BMC Genomics

DATABASE Open Access

Global catalogue of microorganisms (gcm): a comprehensive database and information retrieval, analysis, and visualization system for microbial resources

Linhuan Wu[1,2], Qinglan Sun[1,2], Hideaki Sugawara[4], Song Yang[1,2], Yuguang Zhou[1], Kevin McCluskey[5], Alexander Vasilenko[6], Ken-Ichiro Suzuki[7], Moriya Ohkuma[8], Yeonhee Lee[9], Vincent Robert[10], Supawadee Ingsriswang[11], François Guissart[3], Desmeth Philippe[3*] and Juncai Ma[1,2*]

Abstract

Background: Throughout the long history of industrial and academic research, many microbes have been isolated, characterized and preserved (whenever possible) in culture collections. With the steady accumulation in observational data of biodiversity as well as microbial sequencing data, bio-resource centers have to function as data and information repositories to serve academia, industry, and regulators on behalf of and for the general public. Hence, the World Data Centre for Microorganisms (WDCM) started to take its responsibility for constructing an effective information environment that would promote and sustain microbial research data activities, and bridge the gaps currently present within and outside the microbiology communities.

Description: Strain catalogue information was collected from collections by online submission. We developed tools for automatic extraction of strain numbers and species names from various sources, including Genbank, Pubmed, and SwissProt. These new tools connect strain catalogue information with the corresponding nucleotide and protein sequences, as well as to genome sequence and references citing a particular strain. All information has been processed and compiled in order to create a comprehensive database of microbial resources, and was named Global Catalogue of Microorganisms (GCM). The current version of GCM contains information of over 273,933 strains, which includes 43,436bacterial, fungal and archaea species from 52 collections in 25 countries and regions.
A number of online analysis and statistical tools have been integrated, together with advanced search functions, which should greatly facilitate the exploration of the content of GCM.

Conclusion: A comprehensive dynamic database of microbial resources has been created, which unveils the resources preserved in culture collections especially for those whose informatics infrastructures are still under development, which should foster cumulative research, facilitating the activities of microbiologists world-wide, who work in both public and industrial research centres. This database is available from http://gcm.wfcc.info.

Keywords: Microbial resources, Data management, Data sharing

* Correspondence: Philippe.DESMETH@belspo.be; ma@im.ac.cn
[3]Belgian Coordinated Collections of Micro-organisms Programme, Belgian Science Policy Office, avenue Louise, 231 1050 Brussels, Belgium
[1]Institute of Microbiology, Chinese Academy of Sciences, Beijing, China
Full list of author information is available at the end of the article

Background

Microbial culture collections play an important and essential role in collecting, maintaining, and distributing quality assured living microbial strains. The Word Federation for Culture Collections (WFCC) is a Multidisciplinary Commission of the International Union of Biological Sciences (IUBS) and a Federation within the International Union of Microbiological Societies (IUMS).The WFCC promotes the interests of culture collections, develops shared resources, and organizes the International Conference on Culture Collections every three years. As one of its longstanding activi ties, the WFCC participated in the development of the WFCC World Data Centre for Microorganisms (WDCM) in the late 1960s [1]. With additional input from the United Nations Educational, Scientific and Cultural Organization Microbial Resources Centers (MIRCEN) project, the WDCM was maintained as the WFCC-MIRCEN WDCM and become accessible as an internet page in 1997. The WDCM serves as the data center of the WFCC and provides an important information resource for all microbiological activities. Additionally, the WDCM acts as a coordination center for data activities among WFCC members. As one of the main databases in WDCM, CCINFO (Culture Collection INFOrmation database) lists 652 culture collections from 70 countries maintain more than 1.9 million strains. (http://www.wfcc.info/ccinfo/, accessed 12/3/2013).

Increasing demands on culture collections for authenticated, reliable biological material and associated information were accompanied by the growth of biotechnology and basic science. The WFCC guidelines recommend that every collection publish an online or printed catalogue regularly, both to disseminate information about strains and to promote scientific and industrial usage of materials held in their collection. However, according to the available statistics, fewer than one-sixth of collections registered in CCINFO post their catalogue online and this greatly hinders the visibility and hence the accessibility of strains in these collections without public electronic catalogs.

To help all collections establish an online catalog, the WDCM has constructed a data management system and a global catalogue to organize, make public, and explore the data resources of its member collections. This data management system, called the WFCC Global Catalogue of Microorganisms (GCM) is a scalable, reliable, dynamic and user-friendly system that helps culture collections manage, disseminate and share the information related to their holdings. It also provides a uniform interface for the scientific and industrial communities to access the comprehensive microbial resource information.

Construction and content

Data sources

The Global Catalog of Microorganisms database contains information from a variety of sources:

- Information provided by culture collection staff
- Data from public data sources such as the US National Library of Medicine (PubMed) and the Patent database
- Links to external databases
- Tools for bioinformatics analysis including a search engine to enhance exploration of GCM data.

By the end of August 2013, the GCM contains strain information from 52 collections (Table 1) located in 25 different countries and regions. While the project is still in its construction phase, preliminary statistics describing the participating collections are unique and informative (Table 2).

The GCM implements the WDCM Minimum Data Sets (MDS) and Recommended Data Sets (RDS) based on widely applied standards such as the OECD Best Practice Guidelines for Biological Resource Centres [2], the Microbial Information Network Europe (MINE) [3], as well as the Common Access to Biological Resources and Information (CABRI) [4]. A detailed description, together with examples of 15 WDCM MDS items can be found at http://gcm.wfcc.info/datastandards/index.jsp (last accessed 12/3/2013).To build the GCM, each participating collection transferred their catalogue information by one of several pathways. Some collections sent Excel or XML files while others provided direct access to their database files. WDCM integrated the data into a global dataset, processed the data to identify relationships among collections (for example strains held in multiple collections), and published the strain information on the GCM web page (http://gcm.wfcc.info). Because not all collections use the same data schema, some of the data items provided by culture collection staff were manually reclassified by GCM staff to allow for an easier integration of catalogue information.

Publications concerning strains are collected from PubMed using both strain number and species name for keyword queries. Nucleotide sequences are extracted from GenBank [5], protein sequence data are collected from UniProt [6], and information about protein 3D structure are extracted from the PDB database [7]. Genome sequencing information is collected from NCBI Microbial Genomes Resources (NCBI).

