

# The Global Catalogue of Microorganisms (GCM) 10K type strain sequencing project: providing services to taxonomists for standard genome sequencing and annotation

Linhuan Wu<sup>1,2</sup> and Juncai Ma<sup>1,2,\*</sup>

## Abstract

The World Federation of Culture Collections and the World Data Center for Microorganisms (WDCM) initiated an international community-led project to sequence and annotate newly described prokaryotic taxa. This sequencing project aims to cooperate with international culture collections and the *International Journal of Systematic and Evolutionary Microbiology* and contribute to the expansion of whole genome sequencing databases for type strains. It will provide global microbial taxonomists with free standard genome sequencing and annotation services. Taxonomists are encouraged to contact the WDCM and participant culture collections to submit a type strain sequencing proposal.

## INTRODUCTION

Since the late nineteenth century, the taxonomy of microorganisms has evolved from classical phenotypic methods to polyphasic approaches. However, the current widely used genotypic criteria, including DNA G+C content, DNA–DNA hybridization (DDH), 16S rRNA and other house-keeping gene sequence nucleotide similarity analyses, have various limitations [1].

Results obtained from genomic DDH experiments and the sequence analysis of small-subunit rRNA are widely used and have come to be recognized as the so-called, ‘gold standard’ for species delineation. In prokaryotic taxonomy, strains with DDH values greater than 70 % similarity under standardized conditions are generally considered to belong to the same species [2]. However, this method is rather personnel-consuming and does not allow the construction of cumulative databases [3].

The various domains of the rRNA gene serve as useful markers in prokaryotic taxonomy and biodiversity assessments, due to its ubiquitous presence and conserved properties. However, this gene is highly conserved and often not able to provide discrimination among species within the same genus, whereas multiple-gene and whole-genome-derived approaches often provide the required resolution [4]. Since 1987, well before genome sequencing of

prokaryotes was commonplace, the search for an ideal taxonomy recognized that the complete deoxyribonucleic acid (DNA) sequence would be the reference standard to determine phylogeny [5].

The rapid development of next-generation sequencing technology has made high-throughput sequencing a key tool for many microbiology laboratories. Average nucleotide identity and average amino acid identity, as well as core- and pan-genome cluster analyses-based approaches have been demonstrated to be robust measurements of genomic relatedness between organisms and accurate approaches for a genome-based taxonomy. [6–8].

As a result, the *International Journal of Systematic and Evolutionary Microbiology* (IJSEM) announced that, as of January 2018, authors of new bacterial and archaeal taxa descriptions are requested to provide genome sequence data with their description of novel taxa (including novel species) [9], which means that genomic data should be made available at the time of publication.

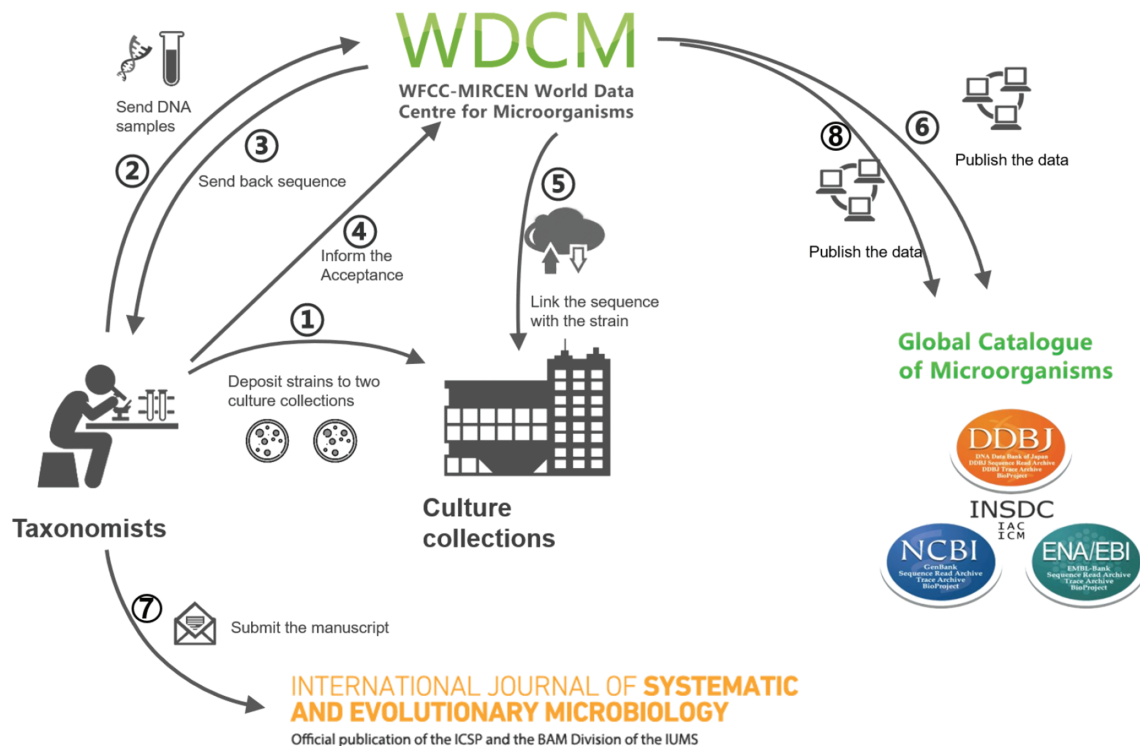
For many decades, culture collections around the world have taken responsibilities for isolation, identification, distribution and long-term preservation of microbial resources to document and preserve, *ex situ*, Earth’s microbial diversity and to provide reference materials to support research and biotechnology. There are currently approximately

