

Chapter 10

Microbiome and Big-Data Mining



Kang Ning

Abstract Microbiome samples are accumulating at a very fast speed, representing microbial communities from every niche (biome) of our body as well as the environment. The fast-growing amount of microbiome samples, as well as the diversified sources from where the samples are collected, have provided us with an unprecedented scene from where we could obtain a better understanding of the microbial evolution and ecology. While all of these represent profound biological patterns and regulation principles, the understanding of them is heavily dependent on data integration and big-data mining, including the data-driven microbiome marker identification, non-linear relationship mining, dynamic pattern discovery, regulation principle discovery, etc.

In this chapter, we first introduce several terminologies in microbiome research, followed by the introduction of microbiome big-data. Then we emphasize the microbiome databases, as well as mainstream microbiome data mining techniques. We have provided several microbiome applications to showcase the power of microbiome big-data integration and mining for knowledge and clinical applications. Finally, we have summarized the current status of microbiome big-data analysis, pointed out several bottlenecks, and illustrated prospects in this research area.

Keywords Microbiome · Big-data · Evolution · Ecology · Database · Data mining · Knowledge discovery · Clinical applications

K. Ning (✉)

Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-Imaging, Center of AI Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China

e-mail: ningkang@hust.edu.cn

Microbiome samples are accumulating at a very fast speed, representing microbial communities from every niche (biome) of our body as well as the environment (Mitchell et al. 2020; Integrative HMP (iHMP) Research Network Consortium 2019; Thompson et al. 2017; Sunagawa et al. 2015). The fast-growing amount of microbiome samples, as well as the diversified sources from where the samples are collected, have provided us with an unprecedented scene from where we could obtain a better understanding of the microbial evolution and ecology (Mitchell et al. 2020; Segata et al. 2013; Integrative Human Microbiome Project 2019). While all of these represent profound biological patterns and regulation principles, the understanding of them is heavily dependent on data integration and big-data mining (Knight et al. 2018), including the data-driven microbiome marker identification (Segata et al. 2011), non-linear relationship mining (Surana and Kasper 2017), dynamic pattern discovery (Halfvarson et al. 2017; Ren et al. 2017; Bashan et al. 2016; Backhed et al. 2015; Liu et al. 2019), regulation principle discovery (Han et al. 2020), etc.

In this chapter, we will first introduce several terminologies in microbiome research, followed by the introduction of microbiome big-data. Then we will emphasize the microbiome databases, as well as mainstream microbiome data mining techniques. We will provide several microbiome applications to showcase the power of microbiome big-data integration and mining for knowledge and clinical applications. Finally, we will summarize the current status of microbiome big-data analysis, point out several bottlenecks, and illustrate prospects in this research area.

10.1 Microbial Communities, Metagenome, and Microbiome

As a ubiquitous and important organism in nature, microorganisms usually coexist in the form of a “microbial community” (Thompson et al. 2017; Sunagawa et al. 2015; Segata et al. 2013; Integrative HMP (iHMP) Research Network Consortium 2014). A microbial community usually contains dozens to thousands of different microorganisms, these species cooperate with each other to adapt to the changes in the environment, and their life activities also have a long-term and profound impact on the environment (Thompson et al. 2017; Integrative HMP (iHMP) Research Network Consortium 2014). With the deepening of human understanding of microorganisms, the basic research of microbial community and its application in the fields of health and environment have become increasingly important (Integrative Human Microbiome Project 2019; Biteen et al. 2016). The main research objects of microbiome include all the genetic materials of microbial communities, related environmental parameters and metabolites, as well as their complex relationships and dynamic changes.

In the microbiome research area, several terms need to be explained clearly, including microbiota, metagenome, and microbiome (Whiteside et al. 2015). A microbial community is a mixture of microbial species living, adapting, and evolving in a certain environment. Metagenome refers to the total genetic materials in the

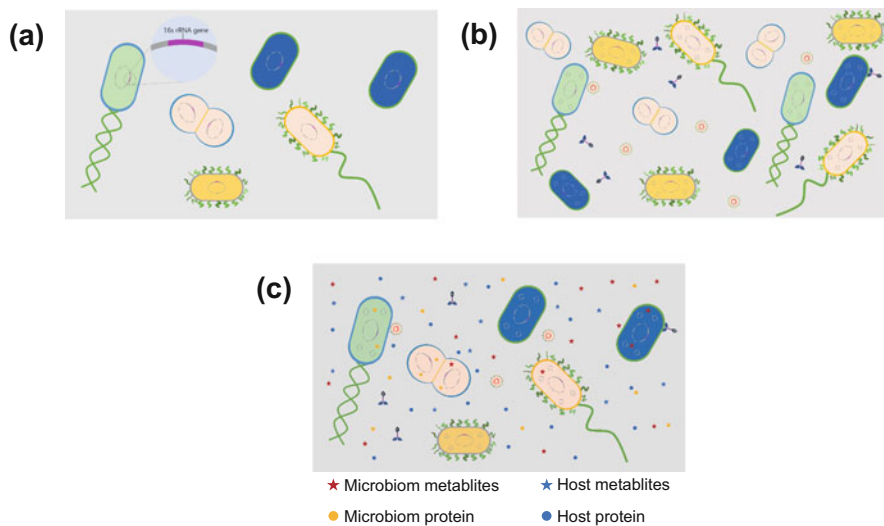


Fig. 10.1 The definitions of microbiota, metagenome, and microbiome. The same shape and color represent the same species, while different symbols represent different entities. (a) Microbiota: identification of all species in the microbial community using 16S rRNA sequencing. (b) Metagenome: all genetic materials in the microbial community. (c) Microbiome: all genetic materials, environmental factors, and metabolites in the microbial community

microbial community, while metagenome could be obtained by shotgun sequencing, many projects are still conducted by 16 s rRNA amplicon sequencing that could only quantitatively profile the species in the community. Microbiome refers to all genetic and non-genetic information contained in the microbial community, including metagenome, as well as all environmental factors and metabolites in the community. A brief illustration of the definitions and relationships of microbial communities, metagenome, and microbiome is provided in Fig. 10.1.

The microbiome research is mostly conducted by the omics approach (Mitchell et al. 2020; Segata et al. 2013). Firstly, samples are collected from niches, stored in a -20°C tube, before DNA extraction and amplification and sequencing. Then high-throughput sequencing is conducted, by means of 16S rRNA sequencing or metagenomic sequencing, and sequencing data are transferred for analysis (Knight et al. 2018).

10.1.1 The Differences Between 16S and Metagenomes

The sequencing principles: 16S rDNA contains nine hypervariable regions and ten conserved regions. A segment of hypervariable region sequence was amplified by PCR and sequenced. Metagenomic sequencing is similar to conventional DNA

library in that it randomly breaks microbial genomic DNA into small fragments and then inserts joints at both ends of the fragments for high-throughput sequencing.

Different fields of study: 16S rRNA sequencing mainly studies the species composition, the evolutionary relationship among species, and the diversity of communities. Besides, metagenomic sequencing can also be used for further research at the genetic and functional levels.

Different degree of species identification: Much of the 16S sequencing results are below species level, while metagenomic sequencing identifies microbes to species level and even to strain level.

The advantages and disadvantages of 16S rRNA and metagenomic sequencing methods for microbial community research have been summarized in (Knight et al. 2018), and we have provided key points in Table 10.1.

