Biology and Fertility of Soils
https://doi.org/10.1007/s00374-026-02004-9

PERSPECTIVE

Digging deeper into soil metagenomics: Opportunities and limitations for studying the genomic potential of soil bacteria and archaea

Itır Geydirici[2] · Rubén Martínez-Cuesta[1,2] · Sebastian Bibinger[2] · Pamela Espíndola-Hernández[2] · Silvia Gschwendtner[2] · Natalia Rodríguez-Berbel[1,2] · Stefanie Schulz[2] · Ann-Christin Wicht[2] · Michael Schloter[1,2]

Received: 24 November 2025 / Revised: 7 March 2026 / Accepted: 16 March 2026

Abstract
Soil microbial ecology has been transformed by recent progress in sequencing and metagenomics, reshaping our understanding of soil microbial functions. This article examines how metagenomics can be used to study the functional potential of soil microbiomes. We highlight opportunities and limitations of current tools and workflows utilized for the analyses of metagenomic sequencing data, providing guidance for methodological thoroughness and transparency, and emphasizing the need for standardized metadata, data availability and workflow reproducibility in metagenomic studies. We encourage authors to connect microbial functional potential to genomic context, functional redundancy and taxonomic diversity. By following these practices, the scientific community can ensure that soil metagenomics delivers robust, reproducible, and ecologically meaningful insights into the 'living engine' of fertile soils.

Keywords Soil microbiome · Metagenomics · Functional annotation · MAGs · Bioinformatics standards · Environmental DNA

Introduction

Soils host billions of microorganisms per gram and are therefore one of Earth's most complex and diverse microbial habitats. These microbial communities support vital ecosystem functions such as carbon storage and nutrients cycling. Despite their importance, the majority remains unknown, as fewer than 3% can be cultivated under laboratory conditions, even with recent advances in cultivation techniques (Louca et al. 2019), which directly affects downstream bioinformatic analyses because of limited representation of environmental strains in databases. This "cultivation barrier" has restricted our understanding of the microbial communities that shape soil ecosystems. Moreover, many microorganisms rely on intricate ecological networks that involve nutrient exchanges, signalling molecules and co-metabolic interactions which are nearly impossible to reproduce under laboratory conditions (Nichols et al. 2010).

The use of molecular approaches in combination with high throughput sequencing has revolutionized microbial ecology by allowing direct access to the genetic content of entire microbial communities (Tyson et al. 2004). In the early times of molecular analyses of soil microbiomes, selected marker genes, such as the 16 S rRNA gene, were targeted for the identification of microbiome members via metabarcoding. Linking functional potential with the respective taxonomic diversity in environmental DNA became possible thanks to the introduction of metagenomics, revealing how the microbiome contributes to soil resilience and health. Continuous advances in sequencing technologies and analysis tools have further improved taxonomic assignment by providing genomic context, characterizing operon structures and gene regulation under different environmental conditions. In addition, this provides the

Itır Geydirici, Rubén Martínez-Cuesta and Stefanie Schulz contributed equally to this work.

✉ Michael Schloter
schloter@tum.de; michael.schloter@helmholtz-munich.de

[1] Chair of Environmental Microbiology, TUM School of Life Sciences, Technical University of Munich, Emil-Ramann-Straße 2, Freising 85354, Germany

[2] Research Unit Comparative Microbiome Analysis, Helmholtz Zentrum Munich, Ingolstädter Landstraße 1, Neuherberg 85764, Germany

Published online: 31 March 2026

necessary information for targeted isolation, which enables biobanking and microbiome engineering with potential bioinoculants. Thus, metagenomics represents a powerful complement to traditional cultivation and metabarcoding-based techniques, by enabling the identification of genetic determinants underlying soil functions, thereby supporting informed monitoring, management, and restoration strategies without requiring cultivation of individual community members (Albertsen et al. 2013).

Linking these genes back to their genetic context in the genome of the corresponding microorganism is extremely valuable and paves the way for an improved understanding of ecological concepts including functional redundancy, where multiple species may possess the same genomic trait, as shown by Mendes et al. (2015) in the transition from forest to agricultural soils. However, distinct lineages may deploy different regulatory mechanisms, life strategies (e.g., oligotroph vs. copiotroph), or occupy unique ecological niches (Louca et al. 2018) which strongly impact functional redundancy in real-world settings. For example, Redondo et al. (2025) used the phylogeny to predict the functional potential and niche of ammonia oxidizing archaea. Thus, resolving the specific genomic architecture of a function allows us to predict the environmental conditions under which that function is likely to be expressed, such as in Yang et al. (2024), in which nitrogen addition promoted horizontal gene transfer (HGT) events, despite a decrease in microbial diversity. Through the characterization of these functions within their ecological context, metagenomics provides a bridge between microbial biodiversity, soil fertility and resilience. In this context, genome-resolved metagenomics enables the investigation of functional redundancy by resolving how many and which taxa encode genes for specific transformation steps, and how these functions are distributed across genomes and microbial communities (Albertsen et al. 2013; Howe et al. 2014; Mendes et al. 2015).

