Genomics Proteomics Bioinformatics 20 (2022) 867–881

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

REVIEW

Microbial Dark Matter: from Discovery to Applications

Yuguo Zha [#], Hui Chong [#], Pengshuo Yang [#], Kang Ning [*]

MOE Key Laboratory of Molecular Biophysics, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Center of Artificial Intelligence Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China

Received 6 July 2021; revised 28 September 2021; accepted 22 March 2022
Available online 26 April 2022

Handled by Feng Gao

KEYWORDS

Microbiome;
Dark matter;
Artificial intelligence;
Knowledge discovery;
Application

Abstract With the rapid increase of the **microbiome** samples and sequencing data, more and more knowledge about microbial communities has been gained. However, there is still much more to learn about microbial communities, including billions of novel species and genes, as well as countless spatiotemporal dynamic patterns within the microbial communities, which together form the microbial **dark matter**. In this work, we summarized the dark matter in microbiome research and reviewed current data mining methods, especially **artificial intelligence** (AI) methods, for different types of **knowledge discovery** from microbial dark matter. We also provided case studies on using AI methods for microbiome data mining and knowledge discovery. In summary, we view microbial dark matter not as a problem to be solved but as an opportunity for AI methods to explore, with the goal of advancing our understanding of microbial communities, as well as developing better solutions to global concerns about human health and the environment.

Introduction

Microbial communities from diverse global environments have been investigated, revealing abundant novel species and genes, in addition to unique spatiotemporal dynamics across environments [1–3]. Nevertheless, a substantial amount of microbial biodiversity remains to be discovered. These novel community structures and functions constitute an enormous reservoir of diversity that has been referred to as microbial dark matter. Microbial dark matter comprises several different components (**Figure 1**). 1) There are millions of biomes (niches) that microbial communities inhabit [4–6], including general environments such as freshwaters and soils, in addition to context-dependent biomes or understudied biomes such as the gut microbiomes of patients with different diseases. 2) In addition, tens of millions of microbial species are known that span several life kingdoms, including bacteria [1,3,7], archaea [2,8], viruses [9–14], and protists [15]. 3) Furthermore, billions of functional genes are encoded by genomes within microbial communities [2,16,17]. 4) Finally, there are countless dynamic ecological and evolutionary patterns that influence microbial community

* Corresponding author.
E-mail: ningkang@hust.edu.cn (Ning K).
# Equal contribution.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2022.02.007

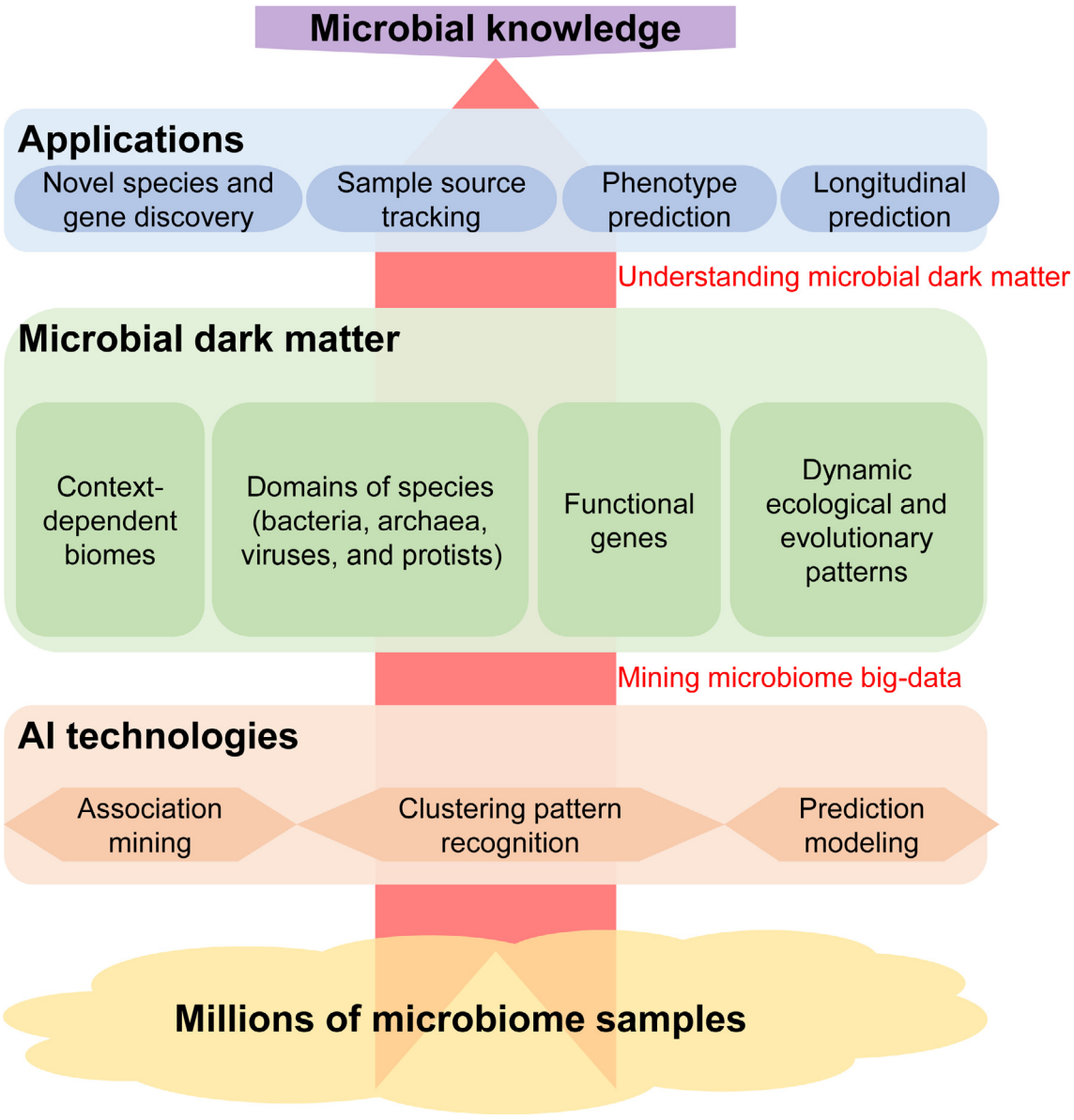

Figure 1 The microbial dark matter and the techniques to better understand such dark matter toward better solutions in applications
There are three key steps for microbiome knowledge discovery from millions of microbiome samples, including the development of AI technologies and microbiome analysis tools, the sets of microbial dark matter to be unearthed, and countless applications. Among these, the microbial dark matter represents the core resource to be discovered. The major types of microbial dark matter introduced in this review include: more than a million context-dependent biomes in which microbial communities could reside; more than a million species, including bacteria, archaea, viruses, and protists; more than a billion functional genes; and the countless number of dynamic ecological and evolutionary patterns. AI, artificial intelligence.

compositions [15,18–22]. All of these areas of microbial dark matter hold great potential for a better understanding of the microbial world, but many of these areas remain understudied [7].

Big data and artificial intelligence (AI) technologies have enabled a more efficient mining of microbial dark matter to generate a better understanding of microbial communities and their potential applications [7,23]. Microbiome big data are derived from millions of microbial community samples, wherein each sample could comprise a few hundred megabytes of 16S rRNA gene sequencing data. In addition, whole-genome sequencing (WGS) can yield over 10 gigabytes of sequencing data per sample. Thus, a typical study including a few thousand samples could comprise over 10 terabytes of sequencing data [24,25]. AI refers to computer systems that leverage computers and machines to mimic the problem-solving and decision-making capabilities of the human mind. In this manuscript, AI includes deep learning, and AI methods in this context mainly refer to methods that use deep learning (*e.g.*, neural networks). Typical AI technologies used in microbiome big data analyses include those for association mining, cluster pattern recognition, and prediction modeling [26–31] (Figure 1). Mining of microbiome big data can help generate

knowledge and models for many applications, including the discovery of novel species and genes, sample source tracking, phenotype prediction (especially for disease diagnosis), and prediction models for longitudinal studies.

In the following text, we first introduce big data and AI technologies used for microbial dark matter analysis. We then introduce the primary types of microbial dark matter that are investigated along with current computational solutions for mining such dark matter. Representative studies are also highlighted that have leveraged AI technologies to generate profound insights across a broad spectrum of applications. Finally, we summarize the advantages of AI technologies in solving microbial dark matter problems, while also describing current bottlenecks and possible future solutions for microbial dark matter mining.

Microbiome big data + AI: venue for microbiome knowledge discovery

The rapidly increasing number of microbiome samples from a variety of global environments (also referred to as biomes), in addition to the massively increased level of sequencing data generated from these samples, has led to the formation of microbiome big data, which represents an important resource pool for knowledge discovery [7,24]. Concomitantly, AI has become an important method, if not the most important method, for mining microbiome big data to generate deeper understanding of microbial communities [23,27,29–32]. Indeed, data integration and data mining are two key modules in most microbiome analyses, although these modules are context-dependent in different analytical applications.

