Nucleic Acids Research, 2025, **53**, D763–D771
https://doi.org/10.1093/nar/gkae1057
Advance access publication date: 5 November 2024
Database issue

GCM and gcType in 2024: comprehensive resources for microbial strains and genomic data

Guomei Fan[1,2,3,4,†], Qinglan Sun[1,2,3,4,†], Yan Sun[1,2,3,4,†], Dongmei Liu[1,2,3,4], Shiwen Li[1,2,3,4], Min Li[1,2,3,4], Qi Chen[1,2,3,4], Fang Wang[1,2,3,4], Ohkuma Moriya[5], Takashi Itoh[5], Hiroko Kawasaki[6], Yajing Yu[7], Man Cai[7], Song-Gun Kim[8], Jung-Sook Lee[8], Juncai Ma[1,2,3,4,*] and Linhuan Wu[1,2,3,4,*]

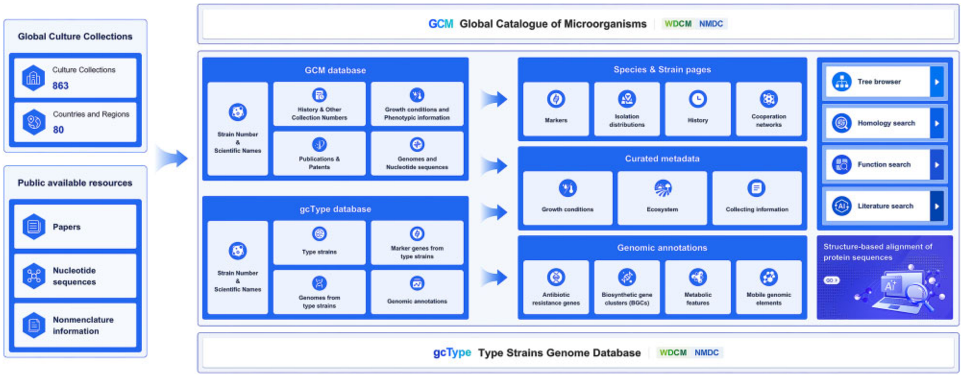
[1]Microbial Resource and Big Data Center, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China
[2]State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China
[3]World Data Center for Microorganisms, Beijing 100101, China
[4]China National Microbiology Data Center (NMDC), Beijing 100101, China
[5]Japan Collection of Microorganisms/ Microbe Division, RIKEN BioResource Research Center, Koyadai 3-1-1, Tsukuba, Ibaraki 305-0074, Japan
[6]NITE Biological Resource Center (NBRC), National Institute of Technology and 24 Evaluation, 2-5-8 Kazusakamatari, Kisarazu, Chiba 292-0818, Japan
[7]China General Microbiological Culture Collection Center (CGMCC), Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China
[8]Korean Collection for Type Cultures (KCTC), Korea Research Institute of Bioscience and 30 Biotechnology (KRIBB), 181 Ipsin-gil, Jeongeup-si, Jeollabuk-do 56212, Republic of Korea

[*]To whom correspondence should be addressed: Tel: +86 10 64807385; Fax: +86 10 64807426; Email: wulh@im.ac.cn
Correspondence may also be addressed to Juncai Ma. Email: ma@im.ac.cn
[†]The first three authors should be regarded as Joint First Authors.

Abstract

Microbial culture collections play a crucial role in the collection, maintenance, and distribution of quality-assured living microbial strains, along with their associated phenotypic and omics data. To enhance the find-able, accessible, interoperable, and re-usable (FAIR) data utilization of microbial resources, the World Data Center for Microorganisms (WDCM) has developed the Global Catalogue of Microorganisms (GCM) and the Global Catalogue of Type Strains (gcType). These platforms provide interactive interfaces for cataloging the holdings of collections, along with detailed annotations of type strain genomes and curated metadata, including ecosystems, growth conditions, and collection locations. The system maximizes the scientific impact of microbial resources and culture collections through an integrated data mining tool that links strain- and species-related information from various public resources. Currently, the GCM and gcType include 574 422 strains from 154 culture collections across 51 countries and regions, along with 25 980 genomes from type species. Additionally, 2 702 655 articles and 103 337 patents are integrated with these microbial resources. The system supports microbial taxonomic research and provides evidence for implementing the Nagoya Protocol in the field of microbial resources and their digital sequence information (DSI). Access is freely available at gcm.wdcm.org and gctype.wdcm.org.

Graphical abstract


Global Culture Collections
Culture Collections
863
Countries and Regions
80
Public available resources
Papers
Nucleotide sequences
Nonmenclature information
GCM Global Catalogue of Microorganisms
WDCM
NMDC
GCM database
Strain Number & Scientific Names
History & Other Collection Numbers
Growth conditions and Phenotypic information
Publications & Patents
Genomes and Nucleotide sequences
gcType database
Strain Number & Scientific Names
Type strains
Marker genes from type strains
Genomes from type strains
Genomic annotations
Species & Strain pages
Markers
Isolation distributions
History
Cooperation networks
Curated metadata
Growth conditions
Ecosystem
Collecting information
Genomic annotations
Antibiotic resistance genes
Biosynthetic gene clusters (BGCs)
Metabolic features
Mobile genomic elements
Tree browser
Homology search
Function search
Literature search
Structure-based alignment of protein sequences
gcType Type Strains Genome Database
WDCM
NMDC

Received: September 14, 2024. **Revised:** October 16, 2024. **Editorial Decision:** October 17, 2024. **Accepted:** October 22, 2024

Introduction

Microbial culture collections are established to preserve a diverse range of microbial species worldwide. They play a vital role in providing authentic material for research, agriculture, medicine, and industry. Ensuring high-quality material alongside associated phenotypic and omics data is one of the primary responsibilities of culture collections, serving as modern microbial biological resource centers (mBRCs) (1).

The World Data Center for Microorganisms (WDCM) initiated the Global Catalogue of Microorganisms project and its database (gcm.wdcm.org) in 2012, starting with a demo version involving 20 collections. The full-scale project has since opened to the entire WFCC culture collections community, maximizing visibility and promoting data access according to the FAIR principles. As of August 2024, the number of participating culture collections has reached 154. The updated version of GCM and its partner databases aim to offer new functions to accommodate the ongoing development of culture collections, address growing demands for integrated data, evaluate contributions to the community, ensure compliance with the Nagoya Protocol for benefit sharing, and connect with industry (2).

One of the most important resources in culture collections is type strains. The taxonomy of prokaryotes has evolved from phenotype-based classification to polyphasic methods, ultimately leading to genome-based taxonomic approaches (3). Since 2018, genome sequence data have been required by the International Journal of Systematic and Evolutionary Microbiology (IJSEM) when publishing a new species. The increase in type strain genomes is essential for investigating evolutionary relationships in prokaryotic systematics and is invaluable for synthetic biology and gene editing. The widespread use of third-generation next-generation sequencing (NGS) technologies (e.g. PacBio and Oxford Nanopore) has increased the availability of complete genomes, enabling in-depth studies of functional genetic elements in type strains, such as defense systems, prophages, and mobile genetic elements(MGE).

