International Cooperation Program for Major Microbial Data Resources: Global Catalogue of Microorganisms (GCM)

Juncai Ma, Linhuan Wu, and Jianyuan Zhang

Abstract Established in the 1960s, WFCC-MIRCEN World Data Centre for Microorganisms, hereinafter referred to as WDCM, is the most important physical resource data platform in the field of microorganisms throughout the world. Getting hosted by the Institute of Microbiology, Chinese Academy of Sciences in 2010, WDCM is the first world data center in the field of life sciences in China. Taking WDCM as a platform, the Institute of Microbiology, Chinese Academy of Sciences insists on developing "self-reliant" international cooperation and advocates Global Catalogue of Microorganisms, hereinafter referred to as GCM, which is an international cooperation program for major microbial data resources. GCM program aims to provide a globally uniform data warehouse for valuable microbial resources scattered in various culture collections and in the hands of the scientists around the world. Currently, 127 microbial resource collection agencies in 46 countries and regions have officially participated in the program, providing effective data support for all aspects of physical resources of microorganisms such as gathering, collection, transnational transfer, academic and commercial applications and benefit sharing, and offering the most important support to the implementation and enforcement of the *Convention on Biological Diversity* in the field of microorganisms. On the basis of GCM international cooperation program, WDCM launched GCM2.0 international cooperation program–Global Microbial Type Strain Genome and Microbiome Sequencing Project–for complete coverage of microbial genomes, and established cooperation network for genome sequencing and function exploring of microbial resources covering 30 major culture collections in more than 20 countries, which is expected to complete genome sequencing of more than 10,000 microbial type strains, establish a set of international standard system in microbial resource sharing and exploration, and set up a globally authoritative reference database and data analysis platform of microbiome.

Keywords Microorganisms · Genomes · Sequencing

J. Ma (✉) · L. Wu · J. Zhang
Institute of Microbiology, Chinese Academy of Sciences, Beijing, China
e-mail: ma@im.ac.cn

China's e-Science Blue Book 2020,
https://doi.org/10.1007/978-981-15-8342-1_10

1 The Context of GCM Program

WFCC-MIRCEN World Data Centre for Microorganisms (WDCM Fig. 1), was set up by WFCC in the 1960s, and it is the most important microbial resource data platform in the field of microorganisms throughout the world. In 2010, WDCM got hosted by the Institute of Microbiology, Chinese Academy of Sciences. It is the first world data center in the field of life sciences in China. Ma Juncai, director of The Center for Microbial Resource and Big Data, Institute of Microbiology, Chinese Academy of Sciences, serves as the chairman of the WDCM in China and presides over the work of the center [2]. Taking WDCM as a platform, the Institute of Microbiology, Chinese Academy of Sciences insists on developing "self-reliant" international cooperation and promotes the global informatization construction of microbial resources to a new level by advocating Global Catalogue of Microorganisms (GCM Fig. 3), which is an international cooperation program for major microbial data resources (Fig. 2).

