D694–D705 *Nucleic Acids Research, 2021, Vol. 49, Database issue* *Published online 29 October 2020*
doi: 10.1093/nar/gkaa957

gcType: a high-quality type strain genome database for microbial phylogenetic and functional research

Wenyu Shi[1,2,†], Qinglan Sun[1,2,3,†], Guomei Fan[1,2], Sugawara Hideaki[4], Ohkuma Moriya[5], Takashi Itoh[5], Yuguang Zhou[6], Man Cai[6], Song-Gun Kim[7], Jung-Sook Lee[7], Ivo Sedlacek[8], David R. Arahal[9], Teresa Lucena[9], Hiroko Kawasaki[10], Lyudmila Evtushenko[11], Bevan S. Weir[12], Sarah Alexander[13], Dlauchy Dénes[14], Somboon Tanasupawat[15], Lily Eurwilaichitr[3,16], Supawadee Ingsriswang[3,16], Bruno Gomez-Gil[17], Manzour H. Hazbón[18], Marco A. Riojas[18], Chatrudee Suwannachart[19], Su Yao[20], Peter Vandamme[21], Fang Peng[22], Zenghui Chen[1,2], Dongmei Liu[1,2], Xiuqiang Sun[1,2], Xinjiao Zhang[1,2], Yuanchun Zhou[23], Zhen Meng[23], Linhuan Wu[1,2,24,*] and Juncai Ma[1,2,3,24,*]

[1]Microbial Resource and Big Data Center, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China, [2]World Data Center for Microorganisms, Beijing 100101, China, [3]China-Thailand Joint Laboratory on Microbial Biotechnology, Beijing 100190, China, [4]National Institute of Genetics, Yata, Mishima 411-8540, Japan, [5]Japan Collection of Microorganisms (JCM)/ Microbe Divion, RIKEN BioResource Center, Koyadai 3-1-1, Tsukuba, Ibaraki 305-0074, Japan, [6]China General Microbiological Culture Collection Center (CGMCC), Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China, [7]Korean Collection for Type Cultures (KCTC), Korea Research Institute of Bioscience and Biotechnology (KRIBB), 181 Ipsin-gil, Jeongeup-si, Jeollabuk-do, 56212, Republic of Korea, [8]Czech Collection of Microorganisms, Masaryk University, Kamenice 5, building A25, 625 00 Brno, Czech Republic, [9]Colección Española de Cultivos Tipo (CECT), and Departamento de Microbiología y Ecología, University of Valencia, 46100 Burjassot (Valencia), Spain, [10]NITE Biological Resource Center (NBRC), National Institute of Technology and Evaluation, 2-5-8 Kazusakamatari, Kisarazu, Chiba 292-0818, Japan, [11]All-Russian Collection of Microorganisms (VKM), G.K. Skryabin Institute of Biochemistry and Physiology of Microorganisms RAS, Pushchino, Moscow region 142290, Russia, [12]Mycology & Bacteriology Systematics, Manaaki Whenua – Landcare Research, Auckland, New Zealand, [13]National Collection of Type Cultures (NCTC), Public Health England (PHE), UK, [14]National Collection of Agricultural and Industrial Microorganisms, Faculty of Food Science, Szent István University, H-1118, Budapest, Somlói út 14-16, Hungary, [15]Faculty of Pharmaceutical Sciences, Chulalongkorn University (PCU), Bangkok 10330, Thailand, [16]Thailand Bioresource Research Center (TBRC), National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Thailand, [17]CIAD, A.C., Collection of Aquatic Important Microorganisms (CAIM). AP 711 Mazatlán, Sinaloa, Mexico, [18]American Type Culture Collection(ATCC), 10801 University Boulevard, Manassas, VA 20110, USA, [19]Biodiversity Research Centre, Thailand Institute of Scientific and Technological Research (TISTR), 35 M 3 Technopolis Khlong 5 Khlong Luang Pathum Thani 12120, Thailand, [20]China Center of Industrial Culture Collection (CICC), Beijing, China, [21]BCCM/LMG Bacteria Collection, Laboratory of Microbiology, Faculty of Sciences, Ghent University, K. L. Ledeganckstraat 35, 9000 Ghent, Belgium, [22]China Center for Type Culture Collection (CCTCC), College of Life Sciences, Wuhan University, Wuhan 430072, China, [23]Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China and [24]State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China

Received August 11, 2020; Revised October 06, 2020; Editorial Decision October 07, 2020; Accepted October 28, 2020

[*]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. Tel: +86 10 64807422; Fax: +86 10 64807426; Email: ma@im.ac.cn
[†]The authors wish it to be known that the first two authors should be regarded as joint first authors.

ABSTRACT

Taxonomic and functional research of microorganisms has increasingly relied upon genome-based data and methods. As the depository of the Global Catalogue of Microorganisms (GCM) 10K prokaryotic type strain sequencing project, Global Catalogue of Type Strain (gcType) has published 1049 type strain genomes sequenced by the GCM 10K project which are preserved in global culture collections with a valid published status. Additionally, the information provided through gcType includes >12 000 publicly available type strain genome sequences from GenBank incorporated using quality control criteria and standard data annotation pipelines to form a high-quality reference database. This database integrates type strain sequences with their phenotypic information to facilitate phenotypic and genotypic analyses. Multiple formats of cross-genome searches and interactive interfaces have allowed extensive exploration of the database's resources. In this study, we describe web-based data analysis pipelines for genomic analyses and genome-based taxonomy, which could serve as a one-stop platform for the identification of prokaryotic species. The number of type strain genomes that are published will continue to increase as the GCM 10K project increases its collaboration with culture collections worldwide. Data of this project is shared with the International Nucleotide Sequence Database Collaboration. Access to gcType is free at http://gctype.wdcm.org/.

INTRODUCTION

Microorganisms are considered the most abundant organisms in the world. It is estimated that ~4–6 × 10^{30} prokaryotic cells exist on Earth, comprising a biomass of 350–550 × 10^{15} g of carbon (1). The total number of prokaryotic species is up to 10^{9} (2). Approximately 1800 bacterial and archaeal species names were published in approved lists of bacterial names in 1980 (3). Thereafter, names published in original articles or in the 'Validation Lists' of the *International Journal of Systematic and Evolutionary Microbiology* (IJSEM) have been validated. As of September 2020, this number increased to 16 763.

The description of a prokaryotic species needs to designate its type strain, whose phenotypes and genotypes are often well characterised and described. 16S rRNA gene and whole-genome sequences derived from type strains, together with phenotypic and chemotaxonomic characteristics are used for taxonomic identification. Thus, type strains are critical references for the characterisation of species and the identification of isolates and strains for taxonomic purposes (4). Currently, the type strains from the 16 763 listed species are available as 67 331 catalogue numbers from over 130 culture collections.

For several decades, it has been recognised that the complete deoxyribonucleic acid (DNA) sequence of a species would be the standard reference to determine their phylogeny, which in turn determines their taxonomic classification (5). With the increasing availability of genome sequences, genome-based methods such as average amino acid identity, average nucleotide identity (ANI) and digital DNA–DNA hybridization (dDDH) have been developed as important measurements for prokaryotic taxonomy (6–8). Since January 2018, IJSEM requires authors of new taxa to provide genome sequences with descriptions of the novel taxa for their manuscripts to be eligible for publication. Data on 16S rRNA similarity, overall genome similarity or distance and phenotypic and physiological information are used in combination to identify a new species (9).

Another essential element of microbial taxonomy is the correct assessment of phylogenetic relationships via the reconstruction of phylogenetic trees. Although phenotypic, chemotaxonomic and genotypic information are useful for identifying microorganisms, such information is often insufficient for reconstructing accurate phylogenies. The increasing availability of microbial genome sequences allows for more comprehensive and accurate depictions of phylogenetic relationships to study the origin and evolution of prokaryotic organisms.