Despite these advancements, a significant gap remains between the current sequenced type species and validly published prokaryotic names. Many collections face limitations in funding or human resources, hindering their ability to fully support the transition of taxonomy requirements or benefit from advancements in next-generation sequencing technology. In response, WDCM initiated the second stage of the Global Catalogue of Microorganisms (GCM), the type strain sequencing project, and the first version of the gcType database in 2019.

The updated version of gcType significantly increases the number of sequenced species and features abundant functional annotations, an advanced search interface, and curated metadata for all validly published type strains. We also built an indexable reference database from gcType's unique protein sequences and incorporated deep learning methods to search for functional components with structure-structure similarities but low sequence similarity.

Contents

The database schema of GCM, gcType and associated WDCM databases

GCM and gcType serve as public portals for a series of WDCM databases (Figure 1) that collect, process, integrate, and publish data from culture collections, taxonomists, and other sources to generate knowledge for the microbial community. Culture collections first register or update their metadata, including acronyms, contacts, services, holdings, and human resources, in the Culture Collection Information Worldwide (CCINFO) database. Each collection is assigned a unique WDCM number, which is used to generate a list of unique strain numbers for all holdings. FungalNames serves as a registry for fungal species names and their statuses. The GCM database maintains the holdings of culture collections and establishes cross-links with CCINFO through acronyms. The list of type strains and their taxonomic and functional annotation results are stored in the gcType database.

Both GCM and gcType integrate publicly available articles, patents, nucleotide sequences, and genomes, which are automatically extracted and processed by the ABC (Analyzer of Bio-Resources Citation) (4), a data mining engine for microbial resources. The ABC was developed based on manually curated regular expressions for species names, culture collection acronyms, phenotypic ontology, isolation sources, and locations to ensure accurate, comprehensive, and efficient data processing. For example, we used culture collection acronyms to search for strain numbers contained in papers, such as 'JCM 10005'. However, sometimes, acronyms should be manually curated, e.g. 'DSM' is applicable for strain preserved in the culture collection with acronym of 'DSMZ'. The isolated locations also needed for manual corrected if they are written in irregular way and could not be automatically mapped onto a world map.

Recently, GCM developed a list of its satellite databases, including the Global Catalogue of Pathogens, which contains genomic and epidemiological information on human pathogenic microorganisms, and the Global Catalogue of Phage, which provides comprehensive information on isolated phages and prophages predicted from taxonomically diverse host organisms. The Reference Strains Database maintains recognized reference strains listed in ISO or other standards. Together, these databases form a cross-linked network that provides comprehensive data services for both mBRCs and global microbiologists.

Data contents of GCM and gcType

As of August 2024, the WDCM CCINFO database registers 861 culture collections from 80 countries, with 294 belonging to universities, 230 hosted by public research institutions, and the remainder in industry or private sectors. The data in GCM and gcType have significantly increased (Figure 2) to keep pace with the growth of global culture collections and genome sequencing efforts. Currently, there are 574 422 strains from 154 culture collections in 51 countries and regions, including 90 257 type strains covering 21 983 validly published species, with 25 980 genomes from sequenced type species included in gcType. Additionally, 2 702 655 articles and 103 337 patents are integrated with these microbial resources.

Improved standardization and data accessibility

Culture collections (CCs) that comply with a Quality Management System (QMS) and follow the Best Practice Guidelines proposed by the WFCC should employ standardized data management for exchange and publication. However, a universal data standard covering items relevant to microbial resources and their management, including scientific names,

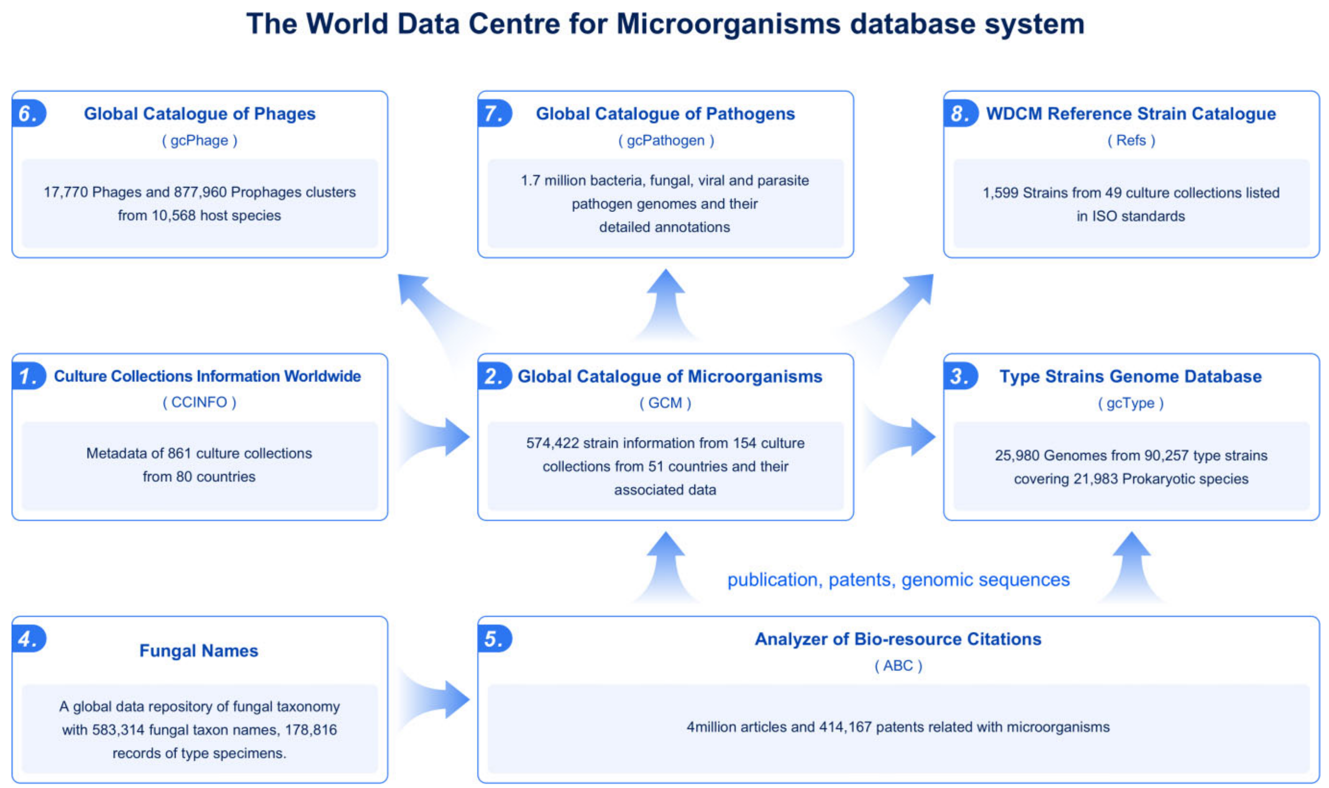

The World Data Centre for Microorganisms database system
6. Global Catalogue of Phages
(gcPhage)
17,770 Phages and 877,960 Prophages clusters from 10,568 host species
7. Global Catalogue of Pathogens
(gcPathogen)
1.7 million bacteria, fungal, viral and parasite pathogen genomes and their detailed annotations
8. WDCM Reference Strain Catalogue
(Refs)
1,599 Strains from 49 culture collections listed in ISO standards
1. Culture Collections Information Worldwide
(CCINFO)
Metadata of 861 culture collections from 80 countries
2. Global Catalogue of Microorganisms
(GCM)
574,422 strain information from 154 culture collections from 51 countries and their associated data
3. Type Strains Genome Database
(gcType)
25,980 Genomes from 90,257 type strains covering 21,983 Prokaryotic species
publication, patents, genomic sequences
4. Fungal Names
A global data repository of fungal taxonomy with 583,314 fungal taxon names, 178,816 records of type specimens.
5. Analyzer of Bio-resource Citations
(ABC)
4million articles and 414,167 patents related with microorganisms

