Published online 28 September 2022

Nucleic Acids Research, 2023, Vol. 51, Database issue ***D767–D776***
https://doi.org/10.1093/nar/gkac832

ProPan: a comprehensive database for profiling prokaryotic pan-genome dynamics

Yadong Zhang[1,2,†], Hao Zhang[1,2,3,†], Zaichao Zhang[4], Qiheng Qian[1,2,3], Zhewen Zhang[1,2,*] and Jingfa Xiao[1,2,3,*]

[1]National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China, [2]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, Beijing 100101, China, [3]University of Chinese Academy of Sciences, Beijing 100049, China and [4]Department of Biology, The University of Western Ontario, London, Ontario N6A 5B7, Canada

Received August 14, 2022; Revised September 09, 2022; Editorial Decision September 14, 2022; Accepted September 16, 2022

ABSTRACT

Compared with conventional comparative genomics, the recent studies in pan-genomics have provided further insights into species genomic dynamics, taxonomy and identification, pathogenicity and environmental adaptation. To better understand genome characteristics of species of interest and to fully excavate key metabolic and resistant genes and their conservations and variations, here we present ProPan (https://ngdc.cncb.ac.cn/propan), a public database covering 23 archaeal species and 1,481 bacterial species (in a total of 51,882 strains) for comprehensively profiling prokaryotic pan-genome dynamics. By analyzing and integrating these massive datasets, ProPan offers three major aspects for the pan-genome dynamics of the species of interest: 1) the evaluations of various species' characteristics and composition in pan-genome dynamics; 2) the visualization of map association, the functional annotation and presence/absence variation for all contained species' gene clusters; 3) the typical characteristics of the environmental adaptation, including resistance genes prediction of 126 substances (biocide, antimicrobial drug and metal) and evaluation of 31 metabolic cycle processes. Besides, ProPan develops a very user-friendly interface, flexible retrieval and multi-level real-time statistical visualization. Taken together, ProPan will serve as a weighty resource for the studies of prokaryotic pan-genome dynamics, taxonomy and identification as well as environmental adaptation.

INTRODUCTION

The striking advancement of sequencing technologies is leading to an explosive accumulation of high-throughput genomic datasets, especially in outputting numerous prokaryotic genomes. Conventionally, comparative genomics in prokaryotes analyzes one single species genome in contrast to the representative genome of that population (1). Nowadays, however, with the erupting prokaryotic genome datasets, it gets difficult to reveal further characteristics for the species of interest. The first conceiving concept of the pan-genome provides a new framework for researchers (2). The pan-genome represents the entire set of genes within a certain species and can be divided into core genes, dispensable genes and unique genes. Since then, researchers have used pan-genomics method to study many prokaryotic genomes. For instance, pan-genomics offers more insights into genomic diversity, gene conservation and nucleotide diversity (3–5). The evidence from pan-genome (*i.e.* core genes, gene gain and loss) also provides extra approaches for species taxonomy (6–8). The dispensable genes are often linked to pathogenic features and the horizontal transfer of key genes changes species' phenotypic variance and consequently, enhances their infectivity and adaptability to host (9–11). In all, pan-genome studies have efficiently excavated novel biological insights in dynamics, taxonomy, pathogenic and many other research fields (12–14).

To exhibit prokaryotic genome analyses and important functional genes, several related databases have been established. A majority of them, such as MBGD (15), MetaRef (16), MicrobesOnline (17) and proGenomes2 (18), were constructed by the orthologs of species and integrated with comparative genomic analyses as well as functional annotations. Some others were designed with a focus on predicting orthologs and providing support for compu-

[*]To whom correspondence should be addressed. Tel: +86 10 8409 7443; Fax: +86 10 8409 7720; Email: xiaojingfa@big.ac.cn
Correspondence may also be addressed to Zhewen Zhang. Email: zhangzw@big.ac.cn
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

tational analyses of evolutionary traits, for example, OrtholugeDB (19) and OrthoDB (20). Alternatives like MicroScope (21) and IMG/M (22), were equipped with comparative analyses and annotations for genome and metagenome datasets. Nevertheless, these databases mentioned above only emphasized on gene orthologs rather than representing the pan-genome dynamics of the species. Pan-genome dynamics mainly focuses on studying the drivers of genome changes, gene gain and loss, and pan-genome openness characteristics in a specified species. Besides, some of them just included fractional prokaryotic genomic datasets (*e.g.* MetaRef and MicrobesOnline) and provided functions barely at the single genome level. Plus, they overlooked that functional annotation for a single genome can neither completely excavate the dynamics of key functional genes, nor inclusively reflect species' environmental adaptation.

To address these issues, here we present ProPan version 1.0, a public database for comprehensively profiling the characteristics in prokaryotic pan-genome dynamics (Supplementary Table S1). ProPan stores 23 archaeal and 1,481 bacterial species, a total of 51,882 strains genome datasets. Through rigorous data quality control and large-scale data profiling, ProPan provides profiled characteristics in pan-genome dynamics and compositions across 1,504 species. Additionally, ProPan integrates visualization of map association, functional annotations and presence/absence variation (PAV) of all gene clusters for these species. Besides, it includes the resistance genes prediction of 126 substances (biocide, antimicrobial drug and metal) and evaluation of 31 metabolic cycle processes (*e.g.* organic carbon oxidation). Furthermore, ProPan has a very user-friendly web interface with four main modules: browse, search, statistics and downloads. To sum up, ProPan covers extensive characteristics in pan-genome dynamics of diverse prokaryotic species to offer further insights into metabolism and resistance studies.

DATA COLLECTION AND PROCESSING

Prokaryotic whole genome sequence collection and quality control

All genomic datasets in ProPan database were retrieved from NCBI (23), including genome sequence, nucleotide sequence, amino acid sequence, assembly statistic report information, *etc*. Given that many factors, such as genome sequencing methods and assembly quality, incorrect classification, chimeric strains and engineered genome-reduced, may have an impact on pan-genome analyses (24), we initially performed data quality control for the retrieved prokaryotic genome datasets by in-house scripts, using the genome assembly statistic report to filter out low-quality and incomplete strains. The filter criteria are: 1) keeping full strain genome representation instead of partial; 2) excluding strains with assembly anomaly such as chimeric, contaminated, misassembled; 3) filtering out strains with abnormal genome length, low-quality sequences, untrustworthy as types, unverified source organisms, many frameshifted proteins, abnormal gene to sequence ratio, *etc*. All these filtered terms were guided by NCBI web page (https://www.ncbi.nlm.nih.gov/assembly/help/anomnotrefseq/). We also filtered out strains with abnormal protein sequence number (SN): $\mu - 2\sigma \leq SN \leq \mu + 2\sigma$ (μ is the average number of species protein sequences and σ is the standard deviations) (25). Mash v2.3 was used to calculate the mutational distance of strain within the same species (26). To select strains for each species, FastANI v1.32 and MCL v14-137 were used to analyze strain average nucleotide identity (ANI) and to perform clustering, respectively (27,28). Strains with ANI greater than or equal to 0.95 were retained. Species with equal to or more than five strains were selected as the datasets to be analyzed. Finally, a total of 51,882 strains related to 23 archaeal species and 1,481 bacterial species were retained in ProPan.

