synthesis of fluorescently labeled dye-terminator and clonal amplification of adapters of glass slide surface or flow cell. Through reversible cyclic termination, bases are identified by imaging. It is the widely used high throughput instrument with the lowest error rate and cost-efficiency among currently available HTS platforms (Shuikan et al., 2020). NovaSeq 6000 is the currently released sequencer of this type of platform.

Applied Biosystems released SOLiD in 2007 and it merged with Invitrogen in 2008. Later, it outplayed the rest of the techniques and become a life technology. In 2014, it was acquainted by the Thermo Fisher Scientific company (Malla et al., 2019). It utilizes repeated cycling of DNA ligation

and the fluorescently labeled dye determines the nucleotide sequences twice. It has been regarded as the second-highest throughput provider.

Ion torrent (2010) amplifies adaptor-ligated fragments on beads using emulsion PCR similar to Roche 454 sequencer (Tripathi et al., 2019). But it differs from pyrosequencing in a base determination by changing pH which is released from the hydrogen ion during base incorporation. In this type Proton 1, is the latest instrument capable of producing high throughputs.

- ***Third-generation sequencing platforms***

PacBio (now Illumina) released in late 2010, is the most widely used technology in this class. It stands alone as the only platform providing both second and third-generation sequencing capability. It uses hairpin adaptors in addition to single-molecule sequencing to form closed ssDNA (Rhoads & Au, 2015). The key feature of this sequencer is the detection of fluorescent-labeled nucleotides at the phosphate group generated as Zero-Mode Waveguide (ZMW) of SMRT bell (Single-Molecule Real-Time) template. The latest instrument Sequel generates the highest read length but also has a high error rate.

Table 2: Comparison of High throughput sequencing platforms (Dai & I, 2016)

Platforms	Generation	Method of detection	Proximate source of base	Sequencing chemistry	Read length(bp)	Reads per run
Sanger Sequencing	First	Optical	ssDNA	Uses Polymerase solution	600-800	96
Illumina	Second	Optical	Clonally amplified DNA	Uses Polymerase solution	2×125- 2×250	8×10^{9}
PacBio	Third	Optical	Single-molecule	Uses Polymerase solution	2×125	$3.5\text{-}7.5\times10^{4}$
SOLiD	Second	Optical	Clonally amplified DNA	Ligation	2×60	8×10^{8}
454	Second	Optical	Clonally amplified DNA	Uses Polymerase solution	700	1×10^{6}
Ion torrent	Second	Non-optical	Clonally amplified DNA	Uses Polymerase solution	200	8.2×10^{7}
Nanopore	Third	Optical	Single-molecule	Direct sequence determination	$2\text{-}5\times10^{3}$	$1.1\text{-}4.7\times10^{4}$

Nanopore sequencer by oxford nanopore in 2014, is a scalable unique technology. It characterizes the changes in current induced to detect bases that pass through the biologic nanopore by anchoring on to it by a molecular motor protein (van Dijk et al., 2018). Comparatively, the PromethION device offers the highest throughput of the other three instruments available on this platform. Still, MinION like the platform is a cost-effective and real-time sequencer if there is no need for large datasets.

As a result of decades of laborious work and the advent of sequencing technologies, the number of genomes sequenced continued to increase throughout the years. It expands the interest in genomic research on human and environmental microbiomes. This expensive international endeavor relied entirely on software and hardware automation.

2.1.3. Barcodes in Metagenomics

The assessment of a combination of the microbiome in terms of species identified in the processed sample as sequenced data of specific gene fragments that are shared with several species is

done by employing barcodes (Rotimi et al., 2018). The *K-mer* frequency (Basepair) distribution is unique to each genome that is termed as *barcodes* of life (Tamames et al., 2019). With these known *K-mer* short sequence fragments, the problem in binning and in identifying the horizontal gene transfer in the organism which are under study can be addressed. The wide variety of barcode combinations used to analyze the wide spectrum of species diversity in the sample is called metabarcoding (Piombo et al., 2021). The most frequently used marker or barcode of life for identifying bacteria, fungi, Algae, and Viruses at their species level are 16s rRNA, Internally Transcribed Spacer regions (ITS), Large Subunits Divergent Domains (LSU D), House Keeping genes respectively. An ideal meta barcode or marker in metagenomic should 1) be *present* in all the organisms (in all the cells) and without the barcodes, the translation process is not possible 2) have *variable sequences* among different species 3) be *conserved* among individuals of the same species 4) be *easy to amplify* and not too long for sequencing (Porter & Hajibabaei, 2018). The barcode match/ similarity is directly proportional to the genome phylogenetic relatedness of a species.