Table 10.1 Advantages and disadvantages of 16S rDNA and metagenomic sequencing methods for microbial community research

Method	Advantage	Disadvantage
Marker gene analysis	<ul style="list-style-type: none"> • Fast, simple, and inexpensive sample preparation and analysis • Closely related to genome content • Suitable for samples with low biomass • Could be compared with existing large public data sets 	<ul style="list-style-type: none"> • Affected by amplification bias • Selection of primers and variable regions will amplify the deviation • Usually need prior knowledge of the microbial community • Resolution is usually only to genus • Need for proper negative control • Limited functional information
Metagenomic analysis	<ul style="list-style-type: none"> • The relative abundance of microbial functional genes can be directly inferred • For known organisms, microbial classification and phylogenetic identity can be achieved at the species and strain level • It is not assumed to understand the microbial community • No biases associated with PCR • The in situ growth rate of target organisms with sequenced genomes can be estimated • It is possible to assemble a population-average microbial genome • Can be used for new gene families 	<ul style="list-style-type: none"> • Relatively expensive, laborious, and complicated sample preparation and analysis • The default pipeline usually does not annotate viruses and plasmids well • Due to assembly artifacts, population-average microbial genomes are often inaccurate

10.2 The Microbiome Research Is Heavily Dependent on Big-Data

As the number of microbiome samples easily exceeds tens of thousands in a medium-sized data collection (Mitchell et al. 2020), the efficiency and accuracy of sample comparison and search become a critical bottleneck (Knight et al. 2018), not to mention millions of samples from the rapidly diversified biomes from less than a hundred to more than three hundred in public databases (Fig. 10.2). The rapidly increasing number of samples from various niches on the planet has thus created a difficult huddle for knowledge discovery from these samples (Mitchell et al. 2020).

Microbiome research is heavily dependent on big-data, largely due to three reasons: (1) As traditional microbial research strategies could not identify the species in the community, current species identification and quantification is mostly done by sequencing techniques plus data analysis techniques. (2) As heterogeneous microbial community samples are collected from hundreds of different niches around the world, the comparison of these communities could only be performed using big-data mining techniques. (3) The mining of millions to trillions of functional genes from microbial communities is also a data-driven task nowadays.

Big-data technology and machine learning technology are very suitable for the organization, integration, and in-depth analysis of microbiome data (Li et al. 2019; Cheng et al. 2019; Tang et al. 2019; Microbiota meet big data 2014). First of all, microbiome data has all the 4 V characteristics of big data (Volume, Velocity, Variety, Veracity): large **Volume**, a large amount of data, including the amount of collection, storage, and calculation. The starting measurement unit of big data is at least p (1000 t), e (one million T), or Z (1 billion T). There are various types

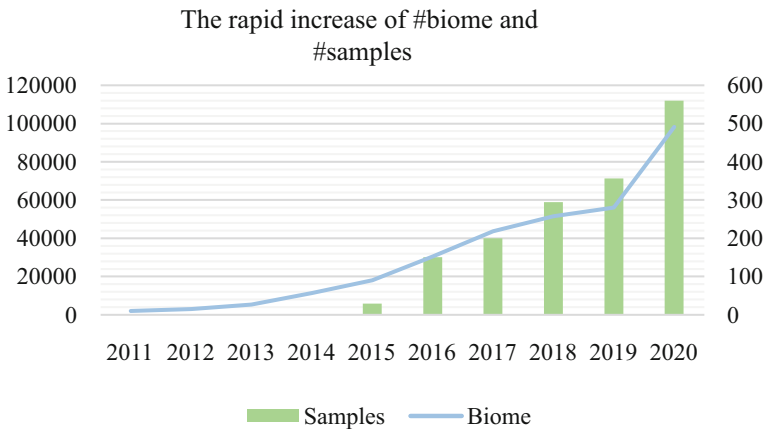


Fig. 10.2 The fast increasing number of microbiome samples, and the rapidly diversified biomes from where they are collected. Results are based on assessment of EBI MGnify database from year 2011 to year 2020

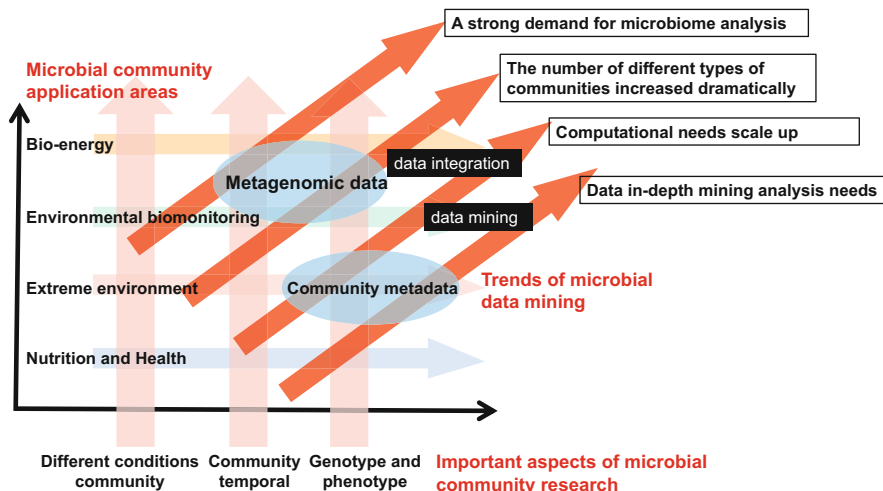


Fig. 10.3 The characteristics and urgent needs in multi-omics researchers

and sources (**Variety**). Including structured, semi-structured, and unstructured data, multi-types of data put forward higher requirements for data-processing ability. The **Value** density is low, and the data value density is relatively low. In other words, it is valuable to wash sand in waves. Information is massive, but the value density is low. How to mine the value of data through powerful machine algorithms is the most important problem to be solved in the era of big data. **Velocity**: this is a significant feature that big data is different from traditional data mining. Secondly, microbiome big data needs to be deeply mined: Data mining (DM) is an emerging interdisciplinary subject that gathers multiple disciplines. It is an extraordinary process, that is, the process of extracting unknown, implied, and potentially valuable information from huge data (Fig. 10.3).

10.3 Microbiome Data Integration and Databases

The development of microbiome research has profoundly boosted the data accumulation as well as the output of the researches. In the past 10 years, an exponential number of publications have been output (Fig. 10.4a), based on more than 100 TB per year of microbiome data accumulated (Fig. 10.4b).

Currently, there are already databases dedicated to microbiome researches (Table 10.2), including MG_RAST ((Meyer et al. 2008), <http://metagenomics.anl.gov/>), CAMERA ((Seshadri et al. 2007), <http://camera.calit2.net/>) as specialized databases, and NCBI SRA (<http://www.ncbi.nlm.nih.gov/sra>) as general databases. Among these databases, NCBI SRA (Kodama et al. 2012), MG-RAST (Meyer

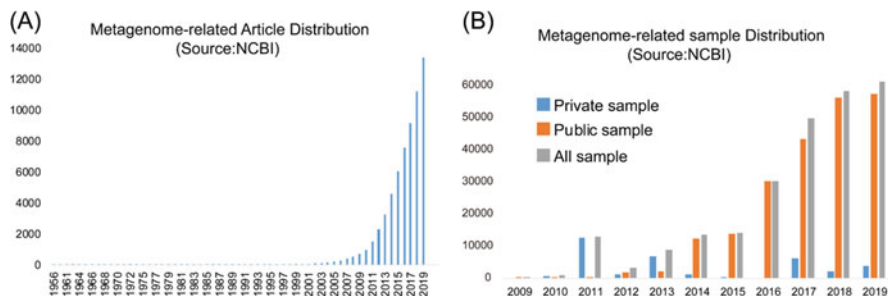


Fig. 10.4 In the past 10 or more years, (a) an exponential number of publications have been output, (b) based on more than 100 TB per year of microbiome data accumulated

Table 10.2 Commonly used microbiome databases (Zhang et al. 2017)

Database name	Description	Database website	References
NCBI SRA	General database that contains microbiome data of all kinds and formats	http://www.ncbi.nlm.nih.gov/sra	Kodama et al. (2012)
EBI MGnify	Specialized microbiome database with a comprehensive collection of samples, and with a unified analytical pipeline	www.ebi.ac.uk/metagenomics/	Mitchell et al. (2020)
MG-RAST	Specialized microbiome database with a unified analytical pipeline	Metagenomics.anl.gov	Paczian et al. (2019), Meyer et al. (2019)
IMG/M	Specialized microbiome database with a unified analytical pipeline	img.jgi.doe.gov	Markowitz et al. (2008)
Qiita	Specialized microbiome database with a unified analytical pipeline and comprehensive meta-data information	http://qiita.ucsd.edu/	Gonzalez et al. (2018)
CAMERA	Specialized microbiome database, data collection not comprehensive	http://camera.calit2.net/	Seshadri et al. (2007)

et al. 2008), and CAMERA2 (Seshadri et al. 2007) each has more than 10,000 microbiome projects, representing hundreds of thousands of samples and several TB of sequencing data.