Organization of data

The GCM database contains the following fields for each strain entry: strain number, other collection numbers, name, organism type, history of deposition, date of

Table 1 Participant list of GCM collections

Acronym	Full name	Country
BCC	BIOTEC Culture Collection	Thailand
BCCM/DCG	BCCM Diatom Collection Gent	Belgium
BCCM/IHEM	Belgian Coordinated Collections of Microorganisms / IHEM Fungi colleciton	Belgium
BCCM/LMBP	Belgian Coordinated Collections of Microorganisms / LMBP Plasmid Collection	Belgium
BCCM/LMG	Belgian Coordinated Collections of Microorganisms/ LMG Bacteria Collection	Belgium
BCCM/MUCL	Mycotheque de l'Universite catholique de Louvain	Belgium
BCCM/ULC	BCCM/ULC Culture Collection of (sub)polar cyanobacteria	Belgium
BCRC	Bioresource Collection and Research Center	Chinese Taipei
BIM	Belarusian Collection of non-pathogenic microorganisms	Belarus
CBS	Centraalbureau voor Schimmelcultures, Filamentous fungi and Yeast Collection	Netherlands
CCAP	Culture Collection of Algae and Protozoa	U.K.
CCARM	Culture Collection of Antimirobial Resistant Microorganisms	Korea
CCCryo	Culture Collection of Cryophilic Algae	Germany
CECT	Coleccion Espanola de Cultivos Tipo	Spain
CGMCC	China General Microbiological Culture Collectio Center	China
CIP	The Collection of the Institut Pasteur	France
CIRM-CF	Centre International de Ressources Microbiennes - Champignons Filamenteux	France
CIRM-CFBP	Centre International de Ressources Microbiennes - Levures (CLBP)	France
CIRM-Levures	Centre International de Ressources Microbiennes - Levures	France
CM-CNRG	Coleccion de Microorganismos del Centro Nacional de Recursos Geneticos	Mexico
CVCM	Centro Venezolano de Colecciones de Microorganismos	Venezuela
CWU-MACC	Herbarium of Kharkov University (CWU) – Micro Algae Cultures Collection	Ukraine
DMic	Medical importance fungi culture collection	Argentina
DSMZ	Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH	Germany
FACHB	Freshwater Algae Culture Collection, Chinese Academy of Sciences	China
FGSC	Fungal Genetics Stock Center	USA
Fiocruz-CLIOC	Coleção de Leishmania do Instituto Oswaldo Cruz	Brazil
GDMCC	Guangdong Culture Collection Centre of Microbiology	China
HPKTCC	Helicobacter pylori Korean Type Culture Collection	Korea
IMI(CABI)	CABI Genetic Resource Collection	U.K.
ITDI	Industrial Technology Development Institute Microbial Culture Collection	Philippines
ITM	Belgian Coordinated Collections of Microorganisms Mycobacterial Culture Collection	Belgium
JCM	Japan Collection of Microorganisms	Japan
KCTC	KCTC Korean Collection for Type Cultures	Korea
KEMB	Korea national Environmental Microorganisms Bank	Korea
KMMCC	Korea Marine Microalgae Culture Center	Korea
LEF	Korea Lichen & Allied Bioresource Center	Korea
LIPIMC	Lembaga Ilmu Pengetahuan Indonesia , Indonesian Institute for Sciences	Indonesia
MCC-MNH	Microbial Culture Collection - Museum of Natural History, Museum of Natural History (MNH)	Philippines
NBRC	NITE Biological Resource Center	Japan
PNCM	Philippine National Collection of Microorganisms	Philippines
PVGB	Plant Virus GenBank	Korea
TISTR	TISTR Culture Collection, Bangkok MIRCEN	Thailand
UCCAA	Ukrainian Collection of Cholera Aetiological Agents O1 and non O1 serogroups	Ukraine

Table 1 Participant list of GCM collections *(Continued)*

UCDFST	Phaff Yeast Culture Collection	USA
UL	The UNILAB Clinical Culture Collection, United Laboratories	Philippines
UMinho-MUM	Micoteca da Universidade do Minho	Portugal
UOA/HCPF	UOA/HCPF University of Athens/Hellenic Collection of Pathogenic Fungi	Greece
UPCC	Natural Sciences Research Institute Culture Collection	Philippines
UPMC	MICROBIAL CULTURE COLLECTION UNIT	MALAYSIA
VKM	All-Russian Collection of Microorganisms	Russia
VTCC	Vietnam Type Culture Collection	Vietnam

isolation, isolation sources, geographic origin, status, optimal temperature for growth, minimum temperature for growth, maximum temperature for growth, medium, application, and published citations to the use of the strain. In addition to these WFCC MDS entries, the GCM contains extensive citation, patent, and gene or genome information related to each strain. All of this information is available from the strain information page for each strain. A schema of the data flow of GCM is shown in Figure 1.

Strains belonging to the same species as well as subspecies are automatically associated to form a species page (Figure 2). A taxonomic tree of species 2000 [8] is generated to serve as a reference for taxonomic identification. Type strains, indicated by their collections are listed on species page. Data on individual strains are organized by culture collections location, type of strain, isolation sources, and genus and species as well. As a result, all data can be retrieved through the browse option provided in the web server according to these properties.

Metagenome and Microbes Environmental Ontology (Hiroshi Mori [9]) which is an ontology about microbial environment was used for text mining of values of isolation sources. The text contained in this data item was automatically compared with the terms of MEO and then sorted into 13 different categories such as soil, microbial-mat/Biofilm, or host-associated, among others (Table 3). For the values that could not be automatically assigned to a specific category, manual curation is required. Data concerning environmental habitats of the isolates can provide important information about the diversity of organism types that are related with certain isolation source types.

About 48% of the strains have geographic information and these strains are from 164 different countries or

Table 2 Summary of GCM strain data

Organism type	Species number	Strain number	Type strain	Sequences	Publications	Patents
Antibody	7	33	0	0	0	0
Phage	181	239	0	0	1	0
Virus	33	296	0	0	0	0
Cyanobacteria	134	287	0	178	0	0
Protozoa	236	754	0	0	0	0
Actinomycetes	842	1490	0	271	192	9
Archaea	1410	3273	1165	2176	1573	48
Microalgae	1820	5495	4	2	1	1
Plasmid	2030	2030	0	0	5	9
Yeast	3668	34907	4796	54773	2089	98
Bacteria	13714	101395	14233	29304	10975	268
Fungi	18537	121548	29916	94348	1960	65
Diatom	19	242	0	0	0	0
Mycobacteria	50	214	0	0	0	0
Other/Rotifera	755	1730	0	0	0	0
Total	43436	273933	50114	181052	16796	498

Information was as submitted by individual collections. Sequence, publication, and patent data were extracted from Genbank, Pubmed, Patent database using the strain numbers.

