International Journal of Environmental & Agriculture Research (IJOEAR)
ISSN: [2454-1850]
[Vol-12, Issue-3, March- 2026]
Website: www.ijoear.com
Journal DOI: 10.25125/agriculture-journal

Metagenomics: Concepts, Methodologies and Transformative Applications

Amit Kumar Jha[1*]; R. K. Vandre[2]; S. S. Tomar[3]; Baleshwari Dixit[4]; Rajeev Ranjan[5]; Sheikh T. J.[6]; Jitendra Kumar Tripathi[7]; Rashmi Jha[8]

[*1]Department of Animal Genetics &Breeding, College of Veterinary Science & Animal Husbandry, Rewa (NDVSU)
[2-6]Nanaji Deshmukh Veterinary Science University, Jabalpur (Madhya Pradesh)
[7]Assistant Professor, Government Model Science College, Rewa (Madhya Pradesh)
[8]Ph. D. Scholar, Department of Biotechnology, APSU, Rewa (Madhya Pradesh)
*Corresponding Author

Received:- 10 March 2026/ Revised:- 19 March 2026/ Accepted:- 26 March 2026/ Published: 31-03-2026

***Abstract*—** *Metagenomics has emerged as a paradigm-shifting approach in microbiology, enabling direct genomic analysis of entire microbial communities from their natural environments without the constraints of laboratory cultivation. This comprehensive review synthesizes current methodologies, computational challenges, and breakthrough applications of metagenomic approaches. We examine the evolution from single-organism genomics to community-level genomic analysis, highlighting how technological advances in sequencing platforms have overcome traditional cultivation limitations. The review covers critical aspects including environmental sampling strategies, next-generation sequencing technologies, assembly algorithms, taxonomic binning approaches, and functional annotation pipelines. The profound implications for understanding microbial ecology, symbiotic relationships, and the discovery of novel gene families are discussed. Current computational challenges and emerging solutions are evaluated, along with the transformative potential of third-generation sequencing technologies. This review positions metagenomics as a foundational technology driving discoveries across environmental microbiology, clinical diagnostics, biotechnology, and our fundamental understanding of microbial contributions to planetary processes.*

***Keywords*— metagenomics, microbial communities, environmental genomics, next-generation sequencing, taxonomic binning, functional annotation.**

I. INTRODUCTION

The intimate relationship between higher organisms and their associated microbial communities has become increasingly recognized as fundamental to understanding biological systems. Humans and animals harbor more bacterial cells than their own somatic cells, emphasizing the critical importance of microbial genomics in comprehending host-microbe interactions and ecosystem dynamics. Because of the intimate relationship of humans and animals with microbes, sequencing the genomes of microbes is necessary as this would facilitate better understanding of the role of microbes in the biosphere. Traditional microbiology, constrained by the requirement for axenic cultures, has provided insights into only a minute fraction of the microbial world. Only a small percentage of the microbes in nature can be cultured, which means that extant genomic data are highly biased and do not represent a true picture of the genomes of microbial species.

The field of genomics experienced its first revolution with the sequencing of complete microbial genomes, beginning with bacteriophages MS2 and φ-X174 in the late 1970s, followed by the landmark sequencing of *Haemophilus influenzae* in 1995. This single-organism approach, while revolutionary, faced inherent limitations: the cultivation bias that excluded the vast majority of environmental microorganisms, and the failure to capture the complex interactions within natural microbial communities. Metagenomics, literally meaning "beyond the genome," represents a paradigmatic shift that circumvents these

limitations by enabling direct genomic analysis of entire microbial communities. Sequence data taken directly from the environment are called metagenomes, and the study of sequence data taken directly from the environment is metagenomics. This approach harnesses environmental DNA (eDNA) extracted directly from natural habitats, providing unprecedented access to the genetic potential of uncultivated microorganisms and their ecological interactions. New sequencing technologies and the drastic reduction in the cost of sequencing have helped tremendously in overcoming these limitations. We now have the ability to obtain genomic information directly from microbial communities in their natural habitats. Suddenly, instead of looking at a few species individually, we are able to study tens of thousands all together.

