

2Pipe starts with a question: matching you with the correct pipeline for MAG reconstruction

Jeferyd Yepes-García,^{1,2} Laurent Falquet^{1,2}

AUTHOR AFFILIATIONS See affiliation list on p. 18.

ABSTRACT Whole-genome sequencing has boosted our ability to explore microbial diversity by enabling the recovery of metagenome-assembled genomes (MAGs) directly from environmental DNA. As a result, the vast availability of sequencing data has prompted the development of numerous bioinformatics pipelines for MAG reconstruction, along with challenges to identify the most suitable pipeline to perform the analysis according to the user needs. This report briefly discusses the computational requirements of these pipelines; presents the variety of interfaces, workflow managers, and package managers they feature; and describes the typical modular structure. Also, it provides a compacted technical overview of 41 publicly available pipelines or platforms to build MAGs starting from short and/or long sequences. Moreover, recognizing the overwhelming number of factors to consider when selecting an appropriate pipeline, we introduce an interactive decision-support web application, 2Pipe, that helps users to identify a suitable workflow based on their input data characteristics, desired outcomes, and computational constraints. The tool presents a question-driven interface to customize the recommendation, a pipeline gallery to offer a summarized description, and a pipeline comparison based on key factors used for the questionnaire. Beyond this and foreseeing the release of novel pipelines in the near future, we include a quick form and detailed instructions for developers to append their workflow in the application. Altogether, this review and the application equip the researchers with a general outlook of the growing metagenomics pipeline landscape and guide the users toward deciding the workflow that best fits their expectations and infrastructure.

KEYWORDS metagenomics, metagenome-assembled genome, pipeline benchmarking, workflow manager

Metagenomics has advanced the study of microbial communities by diminishing the need for cultivation and enabling direct DNA sequencing from complex environments such as the human body, soil, or aquatic ecosystems (1). This has been possible thanks to the combination of high-quality and high-throughput sequencing technologies and recent advances in bioinformatics tools, increasing the scope and resolution at which the microbiota can be explored (2). Moreover, reconstructing metagenome-assembled genomes (MAGs) has enabled the genomic characterization of uncultured microorganisms, the discovery of previously unknown species, the inference of the community's metabolic and functional potential, the establishing of ecological interactions, and the detection of evolutionary mechanisms (2, 3).

Considering the ecological importance of the MAGs, genomic criteria have been designed to determine whether a recovered bin (draft genome) truly represents a MAG or not. For instance, the Minimum Information about MAGs guidelines establish that MAGs can be classified into three quality tiers: high-quality drafts (HQ), medium-quality drafts (MQ), and low-quality drafts; the specific details regarding the genomic quality metrics used for this classification were introduced by Bowers et al. (4). MAGs can also

Editor Alexander Mahnert, Medizinische Universität Graz, Graz, Austria

Address correspondence to Laurent Falquet, laurent.falquet@unifr.ch.

The authors declare no conflict of interest.

See the funding table on p. 18.

Published 29 January 2026

Copyright © 2026 Yepes-García and Falquet. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](#).

be divided into species-assigned MAGs (SMAGs), that is, MAGs for which a species can be assigned, and hypothetical MAGs (HMAGs), that is, MAGs that are supposedly genomes of novel species, according to the genome heterogeneity spectrum proposed by Setubal (5).

In a simplified manner, MAGs are obtained through bioinformatics pipelines that include quality control, assembling and binning the sequences, and the annotation of each recovered genome (6) (Fig. 1). These pipelines are then responsible for the correct MAG assembly and have a key role at extracting meaningful information about the structure and function of microbial communities (1). Through their orchestrated workflow, they simplify and standardize the common tasks that are required to achieve HQ MAGs, reducing the occurrence of manual errors by improving reproducibility (7). Nonetheless, pipeline choice may not be a trivial decision, given that it should be based on the alignment between user needs and workflow key factors such as the type of sequencing data they handle (short or long reads, or both), analytical functions (i.e., co-assembly, sequential co-assembly, taxonomic profiling, and eukaryotic recovery), and computational environment (e.g., availability of local resources, high-performance computing [HPC] infrastructure, or web-based tools). Therefore, pipeline selection can quickly become an overwhelming process and challenge researchers with a vast landscape of options, delaying the start of the analysis or even not obtaining the expected results since the incorrect workflow was chosen.

Here, we describe the general workflow followed by bioinformatics pipelines to recover MAGs directly from metagenomics data, discussing important aspects the pipelines feature, such as the tools they encompass and the type of data they can handle. We also succinctly highlight major considerations regarding pipeline execution, storage needs, and computational infrastructure. Likewise, we provide a compact overview of 41 publicly available pipelines, suites, or platforms that enable MAG reconstruction and/or annotation starting from short and/or long sequences. Finally, considering the main practical features of each pipeline and aiming at aiding researchers in navigating the ecosystem of workflows, we also introduce 2Pipe, a decision-support web application designed to match metagenomics community users with the most suitable MAG pipeline based on their input data, technical requirements, bioinformatics experience, and preferred interface.

PIPELINE WORKFLOW, TOOLS, AND BENCHMARKS

The traditional computational workflow to build and annotate MAGs involves several steps (6); Fig. 1 introduces the general series of steps to potentially achieve MQ or HQ MAGs, along with some common software integrated by the pipelines. In brief, it begins with quality control, where low-quality reads and contaminants are removed (8, 9); when required, some pipelines include the option to discard host organism sequences (10). This is followed by the assembly step, where reads are extended to create

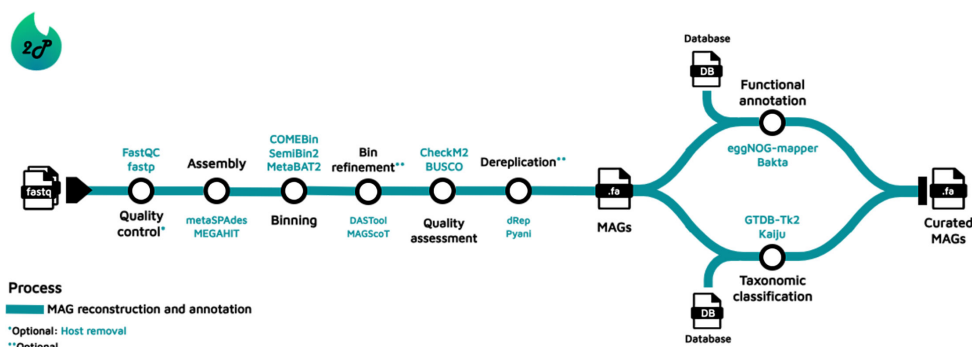


FIG 1 Usual bioinformatics workflow followed to perform MAG recovery, classification, and annotation. Some common tools incorporated by the pipelines are highlighted.

contiguous sequences, also called contigs. The contigs are then grouped into bins that ideally represent individual genomes, based on the sequence composition and coverage patterns, among other genomic features (11). Optionally, the bins are subjected to a process of refinement when researchers consider it necessary (12, 13). Afterward, these bins are evaluated for common metrics such as completeness and contamination to assess their quality and hence determine whether they constitute MAGs or not, using the criteria previously mentioned (14). In some cases, the workflows can encompass dereplication tools or modules that attempt to curate the MAG set by clustering them according to their genomic similarity and thus selecting a representative MAG from each cluster (15). To conclude with the workflow, the MAGs are then taxonomically affiliated and functionally annotated to assign biological meaning, extracting insights related to their identity and potential roles within their microbial communities (16, 17). A detailed description of the tools for each step of the workflow is provided by Yang et al. (6), and Wajid et al. (18) present an overview of the typical analysis pipeline and software using an interesting music analogy.

We present on Table 1 the tools and third-party software for quality control, assembly, binning, refinement, taxonomic classification, and functional annotation that each of the pipeline documented here encompasses. Additionally, a detailed description of the main workflow for each of them can be found in File S1, where important technical considerations such as the type of input (short reads, long sequences, or both), tools employed at each step, advantages, limitations, and/or special features they depict are presented.

As previously mentioned, the MAG reconstruction workflow is triggered with the quality control of the raw reads to ensure the accuracy and integrity of downstream analyses. Usually, the reads received from the sequencing facility contain sequencing errors, low-quality bases, adapters, and contaminant sequences (e.g., host or environment DNA) that can lead to fragmented assemblies or chimeric bins if not properly removed (6, 10). These issues are addressed by filtering and trimming, if required, the raw reads using tools like Trimmomatic (9), fastp (8), Cutadapt (139), or BBTools (21). In the case of contamination removal, tools such as KneadData (82), Bowtie2 (55), Minimap2 (88), BWA (75), or Kraken (either v1 or v2) (62, 108) are commonly used to screen and remove host-derived or non-target reads. For long-read data (Oxford Nanopore known as ONT or Pacific Biosciences known as PacBio), Filtlong (42), Nanofilt (43), and Porechop (90) are used for length filtering, quality trimming, and adapter removal. The pipeline quality control and contamination removal modules are often complemented by FastQC (20) or MultiQC (103), the standard methods to evaluate the overall quality and report it; NanoPack2 and pycoQC (215) provide detailed quality summaries for long reads. In a recent report, Gao et al. (10) compared many available tools for removing host contamination, namely, KneadData, Bowtie2, KMCP (260), BWA, KrakenUniq (35), and Kraken2, highlighting the superior performance depicted by Bowtie2 in terms of resource usage, while Kraken2 demonstrated the shortest execution times; the accuracy of Bowtie2, KneadData, and BWA outperformed the rest of the tools.

Furthermore, the assembly step represents the core of the process since it reconstructs longer contiguous sequences from the high-quality reads. Notably, assembling metagenomics data sets faces complex challenges due to varying species abundance, uneven coverage, and the presence of closely related organisms (261). The short-read assemblers rely mainly on two strategies: overlap-layout-consensus, which aligns overlapping reads to build contigs, and the more widely used De Bruijn graph method, which decomposes reads into k-mers and represents them as nodes and edges in a graph (261). MEGAHIT (22), metaSPAdes (30), and IDBA-UD (31) are examples of tools that implement the De Bruijn graph approach, incorporating heuristics to address the coverage variation and strain complexity. In contrast, assemblers for long-read data such as metaFlye (45), Canu (111), and hifiasm (157) are designed to apply graph-based algorithms optimized for higher error rates and uneven depth. In some cases, hybrid strategies are employed, combining long reads for structural resolution with accurate

TABLE 1 Software and tools incorporated by each pipeline or web-based platform

No.	Pipeline/platform	Quality control preprocessing	Assembly ^a	Binning	Quality assessment	Bin refinement	Taxonomic annotation ^b	Functional annotation ^b	Other
1	Ancient DNA (19)	FastQC (20), fastp (8), and BBTools (21)	Bowtie2 and MEGAHIT (22)	CONCOCT (23), MaxBin (24), and MetaBAT (25)	CheckM (26)	DASTool (12)	GTDB-Tk (17)		mapDamage2 (27)
2	Anvi'o (28)	Illumina-utils (29)	metaSPAdes (30), MEGAHIT, and IBDA-UD (31)	MetaBAT2 (32), CONCOCT, MaxBin2 (33), and BinSanity (34)		DASTool	KrakenUniq (35) and Centrifuge (36)	DIAMOND (37) (NCBI Cluster of Orthologs Groups (COG) [38]), Pyrodigal (39), and HMMER (40)	
3	Aviary (41)	FastQC, Filtlong (42), NanoPack2 (43), and SingleM (44)	metaSPAdes, MEGAHIT, metaFlye (45), and Unicycler (46)	MetaBAT2, MetaBAT, MaxBin2, VAMB (47), CONCOCT, and Rosella (48)	CheckM, metaQUAST (49), and CoverM (50)	DASTool	GTDB-Tk	Prodigal (51) and DIAMOND (eggNOG [52])	Lorikeet (53)
4	BugBuster (54)	fastp and Bowtie2 (55)	MEGAHIT	METABAT2, SemiBin2 (56), and COMEBin (57)	CheckM2 (58)	MetaWRAP-native module (59)	GTDB-Tk2 (60)	Prodigal and MetaCerberus (61)	Kraken2 (62), Sourmash (63), and deepARG (64)
5	BV-BRC ^c (65)	TrimGalore (66), BBTools, and BLAST (67)	metaSPAdes and MEGAHIT	PATRIC metagenome binning service (68)	EvalG and EvalCon (69)		RASThK (70)		VIGOR4 (71) and Mat_Peptide (72)
6	DATMA (73)	Trimmomatic (9), FastQC, FLASH2 (74), and BWA (75)	metaSPAdes, Velvet (76), and MEGAHIT	CLAME (77)	CheckM		BLAST and Kaiju (78)	Prodigal and GeneMark (79)	Krona (80)
7	EasyMetagenome (81)	KneadData (82), HostPurge (81), and FastQC	metaSPAdes and MEGAHIT	MetaWRAP-native module	CoverM and CheckM2	MetaWRAP-native module	GTDB-Tk2	MetaProdigal (83) and eggNOG-mapper (84)	dRep (85), Kraken2, Bracken (86), and HUMAnN3 (82)
8	EasyNanoMeta (87)	fastp, Minimap2 (88), SAMtools (89), Porechop (90), and BEDTools (91)	metaFlye, OPERA-MS (92), metaSPAdes, MetaPlatanus (93), and NextPolish (94)	SemiBin2, MetaBAT2, MaxBin2, CONCOCT, and VAMB	CheckM2		GTDB-Tk2 and PhyloPhlAn (95)	Prokka (96)	Kraken2 and Centrifuge
9	Eukfinder (97)	Bowtie2 and Trimmomatic	metaSPAdes	MyCC (98) and METAXA2 (99)			Centrifuge and PLAST (100)		
10	EURYALE (MEDUSA) (101, 102)	FastQC, fastp, Bowtie2, and MultiQC (103)	MEGAHIT				Kaiju and Kraken2	DIAMOND (NCBI nr [104])	Krona
11	Galaxy ^d (105)	FastQC, Seqtk (106), and Trimmomatic	metaSPAdes	MaxBin2			GTDB-Tk2 and Contig Annotation Tool (CAT) (107)	Prokka	Kraken (108)
12	GEN-ERA (109)	fastp and FastQC	SPAdes (110), metaSPAdes, Canu (111), metaFlye, Pilon (112), and RagTag (113)	MetaBAT2 and CONCOCT	CheckM, GUNC (114), CheckM2, EukCC (115), BUSCO (116), Physeter (117), Kraken, and QUAST (118)	AMAW (119), BRAKER2 (120), and GTDB-Tk	Prodigal, Mantis (121), and Anvi'o scripts (Kyoto Encyclopedia of Genes and Genomes, KEGG [122])	OrthoFinder (123)	
13	HIFI-MAG (124)			MetaBAT2 and SemiBin2	CheckM2	DASTool	GTDB-Tk2		
14	IDseq ^e (125)	Trimmomatic, STAR (126), Bowtie2, and CD-HIT (127)	SPAdes and Bowtie2				GSNAPL (128) and RAP-search2 (129)		
15	IMG/M ^f (130)			SemiBin2	CheckM		GTDB-Tk	Prodigal, GeneMarkS-2 (131), and HMMER (NCBI COG, Pfam [132], and TIGRFAMs [133])	EukCC, SignalP (134), and TMHMM (135)
16	JAMS (136)	Trimmomatic and Bowtie2	MEGAHIT and SPAdes				Kraken2	Prokka and Samtools	