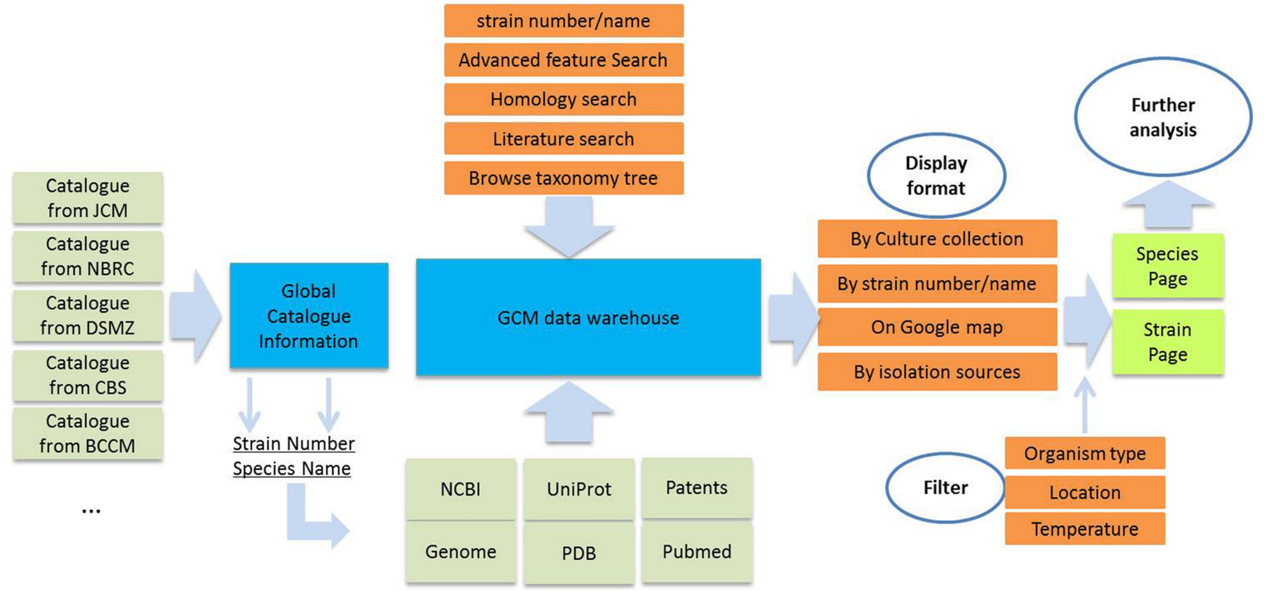


Figure 1 Scheme of the workflow of GCM. Catalogue information from each of collection, shown on the left, is used to construct the framework of the global catalogue database. Species name and strain numbers are collected from the catalogue information, and are further used to identify and extract information from public database such as Genbank, NCBI Genome, SwissProt, PDB, Pubmed and the patent database. The data warehouse is built in a SQL database, and can be accessed via a web interface through different search options. Search results may be displayed in different formats allowing users to refine the results by using filters. The final results are displayed either as a strain page or they can be gathered into a species page depending on the query. BLAST and ClustalW are provided for further analysis of the results.

regions. Data on the geographic origin of isolates (Table 4) is complementary to the habitat and can provide useful information on relative biodiversity and sampling efforts for different countries and regions. These data will ultimately be integrated into the Global Biodiversity Information Facility (GBIF) database through planned activities of GCM (Éamonn [10]).

Data quality control

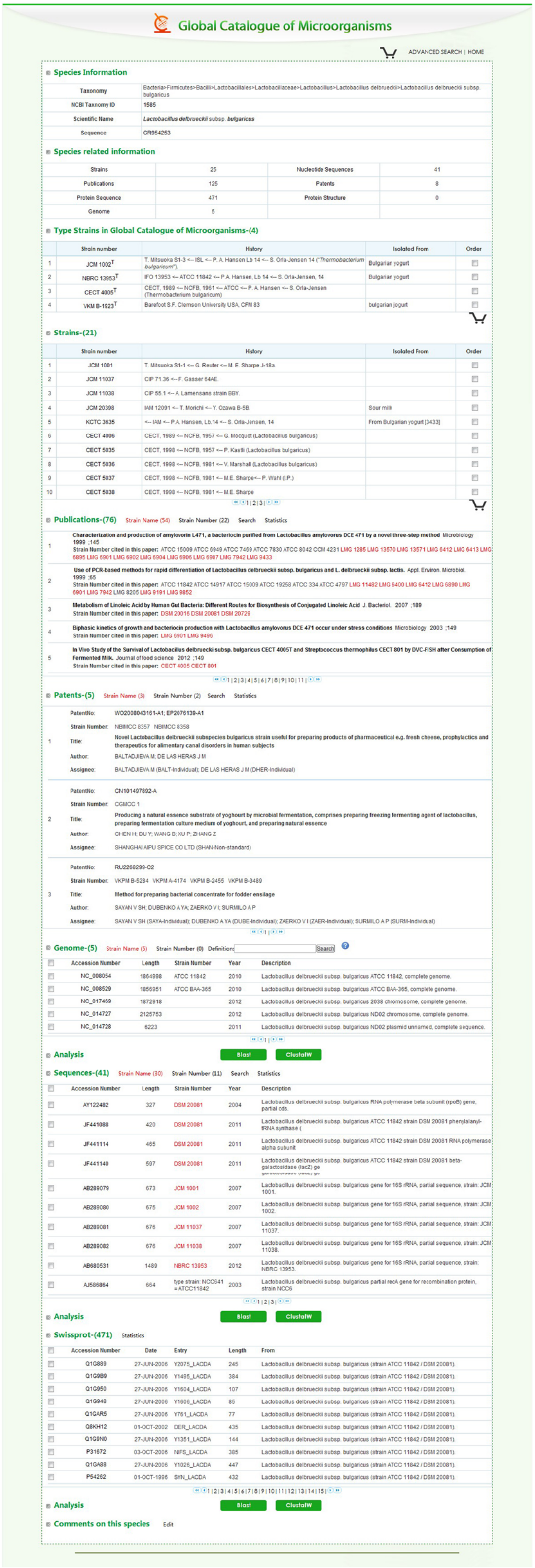
Because original catalogue data are sometimes non-validated, quality control measures are necessary before data can be published in GCM online. The most frequent quality problem is the misspelling of species name or non-standard naming of species. For example, "*Absidiapsychrophilia*" was wrongly spelled as "*Absidiapsychrophila*" in certain collections. In such cases, GCM uses standard microbial nomenclature databases to perform a quality check of its taxonomic data. Databases include the List of Prokaryotic names with Standing in Nomenclature (LPSN) [11], "Species 2000", and NCBI taxonomy [12] for bacteria and archaea, MycoBank [13] for fungi and yeast.

A programming script was written (in the Java™ language) to automatically compare species names between the GCM catalogue and the nomenclature databases cited above. The comparison showed that from the 36,340 different archaea, bacteria, fungi and microalgae contained in GCM, 2188 could not be found in any of the nomenclature databases above. The average mismatching is 6% (Table 5). When conflicts are identified, GCM sends the results of these comparisons to curators at the relevant collections to allow them to edit their catalogue information online. When mismatches occur, the system provides the probably correct species name based on character string similarity. Following such comparison, the majority of spelling mistake is corrected.

The second type of problems with the quality of information is related to data content. For example, some "*Escherichia coli*" strains were wrongly assigned as "Fungi" in the host collection databases. The GCM system collects and compares the lists of differences in the description of cultures in one collection with cultures of the same strains in other collections.

History information was used to do the quality check for species name as well. Totally 12147 strains contain detailed history information in GCM. The system listed all of species name and compared with their history species name in other collections. The result indicated that among 12147 strains, 1746 strains had different species name with their history strains. Further analysis on the result showed that, among the mismatch, 267 belonged to misspelling problems such as "*Candida viswannathii*" was wrongly spelled to "*Candida viswanathii*". However, the left were mistakes or name changes occur during the strain transfer between collections.

