D644–D650 *Nucleic Acids Research, 2021, Vol. 49, Database issue* *Published online 3 October 2020*
doi: 10.1093/nar/gkaa821

BacWGSTdb 2.0: a one-stop repository for bacterial whole-genome sequence typing and source tracking

Ye Feng[1,2], Shengmei Zou[2], Hangfei Chen[1], Yunsong Yu[1,*] and Zhi Ruan[1,*]

[1]Sir Run Run Shaw Hospital, Zhejiang University School of Medicine, Hangzhou 310016, China and [2]Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou 310029, China

Received August 05, 2020; Revised September 15, 2020; Editorial Decision September 16, 2020; Accepted September 17, 2020

ABSTRACT

An increasing prevalence of hospital acquired infections and foodborne illnesses caused by pathogenic and multidrug-resistant bacteria has stimulated a pressing need for benchtop computational techniques to rapidly and accurately classify bacteria from genomic sequence data, and based on that, to trace the source of infection. BacWGSTdb (http://bacdb.org/BacWGSTdb) is a free publicly accessible database we have developed for bacterial whole-genome sequence typing and source tracking. This database incorporates extensive resources for bacterial genome sequencing data and the corresponding metadata, combined with specialized bioinformatics tools that enable the systematic characterization of the bacterial isolates recovered from infections. Here, we present BacWGSTdb 2.0, which encompasses several major updates, including (i) the integration of the core genome multi-locus sequence typing (cgMLST) approach, which is highly scalable and appropriate for typing isolates belonging to different lineages; (ii) the addition of a multiple genome analysis module that can process dozens of user uploaded sequences in a batch mode; (iii) a new source tracking module for comparing user uploaded plasmid sequences to those deposited in the public databases; (iv) the number of species encompassed in BacWGSTdb 2.0 has increased from 9 to 20, which represents bacterial pathogens of medical importance; (v) a newly designed, user-friendly interface and a set of visualization tools for providing a convenient platform for users are also included. Overall, the updated BacWGSTdb 2.0 bears great utility in continuing to provide users, including epidemiologists, clinicians and bench scientists, with a one-stop solution to bacterial genome sequence analysis.

INTRODUCTION

The history of the world is intertwined with the impact that bacterial infectious diseases have had on humans. The advent of antimicrobials has fostered the belief that, as Sir McFarland Burnett stated in 1962, 'Almost all of the major practical problems of dealing with infectious disease had been solved' (1). However, this was soon led to disillusion due to the later emergence and rapid dissemination of antimicrobial resistance (AMR). In particular, in the current era of globalization, increasing international travel and food transportation have led to cross-board bacterial transmission events and even pandemics (2). A well-known example in the recent decade is the large outbreak caused by Shiga-toxin-producing *Escherichia coli* O104:H4, which started in Germany in the summer of 2011 with the consumption of sprouts and later spread in only two months to at least 16 countries (3). Although relatively rare, such mode of transmission also take place with hospital-acquired infections. For example, *Acinetobacter baumannii,* a notorious multi-drug resistant nosocomial bacteria, was believed to have initially infected American soldiers in Iraq and from them to have been brought back to military hospitals in the United States, after which it disseminated rapidly throughout the entire nation (4). Given the possible cross-border transmission nature of pathogenic bacteria and the consequent threat to global public health, there is a pressing need for global surveillance and the early detection of infectious disease outbreaks.

Traditional epidemiological studies have usually focused initially on patients with certain epidemiological links, recovered bacterial isolates from these subjects' clinical samples and finally determined whether their isolates had clonal relationships through bacterial typing techniques (5). However, such suspected links are often missing due to untimely or incomplete epidemiological surveys or asymptomatic carriers (6). Therefore, it is imperative to establish a reverse strategy, i.e. when bacterial isolates are found to be sufficiently similar by high-resolution typing techniques, they are deemed to be a consequence of infection from a common source (7). Due to its ultimate single base pair resolu-

*To whom correspondence should be addressed. Email: r_z@zju.edu.cn
Correspondence may also be addressed to Yunsong Yu. Email: yvys119@zju.edu.cn