II. METHODOLOGICAL FRAMEWORK

2.1 Environmental Sampling Strategies:

The foundation of any metagenomic investigation lies in representative sampling of the target microbial community. Unlike traditional microbiological approaches where target organisms are visible, metagenomic sampling must account for the invisible microbial world and its inherent heterogeneity.

Sample Size and Replication: To estimate the fraction of species sequenced, rarefaction curves are typically used. A rarefaction curve plots the number of species as a function of the number of individuals sampled. The curve usually begins with a steep slope, which at some point begins to flatten as fewer species are being discovered per sample. For microbial samples, different operational taxonomic units (OTUs) are typically characterized by 16S (prokaryotic) or 18S (eukaryotic) rDNA, also referred to as ribotypes.

Filtration and Size Fractionation: When filtering an environmental sample, the goals are: (1) obtain as much as possible of what is desired and (2) exclude as much as possible of what is not desired. Computational filtering can be used after sequencing. Genomic material that is obviously within the clades of interest can be filtered in using similarity searches against annotated sequence databases. Care must be taken with false negatives: relevant genomic material may be filtered out simply because homologues have never been deposited in existing databases.

Metadata Collection and Standardization: Keeping strict and comprehensive records of metadata is as important as the sequence data. Metadata are the "data about the data": where the samples were taken from, when, and under which conditions. Many metagenomic studies are driven by discovery and data mining, rather than by hypothesis. These studies seek statistically significant correlations between the metagenomic data and the habitat-associated metadata, which may lead to biologically significant discoveries.

2.2 Sequencing Technologies and Their Evolution:

First-Generation Sanger Sequencing: Until recently, prokaryotic genomes have been typically sequenced using Sanger shotgun sequencing. The first step is shearing the DNA content of a genomic clone into random fragments, hence the "shotgun." The fragments are then cloned into plasmid vectors that are grown in monoclonal libraries to produce enough genomic material for sequencing. Another disadvantage of shotgun sequencing is the "cloning bias." Some genes cannot be incorporated into the library vector, usually because of toxicity to the vector expressing them.

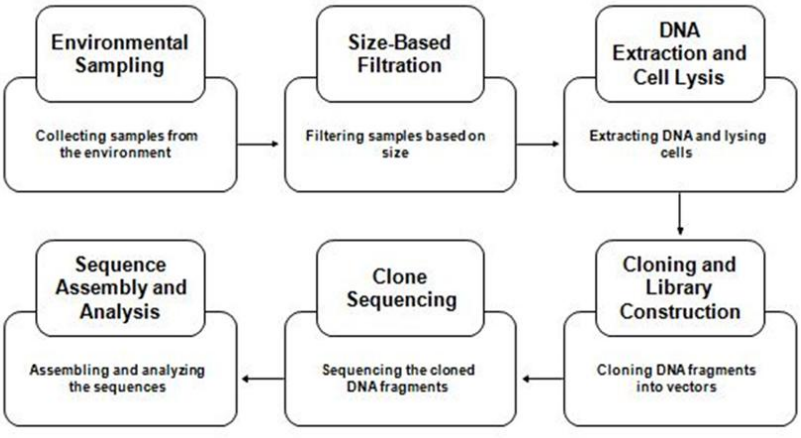

FIGURE 1: Environmental shotgun sequencing workflow: (A) Environmental sampling; (B) Size-based filtration; (C) DNA extraction and cell lysis; (D) Cloning and library construction; (E) Clone sequencing; (F) Sequence assembly and analysis.

Second-Generation High-Throughput Sequencing: Second-generation sequencing methods have been rapidly gaining ground and are replacing Sanger sequencing for small-sized genomes and environmental genomics. A common denominator among second-generation methods is the generation of "polymerase colonies" or polonies. Polonies are PCR amplicons derived from a single molecule of nucleic acid.