Divergent results are forwarded to the curators of the respective collections for corrections. Performing such



Global Catalogue of Microorganisms

ADVANCED SEARCH | HOME

Species Information

Taxonomy	Bacteria>Firmicutes>Bacilli>Lactobacillales>Lactobacillaceae>Lactobacillus>Lactobacillus delbrueckii>Lactobacillus delbrueckii subsp. bulgaricus
NCBI Taxnomy ID	1585
Scientific Name	*Lactobacillus delbrueckii* subsp. *bulgaricus*
Sequence	CR954253

Species related information

Strains	25	Nucleotide Sequences	41
Publications	125	Patents	8
Protein Sequence	471	Protein Structure	0
Genome	5		

Type Strains in Global Catalogue of Microorganisms-(4)

	Strain number	History	Isolated From	Order
1	JCM 1002^T	T. Mitsuoka S1-3 <-- ISL <-- P. A. Hansen Lb 14 <-- S. Orla-Jensen 14 ("*Thermobacterium bulgaricum*").	Bulgarian yogurt	□
2	NBRC 13953^T	IFO 13953 <-- ATCC 11842 <-- P.A. Hansen, Lb 14 <-- S. Orla-Jensen, 14	Bulgarian yogurt	□
3	CECT 4005^T	CECT, 1989 <-- NCFB, 1961 <-- ATCC <-- P. A. Hansen <-- S. Orla-Jensen (Thermobacterium bulgaricum)		□
4	VKM B-1923^T	Barefoot S.F. Clemson University USA, CFM 83	bulgarian jogurt	□

Strains-(21)

	Strain number	History	Isolated From	Order
1	JCM 1001	T. Mitsuoka S1-1 <-- G. Reuter <-- M. E. Sharpe J-18a.		□
2	JCM 11037	CIP 71.36 <-- F. Gasser 64AE.		□
3	JCM 11038	CIP 55.1 <-- A. Lamensans strain BBY.		□
4	JCM 20398	IAM 12091 <-- T. Morichi <-- Y. Ozawa B-5B.	Sour milk	□
5	KCTC 3635	<-- IAM <-- P.A. Hansen, Lb.14 <-- S. Orla-Jensen, 14	From Bulgarian yogurt [3433]	□
6	CECT 4006	CECT, 1989 <-- NCFB, 1957 <-- G. Mocquot (Lactobacillus bulgaricus)		□
7	CECT 5035	CECT, 1998 <-- NCFB, 1957 <-- P. Kastli (Lactobacillus bulgaricus)		□
8	CECT 5036	CECT, 1998 <-- NCFB, 1981 <-- V. Marshall (Lactobacillus bulgaricus)		□
9	CECT 5037	CECT, 1998 <-- NCFB, 1981 <-- M.E. Sharpe<-- P. Wahl (I.P.)		□
10	CECT 5038	CECT, 1998 <-- NCFB, 1981 <-- M.E. Sharpe		□

1 | 2 | 3 |

Publications-(76) Strain Name (54) Strain Number (22) Search Statistics

1 **Characterization and production of amylovorin L471, a bacteriocin purified from Lactobacillus amylovorus DCE 471 by a novel three-step method** Microbiology 1999 ;145
Strain Number cited in this paper: ATCC 15009 ATCC 6949 ATCC 7469 ATCC 7830 ATCC 8042 CCM 4231 LMG 1285 LMG 13570 LMG 13571 LMG 6412 LMG 6413 LMG 6895 LMG 6901 LMG 6902 LMG 6904 LMG 6906 LMG 6907 LMG 7942 LMG 9433

2 **Use of PCR-based methods for rapid differentiation of Lactobacillus delbrueckii subsp. bulgaricus and L. delbrueckii subsp. lactis.** Appl. Environ. Microbiol. 1999 ;65
Strain Number cited in this paper: ATCC 11842 ATCC 14917 ATCC 15009 ATCC 19258 ATCC 334 ATCC 4797 LMG 11482 LMG 6400 LMG 6412 LMG 6890 LMG 6901 LMG 7942 LMG 8205 LMG 9191 LMG 9852

3 **Metabolism of Linoleic Acid by Human Gut Bacteria: Different Routes for Biosynthesis of Conjugated Linoleic Acid** J. Bacteriol. 2007 ;189
Strain Number cited in this paper: DSM 20016 DSM 20081 DSM 20729

4 **Biphasic kinetics of growth and bacteriocin production with Lactobacillus amylovorus DCE 471 occur under stress conditions** Microbiology 2003 ;149
Strain Number cited in this paper: LMG 6901 LMG 9496

5 **In Vivo Study of the Survival of Lactobacillus delbruecki subsp. bulgaricus CECT 4005T and Streptococcus thermophilus CECT 801 by DVC-FISH after Consumption of Fermented Milk.** Journal of food science 2012 ;149
Strain Number cited in this paper: CECT 4005 CECT 801

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |

Patents-(5) Strain Name (3) Strain Number (2) Search Statistics

1
PatentNo: WO2008043161-A1; EP2076139-A1
Strain Number: NBIMCC 8357 NBIMCC 8358
Title: **Novel Lactobacillus delbrueckii subspecies bulgaricus strain useful for preparing products of pharmaceutical e.g. fresh cheese, prophylactics and therapeutics for alimentary canal disorders in human subjects**
Author: BALTADJIEVA M; DE LAS HERAS J M
Assignee: BALTADJIEVA M (BALT-Individual); DE LAS HERAS J M (DHER-Individual)

2
PatentNo: CN101497892-A
Strain Number: CGMCC 1
Title: **Producing a natural essence substrate of yoghourt by microbial fermentation, comprises preparing freezing fermenting agent of lactobacillus, preparing fermentation culture medium of yoghourt, and preparing natural essence**
Author: CHEN H; DU Y; WANG B; XU P; ZHANG Z
Assignee: SHANGHAI AIPU SPICE CO LTD (SHAN-Non-standard)

3
PatentNo: RU2268299-C2
Strain Number: VKPM B-5284 VKPM A-4174 VKPM B-2455 VKPM B-3489
Title: **Method for preparing bacterial concentrate for fodder ensilage**
Author: SAYAN V SH; DUBENKO A YA; ZAERKO V I; SURMILO A P
Assignee: SAYAN V SH (SAYA-Individual); DUBENKO A YA (DUBE-Individual); ZAERKO V I (ZAER-Individual); SURMILO A P (SURM-Individual)

1 |

Genome-(5) Strain Name (5) Strain Number (0) Definition: Search

□	Accession Number	Length	Strain Number	Year	Description
□	NC_008054	1864998	ATCC 11842	2010	Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842, complete genome.
□	NC_008529	1856951	ATCC BAA-365	2010	Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365, complete genome.
□	NC_017469	1872918		2012	Lactobacillus delbrueckii subsp. bulgaricus 2038 chromosome, complete genome.
□	NC_014727	2125753		2012	Lactobacillus delbrueckii subsp. bulgaricus ND02 chromosome, complete genome.
□	NC_014728	6223		2011	Lactobacillus delbrueckii subsp. bulgaricus ND02 plasmid unnamed, complete sequence.

