Computational and Structural Biotechnology Journal 19 (2021) 6301–6314

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data

Chao Yang [a,1], Debajyoti Chowdhury [b,c,1], Zhenmiao Zhang [a], William K. Cheung [a], Aiping Lu [b,c], Zhaoxiang Bian [d,e], Lu Zhang [a,b,*]

[a] Department of Computer Science, Hong Kong Baptist University, Hong Kong Special Administrative Region
[b] Computational Medicine Lab, Hong Kong Baptist University, Hong Kong Special Administrative Region
[c] Institute of Integrated Bioinformedicine and Translational Sciences, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong Special Administrative Region
[d] Institute of Brain and Gut Research, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong Special Administrative Region
[e] Chinese Medicine Clinical Study Center, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong Special Administrative Region

ARTICLE INFO

Article history:
Received 30 August 2021
Received in revised form 17 November 2021
Accepted 17 November 2021
Available online 23 November 2021

Keywords:
Metagenomic sequencing
Metagenome-assembled genomes
Genome assembly
Metagenome binning
Gene prediction
Gene functional annotation
Taxonomic classification
Microbial abundance profiling

ABSTRACT

Metagenomic sequencing provides a culture-independent avenue to investigate the complex microbial communities by constructing metagenome-assembled genomes (MAGs). A MAG represents a microbial genome by a group of sequences from genome assembly with similar characteristics. It enables us to identify novel species and understand their potential functions in a dynamic ecosystem. Many computational tools have been developed to construct and annotate MAGs from metagenomic sequencing, however, there is a prominent gap to comprehensively introduce their background and practical performance. In this paper, we have thoroughly investigated the computational tools designed for both upstream and downstream analyses, including metagenome assembly, metagenome binning, gene prediction, functional annotation, taxonomic classification, and profiling. We have categorized the commonly used tools into unique groups based on their functional background and introduced the underlying core algorithms and associated information to demonstrate a comparative outlook. Furthermore, we have emphasized the computational requisition and offered guidance to the users to select the most efficient tools. Finally, we have indicated current limitations, potential solutions, and future perspectives for further improving the tools of MAG construction and annotation. We believe that our work provides a consolidated resource for the current stage of MAG studies and shed light on the future development of more effective MAG analysis tools on metagenomic sequencing.

Contents

CNN, convolutional neural network; DBG, De Bruijn graph; GTDB, Genome Taxonomy Database; HMM, Hidden Markov Model; KEGG, Kyoto Encyclopedia of Genes and Genomes; LCA, lowest common ancestor; LPA, label propagation algorithm; MAGs, metagenome-assembled genomes; OLC, overlap-layout consensus; ONT, Oxford Nanopore Technologies; ORFs, open reading frames; PacBio, Pacific Biosciences; QC, quality control; SLR, synthetic long reads; TNFs, tetranucleotide frequencies.

* Corresponding author at: Department of Computer Science, Hong Kong Baptist University, Hong Kong Special Administrative Region.
E-mail address: ericluzhang@hkbu.edu.hk (L. Zhang).

[1] Co-first authors and these authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2021.11.028

1. Introduction

Microbes are essential for nutrient cycling and metabolic processes in living beings and the environment [1,2]. Despite their crucial associations with physiology, their genomic characteristics and co-existence with other species have been poorly characterized. Microbes can be characterized using traditional culture-dependent approaches (Fig. 1**A**) that involve the isolation and sequencing of individual microbes from lab-based cultures [3,4]. However, some microbes cannot be easily isolated and cultured *in vitro* due to their complex natural habitats, such as unreproducible environmental, temporal, physical, biochemical and genetic complexities [5]. Traditional methods are limited to identifying a narrow spectrum of microbes, and they leave many microbes uncharacterized [6,7]. However, these microbes may be identified by metagenomic sequencing, which enables the retrieval of genomic sequences from a mixture of microbial DNA by next-generation sequencing in a culture-independent way [8–10]. Thus, this method is broadly known as the culture-independent method (Fig. 1**A**). Many studies have adopted metagenomic sequencing to explore the effects of microbes on human health. This has provided new research opportunities in biomedical science and revealed a great number of novel associations between the host microbiome and disease. For instance, a *meta*-analysis identified several associations between gut microorganisms, such as *Fusobacterium nucleatum*, *Parvimonas micra* and *Gemella morbillorum*, and colorectal cancer [11]. Thingholm et al. examined several bacterial genera, including *Akkermansia*, *Faecalibacterium*, *Oscillibacter* and *Alistipes*, those significantly depleted in obese individuals. These genera were found to be associated with producing short-chain fatty acids and were linked with alterations in serum metabolite concentrations [12].

Most previous studies on the microbiome have relied heavily on the availability of reference genomes. They have involved the direct alignment of sequencing data against reference genomes, marker gene sets or species-specific sequences for taxonomic assignments. The *k*-mer frequencies and read depth of universal single-copy genes are commonly used to estimate taxonomic abundance [13]. However, as reference genomes are incomplete, it is currently only possible to explore associations for microbes with high-quality reference genomes. This inevitably introduces a technical challenge to the identification of associations involving novel genes or microbes. For instance, approximately 40–50% of human intestinal microbes lack a high-quality reference genome [14,15]. Therefore, a well-characterized collection of reference genomes generated from metagenomic sequencing is required.

With recent advances in sequencing technologies and computational tools, this challenge has been attempted to be addressed by constructing metagenome-assembled genomes (MAGs). A MAG refers to a group of scaffolds with similar characteristics from a metagenome assembly that represent the microbial genome. In this approach, sequencing reads are assembled into scaffolds and then the scaffolds are grouped into candidate MAGs based on tetranucleotide frequencies (TNFs), abundances, complimentary marker genes [16], taxonomic alignments [17] and codon usage [18]. The MAGs with high completeness and low levels of contamination are then used for further taxonomic annotation and gene prediction (Fig. 1**B**). Many studies have generated valuable reference genomes for the human microbiome using MAGs. Using genome assemblies from large-scale metagenomic sequencing data, Pasolli et al. identified more than 150,000 MAGs, of which more than 50% described previously uncharacterized microbes [19]. This discovery has increased the average read mapping rates from 67.76% to 87.51% for the human gut microbiome. Almeida et al. [20] constructed the Unified Human Gastrointestinal Genome (UHGG) by integrating previously published MAGs and microbial reference genomes from publicly available databases. The UHGG identified 4,644 gut prokaryotes from 204,938 nonredundant MAGs. These newly identified genomes have been shown to have distinct functional properties and be associated with complex human diseases, which may improve the current predictive models [10,21].

Here, we review advances in computational tools and techniques used to construct MAGs. They are primarily categorized and described in two sections: upstream (Section 2) and downstream (Section 3) analyses. Upstream analyses are designed to construct MAGs, and they include sequence quality control (QC), metagenome assembly, assembly QC and metagenome binning.

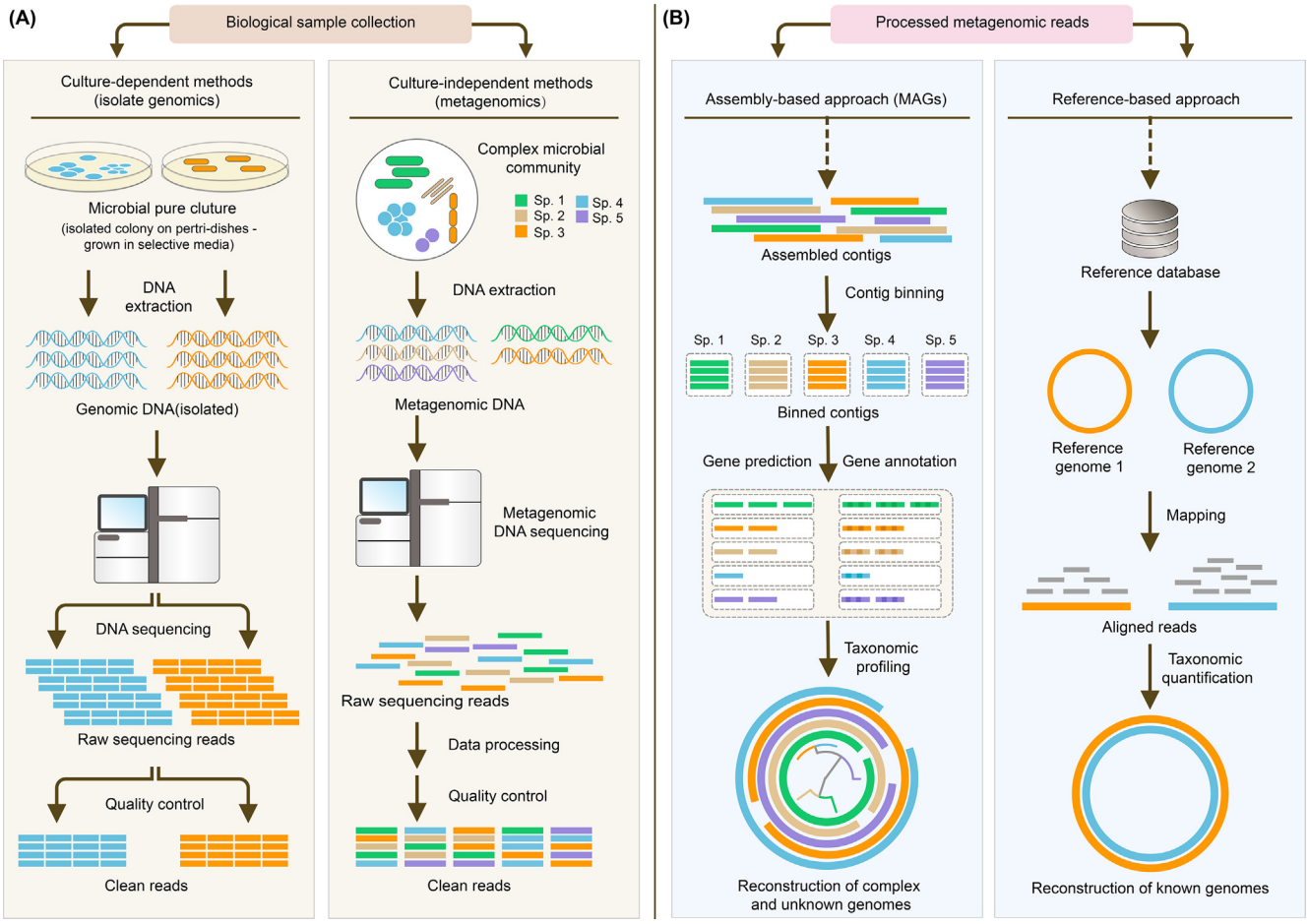

Fig. 1. Schematic representation of the different approaches used in the metagenomic research field. (A) A schematic contrast between culture-independent (metagenomics) methods, and culture-dependent methods. The process for generating sequencing data for the two strategies have been illustrated. (B) A schematic contrast between assembly-based and reference-based approaches on metagenomic sequencing data.

