Published online 26 October 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D637–D648***
doi: 10.1093/nar/gky1008

gcMeta: a Global Catalogue of Metagenomics platform to support the archiving, standardization and analysis of microbiome data

Wenyu Shi[1,†], Heyuan Qi[1,†], Qinglan Sun[1], Guomei Fan[1], Shuangjiang Liu[1,2], Jun Wang[3], Baoli Zhu[3,4,5,6], Hongwei Liu[7], Fangqing Zhao[8], Xiaochen Wang[1], Xiaoxuan Hu[1], Wei Li[1], Jia Liu[9], Ye Tian[9], Linhuan Wu[1,2,*] and Juncai Ma[1,2,*]

[1]Microbial Resource and Big Data Center, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China, [2]State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China, [3]CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Science, Beijing 100101, China, [4]University of Chinese Academy of Sciences, Beijing 100049, China, [5]Collaborative Innovation Centre for Diagnosis and Treatment of Infectious Diseases First Attainted Hospital, College of Medicine, Zhejiang University, Hangzhou 310058, China, [6]Beijing Key Laboratory of Antimicrobial Resistance and Pathogen Genomics, Beijing 100101, China, [7]State Key Laboratory of Mycology, Institute of Microbiology, Chinese Academy of Science, Beijing 100101, China, [8]Computational Genomics Lab, Beijing Institutes of Life Science, Chinese Academy of Sciences, Beijing 100101, China and [9]Internet of Things Information Technology and Application Laboratory, Computer Network Information Center, Chinese Academy of Sciences. Beijing 100101, China

Received August 14, 2018; Revised September 18, 2018; Editorial Decision October 10, 2018; Accepted October 13, 2018

ABSTRACT

Meta-omics approaches have been increasingly used to study the structure and function of the microbial communities. A variety of large-scale collaborative projects are being conducted to encompass samples from diverse environments and habitats. This change has resulted in enormous demands for long-term data maintenance and capacity for data analysis. The Global Catalogue of Metagenomics (gcMeta) is a part of the 'Chinese Academy of Sciences Initiative of Microbiome (CAS-CMI)', which focuses on studying the human and environmental microbiome, establishing depositories of samples, strains and data, as well as promoting international collaboration. To accommodate and rationally organize massive datasets derived from several thousands of human and environmental microbiome samples, gcMeta features a database management system for archiving and publishing data in a standardized way. Another main feature is the integration of more than ninety web-based data analysis tools and workflows through a Docker platform which enables data analysis by using various operating systems. This platform has been rapidly expanding, and now hosts data from the CAS-CMI and a number of other ongoing research projects. In conclusion, this platform presents a powerful and user-friendly service to support worldwide collaborative efforts in the field of meta-omics research. This platform is freely accessible at https://gcmeta.wdcm.org/.

INTRODUCTION

'Meta-omics' (e.g. metataxonomics, metagenomics and metatranscriptomics) approaches have been increasingly used to study the structure, function and intercellular interactions of the microbial communities and the fundamental mechanisms of microbial life and evolution. Dramatic progress in the next generation sequencing technology has made large-scale sampling and sequencing possible, even for individual laboratory. Meta-omics has also promoted collaborative efforts in a grand vision across the international research community, as exemplified by the Earth Microbiome Project (EMP) (1) and Human Microbiome Project (HMP) (2). These collaborative projects have produced large volume of data and hence generated meaningful interpretations from a full spectrum of sources which are impossible with a single independent study. Along with these changes, microbiome research is becoming a data driven science (3) and rapid advances in this area have

[*]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, in their opinion, the first two authors should be regarded as joint First Authors.

brought about significant challenges. Firstly, comparing data from independent research groups becomes difficult if standard operating procedures (SOPs) and reporting standards are not followed. Adhering to universal standards in every step of the study, including sampling, sequencing, data submission, data analysis, and data publication, is necessary to understand the results of a single study within a broader context (4). Recently, significant efforts have been made for the development of universal protocols and standards (5). Although, well-organized and reputable collaborative projects often have built-in standards (http://www.microbiome-standards.org/) and SOPs (6), sometimes it becomes difficult for different projects to implement identical standards. For example, human and environmental microbiota data from different projects are not readily comparable due to inconsistencies in standards, protocols as well as workflows (7). Further developments, and more importantly, adoption of these SOPs and standards by all projects and labs worldwide is crucial for the scientific community. The second challenge is long-term preservation and open access of the data. Integration of all relevant publicly available data is a prerequisite for future cross-studies. A stable and robust data infrastructure is needed that would provide a reliable data archive and rational data organization, thus ensuring data reproducibility and allowing data reinterpretation. The third challenge is to analyze Gigabyte (GB) to Terabyte (TB) scale data on a single computer. Despite the availability of a variety of stand-alone tools, it is almost impossible for any given individual lab to have sufficient infrastructure for data storage and to maintain multi-computer network clusters resources (8). Currently, there are several public resources, including the European Bioinformatics Institute (EBI) metagenomics (9), the Metagenomics RAST server (MG-RAST) (10) and the JGI IMG Integrated Microbial Genomes & Microbiomes (IMG/M) (11). However, considering the rapid increase in data volume and growing demands for data analysis capacity, more public services with the ability to provide data archiving and cloud-based data analysis are required (12).

Because of the wide geographical coverage, rich ecosystem, as well as diverse ethnicity and lifestyles of the people, China harbors enormous diversity of microbial communities. In comparison to developed countries, however, the microbial communities in China are less comprehensively and systematically studied. China also lacks nationwide collaborative projects. The Chinese Academy of Sciences Initiative of Microbiome (CAS-CMI) is one of the leading projects organized at the national level to find solutions to the current challenges in human and environmental health, agriculture and industrial developments. At the same time, it will establish the biobanks (samples, strains and data) of Chinese Microbiome Initiative, and support long term preservation and reuse of data worldwide in a free and open way.

The Global Catalogue of Metagenomics (gcMeta) platform is a part of the CAS-CMI. As a partner database of the World Data Center for Microorganisms (WDCM) (13) as well, gcMeta has two features: firstly, designing and implementing as a standardized and state-of-art database management system to support long-term preservation and integration data from the CAS-CMI project as well as from microbiome research projects worldwide. Secondly, the platform provides web-based tools and pre-defined workflows, along with computing resources for massive data analysis requirements globally.

PLATFORM DESIGN AND IMPLEMENTATION

How to use the platform

The platform supplies management, analysis and publication services for microbiome related data, including genomes, marker genes, metagenomes, metatranscriptomes and their associated metadata (Figure 1). The users can upload the raw data and their metadata into the system via a web submission interface or a data upload web application. After data quality check, the data can be browsed in the system under the user's account. Currently, we provide web-based analysis workflows for marker genes and whole-genome shotgun sequencing (WGS) data. The users can use these workflows or individual tools for data analysis and visualization. The Global Unique Persistent Identifier (GUID) system is used for the open data. To publish the data, users should change the status of their data from 'private' to 'public', then, the system assigns a persistent identifier (PID) (http://www.pidconsortium.eu/) to the each of the records. The PID is used to cite the data elsewhere and provide a report to the users.

For data protection, login is required before data submission and exploring the full functions. We provide a temporary guest account effective for 24 h along with any submissions, uploaded files and analysis results. All the public available primary raw data or metadata could be downloaded. Access to gcMeta is free at https://gcmeta.wdcm.org/.

Database design

The database hosts information on samples and their associate metadata, and primary 'raw' data. A relational database is used to host all relevant data. Schema of the database is shown in Figure 2. The major data record types are 'Study', 'Sample', 'Experiment' and 'Sequence'. 'Study' could include several 'sub-studies' and is related to 'Sample' by the 'Study ID'. The samples and their associated metadata are recorded. 'Sample' is referenced to 'Experiment'. 'Experiment' is further referenced to sequence information. The sequence information contains the sequencing methods and strategies, as well as the processing of the sequencing results, including data quality control and assembly. The gcMeta platform is implemented by an open source database system PostgreSQL.