In this article, we discuss how and which metagenomic workflows and tools can be used to link microbial diversity to functional potentials and genomic traits, allowing researchers to address questions related to: (1) functional redundancy from communities to single strains, (2) genomic traits, (3) gene regulation including operon structure, (4) to identify horizontal gene transfer (HGT). While general standardization guidelines exist within the broader bioinformatics community (Ten Hoopen et al. 2017; Vuong et al. 2022), these recommendations are rarely contextualized for soil systems. Soil represents one of the most complex microbial habitats, characterized by high matrix heterogeneity and exceptional taxonomic and functional diversity, which pose analytical challenges often underrepresented in general protocols. Consequently, methodological considerations that may be sufficient in low-diversity or host-associated systems may require adaptation when applied to soils. Here, we therefore highlight caveats specific to soil metagenomics and provide a practical overview of commonly used and accessible pipelines that accommodate different sequencing strategies and data inputs, thereby offering guidance tailored to soil-focused studies.

Sequencing strategies for soil metagenomes

Modern metagenomic studies rely on short and long-read sequencing, two complementary sequencing technologies that can be used both standalone and in combination. Short reads are often generated by Illumina platforms with lengths between 50 and 250 base pairs, delivering higher accuracy and coverage compared to long-read metagenomics at similar throughput, making them ideal for taxonomic profiling and the detection of rare taxa (Quince et al. 2017). However, they limit the reconstruction of complex and repetitive genomic regions due to insufficient read length, in combination with the high diversity of soil microbial communities.

In contrast, long-read sequencing technologies, such as nanopore and single molecule real-time (SMRT) sequencing, can produce reads spanning thousands of base pairs if high quality DNA (e.g. high purity) and large fragments are used as input. However, DNA extraction protocols for soil samples based on bead-beating often fragment the DNA, which potentially reduces the potential mean length to hundreds of base pairs, or less. Therefore, the quality control of the input DNA fragment size is crucial. These technologies facilitate the recovery of more contiguous metagenome-assembled genomes (MAGs), which provide the necessary genomic context for a higher taxonomic resolution. Long-read sequencing also enables the investigation of full operons, where specific functional traits can be linked to distinct taxonomic lineages and mobile genetic elements, which are often the primary drivers of functional adaptation in soil (Sereika et al. 2022; Yang et al. 2024). This structural continuity circumvents the chimerism frequently observed in short-read assemblies (Bickhart et al. 2022). Nevertheless, these technologies have higher error rates, which can be mitigated through hybrid approaches, where long reads are analysed in combination with short reads to improve accuracy as well as coverage (Tully et al. 2018).

Consequently, combining both sequencing approaches is increasingly considered the best practice for soil metagenomic studies. Hybrid assemblies enhance the recovery of MAGs and provide a more comprehensive view of microbial identity and encoded functional potentials. As bioinformatic tools evolve, the synergy of short and long-read sequencing provides near-complete genome reconstructions. This allows to expand the scope of metagenomic sequencing beyond functional or taxonomic profiling under changing

environmental conditions, by linking functional potential to genomic traits, such as regulon and operon structures, and further integrating more robust taxonomic assignment. Advances in metagenomics further contribute to overcoming the cultivation bottleneck. Predictive tools can now guide the targeted cultivation of previously uncultured taxa by approximating microbial growth conditions directly from MAGs and their regulon and operon structures (Barnum et al. 2024). This progress allows the shift from purely descriptive metagenomics towards an integrative method in soil microbiology that links genetic potential, ecological role, and cultivation feasibility (Espíndola-Hernández et al. 2026).

Despite advances in sequencing and bioinformatics, the high diversity of the soil microbiome remains one of the major challenges in the recovery of complete MAGs. This complexity is partly driven by the persistence of extracellular DNA, which might complicate data interpretation. While this DNA provides a long-term record of environmental shifts, it acts as biological noise (Carini et al. 2017) that can be removed using methods such as propidium monoazide (Du et al. 2025). Diversity is not the only bottleneck; factors such as uneven abundance, variable genome sizes, and widespread repetitive genomic elements further complicate the recovery of complete genomes. Additionally, the presence of closely related strains creates complexities in assembly graphs, often resulting in the fragmentation of highly conserved sequences, including single-copy genes (SCGs) used to estimate completeness, and preventing their accurate assignment to specific bins (Eisenhofer et al. 2023). This can lead to a loss of biological context, where functional genes are detected without a clear taxonomic attribution. To mitigate these challenges, deeper metagenomic sequencing and assembly strategies optimized for uneven coverage can improve the recovery of genomes from rare taxa that would otherwise remain fragmented (Bağcı et al. 2025).

From sequence to insight: streamlining soil metagenomic workflows

The analysis of metagenomic data consists of a series of consecutive steps, ranging from raw read processing to functional interpretation.