Downstream analyses are designed to annotate MAGs, and they include gene prediction, gene functional annotation, taxonomic classification and profiling. In Section 4, we introduce the computational requisition, comparative performance and selective guidance of these tools. We also discuss the potential challenges and some plausible strategies to address them in Section 5. This review thoroughly investigates the computational tools used to identify microbes using MAGs, based on metagenomic sequencing, and it provides reasonable solutions to overcome the current challenges associated with the technical limitations in this field.

2. Tools for upstream analyses to construct MAGs

2.1. Tools for sequence QC

The first inevitable step in constructing MAGs is to perform QC of the metagenomic sequencing data and to remove spurious and contaminating reads. For short-read sequencing, the QC step includes filtering low-quality nucleotides/reads, removing adapter sequences, processing enrichment bias and generating quality assessment metrics [22,23]. There are several popular tools available for sequence QC, such as FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), fastp [24], Trimmomatic [25] and SOAPnuke [26] (Table 1). FastQC provides a visual illustration of base quality, the distribution of GC content and nucleotide bias. fastp enhances the QC processing speed by using advanced multi-threading implementation. Trimmomatic and SOAPnuke are the two most frequently used tools that are optimized for Illumina and BGISEQ sequencing platforms, respectively. Eliminating contaminating reads from host genomes is another essential process, as the contaminants may get carried forward through the DNA extraction process and they may introduce bias in subsequent analyses. This step may be performed to remove reads that can be aligned to the host genome.

Long-read sequencing technologies, such as single-molecule real-time sequencing by Pacific Biosciences (PacBio), nanopore sequencing by Oxford Nanopore Technologies (ONT), synthetic long reads (SLR) and linked-read sequencing, rely on different QC principles, and the corresponding tools are also different. For example, SequelTools [27] can be used to examine the quality of PacBio long reads by filtering out the low-quality reads and producing multiple statistical plots. Another tool, proovread [28], has been developed to iteratively correct the base errors of long reads using high-quality short-reads. Some programs, such as NanoPack [29] and MinIONQC [30], are designed for ONT sequencing data. NanoPack focuses on processing and statistically evaluating raw ONT long reads, whereas MinIONQC enables the rapid comparison of multiple ONT flow cells. Recently, an integrated tool, LongQC [31], has been developed to process long reads from both PacBio and ONT platforms. It can automatically produce multiple statistical insights, including adapter statistics, quality statistics, GC content and a per-read base error estimation. SLR and Linked-read sequencing data allow the use of QC tools designed for short-read sequencing [32–34].

Table 1
Tools for sequence quality control. For each tool, the sequencing technologies (column 2), the original publications (column 3), Characteristics (column 4) and the websites to download these tools (column 5) are illustrated. The sequence quality control tools and related content are explained in Section 2.1.

Tools	Technologies	Publications	Characteristics	Websites
fastp	short reads, SLR and linked reads	Chen et al. 2018 [24]	ultra-fast; exhaustive functions	https://github.com/OpenGene/fastp
FastQC	short reads, SLR and linked reads		excellent visualization; exhaustive functions	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trimmomatic	short reads, SLR and linked reads	Bolger et al. 2014 [25]	flexible and exhaustive functions	http://www.usadellab.org/cms/index.php?page=trimmomatic
SOAPnuke	short reads, SLR and linked reads	Chen et al. 2018 [26]	reduced memory; predefined modules	https://github.com/BGI-flexlab/SOAPnuke
SequelTools	long reads	Hufnagel et al. 2020 [27]	user-friendly; exhaustive functions	https://github.com/ISUgenomics/SequelTools
proovread	long reads	Hackl et al. 2014 [28]	iterative consensus; computationally efficient	https://github.com/BioInf-Wuerzburg/proovread
NanoPack	long reads	Coster et al. 2018 [29]	exhaustive functions	https://github.com/nanopack
MinIONQC	long reads	Lanfear et al. 2019 [30]	suitable for large projects referring to multiple samples.	https://github.com/roblanf/minion_qc
LongQC	long reads	Fukasawa et al. 2020 [31]	platform-independent, computationally efficient and user-friendly	https://github.com/yfukasawa/LongQC

Table 2
Tools for metagenome assembly. For each assembler, the sequencing technologies (column 2), the original publications (column 3), and summaries of the core algorithms (column 4) and the websites to download these tools (column 5) are illustrated. The assemblers and related algorithms are explained in Section 2.2. DBG: De Bruijn graph; OLC: overlap-layout consensus.

Tools	Technologies	References	Core algorithms	Websites
Omega	short reads	Haider et al. 2014 [35]	OLC	http://omega.omicsbio.org
MetaVelvet	short reads	Namiki et al. 2012 [36]	DBG	http://metavelvet.dna.bio.keio.ac.jp
MetaVelvet-SL	short reads	Afiahayati et al. 2015 [38]	DBG	http://metavelvet.dna.bio.keio.ac.jp/MSL.html
MetaVelvet-DL	short reads	Liang et al. 2021 [39]	DBG	http://www.dna.bio.keio.ac.jp/metavelvet-dl/
IDBA-UD	short reads	Peng et al. 2012 [40]	DBG	http://www.cs.hku.hk/~alse/idba_ud
MEGAHIT	short reads	Li D et al. 2015 [41]	DBG	https://github.com/voutcn/megahit
metaSPAdes	short reads	Nurk et al. 2017 [43]	DBG	https://github.com/ablab/spades
Ray Meta	short reads	Boisvert et al. 2012 [45]	DBG	http://denovoassembler.sf.net.
Athena-meta	linked reads	Bishara et al. 2018 [46]	DBG	https://github.com/abishara/athena_meta
cloudSPAdes	linked reads	Tolstoganov et al. 2019 [47]	DBG	https://github.com/ablab/spades/releases/tag/cloudspades-paper
Nanoscope	SLR	Kuleshov et al. 2016 [34]	DBG	https://github.com/kuleshov/nanoscope
Canu	long reads	Koren et al. 2017 [51]	OLC	https://github.com/marbl/canu
NECAT	long reads	Chen et al. 2021 [52]	String Graph	https://github.com/xiaochuanle/NECAT
wtdbg2	long reads	Ruan et al. 2020 [53]	Fuzzy Bruijn Graph	https://github.com/ruanjue/wtdbg2
metaFlye	long reads	Kolmogorov et al. 2020 [54]	OLC	https://github.com/fenderglass/Flye
DBG2OLC	short and long reads	Ye et al. 2016 [56]	DBG and OLC	https://github.com/yechengxi/dbg2OLC
OPERA-MS	short and long reads	Bertrand et al. 2019 [57]	DBG	https://github.com/CSB5/OPERA-MS
Unicycler	short and long reads	Wick et al. 2017 [58]	DBG	https://github.com/rrwick/Unicycler

2.2. Tools for metagenome assembly

2.2.1. Metagenome assemblers for short-read sequencing

The conventional metagenome assemblers for short-read sequencing were designed using overlap-layout consensus (OLC) approaches (Table 2). For example, Omega [35] stores the prefix and suffix sequences of each read within the hash tables, and these sequences are then used to construct a bi-directed graph after linking the reads with their overlapping sequences. This graph is later simplified by removing transitive edges to explore the paths with minimum cost. Due to inherent issues with the OLC approach, it is difficult for Omega to process a large number of short reads, and it is also unable to distinguish chimeric contigs. There are several other assemblers designed using the De Bruijn graph (DBG), which splits the reads into *k*-mers and reduces the computer memory requirement (Table 2). MetaVelvet [36] constructs a DBG with Velvet [37] and partitions it into subgraphs using coverage peaks of the *k*-mers to divide different microbial genomes. The chimeric contigs and the contigs with repetitive sequences are then identified and split using paired-end information and local read depth divergence. MetaVelvet-SL [38] detects potential chimeric nodes in the DBG and deconvolves them using support vector machines. MetaVelvet-DL [39] constructs an end-to-end deep learning model with convolutional neural networks (CNNs) and Long Short-Term Memory. It has been shown to be more powerful than MetaVelvet-SL at deciphering chimeric contigs. A common problem when constructing DBGs is the selection of *k*-mer size, as it has a substantial impact on the ability to deal with repetitive sequences and uneven node depths [39]. To optimize the choice of *k*-mers, IDBA-UD [40] attempts to prune the graph iteratively and merge bubbles with increasing *k*-mer sizes. The *k*-mer size is determined if significantly different depths of the components of the graph are observed. MEGAHIT [41,42] couples the process of selecting *k*-mer sizes with a succinct DBG and shows strong computational efficiency. metaSPAdes [43] is a commonly used metagenomic assembler that improves the SPAdes tool [44] by introducing a novel heuristic strategy to differentiate interspecies repeats. It assumes an uneven depth in the assembly graph and builds multiple DBGs with different *k*-mer sizes. The hypothetical *k*-mers are designed to identify chimeric contigs. Another advanced tool, Ray Meta [45], generates the local depth distribution for each seed path in a DBG. It can be deployed explicitly on

Table 3
Tools for assembly quality control. List of tools for assembly quality control. For each tool, requires reference genomes or not (column 2), the original publications (column 3) and the websites to download these tools (column 4) are illustrated. The quality control tools and related descriptions are presented in Section 2.3.

Tools	Require reference genome	Publications	Websites
MetaQUAST	Yes	Mikheenko et al. 2016 [60]	http://bioinf.spbau.ru/metaquast
REAPR	No	Hunt et al. 2013 [62]	https://www.sanger.ac.uk/tool/reapr/
VALET	No	Olson et al. 2019 [63]	https://github.com/marbl/VALET
DeepMAsED	No	Mineeva et al. 2020 [64]	https://github.com/leylabmpi/DeepMAsED
CheckM	No	Parks et al. 2015 [81]	https://github.com/Ecogenomics/CheckM

the distributed computing system and enables metagenome assembly on computer clusters without large memory requirements.

2.2.2. Metagenome assemblers for long-read sequencing

Due to the lack of long-range genome connectedness, the assemblers designed for short-read sequencing are generally limited to processing intra- and inter-species repeats. Assemblers designed for both virtual (SLR and linked reads) and physical (ONT and PacBio) long reads have shown great promise in generating assemblies with excellent continuity. For virtual long reads, Bishara et al. [46] developed Athena to improve metagenomic assembly by considering the co-barcoded linked reads between contigs. It constructs a scaffold graph by linking the contigs from metaSPAdes based on the support of paired-end reads. The local assembly is performed by recruiting the co-barcoded linked reads shared by the contig pairs connected in the scaffold graph. cloudSPAdes [47] builds upon the assembly graph of metaSPAdes and evaluates the similarities between the barcode sets of two edges, which measures the probability that the sequences are derived from the same genomic region. The edges with high similarity are then connected to simplify the graph. Nanoscope [34] integrates SOAPdenovo [48] and Celera [49] to assemble short reads and SLRs independently and merge their contigs using Minimus2 [50]. As the physical long reads generated by PacBio and ONT platforms have much higher base error rates than short reads, the developers of long-read assemblers have implemented dedicated modules for base error correction. Some tools use pre-assembly error correction. Canu [51] and NECAT [52] correct the sequencing errors in long reads before genome assembly using the OLC approach. Rather than correcting error-prone long reads, wtdbg2 [53] enables the process of inexact sequence matches to build a consensus from the intermediate contigs. metaFlye [54] has been built on Flye [55] and has dedicated features to process long reads from a mixture of microbial genomes. It combines the long reads into error-prone disjointigs and collapses the repetitive sequences into a repeat graph.