Figure 1. Structure of the WDCM database. The WDCM database system consists of a series of databases. Culture collection users firstly register their unique acronym and metadata in CCINFO database. After that, they register their digitalized catalogue information with the GCM database. If type strains of their holdings were sequenced by WDCM or other sequencing efforts, genomic data are shared through gcType. FungalNames is a registry for fungal species. The ABC data mining tools use species taxon data from FungalNames and other resources to extract publicly available articles, patents, nucleotide sequences and genomes. All these data has been integrated in GCM and gcType. GCM also supports a list of satellite databases, including the Global Catalogue of Pathogens, the Global Catalogue of Phage and the Reference Strains Database.

taxonomic statuses, growth conditions, and practical management details (e.g. storage conditions, methods of cultivation, personnel responsible for data entry/update), had not yet been released. In 2020, the WDCM team and its global partners published a new ISO standard, 'ISO 21710:2020 Biotechnology—Specification on data management and publication in microbial resource centers', by the Biotechnology Technical Committee (TC 276) (https://www.iso.org/standard/71384.html). This document specifies requirements for in-house data management and online data publication in microbial resource centers (mBRCs) to enable consistent formatting and establish a quality control workflow to improve data quality. It provides recommendations for mBRCs to enhance data sharing and integration of microbial material and associated data, as well as a list of minimum and recommended data-set items. All GCM and gcType data items adhere to these ISO standards.

The mission of Open Science is to make research results accessible to the entire scientific community. Culture collections can serve as Open Science hubs by supplying and distributing microbial materials for scientific studies (5). A globally unique identifier for strains and their associated metadata can ensure efficient retrieval and proper citation. Ideally, the strain number would serve as a global unique identifier, typically combining the culture collection's acronym with a digit assigned by each collection. However, these strain numbers are not globally unique (since a plant specimen, a reagent or any kind of other items could have this number, e.g. the strain number of DSM 962 refer to the strain of '*Conidiobolus coronatus*' and also refer to a kind of Scanning Electron Microscope 'Zeiss DSM 962″ produced by Carl Zeiss), which can lead to confusion when tracking the use of microbial resources in publications and patents, resulting in numerous false-positive results. To address this issue, the GCM database has assigned Digital Object Identifiers (DOIs) to all 574 422 strains, linking them to a unique strain number in the microbial world. The DOI should be used not only when citing a strain in an article or patent but also, according to the Nagoya Protocol (NP), in accompanying documents for strain exchanges, including prior informed consent (PIC), mutually agreed terms (MAT), and internationally recognized certificates of compliance (IRCC). Recording the DOI in these documents ensures traceable and legitimate exchanges of microbial resources and their digital sequence information (DSI) (6).

History information and citation integration for implementing the Nagoya Protocol and Access and Benefit sharing (ABS)

Due to the frequent exchange of microbial strains between culture collections, 'equivalent strain numbers' refer to strains with documented exchange histories. Identifying these equivalent strains is crucial for comprehensively understanding the strains. The strainInfo database has previously contributed significantly to this effort (7,8). However, strainInfo has been inactive for some time, prompting WDCM to construct a strain history tracking system (Figure 3) using a semantic web schema based on a more comprehensive strain data volume. In

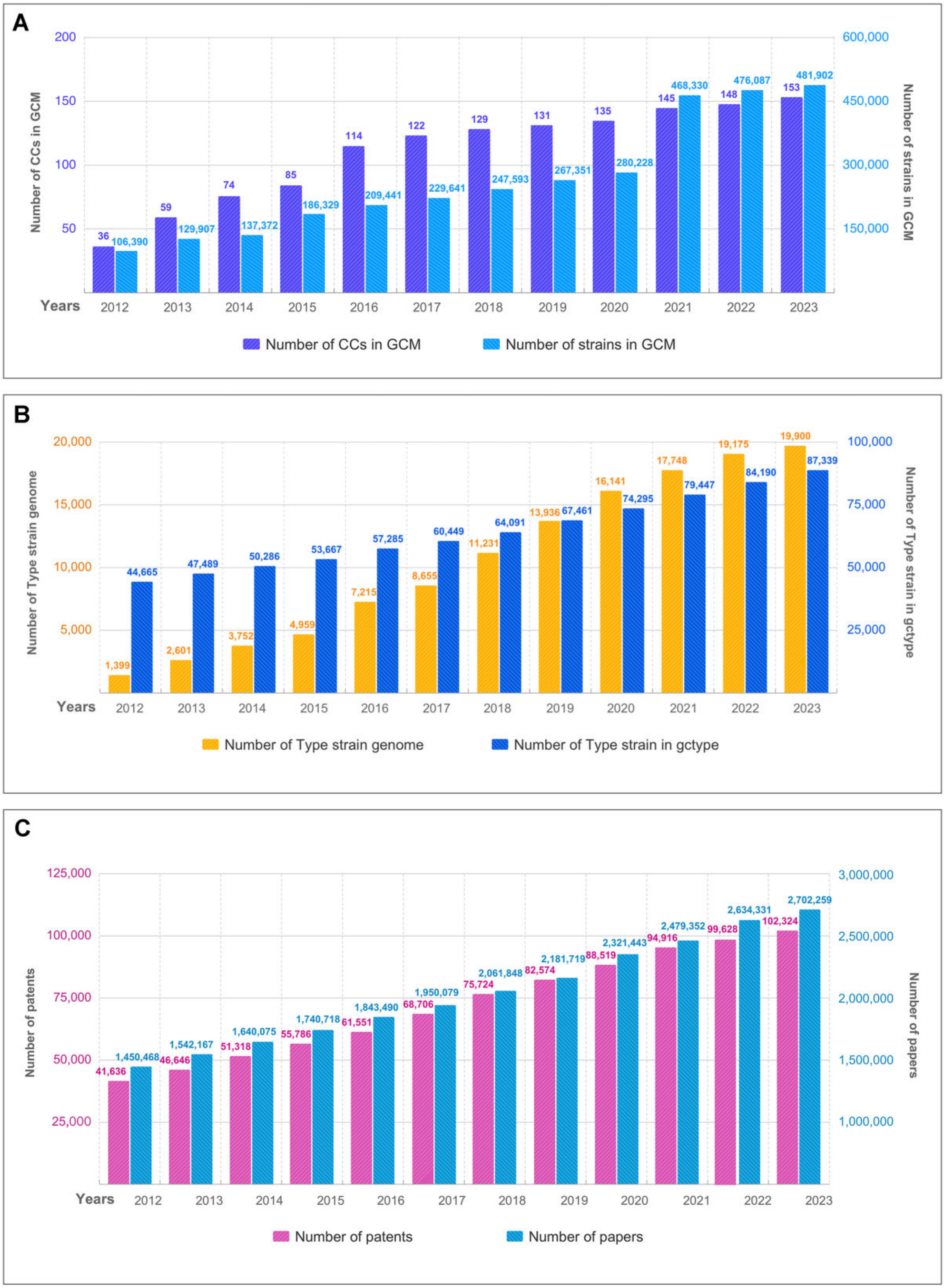


Figure 2. Increasing in data volume of GCM and gcType. (**A**) Number of culture collections and their holdings registered in GCM. (**B**) Number of type strains and their associated genomes in gctype database, the number of type strain has increased steadily while their genomes has increased faster. (**C**) Number of patents and papers related with strains registered in GCM.

