Hindawi Publishing Corporation
Biotechnology Research International
Volume 2015, Article ID 121735, 13 pages
http://dx.doi.org/10.1155/2015/121735

Review Article

Metagenomics: Retrospect and Prospects in High Throughput Age

Satish Kumar,[1] Kishore Kumar Krishnani,[1] Bharat Bhushan,[2] and Manoj Pandit Brahmane[1]

[1] *ICAR-National Institute of Abiotic Stress Management, Baramati, Pune, Maharashtra 413115, India*
[2] *ICAR-Central Institute of Post-Harvest Engineering and Technology, Abohar Station, Punjab 152116, India*

Correspondence should be addressed to Bharat Bhushan; buddingbiochemist@gmail.com

Received 13 July 2015; Accepted 26 October 2015

Academic Editor: Manuel Canovas

In recent years, metagenomics has emerged as a powerful tool for mining of hidden microbial treasure in a culture independent manner. In the last two decades, metagenomics has been applied extensively to exploit concealed potential of microbial communities from almost all sorts of habitats. A brief historic progress made over the period is discussed in terms of origin of metagenomics to its current state and also the discovery of novel biological functions of commercial importance from metagenomes of diverse habitats. The present review also highlights the paradigm shift of metagenomics from basic study of community composition to insight into the microbial community dynamics for harnessing the full potential of uncultured microbes with more emphasis on the implication of breakthrough developments, namely, Next Generation Sequencing, advanced bioinformatics tools, and systems biology.

1. Introduction

Despite the exhaustive knowledge of intricate molecular mechanisms of most of the cellular processes and the availability of complex culture media, scientists are still able to culture less than 1% of all microorganisms present in diverse natural habitats. This leaves scientists unable to study more than 99% of the biological diversity in the environment with conventional techniques. Metagenomics is the function-based or sequence-based culture independent analysis of metagenomes trapped from a wide range of habitats. A typical metagenomic study combines the potential of genomics, bioinformatics, and systems biology in exploring the collective microbial genomes isolated directly from environmental samples. Course changing developments in recent times, like inexpensive Next Generation Sequencing (NGS) technologies, advanced bioinformatics tools, and high throughput screening (HTS) methods for metagenomic libraries, have left greatest impact on the science of metagenomics. These breakthrough developments have set a wave of excitement among large number of research groups all across the globe, triggering strong quest about the concealed potential of the existing microbial world beyond Petri dish. The cost of the large scale sequencing has reduced dramatically in the last few years. Using NGS, now it has become routine to generate hundreds of megabases of sequence data for expense of well under \$20,000 bringing metagenomics in reach of many laboratories across the globe [1]. These advances in sequencing technologies have fuelled the research on metagenomics and have laid the way for the scientific community to undertake mammoth projects generating huge amount of sequence data. Dinsdale et al. [2] in their study on metagenomic comparison of 45 distinct microbiomes and 42 viromes generated 15 million sequences employing Next Generation Sequencing (NGS) and revealed strong discriminatory metabolic profiles across all the investigated microbiomes. Although the large scale sequencing studies in the pilot project on Sargasso Sea [3] and its extension, the Sorcerer II Global Ocean Sampling expedition [4], were carried out using Sanger sequencing based ABI 3750XL sequencer, Sanger sequencing is no longer the main source of metagenomic sequence data. The impact of NGS technologies on metagenomics has been so profound that a typical metagenomic project in the recent times generates large amounts of sequence data and due to this dominance of sequence-based projects, Kunin et al. [1] have redefined the metagenomics as "application of shotgun sequencing to DNA obtained directly from environmental sample producing at least 50 Mbp randomly sampled sequence data." Metagenomic tools have allowed us the unprecedented access to the natural microbial communities and their potential activities. Metagenomics is now an established and prospered research arena and has completely suppressed the once prevailed erroneous notion that microorganisms did not exist unless they could be cultured. Initially, the research endeavours of most of the groups were primarily focused on answering the questions investigating "who are there" and have now shifted to finding key aspects of "what they are doing and how exactly they do it." The present review summarizes the historic landmarks critical in the progression of the science of metagenomics and also highlights the progress made during the last two decades for trickling novel functions in metagenomes. This review also encompasses the impact of course changing developments in DNA sequencing and bioinformatics in the progression of science of metagenomics.

2. Metagenomics: Inception, Landmarks, and Progression

Though the term metagenome came off late in 1998 [5], the reports about unculturability of microbes go hundred years back to 1898, when Heinrich Winterberg first reported about microbial unculturability, the so-called great plate count anomaly. Owing to the lack of culture methods for a major segment of the microbes, their genetic potential remained unutilised for a longer time. Before 1985, most of what was known to us about the existence of microbial world was derived from cultured microbes. The studies of Staley and Konopka [6] in 1985 regarding the existing data of that time on "great plate count anomaly" highlighted first time the level of ignorance about microbial world and affirmed the fact that larger spectrum of microbes was left unaccessed. This affirmation of Staley and Konopka did not prove convincing to microbiologists of that time. Later, in 1990, studies of DNA-DNA reassociation kinetics of soil DNA by Torsvik et al. [7] provided the compelling evidence that culturing did not capture the complete spectrum of microorganism because the majority of microbial cells that could be seen in a microscope with various staining procedures could not be induced to produce colonies on Petri plates or cultures in test tubes. During this decade of 1980s, evidence started accumulating which drew attention of the scientific community towards uncultured microbial world, and the belief that microbial world had been conquered was laid to rest.

The pioneering work of Woese [8] in 1985 explicated that the 16S rRNA gene provides evolutionary chronometer and this proposal of Woese changed the whole progression of microbiology at that time. Development of PCR technology and primer designed to amplify the complete 16S rRNA gene left a catalytic effect and 16S rRNA gene became a phylogenetic marker of choice. Owing to its universal presence in all bacteria, its multigene nature, and its large enough size (1500 bp) for informatics purpose, the 16S rRNA gene marker has been employed most extensively for characterization of naturally occurring microbiota.

The idea that 16S rRNA gene from the environmental samples can directly be cloned was first put forward by Pace et al. in 1985 [9]. Later, in 1991, Schmidt et al. [10] reported successful cloning of 16S rRNA gene sequences from marine picoplankton communities using bacteriophage lambda vector. Though the cloning of 16S rRNA gene by Schmidt et al. was a breakthrough, the hidden metabolic potential of the community members could only be achieved by functional screening of cloned genes of metagenomic origin. Later, in 1995, Healy et al. [11] recovered the cellulose and xylosidase encoding genes by functional screening of metagenomic libraries from environmental DNA isolated from the mixed liquor of thermophilic, anaerobic digesters.