However, the microbiome data in several major databases have not been well sorted out, whether in terms of the unification and integration of microbiome data format, or the matching environmental parameters (metadata). One of the key points is that the microbiome data has not been effectively classified and organized, resulting in a bottleneck for sample classification and comparison. Microbial community samples and relevant sequencing data are organized according to the biome ontology organization structure by hierarchical structures. For example: at the end of 2019, EBI MGnify contains sub-millions samples from 491 biomes (<https://www.ebi.ac.uk/metagenomics/biomes>) (Mitchell et al. 2020), in which the samples from human fecal have the exact biome position at “root > Host-associated > Human > Digestive system > Large intestine > Fecal.” This ontology structure is very beneficial to the classification of samples. However, the hierarchical organization structure of the current ontology is not completely tree-like, but has the feature that an ontology belongs to the direct sub-ontology of multiple ontologies. For example, “Fecal” has more than five upper level ontology information. Therefore, the relevant living environment ontology of each microbiome data is likely to have multi-label. On the one hand, the multi-label nature of microbiome data is not conducive to the simple classification of samples, resulting in the bottleneck of sample classification and comparison. On the other hand, the multi-tag attribute of microbiome data conforms to the characteristics of big-data research, and better results are expected to be obtained when processed by machine learning or deep learning.

10.4 Mainstream Microbiome Data Mining Techniques

As regard to microbiome data mining tools, current methods could be categorized according to their purposes (Table 10.3):

- 1. Identification of microbial species based on microbiome:** Based on the metagenome sequencing data, the species contained in the metagenome can be assigned to different taxonomic levels, such as phylum, class, order, family, genus, etc. At present, metagenome-based microbial species identification can be categorized into alignment-based and alignment-free sequence classification methods, both of which are based on the assumption that similar sequences originate from similar species. Sequence alignment identifies the species corresponding to the target genome sequence by comparing it with the existing database. Alignment-free sequence classification methods use the characteristics of the sequences themselves, such as GC content, codon usage frequency, etc., to classify them into the species corresponding to the most similar sequences. Typical examples of species identification methods include Megan (Huson et al. 2007), QIIME2 (Bolyen et al. 2019), etc. However, these methods are mostly limited to sequences of known classes and functions in databases (sequences in databases are mostly from model organisms or culturable microorganisms), so

the exact species of the majority of microorganisms in the microbial community remain largely unclear.

2. **Tools for microbial community structure decoding:** Tools for microbial community structure decoding and comparison include those for species composition analysis such as Phyloshop (Shah et al. 2011), Parallel-Meta (Su et al. 2012), MEGAN (Huson et al. 2007), etc., and those for microbial community comparison including UniFrac (Lozupone and Knight 2005) and Fast UniFrac (Hamady et al. 2010). However, these tools still have limitations: MEGAN (Huson et al. 2007) and STAMP (Parks and Beiko 2010) have provided an approach for microbial community sample comparison based on species composition, while such method is largely limited by the ignorance of evolutionary relationships among species (Hamady and Knight 2009). UniFrac (Lozupone and Knight 2005) and Fast UniFrac (Hamady et al. 2010) have taken phylogeny information into consideration, yet they could hardly handle thousands of samples due to large time cost. There is still a lack of efficient and accurate sample comparison and search methods, especially for model-based method.
3. **Microbial-based functional profiling and regulation model generation:** In terms of predicting the main functions of species, the current research is still in its infancy. Methods such as PICRUSt (Langille et al. 2013), based on 16S rRNA data, could analyze differences between samples by inferring the composition of functional genes in the samples. However, this prediction method cannot fully reflect the detailed functional composition and metabolic pathways of different species in a sample. Functional genes in microbial community analysis level, in view of the biosynthesis gene cluster (BGC) and antibiotic resistance gene cluster (ARG) gene functions such as group analysis, in addition to the typical antiSMASH (Medema et al. 2011) and NaPDoS (Ziemert et al. 2012) analysis platform and IMG-ABC (Hadjithomas et al. 2015), DoBISCUIT (Ichikawa et al. 2013), ClusterMine360 (Conway and Boddy 2013) database. Functional annotation and enrichment analysis of microbiome genes can deepen the understanding of microbial community functions and the analysis of key metabolic pathways and microbiome-host metabolic regulation mechanisms. However, the microbiome contains a large number of genes, and the functions of most genes are unknown.
4. **Microbial gene mining from metagenomics data:** At present the main database and the software including DoBISCUIT (Ichikawa et al. 2013) system (<http://www.bio.nite.go.jp/pks/>) based on manual selection of data, and the databases designed for specific types of metabolites, such as ClusterMine360 (Conway and Boddy 2013) database system, NaPDoS (Ziemert et al. 2012) analysis system (<http://napdos.ucsd.edu/>) for secondary metabolism genes, COBRA (Becker et al. 2007) for intestinal flora metabolism modeling analysis system, as well as antiSMASH (Medema et al. 2011) biosynthesis gene cluster (BGC) analysis system, etc. Relevant methods, however, largely depend on the reference sequence, known species in the microbial community species reference sequence under the condition of the lack of its completeness is not very ideal. The genes

Table 10.3 Representative analytical platforms for microbiome researches

Name	Description	Website	References
QIIME	Most frequently used package, with comprehensive sets of tools, discontinued in 2018	http://qiime.org	Caporaso et al. (2010)
QIIME 2	QIIME version 2, with a full set of command line and visualized interfaces for interactive and reproducible microbiome analysis	https://qiime2.org	Bolyen et al. (2019)
USEARCH	Fast sequence search and clustering toolset	http://www.drive5.com/usearch	Edgar (2010)
VSEARCH	Fast sequence search and clustering toolset specifically designed for metagenomics sequence analysis	https://github.com/torognes/vsearch	Rognes et al. (2016)
Trimmomatic	Quality control tool for metagenome sequences	http://www.usadellab.org/cms/index.php?page=trimmomatic	Bolger et al. (2014)
Bowtie2	Sequencing data alignment tool	http://bowtie-bio.sourceforge.net/bowtie2	Langmead and Salzberg (2012)
MetaPhlAn2	Microbial community structure profiling tool for k-mer based metagenomic sequence classification tool	http://segatalab.cibio.unitn.it/tools/metaphlan2	Truong et al. (2015)
Kraken2	Species-level functional profiling for microbial communities	https://ccb.jhu.edu/software/kraken2	Wood and Salzberg (2014)
HUMAnN2	Interactive microbial community profiling tool	http://www.huttnerhoyer.org/humann2	Franzosa et al. (2018)
MEGAN	Ultrafast metagenome assembly tool	https://www.wsi.uni-tuebingen.de/lehrestuehle/algorithms-in-bioinformatics/software/megan6/	Huson et al. (2007)
MEGAHIT	High-quality metagenome assembly tool	https://github.com/vouten/megahit	Li et al. (2015)
metaSPAdes	Tool for metagenome sequence assembly quality evaluation	http://cab.spbu.ru/software/spades	Nurk et al. (2017)
MetaQUAST	Gene prediction from metagenomics sequence	http://quast.sourceforge.net/metaquast	Mikheenko et al. (2016)
MetaGeneMark		http://exon-gatech.edu/GeneMark/meta_gmhmmmp.cgi	Zhu et al. (2010)