Because of their genomic and hence metabolic and functional diversity, microorganisms serve as ideal models for biotechnology studies. Combined with comprehensive phenotypic and physiological information, genome sequences of type strains enable the connection of genes with functions and provide insights into the metabolic and functional potential of microorganisms. Therefore, accruing data on genomes of microorganisms will greatly promote biotechnology studies.

Given the immense efforts of microbiologists and community sequencing projects such as The Genomic Encyclopedia of Bacteria and Archaea (GEBA) (10), the number of publicly available genome sequences of type strains continues to increase rapidly. Currently, there are >12 000 genome sequences in the International Nucleotide Sequence Database Collaboration (INSDC). However, a large number of microbial type strains remain to be sequenced. Therefore, the World Data Centre for Microorganisms (WDCM) GCM 10K project (11) is cooperating with culture collections across the world to fill the current gap in whole-genome databases for validly published species as well as with IJSEM to provide free sequencing and genome annotation services for newly described species (12).

With increasing microbial genomic information, the databases and servers which host and analyse these data continue to expand. The Genomes Online Database (GOLD) (13) in conjunction with the Integrated Microbial Genomes (IMG) (14) provide a comprehensive catalogue of microbial genomes and platform for the analysis of microbial genomes and microbiomes. The Type (Strain) Genome Server (15), which is connected to the comprehensive prokaryotic metadata resources BacDive (16) and LPSN (17), is considered a high-throughput web server for genome-based prokaryotic taxonomy. However, there is still a need for a database that provides not only up-to-date type strains and the associated comprehensive genome information but also user-friendly searchable and comparable functions.

To facilitate access to and maximise the value of genome sequences of type strains, the GCM 10K type strain sequencing project developed the gcType platform. This platform integrates publicly available information from other databases along with the sequencing efforts of GCM 10K project according to strict quality control standards, followed by a powerful standard data-processing pipeline to yield a high-quality reference database that provides web-based data analysis pipelines for genomic analysis and genome-based species identification. Moreover, it associates taxonomic, phenotypic and physiological information with the type strains to enable users to conduct comprehensive genomic and functional analyses. On the whole, gcType is a unique and useful resource that permits microbial taxonomists and microbiologists to gather up-to-date information on microbial type strain sequences.

DATABASE DESIGN AND IMPLEMENTATION

gcType portal and search functions

Users of the platform can query the database and perform genomic analyses using the gcType portal. gcType is the GCM 10K sequencing project data portal for the dissemination of information and publication of updated sequencing results. It currently provides multiple, flexible search functions for users to explore its resources (Figure 1). A text-based advanced search option allows users to conduct cross-genome searches through a single or combination of metadata information. The input query retrieves all data containing the corresponding keywords in the metadata fields, such as the sequencing status, library or contig or scaffold numbers.

All validly published species have been mapped onto the National Center for Biotechnology Information (NCBI) taxonomy ID (18), and all genome sequences have been mapped onto the Genome Taxonomy Database (GTDB) system using GTDB-Tk (19). Users can browse through these taxonomic trees to search for sequenced and unsequenced species. Sequence-based searches against pregenerated type strain 16S rRNA sequence databases using the Basic Local Alignment Search Tool (BLAST) are provided as well. The resultant hits of the query sequences are displayed as alignments and links to the matched type strains.

A statistics page displaying tables with genomic characteristics provides gcType users with an overview of the diversity of microbial genomic data. In particular, these tables feature the guanine-cytosine (GC) content, average genome size, number of predicted genes and functional annotation results among different phyla. Interactive interfaces allow users to further explore the features of various taxonomic groups.

Integrated information for type strains

Comprehensive taxonomic, phenotypic, physiological and genomic information is organised by type strain species (Figure 2A-2D). The taxonomic status, 16S rRNA gene sequence and NCBI taxonomy ID are provided. One or several 'genome sequence project' pages are linked with the type strain page. For genome sequences extracted from public resources, GenBank (20) or GOLD 'Bioproject', 'Biosample' and 'Assembly IDs' are provided as links to the original sites in GenBank. Simultaneously, a 'GCM project number' is assigned to the sequence and linked to the annotation results generated by the GCM microbial genomic annotation pipeline. Metadata from the genome sequences (e.g. library, N50 and GC content) as well as annotation results of the genome are listed by hits from various reference databases such as COG (21), KEGG (22) and CARD (23). Finally, It also incorporates open-source web applications: JBrowse (24), a linear genome viewer, and CGView (25), a circular genome viewer.

Data sources

Type strain information. Lists of validly published species, a list of type strains, a list of type strains by culture collection, and a comprehensive list of 16S rRNA gene and genome sequences are provided. New prokaryotic taxa were considered validly published only if their names were published in the *International Journal of Systematic Bacteriology* (until October 1999) or the renamed IJSEM (from January 2000 to present). Names published in an original article, or the 'Validation Lists' are considered to be valid (26). The list of species with validly published names were manually collected from IJSEM novel species articles and the validation lists. All data were collected until September 2020. IJSEM and the International Committee on Systematics of Prokaryotes require the type strains of new taxa to be deposited in at least two recognised culture collections in two countries. The deposition of type strains and related metadata information has been described in IJSEM articles. For candidate type strains of novel species, taxonomists who request free genome sequencing services from WDCM are asked to provide detailed metadata of the type strains before WDCM formally accepts their proposal. The metadata fields are following the description recommended by the minimum information about a genome sequence (MIGS) specification (27).

16S rRNA gene sequence data. 16S rRNA gene sequences similarity comparison remains the initial step in the identification of prokaryotic species workflow; therefore, a high-quality 16S rRNA gene sequence database of species with validly published names is fundamental for conducting taxonomy studies. Some reference databases for 16S rRNA gene sequences, e.g. EzBiocloud (28), SILVA (29) and RDP (30), provide online resources based on different data integration strategies and filtering criteria (31). Because gcType is designed for type strain-based taxonomic studies, we collected 16S rRNA gene sequences from two sources. The first contains 16S rRNA gene sequences from publicly available resources, including 16S rRNA genes or sequences derived from completed or partial genomes by RNAmmer (32). Their accession numbers were obtained from GenBank, and the associated sequence data were then extracted. Second, for the GCM 10K sequencing project, prior to proceeding with whole-genome sequencing, submitted strains are validated via 16S rRNA sequencing. The sequences obtained from the GCM 10K project were added to the database to allow the quality of type strains to be examined