The initial step of the analysis is to demultiplex, quality control, and trim the raw reads based on parameters like length, quality, and adapter presence. After quality control, reads are assembled into contiguous sequences, or contigs, using assemblers optimized for long or short reads. Short-read assemblers such as *MEGAHIT* (Li et al. 2015) or *metaSPAdes* (Nurk et al. 2017) often deploy De Bruijn graph-based assembly approaches, while long-read assemblers rely on overlap layout consensus, like *Canu* (Koren et al. 2017) or *metaFlye* (Kolmogorov et al. 2020), optimized to handle the higher error rates of long-read sequencing. Additionally, hybrid assemblers, such as *Unicycler* (Wick et al. 2017) or *OPERA-MS* (Bertrand et al. 2019), can merge long and short reads, maximizing genomes' recovery and completeness. In the assembly step, the choice of the assembly mode depends on the sampling plan and the presence of replicates. Modes such as *merged* or *co-assembly*, as seen in Fig. 1, aim to increase the recovery of high-quality assemblies and subsequent bins by pooling sequences from multiple samples. Specifically, the merged mode (where samples are assembled individually before merging resulting contigs) serves as a vital alternative when memory constraints make standard co-assembly unfeasible (Tully et al. 2018), although with a slightly higher risk of chimerism (Tamames and Puente-Sánchez 2019). However, when samples originate from different treatments or conditions, they should be analysed in *individual* mode to avoid conflating distinct biological signals (Liu et al. 2025).

Contigs can be further structurally annotated by detecting open reading frames (ORF), with tools such as *Prodigal* (Hyatt et al. 2010) and taxonomically and functionally annotated by mapping them against reference databases such as KEGG (Kanehisa et al. 2017) or Pfam (Finn et al. 2014) using tools like *DIAMOND* (Buchfink et al. 2015). Because annotation stringency depends on parameters like identity and E-value thresholds, reporting these parameters is crucial for reproducibility. These annotations provide a catalogue of microbial taxa and their metabolic capacities. Hence enabling exploration of microbiome functional and taxonomic composition (Liu et al. 2025).

The resulting annotated contigs are subsequently binned into draft genomes. Binning algorithms cluster assembled contigs based on oligonucleotide composition and differential coverage profiles (Kang et al. 2019) using parameters like k-mer frequencies, coverage, taxonomy or GC content. Modern binning tools like *VAMB* also incorporate machine learning algorithms to improve accuracy, essential for complex microbial communities (Nissen et al. 2021). These bins are typically evaluated by tools such as *CheckM2* based on completeness, defined as the fraction of expected single-copy genes (SCGs) recovered, and contamination, which assesses the redundancy or conflicting taxonomy of these markers (Chklovski et al. 2023). Complementary tools such as *GUNC* (Genome UNClutterer) can be used to assess chimerism and non-redundant contamination by detecting phylogenetic inconsistencies across contigs within a bin, thereby helping to identify misbinning and assembly artifacts (Orakov et al. 2021). SCGs are essential genes found once per bacterial genome and are often standardized using the Bac120 set; examples include genes involved in DNA replication, such as *gyrB*, and translation, such as *rplA* (Mallawaarachchi et al. 2024; Parks et al. 2018).

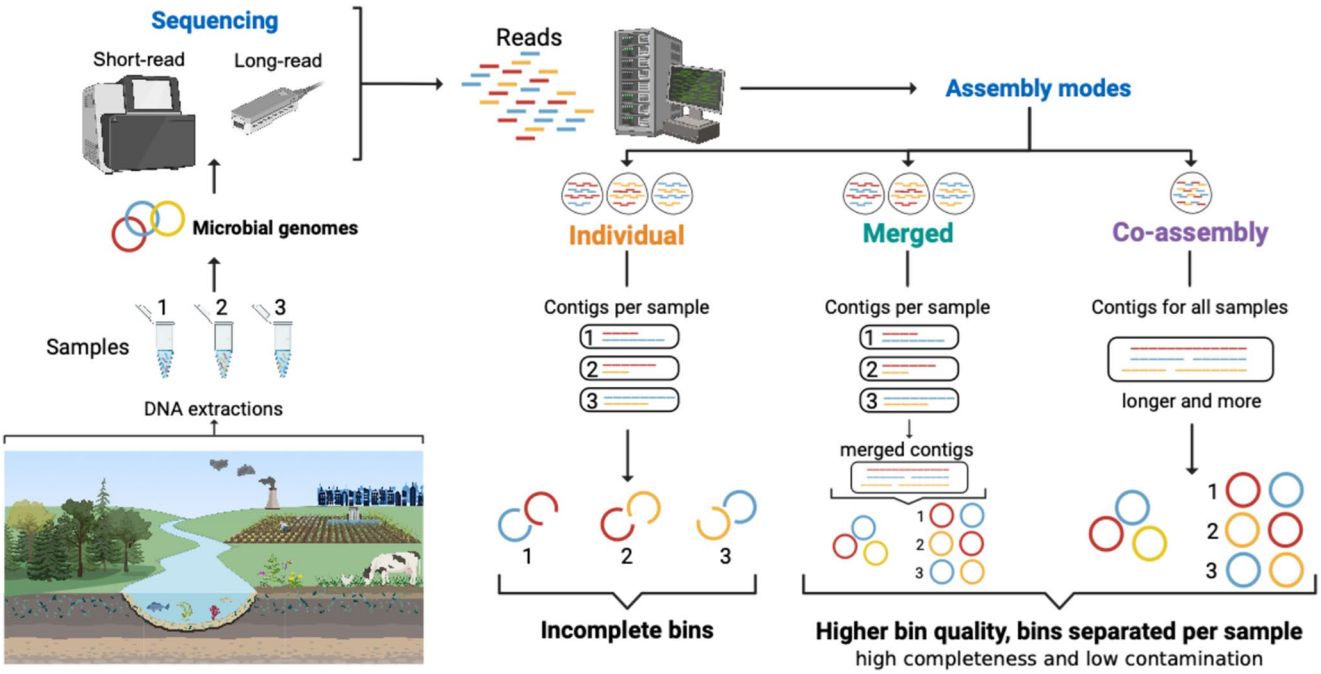

Fig. 1 Schematic description of metagenomic analysis and the different assembly options. Created with BioRender by Rodríguez-Berbel (2025)