Table 1. Name of species and number of isolates in BacWGSTdb

Species	Version 1	Version 2	Number of isolates (Version 1)	Number of isolates (Version 2)
Acinetobacter baumannii	√	√	1026	4260
Bacillus anthracis	√	√	158	227
Bacillus cereus		√		1017
Campylobacter coli		√		820
Campylobacter jejuni		√		1621
Clostridioides difficile		√		2322
Enterococcus faecalis		√		1523
Enterococcus faecium		√		1885
Escherichia coli	√	√	6328	21 733
Klebsiella pneumoniae	√	√	3279	9020
Listeria monocytogenes		√		3243
Mycobacterium abscessus		√		1611
Mycobacterium tuberculosis	√	√	2532	6512
Salmonella enterica	√	√	5276	14 658
Staphylococcus aureus	√	√	4890	11 505
Streptococcus agalactiae		√		1398
Streptococcus pneumoniae	√	√	3018	8347
Streptococcus suis		√		1252
Vibrio cholerae		√		818
Yersinia pestis	√	√	257	369

tion, whole-genome sequencing (WGS) has fulfilled this demand and gradually replaced pulsed-field gel electrophoresis (PFGE) and conventional seven-locus multi-locus sequence typing (MLST) as the new 'gold standard' typing technique (8,9). Concomitantly, there is a pressing need to organize, standardize and share bacterial genome sequences in a worldwide accessible database to fully exert the power of this reverse strategy on international surveillance and the early outbreak detection of bacterial infections.

In the year 2015, we initially introduced BacWGSTdb, a bacterial whole genome sequence typing and source tracking database designed for nine bacterial organisms of medical importance (10). This database incorporates extensive resources from bacterial genome sequence data as well as the corresponding metadata retrieved from the NCBI GenBank and BioSample database (11,12). By implementing a reference genome-based single-nucleotide polymorphism (SNP) approach, BacWGSTdb provides instant comparisons between user-uploaded genomes and an unprecedentedly large global set of genomes. Thus, clinicians, microbiologists and epidemiologists who work in medical facilities or public health institutions with no specialist knowledge of bioinformatics can use the database for the determination of clonal relationships and source tracking. Bench scientists can also use BacWGSTdb as a convenient tool for preliminary evolutionary and comparative genomic analyses.

Since the first version of BacWGSTdb, a vast number of additional genomic sequences was determined experimentally and, more importantly, a variety of newly discovered phenotypic traits (e.g. AMR and virulence) could be associated with the WGS data in the public domain (13,14). The demand for database updates and a uniform platform for real-time *in silico* prediction of these phenotypic traits based on genomic sequence data has become urgent. Here, we updated BacWGSTdb to version 2.0, which reflects not only a large increase in the curated dataset of bacterial genome sequencing data for the existing bacterial organisms, but also the new addition of eleven common pathogenic species. In addition to its updated content, a novel module for tracking the source of newly sequenced plasmids has been incorporated. AMR has challenged the treatment of infectious diseases which pose a serious threat to public health. As the major vector carrying AMR genes, plasmids are prone to horizontal transfer and offer AMR to originally antimicrobial susceptible bacteria, making treatment even more difficult. In this sense, tracing the transmission of AMR-carrying plasmids is of equal importance to tracing that of bacteria. Furthermore, the phylogenetic analysis in BacWGSTdb 2.0 is enhanced by adding the core genome MLST (cgMLST) approach, thereby meeting different genotyping demands. Taken together, we expect that BacWGSTdb 2.0 will continue to benefit users by providing a one-stop repository for bacterial whole-genome sequence typing and source tracking.

DATABASE UPDATE AND ENHANCEMENTS

BacWGSTdb 1.0 included nine bacterial species, while in this update, eleven more species have been newly added. The detailed species and the number of isolates included in BacWGSTdb are listed in Table 1. Thereby, BacWGSTdb has covered almost all common nosocomial, community and foodborne bacterial pathogens.