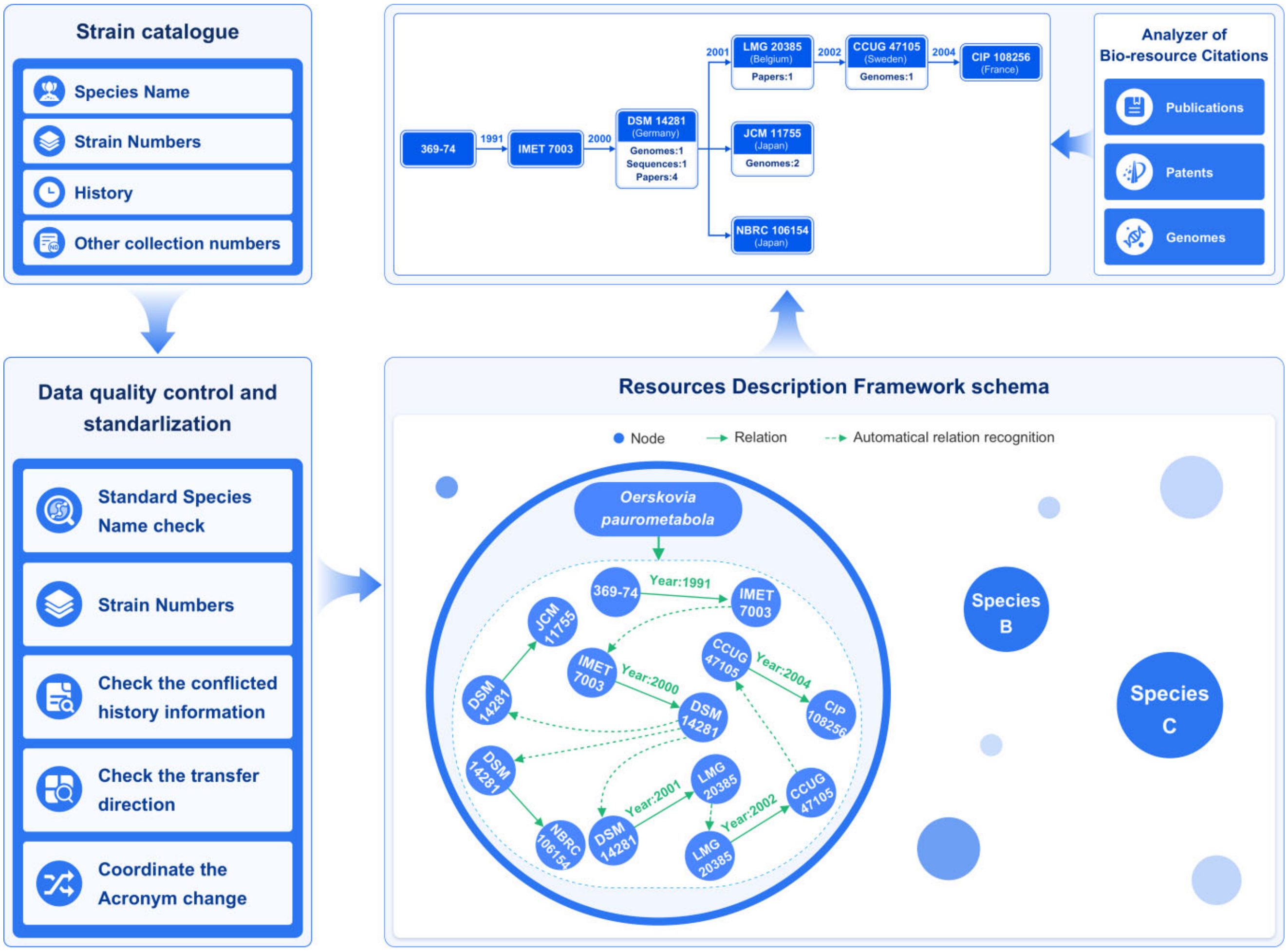


Figure 3. Schematic representation of the data processing workflow for the history tracking system. We firstly process values of data items 'species name', 'strain number', 'history' and 'other collection numbers' contained in strain catalogue into separated values. Then, these separated values are performed quality control and standardization. After that, the separated values and the 'predecessor-successor' relationships (e.g. IMET 7003 < 369–74) are imported into a RDF schema, where strain numbers and original depositors are considered as 'nodes'(blue nodes), and the 'predecessor-successor' relationships are depicted as 'relations' (green lines with years). Nodes with identical values are automatically combined(green dashed lines). Strains belong to the same species are grouped into a cluster (blue circle as a example). Then the RDF schema could automatically generate a history transfer flow above. Number of genomes and papers are extracted by ABC from public resources.

this system, data from the 'History' 'other collection numbers' data item of each strain are separated into individual strain numbers or original depositors. The separated values and the 'predecessor-successor' relationships are imported into an RDF schema, where strain numbers and original depositors are considered as 'nodes', and the 'predecessor-successor' relationships are depicted as 'relations' in the semantic web.

Nodes with identical values are automatically combined, and similar values are manually verified to check whether they refer to the same strains. In cases of conflicting relationships between nodes, verification against the original catalog or reference publications is performed. Changes in culture collection acronyms and strain numbers are also considered. Relationships between nodes automatically extend the history chains and form clusters within a species. The online interface provides multiple options to select cluster strain numbers, transfer routes, citations, and more. Strains are displayed alongside their publications, patents, and genome sequences, which are valuable for tracking utilization. This history and transfer routes between countries and citations provide information on the origin and transfer of strains together with the further utilization of strains. The implementation of DOIs allows for tracking by a global unique identify system in further citations. As a result, they offer support and a feasible solution for implementing the Nagoya Protocol and Access and Benefit Sharing (ABS).

Expansion of functional annotation of type strain genomes with manually curated metadata

In addition to protein annotation with COG (9), KEGG (10), Antibiotic Resistance Gene(ARGs) (11), Virulence Factors (12), and CAZy (13), the updated version significantly expands functional annotation to include mobile genetic elements, biosynthetic gene clusters (BGCs), defense systems, and prophages, exploring the diversity of isolated strains preserved in culture collections. Core genes analysis was performed using Roary (14) (v3.13.0, parameters: -i 70 -ap) at the genus level for genera with no less than 50 type strain genomes. The annotated results cover a wide range of taxonomic groups, serving as reference datasets for studies on gene distribution across wide taxonomic categories. As a result, we provide downloadable files for various kinds of annotation results.