2.1.4. Screening of the library

The raw library is found to be complex and technically demanding. So, the constructed vector library is screened based on its functional and structural sequences.

Table A3: Screening of sequences between functional and structural basis (Bharti & Grimm, 2021)

Metagenomics	Functional	Structural
Involves	Gene constructions → Screening → heterologous expression → bioinformatics analysis and protein product characterization	Assembly → binning → Microbial community analysis
Process	cloning DNA fragments, expressing genes in a surrogate host, and screening for enzymatic activities	Novel product elucidation, cloning DNA fragments and screening for Biosynthesis Gene Clusters, and identification
Analysis	Sequence analysis, farm analysis, structural prediction, Phylogenetic analysis, Protein activity, optimum temperature, and pH	Taxonomic profiling, Whole-genome prediction, Species diversity (α, β, γ), and metabolic pathway
Makes	Better understanding the gene function and biochemical pathways	A better understanding of ecology and biological information

2.2. Assembly

A small part of the data analysis process is the assembly approach. The increased use of metagenomic analysis in biological research has led to the development of integrated pipelines. Such pipelines include MetAmos and MOCAT which are stand-alone assembly packages, as well as ClovR a framework that enables metagenomic analyses on cloud computing frameworks (Dubey et al., 2020). The stitching together of individual DNA sequences into genes or organisms is the critical stage of metagenomic assembly. Genome assembly is the reconstruction of genomes from the smaller DNA segments called reads (pair or mate pairs) generated by a sequencing experiment (Ghurye et al., 2016). The ambiguities caused by repetitive sequences during assembly contigs-the genome of the fragment stitched together from the set of reads will be resolved. The assembly is of two types viz., *De nova* assembly involves reconstructing genomes directly from read data and comparative assembly uses the previously sequenced closely related organisms (Kieser et al., 2019). The most frequently used assemblers include Metavelvet, Meta-IDBA, MeGAHIT, and Ray. These assemblers use the de Bruijn graph approach. After filtering and correcting the error, many errors and poly- morphisms remain in the data, causing an increase in the size of the resulting size of the de Bruijn graph (Escobar-Zepeda et al.,

2015). Bloom filters (trims) with an inexact data structure that trades off accuracy for memory size, are introduced as an extension approach that also compactly represents the original information without losing the space efficiency (Alves et al., 2018).

2.3. Annotation

Annotation is the final step of automated computational processing of the metagenomic dataset. Annotation, the post-read analysis identifies the genes, encoded proteins (ORF- Open Reading Frames), and those encoded rRNA or tRNA molecules of orthologs accurately (Dong & Strous, 2019). This step is regarded as the beginning of biology with the possible biological function that is correctly matched. The best annotation pipeline depends on the data available, computational resources, and the research problem taken into consideration. The options for genome annotation exist in two flavors: online platforms and standalone pipelines. IMG, MG-RAST, MicroScope, Magnify, and Edge is the online platforms whereas MetaErg, Python or Perl, Scratch, Prokka, DFAST-Core, and PGAP are the standalone type used for annotating (Dudhagara et al., 2015). The genome-centric data provide challenges in annotating like 1) Poor assembly quality with contamination 2) user need to make sense of many annotated genomes simultaneously and 3) close reference genomes are not yet available.

Table 4: The tools mostly employed in the operation, storing, and sharing of metadata (Oulas et al., 2015)

Tools	**Function**
MED	Partitions the data set of amplicon sequences into homogenous OTUs for alpha- and beta-diversity analyses. Solves the limitations of fine-scale resolution descriptions of microbial communities
UPRASE	Operational taxonomic units (OTUs) can be generated. Filtering and trimming reads into equal lengths, removing singleton reads, and clustering the remaining reads
QIIME	Quantitative Insights Into Microbial Ecology (QIIME) generates the data on the Illumina or other NGS platforms via graphics and statistics which demultiplex and quality filters, OTU assignment, phylogenetic reconstruction, and diversity analyses and visualizations. It depends on the use of the PyCogent toolkit -identify misinterpretations and database deposition using raw sequencing results
MOTHUR	Annotates the community sequence data
DADA2	Correcting amplicon errors with no option to generate OTUs. It uses a new quality-aware model of Illumina amplicon errors to improve the DADA algorithm
MGRAST	For species-level metagenomic data analysis To identify the microbial profile and score their abundance
MetaPhlAn2	
Kraken	
CLARK	
FOCUS	
SUPERFOCUS	

