Title

Metagenomic tools in microbial ecology research

Authors

Taş, Neslihan

de Jong, Anniek Ee

Li, Yaoming

et al.

Publication Date

2021-02-01

DOI

10.1016/j.copbio.2021.01.019

Peer reviewed

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Metagenomic tools in microbial ecology research

Neslihan Taş[1,2], Anniek EE de Jong[3,4], Yaoming Li[5,6], Gareth Trubl[7], Yaxin Xue[8] and Nicholas C Dove[9]

Ability to directly sequence DNA from the environment permanently changed microbial ecology. Here, we review the new insights to microbial life gleaned from the applications of metagenomics, as well as the extensive set of analytical tools that facilitate exploration of diversity and function of complex microbial communities. While metagenomics is shaping our understanding of microbial functions in ecosystems via gene-centric and genome-centric methods, annotating functions, metagenome assembly and binning in heterogeneous samples remains challenging. Development of new analysis and sequencing platforms generating high-throughput long-read sequences and functional screening opportunities will aid in harnessing metagenomes to increase our understanding of microbial taxonomy, function, ecology, and evolution in the environment.

Addresses
[1] Earth and Environmental Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

[2] Biosciences Area, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[3] Deltares, Daltonlaan 600, 3584 BK Utrecht, The Netherlands
[4] Department of Microbiology, Institute for Water and Wetland Research, Radboud University, Heyendaalseweg 135, 6525 AJ, Nijmegen, The Netherlands
[5] School of Grassland Science, Beijing Forest University, Beijing, 100083, China
[6] Zhejiang Tiantong Forest Ecosystem National Observation and Research Station, Shanghai, 200241, China
[7] Physical and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA
[8] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, N-5008, Norway
[9] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Corresponding author: Taş, Neslihan (ntas@lbl.gov)

Current Opinion in Biotechnology 2021, **67**:184–191

This review comes from a themed issue on **Environmental biotechnology**

Edited by **Robbert Kleerebezem** and **Diana Machado de Sousa**

https://doi.org/10.1016/j.copbio.2021.01.019

Introduction

Over 20 years ago the term 'metagenomics' [1] emerged, announcing a new frontier in exploring the cross-section of biology and chemistry. From its first application in soil, metagenomics has become a driving force for discoveries in microbial ecology and biotechnology and a key method in exploring the microbial universe. As sequencing technologies became cheaper, faster, easier to use and higher-throughput, our ability to survey microbial diversity and functional potential in any ecosystem increased dramatically. Successful applications of metagenomics are closely tied to the availability and capability of the computational methods. In this review, we highlight the new knowledge gleaned from applications of metagenomics in Earth's different ecosystems as well as the analytical tools that enable such discoveries.

A journey from genes to ecosystem functions

Metagenomics, direct sequencing and analysis of DNA from microbial assemblages, has rapidly become a routinely employed method to characterize the functional potential of microbial communities. In its most straightforward application, DNA is extracted, prepared into libraries, and sequenced either on short-read (Illumina, Roche 454, Ion Torrent) or long-read (PacBio, Oxford Nanopore) platforms. All metagenomics analyses start with quality control of the sequence reads, which aims to minimize sequence bias and artifacts by removing adapter sequences, low quality bases calls, and contaminant sequences that are not from the source environment (Table 1). Earlier applications of metagenomics relied largely on a gene-centric approach to quantify the relative abundance of genes of interest and their function within a metagenome [2], which requires gene detection [3] and annotation of short reads. These efforts immediately increased the number of gene clusters in databases [2,4] and spurred greater interest in the applications of metagenomics. Various tools enable this analysis (Tables 1 and 2); however, almost 50% of genes in environmental microbiomes lack annotated functions. This parallels the fact that one-third of protein-coding genes in microbial isolate genomes are unannotated [5]. Hence, our ability to identify functional genes is closely tied to the completeness of gene databases and improvements to our collective knowledge of gene functions [5].

Gene-centric metagenome analysis can be performed via stand-alone tools or web-based applications. Read based annotations require aligning predicted gene sequences to known genes to infer functional gene abundances and

Table 1

Bioinformatic programs used in sequence read quality control, assembly, binning and metagenome assembled genome (MAG) refinement