Species pan-genome profiling and multi-dimensional datasets analyses

After data quality control, a series of tools and methods were employed to primarily analyze the datasets. In brief, Prokka v1.14.5 was used to annotate genomes that lacked annotation information (29). Roary v3.13 was used for pan-genome ortholog clustering analyses (30). The R package micropan v2.1 was used to estimate whether the species had an open or closed pan-genome using Heaps' law ($n = k\ N^{-\alpha}$) with a permutation value of 1000 and repeat 100 (31,32). The average value of exponent α was applied to determine whether the pan-genome is open ($\alpha \leq 1$) or closed ($\alpha > 1$). Subsequently, based on gene clustering results, VariScan v2.0.3 was used to calculate the nucleotide diversity of core gene clusters and variable gene clusters (33). Then, eggNOG-mapper v2.1.6 and eggNOG database v5.0 were used for gene cluster annotation (34,35). The METABOLIC-G module in METABOLIC v4.0 software was employed to dissect the metabolic cycle characteristics of species (36). In addition, we analyzed the resistance characteristics of species by five databases, NCBI AMRFinderPlus (37), CARD (38), Resfinder (39), ARG-ANNOT (40) and MEGARes (41), to constitute the resistance seed dataset and finally, we applied BLAST+ v2.12.0 for the alignment retrieval of target sequences (42). In terms of visualization, the R package ComplexHeatmap v2.6.2 was used to visualize the PAV of species resistance (43). The protein-protein interaction (PPI) network of homologous proteins of gene clusters was dynamically visualized using the STRING database (44). The R script in METABOLIC software was used for mapping metabolic pathway networks (36). An overview of the datasets processing workflow is shown in Figure 1.

DATABASE IMPLEMENTATION

ProPan was constructed by universal popular database development techniques. AJAX, Bootstrap, CSS, HTML5, JQuery and Thymeleaf (a Java template engine) were used for data rendering and interactive operations of front-end pages. Spring Boot was used as the basic main architecture of the back-end system. MySQL served as a container for data storage and management. The framework Mybatis was employed for quick and robust access to datasets. Echarts.js and plotly.js were adopted for building interactive statistical graphs. Bootstrap Table was employed to dynamically tab-

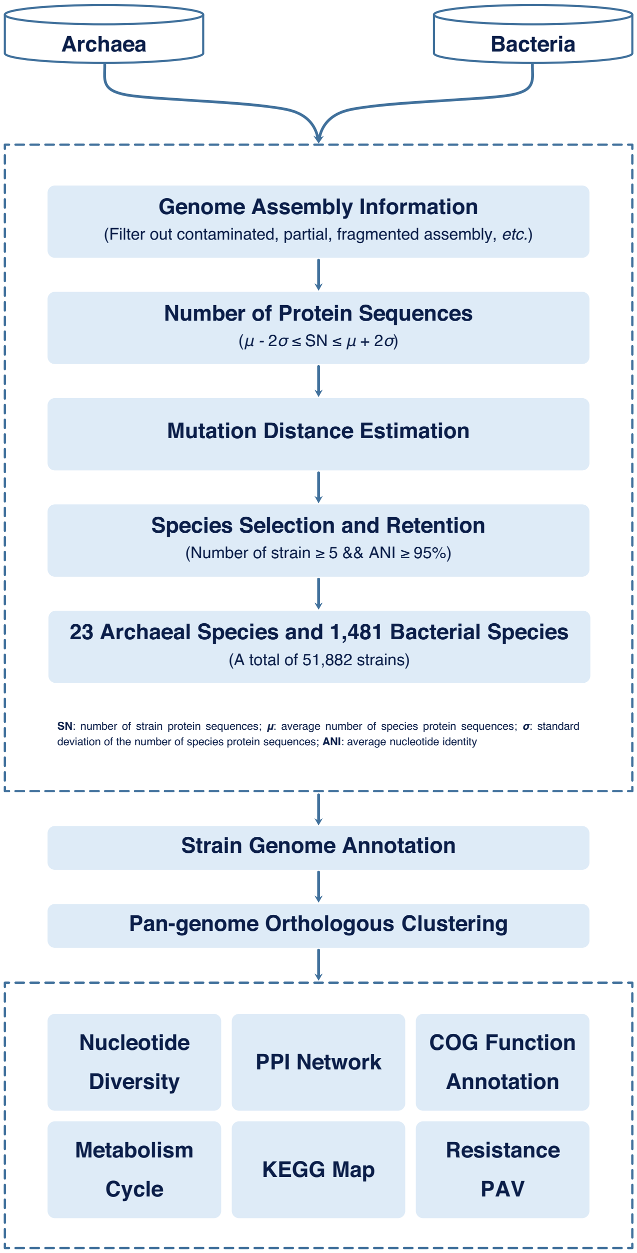

Archaea
Bacteria
Genome Assembly Information
(Filter out contaminated, partial, fragmented assembly, *etc.*)
Number of Protein Sequences
(μ - 2$\sigma \leq$ SN $\leq \mu$ + 2σ)
Mutation Distance Estimation
Species Selection and Retention
(Number of strain ≥ 5 && ANI ≥ 95%)
23 Archaeal Species and 1,481 Bacterial Species
(A total of 51,882 strains)
SN: number of strain protein sequences; ***μ***: average number of species protein sequences; ***σ***: standard deviation of the number of species protein sequences; **ANI**: average nucleotide identity
Strain Genome Annotation
Pan-genome Orthologous Clustering
Nucleotide Diversity
PPI Network
COG Function Annotation
Metabolism Cycle
KEGG Map
Resistance PAV

Figure 1. An overview of the datasets processing workflow in ProPan. All genomic datasets used to construct the ProPan database were retrieved from NCBI and processed uniformly. Multi-dimensional data analyses were performed: nucleotide diversity, PPI network, resistance PAV, *etc*.

ulate datasets. The virtual environment created by Docker acted as a container to deploy the database.

DATABASE CONTENTS AND FEATURES

Currently, ProPan version 1.0 integrates 51,882 strain genome datasets related to 1,504 prokaryotic species. A variety of multi-dimensional analyses were carried out from the population perspective based on the gene clusters, including nucleotide diversity assessment of gene clusters, COG functional enrichment analysis, 31 metabolic cycle processes analyses and map construction, resistance genes prediction and PAV analyses of 126 substances (biocide, antimicrobial drug and metal), KEGG pathway association, PPI network map construction and others. The multi-dimensional analysis results show users extensive functional annotations of species of interest. ProPan is also equipped with major modules (browse, search, statistics and downloads) for the accessibility, visualization and downloads of analytical results.