(Continued on next page)

TABLE 1 Software and tools incorporated by each pipeline or web-based platform (Continued)

No.	Pipeline/platform	Quality control preprocessing	Assembly ^a	Binning	Quality assessment	Bin refinement	Taxonomic annotation ^b	Functional annotation ^b	Other
17	KBase ^c (138)	FastQC, Trimmomatic, and Cutadapt (139)	metaSPAdes, MEGAHIT, and IBDA-LUD	MetaBAT2, CONCOCT, and MaxBin2	CheckM	DASTool	RASTK and GTDB-Tk	InterProScan (137) Prokka, dbCAN3 (140), and DRAM (141)	BEDTools OMEGA (142), ModelSEED2 (143), Kaiju, FastANI (144), dRep, FastTree2 (145), and Muscle5 (146)
18	MAGNETO (147)	fastp, Bowtie2, and FastQscreen (148)	MEGAHIT and Simka (149)	MetaBAT2	CheckM		GTDB-Tk	Prodigal, Linclust (150), CD-HIT, eggNOG-mapper	mOTUs (151) and dRep
19	MAGO (152)	FastQC and fastp	metaSPAdes, MEGAHIT, and IBDA-LUD	MaxBin2, MetaBAT, CONCOCT, and BinSanity	CheckM		GTDB-Tk	Prokka	Roary (153), eZTree (154), and FastANI
20	Mapler (155)	FastQC	metaMDBG (156), hifiasm (157), metaFlye, OPERA-MS, and Minimap2	MetaBAT2	CheckM2 and metaQUAST		GTDB-Tk2 and Kraken2		KAT (158)
21	MetaGEM (159)	fastp	MEGAHIT and BWA	MetaBAT2, CONCOCT, and MaxBin2	MetaWRAP-native module		GTDB-Tk	Prokka	Roary, CarveMe (160), SMETANA (161), MEMOTE (162), and GRID (163)
22	MetaGenePipe (164)	Trimmomatic, TrimGalore, and FastQC	MEGAHIT				DIAMOND (SwissProt [165])	Prodigal and HMMER (166) (Kofam [167])	BLAST
23	Metagenome-Atlas (168)	BBDTools	MEGAHIT and metaSPAdes	MetaBAT2, MaxBin2, and VAMB	BUSCO, CheckM, and CheckM2	DASTool	GTDB-Tk	Prodigal, eggNOG, and Distilled and Refined Annotation of Metabolism pipeline (DRAM)	dRep
24	Metagenomics-Toolkit (169)	fastp, Porechop, Filtlong, NanoPack2, KMC (170), and Nonpareil (171)	metaFlye, metaSPAdes, MEGAHIT, and Assembler Resource Estimator (169)	MetaBAT2, MetaCoAG (172), and MetaBinner (173)	CheckM	MAGS-coT (13)	MMSeqs2 taxonomy (174) and GTDB-Tk2	Prodigal, Prokka, and RGI (175)	CarveMe, SMETANA, MEMOTE, gapseq (176), Pyani (177), and SANS (178)
25	Metaphor (179)	FastQC, fastp, and MultiQC	MEGAHIT	VAMB, MetaBAT2, and CONCOCT	metaQUAST	DASTool	DIAMOND (NCBI COG)	Prodigal and Prokka	
26	metagWGS (180)	FastQC, Cutadapt, Sickle (181), SAMtools, and BWA	metaSPAdes, MEGAHIT, hifiasm, and metaFlye	MetaBAT2, CONCOCT, and MaxBin2	metaQUAST	Binette (182)	GTDB-Tk2	Prodigal and eggNOG-mapper	dRep and Kaiju
27	MetaWRAP (59)	FastQC and TrimGalore	metaSPAdes and MEGAHIT	MetaBAT2, CONCOCT, and MaxBin2	CheckM	MetaWRAP-native module	Kraken and BLAST	Prokka	Kraken and Blobology (183)
28	MG-TK (184)	Trimmomatic, Porechop, Kraken, Kraken2, and SDM (185)	SPAdes, MEGAHIT, Flye (186), and metaMDBG	MetaBAT2, SemiBin2, and MetaDecoder (187)	CheckM and CheckM2		GTDB-Tk	Prodigal and DIAMOND (KEGG Carbohydrate-Active enZymes, CAZy [188] and eggNOG)	mOTUs2 (189), MetaPhlAn (190), FreeBayes (191), riboFinder (192), and BCFtools (89)
29	MGenify ^c (193)	Trimmomatic and Biopython (194)	metaSPAdes				DIAMOND (UniRef90 [195])	Prodigal, FragGeneScan (196), InterProScan, eggNOG-mapper, and HMMER (40)	mOTUs2 and antiSMASH (197)

(Continued on next page)

TABLE 1 Software and tools incorporated by each pipeline or web-based platform (Continued)

No.	Pipeline/platform	Quality control preprocessing	Assembly ^a	Binning	Quality assessment	Bin refinement	Taxonomic annotation ^b	Functional annotation ^b	Other
30	MOSHPIIT ^c (198)	Cutadapt and Bowtie2	SPAdes and MEGAHIT	MetaBAT2	QUAST and BUSCO	Sourmash	Kraken2 and Kaiju	eggNOG-mapper and DIAMOND (eggNOG and CAZY)	
31	MUFFIN (199)	fastp and FilTlong	SPAdes, Flye, and Unicycler	MetaBAT2, CONCOCT, and MaxBin2	CheckM	MetaWRAP-native module	Sourmash (Genome Taxonomy Database, GTDB [200])	eggNOG-mapper	Salmon (201) and Trinity (202)
32	NanoPhase (203)	Filtlong	metaFlye, Racon (204), and medaka (205)	MetaBAT2 and MaxBin2	CheckM and QUAST	MetaWRAP-native module	GTDB-Tk	Prodigal and DIAMOND (UniProtKB [206])	
33	nF-core/mag (207)	fastp, AdapterRemoval (208), Bowtie2, BBTools, Trimmomatic, FastQC, Porechop, FilTlong, and NanoPack2	MEGAHIT, metaSPAdes, Flye, metaMDBG, and hybridSPAdes (209)	MetaBAT2, CONCOCT, and MaxBin2	BUSCO, CheckM, CheckM2, GUNC, and QUAST	DASTool	GTDB-Tk2 and CAT	Prodigal, Prokka, and MetaEuk (210)	Kraken2, MultiQC, Centrifuge, PyDamage (211) geNomad (212), and Tiara (213)
34	ngs-preprocess MpCap Bacannot (214)	Porechop, Nanopack2, pycoQC (215), and fastp	SPAdes, Flye, Canu, Unicycler, Shovill (216), HASLR (217), Raven (218), Shasta (219), wrdbg2 (220), and Pilon	MetaBAT2	CheckM, GUNC, and CoverM		GTDB-Tk2	Prokka, antiSMASH, KofamScan (167), KEGGDecoder (221), Bakta (16), and Barnmap (222)	AMRFinderPlus (223), CARD-RGI, BEDTools, Phigaro (224), VFDB (225), PlasmidFinder (226), MLST (227), Platon (228), PHASTER (229), ARGminer (230), and ResFinder (231) mOTUs, MultiQC, MetaPhlAn4 (82), Salmon, gffquant (234), and kallisto (235)
35	nIMP3 (232)	BWA, Samtools, BBTools, FastQC, Kraken2, and SortMeRNA (233)	MEGAHIT	MetaBAT2	CheckM, GUNC, and GUNC		GTDB-Tk2	Prodigal and eggNOG-mapper	Barnap, RGI (175), ABRIcate (239) (MEGARes [240] and VFDB), Seqtk, Macrel (241), and Mash (242)
36	SnakeMAGs (236)	Illumina-utis, Trimmomatic, and Bowtie2	MEGAHIT	MetaBAT2	CheckM, CheckM2, and CompareM 245		GTDB-Tk2	Prodigal, MUMmer 246, HMMER, and Barnap	DIAMOND (NCBI COG, KEGG), SQTtools 247, and POGENOM 248
37	SPIRE (237)	NGLess (238)	MEGAHIT, BWA, and Samtools	MetaBAT2			GTDB-Tk2		Kraken
38	SqueezeMeta (243)	PRINSEQ 244, Trimmomatic, and SAMtools	MEGAHIT, SPAdes, Canu, and Flye	MetaBAT2, CONCOCT, and MaxBin2	CheckM, CheckM2, and CompareM 245	DASTool	GTDB-Tk2		
39	Sunbeam (249)	Trimmomatic, Cutadapt, Komplexity (249), and BWA	MEGAHIT						

(Continued on next page)

TABLE 1 Software and tools incorporated by each pipeline or web-based platform (Continued)

No.	Pipeline/platform	Quality control	preprocessing	Assembly ^a	Binning	Quality assessment	Bin refinement	Taxonomic annotation ^b	Functional annotation ^b	Other
40	VEBA (250)	KneadData, fastp, BBTools, Bowtie2, NanoPack2, and Minimap2		metaSPAdes, SPAdes, rnaSPAdes (251), MEGAHIT, Flye, and metaFlye	MetaBAT2, CONCOCT, MaxBin2, and SemiBin2	CheckM2, Tiara, CheckV (252), BUSCO, and CoverM	Binette	GTDB-Tk2, MetaEuk, geNomad, and VirFinder (253)	Prodigal, DIAMOND (UniRef50/90, MIBIG [254], VFDB, and CAZy) HMMER (Pfam, NCBIfam-AMR [223], Antifam [255], and KOfam), and MicrobeAnnotator (256)	antiSMASH, Muscle5, FastTree2, FastANI, sylph (257), and HUMAnN3
41	WGS2+/LoRA ^c (258)	KneadData, fastp, and Kraken2		metaSPAdes, metaFlye, MiniMap2, and Samtools	MetaBAT2	CheckM and CheckM2		GTDB-Tk2	Prodigal, eggNOG-mapper, and MinPath (259)	SortMeRNA, Krona, Trinity, and AMRFinderPlus

^aNot all the tools included here are assemblers; some of them are alignment or polishing tools that the pipeline's assembly module includes.

^bIf a database is necessary, it is mentioned in parentheses.

^cWe describe here the main workflow to recover MAGs published on these platforms or suites. However, they may offer many more services, tools or pipelines to meet any other need that the users demand.

short reads for initial graph assembly, as implemented in tools like OPERA-MS (92) and hybridSPAdes (209).

To this date, some authors have attempted to provide a comprehensive and unbiased benchmark of the most popular assemblers using different data sets that vary in complexity. For instance, Goussarov et al. (262) developed a comparison among short, long, and hybrid assemblers using a complex mock metagenome with more than 200 bacterial strains, demonstrating that metaSPAdes can achieve superior performance in terms of assembly fragmentation and chimerism when using Illumina reads, while Canu depicted the best metrics (chimerism and fragmentation) for ONT data. A similar conclusion regarding short-read assemblers was presented by Meyer et al. (263), where although MEGAHIT and metaSPAdes showed similar performance, metaSPAdes delivers fewer fragmented assemblies using simulated mouse gut sequences that enclosed more than 540 species. During the analysis of data sets enclosing mixed real metagenomic reads and reads from known genomes, Wang et al. (264) reported MEGAHIT as the most efficient assembler, while metaSPAdes outperformed MEGAHIT, IDBA-UD, and Faucet (265) in terms of integrity and continuity at the species level, and it showed the overall best performance at the strain level.

In the case of hybrid assembly, Brown et al. (266) showed boosted contiguity and reduced assembly errors with either hybridSPAdes or OPERA-MS, although yielding frequent misassemblies during *in silico* spike-in experiments using real and simulated reads. Nevertheless, assemblies obtained with these hybrid same tools were less complete and more fragmented than long-read only assemblies using the same data set of more than 200 bacterial strains mentioned above (262). As a result, Goussarov et al. suggest constructing the assembly using long reads complemented with short-read polishing, when the coverage is sufficient.

Accompanying the core of the pipelines, binning tools also represent an important step to reconstruct as accurately as possible the genomes present in the microbial communities. Classical binning strategies can be divided into different categories: (i) algorithms based on the genomic composition (mainly k-mer frequencies and GC content), (ii) approaches using read depth (coverage) profiles across multiple samples to link contigs with similar abundance patterns, and (iii) combined strategies that integrate both sequence composition and coverage signals (6). Classical tools based on these strategies such as MetaBAT2 (32), MaxBin2 (33), and CONCOCT (23) have been widely incorporated into the workflows given their efficiency and robustness. Nevertheless, more recent methods leverage machine learning and semi-supervised approaches to improve the resolution in more complex environments such as soil or ocean (267). SemiBin2 (56) represents an example of these recent strategies as it uses deep learning with semi-supervised contrastive learning to incorporate both intrinsic sequence information and external reference genomes. Another example is represented by COMEBin (57), which employs graph neural networks to integrate contrastive multiview representation learning, coverage, and a clustering algorithm.

Similar to the assembly case, there have been efforts to benchmark the performance of the available binning tools. In a recent report, Han et al. (11) used different combinations of short, long, and hybrid data to compare the outcomes from 10 bidders, finding that deep-learning-based tools (COMEBin and SemiBin2) were almost always among the top three high-performance bidders regardless of the combination of the contig provenance. Through comparisons among less tools, Cansdale & Chong (268) showed that CONCOCT generated more high-quality bins than MetaBAT2 using a simple gut metagenome, while Meyer et al. (263) reported homogeneous results among CONCOCT, MetaBAT2, and MaxBin2, with MAG completeness slightly increased by CONCOCT at the expense of genome purity. Contrastingly, Groopm2 (269) and MetaBAT2 provided the best performance metrics in recall, purity, and the number of high-quality genome bins at recovering MAGs from Critical Assessment of Metagenome Interpretation (CAMI) data sets (270). In addition, Yepes-García and Falquet (271) used environmental metagenomic samples (rice soil) to show how MetaBinner stands out for the greater number of bins

recovered as compared with MetaBAT2 and SemiBin2, albeit only 10% of these were at least MQ MAGs.

Moreover, the inclusion (or enabling) of tools within the workflows to recover a non-redundant and high-quality MAG set is determinant. Several pipelines incorporate bin refinement modules or tools to improve the quality of the bin set as they reduce contamination, increase completeness, and may recover mis-binned contigs (12, 13, 85). The tools in charge of this task take as input the bins from different binning software to provide the best possible version of each bin and potential MAG. Among the existing tools for bin refinement, MAGScoT (13) is claimed by the developers as the piece of software with the best performance, as compared to DASTool (12) and the MetaWRAP-binning module (59), in terms of MAG quantity and quality using simulated marine and human gut data sets. Nonetheless, Han et al. (11) showed how MetaWRAP achieved the highest rank score (custom ranking score developed for the study) followed closely by MAGScoT, although this former tool demanded 10 times less memory and carried the bin refinement in one-tenth of a fraction of the time required by MetaWRAP.

Contamination estimation tools aid in the main goal of ensuring the reliability of the MAGs, with representative tools such as CheckM (26), BUSCO (116), and CheckM2 (58) that infer completeness and contamination based on single-copy marker genes from specific lineages or deep learning models. Notwithstanding, a benchmarking study (14) showed that CheckM may underestimate contamination, mainly if sequences from distantly related taxa are present, as it reported contamination values between 1% and 2% when the true contamination introduced by the researchers was 11%. In contrast, in the same study, the authors found that tools integrating phylogenomic signals or read classification strategies like GUNC (114), Kraken2 (62), Physeter (117), and Forty-Two (272) achieved contamination estimations closer to the true values and performed overall better at detecting inter-domain contamination. Further, within the CheckM2 paper itself, the developers demonstrated its greater accuracy to detect genome contamination conferred by unusual lineages and to predict genome completeness.