Fig. 1 The website of WDCM

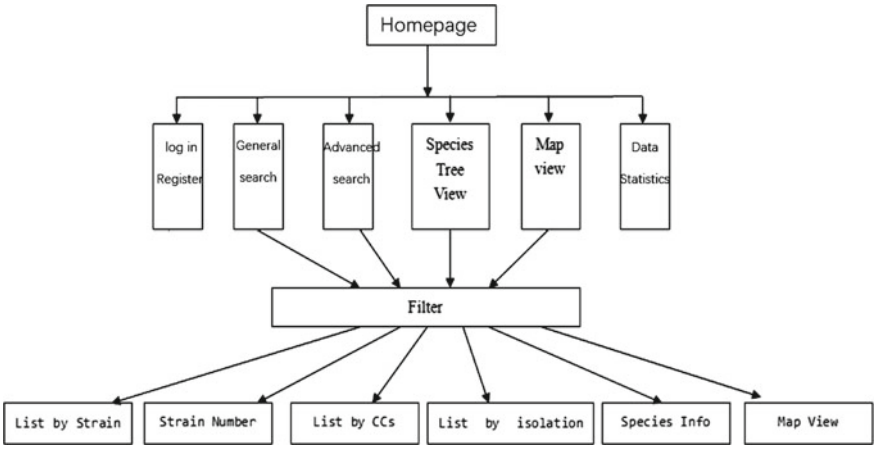

Fig. 2 Data management and service of GCM data platform

Fig. 3 The Website of GCM

2 GCM1.0 Global Catalogue of Microorganisms

The GCM program aims to provide a globally uniform data warehouse for the valuable microbial resources scattered in various culture collections and held by scientists around the world, and to offer information services of microbial stain resources to scientific and industrial circles and industries around the world in the form of a uniform data portal. So far, GCM has integrated detailed information of more than 440,000 microbial physical resources, including microorganisms from special ecological environment with great scientific research and industrial application value. As a big data platform for the integration of microbial digital resources, GCM has also adopted advanced data exploring methods to further extract information on the subsequent research and utilization of microbial resources from more than 6 million published microbial documents and patents worldwide. Therefore, the information platform could provide effective data support for all aspects of physical resources of microorganisms such as gathering, collection, transnational transfer, academic and commercial applications and benefit sharing, and it offers China's most important support to the implementation and enforcement of the *Convention on Biological Diversity* in the field of microorganisms. The integration development framework of the platform adopts java MVC framework and separates the front and back of the integration website to minimize the mutual interference between information storage and information retrieval. MySQL database is used for data storage. In data table design, the basic information and exploring information of strains are differentiated regularly. Currently, the size of the database file is about 20 GB. In the aspect of website monitoring, the monitoring information is updated on a daily basis by using the method of web container access log statistics. Those contents that are monitored are as follows: page view, IP visits, access area, download amount, zone time, etc. GCM relies on strain information offered by culture collections around the world to provide Internet users with query and retrieval, data statistics, literature association, separated sources and indexing of collection sites, etc.

Up till now, a total of 127 microbial resource collection institutions from 48 countries and regions such as the United States, France, Germany, the Netherlands, etc. have officially participated in the program. Meanwhile, it has also established substantive cooperation with regional networks such as ACM, ANRRC, EMbaRC and national networks of Russia, Thailand and Portugal to provide regional data management and sharing with the GCM platform. In order to cater to the implementation of the national strategy of "The Belt and Road Initiative", we have proposed "The Belt and Road Initiative" microbial data resource sharing program on the basis of the GCM platform. The program has been supported by a number of microbial resource institutions. The implementation of the cooperation program could help improve the informatization management level of microbial resources, fulfill the exploration and utilization of the microorganisms with important functions, and promote the development of biotechnology and bio-industry.

3 GCM2.0 Global Microbial Type Strain Genome and Microbiome Sequencing Project

In October 2017, led by the Institute of Microbiology, Chinese Academy of Sciences, GCM2.0 Global Microbial Type Strain Genome and Microbiome Sequencing Project was jointly launched by 12 countries in the world. The launched program is the second phase of the GCM program, including type strain genome sequencing and information analysis, sharing and application of the sequencing data, which is a mighty supplement to the existing GCM1.0 strain data information. The good cooperation between the program and various microbial resource culture collections around the world accumulated in the earlier stage has provided a solid foundation for the resource acquisition of the program, which is also an important condition for us to lead the program.

On October 12, 2017, at the "7th WDCM Academic Seminar", Ma Juncai, the director of WDCM and also the director of the Center for Microbial Resource and Big Data, Institute of Microbiology, Chinese Academy of Sciences, announced the official launch of the Global Microbial Type Strain Genome and Microbiome Sequencing Project led by WDCM and the Institute of Microbiology, Chinese Academy of Sciences and jointly initiated by the culture collections of 12 countries in the world (Fig. 4). At present, 20 culture collections from more than 12 countries such as ATCC of the United States, JCM and NBRC of Japan, KCTC of South Korea, etc.

Fig. 4 The official launch of GCM2.0 global microbial type strain genome and microbiome sequencing project

have joined in the program. The strain resources collected by these culture collections have covered more than 90% of the already known type microbe, of which ATCC and JCM are the largest and most influential type microbe culture collections in the world, which could ensure the availability of resources.

3.1 Whole Genome Sequencing of Type Strains: An Important Breakthrough Point for Decoding the Correlation Between Gene Composition and Function

Type strains are strains that are preserved in pure (reproducible) state as criteria for classification concepts in the process of naming, classifying, recording and publishing microorganisms. At present, there are more than 8,000 sequenced microbial genomes, but the coverage of species is uneven in that a large number of type strains have not been covered, and the data quality is varied, making it impossible for the data to be referred to. It has resulted in a large number of gaps in the systematic classification and genome annotation, etc. of microbial data, which makes the analysis fail to come to an end.