Ontologies and data standards are crucial to ensure reusability and interoperability of data. To ensure data comparability and consistency between CAS-CMI and public data resources, gcMeta adopts the Minimum Information about Metagenomic Sequence (MIMS) (14) and Minimum Information about a MARKer gene Sequence (MIMARKS) (15). It also uses the Minimum Information of any(x) Sequence (MIxS), which describes 15 different environmental packages to specify the environmental context of a microbial sample. The Environment Ontology (ENVO) (16) for the three environmental metadata fields including

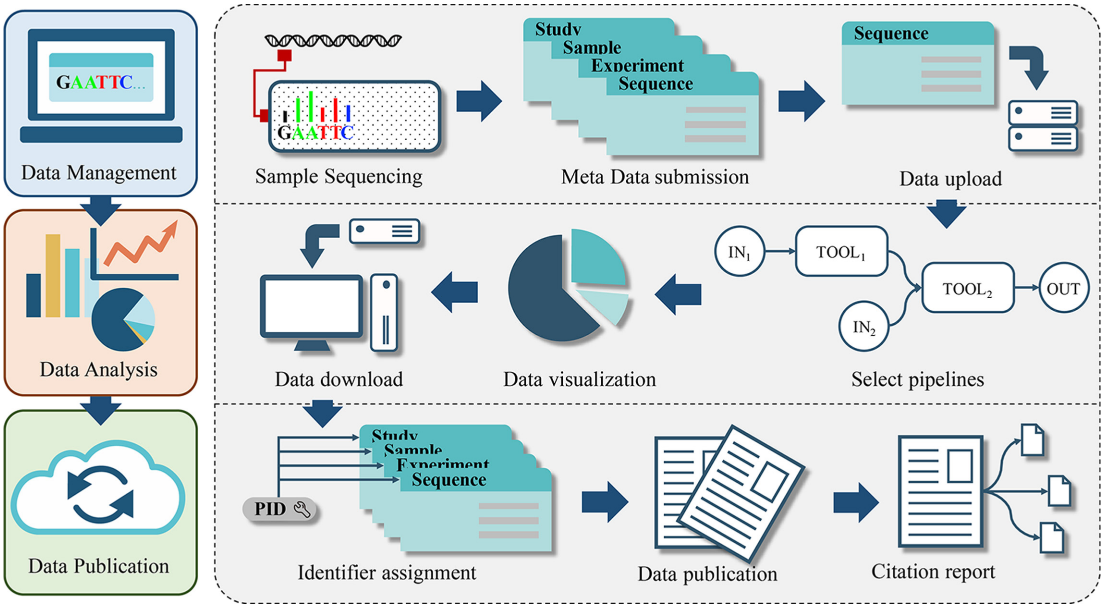


Figure 1. General pipeline of the gcMeta platform. The functional services of gcMeta can be described in three parts: data management, data analysis and data publication. Users submit the meta-data and primary raw data into the system under their own accounts. Users are allowed to analyze the data by preinstalled tools and workflows. Data and results could be downloaded for further analysis. A unique identifier PID would be assigned to each record before the data is public. If the data is further cited in other resources with the PID, the citation could be traced automatically.

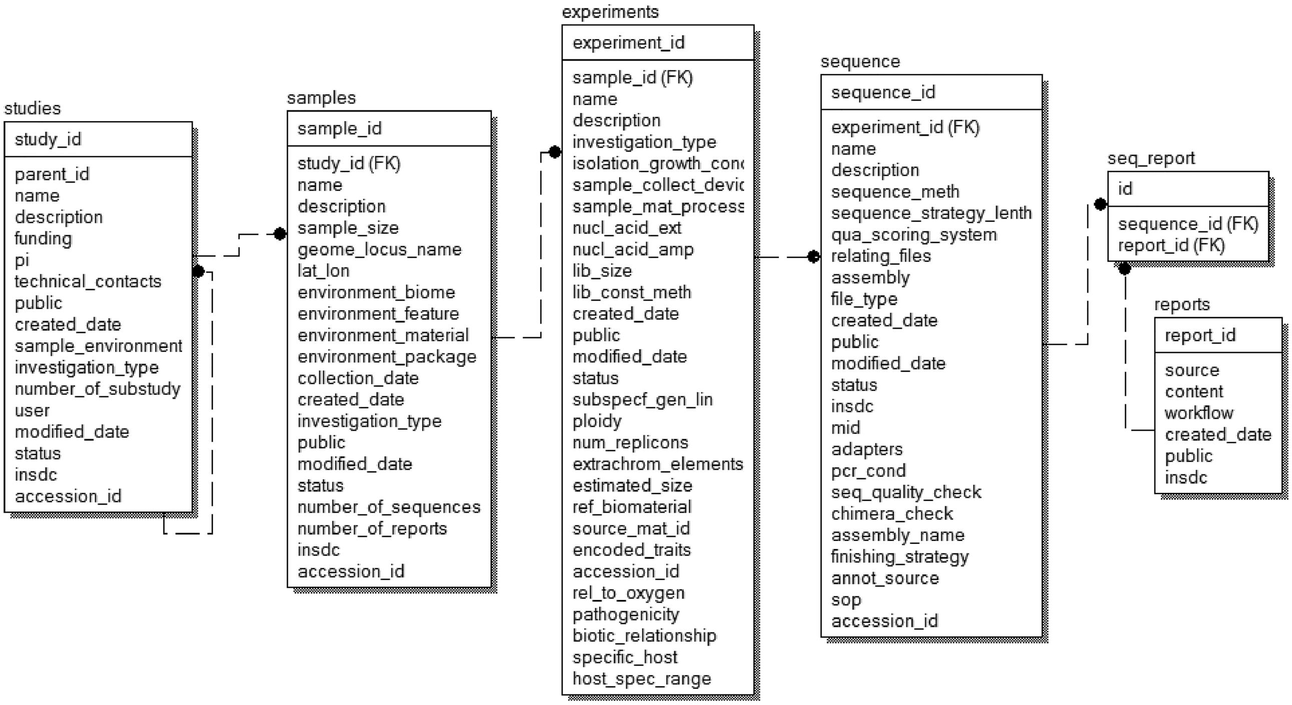


Figure 2. Database schema of gcMeta. Main data structure and relationships between the different tables are illustrated.

environmental biome, feature and material is used to describe the sampling in the system, using a total of 95 controlled terms.

Data sources

As of August 2018, gcMeta has archived a total of 3053 studies and 124 052 samples, hosting more than 120 TB of sequencing data. We have two major data sources. The first is publicly available data (e.g. MG-RAST, EBI metagenomics and HMP). Publicly available data are integrated based on isolation sources, environmental features and experiment types, and thus allowing data comparison across different data resources. Efforts were made to overcome varying data quality level, format and (lack of) metadata description. Expression was unified according to environmental ontologies. Secondly, gcMeta serves as a data catalogue for the CAS-CMI project and some other ongoing projects in China. This platform has been rapidly expanding, and now hosts CAS-CMI project sample data from waste water, human gut, characteristic Chinese fermented food and so on. The number of samples from these projects is more than 2000 up to now.

Data management services

The system allows two routes of data submission. Users can establish a record of their studies and associated samples and experiments online one by one through a web form as indicated in Figure 3. For raw sequence data, users can upload data to the system using a web application. To submit high volume of data in a single session, users can choose a simple tab-delimited file format such as Microsoft Excel. Implementation of data standards occurs during database design as well as prior to the generation of the data in the sampling stage. We also use data validation tools during the on-batch data submission. The system is able to accept data submission from all over the world and offers standardized quality control for the submitted data.

The platform under the CAS-CMI project coordinates with other research institutes, universities, and hospital across China for data repository. Since the ongoing projects are one of the forms of data sources, some data are currently not available to the general public at this stage; data can be accessed via project members only, but it will eventually be publicly available. Data submitters can upload and share their pre-publication data with their research collaborators. Only controlled-access is available for pre-publication data. When required by journals, the data status could be seamlessly switched from 'private' to 'public' in the system. In this way, we encourage data sharing while protecting data privacy and security. We limit the number of mandatory fields to keep a balance between the burden on contributing scientists and reusability of the data for future analysis. The data could be searched and browsed online after it is submitted.

