Establishing a Microbiome Big Data Center for Long-term Scientific Impact (Postprint)

Authors: Zhang Guoqing, Ning Kang, Zhi Xiaoyang, Liu Wan, Xu Ping, Zhou Haokui, Hu Qiannan, Zhao Guoping

Date: 2017-03-22T00:00:00+00:00

Abstract

The concepts and technologies of metagenomic research have propelled the rise of microbiome science, accumulating extensive microbial genomic data and microbiome data related to health, animals, plants, and the environment, and establishing databases, standardization methods, and analytical tools of considerable scale and influence. Most platforms focus on providing data support for projects or specific types of microbial communities, making it difficult to meet the demands of more in-depth and comprehensive microbial biology research. This article proposes adopting the concepts of a microbial phylogenome that comprehensively focuses on the aggregate of microbial taxonomic units and a microbiome that focuses on the aggregate of microbial populations in specific ecological niches, to construct a comprehensive microbiome data warehouse that integrates microbial taxonomy, evolution, ecology, and related 'omics' data and information. On this basis, further integration of data from basic life sciences research and systems synthetic biology research would support the research and development of comprehensive reference databases with high-level quality control, standardized assembly and annotation, and first-class methods for data deposition, search and sharing, deep learning, and analytical mining. Consequently, this will further integrate metadata and data from large-scale microbiome projects, forming a microbiome big data center with comprehensive and integrated data, safe and efficient management, and complete service functions.

Full Text

Development of Comprehensive Microbiome Big Data Warehouse/Center for Long-term Scientific Impact

Zhang Guoqing[1], Ning Kang[2], Zhi Xiaoyang[3], Liu Wan , Xu Ping , Zhou Haokui , Hu Qiannan[1], Zhao Guoping[1]

[1]Bio-Med Big Data Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[2]School of Life Science and Technology, Huazhong University of Science & Technology, Wuhan 430074, China
[3]Shanghai Center for Bioinformation Technology, Shanghai 201203, China
Shanghai Information Center for Life Sciences, Chinese Academy of Sciences, Shanghai 200031, China
Center for Synthetic Biology Engineering Research, Shenzhen Institutes of Advanced Technology, Shenzhen 518055, China

Abstract

The concept and technology of metagenomics have catalyzed the rise of microbiome research, generating vast amounts of microbial genome data related to human health, animals, plants, and various environments. This has led to the establishment of databases, standardized methods, and analytical tools with considerable scale and influence. However, most existing platforms focus on providing data support for specific projects or particular types of microbial communities, making it difficult to meet the needs of more comprehensive and in-depth microbiological research. This paper proposes integrating two conceptual frameworks: the "microbiophylome," which encompasses the entirety of microbial taxonomic units, and the "microbiome," which focuses on microbial populations within specific ecological niches. We recommend constructing a comprehensive microbiome data warehouse that integrates microbial classification, evolution, ecology, and related "omics" data and information. Building upon this foundation, the warehouse should further incorporate data from basic life sciences research and systems/synthetic biology to support the development of comprehensively quality-controlled reference databases, standardized assembly and annotation protocols, and state-of-the-art methods for data deposition, search and sharing, deep learning, and analytical mining. This initiative will also facilitate the integration of metadata and data from large-scale microbiome projects, ultimately forming a microbiome big data center characterized by comprehensive data integration, secure and efficient management, and complete service functionalities.

Keywords: microbiome, microbiophylome, classification, ecology, synthetic biology

Since the establishment of the International Metagenomics Consortium in October 2005, multiple countries have launched human microbiome research initiatives, including the Human Microbiome Project (HMP) and its successor iHMP in the United States [1,2], the EU' s MetaHIT and its follow-up project MetaGenoPolis, the standards project IHMS [4], and microbiome diversity programs in South Korea. Additionally, numerous international research projects have been conducted focusing on human health, disease, environmental monitoring, and remediation. These initiatives have driven the development of specialized databases, reference databases, standardization and quality control protocols,

and data analysis and mining tools.

Currently, most international microbiome data platforms center on reference data catalogs and metagenomic datasets, providing basic analytical services such as functional annotation and community species structure analysis for microbiome studies in specific environments (primarily project-oriented). Consequently, these platforms essentially serve as research platforms for the "molecular ecology" and metagenomics of particular environments. For other microbiological research areas—such as microbial systematics, comprehensive ecological analysis, reference database construction, and cell factory design—they offer only low-level support like data downloads and lack direct online computational environments or algorithm testing platforms. Furthermore, some platforms cover only specific projects, resulting in insufficient data coverage and integration, forcing researchers to spend considerable time discovering, acquiring, and integrating data, information, and knowledge resources, which hinders straightforward and convenient research.