Notably, highly conserved and repetitive sequences such as transposons often fail to be assembled or binned (Mallawaarachchi et al. 2024; Moss et al. 2020). Also, the 16 S rRNA gene, which is often considered as the "gold standard" for phylogenetic assignments if metabarcoding approaches are used (Bartoš et al. 2024), may be present in multiple different copies within the same genome (Coenye and Vandamme 2003) causing intragenomic operon heterogeneity and artificially lowering the perceived completeness. Nevertheless, conflicting read patterns generally originate during the assembly process rather than along the binning step (Trigodet et al. 2026). Additional tools integrate phylogenomic information and contig-level binning based on k-mer composition, coverage profiles, and genomic neighbourhood structure to further improve annotations. For example, tools such as *DRAM* (Shaffer et al. 2020) summarize gene-level annotations into genome-resolved metabolic profiles, while *Argo* (Chen et al. 2025) leverages long-read context to resolve the taxonomic hosts of mobile functional elements. Specialized features such as *MetaRON* enable operon prediction directly from metagenomic assemblies, while *antiSMASH* facilitates the identification of biosynthetic gene clusters, further strengthening the functional interpretation of gene organization within reconstructed genomes (Blin et al. 2025; Du et al. 2023; Zaidi et al. 2021).

While many metagenomic studies only consider bins above 50% completeness and below 10% contamination as MAGs (Nayfach et al. 2021; Parks et al. 2017), the Minimum Information about a Metagenome-Assembled Genome (MIMAG) standard classifies these as 'medium-quality', those below these thresholds as 'low-quality' and 'high-quality' as those having ≥ 90% completeness and ≤ 5% contamination (Bowers et al. 2017). However, the recovery of high-quality MAGs from soil environments is a challenging task, as exemplified by Ma et al. (2023), in which 3,641 (9.1%) high-quality MAGs were retrieved from a total of 40,039 MAGs from 9 different soil ecosystems. Soil samples typically exhibit high taxonomic richness; in such scenarios, MAG quality may vary substantially (Liu et al. 2025; Riley et al. 2023), largely due to a probabilistic sequencing bias that favours more abundant taxa and results in lower coverage for rare species at a given sequencing depth.

High-quality and near-complete MAGs are crucial for moving beyond simple associations and for assigning specific metabolic pathways to distinct taxonomic lineages, thereby improving mechanistic interpretation of soil processes (Shaffer et al. 2020). Ultimately, a full meta-analysis study would be required to ascertain how diversity levels influence the retrieval of high-quality MAGs and if increased sequencing depth solves this issue.

In addition, curated MAGs may be deposited in public repositories, such as the MGnify Genomes Catalogue (Gurbich et al. 2023), which facilitates comparative analyses within an ecological framework by mapping user-provided MAGs against biome-specific catalogues, allowing for a precise assignation of the metabolic potential.

Since metagenomic analyses involve various steps and tools, the use of end-to-end pipelines allows the analyses to be reproducible, scalable, accessible, and standardized. Pipelines such *as nf-core/mag*, which uses the workflow

Table 1 Evaluation of different pipelines for complete and specific metagenomic analyses of prokaryotes based on analysis type and usage criteria

Pipeline	nf-core/mag	SqueezeMeta	MetaWRAP	ATLAS	MetaErg	HUMAnN3	Anvi'o	MGnify	
Purpose	MAG construction	Functional profiling and MAG construction	MAG construction	MAG construction	Functional profiling	Functional profiling	Functional profiling and MAG construction	Functional profiling and MAGs construction	
Sequencing input	Short & long reads	Short & long reads	Short reads	Short reads	Short reads	Short reads	Short reads	Short reads	
Data input	Raw reads	Raw reads	Raw reads	Raw reads	Assembled contigs	Paired-end reads	Raw reads	Raw reads	
Last version	5.1.0 10.2025	1.7.2 06.2025	1.3.0 08.2020	2.19.0 07.2024	1.2.3 02.2020	3 07.2022	8 09.2023	5.0.7 | 1.2.0 06.2022 | 08.2025	
Platform	CLI	CLI	CLI	CLI	CLI	CLI	GUI + CLI	CLI	Web
User-friendliness	Medium	Medium	Low	Low	Low	High	Medium	Medium	High
Flexibility / control	High	High	High	High	Low	Low	Medium	Medium	Low
Sample amount (scale)	High	High	Medium	High	Low	Medium	Medium	High	Low
Binning	Yes	Yes	Yes	Yes	Partial	No	Yes	Yes	Yes
Metagenome annotation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Function assignment	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Taxonomy assignment	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Co-assembly	Yes	Yes	Yes	Yes	No	No	Yes	No	No
Recommended references	(Krakau et al. 2022)	(Tamames and Puente-Sánchez 2019)	(Uritskiy et al. 2018)	(Kieser et al. 2020)	(Dong and Strous 2019)	(Beghini et al. 2021)	(Eren et al. 2015)	(Richardson et al. 2023)	