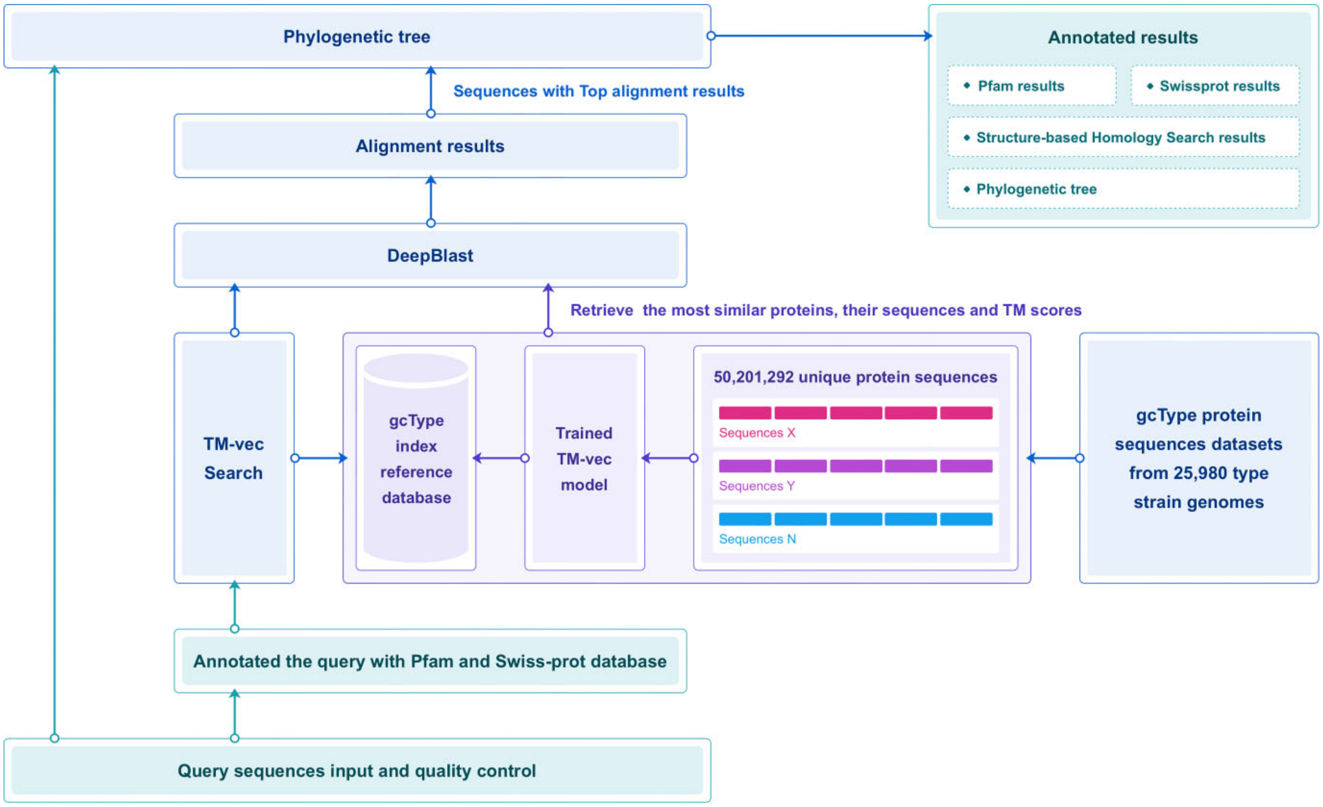

Phylogenetic tree
Sequences with Top alignment results
Alignment results
DeepBlast
Retrieve the most similar proteins, their sequences and TM scores
TM-vec Search
gcType index reference database
Trained TM-vec model
50,201,292 unique protein sequences
Sequences X
Sequences Y
Sequences N
gcType protein sequences datasets from 25,980 type strain genomes
Annotated the query with Pfam and Swiss-prot database
Query sequences input and quality control
Annotated results
Pfam results
Swissprot results
Structure-based Homology Search results
Phylogenetic tree

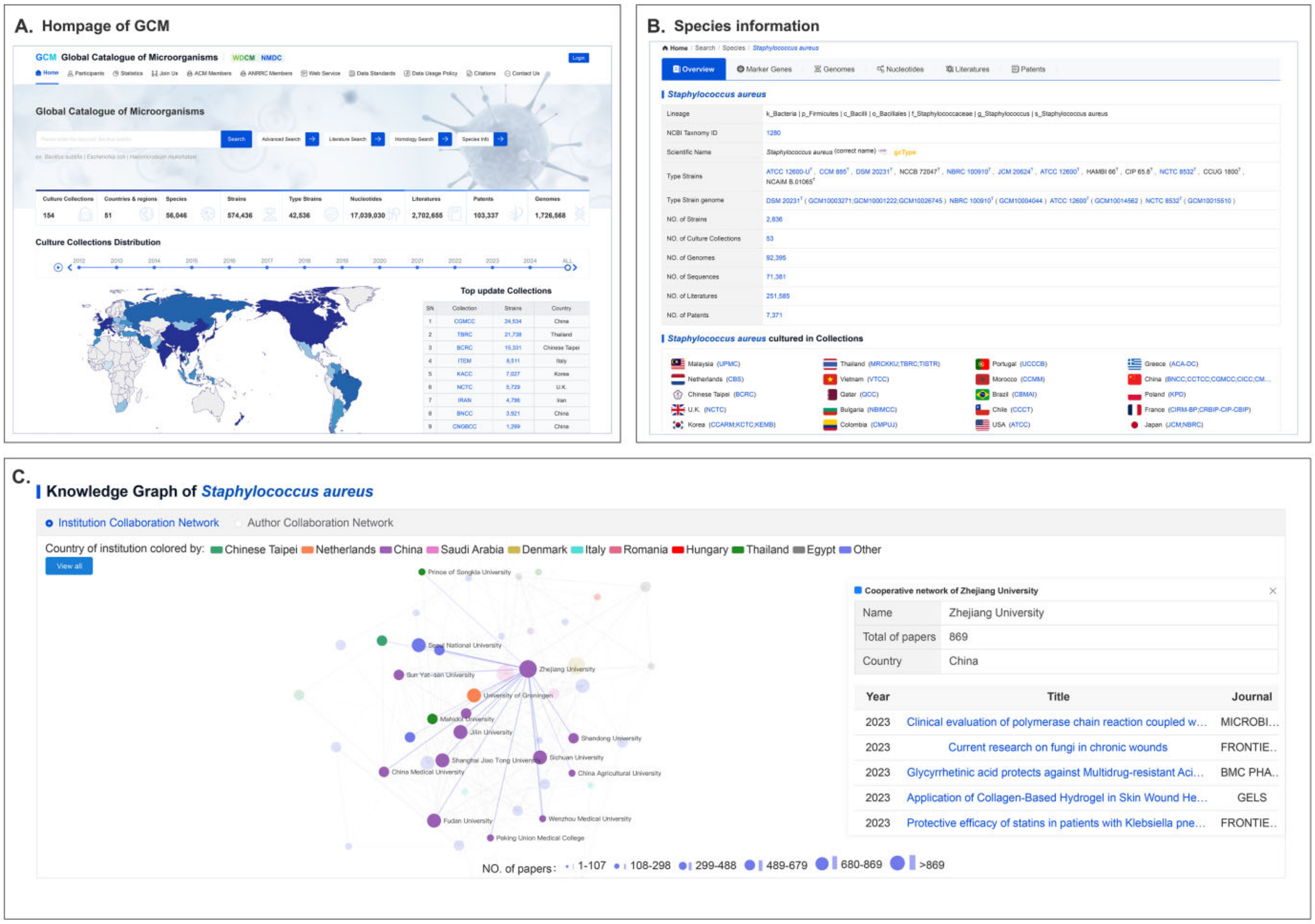
Figure 4. Schematic representation of remote protein searches implementation. We provide an online service on the website. User submitted query is firstly performed quality control and then annotated with Pfam and UniProt database. Then the query sequence will be searched using TM-Vec search model against an indexable reference databases from gcType's unique protein sequences. The top 10 sequences with TM values >0.5 are combined with query sequences to run DeepBLAST, producing structure-aware alignments. Finally, a phylogenetic tree is generated. The TM-score (25) is a metric for assessing the topological similarity of protein structures. TM-score has the value in (0, 1], where 1 indicates a perfect match between two structures. Following strict statistics of structures in the PDB, scores below 0.17 correspond to randomly chosen unrelated proteins where structures with a score higher than 0.5 assume generally the same fold in SCOP (26)/CATH (27).