1 |

Analysis Blast ClustalW

Sequences-(41) Strain Name (30) Strain Number (11) Search Statistics

□	Accession Number	Length	Strain Number	Year	Description
□	AY122482	327	DSM 20081	2004	Lactobacillus delbrueckii subsp. bulgaricus RNA polymerase beta subunit (rpoB) gene, partial cds.
□	JF441088	420	DSM 20081	2011	Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842 strain DSM 20081 phenylalanyl-tRNA synthase (
□	JF441114	465	DSM 20081	2011	Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842 strain DSM 20081 RNA polymerase alpha subunit
□	JF441140	597	DSM 20081	2011	Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842 strain DSM 20081 beta-galactosidase (lacZ) ge
□	AB289079	673	JCM 1001	2007	Lactobacillus delbrueckii subsp. bulgaricus gene for 16S rRNA, partial sequence, strain: JCM 1001.
□	AB289080	675	JCM 1002	2007	Lactobacillus delbrueckii subsp. bulgaricus gene for 16S rRNA, partial sequence, strain: JCM 1002.
□	AB289081	676	JCM 11037	2007	Lactobacillus delbrueckii subsp. bulgaricus gene for 16S rRNA, partial sequence, strain: JCM 11037.
□	AB289082	676	JCM 11038	2007	Lactobacillus delbrueckii subsp. bulgaricus gene for 16S rRNA, partial sequence, strain: JCM 11038.
□	AB680531	1489	NBRC 13953	2012	Lactobacillus delbrueckii subsp. bulgaricus gene for 16S rRNA, partial sequence, strain: NBRC 13953.
□	AJ586864	664	type strain: NCC641 = ATCC11842	2003	Lactobacillus delbrueckii subsp. bulgaricus partial recA gene for recombination protein, strain NCC6

1 | 2 | 3 |

Analysis Blast ClustalW

Swissprot-(471) Statistics

□	Accession Number	Date	Entry	Length	From
□	Q1G889	27-JUN-2006	Y2075_LACDA	245	Lactobacillus delbrueckii subsp. bulgaricus (strain ATCC 11842 / DSM 20081).
□	Q1G9B9	27-JUN-2006	Y1495_LACDA	384	Lactobacillus delbrueckii subsp. bulgaricus (strain ATCC 11842 / DSM 20081).
□	Q1G950	27-JUN-2006	Y1604_LACDA	107	Lactobacillus delbrueckii subsp. bulgaricus (strain ATCC 11842 / DSM 20081).
□	Q1G948	27-JUN-2006	Y1606_LACDA	85	Lactobacillus delbrueckii subsp. bulgaricus (strain ATCC 11842 / DSM 20081).
□	Q1GAR5	27-JUN-2006	Y761_LACDA	77	Lactobacillus delbrueckii subsp. bulgaricus (strain ATCC 11842 / DSM 20081).
□	Q8KH12	01-OCT-2002	DER_LACDA	435	Lactobacillus delbrueckii subsp. bulgaricus (strain ATCC 11842 / DSM 20081).
□	Q1G9N0	27-JUN-2006	Y1351_LACDA	144	Lactobacillus delbrueckii subsp. bulgaricus (strain ATCC 11842 / DSM 20081).
□	P31672	03-OCT-2006	NIFS_LACDA	385	Lactobacillus delbrueckii subsp. bulgaricus (strain ATCC 11842 / DSM 20081).
□	Q1GA88	27-JUN-2006	Y1026_LACDA	447	Lactobacillus delbrueckii subsp. bulgaricus (strain ATCC 11842 / DSM 20081).
□	P54262	01-OCT-1996	SYN_LACDA	432	Lactobacillus delbrueckii subsp. bulgaricus (strain ATCC 11842 / DSM 20081).

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |

Analysis Blast ClustalW

Comments on this species Edit

Figure 2 Example of Species page of *Lactobacillus delbrueckii* subsp. *bulgaricus*.

controls for all fields of the database greatly assist collections in correcting existing mistakes.

Utility

Interface and web tools

The database homepage contains a world map which indicates the countries and regions that have already joined the GCM project. Statistics and graphics indicate the continuing acquisition of data into the GCM. A simplified search interface allows the querying of the database by using the strain number and species name. In addition, a variety of tools have been implemented to enhance its use. The main web tools that were integrated into the GCM are the following:

Advanced search

Three query options are available in the advanced search section. Users may search strains within a range of values for one or several properties, including cultivation temperature, substrate, or application, before retrieving the retrieve corresponding results.

Since GCM maintains nucleotide sequences data associated with individual strains, a sequence alignment tool based on the Basic Local Alignment Search Tool (BLASTN) [14] is included. Results are ordered by similarity.

Bibliographic and patents queries are also possible and allow users to search by keywords in titles, abstracts of articles or patents. Search results are listed as strain numbers, strain names, publication abstracts and titles and can be exported in text file format.

With the advanced search tools, the system can perform the following searches

- Searching for type strains for some taxa in certain culture collections
- Searching for strains with specific characteristics in the list of Culture Collection (CC) or Biological Resource Center (BRC), such as range of growth temperature, transfer history, collected location and others
- Searching for strains with specific properties
- Searching strains isolated from various substrates, including sludge or wastewater, soils, sediment, fermentation products. Results are listed in table format, with the type of organism type used as column name;
- Searching strains with particular protein coding genes

Results are listed by strain number, species name, culture collections, and isolation sources. A few filter windows are provided in the result page to allow users to refine the results by collections, growth temperature, isolation sources or organism type.

Table 3 Isolation sources of Strains sorted by type of organism

Isolation source type	Fungi	Bacteria	Yeasts	Actinomycetes	Archaea	Phage	Microalgae	Total
Sludge/Wasterwater	1	1091	6	-	9	-	2	1109
Soil	1708	3468	484	264	95	1	1	6021
Sediment	4	46	17	-	14	-	-	81
Fermentation products	123	358	327	-	1	-	-	809
Plant-associated	405	314	644	-	1	-	2	1366
Host-associated	139	480	180	-	3	-	-	802
Human-associated	18	11167	55	-	15	-	2	11257
Water	4	398	50	-	48	-	-	500
Microbial-mat/Biofilm	-	-	-	-	1	-	-	1
Air	6	20	29	-	1	-	-	56
Genetic engineering strain	22698	-	-	-	-	-	-	-
Food	193	83	69	-	2	-	-	347
Others	135	728	76	-	25	-	1	965
Total	2736	18153	1937	264	215	1	8	23314

Species tree viewer

A species2000 taxonomy tree is used for the organization of strain information. Species names are used to map between GCM data and species2000 name (http://www.sp2000.org/), and then a taxonomic tree containing the number of strains for each genus is constructed. User can then browse the taxonomy tree itself, or search a species name within it.

Map viewer

While geographic origins of strains are usually provided as rural location, national park or cities, GCM can automatically translate such locations into more precise information of longitude and latitude. Strains are then displayed on a map using the Google maps API. In some cases, the location information is a more specific place such as a university or an institute, which could not be translated directly into longitude and latitude values. In such cases, manual annotation by the administrator of GCM will then use the value of the located city as an approximation. An example strain information page is displayed in Figure 3.