In pyrosequencing methods such as Roche 454 sequencing, sequencing is performed by polymerase extension of a primed template. Single nucleotide species are added at each cycle. If the particular nucleotide species added to the polymerase reaction pairs with the one on the template, the incorporation causes a luciferase-based light reaction.

FIGURE 2: Pyrosequencing mechanism: Single stranded DNA template is first hybridized with the sequencing primer and mixed with the two substrates adenosine 5'-phosphosulfate (APS) and luciferin. In each cycle, (1) one of four nucleotides is added to the reaction. (2) If the nucleotide is complementary to the base in the template strand, then the DNA polymerase incorporates it into the growing strand. (3) Pyrophosphate (PPi) is released and converted to ATP by sulfurylase. (4) ATP serves as a substrate to luciferase, causing a light reaction. (5) Excess nucleotides are degraded by apyrase.

Sequencers such as Illumina produce shorter reads (50-300 bp) but generate very large volumes of DNA per sequencing run. Despite the individual short read lengths, these technologies provide a viable alternative for sequencing whole genomes, due to the large volume of DNA sequenced.

Emerging Third-Generation Technologies: Third-generation sequencing, loosely defined as technology that is capable of sequencing long sequences without amplification, represents an advanced development. Long-read sequencing technologies promise to address current assembly limitations by generating contiguous sequences spanning repetitive regions and complete

operons. These technologies, including platforms from PacBio and Oxford Nanopore, enable the sequencing of DNA molecules up to tens of kilobases in length, facilitating the assembly of complex microbial communities and the resolution of repetitive genomic regions.

2.3 Assembly Algorithms and Computational Approaches:

Challenges in Metagenomic Assembly: For sequencing a whole genome, the reads are assembled into progressively longer contiguous sequences or contigs and finally to the whole genome. The incomplete and fragmentary nature of metagenomic data presents challenges. In contrast, in all but the most species-poor metagenome, a full assembly is not possible—first, because the sampling is incomplete and many if not all species' genomes are partially sampled, if at all; second, because the species information itself is incomplete, and it is difficult to map individual reads to their species of origin.

Graph-Based Assembly Strategies: Phrap, Forge, Arachne, JAZZ, and the Celera Assembler are the assembly programmes developed for single genome assembly from Sanger sequencing. They seem to provide good results even when assembling metagenomic sequence data from Sanger sequencing. Most of these assembly algorithms use mate-pair information to check the scaffolds or the assembled intermediaries between raw reads and whole chromosomes.

For short reads, these techniques are not suitable. The solution to a Hamiltonian path is an NP-complete problem, meaning that the time necessary for a solution grows exponentially with the number of nodes. The EULER assembler was the first to present an alternative technique using de Bruijn graphs, which represent sequences as overlaps of k-mers rather than individual reads, making assembly of short reads computationally feasible.

Specialized Assembly Approaches: Recently, assembly methods have been developed that find putative open reading frame (ORF) regions first, and then assemble those regions. This method, dubbed ORFome assembly, increases assembly accuracy for ORF regions at the expense of losing noncoding regions.

III. ANALYTICAL METHODOLOGIES

3.1 Coverage Assessment and Genome Size Estimation:

Coverage Calculations: Coverage of a genome is defined as the mean number of times a nucleotide is sequenced. If DNA shearing and sequencing are treated as random events, then the Poisson distribution model can be used to estimate the number of reads required to sequence an entire genome. This model is given by the Lander-Waterman equation:

$$C = \frac{L \times N}{G} \tag{1}$$

Where L is the read length, N is the number of reads, G is the genome length, and C is coverage. The fraction of sequence covered would be given as:

$$P_0 = 1 - e^{-C} = 1 - e^{-(LN/G)} \tag{2}$$

Effective Genome Size (EGS) Estimation: An effective genome size (EGS) measure has been suggested that includes multiple plasmid copies, inserted sequences, and associated phages and viruses. EGS uses the density of single copy marker genes to extrapolate the EGS:

$$EGS = \frac{a + b \times L^{-c}}{x} \tag{3}$$

Where L is the read length, x is the marker gene density, and a, b, and c are empirical parameters.