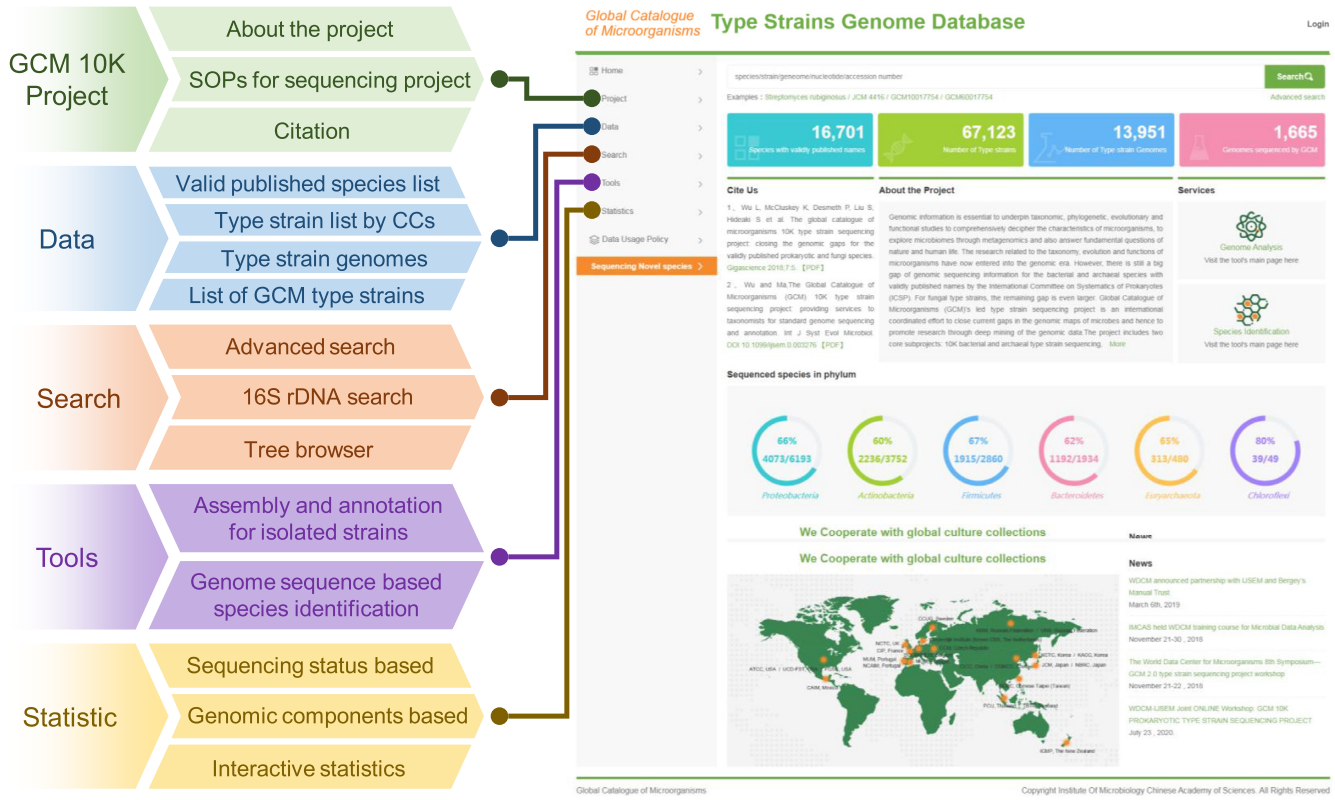

Figure 1. Features of the gcType portal: a map of the main pages and list of the subordinate pages on the gcType website.

because a small percentage the strains are incorrect, likely due to deposition or preservation errors. Data were further filtered and evaluated based on quality control criteria such as sequence length, number of ambiguous bases, completeness and the accuracy of the metadata. After alignment with sequences in infernal (33) and Rfam (34) databases and trimming of the 5′ and 3′ ends the two sources of 16S rRNA gene sequences were integrated to construct a reference dataset for the new species identification pipeline. Evaluation scores related to data quality are displayed on the webpage and can be selected by users. Data processing is described in Figure 3.

Genome sequencing information and sequence data. gcType publishes genome sequences and their annotation results from the GCM 10K sequencing project, which utilises the next-generation sequencing (NGS) platform. To improve the quality of these sequences, we employ third-generation sequencing (TGS) technologies (we are currently using Pacific Biosciences) as a complementary, allowing for the generation of completed or nearly completed bacterial genomes. Importantly, this combination strategy using second- and third-generation sequencing platforms is used for genome sequences that have been poorly assembled in the second-generation sequencing platform alone (>50 contigs).

Raw sequence data that pass strict quality control criteria are run through the GCM microbial genome annotation pipeline. Annotated sequences are then added to the gcType type strain genome database.

Besides the genome sequences in the GCM 10K project, publicly available genome sequence data are extracted from GenBank via their unique type strain numbers. Genome sequences with gene prediction results provided by GenBank in FASTA format were used to perform a GCM microbial genome annotation pipeline; genome sequences without gene prediction results were used to perform a genomic component analysis followed by annotation using the GCM microbial genome annotation pipeline. These newly annotated type strain sequences are then incorporated into the gcType Type strain genome database.

To create a non-redundant set of representative genomes, genome sequences are further filtered by the number of contigs, genome sizes and N50 statistics. Sequences with number of contigs larger than 500 are filtered in the reference database. A single, high-quality representative sequence from each species is selected to create the reference dataset for the new species identification pipeline. Scores pertaining to the quality of the genome are displayed on the webpage and can be selected if there is more than one sequence. A schematic representation of data processing is shown in Figure 3.

As of September 2020, gcType contains 13 962 prokaryotic type strain genomes, from which 1049 were produced by the GCM 10K sequencing project.

Database design

gcType uses the open-source MySQL relational database management system to manage the data. This database con-

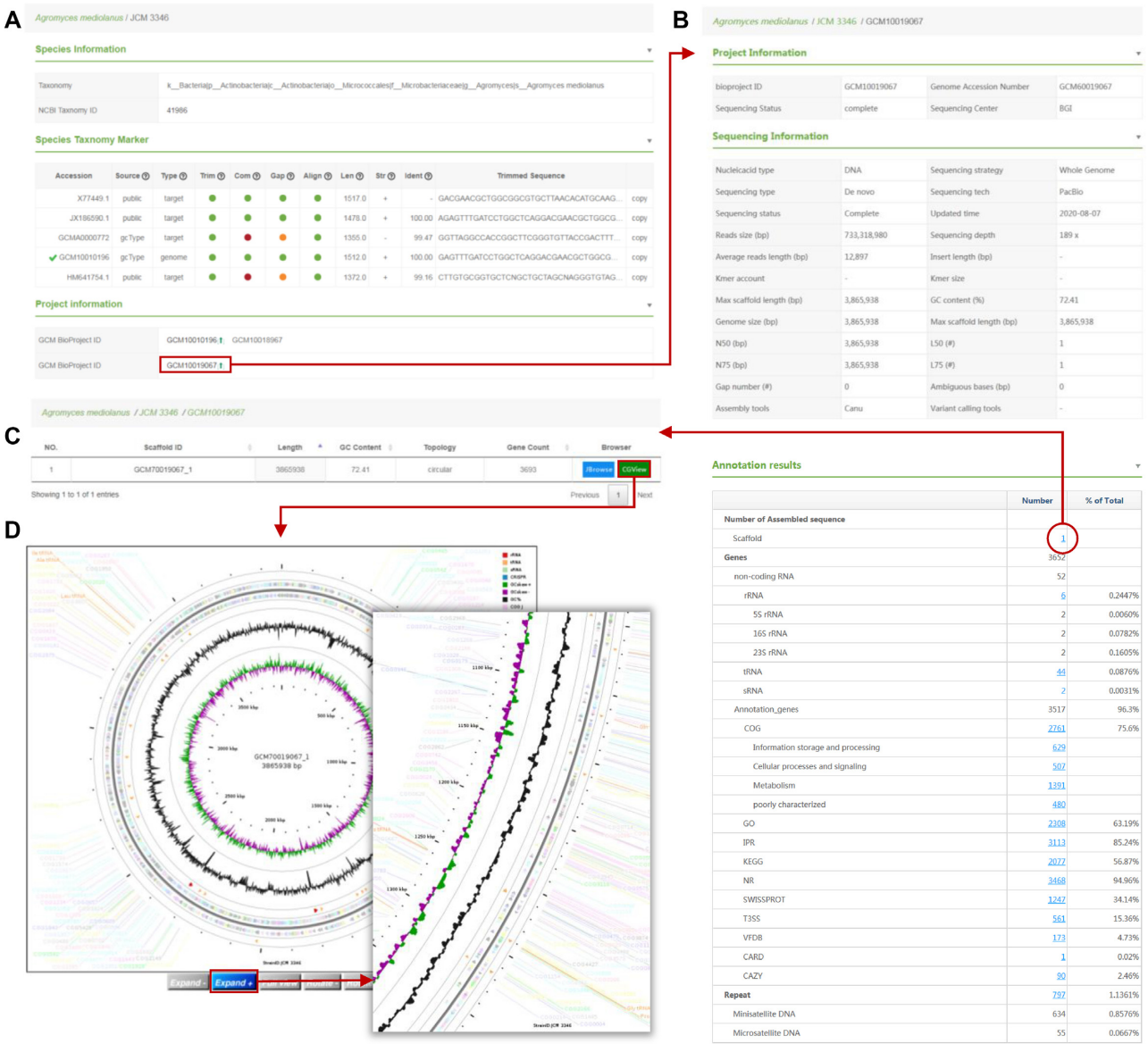

Figure 2. Comprehensive information about a type strain and its associated genome sequences. (**A**) Species information and marker genes. The NCBI taxonomy ID is linked to the NCBI taxonomy database. (**B**) Project and sequencing information. (**C**) Genome annotation results with the associated statistics. (**D**) Interactive circular view of type strain genome.