2.4. Binning

The process of grouping reads or contigs thereby assigning the individual genome is called binning. The genome recovery domain agglomerates or bins the sequences assembled and annotated reads into individual groups based on compositional or alignment (Thomas et al., 2014). An ideal binning tool should enable a clear distinction of clusters (the visualization of metagenomic data) and automatically produce accurate results. Modern binning techniques use both previously available information independent from the sample and intrinsic information present in the sample. With the diversity and complexity of the sample, a degree of success exists in resolving the sequences up to individual species and also up to very broad taxonomic groups (Knight et al., 2018). Without the consideration of reference sequences, binning enables the comprehensive discovery of new microbial organisms by aiding in microbial genome reconstruction. The most frequently used binning software are TETRA, MEGAN, Phylopthia, SOrt-ITEMS, DiScRIBinATE, ProViDE, PCAHIER, SPHINX, INDUS, and TWARIT (Liu et al., 2021). These binning wares operate the global view of hierarchical classification and function of diverse communities in a supervised or unsupervised manner.

3. Application of metagenomics in the scientific community

Metagenomics reveals the knowledge of microbial communities of uncultivable organisms in environmental niches by using various screening technologies based on sequence and function (Nowrotek et al., 2019). It has vast application in every aspect of life including,

- *The identification of novel organisms or gene clusters encoding for enzymes or drug discovery*

Specialized enzymes (also called natural products or secondary metabolites) derived from bacteria, fungi, marine organisms and plants are an important source of antibiotics, anti-cancer agents, insecticides, immunosuppressants, and herbicides. Many secondary metabolites in bacteria and fungi are biosynthesized via metabolic pathways whose enzymes are encoded by clustered genes on a chromosome (Chavali & Rhee, 2018). Metabolic gene clusters comprise a group of physically co-localized genes that together encode enzymes for the biosynthesis of a specific metabolite. Although metabolic gene clusters are generally not known to occur outside of microbes, several plant metabolic gene clusters have been discovered through the metagenomic studies in recent years.

- *Diagnose of disease*

The majority demonstrates mNGS that has sensitivity similar to PCR assays and identifies more potential pathogens simultaneously than conventional methods. The study offers the gateway to exploring and characterizing hidden microbial communities through a culture-independent mode by direct DNA isolation and sequencing. The mechanistic details of numerous microbes and their interaction with the niche. The major constraint is that the data obtained by the study is highly complex and multi-dimensional, it requires accurate analytical tools to evaluate and interpret the data (Wani et al., 2022).

- *Bioremediation and pollution monitoring*

Metals are persisting and non-biodegradable, which can enter the food chain via crop plants. The heavy metals might accumulate in the animal body through biomagnification. The microbial treatment of metals provides an excellent and new perspective in preventing environmental pollution through their specific biodegradation mechanisms at the molecular level. The metagenomic analysis of a populated environment might be used to simplify the processing and examination of specific genomic information in bioremediation experiments. Development in NGS demands detailed metagenomic analysis of environmental microbes offering unparalleled perspectives through key biosorption mechanisms (Sharma et al., 2021).

3.1. Limitations

Metagenomics is a global tool that consists of too much data making the analysis so complex. Through the advancement of many Omic tools, the relation between genes and survival through metagenomes can't be established (Garlapati et al., 2019). As adding support to the great plate count anomaly theory there exists many unidentifiable genes even through metagenomics. Biocatalyst's discovery remains a challenge even with the increased functional screening capabilities. Requiring high throughput sequencing instrumentation, tedious extraction, and contamination/chimera sequences are some of the constraints. It is very hard to predict phenotypic characters through metagenomic studies. The technique needs more precise information and a model organism is also welcome.

