frontiers | Frontiers in Bioinformatics

TYPE Review
PUBLISHED 04 February 2026
DOI 10.3389/fbinf.2025.1721028

OPEN ACCESS

EDITED BY
Bruno César Feltes,
Federal University of Rio Grande do Sul, Brazil

REVIEWED BY
Maria Suarez-Diez,
Wageningen University and Research,
Netherlands
Jaspreet Kaur,
University of Delhi, India

*CORRESPONDENCE
Andrew J. Maritan,
✉ amaritan@mpi-bremen.de
Frank J. Stewart,
✉ frank.stewart@montana.edu

RECEIVED 08 October 2025
REVISED 16 December 2025
ACCEPTED 24 December 2025
PUBLISHED 04 February 2026

CITATION
Maritan AJ and Stewart FJ (2026) Essential nucleic acid omics: a theoretical foundation for early-stage users.
Front. Bioinform. 5:1721028.
doi: 10.3389/fbinf.2025.1721028

Essential nucleic acid omics: a theoretical foundation for early-stage users

Andrew J. Maritan[1,2]* and Frank J. Stewart[1,3]*

[1]Montana State University, Department of Microbiology & Cell Biology, Bozeman, MT, United States, [2]Max Planck Institute for Marine Microbiology, Bremen, Germany, [3]Georgia Institute of Technology, School of Biological Sciences, Center for Microbial Dynamics and Infection, Atlanta, GA, United States

Modern biology often relies on the analysis of entire sets of molecules (omics). A subset of omics uses nucleic acid sequencing to reconstruct genomes and profile gene expression. Novel findings and existing data are contextualized by databases, which have been growing exponentially due to falling sequencing costs and increased computing access. The increasing accessibility of omics has led to rapid adoption and widespread self-training via open-access tools. In this training environment new users (many of whom are students also applying computing for the first time) are confronted with Terabytes of sequence data and an ocean of topic-specific computing guides (often directed at high-level users). This flood of information creates an initial barrier of confusion and frustration, where it is challenging to identify the overarching goals of omics analyses through the details of computing. We believe this confusion is understandable but not pre-destined, as omics is–at its core–simple. This simplicity comes from its modular nature, where any analysis requires familiarity with only a few consistent steps. Here, we identify core elements of all omics analyses–data products, tools, and workflows–using microbiology applications to ground the discussion. This structure is informed by first-hand experience training early-stage omics users, where covering omics theory provides a foundation for practical implementation.

KEYWORDS

beginner, early career, FAIR, guide, ISA, MAGS, metabarcoding, pipeline

1 Introduction

Analyzing nucleic acid sequences (omics) is a universal tool in contemporary biology. In this world, new biologists benefit from understanding the foundational motivations and methods of omics analyses whether or not they intend to apply these tools themselves. In our experience teaching omics to students for both future use and background context, we see that most are competent biologists lacking computing experience. For these students, the technical details of computing often obscure the fact that omics analyses are simple arrangements of a few modular tools, producing a few consistent outputs. To highlight the simplicity of omics, we focused this review on broadly applicable theory to provide a view of the omics "bigger picture". Further, to avoid distracting from this perspective, we limited our discussion of the technical details of computing, which is available in other excellent guides as needed (see "Section 5.1.1 Opportunities for Further Training").

We organized this review in four parts, providing an increasingly granular understanding of nucleic acid omics. Part 1 (Section 2) describes the biological goals of different types of omics by discussing the history of their development. Part 2 (Section 3) identifies the analytical goals of omics by identifying the core data products. Part 3 (Section 4) describes the repeated modular steps of omics analyses that are used to generate core data products. Part 4 (Section 5) concludes with computational and non-computational tips for new users. The review's structure ensures that a student can read it completely for a top-to-bottom guide to nucleic acid omics, or individual sections to clarify specific questions. For some readers, this review will be sufficient to understand the "methods" sections of manuscripts, while others will want to continue with more specific training to run omics analyses independently. For both groups, this review should make nucleic acid omics more tractable, providing a foundation for engaging more deeply with omics literature and/or code.

2 The development of omics: a short history

2.1 What is omics?

The term omics describes analyzing a system (single cells, organs, organisms, or communities of organisms) using its biological molecules (DNA, RNA, proteins, metabolites). The biological molecule in-question determines the name of the omics analysis (DNA: gen-omics, RNA: transcript-omics, proteins: prote-omics, metabolites: metabol-omics), while the system's scope determines the prefix ("meta-" applies to community studies: meta-genomics, meta-transcriptomics, while no prefix is applied to single-species studies). In this review, we will focus on nucleic acid-based omics techniques, using the terms "genomics" and "transcriptomics" for consistency, though the topics are applicable to community scale meta-omics (Box 1). Nucleic acid sequences are the foundation of omics as they (especially genomes) provide the near-complete repertoire of life's function-encoding units (protein coding genes and transcripts, rRNA, tRNA, …). Many of these coding units are conserved across the domains of life allowing researchers to assign likely function and taxonomic identity to newly acquired sequences by comparing them to ever-expanding reference databases. These modern technical capacities were developed over decades, and understanding this history–especially past limitations–is essential for adding new data into a historic literature (Brock, 1999). To that aim, we will cover a brief history of the development of modern nucleic acid omics.

2.1.1 Nucleic acids and the central dogma

The groundwork for modern omics was laid by identifying DNA as the molecule of genetic inheritance (Avery et al., 1944), description of DNA's structure (Watson and Crick, 1953), and the articulation (Crick, 1958) and experimental support (Gros et al., 1961; Brenner et al., 1961) of the Central Dogma of molecular biology: genetic information stored in DNA, transmitted by RNA, and manifested as proteins. This biochemical linkage means that the study of any of these molecules informs the understanding of their precursors or derivatives.

2.1.2 Marker genes

Informed by the Central Dogma, particular gene sequences (DNA and RNA) proved to be especially predictive (e.g., phenotype, evolution, heredity, behavior) and are termed "marker genes" (Box 1). Study of marker gene distribution and variance to understand biological phenomena was the direct precursor to omics approaches (marker gene analyses are not always considered "omics" because they do not capture "entire subsets of molecules", though we discuss marker genes throughout this review as they are related to other DNA and RNA-based approaches).

A landmark example of marker gene analyses used the conserved and abundant ribosomal RNA (rRNA; Figure 1A) to study evolution across the tree of life. Using rRNA digestion fragmentation patterns on gels, Archaea were discovered in 1977, upending conceptions of the origin of Eukaryotes (Woese and Fox, 1977). Purified rRNA remained a popular molecule for reconstructing phylogenies, with methods developed to sequence it directly (Stahl et al., 1984; Lane et al., 1985; Figure 1A; Box 1). In parallel, methods developed to multiply and sequence the less abundant DNA fraction, opening the possibility of examining other marker genes, though rRNA remained a popular target to study evolution. Initial sequencing of DNA marker genes required cloning target genes into viral vectors–notably used to place the bacterial origin of the mitochondrion (Yang et al., 1985) – a labor and resource intensive effort. Acquiring enough DNA to sequence marker genes was greatly simplified by the invention of PCR in 1985 (Saiki et al., 1985), allowing near-direct sequencing of low abundance DNA encoded genes. This method was soon applied to study rRNA genes in mixed microbial communities in 1990 revealing previously unknown diversity (Giovannoni et al., 1990).

Beyond the use of rRNA to assign taxonomy and reconstruct evolutionary lineages, marker genes can also be used to screen organisms for pre-selected functions ranging from diagnosing sickle-cell anemia in humans (Kan and Dozy, 1978) to the identification of nitrogen fixing bacteria in the ocean (Tschitschko et al., 2024).

Marker gene analyses upended evolutionary biology, radically increased our catalogue of biodiversity, and made it possible to understand phenotypes without direct observation. Marker gene analyses continue to be useful in studying genes already identified as important, though they only provide "snapshots of organisms" (Pace, 1997). Organisms are constructed from hundreds to thousands of genes (Hou and Lin, 2009), and a single marker gene only explains a tiny percentage of any organism's genetic potential. Genome sequencing is required to understand co-occurring genes in a single organism and is the domain of genomics.

2.1.3 Genomics

Genomics (Box 1) provides greater coding context than single marker genes and emerged with the publication of the first genome in 1977 (the virus phiX; Sanger et al., 1977a; Box 1) and ushered in the age of analyzing large collections of genes (Figure 1B). This first viral genome was small (5,386 bases) but was followed up with genomes from bacteria (*Haemophilus influenzae*, ~1.83 Mbp; Fleischmann et al., 1995) and eventually humans (~6.3 Gbp, diploid; (International Human Genome Sequencing Consortium et al., 2001; Venter et al., 2001; Figure 1B). Larger genomes were increasingly complex, but all possessed the conceptual simplicity that all sequences originated from a single organism. The concepts of

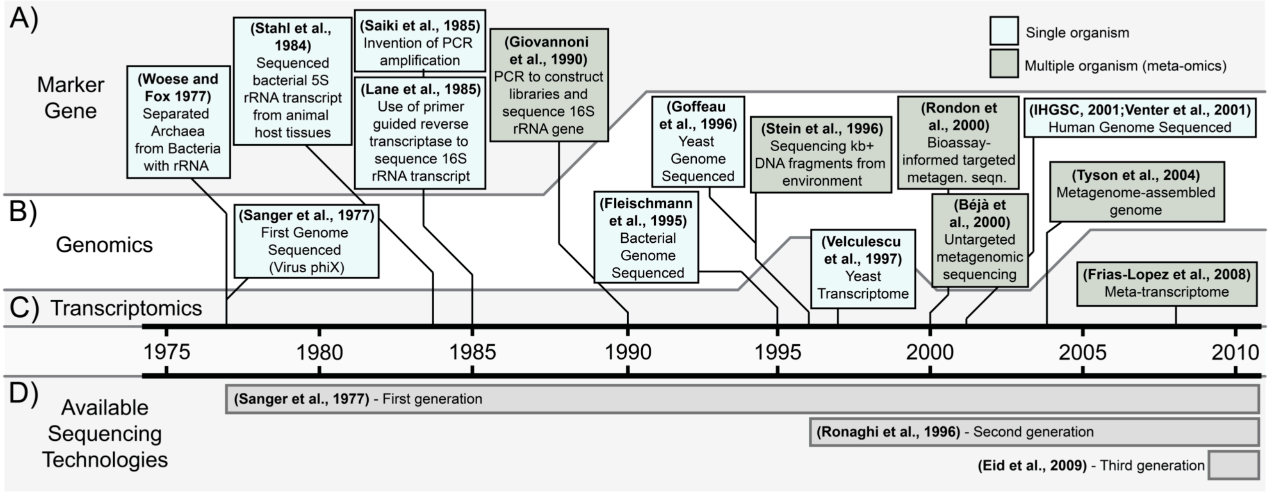
A)
Marker Gene
B)
Genomics
C) Transcriptomics
D)
Available Sequencing Technologies
□ Single organism
▨ Multiple organism (meta-omics)
(Woese and Fox 1977) Separated Archaea from Bacteria with rRNA
(Stahl et al., 1984) Sequenced bacterial 5S rRNA transcript from animal host tissues
(Saiki et al., 1985) Invention of PCR amplification
(Lane et al., 1985) Use of primer guided reverse transcriptase to sequence 16S rRNA transcript
(Giovannoni et al., 1990) PCR to construct libraries and sequence 16S rRNA gene
(Goffeau et al., 1996) Yeast Genome Sequenced
(Stein et al., 1996) Sequencing kb+ DNA fragments from environment
(Rondon et al., 2000) Bioassay-informed targeted metagen. seqn.
(IHGSC, 2001;Venter et al., 2001) Human Genome Sequenced
(Tyson et al., 2004) Metagenome-assembled genome
(Sanger et al., 1977) First Genome Sequenced (Virus phiX)
(Fleischmann et al., 1995) Bacterial Genome Sequenced
(Velculescu et al., 1997) Yeast Transcriptome
(Béjà et al., 2000) Untargeted metagenomic sequencing
(Frias-Lopez et al., 2008) Meta-transcriptome
1975
1980
1985
1990
1995
2000
2005
2010
(Sanger et al., 1977) - First generation
(Ronaghi et al., 1996) - Second generation
(Eid et al., 2009) - Third generation

FIGURE 1
Illustrative timeline of the development of nucleic acid omics. Included are the first examples of technologies that enabled–or are–modern omics between 1975 and 2010, covering **(A)** marker gene sequencing, **(B)** genomics, **(C)** transcriptomics, and **(D)** the availability of different sequencing technologies–shaded from their announcement to the end of the timeline. IHGSC = International Human Genome Sequencing Consortium.

single organism sequencing were quickly applied to sequencing genomic material from complex microbial communities. Multi-species microbiomes were sequenced by first inserting large (50+ kbp) fragments of DNA into *Escherichia coli* and later sequencing these clone libraries either randomly to survey the community (Béjà et al., 2000), or deliberately to capture specific taxa (Stein et al., 1996) or functions (Rondon et al., 2000). This sequencing of genomic DNA from multi-species communities were the first metagenomes, expanding the knowledge of protein-coding gene diversity and the environmental distribution of metabolic functions and taxa (DeLong et al., 2006; Yooseph et al., 2007). Further, genes encoding taxonomy and function often co-occurred on a single large fragment (Stein et al., 1996; Béjà et al., 2000) allowing researchers to describe an organism solely through molecular data–first identifying it (see marker gene above) and then hypothesizing its "functional potential". This work linking taxonomy-to-function in metagenomes advanced when metagenomic sequences were used to reconstruct individual complete (or near-complete) microbial genomes (Tyson et al., 2004). These metagenome-assembled genome (MAG; Parks et al., 2017; Yutin et al., 2021; Box 1) and non-MAG (Dragone et al., 2022; Bertagnolli et al., 2023) approaches are now widely applied on diverse uncultured microorganisms to understand their taxonomies and functional potential. The distinction of "functional potential" is essential, as genomes provide evidence that a function might be performed, but do not demonstrate activity (Hatzenpichler et al., 2020). Activity can be more accurately approximated by studying gene expression–a sort of "metabolic intention" – which is the domain of transcriptomics (Box 1).

2.1.4 Transcriptomics

To understand "metabolic intention" the sequencing methods of DNA pools were applied to RNA (following reverse transcription of RNA to cDNA), creating the field of transcriptomics. Early sequencing of untargeted RNA provided initial insights into the diversity of expressed genes in different cell types, requiring the cloning of individual cDNA transcripts into *E. coli* clones (Adams et al., 1991). Accurately quantifying RNA expression–allowing rigorous comparison between cell types–became possible with the advent of microarrays. There, cDNA from thousands of pre-selected gene targets were attached to glass slides and then hybridized with experimentally sourced (extracted and reverse transcribed) cDNA, creating fluorescence proportional to the sample cDNA, allowing quantification (Schena et al., 1995). Derivatives of these technologies are still in-use today and set the stage for "transcriptomics" where RNA sequencing was used in-concert with existing genomes to identify expressed genomic regions (Velculescu et al., 1997), which began using "Serial Analysis of Gene Expression" (SAGE). In SAGE-based transcriptomics, cDNA was sequenced by cleaving each cDNA transcript into a short tag (9–11 bp), the tags concatenated into a longer sequence, cloned into *E. coli*, PCR amplified, and sequenced. These tags were then extracted bioinformatically, aligned against a reference genome, where alignment (Box 1) of an RNA tag to a DNA sequence indicated gene expression and the number of tags aligning to any DNA sequence used to quantify expression (Velculescu et al., 1995; Velculescu et al., 1997). SAGE transcriptomics was first used in 1997 on yeast cultures with RNA tags aligned to the new yeast genome (Goffeau et al., 1996) to generate maps of thousands of expressed genes (Velculescu et al., 1997; Figure 1C). Advances in sequencing (RNA-seq; Bainbridge et al., 2006; Nagalakshmi et al., 2008), resulted in more and longer RNA sequences (beyond SAGE's 10s of bp tag approach) increasing its sensitivity (identifying splicing and lowly expressed genes), read coverage, and data volume. As transcriptomics advanced, it was applied to mixed-species microbiomes, revealing gene expression by dominant (Frias-Lopez et al., 2008; Hewson et al., 2010; Stewart et al., 2012) and less abundant taxa (Stewart et al., 2012) in the environment.

2.1.5 Perspectives

In the last century, biologists have learned that DNA is the molecule of trait inheritance and can now measure gene expression from nanogram quantities of RNA in the wild. These new approaches have enabled sequence-based investigations of diversity and function across earth (Sunagawa et al., 2020; Nayfach et al., 2021; Shaffer et al., 2022) and into space (International Space Station; Castro-Wallace et al., 2017), surpassing the dreams of early omics scientists (Pace, 1997).

These global surveys of commonplace (seawater, soils, human skin) and extreme (hot springs, alkaline lakes) environments were essential to fill the complete vacuum of information about the diversity and distribution of uncultured microorganisms. However, the number of unexplored ecosystems shrinks daily, and as-such, modern microbiologists should not expect that sequence-based surveys will provide the acclaim of the early days of sequencing.

Today's omics researchers should follow the example of early scientists to answer specific biological questions (reconstructing the tree of life, Woese and Fox, 1977; reconstructing the evolution of symbionts; Lane et al., 1985; Yang et al., 1985) using available tools. The basic toolkit of omics is well established (at least since 2008; Figure 1), but advances in rapid cheap sequencing and ~50 years of archived sequencing data have produced opportunities to answer new biological questions with global samples (The Earth Microbiome Project, Thompson et al., 2017; TARA Oceans; Sunagawa et al., 2020) and replication across space and time (the National Science Foundation's: National Ecological Observatory Network, Dantzer et al., 2023; and Long Term Ecological Research Network; Knapp et al., 2012). Zooming-in, omics now has the capacity to sequence the genomes (Raghunathan et al., 2005; Woyke et al., 2009) and transcriptomes (Ma et al., 2023) of single cells; part of a broader interest in understanding heterogeneity between single cells (Hatzenpichler et al., 2020; Kitzinger et al., 2020; Marlow et al., 2020). Microbiology's newfound acquisition of spatially resolved (micron to global) and longitudinal (decades) sequencing data is one exciting new frontier for omics research (Eren and Banfield, 2024).

2.2 Nucleic acid sequence analyses are everywhere

Omics use has grown exponentially since its inception (Gauthier et al., 2019). One indicator of omics use is the rate of sequence deposits into reference databases. The NCBI Sequence Read Archive (the major public repository for unprocessed sequence data globally) has added 25.6 Petabase pairs (2.56×10^{16} base pairs - the data equivalent of ~6,500,000 human genomes; Nurk et al., 2022) from 2012 to 2021 (Katz et al., 2022). The NCBI GenBank (a repository for assembled sequence data) has doubled in size every ~2 years from 2013 to 2024, to a total of 3.4×10^{13} bp (Sayers et al., 2025). The number of available reference genomes also indicate use, with the number of human genomes doubling every 7 months from 2001 to 2015 (Stephens et al., 2015). This exponential data production has been accompanied with a proportionate development of new bioinformatics approaches (Gauthier et al., 2019), with a conservative estimate of 25,000 unique bioinformatics tools produced between 1990 and 2017 (Clément et al., 2018). This flood of data and tools has created an application bottleneck, where many omics practitioners simplify analytical decisions by focusing on the straightforward aim of recovering genomes to describe the metabolism of focal taxa. This simplifying approach makes sense in-light of abundant data and tool options, but we believe that reducing complexity through genome-only analyses is unnecessary. The apparent complexity of nucleic acid omics is illusory, with all omics analyses built on a simple and consistent set of data products and methodologies. In this review we will distill diverse omics analyses–extending beyond genomes–into their shared data products, the classes of tools to generate them, and how these tools and data are strung together into workflows to answer biological questions. We will begin by describing the five core omics data products.

