D682–D687 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 3 October 2015*
doi: 10.1093/nar/gkv1004

BacWGSTdb, a database for genotyping and source tracking bacterial pathogens

Zhi Ruan[1] and Ye Feng[1,2,*]

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

Received August 13, 2015; Revised September 19, 2015; Accepted September 23, 2015

ABSTRACT

Whole genome sequencing has become one of the routine methods in molecular epidemiological practice. In this study, we present BacWGSTdb (http://bacdb.org/BacWGSTdb), a bacterial whole genome sequence typing database which is designed for clinicians, clinical microbiologists and hospital epidemiologists. This database borrows the population structure from the current multi-locus sequence typing (MLST) scheme and adopts a hierarchical data structure: species, clonal complex and isolates. When users upload the pre-assembled genome sequences to BacWGSTdb, it offers the functionality of bacterial genotyping at both traditional MLST and whole-genome levels. More importantly, users are told which isolates in the public database are phylogenetically close to the query isolate, along with their clinical information such as host, isolation source, disease, collection time and geographical location. In this way, BacWGSTdb offers a rapid and convenient platform for worldwide users to address a variety of clinical microbiological issues such as source tracking bacterial pathogens.

INTRODUCTION

Based on the premise that similar isolates may share similar medical trait, a *prima facie* concern shared by the global medical community is: 'Have we seen the particular pathogen before? Where, when and what kind of disease is it associated with?' To address this concern, a variety of genotyping methods have been developed, of which multi-locus sequence typing (MLST) is considered the gold standard for many bacterial pathogens for over a decade (1,2). While the performance of MLST is good enough in inter-lineage genotyping, this technology lacks enough discriminatory capability to differentiate tightly linked bacterial isolates (3,4). The advent of next-generation sequencing technologies has made it possible to obtain the entire bacterial genome at relatively modest cost and effort. Because its extremely high resolution could afford source tracking of the same bacterial clone isolated from different patients, areas or periods, whole genome sequencing (WGS) has been increasingly used to solve a wide range of research problems concerning bacterial epidemiology, drug resistance, pathogenicity and evolution (5–9). To date, US Food and Drug Administration (FDA) has granted WGS the marketing authorization for investigating food-borne outbreaks in USA. It is therefore expected in the near future that WGS would become a routine tool not only for basic research but also for clinical diagnostics and surveillance.

Nevertheless, the development of sequencing technology itself is not sufficient to achieve this goal. An easy-to-use public database is also required in order for international exchange of whole genome sequence typing (WGST) information of bacteria. Although Sequence Read Archive (SRA) and European Nucleotide Archive (ENA) have already offered a platform for storing the raw reads of WGS (10,11), it is far from convenient for investigators to deploy the raw data, especially those clinicians with limited bioinformatics skills. More importantly, the provenance for many isolates, such as host, isolation source, disease, collection time and geographical location, has not always been submitted along with the genome sequences. Consequently, medical significance could not be predicted even if a highly similar genome can be found. From this perspective, a new tool that is designed for the clinicians, clinical microbiologists and hospital epidemiologists monitoring the emergence and outbreak of important bacterial pathogens is urgently needed.

Generally, two strategies scale well to handling the genomic comparison of thousands of bacterial isolates: gene-by-gene genomic analysis and a reference genome-based single nucleotide polymorphism (SNP) strategy (12). The former is actually a MLST-like approach called whole-genome MLST (wgMLST), which is based on indexing alleles for all coding sequences in the genome, and therefore providing a highly scalable means of studying the sequence variation encoded within it. In theory, the gene-by-gene strategy is suitable for typing bacteria at a wide range of resolutions and might present a much more accurate phylogenetic relationship than traditional MLST (13,14).

*To whom correspondence should be addressed. Tel: +86 571 86006142; Fax: +86 571 86006142; Email: pandafengye@zju.edu.cn

Currently the Bacterial Isolate Genome Sequence Database (BIGSdb), which was developed following this strategy, has been integrated into the PubMLST database (http://pubmlst.org) for a few species. Users can choose either the traditional MLST scheme or the new whole genome-based one according to their specific typing purpose (14–16). However, the phylogenetic analysis performed by this strategy is quite time-consuming and usually demands considerable computational resources, especially when manipulating hundreds of genome sequences together.