Similarly, some pipelines could include dereplication strategies after quality assessment, typically based on Average Nucleotide Identity with the aim of curating the MAG set and selecting the *best* representative MAG in each cluster of MAGs. Nonetheless, enabling the execution of these dereplication tools (85, 144, 177), as well as the parameter configuration, should always be thought thoroughly as discussed by Evans and Denef (15), who analyzed the advantages and drawbacks of running de-replication procedures. Briefly, these authors highlighted how dereplication maintains high quality of genomic databases and enhances coverage pattern estimations; however, dereplication may lead to a loss of information on variability in the auxiliary gene content among representatives from the same species.

One of the final stages when building MAGs is represented by reporting the taxonomic affiliation of each genome. The most common tool included within the workflows (Table 1) is GTDB-Tk (17) since it demonstrated that its phylogeny-based approach achieves high agreement (around 90%) with manually curated classifications in the GTDB, while GTDB-Tk v2 (GTDB-Tk2) is further optimized to reduce memory requirements without compromising the accuracy. Beyond this, the report describing the capabilities of CAT and BAT (107) included a benchmark against GTDB-Tk that demonstrated very similar performance as BAT and GTDB-Tk provided the same final MAG annotations.

Other classifiers not particularly designed to annotate MAGs can be included within the workflows such as MetaPhlAn4 (190), Kraken (108), Kraken2 (62), Centrifuge (36), and Kaiju (78) through the re-formatting of the draft genomes to make them suitable as input for these tools. There have been several efforts to benchmark taxonomic classifiers in a wide variety of scenarios and using different types of data (10, 273–279); however, these studies contrasting their performance and precision have shown variable results. For instance, Kraken2 in combination with Bracken exhibited superior precision, sensitivity, F1 score, and overall sequence classification of a custom *in silico* mock community

within a comparison against MetaPhlAn and Kaiju (273); similar results were described by Timilsina et al. (274), who reported the highest accuracy and broad sensitivity achieved by Kraken2/Bracken (86) in simulated microbial communities as compared against MetaPhlAn4 and Centrifuge. Meanwhile, Irankhah et al. (275) observed how MetaPhlAn4 exhibited higher precision in identifying species in a simulated data set, outperforming Kraken2, Bracken, and Centrifuge. In contrast, when attempting to classify long reads (ONT), Kraken2 and Centrifuge demonstrated low to very low precision for all defined mock communities considered in the study (276). Similarly, Centrifuge depicted the worst performance at classifying sequences belonging to a mock community built from human fecal samples, within the study that introduced the tool DeepMicrobes (277).

To complete the final stages of the MAG reconstruction, functional annotation serves to reveal metabolic potential and ecological roles of microbial communities, with a remarkably high number of options available (280). The selection of these tools depends on the study goal, and it is usually a conscious decision made by the researchers. For more than 10 years, Prokka (96) has remained as standard for rapid genome annotation, predicting coding sequences, rRNAs, and tRNAs and assigning functions through curated databases. Nevertheless, more elaborated tools like eggNOG-mapper (84) have emerged to provide large-scale functional annotation, and the DRAM pipeline (141) offers detailed metabolic summaries. Web-based systems like RASTtk (70) (implemented within the Bacterial and Viral Bioinformatics Resource Center, BV-BRC [65]) and MGnify (193) can achieve quick and reliable annotations, while for specialized functional insights, tools like antiSMASH (197), KOfamKOALA (167), and dbCAN3 (140) are often incorporated into the workflows.

As shown on Table 1, taxonomic and functional annotation steps heavily rely on existing databases, highlighting the importance of these information resources. In the case of taxonomic classification, the GTDB (200) provides a phylogenetically consistent framework for prokaryotic and archaeal taxonomy, while nucleotide and protein repositories like UniRef (195) and Swiss-Prot (165) offer curated sequences that serve reliable standards for accurate assignments. On the functional prediction side, the KEGG (122) and its ortholog collection (KOfam [167]) enables the reconstruction of metabolic pathways, while Pfam (132) catalogs protein domains and families that help identify conserved protein functions. In the same sense, the database for evolutionary genealogy of genes: non-supervised Orthologous Groups (eggNOG) (52) covers orthologous groups linked to functional categories including COG (38), KEGG, and Gene Ontology terms (281). Other specialized databases are represented by the CAZy (188) and the database of proteolytic enzymes, their substrates, and inhibitors (MEROPS) (282). Please note that this is not a comprehensive review, and hence we suggest further reading of the works by Zeller and Huson (283) and Lin et al. (280), who explored and compared computational methods and classification systems, including databases, for protein function prediction.

Finally, benchmarking entire pipelines can be more challenging as they include many pieces of software which difficulties setting a groundline for comparisons. Notwithstanding, there are a few works where the whole pipeline execution has been benchmarked, for instance, Churchward et al. (147), who tested their pipeline performance (MAGNETO) against similar workflows such as nf-core/mag, Metagenome-Atlas, and MetaWRAP. These authors recovered a superior number of HQ MAGs from human gut microbiomes (Integrative Human Microbiome Project) through MetaWRAP operated in either single-assembly with single binning or co-assembly with a co-binning approach (see the next section for a detailed explanation of these approaches). Meanwhile, Yepes-García and Falquet (271), starting from sequences belonging to a mock community, depicted slight differences in terms of genome completeness, contamination, and number of MAGs taxonomically annotated at species level among MetaWRAP, nf-core/mag, SnakeMAGs, and Metagenome-Atlas. nf-core/mag reached the highest percentages of MQ and HQ MAGs, whilst DATMA, also included in this study, performed poorly as only 40% of the MAGs were assigned a proper taxonomic classification and not a single MQ or HQ MAG was recovered.

PRACTICAL AND TECHNICAL CONSIDERATIONS FOR PIPELINE EXECUTION

As high-throughput sequencing technologies have grown in the past years, the availability of MAG-centered pipelines has been quickly expanded to handle and integrate different data types and computational strategies (169, 180, 250). Specifically, recent pipelines have been designed or have evolved to assemble and bin short reads (normally Illumina), long reads (mainly ONT and PacBio), or a blend of both technologies to maximize base calling, depth, contiguity, and structural information (180, 250). Short reads synthesized through DNA nanoball sequencing (284) or long reads derived from CycloneSEQ (285) can be eventually processed by some pipelines (207, 214). Differences or similarities among these MAG-reconstruction approaches based on the type of sequence used as input have been studied by Goussarov et al. (262), and Kim et al. (286) analyzed the variations in terms of genome recovery between Illumina and MGI platforms.

Among the several tools that compose a pipeline (Fig. 1), assembly and binning tools are mainly responsible for the scaling up in the hardware demands, especially when handling data sets with several samples encompassing millions of short-read sequences (6). Moreover, these tools can be executed in different configurations such as co-assembly and co-binning, as these strategies can increase the overall MAG recovery rate and quality (287). Briefly, co-assembly refers to the possibility of performing the metagenome assembly after merging user-specified samples to enhance the coverage, capturing a higher fraction of the diversity (287), while co-binning establishes the possibility of binning contigs using coverage information across multiple samples simultaneously after single or co-assembly (11). Co-binning is advantageous at exploring coverage across samples and improving separation of closely related genomes (47). Despite the desirable benefits co-assembly can bring to the analysis, it is computationally intensive and increases the probability of generating fragmented assemblies (147), although sequential co-assembly has emerged recently as an efficient alternative that enhances both time and memory requirements by the assembler (288). Similarly, co-binning can be sensitive to uneven sequencing depth, requires high-quality coverage profiles, and can be affected by low diversity among samples (147). Vosloo et al. (287) and Han et al. (11) have demonstrated how superior performance can be achieved by applying co-assembly and/or co-binning.

On the other hand, the workflow execution varies in terms of computational demands, where small-scale data sets can be processed on high-end workstations, while large or complex metagenomes often require access to HPC clusters or cloud-based environments (Azure, Amazon Web Services or AWS, Google Cloud, and Terra, among others). Beyond sample-specific computational requirements, and as mentioned before, most metagenomics pipelines rely on external reference databases to perform taxonomic classification, functional annotation, and quality assessment of MAGs. Commonly used databases, namely, RefSeq (289), GTDB, UniProt (206), KEGG, and eggNOG, are large and require substantial local storage that ranges from tens to hundreds of gigabytes. For instance, the latest GTDB release (R226) exceeds 140 GB, while comprehensive functional annotation pipelines like DRAM can demand up to 500 GB to exploit its full potential. Being so, MAG building is a demanding process that needs adequate disk space, CPU capacity, and memory availability.

For researchers without access to HPC resources, web-based platforms such as KBase (290), MGnify (193), Galaxy (105), and BV-BRC (65), among others, can assist them by carrying out analysis execution in their servers. In addition, these platforms aid users without a strong experience in command line interface (CLI) interaction since they provide user-friendly interfaces where users can upload raw reads and run predefined workflows. As a result, these platforms eliminate the need for CLI proficiency and offer built-in visualization applications and databases for downstream interpretation; a complete landscape of web-based applications is compiled by Achudhan et al. (291) and Chivian et al. (138).

Furthermore, given the MAG pipeline evolution in complexity, involving multiple tools, dependencies, and steps, the use of workflow managers has become the standard to ensure reproducibility, scalability, and portability (292). Specifically, workflow managers ease pipeline step definition in a modular and automated architecture to orchestrate entire analyses, tracking software versions, managing intermediate files, restarting the process if interrupted, handling multiple samples as input, and enabling parallel processing in a reproducible manner. Some representatives of these helpful orchestrators are Snakemake (293), Nextflow (294), and Workflow Definition Language (WDL) (295) whose design, implementation, benefits, and scope have been reviewed in some reports (292–294, 296); also, important guidelines for pipeline design based on workflow managers have been published by Roach et al. (297), Reiter et al. (298), and Ahmed et al. (7). Advantageously, containerization platforms such as Docker, Singularity, and Sequera Containers, or package managers like Conda or the Python Package Index complement workflow orchestrators by offering a flexible and reproducible solution for software and dependency management (299). As a result, this combination allows users to run the analysis without system conflicts, specific versions of the software, and libraries.

In contrast, beyond the MAG assembly and annotation, some pipelines feature interesting options that complement the analysis and provide a wider understanding of the microbial community. The range of these special options is wide, and therefore they must be carefully selected. In this sense, read-based taxonomic profiling (1) is one of the most common offerings by the pipelines as this process does not rely on the main workflow and can be executed in parallel. Furthermore, some pipelines can incorporate tools or modules to recover viral or eukaryotic MAGs (250), and it is even possible to find pipelines mostly focused on this type of MAGs (97). Another popular extra option is represented by the possibility of establishing genome-scale metabolic models among the built MAGs (159, 169). However, in many cases, some workflows can be considered unique since they include options that no other pipeline encompasses. Examples of these rare features are the possibility to assemble plasmids (169), genotype recovery (41), controlled resource allocation (169), and an alternative assembly and binning order, where the reads are first grouped (binning) and then assembled in batches (73).

On Table 2, we present a summarized overview of the technical features and methodological factors each workflow presents, and hence these same pipeline aspects are also the basis for the questionnaire presented on 2Pipe. Methodological factors include the ability to assemble short reads, long sequences, or both in a hybrid approach; the possibility to request a co-assembly and/or co-binning natively; whether the user can input multiple samples or not; if the pipeline includes a bin refinement tool; and special functionalities they may incorporate. In the same sense, technical features are described through factors like which kind of resources the user is planning to use for the pipeline execution, the interface they feel more comfortable working with, the workflow manager they expect to orchestrate the data flow, and the software/package technology management available within each workflow. We assigned one of the following (non-mutually exclusive) labels in order to classify them: short-read-centered or long-read-focused (if their main input is short or long reads), dual (if they can handle both long and short reads, but they do not perform hybrid assembly), hybrid (pipelines able to assemble short and long reads together), web-based (pipelines offered by online platforms or suites), or special (pipelines designed for a specific purpose).

2PIPE: IT STARTS WITH A QUESTION

Considering the pipeline landscape identified in this review, we have developed a decision-support application that concatenates most of the features described for each workflow. 2Pipe is an interactive web application designed to help researchers identify the most suitable metagenomics pipeline for reconstructing and annotating MAGs. 2Pipe can be used by users with different expertise levels and computational access, simplify-

TABLE 2 Technical and operational features for each pipeline or web-based platform

No.	Pipeline/ Platform	Category	Short reads	Long reads ^a	Hybrid assembly	Multiple samples	Co-assembly and/or co-binning ^b	Bin refinement	Infrastructure ^c	Interface ^d	Workflow manager	Software execution	Special features	Last update ^e	Number of citations ^f	License ^g
1	Ancient DNA (19)	Special	Yes	No	No	No	No	Yes	Local and HPC	CLI		Local	Ancient DNA identification	2024	0	Not specified
2	Anvi'o (28)	Short-read- centered	Yes	No	No	Yes	Yes	Yes	Local and HPC	CLI/graphical user interface (GUI)		Conda	Visualization module	2025	678	GNU GPL v3
3	Aviary (41)	Hybrid	Yes	Yes	Yes	No	No	Yes	Local, HPC, and CC		Snakemake	Conda	Genotype recovery	2025	Not found	GNU GPL v3
4	BugBuster (54)	Short-read- centered	Yes	No	No	Yes	No	Yes	Local, HPC, and CC	CLI	Nextflow	Docker	Taxonomic profiling and antimicrobial resistance gene prediction	2025	0	Not specified
5	BV-BRC (65)	Web-based	Yes	No	No	Yes	No	No	External	GUI		External	Taxonomic profiling and viral MAGs	2024	783	MIT License
6	DATMA (73)	Short-read- centered	Yes	No	No	No	No	No	Local and HPC	CLI	COMP	Local	Reads first grouped (binning) and assembled in batches	2020	4	GNU GPL v3
7	EasyMetagenome (81)	Short-read- centered	Yes	No	No	Yes	Yes	Yes	Local and HPC	CLI		Conda	Taxonomic profiling	2024	14	GNU GPL v3
8	EasyNanoMeta (87)	Long-read- focused	No	Yes (ONT)	Yes	No	No	No	Local and HPC	CLI		Conda, Singularity	Taxonomic profiling	2024	0	GNU GPL v3
9	Eukfinder (97)	Special	Yes	Yes	No	No	No	No	Local and HPC	CLI		Conda	Eukaryotic MAGs	2025	1	MIT License
10	EURYALE (MEDUSA) (101, 102)	Short-read- centered	Yes	No	No	Yes	No	No	Local, HPC, and CC	CLI	Nextflow	Conda, Singularity, ty, Docker		2024	7	MIT License
11	Galaxy (105)	Web-based	Yes	Yes	Yes	No	No	Yes	External	GUI		External	Taxonomic profiling	2024	1168	Academic Free License v3
12	GEN-ERA (109)	Dual	Yes	Yes (ONT)	No	Yes	No	No	Local, HPC, and CC	CLI	Nextflow	Singularity	Metabolic modeling	2024	7	GNU GPL v3
13	HIFI-MAG (124)	Long-read- focused	No	Yes (PacBio)	No	Yes	No	Yes	Local, HPC, and CC	CLI	Snakemake	Conda		2025	8	BSD-3-Clause-Clear License
14	IDseq (125)	Web-based	Yes	Yes (ONT)	No	No	No	No	External	GUI		External	Viral MAGs	2025	347	MIT License
15	IMG/M (130)	Web-based	NA	NA	NA	No	No	No	External	GUI		External	Eukaryotic MAGs	2025	268	IMG Expert Review Submission Agreement