After the data is set public, PID will be assigned to each 'Study' 'Sample' 'Experiment' and 'Sequence' record. The PID is a Global Unique Persistent Identifier system that provides long-term identifying service similar to Digital Object Identifier (DOI) (http://www.doi.org/). An example of PID in gcMeta for 'Study' is 21.86101/gcm.study.88c292e8f67c11e7b172b49691092464, where '21' is the handle prefix for PID. 'gcm.study' is identifier for 'Study' records in the gcMeta database and '88c292e8f67c11e7b172b49691092464' is a randomly assigned series code for the record. The PID can be used to search the Handle (http://hdl.handle.net) site.

Data analysis and visualization workflows

Metagenomics data are often referred as 'marker gene amplification metagenomics' and 'full shotgun metagenomics'. Depending on the data types, general workflows for data analysis include two different categories. A wide array of tools are currently available to carry out each step of the workflows (17). In gcMeta, we supply analysis tools and workflows which are installed based on Docker technologies, and thus allow users with limited computational resources to perform analysis of metagenomic samples.

Dockerized analysis tools

More than 90 tools could be used for microbial genomic and metagenomic analysis in web-based interactive mode. These tools are grouped into 6 different categories according to their functions: raw reads preprocessing, sequence assembly, genome structural analysis, database annotation, community profiling and sequence alignment (shown in Table 1). (1) Raw reads preprocessing contains formatting, trimming, filtering, error-correcting and other tools, which are used to reformat or filter the raw data for downstream analysis. (2) Sequence assembly includes assembly, extension and validation tools for both DNA and RNA sequences. Short or long reads become contigs, scaffolds, draft genomes or even complete genomes after this process. (3) Genome structural analysis contains tools for gene prediction, tandem or interspersed repeat detection, CRISPR array detection, etc. The outputs of some tools can be used for further annotation with various databases. (4) Database annotation groups some of the automatic gene annotation tools such as Prokka, DFAST and InterProScan. Formatted databases for annotation are stored in gcMeta for sequence alignment. (5) Community profiling includes tools for classification and quantification, *de novo* binning, community function prediction and downstream analysis both from short reads and contigs of metagenomic data. (6) Sequence alignment contains mapping and alignment, phylogenetic and evolution analysis tools.

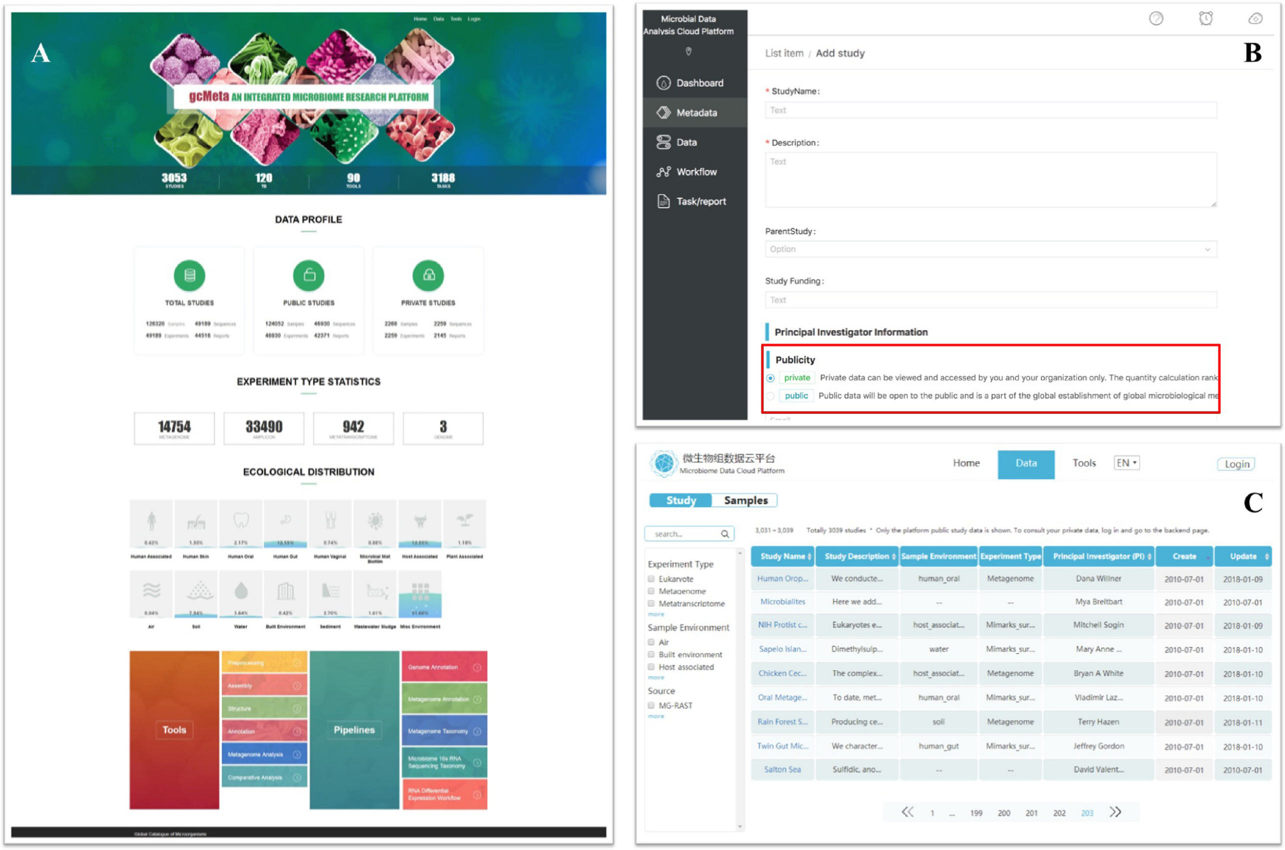
Integrated analysis workflows

Since outputs of upstream tools can be severed as the inputs of downstream tools, tools can be easily concatenated to achieve a predefined workflow in this platform. Till date, there has been no generally accepted 'analysis standard' for a metagenomic analysis workflow. Most workflows involve aspects such as quality control, assembly, binning, taxonomic assignment and functional annotation. However, each workflow is tailored for specific computing resources, aims of analysis and characteristics of the data. In gcMeta, we provide five main workflows for genomes, marker genes, metagenomes analysis. All merged in the workflow overview in Figure 4, and they are:

1) Metagenome assembly and annotation (microbiome sample - NGS reads—quality control—assembly and validation—binning—genome structural analysis—database annotation): In this workflow, we assemble the short reads into contigs. These contigs can be further sorted or binned by similarity to assemble partial to full genomes of microorganisms. The assembled sequences are used for subsequent structural and functional analysis. Firstly, NGS reads, as input, are trimmed into clean reads after performing quality control (host genome contamination removal with Bowtie using parameter 'very-sensitive', quality

Table 1. Tools embedded in the gcMeta platform. The tools belong to the group raw reads preprocessing, sequence assembly, genome structural analysis, database annotation, community profiling and sequence alignment are set as red, blue, purple, orange green and yellow respectively. BBtools software suite (http://jgi.doe.gov/data-and-tools/bbtools/), FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/), fastp (https://github.com/OpenGene/fastp/), Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), minced (https://github.com/ctSkennerton/minced/tree/master) and RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker) are all referenced to their websites