**Author affiliations:** <sup>1</sup>Microbial Resource and Big Data Center, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, PR China; <sup>2</sup>WFCC-MIRCEN World Data Center for microorganisms, WDCM, Beijing 100101, PR China.

**\*Correspondence:** Juncai Ma, ma@im.ac.cn

**Keywords:** type strain; genome sequencing; culture collections; prokaryote.

**Abbreviations:** AAI, average amino acid identity; ANI, average nucleotide identity; DDH, DNA–DNA hybridization (DDH); DNA, deoxyribonucleic acid; GCM, global catalogue of microorganisms; GEBA, genomic encyclopedia of bacteria and archaea; GOLD, genomes online database; IJSEM, *International Journal of Systematic and Evolutionary Microbiology*; NGS, next-generation sequencing; SOPs, standard operating procedures; USDOE JGI, US Department of Energy Joint Genome Institute; WDCM, World data center for microorganisms; WFCC, World federation of culture collections.



**Fig. 1.** Workflow of the new type strain sequencing. 1, Taxonomists deposit the strains, to be proposed as type strains, in culture collections in at least two different countries (at least one culture collection listed in Table 1) and obtain receiving numbers (a pre-assigned number by culture collections before formally accepting the type strain deposition) for the strains. 2, Using the receiving number (declaring that the strains have been deposited in culture collections), taxonomists register with the WDCM Global Catalogue of Microorganisms project for candidate type strain sequencing and send DNA samples to the WDCM. 3, the WDCM sends the sequencing results to the taxonomists within 2 months (including assembly, G+C content, OGRI, N50 and annotation results required by the IJSEM). 4, The taxonomists inform the WDCM of the culture collection accession numbers when they receive formal certificates from the respective culture collections. 5, The WDCM substitutes the receiving numbers for the culture collection strain accession numbers and links these numbers with genome sequence data. 6, Taxonomists request the WDCM to publish the raw data and annotation data within the NCBI/EMBL/DDBJ/GCM. 7, Taxonomists submit the manuscript to the IJSEM with the genome sequence accession number. 8, If the paper is not published in the IJSEM (including without submission to journal, withdrawal by the authors or rejection by the editors), sequencing data will be published in the NCBI/EMBL/DDBJ/GCM databases 24 months after sequencing.

2 693 000 microbial strains available from the 746 collections in 76 countries and regions referring to the CCINFO and STRAIN database of the World Data Center for Microorganisms (WDCM)[10]. The Genomes OnLine Database (GOLD) [11] maintains an open, public genome database of 16 112 bacterial and archaeal type strains covering more than 6000 species. Comparing with the more than 15 000 species described in the IJSEM with validly published names, additional efforts to continue sequencing type strains are necessary to fill in the gap.

Addressing the gap in genome data, the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project led by the US Department of Energy Joint Genome Institute (US DOE JGI) has published the outcomes of its first stage: 56 whole-genome sequences at year 2009 [12] and 1 003 [13] at year 2017, for a total of 1059 type strain whole-genome sequences. The newly started third-stage of the GEBA focuses on the genomes of soil, plant-associated and newly described type strains [14].

The World Federation of Culture Collections (WFCC) [15] and the WDCM have initiated an international community-led project towards genome sequencing and data annotation of microbial type strains, the Global Catalogue of Microorganisms (GCM) 10K type strain sequencing project. This project will include all bacterial and archaeal 80 type strains with validly published species names, as well as selected reference fungal type strains frequently used for functional or phylogenetic studies.