MobileElementFinder (15) (v1.1.1) is used to detect mobile genetic elements, e.g. plasmid, insertion sequence (IS) and integrative and conjugative elements (ICE). Results are filtered with criteria: e_value ≤0.00001, identity >90% and coverage >80%. Plasmid and integron predictions are based on Platon (16) (v1.6) with default parameters and BacAnt (17) (v3.4.0, parameters: -c 80,80,80 -i 90,90,90), respectively. ARGs or virulence factors were considered to co-localize with an MGE if they shared a contig with an MGE gene in a nearby area (<10 kilo-bases). Defense systems are analyzed using Defense-finder (18) (v 1.2.0), and prophage sequences are predicted with geNomad (19) (v1.7.1, parameters: –enable-score-calibration and –max-fdr 0.02). Initial prophage sequences are filtered using a length cutoff of 2000 bp. CheckV (20) (v1.0.1) is utilized to assess completeness and refine prophage boundaries, followed by inspection of sequences with significant abnormalities, such as anomalous GC content. Secondary metabolite BGCs are predicted using AntiSMASH (v7.0.0) (21). To categorize BGC biochemistry, families generated by AntiSMASH are classified into eight broader categories: PKSI, PKS-NPR_Hybrids, PKSothers, NRPS, RiPPs, Terpene, Saccharides and Others, based on BiG-SCAPE (22). Pan-genome analysis at the genus level is conducted using Roary (14) (v3.13.0, parameters: -i 70 -ap).

Curated metadata, such as ecosystem and growth conditions (temperature, pH values), and collection locations, are essential features. Habitat information when species were first described can often be irregular, often falling to be assigned into formatted ontology categories automatically. Consequently, in the current version of gcType, different habitat values are manually assigned to an environment ontology (23) according to which level the values belong to. Growth conditions are categorized by temperature range, oxygen tolerance, Gram type, and cell shape of type strains. An online search interface allows users to select type strain genomes based on various metadata categories. These curated metadata, combined with richer annotation results, enhance comparative genomic studies and facilitate analysis of metabolic, biosynthetic, antibiotic resistance, and pathogenic features within specific groups.

Structure-based alignment method for searching the gcType reference protein database

Current annotation approaches primarily rely on alignment-based sequence homology methods (24). Similarly, sequence homology methods are commonly used to identify proteins with potentially similar functions. However, proteins perform their functions based on structure, and current sequence homology methods may fail to recognize proteins with remote evolutionary distances and low sequence similarity. Structure-based alignment methods are needed to align proteins with the gcType protein structure database to fully exploit the potential of type strain protein diversity. We utilize TM-Vec and



Culture Collections	Countries & regions	Species	Strains	Type Strains	Nucleotides	Literatures	Patents	Genomes
154	51	56,046	574,436	42,536	17,039,030	2,702,655	103,337	1,726,568

SN	Collection	Strains	Country
1	CGMCC	24,534	China
2	TBRC	21,738	Thailand
3	BCRC	15,331	Chinese Taipei
4	ITEM	8,511	Italy
5	KACC	7,027	Korea
6	NCTC	5,729	U.K.
7	IRAN	4,796	Iran
8	BNCC	3,921	China
9	CNGBCC	1,289	China

Lineage	k_Bacteria | p_Firmicutes | c_Bacilli | o_Bacillales | f_Staphylococcaceae | g_Staphylococcus | s_Staphylococcus aureus
NCBI Taxonomy ID	1280
Scientific Name	*Staphylococcus aureus* (correct name) gcType
Type Strains	ATCC 12600-U^T, CCM 885^T, DSM 20231^T, NCCB 72047^T, NBRC 100910^T, JCM 20624^T, ATCC 12600^T, HAMBI 66^T, CIP 65.8^T, NCTC 8532^T, CCUG 1800^T, NCAIM B.01065^T
Type Strain genome	DSM 20231^T (GCM10003271;GCM10001222;GCM10026745) NBRC 100910^T (GCM10004044) ATCC 12600^T (GCM10014562) NCTC 8532^T (GCM10015510)
NO. of Strains	2,836
NO. of Culture Collections	53
NO. of Genomes	92,395
NO. of Sequences	71,381
NO. of Literatures	251,585
NO. of Patents	7,371

Staphylococcus aureus cultured in Collections

Malaysia (UPMC)	Thailand (MRCKKU;TBRC;TISTR)	Portugal (UCCCB)	Greece (ACA-DC)
Netherlands (CBS)	Vietnam (VTCC)	Morocco (CCMM)	China (BNCC;CCTCC;CGMCC;CICC;CM...
Chinese Taipei (BCRC)	Qatar (QCC)	Brazil (CBMAI)	Poland (KPD)
U.K. (NCTC)	Bulgaria (NBIMCC)	Chile (CCCT)	France (CIRM-BP;CRBIP-CIP-CBIP)
Korea (CCARM;KCTC;KEMB)	Colombia (CMPUJ)	USA (ATCC)	Japan (JCM;NBRC)

C.
Knowledge Graph of Staphylococcus aureus
● Institution Collaboration Network
Author Collaboration Network
Country of institution colored by: Chinese Taipei Netherlands China Saudi Arabia Denmark Italy Romania Hungary Thailand Egypt Other
View all
Prince of Songkla University
Seoul National University
Zhejiang University
Sun Yat-sen University
University of Groningen
Mahidol University
Jilin University
Shandong University
Shanghai Jiao Tong University
Sichuan University
China Medical University
China Agricultural University
Fudan University
Wenzhou Medical University
Peking Union Medical College
NO. of papers: 1-107 108-298 299-488 489-679 680-869 >869

Cooperative network of Zhejiang University

Name	Zhejiang University
Total of papers	869
Country	China

Year	Title	Journal
2023	Clinical evaluation of polymerase chain reaction coupled w...	MICROBI...
2023	Current research on fungi in chronic wounds	FRONTIE..
2023	Glycyrrhetinic acid protects against Multidrug-resistant Aci...	BMC PHA..
2023	Application of Collagen-Based Hydrogel in Skin Wound He...	GELS
2023	Protective efficacy of statins in patients with Klebsiella pne...	FRONTIE..