In contrast, the reference genome-based SNP strategy borrows the population structure from the current MLST schemes and requires a reference genome for each of the clonal complexes. The genomes of bacterial isolates are compared against the reference genome and the derived SNP data are used for further phylogenetic analysis. This algorithm assigns a higher identity between sequences differing in only one single nucleotide and a lower identity between sequences with multiple differences. Although it is unsuited to some highly diverse bacterial species such as *Pseudomonas aeruginosa*, or for comparing relationship among remotely related lineages, the reference genome-based SNP strategy offers a satisfactory resolution to differentiate isolates belonging to the same clone and is therefore suitable for analyzing the clonal structure of isolates emerging in an outbreak.

In this study, we introduced a new database which offers the ability to extract MLST information from a bacterial genome sequence. More importantly, it would also help find isolates in the public database that are phylogenetically close to the query isolate, along with their clinical information. We believe this function is vital for source tracking bacterial pathogens during outbreak investigation in the era of genomic epidemiology.

DATABASE DESCRIPTION

Bacterial Whole Genome Sequence Typing Database (BacWGSTdb, http://bacdb.org/BacWGSTdb) aims to provide genotyping at both traditional MLST and WGST level. For this purpose, we borrowed the population structure from the current MLST schemes and specified a reference genome for each of the clonal complexes. The clonal complex is defined herein to be a set of Sequence Types (STs) that differ by one or two alleles. The isolates stored in our database are firstly genotyped according to the MLST scheme, and an appropriate reference genome is chosen by the typing result. The complete or draft genome of the isolates continues to be compared against the specified reference genome. The obtained SNP information is stored in the database as well as the clinical information of isolates, including host, isolation source, disease, collection time and geographical location. Figures 1 and 2 list the infrastructure and the general workflow of data processing of BacWGSTdb, respectively.

Construction of the phylogenetic tree in BacWGSTdb relies on Neighbor-Joining (NJ) algorithm. Although NJ is less accurate than other algorithms such as Maximum-Parsimony or Maximum-Likelihood, this disadvantage is minimized to a great extent when the compared strains are very close to each other. Meanwhile, the NJ algorithm runs significantly faster than the others, especially when manipulating the genome of hundreds of isolates together (17,18). Thus, the speed of retrieval in BacWGSTdb is fast, which fulfills the need of real-time monitoring and identification of bacterial outbreaks.

BacWGSTdb has been implemented using MySQL 5.6 (http://www.mysql.org), PHP 5.5 (http://www.php.net) and Apache 2.4 (http://www.apache.org) on a Red Hat Enterprise Linux Server 6.0. The interface component consists of webpages designed and implemented in HTML/CSS in a Linux environment. It has been tested in the Google Chrome, Mozilla Firefox, Apple Safari, Internet Explorer and Microsoft Edge web browsers.

At the background of BacWGSTdb, BLAST 2.2.26 is used for comparing the query genome with the MLST allele sequences (19). MUMmer 3.22 is used for alignment with the reference genome and the subsequent SNP identification (20). Indels and adjacent mismatches are not considered as true SNPs and are pruned by self-developed Perl scripts. The phylogenetic tree was generated by Clearcut 1.0, a fast and open source implementation for the relaxed NJ algorithm and displayed as Scalable Vector Graphics (SVG) in the web browser using TreeVector (21,22).

BacWGSTdb currently encompasses nine bacterial organisms of medical importance, i.e. *Acinetobacter baumannii*, *Bacillus anthracis*, *Escherichia coli*, *Klebsiella pneumoniae*, *Mycobacterium tuberculosis*, *Salmonella enterica*, *Staphylococcus aureus*, *Streptococcus pneumoniae* and *Yersinia pestis*, all of which can be described by a clonal population structure. At the present stage, genome sequences from GenBank and PATRIC databases have been deployed for preparing BacWGSTdb (23,24). The genome sequences from SRA and ENA databases with detailed strain information have also been incorporated (10,11). When the genome assembly is not available, the raw sequence reads were *de novo* assembled into the draft genome, which was further mapped to the reference genome. The obtained SNP data were stored in BacWGSTdb. The database will be updated periodically and new species can be easily added.