Profiling species' characteristics by pan-genome analyses

We constructed a gene family matrix with uniform processes for 1,504 species with the filtered criteria. According to the different pan-genome conservations, all gene clusters were divided into three categories: 1) core gene clusters that consist of homologs across all the analyzed strains of one species; 2) unique gene clusters that only contain a single analyzed strain; 3) dispensable gene clusters other than the core and unique genes clusters. Based on the three categories, we calculated the pan-genome composition for each species. The statistical results indicate that there is extensive variation in the core genome proportion of species from different phyla (Figure 2A). This variation may be influenced by number differences of the analyzed strains within the species. An open pan-genome indicates that species have a high capacity to exchange genetic material while a closed pan-genome indicates a limited capacity to acquire foreign genes (9). We estimated the pan-genome openness level (open or closed) for the 1,504 species using Heaps' law (Figure 2B and C). It shows that, in a total of 23 archaeal species, 13 species have open pan-genome ($\alpha \leq 1$) (*e.g. Methanosarcina mazei*) and 10 species have closed pan-genome ($\alpha > 1$) (*e.g. Metallosphaera sedula*), respectively. However, among the 1,481 bacterial species, 1,205 species exhibited open pan-genome (*e.g. Escherichia coli*) and only 276 species exhibited closed pan-genome (*e.g. Brucella canis*). The pan-genome openness level of all the 1,504 species is listed in Supplementary Table S2. Regardless of archaea or bacteria, on average, we observe that species with open pan-genome have lower core genome proportion, whereas species with closed pan-genome have higher core genome proportion (Figure 2D).

Pan-genome metabolism exploration and map construction

Carbon, nitrogen and sulfur are very important elements for life. Prokaryotes play vital roles in carbon, nitrogen and sulfur cycles (45,46). By combining species' gene pool, we analyzed the 31 metabolic cycle processes and divided them into four major categories: carbon cycle, nitrogen cycle, sulfur cycle and other cycle (see Supplementary Table S3 for category information of these metabolic cycles). In our results, although both archaea and bacteria hold most of the metabolic cycle processes (*e.g.* organic carbon oxidation, nitrogen fixation, sulfide oxidation and arsenate reduction), there are still some different preferences between archaea and bacteria (Figure 3). In the carbon cycle, 7 out of 23 archaeal species have methanogenesis, which is absent in 1,481 bacterial species. The ethanol oxidation gene clusters are dispensable in 23 archaeal species, while in bacteria, the gene clusters are mainly represented as cores (211 out of

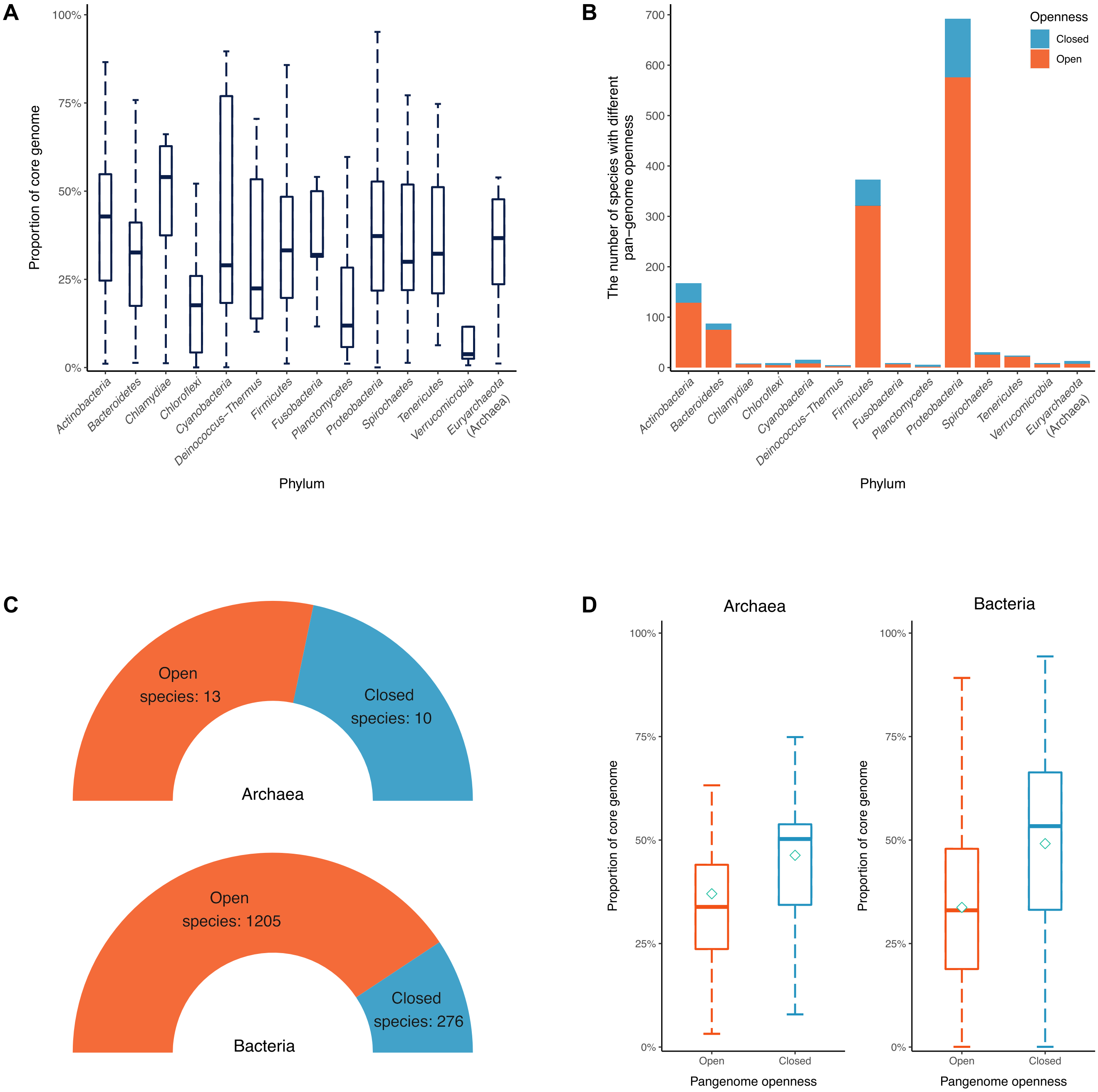


Figure 2. Statistics of pan-genome openness level and core genome proportion distribution. (**A**) The proportion of core genomes in different phyla (species' counts over 10 in ProPan). (**B**) Pan-genome openness level statistics in different phyla (species' counts over 10 in ProPan). (**C**) The proportions of species with different pan-genome openness level in archaea and bacteria, respectively. (**D**) The distribution relationship between pan-genome openness level and core genome proportions in archaeal and bacterial species. Orange means open and blue means closed. The diamonds represent the average for each category.