TYPE	TOOL	TYPE	TOOL	TYPE	TOOL
formatting	SRAtoolkit (18)	assembly	SOAPdenovo2 (46)	gene	GeneMark (78)
simulation	ART (19)	assembly	SPAdes (47,48)	gene	FragGeneScan (79)
simulation	pIRS (20)	assembly	MetaVelvet (49)	repeat	XSTREAM (80)
package	BBtools	assembly	ALLPATH-LG (50)	repeat	RepeatMasker
checking	fastQC	assembly	Meta-IDBA (51)	variation	PRISM (81)
trimming	cutadapt (21)	assembly	MEGAHIT (52)	variation	LUMPY (82)
trimming	Trimmomatic (22)	assembly	RayMeta (53)	annotation	Prokka (83)
trimming	fastp	assembly	CANU (54)	annotation	DFAST (84)
trimming	dustmasker (23)	extension	CAP3 (55)	annotation	InterProScan (85)
trimming	DRISEE (24)	extension	SSPACE (56)	annotation	PfamScan (86)
correction	Musket (25)	scaffolding	OPERA (57)	packages	QIIME (87)
correction	SOAPec (26)	validation	QUAST (58,59)	marker	LEfSe (88)
correction	LoRDEC (27)	validation	REAPR (60)	function	PICRUSt (89)
correction	proovread (28)	validation	CheckM (61)	profiling	MetaCV (90)
polishing	Quiver (29)	validation	BUSCO (62)	profiling	k-SLAM (91)
merging	FLASH (30)	assembly	cufflinks/cuffdiff (63,64)	profiling	Kaiju (92)
trimming	Trim Galore	assembly	StringTie (65)	profiling	Centrifuge (93)
distance	orthoANI (31)	expression	Sailfish (66)	profiling	DUDes (94)
clustering	CD-hit (32)	expression	Kallisto (67)	profiling	mOTU (95)
alignment	MUMmer (33)	expression	DESeq2 (68)	profiling	StrainEst (96)
mapping	BWA (34,35)	expression	Ballgown (69)	distance	Mash (97)
mapping	Bowtie2 (36)	assembly	Trinity (70)	profiling	sourmash (98)
formatting	samtools (37)	assembly	Oases (71)	profiling	MetaPhlAn2 (99)
alignment	BLAST (38,39)	assembly	SOAPdenovoTrans (72)	function	HUMAnN2 (100)
alignment	BLAT (40)	CRISPR	PILER-CR (73)	bining	CONCOCT (101)
alignment	diamond (41)	CRISPR	minced	bining	MaxBin2 (102)
alignment	STAR (42)	RNA	tRNAscan SE (74)	bining	MetaBAT2 (103)
alignment	Tophat2 (43)	RNA	RNAmmer (75)	bining	AbundanceBin (104)
alignment	hisat2 (44)	gene	Prodigal (76)	virus	VirFinder (105)
mapping	BLASR (45)	gene	Glimmer (77)	virus-host	VirHostMatcher (106)



3053 STUDIES | 120 TB | 90 TOOLS | 3188 TASKS

DATA PROFILE

TOTAL STUDIES: 126320 Samples, 49189 Sequences, 49189 Experiments, 44516 Reports

PUBLIC STUDIES: 124052 Samples, 46930 Sequences, 46930 Experiments, 42371 Reports

PRIVATE STUDIES: 2268 Samples, 2259 Sequences, 2259 Experiments, 2145 Reports

EXPERIMENT TYPE STATISTICS

14754 METAGENOME | 33490 AMPLICON | 942 METATRANSCRIPTOME | 3 GENOME

ECOLOGICAL DISTRIBUTION

Human Associated	Human Skin	Human Oral	Human Gut	Human Vaginal	Microbial Mat Biofilm	Host Associated	Plant Associated
0.62%	1.30%	2.17%	12.55%	0.74%	0.88%	12.55%	1.18%

Air	Soil	Water	Built Environment	Sediment	Wastewater Sludge	Misc Environment
0.04%	7.94%	3.64%	0.42%	2.70%	1.61%	51.66%

List item / Add study

* StudyName: Text

* Description: Text

ParentStudy: Option

Study Funding: Text

Principal Investigator Information

Publicity

● private Private data can be viewed and accessed by you and your organization only. The quantity calculation rank

○ public Public data will be open to the public and is a part of the global establishment of global microbiological me

Experiment Type
□ Eukaryote
□ Metagenome
□ Metatranscriptome
more

Sample Environment
□ Air
□ Built environment
□ Host associated
more

Source
□ MG-RAST
more

3,031 ~ 3,039 Totally 3039 studies * Only the platform public study data is shown. To consult your private data, log in and go to the backend page.

Study Name	Study Description	Sample Environment	Experiment Type	Principal Investigator (PI)	Create	Update
Human Orop...	We conducte...	human_oral	Metagenome	Dana Willner	2010-07-01	2018-01-09
Microbialites	Here we add...	--	--	Mya Breitbart	2010-07-01	2010-07-01
NIH Protist c...	Eukaryotes e...	host_associat...	Mimarks_sur...	Mitchell Sogin	2010-07-01	2018-01-09
Sapelo Islan...	Dimethylsulp...	water	Mimarks_sur...	Mary Anne ...	2010-07-01	2018-01-10
Chicken Cec...	The complex...	host_associat...	Metagenome	Bryan A White	2010-07-01	2018-01-10
Oral Metage...	To date, met...	human_oral	Mimarks_sur...	Vladimir Laz...	2010-07-01	2018-01-10
Rain Forest S...	Producing ce...	soil	Metagenome	Terry Hazen	2010-07-01	2018-01-11
Twin Gut Mic...	We character...	human_gut	Mimarks_sur...	Jeffrey Gordon	2010-07-01	2018-01-10
Salton Sea	Sulfidic, ano...	--	--	David Valent...	2010-07-01	2010-07-01

Figure 3. Screenshots and examples of user cases in gcMeta. (**A**) Homepage of the gcMeta. Statistic number of public and private studies, samples, experiments and runs are showed in the homepage. (**B**) A screenshot of data submission by web table. Each entry could be set 'private' or 'public' as highlighted in the red box. (**C**) A screenshot of database browser. In the search interface, search results could be filtered by 'experiment type', 'sample environment' and 'data sources'.

viewing with FastQC and trimming with cutadapt and Trimmomatic). During the host contamination removal process, users can upload the host reference genome or use the index file we provide. Next, clean reads are assembled into contig and scaffold (MEGAHIT). After assembly, contigs and scaffolds are clustered into different bins (MaxBin) and used to perform genome structural analysis (CRISPR detection with PILER-CR, gene prediction with Prodigal, RNA identification with tRNAscan). Then, genes are used to perform annotation (annotation with all referred annotation and alignment tools).

2) Metagenomic 16S rRNA sequencing amplicon taxonomic assignment (microbiome sample—NGS reads—quality control—taxonomic assignment—downstream analysis): As shown in Figure 4, NGS reads, as input, are trimmed and demultiplexed (cutadapt, dada2, demux plugins) with QIIME2 (https://qiime2.org/), and used to perform taxonomic assignment, diversity analysis (diversity analysis, feature-classifier, feature-table, taxa, composition plugins) and phylogenic analysis (phylogeny plugins) with QIIME2 and other downstream analysis (function prediction with PICRUSt and biomarker discovery with LEfSe).

3) Reference based metagenome taxonomic assignment (microbiome sample—NGS reads—quality control—taxonomic assignment—downstream analysis): Read-based taxonomic assignment uses the unassembled DNA or mRNA sequence reads directly and compares them against reference databases to assign taxonomy name to the sequence. NGS reads, as input, are trimmed into clean reads as depicted in workflow 1. Clean reads are then used to perform taxonomic (MetaPhlAn2) and functional assignment (HUMAnN2).

4) Genome assembly and annotation (isolated sample—NGS/TGS reads—quality control—assembly and validation—Genome structural analysis - database annotation): For NGS reads, the workflow is the same as workflow 1, except that there is no contamination removal step in quality control process and contig binning step in assembly process in workflow 4. For TGS reads, as input, are trimmed into clean reads and assembled into contigs and scaffolds with CANU. The draft genomes are then polished with Quiver. Structure analysis and annotation process are the same as in workflow 1.