Table 4 Top 20 countries from which strains were collected

Order	Country	Counts	Order	Country	Counts
1	Japan	8248	11	China	3429
2	France	8070	12	India	2907
3	United States	7701	13	Russian Federation	2872
4	Netherlands	6709	14	South Africa	2419
5	Korea	6270	15	Italy	2009
6	Germany	6051	16	Canada	1848
7	Thailand	5894	17	VietNam	1818
8	United Kingdom	5717	18	Sweden	1786
9	Belgium	5177	19	Australia	1695
10	Spain	3869	20	Switzerland	1466
	Total			85955	

114,578 of 273,933 strains contain information regarding their geographic origins. The strains were collected from 164 countries and regions, of which, 85,955 strains were collected from only 20 countries. This takes up approximately 74% of total strains, which indicates a relatively high sampling effort in these countries.

Data analysis

A variety analysis tools are also employed on both the strain information and species page. The BLAST program (Altschul SF [13]) was used for sequence homology searches within the database. For sequences related to the same strain or species, the ClustalW [15] program is provided to perform multiple sequence alignment analysis.

Data update and management

To provide the greatest benefit to partner collections, a database management function was provided to GCM participating collections (Figure 4). After registration with the GCM project and filling out a metadata form, a user account will be given to the collection. Curators can then either export catalogue information in batch or add strain information individually. The system automatically records every operation, including updates, additions or deletions and after approval by the administrators in charge, the updated records are published online.

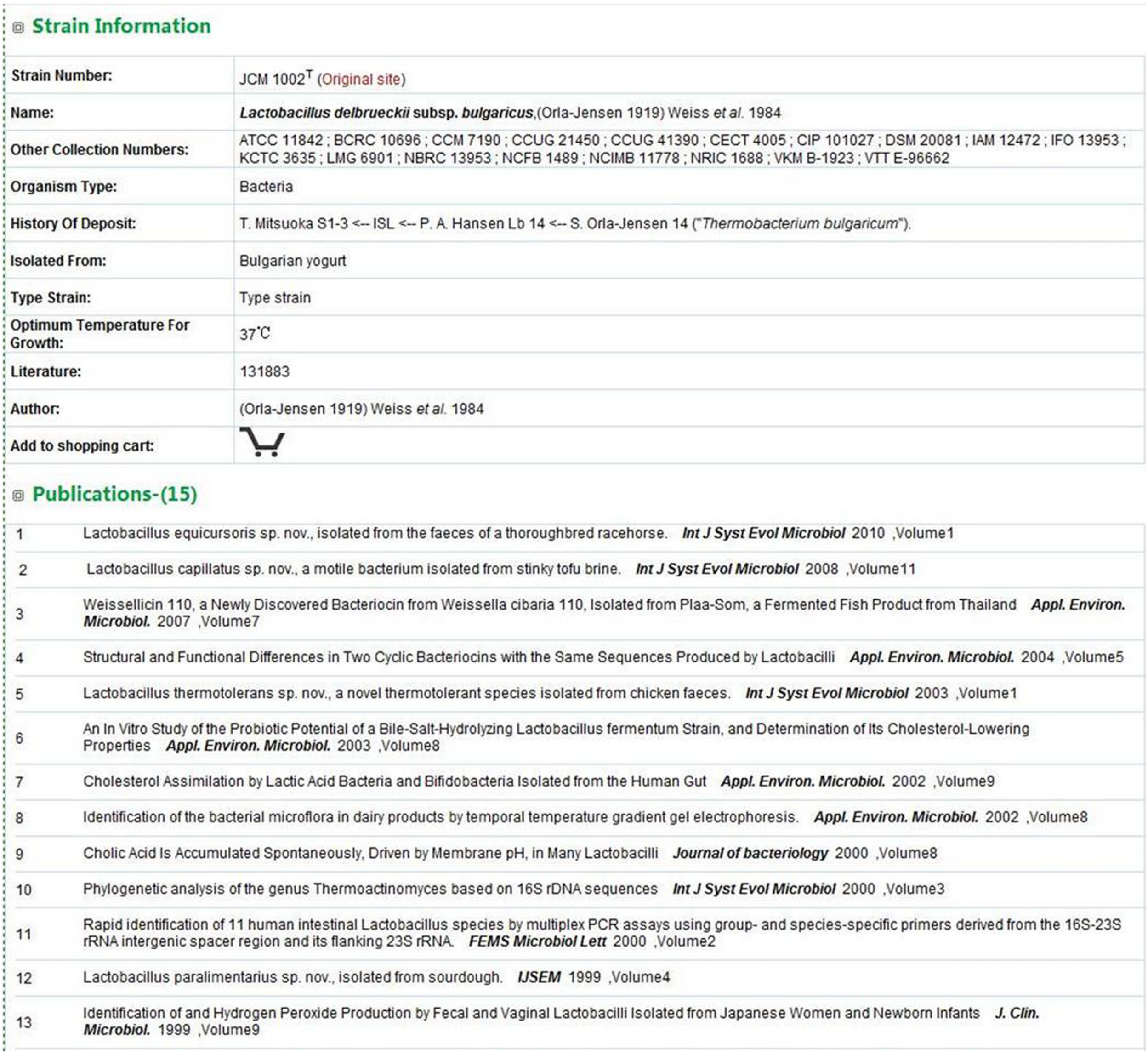
Discussion and conclusion

A large amount of microbial resources are preserved as living strains in collections, however, information describing these strains is often unavailable. Each culture

Table 5 Result summary of species name check

Organism type	Species names	Un-matched species name	Percentage of un-match
Archaea	1399	32	2.30%
Microalgae	1457	360	24.70%
Fungi	20719	698	3.40%
Bacteria	12855	1098	8.50%
Total	36430	2188	6.00%

This table provides comparative results of the species names within GCM with public microbial nomenclature database. Species2000, NCBI taxonomy, LPSN and Mycobank were used as reference databases. The average percentage of unmatched names is 6%, while the archaea and fungi showed lower than average percentage of unmatched names. The percentage of unmatched names is relatively high for microalgae, possibly due to the irregular naming for microalgae.