1,481 species). In the nitrogen cycle, we find that 773 out of 1,481 bacterial species have nitrite ammonification process while none in 23 archaeal species. In addition, the gene clusters of nitrite oxidation and nitrate reduction show low pan-genome conservation in archaea. In the sulfur cycle, 802 out of 1,481 bacterial species have at least one sulfur cycle process, while in all 23 archaeal species, there is no sulfurite reduction, thiosulfate oxidation, thiosulfate disproportionation 1 and thiosulfate disproportionation 2. As for metal reduction, arsenate reduction, arsenite oxidation and selenate reduction, only arsenate reduction is observed in 11 out of 23 archaeal species, but all of them are observed in 1,225 out of 1,481 bacterial species. In addition, there are also metabolic differences between different phyla. In archaea, *Euryarchaeota* has a variety of metabolic processes (Supplementary Figure S1A), but compared with other phyla in archaea (Supplementary Figure S2), the metabolism of sulfur is relatively deficient. In bacteria, *Proteobacteria* covers

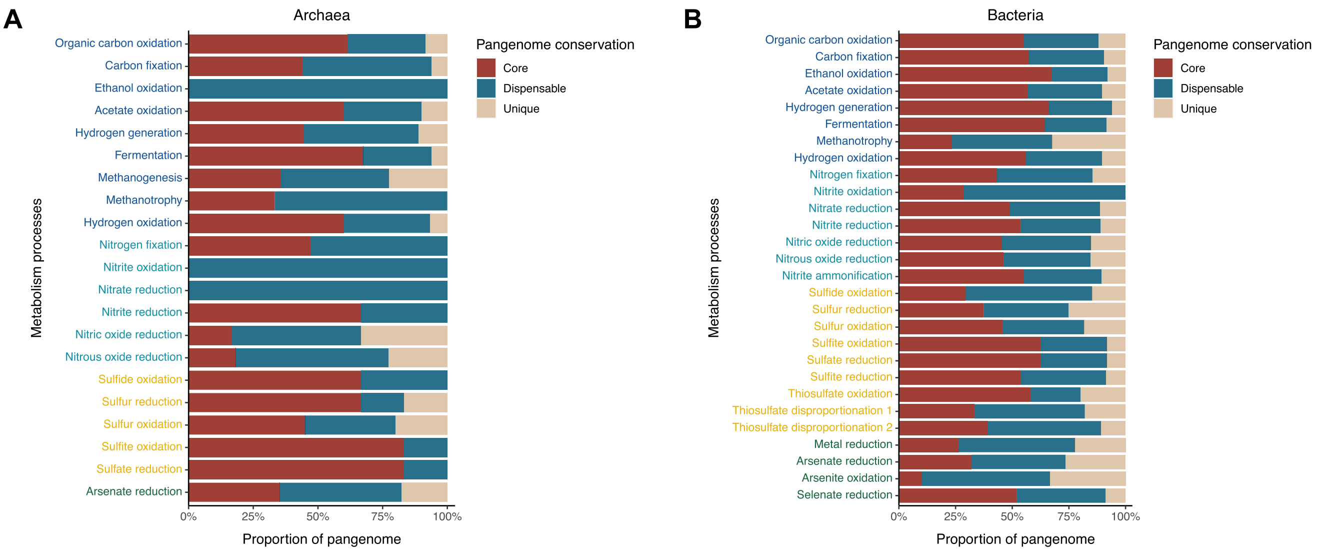


Figure 3. Pan-genome conservation statistics of metabolic cycle process genes for archaea (A) and bacteria (B). The y-axis represents different metabolic cycle processes. The dark blue represents carbon metabolism, cyan represents nitrogen metabolism, yellow represents sulfur metabolism and the green represents other metabolisms. The x-axis represents pan-genome proportion.

almost all metabolic cycle processes (Supplementary Figure S1B), while other phyla exhibit various metabolic differences (Supplementary Figure S2). In sum, for cycling of these important elements, prokaryotic species show different preferences and variations of metabolic genes in the pan-genome conservation.

Prediction of pan-genome resistance genes

The PAV of resistance genes and their conservation among species play a key role in the survival of species in diverse environments (47,48). Therefore, we predicted multiple resistance genes (15 biocides, 97 antimicrobial drugs and 14 metals) in 1,504 species (Supplementary Table S3 shows categorical information for these resistance genes). Combined with pan-genome analyses, we also further explored the conserved distribution of resistance genes in species. The results show that all 23 archaeal species in ProPan have fewer resistance genes (mainly from phylum *Euryarchaeota* and *Crenarchaeota*) and surprisingly, no biocide resistance genes were observed (Figure 4). In addition, copper resistance genes are highly conserved and shared by seven species' core gene clusters in the pan-genome conservation (*e.g. Methanosarcina mazei*, *Sulfolobus islandicus*, *etc.*). These strains are mainly isolated from environments with high mineralization, such as hot springs, salt lakes and river sediment. The 1,481 bacterial species have a wide variety of biocide, antimicrobial drug and metal resistance genes, but several resistance genes show high conservation (in phyla of *Actinobacteria*, *Bacteroidetes*, *Firmicutes* and *Proteobacteria*). For instance, in biocide, peroxide and phenolic compound resistance genes are classified as core gene clusters in 572 and 594 species (*e.g.* peroxide: *Burkholderia stagnalis*; phenolic compound: *Listeria seeligeri*), respectively. Most of these strains come from the soil and water environments. In antimicrobial drug, core fluoroquinolone resistance genes and core tetracycline resistance genes were found in 665 (*e.g. Burkholderia multivorans*) and 635 species (*e.g. Pseudomonas brassicacearum*), respectively. Strains containing core antimicrobial drug resistance genes were isolated from various environments, such as soil, human sputum and blood. In metal, core arsenic and copper resistance genes were found in 617 and 767 species (*e.g.* arsenic: *Leclercia adecarboxylata*; copper: *Kosakonia radicincitans*), respectively. The proportion of core resistance gene clusters in different phyla are shown in Supplementary Figure S3. The analyzed strains were isolated from a variety of environments, such as lake water and human feces. The diversity of the living environment of these species and the high conservation of different resistance genes reflect the species' strong adaptability to the various living environments.

Four principal modules: browse, search, statistics and downloads

ProPan has a very user-friendly web interface and consists of four main modules.

Browse. An overview table of all collected prokaryotic species with species' name, taxonomy ID, processed strain counts, pan-genome composition, metabolism and resistance information. Users can consult different items such as pan-genome openness level, resistance and metabolism type. Respectively, the *Species* and *Analysis* columns can further direct users to the detailed information of the species in NCBI and the analyzed results.