In the last two decades, all sorts of natural environments, for example, soils [12–17], marine picoplankton [18–20], hot springs [21–25], surface water from rivers [26], glacier ice [27], Antarctic desert soil [28], and gut of ruminants [29], have been targeted for metagenomic analysis. Initially, most of the studies carried out on metagenomic diversity analysis targeted at various sample types were based on traditional approaches, such as denaturing gradient gel electrophoresis (DGGE) [30], terminal restriction fragment length polymorphism (T-RFLP) analysis [31], or Sanger sequencing of 16S rRNA gene clone libraries [32]. Sanger sequencing of 16S rRNA gene was dominant approach from 1990 onwards and has been used extensively to access microbial community from almost every harsher environment. Widespread sequencing of ribosomal RNA genes has resulted in the generation of large reference databases, such as the ribosomal database project (RDP) II [33], Greengenes [34], and SILVA [35]. These comprehensive databases allow classification and comparison of environmental 16S rRNA gene sequences. Traditional surveys of environmental prokaryotic communities are based on amplification and cloning of 16S rRNA genes followed by sequence analysis. In the case of some bacterial communities which are amorphous in terms of phylogenetic relationship, 16S rRNA gene based studies have found that unsuitable and functional genes have been used for detection of such functional groups of microbes [36]. As compared to 16S rRNA genes, functional genes are shown to provide a greater resolution for the study of genetic diversity in natural populations of these bacterial communities. Whole community DNA based studies have been used to reveal microbial diversity of particular functional groups of microbes in environmental samples on the basis of functional gene markers. Many functional gene markers, namely, gene *soxB* (unique gene to sulphur oxidizing bacteria) [37] and ammonia monooxygenase, *amoA* (unique to ammonia oxidizing microbes) [38], have been applied to ascertain the diversity of these functional groups of microbes in environmental samples.

3. Prospecting Metagenomes: Towards Unlocking the Concealed Microbial Potential

Unculturable microbes cannot be isolated; hence their tremendous genetic potential can only be exploited by functional metagenomic approaches. Absence of an appropriate biocatalyst has been an impeding factor for many biotransformation processes. With advancement in basic molecular biology techniques, it is now possible to put metagenomics gene sequences from uncultured microbes into expression vectors which on subsequent expression produce novel peptides inside the host cells. Presence of novel proteins can be confirmed by screening the metagenomics clones displaying desired biological activity (function-based screening). Screening of metagenomic clones often involves a simple colour reaction mediated by the enzyme/biomolecule sought (product of cloned gene), which acts on a substrate linked to chromophores leading to the development of a certain colour pattern which is detected either visually or spectrophotometrically.

In the last two decades, many novel antibiotics, drugs, and enzymes/isozymes have been recovered from metagenomic libraries constructed from various environmental samples (Table 1). Constructing metagenomic libraries from environmental samples and subsequent cloning into the expression vectors followed by activity-based screening has endless possibilities of unlocking concealed potential in uncultured microbial world. The activity-based screening of metagenomic libraries initially suffered from low sensitivity and low throughput. Development of high throughput functional screen methods, namely, SIGEX (substrate induced gene expression) [39], METREX (metabolite regulated expression) [40], and PIGEX (product induced gene expression) [41], has accelerated isolation of novel biocatalysts from the environmental samples in last eight years. These high throughput screening methods employ the resolving power of FACS (fluorescence-activated cell sorting) or fluorescence microscopy. The fluorescence-activated cell sorting (FACS) is having wide application for high throughput screening of metagenomic clones, as it can be used to identify the biological activity within a single cell [42].

Limited availability of enzyme activity assay and narrow choice of host for transformation (most often *E. coli*) have been a main constraint in functional metagenomics research. In recent years, new transformation systems have been reported which use different microbes with alternative gene expression system and wide range of protein secretion mechanisms. Development of new host systems using microbes, namely, *Streptomyces* spp. [43], *Thermus thermophilus* [44], *Sulfolobus solfataricus* [45], and Proteobacteria [46], has widened the choice of host and compatible enzyme assay systems. *E. coli*, owing to its ease of transformation and being the best genetically characterised bacterium, has been the choice host for heterologous gene expression in metagenomic studies. With synchronised advances in the HTS (high throughput screening) methods and the choice of transformation systems with wide available range of hosts for heterologous gene expression, the field of functional metagenomics got tremendous momentum. It is now possible to screen up to 50,000 clones per second or over one billion clones per day using system developed by Diversa Corp. (now the part of BASF) which integrates laser with various wavelength capabilities, enabling mass screening of metagenomic clones [47].

These advances in functional metagenomics have paved industry with an unprecedented chance to bring biomolecules of metagenomic origin into a commercial success. Diversa Corp. remained the most prominent biotech company up to 2006 for commercialisation of technologies that evolved out of metagenomic research which was later merged with Celunol Corp. to create Verenium which was further merged with BASF. BASF and other major players like DSM, Syngenta, Genencor International, and BRAIN AG collaborated with different research groups and have commercialised many biological molecules of commercial interest (for details readers are directed to read review by Cowan et al. [48]). Expressing cloned genes of metagenomic origin in heterologous host enables researchers to access the tremendous genetic potential in a microbial community without knowing anything about the original gene sequence, the structure and composition of the desired protein, or the origin of microbe. Functional screening of metagenomic libraries constructed from environmental samples has been found to express interesting moonlighting protein (proteins having two different functions within a single polypeptide chain). Jiang et al. [49] in 2011 reported a novel β-glucosidase gene (bgl1D) with lipolytic activity (thus renamed as Lip1C) which was identified through function-based screening of a metagenomic library constructed from soil. Lipase and esterase remain the most targeted enzyme activities using functional screening of metagenomic libraries of diverse origin [50–55].

4. High Throughput Sequencing and Bioinformatics Tools: Adding New Dimensions to Metagenomics

The arrival of NGS (Next Generation Sequencing) technologies has left most profound impact on the metagenomics and has expanded the scale and scope of metagenomic studies in a way never imagined before. The first NGS technology, which could be materialized due to incredible amalgam of nanotechnology, organic chemistry, optical engineering, enzyme engineering, and robotics, became a viable commercial offering in 2005. The NGS platforms have been used for standard sequencing applications, such as genome sequencing and resequencing, and also for novel applications previously unexplored by Sanger sequencing. Before arrival of NGS platforms, Venter et al. [3] in 2004 generated high magnitude metagenomics sequence data to the tone of 1.66 million reads, comprised of 1.045 billion base pairs with an average read length of 818 bp from metagenomic samples collected from Sargasso Sea. In a further extension of the same endeavour during Sorcerer II Global Ocean Sampling expedition, Rusch et al. [4] generated 7.7 billion sequencing reads, comprising

Table 1: Biological functions derived from the metagenomes from diverse habitats.