The usage of BacWGSTdb includes two function modules, 'Tools' and 'Browse'. The former allows users to upload their own FASTA-formatted genome sequence(s) to find closely related records in the database. The latter allows users to browse the isolates in the database, compare their clinical and microbiological traits, and investigate their phylogenetic relationships. With the release of BacWGSTdb 2.0, we are introducing multiple major changes in both the backend and the web interface that provide users with more effective and efficient ways to browse and query the WGS data and to quickly cluster and identify related sequences for uncovering potential transmission sources. This helps clinicians or public health scientists investigate hospital-acquired or foodborne infection outbreaks. In addition, BacWGSTdb 2.0 also provides the online analyses of conventional seven-locus MLST, predictions of AMR and vir-

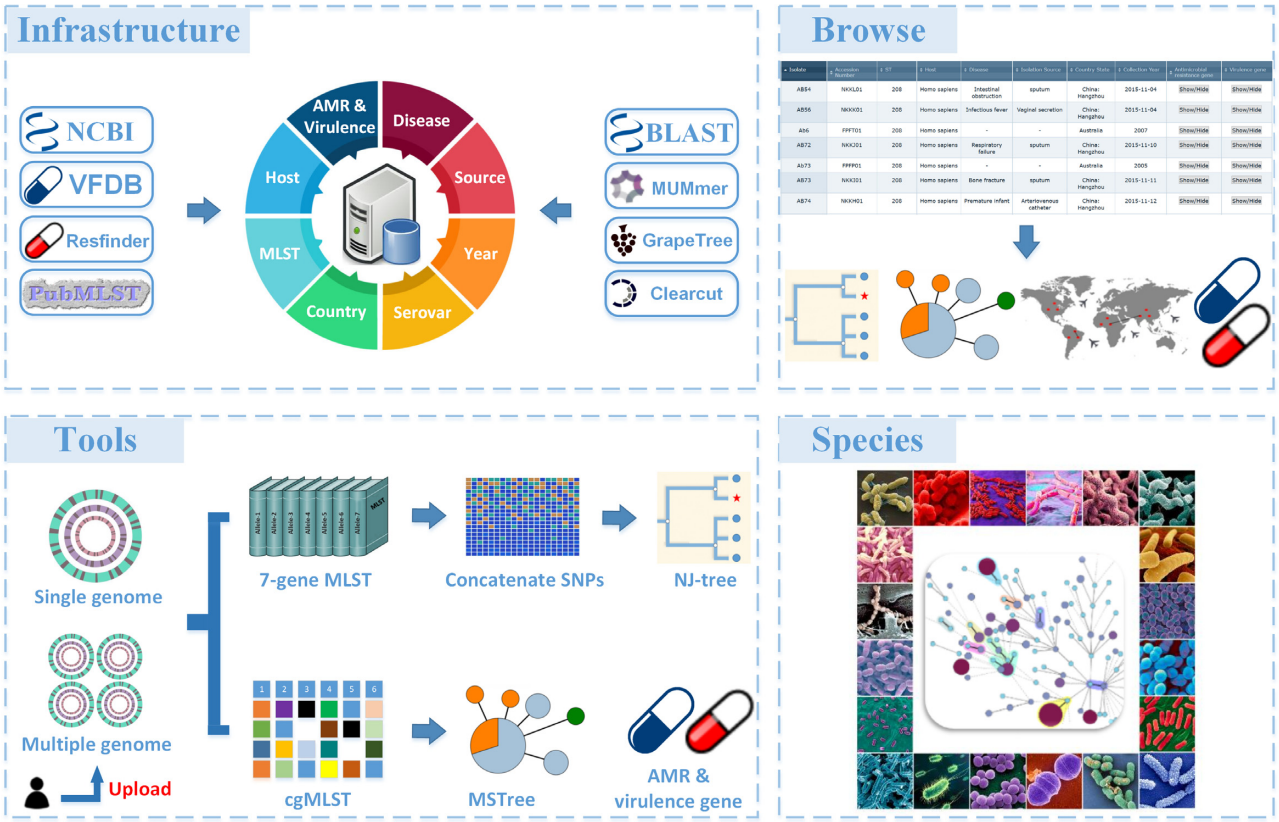

Figure 1. Overview of the content and function modules of BacWGSTdb 2.0. '**Infrastructure**' lists the public database and tools integrated in BacWGSTdb 2.0. '**Browse**' functions to visualize and compare the genetic relationships among isolates deposited in BacWGSTdb 2.0. '**Tools**' functions for whole genome sequence typing and source tracking based on user uploaded sequence(s). '**Species**' represents 20 bacterial species currently supported by BacWGSTdb 2.0.

ulence genes, and source tracking of plasmid sequences. An overview of the database schematic is shown in Figure 1.