With the diversity of its genes and metabolisms, microorganism is an ideal tool for biotechnology research. In recent years, CRISPR/Cas9 gene editing system from bacterial immune system has rapidly become the most popular technology in life sciences. The whole genome decoding of type strains will make it possible to study the correlation between gene composition and function (such as metabolic activity, virulence, antibiotic production, biomass synthesis, biological fixation of nitrogen, etc.), making great contribution to the research of ecology and biochemistry, and will also further accelerate the discovery of new natural products and drugs. Because it is difficult to be cultured, a large number of valuable microorganisms have not been studied, developed and utilized, and the research methods for the composition and function of environmental and human related microbiome need to be developed urgently. The key of using metagenomic method to analyze microbiome is to obtain high quality genomic reference data. Therefore, the sequencing of type strains will also become an important breakthrough point for microbiome study. Meanwhile, with the reduction of sequencing cost and the improvement of massive data analysis ability, it has become the general trend to launch large-scale sequencing plans and carry out researches based on sequence analysis and function exploration. By taking advantage of our superiority in organization, cost, human resources and technologies, we will seize the highland of international strategic biological resources. Through leading the organization of the program, we will also realize a genuine China-led international cooperation guided by Chinese standards, Chinese databases and Chinese scientists.

3.2 The Network Establishment Program of the International Microbial Type Strain Genome and Microbiome Sequencing Cooperation

The program will complete genome sequencing of more than 10,000 bacteria, fungi and archaeobacteria type strains within 5 years, covering all currently known bacteria and archaeobacteria type strains and important fungi type strains, and establish the international microbial type strain genome and microbiome sequencing cooperation network, covering 30 major culture collections in more than 20 countries [1]. It will also select type strains of microorganisms (including bacteria, archaeobacteria and culturable fungi) that have not been sequenced up till now from the microbial resource culture collections throughout the world and complete genome sequencing of more than 90% of the total type strains of microorganisms (Fig. 5).

1. Deposit strains in culture collections in at least two different countries and obtain receiving numbers.
2. Register the receiving number in GCM for the type strain to be a candidate for sequencing, and send the DNA sample to WDCM.
3. The sequencing results are returned.
4. Inform WDCM of the storage number of the sequenced strain confirmed by the culture collection.
5. The sequencing data is associated with the storage number of the culture collection.
6. Authoritative release of original data and annotation data.

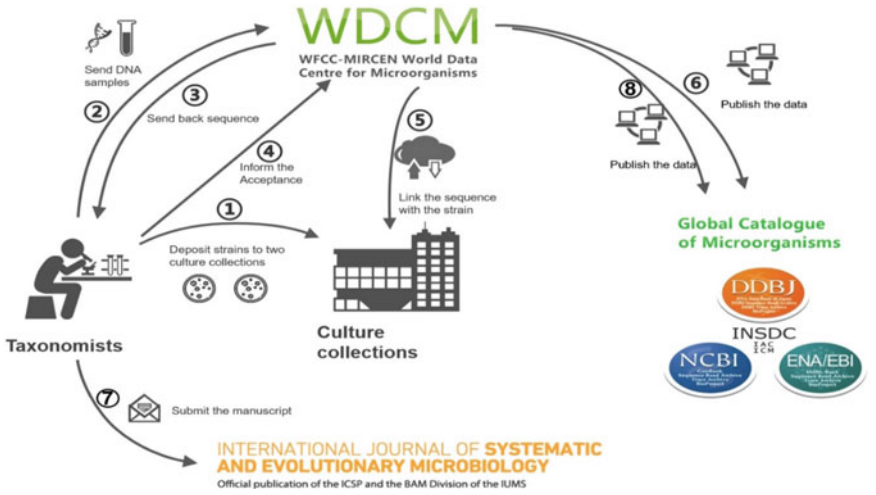

Fig. 5 The flow chart of the international microbial type strain genome and microbiome sequencing cooperation network

7. Submit the paper containing the storage number of the type strain to an authoritative journal for public publication.
8. The sequencing data is published publicly by the authoritative database.

Data standard is the key to successful implementation of a program. The *Data Management and Data Catalog Standard for Microbial Strain Resources*, the establishment of which was led by WDCM, has been officially approved by ISO TC 276 Biotechnology committee, and it is expected to be officially released within two years. It will be the first international data standard in the field of microorganisms. On the basis of data standard, the data generated by this program will also be integrated and shared on the GCM platform. Therefore, we will also form an internationally authoritative microbial data platform.