3 Omics data products: a few goals

Assuming the omics researcher has formulated a scientifically meaningful guiding question, the next step is to identify tractable computation goals: "Are we surveying functional and/or taxonomic content?", "Do we need to contextualize these data phylogenetically?", "Do we want genomes?", "Do we need to quantify or statistically test our findings?", etc. These procedural endpoints allow a bioinformatician to work backwards to construct an analytical workflow, identifying midpoint questions and target data products. These data products of omics (here: genomic, transcriptomic, and marker gene) fall into one of five classes: 1) sequence files, 2) sequence statistics, 3) taxonomy tables, 4) function tables, and 5) count tables. These data products are necessary to any omics analysis and must be incorporated into a larger biological narrative to be useful. With that in mind, we describe the general structure and uses for each of these data classes (summarized in Figure 2).

3.1 Sequence files

Digitized biological sequences are the foundation of all omics (sequencing described below; "Section 4.1.1 Sequencing Technologies") and commonly follow the FASTA or FASTQ formats. The FASTA format contains both a sequence identifier and sequence data (nucleic acid residues; Figure 2A) while FASTQ contains the same information as FASTA files as well as quality scores for each nucleic acid residue (Q-scores). These quality scores allow the user to remove low-confidence sequences and/or bases (which then produces quality filtered FASTA files) before further analysis (Figure 2A). There are three major types of sequence files: 1) reads, 2) assembled contiguous sequences (contigs; Box 1), and 3) genomes. Each of these classes are used to generate the next (reads are used to make contigs, reads and contigs to make genomes) resulting in increasingly long context-rich sequences. We will now describe the characteristics of each of these sequence classes.

3.1.1 Reads

Reads are the raw product of the sequencing platform and the basis of marker gene (16S rRNA gene), genomic, and transcriptomic studies. Reads are classified as "short" or "long" depending on the sequencing technology used to generate them and length of output reads. Short reads are tens to hundreds of bases long, while it is possible for long reads to be thousands to millions of bases (Satam et al., 2023). These reads contain all the (relative) abundance information in a sequence

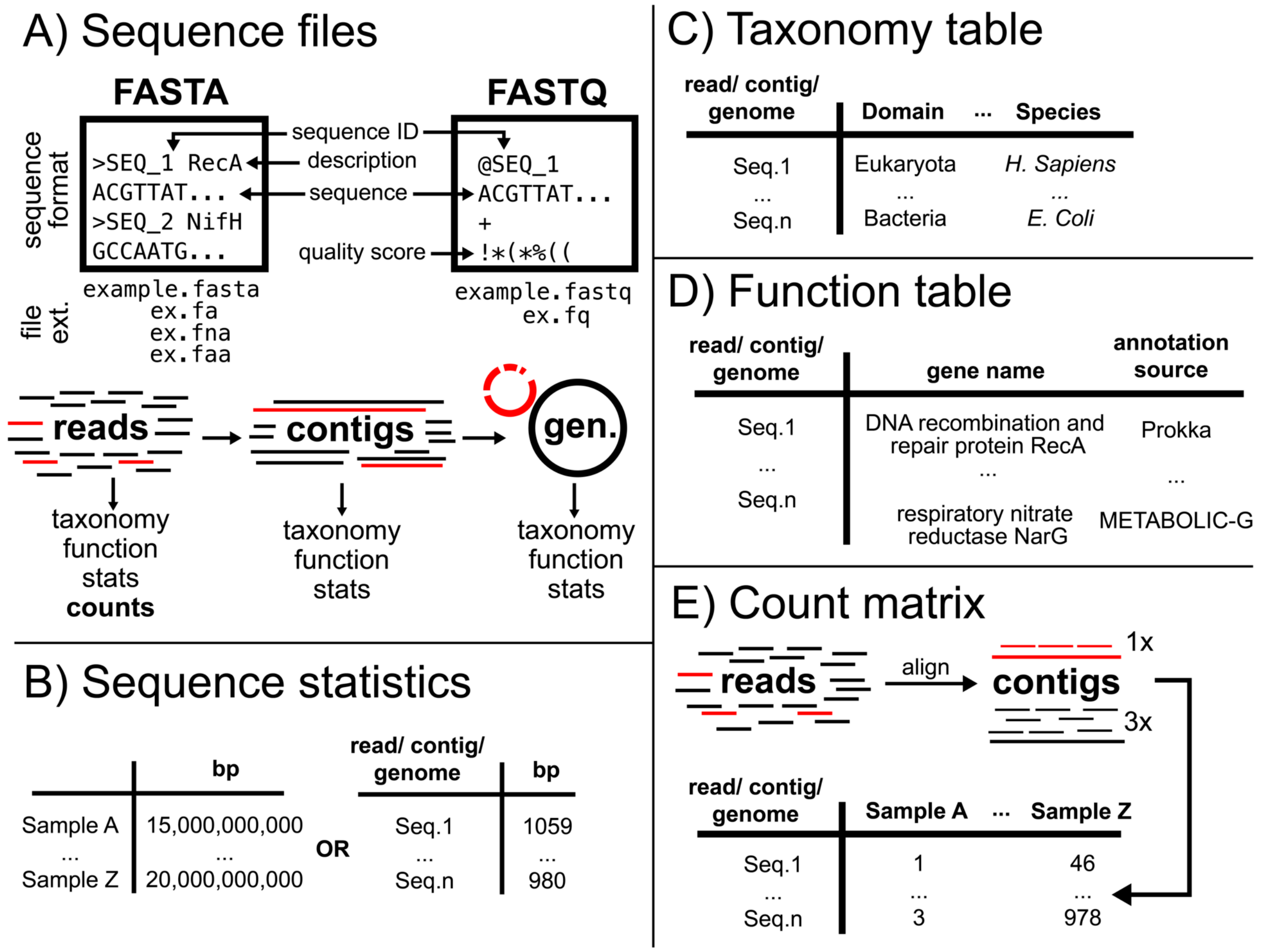

FIGURE 2
All nucleic acid sequence-based omics analyses rely on five core data products. Simplified examples of the core data products. **(A)** Sequence files (two common formats depicted) are used to encode reads, contigs, and genomes. They contain nucleic acid bases and some supporting metadata. Sequence files are used to derive all other data products. **(B)** Sequence statistics are used to describe the contents of a sequence file (often total number of sequences/base pairs or sequence lengths). **(C)** Taxonomy tables link sequences to likely source taxa. **(D)** Function tables link sequences to their likely functions. **(E)** Count tables are generated using read data and contain the estimated abundances of each sequence. Bp = base pairs; cond. = condition; ex. = example; ext. = extension; gen. = genome; seq. = sequence; stats = statistics.

library (Gloor et al., 2017; Box 1), with derivative sequences (contigs and genomes) requiring reads for quantification.

Reads are the functional unit for marker gene studies, using fragments to full-length genes to identify microbial taxa (16S rRNA, *rpoB*; Thompson et al., 2017) or putative functions (*pmoA*, *narG*; Yu et al., 2024). In genomics and transcriptomics, read data are generally treated as a steppingstone to assembling contigs and recovering genomes. However, analysis of unassembled reads can be valuable as it uses the maximum amount of available data and therefore provides a relatively unbiased representation of microbiome gene content (Hauptfeld et al., 2024). Assuming individual reads are of a length sufficient for confident identification of homologous sequences (homologs) in a reference database, unassembled read datasets can be searched to identify taxonomically (Meier et al., 2017; Dragone et al., 2022) and functionally (Ortiz et al., 2021; Täumer et al., 2022; Maritan et al., 2025) informative marker genes. Read-based approaches can also be used to sift through reference databases to identify only the datasets that include metabolisms or taxa of interest (Speth and Orphan, 2018). Though read-based analyses have (often untapped) potential, the most common use of read data is the reconstruction of contigs, which we discuss next.

3.1.2 Contigs

Contigs are generated by assembling reads into longer nucleic acid sequences. This approach is used in genomics to create genomic scaffolds (Prjibelski et al., 2020), and assembly-based transcriptomics to generate transcripts (Grabherr et al., 2011). Assembled contigs are, by definition, a subset of total sequencing effort, as not all reads can be placed into a contig (Hauptfeld et al., 2024). Despite data loss in assembly, contig sequences are useful for community taxonomic and functional reconstruction because their length enables more accurate identification of homologs compared to reads. If an assembled contig contains multiple protein coding sequences (genes of the same operon), this 'genomic neighborhood' (Wei et al., 2024) can be used to increase confidence in assigning gene function (Mihelčić et al., 2019) or taxonomy (Mirdita et al., 2021). Contig-derived genes are also potential inputs for phylogenetic reconstruction, enabling contig-based evolutionary and taxonomic analysis of a microbiome. Though analysis of standalone contigs is useful, the most common use of contigs is to reconstruct genomes.

3.1.3 Genomes

Genomes are made by grouping (i.e., binning) contigs with similar features (details discussed below) into a single sequence file. In this review, we use the term "genome" to discuss a collection of sequences that likely come from the same organism, encompassing genomes recovered from pure cultures and mixed species consortia (termed: metagenome-assembled genomes; MAGs). Genome binning, like contig assembly, results in data loss (Hauptfeld et al., 2024) only examining a subset of the total microbiome. Despite this, genome-based analyses are appealing because they create a meaningful association among contigs, by which taxonomy or functional potential assigned to any contigs is passed onto all other contigs in the genome. This analysis allows a researcher to characterize the metabolic potential of individual microbes (Kohtz et al., 2024) and communities (Shoemaker et al., 2024; Ricci et al., 2025), even if the organisms containing these genomes have never previously been observed (Evans et al., 2015; Wurch et al., 2016). Genomes can also be used as references for aligning transcriptome sequences recovered from the same environment, thereby identifying expression patterns in individuals or communities across environmental gradients (Kitzinger et al., 2020; Porras et al., 2024). Beyond community description and reconstruction, genomes (Tripp et al., 2008) and transcriptomes (Bomar et al., 2011) can also be used to optimize cell culture (Wurch et al., 2016). The methods for generating and processing each of these types of sequence data are discussed in greater detail below.

3.2 Sequence statistics

Sequence statistics are derived from sequence files and have two major purposes: 1) contextualizing a narrative (describing dataset size/complexity, sampling effort, and/or similarity between sequences) and 2) normalizing count data. Viewing and analyzing these statistics typically involves generating tables of total bases in each sequence library or of individual sequences (reads, contigs, or genomes; Figure 2B). The methods for generating sequence statistics are discussed in detail below ("Section 4.2 Sequence Statistics").

3.3 Taxonomy tables

Taxonomic classification (Box 1) aims to generate tables that relate a sequence identifier to a taxonomic lineage (Figure 2C). Taxonomic lineage is assigned to sequence data (reads, contigs, or genomes) by comparing unknown query sequences against reference sequences with known taxonomic origin. If the query sequence is sufficiently similar to a reference, the query is assigned the taxonomy of the reference (Goris et al., 2007; Jain et al., 2018; Parks et al., 2020). The methods for generating taxonomy tables are discussed in detail below ("Section 4.8 Taxonomic Classification").

3.4 Function tables

Function annotation aims to generate tables that relate a sequence identifier to a descriptor of putative cellular function, often relating to metabolism, physiology, or behavior (Figure 2D; Box 1). Functional annotation of sequence data (reads, contigs, or genomes) compares an unknown query sequence against annotated (and potentially experimentally validated, although this is not always possible) reference sequences. There are at least two important caveats regarding using and interpreting functional annotations. First, the quality of any annotation is tied to the completeness and annotation accuracy of the reference database. Sequences from well represented model organisms (*E. coli*, *Pseudomonas* sp., etc.) and their close relatives can typically be annotated with high confidence, while genes in non-model organisms will be less confidently annotated (Goodacre et al., 2014). Second, while functions identified in genomic data indicate metabolic potential and functions identified in transcriptomic data indicate gene expression, neither genomic nor transcriptomic evidence of putative function proves that amino acids were translated or their proteins were active. The methods for generating function tables are discussed in detail below ("Section 4.9 Function Annotation").

3.5 Count matrices

Sequence quantification (reads, contigs, or genomes) aims to generate tables that relate a sequence identifier to an estimate of its relative abundance in a sample, thereby providing a loose indication of a gene or organism's biological significance (Figure 2E). Read quantification often involves simple counting, while quantifying longer sequences (contigs and genomes) requires aligning the source reads to the longer sequences (Aroney et al., 2025). Counts can be used as a descriptor of community composition (Bollati et al., 2024), to test hypotheses of differences in abundance of functional potential or taxa (Maritan et al., 2025), to identify associations between taxa and environment (Mitchell et al., 2024), or as input for quantitative modeling (Louca et al., 2016). The methods for generating count matrices are discussed in detail below ("Section 4.10 Count Data").

3.6 Putting it all together: merging and using omics data

These data products are often the midpoint and endpoint goals of an omics workflow. Once data tables are generated (if all samples, metadata, and sequences have consistent naming), they can be merged into a "master table" for downstream filtering, plotting, phylogenetic inference, statistical tests, or other direct comparisons. However, merging tables without a specific goal may not be useful as it can create unwieldy tables with millions of columns or rows.

By clearly describing the core data products of all omics, we hope to make the endeavor less abstract. Thus far we have covered *what* is generated from omics (the five core data products), below we address *how* omics is executed via specific tools and workflows.

4 The omics toolkit: descriptions of common approaches, their purposes, and connections

The toolkit of nucleic acid omics involves extraction and sequencing of nucleic acids with subsequent processing of generated sequences to make the data products outlined in Section 3. We now explore the methods available to do this, with each major computational step summarized in Figure 3.

4.1 Acquiring sequences

Sequencing is the basis of omics analyses with sequences generated *de novo* or downloaded from public databases. In either case, the quality and utility of any sequence dataset is underpinned by the quantity and length of output reads and confidence in the constituent bases–more and longer reads, with high confidence bases are markers of quality. These qualities are largely determined by the choice of nucleic acid extraction and sequencing technology.

4.1.1 Sequencing technologies

Nucleic acid sequencing has had three major technological generations, each of which are still in-use and have pros and cons (reviewed in; Cheng et al., 2023; Satam et al., 2023). First generation sequencing is often referred to as "sequencing by termination" or "Sanger sequencing" after Frederick Sanger, its inventor and publisher of the first genome (Sanger et al., 1977a; 1977b; Figure 1D). This technology sequences one DNA molecule at a time, producing long sequences with low error rates (Cheng et al., 2023), and was used to achieve other genome "firsts" (bacterial, Fleischmann et al., 1995; yeast; Goffeau et al., 1996; human; International Human Genome Sequencing Consortium et al., 2001; Venter et al., 2001; Figure 1B). Today, Sanger sequencing is still in wide use: cheaply characterizing PCR amplicons from pure cultures and cloned genes, or in sequencing across gaps between contigs in draft genome assemblies (Drevinek et al., 2023; Katara et al., 2024). However, it is inefficient for processing dozens to hundreds of samples simultaneously (Panahi et al., 2024) – a need for efficient microbiome surveys–solved by later generations of sequencing.

Second generation sequencing is also known as "next-generation" or "short-read" sequencing" (Figure 1D) and is largely synonymous with the most prominent producer of short-read sequencers: "Illumina" (though expiring patents and new competitors are driving innovation and price reductions; Eisenstein, 2023; De Ronne et al., 2025). Short read sequencing generally entails spatially separating DNA fragments and observing the synthesis of bases (via fluorescence or pH change), producing short (25–300 bp) reads (Cheng et al., 2023; Satam et al., 2023). These reads can be analyzed with minimal processing (16S rRNA marker gene sequencing and transcriptomics) or assembled into contigs and binned into genomes. Genomes can be challenging to complete using short read data, because complex genomic regions are often longer than the technology's maximum read length (~500 bp; Satam et al., 2023; Panahi et al., 2024), preventing their reconstruction (Mise and Iwasaki, 2022). The read length limitation has been addressed by third generation sequencing.

Third generation sequencing is also known as "long-read sequencing" or by the trade names of the most prominent producers of long-read sequencers: Oxford Nanopore Technologies (ONT) or Pacific Biosciences (PacBio; Figure 1D). As the name suggests, long-read sequencing generates longer reads than second generation (10 kb+; Satam et al., 2023) which allows each read to capture greater genomic context (e.g., full length 16S rRNA genes, near-complete genomes). This technology passes (near-) full length nucleic acid molecules through a fixed sequencing unit (conductive pore or modified DNA polymerase), recording bases as they pass through. This technology can have higher error rates than short-read sequencing (Panahi et al., 2024), though it is possible to combine the higher quality short reads and the greater genomic context of long-reads to create long, high quality contigs (Antipov et al., 2016; see "Section 4.3.3 Contigs" below). Additionally, it is important to note that the name "long-read sequencing" indicates only a technological capacity–not a guarantee–to produce long reads. Sequenced read length depends on the length of the nucleic acids provided to the sequencer, which in-turn depends on minimally fragmenting nucleic acids during extraction, which we will discuss below.

4.1.2 Nucleic acid extraction and sequencing

Generating new sequence data proceeds via two steps: 1) nucleic acid extraction and 2) sequencing, where the intended sequencing technology should inform extraction method. All nucleic acid extractions aim to lyse cells, expose nucleic acids, remove non-nucleic acid lysate, and collect enough nucleic acids to sequence anything. Though all these steps are important, the method of initial lysis largely determines sequencer compatibility.

Long-read sequencing requires minimally fragmented (high molecular weight) nucleic acids to produce long reads, whereas short read sequencing is less sensitive to fragmentation (Table 1; Zhang et al., 2022). For this reason, extraction for long-read sequencing should use "gentler" chemical lysis (detergents: SDS, solvents: Phenol-chloroform, TRIzol, or enzymes: lysozyme; Trigodet et al., 2022), while short read sequencing can combine chemical and mechanical (bead beating, freeze-thaw; Hamilton et al., 2011) lysis to maximize nucleic acid yields.

4.1.3 Data mining

Sample collection, extraction, and sequencing are all costly and can be reduced by using publicly available sequence datasets (NCBI SRA, Leinonen et al., 2011; NCBI nt/nr Sayers et al., 2024; EMBL UniProt; The UniProt Consortium et al., 2025). These datasets, combined with robust questions can be impactful (see "The Parasite Awards", awarded "for rigorous secondary analysis of data"; https://researchparasite.com/). As examples, Kumagai et al. (2018) leveraged both public and newly sequenced genomes to explain the distribution of light harvesting proteins in marine bacteria.