2.2.3. Metagenome assemblers for hybrid assembly

Short-read and long-read sequencing techniques are somewhat complementary to each other, as short reads have high base quality and long reads provide long-range genome connectedness. Several algorithms have attempted to make better use of the superiorities of these two techniques. metaSPAdes [43] considers error-prone long reads as "untrusted contigs" and applies them to thread the complex structure of the assembly graph from short reads. DBG2OLC [56] aligns the contigs from short-read assemblies to error-prone long reads and applies the OLC strategy to concatenate the long reads into contigs. OPERA-MS [57] uses long reads with shallow coverage to link the contigs from short reads into a scaffold graph and then groups them into species-specific clusters. It uses a novel Bayesian clustering algorithm to produce strain-resolved assemblies using contig read depth and connected information from long reads. Unicycler [58] was initially designed to assemble a single bacterial genome, but it was later applied to metagenome assembly [59]. It only uses long reads to choose the paths for ambiguous contigs in the assembly graph and uses multiple sequence alignment (MSA) to correct the sequencing errors of long reads.

2.3. Tools for assembly QC

As shown in Table 3, there are many tools available to evaluate the accuracy and continuity of the contigs and scaffolds generated by metagenome assemblers. MetaQUAST [60] rapidly calculates the basic statistics for the contigs and scaffolds, such as assembly length, N50 values and contig length distribution. It also supports a reference-based assessment in the "meta" mode that aligns the sequences against the given reference genomes or the SILVA 16S rRNA database [61] to calculate reference-based statistics, such as genome coverage, NA50 and NGA50 values. REAPR [62] precisely identifies errors in genome assemblies without relying on a reference sequence. It offers a quantitative comparison across multiple assemblies using the inherent information within the sequencing reads. VALET [63] performs metagenome binning before QC to reduce the number of false positives and false negatives due to uneven read depth. Recently, DeepMAsED [64] has been developed to detect misassembled contigs, without the need for reference genomes, using a deep learning model.

2.4. Tools for metagenome binning

Most of the current assemblers do not represent complete microbial genomes with single scaffolds. Many metagenome binning tools have been developed to group the scaffolds into clusters to represent the whole genome of an organism (Table 4). The existing tools for short reads rely mainly on TNFs, *k*-mer frequencies and read depth. GroopM [65] uses Hough partitioning and a two-way clustering method to group scaffolds with similar read depths and TNFs. MaxBin2 [66] jointly considers TNFs and read depths to estimate the distances between scaffolds and then uses an expectation–maximization algorithm to group the scaffolds co-assembled from multiple metagenomic samples. CONCOCT [67] combines TNFs and read depths to cluster the scaffolds using Gaussian mixture models. MetaBAT2 [68] also uses TNFs and read depths to compute scaffold similarities and constructs a graph by representing their similarities as the weights of the edges. This graph is further partitioned into subgraphs or bins based on a modified label propagation algorithm (LPA). Besides TNFs and read depths, some additional information may also be taken into consideration to link scaffolds. MyCC [16] implements an affinity propagation algorithm to use complementary marker genes between scaffolds. SolidBin [17] performs spectral clustering on taxonomic alignments to connect scaffolds, and BMC3C [18] uses ensemble clustering on codon usage inferred from the scaffolds. Mallawaarachchi et al. found that short scaffolds were commonly neglected by previous metagenome binning tools, and therefore, they developed GraphBin [69], which rescues the short scaffolds using an LPA on the assembly graph. METAMVGL [70] integrates the assembly and paired-end graphs and applies a multi-view

Table 4
Tools for metagenome binning. For each tool, the adopted technologies (column 2), the original publications (column 3), summaries of the core algorithms (column 4), and the websites to download these tools (column 5) are illustrated. The metagenome binning tools and related descriptions are presented in Section 2.4.

Tools	Technologies	Publications	Core algorithms	Websites
GroopM	short reads	Imelfort et al. PeerJ.2014 [65]	Two-way clustering and Hough partitioning	http://ecogenomics.github.io/GroopM/
MaxBin2	short reads	Wu et al. 2016 [66]	Expectation-maximization	http://sourceforge.net/projects/maxbin/
CONCOCT	short reads	Alneberg et al. 2014 [67]	Gaussian Mixture Models	https://github.com/BinPro/CONCOCT
MetaBAT2	short reads	Kang et al. 2019 [68]	Label propagation	https://bitbucket.org/berkeleylab/metabat
MyCC	short reads	Lin et al. 2016 [16]	Affinity propagation	http://sourceforge.net/projects/sb2nhri/files/MyCC/
SolidBin	short reads	Wang et al. 2019 [17]	Spectral clustering	https://github.com/sufforest/SolidBin
BMC3C	short reads	Yu G et al. 2018 [18]	Ensemble clustering	http://mlda.swu.edu.cn/codes.php?name=BMC3C
GraphBin	short reads	Mallawaarachchi et al. 2020 [69]	Label propagation	https://github.com/Vini2/GraphBin
METAMVGL	short reads	Zhang et al. 2021 [70]	Label propagation	https://github.com/ZhangZhenmiao/METAMVGL
VAMB	short reads	Nissen et al. 2021 [71]	Variational Autoencoders	https://github.com/RasmussenLab/vamb
MAGO	short reads	Murovec et al.2020 [73]	Ensemble learning	http://mago.fe.uni-lj.si/
MetaWRAP	short reads	Uritskiy et al. 2018 [74]	Ensemble learning	https://github.com/bxlab/metaWRAP
DAS Tool	short reads	Sieber et al. 2018 [75]	Ensemble learning	https://github.com/cmks/DAS_Tool
ProxiMeta	Hi-C	Press et al. 2017 [76]	Graph-based clustering	https://github.com/phasegenomics/proxiphage_paper
bin3C	Hi-C	DeMaere et al. 2019 [77]	Network clustering	https://github.com/cerebis/bin3C
HiCBin	Hi-C	Du et al. 2021 [79]	Leiden algorithm	https://github.com/dyxstat/HiCBin

Table 5
Tools for gene prediction. For each tool, the method types (column 2), the original publications (column 3), summaries of the core algorithms (column 4) and the websites to download these tools (column 5) are illustrated. The gene prediction tools and related descriptions are presented in Section 3.1.

Tools	Types	Publications	Core algorithms	Websites
MetaGeneMark	model based	Zhu et al. 2010 [84]	Hidden Markov Model	http://exon.gatech.edu/meta_gmhmmp.cgi
Glimmer-MG	model based	Kelley et al. 2012 [85]	Interpolated Markov Model	https://github.com/davek44/Glimmer-MG
FragGeneScan	model based	Delcher et al. 2007 [86]	Hidden Markov Model	https://omics.informatics.indiana.edu/FragGeneScan/
Prodigal	model based	Hyatt et al. 2010 [87]	Dynamic Programming	https://github.com/hyattpd/Prodigal
MetaGene	model based	Noguchi et al. 2006 [88]	Dynamic Programming	http://metagene.nig.ac.jp/metagene/metagene.html
MetaGeneAnnotator	model based	Noguchi et al. 2008 [89]	Dynamic Programming	http://metagene.nig.ac.jp/
Meta-MFDL	machine learning	Biomed Res et al. 2017 [90]	Deep Neural Network	https://github.com/nwpu903/Meta-MFDL
CNN-MGP	machine learning	Al-Ajlan et al. 2019 [91]	Convolutional Neural Network	https://github.com/rachidelfermi/cnn-mgp
Balrog	machine learning	Sommer et al. 2021 [92]	Convolutional Neural Network	https://github.com/salzberg-lab/Balrog

LPA to connect dead ends to the main assembly graph. VAMB [71] embeds the scaffolds in low dimensions using a variational autoencoder [72]. Despite many tools that have been developed for metagenome binning, there is no single best choice, and the ensemble-based tools, such as the binning module in Metagenome Assembled Genomes Orchestra (MAGO) [73], MetaWRAP [74] and DAS Tool [75], provide a promising way to integrate the binning results from different tools.

Hi-C is another sequencing technology that has been used for metagenome binning by introducing genome-wide spatial proximity. ProxiMeta [76] can devolve plasmid genomes and generate high-quality bins without relying on prior information. bin3C [77] has an effective pipeline for contact map generation, bias removal and interaction strength normalization, and it uses the Louvain algorithm [78] for scaffold community detection. HiCBin [79] uses HiCzin [80] to normalize the interaction map and applies the Leiden community detection algorithm to group scaffolds. It also includes a module to detect spurious contacts.

2.5. Quality evaluation of MAGs

The designation of metagenome bins as MAGs relies on several parameters, such as the completeness of marker genes and the contamination of single-copy genes. CheckM [81] is commonly used to determine the quality of each bin. Only the bins with relatively high quality are then selected as the MAGs for subsequent annotation. The bins are commonly classified as finished, high-quality, medium-quality or low-quality drafts based on their completeness, level of contamination and rRNA/tRNA prediction [82]. Due to known issues in assembling rRNA/tRNA sequences [61,83], it is well accepted to select high-quality (completeness greater than 90% and contamination < 5%) and medium-quality (completeness greater than 50%, contamination < 10% and completeness − [5 × contamination] greater than 50) bins as MAGs [21].

3. Tools for downstream analyses to annotate MAGs

3.1. Gene prediction tools

Gene identification and annotation are the next critical steps after carefully selecting MAGs from the metagenome assembly. In this section, we discuss the approaches used to identify and predict genes by recognizing potential coding sequences within MAGs (Table 5).

3.1.1. Model-based gene prediction tools

Hidden Markov Model (HMM)-based tools are the most commonly used tools for model-based gene prediction. There are several tools available under this category, such as MetaGeneMark [84], Glimmer-MG [85] and FragGeneScan [86]. MetaGeneMark extracts oligonucleotide frequencies and their compositions from the genomes of known prokaryotic species to train the HMM. Glimmer-MG clusters the input sequences that are most likely to share the same origin and trains the interpolated Markov model within each cluster to optimize the probabilistic inference. FragGeneScan incorporates the sequencing error models into six-

Table 6
Tools for gene annotation. For each tool, the method types (column 2), the original publications (column 3), summaries of the core algorithms/programs (column4) and the websites to download these tools (column 5) are illustrated. The gene annotation tools and related descriptions are presented in Section 3.2.

Tools	Types	Publications	Core algorithms	Websites
eggNOG-mapper	Homology-based	Huerta-Cepas et al. 2017 [94]	Hidden Markov Model	http://eggnog-mapper.embl.de
GhostKOALA	Homology-based	Kanehisa et al. 2016 [95]	GHOSTX (seed search method)	http://www.kegg.jp/blastkoala/
MG-RAST	Homology-based	Keegan et al. 2016 [96]	Parallelized BLAT	http://api.metagenomics.anl.gov/api.html
PANNZER2	Homology-based	Törönen et al. 2018 [97]	Sansparallel (suffix array neighborhood search)	http://ekhidna2.biocenter.helsinki.fi/sanspanz/
InterProScan	Motif-based	Quevillon et al. 2005 [107]	Phobius (Hidden Markov Model)	http://www.ebi.ac.uk/InterProScan/
GeConT	Gene context based	Ciria et al. 2004 [110]	Blastp	http://www.ibt.unam.mx/biocomputo/gecont.html
FunGeCo	Gene context based	Anand et al. 2020 [112]	Hidden Markov Model	https://web.rniapps.net/fungeco
FlaGs	Gene context based	Saha et al. 2021 [114]	Jackhmmer (Hidden Markov Model)	https://github.com/GCA-VH-lab/FlaGs

periodic inhomogeneous Markov models, enabling the identification of genes with frameshifts.

