Thoughts on the Construction of China Microbiome Data Center (Postprint)

Authors: Ma Juncai, Zhao Fangqing, Su Xiaoquan, Xu Jian, Wu Linhuan

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

Abstract

In recent years, the United States and the European Union have successively launched microbiome-related research projects. However, the collection, storage, functional mining, and development and utilization of microbiome big data have always been core issues constraining the development of microbiome research. This article analyzes that China' s current microbiome data management suffers from problems such as non-uniform standards, lack of cross-domain data integration, high-quality reference databases, and deep data mining technologies, and proposes timely launching the "China Microbiome" initiative, establishing a China Microbiome Data Center, and on the basis of microbiome data standardization, building a microbiome big data computing, storage, and sharing platform, developing new methods for microbiome big data mining, and achieving systematic management and efficient utilization of China' s microbiome data resources.

Full Text

Strategies on Establishment of China' s Microbiome Data Center

Ma Juncai[1], Zhao Fangqing[2], Su Xiaoquan[3], Xu Jian[3], Wu Linhuan[1]

[1]Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China
[2]Beijing Institute of Life Sciences, Chinese Academy of Sciences, Beijing 100101, China
[3]Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao 266101, China

Abstract

In recent years, the United States and European Union have successively launched microbiome research initiatives. However, the collection, storage, functional mining, and utilization of microbiome big data have remained core issues constraining microbiome development. This paper analyzes current problems in China' s microbiome data management, including non-uniform standards, lack of cross-domain data integration, high-quality reference databases, and deep data mining technologies. We propose the timely launch of a "China Microbiome Initiative" to establish a China Microbiome Data Center. Based on microbiome data standardization, this center would build a microbiome big data computing, storage, and sharing platform, develop novel methods for microbiome big data mining, and achieve systematic management and efficient utilization of China' s microbiome data resources.

Keywords: microbiome, standardization, big data, China Microbiome Data Center

DOI: 10.16418/j.issn.1000-3045.2017.03.010

Microbiome refers to the total microbial community in a specific environment. Through interactions and equilibrium within a defined environmental space, these communities form relatively stable ecological systems with certain physiological functions. For a long time, microbial communities have been recognized as playing crucial roles in nutrient metabolism, pollutant degradation, and maintaining ecosystem balance in animals, plants, and humans, yet the underlying mechanisms have remained unclear. The widespread application of high-throughput sequencing technology has opened new avenues for studying microbial functions and mechanisms at the community level, enabling us to investigate the composition and functions of microorganisms in natural and human environmental samples from a whole-genome perspective. This provides important means for discovering new genes, developing novel bioactive substances, and studying microbial diversity and evolution in the environment, rapidly transforming this field into a research hotspot.

The massive generation of sequencing data has made microbiomics a true big data science. Taking the human microbiome as an example, it contains trillions of cells, accounting for over 90% of total human cells, encompassing thousands of species and at least 20 million unique microbial genes—far exceeding the number of human genes (approximately 20,000 to 25,000 genes [1]). The Human Microbiome Project (HMP), from its launch in 2008 to the end of its first phase in 2012, completed 5,177 16S rDNA samples, 681 whole genome sequences (WGS), and over 3,000 high-quality reference genomes [2]. However, when sequencing cost is no longer the main limiting factor for microbiome research, data analysis has become the greatest challenge. Even for well-studied human microbiomes, nearly half of the predicted open reading frames (ORFs) cannot find corresponding similar sequences for functional studies [3]. For metagenomic studies in new environments, effective experimental and computational methods are even more

lacking. Currently, high-quality reference databases for environment-related microbiomes, such as soil microbiomes [4] and fermented foods [5], have been gradually established internationally, providing important references for functional annotation and greatly facilitating data integration.

This paper focuses on the critical issue of microbiome data management and analysis, examining current status and needs, summarizing international development trends and challenges, and proposing thoughts and recommendations for constructing China' s microbiome data center.

1. Current Status and Needs of Microbiome Data Management and Analysis

The lack of standardized protocols throughout current metagenomic research processes—from sample collection, extraction, and measurement methods (such as high-throughput sequencing, mass spectrometry, and nuclear magnetic resonance) to data analysis and integration—represents a fundamental challenge. The non-uniformity of metagenomic data standards and the absence of integration technologies mean that sample data from different research projects, sampling sources, and data platforms can only be simply aggregated based on sampling information, but cannot be integrated and uniformly mined according to structural features and functions. Consequently, the biological significance embedded in large-scale datasets cannot be extracted.