Search. A series of searching approaches were equipped in ProPan: 1) a quick search box on the home page was designed for real-time queries by specifying species names, resistance or metabolism cycle types; 2) three advanced search modules on the search page were also integrated: in terms

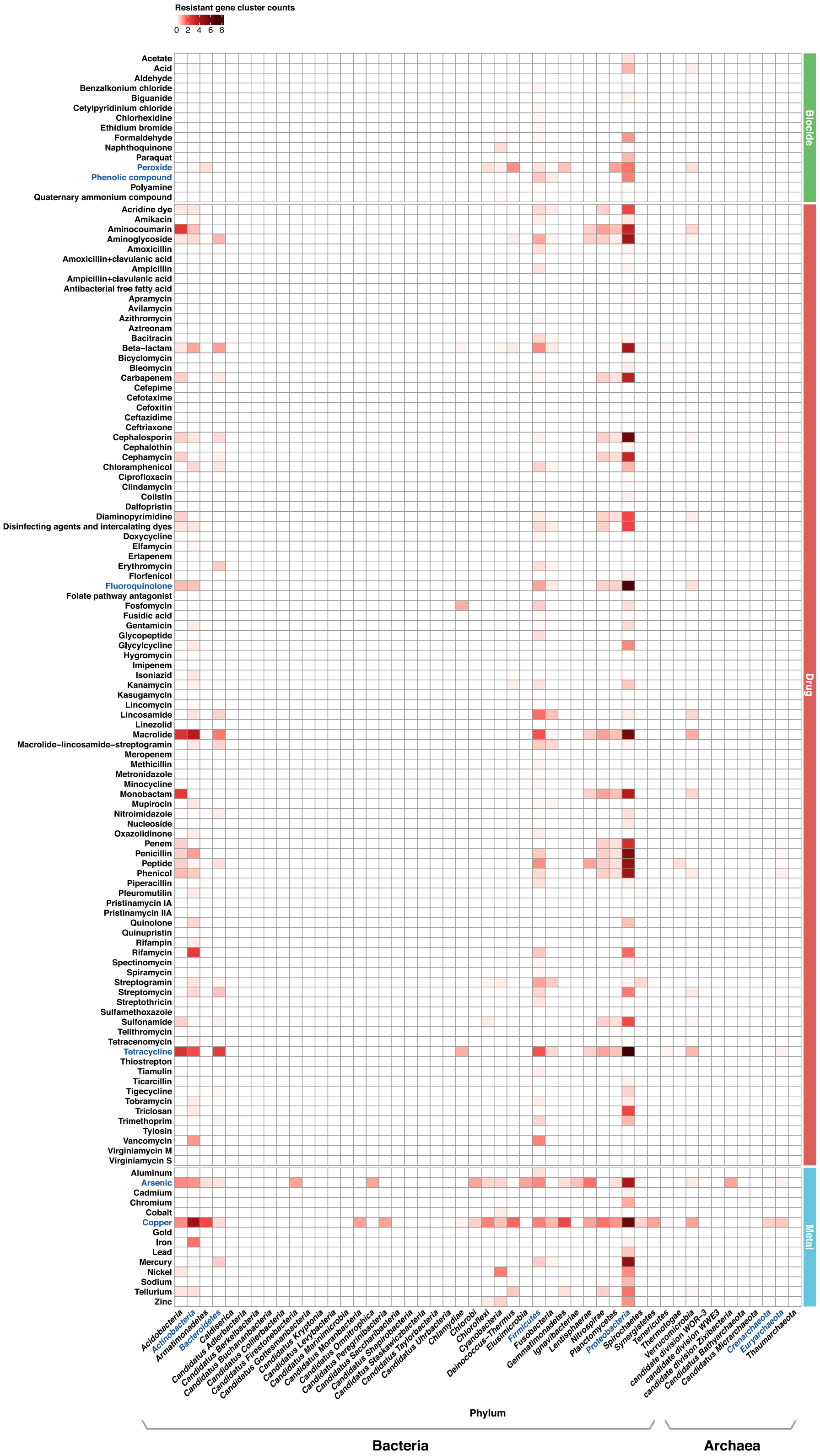
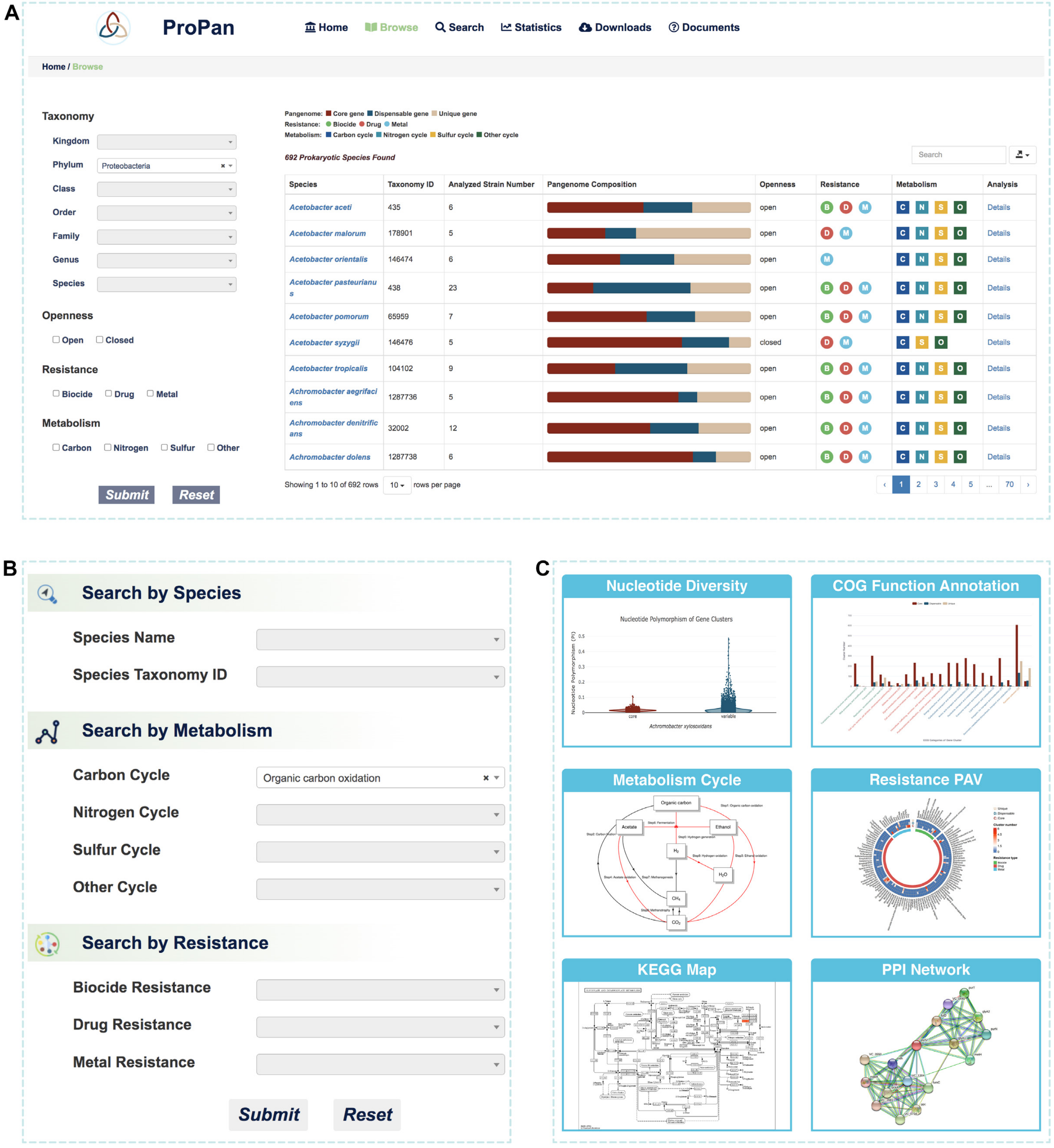