Tools	Features	Website
Sequence Read Quality Control		
FastQC	Provides several graphic QC statistics information	[link]
MultiQC	Aggregates results from multiple samples into one single report	[link]
FastQ Screen	Screens sequences against a set of reference database	[link]
BBDuk	Decontaminates sequences using Kmer-based operations	[link]
Khmer	Trims and normalizes sequences for Kmer-based analysis	[link]
Read Assembly		
CLC Assembler	A De-bruijn graph-based assembly tool that integrates in commercial CLC workbench developed by QIAGEN	[link]
Meta-IDBA	Attempts to cover for both high and low abundant genomes by iterating with multiple k-mer size	[link]
MetaVelvet-SL	An extension of Velvet assembler hat integrating a Support Vector Machine (SVM) – is trained by a similar population of samples – to increase the performance	[link]
MEGAHIT	Uses increasing k-mer strategy with succinct de Bruijn data structure to reduce computational cost	[link]
Metaspades	A mode of the assembly software SPAdes for metagenomic assembly, using a heuristic method to distinguish interspecies repeats. Single cell mode is recommended for viromes.	[link]
Assembly Quality Check		
Quast	Evaluates genome/metagenome assemblies by computing various metrics such as contig length, N50, GC content	[link]
dnAQET	A Java package designed to evaluate scaffolds/contigs against a reference genome	[link]
GenomeQC	A toolkit that integrates multiple metrics to characterize both assembly and gene annotation quality across multiple data	[link]
Binning and Metagenome Assembled Genome Refinement		
MetaBAT2	Uses a k-medoid clustering method to bin contigs by calculating pairwise distance based on tetranucleotide frequency	[link]
Maxbin2	Employs an Expectation-Maximization (EM) algorithm to cluster contigs after co-assembly of multiple metagenomic datasets	[link]
CONCOCT	An unsupervised binning approach for metagenomic contigs by using nucleotide composition - kmer frequencies - and coverage data	[link]
GroopM	An automated binning tool that uses differential coverage (spatio-temoral model) to obtain high quality bins genomes from multi-sample metagenomes.	[link]
DASTool	Integrates results from various binning algorithms to calculate an optimized, non-redundant set of bins	[link]
CheckM	A common tool used to evaluate the quality of recovered MAGs, like completeness and contamination, based on the frequency of single-copy marker genes	[link]
Gene prediction		
FragGeneScan	Predicts genes from short reads incorporating a sequence error model and codon usage statistics.	[link]
Glimmer-MG	Uses interpolated Markov models (IMMs) to identify the coding regions and distinguish them from noncoding DNA.	[link]
Prodigal	Provides fast gene prediction from prokaryotic genomes, includes normal mode (reference-based) and anonymous mode (metagenomes).	[link]
MetaGeneMark	Predicts protein coding genes in metagenomic data using ab initio approaches.	[link]
Prokka	Identifies genes against series external databases that can annotate bacterial, archaeal and viral genomes.	[link]

distribution. Developing a stand-alone analysis capability requires local computational resources and proficiency in bioinformatics, but in return, users can rapidly assess the sensitivity of different parameters and analytical approaches to improve annotations. Web servers (Table 3) provide a user-friendly analysis platform that is accessible to researchers from all experience levels, but they are limited to small data sizes, provide results from a select list of analysis tools and analysis completion can take weeks to months (depending on the server load). Despite these limitations, services such as JGI IMG/M [6], MG-RAST [7], CyVerse [8] and KBase [9] serve a wide research community and support development and deployment of new analytical tools. Metagenomics now is a well-established technology, where a growing number of datasets [10] and analysis tools are accessible to many researchers [11].

Information gleaned from gene-centric metagenomics, at times in combination with RNA sequencing (metatranscriptomics) and protein identification (metaproteomics), provides greater understanding of microbial processes governing biogeochemical cycles in ecosystems. Metagenomic analysis of taxonomic and functional diversity in prairie soil microbiomes, for example, showed how long-term agricultural practices can result in the loss of

Table 2

Commonly used software for discovery of phylogeny and functional potential based on search against database

Database	Tools	Key features	Website
Nucleotide	Kraken2	Exacts k-mer search in memory	[link]
	Bracken	Computes relative abundance of species using Bayesian estimation	[link]
	CLARK	Supervised sequence classification using discriminative k-mers	[link]
	k-SLAM	K-mer search with additional validation using pseudo-assembly	[link]
	Centrifuge	Fast and memory-efficient tools for taxonomic profiling using BWT	[link]
Protein	DIAMOND	Protein homology search using spaced seeds with a reduced amino acid alphabet	[link]
	Kaiju	Fast for large-scale profiling in protein database	[link]

