Published online 5 November 2018 *Nucleic Acids Research, 2019, Vol. 47, Database issue* ***D687–D692***
doi: 10.1093/nar/gky1080

VFDB 2019: a comparative pathogenomic platform with an interactive web interface

Bo Liu, Dandan Zheng, Qi Jin, Lihong Chen* and Jian Yang*

MOH Key Laboratory of Systems Biology of Pathogens, Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100176, China

Received September 12, 2018; Revised October 17, 2018; Editorial Decision October 18, 2018; Accepted October 19, 2018

ABSTRACT

The virulence factor database (VFDB, http://www.mgc.ac.cn/VFs/) is devoted to providing the scientific community with a comprehensive warehouse and online platform for deciphering bacterial pathogenesis. The various combinations, organizations and expressions of virulence factors (VFs) are responsible for the diverse clinical symptoms of pathogen infections. Currently, whole-genome sequencing is widely used to decode potential novel or variant pathogens both in emergent outbreaks and in routine clinical practice. However, the efficient characterization of pathogenomic compositions remains a challenge for microbiologists or physicians with limited bioinformatics skills. Therefore, we introduced to VFDB an integrated and automatic pipeline, VFanalyzer, to systematically identify known/potential VFs in complete/draft bacterial genomes. VFanalyzer first constructs orthologous groups within the query genome and preanalyzed reference genomes from VFDB to avoid potential false positives due to paralogs. Then, it conducts iterative and exhaustive sequence similarity searches among the hierarchical prebuilt datasets of VFDB to accurately identify potential untypical/strain-specific VFs. Finally, via a context-based data refinement process for VFs encoded by gene clusters, VFanalyzer can achieve relatively high specificity and sensitivity without manual curation. In addition, a thoroughly optimized interactive web interface is introduced to present VFanalyzer reports in comparative pathogenomic style for easy online analysis.

INTRODUCTION

The emergence of new pathotypes of pathogenic bacteria, often with drug-resistant phenotypes, poses a significant threat to public health worldwide despite our array of antibiotics. The pathogenesis of bacterial pathogens is determined by the virulence factors (VFs) that enable them to cause infection. The potential horizontal transfer of VFs between different strains or species of bacteria makes the emergence of new pathotypes of bacteria almost inevitable. Comprehensive characterization of the VFs carried by the new pathotypes of emerging bacterial pathogens is critical for the effective prevention and control of infectious diseases.

Whole-genome sequencing (WGS) is an effective method for identifying and comprehensively characterizing infectious agents during outbreaks. For example, the year 2011 witnessed the outbreak of hemolytic-uremic syndrome caused by an emerging strain of *Escherichia coli* O104:H4 in Europe. Two groups independently determined the bacterial genome and characterized the key VFs of the new pathotype of *E. coli*, which likely originated from an enteroaggregative *E. coli* ancestor by acquiring the Shiga-toxin encoding phage from enterohemorrhagic *E. coli* donor strains via horizontal gene transfer (1,2). The success of the fast determination and characterization of the emerging *E. coli* strain illustrated the power of bacterial WGS followed by comparative pathogenomic analysis for efficient control of emerging infectious diseases.

The virulence factor database (VFDB, http://www.mgc.ac.cn/VFs/) aims to provide up-to-date knowledge of VFs from various bacterial pathogens and serves as a comprehensive warehouse of bacterial pathogenesis knowledge for the scientific community (3). Since the introduction of comparative pathogenomic approaches to the database in 2008, VFDB has further provided an online platform for exploring the diversity of bacterial genomes in terms of virulence by highlighting common and species- or strain-specific VFs (4). Nevertheless, the originally designed comparative pathogenomic results still depend on manual inspections by database curators. Therefore, it is difficult to fulfill the analytical demand resulting from the dramatically increasing number of bacterial genomes contributed by the scientific community in recent years.

As a consequence, we recently developed a comprehensive pipeline, named VFanalyzer, for the automatic analysis

*To whom correspondence should be addressed. Tel: +86 10 6787 5146; Fax: +86 10 6787 5146; Email: yangj@ipbcams.ac.cn
Correspondence may also be addressed to Lihong Chen. Email: chenlh@ipbcams.ac.cn

of known/potential VFs in given complete/draft bacterial genomes. Beyond commonly used simple BLAST searches, VFanalyzer leverages the well-curated datasets of VFDB and a comparative pathogenomic strategy to accomplish the efficient identification of bacterial VFs in a curation-free manner. Moreover, the JavaScript-enhanced user interface of VFDB was thoroughly optimized to offer an improved user experience for all contents of the database, including VFanalyzer.