Microbiome data and its analytical characteristics impose high demands on complex data integration. Microbiome research generates massive amounts of complex data, including metadata describing environments and samples, original sequencing files, and variously formatted data from sequence annotation and functional studies. This creates enormous challenges for organizing, storing, accessing, sharing, and integrating such large-scale complex data with associated datasets. Furthermore, the integration and comparative analysis of data from different ecosystems (such as gut, soil, and marine environments) with different structural and functional features holds significant value for cross-ecosystem analysis and understanding interaction mechanisms between species distribution and environmental factors.

Microbiome data analysis also suffers from a lack of high-quality reference sequences. Species identification and gene annotation in metagenomic research both depend on known reference genomes and related annotation information. Even for human microbiomes with extensive systematic research, nearly half of predicted ORFs lack functional annotation [3]. For new environmental metagenomes, effective experimental and computational approaches are even more scarce. Additionally, the absence of rapid metagenomic comparison and massive data search technologies, coupled with the storage and analysis costs and computational demands of explosively growing metagenomic data, urgently require innovative solutions combining novel hardware (such as GPUs), cloud computing, associated data integration methods, and efficient search

algorithms.

2. International Microbiome Data Platform Construction

On May 13, 2016, the U.S. government announced the $521 million "National Microbiome Initiative,"aiming to comprehensively study microbial ecosystems in various environments to reveal microbiome composition, structure, and function, and to promote the protection and restoration of healthy microbiome functions. By 2016, 13 human health-related microbiome projects supported by the U.S. National Institutes of Health (NIH), including the Human Microbiome Project (HMP) and the European Union-supported Human Gut Microbiome (MetaHIT), as well as nine environmental microbiome research programs including the Earth Microbiome Project (EMP) and Marine Microbial B3 Plan (Micro B3 Biodiversity, Bioinformatics, and Biotechnology) had been launched internationally [6]. Most of these projects have established robust data integration mechanisms and data management platforms to comprehensively understand microbial community diversity and functions through sequencing analysis of human and environmental samples.

The HMP, from 2008 to 2012 (with Phase II beginning in 2014), aimed to explore the relationship between the human microbiome and human health and disease, focusing on five areas: respiratory tract, oral cavity, skin, gut, and vagina. The project collected thousands of samples from 242 individuals at two clinical centers (Baylor College of Medicine and Washington University School of Medicine) and performed 16S and WGS sequencing at four sequencing centers (Baylor College of Medicine Human Genome Sequencing Center, MIT Broad Institute, J. Craig Venter Institute, and Washington University School of Medicine). Since sample collection and sequencing were conducted by different institutions, the project developed standardized protocols and quality control procedures for sequencing and data analysis. HMP also established a Data Analysis and Coordination Center (DACC) to store all project-generated 16S, WGS, and reference genome sequences. DACC also published news, announcements, and project statistics, and collaborated with sequencing centers on data analysis and annotation. All project data were simultaneously submitted to NCBI for public release.

In August 2010, the Earth Microbiome Project (EMP) officially launched, aiming to comprehensively analyze microbial community diversity and functions through metagenomic sequencing of typical global environmental samples, including soil, marine, air, and freshwater ecosystems. From its inception, establishing an integrated database of samples, genes, and proteins to address fundamental questions about Earth' s ecosystems was defined as one of its three main objectives [8]. To achieve quality control of metadata and data, EMP recommended using the Minimum Information about a Genome Sequence (MIGS) specification [9] and Minimum Information about an Environmental Sequence (MIENS) specification [10] as data standards, and defined standards and protocols for metadata, DNA extraction, and different sequencing targets including

16S, 18S, and ITS [11]. Project data are managed and shared through the Quantitative Insights into Microbial Ecology (QIIME) database. As of August 2014, the project had over 200 collaborators providing data covering more than 40 different ecological environments [12].