3.2 Gene Identification and Functional Prediction:

Gene Calling Challenges: Genes are the basic functional unit in the genome, which may constitute larger functional units such as operons, transcriptional units, and functional networks. In the Global Ocean Sampling (GOS) data, which were Sanger-sequenced, the mean number of whole reading frames per assembly is 4.7.

For a high complexity metagenomic dataset, gene prediction on assemblies can be as accurate as 85% of the originally predicted genes in the constituting genomes, and for a low complexity set this goes up to 90%. For genes with known homologs, BLASTing against known databases is a common approach. For new families and new genes that have no homologs in known databases, *ab initio* gene prediction tools are used. GeneMark.hmm is a programme that uses inhomogeneous Markov models based on monocodon frequency analysis for gene calling.

Innovative Gene Discovery Approaches: A different approach to gene finding involves beginning with simple ORF identification of consecutive translatable regions that translate to at least 60 amino acids and then clustering those sequences using an all-against-all BLAST search.

3.3 Taxonomic Classification and Binning:

Composition-Based Binning: There are two binning strategies: composition-based binning and similarity-based binning. GC content of bacterial genomes is used routinely for higher-level systematics. Tetranucleotides are used by the TETRA program. PhyloPythia is a supervised method that trains a set of support vector machines (SVMs) to bin sequences of a length greater than 1 kb. Growing Self Organizing Maps and Seeded GSOM (S-GSOM) use a variant of the machine learning algorithm self-organizing maps.

A composite supervised method that uses both TETRA and codon usage statistics has been developed to classify fragments in the 100-300-bp range.

Similarity-Based Binning: MEGAN implements similarity-based binning by reading a BLAST file output. MEGAN assigns each read to the lowest common ancestor on the phylogenetic tree. Phymm uses interpolated Markov models to characterize variable length DNA sequences by their phylogenetic groupings.

3.4 Diversity Assessment and Community Structure Analysis:

Alpha Diversity Quantification: Shannon's index is used to calculate α-diversity:

$$H'_{\alpha} = -\sum_{i=1}^{S} p_i \ln p_i$$

Where $p_i = n_i/N$, S is the total number of OTUs, n_i is the number of clones in each OTU, and N is the total number of individuals.