tains tables that are continuously updated with information on species' taxonomy and nomenclature. Phenotypic and physiological metadata descriptions are linked with proper taxa status. The sequencing projects and their associated metadata and annotation results are linked to the type strains. Scaffolds and predicted features from genomic analyses are related to the assembled genome.

gcType follows an identifier system similar to that of INSDC, in which a 'GCM Biosample' number is assigned to each type strain and a 'GCM Bioproject' number to each sequencing project. For type strains that have been sequenced multiple times and thus have several sets of sequencing results, different 'GCM Bioproject' numbers are assigned to each sequencing effort, a 'locus tag' is assigned to each predicted gene and an 'assembly number' is assigned to the assembled genome.

DATA ANALYSIS PIPELINES

Currently, owing to difficulties in using multiple bioinformatics tools and programming scripts for in-house analyses, taxonomists often rely on commercial sequencing and data analysis services to generate genome sequences. However, due to a wide range of data models, annotation pipelines and versions of reference databases, the results of analysing the same genome vary greatly. Functional assignments generated from the same gene using different resources may generate very different results (35).

High-quality valid databases that include the 16S rRNA gene and genome sequences are prerequisites for taxonomic classification and identification. Because these pipelines are designed for taxonomic purposes, only data of type strains with validly published names are included. Therefore, gc-

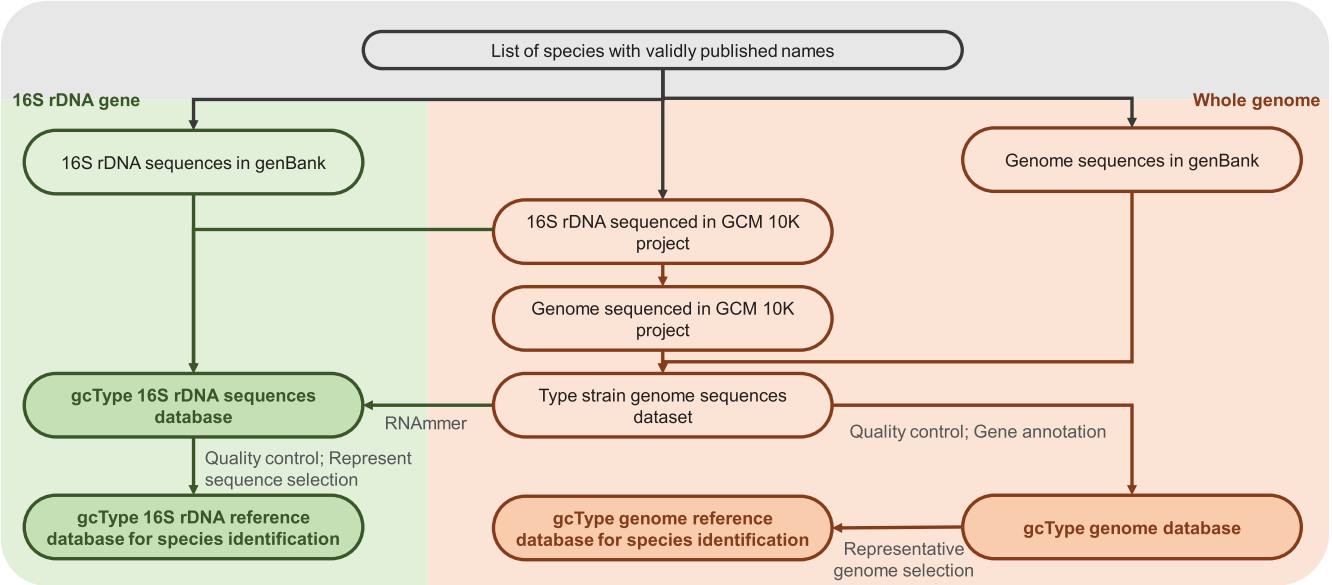

Figure 3. Schematic representation of gcType for data processing: Two sources of 16S rRNA gene sequences are integrated to form the gcType 16S rRNA gene database. Sequence data is further aligned and trimmed to form a reference database for a new species identification pipeline. Publicly available genome sequences and GCM 10K genome sequences are processed by the GCM microbial genome annotation pipeline to form a type strain genome database. The assembled genomes from these two sources are integrated after quality control to form a type strain genome reference database for a new species identification pipeline.

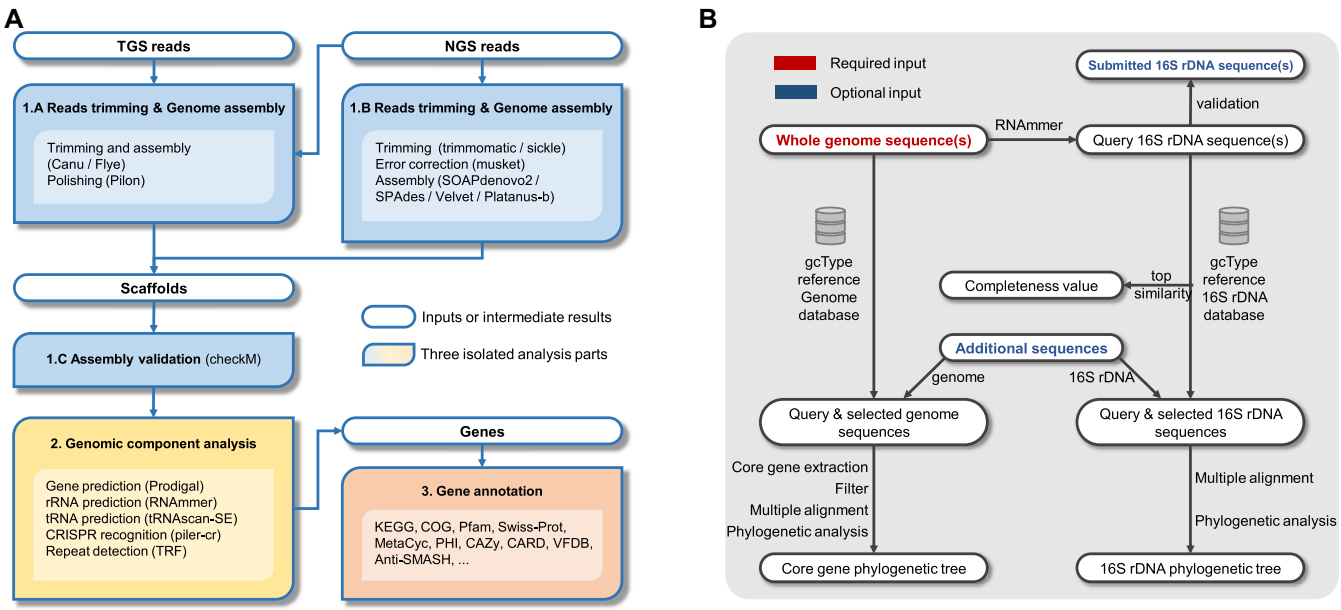

Figure 4. Workflow of gcType data analysis pipelines. (**A**) GCM microbial genome assembly and annotation pipeline. (**B**) New species identification pipeline.