USAGE OF BacWGSTdb

Use of BacWGSTdb includes two major parts: TOOLS and BROWSE. One of the most important tools, *Typing & Tracking*, is designed for users who have sequenced the genome of their query isolates. After uploading a pre-assembled complete or draft genome sequence, users are told the MLST information of the isolate as well as the recommended reference genome. The query genome is then aligned against the reference genome and the SNP data are provided for download, which is in standard Variant Call Format (VCF). Next, the SNP data would be automatically compared with those deposited in the database, and the most similar isolates to the query one will be displayed (Figure 3). At the bottom of the resulting page, a phylogenetic tree is displayed in order to better reveal the phylogenetic relationship between the query isolate and the listed close isolates.

We have also developed a series of tools to serve as effective supplementations to the key tool *Typing & Tracking*, including:

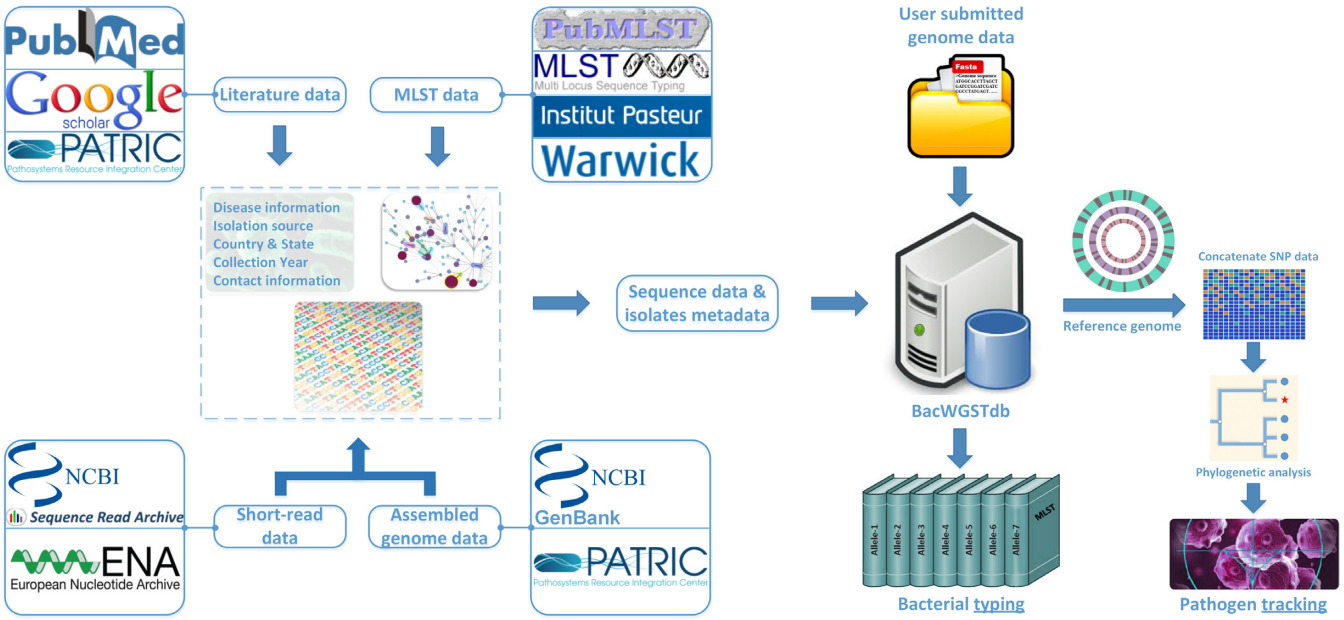

Figure 1. Database structure.