*including a combination of the Anvi'o contigs and metagenomics workflows

** including a combination of the MGnify assembly analysis v5.0.7 (https://github.com/EBI-Metagenomics/assembly-analysis-pipeline) and genome-generation v1.2.0 pipelines (https://github.com/EBI-Metagenomics/genomes-generation)

†*CLI* command line interface, *GUI* graphical user interface

management system *Nextflow*, and *SqueezeMeta*, automate the processing of metagenomic data, improve consistency and enable large-scale comparisons across many samples (Krakau et al. 2022; Tamames and Puente-Sánchez 2019). The use of these pipelines requires access to a high-performance computing system, whereas online platforms such as *MGnify* (Richardson et al. 2023) allow metagenomic analyses in the absence of such resources, increasing accessibility. Some of the available pipelines and their characteristics are summarized in Table 1. Once the analysis has been completed and the data are ready to be published, it is essential to deposit raw sequencing reads in public repositories such as the NCBI Sequence Read Archive (SRA) or the European Nucleotide Archive (ENA) to enable independent reanalysis, comparative studies and long-term data accessibility (Leinonen et al. 2011; Sayers et al. 2025).

Conclusion and outlook

Metagenomics has become an essential tool for investigating the microbial drivers of key environmental processes such as nutrient turnover, stress response or plant-microbe interactions. By linking taxonomic diversity to functional potential within genomic context, genome-resolved approaches move beyond simple fingerprinting. This enables a more detailed assessment of functional redundancy, metabolic specialization and regulatory genomic structures of complex soil communities. At the same time, the high diversity and genomic complexity of soil microbiomes pose substantial analytical challenges that require careful methodological decisions and transparent reporting. Comprehensive metadata reporting, public availability of raw and processed data, reproducible analytical workflows, and standardized quality assessment of assemblies and MAGs are central to improving comparability and robustness across studies. As soil metagenomic datasets continue to increase in size and complexity, methodological consistency and a critical ecological interpretation remain essential for translating metagenomic information from isolated case studies into meaningful biological insights and to continuously improve the quality of future annotations. A comprehensive understanding of the consequences of microbial functional redundancy and resilience requires further studies including metatranscriptomics, proteomics and targeted chemical measurements, which operate on different temporal and spatial scales than DNA-based methods, as well as isolation of the respective microbes and targeted genome modifications (Orellana et al. 2019; Podlesny et al.

2026). While the bioinformatic tools and pipelines described in this article represent the current state-of-the-art, the field of metagenomics is evolving rapidly. We strongly encourage to consult the latest software documentation, repository and database updates, as new algorithms and optimized workflows are frequently released.

Author contributions I. G., R. M. C. and S.S. made substantial contributions to the conception of the work and wrote the main manuscript. N. R. B. prepared Figure 1, P. E. H., S. S. and M. S. contributed to the conception of the work. All authors reviewed and critically evaluated the manuscript, provided their intellectual properties during the revision process and approved the final version for publication.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Conflict of interest The authors declare no conflicting interests.

References

Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KL, Tyson GW, Nielsen PH (2013) Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nat Biotechnol 31:533–538. https://doi.org/10.1038/nbt.2579

Bağcı C, Negri T, Buena-Atienza E, Gross C, Ossowski S, Ziemert N (2025) Ultra-deep long-read metagenomics captures diverse taxonomic and biosynthetic potential of soil microbes. Gigascience 14:1–13. https://doi.org/10.1093/gigascience/giaf135

Barnum TP, Crits-Christoph A, Molla M, Carini P, Lee HH, Ostrov N (2024) Predicting microbial growth conditions from amino acid composition. https://doi.org/10.1101/2024.03.22.586313

Bartoš O, Chmel M, Swierczková I (2024) The overlooked evolutionary dynamics of 16S rRNA revises its role as the gold standard for bacterial species identification. Sci Rep 14:9067. https://doi.org/10.1038/s41598-024-59667-3

Beghini F, McIver LJ, Blanco-Míguez A, Dubois L, Asnicar F, Maharjan S, Mailyan A, Manghi P, Scholz M, Thomas AM, Valles-Colomer M, Weingart G, Zhang Y, Zolfo M, Huttenhower C, Franzosa EA, Segata N (2021) Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. https://doi.org/10.7554/eLife.65088

Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, Li C, Dvornicic M, Soldo JP, Koh JY, Tong C, Ng OT, Barkham T, Young B, Marimuthu K, Chng KR, Sikic M, Nagarajan N (2019) Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nat Biotechnol 37:937–944. https://doi.org/10.1038/s41587-019-0191-2

Bickhart DM, Kolmogorov M, Tseng E, Portik DM, Korobeynikov A, Tolstoganov I, Uritskiy G, Liachko I, Sullivan ST, Shin SB, Zorea A, Andreu VP, Panke-Buisse K, Medema MH, Mizrahi I, Pevzner PA, Smith TPL (2022) Generating lineage-resolved, complete metagenome-assembled genomes from complex microbial communities. Nat Biotechnol 40:711–719. https://doi.org/10.1038/s41587-021-01130-z