Therefore, we argue that the current general concept of "microbiome" —defined as the collection of all members of a multi-species microbial community (microbiota) and their genetic information (primarily meta/megagenome) and life functions within a specific environment (ecological niche or biotype)—needs to be connected to a broader concept of "microbiophylome." The microbiophylome represents the collection of genetic information (mainly various "omics" data) and related biological structure-function information for all microbial individuals and various microbial populations. By integrating data under both concepts and constructing a comprehensive microbiome data warehouse with global search capabilities, we can achieve high-quality control through continuously improved assembly, annotation technologies, and research service platforms. This will enable secure and effective deposition, sharing, and mining of high-throughput data.

The proposed microbiome data warehouse should encompass three categories of data: taxonomy, evolution, and ecology, serving as a management and sharing platform for these major data types. This foundation is essential for establishing high-quality data standards and quality control procedures for long-term research development. Based on these standards and procedures, the warehouse will systematically collect metadata from large-scale international projects and integrate raw data as needed, forming a microbiome big data center with complete data integration, secure and efficient management, and comprehensive service functions. This big data center will serve as a fundamental platform for all microbiology research, including microbiome projects, providing tool systems such as the Microbiome Systematics Toolkit, Microbiome System Ecology Toolkit, and Microbiome System Synthetic Biology Toolkit (Figure 1 [Figure 1: see original paper]).

1. Microbiome System Taxonomy

Microbial systematics represents one of the theoretical foundations of microbiome research, having evolved through morphological, chemical, and molecular classification stages to form the polyphasic taxonomy technical and theoretical system [13]. While 16S rRNA phylogeny provides the foundational framework for current classification systems, its low resolution at lower taxonomic levels has long been criticized by the academic community. New-generation high-throughput sequencing technologies are profoundly impacting microbial systematics, and integrating genomic data into polyphasic classification represents the future trend [14]. However, the high structural and genetic diversity of microbial genomes, along with horizontal gene transfer, makes evolutionary history reconstruction extremely difficult and complex [15]. Incorporating genomes into the polyphasic classification system remains an ongoing process [16].

Marker molecules such as 16S rRNA have enabled the development of comprehensive feature sequence databases like RDP [17], Greengenes [18], and Silva [19]. However, numerous unculturable microorganisms exist in nature, and substantial sequencing data still cannot be assigned taxonomic status. Understanding target microbial functions still relies on reference data from pure-culture microbial genomes. To some extent, the theoretical bottleneck in microbiome research lies in microbial isolation and pure-culture techniques and a comprehensive classification system. Data from numerous microbiome projects constitute reference databases for meta-analysis, providing researchers with background information to derive new conclusions through comparative studies.

The 16S rRNA gene has become the gold standard for modern molecular classification due not only to its functional conservation and moderate evolutionary rate but also, more importantly, to decades of accumulated database resources —most currently identified microbial species have corresponding gene sequence information [17-20]. Applying genome data in taxonomy also requires a data platform for storage and analysis. Existing international genome sub-databases contain substantial redundant data, hindering comparative genome analysis and lacking taxonomic data support. Therefore, an integrated database platform is urgently needed to effectively describe species genomes and taxonomic identification phenotype data in taxonomy, while providing data analysis pipelines. Through unified optimization standards and analysis workflows, each genome sequence should undergo gene prediction, functional annotation, and metabolic network reconstruction to support in-depth genome information mining. Integrating taxonomic phenotype data will facilitate researchers' queries and comparative analyses of different species' characteristics, enabling comparative genomics studies on the genetic mechanisms of phenotypes.

2. Microbiome System Ecology

Metagenomics is a representative method widely used in current research on unculturable microorganisms, with full-length DNA sequencing and 16S rDNA

being two typical technical approaches. The combination of these two methods has rapidly advanced our understanding of the structure and function of important microbial communities [21,22]. For data generated by these two methods, bioinformatics analyses can be performed in quality control, sequence classification and functional partitioning, and integrated methods (Table 2).