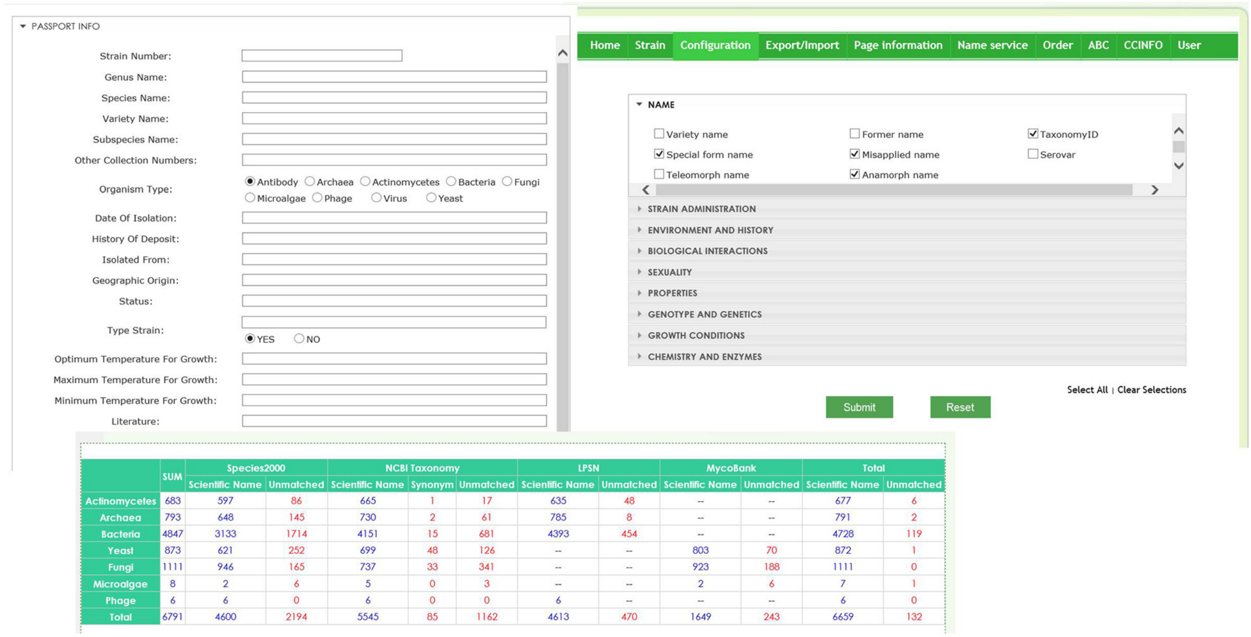
⊡ Strain Information

Strain Number:	JCM 1002^T (Original site)
Name:	***Lactobacillus delbrueckii* subsp. *bulgaricus***,(Orla-Jensen 1919) Weiss *et al.* 1984
Other Collection Numbers:	ATCC 11842 ; BCRC 10696 ; CCM 7190 ; CCUG 21450 ; CCUG 41390 ; CECT 4005 ; CIP 101027 ; DSM 20081 ; IAM 12472 ; IFO 13953 ; KCTC 3635 ; LMG 6901 ; NBRC 13953 ; NCFB 1489 ; NCIMB 11778 ; NRIC 1688 ; VKM B-1923 ; VTT E-96662
Organism Type:	Bacteria
History Of Deposit:	T. Mitsuoka S1-3 <-- ISL <-- P. A. Hansen Lb 14 <-- S. Orla-Jensen 14 ("*Thermobacterium bulgaricum*").
Isolated From:	Bulgarian yogurt
Type Strain:	Type strain
Optimum Temperature For Growth:	37℃
Literature:	131883
Author:	(Orla-Jensen 1919) Weiss *et al.* 1984
Add to shopping cart:	🛒

⊡ Publications-(15)

1	Lactobacillus equicursoris sp. nov., isolated from the faeces of a thoroughbred racehorse. ***Int J Syst Evol Microbiol*** 2010 ,Volume1
2	Lactobacillus capillatus sp. nov., a motile bacterium isolated from stinky tofu brine. ***Int J Syst Evol Microbiol*** 2008 ,Volume11
3	Weissellicin 110, a Newly Discovered Bacteriocin from Weissella cibaria 110, Isolated from Plaa-Som, a Fermented Fish Product from Thailand ***Appl. Environ. Microbiol.*** 2007 ,Volume7
4	Structural and Functional Differences in Two Cyclic Bacteriocins with the Same Sequences Produced by Lactobacilli ***Appl. Environ. Microbiol.*** 2004 ,Volume5
5	Lactobacillus thermotolerans sp. nov., a novel thermotolerant species isolated from chicken faeces. ***Int J Syst Evol Microbiol*** 2003 ,Volume1
6	An In Vitro Study of the Probiotic Potential of a Bile-Salt-Hydrolyzing Lactobacillus fermentum Strain, and Determination of Its Cholesterol-Lowering Properties ***Appl. Environ. Microbiol.*** 2003 ,Volume8
7	Cholesterol Assimilation by Lactic Acid Bacteria and Bifidobacteria Isolated from the Human Gut ***Appl. Environ. Microbiol.*** 2002 ,Volume9
8	Identification of the bacterial microflora in dairy products by temporal temperature gradient gel electrophoresis. ***Appl. Environ. Microbiol.*** 2002 ,Volume8
9	Cholic Acid Is Accumulated Spontaneously, Driven by Membrane pH, in Many Lactobacilli ***Journal of bacteriology*** 2000 ,Volume8
10	Phylogenetic analysis of the genus Thermoactinomyces based on 16S rDNA sequences ***Int J Syst Evol Microbiol*** 2000 ,Volume3
11	Rapid identification of 11 human intestinal Lactobacillus species by multiplex PCR assays using group- and species-specific primers derived from the 16S-23S rRNA intergenic spacer region and its flanking 23S rRNA. ***FEMS Microbiol Lett*** 2000 ,Volume2
12	Lactobacillus paralimentarius sp. nov., isolated from sourdough. ***IJSEM*** 1999 ,Volume4
13	Identification of and Hydrogen Peroxide Production by Fecal and Vaginal Lactobacilli Isolated from Japanese Women and Newborn Infants ***J. Clin. Microbiol.*** 1999 ,Volume9

Figure 3 Example of strain information of *Lactobacillus delbrueckii* subsp. *bulgaricus.*



PASSPORT INFO

Strain Number:
Genus Name:
Species Name:
Variety Name:
Subspecies Name:
Other Collection Numbers:
Organism Type: ● Antibody ○ Archaea ○ Actinomycetes ○ Bacteria ○ Fungi ○ Microalgae ○ Phage ○ Virus ○ Yeast
Date Of Isolation:
History Of Deposit:
Isolated From:
Geographic Origin:
Status:
Type Strain: ● YES ○ NO
Optimum Temperature For Growth:
Maximum Temperature For Growth:
Minimum Temperature For Growth:
Literature:

NAME

□ Variety name
☑ Special form name
□ Teleomorph name
□ Former name
☑ Misapplied name
☑ Anamorph name
☑ TaxonomyID
□ Serovar

STRAIN ADMINISTRATION
ENVIRONMENT AND HISTORY
BIOLOGICAL INTERACTIONS
SEXUALITY
PROPERTIES
GENOTYPE AND GENETICS
GROWTH CONDITIONS
CHEMISTRY AND ENZYMES

	SUM	Species2000		NCBI Taxonomy			LPSN		MycoBank		Total	
		Scientific Name	Unmatched	Scientific Name	Synonym	Unmatched	Scientific Name	Unmatched	Scientific Name	Unmatched	Scientific Name	Unmatched
Actinomycetes	683	597	86	665	1	17	635	48	--	--	677	6
Archaea	793	648	145	730	2	61	785	8	--	--	791	2
Bacteria	4847	3133	1714	4151	15	681	4393	454	--	--	4728	119
Yeast	873	621	252	699	48	126	--	--	803	70	872	1
Fungi	1111	946	165	737	33	341	--	--	923	188	1111	0
Microalgae	8	2	6	5	0	3	--	--	2	6	7	1
Phage	6	6	0	6	0	0	6	--	--	--	6	0
Total	6791	4600	2194	5545	85	1162	4613	470	1649	243	6659	132

Figure 4 Database management system for collections. Users can import data by generating an EXCEL file to meet the WDCM RDS. After the data are imported into the system, users can either update or edit the catalogue information online. A species name check result is provided in the database management system to provide an overview of data quality and allow for further modifications.

collection is independently responsible for the maintenance of data associated with their microbes, there is presently no enforced data harmonization and information sharing mechanism is available. Such situation hinders both the efficient management of collections and the ability to explore statistics about world microbial resources. Therefore, there is great demand for developing a mechanism for digital, online resource sharing, which provides a fundamental tool for best practices in information management.