(Continued on next page)

TABLE 2 Technical and operational features for each pipeline or web-based platform (Continued)

No.	Pipeline/ Platform	Category	Short reads	Long reads ^a	Hybrid assembly	Multiple samples	Co-assembly and/or co-binning ^b	Bin refinement	Infrastructure ^c	Interface ^d	Workflow manager	Software execution	Special features	Last update ^e	Number of citations ^f	License ^g
16	JAMS (136)	Short-read- centered	Yes	No	No	No	No	No	Local and HPC	CLI		Conda	Direct sample comparison	2025	7	GNU GPL v3
17	KBase (138)	Web-based	Yes	Yes	Yes	Yes	Yes	Yes	External	GUI		External	Taxonomic profiling and metabolic modeling	2024	63	MIT License
18	MAGNETO (147)	Short-read- centered	Yes	No	No	Yes	Yes	No	Local, HPC, and CC	CLI	Snakemake	Conda	Taxonomic profiling	2025	13	GNU GPL v3
19	MAGO (152)	Short-read- centered	Yes	No	No	No	No	Yes	Local and HPC	CLI		Singularity,Phylogenetic tree generation Docker and pangenome analysis	2020	21	Creative Commons BY 4.0	
20	Mapler (155)	Long-read- focused	No	Yes (PacBio)	No	Yes	No	No	Local, HPC, and CC	CLI	Snakemake	Conda	Visualization module	2025	0	GNU AGPL v3
21	MetaGEM (159)	Short-read- centered	Yes	No	No	Yes	No	Yes	Local, HPC, and CC	CLI	Snakemake	Conda	Eukaryotic MAGs and metabolic modeling	2023	99	MIT License
22	MetaGenePipe (164)	Short-read- centered	Yes	No	No	Yes	Yes	No	Local, HPC, and CC	CLI	WDL (295)	Singularity		2023	1	Apache License 2.0
23	Metagenome-Atlas (168)	Short-read- centered	Yes	No	Yes	Yes	Yes	Yes	Local, HPC, and CC	CLI	Snakemake	Conda		2024	159	BSD-3-Clause-Clear
24	Metagenomics- Toolkit (169)	Dual	Yes	Yes (ONT)	No	Yes	No	Yes	Local, HPC, and CC	CLI	Nextflow	Docker	Plasmid assembly, metabolic modeling and controlled resource allocation	2025	0	GNU AGPL v3
25	Metaphor (179)	Short-read- centered	Yes	No	No	Yes	Yes	Yes	Local, HPC, and CC	CLI	Snakemake	Conda	Visualization module	2024	13	MIT License
26	metagWGS (180)	Dual	Yes	Yes (PacBio)	No	Yes	Yes	Yes	Local, HPC, and CC	CLI	Nextflow	Singularity	Taxonomic profiling	2025	2	GNU GPL v3
27	MetaWRAP (59)	Short-read- centered	Yes	No	No	Yes	Yes	Yes	Local and HPC	CLI		Conda and Docker	Taxonomic profiling	2020	1917	MIT License
28	MG-TK (184)	Dual	Yes	No	No	Yes	Yes	No	Local and HPC	CLI		Conda	Taxonomic profiling and strain delineation	2025	99	GNU GPL v2
29	Mgnify (193)	Web-based	Yes	Yes	Yes	Yes	Yes	No	External	GUI		External	Taxonomic profiling	2025	286	Apache License 2.0
30	MOSHPIIT (198)	Short-read- centered	Yes	No	No	Yes	No	Yes	Local and HPC	CLI		Conda	Taxonomic profiling	2025	1	BSD-3-Clause-Clear
31	MUFFIN (199)	Hybrid pipelines	No	Yes (ONT)	Yes	Yes	No	Yes	Local, HPC, and CC	CLI	Nextflow	Conda, Docker, and Singularity	Metatranscriptome support	2022	34	GNU GPL v3

(Continued on next page)

TABLE 2 Technical and operational features for each pipeline or web-based platform (Continued)

No.	Pipeline/ Platform	Category	Short reads	Long reads ^a	Hybrid assembly	Multiple samples	Co-assembly and/or co-binning ^b	Bin refinement	Infrastructure ^c	Interface ^d	Workflow manager	Software execution	Special features	Last update ^e	Number of citations ^f	License ^g
32	NanoPhase (203)	Long-read- focused	No	Yes (ONT)	Yes	No	No	Yes	Local and HPC	CLI		Conda		2023	73	MIT License
33	nf-core/mag (207)	Hybrid	Yes	Yes (ONT or PacBio)	Yes	Yes	Yes	Yes	Local, HPC, and CC	CLI	Nextflow	Conda, Docker, Singularity	Ancient DNA Identification	2025	57	MIT License
34	ngs-preprocess MpGap Bacannot (214)	Hybrid	Yes	Yes	Yes	No	No	No	Local, HPC, and CC	CLI	Nextflow	Conda, Docker, Singularity and Others	Antimicrobial resistance gene prediction, virulence factor annotation, and plasmid assembly	2025	2	GNU GPL v3
35	nIMP3 (232)	Short-read- centered	Yes	No	No	Yes	No	No	Local, HPC, and CC	CLI	Nextflow	Docker, Singularity	Metatranscriptome support and taxonomic profiling	2024	150	MIT License
36	SnakeMAGs (236)	Short-read- centered	Yes	No	No	Yes	No	No	Local, HPC, and CC	CLI	Snakemake	Conda		2024	6	CeCILL Free Software License Agreement v2.1
37	SPIRE (237)	Short-read centered	Yes	No	No	Yes	No	No	Local, HPC, and CC	CLI	Nextflow		Antimicrobial resistance gene prediction and virulence factor annotation	2025	41	MIT License
38	SqueezeMeta (243)	Hybrid	Yes	Yes	Yes	Yes	Yes	Yes	Local and HPC	CLI		Conda	Taxonomic profiling, metatranscriptome support, and visualization module	2025	400	GNU GPL v3
39	Sunbeam (249)	Short-read- centered	Yes	No	No	Yes	No	No	Local and HPC	CLI	Snakemake	Conda and Taxonomic profiling		2025	184	GNU GPL v3
40	VEBA (250)	Dual	Yes	Yes (ONT or PacBio)	No	Yes	Yes and pseudo- coassembly	Yes	Local and HPC	CLI	GenoPype (301)	Conda and Eukaryotic or viral MAGs, Docker	antimicrobial resistance gene prediction, and virulence factor annotation	2025	23	GNU AGPL v3
41	WGS2+/-LoRA (258)	Web-based	Yes	Yes (ONT or PacBio)	No	Yes	No	No	External and CC	GUI	AWS environment	External	Visualization module, metatranscriptome support,	2025	138	CC0 1.0 Universal

(Continued on next page)

TABLE 2 Technical and operational features for each pipeline or web-based platform (Continued)

No.	Pipeline/ Platform	Category	Short reads	Long reads ^a	Hybrid assembly	Multiple samples	Co-assembly and/or co-binning ^b	Bin refinement	Infrastructure ^c	Interface ^d	Workflow manager	Software execution	Special features	Last update ^e	Number of citations ^f	License ^f
													and antimicrobial resistance gene prediction			

^aLong reads: ONT, Oxford Nanopore Technology; PacBio, Pacific Biosciences.
^bCo-assembly and/or co-binning: it highlights if the pipeline counts with options to control co-assembly and/or co-binning.
^cInfrastructure: it refers to the computational infrastructure where the pipeline can be executed natively. HPC, high-performance cluster; CC, cloud computing; External, pipelines controlled by the platform or suite and use third-party resources.
^dInterface: CLI, command line interface; GUI, graphical user interface.
^eLast update and number of citations: at the moment of writing this report.
^fLicense: these licenses cover the pipeline code and platforms; the third-party software and tools they encompass may be covered by a different license.

ing the often-complex selection process by mapping user needs to a curated database of available pipelines.

At the core of 2Pipe, there is a dynamic and question-driven interface that guides users step by step through a personalized questionnaire. This adaptive form collects information related to the methodological factors and technical features detailed on Table 2. Therefore, every response is used to assign a score to each pipeline based on the presence or absence of specific features that align with the user's input. The recommendation system will then suggest the pipeline with the highest score, as well as the second "best hit" for the user to check in case the first option does not fulfill their requirements; these suggestions can be as well the starting point for the user to dig into the other sections of 2Pipe. It is worth mentioning that the scoring is weighted, and some features have prevalence as they are definitive for the pipeline suggestion. Specifically, all matching features presented in the questions add one point to the final score, excepting type of reads to analyze (2 points), the need for a GUI (3 points), and the requirement for external computational resources (3 points). These features are prioritized then, and the recommendation must reflect them as they cannot simply be bypassed with any other pipeline. The system also includes a protection for cases when the users do not provide at least three answers, asking them to restart the questionnaire. Likewise, in case of a tie among more than two pipelines, the recommendation system will show all of them with the respective matching features.

Aside from the accession to the questionnaire and the response-based recommendation at the end of this, 2Pipe as well encompasses a pipeline gallery, where a visual catalog is displayed, offering individual summaries of each pipeline and a direct access to the source code or to the publication that documents the pipeline. Additionally, 2Pipe makes available an interactive view of Table 2 that includes the possibility of filtering by each feature or by a combination of them, allowing users to directly tailor the search for the pipeline that best suits their needs; the displayed categories are the same key attributes the question-based suggestion system relies on. 2Pipe also incorporates the features presented in Table 1, assisting the user when comparing the pipelines beyond technical aspects. Also, these tools and external software are organized in a gallery that allows the user to match pipelines that use them, which is useful if the user is looking for a specific software combination that a specific pipeline can offer.

On the other hand, given the importance pipeline and tool benchmarking represents, 2Pipe provides an exclusive page where the reports cited in this work comparing performance and/or technical features are introduced. This page is divided into sections according to the tools benchmarked in the papers, namely, assemblers, binners, bin-refinement tools, contamination-estimation software, complete pipelines, workflow managers, and taxonomic classifiers. Moreover, we include sections for reviews, tutorials, and protocols for manual MAG reconstruction and key papers that set interesting discussions around MAG recovery.

The source code for 2Pipe is available at the repository <https://github.com/jeffe107/2pipe>, and foreseeing the possibility of new pipelines being released in the near future, we provide a quick form for developers to include their workflow into 2Pipe's recommendation system, pipeline gallery, and table comparison. Also, at the GitHub repository, developers can find a simple template and detailed instructions for the inclusion of their pipeline through a pull request.

CONCLUSION

The rapid evolution of sequencing technologies has broadened the availability of metagenomics data sets that demand bioinformatics tools adjusted to the user requirements to achieve cutting-edge analysis, including MAG reconstruction. As a result, in the past 10 years, a rise in the number of MAG reconstruction pipelines available has been observed, and the selection of the proper pipeline for the analysis has become an essential step during the execution of metagenomics projects. This review offers a compact description of 41 publicly available pipelines or platforms, with special focus on

their capabilities and distinctive features to serve as a valuable resource for researchers navigating this overwhelming landscape. Beyond the scope of a classical review, we streamlined the selection process by introducing 2Pipe, an interactive decision-support web application that aligns the user needs with the most convenient workflow for their analysis and allows a general overview of the pipeline universe with its gallery and pipeline-comparison sections. Finally, this review and its accompanying application provide a unified framework that simplifies the decision-making process, releasing part of the burden and uncertainty when setting a metagenomics data analysis project.

ACKNOWLEDGMENTS

J.Y.G. especially thanks the Federal Commission for Scholarships for Foreign Students (FCS) for their support through the Swiss Government Excellence Scholarship.

AUTHOR AFFILIATIONS

¹Department of Biology, University of Fribourg, Fribourg, Canton of Fribourg, Switzerland

²Swiss Institute of Bioinformatics, Lausanne, Vaud, Switzerland

AUTHOR ORCIDs

Jeferyd Yepes-García  <http://orcid.org/0000-0002-3278-3332>

Laurent Falquet  <http://orcid.org/0000-0001-8102-7579>

FUNDING

Funder	Grant(s)	Author(s)
Federal Commission for Scholarships for Foreign Students		Jeferyd Yepes-García

AUTHOR CONTRIBUTIONS

Jeferyd Yepes-García, Conceptualization, Software, Writing – original draft, Writing – review and editing | Laurent Falquet, Conceptualization, Supervision, Writing – original draft, Writing – review and editing

DATA AVAILABILITY

2Pipe is hosted under the domain <https://2pipe.app/>. The source code is available at <https://github.com/jeffe107/2pipe>, along with a template to include new pipelines. The quick form to add a new pipeline can be found at <https://form.jotform.com/jeffe10789/2pipe-form>. For version tracking, 2Pipe v.2.0 release has been deposited at Zenodo, and it can be followed with the identifier <https://doi.org/10.5281/zenodo.17334924>.

ADDITIONAL FILES

The following material is available [online](#).

Supplemental Material

File S1 (mSystems00844-25-S0001.pdf). Detailed summary description for each pipeline considered in this review.