(1) Metagenomic sequence quality control involves a series of short-read quality analyses and filtering. The most critical quality control steps include short-read quality analysis, trimming, and removal of chimeric sequences. Quality control can be implemented using software packages such as Mothur [23] and QC-Chain [24,25].

(2) Based on quality differences and sequence lengths, sequencing data can be classified into different precision levels of "phylum," "class," "order," "family," and "genus." Sequence classification can be divided into similarity-based analysis and composition-based analysis. Similarity-based analysis is limited by known classified and functional sequences, with over 90% of microbial community sequencing data unable to be classified through this method. Composition-based methods rely on specific features such as GC content and coding region proportion, comparing feature vectors in gene groups to determine sequence classification [26]. MEGAN [27] utilizes the NCBI taxonomy database to display the evolutionary positions and taxonomic composition of components in single or multiple metagenomes.

(3) Research objects in metagenomics can be divided into community species structure and community functional structure. In recent years, analysis techniques based on 16S rRNA biomarkers, such as MOTHUR [23], QIIME [28], and Parallel-META [29], have expanded research scope on microbial community structure. However, their high conservation and multicopy nature also limit their applications. In functional structure, the basic research strategies of metagenomics include large-fragment DNA assembly, gene prediction, gene annotation, and metabolic pathway analysis. Additionally, considering the characteristics of different microbial communities (based on community metadata), all data can be divided into two or more classes to identify and characterize community biomarkers and conduct mining of community functional feature markers based on whole-genome sequencing data of microbial communities.

(4) Mining based on single data types increasingly fails to meet the needs of microbial community research. Metabolomics and single-cell data are gradually being integrated with metagenomic data. In community metabolomics, small molecule metabolites, NMR standard spectra, standard mass spectra, and physicochemical information of various metabolites are being developed, along with related data analysis methods. In single-cell data analysis, the main components include single-cell genomics, transcriptomics, and characterization signals. These data provide multi-angle support for microbial community research from global and individual, genetic and epigenetic, structural and functional

perspectives.

From the perspective of tool research related to microbiome system ecology represented by metagenomics, beyond algorithms and performance, the scope and quality of datasets behind tools significantly affect accuracy. Constructing a unified data warehouse to compile complete feature sequences, reference genomes, functional genomes, and other microbiophylome data, as well as typical microbial ecological community metagenomes and metabolomes, will form a three-dimensional and complete microbiome data matrix. This will enable research on metagenome assembly, annotation, population interactions and networks, and microbial ecological comparisons, forming microbiome system ecology tools and developing specialized microbial ecological data platforms for ecological microbiomes, microbiome physiology and metabolism, evolutionary microbiomes, and chemical microbiomes.

3. Microbiome System Synthetic Biology

Mining natural product biosynthetic gene clusters from metagenomic data through high-throughput sequencing is conducted through three main approaches: (1) annotation of gene clusters for single compound classes, such as PRISM [30] and GRAPE [31]; (2) annotation of gene clusters for single species, such as StreptomeDB [32]; and (3) annotation of gene clusters across multiple species and compounds, such as antiSMASH [33]. Gene cluster functional annotation at the gene scale represents only a portion of gene data. Genome-scale mining can not only discover novel natural products but also identify related biosynthetic pathways, providing a data foundation for synthetic biology research.

Multiple research groups have developed various databases for molecular structure biotransformation and catalytic elements related to microbial cell factories, including the biosynthetic reaction and catalytic element database BRENDA [34], the molecular structure transformation database Rhea [35], and KEGG molecular structure transformation data [36]. The Rxnfinder research group has developed a data-driven, one-stop technology system for designing novel biosynthetic reactions, enzymes, and pathways based on literature [37]. These databases contain rich synthetic biology resources from different levels and perspectives, but comprehensive biological databases for single microorganisms as research units have not yet been fully established.

To efficiently and rationally develop and design biosynthetic pathways for target compounds, multiple pathway design methods based on atomic matching or reaction rules have been developed [38-41]. However, few studies focus on chassis cells as research units. Microbial cell factory design has become a research focus in biosynthesis [42], generating multiple methods for designing target compound synthesis pathways based on chassis cells, such as FMM [43], PHT [44], and MRE [45]. With continuous development of genome sequencing technologies, another important research direction in simulating cell factory design in-

volves genome-scale vector metabolic network models and flux balance analysis (FBA) [46]-based model optimization methods for designing, optimizing, and evaluating yields of synthetic target compound pathways.