There are currently several databases available for microbiome data integration, including specialized databases such as MetaGenomic Rapid Annotations using Subsystems Technology (MG-RAST) [33], European Bioinformatics Institute (EBI) MGnify [4], and Qiita [34], in addition to general databases like the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) [35–37]. MG-RAST is a public resource for automated phylogenetic and functional analysis of metagenomes. It produces automated functional assignments for metagenomic sequences by comparison against both protein and nucleotide databases. Likewise, Qiita is an open-source web-based platform that enables non-bioinformaticians to easily conduct their own analyses and meta-analyses. Among these databases, the EBI MGnify database is a typical data resource for microbiome studies [4], containing sub-millions of microbiome samples and their sequencing data, as well as analytical results associated with samples. Nearly a million microbiome samples, sequencing data, and meta-data (*e.g.*, data for environmental factors, phenotypic characteristics, and other features) have already been deposited in these databases, and most are publicly available, thereby representing an enormous pool of resources for knowledge discovery [4].

There are currently hundreds of tools for microbiome data analysis in microbiome data mining that can be used for different analytical approaches and at different stages. For example, Mothur [38] is a popular traditional tool for the quality control of 16S rRNA sequencing data, while QIIME 2 [39] is also a widely used traditional tool for microbial community structure profiling. Likewise, many more analytical tools are available for functional profiling of microbial communities. These tools enable the rapid transformation of microbiome sequencing data to community structures and functional profiles [24]. In addition, more advanced traditional analytical tools are also available, including HUMAnN2 [40] and MetaPhlAn2 [41] that enable in-depth analysis of community functions. Further, the traditional bayesian-based method SourceTracker [5] can be used for microbial source tracking, while the traditional method antiSMASH [42] and the AI method DeepARG [43] can be used to mine functional genes.

However, the currently available millions of microbiome samples across hundreds of global environments have led to a shortage of AI methods for mining this volume of data, either for sample comparison and source tracking, functional gene mining, or discovery of dynamic patterns. 1) Current methods in sample comparison and source tracking [5,6] are either based on distance calculations or unsupervised learning and exhibit a tradeoff between accuracy and efficiency, wherein they are only able to accurately source track a few tens of samples. 2) Current methods in functional gene mining [23,42,44] are based on database searches that are not able to find novel genes, while reference-free methods have a high false-positive rate. 3) Current methods and pipelines for context-dependent analytical applications are also limited by their inability to mine intrinsic patterns hidden among thousands to millions of samples [18,20,22,45,46]. All of these limitations have necessitated the development of AI methods that could help in microbial dark matter knowledge discovery.

Dark matter in the microbiome and the computational mining techniques

Context-dependent biomes

Hundreds of biomes, and countless context-dependent biomes, have been annotated or investigated [4]. These include hundreds of general biomes (*e.g.*, soils and freshwaters) and context-dependent biomes that are more specific than general biomes [4]. Context-dependent biomes are involved in many concrete applications related to microbiome knowledge, such as population-specific [1,19] and disease-related patterns [47,48].

Although context-dependent biomes are directly related to various microbiome applications, most remain understudied [1,49]. For example, it remains unknown how gut microbial communities can reflect the progression of colorectal cancer (CRC) in patients [47,48]. Current studies have only indicated that gut microbial communities change with CRC progression. Although there is evidence that gut microbiota could be used to diagnose CRC, it is still not mature for using gut microbial communities as indicators for CRC progression. This is partially due to a lack of understanding of how gut microbial communities mediate CRC, in addition to the lack of an accurate model for prediction. These problems are exacerbated due to the limited accuracy of many disease models because of regional variation [50] or are confounded by host variables such as body mass index (BMI) and age [51].

Numerous gut microbial community datasets have been accumulated, despite that gut microbiomes remain largely understudied, and these datasets are numerous and diverse enough to enable accurate predictions. Indeed, this high

abundance of gut microbiome datasets has been useful for microbial source tracking [52–56]. The SourceTracker program uses a bayesian approach and has been used to differentiate samples from the human mouth, gut, and skin, in addition to monitoring the progression of gut microbial community development in infants [21]. The random forest approach is more widely used to identify microbial community sources via application toward the prediction of locations and times for forensic studies [7,57,58], in addition to application in predicting sources of contamination [59,60]. ONN4MST [61] is a deep learning method which employs a neural network model to source track microbial communities at high efficiency and accuracy without any prior knowledge about the microbial communities to be estimated. Its pre-built biome ontology includes 60 environmental biomes, 25 host-associated biomes, and 10 engineered biomes, which represent the most comprehensive potential sources utilized for source tracking. However, ONN4MST is limited to searching biomes contained in the pre-built biome ontology, but cannot search understudied biomes. EXPERT [62] is also a deep learning method for microbial source tracking, which employs neural network models and acquired flexibility by applying a transfer learning approach, enabling adaptation to newly introduced biomes. It pre-built three neural network models for source tracking among 1) all possible sources from diverse environments, 2) human-associated sources, and 3) human gut-associated sources. Therefore, EXPERT enabled source tracking in many related contexts, such as characterizing disease or time-related compositional shifts of the human gut microbiome. These methods and tasks can contribute to a better understanding of microbial communities. Several databases and data mining methods have been previously reported for microbial source tracking, with representative databases and analytical methods shown in **Table 1**.

Domains of species

Traditional microbiome studies have primarily focused on bacteria, although bacteria only represent a small fraction of all microorganisms. In addition to bacteria, archaea, viruses, and protists are also often abundant in environments. Archaea have distinct molecular characteristics from bacteria, despite often being considered "prokaryotes". Archaea are commonly found in extreme environments and define the limits of life on Earth in many cases [63]. Archaea were originally discovered and described in extreme environments including in high salinity [64], extremely acidic [65], and anaerobic environments [66]. Many unique archaeal genes have been implicated in the adaptations to these extreme environments [65,66]. In addition, viruses are not strictly defined as microbial organisms, because they only harbor a small number of genes and are surrounded by a protein coat. Viruses, as very small infectious agents, rely on living cells to multiply and are the smallest and most abundant of all microorganisms [67]. Protists are unicellular eukaryotic microorganisms that exhibit less complex physiological structures than other eukaryotes. Protists are not necessarily phylogenetically similar but are considered a single group because they do not fit into other taxonomic kingdoms [68].

All microorganisms, including bacteria, archaea, viruses, and protists, are representatives of billions of years of

Table 1 Microbiome sample source tracking methods

Method type	Algorithm	Data pre-processing	Computational model	Computational resource	Representative tool	Ref.
Distance-based	Pair-wise sample distance or similarity	No feature selection	Pair-wise distance calculation	Multi-thread, GPU acceleration	JSD	[52]
					UniFrac	[53]
					Meta-Storms	[54]
					Meta-Prism	[55]
Unsupervised	Bayesian; EM	No feature selection	Model-free unsupervised learning	Multi-thread	SourceTracker	[5]
					FEAST	[6]
Supervised	Ensemble learning; deep learning	Feature selection before source tracking	Model-based supervised learning	Multi-thread, GPU acceleration	Random forest	[56]
					ONN4MST	[61]
					EXPERT	[62]

Note: GPU, graphics processing unit; EM, expectation maximization.

evolution, in addition to the adaptations required to live in specific environments [69]. For example, phylogenetic analyses of these microorganisms have revealed that the composition of human gut microbiomes is affected by hosts [70], while additional research has illustrated dynamic changes and the robustness of gut microbiota in the adaptations to their hosts [71]. Although microorganisms harbor very important functional genes, most of their genomic contents remain poorly understood. For example, several families of archaea were only recently characterized, with their novel evolutionary positions only being recently determined, while phylogenetic positions of most protists have yet to be determined [72]. Moreover, over 60,000 protistan species have been identified in the NCBI taxonomy system, while many have also yet to be identified [73]. Deeper insights into these poorly understood taxa could reveal an enormous amount of important, but currently unknown genes. Several databases and data mining methods have been previously reported for the analysis of bacteria, archaea, viruses, and protists [4,33,74–79], with representative databases and analytical methods shown in **Table 2**.

Functional genes from microbial communities

Billions of functional genes have been annotated. In addition, advancements in sequencing technologies and the development of microbiome culture strategies have led to several microbiome projects that focus on distinct types of biomes. For example, the human microbiome project [1] for identifying the human gut microbiome, the Tara Oceans project [80] for identifying the global ocean microbiome, and the Earth microbiome project [2] that focuses on identifying global soil microbiomes. These projects have generated a massive number of microbial genomes and provide significant reservoirs of functional genes.