Prokka	Fast prokaryotic genome annotation tool	http://www.vicbioinformatics.com/software/prokka.shtml	Seemann (2014)
CD-HIT	Generation of non-redundant gene set	http://weizhongli-lab.org/cd-hit	Fu et al. (2012)
Salmon	k-mer based gene quantification tool	https://combine-lab.github.io/salmon	Patro et al. (2017)
MetaWRAP	Meta tool for metagenomics sequence binning	https://github.com/bxlab/metaWRAP	Uritskiy et al. (2018)
DAS tool	Another tool for metagenomics sequence binning	https://github.com/cmks/DAS_Tool	Sieber et al. (2018)
MOCAT2	A metagenomic assembly, annotation, and profiling framework	https://mocat.embl.de/index.html	Kultima et al. (2016)
ConStrains	Sub-species identification tool for microbial communities	https://bitbucket.org/tuo-chengwei/constrains/src	Luo et al. (2015)
MetaPhiAn	A high-resolution microbial community profiling tool based on metagenomics sequences	http://segatalab.cibio.unitn.it/tools/metaphlan/index.html	Truong et al. (2015)
PICRUSt	Functional profile prediction based on species composition profile of the microbial communities	http://picrust.github.io/picrust	Langille et al. (2013)
antiSMASH	Resource and analytical tool on secondary metabolite biosynthetic gene clusters	https://antismash-db.secondarymetabolites.org	Medema et al. (2011)
CARMA	Taxonomic classification of metagenomic sequences	https://www.cebitec.uni-bielefeld.de/webcarma.cebitec.uni-bielefeld.de/	Gerlach and Stoye (2011)
Sort-ITEMS	Metagenomic sequence analysis tool	http://metagenomics.atc.tcs.com/binning/SOrt-ITEMS	Monzoorul Haque et al. (2009)

(continued)

Table 10.3 (continued)

Name	Description	Website	References
PHYLOSHOP	Microbial community profiling tool	https://omics.informatics.indiana.edu/mg/phyloshop/	Shah et al. (2011)
UniFrac	Tool for microbial community species composition comparison	http://bmf.colorado.edu/unifrac	Lozupone and Knight (2005), Hamady et al. (2010)
PhyloPythia	Accurate phylogenetic classification analysis tool		McHardy et al. (2007)
MG-RAST	Analytical platform for microbiome research	https://www.mg-rast.org/	Meyer et al. (2008)
CAMERA	Tool set for metagenomics data analysis	http://camera.calit2.net/	Seshadri et al. (2007)
IBDsite	An integrated package for metagenomic sequence analysis for IBD	https://www.itb.cnr.it/ibd/	Merelli et al. (2012)

around the “environment–microbial community–metabolism” chain are largely unannotated, leaving large room for improvement.

5. **Microbiome data analysis platform:** There are currently several analytics platforms that cover the main steps of microbiome data analysis, such as QIIME (Caporaso et al. 2010), MG-RAST (Glass et al. 2010; Keegan et al. 2016), Camera (Seshadri et al. 2007), and EBI Metagenomics (now known as EBI Mgnify) (Mitchell et al. 2020). These sites often contain large datasets and data-processing platforms. At present, the biggest bottleneck in this regard is that the development of metagenomic data analysis platforms is far behind the rapid accumulation of metagenomic data. In particular, the integration analysis and deep mining of massive metagenomic data and other omics data are in urgent need.

10.5 Integration of Metagenome and Pan-Genome Towards Holistic Analysis of Microbial Communities

The microbiome data is mostly analyzed by the metagenome approach (Fig. 10.5). Metagenomics has been utilized for the studies of changes in community organization and microbial inhabitants, resulting in the discovery of a remarkable amount of genomic diversity and the characterization of new bacterial members (Integrative HMP (iHMP) Research Network Consortium 2014; Riesenfeld et al. 2004). A series of metagenome analysis tools, such as MEGAHIT (Li et al. 2015), MEGAN (Huson et al. 2007), and MetaPhlan2 (Truong et al. 2015) have been proposed allowing for metagenomics assembly, taxonomy, and functional analysis. The analyses of microbiome composition and function in different sites of human body including skin, oral, and gut show great differences in the microbial structure (Koren et al. 2011; Costello et al. 2009). For example, the taxonomic representation of bacteria on the human skin includes *Staphylococcus*, *Micrococcus*, and *Corynebacterium* (Fredricks 2001; Grice et al. 2009), while the dominant microorganisms in oral are *Streptococci*, *Lactobacillus*, and *Fusobacterium* (Dewhirst et al. 2010; Teng et al. 2015). In addition, the main components of microorganisms in the human gut are *Bacteroides* and *Prevotella* (Costea et al. 2018; Wu et al. 2011). These microbes in human body have coevolved with their hosts, which is also related to human health and disease (Costello et al. 2009; Clemente et al. 2012). The composition of microbes in different hosts varies greatly, and there are dynamic changes under different environmental factors (Costello et al. 2009). For example, Sonnenburg et al. revealed a seasonal cycle of gut microbiota corresponding to the enrichment of functions of the Hadza hunter-gatherers, especially *Bacteroides*, varies with the season, especially between the dry season and the wet season (Smits et al. 2017). Such studies revealed the succession of microbial community that changes with season in human gut. In addition, studies of microbial communities in natural environments such as soil (Daniel 2004), deep-sea (Mason et al. 2014), and wastewater (Guo et al. 2017) have uncovered hundreds of microbes, new genes,

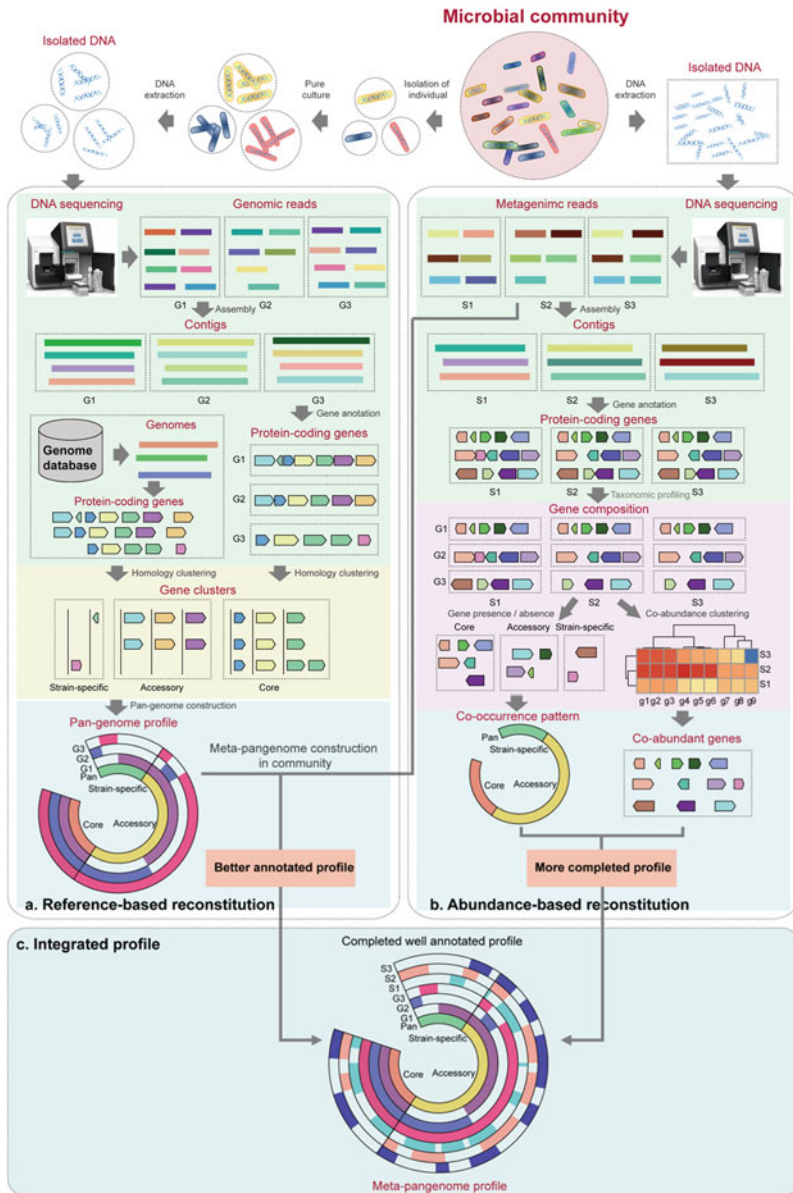


Fig. 10.5 Scheme of integrative pan-genome with metagenome studies on microbial community. (a) Using pan-genome of a set of genomes from isolates as a reference to recruit reads from metagenomes to quantify relative frequency of each gene sequence in community. (b) Binning co-abundant genes obtained from de novo assembly across metagenomic samples to reconstitute metagenomic species pan-genomes. Co-abundant with core or accessory genes of microbial species co-occurrence in samples and yield co-abundance. This figure was adapted from a previous published work [Integrating pan-genome with metagenome for microbial community profiling. Computational and Structural Biotechnology Journal, 2021, 19:1458–1466] with permission of authors

and uncharacterized metabolism, revealing an incredible microbial diversity and complexity.