Figure 4. Distribution of resistance gene clusters (biocide, antimicrobial drug and metal) in different phyla. The y-axis represents different resistant substances. The x-axis represents different phyla in ProPan. The genes and phyla mentioned in the main text are marked in blue.



ProPan

Home Browse Search Statistics Downloads Documents

Home / Browse

Taxonomy

Kingdom
Phylum Proteobacteria
Class
Order
Family
Genus
Species

Openness

□ Open □ Closed

Resistance

□ Biocide □ Drug □ Metal

Metabolism

□ Carbon □ Nitrogen □ Sulfur □ Other

Submit Reset

Pangenome: 🟥 Core gene 🟦 Dispensable gene 🟫 Unique gene
Resistance: ● Biocide ● Drug ● Metal
Metabolism: 🟦 Carbon cycle 🟦 Nitrogen cycle 🟨 Sulfur cycle 🟩 Other cycle

692 Prokaryotic Species Found

Search

Species	Taxonomy ID	Analyzed Strain Number	Pangenome Composition	Openness	Resistance	Metabolism	Analysis
Acetobacter aceti	435	6		open	B D M	C N S O	Details
Acetobacter malorum	178901	5		open	D M	C N S O	Details
Acetobacter orientalis	146474	6		open	M	C N S O	Details
Acetobacter pasteurianus	438	23		open	B D M	C N S O	Details
Acetobacter pomorum	65959	7		open	B D M	C N S O	Details
Acetobacter syzygii	146476	5		closed	D M	C S O	Details
Acetobacter tropicalis	104102	9		open	B D M	C N S O	Details
Achromobacter aegrifaciens	1287736	5		open	B D M	C N S O	Details
Achromobacter denitrificans	32002	12		open	B D M	C N S O	Details
Achromobacter dolens	1287738	6		open	B D M	C N S O	Details

Showing 1 to 10 of 692 rows 10 rows per page

‹ 1 2 3 4 5 ... 70 ›

Search by Species

Species Name
Species Taxonomy ID

Search by Metabolism

Carbon Cycle: Organic carbon oxidation
Nitrogen Cycle
Sulfur Cycle
Other Cycle

Search by Resistance

Biocide Resistance
Drug Resistance
Metal Resistance

Submit Reset

Figure 5. Screenshots of ProPan version 1.0. (**A**) Browse page: including filtered items on the left and an overview table on the right. (**B**) Search page: three advanced search modules (species, metabolism and resistance). (**C**) Analysis results: nucleotide polymorphism of gene clusters, COG functional annotation histogram, carbon metabolism cycle map (a red arrow means the presence of a pathway process and a black arrow means absence), heatmap of PAV in resistance genes, KEGG map of a gene cluster (red represents the target gene) and PPI networks (red indicates the selected gene).

of species, users can view the analyzing results by specifying a name or taxonomy ID; in terms of metabolism, users can search by different metabolic cycles (carbon cycle, nitrogen cycle sulfur cycle and other cycle) to obtain the species' information and pan-genome conservation of the retrieved metabolic cycle processes; in terms of resistance, users can specify the name of biocide, antimicrobial drug or metal to access to the results of interest.

Statistics. Four aspects of all species statistics were provided for visualization: 1) the proportion and composition of the pan-genome; 2) the distribution and counts for resistance gene clusters; 3) the distribution and counts for metabolism cycle gene clusters; 4) the proportion of gene clusters in metabolism and resistance across *archaea* and *bacteria* kingdoms.

Downloads. Gene cluster annotation matrixes for each species are available for download, including pan-genome conservation, COG functional annotation, enzyme category, metabolism, resistance and others. In addition, the nucleotide and protein sequences of each gene cluster can also be downloaded.

We show a quick glimpse of how ProPan works in Figure 5. In brief, the detailed information of analyses can be accessed by both the browse page and search page (Figure 5A and Figure 5B). For analyzed strain information, all the related items are listed in the table, including GenBank number, CDS counts, GC contents and others. For the nucleotide diversity, to evaluate nucleotide polymorphism and divergence levels of different gene clusters, the mutation parameter (θ) and nucleotide diversity (π) of the core gene clusters and variable gene clusters were calculated, respectively. For COG function annotation, the COG functions of gene clusters were annotated and visualized in combination with pan-genome conservation, providing users insights into the preference and conservation of protein functions in species. For metabolism cycle, a variety of maps were constructed for users to obtain information on the PAV and pan-genome conservation of metabolic process genes. For resistance PAV, heatmaps were generated to provide insights into PAV and pan-genome conservation of resistance genes. For gene cluster annotation, incorporating a variety of annotation information for each gene cluster, such as pan-genome conservation, KEGG pathway and PPI networks. Besides, users can exhaustively understand the biological function of gene clusters. Examples of the analyses are shown in Figure 5C.

DISCUSSION AND FUTURE DIRECTIONS

Given the explosive accumulation of prokaryotic genome datasets, conventional one-to-one comparative genomics has been incompetent for the analyses and comparison of massive datasets and thus, it becomes laborious to uncover more characteristics for the species of interest. The introduction of pan-genomics has shown unprecedented advantages in analyzing various characteristics of target species, for example, in genomic dynamics, gene conservation and species taxonomy. To comprehensively profile the characteristics of prokaryotic pan-genome dynamics as well as to adequately explore the PAV and conservation for pan-genome metabolic and resistant functional genes, we developed the database ProPan, a user-friendly portal with multiple flexible retrievals. ProPan integrates multi-dimensional analyses to profile the characteristics in pan-genome dynamics of 1,504 prokaryotes and furtherly excavate crucial functional genes for metabolic cycles (carbon cycle, nitrogen cycle, sulfur cycle and other cycle) and resistances (biocide, antimicrobial drug and metal).