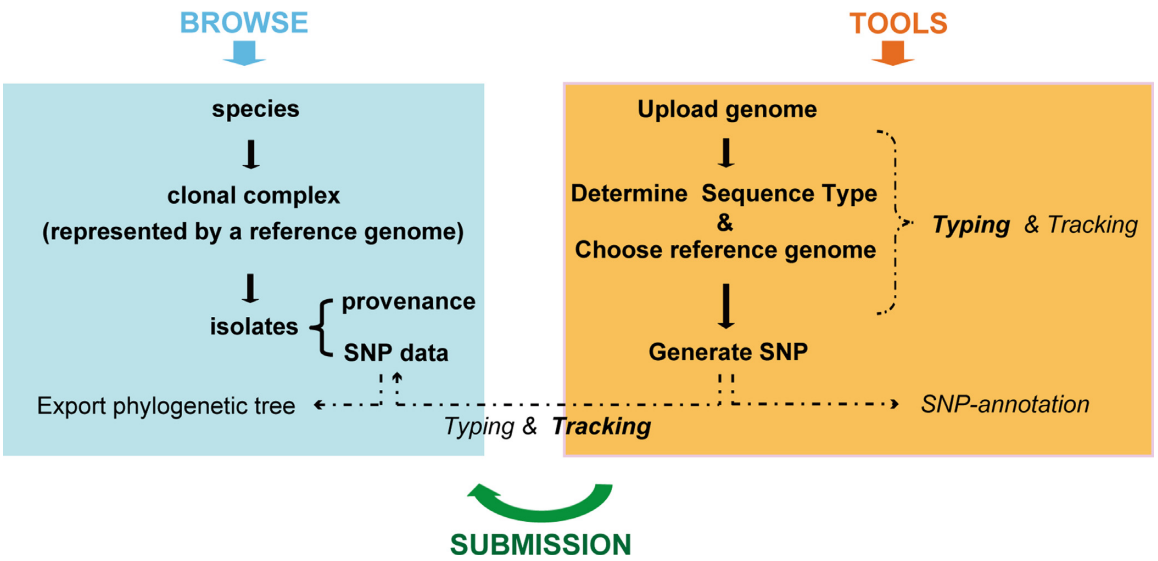

Figure 2. Workflow of data processing. BacWGSTdb contains three main sections: Browse, Tools and Submission. They all adopt a hierarchical infrastructure: species, clonal complex (represented by a reference genome) and isolates. SNP data is the key component of the database and connects the three sections together.

(i) *SNP-annotation*: this tool facilitates users to predict the outcome of SNPs. Based on the genomic annotation stored within the database, each of the SNPs uploaded by users will be judged whether it is synonymous, non-synonymous or intergenic.

(ii) *Choose-refgenome-by-ST*: by using this tool, users input the ST number and are told which reference genome they should use.

(iii) *Generate-SNP-by-genome*: when users are willing to choose other reference genomes instead of the recommended one, they can upload both the reference and the query genome and download the resulting SNP data.

(iv) *Coordinate-conversion*: if users don't generate the SNP data by the recommended reference genome at the very beginning, but they are willing to use or submit the SNP data into our database, they can make the coordinate conversion with this tool. Users are required to upload their SNP data and specify the two reference genomes. Pairwise alignment between the two genomes runs at the backend server and then the converted coordinates will be provided.

The BROWSE function is designed for visualizing and comparing isolates deposited in the database. When users browse BacWGSTdb, they need to choose a reference