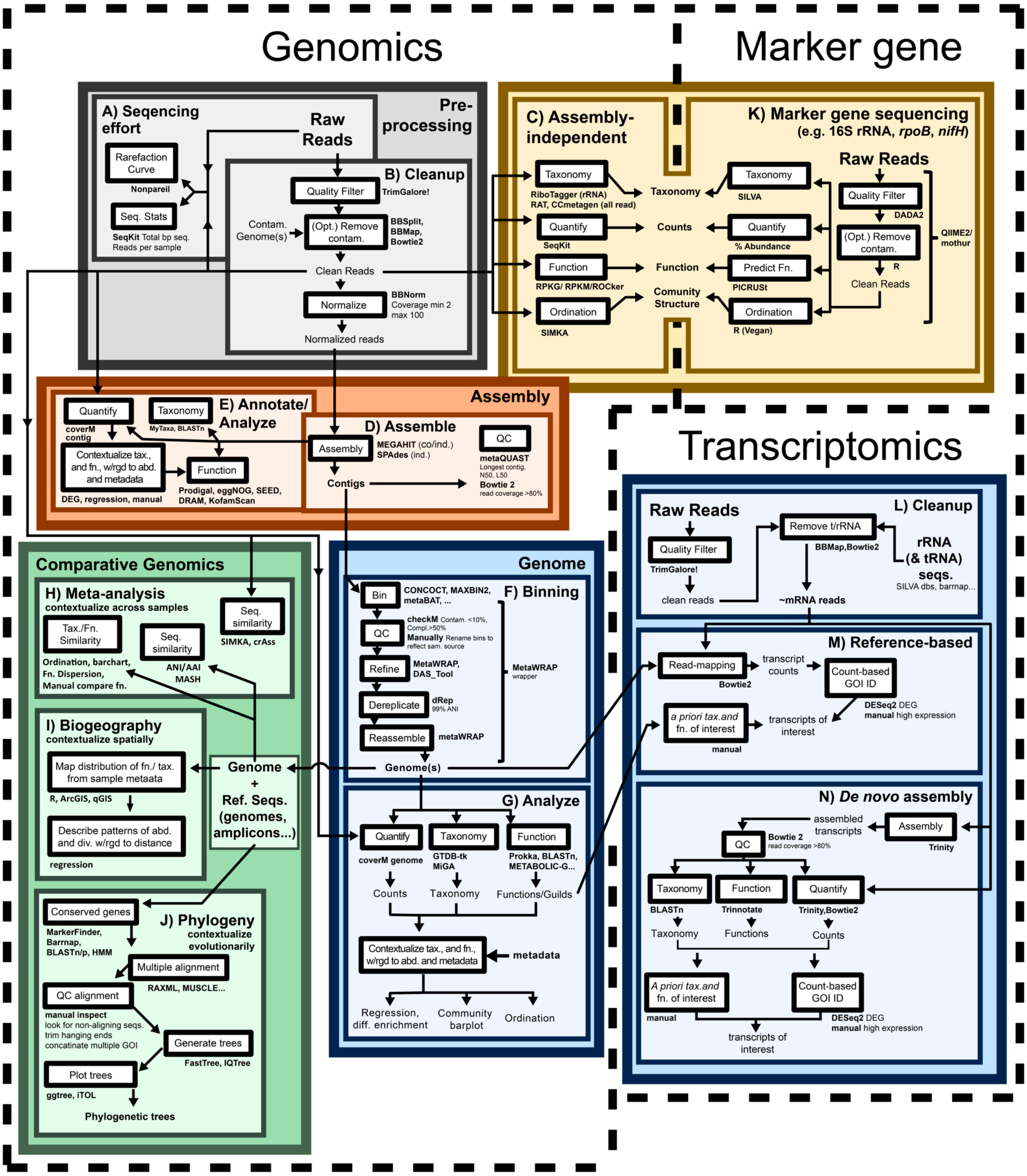

FIGURE 3
Nucleic acid sequence-based omics analyses are modular and complementary. Generalized workflow for performing marker gene, genomics, and transcriptomics analyses, with each demarcated by dashed lines. Within a single approach major processing steps are grouped within colored boxes, with inset, lighter boxes indicating subsidiary tasks. Discrete processing steps are named on white boxes (with example tools for performing the task alongside), colored boxes are used for clarity, but do not indicate importance. Data inputs and outputs are connected to processing steps with arrows. All pipelines begin with "Raw Reads" at the top of their respective approach. Marker gene: Marker gene analyses (from targeted amplification and untargeted genomics/transcriptomics) begin with quality filtering **(B,K)** and are then used to immediately generate counts and predict taxonomy and function **(C,K)**. Genomics: All genomic analyses generally begin by quantifying sequencing effort and calculating read statistics **(A)** and quality filtering reads **(B)**. Contig-based analyses assemble reads into contigs to generate count, taxonomy, and function tables **(D)**. Genome-based analyses use these contigs to generate genomes for subsequent generation of count, taxonomy, and function tables **(F,G)**. All genomic analyses are well suited for comparison against existing sequence databases **(H-J)**. Transcriptomics: Transcriptomic analyses begin by quality filtering reads **(L)**. There is then a split where some transcriptomics use genomics as a reference (reference-based; **(M)**) while others proceed independently and assemble RNA contigs (transcripts) requiring functional annotation and taxonomic classification of each transcript (*de novo* assembly; **(N)**) Both reference-based and *de novo*

(*Continued*)

FIGURE 3 (Continued)
transcriptomics quantify expression by aligning RNA reads against the longer reference (genome or assembled transcripts) with the transcript aligned to determining taxonomy and function, with the number of reads aligning indicating expression (count table; **(G,N)**) Abd. = abundance, asse. = assembly; compl. = completion; contam. = contamination; db = database; DEG = differentially expressed genes; diff. = differential; div. = diversity; fn. = function; GOI = gene of interest; ind. = individual; opt. = optional; QC = quality control; ref. = reference; sam. = sample; seq(s). = sequence(s); tax. = taxon/taxonomy; w/wgd. = with regard to.

TABLE 1 Summary characteristics of short read and long read sequencers. Highlighting differences in read lengths and nucleic acid extraction methods.

Features	Short read ("second generation")	Long read ("third generation")
Usual Read Lengths (bp)	100s	10,000–1,000,000s
Common Manufacturers	Illumina, Element Biosciences	Oxford Nanopore, Pacific Biosciences
Optimal Nucleic Acid Extraction Methods	Mechanical (bead beating, freeze thaw) and/or Chemical (detergents, enzymes)	Chemical (detergents, enzymes)

While Henriques et al. (2024) used publicly available vertebrate genomes to reconstruct the evolutionary trajectories of endogenous viral genes domesticated for host function in placental mammals. Though studies based on data mining are useful, it should be noted that papers centered around data mining are always limited by available resources. It should be noted that the use of others' data requires careful attribution of the datasets used (citations, accession numbers) and potentially the consent of those who generated the data. Best practices for using and sharing public data should always be followed and are described in publishing policies in academic journals, or in review papers (Sielemann et al., 2020; Hug et al., 2025).

4.1.4 Sequencing effort

For both new and mined read data, it is essential to consider sequencing effort. When sequencing genome(s), it is essential to sequence enough to capture the sequence diversity present in a sample. The relationship between sequencing effort and new information obtained follows a logarithmic relationship, where more sequencing recovers more and more novelty, until enough sequencing has been performed and novelty saturates. Identifying where a sample lies on the sequencing effort-to-novelty plot is a measure of sequence "coverage" which describes the fraction of the genome(s) represented by sequenced reads (Rodriguez-R and Konstantinidis, 2014). The number of reads needed to achieve high (>90%) coverage varies by system (Rodriguez-R et al., 2018), with larger genomes (human) and diverse microbiomes (sediments) requiring more sequencing effort than small genomes (phiX) or simple microbiomes (hot springs). A sample with suboptimal coverage can still be analyzed, but with the caveat that the analysis will be incomplete due to unidentified sequences.

4.1.5 Perspectives

Since its invention ~50 years ago, sequencing quality has improved while costs have decreased (Cheng et al., 2023; Satam et al., 2023). Sequencing will continue to evolve as existing technologies mature and new ones emerge, leaving technology selection a constantly evolving decision. Given the increased use of sequencing, public sequence data will likely continue to expand, a monumental resource to scientific discovery by secondary analyses. Now, after acquiring nucleic acid sequences, omics analysis can begin.

4.2 Sequence statistics

The first products of any omics analysis are generally sequence statistics, used for narrative or quantitative purposes (Figures 2B, 3A). For narrative purposes, read statistics are used to show that there is sufficient sampling to test a hypothesis (total bp sequenced per sample; Liu et al., 2015). For contigs, descriptive statistics are used to summarise assembly success: longest contig, total contig counts, N50, L50 metrics (Mikheenko et al., 2016). Genome statistics can indicate binning success: contamination in genomes (Bowers et al., 2017), genome size (Chklovski et al., 2023), and how representative the genomes are of a sampled community (percent of reads mapping to all genomes; Hauptfeld et al., 2024). These statistics should be generated any time that a sequence file is acquired or produced, with multiple tools available to streamline these calculations (reads: Nonpareil3, Rodriguez-R et al., 2018; contigs: QUAST; Mikheenko et al., 2016; genomes: BUSCO; Seppey et al., 2019; CheckM2; Chklovski et al., 2023). For quantitative purposes, sequence statistics are generally used to normalize count data against the length of the sequence and sequence library size, or to compare counts across or within datasets (simplified by efficient tools; SeqKit, Shen et al., 2024). For there to be any sequences to statistically summarise, we must first apply quality control standards.

4.3 Quality control

The quality of sequence files should be examined at the levels of reads (Figures 3B,K), contigs (Figures 3D,N) or genomes (Figure 3F) to increase confidence in any results.

4.3.1 Contaminant removal

Non-target sequences (contaminants) should be removed from a sequence library before downstream analysis (Figures 3B,K,L). Contaminating sequences can originate from several sources, including defined organisms expected to be in the sample but not the target of inquiry (human DNA sequences in an analysis of the human skin microbiome) or incidental organisms that should not be in the sample (plasmid or bacterial DNA in a reagent solution). Defined contaminants can be removed by aligning the new reads against a contaminant's genome or transcriptome and removing reads that align to the contaminant (Lataretu et al., 2025). Incidental contaminants are ideally detected by sequencing negative controls (where no sequences are expected) from various steps of sampling and sequencing preparation, with any recovered sequences representing potential contaminants (Fierer et al., 2025). These sequences can be classified as contaminants based on statistical probabilities (Davis et al., 2018) or taxonomy (based on taxa known to contaminate molecular biology reagents; De Goffau et al., 2018). In cases of limited contamination, it is recommended to remove (and report) potential contaminant sequences (Clum et al., 2021) while samples with rampant contamination may need to be discarded entirely (Fierer et al., 2025). These removed sequences may represent true biological signals, as knowledge of any biological system is often incomplete–with any decontamination balancing description of true novelty and cautious interpretation of data. Now we discuss standard quality control methods in read data.

4.3.2 Reads

Evaluations of read quality should consider whether sequences are of: 1) high quality and 2) sufficient quantity for the planned analyses.

Assessments of read quality should consider both sequence length and confidence in base assignment. A sequence substantially shorter than expectations (relative to sequencing technology) may indicate a poorly sequenced molecule and should be removed (Martin, 2011). Base confidence (in FASTQ sequence files) is encoded by the quality-score (Q-score), estimating the probability that a single base in a sequence is correctly assigned (A,C,G, or T), with higher Q-scores indicating higher confidence (O'Rawe et al., 2015). Quality filtering reads first trims sequences to remove low quality (user specified Q-score) bases, with the whole sequence discarded if trimming shortens it past a minimum length (TrimGalore! – https://github.com/FelixKrueger/TrimGalore), a process that should be performed before assessing sequence quantity or performing other omics analyses.

The necessary number of reads is dependent on the type of sample, with more complex microbiomes requiring more sequencing than simple ones (Rodriguez-R and Konstantinidis, 2014). Sampling sufficiency can be assessed using rarefaction analysis, where cleaned reads are randomly subsampled, a metric of novelty calculated at each increment, and then plotted against one another (sequence diversity vs. read number; Rodriguez-R et al., 2018). If novelty saturates (asymptotes) with increased read number, most of the sequence diversity was captured, whereas a linear relationship–without saturating–indicates unsequenced diversity. Unrepresentative samples can be resolved with more sequencing, but if this is not possible, such samples can still provide useful–though caveated–information. Next, we discuss contigs.

4.3.3 Contigs

Evaluations of contig quality should consider 1) assembly quality and 2) if the assembled contigs represent the sequenced reads.

Assessments of assembly quality typically consider the number of contigs, length of the longest contig and the metrics: L50 and N50. Acceptable values for the number of contigs and longest contig can vary depending on study goals and sequencing technology used, but large values for both metrics indicate a better assembly. The metric N50 calculates the length of the shortest contig at which all contigs as long or longer than the N50 value encompass 50% of the contigs–a weighted median contig length. Larger N50 values indicate that an assembly consists of longer contigs, generally indicating assembly success (International Human Genome Sequencing Consortium et al., 2001). The metric L50 represents the smallest number of contigs whose summed length constitutes half the total length of the assembly. A large N50 value combined with a smaller L50 indicates that the assembly is composed of a few long (likely data-rich) sequences (Bradnam, 2015). Though informative, using contig length to infer assembly quality requires caution, as these metrics are useful to compare assemblies against one another–especially when assembling a single genome–but in a mixed microbiome, small contigs are not necessarily a problem. Short (<2,000 bp) contigs can still provide valuable information and can be common for communities enriched in plasmids, viruses, mobile genetic elements, or low-abundance microbes (Maguire et al., 2020; Kieft and Anantharaman, 2022). These short contigs are often removed by default when binning genomes (Alneberg et al., 2014), with the justification that the average bacterial gene is ~1,000 bp (Xu et al., 2006) and shorter contigs are unlikely to contain complete genes. This removal can discard valuable genetic context and should be performed with full knowledge of the risks for data loss.

Assessing the representativeness of assembled contigs for a microbial community often involves aligning the un-assembled reads to the assembled contigs (Aroney et al., 2025). The percentage of reads aligning to the contigs indicates how much of the original information is present in the derived contigs. If the percentage of reads aligning to contigs is high (>90%) then the contigs can be considered representative of the community while a low value (<50%) indicates that the contigs are not representative. In cases of unrepresentative contigs, the assemblies still contain useful information for individual genomes, but community-scale inference may require refining contig assembly (parameter optimization, more sequencing) or performing read-only analyses (see Section: 4.4 Read-based Marker Gene Analyses"). Next, we discuss quality control in genomes.

4.3.4 Genomes

Genome quality is most often assessed by the metrics of "completeness" and "contamination", for which there are published quality standards (Bowers et al., 2017). User-friendly tools exist to calculate both metrics, with new versions accounting for whole-genome features (Chklovski et al., 2023). For illustrative purposes, we will describe quality estimation using older methods–that use a constrained number of single copy marker genes–as they are more tractable for beginners (Parks et al., 2015). In the marker gene-based approach, genomes are screened for the presence of a set of single copy marker genes, with the assumption that a complete genome should have only one copy of each gene in the set. In this approach, completeness is estimated as the percentage of marker genes detected in the genome with contamination based on how many times a single copy gene was duplicated–potentially indicating errors in assembly or binning. Ideally, genome analysis would be performed on whole uncontaminated genomes, but this is often impossible due to limited sequencing data, requiring the use of incomplete or contaminated genomes. These imperfect genomes can still provide useful insights, if caveats in the interpretation of these data are acknowledged.

Despite the utility of standard quality metrics, there are cases where they are misleading. First, relying on simplifying metrics obscures reality. Computationally, a "complete" genome does not mean that the genome is "closed" or "finished" (i.e., represented by a single contig without gaps; Bowers et al., 2017) which is an even higher standard of quality. Further, many studies only analyze "high to medium quality" genomes (Bowers et al., 2017), potentially discarding other genomic data that does not conform to the expectations of "completeness", including endosymbionts, plasmids, and viruses, all essential components of a system. Combining both challenges, obtaining a "complete" or "closed" genome cannot assess if a single organism contains multiple chromosomes (*Rhizobium*, Landeta et al., 2011) and/or nucleic acids from other sources (viruses, plasmids, endosymbionts). These shortcomings are systemic but can be overcome with intentional analysis. Obtaining closed genomes often requires focused efforts (long read and/or deep sequencing), while recovering overlooked plasmids and viruses can come from otherwise discarded data (Fogarty et al., 2024). Identifying which chromosomes, endosymbionts, plasmids, and viruses reside inside one organism requires sequencing single cells (i.e., single amplified genomes; Labonté et al., 2015) to gain a fuller understanding of their importance and functions. Now, we move on to discuss the use of marker gene surveys in omics.

4.4 Read-based marker gene analyses

Quality controlled reads can be used to provide insight into the taxonomy or functional potential of an organism or community through the analysis of marker genes. Genes are considered "markers" if they are involved in metabolisms of interest (e.g., *nifD*: encoding the nitrogen fixing Nitrogenase molybdenum-iron protein alpha chain or *mcrA*: encoding the methane producing Methyl-coenzyme M reductase I subunit alpha) or can be used to reconstruct evolutionary relationships (e.g., 16S rRNA gene or *rpoB* encoding the beta subunit of bacterial RNA polymerase). Such genes are typically well represented in existing databases, serving as useful references for comparison (phylogenetic analysis). Marker gene analyses most commonly use targeted amplification and sequencing (Figure 3C). As an example of taxonomically informative marker genes, Lozupone et al. (2013) used 16S rRNA gene amplicons sourced from global sequencing of human microbiomes to identify forces structuring communities including disease status and body site. Surveying function, Dumont et al. (2014) used gene amplification for *pmoA* (particulate methane monooxygenase, beta subunit) to identify the presence of methanotrophic bacteria and delineate phylogenetic clusters. Somewhat less commonly, marker genes can be recovered from untargeted metagenomic and transcriptomic sequencing (Figure 3K), where reads are aligned to marker gene databases, with confident read alignments to a gene indicating its presence and abundance. For example, Maritan et al. (2025) searched marine sediment metagenomes for metabolic marker genes involved in aerobic and anaerobic metabolisms in coral reef sediments. For taxonomy, Urayama et al. (2024) surveyed the prokaryotic taxonomic composition of metagenomes from several hot springs using fragmentary rRNA sequences before digging deeper into the sequences of co-existing viruses. Both amplification and genome/transcriptome applications are appropriate for targeted questions that involve the constrained goals of identifying specific metabolisms or taxa. While the genomic/transcriptomic approaches have the advantage of being able to initially query the whole dataset (Hauptfeld et al., 2024) for specific genes and later studying more detail though assembly and binning, which we will discuss next.

4.5 Contig assembly and analysis

Contig assembly aims to reconstruct longer and more information-rich sequences from shorter reads. This process entails two steps: 1) normalization and 2) assembly.

4.5.1 Read normalization

Read normalization (Figure 3B) reduces the computational burden of contig assembly by limiting the amount of data passed to the algorithm. This is achieved by subsampling redundant sequences and removing low abundance sequences (that are unlikely to assemble). Normalization is appropriate for diverse (e.g., sediments; Maritan et al., 2025) and simpler (e.g., hot springs; Colman et al., 2024) samples. Read normalization is straightforward to implement with tools like bbnorm, developed by the Joint Genome Institute ("https://sourceforge.net/projects/bbmap/"), where read data is input and normalized, with the output generally ready for contig assembly.

4.5.2 Contig assembly

Assemblers (Figures 3D,N) use short reads to reconstruct longer sequences (DNA or RNA). Detailing assembly algorithms is beyond the scope of this review (described in Ekim et al., 2021; Yang et al., 2021), but illustratively, assemblers look for overlap between reads and use this overlap to create longer and longer sequences (Ayling et al., 2020). There are two major classes of assembly: 1) guided and 2) *de novo*.

Guided (i.e., reference-based) assembly aligns reads to sequences from related organism(s), serving as a scaffold to guide placement of the read data. These guide sequences should be sourced from organisms closely related to those in the reads and can be a reference genome (i.e., reference guided assembly Lischer and Shimizu, 2017); or long read data from the same sample (i.e., hybrid assembly, Antipov et al., 2016).