Single genome analysis

In BacWGSTdb 1.0, the module Single Genome Analysis performs conventional seven-locus MLST analysis and searches of the genetically closest relatives in the database following the SNP approach upon the user's uploaded pre-assembled single genome. In detail, MUMmer 3.22 is used for the alignment with the reference genome and the subsequent SNP identification; and the phylogenetic tree is generated by Clearcut 1.0 (15,16). The most important update to this module in BacWGSTdb 2.0 is that it now offers both SNP and cgMLST analysis to compare the user uploaded genome sequence against those deposited in the database. The SNP approach compares single nucleotide differences between isolates to a designated reference genome, which can be used to investigate the clonal relationship among isolates sharing a high genetic relatedness (e.g. collected from outbreaks of hospital or foodborne infections). By comparison, the cgMLST approach, another widely used sequence typing approach for bacterial genomes, is an extension of conventional seven-locus MLST scheme that expands the range of target genes to whole genome level and is often used as a solution to provide highly detailed phylogenetic relatedness of a species and is suitable for investigating the middle/long-term evolutionary history of bacterial pathogens. Thus, the two approaches are complementary and meet different genotyping demands (17).

To perform a cgMLST analysis, a pre-defined reference database (cgMLST scheme) for each species, which contains all known allelic variants in the coding regions for all genomes deposited in BacWGSTdb, is prepared at the backend of BacWGSTdb 2.0 by LOCUST 1.0 (18). The user uploaded genomic sequence is compared to the species-specific cgMLST scheme using BLASTn (19). The identified allelic profile continues to make comparisons with that of each isolate deposited in BacWGSTdb 2.0 for determination of the closely related isolates according to their number of pairwise allelic differences. GrapeTree is used to construct and visualize the minimal spanning tree generated based on the cgMLST allelic profiles, which supports manipulations of both tree layout and the user specified metadata attributes (20) (Figure 2).

Another important update in this module is the integration of a novel function for typing and source tracking of plasmids. BacWGSTdb 2.0 deposits all complete plasmid sequences from the NCBI GenBank database as well as their metadata. When a draft bacterial genome, which usually contains the plasmid sequence of the sequenced isolate, is uploaded, its plasmid replicon types will be determined by searching the genomic sequence against the database of plasmid replicon genes with BLASTn (21). Moreover,

Figure 2. Screenshots of the updated web interface of the **Single Genome Analysis** module and the detailed outputs using a *Salmonella enterica* isolate as an example. After uploading the preassembled genomic sequence and setting the appropriate parameters, the analytical results return and can subsequently be classified into three sections: conventional seven-locus MLST, identification of AMR and virulence genes, and source tracking of plasmids and bacteria. In particular, the phylogenetic analysis revealed that the query isolate shows high degree of genetic relatedness in the database, suggesting that these isolates might have originated from the same source. The entire analysis process takes 3–5 min.

pairwise distances between the users' sequence and those of each of the plasmids in BacWGSTdb are computed using Mash 2.2 with maximal *P*-value set to 0.1 and minimal identity set to 0.9 (22). The analytical results include a table listing the records of similar plasmids based on Mash distance, together with a world map displaying the records with available geographical location data (Figure 2). This update makes it possible to trace the transmission routes of plasmids, which we believe are of at least equal importance with that of chromosomes.

Furthermore, the *in silico* identification of acquired AMR and virulence genes are also incorporated into BacWGSTdb 2.0. The user uploaded sequence is searched against the ResFinder 3.2 and VFDB 2019 database by BLASTn, with minimal identity and a threshold coverage of 90% (23–25).