There are also several gene prediction tools for bacterial and archaeal genomes using dynamic programming. For example, Prodigal [87] uses dynamic programming based on frame bias scores to train gene prediction models in the preliminary phase. In the final phase, the same algorithm is used in Prodigal to process hexamer coding scores for each gene to predict their potential protein-encoding abilities. MetaGene [88] calculates two types of scores for all possible open reading frames (ORFs) to measure their intrinsic (including base composition and length) and extrinsic (including orientations and distance of neighboring genes) characteristics. These scores are further combined and serve as inputs for the dynamic programming to further estimate the optimal paths of ORFs. MetaGeneAnnotator [89] improves the prediction of lateral gene transfers and translation start sites of genes by adopting the prophage gene and ribosome binding site models, respectively.

3.1.2. Deep learning-based gene prediction tools

Recently, various deep learning tools have gained considerable attention for gene prediction. Meta-MFDL [90] is a commonly used tool that constructs a representative vector by fusing multiple features, such as monocodon usage, mono-amino acid usage, ORF length coverage and Z-curve features and then trains a deep stacking network to distinguish between coding and noncoding ORFs. CNN-MGP [91] uses a CNN model to automatically learn the characteristics of ORFs in training datasets and predict the probability of ORFs in MAGs. Balrog et al. [92] used a temporal CNN for gene prediction based on a large and diverse set of microbial genomes.

3.2. Gene functional annotation tools

Metagenomic sequencing enables the evaluation of the functional characteristics of microbial communities. Gene functional annotation tools can be classified into two categories: 1) tools with broad scopes to evaluate full functional potential and 2) tools with narrow scopes focusing on one or a few specific biological processes. This review focuses only on the tools designed to provide a full functional overview (Table 6).

3.2.1. Homology-based tools

Homology-based tools usually rely on different variants of BLAST [93] to compare the sequences of predicted genes with those of known genes. They are often very slow to process the large number of genes predicted from MAGs. However, modern methods, such as eggNOG-mapper [94], GhostKOALA [95], MG-RAST [96] and PANNZER2 [97], employ optimized alignment strategies that enable 100- to 1,000-fold faster alignment of gene sequences to databases. eggNOG-mapper performs ultra-fast alignments against orthologies based on pre-computed sequence clusters and phylogenic information from eggNOG [98]. It uses HMM to search the most closely matching reference sequences to each query protein. eggNOG-mapper is much faster than traditional BLAST-based approaches. GhostKOALA [95] is an automatic annotation pipeline relying on GHOSTX [99] and also assigns Kyoto Encyclopedia of Genes and Genomes (KEGG) orthologies and pathways to each gene [100]. MG-RAST [96] provides an online metagenomic analysis interface that includes data uploading, QC and alignment with M5nr reference databases [101]. PANNZER2 [97] incorporates SANSparallel [102], a message passing interface implementation of a suffix array neighborhood searching approach, to enable rapid homology searches of Gene Ontology [103] annotations.

3.2.2. Motif-based tools

Protein sequences are sometimes partially assembled from metagenomic sequencing data, and this may adversely influence homology-based annotation mainly due to incomplete and misassembled scaffolds in MAGs. In such cases, despite poor alignment homologies, the predicted proteins tend to perform analogous functions to those sharing common sequences, patterns or specific motifs. Databases such as InterPro [104], PROSITE [105] and PRINTS [106] have collected such patterns or motifs based on statistical inference. Given an input sequence, InterProScan [107] performs a systematic search of the InterPro database using Phobius [108] and predicts protein domains, active sites and potential functional annotations. However, as MAGs inherently contain novel sequences, it is always recommended to perform both motif-based analysis and other homology-based approaches for better functional annotations.

3.2.3. Gene context-based tools

Metagenomic sequencing enables the recognition of a large number of novel genes that may share no homology with known genes and thus are not suitable to be annotated using the aforementioned approaches. To address such limitations, gene context-based tools have been introduced. Harrington et al. combined homology-based approaches and tailored gene neighborhood methods to perform gene annotation for complex metagenomic datasets [109]. Their tool infers the specific functions of 76% and non-specific functions of 83% of the predicted genes and outperforms standard BLAST-based methods. Ciria et al. developed GeConT [110] to visualize the genomic context of target genes and extract their orthologs from the COG database [111] for gene function prediction. FunGeCo [112] uses HMM to align the predicted genes to the Pfam database [113] and record their genomic locations. Based on this information, FunGeCo infers the significant enrichment of domains in the gene context and visualizes them on a web server. FlaGs [114] extracts upstream and downstream genes for a given gene of interest and annotates and further clusters the flanking genes using a sensitive HMM-based method. A

Table 7
Tools for MAG taxonomic classification. For each tool, the method types (column 2), the original publications (column 3), summaries of the core algorithms (column 4) and the websites to download these tools (column 5) are illustrated. The detailed description is presented in Section 3.3.

Tools	Types	Publications	Core algorithms	Websites
GTDB-Tk	concatenated protein	Chaumeil et al. 2019 [115]	Likelihood-based phylogenetic inference	https://github.com/Ecogenomics/GTDBTk
ezTree	concatenated protein	Wu et al. 2018 [119]	Maximum likelihood	https://github.com/yuwwu/ezTree
PhyloPhlAn3	concatenated protein	Asnicar et al. 2020 [121]	Maximum likelihood	https://huttenhower.sph.harvard.edu/phylophlan/
MiGA	genome-based relatedness	Rodriguez-R et al. 2018 [122]	Markov clustering	http://microbial-genomes.org/

Table 8
Tools for profiling MAG abundance. For each tool, the method types (column 2), the original publications (column 3), summaries of the core algorithms (column 4) and the websites to download these tools (column 5) are illustrated. The gene prediction tools and related descriptions are presented in Section 3.4.

Tools	Types	Publications	Core algorithms	Websites
Kaiju	translated protein based	Menzel P et al. 2016 [123]	Backwards search	http://kaiju.binf.ku.dk
Kraken	*k*-mer based	Wood DE et al. 2014 [126]	Classification tree	https://ccb.jhu.edu/software/kraken/
Kraken2	*k*-mer based	Wood DE et al. 2019 [127]	Spaced seed	https://ccb.jhu.edu/software/kraken2/
Bracken	*k*-mer based	Jennifer Lu et al. 2017 [128]	Bayesian probability algorithm	https://ccb.jhu.edu/software/bracken/
CLARK	*k*-mer based	Ounit R et al. 2015 [129]	Spectral decomposition	http://clark.cs.ucr.edu/
k-SLAM	*k*-mer based	Ainsworth D et al. 2017 [130]	Pseudo-assembly	https://github.com/aindj/k-SLAM
MetaPhlAn3	marker gene based	Beghini F et al. 2021. [131]	Comprehensive pipeline	https://huttenhower.sph.harvard.edu/metaphlan/
PanPhlAn3	marker gene based	Beghini F et al. 2021. [131]	Comprehensive pipeline	http://segatalab.cibio.unitn.it/tools/panphlan/
IGGsearch	marker gene-based	Nayfach S et al. 2019 [10]	Comprehensive pipeline	https://github.com/snayfach/IGGsearch
ConStrain	SNP based	Luo C et al. 2015 [132]	SNP-flow algorithm	https://bitbucket.org/luo-chengwei/constrains
StrainFinder	SNP based	Smillie CS et al. 2018 [133]	Expectation-maximization	https://github.com/cssmillie/StrainFinder
StrainEst	SNP based	Albanese D et al. 2017 [134]	Penalized optimization	https://github.com/compmetagen/strainest
StrainPhlAn3	SNP based	Beghini F et al. 2021. [131]	Comprehensive pipeline	http://segatalab.cibio.unitn.it/tools/strainphlan/

phylogenetic tree is then constructed for those flanking genes to examine their conservation.

3.3. Tools for MAG taxonomic classification

Another critical task when annotating MAGs is to determine their taxonomic classifications. Traditional methods based on 16S rRNA small subunit genes have been successfully established to understand the diversity of MAGs in prokaryotic communities, but these methods offer limited resolution and 16S rRNAs are poorly represented in MAGs [20,61]. In contrast, methods using single-copy marker genes have gained popularity due to their improved resolution (Table 7). GTDB-Tk [115] uses HMMER [116] to identify marker genes in the reference genomes from the Genome Taxonomy Database (GTDB) [117]. These marker genes are concatenated for MSA and then used to construct a reference tree using a likelihood-based phylogenetic inference algorithm [118]. GTDB-Tk performs taxonomic classification for a queried MAG based on its position in the reference tree, relative evolutionary divergence and average nucleotide identity to the reference genome. ezTree [119] is designed to automatically search for single-copy marker genes in a pre-defined database and build a phylogenetic tree based on maximum-likelihood [120]. To improve the classification of closely related MAGs, Asnicar et al. developed PhyloPhlAn 3.0 [121], which has been used to extract species-specific marker genes from the integration of more than 150,000 MAGs and 80,000 reference genomes. PhyloPhlAn 3.0 uses a novel strategy to place the queried MAGs in the phylogenetic tree. The steps include identifying the marker genes from the MAGs, performing MSA for the marker genes, concatenating the MSAs into a unique MSA and reconstructing the phylogeny using a maximum-likelihood approach. In addition, the Microbial Genomes Atlas [122] uses a unique method to evaluate the similarities between the queried MAGs and sequences from a reference database, which are calculated based on genome-aggregated average nucleotide and amino acid identities. The reference genome with the highest matching score is then selected using the Markov cluster algorithm.

3.4. Tools for profiling MAG abundance

There are several available tools to estimate the abundance of MAGs in metagenomic sequencing data. These tools have been divided into four categories: 1) protein-based tools, 2) *k-mer*-based tools, 3) marker gene-based tools and 4) single nucleotide polymorphism (SNP)-based tools. All of these four methods estimate MAG abundance, but they tend to perform distinctly and have different resolutions. For example, *k-mer*-based tools calculate the abundance of specific sequences of MAGs, whereas marker gene-based tools report their taxonomic abundance. Here, we discuss these tools and their potential roles in MAG profiling (Table 8).