Type of activity exhibited by the metagenomic clone	Library type	Number of clones screened/size of DNA used for library construction	Sampling site	Screening method	Reference
Lipase	Plasmid and fosmid	29.3 Gb of cloned soil DNA	German forest soil (horizon A)	Phenotypic detection (tributyrin hydrolysis)	[50]
	Fosmid	200,000 clones	Qiongdongnan basin, South China Sea (water depth 778.5 m)	Phenotypic screening (tributyrin hydrolysis)	[51]
	Fosmid	15,000 clones	Peat-swamp forest soil from Narathiwat Province, Thailand	Phenotypic detection (tributyrin hydrolysis)	[52]
Esterase	Plasmid	20,000 clones	High Andean forest soil	Phenotypic detection (tributyrin hydrolysis)	[53]
	Fosmid	20000	Deep-sea sediment	Phenotypic detection (tributyrin hydrolysis)	[54]
	Fosmid	142,900	Red pepper plant rhizosphere and strawberry plant rhizosphere	Phenotypic detection (tributyrin hydrolysis)	[55]
Protease	Fosmid	17000	Surface sand from the Gobi and Death Valley deserts	Phenotypic detection (skimmed milk)	[74]
	Plasmid	70,000	Goat skin surface	Phenotypic detection (skimmed milk)	[75]
Laccase	Plasmid	8000	Mangrove soil	Phenotypic detection (hydrolysis of guaiacol)	[76]
	Phagemid	Not mentioned	Bovine rumen microflora	Phenotypic detection (oxidation of syringaldazine)	[77]
Agarase	Cosmids	1,532	Soil from uncultivated field (Germany)	Phenotypical detection (hydrolysis of low melting point agarose)	[78]
Amidase	Plasmids	193,000	Soil and enrichment cultures from marine sediment, goose pond, lakeshore, and an agricultural field (Netherlands)	Heterologous complementation	[79]
Alcohol oxidoreductase	Plasmids	900,000 and 400,000	Soil and enrichment cultures from a sugar beet field (Germany), river sediment (Germany), sediments from Solar Lake (Egypt), and sediment from the Gulf of Eilat (Israel)	Phenotypic detection (NAD(P)H-dependent reduction of carbonyls or by measuring the NAD(P)-dependent oxidation of alcohols)	[80]
Antibiotics and bioactive compounds with anti-infective properties	Fosmid	80,500 clones from Yuseong and 33,200 clones from Jindong Valley forest soil	Forest soil from Jindong Valley	Phenotypic detection	[81]
	Cosmids	Not mentioned	Bromeliad tank water (Costa Rica)	Phenotypical detection	[82]
	BAC	24,546	Soil	Phenotypic detection	[83]
DNA polymerase 1	Plasmid	21,198 Sanger sequence reads were analyzed	Octopus hot spring (93°C) in Yellowstone National Park	Activity-based screening (primer extension assay)	[84]

Table 1: Continued.

Type of activity exhibited by the metagenomic clone	Library type	Number of clones screened/size of DNA used for library construction	Sampling site	Screening method	Reference
Na^+/H^+ antiporters	Plasmid	8,000	Chaerhan Salt Lake, China	Heterologous complementation	[85]
Cellulases and xylanases	Fosmid library	Not mentioned	Hindgut of wood-feeding termite	AZCL-HE cellulose and AZCL-Xylan based assay	[86]
Phytases	Fosmid library	14,440	Soil	Functional screening (by supplying only the phytate as the sole P source in the growth medium and selecting only clones with strong growth rate)	[87]

6.3 billion base pairs using Sanger sequencing. This large amount of sequence data using Sanger sequencing was a great endeavour but the magnitude of data which are produced in a single run of NGS machine is severalfold higher. The large scale sequencing projects and consortia have already produced NGS derived huge sequence data sets, namely, The ENCODE project (over 15 trillion bases of raw data) [56], 1000 Genomes (over 20,000 Gb bases of raw data with about 5x coverage) [57], Human Microbiome Project (over 5 terabytes of genomic data) [58], and Earth Microbiome Project (envisage to produce over two petabytes of sequence data) [59]. The NGS platforms have paved the way to directly sequence the metagenomic DNA circumventing the need for tedious steps of cloning and library preparation. NGS platforms allow massive parallel sequencing where hundreds of thousands to hundreds of millions of sequencing reactions are performed and detected simultaneously, resulting in very high throughput. As multiple NGS platforms coexist in the market place with the unique chemistry of each, the decision about the suitability of a particular type of NGS platform for a metagenomic project is most critical in deciding the outcome of metagenomic studies. Hence, the selection of a particular NGS platform has to be made on the basis of varying features of NGS platforms like read length, degree of automation, throughput per run, data quality, ease in data analysis, and cost per run as compared in Table 2 (for details readers are directed to read the review by Liu et al., 2012 [60]).

454/Roche Life Sciences (pyrosequencing technology) and the Illumina/Solexa system are two most extensively applied sequencing platforms for metagenomic studies carried out in the last eight years followed by ABI SOLiD. The longer read length resulting due to Roche chemistry allows unambiguous mapping of reads to complex targets, giving Roche 454 platform an upper edge over other competitors. The another major player Illumina's (earlier Solexa) offerings, HiSeq 1500/2500, HiSeq 2000/1000, and Genome Analyzer IIX are widely used NGS platforms for metagenomic research. One of the latest additions of Illumina, that is, HiSeq 1500/2500, offers two run modes (rapid run and high throughput run mode). This high throughput run mode is perfect for larger studies with more samples and hence is best suited for metagenomics investigations. It requires only 1 ng of community DNA to get complete metagenomic sequence data using reversible terminator chemistry of Illumina for their HiSeq 2500 which is able to generate 270–300 GB of sequence data with read length of up to 200 bp and very high coverage in a short period of less than 5 days. Illumina's recently launched NGS platform HiSeq X Ten has more than 1.5 Tb data output with more than 3 billion reads (above 150 bp size) per flow cell. After Roche 454 and Illumina's NGS platforms, the polony sequencing based ABI (now Life Technologies) SOLiD platforms with highest accuracy (99.99%) are frequently applied in metagenomic research. These NGS platforms are amenable for deep sequencing which makes it possible to detect very low abundant members of complex populations in metagenomic samples. The actual read length and depth required will depend on the desired sensitivity and complexity of the population. NGS technologies have led the way for shotgun metagenomics to reconstruct whole bacterial and archaeal genomes without presence of a reference genome (or their genome sequence) by using powerful assembly algorithms that join short overlapping DNA fragments generated by the NGS sequencers. As each NGS platform differs substantially in read length, coverage, and accuracy, whether these platforms recover the same diversity from a sample remains a fundamental question. Luo et al. [61] carried out direct comparison of the two most widely used NGS platforms, that is, Roche 454 FLX Titanium and Illumina Genome Analyzer (GA) II, on the same DNA samples obtained from Lake Lanier, Atlanta. They inferred ~90% assembly overlap of total sequences and high correlation ($R^2 > 0.9$) for the *in situ* abundance of genes and genotypes between two platforms and sequence assemblies produced by Illumina were of equivalent quality to Roche 454 as evaluated on the basis of base call error, frame shift frequency, and contig length. Ion Torrent (and more recently Ion Proton), Pacific Biosciences (PacBio) SMRT sequencing, and Complete Genomics offering DNA nanoball sequencing are few other emerging sequencing technologies, but none of these emerging sequencing technologies have been thoroughly applied and tested with metagenomic samples. NGS platforms are amenable to multiplexing where hundreds

Table 2: Comparison of the unique features of NGS platforms widely applied in metagenomic research.

Sequencer	Roche/454 GS FLX Titanium	HiSeq 2000	SOLiDv4
NGS chemistry	Pyrosequencing	Sequencing by synthesis	Sequencing by ligation and exact call chemistry
Library/template preparation	Emulsion PCR (emPCR)	Solid phase amplification	Emulsion PCR for fragment/mate-pair end sequencing
Average read length	250–310 bp (highest among the NGS platforms) Now approaching 400–500 (titanium) pyroreads	Initially it was 36, now approaching 150	35
Run time (days)	24 hours (fastest of all)	4 days (fragment run) 9 days (mate-pair run)	7 days (fragment run) 14 days (mate-pair run)
Output data/run	0.7 Gb	600 Gb (over 1 Tb with Illumina's HiSeq X Ten)	120 Gb
Advantage	Longer reads Least time for one run Amenable to multiplexing allowing many samples in single run	High throughput Most widely used platform	Highest accuracy due to ECC (exact call chemistry)
Limitations	High error rate in homopolymer region High cost of reagents Low in throughput Artificial replicate sequences during ePCR [88]	Short read length Low multiplexing capability of samples Single base error with GGC motifs High error rate at tail end reads [89]	Long run time Short read length

to thousands of samples can be sequenced in parallel by adding 9–12 bp DNA tag to each DNA fragment prior to sequencing. Later, this tag is used to identify the origin of the fragment from pooled samples permitting the simultaneous exploration of thousands of bacterial communities in a highly cost-effective manner [62].