The major target group for such system are culture collections staff, as well as academic and industrial microbiologists. We believe that GCM will assist collections, which lack the required human resources and information technology, to publish their stock information in an efficient and standardized way that is most useful for scientific and industrial communities. Database queries via a user-friendly and web-based interface should greatly promote the sharing and use of microbial resources.

While this project is still in its early stage, we are confident that it will continue to grow with the further addition of data, analytical tools and other functionalities. In the future, additional database management tools will be provided to allow more culture collections to share their data via GCM. These tools will lead to the increased availability of accessible data pertaining to microbial strains held in public collections and their utilization for bioindustry, medicine, and research. As it grows, GCM will incorporate information related to enzymatic and metabolic pathways using developing genomics and bioinformatics tools. Ultimately, GCM is a comprehensive data platform on microbial resources that is available to the public.

Availability and requirements

The GCM database runs on a platform with both Java and MySQL server. Catalogue information gathered from associated collections is centralized within WDCM servers, which is hosted at the Institute of Microbiology, of the Chinese Academy of Sciences.

The Blast program is used for the sequence homology search in the database (BLASTN 2.2.25). Multiple sequence alignments are performed using the ClustalW program (version2.1). GCM is available at http://gcm.wfcc.info.

Competing interest

The authors declare that they have no competing interest.

Authors' contributions

LW designed database and web services. QS integrated data resources from Pubmed and Genbank. SY developed the database, made the webpages and wrote the database queries. HS designed and supervised construction of the database. YZ, KM, AV, SKI, MO, YL, VR, SI and FG provided the catalogue information and participated in the design of the database. LW and KM wrote the manuscript. DP and JM designed the database and supervised construction of the project. All authors read and approved the final manuscript.

Acknowledgement

GCM project was initiated by WDCM and approved by the WFCC board. WDCM acknowledged the contributions of all participating collections to the

GCM project. At the time of writing this article 52 collections from 25 countries have already joined the effort.

Author details
[1]Institute of Microbiology, Chinese Academy of Sciences, Beijing, China. [2]World Data Centre for Microorganisms(WDCM), Beijing, China. [3]Belgian Coordinated Collections of Micro-organisms Programme, Belgian Science Policy Office, avenue Louise, 231 1050 Brussels, Belgium. [4]National Institute of Genetics, Yata, Mishima 411-8540 Japan. [5]Fungal Genetics Stock Center, University of Missouri, Kansas City, Missouri, USA. [6]G.K. Skryabin Institute of Biochemistry and Physiology of Microorganisms RAS, Pushchino, Moscow region, Russia. [7]Culture Collection Division Biological Resource Center (NBRC), National Institute of Technology and Evaluation (NITE), 2-5-8 Kazusakamatari, Kisarazu-shi, Chiba 292-0818 Japan. [8]Japan Collection of Microorganisms/ Microbe Divion, RIKEN BioResource Center, Koyadai 3-1-1, Tsukuba, Ibaraki 305-0074 Japan. [9]Seoul Women's Unviersity/Korea National Research Resource Center, Seoul, Korea. [10]CBS-KNAW, Fungal Biodiversity Centre, Uppsalalaan 8, Utrecht, The Netherlands. [11]Bioresources Technology Unit, National Center for Genetic Engineering and Biotechnology, 113 Thailand Science Park, KlongLuang, Prathumthani 12120 Thailand.

Received: 4 September 2013 Accepted: 27 December 2013
Published: 30 December 2013

References

1. Satoru Miyazaki HS: **Networking of biological resource centers: WDCM Experiences.** *Data Science Journal* 2002, **1**(2):102–107.
2. *OECD Best Practice Guidelines for Biological Resource Centres*; 2007. http://www.oecd.org/health/biotech/oecdbestpracticeguidelinesforbiologicalresourcecentres.htm.
3. Gams W, Hennebert GL, Stalpers JA, Janssens D, Schipper MA, Smith J, Yarrow D, Hawksworth DL: **Structuring strain data for storage and retrieval of information on fungiand yeasts in MINE, the microbial information network Europe.** *J Gen Microbiol* 1998, **134:**1667–1689.
4. CABRI: *Common Access to Biological Resources and Information (CABRI)*, GUIDELINES FOR CATALOGUE PRODUCTION; 1998. http://www.cabri.org/guidelines/catalogue/CPcover.html.
5. Benson DA CM, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2013, **41**(D1):D36–D42.
6. Consortium TU: **Reorganizing the protein space at the Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2012(40):71–75.
7. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE: **The RCSB protein data bank: redesigned web site and web services.** *Nucleic Acids Res* 2011, **39:**D392–401.
8. Cachuela-Palacio M: **Towards an index of all known species: the Catalogue of Life, its rationale, design and use.** *Integrative Zoology* 2006, **1**(1):418–421.
9. Hiroshi M: *Metagenome and Microbes Environmental Ontology*; 2013. http://bioportal.bioontology.org/ontologies/ME.
10. Eamonn OT: **Meeting report: hackathon-workshop on Darwin Core and MIxS standards alignment.** *Stand Genomic Sci* 2012, **7**(1):166–170.
11. Euzéby JP: **List of bacterial names with standing in nomenclature: aFolder available on the internet.** *Int J Syst Evol Microbiol* 1997, **47**(2):590–592.
12. Federhen S: **The NCBI Taxonomy database.** *Nucleic Acids Res* 2012, **40**(1): D136–D143.
13. Crous PW, Walter G, Stalpers JA, Vincent R, Gerrit S: **MycoBank: an online initiative to launch mycology into the 21st century.** *Stud Mycol* 2004, **50:**19–22.
14. Altschul SF GW, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of molecular biology* 1990, **215:**403–410.
15. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: **ClustalW and ClustalX version 2.** *Bioinformatics* 2007, **23**(21):2947–2948.

doi:10.1186/1471-2164-14-933
Cite this article as: Wu *et al.*: Global catalogue of microorganisms (gcm): a comprehensive database and information retrieval, analysis, and visualization system for microbial resources. *BMC Genomics* 2013 **14**:933.