Type features two reference database integrated pipelines: genome assembly and annotation pipeline and the new species identification pipeline.

Genome assembly and annotation pipeline

The genome assembly and annotation pipeline comprises three analytical procedures (Figure 4A): (i) processing of raw reads and assembly, (ii) genomic component analysis and (iii) gene annotation.

(1) **Processing of raw reads and assembly**

A) If TGS long reads (PacBio or Nanopore reads) are provided as the input, the raw sequencing reads are trimmed and assembled into contigs or scaffolds using Canu (36) or Flye (37). If NGS short reads (Illumina paired-end reads) are also provided, they will be used to enhance contigs or scaffolds using Pilon (38).

B) If only NGS short reads are provided, raw reads are trimmed into clean reads using Sickle (https://github.com/najoshi/sickle) or Trimmomatic (39), corrected us-

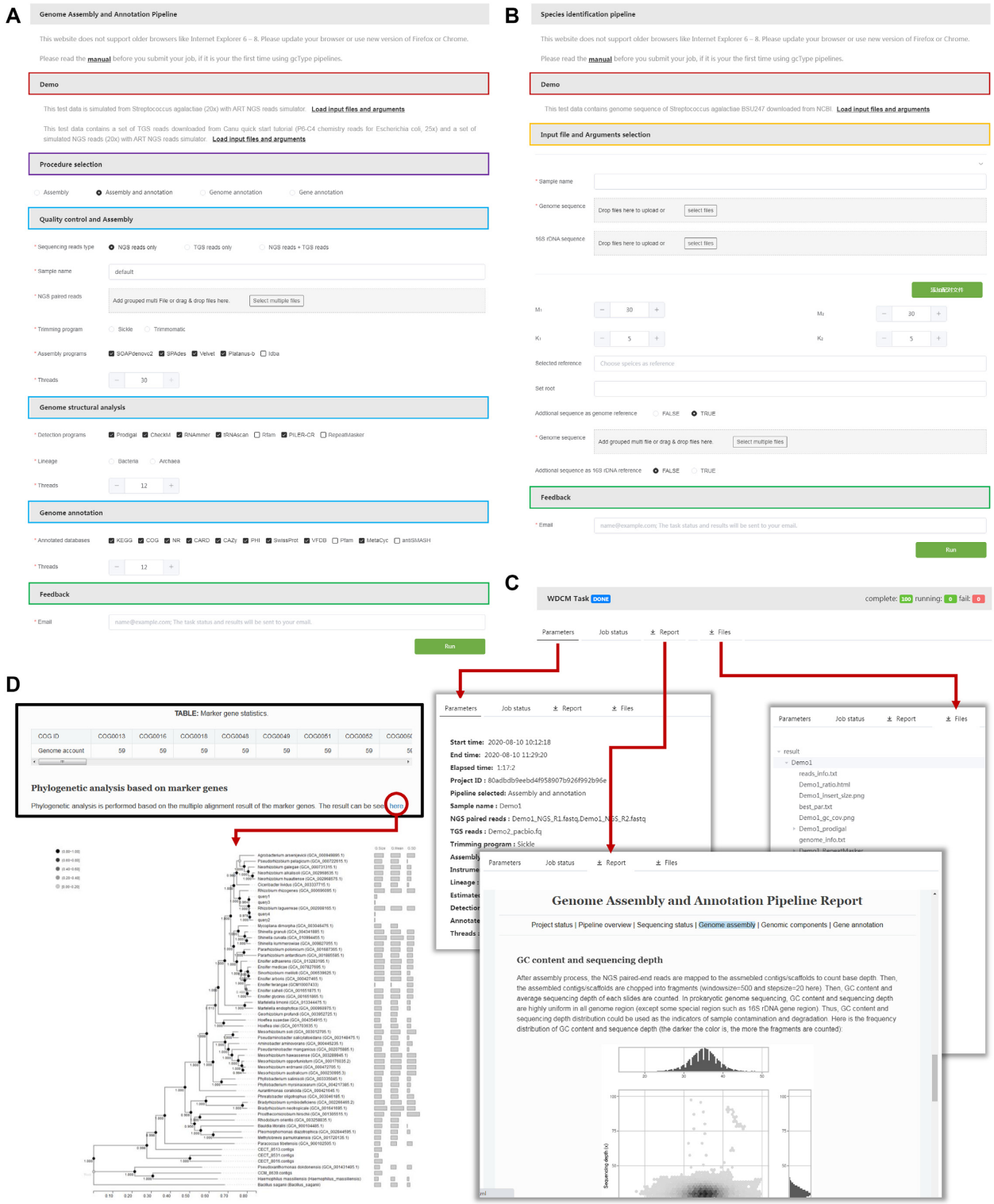

Figure 5. Submission pages and results of the analysis for the two pipelines, provided by gcType. (**A**) Submission page of a genome assembly and annotation pipeline. Different analytical tools and annotation databases are provided. (**B**) Submission page of a species identification pipeline. (**C**) Pages of the online results include parameters selected by the user, logs, report and output file. (**D**) An example of a phylogenetic analysis based on marker genes from a species identification pipeline.

ing Musket (40) and independently assembled into contigs or scaffolds using multiple assemblers (e.g. SOAPdenovo2 (41), SPAdes (42), Velvet (43) and Platanus (44)). Then, the best assembly result is selected according to N50, N75, the contig number, the length of the largest contig and the number of total bases and ambiguous bases. Thereafter, the reads are mapped to the best assembly result to check for mis-assemblies and evaluate the reads coverage.

C) The final assembly result (i.e. best assembly) is used to estimate the completeness and contamination of the genome using checkM (45) and to perform further genomic component analysis.

(2) **Genomic component analysis:** Genomic component analysis involves CRISPR array recognition using PILER-CR (46), repetitive structure detection using TRF (47), non-coding RNA prediction using tRNAscanSE (48) and RNAmmer and gene prediction using Prodigal (49). The analysis is performed based on the final assembly result.

(3) **Gene annotation**: Predicted genes are annotated using several databases, including KEGG, GO (50), COG, NR (51), Swiss-Prot (52), AntiSMASH (53), MetaCyc (54), PHI (55), Pfam (56), CARD and VFDB (57).

New species identification pipeline

For the new species identification pipeline (shown in Figure 4B), the new type strain genome sequence is used as the query in a similarity search against the gcType 16S rRNA gene and genome sequence reference database following the recommendations for the use of genome data in IJSEM.

First, 16S rRNA gene sequence(s) are extracted from the submitted query genome sequence. If a full-length 16S rRNA sequence of the same type strain sequenced using the Sanger method is available, it is compared with the 16S rRNA sequence extracted from the whole-genome assembly to ensure authenticity of the data.

Second, the 16S rRNA gene sequence is aligned to the gcType 16S rRNA gene database using the BLAST tool. Sequences with the highest similarity are used to estimate 16S rRNA gene completeness (58). Users are allowed to select a set of 16S rRNA gene sequences for further phylogenetic analyses from neighbouring sequences or sequences with a remote distance to serve as the reference.

Third, the submitted genome sequence is aligned to the gcType whole-genome reference database using Mash (59) to calculate the genome distance. Various genome similarity metrics, including ANIb (60), FastANI (61), orthoANIb and orthoANIu (62), are provided for the calculation of similarity between a submitted genome sequence and the selected sequences.