The sequence reads generated in NGS based sequencing are typically shorter (except for Pacific Biosciences) than traditional Sanger sequencing reads and have origin from genome of different organisms, which makes the assembly and analysis of metagenomic NGS sequence data extremely challenging. Apart from the problem of assembly of short DNA sequence reads, terabyte-sized data files are generated with each run of instrument, which greatly increases the computer resource requirements of the sequencing laboratories. In a typical sequencing based metagenomic project, post-sequencing steps such as metagenomic sequence assembly, functional annotation, binning of sequences, variant analysis, gene/ORF prediction, community taxonomic profile, and metabolic reconstruction are the most critical steps which decide the outcome of any investigation. The majority of current assembly programs are designed to assemble the sequences coming from single genome and hence not equally effective for a typical metagenomic sequence data set having sequences of different origin. Absence of any reference genome for assembly of genome sequences from unculturable representatives of metagenomic sequence pool makes the task more challenging.

Although several bioinformatics tools for sequence assembly of sequences of metagenomic origin have been developed in past few years, which have simplified the task to some extent, still postsequencing analysis is most challenging. Constant efforts are underway to improve the accuracy of alignment of NGS data in several laboratories all across the globe. Development of sequence assemblers like MetaVelvet [63] and Meta-IDBA [64] which are specifically designed for *de novo* assembly of metagenomic sequence reads and metagenomic analysis and data storage pipelines such as MG-RAST [65], MetAMOS [66], MEGAN, IMG/M [67], CAMERA [68], and GALAXY web server [69] has enabled the researchers with limited expertise in bioinformatics to undertake elaborative projects in metagenomics. A brief account of these bioinformatics tools commonly used for postsequencing analysis of metagenomic data is described in Table 3, in order to provide instant information for researchers having limited expertise in bioinformatics.

Longer read length results in better assembled contigs, which further results in quality scaffolds. Sequencing errors remain major issue and extent of sequencing error is different for different sequencing platforms as mismatches are reported more frequently on Illumina platform, and homopolymer issues resulting in insertion/deletions are often reported with Roche 454 platform. Intrinsic sequencing coverage bias of different platforms can complicate subsequent analysis. There exists no gold standard for metagenomic data analysis and inadvertent errors have to be taken care of at each core step of metagenomic investigation.

Table 3: A brief description of bioinformatic tools commonly employed for postsequencing analysis of metagenomic sequence data.

Postsequencing task	Bioinformatic tool	Brief description	URL	Reference
Metagenomic assembly tool	MetaVelvet	Decomposes a de Bruijn graph into individual subgraphs on the basis of coverage (abundance) difference and graph connectivity. Overcomes the limitation of a single-genome assembler to misidentify sequences from highly abundant species as repeats. Results in higher N50 scores than any single-genome assembler.	http://metavelvet.dna.bio.keio.ac.jp/	[63]
	Meta-IDBA	Implies partitioning the de Bruijn graph into isolated components of different species by grouping similar regions of similar subspecies and partitioning the graph into components based on the topological structure of the graph.	http://i.cs.hku.hk/~alse/hkubrg/projects/metaidba/	[64]
	Genovo	Uses Bayesian approach and generative probabilistic model of read generation which works by discovering likely sequence reconstructions under the model. Algorithm used is iterated conditional modes (ICM) algorithm, which maximizes local conditional probabilities sequentially.	http://cs.stanford.edu/group/genovo/	[90]
	Bambus 2	Uses mate-pair information during the assembly process which is not used by Meta-IDBA, MetaVelvet, and Genovo. Algorithms operate on a contig graph generation followed by orientation, positioning, and simplification for proper scaffolding.	http://amos.sf.net.	[91]
Short read alignment and mapping to reference genome	Bowtie	An ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences which employs Burrows-Wheeler index based on the full-text minute-space (FM) index having low memory footprint (1.3 GB only) also supports gapped, local, and paired-end alignment modes.	http://bowtie-bio.sourceforge.net/index.shtml	[92]
	BWA	Employed for mapping low-divergent sequences against a large reference genome. Has three-algorithm mode for different read length. For Illumina sequence reads up to 100 bp size algorithm BWA-backtrack is used, while algorithms, BWA-SW and BWA-MEM, meant for longer sequences ranged from 70 bp to 1 Mbp.	http://bio-bwa.sourceforge.net/	[93]
	SOAP 3	Fast, accurate, and sensitive GPU-based short read aligner which delivers high speed and sensitivity simultaneously. Found to take less than 30 seconds to align one million read pairs onto the human reference genome, much faster than BWA and Bowtie.	http://www.cs.hku.hk/2bwt-tools/soap3-dp/	[94]
	mrsFAST	A cache oblivious mapper that is designed to map short reads to reference genome. mrsFAST maps short reads with respect to user defined error threshold.	http://sfu-compbio.github.io/mrsfast/	[95]

Table 3: Continued.

Postsequencing task	Bioinformatic tool	Brief description	URL	Reference
Microbial diversity analysis	MLST	Exploits unambiguous nature and electronic portability of nucleotide sequence data for the characterization of microorganisms.	http://www.mlst.net/	[96]
	Axiome	Streamlines and manages analysis of small subunit (SSU) rRNA marker data in QIIME and mothur. Has a companion graphical user interface (GUI) and is designed to be easily extended to facilitate customized research workflows.	http://neufeld.github.com/axiometic	[97]
	PHACCS	Uses the contig spectrum from shotgun DNA based on modified Lander-Waterman algorithm sequence assemblies to predict structure of viral communities and make predictions about diversity.	http://phaccs.sourceforge.net/	[98]
Functional annotation	RAMMCAP	An ultrafast method that can cluster and annotate one million metagenomic reads in only hundreds of CPU hours.	http://weizhong-lab.ucsd.edu/rammcap/cgi-bin/rammcap.cgi	[99]
Gene annotation/gene calling	FragGeneScan	Combines sequencing error models and codon usages in a hidden Markov model to improve the prediction of protein-coding region in short reads.	http://omics.informatics.indiana.edu/FragGeneScan/	[100]
	MetaGeneMark	An ab initio gene prediction tool with updated heuristic models designed for metagenomic sequences.	http://exon.gatech.edu/meta_gmhmmp.cgi	[101]
	MetaGeneAnnotator	Precisely predicts all kinds of prokaryotic genes from a single or a set of anonymous genomic sequences having a variety of lengths. Integrates statistical models of prophage genes in addition to those of bacterial and archaeal genes and also uses a self-training model from input sequences for predictions.	http://metagene.cb.k.u-tokyo.ac.jp/	[102]
Binning	TETRA	Based on statistical analysis of tetranucleotide usage patterns in genomic fragments which automate the task of comparative tetranucleotide frequency analysis and outperform (G+C) content based analysis.	http://www.megx.net/tetra/index.html	[103]
	MetaCluster 5.0	A two-round binning method that separates reads of high-abundance species from those of low-abundance species in two different rounds and aims at identifying both low-abundance and high-abundance species in the presence of a large amount of noise due to many extremely low-abundance species. Uses a filtering strategy to remove noise from the extremely low-abundance species.	http://i.cs.hku.hk/~alse/MetaCluster/	[104]
	Phymm	Uses interpolated Markov models (IMMs) to characterize variable-length oligonucleotides typical of a phylogenetic grouping.	http://www.cbcb.umd.edu/software/phymm/	[105]

Table 3: Continued.