DATABASE UPDATES

VFanalyzer: a comparative pathogenomics-based VF analysis pipeline

The conventional method for the identification of VFs in bacterial genomes usually relies on sequence similarity searches (such as BLAST) among datasets of known VFs. However, this solution is far from complete due to the well-known dilemma in genome annotation that any blanket rule relies on fixed criteria for propagating functional annotation may result in missed annotations (by stringent cutoffs) or misannotations (by loose cutoffs) (5). In addition, many bacterial VFs are encoded by multiple components (i.e. genes) within a genomic region, such as secretion systems. The accurate identification of such gene clusters should depend not only on sequence similarities but also on the genomic context of each component. Therefore, we implement the VFanalyzer pipeline with multiple analytical processes to enable the discovery of highly divergent VFs with low sequence similarities while avoiding potential false positives due to close paralogs. The overall data-processing procedure of VFanalyzer is shown in Figure 1.

Whole-genome ortholog identification. Homologs that evolved from a common ancestor by speciation rather than duplication events are called orthologs and are likely to retain identical functions over evolutionary time (6). Thus, the identification of orthologous groups (OGs) between multiple genomes is a valuable tool for the functional annotation of conserved genes. To take full advantage of comparative pathogenomics, VFanalyzer allows only complete or nearly complete draft bacterial genomes as initial queries: open data available from public domains, private data as raw FASTA sequence(s), preannotated genomes in GenBank format or predicted proteins (Figure 1). For queries by raw genome sequence(s), the software GLIMMER3 (7) is applied in self-training mode to predict protein-coding genes in the given genome prior to follow-up analysis. After necessary validation of the query data, the software OrthoMCL (8) is employed to identify OGs with default parameters among the query genome and the preanalyzed reference genome(s) of the same genus available from VFDB. Genes of the query genome that are explicitly assigned to OGs shared with any reference genome are tagged as potential VF-related genes or non-VF genes based on the annotation of individual orthologs from the reference genome. Untagged genes are then collected for further analysis.

Hierarchical and iterative homolog searches. The predefined reference genome(s) of each genus are selected representatives carrying known typical VFs of each pathogen (usually the type strains). However, the number of references was kept to a minimum for efficiency, since the identification of OGs includes a time-consuming all-to-all BLAST procedure. Thus, untypical or strain-specific VFs are expected to be missed by the previous step. The untagged collection of genes is therefore subjected to an iterative BLAST search against the hierarchical VF datasets available from VFDB: (i) the core dataset of experimentally verified VFs from the same genus as the query genome; (ii) the full dataset of predicted VFs from the same genus as the query genome; and iii) the full dataset of known/predicted VFs from other genera. A relatively strict cutoff is applied to filter the output of each iteration, and only genes outside the filtering criteria are collected for the next cycle or follow-up steps (Figure 1). Although the majority of bacterial VFs are generally pathogen-specific in terms of sequence similarities, intergenus comparative analysis can indeed help to identify previously undiscovered VFs. For instance, the discovery of the *Neisseria* ADP-ribosylating enzyme was initially based on its distant homology with *E. coli* heat-labile enterotoxin and cholera toxin (9). Therefore, the extended sequence similarity search among the full dataset of VFDB is valuable for revealing potentially overlooked VFs. Nevertheless, intergenus predicted VFs could also be false positive results due to the diversity of bacterial pathogenesis; such results require further careful experimental validation. Regardless, the iterative procedures based on hierarchical datasets can significantly improve the BLAST-based sequence similarity searches in terms of both specificity and sensitivity.

For highly divergent proteins, sequence similarity-based BLAST searches may fail to recognize distinct homologs. Conserved protein domains are widely used by the scientific community to identify conserved biological functions among different unrelated species (10). To enable protein domain searches for divergent VFs, we leveraged the comparative pathogenomic datasets of VFDB to generate conserved domain profiles for each component of all VFs using hidden Markov models implemented in hmmbuild from the HMMER3 package (11). Thus, the untagged set of genes produced in the previous step is further screened for potential conserved protein domains by the hmmsearch program. To exclude potential false positive matches, a stringent empirical cutoff is applied to the enrollment of valid results.