To date, ProPan has covered 23 archaeal species and 1,481 bacterial species (in total 51,882 strain genome datasets). As increasingly prokaryotic genomic datasets are being released, we collect, polish and integrate accumulative prokaryotic datasets from more resources, such as IMG/M (22), to enrich the coming versions. Standard data quality control and consistent genome annotation also need to be constructed regarding the dataset from diverse resources. Plus, genus-level pan-genome analyses have received more frequent attention (49,50), thus this will be one of our priorities. Meanwhile, we are developing extensive modules and plugins in ProPan to enhance its functionality. For instance, visualization for pan-genome graphs can intuitively display the gain and loss of functional genes, structural variation and gene collinearity (51–53), thus it will be among our top priorities by adding this module in the short future. In addition to the metabolic cycles already included in ProPan, there are still a few other cycles in prokaryotes to be added, such as the phosphorus cycle (54,55) and potassium cycle (56,57). These metabolic cycles will continue to be integrated into coming releases. Taken together, as one of the essential databases in the National Genomics Data Center (58), we believe that ProPan will be a crucial resource for the studies of prokaryotic genome dynamics, species taxonomy and identification, environmental adaptation and further beyond.

DATA AVAILABILITY

ProPan is a comprehensive database for profiling prokaryotic pan-genome dynamics (https://ngdc.cncb.ac.cn/propan).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank a number of users for reporting bugs and providing suggestions.

FUNDING

National Key Research Program of China [2020YFA0907001]; Strategic Priority Research Program of the Chinese Academy of Sciences [XDB38030400]; National Natural Science Foundation of China [32170669]; China Postdoctoral Science Foundation [2021M693109]; Specialized Research Assistant Program of the Chinese Academy of Sciences [202044]. Funding for open access charge: Strategic Priority Research Program of the Chinese Academy of Sciences [XDB38030400].

Conflict of interest statement. None declared.

REFERENCES

1. Méric,G., Yahara,K., Mageiros,L., Pascoe,B., Maiden,M.C., Jolley,K.A. and Sheppard,S.K. (2014) A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic *Campylobacter*. *PLoS One*, **9**, e92798.
2. Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S.V., Crabtree,J., Jones,A.L., Durkin,A.S. *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 13950–13955.
3. Vernikos,G., Medini,D., Riley,D.R. and Tettelin,H. (2015) Ten years of pan-genome analyses. *Curr. Opin. Microbiol.*, **23**, 148–154.
4. Doron,S., Melamed,S., Ofir,G., Leavitt,A., Lopatina,A., Keren,M., Amitai,G. and Sorek,R. (2018) Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, **359**, eaar4120.
5. Cummins,E.A., Hall,R.J., McInerney,J.O. and McNally,A. (2022) Prokaryote pangenomes are dynamic entities. *Curr. Opin. Microbiol.*, **66**, 73–78.
6. Rouli,L., Merhej,V., Fournier,P.E. and Raoult,D. (2015) The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes New Infect*, **7**, 72–85.
7. Bazinet,A.L. (2017) Pan-genome and phylogeny of *Bacillus cereus sensu lato*. *BMC Evol. Biol.*, **17**, 176.
8. Caputo,A., Fournier,P.E. and Raoult,D. (2019) Genome and pan-genome analysis to classify emerging bacteria. *Biol. Direct*, **14**, 5.
9. Medini,D., Donati,C., Tettelin,H., Masignani,V. and Rappuoli,R. (2005) The microbial pan-genome. *Curr. Opin. Genet. Dev.*, **15**, 589–594.
10. Kim,Y., Gu,C., Kim,H.U. and Lee,S.Y. (2020) Current status of pan-genome analysis for pathogenic bacteria. *Curr. Opin. Biotechnol.*, **63**, 54–62.
11. Bryant,J.M., Brown,K.P., Burbaud,S., Everall,I., Belardinelli,J.M., Rodriguez-Rincon,D., Grogono,D.M., Peterson,C.M., Verma,D., Evans,I.E. *et al.* (2021) Stepwise pathogenic evolution of *Mycobacterium abscessus*. *Science*, **372**, eabb8699.
12. Mira,A., Martín-Cuadrado,A.B., D'Auria,G. and Rodríguez-Valera,F. (2010) The bacterial pan-genome:a new paradigm in microbiology. *Int. Microbiol.*, **13**, 45–57.
13. Brockhurst,M.A., Harrison,E., Hall,J.P.J., Richards,T., McNally,A. and MacLean,C. (2019) The Ecology and Evolution of Pangenomes. *Curr. Biol.*, **29**, R1094–R1103.
14. Tettelin,H. and Medini,D. (2020) In: *The Pangenome: Diversity, Dynamics and Evolution of Genomes*. Springer International Publishing.
15. Uchiyama,I., Mihara,M., Nishide,H., Chiba,H. and Kato,M. (2019) MBGD update 2018: microbial genome database based on hierarchical orthology relations covering closely related and distantly related comparisons. *Nucleic Acids Res.*, **47**, D382–D389.
16. Huang,K., Brady,A., Mahurkar,A., White,O., Gevers,D., Huttenhower,C. and Segata,N. (2014) MetaRef: a pan-genomic database for comparative and community microbial genomics. *Nucleic Acids Res.*, **42**, D617–D624.
17. Dehal,P.S., Joachimiak,M.P., Price,M.N., Bates,J.T., Baumohl,J.K., Chivian,D., Friedland,G.D., Huang,K.H., Keller,K., Novichkov,P.S. *et al.* (2010) MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res.*, **38**, D396–D400.
18. Mende,D.R., Letunic,I., Maistrenko,O.M., Schmidt,T.S.B., Milanese,A., Paoli,L., Hernández-Plaza,A., Orakov,A.N., Forslund,S.K., Sunagawa,S. *et al.* (2020) proGenomes2: an improved database for accurate and consistent habitat, taxonomic and functional annotations of prokaryotic genomes. *Nucleic Acids Res.*, **48**, D621–D625.
19. Whiteside,M.D., Winsor,G.L., Laird,M.R. and Brinkman,F.S. (2013) OrtholugeDB: a bacterial and archaeal orthology resource for improved comparative genomic analysis. *Nucleic Acids Res.*, **41**, D366–D376.
20. Zdobnov,E.M., Kuznetsov,D., Tegenfeldt,F., Manni,M., Berkeley,M. and Kriventseva,E.V. (2021) OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res.*, **49**, D389–D393.
21. Vallenet,D., Calteau,A., Dubois,M., Amours,P., Bazin,A., Beuvin,M., Burlot,L., Bussell,X., Fouteau,S., Gautreau,G. *et al.* (2020) MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Res.*, **48**, D579–D589.
22. Chen,I.A., Chu,K., Palaniappan,K., Ratner,A., Huang,J., Huntemann,M., Hajek,P., Ritter,S., Varghese,N., Seshadri,R. *et al.* (2021) The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. *Nucleic Acids Res.*, **49**, D751–D763.
23. Sayers,E.W., Beck,J., Bolton,E.E., Bourexis,D., Brister,J.R., Canese,K., Comeau,D.C., Funk,K., Kim,S., Klimke,W. *et al.* (2021) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **49**, D10–D17.
24. Wu,H., Wang,D. and Gao,F. (2021) Toward a high-quality pan-genome landscape of *Bacillus subtilis* by removal of confounding strains. *Brief Bioinform*, **22**, 1951–1971.
25. Inglin,R.C., Meile,L. and Stevens,M.J.A. (2018) Clustering of Pan- and Core-genome of *Lactobacillus* provides novel evolutionary insights for differentiation. *BMC Genomics*, **19**, 284.
26. Ondov,B.D., Starrett,G.J., Sappington,A., Kostic,A., Koren,S., Buck,C.B. and Phillippy,A.M. (2019) Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome Biol.*, **20**, 232.
27. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
28. Jain,C., Rodriguez,R.L., 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**, 5114.
29. Seemann,T. (2014) Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
30. Page,A.J., Cummins,C.A., Hunt,M., Wong,V.K., Reuter,S., Holden,M.T., Fookes,M., Falush,D., Keane,J.A. and Parkhill,J. (2015) Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, **31**, 3691–3693.
31. Tettelin,H., Riley,D., Cattuto,C. and Medini,D. (2008) Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.*, **11**, 472–477.
32. Snipen,L. and Liland,K.H. (2015) micropan: an R-package for microbial pan-genomics. *BMC Bioinf.*, **16**, 79.
33. Hutter,S., Vilella,A.J. and Rozas,J. (2006) Genome-wide DNA polymorphism analyses using VariScan. *BMC Bioinf.*, **7**, 409.
34. Huerta-Cepas,J., Szklarczyk,D., Heller,D., Hernández-Plaza,A., Forslund,S.K., Cook,H., Mende,D.R., Letunic,I., Rattei,T., Jensen,L.J. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.
35. Cantalapiedra,C.P., Hernández-Plaza,A., Letunic,I., Bork,P. and Huerta-Cepas,J. (2021) eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.*, **38**, 5825–5829.
36. Zhou,Z., Tran,P.Q., Breister,A.M., Liu,Y., Kieft,K., Cowley,E.S., Karaoz,U. and Anantharaman,K. (2022) METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome*, **10**, 33.
37. Feldgarden,M., Brover,V., Gonzalez-Escalona,N., Frye,J.G., Haendiges,J., Haft,D.H., Hoffmann,M., Pettengill,J.B., Prasad,A.B., Tillman,G.E. *et al.* (2021) AMRFinderPlus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci. Rep.*, **11**, 12728.
38. Alcock,B.P., Raphenya,A.R., Lau,T.T.Y., Tsang,K.K., Bouchard,M., Edalatmand,A., Huynh,W., Nguyen,A.V., Cheng,A.A., Liu,S. *et al.* (2020) CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, **48**, D517–D525.
39. Bortolaia,V., Kaas,R.S., Ruppe,E., Roberts,M.C., Schwarz,S., Cattoir,V., Philippon,A., Allesoe,R.L., Rebelo,A.R., Florensa,A.F. *et al.* (2020) ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.*, **75**, 3491–3500.
40. Gupta,S.K., Padmanabhan,B.R., Diene,S.M., Lopez-Rojas,R., Kempf,M., Landraud,L. and Rolain,J.M. (2014) ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother.*, **58**, 212–220.