10.6 Deep Learning Techniques for Microbiome Research

In recent years, more and more deep learning techniques have been developed for mining microbiome big-data (Li et al. 2019; Tang et al. 2019; Lan et al. 2018; Min et al. 2017; Wang and Gao 2019). These techniques essentially solved the functional gene mining, dynamic pattern discovery, and phenotype prediction problems.

1. **For sample comparison and search:** In microbial community source tracking, the traditional unsupervised learning method SourceTracker (Knights et al. 2011) and FEAST (Shenhav et al. 2019) could achieve very high accuracy when there are hundreds of samples and handful of biomes, while when the number of samples and biomes increase, the running time would increase very rapidly, preventing them from large-scale source tracking. This dilemma could be solved by deep learning solutions: by utilizing model-based methods such as neural network, both speed and accuracy could be achieved for the source tracking problem.
2. **For gene mining:** An example is ARG gene mining, for which traditional BLAST method could find the candidate ARG genes when they could match to those in the database. However, such an approach is limited to known ARG genes, and the search time could be short when faced with millions of candidates to be screened. Again, the deep learning approach has led to the model-based method that could mine novel ARG genes out of millions of candidates in an efficient manner.

All of these limitations have been calling for AI techniques that could discover more knowledge from microbiome dark matters. AI techniques are advantageous in generation of the models from a massive amount of samples, which are representative of the global profile of the context-dependent subjects (Kodama et al. 2012). AI techniques are therefore suitable for accurate and fast search when new samples (either a community, a gene, or a pattern) are searched against the models (Paczian et al. 2019; Markowitz et al. 2008; Daniel 2004). Therefore, AI techniques are especially suitable for microbiome dark matter mining, especially when facing the tradeoff between accuracy and efficiency.

The solutions for eliminating current methods' tradeoffs rely on deep learning approaches (Kodama et al. 2012; Paczian et al. 2019; Meyer et al. 2019; Markowitz et al. 2008; Gonzalez et al. 2018). First of all, model-based methods such as neural networks could be very fast for source tracking: once a rational model has been built, the source tracking could be very fast, and the source tracking accuracy could also be achieved, comparable with or even better than existing distance-based and unsupervised methods. The same approach is suitable for the gene mining problem.

For the spatial-temporal dynamic pattern mining, the deep learning method could also discover the intrinsic patterns out of the cross-section or longitudinal cohorts.

10.7 Representative Microbiome Applications

10.7.1 Case Study 1: Enterotype Analysis (Costea et al. 2018)

In 2011, three sequencing technologies (Illumina, 454, and Sanger) were used to sequence 16S rRNA genes in human fecal samples from three countries (Denmark, Spain, and the USA), and the result was that there were three enterotypes (Costea et al. 2018). The enterotypes were described as “a dense cluster of samples in a multidimensional space composed of communities” and were not affected by age, sex, cultural background, or geographical location. For each enterotype, an indicator/driver group was found at the center of the co-existing microbial network that was most profoundly associated with the enterotype. For example, enterotypes 1 can also be expressed as ET B, and *Bacteroides* is the best indicator group. Enterotype 2, which can also be expressed as ET P, is driven by *Prevotella* and its abundance is usually inversely proportional to the abundance of *Bacteroides*. Enterotype 3, which can also be expressed as ET F, is distinguished by the proportion of Firmicutes, among which the main group is *Ruminococcus*. All of the above analyses are based on the classification at the genus level, because the genus level can better reflect the ecological niche changes (Costea et al. 2018). Although some genera show functional heterogeneity, such as *Streptococci*, which contains both common symbiotic and lethal pathogens and groups that can be used for food fermentation, genera level analysis is generally reliable.

10.7.2 Case Study 2: Gene Mining (Qin et al. 2010)

10.7.2.1 Human Intestinal Microbiome Reference Gene Set

The authors describe the assembly and characterization of 3.3 million non-redundant microbial genes from fecal samples of 124 European individuals by Illumina-based metagenomic sequencing. This gene set is 150 times larger than the human gene complement, contains the vast majority of the (more common) microbial genes in the cohort, and probably includes the majority of the human gut microbial genes. These genes are shared to a large extent between individuals in this cohort. More than 99% of the genes were bacterial, suggesting that the entire cohort contained between 1000 and 1150 endemic bacterial species, with each individual containing at least 160 such species, and that they were also largely shared. The authors define and describe the minimum intestinal metagenome and the minimum

intestinal bacterial genome in terms of the functions of all individuals and most bacteria, respectively.

Most of the microbes that live in the gut have profound effects on human physiology and nutrition and are essential to human life. The content, diversity, and function of the gut microbiome are studied in order to understand and utilize the influence of gut microbiome on human health. Methods based on 16S ribosomal RNA gene (rRNA) sequences revealed that two families of bacteria, the Bacteroidaceae and the Antimicrobiaceae, make up more than 90% of the known phylogenetic categories and dominate the distal intestinal flora. Studies have also shown that there is great diversity in the gut microbiome between healthy individuals.

10.7.2.2 Metagenomic Sequencing of the Intestinal Microbiome

As part of the Metahit (Human Intestinal Genomics) project, the authors collected fecal samples from 124 healthy, overweight, and obese adult individuals and patients with inflammatory bowel disease (IBD) in Denmark and Spain. Total DNA was extracted from the fecal samples.

To generate an extensive catalogue of microbial genes from the human gut, the authors first assembled short Illumina readings into longer overlapping clusters, which could then be analyzed and annotated using standard methods. Using SoapDeNovo, the authors assembled all Illumina GA sequence data from scratch. Up to 42.7% of Illumina GA reads were assembled into a total of 6.58 million overlap groups, and nearly 35% of readings from any one sample could map to overlap groups from other samples, indicating the presence of a common sequence core.

To accomplish the overlapping group setup, the authors combined the unassembled reads from all 124 samples and repeated the de novo assembly process. Thus, about 400,000 overlapping groups with a length of 370 Mb and N50 939 bp are generated. Therefore, the total length of the author's final overlap group is 10.7 GB. Approximately 80% of the 576.7 Gb sequences of Illumina GA sequences were able to be compared with the overlap group at a 90% identity threshold to adapt to sequencing errors and strain variability in the gut, almost double the 42.7% of sequences. Soap de novo assembles them into overlapping clusters because the assembly uses more stringent criteria. This indicates that the author's overlap group represents the vast majority of Illumina sequences.

10.7.2.3 Genome Sets of the Human Intestinal Microbiome

To establish a non-redundant human gut microbiome genome, the authors first used the Metagene program to predict ORFs in overlapping populations and found 14,048,045 ORFs longer than 100 bp. They accounted for 86.7% of the overlap, comparable to the 86% found in fully sequenced genomes. Two-thirds of the ORFs appear to be incomplete, possibly due to the size of the author overlap group (N50 is

2.2 KB). Next, the authors remove the excess ORFs by pair comparison using very strict criteria that 95% conformance exceeds 90% of the shorter ORF length, which can be fused with direct homologues but can avoid dataset bloat due to possible sequencing errors.