Table 3

Metagenomic analysis platforms enabling gene- and genome-centric analysis

Platforms	Key features	Website
Web-based		
EBI metagenomics	A comprehensive platform for the assembly, analysis and archive microbiome data	[link]
MG-RAST	An open source web application for gene-centric analysis that offer automated quality control, annotation, comparative analysis and archiving services.	[link]
KBase	A suite of microbiome analysis apps for gene- and genome-centric analysis with a graphic interface. User friendly import, export, and data edits and metabolic modelling capability	[link]
IMG/M	A platform for comparative analysis and functional annotation for public available genomes, metagenomes and metatranscriptomes.	[link]
Local installation		
MetAMOS	A modular framework for metagenomic assembly, taxonomic and functional annotations, and integrated HTML report.	[link]
MOCAT2	A toolkit to generate assembly, gene predictions, gene catalogs, gene catalog annotations, functional or taxonomic profiles for metagenomics.	[link]
Anvi'o	Provides integrated analysis strategies for genomics, metagenomics, metatranscriptomics, pangenomics, metapangenomics, phylogenomics, and microbial population genetics in an integrated and has extensive interactive visualization capabilities.	[link]
Metawrap	Metagenomic wrapper suite that accomplishes the core tasks of metagenomic analysis from start to finish: read quality control, assembly, visualization, taxonomic profiling, extracting and refining draft genomes (binning), and functional annotation.	[link]
METABOLIC	This software enables the prediction of metabolic and biogeochemical functional trait profiles to any given genome datasets. These genome datasets can either be MAGs, single-cell amplified genomes or pure culture genomes.	[link]
MetaSanity	Provides a unified workflow for genome assessment and functional annotation that combines all outputs into a single queryable database	[link]
DRAM/ DRAM-v	A tool for annotating MAGS and VirSorter identified viral contigs. DRAM annotates MAGs and viral contigs using KEGG (if provided by the user), UniRef90, PFAM, dbCAN, RefSeq viral, VOGDB and the MEROPS peptidase database as well as custom user databases.	[link]

keystone species leading to loss of functional diversity [12]. Combined use of metagenomics, metatranscriptomics and metaproteomics allows elucidation of microbial functions regulating greenhouse gas emissions [13] and their distribution [14] in climate sensitive arctic tundra soils. Furthermore, biochemical and environmental factors impacting microbial functions in sandy sediments were assessed via metagenomics identifying the importance of temporal processes resulting in frequent shifts between H_2-fermentation and H_2-respiration processes [15•]. Soil metagenomes can be further analyzed to find new biologically and environmentally important enzymes. In combination with empirical testing, this approach can be used to extend the categorization of known enzymes in databases [16].

Beyond the highly complex and diverse soil and sediment microbiomes, gene-centric analyses of rarely accessed natural and engineered environments expand our knowledge of fundamental microbial processes. Metagenomics of dust microbiomes showed that genes encoding proteins involved in repairing UV-induced DNA damage along with chemotaxis, germination, and heat-shock proteins were ubiquitous across different sampling locations in North Africa [17]. Patterns in zoonotic protist diversity in raw sewage were studied to understand their distribution in urban environments showing that functionally similar but phylogenetically diverse protist community were inhabiting New York City sewers [18]. Microbes inside or on the surface of plant tissues (roots, stems, and leaves) can impact plant productivity and health [19]. Plant-

associated microorganisms shown to contain extensive collection of carbohydrate metabolism functions and fewer motility genes, suggesting endophytic microorganisms have access to varied and widely-dispersed carbon substrates [20••]. As with other microbe-host systems, host DNA contamination can reduce the ability to sequence microbial reads in plant tissues. However, this can be leveraged to quantify microbial concentrations [21,22]. Interrogating Earth's ecosystems with metagenomic analyses is an arduous task, but such analyses may illuminate the full breadth of microbial function and ecology.

Back to the cell: genome recovery from metagenomes

The major assumption of gene-centric metagenomics is that the genes exist in a well homogenized and cellfree environment where all potential reactions can interactively occur [23]. This assumption, while fundamentally flawed, was a result of our inability to solve short read sequence puzzles into genomes. In earlier attempts, short reads could be organized to infer genome content in low diversity environments like acid mine drainage biofilms [24], but, until recently, characterizing genomes in complex communities, like soils, remained a challenge. Genome recovery from metagenomes in complex communities became possible as sequence read depth per sample increased. Sequencing deep enough to have high coverage, development of methods to reconstruct the long consensus region of DNA (contigs) from a pool of short sequence reads, and coverage-based binning of assembled contigs into population genomes gave rise to genome-centric metagenomics and metagenome assembled genomes (MAGs) [25]. Short-read assembly has unique challenges, notably due to varying abundances of bacteria and archaea within a community, high diversity, and strain-level variance. New generations of assemblers were designed to account for and leverage these distinctive aspects of the data (Table 1). Currently, there is no universally accepted assembler, thus, assembly quality is often evaluated by comparing different methods through summary statistics from single-genome assembly methods like size, contig N50 (i.e. the sequence length of the shortest contig at 50% of the total genome length), and maximum contig length. Contigs derived from assembly are still highly fragmented and redundant, and they cannot be directly grouped into genomes. Binning algorithms use a variety of genome properties such as DNA composition, GC content, tetranucleotide frequency in combination with depth of sequencing coverage, and abundance to group contigs into MAGs.