Some functional genes represent community-specific housekeeping genes. These genes are essential for individual microorganisms (*e.g.*, genes responsible for DNA replication and RNA transcription that are present in almost all species) but are also necessary for the homeostasis of the entire microbial ecosystem. For example, genes that participate in the nitrogen-cycling process and carbon-cycling process in soil bacterial communities have been detected in all community members and identified as community-specific housekeeping genes [2]. Since nitrogen availability is one of the most common environmental limitations in soils ecosystem, these housekeeping genes could aid in the depletion of excess nitrogen and help degrade recalcitrant soil organic matter, thereby maintaining ecosystem homeostasis [81].

Many functional genes of microorganisms are niche-specific and play important roles in stabilizing microbial community structures, while providing insights into the adaptations of specific microbial populations [2,81]. These genes may only exist in a specific biome but participate in important metabolic pathways, while allowing adaptations to environments by degrading harmful substances [82], adapting to external disturbances [22], and adaptation to hosts [83]. One example is metal resistance genes that are enriched in soil biomes. In soil biomes, metals are major abiotic stressors [84], and many soil taxa have developed full sets of functional genes to adapt to metal stress, such as energy metabolism, integral components of membranes, ion transport/chelation, protein/amino acid

Table 2 Databases and methods for the analysis of bacteria, archaea, viruses, and protists

Domain of species	Type	Name	Description	Website	Ref.
Bacteria and archaea	Database	EBI MGnify	A platform to submit, analyze, discover, and compare microbiome data	https://www.ebi.ac.uk/metagenomics/	[4]
	Database	MG-RAST	A metagenomics service for analysis of microbial community structure and function	https://www.mg-rast.org/	[126]
	Software	QIIME 2	A microbiome bioinformatics platform for processing and analyzing the microbiome	https://qiime2.org/	[74]
Virus	Database	Refseq	Virus genome annotation and curation	https://ftp.ncbi.nlm.nih.gov/genomes/	[75]
	Software	Virfinder	A *k*-mer-frequency-based tool for virus contig identification	https://github.com/jessieren/VirFinder	[76]
	Software	Virsorter	A tool designed to detect viral signals in these different types of microbial sequence data	https://github.com/simroux/VirSorter.git	[77]
Protist	Database	Protist Ribosomal Reference database	Provide a reference database of carefully annotated Protist genomes	https://pr2database.github.io/pr2database/index.html	[78]
	Software	OrthoDB	OrthoDB provides evolutionary and functional annotations of orthologs	https://www.orthodb.org	[79]

Note: EBI, European Bioinformatics Institute; MG-RAST, MetaGenomic Rapid Annotations using Subsystems Technology.

metabolism, carbohydrate/fatty acid metabolism, signal transduction, and DNA binding [85]. Another example is functional genes that facilitate cellular motility that are enriched in lake biomes. Cells in these environments live in highly fluid habitats, and functional genes that enable cellular motility (for example, flagellum-formation proteins) are enriched in community members of water biomes [86].

Current focuses on functional genes among microbial communities primarily include antibiotic resistance genes (ARGs) and biosynthetic gene clusters (BGCs). ARGs are critical for maintaining ecological stability within communities, especially by enabling the resistance to outside stresses. In addition, BGCs are directly associated with important metabolic products of communities. Existing tools for ARG mining include ARGMiner and DeepARG, which have provided both data resources and data analysis methods for ARG analysis [43,87]. Many tools have been proposed to detect ARG sequences from genomic or metagenomic sequence libraries. For instance, ResFinder [88] and SEAR [89] both specifically predict plasmid-borne ARGs, while PATRIC [90] has been developed to identify ARGs that encode resistance to carbapenem, methicillin, and beta-lactam antibiotics. However, these tools exhibit limited efficiency or accuracy, especially for identifying novel ARGs. To better understand microbial functional genes and their effects on microbial communities and environments, more powerful tools using deep learning are urgently needed.

antiSMASH 6.0 is a commonly used tool for BGC data mining and is capable of providing microbial BGC resources for comparison, while also providing machine learning models for identifying novel BGCs from microbial communities [42]. Moreover, antiSMASH 6.0 features improved speed and interactive visualization functionalities that provide a more user-friendly BGC mining platform.

In addition, the functions of many genes identified in microbial communities are unknown. For example, a recent study of rumen metagenome-assembled genomes identified 3535 potentially new species and a total of 442,917 encoded proteins involved in carbohydrate metabolism [91]. Moreover, a recent study identified 13 novel TII-PKS BGCs that are uncommon but likely have high clinical medicinal value as bacterial BGCs [92]. Several databases and analytical methods have been proposed for the analysis of functional genes from microbial communities [93–100], with representative databases and analytical methods shown in **Table 3**.

Microbial ecological and evolutionary patterns

Niche-specific spatiotemporal dynamics within microbial communities, in addition to the consequences of these spatiotemporal dynamics on species evolution, are key determinants for the formation, development, stability, and dynamics of microbial communities [21,101–103]. However, many microbial ecological and evolutionary patterns remain to be discovered.

For example, the discovery of human gut microbial community enterotypes has enabled hundreds of projects to determine the "stable" status of both human and animal gut microbial communities [71,104–107]. Further, the existence of enterotypes for all humans on Earth has only been recognized in the last 10 years, while such patterns dynamically change with environments and host diets [18,20,22,46]. Variation in human gut microbial communities has also been extended to the analysis of animals, leading to the identification of variation in other types of gut microbial communities [105,108].

Another example of ecological patterns in microbial community analyses is the temporal dynamics of human gut

Table 3 Databases and methods for the analysis of functional genes from microbial communities

Type	Name	Description	Website	Ref.
Database	ARGminer	Antibiotic resistance gene database	https://bench.cs.vt.edu/argminer	[87]
	CARD	Comprehensive antibiotic resistance database	https://card.mcmaster.ca/	[93]
	SEED	A database to support effective comparative genome analysis	https://www.theseed.org/wiki/Home_of_the_SEED	[94]
	Pfam	A large collection of protein families	https://pfam.xfam.org/	[95]
	EggNOG	A database of orthology relationships, gene evolutionary histories, and functional annotations	https://eggnog5.embl.de/#/app/home	[96]
	Uniref	A database provides a comprehensive protein information	https://www.uniprot.org/help/uniref	[97]
	MetaCyc	A comprehensive reference database of metabolic pathways and enzymes from all domains of life	https://MetaCyc.org	[98]
	KEGG	Kyoto encyclopedia of genes and genomes	https://www.genodme.jp/kegg/	[99]
Software	DeepARG	A tool with a fully automated data analysis pipeline for antibiotic resistance annotation of raw metagenomic samples	https://bench.cs.vt.edu/deeparg	[43]
	AntiSMASH	A tool for the rapid genome-wide identification, annotation, and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genomes	https://antismash.secondarymetabolites.org/	[42]
	HUMAnN2	A pipeline for profiling the microbial pathways	https://huttenhower.sph.harvard.edu/humann2	[40]
	PICRUSt2	Provide information about the functional composition of sampled communities	https://github.com/picrust/picrust2	[100]

microbial communities. Human gut microbiota rapidly responds to changes in diet [103,109,110], and the composition of an individual's gut microbiota is predominantly determined by dietary habits over the long term (*i.e.*, more than 1 year) [107,111]. However, these dynamics are highly variable among individuals [112,113]. Over short-term time scales (*i.e.*, less than 1 month), human gut microbiota can drastically change during dietary shifts, while such changes can also be quickly reversed after shifts in diets [22]. In addition, strong "plastic" patterns can be observed over mid-term time scales (*i.e.*, between a month and a year) [22] (**Figure 2**).

However, these examples only represent a few of the many ecological and evolutionary patterns that remain to be discovered. For example, context-dependent patterns such as microbiome-related disease patterns [114] remain understudied, especially for cancer disease-microbiome patterns [115], longitudinal microbiome patterns [21], large-scale context-independent ecological patterns [2], and the evolutionary patterns of specific genes [116]. Among these areas, cancer disease-microbiome patterns are of particular importance [115], because they could provide evidence for the roles of bacteria in cancer states. Such studies could lead the way for the next generation of cancer prediction and therapeutic strategies [117]. Another area that lacks resolution is a basic and theoretical framework for how various genes act in concert to enable microorganisms to inhabit specific niches and how they may correspond to changes in the environment [116]. Homogeneous selection, homogeneous dispersal, and neutral theory are all ecological concepts that may help our understanding of these processes [118]. Likewise, the evolutionary process of genetic drift, natural selection, and homologous recombination may aid in developing the aforementioned framework [119]. Overall, investigations into these problems could help develop a better understanding of the ecological and evolutionary patterns ranging from small to large scales [22,107,111–117,120].