Finally, selected 16S rRNA gene sequences are aligned using MAFFT (63) or MUSCLE (64). Then, MEGA (65), FastTree (66) and RAxML (67) are used to perform phylogenetic analyses. For the selected genome sequences, 56 marker genes (68) are extracted and used to perform phylogenetic analyses.

Online services and results of analyses of these two pipelines are displayed in Figure 5A-5D.

Table 1. Top 10 culture collections with the largest number of type strains

No.	Culture collection	Country	Preserved type strains[a]	Sequenced type strains
1	DSMZ	Germany	9106	3963
2	JCM	Japan	7200	824
3	ATCC	United States	4477	911
4	BCCM/LMG	Belgium	3513	385
5	NBRC	Japan	3369	716
6	KCTC	Korea	3243	220
7	CCUG	Sweden	3030	176
8	CIP	France	2784	61
9	NRRL	United States	1592	606
10	CGMCC	China	1558	145

[a]The number of type strains preserved in each culture collection was manually extracted from IJSEM. Type strains which were obtained from other culture collections by exchange of strains were not included.

RESULTS AND DISCUSSION

The ability to accurately identify and organise microorganisms into appropriate taxonomic groups is essential for the functional research of microorganisms. With considerable advancements in genome sequencing and data analytics, genomics is the most powerful and efficient method for studying the origins, evolution and interactions of a species and the diversity of the microbial world. As the taxonomic representatives of various species, type strains are very essential resources for genome sequencing. Currently, type strains are preserved in internationally renowned culture collections such as Leibniz-Institut Deutsche Sammlung von Mikroorganismen und Zellkulturen GmbH (DSMZ), Japan Collection of Microorganisms (JCM)/RIKEN BioResource Center and American Type Culture Collection. The number of type strains and the diversity of species are increasing in culture collections across the world. The top five culture collections with the largest type strain deposition account for 88.02% of all validly published species, and 97.13% of all published species are represented in ten culture collections.

The operations of microbial collections have changed enormously over the past 20 years owing to the availability of multidisciplinary data as well as advances in analytical methods and bioinformatics. Traditional culture collections have made great efforts to explore the diversity of microorganisms and collect information on their genes, properties and products. For instance, with the joint effort of the Department of Energy (DOE) Joint Genome Institute (JGI) (69), DSMZ has already published 3,963 type strain genome sequences (Table 1). Large lists of published strain sequences are limited to culture collections that allocate large budgets to sequence, analyse and publish these sequences. Because this is not feasible for most collections, the GCM 10K project can assist them in getting their items sequenced and published while still being available for their repositories.

The number of sequenced bacterial and archaeal genomes has grown exponentially in recent years. As a sequencing center, DOE JGI is currently the largest generator of type strain sequences and has published 3066 sequences to date, followed by WDCM, which has

Table 2. Sequencing efforts of global sequencing centres

No.	Sequencing centres	Number of publicly available type strain genomes
1	DOE Joint Genome Institute (JGI)	3066
2	WDCM GCM 10K project	1049
3	National Institute of Technology and Evaluation	386
4	University of Tokyo	272
5	Shanghai Majorbio Bio-pharm Technology Co.	183
6	J. Craig Venter Institute	147
7	Broad Institute	131
8	Washington University in St. Louis	125
9	Wellcome Trust Sanger Institute	97
10	Baylor College of Medicine	95
	Others	7784

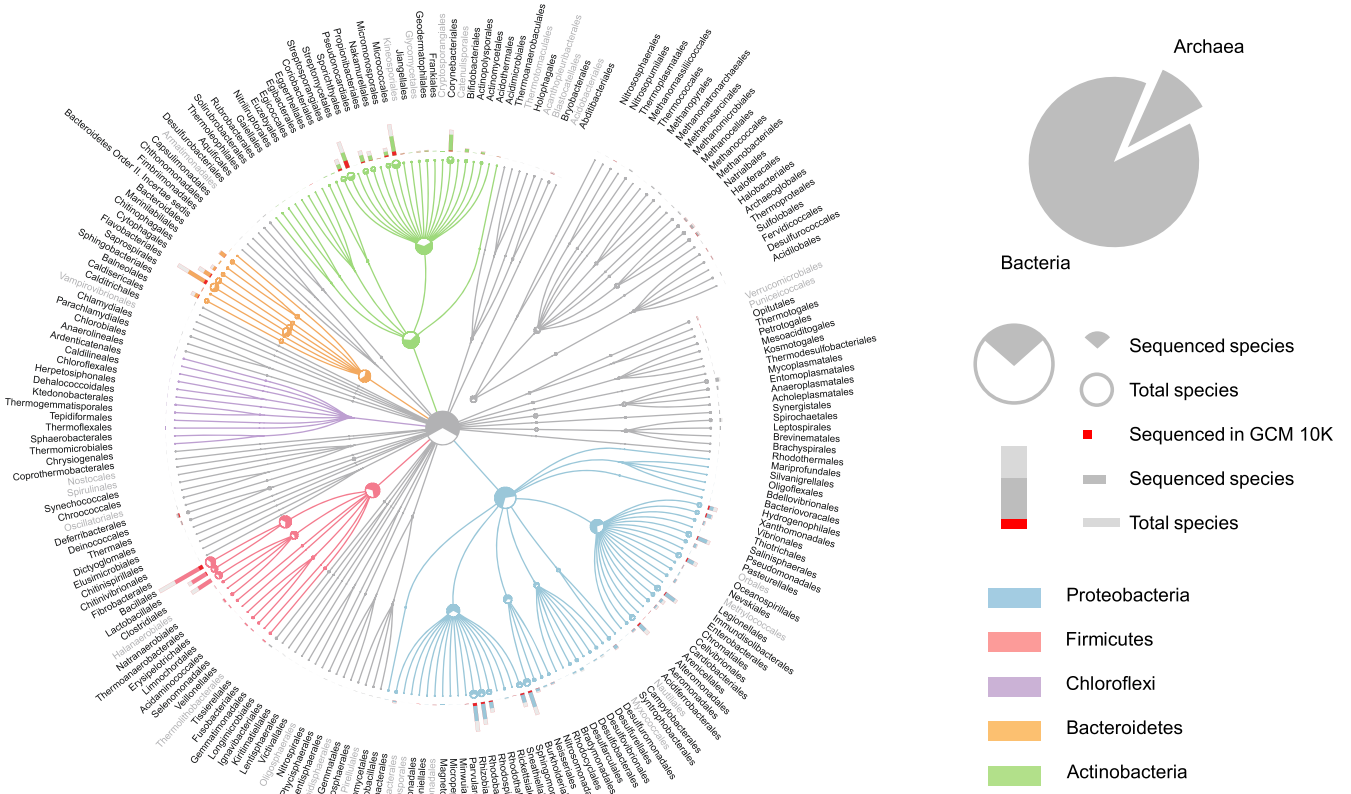

Figure 6. Current sequencing status of prokaryotic organisms by different orders. Top five phyla with the largest number of orders are highlighted. Published type strain genomes of the GCM 10K project are widely distributed in various orders. For orders that have more than nine species (nine is the median number of species of all orders), the order names are coloured in grey if <50% of species have type strain genome sequences, indicating more sequencing efforts should be focussed on these orders.

published 1049 genome sequences of type strains after two year of working with the GCM 10K sequencing project. The data collected in the GCM 10K project is the result of a collaboration of 22 global culture collections. The National Institute of Technology and Evaluation and the University of Tokyo rank third and fourth, generating 386 and 272 type strain genome sequences, respectively, with most of these strains coming from the Biological Resource Center/National Institute of Technology and Evaluation (NBRC) and JCM. The top ten sequencing centres contribute 42% of the total number of sequences, while the remaining 58% is contributed by 477 sequencing centres (Table 2).