5) RNA-seq analysis (isolated sample - NGS reads—quality control—alignment—assembly and

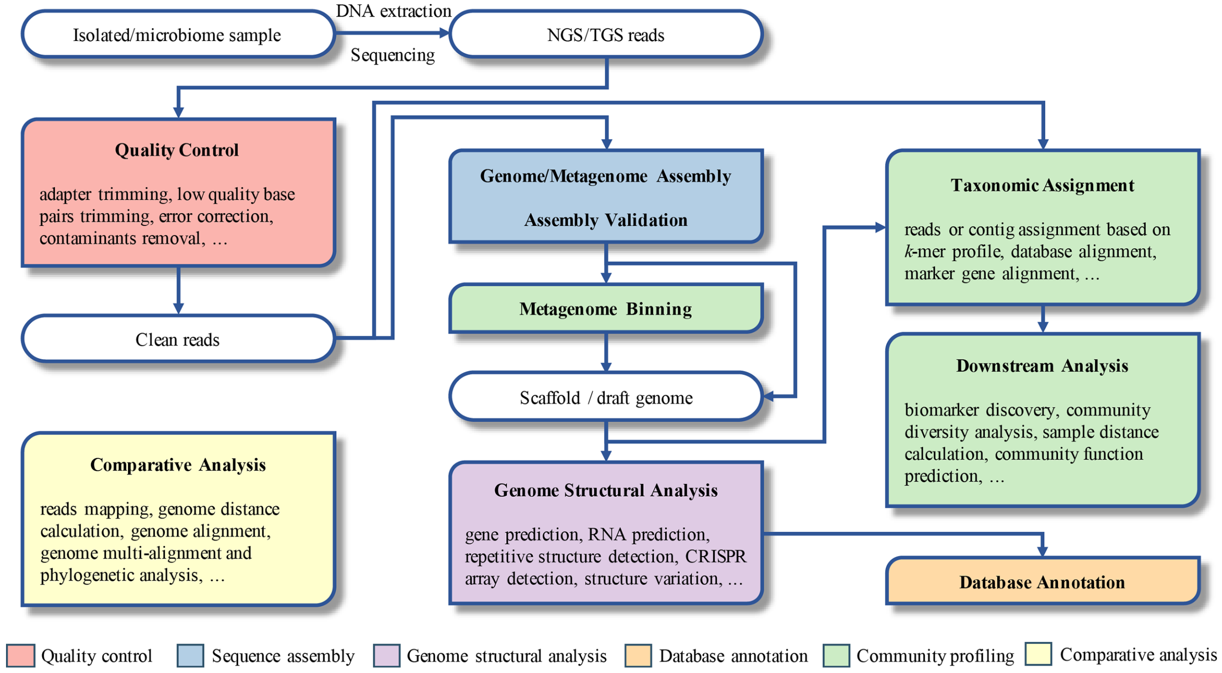
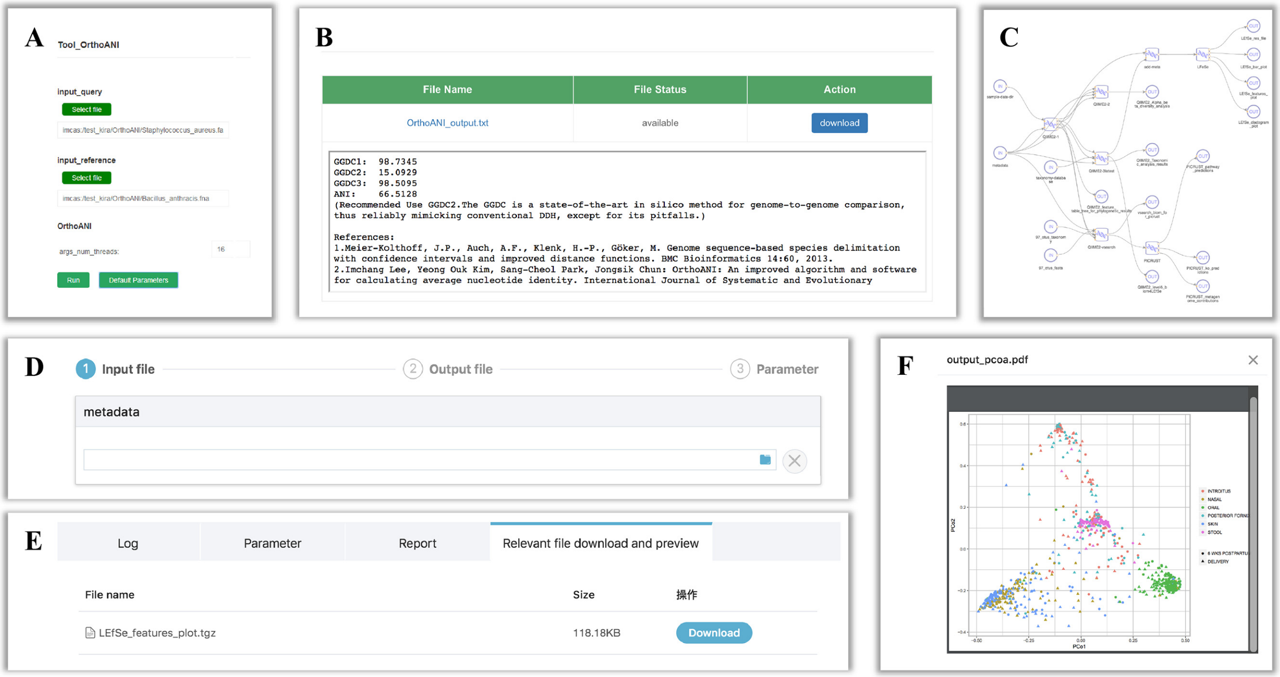

Isolated/microbiome sample
DNA extraction
Sequencing
NGS/TGS reads
Quality Control
adapter trimming, low quality base pairs trimming, error correction, contaminants removal, …
Clean reads
Genome/Metagenome Assembly
Assembly Validation
Metagenome Binning
Scaffold / draft genome
Genome Structural Analysis
gene prediction, RNA prediction, repetitive structure detection, CRISPR array detection, structure variation, …
Taxonomic Assignment
reads or contig assignment based on *k*-mer profile, database alignment, marker gene alignment, …
Downstream Analysis
biomarker discovery, community diversity analysis, sample distance calculation, community function prediction, …
Database Annotation
Comparative Analysis
reads mapping, genome distance calculation, genome alignment, genome multi-alignment and phylogenetic analysis, …
🟥 Quality control
🟦 Sequence assembly
🟪 Genome structural analysis
🟧 Database annotation
🟩 Community profiling
🟨 Comparative analysis

Figure 4. Integrated workflows on gcMeta. The tools can be grouped into 6 clusters shown in different colors (metagenome binning, taxonomic assignment and downstream analysis are all belong to the group community profiling shown in green color). Tools from different functional groups are connected in proper sequence to create workflows. Five main workflows covering different tools according to analysis aims are accessible from a unified user interface exemplified. Comparative analysis tools (shown in yellow) are widely involved in all the workflows. NGS and TGS stands for next-generation sequencing and third-generation sequencing, respectively.



Tool_OrthoANI

input_query: imcas:/test_kira/OrthoANI/Staphylococcus_aureus.fa

input_reference: imcas:/test_kira/OrthoANI/Bacillus_anthracis.fna

OrthoANI

args_num_threads: 16

File Name	File Status	Action
OrthoANI_output.txt	available	download

```
GGDC1:  98.7345
GGDC2:  15.0929
GGDC3:  98.5095
ANI:    66.5128
(Recommended Use GGDC2.The GGDC is a state-of-the-art in silico method for genome-to-genome comparison,
thus reliably mimicking conventional DDH, except for its pitfalls.)

References:
1.Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.-P., Göker, M. Genome sequence-based species delimitation
with confidence intervals and improved distance functions. BMC Bioinformatics 14:60, 2013.
2.Imchang Lee, Yeong Ouk Kim, Sang-Cheol Park, Jongsik Chun: OrthoANI: An improved algorithm and software
for calculating average nucleotide identity. International Journal of Systematic and Evolutionary
```

1 Input file — 2 Output file — 3 Parameter

metadata

Log | Parameter | Report | Relevant file download and preview

File name	Size	操作
LEfSe_features_plot.tgz	118.18KB	Download

Figure 5. Screenshots of the utility of the tool and workflow. (**A**, **B**) The ANI (average nucleotide identity) and dDDH (digital DNA-DNA hybridization) calculation tool which can be used by guest users. (**A**) Screenshots of the job submission including file upload module and necessary arguments setting. (**B**) The results of the job. (**C**–**F**) Metagenomic 16S rRNA sequencing taxonomic assignment workflow. (**C**) A screenshot of the sketch of the workflow. (**D**) The screenshots of the inputs, ouputs and arguments settings. (**E**) The result of the workflow. (**F**) The screenshots of the visualization of the analysis result. The example shows PCoA plot generated by ggplot package.