REFERENCES

1. Navgire GS, Goel N, Sawhney G, Sharma M, Kaushik P, Mohanta YK, Mohanta TK, Al-Harrasi A. 2022. Analysis and Interpretation of metagenomics data: an approach. *Biol Proced Online* 24:18. <https://doi.org/10.1186/s12575-022-00179-7>
2. Kim N, Ma J, Kim W, Kim J, Belenky P, Lee I. 2024. Genome-resolved metagenomics: a game changer for microbiome medicine. *Exp Mol Med* 56:1501–1512. <https://doi.org/10.1038/s12276-024-01262-7>

3. Lemos LN, Mendes LW, Baldrian P, Pylro VS. 2021. Genome-resolved metagenomics is essential for unlocking the microbial black box of the soil. *Trends Microbiol* 29:279–282. <https://doi.org/10.1016/j.tim.2021.01.013>
4. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Elae-Fadrosch EA, 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. <https://doi.org/10.1038/nbt.3893>
5. Setubal JC. 2021. Metagenome-assembled genomes: concepts, analogies, and challenges. *Biophys Rev* 13:905–909. <https://doi.org/10.1007/s12551-021-00865-y>
6. Yang C, Chowdhury D, Zhang Z, Cheung WK, Lu A, Bian Z, Zhang L. 2021. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput Struct Biotechnol J* 19:6301–6314. <https://doi.org/10.1016/j.csbj.2021.1.1028>
7. Ahmed AE, Allen JM, Bhat T, Burra P, Fliege CE, Hart SN, Heldenbrand JR, Hudson ME, Istanto DD, Kalmbach MT, Kapraun GD, Kendig KI, Kendzior MC, Klee EW, Mattson N, Ross CA, Sharif SM, Venkatakrishnan R, Fadlilmola FM, Mainzer LS. 2021. Design considerations for workflow management systems use in production genomics research and the clinic. *Sci Rep* 11:1–18. <https://doi.org/10.1038/s41598-021-99288-8>
8. Chen S, Zhou Y, Chen Y, Gu J. 2018. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34:i884–i890. <https://doi.org/10.1093/bioinformatics/bty560>
9. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
10. Gao Y, Luo H, Lyu H, Yang H, Yousuf S, Huang S, Liu Y-X. 2025. Benchmarking short-read metagenomics tools for removing host contamination. *Gigascience* 14:giaf004. <https://doi.org/10.1093/gigascience/giaf004>
11. Han H, Wang Z, Zhu S. 2025. Benchmarking metagenomic binning tools on real datasets across sequencing platforms and binning modes. *Nat Commun* 16:2865. <https://doi.org/10.1038/s41467-025-57957-6>
12. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 3:836–843. <https://doi.org/10.1038/s41564-018-0171-1>
13. Christoph M, Rühlemann R, Wacker EM, Ellinghaus D, Franke A. 2022. MAGScoT: a fast, lightweight and accurate bin-refinement tool. *Bioinformatics* 38:5430–5433. <https://doi.org/10.1093/bioinformatics/btac694>
14. Cornet L, Baurain D. 2022. Contamination detection in genomic data: more is not enough. *Genome Biol* 23:60. <https://doi.org/10.1186/s13059-022-02619-9>
15. Evans JT, Denef VJ. 2020. To derePLICATE or not to derePLICATE? mSphere 5:e00971-19. <https://doi.org/10.1128/mSphere.00971-19>
16. Schwengers O, Jelonek L, Dieckmann MA, Beyvers S, Blom J, Goesmann A. 2021. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genomics* 7:000685. <https://doi.org/10.1099/mgen.0.000685>
17. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. 2020. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 36:1925–1927. <https://doi.org/10.1093/bioinformatics/btz848>
18. Wajid B, Anwar F, Wajid I, Nisar H, Meraj S, Zafar A, Al-Shawaqfeh MK, Ekti AR, Khatoon A, Suchodolski JS. 2022. Music of metagenomics—a review of its applications, analysis pipeline, and associated tools. *Funct Integr Genomics* 22:3–26. <https://doi.org/10.1007/s10142-021-00810-y>
19. Standeven FJ, Dahlquist-Axe G, Speller CF, Meehan CJ, Tedder A. 2024. An efficient pipeline for creating metagenomic-assembled genomes from ancient oral microbiomes. *bioRxiv*. <https://doi.org/10.1101/2024.09.18.613623>
20. Simon A. 2010. FastQC a quality control tool for high throughput sequence data. FastQC. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc>
21. Bushnell B. 2014. JGI Web Archives. BBMap: a fast, accurate, splice-aware aligner. Available from: <https://archive.jgi.doe.gov/data-and-tools/software-tools/bbtools>
22. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, Yamashita H, Lam TW. 2016. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* 102:3–11. <https://doi.org/10.1016/j.ymeth.2016.02.020>
23. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. *Nat Methods* 11:1144–1146. <https://doi.org/10.1038/nmeth.3103>
24. Wu Y-W, Tang Y-H, Tringe SG, Simmons BA, Singer SW. 2014. MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* 2:26. <https://doi.org/10.1186/2049-2618-2-26>
25. Kang DD, Froula J, Egan R, Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* 3:e1165. <https://doi.org/10.7717/peerj.1165>
26. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25:1043–1055. <https://doi.org/10.1101/gr.186072.114>
27. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* 29:1682–1684. <https://doi.org/10.1093/bioinformatics/btt193>
28. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, Fink I, Pan JN, Yousef M, Fogarty EC, et al. 2020. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* 6:3–6. <https://doi.org/10.1038/s41564-020-00834-3>
29. Eren AM, Vineis JH, Morrison HG, Sogin ML. 2013. A filtering method to generate high quality short reads using illumina paired-end technology. *PLoS One* 8:e66643. <https://doi.org/10.1371/journal.pone.0066643>
30. 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>
31. Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28:1420–1428. <https://doi.org/10.1093/bioinformatics/bts174>
32. 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>
33. Wu YW, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* 32:605–607. <https://doi.org/10.1093/bioinformatics/btv638>
34. Graham ED, Heidelberg JF, Tully BJ. 2017. BinSanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ* 5:e3035. <https://doi.org/10.7717/peerj.3035>
35. Breitwieser FP, Baker DN, Salzberg SL. 2018. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol* 19:198. <https://doi.org/10.1186/s13059-018-1568-0>
36. Kim D, Song L, Breitwieser FP, Salzberg SL. 2016. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* 26:1721–1729. <https://doi.org/10.1101/gr.210641.116>
37. Buchfink B, Reuter K, Drost HG. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 18:366–368. <https://doi.org/10.1038/s41592-021-01101-x>
38. Galperin MY, Vera Alvarez R, Karamycheva S, Makarova KS, Wolf YI, Landsman D, Koonin EV. 2025. COG database update 2024. *Nucleic Acids Res* 53:D356–D363. <https://doi.org/10.1093/nar/gkac983>
39. Larraalde M. 2022. Pyrodigal: python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. *JOSS* 7:4296. <https://doi.org/10.21105/joss.04296>
40. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39:W29–37. <https://doi.org/10.1093/nar/gkr367>
41. Newell RJ, Aroney STN, Zaugg J, Sternes P, Tyson GW, Woodcroft BJ. 2025. ARI: hybrid assembly and genome recovery from metagenomes (0.9.0). Zenodo. <https://zenodo.org/records/15208119>
42. Haveman NJ, Khodadad CLM, Dixit AR, Louyakis AS, Massa GD, Venkateswaran K, Foster JS. 2021. Evaluating the lettuce metatranscriptome with MinION sequencing for future spaceflight food production applications. *NPJ Microgravity* 7:22. <https://doi.org/10.1038/s41526-021-00151-x>

43. De Coster W, Rademakers R. 2023. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* 39:btad311. <https://doi.org/10.1093/bioinformatics/btad311>
44. Woodcroft BJ, Aroney STN, Zhao R, Cunningham M, Mitchell JAM, Nurdiansyah R, Blackall L, Tyson GW. 2025. Comprehensive taxonomic identification of microbial species in metagenomic data using SingleM and Sandpiper. *Nat Biotechnol*:1–6. <https://doi.org/10.1038/s41587-025-02738-1>
45. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, 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>
46. 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>
47. 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>
48. Newell RJP, Tyson GW, Woodcroft BJ. 2023. Rosella: metagenomic binning using UMAP and HDBSCAN (0.5.3). GitHub. <https://github.com/rhysnewell/rosella>
49. Mikheenko A, Saveliev V, Gurevich A. 2016. MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 32:1088–1090. <https://doi.org/10.1093/bioinformatics/btv697>
50. Aroney STN, Newell RJP, Nissen JN, Camargo AP, Tyson GW, Woodcroft BJ. 2025. CoverM: read alignment statistics for metagenomics. *Bioinformatics* 41:btaf147. <https://doi.org/10.1093/bioinformatics/btaf147>
51. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform* 11:1–11. <https://doi.org/10.1186/1471-2105-11-119>
52. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, von Mering C, Bork P. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 47:D309–D314. <https://doi.org/10.1093/nar/gky1085>
53. Newell RJP, McMaster ES, Craig P, Boden M, Tyson GW, Woodcroft BJ. 2023. Lorikeet: strain-resolved metagenome analysis using local reassembly (v0.8.2). Zenodo. Available from: <https://zenodo.org/records/10275469>
54. Fuentes-Santander F, Curiqueo C, Araos R, Ugalde JA. 2025. BugBuster: a novel automatic and reproducible workflow for metagenomic data analysis. *Bioinform Adv* 5:vba152. <https://doi.org/10.1101/2025.02.24.639915>
55. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>
56. Pan S, Zhao XM, Coelho LP. 2023. SemiBin2: self-supervised contrastive learning leads to better MAGs for short- and long-read sequencing. *Bioinformatics* 39:i21–i29. <https://doi.org/10.1093/bioinformatics/btad209>
57. Wang Z, You R, Han H, Liu W, Sun F, Zhu S. 2024. Effective binning of metagenomic contigs using contrastive multi-view representation learning. *Nat Commun* 15:1–14. <https://doi.org/10.1038/s41467-023-44290-z>
58. 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>
59. Uritskiy GV, DiRuggiero J, Taylor J. 2018. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6:1–13. <https://doi.org/10.1186/s40168-018-0541-1>
60. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2022. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 38:5315–5316. <https://doi.org/10.1093/bioinformatics/btac672>
61. Figueroa III JL, Dhungel E, Bellanger M, Brouwer CR, White III RA. 2024. MetaCerberus: distributed highly parallelized HMM-based processing for robust functional annotation across the tree of life. *Bioinformatics* 40:btac119. <https://doi.org/10.1093/bioinformatics/btac119>
62. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol* 20:257. <https://doi.org/10.1186/s13059-019-1891-0>
63. Irber L, Pierce-Ward NT, Abuelanin M, Alexander H, Anant A, Barve K, Baumler C, Botvinnik O, Brooks P, Dsouza D, et al. 2024. Sourmash v4: a multitool to quickly search, compare, and analyze genomic and metagenomic data sets. *JOSS* 9:6830. <https://doi.org/10.21105/joss.06830>
64. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. 2018. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* 6:23. <https://doi.org/10.1186/s40168-018-0401-z>
65. Olson RD, Assaf R, Brettin T, Conrad N, Cucinell C, Davis JJ, Dempsey DM, Dickerman A, Dietrich EM, Kenyon RW, et al. 2023. Introducing the Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC, IRD and ViPR. *Nucleic Acids Res* 51:D678–D689. <https://doi.org/10.1093/nar/gkac1003>
66. Krueger F. 2023. A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data. GitHub. <https://github.com/FelixKrueger/TriM-Galore>
67. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
68. Parrello B, Butler R, Chlenski P, Pusch GD, Overbeek R. 2021. Supervised extraction of near-complete genomes from metagenomic samples: a new service in PATRIC. *PLoS One* 16:e0250092. <https://doi.org/10.1371/journal.pone.0250092>
69. Parrello B, Butler R, Chlenski P, Olson R, Overbeek J, Pusch GD, Vonstein V, Overbeek R. 2019. A machine learning-based service for estimating quality of genomes using PATRIC. *BMC Bioinform* 20:486. <https://doi.org/10.1186/s12859-019-3068-y>
70. Brettin T, Davis JJ, Disz T, Edwards RA, Gerdes S, Olsen GJ, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Thomason JA, Stevens R, Vonstein V, Wattam AR, Xia F. 2015. RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci Rep* 5:8365. <https://doi.org/10.1038/srep08365>
71. Wang S, Sundaram JP, Spiro D. 2010. VIGOR, an annotation program for small viral genomes. *BMC Bioinform* 11:451. <https://doi.org/10.1186/1471-2105-11-451>
72. Larsen CN, Sun G, Li X, Zaremba S, Zhao H, He S, Zhou L, Kumar S, Desborough V, Klem EB. 2020. Mat_peptide: comprehensive annotation of mature peptides from polyproteins in five virus families. *Bioinformatics* 36:1627–1628. <https://doi.org/10.1093/bioinformatics/btz777>
73. Benavides A, Sanchez F, Alzate JF, Cabarcas F. 2020. DATMA: distributed automatic metagenomic assembly and annotation framework. *PeerJ* 8:e9762. <https://doi.org/10.7717/peerj.9762>
74. Magoč T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27:2957–2963. <https://doi.org/10.1093/bioinformatics/btr507>
75. Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. <https://doi.org/10.48550/arXiv.1303.3997>
76. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829. <https://doi.org/10.1101/gr.074492.107>
77. Benavides A, Isaza JP, Niño-García JP, Alzate JF, Cabarcas F. 2018. CLAME: a new alignment-based binning algorithm allows the genomic description of a novel *Xanthomonadaceae* from the Colombian Andes. *BMC Genomics* 19:858. <https://doi.org/10.1186/s12864-018-5191-y>
78. Menzel P, Ng KL, Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun* 7:11257. <https://doi.org/10.1038/ncomms11257>
79. Besemer J, Borodovsky M. 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* 33:W451–4. <https://doi.org/10.1093/nar/gki487>
80. Ondov BD, Bergman NH, Phillippy AM. 2011. Interactive metagenomic visualization in a Web browser. *BMC Bioinform* 12:1–10. <https://doi.org/10.1186/1471-2105-12-385>
81. Bai D, Chen T, Xun J, Ma C, Luo H, Yang H, Cao C, Cao X, Cui J, Deng Y, et al. 2025. EasyMetagenome: a user - friendly and flexible pipeline for shotgun metagenomic analysis in microbiome research. *iMeta* 4:e70001. <https://doi.org/10.1002/imt2.70001>