The authors refer to the genes in the non-redundant set as “epidemic genes” because they are encoded on an overlapping group assembled from the richest read segments. The authors examined the number of prevalent genes found in all individuals, which is a function of the sequencing range and requires at least two gene calls to support reading. An estimate of coverage richness (ICE) based on incidence, determined by 100 people (the maximum number that can be accommodated by the Evaluations21 program), indicates that the authors’ catalog captured 85.3% of the prevalence genes. Although this may be an underestimate, it still suggests that the catalogue contains the vast majority of the prevalent genes in this cohort.

Each person carries $536,112 \pm 12,167$ of the prevalent genes, suggesting that most of the 3.3 million gene pools must be shared. But most of the prevalent genes were found in only a few individuals: 2,375,655 were found in less than 20% of individuals, and 294,110 were found in at least 50% of individuals (these “common” genes, as the authors call them). These values depend on the sampling depth. The sequencing of MH0006 and MH0012 revealed more catalogue genes, which were present in low abundance. Still, even at regular sampling depths, each person still has 204,0566 3603 common genes, suggesting that about 38% of an individual’s total gene pool is shared. Interestingly, patients with IBD carried, on average, 25% fewer genes than those without the disease, which is consistent with the observation that the former had less bacterial diversity than the latter.

10.7.3 Case Study 3: Plasticity of Intestinal Flora (Dynamic Pattern) (Liu et al. 2019)

First of all, at the macroscopic research level of the plasticity of the intestinal flora, the project team and the Capital Medical University have been monitoring the dynamics of the intestinal flora of the foreign aid medical team (volunteer team (VT)) for more than a year, aiming to study diet The influence of factor changes on the structure of human intestinal flora.

In this study, we recruited a team of 10 Chinese volunteers who set out from Beijing, stayed in Trinidad and Tobago (TAT) for 6 months and then returned to Beijing. A high-density longitudinal sampling strategy (average of 19 time points for VT members) was used to collect their stool samples (188 samples) and detailed dietary information. We divided the entire longitudinal study into six stages: when VT stays in TAT, T1 represents the pre-travel stage (20 samples), T2 (28 samples), T3 (60 samples), and T4 (21 samples) represents three time slots. After VT returned to Beijing, T5 (35 samples) and T6 (20 samples) sent two time slots, respectively.

At the same time, we also collected samples of Beijing healthy people (BJN, 57 samples), TAT healthy people (TTN, 28 samples), TAT patients (TTP, six samples), and TAT Chinese (TTC, eight samples) as samples control data set. Finally, we sequenced the V4 hypervariable region of the microbial 16S rRNA gene on 287 stool samples from 41 individuals and analyzed high-quality readings using QIIME (Caporaso et al., 2010).

We found that the microbial community in the intestine has two-way plasticity and elasticity during long-term stay and has a variety of dietary changes. First, BJN and TTN show different microbial community patterns (Fig. 10.6a). However, the microbial community of VT members changed from a microbial community similar to BJN to the TTN mode that accompanied them in TAT and returned to the original mode within 1 month after VT returned to Beijing (Fig. 10.6b–f). In addition, although we found that location and population have a great influence on the differentiation of samples (Fig. 10.6g, h), the dynamic changes of each member of VT show a specific trend (Fig. 10.6i, j), indicating that there may be the plasticity mode depending on the intestinal type among VT members. In addition, the relative abundance of *Sclerotium* and *Bacteroides* showed strong adaptability on the time axis and was negatively correlated on the time axis (Fig. 10.6k, l). Similarly, the relative abundance of *Proteus* and *Actinomycetes* also showed a plasticity pattern (Fig. 10.6m, n). By tracking and comparing at least 10% of the common operational taxonomic units (OTUs) shared by at least 10% of VT members, we found that Firmicutes, Bacteroidetes, Proteobacteria, and Actinobacteria have unique time dynamics during the long-term stay of VT (Fig. 10.6o–r).

10.7.4 Case Study 4: Athletes' Gut Microbiota (Han et al. 2020)

The gut microbiome of athletes and sedentary individuals differs in diversity and in certain taxa; however, it is unclear to what extent the patterns of the gut microbiome differ between the two and whether athletes' potential can be effectively monitored against the microbiome.

This study recruited a total of 306 fecal samples from 19 Chinese professional female rowers and divided them into three groups according to their daily performance: adult elite athletes (AE), young elite athletes (YE), and young non-elite athletes (YN). The differences of intestinal microbiome in different groups were compared to determine the correlation between intestinal microbiome and diet, physical characteristics and sports performance (Fig. 10.7).

Firstly, the intestinal flora of elite athletes and young non-elite athletes were stratified to find that the intestinal flora of elite athletes and young non-elite athletes had different intestinal types. In terms of taxonomic structure and functional composition, it was found that SCFA-producing bacteria were dominant in the microbial community of elite athletes. Secondly, functional analysis showed that

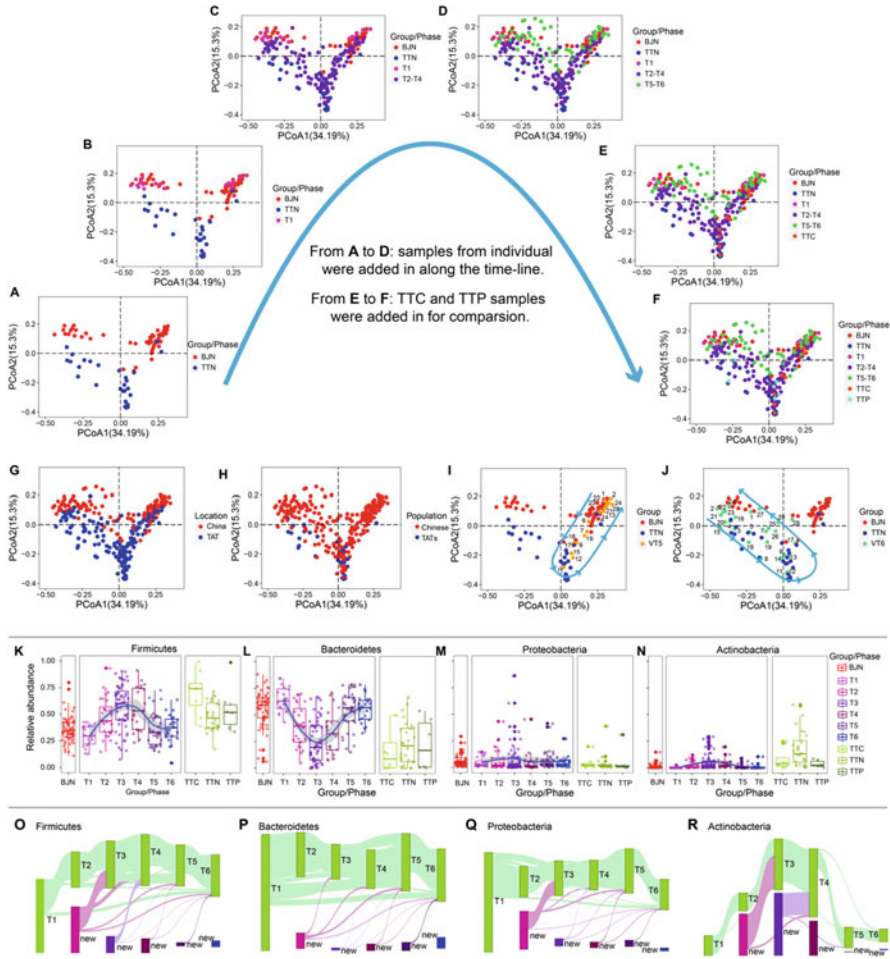


Fig. 10.6 Long-term human gut microbial community pattern and multiple dietary changes (Liu et al. 2019). (Reprinted with permission from authors of Liu et al. (2019))

ATP metabolism, multiple sugar transport systems, and carbohydrate metabolism were enriched in the microbial community of elite athletes. Furthermore, the construction of accurate classifiers based on a combination of taxonomy and functional biomarkers highlights the great potential of monitoring candidate elite athletes from a group of athletes. Finally, it was shown that intestinal flora is closely related to physical characteristics, dietary factors, and exercise-related characteristics. Importantly, the versatility of the athletes' microbiome, which may influence athlete performance by altering the gut microbiome, is associated with dietary factors (29%) and physical characteristics (21%). These findings highlight the complex