De novo assembly has two varieties: 1) individual, and 2) co-assembly. In individual assembly, reads from a single sample are assembled into contigs. In that case, all assembled contigs are by-definition present in that sample. In co-assembly, reads from similar samples (soils from the same site; Riley et al., 2023) are combined and then assembled as a single dataset. This is often done with the aim of generating contigs from lower abundance organisms (Riley et al., 2023). Some resultant contigs from a co-assembly might not be present in all the source samples, but presence/absence can be determined by quantifying abundances of the contigs in the sample (see "Section 4.10 Count Data"). In co-assembly, reads should be normalized after combining samples, thereby potentially retaining low abundance sequences that might have been removed in individual normalization. To maximize contig recovery, it is possible to assemble contigs using both individual and co-assembly approaches and then later remove any duplicated sequences (see "Section 4.7 Sequence dereplication").

These contigs can be used to study standalone genes, plasmids, or viruses. For example, contigs were used by Priest et al. (2025) to identify seasonal patterns of functional potential in the Arctic Ocean, while Fogarty et al. (2024) searched for novel plasmids in human gastrointestinal tracts, and Zhong et al. (2024) identified viruses encoding methane cycling genes. Though these contig-analyses are useful, the most common use of contigs is binning into genomes.

4.6 Genome binning and analysis

Creating genomes from contigs (Figure 3F) involves grouping contigs into distinct, taxonomically coherent "bins". These bins represent draft genomes that must then be evaluated for quality, completeness, taxonomy, and function. Genomes binned using sequences from pure culture isolates or single amplified genomes (SAGs) may be considered "strains" (Conrad et al., 2022). In contrast, genomes binned using sequences from community sequencing are called metagenome-assembled genomes (MAGs) that often represent consensus sequences from multiple closely related strains sharing similar but non-identical genomes (Meziti et al., 2021).

4.6.1 Binning contigs

Binning programs (binners) typically separate contigs into bins based on shared genomic features and read depth (Bowers et al., 2017). Binners assume that intrinsic genomic features, such as GC content and oligonucleotide (i.e., k-mer) frequency, are consistent across a genome (Bussi et al., 2021), allowing an initial univariate (GC) or multivariate (k-mer composition) separation of contigs into clusters. These initial clusters can be refined with the additional assumption that contig read depth–as measured by the number of reads aligning to each assembled contig–is also consistent for all contigs for a given genome (Sharon et al., 2013). Each individual binning tool generally implements all or some of these approaches (Alneberg et al., 2014), creating draft genomes that can be further refined, annotated, and compared.

4.6.2 Genome improvement

Maximizing the accuracy and information content of individual genomes can be done by selecting the highest quality genomes generated using multiple binning programs (refinement) and reassembling high-quality genomes (reassembly).

Refinement starts by using multiple binners to generate somewhat redundant bins. The resulting genomes from each of these binners are compared to find the highest quality genomes with the highest completion and lowest contamination values. Each chosen genome is placed in a final bin set, often with a cleaning step to ensure each contig is only found in a single–highest quality–bin (Uritskiy et al., 2018). Reassembly uses high quality (quality controlled, refined, and/or dereplicated) genomes to try to re-generate these genomes with even better contigs. This involves aligning the original quality-controlled reads to each genome to "isolate" sequences for an organism of interest. These reads can then be re-assembled using a non-metagenome assembler (SPAdes instead of metaSPAdes; Uritskiy et al., 2018), repeating alignment and re-assembly until achieving a genome with the greatest completion and smallest contamination values possible (Kitzinger et al., 2020).

4.6.3 Shortcomings and hazards

While useful and widespread in omics, genome binning does have shortcomings. First, not all sequences can be binned. Binning relies on high quality assemblies that can be grouped based on sequence similarity–which requires that disparate parts of a single genome have similar sequence characteristics (Nelson et al., 2020). This assumption may not be true for genome fragments that have been acquired by horizontal gene transfer (HGT), carrying sequence characteristics different from those of the recipient's genome (Mise and Iwasaki, 2022). Similarly, genetic elements not incorporated into a genome (such as plasmids and viruses; Eren and Banfield, 2024) or second chromosomes (Landeta et al., 2011) do not meet the assumptions of binners. Without contiguity and/or sequence similarity to the focal chromosome, HGT-derived genes, mobile genetic elements, and second chromosomes may be erroneously separated from their true genomic neighbors (Maguire et al., 2020). Second, intergenic or non-protein coding genomic regions (ribosomal RNA operons) and genomic regions with repetitive sequence features, are often challenging to assemble or bin correctly and are underrepresented in genomes (Mise and Iwasaki, 2022; Wilbanks et al., 2022). Third, in samples containing multiple closely related organisms, genome-approaches collapse strain-level microbial diversity, blurring intra-species genomic boundaries (Wilbanks et al., 2022) and obscuring genomic novelty. In these instances when binning excludes sequences or blurs organism boundaries, analysis of binned data may lead to inaccurate measurements of community-level diversity, fail to detect certain taxa or functions, and provide an incomplete view of the genomic environment of cells.

Many of these shortcomings can be minimized. Complex and HGT-derived genomic sequences can be definitively linked to their genomes using long-read sequencing to sequence across ambiguous genome space (Wilbanks et al., 2022). Capturing the diverse genomic material (chromosomes, plasmids, viruses) in a single cell can be achieved with single amplified genome (SAG) sequencing. SAGs also provide strain-level genomes, helping to resolve heterogeneity among closely related genomes. Once reads, contigs, and genomes are generated, they can be simplified by dereplication before analysis.

4.7 Sequence dereplication

Sequence redundancy is common from reads to genomes and can be removed to reduce computing requirements or analytical repetition. Dereplication calculates sequence similarity between sequences (reads, contigs, and genomes) with an array of programs (VSEARCH, Olm et al., 2017; Rognes et al., 2016; MMSeqs2; Steinegger and Söding, 2017; CD-HIT; Fu et al., 2012) and then uses similarity cutoffs to create clusters of similar sequences (sequence clustering; Box 1). Once sequence clusters are identified, the highest quality sequence in each cluster can be extracted and used as a representative for all other sequences in its cluster. In read data, clustering is most frequently seen in taxonomic marker gene analysis using operational taxonomic units (OTUs; Hughes et al., 2001; Box 1). Contig clustering often takes the form of gene catalogues, where protein coding sequences on a contig are clustered, often principally by taxonomy and then by sequence similarity (Muratore et al., 2022; Priest et al., 2025). Finally, genome clustering is most often used for dereplication of entire genomes (Figure 3F). In all these use cases, dereplication by sequence similarity is a powerful and unbiased approach to simplify similar sequences. These sequences are now ready for taxonomic classification and functional annotation.

4.8 Taxonomic classification

4.8.1 Roadmap for implementation

Many analyses aim to connect sequences with taxonomic labels (reads, Figures 3C,K; contigs; Figures 3E,N; and genomes; Figure 3G). Taxonomic classification often relies on aligning an unknown query sequence against a database (untargeted: NCBI nt/nr; or molecule-specific: SILVA rRNA) of sequences with defined taxonomies (i.e., subject sequences), with the query inheriting the taxonomy of its–sufficiently similar–best aligned subject sequence. A common implementation of taxonomy-by-alignment involves using the NCBI BLAST webserver (Camacho et al., 2009; https://blast.ncbi.nlm.nih.gov/Blast.cgi) to align a query against one of multiple databases, providing accessible fast taxonomies. Alignment-based classification is effective (Jain et al., 2018) but can be supported by estimating evolutionary divergence of the query sequence compared to taxonomically resolved homologues. These homologues are selected to include both close and distant relatives of the query and used to construct a phylogenetic tree (see "Section 4.11 Phylogeny"). In this method, the query inherits the taxonomy of its–sufficiently similar–closest neighbor. Implementing phylogenies is straightforward with multiple tools for automated (GTDB-tk, Chaumeil et al., 2022) and semi-automated (PhyloPhlAn, Asnicar et al., 2020; MarkerFinder; Martinez-Gutierrez and Aylward, 2021) phylogenetic classification, providing broad access.

Though both direct-alignment and phylogenetic placement are applicable to read, contig, and genome based-analyses, longer sequences encode more evolutionarily relevant information and thus provide better taxonomic resolution than shorter ones. This is of limited concern for genome-based analyses (containing Megabases to Gigabases; Milo and Phillips, 2015) but can produce less reliable taxonomy for reads (100–250 bp; Hauptfeld et al., 2024). Read length limitation can be overcome by using reads to reconstruct and classify the more informative contigs and genomes, then assigning the constituent reads the taxonomies of their contigs and genomes. Ultimately, this multi-step classification combines the taxonomic clarity of genomes and the community representation of reads (see "Section 4.4 Read-based Marker Gene Analyses") to achieve a high-quality understanding of the sequenced community (and is implemented in open access tools; Hauptfeld et al., 2024).

4.8.2 Shortcomings and hazards

A note for users, the quality of taxonomic classification is dependent on the completeness of the reference database. Under ideal circumstances, database subject sequences originate from an isolated, living specimen providing a confident association between database taxonomy and a living organism. As sequencing captures more diversity than exists in-culture, connecting a query sequence to a type specimen is often not possible, instead requiring comparison to uncultured sequences (MAGs; Murray et al., 2020). This means that assigning taxonomy to divergent organisms requires more effort (phylogenies; Eme et al., 2023) than in organisms closely related to models (*E. coli* and *Staphylococcus aureus*), potentially requiring the creation of new taxonomic groups (Rinke et al., 2013; Murray et al., 2020). Another potential concern for assigning taxonomy is the influence of horizontal gene transfer. The exchange of genes between organisms (bacteria-bacteria, Tschitschko et al., 2024; bacteria-virus; Li et al., 2025; bacteria-eukaryote; Porras et al., 2024) can obscure the evolutionary lineage of any one sequence. Disentangling the current genomic placement–and taxonomy–of any gene generally requires situating it in a complete, contiguous genome.

Taxonomy is a useful, but incomplete classification of living organisms (Aldrich, 1927; Staley, 2009; O'Brien and Luo, 2022). Indeed, ecosystem-scale analyses (biogeochemistry) sometimes pay little to no attention to taxonomy, focusing only on functions encoded in nucleic acids. In aid of both taxonomy-agnostic or -informed analyses of encoded functions, we will next discuss functional annotation.

4.9 Function annotation

The encoded biochemical outputs (expressed RNA and translated proteins) are the focus of many analyses. The act of assigning inferred function to a sequence is called annotation. Functional annotation of reads (Figures 3C,K), contigs (Figures 3E,N), or genomes (Figure 3G) predicts the potential cellular activities of nucleic acid molecules (rRNA, tRNA) or–most commonly–of encoded proteins. Like taxonomic classification, functional annotation compares a query sequence against a database of annotated reference sequences. Under ideal circumstances, prior experimental studies have confirmed the biochemical function of molecules encoded by the reference sequences.

Annotations of non-protein coding regions are identified directly from nucleic acid sequences (rRNA: Barrnap, https://github.com/tseemann/barrnap; tRNA: tRNAscan, Lowe and Eddy, 1997) while protein coding genes are either identified directly from reads or from identified protein coding regions (from reads, contigs, genomes). Identifying protein coding regions (i.e., open reading frames, ORFs; Box 1) searches for their molecular characteristics (e.g., start and stop codons; tools: Prodigal, Hyatt et al., 2010; FragGeneScan; Rho et al., 2010) outputting likely protein-coding sequences for use in homology searches.

Functional annotation is performed as either a targeted or untargeted search. A targeted search focuses on dozens of genes of biogeochemical or ecological significance (Leung and Greening, 2020; Zhou et al., 2022), identifying the potential for a microbiome to perform specific functions of interest. Targeted searches can input short reads (Dragone et al., 2022; Bertagnolli et al., 2023) or open reading frames (Priest et al., 2025). This approach can be used to quantify the presence of catalytic genes in a sample. For example, Dragone et al. (2022) searched Antarctic soil metagenomes for genes involved in trace gas cycling to quantify the genomic potential of the entire microbiome to utilize trace gasses across multiple environments. Targeted searches are also useful as an initial screening of large genomic datasets (reads to genomes) before digging deeper. For example, Speth and Orphan (2018) were interested in the diversity of methanogens across thousands of publicly available metagenome datasets. To save computing time, they pre-screened datasets for the presence of diagnostic methanogen gene *mcrA* (Methyl coenzyme M reductase), only assembling contigs and binning genomes from *mcrA* positive datasets.

Untargeted searches do not have specific genes of interest, instead aiming to annotate as many sequences as possible. This approach is best suited to ORFs because they contain enough genomic content to be confidently annotated against hundreds of thousands of reference genes. This endeavor often starts off semi-targeted, using tools searching for tens of thousands of specific genes (Prokka, Seemann, 2014; KofamScan; Aramaki et al., 2020). The sequences that remain un-annotated after this first pass may still be amenable to annotation and can then be queried against even more comprehensive databases (NCBI nt/nr, UniProtKB). If homologs to these sequences cannot be identified, a cautious approach is to designate such ORFs as "proteins of unknown function", or "hypothetical proteins". The functions of these hypothetical proteins may be inferred based on the functions of nearby sequences (within the same operon; Mihelčić et al., 2019) or demonstrated using non-omics approaches (biochemistry and cell biology; discussed below). An untargeted approach will generate a lot of annotations and is most tractable when constraints are applied to its analysis. One way of constraining the analysis is by examining only a few genomes in-depth. For example, Mitchell et al. (2024) sought to examine gene expression for a single bacterium, using five semi-targeted tools and the NCBI non-redundant protein database to annotate the genome. Another method to constrain the large volume of information from an untargeted analysis is to use an annotation system with a simplifying gene hierarchy (ontology; Box 1). For example, Kelly et al. (2019) annotated seawater metagenomes with the SEED subsystem database–grouping genes by functional categories–which they used to collapse annotations into functional groups, making the analysis of thousands of sequences tractable.

Mechanistically, functional annotation often relies on sequence alignment or Hidden Markov Model (HMM) searches. Alignment compares a query sequence (nucleotide or translated amino acid) to a functionally annotated subject sequence, identifying regions of sequence similarity. If the two sequences are sufficiently similar, the query sequence is assigned the annotation of the subject. The most common sequence aligners are those of NCBI's Basic Local Alignment Search Tool, which work by finding identical sequence fragments (substrings) between a query and reference and then expanding the alignment outward from the region of identity (BLAST, Camacho et al., 2009). BLAST searches can be performed via a web interface that accesses NCBI's servers directly or–more efficiently–using BLAST software installed locally. BLAST-based homolog identification can be computationally intensive but is generally accurate (Al-Fatlawi et al., 2023). In the decades since BLAST was introduced, other alignment alternatives have been developed, with many of these being faster and equally or more accurate (DIAMOND, Buchfink et al., 2015). HMM-based approaches use databases of homologs to build a profile/model for a protein or protein domain of interest. This HMM profile contains features (the probabilities of different amino acids at different positions) intrinsic to the protein or protein family and can be used to search a sequence dataset to identify putative homologs with high confidence (details of HMMs reviewed in Mor et al., 2021). HMMs can be more sensitive than BLAST searches in identifying distant homologues (Kirsip and Abroi, 2019) but require training on high quality sequence data. Fortunately, several repositories of pre-trained HMMs are available (TIGRFAM, Haft, 2001) with some integrated directly into annotation tools (KofamKOALA, Aramaki et al., 2020). Though not widespread yet, attention-based artificial intelligence also holds great potential for functional annotation (Hwang et al., 2024) and prediction (Jumper et al., 2021) but is beyond the scope of this brief overview.

An important note, confidence in any gene's annotation is a balance between effort and confidence. Many genes can be annotated quickly–to a high degree of confidence–but approaching "proving" that an encoded gene can perform a function requires increasing effort. This may require narrowing the focus from many (10,000+) to a few (1–10) genes, eventually departing from omics altogether for the domains of biochemistry and molecular genetics (Table 2). Protein purification or heterologous expression should only be used for absolute proof, as–in most cases–automated gene annotation or simple phylogenies are sufficient to hypothesize the functions of a gene. Solid annotations lay the foundation to compare gene prevalence, abundance, or expression between systems via quantification.

4.10 Count data

Quantifying omic features in a dataset (reads, Figures 3C,K; contigs; Figures 3E,N; or genomes; Figure 3G) uses read data. Read quantification involves counting reads of a given type (reads aligning to marker gene regions), typically followed by normalization to sequencing effort (e.g., Reads per Megabase of sequencing). Contig or genome quantification requires aligning reads to these longer sequences, typically followed by normalizing for contig/genome length and dataset size (e.g., Reads Per Kilobase of Contig per Megabase of sequencing). It should be noted that alignment-based quantification may overestimate sequence abundances with methods developed to counteract this (TAD80; Viver et al., 2021).

Count data are prevalent across genomic studies, estimating the abundance of genes (Dragone et al., 2022; Maritan et al., 2025; Ricci et al., 2025) and microbes (Steinsdóttir et al., 2022;

TABLE 2 Increasing confidence in functional annotation is increasingly time intensive and eventually requires non-computational approaches: A simple workflow for increasingly confident annotations with steps, actions, realistic number of sequences to analyze, interpretations, and examples in the literature.

Steps	Action	Number of sequences analyzed with this technique in one study	If confirmatory, what does this tell you?	Published example
1: Identify likely homologues	Identify candidate homologues (BLASTn/p, HMM)	10,000+	Target sequence is sufficiently similar to known sequence to be a homologue, though may include false positives	Screen thousands of genes (Anantharaman et al., 2016)
2: Contextualize phylogenetically	Phylogenetically place gene of interest against high confidence (SwissProt) gene sequences	10s	Target sequence is situated with other sequences known to perform the function of interest	Tree genes of interest (Graf et al., 2021)
4: Identify essential motifs and structures	Identify key motifs (Pfam) and structures (AlphaFold)	10s	Target sequence possesses necessary architecture for claimed function	Identify functional residues (Porras et al., 2024)
5: Assay with biochemistry and molecular biology	Knockout or clone gene of interest and biochemical assay	1–2	Target sequence performs the assayed function	Clone gene (Tsementzi et al., 2016)

Shoemaker et al., 2024) in a sample. In transcriptomics, cDNA-derived reads are aligned to a reference sequence (genome, Bertrand et al., 2015; or assembled transcript; Sorek et al., 2018) to estimate transcription levels of genes. Metagenome and metatranscriptome studies can also quantify exact numbers of transcript molecules per amount of sample or per gene copy number. This is most precise when mRNA or genomic DNA standards are spiked into samples (Moran et al., 2013; Nowinski et al., 2023) but can also be estimated by normalizing gene expression to measured biochemical properties (expressed mRNA per gram soil; Söllinger et al., 2018; Täumer et al., 2022). This allows precise quantification of omics data and can be especially useful for estimating changes in metabolic activity.

These count data of gene abundances and expression levels provide a basis for hypothesizing about the function of a system but generally require other methodologies for confirmation (quantitative PCR, cell counts, chemical measures, or cell culture).

4.11 Phylogeny

Because phylogenetic inference is essential to taxonomic and functional omics analyses, we will briefly summarise the methods for phylogeny construction here. However, we note that this is only a primer and does not cover all the details needed to correctly perform these analyses. For more in-depth discussion, we direct readers to excellent reviews describing the principles and tools for phylo-genetics/-genomics (Kapli et al., 2020; Steenwyk et al., 2023).