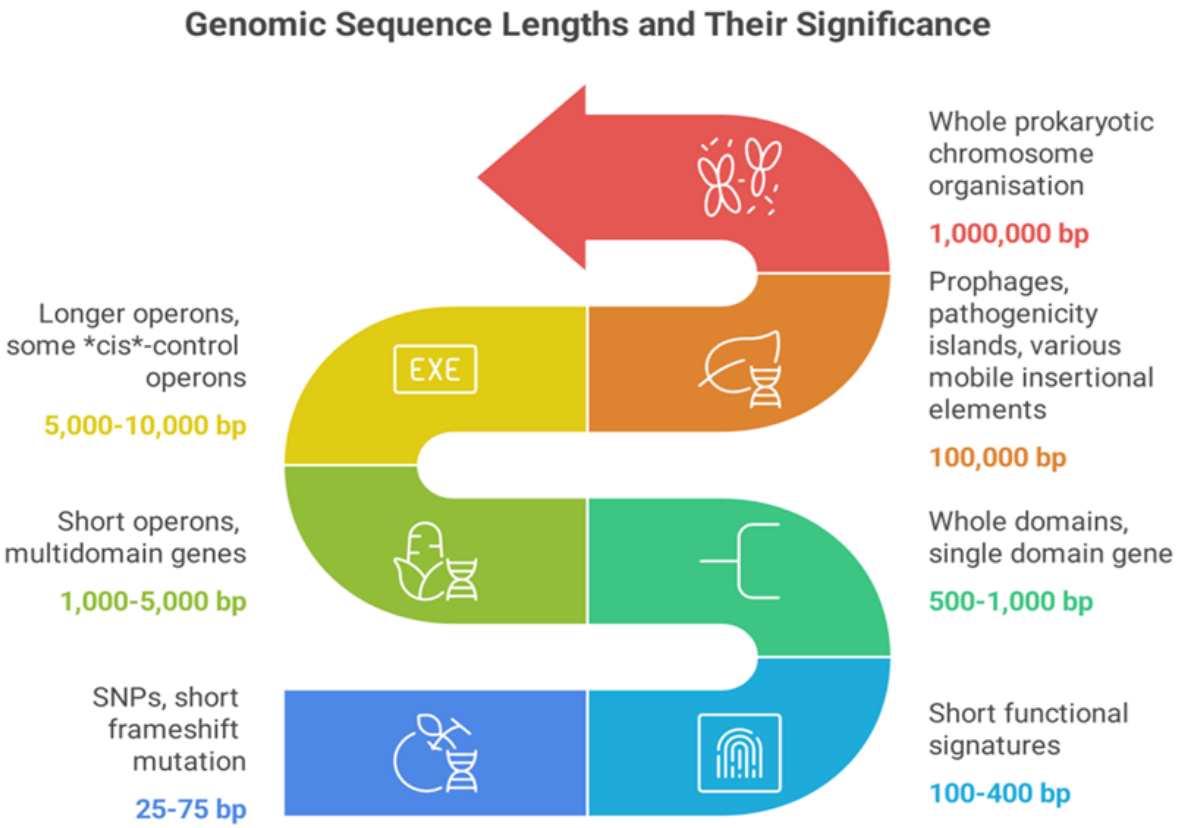

FIGURE 3: Rarefaction curves interpretation: Green curve indicates comprehensive species sampling; blue curve shows incomplete habitat sampling; red curve represents species-rich environments with minimal sampling coverage.

Phylogenetic Marker Considerations: Using 16S/18S rDNA as a proxy for OTU identification is not without limitations. Evidence of horizontal gene transfer involving rDNA may confound its reliability. 16S rDNA may exist in multiple different sequence copies in a single bacterium: the mean number of bacterial ribosomal operons per genome is 4.1, but 16S rDNA gene copy numbers may vary between 1 and 15. Alternative markers, such as single copy housekeeping genes including *rpoB*, *amoA*, *pmoA*, *nirS*, *nirK*, *nosZ*, and *pufM* have been suggested.

Several software packages are useful for biodiversity analysis. EstimateS contains a rich set of biodiversity analysis modules. MOTHUR is tailored towards microbial diversity analysis. QIIME is a very powerful and versatile package for analysis of genomic and metagenomic microbial ecology data. PHACCS is specialized software geared to the analysis of viral metagenomic data.

IV. TRANSFORMATIVE APPLICATIONS

4.1 Environmental Microbiology and Ecosystem Function:

Metagenomic approaches have revolutionized understanding of microbial contributions to global elemental cycles. Functional gene surveys reveal metabolic potential for nitrogen fixation, sulfur oxidation, and carbon cycling across diverse environments. Comparative metagenomics across environmental gradients elucidates how community composition and metabolic capacity respond to geochemical variations.

Many comparative analyses make use of ordination statistics when several metagenomic datasets are involved. Principal component analysis (PCA) and nonmetric multidimensional scaling (NM-MDS) are typically used to visualize the data. Recent studies have suggested how to locate multivariate correlations between metagenomic data and environmental attributes, identifying covariation in amino acid transport and cofactor synthesis in nutrient-poor ocean areas.

4.2 Host-Associated Microbiomes and Human Health:

One notable study examined the connection between the gut microbiome and obesity. Researchers discovered that the metagenome in obese mice was enriched in carbohydrate-active enzymes over that of lean mice. A separate biochemical experiment confirmed that the microbiome in obese mice has a larger energy harvesting capacity than in lean mice and concluded that the gut microbiome contributes to obesity through this feed-forward cycle.

Metagenomic approaches are transforming clinical microbiology by enabling culture-independent pathogen detection and antimicrobial resistance gene surveillance. Recent molecular-based discoveries of highly prevalent viral infections highlight the need for a better understanding of the human viral flora.