Multiple genome analysis

The newly designed multiple genome analysis module can process up to 30 user uploaded genome sequences in a batch mode. *In silico* identifications of conventional seven-locus MLST, AMR and virulence genes are performed for each of the uploaded genomes. A more important purpose for this module is that when users collect multiple isolates which they suspect belong to the same outbreak event, the module can help determine whether there is a clonal relationship among these isolates according to the pairwise comparison

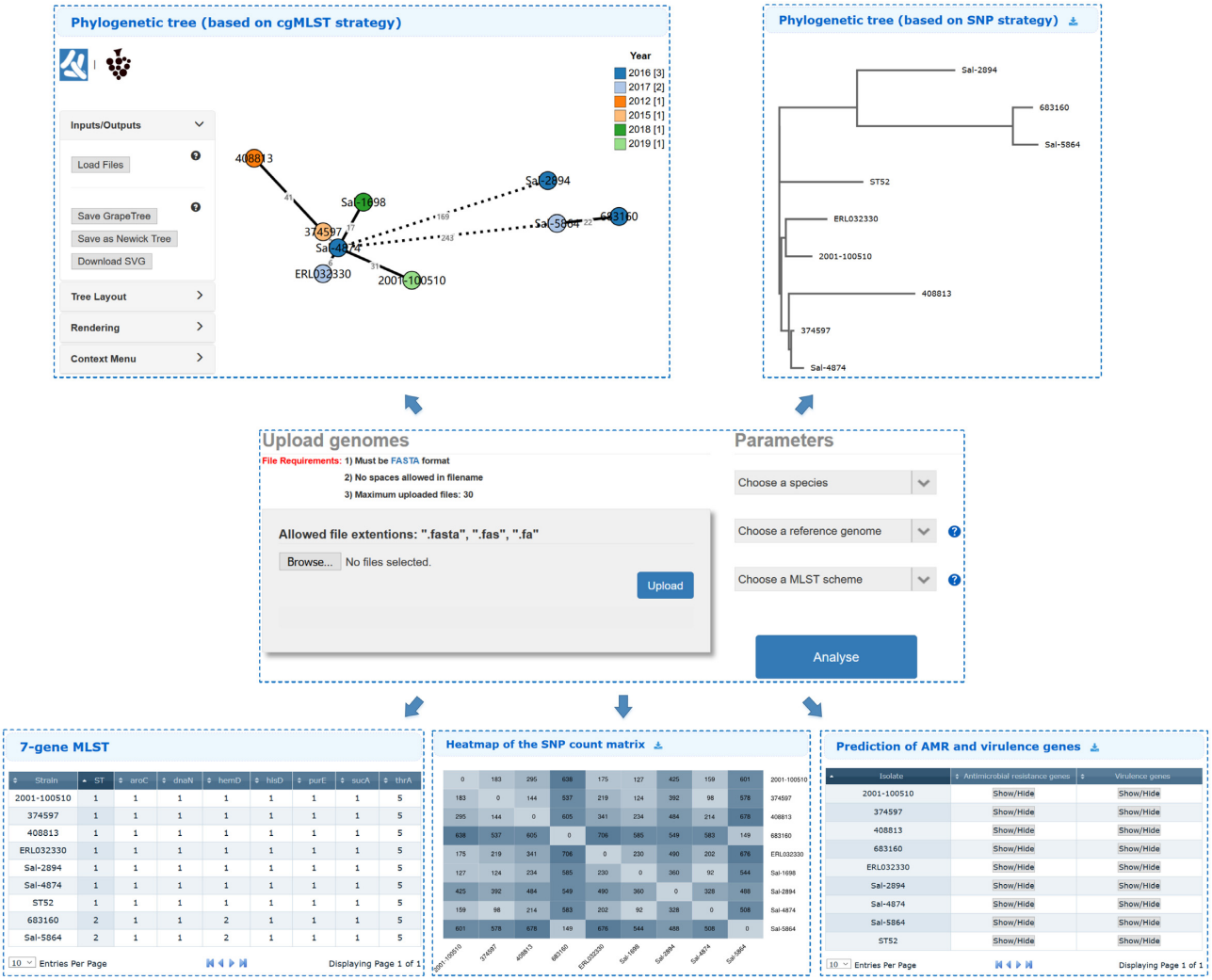

Figure 3. Screenshots of the updated web interface of the **Multiple Genome Analysis** module and the detailed outputs using nine *S. enterica* isolates as an example. The Results page lists for each of the uploaded genomes the conventional seven-locus MLST results and the predictions of AMR and virulence genes. In addition, the phylogenetic trees based on the SNP and cgMLST approaches both reveal that the query isolates did not involve an outbreak event, since they differ from one another by over 100 SNPs or cgMLST loci. The entire analysis process takes 5–8 min.

of the cgMLST alleles or SNP differences. To this end, the phylogenetic relatedness among the user uploaded multiple genomic sequences will be determined following both the SNP and cgMLST approaches (Figure 3).

Browse and Search