Blin K, Shaw S, Vader L, Szenei J, Reitz ZL, Augustijn HE, Cediel-Becerra JDD, de Crécy-Lagard V, Koetsier RA, Williams SE, Cruz-Morales P, Wongwas S, Segurado Luchsinger AE, Biermann F, Korenskaia A, Zdouc MM, Meijer D, Terlouw BR, van der Hooft JJJ, Ziemert N, Helfrich EJN, Masschelein J, Corre C, Chevrette MG, van Wezel GP, Medema MH, Weber T (2025) antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. Nucleic Acids Res 53:W32–W38. https://doi.org/10.1093/nar/gkaf334

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu W-T, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Murat Eren A, Schriml L, Banfield JF, Hugenholtz P, Woyke T (2017) Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725–731. https://doi.org/10.1038/nbt.3893

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

Carini P, Marsden PJ, Leff JW, Morgan EE, Strickland MS, Fierer N (2017) Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. Nat Microbiol 2:16242. https://doi.org/10.1038/nmicrobiol.2016.242

Chen X, Yin X, Xu X, Zhang T (2025) Species-resolved profiling of antibiotic resistance genes in complex metagenomes through long-read overlapping with Argo. Nat Commun 16:1744. https://doi.org/10.1038/s41467-025-57088-y

Chklovski A, Parks DH, Woodcroft BJ, Tyson GW (2023) CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. Nat Methods 20:1203–1212. https://doi.org/10.1038/s41592-023-01940-w

Coenye T, Vandamme P (2003) Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. FEMS Microbiol Lett 228:45–49. https://doi.org/10.1016/S0378-1097(03)00717-1

Dong X, Strous M (2019) An Integrated Pipeline for Annotation and Visualization of Metagenomic Contigs. Front Genet 10:999. https://doi.org/10.3389/fgene.2019.00999

Du R, Xiong W, Xu L, Xu Y, Wu Q (2023) Metagenomics reveals the habitat specificity of biosynthetic potential of secondary

metabolites in global food fermentations. Microbiome 11:115. https://doi.org/10.1186/s40168-023-01536-8

Du Y, Wang Z, Liu K, Chai G, Chi Y, Li T, Duan Y, Xia T, Liu D, Che R (2025) The performance of different methods in characterizing soil live prokaryotic diversity and abundance is highly variable. iMetaOmics 2:e70011. https://doi.org/10.1002/imo2.70011

Eisenhofer R, Odriozola I, Alberdi A (2023) Impact of microbial genome completeness on metagenomic functional inference. ISME Commun 3:12. https://doi.org/10.1038/s43705-023-00221-z

Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319. https://doi.org/10.7717/peerj.1319

Espíndola-Hernández P, Banerjee S, Abdulmalik AO, Andrade-Linares D, Baldi G, Berg G, Brearley FQ, Flocco CG, Galgani L, Gegeckienė L, Gschwendtner S, Hensen T, Kostic T, Ledesma-Amaro R, Maier L, Marciniak A, Ohan J, Overmann J, Rito T, Ryan M, Schulz S, Vieira S, Schloter M (2026) A trait-based framework to identify microbial keystone taxa for microbiome engineering. Cell Rep Sustain 3:100615. https://doi.org/10.1016/j.crsus.2025.100615

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230. https://doi.org/10.1093/nar/gkt1223

Gurbich TA, Almeida A, Beracochea M, Burdett T, Burgin J, Cochrane G, Raj S, Richardson L, Rogers AB, Sakharova E, Salazar GA, Finn RD (2023) MGnify Genomes: A Resource for Biome-specific Microbial Genome Catalogues. J Mol Biol 435:168016. https://doi.org/10.1016/j.jmb.2023.168016

Howe AC, Jansson JK, Malfatti SA, Tringe SG, Tiedje JM, Brown CT (2014) Tackling soil diversity with the assembly of large, complex metagenomes. Proc Natl Acad Sci USA 111:4904–4909. https://doi.org/10.1073/pnas.1402564111

Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45:D353–D361. https://doi.org/10.1093/nar/gkw1092

Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 7:e7359. https://doi.org/10.7717/peerj.7359

Kieser S, Brown J, Zdobnov EM, Trajkovski M, McCue LA (2020) ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. BMC Bioinformatics 21:257. https://doi.org/10.1186/s12859-020-03585-4

Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL, Pevzner PA (2020) metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods 17:1103–1110. https://doi.org/10.1038/s41592-020-00971-x

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive k -mer weighting and repeat separation. Genome Res 27:722–736. https://doi.org/10.1101/gr.215087.116

Krakau S, Straub D, Gourlé H, Gabernet G, Nahnsen S (2022) nf-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning. NAR Genom Bioinform 4:lqac007. https://doi.org/10.1093/nargab/lqac007

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Ten Hoopen P, Vaughan R, Zalunin V, Cochrane G (2011) The European Nucleotide Archive. Nucleic Acids Res 39:D28–D31. https://doi.org/10.1093/nar/gkq967