To date, whole-genome sequences of >13 962 type strains are publicly available. However, over one-fourth of all bacterial type strains have yet to be sequenced (Figure 6). Among the sequenced type strains, less than 21% (2,911) have high-quality, completed genomes. The remaining 11 051 genomes have been published as draft genomes, meaning that each one is composed of numerous contigs or scaffolds rather than a single contiguous and complete sequence. These draft sequences are valuable for some applications but do

not permit the scientific community to fully study topics which require greater detail, such as structure, evolution and comparative genomics. Additional efforts to sequence, catalogue and characterise high-quality type strain genomes to provide a comprehensive coverage of all species with validly published names are urgently needed.

FUTURE DIRECTIONS

Although taxonomic and phylogenetic analyses based on 16S rRNA gene sequences are commonly used methods for bacterial identification, the prevalence of genome-based taxonomy has become more prevalent as a result of the increasing availability of high-quality reference databases and user-friendly analysis pipelines. Taxonomists are required to provide ANI or dDDH comparisons to type strains when identifying isolates, proposing new species, or reclassifying accepted species. However, because many type strain genomes of species with validly published names remain unsequenced, taxonomists require these additional type strain genomes so that their target strains can be analysed with appropriate scientific rigour. To address this critical need, the GCM 10K project is providing free sequencing services to complete these efforts, which are now complemented by the gcType platform.

In the future, gcType will continuously integrate data from various resources and publish updated results of the GCM 10K type strain sequencing project, making it a unique resource for comprehensive type strain genome information. gcType will also integrate data pertaining to genomic and phenotypic characteristics of a taxonomic group to help define associations between genomic and phenotypic characteristics and further predict metabolic features based on the combination of the integrated information. Finally, as the high-quality reference genomic data will provide accurate taxonomic and functional predictions for metagenomic data, genome assemblies from metagenomic samples will be integrated to expand the utility of the database for uncultured prokaryotic organisms.

DATA AVAILABILITY

There are no access restrictions for the academic use of the platform. All public available genomic data or metadata are freely accessible. Both the 16S rRNA gene sequences and whole-genome sequences generated by the GCM 10K project have been continuously submitted to INSDC after data curation. Access to gcType is free at: http://gctype.wdcm.org/.

ACKNOWLEDGEMENTS

We would like to thank the members of the World Federation for Culture Collections for the GCM 10K sequencing project. We would like to thank the supports from National Institute of Genetics (NIG) members, Professor Yasukazu Nakamura, Dr Yasuhiro Tanizawa and Asami Fukuda for their helps on data curation and annotation. We would also like to thank the technical support staff at BGI Shenzhen for their sequencing services and Ronglian for the cloud-based computing platform.

FUNDING

National Key Research Program of China [2017YFD0400302, 2017YFC1201202, 2018YFD0400201]; International Partnership Program of the Chinese Academy of Sciences [153211KYSB, 201900211]; Strategic Priority Research Program of the Chinese Academy of Sciences [XDA19050301]; 13th Five-year Informatization Plan of the Chinese Academy of Sciences [XXH13506, XXH13505]; Key Research Program of the Chinese Academy of Sciences [KFZD-SW-219]; National Science Foundation for Young Scientists of China [31701157, 31801106]; European Union and co-financed by European Social Fund [EFOP-3.6.3-VEKOP-16-2017-00005]. Funding for open access charge: International Partnership Program of Chinese Academy of Sciences [153211KYSB 201900211].

Conflict of interest statement. None declared.

REFERENCES

1. Whitman,W.B., Coleman,D.C. and Wiebe,W.J. (1998) Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. U.S.A.*, **95**, 6578–6583.
2. Curtis,T.P., Sloan,W.T. and Scannell,J.W. (2002) Estimating prokaryotic diversity and its limits. *Proc. Natl Acad. Sci. U.S.A.*, **99**, 10494–10499.
3. Skerman,V.B.D., McGowan,V. and Sneath,P.H.A. (1980) Approved lists of bacterial names. *Int. J. Syst. Bacteriol.*, **30**, 225–420.
4. Tindall,B.J., Rosselló-Móra,R., Busse,H.J., Ludwig,W. and Kämpfer,P. (2010) Notes on the characterization of prokaryote strains for taxonomic purposes. *Int. J. Syst. Evol. Microbiol.*, **60**, 249–266.
5. Wayne,L.G., Brenner,D.J., Colwell,R.R., Grimont,P.A.D., Kandler,O., Krichevsky,M.I., Krichevsky,M.I., Moore,L.H., Moore,W.E.C., Murray,R.G.E. *et al.* (1987) Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Evol. Microbiol.*, **37**, 463–464.
6. Varghese,N.J., Mukherjee,S., Ivanova,N., Konstantinidis,K.T., Mavrommatis,K., Kyrpides,N.C. and Pati,A. (2015) Microbial species delineation using whole genome sequences. *Nucleic Acids Res.*, **43**, 6761–6771.
7. Kim,M., Oh,H.S., Park,S.C. and Chun,J. (2014) Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. *Int. J. Syst. Evol. Microbiol.*, **64**, 346–351.
8. Meier-Kolthoff,J.P., Auch,A.F., Klenk,H.-P. and Göker,M. (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*, **14**, 60.
9. Chun,J., Oren,A., Ventosa,A., Christensen,H., Arahal,D.R., da Costa,M.S., Rooney,A.P., Yi,H., Xu,X.W., De Meyer,S., Trujillo,M.E. *et al.* (2018) Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int. J. Syst. Evol. Microbiol.*, **68**, 461–466.
10. Mukherjee,S., Seshadri,R., Varghese,N.J., Eloe-Fadrosh,E.A., Meier-Kolthoff,J.P., Göker,M., Coates,R.C., Hadjithomas,M., Pavlopoulos,G.A., Paez-Espino,D. *et al.* (2017) 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.*, **35**, 676–683.
11. Wu,L., McCluskey,K., Desmeth,P., Liu,S., Hideaki,S., Yin,Y., Moriya,O., Itoh,T., Kim,C.Y., Lee,J.S. *et al.* (2018) The global catalogue of microorganisms 10K type strain sequencing project: closing the genomic gaps for the validly published prokaryotic and fungi species. *Gigascience*, **7**, 5.
12. Wu,L. and Ma,J. (2019) The Global Catalogue of Microorganisms (GCM) 10K type strain sequencing project: providing services to taxonomists for standard genome sequencing and annotation. *Int. J. Syst. Evol. Microbiol.*, **69**, 895–898.