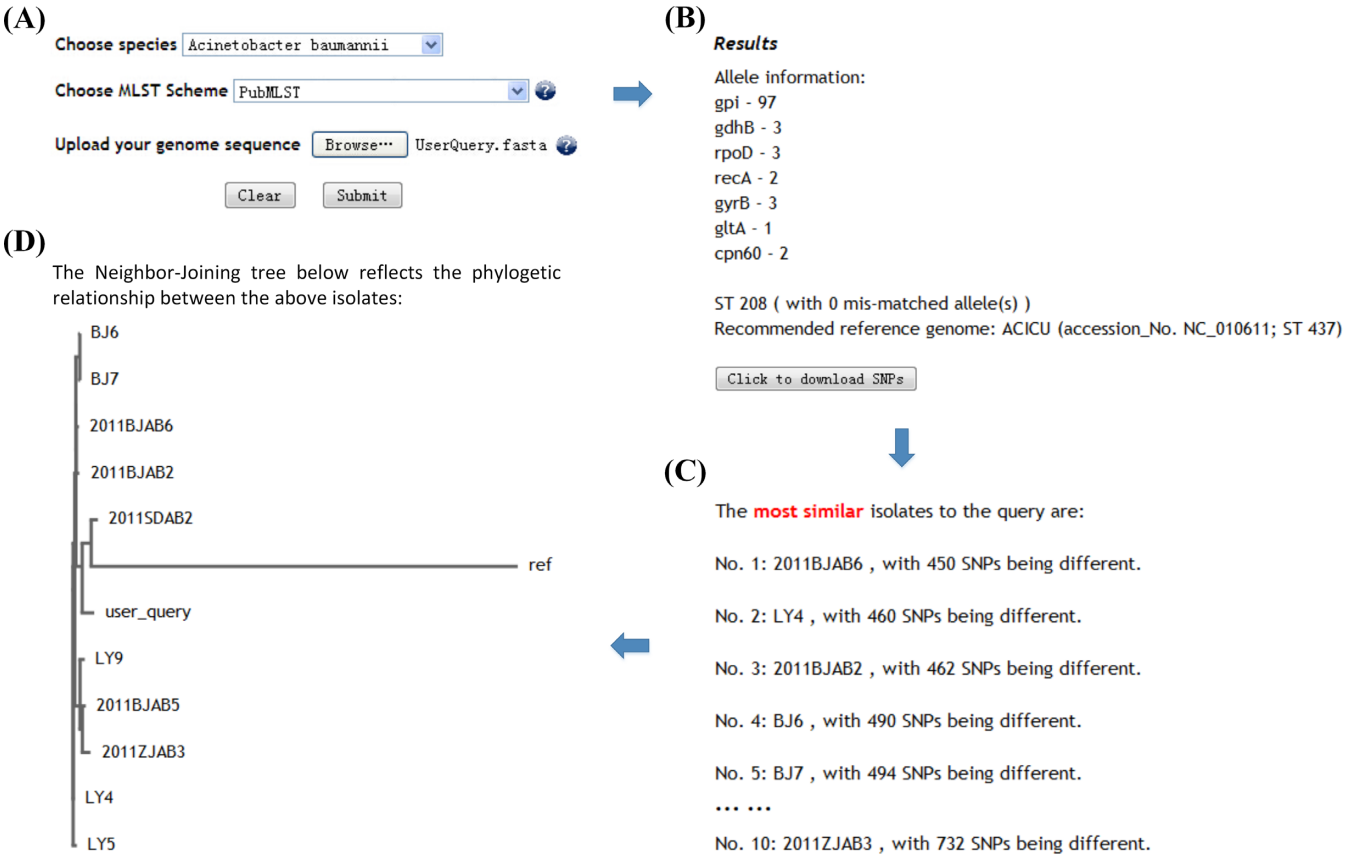

Figure 3. Usage example of the tool *Typing & Tracking*. Panel (**A**) shows the entry page of *Typing & Tracking*, in which users choose species, MLST scheme and upload a query genome sequence. Panel (**B**–**D**) are the results of *Typing & Tracking*: Panel (B) lists the MLST information, the suggested reference genome and the SNP file for download; Panel (C) lists the ten most similar isolates to the query one based on the number of different SNPs; Panel (D) shows the phylogenetic relationship between the query and the ten most similar isolates.

genome first (each reference genome represents a clonal complex) and further choose isolates of interest based on ST, host, clinical outcome, geographical location or any other attributes (Figure 4). According to the SNP data against the same reference genome, a NJ unrooted tree is provided for guidance, which reflects the phylogenetic relationship between the selected isolates. Users can also upload their own SNP data (e.g., produced by *Typing & Tracking*) and compare it with those in the database to figure out the phylogenetic position of their query isolate among the selected isolates. In addition to view the SVG formatted phylogenetic tree in the browser directly, users can also choose to download the Newick-formatted tree files for examination and/or annotation in external tree-drawing applications.

In the SUBMISSION page, users are encouraged to submit their own data to BacWGSTdb. They need to fill in some basic information of their isolate and upload the SNP data (e.g. produced by *Typing & Tracking*). The uploaded SNP data will appear in the BROWSE page 24 h after submission. Users can also contact the administrator to make the curation. The SNP data could be prepared in two ways: users can directly map the raw WGS reads to a reference genome, or align the *de novo*-assembled contigs to the reference genome. We recommend the latter way because the SNP data stored in BacWGSTdb are prepared in this way.