82. 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. *eLife* 10:e65088. <https://doi.org/10.7554/eLife.65088>
83. Hyatt D, LoCascio PF, Hauser LJ, Uberbacher EC. 2012. Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* 28:2223–2230. <https://doi.org/10.1093/bioinformatics/bts429>
84. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol* 38:5825–5829. <https://doi.org/10.1093/molbev/msab293>
85. Olm MR, Brown CT, Brooks B, Banfield JF. 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. <https://doi.org/10.1038/ismej.2017.126>
86. Lu J, Breitwieser FP, Thielen P, Salzberg SL. 2017. Bracken: estimating species abundance in metagenomics data. *PeerJ Comput Sci* 3:e104. <https://doi.org/10.7717/peerj-cs.104>
87. Peng K, Gao Y, Li C, Wang Q, Yin Y, Hameed MF, Feil E, Chen S, Wang Z, Liu Y-X, Li R. 2025. Benchmarking of analysis tools and pipeline development for nanopore long-read metagenomics. *Sci Bull Sci Found Philipp* 70:1591–1595. <https://doi.org/10.1016/j.scib.2025.03.044>
88. Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34:3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
89. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. *Gigascience* 10:giab008. <https://doi.org/10.1093/gigascience/giab008>
90. Wick RR, Judd LM, Gorrie CL, Holt KE. 2017. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microb Genom* 3:e000132. <https://doi.org/10.1099/mgen.0.000132>
91. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842. <https://doi.org/10.1093/bioinformatics/btq033>
92. Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, Li C, Dvorin 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>
93. Kajitani R, Noguchi H, Gotoh Y, Ogura Y, Yoshimura D, Okuno M, Toyoda A, Kuwahara T, Hayashi T, Itoh T. 2021. MetaPlatanus: a metagenome assembler that combines long-range sequence links and species-specific features. *Nucleic Acids Res* 49:e130. <https://doi.org/10.1093/nar/gkab831>
94. Hu J, Fan J, Sun Z, Liu S. 2020. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* 36:2253–2255. <https://doi.org/10.1093/bioinformatics/btz891>
95. Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, Zhu Q, Bolzan M, Cumbo F, May U, Sanders JG, Zolfo M, Kopylova E, Pasolli E, Knight R, Mirarab S, Huttenhower C, Segata N. 2020. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat Commun* 11:2500. <https://doi.org/10.1038/s41467-020-16366-7>
96. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068–2069. <https://doi.org/10.1093/bioinformatics/btu153>
97. Zhao D, Salas-Leiva DE, Williams SK, Dunn KA, Shao JD, Roger AJ. 2025. Eukfinder: a pipeline to retrieve microbial eukaryote genome sequences from metagenomic data. *mBio* 16:e00699-25. <https://doi.org/10.1128/mbio.00699-25>
98. Lin H-H, Liao Y-C. 2016. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Sci Rep* 6:24175. <https://doi.org/10.1038/srep24175>
99. Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, Nilsson RH. 2015. METAXA2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol Ecol Resour* 15:1403–1414. <https://doi.org/10.1111/1755-0998.12399>
100. Van Nguyen H, Lavenier D. 2009. PLAST: parallel local alignment search tool for database comparison. *BMC Bioinform* 10:329. <https://doi.org/10.1186/1471-2105-10-329>
101. Cavalcante JVF, Dantas de Souza I, Morais DAA, Dalmolin RJS. 2024. EURYALE: a versatile Nextflow pipeline for taxonomic classification and functional annotation of metagenomics data. 2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB); Natal, Brazil. <https://doi.org/10.1109/CIBCB58642.2024.10702116>
102. Morais DAA, Cavalcante JVF, Monteiro SS, Pasquali MAB, Dalmolin RJS. 2022. MEDUSA: a pipeline for sensitive taxonomic classification and flexible functional annotation of metagenomic shotgun sequences. *Front Genet* 13:814437. <https://doi.org/10.3389/fgene.2022.814437>
103. Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32:3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>
104. Coordinators NR. 2014. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 42:D7–D17. <https://doi.org/10.1093/nar/gkt1146>
105. Afgan E, Nekrutenko A, Grüning BA, Blankenberg D, Goecks J, Schatz MC, Ostrovsky AE, Mahmoud A, Lonie AJ, Syme A, et al. 2022. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res* 50:W345–W351. <https://doi.org/10.1093/nar/gkac247>
106. Li H. 2025. Toolkit for processing sequences in FASTA/Q formats. GitHub. <https://github.com/lh3/seqtk>
107. von Meijenfeldt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. 2019. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biol* 20:217. <https://doi.org/10.1186/s13059-019-1817-x>
108. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15:R46. <https://doi.org/10.1186/gb-2014-15-3-r46>
109. Cornet L, Durieu B, Baert F, D'hooge E, Colignon D, Meunier L, Lupo V, Cleenwerck I, Daniel H-M, Rigouts L, Sirjacobs D, Declercq S, Vandamme P, Wilmotte A, Baurain D, Becker P. 2022. The GEN-ERA toolbox: unified and reproducible workflows for research in microbial genomics. *Gigascience* 12:1–10. <https://doi.org/10.1093/gigascience/giad022>
110. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. <https://doi.org/10.1089/cmb.2012.0021>
111. 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>
112. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963. <https://doi.org/10.1371/journal.pone.0112963>
113. Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. 2022. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol* 23:258. <https://doi.org/10.1186/s13059-022-02823-7>
114. 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>
115. Saary P, Mitchell AL, Finn RD. 2020. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biol* 21:244. <https://doi.org/10.1186/s13059-020-02155-4>
116. Manni M, Berkeley MR, Seppey M, Zdobnov EM. 2021. BUSCO: assessing genomic data quality and beyond. *Curr Protoc* 1:e323. <https://doi.org/10.1002/cpz1.323>
117. Cornet L, Meunier L, Van Vlierberghe M, Léonard RR, Durieu B, Lara Y, Misztak A, Sirjacobs D, Javaux EJ, Philippe H, Wilmotte A, Baurain D. 2018. Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PLoS One* 13:e0200323. <https://doi.org/10.1371/journal.pone.0200323>

118. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29:1072–1075. <https://doi.org/10.1093/bioinformatics/btt086>
119. Meunier L, Baurain D, Cornet L. 2023. AMAW: automated gene annotation for non-model eukaryotic genomes. *F1000Res* 12:186. <https://doi.org/10.12688/f1000research.129161.1>
120. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinform* 3:lqaa108. <https://doi.org/10.1093/nargab/lqaa108>
121. Queirós P, Delogu F, Hickl O, May P, Wilmes P. 2021. Mantis: flexible and consensus-driven genome annotation. *Gigascience* 10:giab042. <https://doi.org/10.1093/gigascience/giab042>
122. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 44:D457–62. <https://doi.org/10.1093/nar/gkv1070>
123. Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20:238. <https://doi.org/10.1186/s13059-019-1832-y>
124. Portik DM, Feng X, Benoit G, Nasko DJ, Auch B, Bryson SJ, Cano R, Carlin M, Damerum A, Farthing B, Grove JR, Islam M, Langford KW, Liachko I, Locken K, Mangelson H, Tang S, Zhang S, Quince C, Wilkinson JE. 2024. Highly accurate metagenome-assembled genomes from human gut microbiota using long-read assembly, binning, and consolidation methods. *bioRxiv*. <https://doi.org/10.1101/2024.05.10.593587>
125. Kalantar KL, Carvalho T, de Bourcy CFA, Dimitrov B, Dingle G, Egger R, Han J, Holmes OB, Juan Y-F, King R, et al. 2020. IDseq-an open source cloud-based pipeline and analysis service for metagenomic pathogen detection and monitoring. *Gigascience* 9:giaa111. <https://doi.org/10.1093/gigascience/giaa111>
126. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29:15–21. <https://doi.org/10.1093/bioinformatics/bts635>
127. Niu B, Fu L, Sun S, Li W. 2010. Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinform* 11:187. <https://doi.org/10.1186/1471-2105-11-187>
128. Wu TD, Nacu S. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 26:873–881. <https://doi.org/10.1093/bioinformatics/btq057>
129. Zhao Y, Tang H, Ye Y. 2012. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* 28:125–126. <https://doi.org/10.1093/bioinformatics/btr595>
130. Chen I-M, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, Hajek P, Ritter SJ, Webb C, Wu D, Varghese NJ, Reddy TBK, Mukherjee S, Ovchinnikova G, Nolan M, Seshadri R, Roux S, Visel A, Woyke T, Eloe-Fadros EA, Kyrpides NC, Ivanova NN. 2023. The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res* 51:D723–D732. <https://doi.org/10.1093/nar/gkac976>
131. Lomsadze A, Gemayel K, Tang S, Borodovsky M. 2018. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Res* 28:1079–1089. <https://doi.org/10.1101/gr.230615.117>
132. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: the protein families database in 2021. *Nucleic Acids Res* 49:D412–D419. <https://doi.org/10.1093/nar/gkaa913>
133. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. 2013. TIGRFAMs and genome properties in 2013. *Nucleic Acids Res* 41:D387–95. <https://doi.org/10.1093/nar/gks1234>
134. Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* 8:785–786. <https://doi.org/10.1038/nmeth.1701>
135. Möller S, Croning MDR, Apweiler R. 2001. Evaluation of methods for the prediction of membrane spanning regions. *Bioinformatics* 17:646–653. <https://doi.org/10.1093/bioinformatics/17.7.646>
136. McCulloch JA, Badger JH, Cannon N, Rodrigues RR, Valencia M, Barb JJ, Fernandes MR, Balaji A, Crowson L, O'hUigin C, Dzutsev A, Trinchieri G. 2023. JAMS - a framework for the taxonomic and functional exploration of microbiological genomic data. *bioRxiv*:2023.03.03.531026. <https://doi.org/10.1101/2023.03.03.531026>
137. Zdobnov EM, Apweiler R. 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847–848. <https://doi.org/10.1093/bioinformatics/17.9.847>
138. Chivian D, Jungbluth SP, Dehal PS, Wood-Charlson EM, Canon RS, Allen BH, Clark MM, Gu T, Land ML, Price GA, Riehl WJ, Sneddon MW, Sutormin R, Zhang Q, Cottingham RW, Henry CS, Arkin AP. 2023. Metagenome-assembled genome extraction and analysis from microbiomes using KBase. *Nat Protoc* 18:208–238. <https://doi.org/10.1038/s41596-022-00747-x>
139. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j* 17:10. <https://doi.org/10.14806/ej.17.1.200>
140. Zheng J, Ge Q, Yan Y, Zhang X, Huang L, Yin Y. 2023. dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res* 51:W115–W121. <https://doi.org/10.1093/nar/gkad328>
141. 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>
142. Song H-S, Ahamed F, Brown DML, Henry CS, Edirisinghe JN, Kessell AK, Nelson WC, McDermott JE, Hofmockel KS. 2023. OMEGA: a computationally efficient omics-guided global gapfilling algorithm for phenotype-consistent metabolic network reconstruction. Available from: <https://www.genomicscience.energy.gov/abstract/omega-a-computationally-efficient-omics-guided-global-gapfilling-algorithm-for-phenotype-consistent-metabolic-network-reconstruction>
143. Faria JP, Liu F, Edirisinghe JN, Gupta N, Seaver SMD, Freiburger AP, Zhang Q, Weisenhorn P, Conrad N, Zarecki R, Song H-S, DeJongh M, Best AA, Cottingham RW, Arkin AP, Henry CS. 2023. ModelSEED v2: High-throughput genome-scale metabolic model reconstruction with enhanced energy biosynthesis pathway prediction. *bioRxiv*. <https://doi.org/10.1101/2023.10.04.556561>
144. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 9:5114. <https://doi.org/10.1038/s41467-018-07641-9>
145. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLOS One* 5:e9490. <https://doi.org/10.1371/journal.pone.0009490>
146. Edgar RC. 2022. Muscle5: high-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat Commun* 13:6968. <https://doi.org/10.1038/s41467-022-34630-w>
147. Churchward B, Millet M, Bihouée A, Fertin G, Chaffron S. 2022. MAGNETO: an automated workflow for genome-resolved metagenomics. *mSystems* 7:e00432-22. <https://doi.org/10.1128/mSystems.00432-22>
148. Wingett SW, Andrews S. 2018. FastQ Screen: a tool for multi-genome mapping and quality control. *F1000Res* 7:1338. <https://doi.org/10.12688/f1000research.15931.2>
149. Benoit G, Peterlongo P, Mariadassou M, Drezen E, Schbath S, Lavenier D, Lemaître C. 2016. Multiple comparative metagenomics using multiset *k*-mer counting. *PeerJ Comput Sci* 2:e94. <https://doi.org/10.7717/peerj-cs.94>
150. Steinegger M, Söding J. 2018. Clustering huge protein sequence sets in linear time. *Nat Commun* 9:2542. <https://doi.org/10.1038/s41467-018-04964-5>
151. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, Rasmussen S, Brunak S, Pedersen O, Guarner F, de Vos WM, Wang J, Li J, Doré J, Ehrlich SD, Stamatakis A, Bork P. 2013. Metagenomic species profiling using universal phylogenetic marker genes. *Nat Methods* 10:1196–1199. <https://doi.org/10.1038/nmeth.2693>
152. Murovec B, Deutsch L, Stres B. 2020. Computational framework for high-quality production and large-scale evolutionary analysis of metagenome assembled genomes. *Mol Biol Evol* 37:593–598. <https://doi.org/10.1093/molbev/msz237>
153. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31:3691–3693. <https://doi.org/10.1093/bioinformatics/btv421>
154. Wu Y-W. 2018. ezTree: an automated pipeline for identifying phylogenetic marker genes and inferring evolutionary relationships among