The dilemma of traditional methods could be solved by deep learning methods

Several computational solutions have been proposed to solve issues in understanding microbial dark matter [1–3] (Figure 1). However, most of these methods have tradeoffs and especially when considering big data analytical efficiency and accuracy. For example, traditional unsupervised learning methods such as SourceTracker [5] and FEAST [6] can achieve very high accuracy in microbial community source tracking when there are hundreds of samples and a handful of biomes. However, when the number of samples and biomes increases, running time increases rapidly, preventing large-scale source tracking. This problem could be solved by deep learning solutions by utilizing model-based methods such as neural networks that would enable improvements in both speed and accuracy during source tracking [61,62].

Another example of a useful application of deep learning is in ARG mining, in which traditional methods based on Basic Local Alignment Search Tool (BLAST) searches have been used to identify candidate ARGs. However, such an approach is limited to comparison against known ARGs, and search speed is not very fast when using millions of candidates that require screening. The use of deep learning approaches via model-based methods has been shown to more efficiently mine novel ARGs out of millions of candidates [43,121].

The abovementioned limitations suggest that AI techniques could be used to more efficiently uncover knowledge about microbial dark matter. AI techniques are advantageous in that they generate models from a large number of samples that are

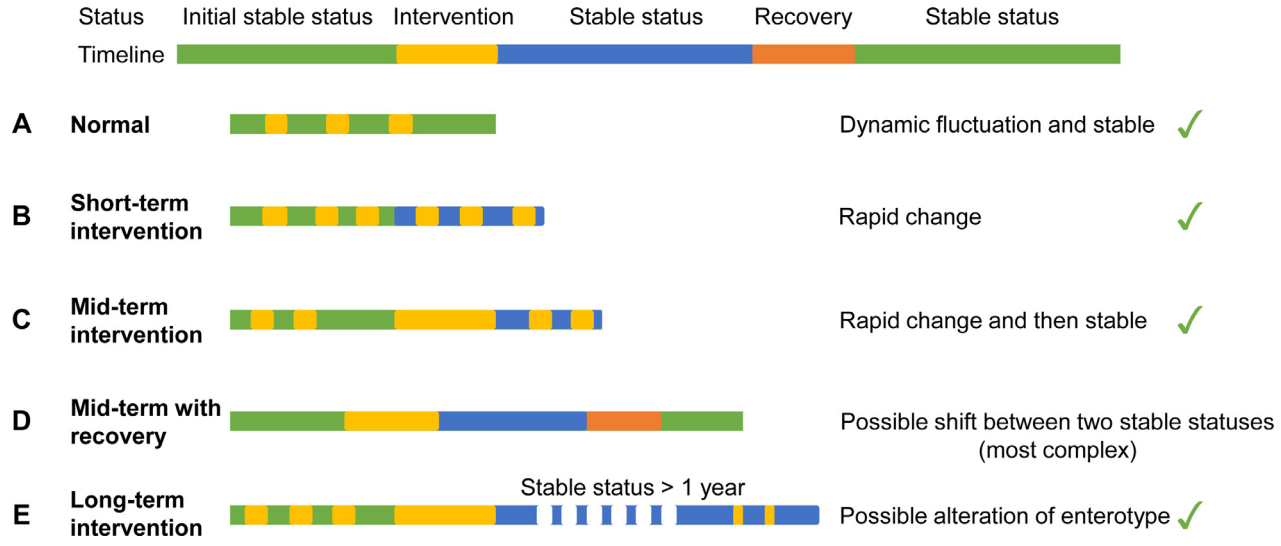

Figure 2 The longitudinal dynamics of the human gut microbial communities have certain patterns
For short-term intervention, it has been demonstrated that dietary intervention is the main driver of the rapid change in the gut microbial community. For mid-term intervention, it has been demonstrated that the dietary intervention could become stable after a month. For long-term intervention, even the enterotype might be changed after one year. The dynamic patterns are based on human gut microbial community samples. And the community profile of each sample is based on the combination of species with different relative abundances.

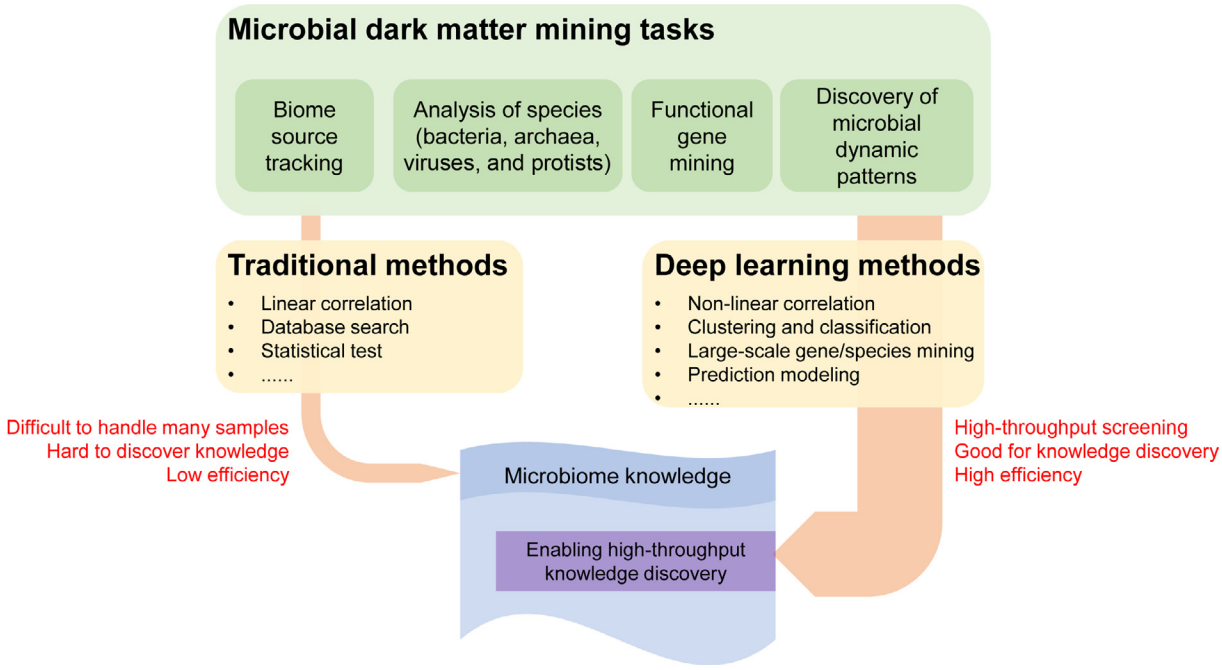

Figure 3 The deep learning approaches for solving the microbial dark matter mining problems
Compared with traditional methods, deep learning methods have enabled high-throughput screening, thus is good for unknown knowledge discovery and has high efficiency.

representative of global profiles within context-dependent subjects [27]. AI techniques are therefore suitable for accurate and fast searches when new sample (either a community, a gene, or a pattern) is searched against established models [28,30,122]. Thus, AI techniques are especially useful for mining microbial dark matter data, particularly when trying to improve tradeoffs between accuracy and efficiency.

Solutions for eliminating tradeoffs in current microbial data mining approaches rely on deep learning techniques [27–31] (**Figure 3**). In particular, model-based methods such as neural networks are advantageous in source tracking. For example, once a rational model has been built, improved efficiency and accuracy of model-based methods can be achieved that is comparable to, or even better than, existing distance-based and unsupervised methods [61,123]. The same approach is suitable for gene mining issues [121]. In spatiotemporal dynamic pattern mining, deep learning approaches could also be used to discover intrinsic patterns out of cross-sections or longitudinal cohorts [124,125].

One example application of the usefulness of these approaches is microbial source tracking. The first model-based method for source tracking, ONN4MST, already outperforms existing methods [61] for source tracking of known biomes. Further, the EXPERT method employs ONN4MST models to source track in different contexts [62] and has exhibited a high potential to facilitate mining of the microbial dark matter data. The EXPERT models are based on fundamental neural network models and transfer learning approaches, and exhibit high speed and accuracy, even when analyzing very few (a few hundred) samples from understudied biomes.

Functional gene mining from metagenome sequences is also an area that could be improved by AI approaches. For example, DeepARG uses a deep learning method that takes sequence alignment similarities as input and employs a neural network to enhance ARG prediction accuracy [43]. DeepARG can achieve a higher precision (0.97) and recall (0.91) than model-free ARG identification methods that exhibit a precision of 0.96 and a recall of 0.51. The hierarchical multi-task deep learning for annotating antibiotic resistance genes (HMD-ARG) method follows a similar approach by curating a comprehensive ARG database and then generating a hierarchical multi-task deep learning model that could help improve novel ARG discovery [121]. The supervised model-based methods are orders of magnitude faster than existing model-free methods.