This project proposes to cooperate with international culture collections and the IJSEM to include type strains of proposed novel species into this project. It will provide global taxonomists with a standard genome sequencing and annotation service. When taxonomists deposit the proposed type strains of newly described species to one of the culture collections (Table 1) that is cooperating with the WDCM for new type strain sequencing, the depositors could arrange to receive genomic sequencing and full annotation by the WDCM as free services. A proposed procedure could be as shown in Fig. 1.

**Table 1.** Culture collections cooperating with the WDCM for new type strain sequencing

Acronym*	Full name/institution	Country	Homepage
1 ATCC	American Type Culture Collection	USA	<a href="http://www.atcc.org/">www.atcc.org/</a>
2 BCRC	Bioresource Collection and Research Center/ Food Industry Research and Development Institute	Chinese Taipei	<a href="http://www.bcrc.firdi.org.tw">www.bcrc.firdi.org.tw</a>
3 CAIM	Collection of Aquatic Important Microorganisms/ CIAD/Mazatlan Unit for Aquaculture and Environmental Management	Mexico	<a href="http://www.ciad.mx/caim">www.ciad.mx/caim</a>
4 CCM	Czech Collection of Microorganisms/ Masaryk University	Czech	<a href="http://sci.muni.cz/ccm/index.html">http://sci.muni.cz/ccm/index.html</a>
5 CCUG	Culture Collection University of Gothenburg	Sweden	<a href="http://www.ccug.se">www.ccug.se</a>
6 CECT	Spanish Type Culture Collection / University of Valencia	Spain	<a href="http://www.cect.org">www.cect.org</a>
7 CGMCC	China General Microbiological Culture Collection Center	China	<a href="http://www.cgmcc.net">www.cgmcc.net</a>
8 CICC	China Center of Industrial Culture Collection	China	<a href="http://www.china-cicc.org">www.china-cicc.org</a>
9 CIP	Collection de l'Institut Pasteur	France	<a href="https://catalogue-crbip.pasteur.fr">https://catalogue-crbip.pasteur.fr</a>
10 ICMP	International Collection of Microorganisms from Plants	New Zealand	<a href="http://www.landcareresearch.co.nz/resources/collections/icmp">www.landcareresearch.co.nz/ resources/collections/icmp</a>
11 JCM	Japan Collection of Microorganisms/ RIKEN BioResource Center	Japan	<a href="http://jcm.brc.riken.jp/">http://jcm.brc.riken.jp/</a>
12 KCTC	Korean Collection for Type Cultures, Korea Research Institute of Bioscience and Biotechnology	Republic of Korea	<a href="http://kctc.kribb.re.kr">http://kctc.kribb.re.kr</a>
13 KMM	G.B. Elyakov Pacific Institute of Bioorganic Chemistry, Far-Eastern Branch, Russian Academy of Sciences	Russia	Not available
14 NBRC	Biological Resource Center/National Institute of Technology and Evaluation	Japan	<a href="http://www.nite.go.jp/en/nbrc/index.html">www.nite.go.jp/en/nbrc/index.html</a>
15 NCAIM	National Collection of Agricultural and Industrial Microorganisms	Hungary	<a href="http://ncaim.uni-corvinus.hu/">http://ncaim.uni-corvinus.hu/</a>
16 NCTC	National Collection of Type Cultures	UK	<a href="http://www.phe-culturecollections.org.uk/">www.phe-culturecollections.org.uk/</a>
17 PCU	Pharmaceutical Sciences Chulalongkorn University Culture Collection/ Chulalongkorn University	Thailand	Not available
18 TBRC	Thailand Bioresource Research Center/National Center for Genetic Engineering and Biotechnology	Thailand	<a href="http://www.tbrcnetwork.org">www.tbrcnetwork.org</a>
19 TISTR	TISTR Culture Collection/Bangkok MIRCEN	Thailand	<a href="http://www.tistr.or.th/tistr_culture">www.tistr.or.th/tistr_culture</a>
20 VKM	All-Russian Collection of Microorganisms	Russian Federation	<a href="http://www.vkm.ru/">http://www.vkm.ru/</a>

\*The abbreviations of culture collections are listed in alphabetical order.

The WDCM has established standard operating procedures for DNA extraction, sample submission, sequencing and data processing to ensure that all genetic resources and associated metadata of type strains are comparable, properly maintained, recorded and stored. One of the aims of this project is to make sequencing data available to the scientific community. If taxonomists choose not to submit an article, the sequence data will be published in the public databases 24 months after sequencing is completed. The details of the GCM 10K type strains sequencing project is described in the previous project announcement [16].