Li D, Liu C-M, Luo R, Sadakane K, Lam T-W (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Liu S, Rodriguez JS, Munteanu V, Ronkowski C, Sharma NK, Alser M, Andreace F, Blekhman R, Błaszczyk D, Chikhi R, Crandall KA, Della Libera K, Francis D, Frolova A, Gancz AS, Huntley NE, Jaiswal P, Kosciolek T, Łabaj PP, Łabaj W, Luan T, Mason C, Moustafa AM, Muralidharan HS, Mutlu O, Ghiasi NM, Rahnavard A, Sun F, Tian S, Tierney BT, Van Syoc E, Vicedomini R, Zackular JP, Zelikovsky A, Zielińska K, Ganda E, Davenport ER, Pop M, Koslicki D, Mangul S (2025) Analysis of metagenomic data. Nat Rev Methods Primers 5:5. https://doi.org/10.1038/s43586-024-00376-6

Louca S, Polz MF, Mazel F, Albright MBN, Huber JA, O'Connor MI, Ackermann M, Hahn AS, Srivastava DS, Crowe SA, Doebeli M, Parfrey LW (2018) Function and functional redundancy in microbial systems. Nat Ecol Evol 2:936–943. https://doi.org/10.1038/s41559-018-0519-1

Louca S, Mazel F, Doebeli M, Parfrey LW (2019) A census-based estimate of Earth's bacterial and archaeal diversity. PLoS Biol 17:e3000106. https://doi.org/10.1371/journal.pbio.3000106

Ma B, Lu C, Wang Y, Yu J, Zhao K, Xue R, Ren H, Lv X, Pan R, Zhang J, Zhu Y, Xu J (2023) A genomic catalogue of soil microbiomes boosts mining of biodiversity and genetic resources. Nat Commun 14:7318. https://doi.org/10.1038/s41467-023-43000-z

Mallawaarachchi V, Wickramarachchi A, Xue H, Papudeshi B, Grigson SR, Bouras G, Prahl RE, Kaphle A, Verich A, Talamantes-Becerra B, Dinsdale EA, Edwards RA (2024) Solving genomic puzzles: computational methods for metagenomic binning. Brief Bioinform 25:bbae372. https://doi.org/10.1093/bib/bbae372

Mendes LW, Tsai SM, Navarrete AA, de Hollander M, van Veen JA, Kuramae EE (2015) Soil-borne microbiome: linking diversity to function. Microb Ecol 70:255–265. https://doi.org/10.1007/s00248-014-0559-2

Moss EL, Maghini DG, Bhatt AS (2020) Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol 38:701–707. https://doi.org/10.1038/s41587-020-0422-6

Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, Wu D, Paez-Espino D, Chen I-M, Huntemann M, Palaniappan K, Ladau J, Mukherjee S, Reddy TBK, Nielsen T, Kirton E, Faria JP, Edirisinghe JN, Henry CS, Jungbluth SP, Chivian D, Dehal P, Wood-Charlson EM, Arkin AP, Tringe SG, Visel A, Abreu H, Acinas SG, Allen E, Allen MA, Alteio LV, Andersen G, Anesio AM, Attwood G, Avila-Magaña V, Badis Y, Bailey J, Baker B, Baldrian P, Barton HA, Beck DAC, Becraft ED, Beller HR, Beman JM, Bernier-Latmani R, Berry TD, Bertagnolli A, Bertilsson S, Bhatnagar JM, Bird JT, Blanchard JL, Blumer-Schuette SE, Bohannan B, Borton MA, Brady A, Brawley SH, Brodie J, Brown S, Brum JR, Brune A, Bryant DA, Buchan A, Buckley DH, Buongiorno J, Cadillo-Quiroz H, Caffrey SM, Campbell AN, Campbell B, Carr S, Carroll J, Cary SC, Cates AM, Cattolico RA, Cavicchioli R, Chistoserdova L, Coleman ML, Constant P, Conway JM, Mac Cormack WP, Crowe S, Crump B, Currie C, Daly R, DeAngelis KM, Denef V, Denman SE, Desta A, Dionisi H, Dodsworth J, Dombrowski N, Donohue T, Dopson M, Driscoll T, Dunfield P, Dupont CL, Dynarski KA, Edgcomb V, Edwards EA, Elshahed MS et al (2021) A genomic catalog of Earth's microbiomes. Nat Biotechnol 39:499–509. https://doi.org/10.1038/s41587-020-0718-6

Nichols D, Cahoon N, Trakhtenberg EM, Pham L, Mehta A, Belanger A, Kanigan T, Lewis K, Epstein SS (2010) Use of Ichip for

High-Throughput In Situ Cultivation of Uncultivable Microbial Species. Appl Environ Microbiol 76:2445–2450. https://doi.org/10.1128/AEM.01754-09

Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, Jensen LJ, Nielsen HB, Petersen TN, Winther O, Rasmussen S (2021) Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol 39:555–560. https://doi.org/10.1038/s41587-020-00777-4

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA (2017) metaSPAdes: a new versatile metagenomic assembler. Genome Res 27:824–834. https://doi.org/10.1101/gr.213959.116

Orakov A, Fullam A, Coelho LP, Khedkar S, Szklarczyk D, Mende DR, Schmidt TSB, Bork P (2021) GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biol 22:178. https://doi.org/10.1186/s13059-021-02393-0