Taken together, these studies have shown that supervised model-based methods are suitable for large-scale microbiome data mining and can facilitate accurate and efficient microbial dark matter discovery. Further, more advanced deep learning techniques such as convolutional neural networks (CNNs) and transfer learning could enable more accurate data mining, while also expanding the scope for knowledge discovery. Representative AI methods for the analysis of microbial dark matter are shown in **Table 4**.

Applications in microbial dark matter analysis

Computational tools, especially machine learning tools, have enabled a diverse set of applications that rely on microbial dark matter mining (**Figure 4**). These tools are described in detail below.

Quality control of sequencing data

Genomic and metagenomic sequencing data commonly contain possible contamination from various environments, yet identification and removal of these contaminants remain

Table 4 Representative methods for the analysis of microbial dark matter

Microbial dark matter	Representative traditional method	Representative AI method	Summary
Context-dependent biomes	SourceTracker [5], FEAST [6]	ONN4MST [61], EXPERT [62]	AI methods are especially suitable for source tracking among thousands to millions of samples in a fast and accurate manner
Domains of species	QIIME2 [74], Virfinder [76], OrthoDB [79]	\	Current methods for bacteria, archaea, virus, and protist analyses are limited to identifying known species
Functional genes	HUMAnN2 [40], antiSMASH [42]	DeepARG [43], HMD-ARG [121]	Current methods could identify novel genes, but with low speed and low fidelity
Dynamic ecological and evolutionary patterns	PCoA, MITRE [120]	\	Current methods are not sensitive to identifying the dynamic ecological and evolutionary patterns

Note: AI, artificial intelligence; "\", not reported.

difficult [126–128]. Machine learning-enabled source tracking and sequence clustering methods can act together to identify and remove contaminants, regardless of known or unknown sources [128]. Indeed, the application of machine learning methods to sequencing data can lead to the removal of most known contaminants [129]. Further, previously unknown or unexpected contaminants can also be identified and removed in an intelligent manner [130], enabling more accurate environmental and clinical data quality-filtering for subsequent studies [131]. For example, machine learning methods have enabled the accurate identification of contaminants in a typical molecular biology laboratory, including from workbenches and floors [6].

Microbial source tracking

Microbial source tracking can be used in multi-faceted applications, including in contaminant sample source identification, forensic studies, and disease prediction [5,6]. Traditional methods for microbial source tracking can generally be categorized into distance-based and unsupervised methods. Distance-based methods compute the distances between each pair of samples (using multiple distance measures) [52–55], while unsupervised methods are limited by pre-defined sets of sources for source tracking [5,6]. Supervised model-based methods [56] accurately quantify the contributions of source biomes for a specific sample but can also adapt to the analysis of samples from less studied biomes [61]. For example, the EXPERT method is able to accurately differentiate CRC stages in patients using a model built from more than 10,000 human microbiome samples from normal individuals [62].

Novel species discovery from different domains of species

It has been estimated that there are more than a million unknown species and more than a billion unannotated microbial genes, providing an expansive opportunity for species and gene discovery. Novel species that live in extreme environments, or those that could generate important metabolites are of interest in clinical and industrial applications [132–135]. Machine learning techniques have enabled novel species discovery from diverse taxa. For example, a recent study revealed thousands of novel protistan species at the global scale and established that protists are distributed discretely, with soil pH as the most important influencing factor on their distribution [73].

Novel functional gene discovery

Functional gene mining from microbial communities, and especially ARG and BGC mining, are a focus of many studies [42,136,137]. Traditional methods for functional gene mining rely on databases comprising ARGs and BGCs that can be searched against, although these approaches are limited in the ability to discover novel functional genes [43,87–90,136,137]. Machine learning methods have, however, made it possible to discover novel functional genes more efficiently. For example, DeepARG has identified thousands of novel ARGs that were previously unannotated [43]. In addition, the use of HMD-ARG has shown that novel ARGs are functional via the combination of computation and wet-lab validation analyses [121].

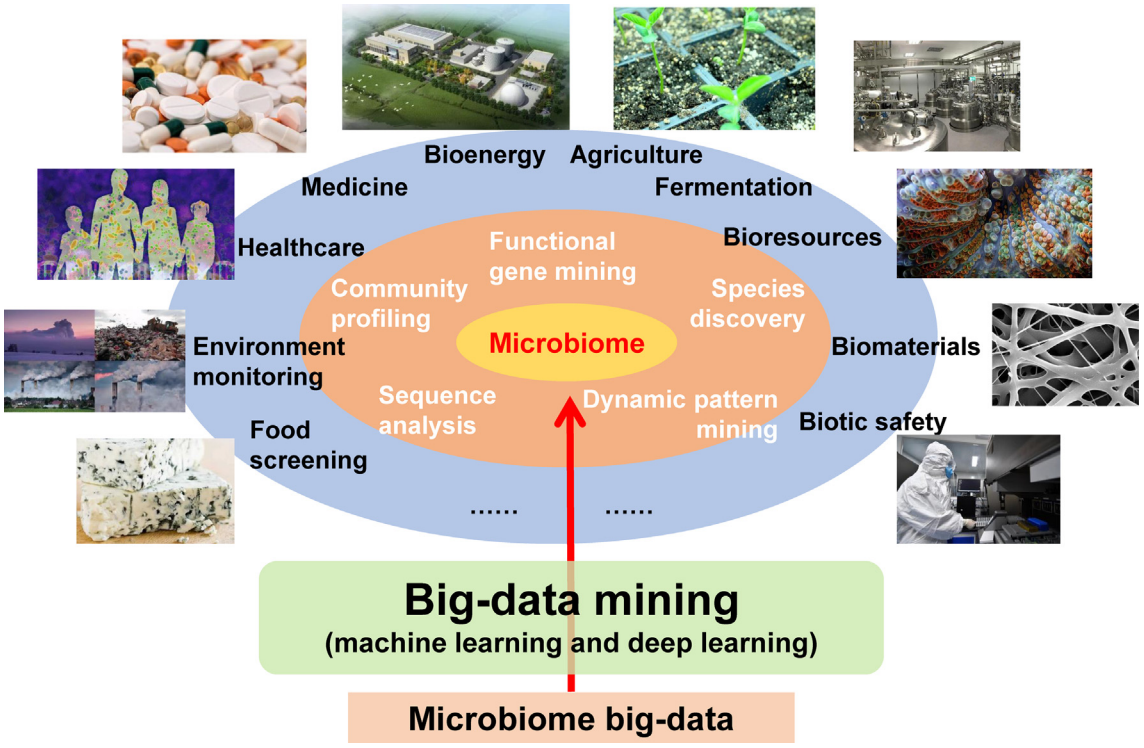

Figure 4 Applications based on computational tools for microbial dark matter mining

Phenotype prediction based on spatiotemporal patterns of the microbial communities

Human microbial communities are intricately linked with the health status of hosts, and it is possible to derive a model for predicting host phenotypes based on host-microbial communities [20]. Indeed, supervised learning methods are suitable for such analyses. For example, EXPERT has been used to accurately differentiate samples from patients with nearly twenty diseases, in addition to monitoring CRC stages in patients using models constructed from over 10,000 human microbiome samples in normal individuals [62].

Longitudinal predictions, as in the prediction of disease progressions, represent another potential area of application [21,138]. Supervised learning has been successfully applied for longitudinal predictions by identifying key events along timelines, in addition to identifying differences among infants at different stages [62]. Machine learning methods have also been used for highly accurate human chronological age predictions [62]. Furthermore, machine learning methods such as random forest classification have been successfully applied in forensic studies leading to the ability to determine times and locations precisely [57]. Phenotype prediction modeling can also be used for identifying environmental indicators [139,140]. For example, machine learning methods have been used to establish several microbial-based lake environment monitoring models using years of freshwater lake samples [139].

Overall, a more comprehensive understanding of microbial dark matter has opened the door for countless applications, with more in-depth applications being possible due to a better understanding of microbial communities. Although most of these applications are context-dependent, they could, in turn, provide large microbial community datasets that can help deepen our understanding of microbial communities across a variety of niches. Such iterative interactions between microbiome knowledge generation and applications could spur a significant improvement in microbiome research.

Conclusion

Understanding microbial dark matter has emerged as a grand challenge for microbial research, and big data mining of microbial dark matter could be a powerful approach to understanding such dark matter. Microbial community niches, species, functional genes, and spatiotemporal dynamics all constitute important components of microbial dark matter. Microbiome studies have gradually produced an abundance of high-quality data that have enabled data mining techniques for large-scale microbiome data mining to promote an in-depth understanding of microbial communities. The rapid development of microbiome data mining could certainly boost the discovery of additional microbial resources and dynamic patterns from these dark matter datasets.

Current microbiome databases and analytical methods are suitable for small-scale microbiome data mining, but two important aspects of data analyses require urgent improvement. First, next-generation microbiome databases that contain sequencing data in addition to metadata, including environmental factors and phenotypic characteristics, are needed. The second is a need for enabling methods in genes mining among millions to billions of samples. Recent updates

in metagenomic databases such as Qiita [34], in addition to model-based methods such as DeepARG [43] and EXPERT [6], have largely solved the large-scale microbiome data mining problem but are only appropriate for specific mining problems.