From the development process of synthetic biology, genome-scale synthetic design is becoming increasingly important. Relying on microbiome data platforms for metagenomes, functional genomes, transcriptomes, metabolomes, etc., we should establish a synthetic biology microbial resource library integrating natural products, regulatory catalysis and transport elements, reaction pathways and networks, and gene circuits and clusters. By studying chassis cells and their metabolic models at the whole-genome scale, we can discover new gene clusters and circuits that bridge high-noise, large-volume microbiome data with high usability, forming microbiome system synthetic biology tools.

4. Microbiome Health Big Data Applications

Metagenomic approaches have penetrated scientific fields aimed at utilizing or overcoming complex microbial communities and their products, including environmental biomonitoring and remediation [47-49], extreme environments [50], and nutrition and health. In the medical field, understanding changes in human microbiome structure and function helps grasp human health dynamics, particularly in oral environments [51,52], intestinal tracts and their digestive mechanisms [53], and skin sensitivity [54]. In bioenergy, complex processes such as cellulosic ethanol conversion and fermentation [55] and biogas production [56] depend on microbial community functions.

When conducting microbiome applications in health, environment, and nutrition, the microbiome big data center, as a neutral third-party service platform, can not only provide support in data resources, analytical mining methods, and knowledge bases but also foster a healthy development model: public data and public methods → private data online analysis and storage on public platforms → point-to-point data exchange with potential collaborators when appropriate → private data public release, feeding back to the public platform.

5. Practice and Considerations

Following the design concept of the microbiome big data center, we have conducted preliminary work. With support from the Omics Data Encyclopedia NODE (http://www.biosino.org/node/), we have established a microbiome data portal (http://www.biosino.org/microbiome) and a microbiome analysis platform (http://www.biosino.org/microap). The microbiome data portal primarily provides browsing, querying, and publication of public microbiome omics data, selecting representative data as reference datasets to support the analysis platform. The microbiome analysis platform relies on reference data from the portal to provide metagenomic functional analysis and ecological community structure analysis. It will subsequently support multiple approaches for user

whole-genome sequencing data analysis and explore models for protecting and utilizing private data on public platforms.

Of course, the microbiome big data center does not encompass all microbiome research work, nor all microbiology research, or even all microbiome research support platforms. Therefore, while building the microbiome big data center, we must 特别强调 emphasize the interaction between the data center and various "physical repositories" and innovative technologies. For the microbiophylome, special attention should be paid to integration and collaborative development with microbial strain banks. For the microbiome, integration and collaborative development with microbiome sample banks (including microbial community banks and DNA banks) are essential. In the process of integrating with synthetic biology parts library construction, the combination of in silico, in vitro, and in vivo technologies, along with the introduction of design-build-test concepts and data from engineering principles, will lay the foundation for entering the era of electronic cells E^2 (electoral/engineered)-cell and X/M-(extensive/multiple)-cell.

In summary, a microbiome big data center with these characteristics will not only ensure the smooth implementation of large-scale scientific research programs but also provide researchers in basic microbiology, population health, environmental science, and other fields with rich datasets and continuously evolving tool platforms over the long term. It will aggregate their wisdom and research achievements, powerfully promoting the development of microbiology as a whole and the extensive and in-depth application of microbial knowledge and technology.

Acknowledgments: We sincerely thank Researcher Ma Yingfei from the Center for Synthetic Biology Engineering Research and Researcher Hao Pei from the Pasteur Institute of the Chinese Academy of Sciences for their strong support.

References

1. McGuire A L, Colgrove J, Whitney S N, et al. Ethical, legal, and social considerations in conducting the Human Microbiome Project. Genome Res, 2008, 18(12): 1861-1864.
2. Integrative HMPRNC. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. Cell Host Microbe, 2014, 16(3): 276-289.
3. Qin J, Li R, Raes J, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature, 2010, 464(7285): 59-65.
4. Cardona S, Eck A, Cassellas M, et al. Storage conditions of intestinal microbiota matter in metagenomic analysis. BMC Microbiol, 2012, 12: 158.
5. Mitchell A, Bucchini F, Cochrane G, et al. EBI metagenomics in 2016

—an expanding and evolving resource for the analysis and archiving of metagenomic data. Nucleic Acids Res, 2016, 44(D1): D595-603.