4.3 Symbiotic Systems and Evolutionary Biology:

In many cases, symbiotic bacteria living in an animal host consist of a small number of species, which are often phylogenetically distant. Researchers sequenced the environmental shotgun sequencing (ESS) data from bacterial symbionts living in the glassy-winged sharpshooter, which is an insect that lives solely on tree sap, a nutrient-poor diet. By binning the ESS data, it has been inferred that one symbiont synthesizes amino acids for the host insect, while another synthesizes cofactors and vitamins.

Another study of the marine gutless worm *Olavius algarvensis* has revealed the different roles of its four symbionts in generating nutrients and processing the worm's waste. None of the symbionts in the insect or in the worm study could be cultured under the reported conditions. Metagenomics thus became the chosen avenue for these studies.

4.4 Biotechnological Innovation and Gene Discovery:

Another type of study enabled by metagenomics is the search for new members of a gene family. The previously small bacterial Eukaryotic Protein Kinase Like (ELK) family has been enriched several folds by the Global Ocean Sampling (GOS) project. Many new members of known families were identified, as well as new families.

4.5 Viral Metagenomics and Genetic Diversity:

Metagenomic studies have enriched our knowledge of viral diversity and the role viruses play as facilitators of microbial genetic diversity. Sequence similarity analyses of viral metagenomic data have shown that approximately 90% of the sequences have no similarity to GenBank sequences, a phenomenon often referred to as "viral dark matter."

The existence of photosynthetic genes in cyanophages—viruses infecting cyanobacteria—has been known for some time. However, metagenomic studies have revealed the extent of this phenomenon: it is estimated that 60% of the *psbA* genes, a component of Photosystem II, in surface water are of phage origin.

V. COMPUTATIONAL TOOLS AND PIPELINES

5.1 Annotation Pipelines:

The versatile and useful annotation pipelines for metagenomics include MG-RAST. MG-RAST accepts a 454 dataset as input, normalizes it, and then performs gene calling and annotation by a variety of sequence similarity searches against various sequence databases, including 16S rDNA.

RAMMCAP uses the fast clustering algorithm CD-HIT to cluster translated ORFs by high sequence similarity. The sequences are then compared to the profile HMM databases TIGRfam using HMMer for functional annotation. Motif Extraction (MEX) is an unsupervised motif creation method that is successful in identifying enzymes in genomic and metagenomic data.

5.2 Comparative Analysis Tools:

Besides using MEGAN as a binning software, it can also be used to compare the OTU composition of two or more frequency-normalized samples. MG-RAST provides a comparative functional and sequence-based analysis for uploaded samples. Other software used for the comparison of microbial populations include UniFrac and MetaStats. Galaxy provides online workbench capabilities for comparative metagenomic analysis. ShotgunFunctionalizeR is a stand-alone analysis tool written in R.

VI. FUTURE DIRECTIONS AND EMERGING TECHNOLOGIES

6.1 Third-Generation Sequencing Integration:

Long-read sequencing technologies, including platforms from PacBio (Single Molecule Real-Time sequencing) and Oxford Nanopore Technologies, promise to address current assembly limitations by generating contiguous sequences spanning repetitive regions and complete operons. These platforms enable the sequencing of native DNA molecules without amplification bias, and some technologies can differentiate between cytosine and methyl-cytosine during sequencing, providing additional epigenetic information directly from environmental samples.

6.2 Data Management Challenges:

A growing problem is that of data management. Sequencing centers are working to equip themselves with computational infrastructure to meet the flow of sequence data. However, many research institutes who request sequencing do not have the computational infrastructure needed to deal with analysis and long-term storage of these data. The development of cloud-based platforms and standardized data repositories represents an ongoing effort to address these challenges.