The potential insights from microbial dark matter discovery are indeed very inspiring. Microbial dark matter comprises novel biomes, species, functional genes, and spatiotemporal patterns. Increased discovery in these areas would certainly lead to the identification of new principles and could be useful for a tremendous number of applications in healthcare, biomedicine, environmental monitoring, and bio-safety, in addition to other areas.

Finally, we emphasize that it has already become clear that microbial communities have been collected from increasingly diverse niches around the world. Nevertheless, it is important to note that the currently sampled niches are far from complete, especially when considering the countless number of application-dependent contexts. Thus, microbial dark matter in a broad sense is almost infinite. However, data mining models could also be updated to cope with increasing diversity in dark matter, although multiple models might be needed to obtain optimal data mining results among different contexts. The arms race between microbial dark matter and AI modeling could lead to a much deeper understanding of microbial communities and their interactions with environments. In addition, it should be emphasized that advances in microbiome technologies will also play important roles in better understanding microbial dark matter. Recent advances in microbiome techniques include the use of third-generation sequencing (*e.g.*, Oxford Nanopore) and the use of metatranscriptomics. For example, a recent study of *in vivo* dental plaques formed on hydroxyapatite disks for 6 h from 74 young adults documented the identification of 21 initial colonizing taxa based on full-length 16S rRNA gene sequences generated with long-read sequencing technology [141]. Metatranscriptomic sequencing can be used to ascertain a gene's activity in a defined environment. Gosalbes et al. [142] conducted a metatranscriptomic analysis of fecal microbiomes from ten healthy humans and discovered that the gut microbiota's primary functional roles were carbohydrate metabolism, energy production, and synthesis of cellular components. This work has proven that the metatranscriptome study could reveal the functions of microbes in amino acid and lipid metabolism.

Taken together, understanding microbial dark matter is not only a challenge, but also an opportunity for computational microbiologists to explore large datasets with the goal of better understanding microbial communities and identifying better solutions for current global concerns in human health and environments. AI technologies have already been applied to microbial dark matter mining problems, and we expect that the increased maturity of AI technologies will lead to increasing in-depth microbiome knowledge that could be mined out of the massive pool of microbial dark matter.

CRediT author statement

Yuguo Zha: Writing - original draft, Writing - review & editing. **Hui Chong:** Writing - original draft. **Pengshuo Yang:** Writing - original draft. **Kang Ning:** Writing - review & editing, Conceptualization, Supervision, Funding acquisition. All authors have read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 32071465, 31871334, and 31671374) and the National Key R&D Program of China (Grant No. 2018YFC0910502).

ORCID

ORCID 0000-0003-3702-9416 (Yuguo Zha)
ORCID 0000-0002-7676-7975 (Hui Chong)
ORCID 0000-0002-2757-3584 (Pengshuo Yang)
ORCID 0000-0003-3325-5387 (Kang Ning)

References

[1] Proctor LM, Creasy HH, Fettweis JM, Lloyd-Price J, Mahurkar A, Zhou W, et al. The integrative human microbiome project. Nature 2019;569:641–8.
[2] Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 2017;551:457–63.
[3] Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Ocean plankton. Structure and function of the global ocean microbiome. Science 2015;348:1261359.
[4] Mitchell AL, Almeida A, Beracochea M, Boland M, Burgin J, Cochrane G, et al. MGnify: the microbiome analysis resource in 2020. Nucleic Acids Res 2020;48:570–8.
[5] Knights D, Kuczynski J, Charlson ES, Zaneveld J, Mozer MC, Collman RG, et al. Bayesian community-wide culture-independent microbial source tracking. Nat Methods 2011;8:761–3.
[6] Shenhav L, Thompson M, Joseph TA, Briscoe L, Furman O, Bogumil D, et al. FEAST: fast expectation-maximization for microbial source tracking. Nat Methods 2019;16:627–32.
[7] Biteen JS, Blainey PC, Cardon ZG, Chun M, Church GM, Dorrestein PC, et al. Tools for the microbiome: nano and beyond. ACS Nano 2016;10:6–37.
[8] Human Microbiome Jumpstart Reference Strains Consortium, Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, et al. A catalog of reference genomes from the human microbiome. Science 2010;328:994–9.
[9] Jonas O, Seifman R. Do we need a global virome project? Lancet Glob Health 2019;7:1314–6.
[10] Carroll D, Daszak P, Wolfe ND, Gao GF, Morel CM, Morzaria S, et al. The global virome project. Science 2018;359:872–4.
[11] Aggarwala V, Liang G, Bushman FD. Viral communities of the human gut: metagenomic analysis of composition and dynamics. Mob DNA 2017;8:12.
[12] Handley SA. The virome: a missing component of biological interaction networks in health and disease. Genome Med 2016;8:32.
[13] Virgin HW. The virome in mammalian physiology and disease. Cell 2014;157:142–50.
[14] Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, et al. The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 2011;21:1616–25.
[15] Gilbert JA, Steele JA, Caporaso JG, Steinbruck L, Reeder J, Temperton B, et al. Defining seasonal marine microbial community dynamics. ISME J 2012;6:298–308.

[16] Surana NK, Kasper DL. Moving beyond microbiome-wide associations to causal microbe identification. Nature 2017;552:244–7.
[17] Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 2010;464:59–65.
[18] Halfvarson J, Brislawn CJ, Lamendella R, Vázquez-Baeza Y, Walters WA, Bramer LM, et al. Dynamics of the human gut microbiome in inflammatory bowel disease. Nat Microbiol 2017;2:17004.
[19] Smits SA, Leach J, Sonnenburg ED, Gonzalez CG, Lichtman JS, Reid G, et al. Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. Science 2017;357:802–6.
[20] Bashan A, Gibson TE, Friedman J, Carey VJ, Weiss ST, Hohmann EL, et al. Universality of human microbial dynamics. Nature 2016;534:259–62.
[21] Backhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva-Datchary P, et al. Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe 2015;17:852.
[22] Liu H, Han M, Li SC, Tan G, Sun S, Hu Z, et al. Resilience of human gut microbial communities for the long stay with multiple dietary shifts. Gut 2019;68:2254–5.
[23] Cheng M, Cao L, Ning K. Microbiome big-data mining and applications using single-cell technologies and metagenomics approaches toward precision medicine. Front Genet 2019;10:972.
[24] Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, et al. Best practices for analysing microbiomes. Nat Rev Microbiol 2018;16:410–22.
[25] Mallick H, Ma S, Franzosa EA, Vatanen T, Morgan XC, Huttenhower C. Experimental design and quantitative analysis of microbial community multiomics. Genome Biol 2017;18:228.
[26] Dhombres F, Charlet J. Formal medical knowledge representation supports deep learning algorithms, bioinformatics pipelines, genomics data analysis, and big data processes. Yearb Med Inform 2019;28:152–5.
[27] Li Y, Huang C, Ding L, Li Z, Pan Y, Gao X. Deep learning in bioinformatics: introduction, application, and perspective in the big data era. Methods 2019;166:4–21.
[28] Tang B, Pan Z, Yin K, Khateeb A. Recent advances of deep learning in bioinformatics and computational biology. Front Genet 2019;10:214.
[29] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform 2017;18:851–69.
[30] Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet 2019;51:12–8.
[31] Lan K, Wang DT, Fong S, Liu LS, Wong KKL, Dey N. A survey of data mining and deep learning in bioinformatics. J Med Syst 2018;42:139.
[32] Wang W, Gao X. Deep learning in bioinformatics. Methods 2019;166:1–3.
[33] Meyer F, Bagchi S, Chaterji S, Gerlach W, Grama A, Harrison T, et al. MG-RAST version 4-lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. Brief Bioinform 2019;20:1151–9.
[34] Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vazquez-Baeza Y, Ackermann G, et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods 2018;15:796–8.
[35] Bernstein MN, Gladstein A, Latt KZ, Clough E, Busby B, Dillman A. Jupyter notebook-based tools for building structured datasets from the sequence read archive. F1000Res 2020;9:376.
[36] Alnasir J, Shanahan HP. Investigation into the annotation of protocol sequencing steps in the sequence read archive. Gigascience 2015;4:23.
[37] Kodama Y, Shumway M, Leinonen R. The sequence read archive: explosive growth of sequencing data. Nucleic Acids Res 2012;40:D54–6.
[38] Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking metagenomics tools for taxonomic classification. Cell 2019;178:779–94.
[39] Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 2019;37:852–7.
[40] Franzosa EA, McIver LJ, Rahnavard G, Thompson LR, Schirmer M, Weingart G, et al. Species-level functional profiling of metagenomes and metatranscriptomes. Nat Methods 2018;15:962–8.
[41] Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods 2012;9:811–4.
[42] Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema Marnix H, et al. AntiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Res 2021;49:W29–35.
[43] Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. Microbiome 2018;6:23.
[44] Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker discovery and explanation. Genome Biol 2011;12:R60.
[45] Jameson E, Doxey AC, Airs R, Purdy KJ, Murrell JC, Chen Y. Metagenomic data-mining reveals contrasting microbial populations responsible for trimethylamine formation in human gut and marine ecosystems. Microb Genom 2016;2:e000080.
[46] Ren T, Boutin S, Humphries MM, Dantzer B, Gorrell JC, Coltman DW, et al. Seasonal, spatial, and maternal effects on gut microbiome in wild red squirrels. Microbiome 2017;5:163.
[47] Jahani-Sherafat S, Alebouyeh M, Moghim S, Ahmadi Amoli H, Ghasemian-Safaei H. Role of gut microbiota in the pathogenesis of colorectal cancer; a review article. Gastroenterol Hepatol Bed Bench 2018;11:101–9.
[48] Zhu Q, Gao R, Wu W, Qin H. The role of gut microbiota in the pathogenesis of colorectal cancer. Tumour Biol 2013;34:1285–300.
[49] Stokholm J, Thorsen J, Blaser MJ, Rasmussen MA, Hjelmso M, Shah S, et al. Delivery mode and gut microbial changes correlate with an increased risk of childhood asthma. Sci Transl Med 2020;12:eaax9929.
[50] He Y, Wu W, Zheng HM, Li P, McDonald D, Sheng HF, et al. Regional variation limits applications of healthy gut microbiome reference ranges and disease models. Nat Med 2018;24:1532–5.
[51] Vujkovic-Cvijin I, Sklar J, Jiang L, Natarajan L, Knight R, Belkaid Y. Host variables confound gut microbiota studies of human disease. Nature 2020;587:448–54.
[52] Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inf Theory 1991;37:145–51.
[53] Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R. UniFrac: an effective distance metric for microbial community comparison. ISME J 2011;5:169–72.
[54] Su X, Xu J, Ning K. Meta-Storms: efficient search for similar microbial communities based on a novel indexing scheme and similarity score for metagenomic data. Bioinformatics 2012;28:2493–501.
[55] Zhu M, Kang K, Ning K. Meta-Prism: ultra-fast and highly accurate microbial community structure search utilizing dual indexing and parallel computation. Brief Bioinform 2021;22:557–67.
[56] Roguet A, Eren AM, Newton RJ, McLellan SL. Fecal source identification using random forest. Microbiome 2018;6:185.
[57] Metcalf JL, Xu ZZ, Weiss S, Lax S, Van Treuren W, Hyde ER, et al. Microbial community assembly and metabolic function during mammalian corpse decomposition. Science 2016;351:158–62.