- uncultivated prokaryotic draft genomes. *BMC Genomics* 19:921. <https://doi.org/10.1186/s12864-017-4327-9>
155. Maurice N, Lemaitre C, Vicedomini R, Frioux C. 2025. Mapler: a pipeline for assessing assembly quality in taxonomically rich metagenomes sequenced with HiFi reads. *Bioinformatics* 41:btaf334. <https://doi.org/10.1093/bioinformatics/btaf334>
 156. Benoit G, Raguideau S, James R, Phillippy AM, Chikhi R, Quince C. 2024. High-quality metagenome assembly from long accurate reads with metaMDBG. *Nat Biotechnol* 42:1378–1383. <https://doi.org/10.1038/s41587-023-01983-6>
 157. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 18:170–175. <https://doi.org/10.1038/s41592-020-01056-5>
 158. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* 33:574–576. <https://doi.org/10.1093/bioinformatics/btw663>
 159. Zorrilla F, Buric F, Patil KR, Zelezniak A. 2021. metaGEM: reconstruction of genome scale metabolic models directly from metagenomes. *Nucleic Acids Res* 49:e126. <https://doi.org/10.1093/nar/gkab815>
 160. Machado D, Andrejev S, Tramontano M, Patil KR. 2018. Fast automated reconstruction of genome-scale metabolic models for microbial species and communities. *Nucleic Acids Res* 46:7542–7553. <https://doi.org/10.1093/nar/gky537>
 161. Zelezniak A, Andrejev S, Ponomarova O, Mende DR, Bork P, Patil KR. 2015. Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proc Natl Acad Sci USA* 112:6449–6454. <https://doi.org/10.1073/pnas.1421834112>
 162. Lieven C, Beber ME, Olivier BG, Bergmann FT, Ataman M, Babaei P, Bartell JA, Blank LM, Chauhan S, Correia K, et al. 2020. MEMOTE for standardized genome-scale metabolic model testing. *Nat Biotechnol* 38:272–276. <https://doi.org/10.1038/s41587-020-0446-y>
 163. Emiola A, Oh J. 2018. High throughput in situ metagenomic measurement of bacterial replication at ultra-low sequencing coverage. *Nat Commun* 9:4956. <https://doi.org/10.1038/s41467-018-07240-8>
 164. Shaban B, Quiroga M del M, Turnbull R, Tescari E, Cao K-AL, Verbruggen H. 2023. MetaGenePipe: an automated, portable pipeline for contig-based functional and taxonomic analysis. *JOSS* 8:4851. <https://doi.org/10.21105/joss.04851>
 165. Poux S, Arighi CN, Magrane M, Bateman A, Wei C-H, Lu Z, Boutet E, Bye-A-Jee H, Famiglietti ML, Roechert B, UniProt Consortium T. 2017. On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics* 33:3454–3460. <https://doi.org/10.1093/bioinformatics/btx439>
 166. Eddy SR. 2008. A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput Biol* 4:e1000069. <https://doi.org/10.1371/journal.pcbi.1000069>
 167. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. 2020. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* 36:2251–2252. <https://doi.org/10.1093/bioinformatics/btz859>
 168. 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>
 169. Belmann P, Osterholz B, Kleinbölting N, Pühler A, Schlüter A, Sczyrba A. 2025. Metagenomics-Toolkit: the flexible and efficient cloud-based metagenomics workflow featuring machine learning-enabled resource allocation. *NAR Genomics Bioinform* 7. <https://doi.org/10.1093/nargab/lqaf093>
 170. Kokot M, Dlugosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. *Bioinformatics* 33:2759–2761. <https://doi.org/10.1093/bioinformatics/btx304>
 171. Rodriguez-R LM, Gunturu S, Tiedje JM, Cole JR, Konstantinidis KT. 2018. Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. *mSystems* 3:e00039-18. <https://doi.org/10.1128/mSystems.00039-18>
 172. Mallawaarachchi V, Lin Y. 2022. MetaCoAG: binning metagenomic contigs via composition, coverage and assembly graphs, p 70–85. In Pe'er I (ed), *Research in computational molecular biology. RECOMB 2022. Lecture notes in computer science*. Springer, Cham.
 173. Wang Z, Huang P, You R, Sun F, Zhu S. 2023. MetaBinner: a high-performance and stand-alone ensemble binning method to recover individual genomes from complex microbial communities. *Genome Biol* 24:1. <https://doi.org/10.1186/s13059-022-02832-6>
 174. Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E. 2021. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics* 37:3029–3031. <https://doi.org/10.1093/bioinformatics/btab184>
 175. Alcock BP, Huynh W, Chalil R, Smith KW, Raphenya AR, Wlodarski MA, Edalatmand A, Petkau A, Syed SA, Tsang KK, et al. 2023. CARD 2023: expanded curation, support for machine learning, and resistance prediction at the comprehensive antibiotic resistance database. *Nucleic Acids Res* 51:D690–D699. <https://doi.org/10.1093/nar/gkac920>
 176. Zimmermann J, Kaleta C, Waschina S. 2021. Gapseq: informed prediction of bacterial metabolic pathways and reconstruction of accurate metabolic models. *Genome Biol* 22:81. <https://doi.org/10.1186/s13059-021-02295-1>
 177. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. 2016. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* 8:12–24. <https://doi.org/10.1039/C5AY02550H>
 178. Witter R. 2020. Alignment- and reference-free phylogenomics with colored de Bruijn graphs. *Algorithms Mol Biol* 15:4. <https://doi.org/10.1186/s13015-020-00164-3>
 179. Salazar VW, Shaban B, Quiroga M del M, Turnbull R, Tescari E, Rossetto Marcelino V, Verbruggen H, Lê Cao K-A. 2022. Metaphor—a workflow for streamlined assembly and binning of metagenomes. *Gigascience* 12:1–12. <https://doi.org/10.1093/gigascience/giad055>
 180. Mainguy J, Vienne M, Fourquet J, Darbot V, Noiro C, Castinel A, Combes S, Gaspin C, Milan D, Donnadiou C, Iampietro C, Bouchez O, Pascal G, Hoede C. 2024. metagWGS, a comprehensive workflow to analyze metagenomic data using Illumina or PacBio HiFi reads. *bioRxiv*. <https://doi.org/10.1101/2024.09.13.612854>
 181. Joshi NA, Fass JN. 2021. Windowed adaptive trimming for fastq files using quality. *GitHub*. <https://github.com/najoshi/sickle>
 182. Mainguy J, Hoede C. 2024. Binette: a fast and accurate bin refinement tool to construct high quality metagenome assembled genomes. *JOSS* 9:6782. <https://doi.org/10.21105/joss.06782>
 183. Kumar S, Jones M, Koutsouvolos G, Clarke M, Blaxter M. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front Genet* 4:237. <https://doi.org/10.3389/fgene.2013.00237>
 184. Hildebrand F, Gossmann TI, Frioux C, Özkurt E, Myers PN, Ferretti P, Kuhn M, Bahram M, Nielsen HB, Bork P. 2021. Dispersal strategies shape persistence and evolution of human gut bacteria. *Cell Host Microbe* 29:1167–1176. <https://doi.org/10.1016/j.chom.2021.05.008>
 185. Hildebrand F, Tadeo R, Voigt AY, Bork P, Raes J. 2014. LotuS: an efficient and user-friendly OTU processing pipeline. *Microbiome* 2:30. <https://doi.org/10.1186/2049-2618-2-30>
 186. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 37:540–546. <https://doi.org/10.1038/s41587-019-0072-8>
 187. Liu C-C, Dong S-S, Chen J-B, Wang C, Ning P, Guo Y, Yang T-L. 2022. MetaDecoder: a novel method for clustering metagenomic contigs. *Microbiome* 10:46. <https://doi.org/10.1186/s40168-022-01237-8>
 188. Drula E, Garron M-L, Dogan S, Lombard V, Henrissat B, Terrapon N. 2022. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res* 50:D571–D577. <https://doi.org/10.1093/nar/gkab1045>
 189. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh HJ, Cuenca M, Hingamp P, Alves R, Costea PI, Coelho LP, Schmidt TSB, Almeida A, Mitchell AL, Finn RD, Huerta-Cepas J, Bork P, Zeller G, Sunagawa S. 2019. Microbial abundance, activity and population genomic profiling with mOTUs2. *Nat Commun* 10:1014. <https://doi.org/10.1038/s41467-019-08844-4>
 190. Blanco-Míguez A, Beghini F, Cumbo F, McIver LJ, Thompson KN, Zolfo M, Manghi P, Dubois L, Huang KD, Thomas AM, et al. 2023. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat Biotechnol* 41:1633–1644. <https://doi.org/10.1038/s41587-023-01688-w>
 191. Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv*. <https://doi.org/10.48550/arXiv.1207.3907>
 192. Cokelaer T, Desvillechabrol D, Legendre R, Cardon M. 2017. “Sequana”: a Set of Snakemake NGS pipelines. *JOSS* 2:352. <https://doi.org/10.21105/joss.00352>

193. Richardson L, Allen B, Baldi G, Beracochea M, Bileschi ML, Burdett T, Burgin J, Caballero-Pérez J, Cochran 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>
194. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423. <https://doi.org/10.1093/bioinformatics/btp163>
195. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH, UniProt Consortium. 2015. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* 31:926–932. <https://doi.org/10.1093/bioinformatics/btu739>
196. Rho M, Tang H, Ye Y. 2010. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res* 38:e191. <https://doi.org/10.1093/nar/gkq747>
197. Blin K, Shaw S, Augustijn HE, Reitz ZL, Biermann F, Alanjary M, Fetter A, Terlouw BR, Metcalf WW, Helfrich EJN, van Wezel GP, Medema MH, Weber T. 2023. antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res* 51:W46–W50. <https://doi.org/10.1093/nar/gkad344>
198. Ziemski M, Gehret L, Simard A, Dau SC, Risch V, Grabocka D, Matzoros C, Wood C, Cabrera PM, Hernández-Velázquez R, Herman C, Evans K, Robeson MS, Bolyen E, Caporaso JG, Bokulich NA. 2025. MOSHPIT: accessible, reproducible metagenome data science on the QIIME 2 framework. *bioRxiv*:2025.01.27.635007. <https://doi.org/10.1101/2025.01.27.635007>
199. van DR, Hölzer M, Viehweger A, Müller B, Bongcam-Rudloff E, Brandt C. 2021. Metagenomics workflow for hybrid assembly, differential coverage binning, metatranscriptomics and pathway analysis (MUFFIN). *PLoS Comput Biol* 17:1–13. <https://doi.org/10.1371/journal.pcbi.1008716>
200. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 50:D785–D794. <https://doi.org/10.1093/nar/gkab776>
201. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 14:417–419. <https://doi.org/10.1038/nmeth.4197>
202. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Muscili E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644–652. <https://doi.org/10.1038/nbt.1883>
203. Liu L, Yang Y, Deng Y, Zhang T. 2022. Nanopore long-read-only metagenomics enables complete and high-quality genome reconstruction from mock and complex metagenomes. *Microbiome* 10:209. <https://doi.org/10.1186/s40168-022-01415-8>
204. Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 27:737–746. <https://doi.org/10.1101/gr.214270.116>
205. Oxford Nanopore Technologies. 2025. Medaka: sequence correction provided by ONT research. GitHub. <https://github.com/nanoporetech/medaka>
206. Bateman A, Martin M-J, Orchard S, Magrane M, Adesina A, Ahmad S, Bowler-Barnett EH, Bye-A-Jee H, Carpentier D, Denny P, et al. 2025. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* 53:D609–D617. <https://doi.org/10.1093/nar/gkac1010>
207. Krakau S, Straub D, Gourel H, Gabernet G, Nahnsen S. 2022. NF-core/mag: a best-practice pipeline for metagenome hybrid assembly and binning. *NAR Genomics Bioinform* 4:lqac007. <https://doi.org/10.1093/nargab/lqac007>
208. Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res Notes* 9:88. <https://doi.org/10.1186/s13104-016-1900-2>
209. Antipov D, Korobeynikov A, McLean JS, Pevzner PA. 2016. hybrid-SPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* 32:1009–1015. <https://doi.org/10.1093/bioinformatics/btv688>
210. Levy Karin E, Mirdita M, Söding J. 2020. MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* 8:48. <https://doi.org/10.1186/s40168-020-00808-x>
211. Borry M, Hübner A, Rohrlach AB, Warinner C. 2021. PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA *de novo* assembly. *PeerJ* 9:e11845. <https://doi.org/10.7717/peerj.11845>
212. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, Chain PSG, Nayfach S, Kyrpides NC. 2024. Identification of mobile genetic elements with geNomad. *Nat Biotechnol* 42:1303–1312. <https://doi.org/10.1038/s41587-023-01953-y>
213. Karlicki M, Antonowicz S, Karnkowska A. 2022. Tiara: deep learning-based classification system for eukaryotic sequences. *Bioinformatics* 38:344–350. <https://doi.org/10.1093/bioinformatics/btab672>
214. de Almeida FM, de Campos TA, Pappas GJ Jr. 2023. Scalable and versatile container-based pipelines for *de novo* genome assembly and bacterial annotation. *F1000Res* 12:1205. <https://doi.org/10.12688/f1000research.139488.1>
215. Leger A, Leonardi T. 2019. pycoQC, interactive quality control for Oxford Nanopore sequencing. *JOSS* 4:1236. <https://doi.org/10.21105/joss.01236>
216. Seemann T. 2020. Shovill: assemble bacterial isolate genomes from Illumina paired-end reads. GitHub. <https://github.com/tseemann/shovill>
217. Haghsheenas E, Asghari H, Stoye J, Chauve C, Hach F. 2020. HASLR: fast hybrid assembly of long reads. *iScience* 23:101389. <https://doi.org/10.1016/j.isci.2020.101389>
218. Vaser R, Šikić M. 2021. Time- and memory-efficient genome assembly with Raven. *Nat Comput Sci* 1:332–336. <https://doi.org/10.1038/s43588-021-00073-4>
219. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, Armstrong J, Tigyi K, Maurer N, Koren S, et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient *de novo* assembly of eleven human genomes. *Nat Biotechnol* 38:1044–1053. <https://doi.org/10.1038/s41587-020-0503-6>
220. Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* 17:155–158. <https://doi.org/10.1038/s41592-019-0669-3>
221. Graham ED, Heidelberg JF, Tully BJ. 2018. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J* 12:1861–1866. <https://doi.org/10.1038/s41396-018-0091-3>
222. Seemann T. 2018. Barnap: bacterial ribosomal RNA predictor. GitHub. <https://github.com/tseemann/barnap>
223. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, Hoffmann M, Pettengill JB, Prasad AB, Tillman GE, Tyson GH, Klimke W. 2021. AMRFinderPlus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci Rep* 11:12728. <https://doi.org/10.1038/s41598-021-91456-0>
224. Starikova EV, Tikhonova PO, Prianichnikov NA, Rands CM, Zdobnov EM, Ilin EN, Govorun VM. 2020. Phigaro: high-throughput prophage sequence annotation. *Bioinformatics* 36:3882–3884. <https://doi.org/10.1093/bioinformatics/btaa250>
225. Dong W, Fan X, Guo Y, Wang S, Jia S, Lv N, Yuan T, Pan Y, Xue Y, Chen X, Xiong Q, Yang R, Zhao W, Zhu B. 2024. An expanded database and analytical toolkit for identifying bacterial virulence factors and their associations with chronic diseases. *Nat Commun* 15:8084. <https://doi.org/10.1038/s41467-024-51864-y>
226. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. 2014. *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* 58:3895–3903. <https://doi.org/10.1128/AAC.02412-14>
227. Jolley KA, Maiden MCJ. 2010. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinform* 11:595. <https://doi.org/10.1186/1471-2105-11-595>
228. Schwengers O, Barth P, Falgenhauer L, Hain T, Chakraborty T, Goessmann A. 2020. Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb Genom* 6:e000398. <https://doi.org/10.1099/mgen.0.000398>
229. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, Wishart DS. 2016. PHASTER: a better, faster version of the PHAST phage search tool. *Nucleic Acids Res* 44:W16–W21. <https://doi.org/10.1093/nar/gkw387>