Postsequencing task	Bioinformatic tool	Brief description	URL	Reference
Automated platforms/servers for comparative and functional analysis of metagenomic sequence data	MG-RAST	MG-RAST (the Metagenomics RAST) server is an automated analysis platform which provides upload, quality control, automated annotation, and analysis for prokaryotic metagenomic shotgun samples.	http://metagenomics.anl.gov	[65]
	MetAMOS	An open source and modular metagenomic assembly and analysis pipeline leveraging over 20 existing tools with some new tools integrated as well. Entire pipeline is built around the unique features provided by the metagenomic scaffolder Bambus 2.	https://github.com/treangen/MetAMOS	[66]
	MEGAN 4	Released in 2011 for taxonomic analysis, comparative analysis, and functional analysis methods based on the SEED and KEGG (Kyoto Encyclopedia for Genes and Genomes)	http://www-ab.informatik.uni-tuebingen.de/software/megan	[106]
	IMG/M	A data management and analysis system for microbial community genomes (metagenomes) hosted at the Department of Energy's (DOE) Joint Genome Institute (JGI). IMG/M consists of metagenome data integrated with isolate microbial genomes from the Integrated Microbial Genomes (IMG) system.	http://img.jgi.doe.gov/cgi-bin/m/main.cgi	[67]
	CAMERA	Provides access to raw environmental sequence data, with associated metadata, precomputed annotation, and analyses. Integrates tools for gene prediction and annotation, clustering, assembly sequence quality control, functional and comparative genomics applications, and many other downstream analysis tools.	http://camera.calit2.net	[68]
	GALAXY	A publicly available web service, with software system that provides support for analysis of genomic, comparative genomic, and functional genomic data through a framework that gives experimentalists simple interfaces to powerful tools while automatically managing the computational details.	http://galaxyproject.org	[69]

Currently, there exist simulation systems (GemSIM [70], MetaSim [71], and Grinder [72]) for NGS sequencing data and they can be applied for metagenomic simulation. MetaSim and Grinder use fixed probabilities of sequencing errors (insertions, deletions, and substitutions) for the same base in different reads, but sequencing coverage biases are not considered by any of these simulators. Jia et al., 2013 [73], have developed Next Generation Sequencing Simulator for Metagenomics (NeSSM) which not only deals with sequencing errors but also deals with sequencing coverage biases effectively. The development of new algorithms for extracting useful information out of metagenomic sequence data is so rapid that new updates and developments are reported every couple of weeks and any comprehensive review of this aspect may appear incomplete due to the continuous upgrade and addition of new algorithms.

5. Conclusion and Future Perspectives

Information from metagenomic libraries has the ability to enrich the knowledge and applications of many aspects of the industry, therapeutics, and environmental sustainability. The last two decades witnessed tremendous progress in function driven screening of metagenomic libraries constructed using community DNA from various, moderate to harsh environments resulting in the discovery of many novel enzymes, bioactive compounds, and antibiotics through heterologous gene expression. Availability of methods to extract DNA from almost any kind of environmental samples, rapidly dropping cost of sequencing, continuously evolving NGS platforms, and readily available computing and analytical power of automated metagenomic servers have brought the science of metagenomics to extremely exciting phase. The perfect stage has been set for executing and implementing the accumulated insights about untapped microbial communities to exploit their concealed potential. Metagenomic data sets are increasingly becoming more complex and comprehensive and *in silico* gene prediction on metagenomic sequence data sets is rocketing. After 2005, enormous information about novel genes/ORFs/operons from diverse environments has accumulated. Now, there is strong need to focus more on validating these novel genes/ORFs of metagenomic origin by putting them in action in real wet lab conditions to search for more novel enzymes and bioactivities for bioprospecting metagenomes; else, we may end up putting all efforts for novel genes/ORFs/operons in dry lab conditions only. Systems biology approach combined with Next Generation Sequencing technologies and bioinformatics is inevitable for achieving these objectives.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] V. Kunin, A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz, "A bioinformatician's guide to metagenomics," *Microbiology and Molecular Biology Reviews*, vol. 72, no. 4, pp. 557–578, 2008.

[2] E. A. Dinsdale, R. A. Edwards, D. Hall et al., "Functional metagenomic profiling of nine biomes," *Nature*, vol. 452, no. 7187, pp. 629–632, 2008.

[3] J. C. Venter, K. Remington, J. F. Heidelberg et al., "Environmental genome shotgun sequencing of the Sargasso Sea," *Science*, vol. 304, no. 5667, pp. 66–74, 2004.

[4] D. B. Rusch, A. L. Halpern, G. Sutton et al., "The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific," *PLoS Biology*, vol. 5, no. 3, article e77, 2007.

[5] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman, "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products," *Chemistry and Biology*, vol. 5, no. 10, pp. R245–R249, 1998.

[6] J. T. Staley and A. Konopka, "Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats," *Annual Review of Microbiology*, vol. 39, pp. 321–346, 1985.

[7] V. Torsvik, J. Goksoyr, and F. L. Daae, "High diversity in DNA of soil bacteria," *Applied and Environmental Microbiology*, vol. 56, no. 3, pp. 782–787, 1990.

[8] C. R. Woese, "Bacterial evolution," *Microbiological Reviews*, vol. 51, no. 2, pp. 221–271, 1987.

[9] N. R. Pace, D. A. Stahl, D. J. Lane, and G. J. Olsen, "Analyzing natural microbial populations by rRNA sequences," *ASM News*, vol. 51, pp. 4–12, 1985.

[10] T. M. Schmidt, E. F. DeLong, and N. R. Pace, "Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing," *Journal of Bacteriology*, vol. 173, no. 14, pp. 4371–4378, 1991.

[11] F. G. Healy, R. M. Ray, H. C. Aldrich, A. C. Wilkie, L. O. Ingram, and K. T. Shanmugam, "Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose," *Applied Microbiology and Biotechnology*, vol. 43, no. 4, pp. 667–674, 1995.

[12] J. D. Coolon, K. L. Jones, T. C. Todd, J. M. Blair, and M. A. Herman, "Long term nitrogen amendment alters the diversity and assemblage of soil bacterial communities in tall grass prairie," *PLoS ONE*, vol. 8, no. 6, Article ID e67884, 2013.

[13] N. Rosenzweig, J. M. Bradeen, Z. J. Tu, S. J. McKay, and L. L. Kinkel, "Rhizosphere bacterial communities associated with long-lived perennial prairie plants vary in diversity composition, and structure," *Canadian Journal of Microbiology*, vol. 59, no. 7, pp. 494–502, 2013.