Orellana LH, Hatt JK, Iyer R, Chourey K, Hettich RL, Spain JC, Yang WH, Chee-Sanford JC, Sanford RA, Löffler FE, Konstantinidis KT (2019) Comparing DNA, RNA and protein levels for measuring microbial dynamics in soil microcosms amended with nitrogen fertilizer. Sci Rep 9:17630. https://doi.org/10.1038/s41598-019-53679-0

Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542. https://doi.org/10.1038/s41564-017-0012-7

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 36:996–1004. https://doi.org/10.1038/nbt.4229

Podlesny D, Kim CY, Robbani SM, Schudoma C, Fullam A, Reimer LC, Koblitz J, Schober I, Iyappan A, Van Rossum T, Schiller J, Grekova A, Kuhn M, Bork P (2026) metaTraits: a large-scale integration of microbial phenotypic trait information. Nucleic Acids Res 54:D835–D841. https://doi.org/10.1093/nar/gkaf1241

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N (2017) Shotgun metagenomics, from sampling to analysis. Nat Biotechnol 35:833–844. https://doi.org/10.1038/nbt.3935

Redondo MA, Jones CM, Legendre P, Guénard G, Hallin S (2025) Predicting gene distribution in ammonia-oxidizing archaea using phylogenetic signals. ISME Commun 5:ycaf087. https://doi.org/10.1093/ismeco/ycaf087

Richardson L, Allen B, Baldi G, Beracochea M, Bileschi ML, Burdett T, Burgin J, Caballero-Pérez J, Cochrane G, Colwell LJ, Curtis T, Escobar-Zepeda A, Gurbich TA, Kale V, Korobeynikov A, Raj S, Rogers AB, Sakharova E, Sanchez S, Wilkinson DJ, Finn RD (2023) MGnify: the microbiome sequence data analysis resource in 2023. Nucleic Acids Res 51:D753–D759. https://doi.org/10.1093/nar/gkac1080

Riley R, Bowers RM, Camargo AP, Campbell A, Egan R, Eloe-Fadrosh EA, Foster B, Hofmeyr S, Huntemann M, Kellom M, Kimbrel JA, Oliker L, Yelick K, Pett-Ridge J, Salamov A, Varghese NJ, Clum A (2023) Terabase-scale coassembly of a tropical soil microbiome. Microbiol Spectr 11:e00200–e00223. https://doi.org/10.1128/spectrum.00200-23

Rodríguez-Berbel N (2025) Schematic description of metagenomic analysis and the different assembly options. https://BioRender.com/vs7anb9

Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Connor R, Feldgarden M, Fine AM, Funk K, Hoffman J, Kannan S, Kelly C, Klimke W, Kim S, Lathrop S, Marchler-Bauer A, Murphy TD, O'Sullivan C, Schmieder E, Skripchenko Y, Stine A, Thibaud-Nissen F, Wang J, Ye J, Zellers E, Schneider VA, Pruitt KD (2025) Database resources of the National Center for Biotechnology Information in 2025. Nucleic Acids Res 53:D20–D29. https://doi.org/10.1093/nar/gkae979

Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M (2022) Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods 19:823–826. https://doi.org/10.1038/s41592-022-01539-7

Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, Liu P, Narrowe AB, Rodríguez-Ramos J, Bolduc B, Gazitúa MC, Daly RA, Smith GJ, Vik DR, Pope PB, Sullivan MB, Roux S, Wrighton KC (2020) DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res 48:8883–8900. https://doi.org/10.1093/nar/gkaa621

Tamames J, Puente-Sánchez F (2019) SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline. Front Microbiol 9:3349. https://doi.org/10.3389/fmicb.2018.03349

Ten Hoopen P, Finn RD, Bongo LA, Corre E, Fosso B, Meyer F, Mitchell A, Pelletier E, Pesole G, Santamaria M, Willassen NP, Cochrane G (2017) The metagenomic data life-cycle: standards and best practices. Gigascience 6:1–11. https://doi.org/10.1093/gigascience/gix047

Trigodet F, Sachdeva R, Banfield JF, Eren AM (2026) Troubleshooting common errors in assemblies of long-read metagenomes. Nat Biotechnol 2:1–10. https://doi.org/10.1038/s41587-025-02971-8

Tully BJ, Graham ED, Heidelberg JF (2018) The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. Sci Data 5:170203. https://doi.org/10.1038/sdata.2017.203

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43. https://doi.org/10.1038/nature02340

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

Vuong P, Wise MJ, Whiteley AS, Kaur P (2022) Ten simple rules for investigating (meta)genomic data from environmental ecosystems. PLoS Comput Biol 18:e1010675. https://doi.org/10.1371/journal.pcbi.1010675

Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Yang J-X, Peng Y, Yu Q-Y, Yang J-J, Zhang Y-H, Zhang H-Y, Adams CA, Willing CE, Wang C, Li Q-S, Han X-G, Gao C (2024) Gene horizontal transfers and functional diversity negatively correlated with bacterial taxonomic diversity along a nitrogen gradient. NPJ Biofilms Microbiomes 10:128. https://doi.org/10.1038/s41522-024-00588-4

Zaidi SSA, Kayani MUR, Zhang X, Ouyang Y, Shamsi IH (2021) Prediction and analysis of metagenomic operons via MetaRon: a pipeline for prediction of Metagenome and whole-genome opeRons. BMC Genomics 22:60. https://doi.org/10.1186/s12864-020-07357-5