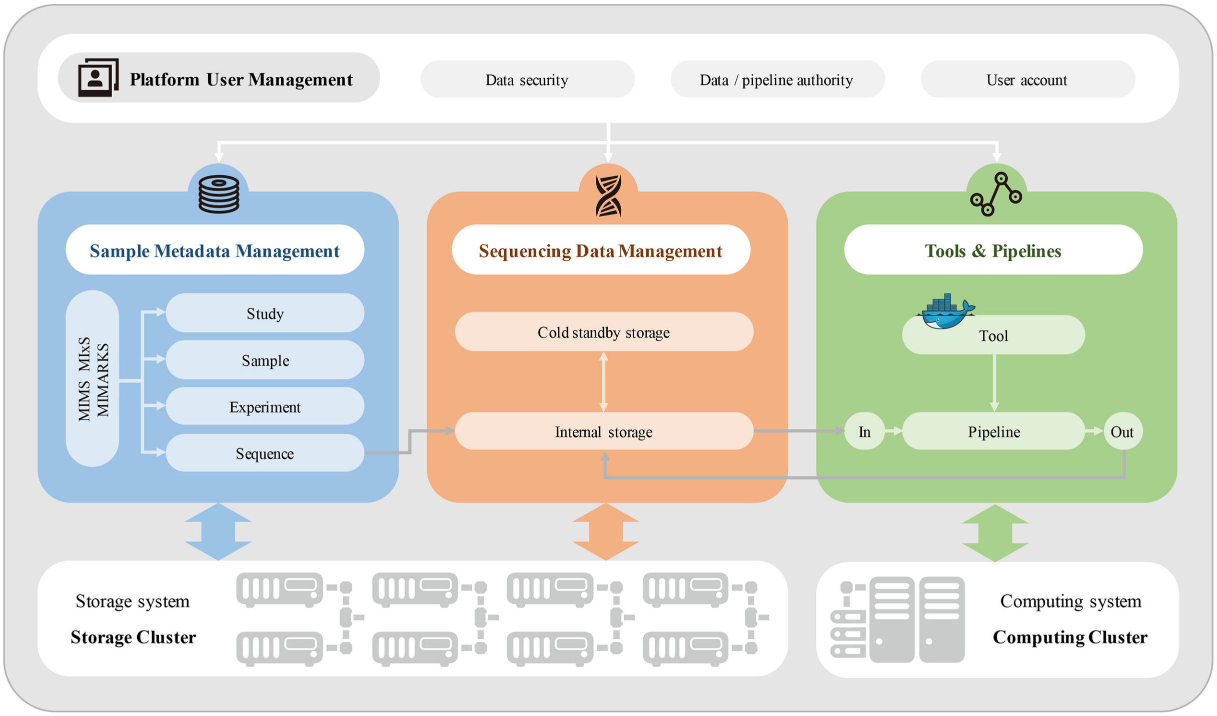


Figure 6. System structure of gcMeta. The platform integrates storage cluster and computing cluster resources with database management system and Docker based tools and workflows to supply comprehensive data archive, publication and analysis service to users.

differential expression analysis): This workflow allows users to identify differentially expressed genes and transcripts by comparing each sample with RNA-seq data. Firstly, NGS reads, as input, are trimmed into clean reads after quality control (quality viewing with FastQC and trimming with TrimGalore). Next, cleaned reads are aligned to the reference genomes with Hisat2. Then, the alignment result is used as an input to assemble transcripts. After assembly, differential expression analysis based on the assembled transcripts will be executed with DESeq2.

Implementation and utility of the tools and workflows

The currently available tools and workflows are developed for different server systems, and often difficult to install, configure and deploy by the microbiologists. Since software typically depends on libraries and other components of the installed environment, a workflow implemented in one environment may not work in another environment without extra configuration. To solve this problem, we use the Docker-based technologies. Docker is a Linux-based container technology that allows tools to run across a wide range of operating systems, and helps users to deploy and reproduce tools and workflows without undue efforts (107).

The tools and workflows provide a web-based interface for the users as indicated in Figure 5. The job can be submitted into the tasks queue after the users submit their files into the system and select a specific tool or workflow and set the essential parameters. If sufficient computing resources are not available, the job is put on waiting schedule. Job status is refreshed in the task page for user view. When the job is completed, results could be downloaded and visualized online.

Details of all the tools and workflows, including input format requirements, arguments setting and examples, are described in the Supplementary file and online manual https://gcmeta.wdcm.org/im/manual/index.jsp.

System design and implementation

The system as indicated in Figure 6 is based on centralized computing and storage resources. The database management system is divided into metadata management, sequencing raw data management and user information management. The current version of the gcMeta database is constructed on the basis of PostgreSQL for metadata and user information, and MongoDB for sequencing raw data index information. The system is operated on Linux servers. The web interface was developed using Python. Tools and workflows were installed with Docker technologies.

DISCUSSION

With the vast diversity of microbial communities and exponentially increasing amount of meta-omics data, we are facing great challenges in organized data management and deep data mining. As a part of the efforts of CAS-CMI, gcMeta provides long-term data preservation and management following the internationally used standards and hence serves as a public data repository. We provide data protection for pre-publication data and GUID for public data, which ensures the reuse of data as required by the scientific community. The platform houses and incorporates data from public and ongoing microbiome projects, and supports comparative analysis of the data collected from distinct projects.

Another key feature of the platform is we offer a set of data analysis and visualization tools and predefined workflows by web interface which facilitates data analysis by microbiologists in an easy way. As the system is based on the Docker technology, it can be run by a variety of operational systems. The analysis application is integrated with the in-house computing resources which provide scientists a robust site for powerful data analysis.

We will keep updating the database and workflows to support the rapidly increasing datasets and complicated studies. gcMeta accepts data submission for single genome, microbiome and transcriptome data. Meanwhile, we will establish connection with other data portals to ensure data properly deposited and preserved within the International Nucleotide Sequence Database Consortium (INSDC). On the other hand, the meta-omics data analysis is becoming more and more diverse. Predefined workflows could not fit for all the analysis aims. Therefore, the future version of gcMeta will provide customized workflows for professional bioinformatics who are interested in the data and computing resources while hope to develop their own analysis workflows.

Additionally, current meta-omics data and associated analysis results are widely dispersed in different kinds of resources from public data archives to specialized databases. Integration of various types of meta-omics data is essential for a comprehensive understanding of the structure, functions and expressions of a specific community, species, strain or gene. Updating our database schema to accommodate diverse data and providing rational links among those data through semantic web technologies are future planned directions as well.

Moreover, with increasing data from microbiome projects in China and worldwide, high quality reference data are needed for accurate data annotation. However, the current type strain genomic data resources in the public database are still unable to fullfill the requirements. The Global Catalogue of Microorganisms (GCM) 2.0 type strains (108) sequencing project and The Genomic Encyclopedia of Bacteria and Archaea (GEBA) project (109) are the ongoing efforts to fill in this gap. We plan to integrate the GCM 2.0 sequencing outputs into our system to provide more accurate annotation of metagenomics data. In conclusion, gcMeta will continuously improve to accommodate the evolving meta-omics researches.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank the former director of WDCM, Hideaki Sugawara for his thoughtful suggestions and continuous supports. We would like to thank the support from GeneDock Co., Ltd., Beijing, China for Docker container technologies. We would like to thank the technical support staff from BGI Shenzhen for their suggestions on data curation.