3.4.1. Protein-based tools

Kaiju is one of the most commonly used protein-based MAG abundance profilers [123]. It is a protein-level classification tool developed for the taxonomic classification of a large number of reads from metagenomic or metatranscriptomic sequencing data. It first compacts the protein sequences predicted from MAGs by Burrows-Wheeler transformation [124] and indexes each sequence by FM-index [125] to reduce the computational time and the memory requirement. Kaiju then translates the query nucleotide sequence into an amino acid sequence, which it aligns against an established database of proteins derived from MAGs and sorts the resulting alignments. Once Kaiju detects sequence homology in the protein database, it outputs the taxonomic identifier of the best match. Sometimes, it also determines the lowest common ancestor (LCA) after recognizing substantial matches among the different taxa. As Kaiju uses protein-level classifications, it has greater sensitivity than methods relying on nucleotide sequences. It can extend its search capacity to fungi and some microbial

eukaryotes depending on the context. Within Kaiju, the reads are translated into amino acid sequences and then searched against a database to recognize the maximum number of exact matches.

3.4.2. k-mer-based tools

Kraken, the most commonly used *k*-mer based tool, replaces sequence alignment with efficient searching of a simple *k*-mer lookup table [126]. In Kraken, the *k*-mers from the sequence database are saved in a compressed lookup table that can be promptly queried for exact matches to *k*-mers found in the reads. For each query read, a tree is constructed using the *k*-mers from the read and the ancestors of the associated taxa, and the tree is then used to determine the final classification with the maximal root-to-leaf path. Kraken2 [127] maintains the accuracy of Kraken but reduces the memory and computational requirements, enabling the inclusion of more reference genomes in the database. Bracken [128], an extension of Kraken, uses a Bayesian probability model to estimate the abundance of MAGs. Another tool, CLARK [129], uses discriminative *k*-mers to perform a supervised sequence classification and then reports the assignments with confidence scores. k-SLAM [130] uses local sequence alignments and pseudo-assembly strategies to generate contigs, leading to more specific assignments of taxonomic classifications. Taxonomic inference is then performed using the LCA technique.

3.4.3. Marker gene-based tools

MetaPhlAn3 [131] is an extensively used marker gene-based tool for MAG profiling. It aligns reads to a marker gene database and normalizes the counts for each gene for a given sample. From such genome-scale information, the abundance of each MAG is estimated. In contrast, PanPhlAn3 [131] maps the reads against the pangenome of a species and offers gene presence or absence information. IGGsearch [10] uses a similar approach to MetaPhlAn3, but it is the first tool that has attempted to extract pools of marker genes from MAGs from 3,810 fecal metagenomes [10]. It stretches the boundary of the metagenomics field to further explore the phylogenetic diversity of different bacteria and other prokaryotes. However, IGGsearch is currently limited to being exclusively used for the analysis of the human gut microbiota.

3.4.4. SNP-based tools

Compared to the other tools, SNP-based tools consider different strains of MAGs and calculate their abundance separately. ConStrain [132] is an effective SNP-based tool at identifying strain-level genotypes. It mainly uses a SNP-flow algorithm to identify all possible SNP types and select the best groups of SNP types to define the mixture of strains. The relative abundance of each strain is further measured using the Metropolis-Hastings Markov Chain Monte-Carlo approach. StrainFinder [133] assumes a multinomial distribution of observed SNPs at a given position to identify multiple strains within a species. It then uses an expectation–maximization algorithm to maximize the likelihood of measuring the frequencies of strains and their genotypes. In contrast, StrainEst [134] uses a penalized optimization procedure to detect all strains within a species of interest. Based on the definition of the dominant strain per species, StrainPhlAn3 [135] uses the SNPs in the marker genes from MetaPhlAn3 to construct a species-specific consensus sequence to infer the strain-level genotypes and abundance.

4. Performance comparison and computational requisites

Constructing MAGs from large metagenomic sequencing datasets is a complex procedure that requires tremendously expensive computational capabilities and manpower. It also demands the careful selection of different tools and a high level of technical expertise (**Supplementary** Table 1). This is challenging and tedious for many biologists and bioinformaticians. Some platforms, such as bioBakery3, MAGO and SqueezeMeta, have integrated an array of commonly used state-of-the-art tools to offer convenient, unified and reproducible methods to construct and annotate MAGs [73,131,135]. Moreover, several container technologies, such as Docker and Singularity, enable the encapsulation of the entire computing environment, including software dependencies, libraries, packages, code and reference data, together to offer a fast and secure option for scalable and reproducible scientific computation [136]. Our comprehensive investigation may prompt further development in this field.

The computational requirements to construct MAGs are very high. For example, a moose rumen microbiome dataset with 280 GB of raw reads from six samples was analyzed using a 120 central processing unit (CPU) and 1.2 TB of random-access memory (RAM). Completing the task using MAGO costs 16,128 CPU hours [73]. In another study, a human pregnancy microbiome dataset with 120 GB of raw reads from 101 samples was analyzed using a 20–28 CPU and 300–480 GB of RAM. Completing this task using MAGO cost 2,880 CPU hours [73]. The computational requirements to construct MAGs also varies among different stages of the process. In the pre-processing stage, fastp [24] is much faster than the other available tools, such as Trimmomatic and FastQC. It can process approximately 100,000 paired reads per second. Genome assembly is the most computationally demanding step. For metagenome assembly from short-reads, MEGAHIT and metaSPAdes have demonstrated excellent performance, with MEGAHIT requiring less time and memory. MEGAHIT uses approximately 5 GB of memory to assemble $\sim$ 4.2 GB of raw reads from Illumina short-read sequencing in 7 h, while metaSPAdes uses 43 GB of memory and 14 h to assemble the same dataset [137]. Of the long-read assemblers, Canu requires much more computational resources to achieve a comparable performance to metaFlye when processing data from the ONT platform [138]. For linked-read sequencing, Athena-meta and cloudSPAdes have shown similar performance, although cloudSPAdes requires less computing time [47]. For the metagenome binning step, the existing tools exhibit variable performance across different datasets. To attain a consensus interpretation, a highly complex and realistic benchmark dataset, the Critical Assessment of Metagenome Interpretation I (CAMI I) dataset, was introduced in 2017 [139]. MetaBAT2 has been reported to demonstrate better performance than other binners, such as BMC3C, CONCOCT and MyCC, when using the CAMI I dataset [140]. However, despite having the quickest running time and lowest memory requirement, the performance of MetaBAT2 is worse than that of MaxBin2 and CONCOCT when using the CAMI II marine datasets, as indicated by a lower F1 score [141]. Notably, some ensemble strategy-based tools, such as MetaWRAP and DAS Tool, show excellent overall performance and generate high-quality scaffold clusters.

In the downstream processes, especially the gene prediction and annotation steps, the computational requirements largely depend on the number of sequences to be processed. Of the different gene prediction tools, Prodigal and MetaGeneMark show high processing speeds [84,92]. Most gene prediction tools provide high-quality prediction of genes. A study by Nicholas et al. offers a comparative insight into the difference between the efficiency of joint prediction by integrating Prodigal, MetaGeneAnnotator and MetaGeneMark and that of using the best tool for each specific organism [142]. Compared with the best tool, the joint prediction model was shown to offer a negligible increase ($\sim$0.47%) in the number of genes predicted. Hence, a single gene prediction tool may be sufficient for organism-specific analyses. Gene function annotation is 2.5 $\times$ faster with eggNOG-Mapper than with Inter-

proscan [94]. The taxonomic profilers take less processing time than the assembly and binning tools. Kraken and Bracken are *k-mer*-based tools that rapidly and accurately estimate taxonomic abundance [143]. However, they require a large amount of memory to load the database. The marker gene-based tools, such as MetaPhlAn3, have significantly reduced memory usage compared with the *k-mer*-based tools, while offering similar accuracy.

5. Outlook, potential challenges and strategies to address them

Genome assembly is an essential step in producing high-quality MAGs. Several state-of-the-art assemblers have been developed (as discussed in Section 2.2) to analyze data from different sequencing technologies. Several studies [10,19] have reported that the quality of the constructed MAGs is high and is similar to the quality of genomes assembled from isolates. However, a recent study [144] showed that some of the MAGs were not as expected. A thorough investigation of a MAG named "HMP_2012_SRS023938_bin.39" [19], which was annotated as the phylum, *Saccharibacteria*, showed that 53.5 kbp of sequence from 11 contigs were contaminated, and the genes represented by these sequences had the best match with *Selenomonas* genes rather than *Saccharibacteria* genes. As a result of these findings, the authors proposed a practical workflow to facilitate the curation of MAGs. Some tools have also been developed to remove contaminating scaffolds from chimeric MAGs. MAGpurify [10] removes scaffolds that are far from those from the same MAG by considering information from multiple sources, such as phylogenetic marker genes, clade-specific markers and GC content. GUNC [145] calculates the clade separation and reference representation scores to quantify genomic chimerism. The construction of MAGs should ideally be performed for all of the high-, medium- and low-abundance microbes. However, the existing tools mostly fail to distinguish between reads from low-abundance microbes and contamination from library preparation and sequencing. Although these tools can generate the contigs and scaffolds for low-abundance species, the contiguity is relatively poor and the sequences only represent partial genomes. Advanced long-read sequencing technologies may be promising to produce complete genomes and detect low-abundance microbes [146]. A previous study showed that selective nanopore sequencing technology enriches specific DNA molecules and enables researchers to focus on DNA fragments of interest [147]. Recently, Kovaka et al. developed UNCALLED to be used while running metagenomic sequencing. This tool enables the depletion of high-abundance species and the enrichment of the remaining low-abundance species [148]. These technologies facilitate the assembly of low-abundance species from complex microbial communities. However, concerns regarding the assembly of long-read sequencing data remain, as they tend to contain a higher rate of sequencing errors and the circularization process to represent the complete microbial genome is often a challenge. This demands the development of efficient algorithms, especially for high base quality assembly. In addition, metatranscriptomic data may provide a new paradigm to detect live microorganisms within a complex environment [149]. Although several metatranscriptome assembly tools, such as IDBA-MTP [150] and Transcript Assembly Graph [151] have been developed, the number of such tools remains limited.

Metagenome binning is another critical step in generating high-quality MAGs. Most of the available binning tools require the scaffolds to be sufficiently long (at least 1 kb) to estimate the stable TNF and read depth. In our previous study [70], we found that most scaffolds were below the required threshold and thus cannot be clustered using most of the existing tools. Graph-based tools have recently been designed to solve this issue by considering the connectedness of scaffolds based on sequence overlap and paired-end constraints [70,152]. These tools have improved the binning performance, but more investigations are required in this area, especially when introducing assembly graphs from error-prone long reads. Some unbinned scaffolds may be derived from interspecies repeats or horizontal gene transfer. They should be carefully assigned into bins, but most existing tools simply assign them to the most likely bins or keep them unbinned. Recently, GraphBin2 [153] was proposed to allow scaffolds belonging to multiple bins by solving subset sum problems with their abundance [153]. The feature space for scaffold binning is usually high due to a large number of tetranucleotides, and most clustering algorithms fail to achieve good performance with high-dimensional data. A recent study proposed a deep learning model, VAMB [71], to use variational autoencoders to embed the scaffolds in low dimensions, thus significantly reducing the hidden noise and increasing the accuracy of the clustering. In the near future, more sophisticated deep learning models are expected to emerge in this field.