The direct output of genome binning often contains false-assignment contigs. Thus, it is common to refine and evaluate MAGs after the binning. Completeness and contamination [26] are two common metrics that are used to assess MAG quality. Completeness seeks to identify sets of single-copy marker genes. Similarly, contamination reports on misbinnings based on multiple detection of single-copy marker gene sets. Both metrics are prone to certain errors, such as insensitivity to strain heterogeneity, transposases, and RNA operons. As studies reporting MAGs are increasing, a set of standards, called the minimum information about a single MAG (MIMAG) [27], has been proposed to standardize reporting, emphasizing manual curation and review. Such standards are important in assuring that published MAGs are of high quality.

Genome-resolved metagenomics has transformed our ability to study uncultured microbes and has led to discoveries in taxonomy, microbial ecology, biogeochemistry, and evolutionary biology. Incorporation of MAGs into the tree of life has increased the number of known microbial phyla, dramatically altering our understanding of microbial phylogeny. For instance, MAG analyses contributed to the discovery of the Candidate Phyla Radiation, which includes over 70 phyla and two superphyla (*Parcubacteria* and *Microgenomates*) [28–30]. Furthermore, such analyses have contributed to the discovery of phylum '*Candidatus* Kryptonia' [31], exclusive to high-temperature pH-neutral geothermal springs. This lineage represented a taxonomic 'blind spot' because of mismatches in the primers commonly used for ribosomal gene surveys. MAGs showed a heterotrophic lifestyle and strong need for symbiosis with other microbes. More recently 52 515 MAGs were generated from over 10 000 metagenomes collected from various habitats covering all of Earth's terrestrial and aquatic environments [32]. This massive effort expanded the known phylogenetic diversity of bacteria and archaea by 44% by generating 12 556 novel candidate species-level operational taxonomic units spanning 135 phyla [32]. For archaea, the discovery of the Asgard superphylum, which also includes the *Lokiarchaeeota*, was a major achievement in enhancing our understanding of archaeal taxonomy [33••]. The cellular structure of these archaea contains many eukaryotic features and provides support for the emergence of eukaryotes from within the archaeal domain of life. In addition, new CRISPR–Cas systems were identified through analysis of MAGs where Cas9, previously found only in bacterial genomes, were also detected in the archaeal domain of life [34], providing new opportunities for testing and applications in biological and clinical research.

The analysis of MAGs has also revealed new insights into microbial metabolic diversity and niche differentiation. For example, *Planctomycetes* MAGs that are abundant in water systems were discovered to be able to perform nitrogen fixation in both the Pacific and the Atlantic Oceans [35•]. This showed the importance of heterotrophic bacteria in the fixation of nitrogen in the surface ocean. A new addition to the nitrogen cycle was the discovery of a complete nitrifying organism from an

engineered system, a process referred to as comammox [36]. Our knowledge of methane metabolisms was advanced via genome-resolved metagenomics where the detection of the key enzyme for methanogenesis (Methyl coenzyme M reductase) in newly discovered *Bathyarchaeota* and *Verstraetearchaeota* MAGs overturned the long-held paradigm that this functional capacity was restricted to *Euryarchaeota* [37,38]. Furthermore, some lineages of *Bathyarchaeota* are suggested to perform homoacetogenesis, the ability to solely use CO_2 and H_2 to generate acetate [39], a metabolic process that was thought to be restricted to the bacterial domain. Genome-resolved metagenomics can unravel complex community potential and interactions involved in organic matter decomposition [40]. Large scale analysis of 1529 MAGs from a permafrost thaw gradient showed previously undescribed fungal pathways for xylose degradation in bacteria. Further pairing of specific microbial populations and biogeochemistry revealed key populations that drive the mineralization of organic matter from plant-derived organic material to simple the greenhouse gases [40]. Explicitly linking microbial function to taxonomy is a major benefit of genome-resolved metagenomics, which will continue to pave the way for new discoveries in microbial ecology. Machine learning and artificial intelligence methods may help to unravel hidden patterns and metabolic capabilities of complex microbial communities and reveal ecological implications.

Viral metagenomics: the new kid on the block

The advent of meta-omic approaches has enabled the study of uncultivated viruses and entirely reshaped our understanding of viruses as major players in many of Earth's biogeochemical cycles. For example, viruses can affect microbial metabolism via lysing their hosts, which stops the host's metabolism while releasing nutrients that may drive other metabolisms (i.e. viral shunt), and during infection, viruses redirect and potentially augment (via auxiliary metabolic genes) host metabolism, changing the function of the host and its metabolic outputs [41]. Viruses can be mined from metagenomes (DNA viruses) and metatranscriptomes (RNA viruses) along with microbes. This allows for characterization of proviruses, viral episomal elements (outside of the genome, but within the host), and virions, as well as revealing viral expression levels and virus-host dynamics [42]. Shotgun metagenomic approaches can characterize viruses in the context of a microbial community, but to obtain rarer viral genomes, a targeted metagenomic approach is needed. Targeted approaches include cell-sorting, where viruses identified within the cells or close-proximity can be sequenced [43], and viral metagenomes (viromes), where viruses (and other entities of virus size) are physically separated from larger organisms via a filter before their nucleic acid is extracted and sequenced [41].