FUNDING

National key Research Program of China [2017YFD0400302, 2017YFC1201202, 2017YFC0907505, 2016YFC1200801, 2016YFB1000605, 2016YFC0901702]; Key Research Program of the Chinese Academy of Sciences [KFZD-SW-219]; Strategic Priority Research Program of the Chinese Academy of Sciences [XDA19050301]; 13th Five-year Informatization Plan of Chinese Academy of Sciences [XXH13506, XXH13505]; National Science Foundation for Young Scientists of China [31701157]; Supporting Program for the Developing Countries Around China [KY201701011]; Major State Basic Research Development Program [2015CB554200]. Funding for open access charge: National key Research Program of China [2016YFC0901702].

Conflict of interest statement. None declared.

REFERENCES

1. Thompson,L.R., Sanders,J.G., McDonald,D., Amir,A., Ladau,J., Locey,K.J., Prill,R.J., Tripathi,A., Gibbons,S.M., Ackermann,G. *et al.* (2017) A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, **551**, 457–463.
2. Lloyd-Price,J., Mahurkar,A., Rahnavard,G., Crabtree,J., Orvis,J., Hall,A.B., Brady,A., Creasy,H.H., McCracken,C., Giglio,M.G. *et al.* (2017) Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*, **550**, 61–66.
3. Kyrpides,N.C., Eloe-Fadrosh,E.A. and Ivanova,N.N. (2016) Microbiome Data Science: understanding our microbial planet. *Trends Microbiol.*, **24**, 425–427.
4. Hoopen,P.T. P., Finn,R.D., Bongo,L.A., Corre,E., Fosso,B., Meyer,F., Mitchell,A., Pelletier,E., Pesole,G., Santamaria,M. *et al.* (2017) The metagenomic data life-cycle: standards and best practices. *Gigascience*, **6**, 1–11.
5. Field,D., Amaral-Zettler,L., Cochrane,G., Cole,J.R., Dawyndt,P., Garrity,G.M., Gilbert,J., Glockner,F.O., Hirschman,L., Karsch-Mizrachi,I. *et al.* (2011) The Genomic Standards Consortium. *PLoS Biol.*, **9**, e1001088.
6. Ten Hoopen,P., Pesant,S., Kottmann,R., Kopf,A., Bicak,M., Claus,S., Deneudt,K., Borremans,C., Thijsse,P., Dekeyzer,S. *et al.* (2015) Marine microbial biodiversity, bioinformatics and biotechnology (M2B3) data reporting and service standards. *Stand. Genomic Sci.*, **10**, 20.
7. Dubilier,N., McFall-Ngai,M. and Zhao,L. (2015) Microbiology: create a global microbiome effort. *Nature*, **526**, 631–634.
8. Niu,S.Y., Yang,J., McDermaid,A., Zhao,J., Kang,Y. and Ma,Q. (2017) Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Brief Bioinform.*, **19**, 360.
9. Mitchell,A.L., Scheremetjew,M., Denise,H., Potter,S., Tarkowska,A., Qureshi,M., Salazar,G.A., Pesseat,S., Boland,M.A., Hunter,F.M.I. *et al.* (2018) EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res.*, **46**, D726–D735.
10. Wilke,A., Bischof,J., Gerlach,W., Glass,E., Harrison,T., Keegan,K.P., Paczian,T., Trimble,W.L., Bagchi,S., Grama,A. *et al.* (2016) The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res.*, **44**, D590–D594.
11. Chen,I.M.A., Markowitz,V.M., Chu,K., Palaniappan,K., Szeto,E., Pillay,M., Ratner,A., Huang,J.H., Andersen,E., Huntemann,M. *et al.* (2017) IMG/M: integrated genome and metagenome comparative data analysis system. *Nucleic Acids Res.*, **45**, D507–D516.
12. Huang,L.R., Kruger,J. and Sczyrba,A. (2018) Analyzing large scale genomic data on the cloud with Sparkhit. *Bioinformatics*, **34**, 1457–1465.
13. Wu,L., Sun,Q., Desmeth,P., Sugawara,H., Xu,Z., McCluskey,K., Smith,D., Alexander,V., Lima,N. and Ohkuma,M. (2016) World data centre for microorganisms: an information infrastructure to

explore and utilize preserved microbial strains worldwide. *Nucleic Acids Res.*, **45**, D611–D618.

14. 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.
15. Yilmaz,P., Kottmann,R., Field,D., Knight,R., Cole,J.R., Amaral-Zettler,L., Gilbert,J.A., Karsch-Mizrachi,I., Johnston,A., Cochrane,G. *et al.* (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.*, **29**, 415–420.
16. Buttigieg,P.L., Pafilis,E., Lewis,S.E., Schildhauer,M.P., Walls,R.L. and Mungall,C.J. (2016) The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *J. Biomed. Semant.*, **7**, 57.
17. Oulas,A., Pavloudi,C., Polymenakou,P., Pavlopoulos,G.A., Papanikolaou,N., Kotoulas,G., Arvanitidis,C. and Iliopoulos,I. (2015) Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinform. Biol. Insights*, **9**, 75–88.
18. Leinonen,R., Sugawara,H., Shumway,M. and Collaboration,I.N.S.D. (2010) The sequence read archive. *Nucleic Acids Res.*, **39**, D19–D21.
19. Huang,W., Li,L., Myers,J.R. and Marth,G.T. (2012) ART: a next-generation sequencing read simulator. *Bioinformatics*, **28**, 593–594.
20. Hu,X., Yuan,J., Shi,Y., Lu,J., Liu,B., Li,Z., Chen,Y., Mu,D., Zhang,H., Li,N. *et al.* (2012) pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics*, **28**, 1533–1535.
21. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.*, **17**, 10–12.
22. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
23. Morgulis,A., Gertz,E.M., Schaffer,A.A. and Agarwala,R. (2006) A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *J. Comput. Biol.*, **13**, 1028–1040.
24. Keegan,K.P., Trimble,W.L., Wilkening,J., Wilke,A., Harrison,T., D'Souza,M. and Meyer,F. (2012) A platform-independent method for detecting errors in metagenomic sequencing data: DRISEE. *PLoS Comput. Biol.*, **8**, e1002541.
25. 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.
26. Li,R., Zhu,H., Ruan,J., Qian,W., Fang,X., Shi,Z., Li,Y., Li,S., Shan,G. and Kristiansen,K. (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, **20**, 265–272.
27. Salmela,L. and Rivals,E. (2014) LoRDEC: accurate and efficient long read error correction. *Bioinformatics*, **30**, 3506–3514.
28. Hackl,T., Hedrich,R., Schultz,J. and Förster,F. (2014) proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics*, **30**, 3004–3011.
29. Chin,C.-S., Alexander,D.H., Marks,P., Klammer,A.A., Drake,J., Heiner,C., Clum,A., Copeland,A., Huddleston,J., Eichler,E.E. *et al.* (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods*, **10**, 563–569.
30. Magoč,T. and Salzberg,S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.
31. 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.
32. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
33. Kurtz,S., Phillippy,A., Delcher,A.L., Smoot,M., Shumway,M., Antonescu,C. and Salzberg,S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
34. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
35. Li,H. and Durbin,R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589–595.
36. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
37. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
38. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
39. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
40. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
41. Buchfink,B., Xie,C. and Huson,D.H. (2014) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
42. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
43. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
44. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
45. Chaisson,M.J. and Tesler,G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics*, **13**, 238.
46. 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.
47. 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.
48. Nurk,S., Meleshko,D., Korobeynikov,A. and Pevzner,P.A. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res.*, **27**, 824–834.
49. Namiki,T., Hachiya,T., Tanaka,H. and Sakakibara,Y. (2012) MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Res.*, **40**, e155.
50. Gnerre,S., Maccallum,I., Przybylski,D., Ribeiro,F.J., Burton,J.N., Walker,B.J., Sharpe,T., Hall,G., Shea,T.P., Sykes,S. *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. U.S.A.*, **108**, 1513–1518.
51. Peng,Y., Leung,H.C., Yiu,S.M. and Chin,F.Y. (2011) Meta-IDBA: a *de novo* assembler for metagenomic data. *Bioinformatics*, **27**, i94–i101.
52. Li,D., Liu,C.-M., Luo,R., Sadakane,K. and Lam,T.-W. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, **31**, 1674–1676.
53. Boisvert,S., Raymond,F., Godzaridis,E., Laviolette,F. and Corbeil,J. (2012) Ray Meta: scalable *de novo* metagenome assembly and profiling. *Genome Biol.*, **13**, R122.
54. Koren,S., 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.
55. Huang,X. and Madan,A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.*, **9**, 868–877.
56. Boetzer,M., Henkel,C.V., Jansen,H.J., Butler,D. and Pirovano,W. (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, **27**, 578–579.
57. Gao,S., Sung,W.K. and Nagarajan,N. (2011) Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *J. Comput. Biol.*, **18**, 1681–1691.
58. Gurevich,A., Saveliev,V., Vyahhi,N. and Tesler,G. (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.