One of the major advantages of MAGs is that they allow the discovery of metagenomic dark matter in complex ecosystems. For instance, Sberro et al. identified thousands of small novel genes from metagenome assemblies from the Human Microbiome Project [154]. These genes were missed by some of the gene finder tools as they usually contain a default minimum ORF length. However, mounting evidence suggests that short genes are widespread and play significant biological roles [155]. Emerging deep learning-based gene predictors without manual feature selection promise to improve prediction efficiency and quality for genes with irregular features [92,142]. According to Almeida et al. [20], approximately 40% of the proteins predicted in MAGs do not have similar sequences in the current databases, such as eggNOG, InterPro, COG and KEGG. Thus, there is a strong demand for new approaches besides gene context-based methods to characterize functional capacities. In the future, the well-annotated protein database will substantially enhance the development of pathway analysis tools, such as HUManN3 [131], to elucidate the impact and underlying mechanisms of human health and diseases associated with microbiota. On the other hand, substantially unexplored genomes have been discovered while constructing MAGs from large-scale metagenomic sequencing data [19,21]. Concatenated protein-based approaches, such as the GTDB [117], are extensively used to characterize those previously unexplored genomes. Compared with conventional 16S rRNA-based approaches, the GTDB provides better consistency and higher resolution of phylogenetic taxonomic annotation for MAGs. In the UHGG Project [20], it is estimated that GTDB-Tk successfully annotated approximately 30% of the constructed MAGs, while more than 60% of the MAGs were not annotated to existing species in the GTDB. Although the number of prokaryotic genomes present in the GTDB has increased to 258,406 [156], the comprehensive annotation of MAGs has remained unsolved. To perform metagenomic profiling, some studies have modified suitable approaches from reference genome-based tools, such as Kraken2 and MetaPhlAn3. For example, IGGsearch optimizes the marker gene-based approach in MetaPhlAn3 to quantify the MAG taxonomic abundance at the species level [10]. However, it is difficult to obtain complete and accurate strain-level profiling using MAGs. Most of the currently available strain-level classification tools, such as StrainFinder, StrainEst and ConStrain, distinguish different strains based on genomic variations, such as SNPs. However, genomic variation across closely related strains is eliminated during the assembly process, which decreases the resolution of strain-level classification. To address this challenge, STRONG has been proposed by using initial assembly graphs prior to variant simplification and extracting additional features for subsequent Bayesian inference [157]. It is expected that strain-level resolution will be further refined with the enhanced phasing ability of long-read sequencing technologies.

A recently published study by Kayani et al. [158] also summarized recent advances and introduced commonly used tools and available pipelines to identify microbial genomes from metagenomic sequencing data. They also offered a comparative landscape of these tools. Our review presents more holistic information for both upstream and downstream analysis tools. We have discussed state-of-the-art tools for genome assembly, metagenome binning and QC for a diverse range of sequencing technologies. We have also discussed a variety of options for downstream analyses, such as gene prediction, gene annotation and taxonomic classification. In addition, we have presented the practical aspects of using these tools, such as platform information, computational requirements and specifications, advantages and limitations to guide the readers toward the most effective selection of tools and/or software applications according to their study objectives. This systematic review may serve as a consolidated community resource to accelerate the research and development of related software, tools and pipelines for use in the field of metagenomics. By leveraging the power of metagenomics, the unprecedented power of the microbiome to influence human health and disease may be decoded from characterization to mechanistic insights.

CRediT authorship contribution statement

Chao Yang: Writing – original draft. **Debajyoti Chowdhury:** Writing – original draft, Visualization. **Zhenmiao Zhang:** Resources. **William K. Cheung:** Supervision. **Aiping Lu:** Supervision. **Zhaoxiang Bian:** Supervision. **Lu Zhang:** Project administration, Writing – review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

L.Z. is supported by a Research Grant Council Early Career Scheme (HKBU 22201419), an IRCMS HKBU (No. IRCMS/19-20/D02), an HKBU Start-up Grant Tier 2 (RC-SGT2/19-20/SCI/007), two grants from the Guangdong Basic and Applied Basic Research Foundation (No. 2019A1515011046 and No. 2021A1515012226).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.11.028.

References:

[1] Phimister EG, Lynch SV, Pedersen O. The Human Intestinal Microbiome in Health and Disease. N Engl J Med 2016;375(24):2369–79. https://doi.org/10.1056/NEJMra1600266.
[2] Giles EM, Couper J. Microbiome in health and disease. J Paediatr Child Health 2020;56(11):1735–8. https://doi.org/10.1111/jpc.v56.1110.1111/jpc.14939.
[3] Andersen SB, Schluter J. A metagenomics approach to investigate microbiome sociobiology. Proc Natl Acad Sci 2021;118(10). https://doi.org/10.1073/pnas.2100934118.
[4] Gulati M, Plosky B. As the Microbiome Moves on toward Mechanism. Mol Cell 2020;78(4):567. https://doi.org/10.1016/j.molcel.2020.05.006.
[5] Stres B, Kronegger L. Shift in the paradigm towards next-generation microbiology. FEMS Microbiol Lett 2019;366:1–9. https://doi.org/10.1093/femsle/fnz159.
[6] Bharti R, Grimm DG. Current challenges and best-practice protocols for microbiome analysis. Brief Bioinform 2021;22:178–93. https://doi.org/10.1093/bib/bbz155.
[7] Berg G, Rybakova D, Fischer D, Cernava T, Vergès M-C, Charles T, et al. Microbiome definition re-visited: old concepts and new challenges. Microbiome 2020;8(1). https://doi.org/10.1186/s40168-020-00875-0.
[8] Lagier J-C, Khelaifia S, Alou MT, Ndongo S, Dione N, Hugon P, et al. Culture of previously uncultured members of the human gut microbiota by culturomics. Nat Microbiol 2016;1(12). https://doi.org/10.1038/nmicrobiol.2016.203.
[9] Browne HP, Forster SC, Anonye BO, Kumar N, Neville BA, Stares MD, et al. Culturing of ‘‘unculturable” human microbiota reveals novel taxa and extensive sporulation. Nature 2016;533(7604):543–6. https://doi.org/10.1038/nature17645.
[10] Nayfach S, Shi ZJ, Seshadri R, Pollard KS, Kyrpides NC. New insights from uncultivated genomes of the global human gut microbiome. Nature 2019;568(7753):505–10. https://doi.org/10.1038/s41586-019-1058-x.
[11] Thomas AM, Manghi P, Asnicar F, Pasolli E, Armanini F, Zolfo M, et al. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. Nat Med 2019;25(4):667–78. https://doi.org/10.1038/s41591-019-0405-7.
[12] Thingholm LB, Rühlemann MC, Koch M, Fuqua B, Laucke G, Boehm R, et al. Obese Individuals with and without Type 2 Diabetes Show Different Gut Microbial Functional Capacity and Composition. Cell Host Microbe 2019;26(2):252–264.e10. https://doi.org/10.1016/j.chom.2019.07.004.
[13] Sun Z, Huang S, Zhang M, Zhu Q, Haiminen N, Carrieri AP, et al. Challenges in benchmarking metagenomic profilers. Nat Methods 2021;18(6):618–26. https://doi.org/10.1038/s41592-021-01141-3.
[14] Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods 2013;10(12):1196–9. https://doi.org/10.1038/nmeth.2693.
[15] Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Res 2016;26(11):1612–25. https://doi.org/10.1101/gr.201863.115.
[16] Lin H-H, Liao Y-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. Sci Rep 2016;6(1). https://doi.org/10.1038/srep24175.
[17] Wang Z, Wang Z, Lu YY, Sun F, Zhu S, Hancock J. SolidBin: Improving metagenome binning with semi-supervised normalized cut. Bioinformatics 2019;35(21):4229–38. https://doi.org/10.1093/bioinformatics/btz253.
[18] Yu G, Jiang Y, Wang J, Zhang H, Luo H, Berger B. BMC3C: binning metagenomic contigs using codon usage, sequence composition and read coverage. Bioinformatics 2018. https://doi.org/10.1093/bioinformatics/bty519.
[19] Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell 2019;176(3):649–662.e20. https://doi.org/10.1016/j.cell.2019.01.001.
[20] Almeida A, Nayfach S, Boland M, Strozzi F, Beracochea M, Shi ZJ, et al. A unified catalog of 204,938 reference genomes from the human gut microbiome. Nat Biotechnol 2021;39(1):105–14. https://doi.org/10.1038/s41587-020-0603-3.
[21] Almeida A, Mitchell AL, Boland M, Forster SC, Gloor GB, Tarkowska A, et al. A new genomic blueprint of the human gut microbiota. Nature 2019;568(7753):499–504. https://doi.org/10.1038/s41586-019-0965-1.
[22] Trivedi UH, CÃ©zard T, Bridgett S, Montazam A, Nichols J, Blaxter M, et al. Quality control of next-generation sequencing data without a reference. Front Genet 2014;5. https://doi.org/10.3389/fgene.2014.00111.
[23] Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol 2020;21(1). https://doi.org/10.1186/s13059-020-1935-5.
[24] Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34(17):i884–90. https://doi.org/10.1093/bioinformatics/bty560.
[25] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170.
[26] Chen Y, Chen Y, Shi C, Huang Z, Zhang Y, Li S, et al. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. GigaScience 2018;7(1). https://doi.org/10.1093/gigascience/gix120.
[27] Hufnagel DE, Hufford MB, Seetharam AS. SequelTools: a suite of tools for working with PacBio Sequel raw sequence data. BMC Bioinf 2020;21(1). https://doi.org/10.1186/s12859-020-03751-8.
[28] Hackl T, Hedrich R, Schultz J, Förster F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics 2014;30(21):3004–11. https://doi.org/10.1093/bioinformatics/btu392.
[29] De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C, Berger B. NanoPack: visualizing and processing long-read sequencing data. Bioinformatics 2018;34(15):2666–9. https://doi.org/10.1093/bioinformatics/bty149.
[30] Lanfear R, Schalamun M, Kainer D, Wang W, Schwessinger B, Hancock J. MinIONQC: Fast and simple quality control for MinION sequencing data. Bioinformatics 2019;35(3):523–5. https://doi.org/10.1093/bioinformatics/bty654.
[31] Fukasawa Y, Ermini L, Wang H, Carty K, LongQC CMS. A quality control tool for third generation sequencing long read data. G3: Genes, Genomes, Genet 2020;10:1193–6. https://doi.org/10.1534/g3.119.400864.
[32] Wang Ou, Chin R, Cheng X, Wu MKY, Mao Q, Tang J, et al. Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping,