4. Conclusion

The rapid development of inexpensive high throughput sequencing technologies spurred the efforts to characterize the microbial communities inhabiting the environment, leading to the development of a new field - metagenomics. The opportunity for developing new algorithms for the analysis of data has been created, accounting for the specific characteristics of metagenomic data. Exploration of unexplored niches enables mining of the novel enzymes with desired characteristics, which aids in performing the biotechnological processes. Like many other technologies, metagenomics is still developing and therefore the technique faces many challenges. However, one cannot rule out the opportunities this technology offers to study the microbial world in particular, and the environment as a whole. Finally, we would like to conclude that tremendous opportunities exist for the development of methods that combine all the different techniques viz., Meta transcriptomics, metabolomics, metaproteomic, and interrogating microbial communities to provide a more complete understanding of the role these communities play in our world. Metagenomics provides a window into the world of unseen microbial diversity that can be explored using biotechnological tools, thereby paving the way to novel scientific, environmental, pharmaceutical, and industrial applications.

Author Contributions: Aravinthkumar A, Dr. Savita Jandaik, Dr. H R. Gautam and Dr. J.Sheela have equally contributed to the work and all the authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

Ahmadi, E., Kowsari, M., Azadfar, D., & Salehi Jouzani, G. (2018). Rapid and economical protocols for genomic and metagenomic DNA extraction from oak (Quercus brantii Lindl.). *Annals of Forest Science*, *75*(2). https://doi.org/10.1007/s13595-018-0705-y

Alves, L. D. F., Westmann, C. A., Lovate, G. L., De Siqueira, G. M. V., Borelli, T. C., & Guazzaroni, M. E. (2018). Metagenomic Approaches for Understanding New Concepts in Microbial Science. *International Journal of Genomics*, *2018*. https://doi.org/10.1155/2018/2312987

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 1–16. https://doi.org/10.1186/s13059-020-1935-5

Bharti, R., & Grimm, D. G. (2021). Current challenges and best-practice protocols for microbiome analysis. *Briefings in Bioinformatics*, *22*(1), 178–193. https://doi.org/10.1093/bib/bbz155

Chatzivassiliou, E. K. (2021). An annotated list of legume-infecting viruses in the light of metagenomics. *Plants*, *10*(7). https://doi.org/10.3390/plants10071413

Chavali, A. K., & Rhee, S. Y. (2018). Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. *Briefings in Bioinformatics*, *19*(5), 1022–1034. https://doi.org/10.1093/bib/bbx020

Chikere, C. B., Mordi, I. J., Chikere, B. O., Selvarajan, R., Ashafa, T. O., & Obieze, C. C. (2019). Comparative metagenomics and functional profiling of crude oil-polluted soils in Bodo West Community, Ogoni, with other sites of varying pollution history. *Annals of Microbiology*, *69*(5), 495–513. https://doi.org/10.1007/s13213-019-1438-3

Csabel, & Tinahones, F. J. (2019). Metagenomics. *Principles of Nutrigenetics and Nutrigenomics: Fundamentals of Individualized Nutrition*, 81–87. https://doi.org/10.1016/B978-0-12-804572-5.00011-2

Dai, X., & I. (2016). Page 1 of 28. *Reproduction*, *5*(June), 1–28.

Dash, H. R., & Das, S. (2018). Molecular Methods for Studying Microorganisms From Atypical Environments. In *Methods in Microbiology* (1st ed., Vol. 45). Elsevier Ltd. https://doi.org/10.1016/bs.mim.2018.07.005

Dong, X., & Strous, M. (2019). An Integrated Pipeline for Annotation and Visualization of Metagenomic Contigs. *Frontiers in Genetics*, *10*(October), 1–10. https://doi.org/10.3389/fgene.2019.00999

Dubey, R. K., Tripathi, V., Prabha, R., Chaurasia, R., Singh, D. P., Rao, C. S., El-Keblawy, A., & Abhilash, P. C. (2020). *Bioinformatics Tools for Soil Microbiome Analysis*. 61–70. https://doi.org/10.1007/978-3-030-15516-2_6

Dudhagara, P., Bhavsar, S., Bhagat, C., Ghelani, A., Bhatt, S., & Patel, R. (2015). Web Resources for Metagenomics Studies. *Genomics, Proteomics and Bioinformatics*, *13*(5), 296–303. https://doi.org/10.1016/j.gpb.2015.10.003

Escobar-Zepeda, A., De León, A. V. P., & Sanchez-Flores, A. (2015). The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. *Frontiers in Genetics*, *6*(DEC), 1–15. https://doi.org/10.3389/fgene.2015.00348

Fanning, S., Proos, S., Jordan, K., & Srikumar, S. (2017). A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. *Frontiers in Microbiology*, *8*(SEP), 1–16. https://doi.org/10.3389/fmicb.2017.01829