In addition to project-established data centers, some major sequencing and research institutions have also built microbiome data platforms. Notable examples include the Integrated Microbial Genomes (IMG) platform established by the U.S. Department of Energy Joint Genome Institute [13] and the Metagenome-RAST (MG-RAST) platform established by Argonne National Laboratory [14], which have gained widespread application. IMG supports annotation, analysis, and management of sequencing data from the Joint Genome Institute and is gradually made freely available to global scientists. For data standards, both IMG and its data management platform Genome Online [FIGURE:1] use specifications developed by the Genomic Standards Consortium (GSC) [16] for describing minimum datasets of environmental sequencing samples, enabling integrated data to be organized and classified by ecosystem, environment, host, or engineering modifications. The platform also provides a series of analysis tools for genomic and metagenomic data.

MG-RAST primarily aims to provide users with phylogenetic and functional annotation analysis workflows for metagenomic data based on high-performance computing resources. For non-bioinformatics specialists, basic annotation information can be obtained simply through a workflow [FIGURE:2]. MG-RAST also provides a data management platform where users can manage their own metadata and sequence files and choose to make data public or keep it private.

China has actively participated in the international EMP initiative. Moreover, since the early 21st century, experts from the Institute of Microbiology, Chinese Academy of Sciences have promoted the "Microbial Earth" research plan, and in 2014, the Chinese Academy of Sciences organized and launched a pilot special research program on soil microorganisms. Chinese scientists have achieved excellent results in human microbiome, brewing microbiome, and microbial data resources. In terms of publication volume, China ranks second globally, only behind the United States, though with a considerable gap

.

Core teams centered around the Chinese Academy of Sciences have a solid foundation in microbiome data platform construction and data analysis. Regarding data platform construction focused on microorganisms, the World Data Center for Microorganisms (WDCM) located at the Institute of Microbiology, Chinese Academy of Sciences is China's first world data center in life sciences. The Global Catalogue of Microorganisms (GCM) platform established by Ma Juncai' s team at the Institute of Microbiology integrates detailed information on over 300,000 microbial physical resources from 110 international microbial resource collection institutions across 43 countries including the U.S., France, Germany, and the Netherlands, many of which come from special ecological environments and have

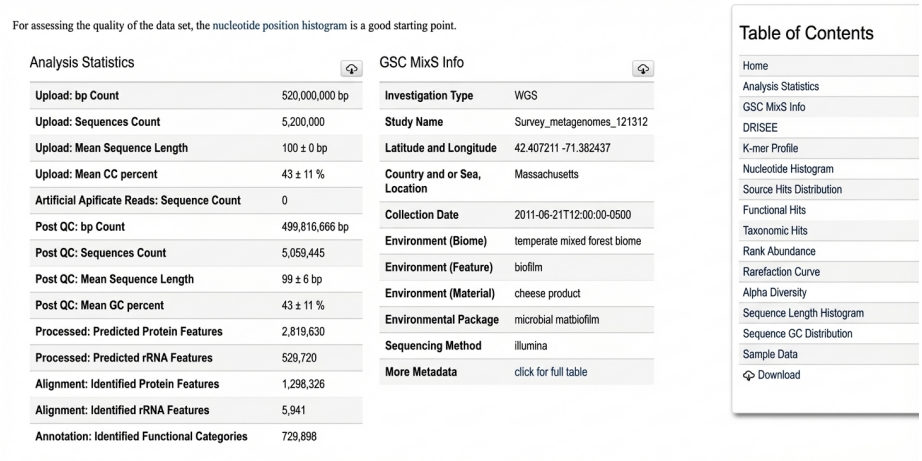

Figure 1: Figure 3

important scientific research and industrial application value [18]. Additionally, Ma Juncai' s team has established high-quality genome reference databases for food-borne pathogenic microorganisms and extremophiles, integrating massive international microbiome data and analysis workflows to form a cloud-based microbiome analysis system.

Recently, Zhao Fangqing' s team at Beijing Institute of Life Sciences, Chinese Academy of Sciences, has developed novel technologies and methods for microbiome research including RiboFR-seq [19], metaSort [20], inGAP-sf [21], and inGAP-CDG [22]. These tools address issues in metagenomic analysis such as assembly, sequence classification and annotation, and microbial interactions, providing new technical means for efficient microbiome interpretation. Su Xiaoquan and Ning Kang' s team at Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, have developed high-performance computational analysis software Parallel-META 3 [23] and metagenomic comparison algorithms Meta-Storms [24] and GPU-Meta-Storms [25], enabling deep, comprehensive, and rapid structural and functional analysis of massive unknown microbiomes and allowing big data-based analysis of microbiome changes under disease or ecological disasters. Xu Jian' s team at the same institute has proposed the concepts of "Ramanome" and "Meta-ramanome," enabling non-label, rapid characterization and measurement of cell population or community states and functions at single microbial cell resolution. These differ essentially from "genotype" data such as metagenomes and offer irreplaceable advantages over existing "phenotype" data like metatranscriptomes, metaproteomes, and metabolomes in terms of single-cell resolution, non-destructiveness, throughput, and cost, representing a novel type of microbiome big data.