The rate of viral discovery from omic approaches is unparalleled and is complemented by the rise of virus-specific bioinformatics and contemporary technologies. New bioinformatics tools have allowed characterization of viral ecosystem impacts [44,45••], detection of obscure viruses [46••,47••], virus taxonomy for uncultured viruses [48], and global comparisons of viruses [45••,46••,49••]. The development of long-read technology allows detection of whole virions [50,51] and when combined with short-reads, allows increased detection and characterization of viral genomes [52,53]. Powerful tools, such as stable isotope probing and nano scale secondary ion mass spectroscopy, are being leveraged to describe virus activity and quantify virus-host interactions [54,55].

There are limitations in characterizing viruses via meta-omic approaches that need attention. Viruses don't have a universal marker gene and most detected viral genes have unknown function, some of which are host-derived. These limitations create challenges for prediction of genome completeness, a complete taxonomic framework, and whether the virus is virulent or temperate [44], all of which impede a complete understanding of viral impacts in an ecosystem. Over the next decade, advancements in methodology and bioinformatics along with increased utilization of tools will push viral metagenomics to move beyond 'stamp collecting' of viral genomes to the quantification of viruses in an ecosystem and evaluation of their ecology particularly as it changes in space and time.

Conclusions and outlook

While microbial ecologists dig deep into the new information metagenomes provide, the metagenomic analyses of microbiomes will continue to evolve by technological and accessibility improvements in DNA and RNA sequencing. Long-read (>10 kb) sequencing technologies hold a great potential to improve genome assemblies and assignment of taxonomy and function. However, these advantages are constrained by a high error rate (10–15%) [56,57]. Because the error rate may be greater than the genetic difference between organisms, especially for low-abundance organisms, the use of long-read data for metagenomics is currently in its developmental stages. Ever growing use of new platforms (e.g. Hi-C [58•] and Tn-seq [59•]) with metagenomics will add to current data generation efforts and create new bottlenecks for data storage and standardization. As long-read sequencing becomes cheaper and more accurate, currently used elaborate methods for MAG discovery will be challenged. Future metagenomics will be closely tied to data analysis solutions that can facilitate, high-speed search and memory-efficient assembly methods that are compatible with terra- to petabytes of data. However, a key companion to these analytic methods is expanding high-quality annotation databases that are pivotal to understand the mechanisms underlying microbiome functions. Moreover, improvements in sample preparation and sequencing

for low DNA and RNA inputs will allow us to sample on smaller scales and enable accessing genomic information from larger spatial scales [60]. Further attempts to adapt current sequencing technologies for absolute quantification [61] of all molecules within a microbial cell can aid in the scaling of core and dynamic functionality complex microbiomes to larger biogeochemical and ecosystem level interactions that drive the Earth's material cycles. Overcoming methodological challenges will continue to increase our understanding of microbial taxonomy, function, ecology, and evolution.

Conflict of interest statement

Nothing declared.

CRediT authorship contribution statement

Neslihan Taş: Writing - original draft. **Anniek EE de Jong:** Writing - original draft. **Yaoming Li:** Writing - original draft. **Gareth Trubl:** Writing - original draft. **Yaxin Xue:** Writing - original draft. **Nicholas C Dove:** Writing - original draft.

Acknowledgements

Funding for this work was provided to Neslihan Taş through by the Office of Biological and Environmental Research in the U.S. Dept. of Energy (DOE) Office of Science - Early Career Research program and to Nicholas Dove through Oak Ridge National Laboratory (ORNL) postdoctoral development funds (ORNL is managed by UT-Battelle, LLC, for the DOE under contract DEAC05-00OR22725). Yaoming Li was supported by the National Natural Science Foundation of China (41871067) and through Zhejiang Tiantong Forest Ecosystem National Observation and Research Station. Anniek E. E. de Jong was supported by the Deltares Strategic Research Program on Water and Health and through the Netherlands Earth System Science Center (NESSC) Gravitation Grant (024.002.001). Work conducted at Lawrence Livermore National Laboratory (LLNL) was supported by the DOE Office of Science, Office of Biological and Environmental Research Genomic Science program award SCW1632 and LLNL LDRD 18-ERD-041 and conducted under the auspices of DOE Contract DE-AC52- 07NA27344.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM: **Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products**. *Chem Biol* 1998, **5**:R245-R249.

2. Tringe SG, Von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC: **Comparative metagenomics of microbial communities**. *Science* 2005, **308**:554-557.

3. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ: **Prodigal: prokaryotic gene recognition and translation initiation site identification**. *BMC Bioinformatics* 2010, **11**:119.

4. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W: **The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families**. *PLoS Biol* 2007, **5**:e16.

5. Antczak M, Michaelis M, Wass MN: **Environmental conditions shape the nature of a minimal bacterial genome**. *Nat Commun* 2019, **10**:3100.