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., Merrick, J. M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J. D., Scott, J., Shirley, R., Liu, L. I., Glodek, A., … Venter, J. C. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, *269*(5223), 496–512. https://doi.org/10.1126/science.7542800

Forbes, J. D., Knox, N. C., Ronholm, J., Pagotto, F., & Reimer, A. (2017). Metagenomics: The next culture-independent game changer. *Frontiers in Microbiology*, *8*(JUL), 1–21. https://doi.org/10.3389/fmicb.2017.01069

Garlapati, D., Charankumar, B., Ramu, K., Madeswaran, P., & Ramana Murthy, M. V. (2019). A review on the applications and recent advances in environmental DNA (eDNA) metagenomics. *Reviews in Environmental Science and Biotechnology*, *18*(3), 389–411. https://doi.org/10.1007/s11157-019-09501-4

Ghurye, J. S., Cepeda-Espinoza, V., & Pop, M. (2016). Metagenomic assembly: Overview, challenges and applications. *Yale Journal of Biology and Medicine*, *89*(3), 353–362.

Handelsman, J. (2005). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, *69*(1), 195–195. https://doi.org/10.1128/mmbr.69.1.195.2005

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Kennedy, J., Flemer, B., Jackson, S. A., Lejon, D. P. H., Morrissey, J. P., O'Gara, F., & Dobson, A. D. W. (2010). Marine metagenomics: New tools for the study and exploitation of marine microbial metabolism. *Marine Drugs*, *8*(3), 608–628. https://doi.org/10.3390/md8030608

Kieser, S., Brown, J., Zdobnov, E. M., Trajkovski, M., & McCue, L. A. (2019). ATLAS: A Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BioRxiv*, 1–8. https://doi.org/10.1101/737528

Knight, R., Vrbanac, A., Taylor, B. C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciolek, T., McCall, L. I., McDonald, D., Melnik, A. V., Morton, J. T., Navas, J., Quinn, R. A., Sanders, J. G., Swafford, A. D., Thompson, L. R.,

Tripathi, A., Xu, Z. Z., … Dorrestein, P. C. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology*, *16*(7), 410–422. https://doi.org/10.1038/s41579-018-0029-9

Köser, C. U., Ellington, M. J., Cartwright, E. J. P., Gillespie, S. H., Brown, N. M., Farrington, M., Holden, M. T. G., Dougan, G., Bentley, S. D., Parkhill, J., & Peacock, S. J. (2012). Routine Use of Microbial Whole Genome Sequencing in Diagnostic and Public Health Microbiology. *PLoS Pathogens*, *8*(8). https://doi.org/10.1371/journal.ppat.1002824

Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K., & Hugenholtz, P. (2008). A Bioinformatician's Guide to Metagenomics. *Microbiology and Molecular Biology Reviews*, *72*(4), 557–578. https://doi.org/10.1128/mmbr.00009-08

Lee, J. H., Yi, H., & Chun, J. (2011). rRNASelector: A computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *Journal of Microbiology*, *49*(4), 689–691. https://doi.org/10.1007/s12275-011-1213-z

Liu, Y. X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X., & Bai, Y. (2021). A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein and Cell*, *12*(5), 315–330. https://doi.org/10.1007/s13238-020-00724-8

Lorenzi, H. A., Hoover, J., Inman, J., Safford, T., Murphy, S., Kagan, L., & Williamson, S. J. (2011). The viral metagenome annotation pipeline (VMGAP): An automated tool for the functional annotation of viral metagenomic shotgun sequencing data. *Standards in Genomic Sciences*, *4*(3), 418–429. https://doi.org/10.4056/sigs.1694706

Malla, M. A., Dubey, A., Kumar, A., Yadav, S., Hashem, A., & Allah, E. F. A. (2019). Exploring the human microbiome: The potential future role of next-generation sequencing in disease diagnosis and treatment. *Frontiers in Immunology*, *10*(JAN), 1–23. https://doi.org/10.3389/fimmu.2018.02868

Nowrotek, M., Jałowiecki, Ł., Harnisz, M., & Płaza, G. A. (2019). Culturomics and metagenomics: In understanding of environmental resistome. *Frontiers of Environmental Science and Engineering*, *13*(3). https://doi.org/10.1007/s11783-019-1121-8

Oulas, A., Pavloudi, C., Polymenakou, P., Pavlopoulos, G. A., Papanikolaou, N., Kotoulas, G., Arvanitidis, C., & Iliopoulos, I. (2015). Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinformatics and Biology Insights*, *9*, 75–88. https://doi.org/10.4137/BBIi.Ss12462.