[58] Smith A, Sterba-Boatwright B, Mott J. Novel application of a statistical technique, random forests, in a bacterial source tracking study. Water Res 2010;44:4067–76.
[59] Li C, Chen J, Wang J, Ma Z, Han P, Luan Y, et al. Occurrence of antibiotics in soils and manures from greenhouse vegetable production bases of Beijing, China and an associated risk assessment. Sci Total Environ 2015;521–2:101–7.
[60] Tong L, Huang S, Wang Y, Liu H, Li M. Occurrence of antibiotics in the aquatic environment of Jianghan Plain, central China. Sci Total Environ 2014;497–8:180–7.
[61] Zha Y, Chong H, Qiu H, Kang K, Dun Y, Chen Z, et al. Ontology-aware deep learning enables ultrafast, accurate and interpretable source tracking among sub-million microbial community samples from hundreds of niches. Genome Med 2022;14:43.
[62] Chong H, Yu Q, Zha Y, Xiong G, Wang N, Huang X, et al. EXPERT: transfer learning-enabled context-aware microbial source tracking. bioRxiv 2021;428751.
[63] Belilla J, Moreira D, Jardillier L, Reboul G, Benzerara K, Lopez-Garcia JM, et al. Hyperdiverse archaea near life limits at the polyextreme geothermal Dallol area. Nat Ecol Evol 2019;3:1552–61.
[64] Yue Y, Shao T, Long X, He T, Gao X, Zhou Z, et al. Microbiome structure and function in rhizosphere of Jerusalem artichoke grown in saline land. Sci Total Environ 2020;724:138259.
[65] Korzhenkov AA, Toshchakov SV, Bargiela R, Gibbard H, Ferrer M, Teplyuk AV, et al. Archaea dominate the microbial community in an ecosystem with low-to-moderate temperature and extreme acidity. Microbiome 2019;7:11.
[66] Wang Y, Feng X, Natarajan VP, Xiao X, Wang F. Diverse anaerobic methane- and multi-carbon alkane-metabolizing archaea coexist and show activity in Guaymas Basin hydrothermal sediment. Environ Microbiol 2019;21:1344–55.
[67] Simmonds P, Adams MJ, Benko M, Breitbart M, Brister JR, Carstens EB, et al. Consensus statement: virus taxonomy in the age of metagenomics. Nat Rev Microbiol 2017;15:161–8.
[68] Caron DA, Alexander H, Allen AE, Archibald JM, Armbrust EV, Bachy C, et al. Probing the evolution, ecology and physiology of marine protists using transcriptomics. Nat Rev Microbiol 2017;15:6–20.
[69] Davenport ER, Sanders JG, Song SJ, Amato KR, Clark AG, Knight R. The human microbiome in evolution. BMC Biol 2017;15:127.
[70] Cheng M, Ning K. Stereotypes about enterotype: the old and new ideas. Genomics Proteomics Bioinformatics 2019;17:4–12.
[71] Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, et al. Enterotypes of the human gut microbiome. Nature 2011;473:174–80.
[72] Baker BJ, De Anda V, Seitz KW, Dombrowski N, Santoro AE, Lloyd KG. Diversity, ecology and evolution of archaea. Nat Microbiol 2020;5:887–900.
[73] Miao W, Song L, Ba S, Zhang L, Guan G, Zhang Z, et al. Protist 10,000 Genomes Project. Innovation (Camb) 2020;1:100058.
[74] Hall M, Beiko RG. 16S rRNA gene analysis with QIIME2. Methods Mol Biol 2018;1849:113–29.
[75] Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, et al. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res 2018;46:D851–60.
[76] Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel *k*-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome 2017;5:69.
[77] Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. PeerJ 2015;3:e985.
[78] Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, et al. The protist ribosomal reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res 2013;41:D597–604.
[79] Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simao FA, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res 2019;47:D807–11.
[80] Sunagawa S, Acinas SG, Bork P, Bowler C, Tara Oceans C, Eveillard D, et al. Tara Oceans: towards global ocean ecosystems biology. Nat Rev Microbiol 2020;18:428–45.
[81] Zhang X, Johnston ER, Wang Y, Yu Q, Tian D, Wang Z, et al. Distinct drivers of core and accessory components of soil microbial community functional diversity under environmental changes. mSystems 2019;4:e00374–19.
[82] Assefa S, Kohler G. Intestinal microbiome and metal toxicity. Curr Opin Toxicol 2020;19:21–7.
[83] Belkaid Y, Harrison OJ. Homeostatic immunity and the microbiota. Immunity 2017;46:562–76.
[84] Narendrula-Kotha R, Theriault G, Mehes-Smith M, Kalubi K, Nkongolo K. Metal toxicity and resistance in plants and microorganisms in terrestrial ecosystems. Rev Environ Contam Toxicol 2020;249:1–27.
[85] Xing C, Chen J, Zheng X, Chen L, Chen M, Wang L, et al. Functional metagenomic exploration identifies novel prokaryotic copper resistance genes from the soil microbiome. Metallomics 2020;12:387–95.
[86] Chaban B, Hughes HV, Beeby M. The flagellum in bacterial pathogens: for motility and a whole lot more. Semin Cell Dev Biol 2015;46:91–103.
[87] Arango-Argoty GA, Guron GKP, Garner E, Riquelme MV, Heath LS, Pruden A, et al. ARGminer: a web platform for the crowdsourcing-based curation of antibiotic resistance genes. Bioinformatics 2020;36:2966–73.
[88] Bortolaia V, Kaas RS, Ruppe E, Roberts MC, Schwarz S, Cattoir V, et al. ResFinder 4.0 for predictions of phenotypes from genotypes. J Antimicrob Chemother 2020;75:3491–500.
[89] Rowe W, Baker KS, Verner-Jeffreys D, Baker-Austin C, Ryan JJ, Maskell DJ, et al. Search engine for antimicrobial pesistance: a cloud compatible pipeline and web interface for rapidly detecting antimicrobial resistance genes directly from sequence data. PLoS One 2015;10:e0133492.
[90] Davis JJ, Wattam AR, Aziz RK, Brettin T, Butler R, Butler RM, et al. The PATRIC bioinformatics resource center: expanding data and analysis capabilities. Nucleic Acids Res 2020;48:D606–12.
[91] Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol 2019;37:953–61.
[92] Sugimoto Y, Camacho FR, Wang S, Chankhamjon P, Odabas A, Biswas A, et al. A metagenomic strategy for harnessing the chemical repertoire of the human microbiome. Science 2019;366:eaax9176.
[93] Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res 2020;48:D517–25.
[94] Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, Cohoon M, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 2005;33:5691–702.
[95] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res 2021;49:D412–9.
[96] Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-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:D309–14.