6. Chen I-MA, Chu K, Palaniappan K, Pillay M, Ratner A, Huang J, Huntemann M, Varghese N, White JR, Seshadri R: **IMG/M v. 5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes**. *Nucleic Acids Res* 2019, **47**:D666-D677.

7. Meyer F, Bagchi S, Chaterji S, Gerlach W, Grama A, Harrison T, Paczian T, Trimble WL, Wilke A: **MG-RAST version 4—lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis**. *Brief Bioinformatics* 2019, **20**:1151-1159.

8. Goff S, Vaughn M, McKay S, Lyons E, Stapleton A, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A *et al.*: **The iPlant collaborative: cyberinfrastructure for plant biology**. *Front Plant Sci* 2011:2.

9. Arkin AP, Stevens RL, Cottingham RW, Maslov S, Henry CS, Dehal P, Ware D, Perez F, Harris NL, Canon S: **The DOE systems biology knowledgebase (KBase)**. *Nat Biotechnol* 2018, **36**:566-569.

10. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A: **Structure and function of the global ocean microbiome**. *Science* 2015, **348**.

11. Mitchell AL, Scheremetjew M, Denise H, Potter S, Tarkowska A, Qureshi M, Salazar GA, Pesseat S, Boland MA, Hunter FMI: **EBI metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies**. *Nucleic Acids Res* 2018, **46**:D726-D735.

12. Fierer N, Ladau J, Clemente JC, Leff JW, Owens SM, Pollard KS, Knight R, Gilbert JA, McCulley RL: **Reconstructing the microbial diversity and function of pre-agricultural tallgrass prairie soils in the United States**. *Science* 2013, **342**:621-624.

13. Hultman J, Waldrop MP, Mackelprang R, David MM, McFarland J, Blazewicz SJ, Harden J, Turetsky MR, McGuire AD, Shah MB: **Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes**. *Nature* 2015, **521**:208-212.

14. Taş N, Prestat E, Wang S, Wu Y, Ulrich C, Kneafsey T, Tringe SG, Torn MS, Hubbard SS, Jansson JK: **Landscape topography structures the soil microbiome in arctic polygonal tundra**. *Nat Commun* 2018, **9**:1-13.

15. • Kessler AJ, Chen Y-J, Waite DW, Hutchinson T, Koh S, Popa ME, Beardall J, Hugenholtz P, Cook PL, Greening C: **Bacterial fermentation and respiration processes are uncoupled in anoxic permeable sediments**. *Nat Microbiol* 2019, **4**:1014-1023
Sandy sediments contain specialized but versatile microbial communities that can frequently adjust their metabolism between fermentation and respiration processes to accommodate environmental conditions.

16. Castillo Villamizar GA, Nacke H, Boehning M, Herz K, Daniel R: **Functional metagenomics reveals an overlooked diversity and novel features of soil-derived bacterial phosphatases and phytases**. *mBio* 2019, **10**.

17. Aalismail NA, Ngugi DK, Díaz-Rúa R, Alam I, Cusack M, Duarte CM: **Functional metagenomic analysis of dust-associated microbiomes above the Red Sea**. *Sci Rep* 2019, **9**:1-12.

18. Maritz JM, Ten Eyck TA, Alter SE, Carlton JM: **Patterns of protist diversity associated with raw sewage in New York City**. *ISME J* 2019, **13**:2750-2763.

19. Bhargava P, Khan M, Verma A, Singh A, Singh S, Vats S, Goel R: **Metagenomics as a tool to explore new insights from plant-microbe interface**. *Plant Microbe Interface*. Springer; 2019:271-289.

20. •• Levy A, Gonzalez IS, Mittelviefhaus M, Clingenpeel S, Paredes SH, Miao J, Wang K, Devescovi G, Stillman K, Monteiro F: **Genomic**

features of bacterial adaptation to plants. *Nat Genet* 2018, **50**:138-150
Extensive analysis of 3837 bacterial genomes results in identification of new plant-associated gene clusters and improve our knowledge on plant-microbe interactions.

21. Guo X, Zhang X, Qin Y, Liu Y-X, Zhang J, Zhang N, Wu K, Qu B, He Z, Wang X: **Host-associated quantitative abundance profiling reveals the microbial load variation of root microbiome**. *Plant Commun* 2020, **1**:100003.

22. Regalado J, Lundberg DS, Deusch O, Kersten S, Karasov T, Poersch K, Shirsekar G, Weigel D: **Combining whole-genome shotgun sequencing and rRNA gene amplicon analyses to improve detection of microbe–microbe interaction networks in plant leaves**. *ISME J* 2020:1-15.

23. McMahon K: **Metagenomics 2.0**. *Environ Microbiol Rep* 2015, **7**:38-39.

24. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF: **Community structure and metabolism through reconstruction of microbial genomes from the environment**. *Nature* 2004, **428**:37-43.

25. Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC, Wilkins MJ, Hettich RL, Lipton MS, Williams KH: **Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla**. *Science* 2012, **337**:1661-1665.

26. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW: **CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes**. *Genome Res* 2015. gr. 186072.186114.

27. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy T, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA: **Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea**. *Nat Biotechnol* 2017, **35**:725-731.

28. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF: **Unusual biology across a group comprising more than 15% of domain Bacteria**. *Nature* 2015, **523**:208-211.

29. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U: **Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system**. *Nat Commun* 2016, **7**:13219.

30. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW: **Recovery of nearly 8000 metagenome-assembled genomes substantially expands the tree of life**. *Nat Microbiol* 2017, **2**:1533-0001542.

31. Eloe-Fadrosh EA, Paez-Espino D, Jarett J, Dunfield PF, Hedlund BP, Dekas AE, Grasby SE, Brady AL, Dong H, Briggs BR *et al.*: **Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs**. *Nat Commun* 2016, **7**:10476.

32. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, Wu D, Paez-Espino D, Chen IM, Huntemann M *et al.*: **A genomic catalog of Earth's microbiomes**. *Nat Biotechnol* 2020.

33. •• Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, Van Eijk R, Schleper C, Guy L, Ettema TJ: **Complex archaea that bridge the gap between prokaryotes and eukaryotes**. *Nature* 2015, **521**:173-179.

34. Burstein D, Harrington LB, Strutt SC, Probst AJ, Anantharaman K, Thomas BC, Doudna JA, Banfield JF: **New CRISPR–Cas systems from uncultivated microbes**. *Nature* 2017, **542**:237-241.

35. • Delmont TO, Quince C, Shaiber A, Esen ÖC, Lee ST, Rappé MS, McLellan SL, Lücker S, Eren AM: **Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes**. *Nat Microbiol* 2018, **3**:804-813
Discovery of one thousand non-redundant microbial population genomes from the TARA Oceans metagenomes via Anvio analysis pipeline.

36. Daims H, Lebedeva EV, Pjevac P, Han P, Herbold C, Albertsen M, Jehmlich N, Palatinszky M, Vierheilig J, Bulaev A: **Complete nitrification by Nitrospira bacteria**. *Nature* 2015, **528**:504-509.

37. Evans PN, Parks DH, Chadwick GL, Robbins SJ, Orphan VJ, Golding SD, Tyson GW: **Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics**. *Science* 2015, **350**:434-438.

38. Vanwonterghem I, Evans PN, Parks DH, Jensen PD, Woodcroft BJ, Hugenholtz P, Tyson GW: **Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota**. *Nat Microbiol* 2016, **1**:1-9.

39. He Y, Li M, Perumal V, Feng X, Fang J, Xie J, Sievert S, Wang F: **Genomic and enzymatic evidence for acetogenesis among multiple lineages of the archaeal phylum Bathyarchaeota widespread in marine sediments**. *Nat Microbiol* 2016, **1**:1-9.

40. Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB, Zayed AAF, Hoelzle RD, Lamberton TO, McCalley CK, Hodgkins SB *et al.*: **Genome-centric view of carbon processing in thawing permafrost**. *Nature* 2018, **560**:49-54.

41. Trubl G, Hyman P, Roux S, Abedon ST: **Coming-of-age characterization of soil viruses: a user's guide to virus isolation, detection within metagenomes, and viromics**. *Soil Syst* 2020, **4**:23.

42. Martinez-Hernandez F, Fornas O, Gomez ML, Bolduc B, de La Cruz Peña MJ, Martínez JM, Anton J, Gasol JM, Rosselli R, Rodriguez-Valera F: **Single-virus genomics reveals hidden cosmopolitan and abundant viruses**. *Nat Commun* 2017, **8**:1-13.

43. Trubl G, Jang HB, Roux S, Emerson JB, Solonenko N, Vik DR, Solden L, Ellenbogen J, Runyon AT, Bolduc B: **Soil viruses are underexplored players in ecosystem carbon processing**. *MSystems* 2018, **3**.

44. Roux S, Adriaenssens EM, Dutilh BE, Koonin EV, Kropinski AM, Krupovic M, Kuhn JH, Lavigne R, Brister JR, Varsani A: **Minimum information about an uncultivated virus genome (MIUViG)**. *Nat Biotechnol* 2019, **37**:29-37.

45. •• Emerson JB, Roux S, Brum JR, Bolduc B, Woodcroft BJ, Jang HB, Singleton CM, Solden LM, Naas AE, Boyd JA: **Host-linked soil viral ecology along a permafrost thaw gradient**. *Nat Microbiol* 2018, **3**:870-880.