[14] J. Han, J. Jung, M. Park, S. Hyun, and W. Park, "Short-term effect of elevated temperature on the abundance and diversity of bacterial and archaeal *amoA* genes in antarctic soils," *Journal of Microbiology and Biotechnology*, vol. 23, no. 9, pp. 1187–1196, 2013.

[15] G. P. Athak, A. Hrenreich, A. Osi, W. R. Sreit, and W. Gärtner, "Novel blue light-sensitive proteins from a metagenomic approach," *Environmental Microbiology*, vol. 11, no. 9, pp. 2388–2399, 2009.

[16] S. Voget, H. L. Steele, and W. R. Streit, "Characterization of a metagenome-derived halotolerant cellulase," *Journal of Biotechnology*, vol. 126, no. 1, pp. 26–36, 2006.

[17] T. Waschkowitz, S. Rockstroh, and R. Daniel, "Isolation and characterization of metalloproteases with a novel domain structure by construction and screening of metagenomic libraries,"

Applied and Environmental Microbiology, vol. 75, no. 8, pp. 2506–2516, 2009.

[18] Z. G. Keresztes, T. Felföldi, B. Somogyi et al., "First record of picophytoplankton diversity in Central European hypersaline lakes," *Extremophiles*, vol. 16, no. 5, pp. 759–769, 2012.

[19] G. Zeidner and O. Béjà, "The use of DGGE analyses to explore eastern Mediterranean and Red Sea marine picophytoplankton assemblages," *Environmental Microbiology*, vol. 6, no. 5, pp. 528–534, 2004.

[20] J. L. Stein, T. L. Marsh, K. Y. Wu, H. Shizuya, and E. F. Delong, "Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon," *Journal of Bacteriology*, vol. 178, no. 3, pp. 591–599, 1996.

[21] B. P. Hedlund, J. A. Dodsworth, J. K. Cole, and H. H. Panosyan, "An integrated study reveals diverse methanogens, Thaumarchaeota, and yet-uncultivated archaeal lineages in Armenian hot springs," *Antonie van Leeuwenhoek*, vol. 104, no. 1, pp. 71–82, 2013.

[22] Q. Huang, H. Jiang, B. R. Briggs et al., "Archaeal and bacterial diversity in acidic to circumneutral hot springs in the Philippines," *FEMS Microbiology Ecology*, vol. 85, no. 3, pp. 452–464, 2013.

[23] W. Hou, S. Wang, H. Dong et al., "A Comprehensive census of microbial diversity in hot springs of Tengchong, Yunnan Province China using 16S rRNA gene pyrosequencing," *PLoS ONE*, vol. 8, no. 1, Article ID e53350, 2013.

[24] J. B. Sylvan, B. M. Toner, and K. J. Edwards, "Life and death of deep-sea vents: bacterial diversity and ecosystem succession on inactive hydrothermal sulfides," *mBio*, vol. 3, no. 1, pp. e00279–e00211, 2012.

[25] J.-K. Rhee, D.-G. Ahn, Y.-G. Kim, and J.-W. Oh, "New thermophilic and thermostable esterase with sequence similarity to the hormone-sensitive lipase family, cloned from a metagenomic library," *Applied and Environmental Microbiology*, vol. 71, no. 2, pp. 817–825, 2005.

[26] C. Wu and B. Sun, "Identification of novel esterase from metagenomic library of Yangtze River," *Journal of Microbiology and Biotechnology*, vol. 19, no. 2, pp. 187–193, 2009.

[27] C. Simon, J. Herath, S. Rockstroh, and R. Daniel, "Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice," *Applied and Environmental Microbiology*, vol. 75, no. 9, pp. 2964–2968, 2009.

[28] C. Heath, X. P. Hu, S. C. Cary, and D. Cowan, "Identification of a novel alkaliphilic esterase active at low temperatures by screening a metagenomic library from antarctic desert soil," *Applied and Environmental Microbiology*, vol. 75, no. 13, pp. 4657–4659, 2009.

[29] X. Gong, R. J. Gruninger, M. Qi et al., "Cloning and identification of novel hydrolase genes from a dairy cow rumen metagenomic library and characterization of a cellulase gene," *BMC Research Notes*, vol. 5, article 566, 2012.

[30] G. Muyzer, E. C. de Waal, and A. G. Uitterlinden, "Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA," *Applied and Environmental Microbiology*, vol. 59, no. 3, pp. 695–700, 1993.

[31] N. Fierer and R. B. Jackson, "The diversity and biogeography of soil bacterial communities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 3, pp. 626–631, 2006.

[32] M. L. Sogin, H. G. Morrison, J. A. Huber et al., "Microbial diversity in the deep sea and the underexplored 'rare biosphere'," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 32, pp. 12115–12120, 2006.

[33] J. R. Cole, B. Chai, T. L. Marsh et al., "The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy," *Nucleic Acids Research*, vol. 31, no. 1, pp. 442–443, 2003.

[34] T. Z. DeSantis, P. Hugenholtz, N. Larsen et al., "Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB," *Applied and Environmental Microbiology*, vol. 72, no. 7, pp. 5069–5072, 2006.

[35] W. Ludwig, O. Strunk, R. Westram et al., "ARB: a software environment for sequence data," *Nucleic Acids Research*, vol. 32, no. 4, pp. 1363–1371, 2004.

[36] G. Braker, A. Fesefeldt, and K.-P. Witzel, "Development of PCR primer systems for amplification of nitrite reductase genes (*nir*K and *nir*S) to detect denitrifying bacteria in environmental samples," *Applied and Environmental Microbiology*, vol. 64, no. 10, pp. 3769–3775, 1998.

[37] K. K. Krishnani, V. Kathiravan, M. Natarajan, M. Kailasam, and S. M. Pillai, "Diversity of sulfur-oxidizing bacteria in greenwater system of coastal aquaculture," *Applied Biochemistry and Biotechnology*, vol. 162, no. 5, pp. 1225–1237, 2010.

[38] K. K. Krishnani, M. S. Shekhar, G. Gopikrishna, and B. P. Gupta, "Molecular biological characterization and biostimulation of ammonia-oxidizing bacteria in brackishwater aquaculture," *Journal of Environmental Science and Health Part A*, vol. 44, no. 14, pp. 1598–1608, 2009.

[39] T. Uchiyama, T. Abe, T. Ikemura, and K. Watanabe, "Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes," *Nature Biotechnology*, vol. 23, no. 1, pp. 88–93, 2005.

[40] L. L. Williamson, B. R. Borlee, P. D. Schloss, C. Guan, H. K. Allen, and J. Handelsman, "Intracellular screen to identify metagenomic clones that induce or inhibit a quorum-sensing biosensor," *Applied and Environmental Microbiology*, vol. 71, no. 10, pp. 6335–6344, 2005.

[41] T. Uchiyama and K. Miyazaki, "Product-induced gene expression, a product-responsive reporter assay used to screen metagenomic libraries for enzyme-encoding genes," *Applied and Environmental Microbiology*, vol. 76, no. 21, pp. 7029–7035, 2010.

[42] M. Podar, C. B. Abulencia, M. Walcher et al., "Targeted access to the genomes of low-abundance organisms in complex microbial communities," *Applied and Environmental Microbiology*, vol. 73, no. 10, pp. 3205–3214, 2007.