13. Galperin,M., Makarova,K., Wolf,Y. and Koonin,E. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
14. Chen,I.-M.A., Chu,K., Palaniappan,K., Pillay,M., Ratner,A., Huang,J., Huntemann,M., Varghese,N., White,J.R. and Seshadri,R. (2019) IMG/M v. 5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.*, **47**, D666–D677.
15. Meier-Kolthoff,J.P. and Göker,M. (2019) TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. *Nat. Commun.*, **10**, 1–10
16. Reimer,L.C., Vetcininova,A., Carbasse,J.S., Söhngen,C., Gleim,D., Ebeling,C. and Overmann,J. (2019) BacDive 2019: bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Res.*, **47**, D631–D636.
17. Parte,A.C., Carbasse,S., Joaquim,M.-K., Jan,P., Reimer,L.C. and Goker,M. (2020) List of prokaryotic names with standing in nomenclature (LPSN) moves to the DSMZ. *Int. J. Syst. Evol. Microbiol.*, doi:10.1099/ijsem.0.004332.
18. Federhen,S. (2015) Type material in the NCBI Taxonomy Database. *Nucleic Acids Res.*, **43**, D1086–D1098
19. Parks,D.H., Chuvochina,A., Maria.,W., David,W., Rinke,C., Skarshewski,A., Chaumeil,P.-A. and Hugenholtz,P.,(2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.*, **36**, 996–1004
20. Sayers,E.W., Cavanaugh,M., Clark,K., Ostell,J., Pruitt,K.D. and Karsch-Mizrachi,I. (2020) GenBank. *Nucleic Acids Res.*, **48**, D84–D86
21. Galperin,M., Makarova,K., Wolf,Y. and Koonin,E. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
22. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
23. Jia,B., Raphenya,A., Alcock,B., Waglechner,N., Guo,P., Tsang,K., Lago,B., Dave,B., Pereira,S., Sharma,A. *et al.* (2017) CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, **45**, D566–D573.
24. Buels,R., Yao,E., Diesh,C.M., Hayes,R.D., Munoz-Torres,M., Helt,G., Goodstein,D.M., Elsik,C.G., Lewis,S.E., Stein,L., Holmes,I.H. *et al.* (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.*, **17**, 66.
25. Stothard,P. and Wishart,D.S. (2005) Circular genome visualization and exploration using CGView. *Bioinformatics*, **21**, 537–539
26. Parker,C.T., Tindall,B.J. and Garrity,G.M. (2019) International code of nomenclature of prokaryotes. *Int. J. Syst. Evol. Microbiol.*, **69**, S1–S111.
27. Field,D., Garrity,G., Gray,T., Morrison,N., Selengut,J., Sterk,P., Tatusova,T., Thomson,N., Allen,M.J., Angiuoli,S.V. *et al.* (2008) The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.*, **26**, 541–547.
28. Yoon,S.H., Ha,S.M., Kwon,S., Lim,J., Kim,Y., Seo,H. and Chun,J. (2017) Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int. J. Syst. Evol. Microbiol.*, **67**, 1613
29. Quast,C., Pruesse,E., Yilmaz,P., Gerken,J., Schweer,T., Yarza,P., Peplies,J. and Glöckner,F.O. (2012) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.*, **41**, D590–D596
30. Cole,J.R., Chai,B., Farris,R.J., Wang,Q., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Bandela,A.M., Cardenas,E., Garrity,G.M. and Tiedje,J.M. (2007) The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res.*, **35**, D169–D172.
31. Park,S.C. and Won,S. (2018) Evaluation of 16S rRNA databases for taxonomic assignments using a mock community. *Genomics Inform.*, **16**, e24.
32. Lagesen,K., Hallin,P., Rodland,E.A., Staerfeldt,H.H., Rognes,T. and Ussery,D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.
33. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
34. Kalvari,I., Argasinska,J., Quinones-Olvera,N., Nawrocki,E.P., Rivas,E., Eddy,S.R., Bateman,A., Finn,R.D. and Petrov,A.I. (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.*, **46**, D335–D342.
35. Chen,I.M., Markowitz,V.M., Chu,K., Anderson,I., Mavromatis,K., Kyrpides,N.C. and Ivanova,N.N. (2013) Improving microbial genome annotations in an integrated database context. *PLoS One*, **8**, e54859.
36. Koren,S.1, Walenz,B.P., Berlin,K., Miller,J.R., Bergman,N.H. and Phillippy,A.M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.*, **27**, 722–736.
37. Kolmogorov,M., Yuan,J., Lin,Y. and Pevzner,P.A. (2019) Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.*, **37**, 540–546.
38. Walker,B.J., Abeel,T., Shea,T., Priest,M., Abouelliel,A., Sakthikumar,S., Cuomo,C.A., Zeng,Q., Wortman,J., Young,S.K. *et al.* (2014) Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, **9**, e112963.
39. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
40. Liu,Y., Schröder,J. and Schmidt,B. (2012) Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics*, **29**, 308–315.
41. Luo,R., Liu,B., Xie,Y., Li,Z., Huang,W., Yuan,J., He,G., Chen,Y., Pan,Q., Liu,Y. *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience*, **1**, 18.
42. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S. and Prjibelski,A.D. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.*, **19**, 455–477.
43. Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
44. Kajitani,R., Toshimoto,K., Noguchi,H., Toyoda,A., Ogura,Y., Okuno,M., Yabana,M., Harada,M., Nagayasu,E., Maruyama,H. *et al.* (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.*, **24**, 1384–1395.
45. Parks,D.H., Imelfort,M., Skennerton,C.T., Hugenholtz,P. and Tyson,G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.
46. Edgar,R.C. (2007) PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, **8**, 18.
47. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
48. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
49. Hyatt,D., Chen,G.L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.
50. The Gene Ontology Consortium2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.*, **45**, D331–D338.
51. Eric,W.S., Richa,A., Evan,E.B., J Rodney,B., Kathi,C., Karen,C., Ryan,C., Nicolas,K.F., Timothy,H. *et al.* (2019) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **47**, D1, D23–D28.
52. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M., Bansal,P., Bridge,A.J., Poux,S., Bougueleret,L. and Xenarios,I. (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt KnowledgeBase: how to use the entry view. *Methods Mol. Biol.*, **1374**, 23–54.
53. Blin,K., Shaw,S., Steinke,K., Villebro,R., Ziemert,N., Lee,S.Y., Medema,M.H. and Weber,T. (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.*, **47**, W81–W87.
54. MetaCyc,C.R., Billington,R., Fulcher,C.A., Keseler,I.M., Kothari,A., Krummenacker,M., Latendresse,M., Midford,P.E., Ong,Q., Ong,W.K. *et al.* (2018) The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **46**, D633–D639.
55. Urban,M., Cuzick,A., Rutherford,K., Irvine,A., Pedro,H., Pant,R., Sadanadan,V., Khamari,L., Billal,S., Mohanty,S. *et al.* (2017)

PHI-base: a new interface and further additions for the multi-species pathogen-host interactions database. *Nucleic Acids Res.*, **45**, D604–D610.
56. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
57. Liu,B., Zheng,D.D., Jin,Q., Chen,L.H. and Yang,J. (2019) VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.*, **47**, D687–D692.
58. Kim,O.S., Cho,Y.J., Lee,K., Yoon,S.H., Kim,M., Na,H., Park,S.C., Jeon,Y.S, Lee,J.H., Yi,H. *et al.* (2012) Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species. *Int. J. Syst. Evol. Microbiol.*, **7**, 16–21.
59. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
60. Richter,M. and Rosselló-Móra,R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 19126–19131.
61. Jain,C., Rodriguez-R,L.M., Phillippy,A.M., Konstantinidis,K.T. and Aluru,S. (2018) High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.*, **9**, doi:10.1038/s41467-018-07641-9.
62. Lee,I., Kim,Y.O., Park,S.C. and Chun,J. (2016) OrthoANI: an improved algorithm and software for calculating average nucleotide identity. *Int. J. Syst. Evol. Microbiol.*, **66**, 1100–1103.
63. Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.*, **9**, 286–298.
64. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
65. Kumar,S., Stecher,G., Li,M., Knyaz,C. and Tamura,K. (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol.*, **35**, 1547–1549.
66. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2-approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
67. Stamatakis,A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.
68. Eloe-Fadrosh,E.A., Paez-Espino,D., Jarett,J., Dunfield,P.F., Hedlund,B.P., Dekas,A.E., Grasby,S.E., Brady,A.L., Dong,H. and Briggs,B.R. (2016) Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat. Commun.*, **7**, 10476.
69. Wu,D., Hugenholtz,P., Mavromatis,K., Pukall,R., Dalin,E., Ivanova,N.N., Kunin,V., Goodwin,L., Wu,M., Tindall,B.J. *et al.* (2009) A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature*, **462**, 1056–1060.