A phylogeny or phylogenetic tree (Figure 3J) shows the evolutionary relationship of a focal sequence relative to reference sequences. Interpreting a phylogeny involves examining two key features: 1) topology and 2) branch length. Topology describes the shape of a phylogenetic tree, including branching patterns and clusters of sequences (Kapli et al., 2020). Assuming there is statistical support (via bootstrapping) for the groupings in the tree, sequences clustering together is often used to support claims that a focal sequence shares evolutionary history with a taxonomic (Eme et al., 2023) or functional (Porras et al., 2024) group, permitting classification or annotation. Conversely, divergence between sequences can be used to delineate new taxonomic groups at coarse (Woese and Fox, 1977; Lane et al., 1985) or fine phylogenetic scales (Tsementzi et al., 2016) and follow up with the question: "what changes have accumulated between two diverging sequences" (e.g., individual sequences, Major et al., 2017; whole genomes, Conrad et al., 2022). Branch length–in a rooted tree–describes the distance from a phylogeny's root to any tip, serving as a proxy for a lineage's age. Time calibrated branch lengths (using dated fossils or geochemical evidence to contextualize divergence) provides insights into the exact timing of diversification (Damsté et al., 2004; LaJeunesse et al., 2018). Beyond describing divergence timing, branch lengths can be used to quantitatively assess how evolution drives ecological associations (Colman et al., 2024).

The creation of a phylogeny comprises three main steps: 1) sequence acquisition, 2) multiple sequence alignment, 3) phylogenetic inference (reviewed in Kapli et al., 2020). Both nucleic acid and amino acid sequences can be used for phylogenetic inference. It is common practice to use nucleic acids to resolve closely related organisms (due to more combinations available for nucleic acids to specify any codon than for amino acids) and amino acid sequences for more distantly related sequences, though nucleic acids and amino acids may provide similar resolution for distant relationships (Kapli et al., 2023).

The first step of sequence acquisition involves identifying sequences for phylogenetic reconstruction. This can be done manually (BLAST genomic sequences against a gene of interest reference database) or automatically (MarkerFinder, Martinez-Gutierrez and Aylward, 2021). If the analytical goal requires comparing homologs only, it may be necessary to remove potentially non-homologous–but similar–sequences identified by sequence searches. This step often requires manual inspection and can be time-intensive (reviewed in Kapli et al., 2020). The product of sequence acquisition–a set of confident homologues–is the starting point for the next step, multiple sequence alignment (MSA). MSA compares sequences to correctly orient homologous base or amino acid positions along the sequence. This results in a matrix in which the rows indicate sequences and columns indicate homologous positions in a sequence, with residues (bases/amino acids) shared among sequences at the same position often indicating shared ancestral patterns of sequence change. If creating a phylogeny from multiple genes, genes may be first combined (concatenated) and then aligned or first aligned separately and then concatenated, in both cases creating an MSA supermatrix. The accuracy of phylogenetic reconstruction depends intrinsically on the accuracy of the MSA. Therefore, any MSA should be manually examined after the (typically) automatic step of alignment, potentially to verify strandedness of homologues (so as not to mistakenly compare palindromic regions), remove sequences that align poorly or with high percentages of gaps, or mask ambiguously aligned regions (Kapli et al., 2020). Finally, phylogenetic inference involves generating a bifurcating tree that estimates evolutionary relationships based on shared residues in the MSA and a model (set of assumptions) about the process of sequence change. This step can be performed by creating and merging multiple trees from each of the aligned genes or a single tree from the gene supermatrix. The methods for constructing trees are diverse and vary in the extent to which they estimate and incorporate parameters describing the evolutionary process and, consequently, the time and computational resources required for the analysis (Kapli et al., 2020).

Phylogeny, and all the previously described tools, were given only a brief treatment. Our aim was to provide a foundation for readers to seek out more in-depth guides as needed. We will end our discussion of tools by highlighting new frontiers for omics application.

4.12 Contextualizing across datasets, time, space, and conditions

Individual tools are essential to produce the core omics data products, but once these data are produced, an omics scientist has the freedom to use these results to answer any number of scientific questions. We suggest that omics users make full use of publicly available databases to place their results into larger contexts (Figures 3H–J). Using public data, a researcher can compare their sequences against other similar (or different) studies to identify: shared or disparate trends (meta-analysis: Thompson et al., 2017; Kumagai et al., 2018; Ruff et al., 2024), reconstruct evolutionary histories (phylogeny: Hug et al., 2016; Eme et al., 2023), spatial distributions (biogeography: Härer and Rennison, 2023; Zhou et al., 2024), or driving environmental factors (modeling: Louca et al., 2016; Lui et al., 2021; Ramoneda et al., 2024; Chuckran et al., 2025). Each of these contexts is a discrete field with its own norms and tools beyond the scope of this review. In any case, situating omics findings in a broader context is almost always a worthwhile exercise that generally increases the utility and impact of omics research. We conclude this review with some final suggestions for new practitioners from our own experiences learning and teaching omics.

5 Tips for new practitioners

Starting to perform bioinformatics is formidable with layers of challenges. First, there are the concrete ones, learning how to code and manage terabyte sized sequencing datasets. Second, there is the conceptual task of designing workflows to generate useful results. Once these challenges are cleared, there remains the most formidable challenge, performing scientifically meaningful "experiments" on the computer. Thus far, this review has focused on the conceptual task of designing workflows for useful results. We will end with some suggestions for future work and literature to handle the fine details of coding and the broader issue of asking meaningful scientific questions.

5.1 Bioinformatics advice

5.1.1 Opportunities for further training

Readers of this review should leave with an understanding of the motivations and methods of nucleic acid omics. For some readers, this review will be sufficient for their goals of digesting the "methods" sections of manuscripts, while others seeking to analyze data independently will need more specific training.

For those looking for additional training, we recommend three types of resources ordered from most accessible to most specialized. First, for guided exposure to using real data to run specific omics analyses, we recommend computing workshops or courses (in-person or online). Such workshops can be broad (binning MAGs) or specific (machine learning for protein prediction), providing an opportunity to develop a range of skills. Second, for self-guided learning of specific computing topics (read mapping to quantify transcripts, identifying viruses in omics data), we recommend online tutorials. These tutorials can be standalone websites (often the online material from a prior workshop) or published as part of a manuscript (e.g., Coenen et al., 2020). Tutorials are incredibly useful for users that can read and write some code (see "Section 5.1.2 Scripting") and want to see how specific analyses are run, often demonstrated by analyzing subsampled real datasets. Finally, when trying to implement a specific tool (often found through a workshop or tutorial), we recommend reading the tool's official documentation. This documentation often exists in two forms. First, many tools are announced with a publication describing their construction and general uses–which is useful for an overview but may overwhelm early omics users with technical information. Second, each tool generally includes a manual written for practical implementation. This may be as simple as a "README" text file included in the downloaded source code or as involved as a dedicated website to explain the uses and functions of the tool.

Building on the foundation of knowledge from this review, early-stage omics users will be able to acquire and integrate additional training to analyze data independently.

5.1.2 Scripting

To interact with omics datasets, users can begin by using software that does not require much coding experience (Genious, Galaxy, Kbase), though accessing the full capacity of omics datasets requires learning to code (but does not require the skills of a professional programmer).

Scripting is the computing equivalent of pipetting in a lab, and aspiring bioinformaticians should be able to write and read code in Bash and R or Python. Bash is the basic language for performing omics analyses, used to interface with high performance computing clusters and run bioinformatics programs. Though Python can be used to write standalone programs, we suggest learning Python or R for their capacities to manage spreadsheets, perform statistics, and make plots. These tools take longer than Excel to master, but they quickly outperform it in flexibility, speed, and reproducibility. We suggest that bioinformaticians learn how to use either R or Python, as it is unlikely that knowing both will be essential, and if more languages are needed, they can be learned (Schloss, 2020). To learn coding, there are lots of online resources, with more available every day. For bash, we recommend chapters from the book *Practical Computing for Biologists* (Haddock and Dunn, 2011) pertaining to bash for a basic overview of some core commands and syntax and immediately applying it on real data to get a feel for its use. For R, the free online textbook *R for Data Science* (Wickham, 2023: https://r4ds.hadley.nz/) is an approachable read, organized to be practical and user friendly. For Python, we recommend the interactive free courses offered on Codeacademy (https://www.codecademy.com/). Though bash, R, and Python have been mainstream tools for decades (and may remain so), the scripting toolkits available and the resources to learn them will change over time, which should inform the training tools selected.

If coding is like pipetting, writing code with AI is like operating an automated liquid handler. Automating lab work may aid a novice wet lab scientist, but scientists with hands-on experience will better understand how to creatively and effectively implement such automation. AI support in bioinformatics works better with the specific vocabulary and perspectives that come from already knowing how to write code and manipulate data. That being said, we wholeheartedly recommend AI coding tools to help write tedious scripts (loops), make existing scripts more efficient, installing tools, debugging and explaining code. In any case AI users should be updated on best practice recommendations and publishing requirements (Buriak et al., 2023; Blau et al., 2024), they will change over time. In any case, a bioinformatician who knows what they want out of a workflow will be better able to get it with whatever basic or advanced tools they bring to bear.

5.1.3 Local data organization

Data organization should be a primary directive. Files will always need naming and directories (folders) will always need organizing. We suggest you adopt a simple filing system (Noble, 2009) and adapt as you see fit. Do not proceed without some kind of system, as impromptu "organization" will eventually accumulate into an unwieldy mess. Two places where poorly organized files can cause major trouble are raw sequencing files and scripts.

All new sequence data should be placed in a clearly marked and backed up location (if working on a team, this original data should be in a shared "group" directory, not on the bioinformatician's personal drive). This directory should have an informative name that contains the elements required to understand what is inside, for example,: sequencing run ID, sampling date, sequence type (genome, transcriptome, amplicon …), geographic source (country, ocean basin, cell culture collection …), and sequencing target (host organism, enrichment culture, soil …). In this directory, it is useful to have two sub-directories for reads. First the minimally processed reads from the sequencer (e.g., "o01_raw_reads") and a second quality-controlled set (per "Section 4.3 Quality Control"; e.g., "o02_trimmed_reads") as these processed reads will be used by multiple steps in any analysis and should be easily accessible. A file name tip: use character delimiters (e.g., " . ", "_") to separate phrases in a file name instead of spaces (" "), to avoid problems later while scripting.

Next, the scripts that are written to analyze an omics dataset should be organized to allow the bioinformatician (or anyone else) to follow the workflow and backed up to prevent loss. Writing individual scripts for each step of an analysis is a good habit to keep the workflow clear and easily debugged. We recommend naming files sequentially so that auto-sorting arranges them logically (e.g., "o01_read_trimming.sh", "o02_read_normalization", "o03_assembly").

Again, the raw sequence data and scripts should be backed up to prevent loss, as they are the minimum information required to regenerate all results.

5.1.4 Public databases

One beautiful aspect of bioinformatics is the interoperability of sequences from diverse sources. The base FASTA format–adopted in 1985 and used today–has ensured that almost all sequence data uses consistent formatting (Wright et al., 2024). This consistency allows straightforward comparison of new data to archived sequences in repositories. Learning the major databases often involves talking to other scientists and looking at the methods of published papers, but once identified, these resources can easily provide sequences to contextualize new data or to supply material for meta-analysis (general: NCBI SRA, Leinonen et al., 2011; task specific: Tara Oceans, Sunagawa et al., 2020). Some tools exist to make these database searches easier (taxonomy browser, Parks et al., 2018; metagenomes pre-screened for community composition; Woodcroft et al., 2025; NCBI webtools; Sayers et al., 2024) though efficient use of these resources generally requires familiarity and practice. This practice is beneficial for any omics scientist, for a bioinformatician with databases is never without samples.

5.1.5 Responsible data sharing and reproducible analyses

Thus far we have discussed scripting, data organization, and public databases from the perspective of how they benefit the reader of this review. Now, we will discuss the responsibilities of omics users to the broader scientific community–focusing on practices that ensure that published data and results are useful as long as possible. At its core, this means ensuring that raw data from published research is usable in the future and the results from a manuscript can be reproduced.

In an omics context, the space available in a manuscript is often too small to provide all the relevant data (sequences or environmental measures) for long-term use, and the researcher must then rely on external resources to ensure the information is accessible. To provide a brief coverage of the issue, we will discuss data archiving and reproducible analyses (though this coverage is necessarily incomplete and the reader should spend more time on these important topics).

5.1.5.1 "FAIR" data archiving

Though sharing published data is a long-standing scientific practice and many journals require that the raw (unprocessed) data supporting the paper is made available, there is a lot of variation in how "available" could be interpreted. To clarify this, Wilkinson et al. (2016) introduced an influential set of guidelines for data management and stewardship summarised in the acronym "FAIR": Findable, Accessible, Interoperable, and Reusable. We will describe some practices for "FAIR" data management in omics, though we will not cover each element of the acronym specifically and readers should spend more time leaning about the specific guidelines, especially before archiving their data (reviewed in Carballo-García and Boté-Vericad, 2022).

The foundation of omics data archiving requires that primary data files (unprocessed reads, assembled sequences, environmental measures) are available to other scientists, which is generally achieved by depositing data in public repositories. To maximize the lifespan of deposited data the repository needs to persist over time. An excellent example of a durable repository is also the most used for accademic omics data. The International Nucleotide Sequence Database Collaboration (INSDC) is an international effort to capture, preserve, and present nucleic acid sequence data for the "permanent scientific record" (Karsch-Mizrachi et al., 2025). This collaboration has operated for over 40 years, supported by the United States of America (National Center for Biotechnology Information; NCBI), Europe (European Molecular Biology Laboratory-European Bioinformatics Institute; EMBL-EBI), and Japan (DNA Data Bank of Japan; DDBJ). In this collaboration, data deposited to any participant (NCBI's Sequence Read Archive, EMBL-EBI's European Nucleotide Archive, and the DDBJ's Sequence Read Archive) is exchanged with the others daily, ensuring global access and redundancy (Karsch-Mizrachi et al., 2025). Though other repositories exist for sequence data, the global support and historic record of the INSDC's repository makes it one of the best options for most sequence data. Now we will turn our attention to the other essential elements of data archival.

One of the most critical elements of data archiving is that all primary data files must have unique and fixed identifiers (names). The exact names do not matter *per se* (though it is useful when these names have some intrinsic meaning; see "Section 5.1.3 Local Data Organization") because all file identifiers should be explained by accompanying metadata. Metadata is essential to explain what is encoded in the primary data, thereby linking the primary data's identifier to relevant descriptions of "what it is". The types of collection descriptors are summarized in the widely used acronym "ISA": Investigation (e.g., principal investigator, institutions), Study (e.g., location, organism, physical conditions, experimental condition), and Assay (e.g., type of nucleic acid sequenced, sequencing technology; Johnson et al., 2021). Lists of the minimum ISA descriptors needed for different types of primary data exist as "minimum information checklists" (examples at: https://fairsharing.org) and are often organized as standard templates (examples at: https://fairsharing.org). When filling out the relevant metadata, it is good practice to describe the data using well-defined hierarchical ontologies (e.g., taxonomic rank; examples at: https://www.ebi.ac.uk/ols4/) to avoid ambiguity, though new labels can be added as needed.

Finally, to be consistent with FAIR principles, both primary data and metadata should be encoded in widely available and commonly used file formats that describe the permissions or restrictions for reuse of the primary data, ensuring that anyone who retrieves the data will be able to read it and know how to use it appropriately.

5.1.5.2 Reproducible analyses

Aside from sharing published data, omics users are expected to follow practices to ensure reproducibility of their results, allowing other scientists to double check their findings.

The most important elements of reproducibility are that other scientists have access to the raw data and the analytical tools used in the analysis. For access to raw sequence data most papers require that research generating new sequence data deposit it publicly (see "Section 5.1.5.1 'FAIR' Data Archiving") while those re-analyzing data list their sources (see "Section 4.1.3 Data Mining"). However, ensuring access to analytical tools is less regulated.

Though it is still common for scientists to use paid tools ("closed-source"; e.g., ArcGIS, MATLAB) for analyses, there has been a concerted effort by the broader scientific community to promote the use of open-source (free) software for research (Schloss, 2020), allowing broader access. For both closed- and open-source tools, it is expected that published manuscripts describe the software names and version used for each step in their analysis.

Finally, to ensure maximum reproducibility, omics users often publish the exact code used to run their analyses. This code is often written and organized with Git (a software for organizing code and handling version control) and made public through GitHub (a cloud-based service built on Git to share code). Publishing the code used for analyses is less critical when the analysis uses the default settings on published tools (which can be easily reported in a paper's "methods" section) but becomes increasingly important when the research builds new tools and performs more complex analyses.

5.1.6 Quantitative results: statistics, modeling, and guesswork

Omics analyses tend to present qualitative or descriptive quantitative results, rather than explanatory or predictive ones. Though the support of statistics or structure of modeling are not necessary for many good omics papers, the field of microbiology is ready to integrate these approaches more fully. There are examples where omics data is used to model biogeochemical processes (Louca et al., 2016; Täumer et al., 2022) and frameworks outlining how sequencing may be used in predictive models (Lui et al., 2021). More informally, scientists should practice making informed guesses, both qualitatively (from a literature base) and quantitatively (using benchmarked biological values; Fox, 2011; Milo and Phillips, 2015). Predictions (especially quantitative ones from rigorous modeling or informal estimations) are an excellent base from which to write clear, falsifiable hypotheses. Such hypotheses–especially when grounded by scientific literature–are foundational to any science (see "Section 5.2.1 Non-omics Literature and Toolkits").

BOX 1 Glossary of key terms

rRNA: The gene encoding, or the transcript that *is*, a part of the ribosome across all domains of life. The 16S rRNA gene and transcript are the most popular target for studying bacterial and archaeal evolution and diversity, though other rRNA subunits have been used and are informative for bacteria, archaea, and eukaryotes.

Assembly: The practice of reconstructing the sequences of the original nucleic acid molecules after their fragmentation before and during sequencing.

Clone Library: The product of inserting genetic material (targeted or untargeted) into a living vector (e.g., *Escherichia coli*), selecting for only the individuals that took up the material while separating genetically distinct clones (e.g., spread plating isolates with an antibiotic screen) allowing the vector to multiply the genetic material of interest sufficiently for further biochemical processing (e.g., sequencing).

Contiguous Sequence (Contig): The shortest output from the assembly of nucleic acid reads (DNA or RNA). DNA contigs can be further refined into **scaffolds** and **chromosomes**.

FAIR Data Practices: A set of guidelines introduced by Wilkinson et al. (2016) outlining practices for ensuring long-term utility of published data, particularly though the promotion of consistent identifiers, rich metadata, and accessible file formats.

Function Annotation: The practice of assigning a sequence (often a predicted **ORF**) a function if it shows sufficient similarity to a sequence with known function (see **marker gene**).

Genomics: The study of life using the untargeted sequencing of DNA.

Genome: Strictly, the name for the complete set of **chromosomes** originating from the organism of study. Loosely, it also refers to assembled genomic material that is grouped into candidate genomes (**bins**, **MAGs**, and **SAGs**).

[Finished, High, Medium, Low] Genome Quality: Classifications of genome quality introduced by Bowers et al. (2017) relying on contiguity and estimates of completion and contamination.

[Genomic] Bin: A type of genome. A draft genome composed of contigs that have been grouped together based on similar characteristics (GC content, base frequency, or coverage), but have not yet been deemed of sufficiently high quality to be considered a usable genome/**MAG.**

[Genomic] Chromosome: A product of genomic sequence **assembly**, the product of combining **scaffolds** to produce a complete gap-free digitized representation of the source chromosome.

[Genomic] Scaffolds: A product of genomic sequence **assembly**, the product of combining multiple **contigs** with consistent orientation and defined gap sizes, though less complete than a **chromosome.**

ISA Abstract Model: Is used to organize metadata collection and distribution, and is a flexible organizing framework describing the metadata necessary to convey key elements of some data's origin, including the Investigation, Study, and Assay that generated it.