[43] G.-Y. Wang, E. Graziani, B. Waters et al., "Novel natural products from soil DNA libraries in a streptomycete host," *Organic Letters*, vol. 2, no. 16, pp. 2401–2404, 2000.

[44] A. Angelov, M. Mientus, S. Liebl, and W. Liebl, "A two-host fosmid system for functional screening of (meta)genomic libraries from extreme thermophiles," *Systematic and Applied Microbiology*, vol. 32, no. 3, pp. 177–185, 2009.

[45] S.-V. Albers, M. Jonuscheit, S. Dinkelaker et al., "Production of recombinant and tagged proteins in the hyperthermophilic archaeon *Sulfolobus solfataricus*," *Applied and Environmental Microbiology*, vol. 72, no. 1, pp. 102–111, 2006.

[46] J. W. Craig, F.-Y. Chang, J. H. Kim, S. C. Obiajulu, and S. F. Brady, "Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria," *Applied and Environmental Microbiology*, vol. 76, no. 5, pp. 1633–1641, 2010.

[47] S. C. Wenzel and R. Müller, "Recent developments towards the heterologous expression of complex bacterial natural product biosynthetic pathways," *Current Opinion in Biotechnology*, vol. 16, no. 6, pp. 594–606, 2005.

[48] D. Cowan, Q. Meyer, W. Stafford, S. Muyanga, R. Cameron, and P. Wittwer, "Metagenomic gene discovery: past, present and future," *Trends in Biotechnology*, vol. 23, no. 6, pp. 321–329, 2005.

[49] C.-J. Jiang, G. Chen, J. Huang et al., "A novel β-glucosidase with lipolytic activity from a soil metagenome," *Folia Microbiologica*, vol. 56, no. 6, pp. 563–570, 2011.

[50] H. Nacke, C. Will, S. Herzog, B. Nowka, M. Engelhaupt, and R. Daniel, "Identification of novel lipolytic genes and gene families by screening of metagenomic libraries derived from soil samples of the German Biodiversity Exploratories," *FEMS Microbiology Ecology*, vol. 78, no. 1, pp. 188–201, 2011.

[51] Y. Hu, C. Fu, Y. Huang et al., "Novel lipolytic genes from the microbial metagenomic library of the South China Sea marine sediment," *FEMS Microbiology Ecology*, vol. 72, no. 2, pp. 228–237, 2010.

[52] B. Bunterngsook, P. Kanokratana, T. Thongaram et al., "Identification and characterization of lipolytic enzymes from a peat-swamp forest soil metagenome," *Bioscience, Biotechnology and Biochemistry*, vol. 74, no. 9, pp. 1848–1854, 2010.

[53] D. J. Jiménez, J. S. Montaña, D. Álvarez, and S. Baena, "A novel cold active esterase derived from Colombian high Andean forest soil metagenome," *World Journal of Microbiology and Biotechnology*, vol. 28, no. 1, pp. 361–370, 2012.

[54] X. Jiang, X. Xu, Y. Huo et al., "Identification and characterization of novel esterases from a deep-sea sediment metagenome," *Archives of Microbiology*, vol. 194, no. 3, pp. 207–214, 2012.

[55] M. H. Lee, K. S. Hong, S. Malhotra et al., "A new esterase EstD2 isolated from plant rhizosphere soil metagenome," *Applied Microbiology and Biotechnology*, vol. 88, no. 5, pp. 1125–1134, 2010.

[56] K. R. Rosenbloom, T. R. Dreszer, J. C. Long et al., "ENCODE whole-genome data in the UCSC genome browser: update 2012," *Nucleic Acids Research*, vol. 40, no. 1, pp. D912–D917, 2012.

[57] T. Lappalainen, M. Sammeth, M. R. Friedländer et al., "Transcriptome and genome sequencing uncovers functional variation in humans," *Nature*, vol. 501, no. 7468, pp. 506–511, 2013.

[58] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "The human microbiome project," *Nature*, vol. 449, no. 7164, pp. 804–810, 2007.

[59] J. A. Gilbert, F. Meyer, D. Antonopoulos et al., "Meeting report: the terabase metagenomics workshop and the vision of an Earth Microbiome Project," *Standards in Genomic Sciences*, vol. 3, no. 3, pp. 243–248, 2010.

[60] L. Liu, Y. Li, S. Li et al., "Comparison of next-generation sequencing systems," *Journal of Biomedicine and Biotechnology*, vol. 2012, Article ID 251364, 11 pages, 2012.

[61] C. Luo, D. Tsementzi, N. Kyrpides, T. Read, and K. T. Konstantinidis, "Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample," *PLoS ONE*, vol. 7, no. 2, Article ID e30087, 2012.

[62] J. G. Caporaso, C. L. Lauber, W. A. Walters et al., "Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms," *ISME Journal*, vol. 6, no. 8, pp. 1621–1624, 2012.

[63] T. Namiki, T. Hachiya, H. Tanaka, and Y. Sakakibara, "MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads," *Nucleic Acids Research*, vol. 40, no. 20, article e155, 2012.

[64] Y. Peng, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "Meta-IDBA: a de Novo assembler for metagenomic data," *Bioinformatics*, vol. 27, no. 13, pp. i94–i101, 2011.

[65] F. Meyer, D. Paarmann, M. D'Souza et al., "The metagenomics RAST server: a public resource for the automatic phylogenetic and functional analysis of metagenomes," *BMC Bioinformatics*, vol. 9, article 386, 2008.

[66] T. J. Treangen, S. Koren, D. D. Sommer et al., "MetAMOS: a modular and open source metagenomic assembly and analysis pipeline," *Genome Biology*, vol. 14, article R2, 2013.

[67] V. M. Markowitz, N. N. Ivanova, E. Szeto et al., "IMG/M: a data management and analysis system for metagenomes," *Nucleic Acids Research*, vol. 36, no. 1, pp. D534–D538, 2008.

[68] S. Sun, J. Chen, W. Li et al., "Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource," *Nucleic Acids Research*, vol. 39, no. 1, pp. D546–D551, 2011.

[69] D. Blankenberg, G. Von Kuster, N. Coraor et al., "Galaxy: a web-based genome analysis tool for experimentalists," in *Current Protocols in Molecular Biology*, unit 19.10, pp. 1–21, John Wiley & Sons, 2010.

[70] K. E. McElroy, F. Luciani, and T. Thomas, "GemSIM: general, error-model based simulator of next-generation sequencing data," *BMC Genomics*, vol. 13, article 74, 2012.

[71] D. C. Richter, F. Ott, A. F. Auch, R. Schmid, and D. H. Huson, "MetaSim: a sequencing simulator for genomics and metagenomics," *PLoS ONE*, vol. 3, no. 10, Article ID e3373, 2008.

[72] F. E. Angly, D. Willner, F. Rohwer, P. Hugenholtz, and G. W. Tyson, "Grinder: a versatile amplicon and shotgun sequence simulator," *Nucleic Acids Research*, vol. 40, no. 12, article e94, 2012.

[73] B. Jia, L. Xuan, K. Cai, Z. Hu, L. Ma, and C. Wei, "NeSSM: a next-generation sequencing simulator for metagenomics," *PLoS ONE*, vol. 8, no. 10, Article ID e75448, 2013.