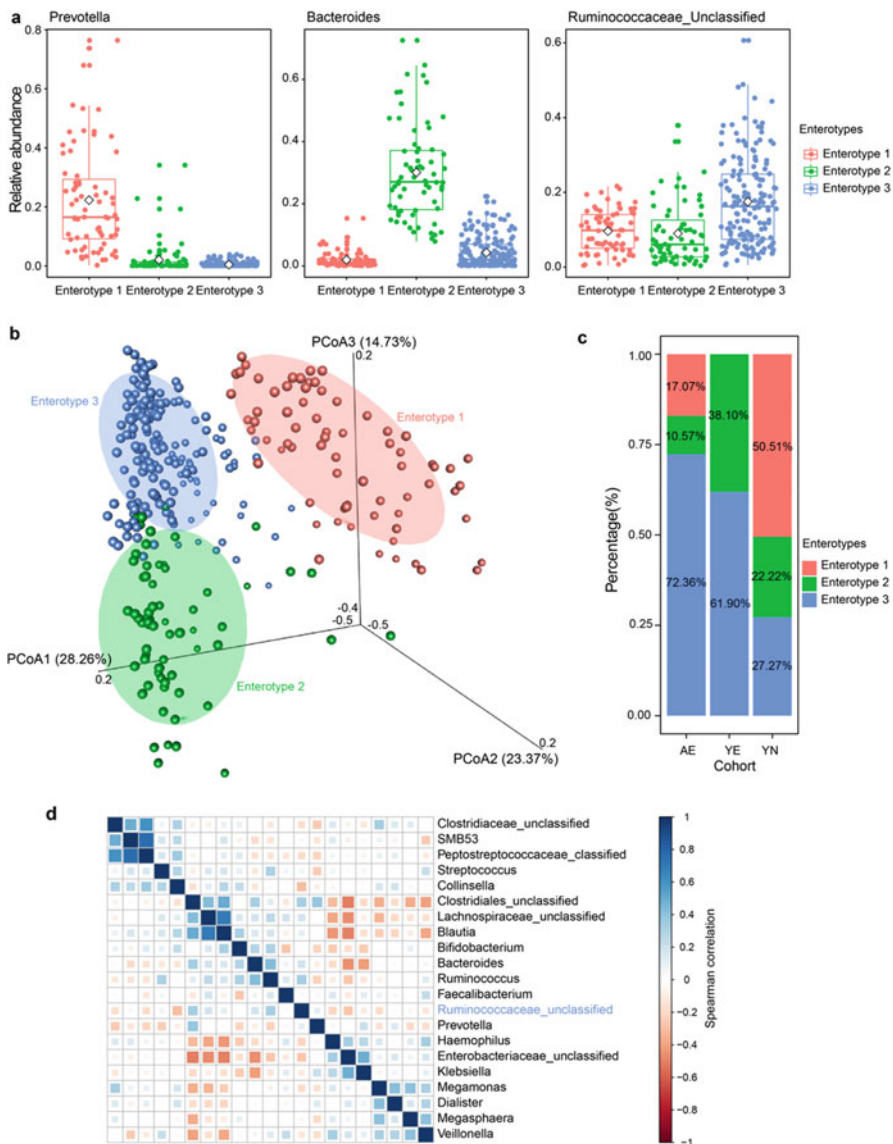


Fig. 10.7 Gut enterotypes in elite and youth non-elite athletes. A total of 306 samples are stratified into three enterotypes. The major contributor in the three enterotypes is *Prevotella*, *Bacteroides*, and Ruminococcaceae_unclassified, respectively. (a) Relative abundances of the top genera (*Prevotella*, *Bacteroides*, and Ruminococcaceae_unclassified) in each enterotype. (b) Three enterotypes were visualized by PCoA of Jensen-Shannon distance at the genus level. (c) The proportion of AE, YE, and YN samples distributed in three enterotypes. 72.3% AE, 61.9% YE, and 27.27% YN samples are found in enterotype 3. (d) Co-occurrence patterns among the dominant genera (average relative abundance >0.01%) across the samples from enterotype 3, as determined by the Spearman correlation analysis. (Reprinted with permission from authors of Han et al. (2020))

interplay of gut flora, dietary factors, and athletes' physical characteristics and performance, with gut flora as a key factor (Han et al. 2020).

10.8 Microbiome Research: Current Status, Bottlenecks, and Prospects

Today, microbiome research is, from many facets, a data-driven science. Firstly, the sequencing techniques have advanced quickly, thus enabling the fast and batch profiling of millions of microbial community samples. Secondly, data mining techniques have also advanced quickly, thus enabling the batch discovery of functional genes, dynamic patterns, as well as prediction of phenotype with high accuracy and fidelity. Thirdly, although data-driven, many discoveries are later verified by we-lab experiments, such as several probiotics (Whiteside et al. 2015; Routy et al. 2018), verified the power and validity of these data-driven approaches.

However, several bottlenecks remain for the microbiome big-data mining researches. One of the most critical bottlenecks is the big-data integration bottleneck (Integrative Human Microbiome Project 2019), and another is the lack of AI techniques for deep mining of important species, functional genes, and community dynamic patterns from a large amount of microbiome data (Microbiota meet big data 2014).

Despite these bottlenecks, microbiome researches are on the sharp rise, and many problems are on the edge of solution, while many more new frontiers are on the horizon. It is foreseeable that with several millions of samples from thousands of niches that have been collected, sequenced, and analyzed, a much better understanding of the microbial community ecology and evolution patterns would be discovered, together with hundreds of clinical or environmental applications made possible.

10.8.1 Microbiome Research as Part of a Multi-Omics Exploration

The multi-omics studies will continue to grow, in at least two directions: first, from multi-omics for single organisms or single species, to single-cell level omics studies, as well as to population and community level studies; second, the tight integration of multi-omics with data science as well as with clinical applications.

From the aspect of expanding the scope of multi-omics for single organisms or single species, single-cell level omics studies, as well as to population and community level studies, we have already seen rapid progress, largely due to the sequencing technical advances. From the aspect of integration of multi-omics with data science as well as with clinical applications, there are very hard challenges

still lying ahead. For example, it remains to be an open problem to determine the concordance of multi-omics along the timeline.

10.9 Summary

Taken together, it has become clear that microbiome research, which represents a rapidly growing omics research area, has already ensured enough high-quality data, as well as enabled data mining techniques, for large-scale microbiome data mining towards an in-depth understanding of microbial communities. The microbial community niches, species, functional genes and their dynamics, have constituted the microbial dark matter, which has been emerged as a grand challenge for microbiome research. The fast development of microbiome data mining would certainly boost the discovery of much more resources and regulation patterns out of these dark matters. And the integration of microbiome and other omics data would lead to a more complete picture of the dynamic patterns as well as regulation principles in the microbiome world.