6. Sunagawa S, Coelho LP, Chaffron S, et al. Ocean plankton. structure and function of the global ocean microbiome. Science, 2015, 348(6237): 1261359.
7. Mukherjee S, Stamatis D, Bertsch J, et al. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. Nucleic Acids Res, 2017, 45(D1): D446-D456.
8. Chen I A, Markowitz V M, Chu K, et al. IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res, 2017, 45(D1): D507-D516.
9. Meyer F, Paarmann D, D' Souza M, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics, 2008, 9: 386.
10. Field D, Garrity G, Gray T, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol, 2008, 26(5): 541-547.
11. Ten Hoopen P, Pesant S, Kottmann R, et al. Marine microbial biodiversity, bioinformatics and biotechnology (M2B3) data reporting and service standards. Stand Genomic Sci, 2015, 10: 20.
12. Vandamme P, Pot B, Gillis M, et al. Polyphasic taxonomy, a consensus approach to bacterial systematics. Microbiol Rev, 1996, 60(2): 407-438.
13. Ramasamy D, Mishra A K, Lagier J C, et al. A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. Int J Syst Evol Microbiol, 2014, 64(Pt 2): 384-391.
14. Yarza P, Richter M, Peplies J, et al. The all-species living Tree project: a 16S rRNA-based phylogenetic tree of all sequenced type strains. Syst Appl Microbiol, 2008, 31(4): 241-250.
15. Cole J R, Wang Q, Fish J A, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res, 2014, 42(Database issue): D633-642.
16. McDonald D, Price M N, Goodrich J, et al. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME J, 2012, 6(3): 610-618.
17. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res, 2013, 41(Database issue): D590-596.
18. Mou X, Sun S, Edwards R A, et al. Bacterial carbon processing by generalist species in the coastal ocean. Nature, 2008, 451(7179): 708-711.

19. Warnecke F, Luginbühl P, Ivanova N, et al. Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. Nature, 2007, 450(7169): 560-565.

20. Schloss P D, Westcott S L, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol, 2009, 75(23): 7537-7541.

21. Zhou Q, Su X, Jing G, et al. Meta-QC-Chain: comprehensive and fast quality control method for metagenomic data. Genomics Proteomics Bioinformatics, 2014, 12(1): 52-56.

22. Zhou Q, Su X, Wang A, et al. QC-Chain: fast and holistic quality control method for next-generation sequencing data. Plos One, 2013, 8(4): e60234.

23. Finotello F, Mastrorilli E, Di Camillo B. Measuring the diversity of the human microbiota with targeted next-generation sequencing. Brief Bioinform, 2016.

24. Huson D H, Auch A F, Qi J, et al. MEGAN analysis of metagenomic data. Genome Res, 2007, 17(3): 377-386.

25. Caporaso J G, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods, 2010, 7(5): 335-336.

26. Su X, Xu J, Ning K. Parallel-META: efficient metagenomic data analysis based on high-performance computation. BMC Systems Biology, 2012, 6(Suppl 1): S16.

27. Skinnider M A, Dejong C A, Rees P N, et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic acids research, 2015, 43(20): 9645-9662.

28. Dejong C A, Chen G M, Li H, et al. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. Nature chemical biology, 2016, 12(12): 1007-1014.

29. Klementz D, Doring K, Lucas X, et al. StreptomeDB 2.0-an extended resource of natural products produced by streptomycetes. Nucleic acids research, 2016, 44(D1): D509-D514.

30. Blin K, Medema M H, Kazempour D, et al. antiSMASH 2.0-a versatile platform for genome mining of secondary metabolite producers. Nucleic acids research, 2013, 41(Web Server issue): W204-W212.

31. Placzek S, Schomburg I, Chang A, et al. BRENDA in 2017: new perspectives and new tools in BRENDA. Nucleic Acids Res, 2017, 45(D1): D380-D388.

32. Morgat A, Lombardot T, Axelsen K B, et al. Rhea: an expert curated resource of biochemical reactions. Nucleic Acids Res, 2017, 45(D1): D415-D418.

33. Kanehisa M, Sato Y, Kawashima M, et al. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res, 2016, 44(D1): D457-D462.

34. Hu Q N, Deng Z, Hu H, et al. RxnFinder: biochemical reaction search engines using molecular structures, molecular fragments and reaction similarity. Bioinformatics, 2011, 27(17): 2465-2471.