[74] J. Neveu, C. Regeard, and M. S. DuBow, "Isolation and characterization of two serine proteases from metagenomic libraries of the Gobi and Death Valley deserts," *Applied Microbiology and Biotechnology*, vol. 91, no. 3, pp. 635–644, 2011.

[75] P. L. Pushpam, T. Rajesh, and P. Gunasekaran, "Identification and characterization of alkaline serine protease from goat skin surface metagenome," *AMB Express*, vol. 1, article 3, 2011.

[76] M. Ye, G. Li, W. Q. Liang, and Y. H. Liu, "Molecular cloning and characterization of a novel metagenome-derived multicopper oxidase with alkaline laccase activity and highly soluble expression," *Applied Microbiology and Biotechnology*, vol. 87, no. 3, pp. 1023–1031, 2010.

[77] A. Beloqui, M. Pita, J. Polaina et al., "Novel polyphenol oxidase mined from a metagenome expression library of bovine rumen: biochemical properties, structural analysis, and phylogenetic relationships," *The Journal of Biological Chemistry*, vol. 281, no. 32, pp. 22933–22942, 2006.

[78] S. Voget, C. Leggewie, A. Uesbeck, C. Raasch, K.-E. Jaeger, and W. R. Streit, "Prospecting for novel biocatalysts in a soil metagenome," *Applied and Environmental Microbiology*, vol. 69, no. 10, pp. 6235–6242, 2003.

[79] E. M. Gabor, E. J. de Vries, and D. B. Janssen, "Construction, characterization, and use of small-insert gene banks of DNA isolated from soil and enrichment cultures for the recovery of novel amidases," *Environmental Microbiology*, vol. 6, no. 9, pp. 948–958, 2004.

[80] A. Knietsch, T. Waschkowitz, S. Bowien, A. Henne, and R. Daniel, "Construction and screening of metagenomic libraries derived from enrichment cultures: generation of a gene bank for genes conferring alcohol oxidoreductase activity on *Escherichia coli*," *Applied and Environmental Microbiology*, vol. 69, no. 3, pp. 1408–1416, 2003.

[81] H. K. Lim, E. J. Chung, J.-C. Kim et al., "Characterization of a forest soil metagenome clone that confers indirubin and indigo production on *Escherichia coli*," *Applied and Environmental Microbiology*, vol. 71, no. 12, pp. 7768–7777, 2005.

[82] S. F. Brady and J. Clardy, "Palmitoylputrescine, an antibiotic isolated from the heterologous expression of DNA extracted from bromeliad tank water," *Journal of Natural Products*, vol. 67, no. 8, pp. 1283–1286, 2004.

[83] D. E. Gillespie, S. F. Brady, A. D. Bettermann et al., "Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA," *Applied and Environmental Microbiology*, vol. 68, no. 9, pp. 4301–4306, 2002.

[84] M. J. Moser, R. A. DiFrancesco, K. Gowda et al., "Thermostable DNA polymerase from a viral metagenome is a potent RT-PCR enzyme," *PLoS ONE*, vol. 7, no. 6, Article ID e38371, 2012.

[85] X. Wang, F. Xu, and S. Chen, "Metagenomic cloning and characterization of Na^+/H^+ antiporter genes taken from sediments in Chaerhan Salt Lake in China," *Biotechnology Letters*, vol. 35, no. 4, pp. 619–624, 2013.

[86] T. Nimchua, T. Thongaram, T. Uengwetwanit, S. Pongpattanakitshote, and L. Eurwilaichitr, "Metagenomic analysis of novel lignocellulose-degrading enzymes from higher termite guts inhabiting microbes," *Journal of Microbiology and Biotechnology*, vol. 22, no. 4, pp. 462–469, 2012.

[87] H. Tan, M. J. Mooij, M. Barret et al., "Identification of novel phytase genes from an agricultural soil-derived metagenome," *Journal of Microbiology and Biotechnology*, vol. 24, no. 1, pp. 113–118, 2014.

[88] T. K. Teal and T. M. Schmidt, "Identifying and removing artificial replicates from 454 pyrosequencing data," *Cold Spring Harbor Protocols*, vol. 5, no. 4, 2010.

[89] K. Nakamura, T. Oshima, T. Morimoto et al., "Sequence-specific error profile of Illumina sequencers," *Nucleic Acids Research*, vol. 39, no. 13, article e90, 2011.

[90] J. Laserson, V. Jojic, and D. Koller, "Genovo: de novo assembly for metagenomes," *Journal of Computational Biology*, vol. 18, no. 3, pp. 429–443, 2011.

[91] S. Koren, T. J. Treangen, and M. Pop, "Bambus 2: scaffolding metagenomes," *Bioinformatics*, vol. 27, no. 21, pp. 2964–2971, 2011.

[92] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome," *Genome Biology*, vol. 10, article R25, 2009.

[93] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 26, no. 5, pp. 589–595, 2010.

[94] R. Luo, T. Wong, J. Zhu et al., "Correction: SOAP3-dp: fast, accurate and sensitive GPU-based short read aligner," *PLoS ONE*, vol. 8, no. 8, Article ID e65632, 2013.

[95] F. Hach, F. Hormozdiari, C. Alkan et al., "MrsFAST: a cache-oblivious algorithm for short-read mapping," *Nature Methods*, vol. 7, no. 8, pp. 576–577, 2010.

[96] M. C. J. Maiden, J. A. Bygraves, E. Feil et al., "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 6, pp. 3140–3145, 1998.

[97] M. D. Lynch, A. P. Masella, M. W. Hall, A. K. Bartram, and J. D. Neufeld, "AXIOME: automated exploration of microbial diversity," *GigaScience*, vol. 2, article 3, 2013.

[98] F. Angly, B. Rodriguez-Brito, D. Bangor et al., "PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information," *BMC Bioinformatics*, vol. 6, article 41, 2005.

[99] W. Li, "Analysis and comparison of very large metagenomes with fast clustering and functional annotation," *BMC Bioinformatics*, vol. 10, article 359, 2009.

[100] M. Rho, H. Tang, and Y. Ye, "FragGeneScan: predicting genes in short and error-prone reads," *Nucleic Acids Research*, vol. 38, no. 20, article e191, 2010.

[101] W. Zhu, A. Lomsadze, and M. Borodovsky, "*Ab initio* gene identification in metagenomic sequences," *Nucleic Acids Research*, vol. 38, article e132, Article ID gkq275, 2010.

[102] H. Noguchi, T. Taniguchi, and T. Itoh, "MetaGeneAnnotator: detecting species specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes," *DNA Research*, vol. 15, no. 6, pp. 387–396, 2008.

[103] H. Teeling, J. Waldmann, T. Lombardot, M. Bauer, and F. O. Glöckner, "TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences," *BMC Bioinformatics*, vol. 5, article 163, 2004.

[104] Y. Wang, H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin, "Metacluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample," *Bioinformatics*, vol. 28, no. 18, pp. i356–i362, 2012.

[105] A. Brady and S. L. Salzberg, "Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models," *Nature Methods*, vol. 6, no. 9, pp. 673–676, 2009.

[106] D. H. Huson, S. Mitra, H.-J. Ruscheweyh, N. Weber, and S. C. Schuster, "Integrative analysis of environmental sequences using MEGAN4," *Genome Research*, vol. 21, no. 9, pp. 1552–1560, 2011.