L50: A metric for assessing an assembly's quality via the number of assembled sequences. The metric of L50 describes how many of the longest sequences are needed to account for 50% of the assembly size. Assuming the assembly is of high quality and sufficient sequencing coverage, smaller L50 values indicate the assembly is composed of only a few sequences, which is considered a good thing.

Machine Learning: A class of tools developed from statistics (Bayesian statistics, game theory, computer vision) where algorithms are trained on existing data to identify patterns in new datasets leading to diverse kinds of artificial learning, including: reinforcement learning, supervised learning, and recently popular unsupervised generative learning (e.g., large language models).

Metagenome-assembled Genome (MAG): A type of genome. A **bin** becomes a MAG after passing quality control standards (see **Genome Quality**).

Marker Gene: Genetic sequences that are strongly associated with a biological process of interest including but not limited to: phenotypes, evolution, or behavior.

Minimum information checklist: Is used to organize metadata collection and distribution and provides guidelines for required data reporting for data arising from specific classes of experiments or assays (see **ISA**) to ensure data usability without mandating exhaustive details.

N50: A metric for assessing an assembly's quality via the length of contigs of an assembly, essentially a weighted median contig length. If all the contigs in an assembly are arranged from longest to shortest and began summing contig lengths one contig at a time, the N50 value would be the length of the contig where 50% of the total length has already been accounted for. Larger N50 values indicate that an assembly consists of longer contigs, generally indicating assembly success.

[Meta]-Omics: Omics is an analytical approach that studies entire sets of biological molecules (DNA, RNA, Proteins, Metabolites). Adding the prefix "meta" indicates that the analysis explicitly considers more than one organism (though non-meta omics may incidentally sequence more than one organism).

Ontology: A hierarchically structure for terminology where each term becomes increasingly specific while still "contained" within its broader term (e.g., a twig on a branch on a limb on a tree") that is often used to describe gene functions (e.g., Gene Ontology) or metadata (e.g., The Environment Ontology).

Open Reading Frame (ORF): An open reading frame is a predicted protein coding region from a nucleic acid sequence predicted due to the presence of genetic features characteristic of experimentally validated protein coding regions.

Operational Taxonomic Unit (OTU): A label for sequences (reads to genomes) that have been deemed to share taxonomy based on **sequence clustering** at a defined percentage similarity threshold (often 95%).

Sequence Clustering: The practice of grouping like-with-like sequences (from reads to genomes), often using the measure of pairwise percentage similarity.

Single amplified genome (SAG): A type of genome. A SAG is produced by sequencing the genomic material from a single cell, ensuring that the genomic environment is represented (distinct from a **MAG,** which may provide an incomplete understanding of the associated mobile genetic material or multiple chromosomes).

Sequence Alignment: The practice of comparing two sequences and searching for shared regions between the two. Often results in metrics describing the length of the aligning region and the percentage similarity.

Sequence Library: The name of the sequences originating from a single sample (e.g., a single metagenome file).

Taxonomic Classification: The practice of assigning a sequence a taxonomic origin based on its similarity to a reference sequence with assigned taxonomy.

Transcriptomics: The study of life using the untargeted sequencing of RNA.

5.1.7 Independent work

A final bit of omics advice is to become independent at performing the entire sample-to-analysis workflow: sample collection and preservation, nucleic acid extraction, sequencing prep (though we suggest out-sourcing sequencing to full-time professionals), to most omics analyses. This capacity simplifies troubleshooting and affords more control over the generation of any omics data. This capacity makes a scientist more independent and useful, ultimately a better hire.

5.2 Non-computational advice

Any omics user exists in a larger scientific environment of non-users. To integrate smoothly into this wider non-omics world, we have included some non-computational tips.

5.2.1 Non-omics literature and toolkits

Though this review is focused on the technical details of computing, omics (like other tools: purifying proteins, culturing cells, or collecting samples) is a means to an end. The end goal of scientific inquiry is to incrementally improve human understanding of how the world works. Studying biology requires specific biological (not just computational) knowledge to guide analysis. This is especially true for omics, as any sequencing produces potentially overwhelming quantities of data, and biological context creates order and provides needed direction.

Analytical direction generally comes in the form of a biological question: "what are the tradeoffs involved in capturing light energy?" (Kumagai et al., 2018), "what pH did life originate in?" (Colman et al., 2024); "how does habitat specificity affect global patterns of speciation?" (Sriswasdi et al., 2017); "how do viruses shape mammalian biology?" (Henriques et al., 2024). In the best-case scenario, a biological question is used to inform data collection and analysis. However, as biological systems are often incompletely characterized, omics must often be exploratory, sequencing poorly understood microbiomes. In these cases, a focused scientific narrative requires crafting a biological question retrospectively.

To this aim, an omics scientist should be comfortable with non-computational biological literature (ecology, redox, stoichiometry, developmental biology, physiology, oceanography, pathogenicity, biochemistry). The goal of any omics scientist talking to an expert in their biological field, should be to be seen as "one of us". Further, it is important to become acquainted with non-omics tools, especially how they fill the gaps left by omics (qPCR, microscopy, rate measurements, stable isotope probing, enrichment cultures, isolation, knockouts, microcosms, transformation, protein purification) and as appropriate, add these tools to one's repertoire. A scientist that understands how omics fits into a constellation of other tools will be better equipped to plan research and identify tractable next steps (even if they never intend to do it themselves, it will help them find collaborators, see "Section 5.2.2 Networking").

5.2.2 Networking

Well-read and self-aware omics scientists should see themselves as a part of a global community with shared questions and aims. Tapping into this community to access the knowledge and skills of other scientists requires networking (i.e., making friends) in one's focal-field and beyond. A network of familiar scientists makes scientific study more efficient, accessible, and enjoyable. Networking can be done anywhere scientists congregate (conferences, workshops, fieldwork, online) and is often more interesting and fruitful when it bridges diverse disciplines (biogeochemistry, ecology, biogeography, organismal biology) and departments (microbiology, ecology, earth sciences, geography, biochemistry, engineering). These connections can be used to identify good colleagues and great collaborators. Collaborating (working on the same projects together) with non-omics scientists will be far easier if the omics-scientist understands diverse methodologies and can effectively communicate what omics can and cannot do (see "Section 5.2.3 Communication").

5.2.3 Communication

Maybe the most important part of any scientist's job is effective communication. Anyone can report sequence statistics, but it is the job of a scientist to distill data into information, take the information and communicate it as a coherent story, thereby creating knowledge about how the world works (Schimel, 2012). These stories are most often told via writing and speaking, reflecting the usual format of the exams and professional products of scientists (grant proposals, manuscripts, and lectures). However, communication can take other forms (videos, animations, infographics) and requires calibration to the level of formality of the medium (popular science magazine articles, general audience public radio interviews). In all cases, the scientist needs to convince the audience that their message is worth listening to, which requires both understanding the message they want to deliver (i.e., biological and bioinformatic literacy) and tailoring it to the interests and knowledge of an audience. For both quantifiable (accepted manuscripts, successful grant applications) and abstract (successfully making and maintaining collaborations, effective lectures) professional achievements, effective communication is the whole product (Hazelett, 2025). Delivering consistent clear messages requires frequent practice, with the best communicators contributing more to the global scientific enterprise.

6 Conclusion

Omics is a glue that connects biological fields–there are few biological questions that could not be enhanced with sequence-based analyses. Though sequencing is expensive, costs have plummeted, with the first human genome costing around $300 million (not-inflation adjusted) in 2001 (Service, 2006) to nearly $100 in 2024 (Liu et al., 2024). This cheaper sequencing has led databases to grow several million times larger in the last 20 years (Hug et al., 2025), increasing access. This access is supported by the development of tools that make data selection (Speth and Orphan, 2018; Maurya et al., 2022; Woodcroft et al., 2025) and use (Wright, 2024) less computationally demanding. Other groups have spent time creating integrated systems to simplify tool use (Anvi'o, Eren et al., 2015; QIIME2; Bolyen et al., 2019; mothur; Schloss, 2020). This combination of accessible data and tools has allowed unprecedented analyses using thousands of samples to identify new biomarkers of human health (Piccinno et al., 2025) and millions of samples to assess global patterns of microbiological distribution (Rodrigues et al., 2025). In this moment, sequence data generation will continue to be exponential, fueling a demand for scientists able to answer biological questions with increasingly large datasets (Stephens et al., 2015). Scientists able to understand and effectively find, analyze, and integrate this sequence data into larger biological narratives are poised to articulate biological processes from micron to global scales, an unprecedented opportunity. We believe this review provides a foundation for just such scientists.

Author contributions

AM: Writing – original draft, Writing – review and editing. FS: Writing – review and editing, Writing – original draft.

Funding

The author(s) declared that financial support was received for this work and/or its publication. Financial support came from the U.S. National Science Foundation (grant nos. OCE 2130185, 1947522, and 2022991 to F.J.S) and the Simons Foundation (to F.J.S).

Acknowledgements

We appreciate the trainees who have been the first-pass recipients of the ideas in this manuscript, Alexis McDonnell for the conversations where this manuscript's organization was formulated, and Jennifer A. Crandall and Onur Adiguzel for reading earlier drafts of this manuscript and providing valuable feedback that shaped the final product. We thank all the scientists who have spent the time to write READMEs on their published tools, written workflows on their GitHubs, created open-access training in print and video, and continued to push the field forward with creative tools. You have collectively built a beautiful foundation for this field. Binit Shah and Sammy Q. Johnson provided useful discussion on machine learning and coding tools respectively. The material in this manuscript first appeared in the dissertation of A.J.M. (Maritan, 2025).

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651–1656. doi:10.1126/science.2047873

Al-Fatlawi, A., Menzel, M., and Schroeder, M. (2023). Is protein BLAST a thing of the past? *Nat. Commun.* 14, 8195. doi:10.1038/s41467-023-44082-5

Aldrich, J. M. (1927). The limitations of taxonomy. *Science* 65, 381–385. doi:10.1126/science.65.1686.381

Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., et al. (2014). Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11, 1144–1146. doi:10.1038/nmeth.3103

Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., et al. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* 7 (1), 13219. doi:10.1038/ncomms13219

Antipov, D., Korobeynikov, A., McLean, J. S., and Pevzner, P. A. (2016). hybrid SPA des: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* 32, 1009–1015. doi:10.1093/bioinformatics/btv688

Aramaki, T., Blanc-Mathieu, R., Endo, H., Ohkubo, K., Kanehisa, M., Goto, S., et al. (2020). KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 36, 2251–2252. doi:10.1093/bioinformatics/btz859

Aroney, S. T. N., Newell, R. J. P., Nissen, J. N., Camargo, A. P., Tyson, G. W., and Woodcroft, B. J. (2025). CoverM: read alignment statistics for metagenomics. *Bioinformatics* 41, btaf147. doi:10.1093/bioinformatics/btaf147

Asnicar, F., Thomas, A. M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., et al. (2020). Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat. Commun.* 11, 2500. doi:10.1038/s41467-020-16366-7

Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.* 79, 137–158. doi:10.1084/jem.79.2.137

Ayling, M., Clark, M. D., and Leggett, R. M. (2020). New approaches for metagenome assembly with short reads. *Brief. Bioinform.* 21, 584–594. doi:10.1093/bib/bbz020

Bainbridge, M. N., Warren, R. L., Hirst, M., Romanuik, T., Zeng, T., Go, A., et al. (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* 7, 246. doi:10.1186/1471-2164-7-246

Béjà, O., Suzuki, M. T., Koonin, E. V., Aravind, L., Hadd, A., Nguyen, L. P., et al. (2000). Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.* 2, 516–529. doi:10.1046/j.1462-2920.2000.00133.x

Bertagnolli, A. D., Maritan, A. J., Tumolo, B. B., Fritz, S. F., Oakland, H. C., Mohr, E. J., et al. (2023). Net-spinning caddisflies create denitrifier-enriched niches in the stream microbiome. *ISME Commun.* 3, 1–6. doi:10.1038/s43705-023-00315-8

Bertrand, E. M., McCrow, J. P., Moustafa, A., Zheng, H., McQuaid, J. B., Delmont, T. O., et al. (2015). Phytoplankton–bacterial interactions mediate micronutrient colimitation at the coastal antarctic sea ice edge. *Proc. Natl. Acad. Sci.* 112, 9938–9943. doi:10.1073/pnas.1501615112

Blau, W., Cerf, V. G., Enriquez, J., Francisco, J. S., Gasser, U., Gray, M. L., et al. (2024). Protecting scientific integrity in an age of generative AI. *Proc. Natl. Acad. Sci. U.S.A.* 121, e2407886121. doi:10.1073/pnas.2407886121

Bollati, E., Hughes, D. J., Suggett, D. J., Raina, J.-B., and Kühl, M. (2024). Microscale sampling of the coral gastrovascular cavity reveals a gut-like microbial community. *Anim. Microbiome* 6, 55. doi:10.1186/s42523-024-00341-4

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi:10.1038/s41587-019-0209-9

Bomar, L., Maltz, M., Colston, S., and Graf, J. (2011). Directed culturing of microorganisms using metatranscriptomics. *mBio* 2, e00012. doi:10.1128/mBio.00012-11

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* 35, 725–731. doi:10.1038/nbt.3893

Bradnam, K. (2015). L50 vs N50: that's another fine mess that bioinformatics got us into. Available online at: http://www.acgt.me/blog/2015/6/11/l50-vs-n50-thats-another-fine-mess-that-bioinformatics-got-us-into.

Brenner, S., Jacob, F., and Meselson, M. (1961). An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* 190, 576–581. doi:10.1038/190576a0

Brock, T. D. (1999). *Milestones in microbiology: 1546 to 1940.* (Washington, D.C.: American Society for Microbiology).

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* 12, 59–60. doi:10.1038/nmeth.3176

Buriak, J. M., Akinwande, D., Artzi, N., Brinker, C. J., Burrows, C., Chan, W. C. W., et al. (2023). Best practices for using AI when writing scientific manuscripts: caution, care, and consideration: creative science depends on it. *ACS Nano* 17, 4091–4093. doi:10.1021/acsnano.3c01544

Bussi, Y., Kapon, R., and Reich, Z. (2021). Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. *PLOS ONE* 16, e0258693. doi:10.1371/journal.pone.0258693

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: architecture and applications. *BMC Bioinforma.* 10, 421. doi:10.1186/1471-2105-10-421

Carballo-García, A., and Boté-Vericad, J. J. (2022). Fair data: history and present context. *Central Eur. J. Educ. Res.* 4 (2), 45–53. doi:10.37441/cejer/2022/4/2/11379

Castro-Wallace, S. L., Chiu, C. Y., John, K. K., Stahl, S. E., Rubins, K. H., McIntyre, A. B. R., et al. (2017). Nanopore DNA sequencing and genome assembly on the international space station. *Sci. Rep.* 7, 18022. doi:10.1038/s41598-017-18364-0

Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., and Parks, D. H. (2022). GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 38, 5315–5316. doi:10.1093/bioinformatics/btac672

Cheng, C., Fei, Z., and Xiao, P. (2023). Methods to improve the accuracy of next-generation sequencing. *Front. Bioeng. Biotechnol.* 11, 982111. doi:10.3389/fbioe.2023.982111

Chklovski, A., Parks, D. H., Woodcroft, B. J., and Tyson, G. W. (2023). CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* 20, 1203–1212. doi:10.1038/s41592-023-01940-w

Chuckran, P. F., Estera-Molina, K., Nicolas, A. M., Sieradzki, E. T., Dijkstra, P., Firestone, M. K., et al. (2025). Codon bias, nucleotide selection, and genome size predict *in situ* bacterial growth rate and transcription in rewetted soil. *Proc. Natl. Acad. Sci.* 122, e2413032122. doi:10.1073/pnas.2413032122

Clément, L., Emeric, D., J, G. B., Laurent, M., David, L., Eivind, H., et al. (2018). A data-supported history of bioinformatics tools. 10.48550/ARXIV.1807.06808.

Clum, A., Huntemann, M., Bushnell, B., Foster, B., Foster, B., Roux, S., et al. (2021). DOE JGI metagenome workflow. *mSystems* 6, e00804–e00820. doi:10.1128/mSystems.00804-20

Coenen, A. R., Hu, S. K., Luo, E., Muratore, D., and Weitz, J. S. (2020). A primer for microbiome time-series analysis. *Front. Genet.* 11, 310. doi:10.3389/fgene.2020.00310

Colman, D. R., Keller, L. M., Arteaga-Pozo, E., Andrade-Barahona, E., St. Clair, B., Shoemaker, A., et al. (2024). Covariation of hot spring geochemistry with microbial genomic diversity, function, and evolution. *Nat. Commun.* 15, 7506. doi:10.1038/s41467-024-51841-5

Conrad, R. E., Viver, T., Gago, J. F., Hatt, J. K., Venter, S. N., Rossello-Mora, R., et al. (2022). Toward quantifying the adaptive role of bacterial pangenomes during environmental perturbations. *ISME J.* 16, 1222–1234. doi:10.1038/s41396-021-01149-9

Crick, F. H. (1958). On protein synthesis. *Symp. Soc. Exp. Biol.* 12, 138–163.

Damsté, J. S. S., Muyzer, G., Abbas, B., Rampen, S. W., Massé, G., Allard, W. G., et al. (2004). The rise of the rhizosolenid diatoms. *Science* 304, 584–587. doi:10.1126/science.1096806

Dantzer, B., Mabry, K. E., Bernhardt, J. R., Cox, R. M., Francis, C. D., Ghalambor, C. K., et al. (2023). Understanding organisms using ecological observatory networks. Integr. Org. *Biol* 5, obad036. doi:10.1093/iob/obad036

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6, 226. doi:10.1186/s40168-018-0605-2

De Goffau, M. C., Lager, S., Salter, S. J., Wagner, J., Kronbichler, A., Charnock-Jones, D. S., et al. (2018). Recognizing the reagent microbiome. *Nat. Microbiol.* 3, 851–853. doi:10.1038/s41564-018-0202-y

De Ronne, M., Boyle, B., and Torkamaneh, D. (2025). AVITI as an alternative to illumina for low-cost genome-wide genotyping. *Genome* 68, 1–4. doi:10.1139/gen-2024-0068

DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N.-U., et al. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 311, 496–503. doi:10.1126/science.1120250

Dragone, N. B., Henley, J. B., Holland-Moritz, H., Diaz, M., Hogg, I. D., Lyons, W. B., et al. (2022). Elevational constraints on the composition and genomic attributes of microbial communities in antarctic soils. *mSystems* 7, e0133021. doi:10.1128/msystems.01330-21

Drevinek, P., Hollweck, R., Lorenz, M. G., Lustig, M., and Bjarnsholt, T. (2023). Direct 16S/18S rRNA gene PCR followed by sanger sequencing as a clinical diagnostic tool for detection of bacterial and fungal infections: a systematic review and meta-analysis. *J. Clin. Microbiol.* 61, e00338. doi:10.1128/jcm.00338-23

Dumont, M. G., Lüke, C., Deng, Y., and Frenzel, P. (2014). Classification of pmoA amplicon pyrosequences using BLAST and the lowest common ancestor method in MEGAN. *Front. Microbiol.* 5. doi:10.3389/fmicb.2014.00034

Eisenstein, M. (2023). Illumina faces short-read rivals. *Nat. Biotechnol.* 41, 3–5. doi:10.1038/s41587-022-01632-4

Ekim, B., Berger, B., and Chikhi, R. (2021). Minimizer-space de Bruijn graphs: whole-genome assembly of long reads in minutes on a personal computer. *Cell Syst.* 12, 958–968.e6. doi:10.1016/j.cels.2021.08.009

Eme, L., Tamarit, D., Caceres, E. F., Stairs, C. W., De Anda, V., Schön, M. E., et al. (2023). Inference and reconstruction of the heimdallarchaeial ancestry of eukaryotes. *Nature* 618, 992–999. doi:10.1038/s41586-023-06186-2

Eren, A. M., and Banfield, J. F. (2024). Modern microbiology: embracing complexity through integration across scales. *Cell* 187, 5151–5170. doi:10.1016/j.cell.2024.08.028

Eren, A. M., Esen, Ö. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L., et al. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* 3, e1319. doi:10.7717/peerj.1319

Evans, P. N., Parks, D. H., Chadwick, G. L., Robbins, S. J., Orphan, V. J., Golding, S. D., et al. (2015). Methane metabolism in the archaeal phylum bathyarchaeota revealed by genome-centric metagenomics. *Science* 350, 434–438. doi:10.1126/science.aac7745

Fierer, N., Leung, P. M., Lappan, R., Eisenhofer, R., Ricci, F., Holland, S. I., et al. (2025). Guidelines for preventing and reporting contamination in low-biomass microbiome studies. *Nat. Microbiol.* 10, 1570–1580. doi:10.1038/s41564-025-02035-2

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* 269, 496–512. doi:10.1126/science.7542800

Fogarty, E. C., Schechter, M. S., Lolans, K., Sheahan, M. L., Veseli, I., Moore, R. M., et al. (2024). A cryptic plasmid is among the most numerous genetic elements in the human gut. *Cell* 187, 1206–1222.e16. doi:10.1016/j.cell.2024.01.039

Fox, J. (2011). Zombie ideas in ecology. *Oikos Blog*. Available online at: http://www.oikosjournal.org/blog/zombie-ideas-ecology.

Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., et al. (2008). Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci.* 105, 3805–3810. doi:10.1073/pnas.0708897105

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152. doi:10.1093/bioinformatics/bts565

Gauthier, J., Vincent, A. T., Charette, S. J., and Derome, N. (2019). A brief history of bioinformatics. *Brief. Bioinform.* 20, 1981–1996. doi:10.1093/bib/bby063

Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., and Field, K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature* 345, 60–63. doi:10.1038/345060a0

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8, 2224. doi:10.3389/fmicb.2017.02224

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 genes. *Science* 274, 546–567. doi:10.1126/science.274.5287.546

Goodacre, N. F., Gerloff, D. L., and Uetz, P. (2014). Protein domains of unknown function are essential in bacteria. *mBio* 5, e00744. doi:10.1128/mBio.00744-13

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* 57, 81–91. doi:10.1099/ijs.0.64483-0

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi:10.1038/nbt.1883

Graf, J. S., Schorn, S., Kitzinger, K., Ahmerkamp, S., Woehle, C., Huettel, B., et al. (2021). Anaerobic endosymbiont generates energy for ciliate host by denitrification. *Nat.* 591 (7850), 445–450. doi:10.1038/s41586-021-03297-6

Gros, F., Hiatt, H., Gilbert, W., Kurland, C. G., Risebrough, R. W., and Watson, J. D. (1961). Unstable ribonucleic acid revealed by pulse labelling of *Escherichia coli*. *Nature* 190, 581–585. doi:10.1038/190581a0

Haddock, S. H. D., and Dunn, C. W. (2011). *Practical computing for biologists*. Sunderland: Sinauer Associates.

Haft, D. H., Loftus, B. J., Richardson, D. L., Yang, F., Eisen, J. A., Paulsen, I. T., et al. (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* 29, 41–43. doi:10.1093/nar/29.1.41

Hamilton, T. L., Boyd, E. S., and Peters, J. W. (2011). Environmental constraints underpin the distribution and phylogenetic diversity of nifH in the yellowstone geothermal complex. *Microb. Ecol.* 61, 860–870. doi:10.1007/s00248-011-9824-9

Härer, A., and Rennison, D. J. (2023). The biogeography of host-associated bacterial microbiomes: revisiting classic biodiversity patterns. *Glob. Ecol. Biogeogr.* 32, 931–944. doi:10.1111/geb.13675

Hatzenpichler, R., Krukenberg, V., Spietz, R. L., and Jay, Z. J. (2020). Next-generation physiology approaches to study microbiome function at single cell level. *Nat. Rev. Microbiol.* 18, 241–256. doi:10.1038/s41579-020-0323-1

Hauptfeld, E., Pappas, N., Van Iwaarden, S., Snoek, B. L., Aldas-Vargas, A., Dutilh, B. E., et al. (2024). Integrating taxonomic signals from MAGs and contigs improves read annotation and taxonomic profiling of metagenomes. *Nat. Commun.* 15, 3373. doi:10.1038/s41467-024-47155-1

Hazelett, D. J. (2025). An open letter to graduate students and other procrastinators: it's time to write. *Nat. Biotechnol.* 43, 447–450. doi:10.1038/s41587-025-02584-1

Henriques, W. S., Young, J. M., Nemudryi, A., Nemudraia, A., Wiedenheft, B., and Malik, H. S. (2024). The diverse evolutionary histories of domesticated metaviral capsid genes in mammals. *Mol. Biol. Evol.* 41, msae061. doi:10.1093/molbev/msae061

Hewson, I., Poretsky, R. S., Tripp, H. J., Montoya, J. P., and Zehr, J. P. (2010). Spatial patterns and light-driven variation of microbial population gene expression in surface waters of the oligotrophic open ocean. *Environ. Microbiol.* 12, 1940–1956. doi:10.1111/j.1462-2920.2010.02198.x

Hou, Y., and Lin, S. (2009). Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. *PloS One* 4 (9), e6978. doi:10.1371/journal.pone.0006978

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., et al. (2016). A new view of the tree of life. *Nat. Microbiol.* 1, 16048. doi:10.1038/nmicrobiol.2016.48

Hug, L. A., Hatzenpichler, R., Moraru, C., Soares, A. R., Meyer, F., Heyder, A., et al. (2025). A roadmap for equitable reuse of public microbiome data. *Nat. Microbiol.* 10, 2384–2395. doi:10.1038/s41564-025-02116-2

Hughes, J. B., Hellmann, J. J., Ricketts, T. H., and Bohannan, B. J. M. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl. Environ. Microbiol.* 67, 4399–4406. doi:10.1128/AEM.67.10.4399-4406.2001

Hwang, Y., Cornman, A. L., Kellogg, E. H., Ovchinnikov, S., and Girguis, P. R. (2024). Genomic language model predicts protein co-regulation and function. *Nat. Commun.* 15, 2880. doi:10.1038/s41467-024-46947-9

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W., and Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* 11, 119. doi:10.1186/1471-2105-11-119

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* 9, 5114. doi:10.1038/s41467-018-07641-9

Johnson, D., Batista, D., Cochrane, K., Davey, R. P., Etuk, A., Gonzalez-Beltran, A., et al. (2021). ISA API: an open platform for interoperable life science experimental metadata. *GigaScience* 10 (9), giab060. doi:10.1093/gigascience/giab060

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Kan, Y. W., and Dozy, A. M. (1978). Antenatal diagnosis of sickle-cell anaemia by DNA analysis of amniotic-fluid cells. *Lancet* 312 (8096), 910–912. doi:10.1016/s0140-6736(78)91629-x

Kapli, P., Yang, Z., and Telford, M. J. (2020). Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* 21, 428–444. doi:10.1038/s41576-020-0233-0

Kapli, P., Kotari, I., Telford, M. J., Goldman, N., and Yang, Z. (2023). DNA sequences are as useful as protein sequences for inferring deep phylogenies. *Syst. Biol.* 72, 1119–1135. doi:10.1093/sysbio/syad036

Karsch-Mizrachi, I., Arita, M., Burdett, T., Cochrane, G., Nakamura, Y., Pruitt, K. D., et al. (2025). The international nucleotide sequence database collaboration (INSDC): enhancing global participation. *Nucleic Acids Res.* 53 (D1), D62–D66. doi:10.1093/nar/gkae1058

Katara, A., Chand, S., Chaudhary, H., Chaudhry, V., Chandra, H., and Dubey, R. C. (2024). Evolution and applications of next generation sequencing and its intricate relations with chromatographic and spectrometric techniques in modern day sciences. *J. Chromatogr. Open* 5, 100121. doi:10.1016/j.jcoa.2024.100121

Katz, K., Shutov, O., Lapoint, R., Kimelman, M., Brister, J. R., and O'Sullivan, C. (2022). The sequence read archive: a decade more of explosive growth. *Nucleic Acids Res.* 50, D387–D390. doi:10.1093/nar/gkab1053

Kelly, L. W., Nelson, C. E., Haas, A. F., Naliboff, D. S., Calhoun, S., Carlson, C. A., et al. (2019). Diel population and functional synchrony of microbial communities on coral reefs. *Nat. Commun.* 10, 1691. doi:10.1038/s41467-019-09419-z

Kieft, K., and Anantharaman, K. (2022). Deciphering active prophages from metagenomes. *mSystems* 7, e00084. doi:10.1128/msystems.00084-22

Kirsip, H., and Abroi, A. (2019). Protein structure-guided hidden markov models (HMMs) as A powerful method in the detection of ancestral endogenous viral elements. *Viruses* 11, 320. doi:10.3390/v11040320

Kitzinger, K., Marchant, H. K., Bristow, L. A., Herbold, C. W., Padilla, C. C., Kidane, A. T., et al. (2020). Single cell analyses reveal contrasting life strategies of the two main nitrifiers in the ocean. *Nat. Commun.* 11, 767. doi:10.1038/s41467-020-14542-3

Knapp, A. K., Smith, M. D., Hobbie, S. E., Collins, S. L., Fahey, T. J., Hansen, G. J. A., et al. (2012). Past, present, and future roles of long-term experiments in the LTER network. *BioScience* 62, 377–389. doi:10.1525/bio.2012.62.4.9

Kohtz, A. J., Petrosian, N., Krukenberg, V., Jay, Z. J., Pilhofer, M., and Hatzenpichler, R. (2024). Cultivation and visualization of a methanogen of the phylum thermoproteota. *Nature* 632, 1118–1123. doi:10.1038/s41586-024-07631-6

Kumagai, Y., Yoshizawa, S., Nakajima, Y., Watanabe, M., Fukunaga, T., Ogura, Y., et al. (2018). Solar-panel and parasol strategies shape the proteorhodopsin distribution pattern in marine flavobacteriia. *ISME J.* 12, 1329–1343. doi:10.1038/s41396-018-0058-4

Labonté, J. M., Swan, B. K., Poulos, B., Luo, H., Koren, S., Hallam, S. J., et al. (2015). Single-cell genomics-based analysis of virus–host interactions in marine surface bacterioplankton. *ISME J.* 9, 2386–2399. doi:10.1038/ismej.2015.48

LaJeunesse, T. C., Parkinson, J. E., Gabrielson, P. W., Jeong, H. J., Reimer, J. D., Voolstra, C. R., et al. (2018). Systematic revision of symbiodiniaceae highlights the antiquity and diversity of coral endosymbionts. *Curr. Biol.* 28, 2570–2580.e6. doi:10.1016/j.cub.2018.07.008

Lander, E. S., Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi:10.1038/35057062

Landeta, C., Dávalos, A., Cevallos, M. Á., Geiger, O., Brom, S., and Romero, D. (2011). Plasmids with a chromosome-like role in rhizobia. *J. Bacteriol.* 193, 1317–1326. doi:10.1128/JB.01184-10

Lane, D. J., Pace, B., Olsen, G. J., Stahlt, D. A., Sogint, M. L., and Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proc. Natl. Acad. Sci. U. S. A.* 82, 6955–6959. doi:10.1073/pnas.82.20.6955

Lataretu, M., Krautwurst, S., Huska, M. R., Marquet, M., Viehweger, A., Braun, S. D., et al. (2025). Targeted decontamination of sequencing data with CLEAN. *Nar. Genomics Bioinforma.* 7, lqaf105. doi:10.1093/nargab/lqaf105

Leinonen, R., Sugawara, H., Shumway, M., and on behalf of the International Nucleotide Sequence Database Collaboration (2011). The sequence read archive. *Nucleic Acids Res.* 39, D19–D21. doi:10.1093/nar/gkq1019

Leung, P. M., and Greening, C. (2020). Greening lab metabolic marker gene databases.

Li, Q., Yang, F., and Zhou, C.-Z. (2025). Cyanophages: billions of years of coevolution with cyanobacteria. *Annu. Rev. Microbiol.* 79, 639–661. doi:10.1146/annurev-micro-042924-095145

Lischer, H. E. L., and Shimizu, K. K. (2017). Reference-guided *de novo* assembly approach improves genome reconstruction for related species. *BMC Bioinforma.* 18, 474. doi:10.1186/s12859-017-1911-6

Liu, X., Zhao, B., Zheng, H.-J., Hu, Y., Lu, G., Yang, C.-Q., et al. (2015). Gossypium barbadense genome sequence provides insight into the evolution of extra-long staple fiber and specialized metabolites. *Sci. Rep.* 5, 14139. doi:10.1038/srep14139

Liu, W., Li, Y., Patrinos, G. P., Xu, S., Thong, M.-K., Chen, Z., et al. (2024). The 1% gift to humanity: the human genome project II. *Cell Res.* 34, 747–750. doi:10.1038/s41422-024-01026-y

Louca, S., Parfrey, L. W., and Doebeli, M. (2016). Decoupling function and taxonomy in the global ocean microbiome. *Science* 353, 1272–1277. doi:10.1126/science.aaf4507

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964. doi:10.1093/nar/25.5.955

Lozupone, C. A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., Vázquez-Baeza, Y., et al. (2013). Meta-analyses of studies of the human microbiota. *Genome Res.* 23, 1704–1714. doi:10.1101/gr.151803.112

Lui, L. M., Majumder, E. L.-W., Smith, H. J., Carlson, H. K., Von Netzer, F., Fields, M. W., et al. (2021). Mechanism across scales: a holistic modeling framework integrating laboratory and field studies for microbial ecology. *Front. Microbiol.* 12, 642422. doi:10.3389/fmicb.2021.642422

Ma, P., Amemiya, H. M., He, L. L., Gandhi, S. J., Nicol, R., Bhattacharyya, R. P., et al. (2023). Bacterial droplet-based single-cell RNA-seq reveals antibiotic-associated heterogeneous cellular states. *Cell* 186, 877–891.e14. doi:10.1016/j.cell.2023.01.002

Maguire, F., Jia, B., Gray, K. L., Lau, W. Y. V., Beiko, R. G., and Brinkman, F. S. L. (2020). Metagenome-assembled genome binning methods with short reads disproportionately fail for plasmids and genomic islands. *Microb. Genomics* 6, mgen000436. doi:10.1099/mgen.0.000436

Major, P., Embley, T. M., and Williams, T. A. (2017). Phylogenetic diversity of NTT nucleotide transport proteins in free-living and parasitic bacteria and eukaryotes. *Genome Biol. Evol.* 9, 480–487. doi:10.1093/gbe/evx015

Maritan, A. J. (2025). *Zooming in and out on coral reef microbiomes: molecular patterns over space and time*. Montana: Montana State University. Available online at: https://scholarworks.montana.edu/handle/1/19337.

Maritan, A. J., Clements, C. S., Pratte, Z. A., Hay, M. E., and Stewart, F. J. (2025). Sea cucumber grazing linked to enrichment of anaerobic microbial metabolisms in coral reef sediments. *ISME J.* 19, wraf088. doi:10.1093/ismejo/wraf088

Marlow, J., Colocci, I., Jungbluth, S. P., Weber, N. M., Gartman, A., and Kallmeyer, J. (2020). Mapping metabolic activity at single cell resolution in intact volcanic fumarole sediment. *FEMS Microbiol. Lett.* 367, fnaa031. doi:10.1093/femsle/fnaa031

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.J.* 17, 10–12. doi:10.14806/ej.17.1.200

Martinez-Gutierrez, C. A., and Aylward, F. O. (2021). Phylogenetic signal, congruence, and uncertainty across bacteria and archaea. *Mol. Biol. Evol.* 38, 5514–5527. doi:10.1093/molbev/msab254

Maurya, A., Szymanski, M., and Karlowski, W. M. (2022). ARA: a flexible pipeline for automated exploration of NCBI SRA datasets. *GigaScience* 12, giad067. doi:10.1093/gigascience/giad067

Meier, D. V., Pjevac, P., Bach, W., Hourdez, S., Girguis, P. R., Vidoudez, C., et al. (2017). Niche partitioning of diverse sulfur-oxidizing bacteria at hydrothermal vents. *ISME J.* 11, 1545–1558. doi:10.1038/ismej.2017.37

Meziti, A., Rodriguez-R, L. M., Hatt, J. K., Peña-Gonzalez, A., Levy, K., and Konstantinidis, K. T. (2021). The reliability of metagenome-assembled genomes (MAGs) in representing natural populations: insights from comparing MAGs against isolate genomes derived from the same fecal sample. *Appl. Environ. Microbiol.* 87, e02593-20. doi:10.1128/AEM.02593-20

Mihelčić, M., Šmuc, T., and Supek, F. (2019). Patterns of diverse gene functions in genomic neighborhoods predict gene function and phenotype. *Sci. Rep.* 9, 19537. doi:10.1038/s41598-019-55984-0

Mikheenko, A., Saveliev, V., and Gurevich, A. (2016). MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 32, 1088–1090. doi:10.1093/bioinformatics/btv697

Milo, R., and Phillips, R. (2015). *Cell biology by the numbers*. New York, NY: Garland Science. doi:10.1201/9780429258770

Mirdita, M., Steinegger, M., Breitwieser, F., Söding, J., and Levy Karin, E. (2021). Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* 37, 3029–3031. doi:10.1093/bioinformatics/btab184

Mise, K., and Iwasaki, W. (2022). Unexpected absence of ribosomal protein genes from metagenome-assembled genomes. *ISME Commun.* 2, 118. doi:10.1038/s43705-022-00204-6

Mitchell, J. H., Freedman, A. H., Delaney, J. A., and Girguis, P. R. (2024). Co-expression analysis reveals distinct alliances around two carbon fixation pathways in hydrothermal vent symbionts. *Nat. Microbiol.* 9, 1526–1539. doi:10.1038/s41564-024-01704-y

Mor, B., Garhwal, S., and Kumar, A. (2021). A systematic review of hidden markov models and their applications. *Arch. Comput. Methods Eng.* 28, 1429–1448. doi:10.1007/s11831-020-09422-4

Moran, M. A., Satinsky, B., Gifford, S. M., Luo, H., Rivers, A., Chan, L.-K., et al. (2013). Sizing up metatranscriptomics. *ISME J.* 7, 237–243. doi:10.1038/ismej.2012.94

Muratore, D., Boysen, A. K., Harke, M. J., Becker, K. W., Casey, J. R., Coesel, S. N., et al. (2022). Complex marine microbial communities partition metabolism of scarce resources over the diel cycle. *Nat. Ecol. Evol.* 6, 218–229. doi:10.1038/s41559-021-01606-w

Murray, A. E., Freudenstein, J., Gribaldo, S., Hatzenpichler, R., Hugenholtz, P., Kämpfer, P., et al. (2020). Roadmap for naming uncultivated archaea and bacteria. *Nat. Microbiol.* 5, 987–994. doi:10.1038/s41564-020-0733-x

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., et al. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320, 1344–1349. doi:10.1126/science.1158441

Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., et al. (2021). A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* 39, 499–509. doi:10.1038/s41587-020-0718-6