59. Mikheenko,A., Saveliev,V. and Gurevich,A. (2015) MetaQUAST: evaluation of metagenome assemblies. *Bioinformatics*, **32**, 1088–1090.
60. Hunt,M., Kikuchi,T., Sanders,M., Newbold,C., Berriman,M. and Otto,T.D. (2013) REAPR: a universal tool for genome assembly evaluation. *Genome Biol.*, **14**, R47.
61. 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.
62. Simão,F.A., Waterhouse,R.M., Ioannidis,P., Kriventseva,E.V. and Zdobnov,E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–3212.
63. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
64. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
65. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.-C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
66. Patro,R., Mount,S.M. and Kingsford,C. (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.*, **32**, 462–464.
67. Bray,N.L., Pimentel,H., Melsted,P. and Pachter,L. (2016) Erratum: near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
68. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
69. Frazee,A.C., Pertea,G., Jaffe,A.E., Langmead,B., Salzberg,S.L. and Leek,J.T. (2015) Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat. Biotechnol.*, **33**, 243–246.
70. Grabherr,M.G., Haas,B.J., Yassour,M., Levin,J.Z., Thompson,D.A., Amit,I., Adiconis,X., Fan,L., Raychowdhury,R. and Zeng,Q. (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnol.*, **29**, 644–652.
71. Schulz,M.H., Zerbino,D.R., Vingron,M. and Birney,E. (2012) Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–1092.
72. Xie,Y., Wu,G., Tang,J., Luo,R., Patterson,J., Liu,S., Huang,W., He,G., Gu,S., Li,S. *et al.* (2014) SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, **30**, 1660–1666.
73. Edgar,R.C. (2007) PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, **8**, 18.
74. Lowe,T.M. and Chan,P.P. (2016) tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.*, **44**, W54–W57.
75. 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.
76. 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.
77. Delcher,A.L., Harmon,D., Kasif,S., White,O. and Salzberg,S.L. (1999) Improved microbial gene identification with GLIMMER. *Nucleic. Acids. Res.*, **27**, 4636–4641.
78. Borodovsky,M. and Mcininch,J. (1993) GENMARK: Parallel gene recognition for both DNA strands. *Comput. Chem.*,**17**, 123–133.
79. Rho,M., Tang,H. and Ye,Y. (2010) FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res.*, **38**, e191.
80. Newman,A.M. and Cooper,J.B. (2007) XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics*, **8**, 382.
81. Jiang,Y., Wang,Y. and Brudno,M. (2012) PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics*, **28**, 2576–2583.
82. Layer,R.M., Chiang,C., Quinlan,A.R. and Hall,I.M. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.
83. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
84. Tanizawa,Y., Fujisawa,T. and Nakamura,Y. (2018) DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. *Bioinformatics*, **34**, 1037–1039.
85. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
86. 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.
87. Caporaso,J.G., Kuczynski,J., Stombaugh,J., Bittinger,K., Bushman,F.D., Costello,E.K., Fierer,N., Pena,A.G., Goodrich,J.K., Gordon,J.I. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, **7**, 335–336.
88. Segata,N., Izard,J., Waldron,L., Gevers,D., Miropolsky,L., Garrett,W.S. and Huttenhower,C. (2011) Metagenomic biomarker discovery and explanation. *Genome Biol.*, **12**, R60.
89. Langille,M.G.I., Zaneveld,J., Caporaso,J.G., McDonald,D., Knights,D., Reyes,J.A., Clemente,J.C., Burkepile,D.E., Vega Thurber,R.L.V., Knight,R. *et al.* (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.*, **31**, 814–821.
90. Liu,J.M., Wang,H.F., Yang,H.X., Zhang,Y.Z., Wang,J.F., Zhao,F.Q. and Qi,J. (2012) Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. *Nucleic Acids Res.*, **41**, e3.
91. Ainsworth,D., Sternberg,M.J.E., Raczy,C. and Butcher,S.A. (2017) k-SLAM: accurate and ultra-fast taxonomic classification and gene identification for large metagenomic data sets. *Nucleic Acids Res.*, **45**, 1649–1656.
92. Menzel,P., Ng,K.L. and Krogh,A. (2016) Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.*, **7**, 11257.
93. Kim,D., Song,L., Breitwieser,F.P. and Salzberg,S.L. (2016) Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res.*, **26**, 1721–1729.
94. Piro,V.C., Lindner,M.S. and Renard,B.Y. (2016) DUDes: a top-down taxonomic profiler for metagenomics. *Bioinformatics*, **32**, 2272–2280.
95. Sunagawa,S., Mende,D.R., Zeller,G., Izquierdo-Carrasco,F., Berger,S.A., Kultima,J.R., Coelho,L.P., Arumugam,M., Tap,J., Nielsen,H.B. *et al.* (2013) Metagenomic species profiling using universal phylogenetic marker genes. *Nat. Methods*, **10**, 1196–1199.
96. Albanese,D. and Donati,C. (2017) Strain profiling and epidemiology of bacterial species from metagenomic sequencing. *Nat. Commun.*, **8**, 2260.
97. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S., Phillippy,A.M., Altschul,S., Gish,W., Miller,W. *et al.* (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
98. Brown,C.T. and Irber,L. (2016) sourmash: a library for MinHash sketching of DNA. *J. Open Source Softw.*, **1**, 27.
99. Segata,N., Waldron,L., Ballarini,A., Narasimhan,V., Jousson,O. and Huttenhower,C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods*, **9**, 811–814.
100. Abubucker,S., Segata,N., Goll,J., Schubert,A.M., Izard,J., Cantarel,B.L., Rodriguez-Mueller,B., Zucker,J., Thiagarajan,M., Henrissat,B. *et al.* (2012) Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput. Biol.*, **8**, e1002358.
101. Alneberg,J., Bjarnason,B.S., de Bruijn,I., Schirmer,M., Quick,J., Ijaz,U.Z., Lahti,L., Loman,N.J., Andersson,A.F. and Quince,C. (2014) Binning metagenomic contigs by coverage and composition. *Nat. Methods*, **11**, 1144–1146.

102. Wu,Y.W., Simmons,B.A. and Singer,S.W. (2015) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, **32**, 605–607.
103. Kang,D.D., Froula,J., Egan,R. and Wang,Z. (2015) MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, **3**, e1165.
104. Wu,Y.W. and Ye,Y. (2011) A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using *l*-tuples. *J. Comput. Biol.*, **18**, 523–534.
105. Ren,J., Ahlgren,N.A., Lu,Y.Y., Fuhrman,J.A. and Sun,F. (2017) VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. *Microbiome*, **5**, 69.
106. Ahlgren,N.A., Ren,J., Young,L.Y., Fuhrman,J.A. and Sun,F. (2017) Alignment-free d_2^* oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.*, **45**, 39–53.
107. Boettiger,C. (2015) An introduction to Docker for reproducible research. *ACM SIGOPS Oper. Syst. Rev.*, **49**, 71–79.
108. 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**, giy026.
109. 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. and Paez-Espino,D. (2017) 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat. Biotechnol.*, **35**, 676–683.