The Browse and Search module is designed to visualize and compare the genetic relationships among isolates deposited in BacWGSTdb. Each 'Browse' page lists the clinical and microbiological metadata of each of the isolates deposited in the database, including their seven-locus MLST sequence type, host, clinical outcome, collection date and geographical location. Data can be sorted by clicking on a specific column header and downloaded as a tab-delimited file. The updated annotations on the AMR and virulence genes for each isolate are also displayed. In BacWGSTdb 1.0, users can select multiple isolates belonging to the same sequence type, and build phylogenetic trees following the SNP approach. This limit has been broken in BacWGSTdb 2.0: users can select and compare any isolate of interest, even from among those belonging to different sequence types. Both the SNP and cgMLST approaches will be applied for establishment of phylogenetic trees. In addition, a newly designed search function enables users to look up keywords (e.g. sequence types) of interest based on varied categories (Figure 4). We therefore believe that the updated Browse and Search module will allow users to retrieve information from the database in a convenient and time-saving manner.

CONCLUSIONS AND FUTURE PERSPECTIVES

Next-generation sequencing and bioinformatics are expediting pathogen characterization, transforming the response to infectious disease outbreaks, and providing new insights into disease emergence and transmission. Standardized and user-friendly online databases make WGS analysis more accessible, even to those lacking bioinformatics expertise. Here, we reported a major update of BacWGSTdb, including the significantly expanded con-