Figure 5. Features in the GCM and gcType web interfaces (**A**) The homepage provides multiple search options and displays the submission collections through a world map. (**B**) The overview page of 'species' shows the integrated information for a species, which includes type strains, marker genes, genomes and gene sequences, related articles and patents. (**C**) Institutes and authors' information of publications could be extracted to construct the cooperation network. The dots represent individual author or institute, the lines between dots indicate the joint publications between authors or institutes.

DeepBLAST (25) to construct the alignment pipeline (Figure 4). TM-Vec is first employed to build indexable reference databases from gcType's unique protein sequences, encompassing datasets from 25 980 type strain genomes and covering 50 million unique protein sequences, including 2.4 million previously unannotated proteins. Query proteins are initially annotated using Pfam (28) and UniProt (29) databases. The TM-Vec search model is then applied against the gcType reference database to identify proteins with the highest predicted structural similarity. The top 10 sequences with TM values > 0.5, along with their sequences and TM scores, are retrieved. The retrieved sequences are combined with query sequences to run DeepBLAST, producing structure-aware alignments. Finally, a phylogenetic tree is generated using Mafft (30) and Fasttree (31), and results are reported alongside Pfam and UniProt annotation feedback.

Database interfaces

Advanced search and metadata search

Users of GCM and gcType can initiate searches through various options on the homepages (Figure 5A). Direct searches by strain numbers, species names, and genome accession numbers are available. Additionally, advanced search options allow for combinations of search criteria. The literature search facilitates the identification of specific strains or species mentioned in article titles or abstracts, indexed by the ABC data mining tools, e.g strains or species which may have the ability to for L-alanine biosynthesis. For the GCM database, homology searches enable queries against an indexed 16S database of all GCM strains. In the gcType database, users can search not only the indexed 16S database of all type strains but also the complete genes or protein sequences of all type strains. Moreover, we offer a metadata search based on curated metadata, including temperature range, oxygen tolerance, Gram staining, cell shapes, and biomes where type strains were collected.

Integrating various sources of information for a species in the GCM database

Species information in GCM is consolidated on a species page. An overview page (Figure 5B) presents general details about the species, including type strains, non-type strain lists, preservation information, collection locations, and history. Standardized nomenclature information were integrated with LPSN (32) and FungalNames (33) Several labels provide access to marker genes, genomes, and nucleotide sequences from both type and non-type strains. The data mining results for publications and patents are also accessible via separate labels, listing publications and patents related to the species or strains, with strain numbers extracted from these sources. We employ the RDF (Resource Description Framework) schema to manage data on publication authors, institutions, countries, titles, years, and other relevant information. This approach enables us to construct collaboration networks based on publication data (Figure 5C).

Type strain genomes with rich annotations

Type strain genomes can be accessed through the GCM database or directly searched in gcType. The ‘data’ section of the gcType database includes several valuable lists: valid published species, type strains categorized by species name and culture collection, type strain genomes, and 16S rDNA sequences from type strains. Functional annotation results for all type strain genomes can be browsed and searched in the ‘function’ section, which includes core genes by genus, prophages, defense system categories, mobile genetic elements, antibiotic resistance genes, and virulence factors. Summary statistics for each type of functional annotation are also provided. For type strain genomes, metadata about the sequencing project and all annotated results are displayed. Jbrowser (34) and gcviewer (35) were used to visualize genomic components in linear and circular formats.

Future directions

GCM and gcType offer unique, comprehensive archives of microbial resources preserved in culture collections, along with high-quality type strain genomes featuring curated metadata and detailed annotation information. These datasets are accessible via dedicated websites. We envision that GCM and gcType, alongside the WDCM database, will serve as invaluable resources for exploring mBRC resources and facilitating in-depth genomic research from various perspectives.

As artificial intelligence (AI) rapidly transforms microbial research, it shows immense potential for knowledge mining across vast volumes of publications. Currently, our publication data mining tools utilize text mining and ontology to automatically extract information from publications and patents. However, efficiency is limited because text mining using ontology often generates false positive/negative or inaccurate information. As a result, manual curation with the original articles to verify ambiguous information remains necessary. With over 3 million publications related to microbial resources and technologies, these can serve as valuable resources for training large language models, significantly enhancing knowledge mining efficiency and content. Furthermore, given the substantial number of unannotated genes and proteins in type strains, we are exploring methods to improve remote homology detection, thereby increasing the percentage of annotated genes and proteins using AI techniques.

In conclusion, WDCM will continue to leverage cutting-edge data mining technologies to integrate microbial resources from global mBRCs, exploring the rich diversity preserved on the Earth and meeting the needs of scientific communities worldwide.

Data availability

There are no access restrictions for academic use of the platform. Access is freely available at gcm.wdcm.org and gctype.wdcm.org. Both GCM and gcType provide free access to API modules without registration.

Acknowledgements

WDCM thanks the supports by WFCC Executive board. WDCM acknowledges the contributions of all participating collections to the GCM and GCM 10K project.

Funding

National key Research Program of China [2022YFC2602205]; Strategic Priority Research Program of the Chinese Academy of Sciences [XDB0830000]; 14th Five-year Informatization Plan of Chinese Academy of Sciences [CAS-WX2021SF-0202]; Key Research Program of the Chinese Academy of Sciences [KFZD-SW-219]; National Science Foundation for International Cooperation [82161148010]. Funding for open access charge: National key Research Program of China [2022YFC2602305].

Conflict of interest statement

None declared.

References

1. Smith,D., McCluskey,K. and Stackebrandt,E. (2014) Investment into the future of microbial resources: culture collection funding models and BRC business plans for biological resource centres. *Springerplus*, **3**, 81.
2. De Vero,L., Boniotti,M.B., Budroni,M., Buzzini,P., Cassanelli,S., Comunian,R., Gullo,M., Logrieco,A.F., Mannazzu,I., Musumeci,R., *et al.* (2019) Preservation, characterization and exploitation of microbial biodiversity: the perspective of the Italian network of Culture collections. *Microorganisms*, **7**, 685.
3. Jiao,J.Y., Abdugheni,R., Zhang,D.F., Ahmed,I., Ali,M., Chuvochina,M., Dedysh,S.N., Dong,X., Göker,M., Hedlund,B.P., *et al.*2024) Advancements in prokaryotic systematics and the role of Bergey’s International Society for Microbial Systematicsin addressing challenges in the meta-data era. *Natl. Sci. Rev.*, **11**, nwae168.
4. Lh,W., Sun,Q.L., Desmeth,P., Sugawara,H., Zh,X., Mccluskey,K., Smith,D., Alexander,V., Lima,N., Ohkuma,M., *et al.* (2016) World data centre for microorganisms: an information infrastructure to explore and utilize preserved microbial strains worldwide. *Nucleic Acids Res.*, **45**, D611–D618.
5. Marino,M., Jacopo,T., Gian,P.A., Maria,S.B., Valentina,B., Annamaria,B., Maria,B.B., Marilena,B., Pietro,B., Stefania,C., *et al.*2024) Treasures of Italian microbial culture collections: an overview of preserved biological resources, offered services and know-how, and management. *Sustainability*, **16**, 3777.
6. Becker,P., Bosschaerts,M., Chaerle,P., Daniel,H.M., Hellemans,A., Olbrechts,A., Rigouts,L., Wilmotte,A. and Hendrickx,M. (2019) Public Microbial Resource Centers: key hubs for findable, accessible, interoperable, and reusable (FAIR) microorganisms and genetic materials. *Appl. Environ. Microb.* **85**, e01444-18.
7. Verslyppe,B., De Smet,W., De Baets,B., De Vos,P. and Dawyndt,P. (2011) Make Histri: reconstructing the exchange history of bacterial and archaeal type strains. *Syst. Appl. Microbiol.*, **34**, 328–336.
8. Bert,V., Renzo,K., Wim De,S., Bernard De,B., De V,P. and Peter,D. (2010) Microbiological Common Language (MCL): a standard for electronic information exchange in the Microbial Commons. *Res. Microbiol.*, **161**, 439–445.