3. Current Status and Needs of China' s Microbiome Data Platform Construction

However, China' s microbiome-related research data resources are scattered across various laboratories, with no national-level microbiome database system or data management mechanism. Concurrent problems include non-uniform standards, disconnected data production and analysis, difficulties in data integration and preservation, incomplete analysis technologies and methods, and lack of deep data mining techniques. The absence of efficient, stable, and usable computing platforms prevents the discovery of valuable biological information from massive data, severely hindering the development of microbiome technologies and applications.

4. Thoughts and Recommendations

Data resources are the key to microbiome research and represent important strategic resources. Compared with genome research, microbiome research is still in its initial stage internationally. We should address key problems in microbiome data management and analysis based on China' s research status and gradually develop our own core advantages. We specifically propose the following recommendations:

(1) **Construct microbiome data standardization and management systems.** Establish a complete set of technical standards for microbiome research (sample collection, preservation, data production, analysis, quality control) and management norms and mechanisms (data sharing, storage, intellectual property, etc.). Implement standardized data interfaces and storage solutions, standardized analysis methods and workflows, evaluation systems for standardized computing and storage solutions, and standardized data security and classification systems. On this basis, develop a microbiome data management system to gradually integrate domestic microbiome data resources from human, environmental, and industrial/agricultural sources, achieving effective management and efficient integration of China' s microbiome data resources.

(2) **Establish microbiome big data computing, storage, and sharing platforms.** Collect and organize massive public microbiome data, integrate multi-omics information of samples, and achieve broad and deep-level integration of microbiome big data. Establish high-quality microbiome reference databases, develop efficient big data search and similarity analysis algorithms, and create efficient microbiome data processing workflows to enable systematic management, efficient analysis, and integrated utilization of microbiome data.

(3) **Develop novel methods for microbiome big data mining.** Establish metagenomic species annotation and whole-genome sequence assembly methods suitable for metagenomes, develop metagenomic assembly and

sequence classification algorithms based on strategies for reducing species complexity, create functional annotation methods for distant metagenomic data based on multiple sequence alignment, develop microbiome big data search engines based on microbial community structure and functional similarity, and combine artificial intelligence to develop microbiome diagnosis and early warning technologies for chronic diseases and ecological disasters. Develop data processing methods suitable for high-performance computing platforms to enable visualization of large-scale data and analysis results.

(4) **Strengthen international cooperation with Chinese leadership.** Based on the data platform, participate in international standard formulation, actively lead international microbiome data cooperation plans that meet China' s major needs, form a larger-scale data sharing system, and enhance China' s international influence and discourse in microbiome research.

Countries have placed microbiome research in an unprecedentedly important position and formed relatively solid working foundations. China has significant advantages in microbial resources and sequencing capabilities, but still faces many weak links in key technologies for microbiome big data collection, storage, functional mining, and utilization, which are critical constraints on China' s microbiome research. Therefore, we recommend the timely launch of a "China Microbiome Initiative"to establish a China Microbiome Data Center and achieve systematic management and efficient utilization of China' s microbiome data resources.

References

1. Grice E A, Segre J A. The Human Microbiome: our second Genome. Annu Rev Genomics Hum Genet, 2012, 13(1): 151-170.
2. Gevers D, Knight R, Petrosino J F, et al. Human Microbiome Project Consortium: A framework for human microbiome research. Nature, 2012, 486(7402): 215-221.
3. Kurokawa K, Itoh T, Kuwahara T, et al. Comparative metagenomics revealed commonly enriched gene sets in human gut microbiomes. DNA Res, 2007, 14: 169-181.
4. Choi J, Yang F, Stepanauskas R, et al. Strategies to improve reference databases for soil Microbiomes. The ISME Journal, 2016, 1-6.
5. Sun Z, Harris H M B, McCann A, et al. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. Nat Commun, 2015, 6: 8322.
6. Stulberg E, Fravel D, Proctor L M, et al. An assessment of US microbiome research. Nat Biotechnol, 2016, 1(1):15015.

7. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. Nature, 2012, 486(7402): 207-214.

8. Gilbert J A, Meyer F, Jansson J, et al. The Earth Microbiome Project: Meeting report of the "1st EMP meeting on sample selection and acquisition" at Argonne National Laboratory. Standards in Genomic Sciences, 2010, 3(3):249-253.

9. Field D, Garrity G, Gray T, et al. The minimum information about a genome sequence (MIGS) specification. Nat Biotechnol, 2008, 26(5): 541-547.

10. Yilmaz P, Kottmann R, Field D, et al. Minimum information about an ENvironmental Sequence (MIENS) specification. Nat Biotechnol, 2011, 29: 415-420.

11. http://www.earthmicrobiome.org/emp-standard-protocols/

12. Gilbert J A, Jansson J K, Knight R, et al. The Earth Microbiome project: successes and aspirations. BMC Biol, 2014, 12(1): 69.

13. 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.

14. 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(1): 386.

15. Mukherjee S, Stamatis D, Bertsch J, et al. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. Nucleic Acids Res, 2017, 45(D): D446-D456.

16. Field D, Sterk P, Kottmann R, et al. Genomic Standards Consortium Projects. Standards in Genomic Sciences, 2014, 9(3): 599-601.

17. Wooley J C, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput Biol, 2010, 6(2): e1000667.

18. Wu L, Sun Q, Desmeth P, et al. World data centre for microorganisms: an information infrastructure to explore and utilize preserved microbial strains worldwide. Nucleic Acids Res, 2017, 45(D): D611-D618.

19. Zhang Y, Ji P, Wang J, et al. RiboFR-Seq: a novel approach to linking 16S rRNA amplicon profiles to metagenomes. Nucleic Acids Res, 2016, 44(10): e99.

20. Ji P, Zhang Y, Wang J, et al. MetaSort untangles metagenome assembly by reducing microbial community complexity. Nat Commun, 2017, 8: 14306.

21. Shi W, Ji P, Zhao F. The combination of direct and paired link graphs can boost repetitive genome assembly. Nucleic Acids Res, 2016. DOI: https://doi.org/10.1093/nar/gkw1191

22. Peng G, Ji P, Zhao F. A novel codon-based de Bruijn graph algorithm for gene construction from unassembled transcriptomes. Genome Biol, 2016, 17(1): 232.

23. Su X, Xu J, Ning K. Parallel-META 3: Comprehensive taxonomical and functional analysis platform for efficient comparison of microbial communities. Sci Rep, 2017, 7:40371.

24. Su X, Wang X, Jing G, et al. Meta-Storms: A new method to assess the similarity of microbial communities based on a novel indexing scheme and similarity score for metagenomic data. Bioinformatics, 2012, 28(19): 2493-2501.

25. Jing G, Sun Z, Wang H, et al. Parallel-META 3: Comprehensive taxonomical and functional analysis platform for efficient comparison of microbial communities. Sci Rep, 2017, 7:40371.

Author Information

Ma Juncai is Director of the Center for Microbial Resources and Big Data at the Institute of Microbiology, Chinese Academy of Sciences; Senior Engineer; Director of the World Data Center for Microorganisms; Executive Board Member of the World Federation for Culture Collections (WFCC); and Member of the Human Genetic Resources Management Expert Committee of the Ministry of Science and Technology. He is Principal Investigator of the National High-Tech "863" Program project "Key Technology Research on Microbial Digital Resource Information System Integration." His main research fields include informatization of microbial resources and biotechnology, and cloud-based microbial big data management and analysis platforms. E-mail: ma@im.ac.cn

Zhao Fangqing is a researcher at the Beijing Institute of Life Sciences, Chinese Academy of Sciences, focusing on bioinformatics and microbiome research.

Su Xiaoquan is a researcher at the Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, focusing on microbiome big data analysis.

Xu Jian is a researcher at the Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, focusing on single-cell microbiology and microbiome research.

Wu Linhuan is a researcher at the Institute of Microbiology, Chinese Academy of Sciences, focusing on microbial informatics.