EXAMPLE

The following is an example of how to use BacWGSTdb. *A. baumannii* has emerged worldwide as an important nosocomial pathogen due to its global occurrence and the ability to develop antimicrobial resistance. Clonal dissemination is characteristic of this important bacterial pathogen as revealed by previous studies (25–30). Currently, there are two MLST schemes available for *A. baumannii*, namely MLST-OD (associated with Oxford Database, http://pubmlst.org) and MLST-IP scheme (developed by Institute Pasteur, http://bigsdb.web.pasteur.fr). The former has a higher resolution and the latter is relatively more conservative. In the year 2014, an outbreak of bacteremia caused by *A. baumannii* was detected in a tertiary hospital in Hangzhou, China. We therefore selected one isolate (ABMDR55) for WGS using Illumina Miseq sequencer and the raw reads were assembled into contigs by using CLC Genomics Workbench 8.0 software. Then we analyzed the draft genome by the tool *Typing & Tracking* in BacWGSTdb. The resulting page confirmed that ABMDR55 belonged to $ST208^{OD}/ST2^{IP}$ according to the two MLST schemes and its appropriate reference genome was ACICU ($ST437^{OD}/ST2^{IP}$). The strains closely related to ABMDR55 in the database were all isolated from different Chinese cities and different time periods. The close relationship among these isolates indicated

(A)

Isolate	PubMLST	Pasteur_ST	Host	Disease	Isolation Source	Country	State Province	Collection Year	Accession No.
ABUH304352	208	2	Human	Pneumonia	Bodily Fluid	USA	Ohio	2009	JWSH02
ABUH319118	208	2	Human	Pneumonia	Bodily Fluid	USA	Ohio	2008	JWRW02
ABUH42783	208	2	Human	Wound infection	Tissues	USA	Ohio	2007	JWRU02
ABUH4837	208	2	Human	Bloodstream infection	Bodily Fluid	USA	Ohio	2007	JWRT02
ABUH497144	208	2	Human	Pneumonia	Bodily Fluid	USA	Ohio	2008	JWRY02
ABUH504227	208	2	Human	Pneumonia	Bodily Fluid	USA	Ohio	2008	JWRZ02

20 Entries Per Page Displaying Page 1 of 34

Functionality works without query SNP file: Download isolate information | Draw Phylogenetic tree | Download Newick-formatted tree file | Reset

Functionality works with query SNP file (.vcf format): Browse··· Output_SNP.txt | Draw phylogenetic tree | Download Newick-formatted tree file | Reset

(B)

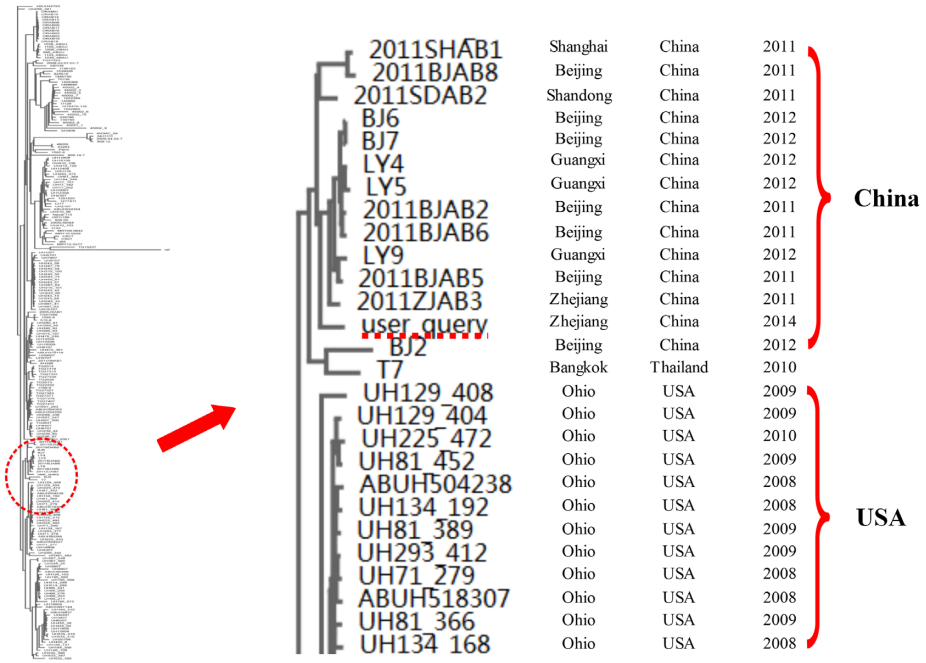

Figure 4. Usage example of Browse. Panel (**A**), a snapshot of isolate information in Browse Page. When users want to incorporate their query SNP data into the phylogenetic analysis, the uploaded SNP file should follow the same reference genome to the selected isolates. Panel (**B**), a phylogenetic tree based on the SNP data, which contains all ST208^{OD} isolates and user query in this case.

they probably belonged to the same clone which had been widely disseminated from a wide spatial and temporal range in China. The SNP file between ABMDR55 and ACICU was downloaded from the resulting page for further analysis. The entire analysis process took <30 s (Figure 3).