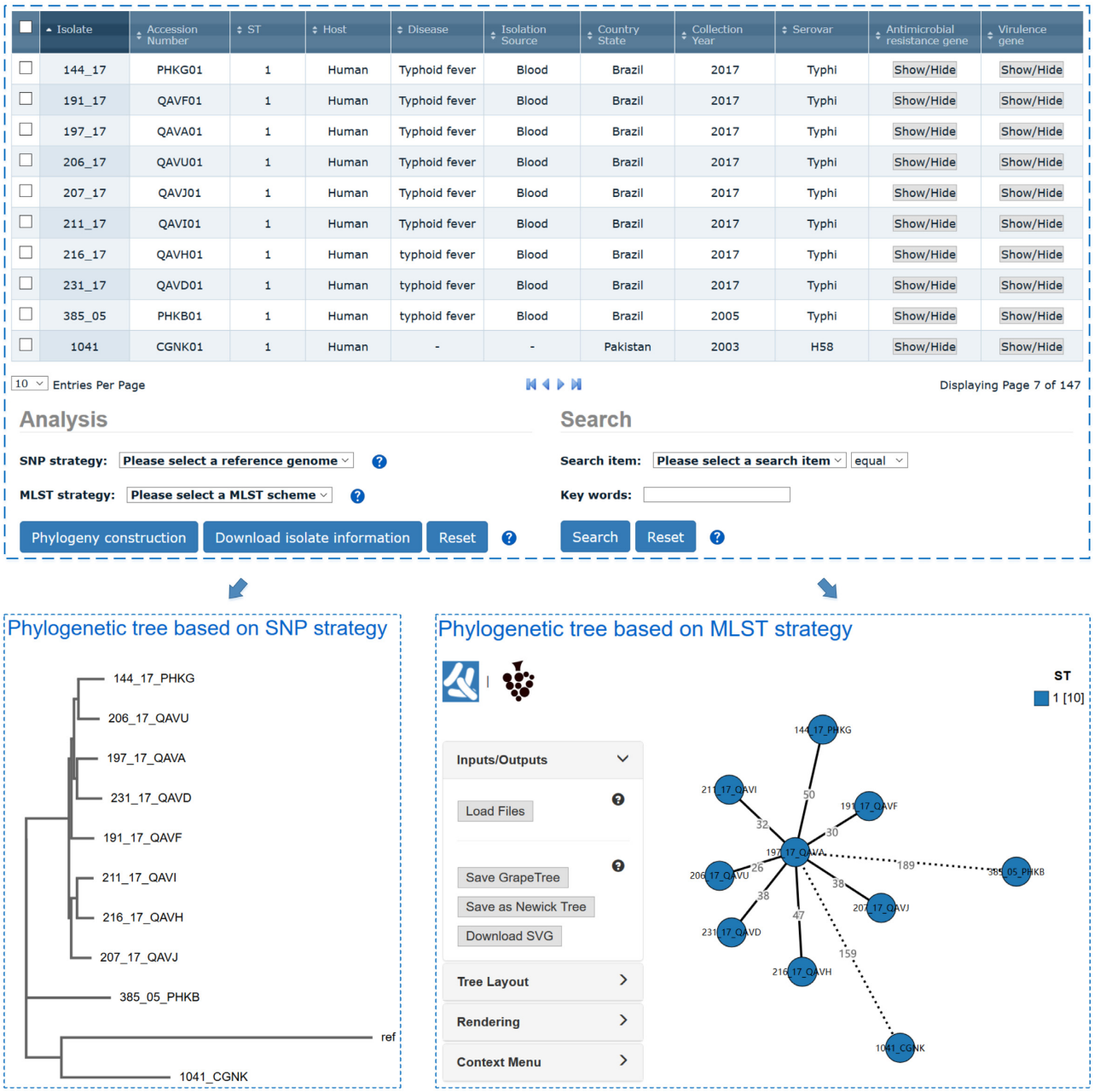

Figure 4. Screenshots of the updated Browse interface. Users can sort and select isolates based on their various attributes. For the selected isolates, phylogenetic trees following both the SNP and cgMLST approaches are provided.

tent of the database, additional analytical and visualization tools, and a newly designed, user-friendly interface, all of which greatly facilitate the genomic epidemiological surveillance of bacterial pathogens. We believe these additions significantly enrich our database, which is expected to provide a one-stop solution to bacterial genome analysis and, more importantly, translate whole genome sequencing from proof-of-concept to routine use in clinical practice.

FUNDING

National Natural Science Foundation of China [31670132, 81401698]; Zhejiang Province Public Welfare Technology Application Research Project [LGF18H190001]. Funding for open charge: National Natural Science Foundation of China [31670132].

Conflict of interest statement. None declared.

REFERENCES

1. Brachman,P.S. (2003) Infectious diseases–past, present, and future. *Int. J. Epidemiol.*, **32**, 684–686.
2. Hernando-Amado,S., Coque,T.M., Baquero,F. and Martinez,J.L. (2019) Defining and combating antibiotic resistance from One Health and Global Health perspectives. *Nat. Microbiol.*, **4**, 1432–1442.
3. Alexander,D.C., Hao,W., Gilmour,M.W., Zittermann,S., Sarabia,A., Melano,R.G., Peralta,A., Lombos,M., Warren,K., Amatnieks,Y.