Since microbial taxonomy is based upon comparisons of living type cultures, the preservation of type strains, not only including the viable, pure and well-characterized strains, but also the curated and comprehensive metadata associated with each strain, are essential for follow-up studies. As a result, a comprehensive description of taxonomic information will be recommended for new strain sequencing registration.

This effort will contribute to establish a high-quality complete genome sequence database and, hence, to form a genome-based prokaryotic taxonomic framework as the

new 'gold standard'. The resulting knowledge and tools generated through this project will not only directly facilitate identification of micro-organisms, but also improve our abilities to predict new gene complexes as well as the functional existence of microbial communities.

#### Funding information

This work was supported by the National key Research Program of China (Grant Nos. 2017YFD0400302, 2016YFC1200801, 2016YFC0901702); the Key Research Program of the Chinese Academy of Sciences (Grant No. KFZD-SW-219); the 13th Five-year Informatization Plan of Chinese Academy of Sciences (Grant Nos. XXH13506, XXH13505).

#### Acknowledgements

We would like to thank the participant collections who have made the cooperation possible. We would like to thank the WFC Executive Board for supports, especially İpek Kurtböke (WFC President), Kevin McCluskey (WFC Vice-President) and Philippe Desmeth (WFC Past-President) for support. We would like to thank Martha E. Trujillo and Fred A. Rainey for suggestions to the cooperation and the article. We would like to thank the former director of the WDCM, Hideaki Sugawara, who gives continuous support to the WDCM.

#### Conflicts of interest

The authors declare that they have no conflicts of interest.

## References

1. Ramasamy D, Mishra AK, Lagier JC, Padhmanabhan R, Rossi M *et al.* A polyphasic strategy incorporating genomic data for the taxonomic description of novel bacterial species. *Int J Syst Evol Microbiol* 2014;64:384–391.
2. Stackebrandt E, Frederiksen W, Garrity GM, Grimont PA, Kämpfer P *et al.* Report of the Ad Hoc Committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol* 2002;52:1043–1047.
3. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci USA* 2009;106:19126–19131.
4. Konstantinidis KT, Tiedje JM. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* 2007;10:504–509.
5. Wayne LG, Moore WEC, Stackebrandt E, Kandler O, Colwell RR *et al.* Report of the Ad Hoc Committee on reconciliation of approaches to bacterial systematics. *Int J Syst Evol Microbiol* 1987;37:463–464.
6. Daubin V, Gouy M, Perrière G. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res* 2002;12:1080–1090.
7. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K *et al.* Microbial species delineation using whole genome sequences. *Nucleic Acids Res* 2015;43:6761–6771.
8. Kim M, Oh HS, Park SC, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int J Syst Evol Microbiol* 2014;64:346–351.
9. Genome sequencing data required with taxonomic descriptions. 2018 [http://ijs.microbiologyresearch.org/content/Genome\\_data\\_required\\_IJSEM](http://ijs.microbiologyresearch.org/content/Genome_data_required_IJSEM). Accessed 12th December.
10. WDCM. <http://www.wdcm.org/>. Accessed 12th December, 2018.
11. JGI GOLD database. 2018 <https://gold.jgi.doe.gov/>. Accessed 12th December.
12. Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E *et al.* A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature* 2009;462:1056–1060.
13. Mukherjee S, Seshadri R, Varghese NJ, Eloë-Fadrosch EA, Meier-Kolthoff JP *et al.* 1003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat Biotechnol* 2017;35:676–683.
14. Whitman WB, Woyke T, Klenk HP, Zhou Y, Lilburn TG *et al.* Genomic encyclopedia of bacterial and archaeal type strains, phase III: the genomes of soil and plant-associated and newly described type strains. *Stand Genomic Sci* 2015;10:26.
15. WFCC. <http://www.wfcc.info/>. Accessed 12th December, 2018.
16. Wu L, McCluskey K, Desmeth P, Liu S, Hideaki S *et al.* The global catalogue of microorganisms 10K type strain sequencing project: closing the genomic gaps for the validly published prokaryotic and fungi species. *Gigascience* 2018;7:5.

**Five reasons to publish your next article with a Microbiology Society journal**

1. The Microbiology Society is a not-for-profit organization.
2. We offer fast and rigorous peer review – average time to first decision is 4–6 weeks.
3. Our journals have a global readership with subscriptions held in research institutions around the world.
4. 80% of our authors rate our submission process as 'excellent' or 'very good'.
5. Your article will be published on an interactive journal platform with advanced metrics.

**Find out more and submit your article at [microbiologyresearch.org](http://microbiologyresearch.org).**