41. Doster,E., Lakin,S.M., Dean,C.J., Wolfe,C., Young,J.G., Boucher,C., Belk,K.E., Noyes,N.R. and Morley,P.S. (2020) MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Res.*, **48**, D561–D569.
42. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinf.*, **10**, 421.
43. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
44. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P. *et al.* (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
45. 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.
46. Liao,W., Tong,D., Li,Z., Nie,X., Liu,Y., Ran,F. and Liao,S. (2021) Characteristics of microbial community composition and its relationship with carbon, nitrogen and sulfur in sediments. *Sci. Total Environ.*, **795**, 148848.
47. Alonso,A., Sánchez,P. and Martínez,J.L. (2001) Environmental selection of antibiotic resistance genes. *Environ. Microbiol.*, **3**, 1–9.
48. Allen,H.K., Donato,J., Wang,H.H., Cloud-Hansen,K.A., Davies,J. and Handelsman,J. (2010) Call of the wild: antibiotic resistance genes in natural environments. *Nat. Rev. Microbiol.*, **8**, 251–259.
49. Zhong,C., Han,M., Yu,S., Yang,P., Li,H. and Ning,K. (2018) Pan-genome analyses of 24 *Shewanella* strains re-emphasize the diversification of their functions yet evolutionary dynamics of metal-reducing pathway. *Biotechnology for Biofuels and Bioproducts*, **11**, 193.
50. Liu,X., Mao,B., Gu,J., Wu,J., Cui,S., Wang,G., Zhao,J., Zhang,H. and Chen,W. (2021) *Blautia*-a new functional genus with potential probiotic properties?*Gut Microbes*, **13**, 1–21.
51. Computational Pan-Genomics Consortium. (2018) Computational pan-genomics: status, promises and challenges. *Brief Bioinform*, **19**, 118–135.
52. Eizenga,J.M., Novak,A.M., Sibbesen,J.A., Heumos,S., Ghaffaari,A., Hickey,G., Chang,X., Seaman,J.D., Rounthwaite,R., Ebler,J. *et al.* (2020) Pangenome graphs. *Annu. Rev. Genomics Hum. Genet.*, **21**, 139–162.
53. Li,H., Wang,S., Chai,S., Yang,Z., Zhang,Q., Xin,H., Xu,Y., Lin,S., Chen,X., Yao,Z. *et al.* (2022) Graph-based pan-genome reveals structural and sequence variations related to agronomic traits and domestication in cucumber. *Nat. Commun.*, **13**, 682.
54. Richardson,A.E. and Simpson,R.J. (2011) Soil microorganisms mediating phosphorus availability update on microbial phosphorus. *Plant Physiol.*, **156**, 989–996.
55. Liang,J.L., Liu,J., Jia,P., Yang,T.T., Zeng,Q.W., Zhang,S.C., Liao,B., Shu,W.S. and Li,J.T. (2020) Novel phosphate-solubilizing bacteria enhance soil phosphorus cycling following ecological restoration of land degraded by mining. *ISME J.*, **14**, 1600–1613.
56. Etesami,H., Emami,S. and Alikhani,H.A. (2017) Potassium solubilizing bacteria (KSB): Mechanisms, promotion of plant growth, and future prospects - a review. *Journal of Soil Science and Plant Nutrition*, **17**, 897–911.
57. Wang,J., Li,R., Zhang,H., Wei,G. and Li,Z. (2020) Beneficial bacteria activate nutrients and promote wheat growth under conditions of reduced fertilizer application. *BMC Microbiol.*, **20**, 38.
58. CNCB-NGDC Members and Partners. (2022) Database resources of the national genomics data center, china national center for bioinformation in 2022. *Nucleic Acids Res.*, **50**, D27–D38.