and de novo assembly. Genome Res 2019;29(5):798–808. https://doi.org/10.1101/gr.245126.118.
[33] Chen Z, Pham L, Wu T-C, Mo G, Xia Yu, Chang PL, et al. Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. Genome Res 2020;30(6):898–909. https://doi.org/10.1101/gr.260380.119.
[34] Kuleshov V, Jiang C, Zhou W, Jahanbani F, Batzoglou S, Snyder M. Synthetic long-read sequencing reveals intraspecies diversity in the human microbiome. Nat Biotechnol 2016;34(1):64–9. https://doi.org/10.1038/nbt.3416.
[35] Haider B, Ahn T-H, Bushnell B, Chai J, Copeland A, Pan C. Omega: an overlap-graph de novo assembler for metagenomics. Bioinformatics 2014;30(19):2717–22. https://doi.org/10.1093/bioinformatics/btu395.
[36] Namiki T, Hachiya T, Tanaka H, Sakakibara Y. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Res 2012;40(20):e155. https://doi.org/10.1093/nar/gks678.
[37] Zerbino DR, Birney E. algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;18(5):821–9. https://doi.org/10.1101/gr.074492.107.
[38] Afiahayati, Sato K, Sakakibara Y. An extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. DNA Res 2015;22(1):69–77. https://doi.org/10.1093/dnares/dsu041.
[39] Ching Liang K, Sakakibara Y. MetaVelvet-DL: a MetaVelvet deep learning extension for de novo metagenome assembly. BMC Bioinf 2021;22. https://doi.org/10.1186/S12859-020-03737-6.
[40] Peng Y, Leung HCM, Yiu SM, Chin FYL. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 2012;28(11):1420–8. https://doi.org/10.1093/bioinformatics/bts174.
[41] Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 2015;31(10):1674–6. https://doi.org/10.1093/bioinformatics/btv033.
[42] Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, et al. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods 2016;102:3–11. https://doi.org/10.1016/j.ymeth.2016.02.020.
[43] Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. metaSPAdes: a new versatile metagenomic assembler. Genome Res 2017;27(5):824–34. https://doi.org/10.1101/gr.213959.116.
[44] Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 2012;19(5):455–77. https://doi.org/10.1089/cmb.2012.0021.
[45] Boisvert S, Raymond F, Godzaridis É, Laviolette F, Corbeil J. Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biol 2012;13(12):R122. https://doi.org/10.1186/gb-2012-13-12-r122.
[46] Bishara A, Moss EL, Kolmogorov M, Parada AE, Weng Z, Sidow A, et al. High-quality genome sequences of uncultured microbes by assembly of read clouds. Nat Biotechnol 2018;36(11):1067–75. https://doi.org/10.1038/nbt.4266.
[47] Tolstoganov I, Bankevich A, Chen Z, Pevzner PA. cloudSPAdes: assembly of synthetic long reads using de Bruijn graphs. Bioinformatics 2019;35(14):i61–70. https://doi.org/10.1093/bioinformatics/btz349.
[48] Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 2010;20(2):265–72. https://doi.org/10.1101/gr.097261.109.
[49] Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of Drosophila. Science 2000;287(5461):2196–204. https://doi.org/10.1126/science:287.5461.2196.
[50] Sommer DD, Delcher AL, Salzberg SL, Pop M. Minimus: a fast, lightweight genome assembler. BMC Bioinf 2007;2007(81):1–11. https://doi.org/10.1186/1471-2105-8-64.
[51] Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 2017;27(5):722–36. https://doi.org/10.1101/gr.215087.116.
[52] Chen Y, Nie F, Xie S-Q, Zheng Y-F, Dai Qi, Bray T, et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. Nat Commun 2021;12(1). https://doi.org/10.1038/s41467-020-20236-7.
[53] Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods 2020;17(2):155–8. https://doi.org/10.1038/s41592-019-0669-3.
[54] Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods 2020;17(11):1103–10. https://doi.org/10.1038/s41592-020-00971-x.
[55] Kolmogorov M, Yuan J, Lin Yu, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 2019;37(5):540–6. https://doi.org/10.1038/s41587-019-0072-8.
[56] Ye C, Hill CM, Wu S, Ruan J, Ma Z(. DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. Sci Rep 2016;6(1). https://doi.org/10.1038/srep31900.
[57] Bertrand D, Shaw J, Kalathiyappan M, Ng AHQ, Kumar MS, Li C, et al. Hybrid metagenomic assembly enables high-resolution analysis of resistance determinants and mobile elements in human microbiomes. Nat Biotechnol 2019;37(8):937–44. https://doi.org/10.1038/s41587-019-0191-2.
[58] Wick RR, Judd LM, Gorrie CL, Holt KE, Phillippy AM. Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 2017;13(6):e1005595. https://doi.org/10.1371/journal.pcbi.1005595.
[59] Liu L, Wang Y, Che Y, Chen Y, Xia Yu, Luo R, et al. High-quality bacterial genomes of a partial-nitritation/anammox system by an iterative hybrid assembly method. Microbiome 2020;8(1). https://doi.org/10.1186/s40168-020-00937-310.21203/rs.3.rs-275906/v1.
[60] Mikheenko A, Saveliev V, Gurevich A. MetaQUAST: evaluation of metagenome assemblies. Bioinformatics 2016;32(7):1088–90. https://doi.org/10.1093/bioinformatics/btv697.
[61] Yuan C, Lei J, Cole J, Sun Y. Reconstructing 16S rRNA genes in metagenomic data. Bioinformatics 2015;31(12):i35–43. https://doi.org/10.1093/bioinformatics/btv231.
[62] Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol 2013;14(5):R47. https://doi.org/10.1186/gb-2013-14-5-r47.
[63] Olson ND, Treangen TJ, Hill CM, Cepeda-Espinoza V, Ghurye J, Koren S, et al. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. Brief Bioinform 2019;20(4):1140–50. https://doi.org/10.1093/bib/bbx098.
[64] Mineeva O, Rojas-Carulla M, Ley RE, Schölkopf B, Youngblut ND, Luigi Martelli P. DeepMAsED: evaluating the quality of metagenomic assemblies. Bioinformatics 2020;36(10):3011–7. https://doi.org/10.1093/bioinformatics/btaa124.
[65] Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ 2014;2:e603. https://doi.org/10.7717/peerj.603.
[66] Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 2016;32(4):605–7. https://doi.org/10.1093/bioinformatics/btv638.
[67] Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nat Methods 2014;11(11):1144–6. https://doi.org/10.1038/nmeth.3103.
[68] Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 2019;7:e7359. https://doi.org/10.7717/peerj.7359.
[69] Mallawaarachchi V, Wickramarachchi A, Lin Yu, Valencia A. GraphBin: refined binning of metagenomic contigs using assembly graphs. Bioinformatics 2020;36(11):3307–13. https://doi.org/10.1093/bioinformatics/btaa180.
[70] Zhang Z, Zhang Lu. METAMVGL: a multi-view graph-based metagenomic contig binning algorithm by integrating assembly and paired-end graphs. BMC Bioinf 2021;22(S10). https://doi.org/10.1186/s12859-021-04284-4.
[71] Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, et al. Improved metagenome binning and assembly using deep variational autoencoders. Nat Biotechnol 2021;39(5):555–60. https://doi.org/10.1038/s41587-020-00777-4.
[72] Kingma DP, Welling M. Auto-Encoding Variational Bayes. 2nd Int Conf Learn Represent ICLR 2014 -. Conf Track Proc 2013.
[73] Murovec B, Deutsch L, Stres B, Rosenberg M. Computational Framework for High-Quality Production and Large-Scale Evolutionary Analysis of Metagenome Assembled Genomes. Mol Biol Evol 2020;37(2):593–8. https://doi.org/10.1093/molbev/msz237.
[74] Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 2018;6(1). https://doi.org/10.1186/s40168-018-0541-1.
[75] Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol 2018;3(7):836–43. https://doi.org/10.1038/s41564-018-0171-1.
[76] Press MO, Wiser AH, Kronenberg ZN, Langford KW, Shakya M, Lo C-C, et al. Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions. BioRxiv 2017;198713. https://doi.org/10.1101/198713.
[77] DeMaere MZ, Darling AE. bin3C: exploiting Hi-C sequencing data to accurately resolve metagenome-assembled genomes. Genome Biol 2019;20(1). https://doi.org/10.1186/s13059-019-1643-1.
[78] Hu J, Schroeder A, Coleman K, Chen C, Auerbach BJ, Li M. Statistical and machine learning methods for spatially resolved transcriptomics with histology. Comput Struct Biotechnol J 2021;19:3829–41. https://doi.org/10.1016/j.csbj.2021.06.052.
[79] Du Y, HiCBin SF. Binning metagenomic contigs and recovering metagenome-assembled genomes using Hi-C contact maps. BioRxiv 2021. https://doi.org/10.1101/2021.03.22.436521.
[80] Du Y, Laperriere SM, Fuhrman J, HiCzin SF. Normalizing metagenomic Hi-C data and detecting spurious contacts using zero-inflated negative binomial regression. BioRxiv 2021. https://doi.org/10.1101/2021.03.01.433489.