35. McClymont K, Soyer O S. Metabolic networks to drive synthetic biology. Biochem Soc Trans, 2013, 41(6): 1453-1458.

36. Carbonell P, Planson A G, Fichera D, et al. A computational framework for the automated construction of metabolic models. Nat Protoc, 2012, 7(6): 983-996.

37. Moriya Y, Shigemizu D, Hattori M, et al. PathPred: an enzyme-catalyzed metabolic pathway prediction server. Nucleic Acids Res, 2010, 38(Web Server issue): W138-143.

38. Pitkanen E, Jouhten P, Rousu J. Inferring branching pathways in genome-scale metabolic networks. BMC Syst Biol, 2009, 3: 103.

39. Tu W, Zhang H, Liu J, et al. BioSynther: a customized biosynthetic potential explorer. Bioinformatics, 2016, 32(3): 446-448.

40. King Z A, Lloyd C J, Feist A M, et al. Next-generation genome-scale models for metabolic engineering. Curr Opin Biotechnol, 2015, 35: 23-29.

41. Chou C H, Chang W C, Chiu C M, et al. FMM: a web server for metabolic pathway reconstruction and comparative analysis. Nucleic Acids Res, 2009, 37(Web Server issue): W129-134.

42. Rahman S A, Advani P, Schunk R, et al. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). Bioinformatics, 2005, 21(7): 1189-1193.

43. Carbonell P, Parutto P, Baudier C, et al. XTMS: pathway design in an eXTended metabolic space. Nucleic Acids Res, 2014, 42(Web Server issue): W389-W394.

44. Thiele I, Palsson B O. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc, 2010, 5(1): 93-121.

45. Narihiro T, Sekiguchi Y. Microbial communities in anaerobic digestion processes for waste and wastewater treatment: a microbiological update. Current opinion in biotechnology, 2007, 18(3): 273-278.

46. Tong X, Xu H, Zou L, et al. High diversity of airborne fungi in the hospital environment as revealed by meta-sequencing-based microbiome analysis. Sci Rep, 2017, 7: 39606.

47. Tringe S G, Zhang T, Liu X, et al. The airborne metagenome in an indoor urban environment. Plos one, 2008, 3(4): e1862.

48. Demain A L. Microbial biotechnology. Trends Microbiol, 2000, 8(8): 353-354.

49. Cebra J J. Influences of microbiota on intestinal immune system development. Am J Clin Nutr, 1999, 69(5): 1046S-1051S.

50. Yang F, Zeng X, Ning K, et al. Saliva microbiomes distinguish caries-active from healthy human populations. The ISME Journal, 2012, 6(1): 1-10.

51. Jumpertz R, Le D S, Turnbaugh P J, et al. Energy-balance studies reveal associations between gut microbes, caloric load, and nutrient absorption in humans. The American Journal of Clinical nutrition, 2011, 94(1): 58-65.

52. Kong H H. Skin microbiome: genomics-based insights into the diversity and role of skin microbes. Trends Mol Med, 2011, 17(6): 320-328.

53. Shi W, Moon C D, Leahy S C, et al. Methane yield phenotypes linked to differential gene expression in the sheep rumen microbiome. Genome Res, 2014, 24(9): 1517-1525.

54. Agler M T, Ruhe J, Kroll S, et al. Microbial hub taxa link host and abiotic factors to plant microbiome variation. PLoS Biol, 2016, 14(1): e1002352.

Zhang Guoqing is Vice Director and Principal Investigator of the Bio-Med Big Data Center at the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. His main research interests include bioinformatics databases and knowledge bases, focusing on the integration and mining of omics data, literature data, and clinical data in fields such as precision medicine, large population cohorts, personalized drug development, microbiome, and synthetic biology. E-mail: gqzhang@picb.ac.cn

Zhao Guoping is an Academician of the Chinese Academy of Sciences, Chief Scientist of the Bio-Med Big Data Center at the Shanghai Institutes for Biological Sciences, Professor at the Institute of Plant Physiology and Ecology, Executive Director of the Chinese National Human Genome Center at Shanghai, and Counselor to the Chinese Society for Microbiology and Chinese Society of Biotechnology. His research focuses on microbial metabolic regulation and the structure-function relationships and reaction mechanisms of enzymes, developing corresponding microbial and protein engineering biotechnologies. E-mail: gpzhao@sibs.ac.cn

Note: Figure translations are in progress. See original paper for figures.