Pavlovic, S., Klaassen, K., Stankovic, B., Stojiljkovic, M., & Zukic, B. (2020). Next-Generation Sequencing: The Enabler and the Way Ahead. In *Microbiomics*. INC. https://doi.org/10.1016/b978-0-12-816664-2.00009-8

Pérez-Cobas, A. E., Gomez-Valero, L., & Buchrieser, C. (2020). Metagenomic approaches in microbial ecology: An update on whole-genome and marker gene sequencing analyses. *Microbial Genomics*, *6*(8), 1–22. https://doi.org/10.1099/mgen.0.000409

Piombo, E., Abdelfattah, A., Droby, S., Wisniewski, M., Spadaro, D., & Schena, L. (2021). Metagenomics approaches for the detection and surveillance of emerging and recurrent plant pathogens. *Microorganisms*, *9*(1), 1–19. https://doi.org/10.3390/microorganisms9010188

Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology*, *27*(2), 313–338. https://doi.org/10.1111/mec.14478

Prasad, R., Kumar, V., Singh, J., & Upadhyaya, C. P. (2021). *Recent Developments in Microbial Technologies*. http://link.springer.com/10.1007/978-981-15-4439-2

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Rotimi, A. M., Pierneef, R., & Reva, O. N. (2018). Selection of marker genes for genetic barcoding of microorganisms and binning of metagenomic reads by Barcoder software tools. *BMC Bioinformatics*, *19*(1), 1–12. https://doi.org/10.1186/s12859-018-2320-1

Sanger, F., Nicklen, S., & Coulson, A. . (1977). DNA sequencing with chain-terminating. *Proc Natl Acad Sci USA*, *74*(12),

5463–5467.

Sharma, P., Kumar, S., & Pandey, A. (2021). Bioremediated techniques for remediation of metal pollutants using metagenomics approaches: A review. *Journal of Environmental Chemical Engineering*, *9*(4), 105684. https://doi.org/10.1016/j.jece.2021.105684

Shuikan, A., Ali Alharbi, S., Hussien M. Alkhalifah, D., & N. Hozzein, W. (2020). High-Throughput Sequencing and Metagenomic Data Analysis. *Metagenomics - Basics, Methods and Applications*. https://doi.org/10.5772/intechopen.89944

Stewart, E. J. (2012). Growing unculturable bacteria. *Journal of Bacteriology*, *194*(16), 4151–4160. https://doi.org/10.1128/JB.00345-12

Tamames, J., Cobo-Simón, M., & Puente-Sánchez, F. (2019). Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes. *BioRxiv*, 1–16. https://doi.org/10.1101/522292

Thomas, T., Gilbert, J., & Meyer, F. (2014). Metagenomics: A guide from sampling to data analysis. *The Role of Bioinformatics in Agriculture*, *Figure 1*, 357–383. https://doi.org/10.1201/b16568

Trindade, M., van Zyl, L. J., Navarro-Fernández, J., & Elrazak, A. A. (2015). Targeted metagenomics as a tool to tap into marine natural product diversity for the discovery and production of drug candidates. *Frontiers in Microbiology*, *6*(AUG), 1–14. https://doi.org/10.3389/fmicb.2015.00890

Tripathi, V., Kumar, P., Tripathi, P., Kishore, A., & Kamle, M. (2019). Microbial genomics in sustainable agroecosystems: Volume 2. In *Microbial Genomics in Sustainable Agroecosystems: Volume 2* (Vol. 2). https://doi.org/10.1007/978-981-32-9860-6

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. *Trends in Genetics*, *34*(9), 666–681. https://doi.org/10.1016/j.tig.2018.05.008

Wani, A. K., Roy, P., Kumar, V., & Mir, T. ul G. (2022). Metagenomics and artificial intelligence in the context of human health. *Infection, Genetics and Evolution*, *100*(March), 105267. https://doi.org/10.1016/j.meegid.2022.105267

Xu, Y., & Zhao, F. (2018). Single-cell metagenomics: challenges and applications. *Protein and Cell*, *9*(5), 501–510. https://doi.org/10.1007/s13238-018-0544-5