Postsearch data validation and recovery. Despite the aforementioned careful, stepwise, exhaustive screening of known/potential VFs in the query genome, possible missed VFs (i.e. false negatives) or misclassified VFs (i.e. false positives) may be inevitable in certain cases due to the intrinsic features of bacterial genomes. Paralogs resulting from duplication events are sometimes difficult to clearly distinguish from true orthologs based solely on sequence similarities. Some components of complex VFs could be highly variable, which will result in failed discovery with common criteria. Nevertheless, possibly due to coregulation for the fine assembly of the complex architectures, many VFs with multiple components are encoded by gene cluster(s). Therefore, the genomic context of each predicted VF component

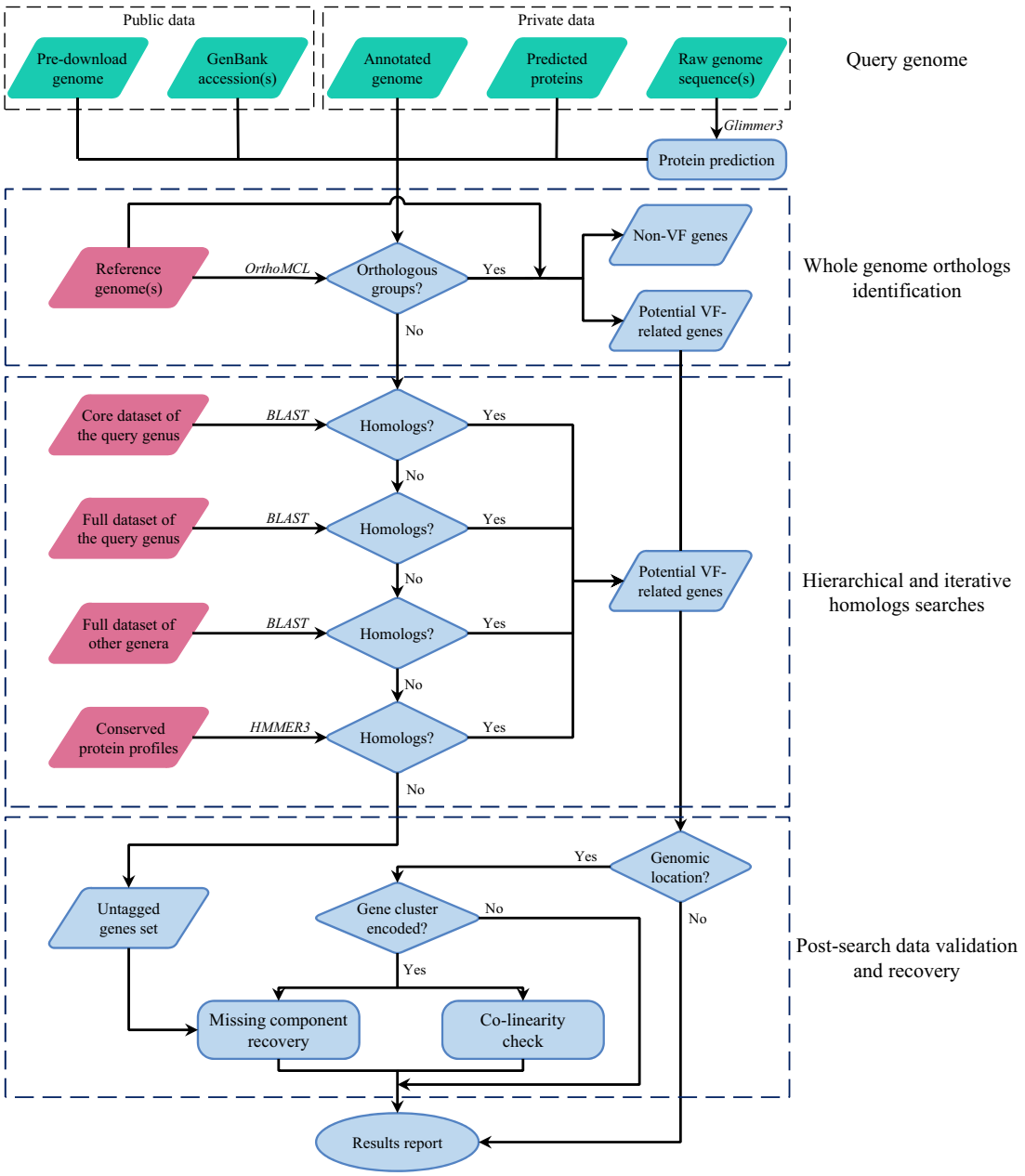

Figure 1. The overall workflow of VFanalyzer data analysis processes. The green parallelograms are user-defined input data, and the red ones are curated datasets from the VFDB database.

could provide valuable information for accurate bacterial VF identification.