VII. CONCLUSIONS AND IMPLICATIONS

We are in the midst of the fastest growing revolution in molecular biology, perhaps in all of life science, and it only seems to be accelerating. Assembly, quality control, binning, and annotation all require ingenious algorithms combined with the latest computational power. Metagenomics represents a transformative technology that has fundamentally altered our understanding of microbial life and its planetary significance. By circumventing cultivation limitations, metagenomic approaches have revealed the vast extent of microbial diversity, complex ecological interactions, and the genomic basis of ecosystem function.

The applications discussed demonstrate metagenomics' broad impact across environmental science, medicine, biotechnology, and evolutionary biology. From understanding symbiotic relationships and discovering novel gene families to tracking viral diversity and uncovering the role of the microbiome in human health, metagenomics has become an indispensable tool in modern biological research.

As computational capabilities continue expanding and sequencing costs decline, metagenomics will undoubtedly drive further revolutionary discoveries in our understanding of life's microbial foundations. The integration of metagenomics with other omics approaches—metatranscriptomics, metaproteomics, and metabolomics—promises to provide even deeper insights into the functional dynamics of microbial communities. Continued development of computational infrastructure, standardized data sharing practices, and accessible analysis tools will be essential to ensure that the potential of metagenomics is fully realized across the global scientific community.

ACKNOWLEDGMENTS

The author acknowledges the contributions of researchers whose foundational work in metagenomics has made this review possible.

CONFLICT OF INTEREST

The author declares no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Savage, D. C. (1977). Microbial ecology of the gastrointestinal tract. *Annual Review of Microbiology, 31*, 107–133.

[2] Kaput, J., Cotton, R. G., Hardman, L., Watson, M., & Al Aqeel, A. (2009). Planning the human variome project: The Spain report. *Human Mutation, 30*(4), 496–510.

[3] O'Hara, A. M., & Shanahan, F. (2006). The gut flora as a forgotten organ. *EMBO Reports, 7*(7), 688–693.

[4] Rappé, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Review of Microbiology, 57*, 369–394.

[5] Gilbert, J. A., & Dupont, C. L. (2011). Microbial metagenomics: Beyond the genome. *Annual Review of Marine Science, 3*, 347–371.

[6] Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., & Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chemistry & Biology, 5*(10), R245–R249.

[7] Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, 74*(12), 5463–5467.

[8] Sorek, R., Zhu, Y., Creevey, C. J., Francino, M. P., & Bork, P. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. *Science, 318*(5855), 1449–1452.

[9] Mitra, R. D., & Church, G. M. (1999). In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic Acids Research, 27*(24), e34.

[10] Ronaghi, M., Uhlén, M., & Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. *Science, 281*(5375), 363–365.

[11] Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., & Gnerre, S. (2002). ARACHNE: A whole-genome shotgun assembler. *Genome Research, 12*(1), 177–189.

[12] Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., & Dehal, P. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science, 297*(5585), 1301–1310.

[13] Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., & Flanigan, M. J. (2000). A whole-genome assembly of *Drosophila*. *Science, 287*(5461), 2196–2204.

[14] Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., & McHardy, A. C. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature Methods, 4*(6), 495–500.

[15] Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences, 98*(17), 9748–9753.

[16] Chaisson, M. J., & Pevzner, P. A. (2008). Short read fragment assembly of bacterial genomes. *Genome Research, 18*(2), 324–330.

[17] Ye, Y., & Tang, H. (2009). An ORFome assembly approach to metagenomics sequences analysis. *Journal of Bioinformatics and Computational Biology, 7*(3), 455–471.

[18] Lander, E. S., & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics, 2*(3), 231–239.

[19] Raes, J., Korbel, J. O., Lercher, M. J., von Mering, C., & Bork, P. (2007). Prediction of effective genome size in metagenomic samples. *Genome Biology, 8*(1), R10.

[20] Yooseph, S., Li, W., & Sutton, G. (2007). Gene identification and protein classification in microbial metagenomic sequence data via incremental clustering. *BMC Bioinformatics, 8*(1), 182.

[21] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research, 25*(17), 3389–3402.

[22] Azad, R. K., & Borodovsky, M. (2004). Probabilistic methods of identifying genes in prokaryotic genomes: Connections to the HMM theory. *Briefings in Bioinformatics, 5*(2), 118–130.