230. Arango-Argoty GA, Guron GKP, Garner E, Riquelme MV, Heath LS, Pruden A, Vikesland PJ, Zhang L. 2020. ARGminer: a web platform for the crowdsourcing-based curation of antibiotic resistance genes. *Bioinformatics* 36:2966–2973. <https://doi.org/10.1093/bioinformatics/btaa095>
231. Florensa AF, Kaas RS, Clausen P, Aytan-Aktug D, Aarestrup FM. 2022. ResFinder – an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microb Genom* 8:000748. <https://doi.org/10.1099/mgen.0.000748>
232. Narayanasamy S, Jarosz Y, Muller EEL, Heintz-Buschart A, Herold M, Kaysen A, Laczny CC, Pintel N, May P, Wilmes P. 2016. IMP: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome Biol* 17:260. <https://doi.org/10.1186/s13059-016-1116-8>
233. Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* 28:3211–3217. <https://doi.org/10.1093/bioinformatics/bts611>
234. Schudoma C. 2023. Gff_quantifier. https://github.com/cschu/gff_quantifier
235. Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 34:525–527. <https://doi.org/10.1038/nbt.3519>
236. Tadrent N, Dedeine F, Hervé V. 2022. *SnakeMAGs*: a simple, efficient, flexible and scalable workflow to reconstruct prokaryotic genomes from metagenomes. *F1000Res* 11:1522. <https://doi.org/10.12688/f1000research.128091.2>
237. Schmidt TSB, Fullam A, Ferretti P, Orakov A, Maistrenko OM, Ruscheweyh H-J, Letunic I, Duan Y, Van Rossum T, Sunagawa S, Mende DR, Finn RD, Kuhn M, Pedro Coelho L, Bork P. 2024. SPIRE: a Searchable, Planetary-scale microbiome REsource. *Nucleic Acids Res* 52:D777–D783. <https://doi.org/10.1093/nar/gkad943>
238. Coelho LP, Alves R, Monteiro P, Huerta-Cepas J, Freitas AT, Bork P. 2019. NG-meta-profiler: fast processing of metagenomes using NGLess, a domain-specific language. *Microbiome* 7:84. <https://doi.org/10.1186/s40168-019-0684-8>
239. Seemann T. 2020. ABRicate: mass screening of contigs for antimicrobial and virulence genes. GitHub. <https://github.com/tseemann/abricate>
240. Doster E, Lakin SM, Dean CJ, Wolfe C, Young JG, Boucher C, Belk KE, Noyes NR, Morley PS. 2020. MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Res* 48:D561–D569. <https://doi.org/10.1093/nar/gkz1010>
241. Santos-Júnior CD, Pan S, Zhao X-M, Coelho LP. 2020. Macrel: antimicrobial peptide screening in genomes and metagenomes. *PeerJ* 8:e10555. <https://doi.org/10.7717/peerj.10555>
242. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* 17:132. <https://doi.org/10.1186/s13059-016-0997-x>
243. Tamames J, Puente-Sánchez F. 2018. SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline. *Front Microbiol* 9:3349. <https://doi.org/10.3389/fmicb.2018.03349>
244. Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 27:863–864. <https://doi.org/10.1093/bioinformatics/btr026>
245. Parks DH. 2020. CompareM: a toolbox for comparative genomics. GitHub. <https://github.com/donovan-h-parks/CompareM>
246. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: a fast and versatile genome alignment system. *PLOS Comput Biol* 14:e1005944. <https://doi.org/10.1371/journal.pcbi.1005944>
247. Puente-Sánchez F, García-García N, Tamames J. 2020. SQMtools: automated processing and visual analysis of 'omics data with R and anvio. *BMC Bioinform* 21:358. <https://doi.org/10.1186/s12859-020-03703-2>
248. Sjöqvist C, Delgado LF, Alneberg J, Andersson AF. 2021. Ecologically coherent population structure of uncultivated bacterioplankton. *ISME J* 15:3034–3049. <https://doi.org/10.1038/s41396-021-00985-z>
249. Clarke EL, Taylor LJ, Zhao C, Connell A, Lee JJ, Fett B, Bushman FD, Bittinger K. 2019. Sunbeam: an extensible pipeline for analyzing metagenomic sequencing experiments. *Microbiome* 7:46. <https://doi.org/10.1186/s40168-019-0658-x>
250. Espinoza JL, Phillips A, Prentice MB, Tan GS, Kamath PL, Lloyd KG, Dupont CL. 2024. Unveiling the microbial realm with VEBA 2.0: a modular bioinformatics suite for end-to-end genome-resolved prokaryotic, (micro)eukaryotic and viral multi-omics from either short- or long-read sequencing. *Nucleic Acids Res* 52:e63–e63. <https://doi.org/10.1093/nar/gkae528>
251. Bushmanova E, Antipov D, Lapidus A, Pribelski AD. 2019. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience* 8:giz100. <https://doi.org/10.1093/gigascience/giz100>
252. Nayfach S, Camargo AP, Schulz F, Eloë-Fadrosch E, Roux S, Kyrpides NC. 2021. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* 39:578–585. <https://doi.org/10.1038/s41587-020-00774-7>
253. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. 2017. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome* 5:69. <https://doi.org/10.1186/s40168-017-0283-5>
254. Zdouc MM, Blin K, Louwen NLL, Navarro J, Loureiro C, Bader CD, Bailey CB, Barra L, Booth TJ, Bozhüyük KAJ, et al. 2025. MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Res* 53:D678–D690. <https://doi.org/10.1093/nar/gkae1115>
255. Eberhardt RY, Haft DH, Punta M, Martin M, O'Donovan C, Bateman A. 2012. AntiFam: a tool to help identify spurious ORFs in protein annotation. *Database (Oxford)* 2012:bas003. <https://doi.org/10.1093/database/bas003>
256. Ruiz-Perez CA, Conrad RE, Konstantinidis KT. 2021. MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. *BMC Bioinform* 22:11. <https://doi.org/10.1186/s12859-020-03940-5>
257. Shaw J, Yu YW. 2025. Rapid species-level metagenome profiling and containment estimation with sylph. *Nat Biotechnol* 43:1348–1359. <https://doi.org/10.1038/s41587-024-02412-y>
258. Weber N, Liou D, Dommer J, MacMenamin P, Quiñones M, Misner I, Oler AJ, Wan J, Kim L, Coakley McCarthy M, Ezeji S, Noble K, Hurt DE. 2018. Nephele: a cloud platform for simplified, standardized and reproducible microbiome data analysis. *Bioinformatics* 34:1411–1413. <https://doi.org/10.1093/bioinformatics/btx617>
259. Ye Y, Doak TG. 2009. A parsimony approach to biological pathway reconstruction/inference for genomes and metagenomes. *PLOS Comput Biol* 5:e1000465. <https://doi.org/10.1371/journal.pcbi.1000465>
260. Shen W, Xiang H, Huang T, Tang H, Peng M, Cai D, Hu P, Ren H. 2023. KMCP: accurate metagenomic profiling of both prokaryotic and viral populations by pseudo-mapping. *Bioinformatics* 39:btac845. <https://doi.org/10.1093/bioinformatics/btac845>
261. Ayling M, Clark MD, Leggett RM. 2020. New approaches for metagenome assembly with short reads. *Brief Bioinform* 21:584–594. <https://doi.org/10.1093/bib/bbz020>
262. Goussarov G, Mysara M, Cleenwerck I, Claesen J, Leys N, Vandamme P, Van Houdt R. 2024. Benchmarking short-, long- and hybrid-read assemblers for metagenome sequencing of complex microbial communities. *Microbiology (Reading)* 170:001469. <https://doi.org/10.1099/mic.0.001469>
263. Meyer F, Lesker TR, Koslicki D, Fritz A, Gurevich A, Darling AE, Sczyrba A, Bremges A, McHardy AC. 2021. Tutorial: assessing metagenomics software with the CAMI benchmarking toolkit. *Nat Protoc* 16:1785–1801. <https://doi.org/10.1038/s41596-020-00480-3>
264. Wang Z, Wang Y, Fuhrman JA, Sun F, Zhu S. 2020. Assessment of metagenomic assemblers based on hybrid reads of real and simulated metagenomic sequences. *Brief Bioinform* 21:777–790. <https://doi.org/10.1093/bib/bbz025>
265. Rozov R, Goldshlager G, Halperin E, Shamir R. 2018. Faucet: streaming de novo assembly graph construction. *Bioinformatics* 34:147–154. <https://doi.org/10.1093/bioinformatics/btx471>
266. Brown CL, Keenum IM, Dai D, Zhang L, Vikesland PJ, Pruden A. 2021. Critical evaluation of short, long, and hybrid assembly for contextual analysis of antibiotic resistance genes in complex environmental metagenomes. *Sci Rep* 11:3753. <https://doi.org/10.1038/s41598-021-83081-8>
267. Herazo-Álvarez J, Mora M, Cuadros-Orellana S, Vilches-Ponce K, Hernández-García R. 2025. A review of neural networks for metagenomic binning. *Brief Bioinform* 26:bbaf065. <https://doi.org/10.1093/bib/bbaf065>

268. Cansdale A, Chong JPJ. 2024. MAGqual: a stand-alone pipeline to assess the quality of metagenome-assembled genomes. *Microbiome* 12:1–10. <https://doi.org/10.1186/s40168-024-01949-z>
269. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. 2014. GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* 2:e603. <https://doi.org/10.7717/peerj.603>
270. Yue Y, Huang H, Qi Z, Dou HM, Liu XY, Han TF, Chen Y, Song XJ, Zhang YH, Tu J. 2020. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. *BMC Bioinformatics* 21:1–15. <https://doi.org/10.1186/s12859-020-03667-3>
271. Yepes-García J, Falquet L. 2024. Metagenome quality metrics and taxonomical annotation visualization through the integration of MAGFlow and BigMAG. *F1000Res* 13:640. <https://doi.org/10.12688/f1000research.152290.2>
272. Simon P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, Roure B, Satoh N, Quéinnec É, Ereskovsky A, Lapébie P, Corre E, Delsuc F, King N, Wörheide G, Manuel M. 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr Biol* 27:958–967. <https://doi.org/10.1016/j.cub.2017.02.031>
273. Edwin NR, Fitzpatrick AH, Brennan F, Abram F, O'Sullivan O. 2024. An in-depth evaluation of metagenomic classifiers for soil microbiomes. *Environmental Microbiome* 19:19. <https://doi.org/10.1186/s40793-024-00561-w>
274. Timilsina M, Chundru D, Pradhan AK, Blaustein RA, Ghanem M. 2025. Benchmarking metagenomic pipelines for the detection of foodborne pathogens in simulated microbial communities. *J Food Prot* 88:100583. <https://doi.org/10.1016/j.jfip.2025.100583>
275. Irankhah L, Khorsand B, Naghibzadeh M, Savadi A. 2024. Analyzing the performance of short-read classification tools on metagenomic samples toward proper diagnosis of diseases. *J Bioinform Comput Biol* 22:2450012. <https://doi.org/10.1142/S0219720024500124>
276. Van Uffelen A, Posadas A, Roosens NHC, Marchal K, De Keersmaecker SCJ, Vanneste K. 2024. Benchmarking bacterial taxonomic classification using nanopore metagenomics data of several mock communities. *Sci Data* 11:864. <https://doi.org/10.1038/s41597-024-03672-8>
277. Liang Q, Bible PW, Liu Y, Zou B, Wei L. 2020. DeepMicrobes: taxonomic classification for metagenomics with deep learning. *NAR Genomics Bioinform* 2:lqaa009. <https://doi.org/10.1093/nargab/lqaa009>
278. Pusadkar V, Azad RK. 2023. Benchmarking metagenomic classifiers on simulated ancient and modern metagenomic data. *Microorganisms* 11:2478. <https://doi.org/10.3390/microorganisms11102478>
279. Marić J, Križanović K, Riondet S, Nagarajan N, Šikić M. 2024. Comparative analysis of metagenomic classifiers for long-read sequencing datasets. *BMC Bioinform* 25:15. <https://doi.org/10.1186/s12859-024-05634-8>
280. Lin B, Luo X, Liu Y, Jin X. 2024. A comprehensive review and comparison of existing computational methods for protein function prediction. *Brief Bioinform* 25:bbae289. <https://doi.org/10.1093/bib/bbae289>
281. The Gene Ontology Consortium. 2019. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res* 47:D330–D338. <https://doi.org/10.1093/nar/gky1055>
282. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. 2018. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res* 46:D624–D632. <https://doi.org/10.1093/nar/gkx1134>
283. Zeller M, Huson DH. 2022. Comparison of functional classification systems. *NAR Genomics Bioinform* 4:lqac090. <https://doi.org/10.1093/nargab/lqac090>
284. Xu Y, Lin Z, Tang C, Tang Y, Cai Y, Zhong H, Wang X, Zhang W, Xu C, Wang J, Wang J, Yang H, Yang L, Gao Q. 2019. A new massively parallel nanoball sequencing platform for whole exome research. *BMC Bioinform* 20:153. <https://doi.org/10.1186/s12859-019-2751-3>
285. Liang H, Zou Y, Wang M, Hu T, Wang H, He W, Ju Y, Guo R, Chen J, Guo F, et al. 2025. Efficiently constructing complete genomes with CycloneSEQ to fill gaps in bacterial draft assemblies. *GigaByte* 2025:gigabyte154. <https://doi.org/10.46471/gigabyte.154>
286. Kim H-M, Jeon S, Chung O, Jun JH, Kim H-S, Blazyte A, Lee H-Y, Yu Y, Cho YS, Bolser DM, Bhak J. 2021. Comparative analysis of 7 short-read sequencing platforms using the Korean Reference Genome: MGI and Illumina sequencing benchmark for whole-genome sequencing. *Gigascience* 10:giab014. <https://doi.org/10.1093/gigascience/giab014>
287. Vosloo S, Huo L, Anderson CL, Dai Z, Sevillano M, Pinto A. 2021. Evaluating *de novo* assembly and binning strategies for time series drinking water metagenomes. *Microbiol Spectr* 9:e01434-21. <https://doi.org/10.1128/Spectrum.01434-21>
288. Lynn HM, Gordon JL. 2025. Sequential co-assembly reduces computational resources and errors in metagenome-assembled genomes. *Cell Rep Methods* 5:101005. <https://doi.org/10.1016/j.crmeth.2025.101005>
289. Goldfarb T, Kodali VK, Pujar S, Brover V, Robbertse B, Farrell CM, Oh D-H, Astashyn A, Ermolaeva O, Haddad D, et al. 2025. NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Res* 53:D243–D257. <https://doi.org/10.1093/nar/gkae1038>
290. Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez F, Canon S, et al. 2018. KBase: The United States department of energy systems biology knowledgebase. *Nat Biotechnol* 36:566–569. <https://doi.org/10.1038/nbt.4163>
291. Achudhan AB, Kannan P, Gupta A, Saleena LM. 2024. A review of web-based metagenomics platforms for analysing next-generation sequence data. *Biochem Genet* 62:621–632. <https://doi.org/10.1007/s10528-023-10467-w>
292. Wratten L, Wilm A, Göke J. 2021. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nat Methods* 18:1161–1168. <https://doi.org/10.1038/s41592-021-01254-9>
293. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, Wilm A, Holtgrewe M, Rahmann S, Nahnsen S, Köster J. 2021. Sustainable data analysis with Snakemake. *F1000Res* 10:33. <https://doi.org/10.12688/f1000research.29032.1>
294. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. *Nat Biotechnol* 35:316–319. <https://doi.org/10.1038/nbt.3820>
295. OpenWDL S. 2025. Guides and reference material for the workflow description language. GitHub. <https://github.com/openwdl/docs>
296. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. 2020. The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* 38:276–278. <https://doi.org/10.1038/s41587-020-0439-x>
297. Roach MJ, Pierce-Ward NT, Suchecki R, Mallawaarachchi V, Papudeshi B, Handley SA, Brown CT, Watson-Haigh NS, Edwards RA. 2022. Ten simple rules and a template for creating workflows-as-applications. *PLoS Comput Biol* 18:e1010705. <https://doi.org/10.1371/journal.pcbi.1010705>
298. Reiter T, Brooks PT, Irber L, Joslin SEK, Reid CM, Scott C, Brown CT, Pierce-Ward NT. 2021. Streamlining data-intensive biology with workflow systems. *Gigascience* 10:giab140. <https://doi.org/10.1093/gigascience/giab140>
299. Kadri S, Sboner A, Sigaras A, Roy S. 2022. Containers in bioinformatics: applications, practical considerations, and best practices in molecular pathology. *J Mol Diagn* 24:442–454. <https://doi.org/10.1016/j.jmoldx.2022.01.006>
300. Badia RM, Conejero J, Diaz C, Ejarque J, Lezzi D, Lordan F, Ramon-Cortes C, Sirvent R. 2015. COMP Superscalar, an interoperable programming framework. *SoftwareX* 3–4:32–36. <https://doi.org/10.1016/j.softx.2015.10.004>
301. Espinoza JL. 2023. GenoPype: architecture for creating bash pipelines, in particular, for bioinformatics. GitHub. <https://github.com/jolespin/genopype>

AUTHOR BIOS

Jeferyd Yepes-García is PhD candidate in bioinformatics at the University of Fribourg and the Swiss Institute of Bioinformatics (SIB). He started his doctoral studies under the Swiss Government Excellence Scholarship program after finishing his master's program in engineering at the University of Antioquia (Colombia). His thesis is focused on the analysis of the rice straw degradation process through a metagenomics approach, with special interest in the process of reconstructing metagenome-assembled genomes from the microbial community. Furthermore, as part of his thesis, he is fine-tuning protein language models with the incorporation of structural information to automatize the protein function prediction from microbiome data. His work has been recognized as SIB Remarkable Output 2024 and the Best Poster in Bioinformatics during Life Sciences Switzerland (LS²) Annual Meeting 2025.



Dr. **Laurent Falquet** researcher's career spans biochemistry, bioinformatics, and metagenomics. Initially, he studied deubiquitinating enzymes, demonstrating that UBP5_HUMAN cleaves di-ubiquitin substrates and requires zinc for activity. During a postdoc, he managed protein profiles in the PROSITE database and discovered new protein domains. At the Swiss Institute of Bioinformatics, he coordinated EMBnet Switzerland and later led the Vital-IT genome assembly group, gaining expertise in assembling bacterial and eukaryotic genomes using NGS. He moved to the University of Fribourg as head of the bioinformatics platform and contributed to a Sinergia project on bacterial toxin-antitoxin systems linked to antibiotic resistance, developing the TASmania database. His current research focuses on metagenomics, including a PhD project on rice straw degradation and the SOLANUM project using native microbiomes to protect potatoes from late blight. He has also been involved in COST actions on statistical and machine learning approaches in microbiome studies and crop microbiome networks.