9. Galperin,M.Y., Kristensen,D.M., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2019) Microbial genome analysis: the COG approach. *Brief. Bioinform.*, **20**, 1063–1070.
10. Kanehisa,M., Furumichi,M., Sato,Y., Kawashima,M. and Ishiguro-Watanabe,M. (2023) KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.*, **51**, D587–D592.
11. Alcock,B.P., Raphyenya,A.R., Lau,T.T.Y., Tsang,K.K., Bouchard,M., Edalatmand,A., Huynh,W., Nguyen,A.V., Cheng,A.A., Liu,S., *et al.* (2020) CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, **48**, D517–D525.
12. Chen,L., Yang,J., Yu,J., Yao,Z., Sun,L., Shen,Y. and Jin,Q. (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.*, **33**, D325–D328.
13. Lombard,V., Golaconda Ramulu,H., Drula,E., Coutinho,P.M. and Henrissat,B. (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.*, **42**, D490–D495.
14. Andrew,J.P., Carla,A.C., Vanessa,K.W., Sandra,R., Matthew,T.G., Maria,F., Daniel,F., Jacqueline,A.K. and Julian,P. (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, **31**, 3691–3693.
15. Johansson,M.H.K., Bortolaia,V., Tansirichaiya,S., Aarestrup,F.M., Roberts,A.P. and Petersen,T.N. (2021) Detection of mobile genetic elements associated with antibiotic resistance in Salmonella enterica using a newly developed web tool: mobileElementFinder. *J. Antimicrob. Chemother.*, **76**, 101–109.
16. Hua,X., Liang,Q., Deng,M., He,J., Wang,M., Hong,W., Wu,J., Lu,B., Leptihn,S., Yu,Y., *et al.* (2021) BacAnt: A combination annotation server for bacterial DNA sequences to identify antibiotic resistance genes, integrons, and transposable elements. *Front. Microbiol.*, **12**, 649969.
17. Xiaoting,H., Qian,L., Min,D., Jintao,H., Meixia,W., Wenjie,H., Jun,W., Bian,L., Sebastian,L., Yunsong,Y., *et al.* (2021) BacAnt: a combination annotation server for bacterial DNA sequences to identify antibiotic resistance genes, integrons, and transposable elements. *Front. Microbiol.* **12**, 649969.
18. Tesson,F., Hervé,A., Mordret,E., Touchon,M., d'Humières,C., Cury,J. and Bernheim,A. (2022) Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat. Commun.*, **13**, 2561.
19. Camargo,A.P., Roux,S., Schulz,F., Babinski,M., Xu,Y., Hu,B., Chain,P.S.G., Nayfach,S. and Kyrpides,N.C. (2024) Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.*, **42**, 1303–1312.
20. Nayfach,S., Camargo,A.P., Schulz,F., Eloe-Fadrosh,E., Roux,S. and Kyrpides,N.C. (2021) CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.*, **39**, 578–585.
21. Blin,K., Shaw,S., Augustijn,H.E., Reitz,Z.L., Biermann,F., Alanjary,M., Fetter,A., Terlouw,B.R., Metcalf,W.W., Helfrich,E.J.N., *et al.* (2023) antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.*, **51**, W46–W50.
22. Navarro-Muñoz,J.C., Selem-Mojica,N., Mullowney,M.W., Kautsar,S.A., Tryon,J.H., Parkinson,E.I., De Los Santos,E.L.C., Yeong,M., Cruz-Morales,P., Abubucker,S., *et al.* (2020) A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.*, **16**, 60–68.
23. Buttigieg,P.L., Morrison,N., Smith,B., Mungall,C.J., Lewis,S.E. and ENVO ConsortiumENVO Consortium. (2013) The environment ontology: contextualising biological and biomedical entities. *J. Biomed. Semantics*, **4**, 43.
24. Flamholz,Z.N., Biller,S.J. and Kelly,L. (2024) Large language models improve annotation of prokaryotic viral proteins. *Nat. Microbiol.*, **9**, 537–549.
25. Hamamsy,T., Morton,J.T., Blackwell,R., Berenberg,D., Carriero,N., Gligorijevic,V., Strauss,C.E.M., Leman,J.K., Cho,K. and Bonneau,R. (2024) Protein remote homology detection and structural alignment using deep learning. *Nat. Biotechnol.*, **42**, 975–985.
26. Andreeva,A., Kulesha,E., Gough,J. and Murzin,A.G. (2020) The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.*, **48**, D376–D382.
27. Sillitoe,I., Bordin,N., Dawson,N., Waman,V.P., Ashford,P., Scholes,H.M., Pang,C.S.M., Woodridge,L., Rauer,C., Sen,N., *et al.* (2021) CATH:CATH: increased structural coverage of functional space. *Nucleic Acids Res.*, **49**, D266–D273.
28. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J., *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
29. The UniProt Consortium (2023) UniProt: the Universal Protein knowledgebase in 2023. *Nucleic Acids Res.*, **51**, D523–D531.
30. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
31. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
32. Meier-Kolthoff,J.P., Carbasse,J.S., Peinado-Olarte,R.L. and Göker,M. (2022) TYGS and LPSN: a database tandem for fast and reliable genome-based classification and nomenclature of prokaryotes. *Nucleic Acids Res.*, **50**, D801–D807.
33. Wang,F., Wang,K., Cai,L., Zhao,M., Kirk,P.M., Fan,G., Sun,Q., Li,B., Wang,S., Yu,Z., *et al.* (2023) Fungal names: a comprehensive nomenclatural repository and knowledge base for fungal taxonomy. *Nucleic Acids Res.*, **51**, D708–D716.
34. Skinner,M.E., Uzilov,A.V., Stein,L.D., Mungall,C.J. and Holmes,I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
35. Grin,I. and Linke,D. (2011) GCView: the genomic context viewer for protein homology searches. *Nucleic Acids Res.*, **39**, W353–W6.

Received: September 14, 2024. **Revised:** October 16, 2024. **Editorial Decision:** October 17, 2024. **Accepted:** October 22, 2024