3.3 GCM Global Cooperation Program Supporting 2019-nCoV Outbreak

On January 24, 2020, Novel Coronavirus National Science and Technology Resource Service System (https://nmdc.cn/nCoV), which was jointly constructed by the Institute of Microbiology, Chinese Academy of Sciences and the Chinese Center for Disease Control and Prevention, was officially launched. The System will timely publish authoritative information on science and technology resources as well as scientific data concerning novel Cov, including the collection of virus strain resources (National Pathogen Microbial Resource Bank), electron micrographs, detection methods, genomes, scientific literature, etc., according to its scientific research progress so as to provide supports for scientific study on 2019-nCoV and special information service of science and technology resources dealing with the current prevention and control of pneumonia caused by 2019-nCoV infection.

On February 18, 2020, Global Coronavirus Data Sharing and Analysis System (https://nmdc.cn/#/coronavirus) was launched by National Microbial Data Center (NMDC). The database includes a total of 3135 coronavirus genomes, including 32,865 nucleic acid sequences from 20,241 strains, separated from 496 different host types and 568 collection sites. The system also provides users with similarity query analysis and phylogenetic analysis based on uploaded genomic sequences, amplified partial sequences or translated protein sequences and data in the database. The system integrates global coronavirus genes and whole genome data, which provides important support and guarantee for Chinese scientists to carry out analysis and research, and promotes the collection and comprehensive analysis and sharing of coronavirus data domestic and abroad. On the other hand, it provides tools such as integrated similarity comparison, phylogenetic analysis, etc., which realizes the integration and standardized analysis and mining process of viromics data, helping scientists to quickly conduct research on virus mutation, traceability, and evolution.

4 Summary and Prospect

At present, the program has already set up the five working groups of bacteria screening, fungi screening, standard operating procedures (SOP), databases, and intellectual property and legal issues, in all of which Chinese scientists have played an important role. The first phase of the program has already started to accept about 800 candidate type strain samples from Belgium, China, Japan, South Korea, the Netherlands, Portugal, Russia, Sweden, Thailand, the United States and Britain. The *Data Management and Data Catalog Standard for Microbial Strain Resources* established by the program has been approved by ISO TC 276 Biotechnology committee as PWI20710. The Chinese Academy of Sciences has already deployed the project of "Research on the Common Technologies of Microbiome in Population and Environmental Health". In 2016, the scientists of the Chinese Academy of Sciences jointly appealed to the state for launching China Microbiome Program, and obtained the instructions of the state leaders. Now the project has been funded, so it is hoped that CAS could take the type microbial genome sequencing program as the starting point, rely on our country's advantages in aspects such as microbial resources research, sequencing technology, and the ability of comprehensive analysis of microbial data, vigorously support the key research and development project of "China Microbiome Program" which covers contents such as human body, agriculture, environment, traditional fermentation, new technologies, etc., and further utilize the international cooperation network established by the program to start the China-led microbiome international cooperation program and capture the strategic commanding point in the field of microorganisms as soon as possible.

Through the implementation of this program, we will take the lead in establishing international standards in the field of microorganisms and set up an internationally authoritative microbial data platform; we will systematically study the physiological functions of microorganisms on a global scale, and establish an integrated research and development application system including biological resource exploration, basic frontier research, technological innovation and industrial development; we will convene a series of brand-name academic conferences and training courses with domain influence to cultivate China's international leading strategic talents and young talents, laying an important foundation for China's talent, resource, technology and industry leadership in the field of microbial resources and even biotechnology.

References

1. Wu L, Ma J (2019) The global catalogue of microorganisms (GCM) 10K type strain sequencing project: providing services to taxonomists for standard genome sequencing and annotation. Int J Syst Evol Microbiol https://doi.org/10.1099/ijsem.0.003276
2. Wu L, Sun Q, Ma J (2017) World data centre for microorganisms: an information infrastructure to explore and utilize preserved microbial strains worldwide. Nucleic Acids Res 45(D1):D611–D618

Juncai Ma, Ph.D., senior engineer, is currently the director of the Center for Microbial Resource and Big Data, Institute of Microbiology, Chinese Academy of Sciences, director of National Microbiology Data Center (NMDC), director of WFCC-MIRCEN World Data Centre for Microorganisms, director of the Biotechnology and Bioindustry Information Center of Chinese Society of Biotechnology, member of Executive Committee of WFCC, chairman of the data management working group of Asian Network of Research Resource Centers (ANRRC) , and Co-chairman of the Working Group of BOLD Mirror, iBOL.