[23] Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., & Glöckner, F. O. (2004). TETRA: A web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics, 5*(1), 163.

[24] McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., & Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nature Methods, 4*(1), 63–72.

[25] Chan, C. K., Hsu, A. L., Halgamuge, S. K., & Tang, S. L. (2008). Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics, 9*, 215.

[26] Tzahor, S., Aharonovich, D. M., Kirkup, B., Yogev, T., & Frank, I. B. (2009). A supervised learning approach for taxonomic classification of core-photosystem-II genes and transcripts in the marine environment. *BMC Genomics, 10*, 229.

[27] Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Research, 17*(3), 377–386.

[28] Brady, A., & Salzberg, S. L. (2009). Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. *Nature Methods, 6*(9), 673–676.

[29] Schouls, L. M., Schot, C. S., & Jacobs, J. A. (2003). Horizontal transfer of segments of the 16S rRNA genes between species of the *Streptococcus anginosus* group. *Journal of Bacteriology, 185*(24), 7241–7246.

[30] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., & Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology, 72*(7), 5069–5072.

[31] Case, R. J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W. F., & Kjelleberg, S. (2007). Use of 16S rRNA and *rpoB* genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology, 73*(1), 278–288.

[32] Klappenbach, J. A., Saxman, P. R., Cole, J. R., & Schmidt, T. M. (2001). rrndb: The Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Research, 29*(1), 181–184.

[33] Achenbach, L. A., Carey, J., & Madigan, M. T. (2001). Photosynthetic and phylogenetic primers for detection of anoxygenic phototrophs in natural environments. *Applied and Environmental Microbiology, 67*(7), 2922–2926.

[34] Colwell, R. K. (2005). *EstimateS: Statistical estimation of species richness and shared species from samples*. http://viceroy.eeb.uconn.edu/estimates/

[35] Angly, F., Rodriguez-Brito, B., Bangor, D., McNairnie, P., & Breitbart, M. (2005). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics, 6*, 41.

[36] Willner, D., Furlan, M., Haynes, M., Schmieder, R., Angly, F. E., & Silva, J. (2009). Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLOS ONE, 4*(10), e7370.

[37] Gianoulis, T. A., Raes, J., Patel, P. V., Bjornson, R., Korbel, J. O., & Letunic, I. (2009). Quantifying environmental adaptation of metabolic pathways in metagenomics. *Proceedings of the National Academy of Sciences, 106*(5), 1374–1379.

[38] Turnbaugh, P. J., Ley, R. E., Mahowald, M. A., Magrini, V., Mardis, E. R., & Gordon, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature, 444*(7122), 1027–1031.

[39] Nishizawa, T., Okamoto, H., Konishi, K., Yoshizawa, H., Miyakawa, Y., & Mayumi, M. (1997). A novel DNA virus (TTV) associated with elevated transaminase levels in posttransfusion hepatitis of unknown etiology. *Biochemical and Biophysical Research Communications, 241*(1), 92–97.

[40] Woyke, T., Teeling, H., Ivanova, N. N., Huntemann, M., Richter, M., & Glöckner, F. O. (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature, 443*(7114), 950–955.

[41] Edwards, R. A., & Rohwer, F. (2005). Viral metagenomics. *Nature Reviews Microbiology, 3*(6), 504–510.

[42] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., & Kubal, M. (2008). The metagenomics RAST server: A public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics, 9*, 386.

[43] Haft, D. H., Selengut, J. D., & White, O. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Research, 31*(1), 371–373.

[44] Kunik, V., Meroz, Y., Solan, Z., Sandbank, B., Weingart, U., & Ruppin, E. (2007). Functional representation of enzymes by specific peptides. *PLOS Computational Biology, 3*(8), e167.

[45] Mitra, S., Klar, B., & Huson, D. H. (2009). Visual and statistical comparison of metagenomes. *Bioinformatics, 25*(14), 1849–1855.