[97] Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 2007;23:1282–8.
[98] Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, et al. The MetaCyc database of metabolic pathways and enzymes. Nucleic Acids Res 2018;46:D633–9.
[99] Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2017;45:D353–61.
[100] Douglas GM, Maffei VJ, Zaneveld JR, Yurgel SN, Brown JR, Taylor CM, et al. PICRUSt2 for prediction of metagenome functions. Nat Biotechnol 2020;38:685–8.
[101] Kang C, Zhang Y, Zhu X, Liu K, Wang X, Chen M, et al. Healthy subjects differentially respond to dietary capsaicin correlating with specific gut enterotypes. J Clin Endocrinol Metab 2016;101:4681–9.
[102] Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature 2012;486:207–14.
[103] Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, et al. A core gut microbiome in obese and lean twins. Nature 2009;457:480–4.
[104] Han M, Yang K, Yang P, Zhong C, Chen C, Wang S, et al. Stratification of athletes' gut microbiota: the multifaceted hubs associated with dietary factors, physical characteristics and performance. Gut Microbes 2020;12:1–18.
[105] Costea PI, Hildebrand F, Manimozhiyan A, Backhed F, Blaser MJ, Bushman FD, et al. Enterotypes in the landscape of gut microbial community composition. Nat Microbiol 2018;3:8–16.
[106] Koren O, Knights D, Gonzalez A, Waldron L, Segata N, Knight R, et al. A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. PLoS Comput Biol 2013;9:e1002863.
[107] Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science 2011;334:105–8.
[108] Wang J, Linnenbrink M, Künzel S, Fernandes R, Nadeau MJ, Rosenstiel P, et al. Dietary history contributes to enterotype-like clustering and functional metagenomic content in the intestinal microbiome of wild mice. Proc Natl Acad Sci U S A 2014;111:E2703–10.
[109] Claesson MJ, Jeffery IB, Conde S, Power SE, O'Connor EM, Cusack S, et al. Gut microbiota composition correlates with diet and health in the elderly. Nature 2012;488:178–84.
[110] David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, et al. Diet rapidly and reproducibly alters the human gut microbiome. Nature 2014;505:559–63.
[111] Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H, Goodman AL, et al. The long-term stability of the human gut microbiota. Science 2013;341:1237439.
[112] Sonnenburg JL, Bäckhed F. Diet-microbiota interactions as moderators of human metabolism. Nature 2016;535:56–64.
[113] Moeller AH, Degnan PH, Pusey AE, Wilson ML, Hahn BH, Ochman H. Chimpanzees and humans harbor compositionally similar gut enterotypes. Nat Commun 2012;3:1179.
[114] Altomare A, Putignani L, Chierico FD, Cocca S, Angeletti S, Ciccozzi M, et al. Gut mucosal-associated microbiota better discloses inflammatory bowel disease differential patterns than faecal microbiota. Dig Liver Dis 2019;51:648–56.
[115] Poore GD, Kopylova E, Zhu Q, Carpenter C, Fraraccio S, Wandro S, et al. Microbiome analyses of blood and tissues suggest cancer diagnostic approach. Nature 2020;579:567–74.
[116] Levin D, Raab N, Pinto Y, Rothschild D, Zanir G, Godneva A, et al. Diversity and functional landscapes in the microbiota of animals in the wild. Science 2021;372:eabb5352.
[117] Sepich-Poore GD, Zitvogel L, Straussman R, Hasty J, Wargo JA, Knight R. The microbiome and human cancer. Science 2021;371:eabc4552.
[118] Kavagutti VS, Andrei AŞ, Mehrshad M, Salcher MM, Ghai R. Phage-centric ecological interactions in aquatic ecosystems revealed through ultra-deep metagenomics. Microbiome 2019;7:1–15.
[119] Woodcroft BJ, Singleton CM, Boyd JA, Evans PN, Emerson JB, Zayed AAF, et al. Genome-centric view of carbon processing in thawing permafrost. Nature 2018;560:49–54.
[120] Bogart E, Creswell R, Gerber GK. MITRE: inferring features from microbiota time-series data linked to host status. Genome Biol 2019;20:186.
[121] Li Y, Xu Z, Han W, Cao H, Umarov R, Yan A, et al. HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. Microbiome 2021;9:40.
[122] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet 2015;16:321–32.
[123] Zha Y, Chong H, Ning K. Microbiome sample comparison and search: from pair-wise calculations to model-based matching. Frontiers Microbiol 2021;12:642439.
[124] Sharma D, Xu W. phyLoSTM: a novel deep learning model on disease prediction from longitudinal microbiome data. Bioinformatics 2021;37:3707–14.
[125] Chen X, Liu L, Zhang W, Yang J, Wong KC. Human host status inference from temporal microbiome changes via recurrent neural networks. Brief Bioinform 2021;22:bbab223.
[126] Keegan KP, Glass EM, Meyer F. MG-RAST, a metagenomics service for analysis of microbial community structure and function. Methods Mol Biol 2016;1399:207–33.
[127] Wilke A, Bischof J, Harrison T, Brettin T, D'Souza M, Gerlach W, et al. A RESTful API for accessing microbial community data for MG-RAST. PLoS Comput Biol 2015;11:e1004008.
[128] Xi W, Gao Y, Cheng Z, Chen C, Han M, Yang P, et al. Using QC-blind for quality control and contamination screening of bacteria DNA sequencing data without reference genome. Front Microbiol 2019;10:1560.
[129] Vesselinov VV, Alexandrov BS, O'Malley D. Contaminant source identification using semi-supervised machine learning. J Contam Hydrol 2017;212:134–42.
[130] Moossavi S, Fehr K, Khafipour E, Azad MB. Repeatability and reproducibility assessment in a large-scale population-based microbiota study: case study on human milk microbiota. Microbiome 2021;9:41.
[131] Eisenhofer R, Minich JJ, Marotz C, Cooper A, Knight R, Weyrich LS. Contamination in low microbial biomass microbiome studies: issues and recommendations. Trends Microbiol 2019;27:105–17.
[132] Oliverio AM, Geisen S, Delgado-Baquerizo M, Maestre FT, Turner BL, Fierer N. The global-scale distributions of soil protists and their contributions to belowground systems. Sci Adv 2020;6:eaax8787.
[133] Lesker TR, Durairaj AC, Galvez EJC, Lagkouvardos I, Baines JF, Clavel T, et al. An integrated metagenome catalog reveals new insights into the murine gut microbiome. Cell Rep 2020;30:2909–22.
[134] 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;39:105–14.
[135] Li J, Jia H, Cai X, Zhong H, Feng Q, Sunagawa S, et al. An integrated catalog of reference genes in the human gut microbiome. Nat Biotechnol 2014;32:834–41.
[136] Blin K, Kim HU, Medema MH, Weber T. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. Brief Bioinform 2019;20:1103–13.
[137] Donia MS, Cimermancic P, Schulze CJ, Wieland Brown LC, Martin J, Mitreva M, et al. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell 2014;158:1402–14.

[138] Wilmanski T, Diener C, Rappaport N, Patwardhan S, Wiedrick J, Lapidus J, et al. Gut microbiome pattern reflects healthy ageing and predicts survival in humans. Nat Metab 2021;3:274–86.
[139] Henry R, Schang C, Coutts S, Kolotelo P, Prosser T, Crosbie N, et al. Into the deep: evaluation of SourceTracker for assessment of faecal contamination of coastal waters. Water Res 2016;93:242–53.
[140] Osterberg A, Graf W, Karlbom U, Påhlman L. Evaluation of a questionnaire in the assessment of patients with faecal incontinence and constipation. Scand J Gastroenterol 1996;31:575–80.
[141] Ihara Y, Takeshita T, Kageyama S, Matsumi R, Asakawa M, Shibata Y, et al. Identification of initial colonizing bacteria in dental plaques from young adults using full-length 16S rRNA gene sequencing. mSystems 2019;4:e00360–19.
[142] Gosalbes MJ, Durbán A, Pignatelli M, Abellan JJ, Jiménez-Hernández N, Pérez-Cobas AE, et al. Metatranscriptomic approach to analyze the functional human gut microbiota. PLoS One 2011;6:e17447.