References

- Backhed F et al (2015) Dynamics and stabilization of the human gut microbiome during the first year of life. *Cell Host Microbe* 17(6):852
- Bashan A et al (2016) Universality of human microbial dynamics. *Nature* 534(7606):259
- Becker SA et al (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. *Nat Protoc* 2(3):727–738
- Biteen JS et al (2016) Tools for the microbiome: nano and beyond. *ACS Nano* 10(1):6–37
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120
- Bolyen E et al (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* 37(8):852–857
- Caporaso JG et al (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335–336
- Cheng M, Cao L, Ning K (2019) Microbiome big-data mining and applications using single-cell technologies and metagenomics approaches toward precision medicine. *Front Genet* 10:972
- Clemente JC et al (2012) The impact of the gut microbiota on human health: an integrative view. *Cell* 148(6):1258–1270
- Conway KR, Boddy CN (2013) ClusterMine360: a database of microbial PKS/NRPS biosynthesis. *Nucleic Acids Res* 41(Database issue):D402–D407
- Costea PI et al (2018) Enterotypes in the landscape of gut microbial community composition. *Nat Microbiol* 3(1):8–16
- Costello EK et al (2009) Bacterial community variation in human body habitats across space and time. *Science* 326(5960):1694–1697
- Daniel R (2004) The soil metagenome – a rich resource for the discovery of novel natural products. *Curr Opin Biotechnol* 15(3):199–204
- Dewhirst FE et al (2010) The human oral microbiome. *J Bacteriol* 192(19):5002–5017

- Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460–2461
- Franzosa EA et al (2018) Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods* 15(11):962–968
- Fredricks DN (2001) Microbial ecology of human skin in health and disease. *J Investig Dermatol Symp Proc* 6(3):167–169
- Fu L et al (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23):3150–3152
- Gerlach W, Stoye J (2011) Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Res* 39(14):e91
- Glass EM et al (2010) Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc* 2010(1):pdb.prot5368
- Gonzalez A et al (2018) Qiita: rapid, web-enabled microbiome meta-analysis. *Nat Methods* 15(10):796–798
- Grice EA et al (2009) Topographical and temporal diversity of the human skin microbiome. *Science* 324(5931):1190–1192
- Guo J et al (2017) Metagenomic analysis reveals wastewater treatment plants as hotspots of antibiotic resistance genes and mobile genetic elements. *Water Res* 123:468–478
- Hadjithomas M et al (2015) IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *MBio* 6(4):e00932
- Halfvarson J et al (2017) Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol* 2:17004
- Hamady M, Knight R (2009) Microbial community profiling for human microbiome projects: tools, techniques, and challenges. *Genome Res* 19(7):1141–1152
- Hamady M, Lozupone C, Knight R (2010) Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J* 4(1):17–27
- Han M et al (2020) Stratification of athletes' gut microbiota: the multifaceted hubs associated with dietary factors, physical characteristics and performance. *Gut Microbes* 12(1):1–18
- Huson DH et al (2007) MEGAN analysis of metagenomic data. *Genome Res* 17(3):377–386
- Ichikawa N et al (2013) DoBISCUIT: a database of secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res* 41(Database issue):D408–D414
- Integrative HMP (iHMP) Research Network Consortium (2014) The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* 16(3):276–289
- Integrative HMP (iHMP) Research Network Consortium (2019) The Integrative Human Microbiome Project. *Nature* 569(7758):641–648
- (2019) After the Integrative Human Microbiome Project, what's next for the microbiome community? *Nature* 569(7758):599
- Keegan KP, Glass EM, Meyer F (2016) MG-RAST, a metagenomics service for analysis of microbial community structure and function. *Methods Mol Biol* 1399:207–233
- Knight R et al (2018) Best practices for analysing microbiomes. *Nat Rev Microbiol* 16(7):410–422
- Knights D et al (2011) Bayesian community-wide culture-independent microbial source tracking. *Nat Methods* 8(9):761–763
- Kodama Y et al (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res* 40(Database issue):D54–D56
- Koren O et al (2011) Human oral, gut, and plaque microbiota in patients with atherosclerosis. *Proc Natl Acad Sci U S A* 108(suppl 1):4592–4598
- Kultima JR et al (2016) MOCAT2: a metagenomic assembly, annotation and profiling framework. *Bioinformatics* 32(16):2520–2523
- Lan K et al (2018) A survey of data mining and deep learning in bioinformatics. *J Med Syst* 42(8):139
- Langille MG et al (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 31(9):814–821

- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359
- Li D et al (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 31(10):1674–1676
- Li Y et al (2019) Deep learning in bioinformatics: introduction, application, and perspective in the big data era. *Methods* 166:4–21
- Liu H et al (2019) Resilience of human gut microbial communities for the long stay with multiple dietary shifts. *Gut* 68(12):2254–2255
- Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71(12):8228–8235
- Luo C et al (2015) ConStrains identifies microbial strains in metagenomic datasets. *Nat Biotechnol* 33(10):1045–1052
- Markowitz VM et al (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res* 36(Database issue):D534–D538
- Mason OU et al (2014) Metagenomics reveals sediment microbial community response to deepwater horizon oil spill. *ISME J* 8(7):1464–1475
- McHardy AC et al (2007) Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods* 4(1):63–72
- Medema MH et al (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res* 39(Web Server issue):W339–W346
- Merelli I, Viti F, Milanesi L (2012) IBDsite: a galaxy-interacting, integrative database for supporting inflammatory bowel disease high throughput data analysis. *BMC Bioinformatics* 13(suppl 14):S5
- Meyer F et al (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386
- Meyer F et al (2019) MG-RAST version 4-lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. *Brief Bioinform* 20(4):1151–1159
- (2014) Microbiota meet big data. *Nat Chem Biol* 10(8):605
- Mikheenko A, Saveliev V, Gurevich A (2016) MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics* 32(7):1088–1090
- Min S, Lee B, Yoon S (2017) Deep learning in bioinformatics. *Brief Bioinform* 18(5):851–869
- Mitchell AL et al (2020) MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res* 48(D1):D570–D578
- Monzoorul Haque M et al (2009) SOrt-ITEMS: sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. *Bioinformatics* 25(14):1722–1730
- Nurk S et al (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 27(5):824–834
- Paccian T et al (2019) The MG-RAST API explorer: an on-ramp for RESTful query composition. *BMC Bioinformatics* 20(1):561
- Parks DH, Beiko RG (2010) Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* 26(6):715–721
- Patro R et al (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* 14(4):417–419
- Qin J et al (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464(7285):59–65
- Ren T et al (2017) Seasonal, spatial, and maternal effects on gut microbiome in wild red squirrels. *Microbiome* 5(1):163
- Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* 38:525–552
- Rognes T et al (2016) VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584
- Routy B et al (2018) Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science* 359(6371):91–97

- Seemann T (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30(14):2068–2069
- Segata N et al (2011) Metagenomic biomarker discovery and explanation. *Genome Biol* 12(6):R60
- Segata N et al (2013) Computational meta'omics for microbial community studies. *Mol Syst Biol* 9:666
- Seshadri R et al (2007) CAMERA: a community resource for metagenomics. *PLoS Biol* 5(3):e75
- Shah N et al (2011) Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics. *Pac Symp Biocomput*:165–176
- Shenhav L et al (2019) FEAST: fast expectation-maximization for microbial source tracking. *Nat Methods* 16(7):627–632
- Sieber CMK et al (2018) Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol* 3(7):836–843
- Smits SA et al (2017) Seasonal cycling in the gut microbiome of the Hadza hunter-gatherers of Tanzania. *Science* 357(6353):802–806
- Su X, Xu J, Ning K (2012) Parallel-META: efficient metagenomic data analysis based on high-performance computation. *BMC Syst Biol* 6(Suppl 1):S16
- Sunagawa S et al (2015) Ocean plankton. Structure and function of the global ocean microbiome. *Science* 348(6237):1261359
- Surana NK, Kasper DL (2017) Moving beyond microbiome-wide associations to causal microbe identification. *Nature* 552(7684):244–247
- Tang B et al (2019) Recent advances of deep learning in bioinformatics and computational biology. *Front Genet* 10:214
- Teng F et al (2015) Prediction of early childhood caries via spatial-temporal variations of oral microbiota. *Cell Host Microbe* 18(3):296–306
- Thompson LR et al (2017) A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* 551(7681):457–463
- Truong DT et al (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 12(10):902–903
- Uritskiy GV, DiRuggiero J, Taylor J (2018) MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* 6(1):158
- Wang W, Gao X (2019) Deep learning in bioinformatics. *Methods* 166:1–3
- Whiteside SA et al (2015) The microbiome of the urinary tract—a role beyond infection. *Nat Rev Urol* 12(2):81–90
- Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 15(3):R46
- Wu GD et al (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334(6052):105–108
- Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* 38(12):e132
- Ziemert N et al (2012) The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One* 7(3):e34064
- Zhang G et al (2017) Development of Comprehensive Microbiome Big Data Warehouse/Center for Long-term Scientific Impact[J]. *Bulletin of Chinese Academy of Sciences* 32(3):280–289