46. •• Roux S, Krupovic M, Daly RA, Borges AL, Nayfach S, Schulz F, Sharrar A, Carnevali PBM, Cheng J-F, Ivanova NN: **Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes**. *Nat Microbiol* 2019, **4**:1895-1906
Inoviruses can leave without killing their hosts. This paper 56 members of the *Inoviridae* are discovered using a machine learning approach showing expansive diversity of toxin–antitoxin and gene expression modulation systems, alongside evidence of both synergistic (CRISPR evasion) and antagonistic (superinfection exclusion) interactions with co-infecting viruses.

47. •• Starr EP, Nuccio EE, Pett-Ridge J, Banfield JF, Firestone MK: **Metatranscriptomic reconstruction reveals RNA viruses with the potential to shape carbon cycling in soil**. *Proc Natl Acad Sci U S A* 2019, **116**:25900-25908
Time-resolved analysis shows that RNA viruses are diverse, abundant, and active in soil. Shifts in eukaryote, RNA phage, and RNA viral abundances over a few-day period reveal that entire soil communities can rapidly respond to altered resource availability.

48. Jang HB, Bolduc B, Zablocki O, Kuhn JH, Roux S, Adriaenssens EM, Brister JR, Kropinski AM, Krupovic M, Lavigne R: **Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks**. *Nat Biotechnol* 2019, **37**:632-639.

49. •• Gregory AC, Zayed AA, Conceição-Neto N, Temperton B, Bolduc B, Alberti A, Ardyna M, Arkhipova K, Carmichael M, Cruaud C: **Marine DNA viral macro-and microdiversity from pole to pole**. *Cell* 2019, **177**:1109-1123
Metagenomic assembly of 145 marine viromes leads to discovery of 195,728 viral populations, shows five distinct ecological zones for marine viruses that are not follow the latitudinal diversity gradient.

50. Akpinar F, Yin J: **Characterization of vesicular stomatitis virus populations by tunable resistive pulse sensing**. *J Virol Methods* 2015, **218**:71-76.

51. Wu H, Chen Y, Zhou Q, Wang R, Xia B, Ma D, Luo K, Liu Q: **Translocation of rigid rod-shaped virus through various solid-state nanopores**. *Anal Chem* 2016, **88**:2502-2510.

52. Boldogkői Z, Moldován N, Balázs Z, Snyder M, Tombácz D: **Long-read sequencing–a powerful tool in viral transcriptome research**. *Trends Microbiol* 2019, **27**:578-592.

53. Warwick-Dugdale J, Solonenko N, Moore K, Chittick L, Gregory AC, Allen MJ, Sullivan MB, Temperton B: **Long-read viral metagenomics captures abundant and microdiverse viral populations and their niche-defining genomic islands**. *PeerJ* 2019, **7**:6800.

54. Pasulka AL, Thamatrakoln K, Kopf SH, Guan Y, Poulos B, Moradian A, Sweredoski MJ, Hess S, Sullivan MB, Bidle KD *et al.*: **Interrogating marine virus-host interactions and elemental transfer with BONCAT and nanoSIMS-based methods**. *Environ Microbiol* 2018, **20**:671-692.

55. Mayali X: **NanoSIMS: microscale quantification of biogeochemical activity with large-scale impacts**. *Annu Rev Mar Sci* 2020, **12**:449-467.

56. Nicholls SM, Quick JC, Tang S, Loman NJ: **Ultra-deep, long-read nanopore sequencing of mock microbial community standards**. *GigaScience* 2019, **8**.

57. Singh H, Sharma P, Kaur RP, Thakur D, Kaur P: **Computational metagenomics**. In *State-of-the-Art, Facts and Artifacts. In Metagenomics: Techniques, Applications, Challenges and Opportunities.* Edited by Chopra RS, Chopra C, Sharma NR. Springer Singapore; 2020:199-227.

58. • DeMaere MZ, Darling AE: **bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes**. *Genome Biol* 2019, **20**:46

An unsupervised analysis method that leverages the hierarchical nature of Hi-C interaction rates to resolve MAGs using a single time point.

59. • Thibault D, Jensen PA, Wood S, Qabar C, Clark S, Shainheit MG, Isberg RR, van Opijnen T: **Droplet Tn-Seq combines microfluidics with Tn-Seq for identifying complex single-cell phenotypes**. *Nat Commun* 2019, **10**:5729

Transposon-insertion sequencing in droplets enables encapsulation of individual transposon mutants in microfluidics and facilitates high-throughput determination of genome-wide bacterial fitness in mixed communities.

60. Fierer N: **Embracing the unknown: disentangling the complexities of the soil microbiome**. *Nat Rev Microbiol* 2017, **15**:579-590.

61. Delogu F, Kunath BJ, Evans PN, Arntzen MØ, Hvidsten TR, Pope PB: **Integration of absolute multi-omics reveals dynamic protein-to-RNA ratios and metabolic interplay within mixed-domain microbiomes**. *Nat Commun* 2020, **11**:4708.