Nelson, W. C., Tully, B. J., and Mobberley, J. M. (2020). Biases in genome reconstruction from metagenomic data. *PeerJ* 8, e10119. doi:10.7717/peerj.10119

Noble, W. S. (2009). A quick guide to organizing computational biology projects. *PLOS Comput. Biol.* 5, e1000424. doi:10.1371/journal.pcbi.1000424

Nowinski, B., Feng, X., Preston, C. M., Birch, J. M., Luo, H., Whitman, W. B., et al. (2023). Ecological divergence of syntopic marine bacterial species is shaped by gene content and expression. *ISME J.* 17, 813–822. doi:10.1038/s41396-023-01390-4

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., et al. (2022). The complete sequence of a human genome. *Science* 376, 44–53. doi:10.1126/science.abj6987

Olm, M. R., Brown, C. T., Brooks, B., and Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* 11, 2864–2868. doi:10.1038/ismej.2017.126

Ortiz, M., Leung, P. M., Shelley, G., Jirapanjawat, T., Nauer, P. A., Van Goethem, M. W., et al. (2021). Multiple energy sources and metabolic strategies sustain microbial diversity in antarctic desert soils. *Proc. Natl. Acad. Sci.* 118, e2025322118. doi:10.1073/pnas.2025322118

O'Brien, S. J., and Luo, S.-J. (2022). Taxonomic species recognition should be consistent. *Natl. Sci. Rev.* 9, nwad022. doi:10.1093/nsr/nwad022

O'Rawe, J. A., Ferson, S., and Lyon, G. J. (2015). Accounting for uncertainty in DNA sequencing data. *Trends Genet.* 31, 61–66. doi:10.1016/j.tig.2014.12.002

Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science* 276, 734–740. doi:10.1126/science.276.5313.734

Panahi, B., Mohammadzadeh Jalaly, H., and Hamid, R. (2024). Using next-generation sequencing approach for discovery and characterization of plant molecular markers. *Curr. Plant Biol.* 40, 100412. doi:10.1016/j.cpb.2024.100412

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25, 1043–1055. doi:10.1101/gr.186072.114

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., et al. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* 2, 1533–1542. doi:10.1038/s41564-017-0012-7

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., et al. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36, 996–1004. doi:10.1038/nbt.4229

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., and Hugenholtz, P. (2020). A complete domain-to-species taxonomy for bacteria and archaea. *Nat. Biotechnol.* 38, 1079–1086. doi:10.1038/s41587-020-0501-8

Piccinno, G., Thompson, K. N., Manghi, P., Ghazi, A. R., Thomas, A. M., Blanco-Míguez, A., et al. (2025). Pooled analysis of 3,741 stool metagenomes from 18 cohorts for cross-stage and strain-level reproducible microbial biomarkers of colorectal cancer. *Nat. Med.* 31, 2416–2429. doi:10.1038/s41591-025-03693-9

Porras, M. Á. G., Assié, A., Tietjen, M., Violette, M., Kleiner, M., Gruber-Vodicka, H., et al. (2024). An intranuclear bacterial parasite of deep-sea mussels expresses apoptosis inhibitors acquired from its host. *Nat. Microbiol.* 9, 2877–2891. doi:10.1038/s41564-024-01808-5

Priest, T., Oldenburg, E., Popa, O., Dede, B., Metfies, K., Von Appen, W.-J., et al. (2025). Seasonal recurrence and modular assembly of an arctic pelagic marine microbiome. *Nat. Commun.* 16, 1326. doi:10.1038/s41467-025-56203-3

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020). Using SPAdes *de novo* Assembler. *Curr. Protoc. Bioinforma.* 70, e102. doi:10.1002/cpbi.102

Raghunathan, A., Ferguson, H. R., Bornarth, C. J., Song, W., Driscoll, M., and Lasken, R. S. (2005). Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* 71, 3342–3347. doi:10.1128/AEM.71.6.3342-3347.2005

Ramoneda, J., Hoffert, M., Stallard-Olivera, E., Casamayor, E. O., and Fierer, N. (2024). Leveraging genomic information to predict environmental preferences of bacteria. *ISME J.* 18, wrae195. doi:10.1093/ismejo/wrae195

Rho, M., Tang, H., and Ye, Y. (2010). FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.* 38, e191. doi:10.1093/nar/gkq747

Ricci, F., Leung, P. M., Hutchinson, T., Nguyen-Dinh, T., Frank, A. H., Hood, A. V. S., et al. (2025). Chemosynthesis enhances net primary production and nutrient cycling in a hypersaline microbial mat. *ISME J.* 19, wraf117. doi:10.1093/ismejo/wraf117

Riley, R., Bowers, R. M., Camargo, A. P., Campbell, A., Egan, R., Eloe-Fadrosh, E. A., et al. (2023). Terabase-scale coassembly of a tropical soil microbiome. *Microbiol. Spectr.* 11, e00200-23. doi:10.1128/spectrum.00200-23

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J.-F., et al. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499, 431–437. doi:10.1038/nature12352

Rodrigues, J. F. M., Tackmann, J., Malfertheiner, L., Patsch, D., Perez-Molphe-Montoya, E., Näpflin, N., et al. (2025). The MicrobeAtlas database: global trends and insights into. *Earth's Microbial Ecosystems*. doi:10.1101/2025.07.18.665519

Rodriguez-R, L. M., and Konstantinidis, K. T. (2014). Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* 30, 629–635. doi:10.1093/bioinformatics/btt584

Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R., and Konstantinidis, K. T. (2018). Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. *mSystems* 3, e00039. doi:10.1128/mSystems.00039-18

Rognes, T., Flouri, T., Nichols, B., Quince, C., and Mahé, F. (2016). VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4, e2584. doi:10.7717/peerj.2584

Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., et al. (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.* 66, 2541–2547. doi:10.1128/AEM.66.6.2541-2547.2000

Ruff, S. E., Schwab, L., Vidal, E., Hemingway, J. D., Kraft, B., and Murali, R. (2024). Widespread occurrence of dissolved oxygen anomalies, aerobic microbes, and oxygen-producing metabolic pathways in apparently anoxic environments. *FEMS Microbiol. Ecol.* 100, fiae132. doi:10.1093/femsec/fiae132

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., et al. (1985). Enzymatic amplification of β-Globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230, 1350–1354. doi:10.1126/science.2999980

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, J. C., et al. (1977a). Nucleotide sequence of bacteriophage φX174 DNA. *Nature* 265, 687–695. doi:10.1038/265687a0

Sanger, F., Nicklen, S., and Coulson, A. R. (1977b). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* 74, 5463–5467. doi:10.1073/pnas.74.12.5463

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., et al. (2023). Next-generation sequencing technology: current trends and advancements. *Biology* 12, 997. doi:10.3390/biology12070997

Sayers, E. W., Beck, J., Bolton, E. E., Brister, J. R., Chan, J., Comeau, D. C., et al. (2024). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 52, D33–D43. doi:10.1093/nar/gkad1044

Sayers, E. W., Cavanaugh, M., Frisse, L., Pruitt, K. D., Schneider, V. A., Underwood, B. A., et al. (2025). GenBank 2025 update. *Nucleic Acids Res.* 53, D56–D61. doi:10.1093/nar/gkae1114

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270 (5235), 467–470. doi:10.1126/science.270.5235.467

Schimel, J. (2012). *Writing science: how to write papers that get cited and proposals that get funded*. (Oxford New York: Oxford University Press).

Schloss, P. D. (2020). Reintroducing mothur: 10 years later. *Appl. Environ. Microbiol.* 86, e02343-19. doi:10.1128/AEM.02343-19

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 2068–2069. doi:10.1093/bioinformatics/btu153

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). "BUSCO: assessing genome assembly and annotation completeness," in *Gene prediction*. Editor M. Kollmar (New York, NY: Springer), 227–245. doi:10.1007/978-1-4939-9173-0_14

Service, R. F. (2006). The race for the $1000 genome. *Science* 311, 1544–1546. doi:10.1126/science.311.5767.1544

Shaffer, J. P., Nothias, L.-F., Thompson, L. R., Sanders, J. G., Salido, R. A., Couvillion, S. P., et al. (2022). Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity. *Nat. Microbiol.* 7, 2128–2150. doi:10.1038/s41564-022-01266-x

Sharon, I., Morowitz, M. J., Thomas, B. C., Costello, E. K., Relman, D. A., and Banfield, J. F. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res.* 23, 111–120. doi:10.1101/gr.142315.112

Shen, W., Sipos, B., and Zhao, L. (2024). SeqKit2: a Swiss army knife for sequence and alignment processing. *iMeta* 3, e191. doi:10.1002/imt2.191

Shoemaker, A., Maritan, A., Cosar, S., Nupp, S., Menchaca, A., Jackson, T., et al. (2024). Wood–ljungdahl pathway encoding anaerobes facilitate low-cost primary production in hypersaline sediments at Great Salt Lake, Utah. *FEMS Microbiol. Ecol.* 100, fiae105. doi:10.1093/femsec/fiae105

Sielemann, K., Hafner, A., and Pucker, B. (2020). The reuse of public datasets in the life sciences: potential risks and rewards. *PeerJ* 8, e9954. doi:10.7717/peerj.9954

Söllinger, A., Tveit, A. T., Poulsen, M., Noel, S. J., Bengtsson, M., Bernhardt, J., et al. (2018). Holistic assessment of rumen microbiome dynamics through quantitative metatranscriptomics reveals multifunctional redundancy during key steps of anaerobic feed degradation. *mSystems* 3, e00038-18. doi:10.1128/msystems.00038-18

Sorek, M., Schnytzer, Y., Waldman Ben-Asher, H., Caspi, V. C., Chen, C.-S., Miller, D. J., et al. (2018). Setting the pace: host rhythmic behaviour and gene expression patterns in the facultatively symbiotic cnidarian aiptasia are determined largely by symbiodinium. *Microbiome* 6, 83. doi:10.1186/s40168-018-0465-9

Speth, D. R., and Orphan, V. J. (2018). Metabolic marker gene mining provides insight in global mcrA diversity and, coupled with targeted genome reconstruction, sheds further light on metabolic potential of the Methanomassiliicoccales. *PeerJ* 6, e5614. doi:10.7717/peerj.5614

Sriswasdi, S., Yang, C., and Iwasaki, W. (2017). Generalist species drive microbial dispersion and evolution. *Nat. Commun.* 8, 1162. doi:10.1038/s41467-017-01265-1

Stahl, D. A., Lane, D. J., Olsen, G. J., and Pace, N. R. (1984). Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science* 224, 409–411. doi:10.1126/science.224.4647.409

Staley, J. T. (2009). Universal species concept: pipe dream or a step toward unifying biology? *J. Ind. Microbiol. Biotechnol.* 36, 1331–1336. doi:10.1007/s10295-009-0642-8

Steenwyk, J. L., Li, Y., Zhou, X., Shen, X.-X., and Rokas, A. (2023). Incongruence in the phylogenomics era. *Nat. Rev. Genet.* 24, 834–850. doi:10.1038/s41576-023-00620-x

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H., and DeLong, E. F. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J. Bacteriol.* 178, 591–599. doi:10.1128/jb.178.3.591-599.1996

Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* 35, 1026–1028. doi:10.1038/nbt.3988

Steinsdóttir, H. G. R., Schauberger, C., Mhatre, S., Thamdrup, B., and Bristow, L. A. (2022). Aerobic and anaerobic methane oxidation in a seasonally anoxic basin. *Limnol. Oceanogr.* 67, 1257–1273. doi:10.1002/lno.12074

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., et al. (2015). Big data: astronomical or genomical? *PLOS Biol.* 13, e1002195. doi:10.1371/journal.pbio.1002195

Stewart, F. J., Ulloa, O., and DeLong, E. F. (2012). Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ. Microbiol.* 14, 23–40. doi:10.1111/j.1462-2920.2010.02400.x

Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Acinas, S. G., Eveillard, D., et al. (2020). Tara oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* 18, 428–445. doi:10.1038/s41579-020-0364-5

Täumer, J., Marhan, S., Groß, V., Jensen, C., Kuss, A. W., Kolb, S., et al. (2022). Linking transcriptional dynamics of CH4-cycling grassland soil microbiomes to seasonal gas fluxes. *ISME J.* 16, 1788–1797. doi:10.1038/s41396-022-01229-4

The UniProt Consortium, Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Adesina, A., et al. (2025). UniProt: the universal protein knowledgebase in 2025. *Nucleic Acids Res.* 53, D609–D617. doi:10.1093/nar/gkae1010

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551, 457–463. doi:10.1038/nature24621

Trigodet, F., Lolans, K., Fogarty, E., Shaiber, A., Morrison, H. G., Barreiro, L., et al. (2022). High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *Mol. Ecol. Resour.* 22, 1786–1802. doi:10.1111/1755-0998.13588

Tripp, H. J., Kitner, J. B., Schwalbach, M. S., Dacey, J. W. H., Wilhelm, L. J., and Giovannoni, S. J. (2008). SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* 452, 741–744. doi:10.1038/nature06776

Tschitschko, B., Esti, M., Philippi, M., Kidane, A. T., Littmann, S., Kitzinger, K., et al. (2024). Rhizobia–diatom symbiosis fixes missing nitrogen in the ocean. *Nature* 630, 899–904. doi:10.1038/s41586-024-07495-w

Tsementzi, D., Wu, J., Deutsch, S., Nath, S., Rodriguez-R, L. M., Burns, A. S., et al. (2016). SAR11 bacteria linked to ocean anoxia and nitrogen loss. *Nature* 536, 179–183. doi:10.1038/nature19068

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., et al. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* 428, 37–43. doi:10.1038/nature02340

Urayama, S., Fukudome, A., Hirai, M., Okumura, T., Nishimura, Y., Takaki, Y., et al. (2024). Double-stranded RNA sequencing reveals distinct riboviruses associated with thermoacidophilic bacteria from hot springs in Japan. *Nat. Microbiol.* 9, 514–523. doi:10.1038/s41564-023-01579-5

Uritskiy, G. V., DiRuggiero, J., and Taylor, J. (2018). MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6, 158. doi:10.1186/s40168-018-0541-1

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science* 270, 484–487. doi:10.1126/science.270.5235.484

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., et al. (1997). Characterization of the yeast transcriptome. *Cell* 88, 243–251. doi:10.1016/S0092-8674(00)81845-0

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001). The sequence of the human genome. *Science* 291, 1304–1351. doi:10.1126/science.1058040

Viver, T., Conrad, R. E., Orellana, L. H., Urdiain, M., González-Pastor, J. E., Hatt, J. K., et al. (2021). Distinct ecotypes within a natural haloarchaeal population enable adaptation to changing environmental conditions without causing population sweeps. *ISME J.* 15, 1178–1191. doi:10.1038/s41396-020-00842-5

Watson, J. D., and Crick, F. H. C. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* 171, 737–738. doi:10.1038/171737a0

Wei, X., Tan, H., Lobb, B., Zhen, W., Wu, Z., Parks, D. H., et al. (2024). AnnoView enables large-scale analysis, comparison, and visualization of microbial gene neighborhoods. *Brief. Bioinform.* 25, bbae229. doi:10.1093/bib/bbae229

Wickham, H. (2023). *R for data science*. 2nd ed. Sebastopol: O'Reilly Media.

Wilbanks, E. G., Doré, H., Ashby, M. H., Heiner, C., Roberts, R. J., and Eisen, J. A. (2022). Metagenomic methylation patterns resolve bacterial genomes of unusual size and structural complexity. *ISME J.* 16, 1921–1931. doi:10.1038/s41396-022-01242-7

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3, 160018. doi:10.1038/sdata.2016.18

Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc. Natl. Acad. Sci.* 74, 5088–5090. doi:10.1073/pnas.74.11.5088

Woodcroft, B. J., Aroney, S. T. N., Zhao, R., Cunningham, M., Mitchell, J. A. M., Nurdiansyah, R., et al. (2025). Comprehensive taxonomic identification of microbial species in metagenomic data using SingleM and sandpiper. *Nat. Biotechnol.* doi:10.1038/s41587-025-02738-1

Woyke, T., Xie, G., Copeland, A., González, J. M., Han, C., Kiss, H., et al. (2009). Assembling the marine metagenome, one cell at a time. *PLoS ONE* 4, e5299. doi:10.1371/journal.pone.0005299

Wright, E. (2024). Accurately clustering biological sequences in linear time by relatedness sorting. *Nat. Commun.* 15, 3047. doi:10.1038/s41467-024-47371-9

Wright, A., Wilkinson, M. D., Mungall, C., Cain, S., Richards, S., Sternberg, P., et al. (2024). FAIR header reference genome: a TRUSTworthy standard. *Brief. Bioinform.* 25, bbae122. doi:10.1093/bib/bbae122

Wurch, L., Giannone, R. J., Belisle, B. S., Swift, C., Utturkar, S., Hettich, R. L., et al. (2016). Genomics-informed isolation and characterization of a symbiotic nanoarchaeota system from a terrestrial geothermal environment. *Nat. Commun.* 7, 12115. doi:10.1038/ncomms12115

Xu, L., Chen, H., Hu, X., Zhang, R., Zhang, Z., and Luo, Z. W. (2006). Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Mol. Biol. Evol.* 23, 1107–1108. doi:10.1093/molbev/msk019

Yang, D., Oyaizu, Y., Oyaizu, H., Olsen, G. J., and Woese, C. R. (1985). Mitochondrial origins. *Proc. Natl. Acad. Sci.* 82 (13), 4443–4447. doi:10.1073/pnas.82.13.4443

Yang, C., Chowdhury, D., Zhang, Z., Cheung, W. K., Lu, A., Bian, Z., et al. (2021). A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput. Struct. Biotechnol. J.* 19, 6301–6314. doi:10.1016/j.csbj.2021.11.028

Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L., Williamson, S. J., Remington, K., et al. (2007). The sorcerer II global ocean sampling expedition: expanding the universe of protein families. *PLoS Biol.* 5, e16. doi:10.1371/journal.pbio.0050016

Yu, L., Jia, R., Liu, S., Li, S., Zhong, S., Liu, G., et al. (2024). Ferrihydrite-mediated methanotrophic nitrogen fixation in paddy soil under hypoxia. *ISME Commun.* 4, ycae030. doi:10.1093/ismeco/ycae030

Yutin, N., Benler, S., Shmakov, S. A., Wolf, Y. I., Tolstoy, I., Rayko, M., et al. (2021). Analysis of metagenome-assembled viral genomes from the human gut reveals diverse putative CrAss-like phages with unique genomic features. *Nat. Commun.* 12, 1044. doi:10.1038/s41467-021-21350-w

Zhang, L., Chen, T., Wang, Y., Zhang, S., Lv, Q., Kong, D., et al. (2022). Comparison analysis of different DNA extraction methods on suitability for long-read metagenomic nanopore sequencing. *Front. Cell. Infect. Microbiol.* 12, 919903. doi:10.3389/fcimb.2022.919903

Zhong, Z.-P., Du, J., Köstlbacher, S., Pjevac, P., Orlić, S., and Sullivan, M. B. (2024). Viral potential to modulate microbial methane metabolism varies by habitat. *Nat. Commun.* 15, 1857. doi:10.1038/s41467-024-46109-x

Zhou, Z., Tran, P. Q., Breister, A. M., Liu, Y., Kieft, K., Cowley, E. S., et al. (2022). METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome* 10, 33. doi:10.1186/s40168-021-01213-8

Zhou, Z., Wang, C., Cha, X., Zhou, T., Pang, X., Zhao, F., et al. (2024). The biogeography of soil microbiome potential growth rates. *Nat. Commun.* 15, 9472. doi:10.1038/s41467-024-53753-w