Then we went to the BROWSE page for obtaining the clinical information of isolates that were close to ABMDR55 (Figure 4). In this case, ST208^{OD} is very likely to be a pandemic lineage since a total of 240 ST208^{OD} *A. baumannii* genomic sequence data were compiled in BacWGSTdb. The strains belonging to ST208^{OD} were isolated from different countries, such as Spain, Denmark, Czech, Iraq, Thailand, China, Japan and USA. The earliest strain was isolated in the year 2002, and the most recent one was in 2014. We selected all of the ST208^{OD} isolates and meanwhile uploaded the SNP file of ABMDR55 to perform the phylogenetic analysis. Generating the phylogenetic tree took <15 s. According to the derived NJ tree, ABMDR55 and the Chinese ST208^{OD} isolates were grouped into an independent branch (Figure 4), which was consistent with a wide clonal dissemination of *A. baumannii* in China (27).

CONCLUDING REMARKS AND PERSPECTIVES

In light of the rising threats of antimicrobial resistance and emerging virulence among bacterial pathogens, BacWGSTdb represents a rapid and convenient tool for monitoring the emergence or dissemination of new clones and also for global collaboration on the molecular epidemiological investigation of medically important bacterial

pathogens. BacWGSTdb will continue to improve, and additional features for analyzing WGS data are also under development.

ACKNOWLEDGEMENT

We thank Dr. Cheng-Hsun Chiu and Dr. Huan Chen for their valuable suggestions on construction of BacWGSTdb.

FUNDING

National Natural Science Foundation of China [81201248, 81401698]. Funding for open access charge: National Natural Science Foundation of China [81201248, 81401698].
Conflict of interest statement. None declared.