[81] Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 2015;25(7):1043–55. https://doi.org/10.1101/gr.186072.114.
[82] Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 2017;35(8):725–31. https://doi.org/10.1038/nbt.3893.
[83] Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2017;2(11):1533–42. https://doi.org/10.1038/s41564-017-0012-7.
[84] Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res 2010;38(12):e132. https://doi.org/10.1093/nar/gkq275.
[85] Kelley DR, Liu Bo, Delcher AL, Pop M, Salzberg SL. Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Res 2012;40(1):e9. https://doi.org/10.1093/nar/gkr1067.
[86] Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res 2010;38(20):e191. https://doi.org/10.1093/nar/gkq747.
[87] Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinf 2010;11(1). https://doi.org/10.1186/1471-2105-11-119.
[88] Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Res 2006;34(19):5623–30. https://doi.org/10.1093/nar/gkl723.
[89] Noguchi H, Taniguchi T, Itoh T. detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Res 2008;15(6):387–96. https://doi.org/10.1093/dnares/dsn027.
[90] Zhang S-W, Jin X-Y, Zhang T. Gene Prediction in Metagenomic Fragments with Deep Learning. Biomed Res Int 2017;2017:1–9. https://doi.org/10.1155/2017/4740354.
[91] Al-Ajlan A, El Allali A. Convolutional Neural Networks for Metagenomics Gene Prediction. Interdiscip Sci 2019;11(4):628–35. https://doi.org/10.1007/s12539-018-0313-4.
[92] Sommer MJ, Salzberg SL, Ouzounis CA. A universal protein model for prokaryotic gene prediction. PLoS Comput Biol 2021;17(2):e1008727. https://doi.org/10.1371/journal.pcbi.1008727.
[93] Boratyn GM, Camacho C, Cooper PS, Coulouris G, Fong A, Ma N, et al. BLAST: a more efficient report with usability improvements. Nucleic Acids Res 2013;41(W1):W29–33. https://doi.org/10.1093/nar/gkt282.
[94] Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. Mol Biol Evol 2017;34(8):2115–22. https://doi.org/10.1093/molbev/msx148.
[95] Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol 2016;428(4):726–31. https://doi.org/10.1016/j.jmb.2015.11.006.
[96] KP K, EM G, F M. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. Methods Mol Biol 2016;1399:207–33. https://doi.org/10.1007/978-1-4939-3369-3_13.
[97] Törönen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation web server. Nucleic Acids Res 2018;46(W1):W84–8. https://doi.org/10.1093/nar/gky350.
[98] Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 2019;47(D1):D309–14. https://doi.org/10.1093/nar/gky1085.
[99] S S, T I, M O, M K, Y A. GHOSTX: A Fast Sequence Homology Search Tool for Functional Annotation of Metagenomic Data. Methods Mol Biol 2017;1611:15–25. https://doi.org/10.1007/978-1-4939-7015-5_2.
[100] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45(D1):D353–61. https://doi.org/10.1093/nar/gkw1092.
[101] Wilke A, Harrison T, Wilkening J, Field D, Glass EM, Kyrpides N, et al. The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. BMC Bioinf 2012;13(1). https://doi.org/10.1186/1471-2105-13-141.
[102] Somervuo P, Holm L. SANSparallel: interactive homology search against Uniprot. Nucleic Acids Res 2015;43(W1):W24–9. https://doi.org/10.1093/nar/gkv317.
[103] Resource TGO. 20 years and still GOing strong. Nucleic Acids Res 2019;47:D330–8. https://doi.org/10.1093/NAR/GKY1055.
[104] R A, TK A, A B, A B, E B, M B, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 2001;29:37–40. https://doi.org/10.1093/NAR/29.1.37.
[105] Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, et al. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 2010;38(suppl_1):D161–6. https://doi.org/10.1093/nar/gkp885.
[106] Attwood TK, Beck ME. PRINTS–a protein motif fingerprint database. Protein Eng 1994;7(7):841–8. https://doi.org/10.1093/protein/7.7.841.
[107] Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. Nucleic Acids Res 2005;33(Web Server):W116–20. https://doi.org/10.1093/nar/gki442.
[108] Kall L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. Nucleic Acids Res 2007;35(Web Server):W429–32. https://doi.org/10.1093/nar/gkm256.
[109] Harrington ED, Singh AH, Doerks T, Letunic I, von Mering C, Jensen LJ, et al. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. Proc Natl Acad Sci U S A 2007;104(35):13913–8. https://doi.org/10.1073/pnas.0702636104.
[110] Ciria R, Abreu-Goodger C, Morett E, Merino E. GeConT: gene context analysis. Bioinformatics 2004;20(14):2307–8. https://doi.org/10.1093/bioinformatics/bth216.
[111] Galperin MY, Wolf YI, Makarova KS, Vera Alvarez R, Landsman D, Koonin EV. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res 2021;49(D1):D274–81. https://doi.org/10.1093/nar/gkaa1018.
[112] Anand S, Kuntal BK, Mohapatra A, Bhatt V, Mande SS, Hancock J. FunGeCo: a web-based tool for estimation of functional potential of bacterial genomes and microbiomes using gene context information. Bioinformatics 2020;36(8):2575–7. https://doi.org/10.1093/bioinformatics/btz957.
[113] Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res 2014;42(D1):D222–30. https://doi.org/10.1093/nar/gkt1223.
[114] Saha CK, Pires RS, Brolin H, Atkinson GC. Predicting Functional Associations using Flanking Genes (FlaGs). BioRxiv 2018. https://doi.org/10.1101/362095.
[115] PA C, AJ M, P H, DH P. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 2019;36:1925–7. https://doi.org/10.1093/BIOINFORMATICS/BTZ848.
[116] SR E. Accelerated Profile HMM Searches. PLoS Comput Biol 2011;7. https://doi.org/10.1371/JOURNAL.PCBI.1002195.
[117] Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol 2018;36(10):996–1004. https://doi.org/10.1038/nbt.4229.
[118] Fa M, Rb K, Ev A. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinf 2010;11:538. https://doi.org/10.1186/1471-2105-11-538.
[119] Wu Y-W. ezTree: an automated pipeline for identifying phylogenetic marker genes and inferring evolutionary relationships among uncultivated prokaryotic draft genomes. BMC Genomics 2018;19(S1). https://doi.org/10.1186/s12864-017-4327-9.
[120] Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 2009;26(7):1641–50. https://doi.org/10.1093/molbev/msp077.
[121] F A, AM T, F B, C M, S M, P M, et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat Commun 2020;11. https://doi.org/10.1038/S41467-020-16366-7.
[122] Rodriguez-R LM, Gunturu S, Harvey WT, Rosselló-Mora R, Tiedje JM, Cole JR, et al. The Microbial Genomes Atlas (MiGA) webserver: taxonomic and gene diversity analysis of Archaea and Bacteria at the whole genome level. Nucleic Acids Res 2018;46(W1):W282–8. https://doi.org/10.1093/nar/gky467.
[123] Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun 2016;7(1). https://doi.org/10.1038/ncomms11257.
[124] Burrows M, Wheeler DJ. A Block-sorting Lossless Data Compression Algorithm, 1994.
[125] Ferragina P, Manzini G. Opportunistic data structures with applications. Annu Symp Found Comput Sci - Proc 2000:390–8. https://doi.org/10.1109/SFCS.2000.892127.
[126] Wood DE, Kraken SSL. ultrafast metagenomic sequence classification using exact alignments. Genome Biol 2014.1–12.;2014(153):15. https://doi.org/10.1186/GB-2014-15-3-R46.
[127] Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 2019;20(1). https://doi.org/10.1186/s13059-019-1891-0.
[128] Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics. PeerJ Comput Sci 2017;3:e104. https://doi.org/10.7717/peerj-cs.104.
[129] Ounit R, Wanamaker S, Close TJ, Lonardi S. fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics 2015;16(1). https://doi.org/10.1186/s12864-015-1419-2.
[130] D A, MJE S, C R, SA B. k-SLAM: accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. Nucleic Acids Res 2017;45:1649–56. https://doi.org/10.1093/NAR/GKW1248.
[131] F B, LJ M, A B-M, L D, F A, S M, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with bioBakery 3. Elife 2021;10. https://doi.org/10.7554/ELIFE.65088.
[132] Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D. ConStrains identifies microbial strains in metagenomic datasets. Nat Biotechnol 2015;33(10):1045–52. https://doi.org/10.1038/nbt.3319.
[133] Smillie CS, Sauk J, Gevers D, Friedman J, Sung J, Youngster I, et al. Strain Tracking Reveals the Determinants of Bacterial Engraftment in the Human Gut Following Fecal Microbiota Transplantation. Cell Host Microbe 2018;23(2):229–40. https://doi.org/10.1016/j.chom.2018.01.003.

[134] Albanese D, Donati C. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nat Commun 2017;2017(81):1–14. https://doi.org/10.1038/s41467-017-02209-5.
[135] Tamames J, Puente-Sánchez F. A Highly Portable, Fully Automatic Metagenomic Analysis Pipeline. Front Microbiol 2019;9. https://doi.org/10.3389/fmicb.2018.0334910.3389/fmicb.2018.03349.s00110.3389/fmicb.2018.03349.s002.
[136] Kurtzer GM, Sochat V, Bauer MW, Gursoy A. Singularity: Scientific containers for mobility of compute. PLoS ONE 2017;12(5):e0177459. https://doi.org/10.1371/journal.pone.0177459.
[137] J V, S W, AK K. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! PLoS One 2017;12. https://doi.org/10.1371/JOURNAL.PONE.0169662.
[138] Latorre-Pérez A, Villalba-Bermell P, Pascual J, Vilanova C. Assembly methods for nanopore-based metagenomic sequencing: a comparative study. Sci Reports 2020.1–14.;2020(101):10. https://doi.org/10.1038/s41598-020-70491-3.
[139] Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software. Nat Methods 2017;14(11):1063–71. https://doi.org/10.1038/nmeth.4458.
[140] Yue Y, Huang H, Qi Z, Dou H-M, Liu X-Y, Han T-F, et al. Evaluating metagenomics tools for genome binning with real metagenomic datasets and CAMI datasets. BMC Bioinf 2020;21(1). https://doi.org/10.1186/s12859-020-03667-3.
[141] Meyer F, Fritz A, Deng Z-L, Koslicki D, Gurevich A, Robertson G, et al. Critical Assessment of Metagenome Interpretation - the second round of challenges. BioRxiv 2021;2021(49):07. https://doi.org/10.1101/2021.07.12.451567.
[142] Dimonaco NJ, Aubrey W, Kenobi K, Clare A, Creevey CJ. No one tool to rule them all: Prokaryotic gene prediction tool performance is highly dependent on the organism of study. BioRxiv 2021. https://doi.org/10.1101/2021.05.21.445150.
[143] Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic Classification. Cell 2019;178(4):779–94. https://doi.org/10.1016/j.cell.2019.07.010.
[144] Chen L-X, Anantharaman K, Shaiber A, Eren AM, Banfield JF. Accurate and complete genomes from metagenomes. Genome Res 2020;30(3):315–33. https://doi.org/10.1101/gr.258640.119.
[145] Orakov A, Fullam A, Coelho LP, Khedkar S, Szklarczyk D, Mende DR, et al. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome Biol 2021;22(1). https://doi.org/10.1186/s13059-021-02393-0.
[146] Moss EL, Maghini DG, Bhatt AS. Complete, closed bacterial genomes from microbiomes using nanopore sequencing. Nat Biotechnol 2020;38(6):701–7. https://doi.org/10.1038/s41587-020-0422-6.
[147] Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. Nat Methods 2016;13(9):751–4. https://doi.org/10.1038/nmeth.3930.
[148] Kovaka S, Fan Y, Ni B, Timp W, Schatz MC. Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. Nat Biotechnol 2021;39(4):431–41. https://doi.org/10.1038/s41587-020-0731-9.
[149] Shakya M, Lo C-C, Chain PSG. Advances and Challenges in Metatranscriptomic Analysis. Front Genet 2019;10. https://doi.org/10.3389/fgene.2019.0090410.3389/fgene.2019.00904.s001.
[150] Leung HCM, Yiu S-M, Chin FYL. IDBA-MTP: A Hybrid Metatranscriptomic Assembler Based on Protein Information. J Comput Biol 2015;22(5):367–76. https://doi.org/10.1089/cmb.2014.0139.
[151] Ye Y, Tang H. Utilizing de Bruijn graph of metagenome assembly for metatranscriptome analysis. Bioinformatics 2016;32(7):1001–8. https://doi.org/10.1093/bioinformatics/btv510.
[152] Al L, Ai K. Metagenomic Data Assembly - The Way of Decoding Unknown Microorganisms. Front Microbiol 2021;12. https://doi.org/10.3389/FMICB.2021.613791.
[153] Mallawaarachchi VG, Wickramarachchi AS, Lin Y. GraphBin2: Refined and Overlapped Binning of Metagenomic Contigs Using Assembly Graphs. DROPS-IDN/12797 2020;172. https://doi.org/10.4230/LIPICS.WABI.2020.8.
[154] Sberro H, Fremin BJ, Zlitni S, Edfors F, Greenfield N, Snyder MP, et al. Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. Cell 2019;178(5):1245–1259.e14. https://doi.org/10.1016/j.cell.2019.07.016.
[155] Andrews SJ, Rothnagel JA. Emerging evidence for functional peptides encoded by short open reading frames. Nat Rev Genet 2014;15(3):193–204. https://doi.org/10.1038/nrg3520.
[156] DH P, M C, C R, AJ M, PA C, P H. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res 2021. https://doi.org/10.1093/NAR/GKAB776.
[157] Quince C, Nurk S, Raguideau S, James R, Soyer OS, Summers JK, et al. STRONG: metagenomics strain resolution on assembly graphs. Genome Biol 2021;22(1). https://doi.org/10.1186/s13059-021-02419-7.
[158] Kayani Masood ur Rehman, Huang W, Feng R, Chen L. Genome-resolved metagenomics using environmental and clinical samples. Brief Bioinform 2021;22(5). https://doi.org/10.1093/bib/bbab030.