et al. (2012) Escherichia coli O104:H4 infections and international travel. *Emerg. Infect. Dis.*, **18**, 473–476.
4. Scott,P., Deye,G., Srinivasan,A., Murray,C., Moran,K., Hulten,E., Fishbain,J., Craft,D., Riddell,S., Lindler,L. *et al.* (2007) An outbreak of multidrug-resistant Acinetobacter baumannii-calcoaceticus complex infection in the US military health care system associated with military operations in Iraq. *Clin. Infect. Dis.*, **44**, 1577–1584.
5. Quainoo,S., Coolen,J.P.M., van Hijum,S., Huynen,M.A., Melchers,W.J.G., van Schaik,W. and Wertheim,H.F.L. (2017) Whole-Genome Sequencing of Bacterial Pathogens: the Future of Nosocomial Outbreak Analysis. *Clin. Microbiol. Rev.*, **30**, 1015–1063.
6. Chisholm,R.H., Campbell,P.T., Wu,Y., Tong,S.Y.C., McVernon,J. and Geard,N. (2018) Implications of asymptomatic carriers for infectious disease transmission and control. *R. Soc. Open Sci.*, **5**, 172341.
7. Ruan,Z., Yu,Y. and Feng,Y. (2020) The global dissemination of bacterial infections necessitates the study of reverse genomic epidemiology. *Brief. Bioinform.*, **21**, 741–750.
8. Boolchandani,M., D'Souza,A.W. and Dantas,G. (2019) Sequencing-based methods and resources to study antimicrobial resistance. *Nat. Rev. Genet.*, **20**, 356–370.
9. Gardy,J.L. and Loman,N.J. (2018) Towards a genomics-informed, real-time, global pathogen surveillance system. *Nat. Rev. Genet.*, **19**, 9–20.
10. Ruan,Z. and Feng,Y. (2016) BacWGSTdb, a database for genotyping and source tracking bacterial pathogens. *Nucleic Acids Res.*, **44**, D682–D687.
11. Sayers,E.W., Cavanaugh,M., Clark,K., Ostell,J., Pruitt,K.D. and Karsch-Mizrachi,I. (2020) GenBank. *Nucleic Acids Res.*, **48**, D84–D86.
12. Courtot,M., Cherubin,L., Faulconbridge,A., Vaughan,D., Green,M., Richardson,D., Harrison,P., Whetzel,P.L., Parkinson,H. and Burdett,T. (2019) BioSamples database: an updated sample metadata hub. *Nucleic Acids Res.*, **47**, D1172–D1178.
13. Su,M., Satola,S.W. and Read,T.D. (2019) Genome-Based Prediction of Bacterial Antibiotic Resistance. *J. Clin. Microbiol.*, **57**, e01405-18.
14. Macesic,N., Bear Don't Walk,O.I., Pe'er,I., Tatonetti,N.P., Peleg,A.Y. and Uhlemann,A.C. (2020) Predicting Phenotypic Polymyxin Resistance in Klebsiella pneumoniae through Machine Learning Analysis of Genomic Data. *mSystems*, **5**, e00656-19.
15. 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.
16. Sheneman,L., Evans,J. and Foster,J.A. (2006) Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics*, **22**, 2823–2824.
17. Schürch,A.C., Arredondo-Alonso,S., Willems,R.J.L. and Goering,R.V. (2018) Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. *Clin. Microbiol. Infect.*, **24**, 350–354.
18. Brinkac,L.M., Beck,E., Inman,J., Venepally,P., Fouts,D.E. and Sutton,G. (2017) LOCUST: a custom sequence locus typer for classifying microbial isolates. *Bioinformatics*, **33**, 1725–1726.
19. Johnson,M., Zaretskaya,I., Raytselis,Y., Merezhuk,Y., McGinnis,S. and Madden,T.L. (2008) NCBI BLAST: a better web interface. *Nucleic Acids Res.*, **36**, W5–W9.
20. Zhou,Z., Alikhan,N.F., Sergeant,M.J., Luhmann,N., Vaz,C., Francisco,A.P., Carrico,J.A. and Achtman,M. (2018) GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res.*, **28**, 1395–1404.
21. 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.
22. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
23. Carattoli,A., Zankari,E., García-Fernández,A., Voldby Larsen,M., Lund,O., Villa,L., Møller Aarestrup,F. and Hasman,H. (2014) In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.*, **58**, 3895–3903.
24. Liu,B., Zheng,D., Jin,Q., Chen,L. and Yang,J. (2019) VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.*, **47**, D687–D692.
25. Zankari,E., Hasman,H., Cosentino,S., Vestergaard,M., Rasmussen,S., Lund,O., Aarestrup,F.M. and Larsen,M.V. (2012) Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.*, **67**, 2640–2644.