REFERENCES

1. Maiden,M.C., Bygraves,J.A., Feil,E., Morelli,G., Russell,J.E., Urwin,R., Zhang,Q., Zhou,J., Zurth,K., Caugant,D.A. *et al.* (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 3140–3145.
2. Perez-Losada,M., Cabezas,P., Castro-Nallar,E. and Crandall,K.A. (2013) Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infect. Genet. Evol.*, **16**, 38–53.
3. Aanensen,D.M. and Spratt,B.G. (2005) The multilocus sequence typing network: mlst.net. *Nucleic Acids Res.*, **33**, W728–W733.
4. Maiden,M.C. (2006) Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.*, **60**, 561–588.
5. Forde,B.M. and O'Toole,P.W. (2013) Next-generation sequencing technologies and their impact on microbial genomics. *Brief. Funct. Genomics*, **12**, 440–453.
6. Sabat,A.J., Budimir,A., Nashev,D., Sa-Leao,R., van Dijl,J., Laurent,F., Grundmann,H. and Friedrich,A.W. (2013) Overview of molecular typing methods for outbreak detection and epidemiological surveillance. *Euro. Surveill.*, **18**, 20380.
7. Bertelli,C. and Greub,G. (2013) Rapid bacterial genome sequencing: methods and applications in clinical microbiology. *Clin. Microbiol. Infect.*, **19**, 803–813.
8. Torok,M.E. and Peacock,S.J. (2012) Rapid whole-genome sequencing of bacterial pathogens in the clinical microbiology laboratory–pipe dream or reality? *J. Antimicrob. Chemother.*, **67**, 2307–2308.
9. Didelot,X., Bowden,R., Wilson,D.J., Peto,T.E. and Crook,D.W. (2012) Transforming clinical microbiology with bacterial genome sequencing. *Nat. Rev. Genet.*, **13**, 601–612.
10. Kodama,Y., Shumway,M., Leinonen,R. and International Nucleotide Sequence Database Collaboration. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
11. Silvester,N., Alako,B., Amid,C., Cerdeno-Tarraga,A., Cleland,I., Gibson,R., Goodgame,N., Ten Hoopen,P., Kay,S., Leinonen,R. *et al.* (2015) Content discovery and retrieval services at the European Nucleotide Archive. *Nucleic Acids Res.*, **43**, D23–D29.
12. Maiden,M.C., Jansen van Rensburg,M.J., Bray,J.E., Earle,S.G., Ford,S.A., Jolley,K.A. and McCarthy,N.D. (2013) MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.*, **11**, 728–736.
13. Jolley,K.A. and Maiden,M.C. (2014) Using multilocus sequence typing to study bacterial variation: prospects in the genomic era. *Future Microbiol.*, **9**, 623–630.
14. Jolley,K.A. and Maiden,M.C. (2013) Automated extraction of typing information for bacterial pathogens from whole genome sequence data: Neisseria meningitidis as an exemplar. *Euro. Surveill.*, **18**, 20379.
15. Jolley,K.A. and Maiden,M.C. (2010) BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinform.*, **11**, 595.
16. Sheppard,S.K., Jolley,K.A. and Maiden,M.C. (2012) A gene-by-gene approach to bacterial population genomics: whole genome MLST of campylobacter. *Genes (Basel)*, **3**, 261–277.
17. Tamura,K., Nei,M. and Kumar,S. (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. U.S.A.*, **101**, 11030–11035.
18. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
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. 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.
21. Sheneman,L., Evans,J. and Foster,J.A. (2006) Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics*, **22**, 2823–2824.
22. Pethica,R., Barker,G., Kovacs,T. and Gough,J. (2010) TreeVector: scalable, interactive, phylogenetic trees for the web. *PLoS One*, **5**, e8934.
23. Benson,D.A., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2015) GenBank. *Nucleic Acids Res.*, **43**, D30–D35.
24. Wattam,A.R., Abraham,D., Dalay,O., Disz,T.L., Driscoll,T., Gabbard,J.L., Gillespie,J.J., Gough,R., Hix,D., Kenyon,R. *et al.* (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res.*, **42**, D581–D591.
25. Adams-Haduch,J.M., Onuoha,E.O., Bogdanovich,T., Tian,G.B., Marschall,J., Urban,C.M., Spellberg,B.J., Rhee,D., Halstead,D.C., Pasculle,A.W. *et al.* (2011) Molecular epidemiology of carbapenem-nonsusceptible Acinetobacter baumannii in the United States. *J. Clin. Microbiol.*, **49**, 3849–3854.
26. Wright,M.S., Haft,D.H., Harkins,D.M., Perez,F., Hujer,K.M., Bajaksouzian,S., Benard,M.F., Jacobs,M.R., Bonomo,R.A. and Adams,M.D. (2014) New insights into dissemination and variation of the health care-associated pathogen Acinetobacter baumannii from genomic analysis. *Mbio*, **5**, doi:10.1128/mBio.00963-13.
27. Ruan,Z., Chen,Y., Jiang,Y., Zhou,H., Zhou,Z., Fu,Y., Wang,H., Wang,Y. and Yu,Y. (2013) Wide distribution of CC92 carbapenem-resistant and OXA-23-producing Acinetobacter baumannii in multiple provinces of China. *Int. J. Antimicrob. Agents*, **42**, 322–328.
28. Lee,H.Y., Chen,C.L., Wu,S.R., Huang,C.W. and Chiu,C.H. (2014) Risk factors and outcome analysis of acinetobacter baumannii complex bacteremia in critical patients. *Crit. Care Med.*, **42**, 1081–1088.
29. Wang,N., Ozer,E.A., Mandel,M.J. and Hauser,A.R. (2014) Genome-wide identification of Acinetobacter baumannii genes necessary for persistence in the lung. *Mbio*, **5**, doi:10.1128/mBio.01163-14.
30. Wen,H., Wang,K., Liu,Y., Tay,M., Lauro,F.M., Huang,H., Wu,H., Liang,H., Ding,Y., Givskov,M. *et al.* (2014) Population dynamics of an Acinetobacter baumannii clonal complex during colonization of patients. *J. Clin. Microbiol.*, **52**, 3200–3208.