In an attempt to increase the overall specificity and sensitivity, VFanalyzer includes two additional postsearch processes for queries with genomic location information available for each gene (i.e. not applicable to queries using predicted proteins) (Figure 1). The first process is a collinearity check for VFs expected to be encoded by gene cluster(s) based on the reference genomes. For multiple homologous copies of each component, the genomic context rather than sequence similarity is the primary criterion to exclude potential false positive copies. The second process is a missing component recovery attempt for clustered genes with missing components to decrease potential false negatives. By using a deliberately loosened sequence similarity cutoff within the specific genomic location defined by the identified upstream and downstream components of the VFs, this process is expected to be able to retrieve potential highly divergent components to improve the sensitivity of VFanalyzer while effectively suppressing the possible false positives from other genomic regions caused by the loosened criteria.

Results retrieval and data presentation. Upon the successful submission of the user-defined VFanalyzer run, a unique job ID is assigned, which can be used to check the progress of the individual query or to retrieve the final results from

the VFDB website at any time within four weeks. We have recruited an independent powerful Linux workstation with 56 CPU cores and 512 GB of memory dedicated solely to the computing jobs of VFanalyzer. However, due to the inherent complexity of the analysis pipeline, a complete VFanalyzer job still takes several to dozens of minutes, depending on the genus and genome size of the query data. Instead of letting users wait for the running jobs, an email notice is sent to the user-defined address once the final results are ready for user inspection.

The VFanalyzer report is presented in a concise table with comparative pathogenomic compositions, as previously described (4). Each column represents a single genome, whereas each row represents a VF component grouped by functional categories. The IDs of VF-related genes from individual genomes are given in respective cells, with hyperlinks to a popup window for inspection of the gene details. The aforementioned intergenus recognized VFs are marked with the original genus of reference to remind users to pay particular attention to the biological interpretation. In principle, the comparison table attempts to present all genomes involved in the analysis, including the query (highlighted in bold) and all related preanalyzed reference genomes from VFDB. To avoid overwhelming the user's browser by a large amount of genomic data for the comparative display, only up to 20 genomes are displayed by default for brevity. More/fewer genomes can be customized later, either for a full comparison view by users with good network speed and local CPU power or for a specialized view of only a subset of genomes of specific interest. All tables can be easily downloaded as Excel files for further offline analysis.

Enhanced JavaScript-rich web interface

JavaScript is one of the modern technologies of the World Wide Web that enables interactive web applications. Since a previous release in 2012, a cross-browser JavaScript library, ExtJS, was used to implement the data presentation of partial VFDB contents (12). By the last release, an alternative JavaScript-based interface was further introduced to present the full contents of the database interactively (13). However, it is now unstable and dysfunctional due to intrinsic bugs in the obsolete version of the JavaScript library. We therefore completely redesigned and rewrote the JavaScript-rich web interface of VFDB with an upgraded version of the ExtJS library to make all web applications fully functional and highly stable for efficient data presentation. Additional features are available for better big data presentation, such as the popular D3-tree to display hierarchical data structures dynamically and efficiently. The new web interface can not only present all contents of VFDB in a uniformed desktop-like web page but can also effectively organize a large amount of tabular data within an Excel-style grid to provide database users with much better experiences than traditional web pages.

For example, the aforementioned grid for the presentation of VFanalyzer results is fully sortable and filterable by a single click on the column title, and each column can also be moved and scaled (or hidden) by dragging and dropping on the title. In cases where many genomes are presented for comparison and cannot be easily handled by the client screen, the locked view is particularly helpful for focusing on the genome(s) of interest by freezing the column(s) to the left. In addition, multiple groupings of VF categories in the form of a tree structure can make the entire comparison grid better organized and easier to understand than a plain 2D table. More online data analysis tools are available from the bottom toolbar of the VFanalyzer results grid, such as tools to transfer the full table to a simplified grid in text and symbolic or schematic mode for an easy overview. Additional web applications will be continuously developed within the framework of the new interface since it is highly extendable and scalable. Moreover, the generic interface design is readily reusable for future studies by our group or others (14–16).

Although all modern web browsers support JavaScript, some users may disable it due to specific concerns. Additionally, our new functional web applications based on JavaScript may require considerably more client resources in terms of network speed, CPU and memory than the original VFDB web pages. Therefore, to keep the database readily accessible by all users worldwide, we will continue to maintain the traditional web pages in the future. However, we highly recommend that VFDB users switch to the new web interface if possible, since novel features such as VFanalyzer will be available only from the new interface with full JavaScript support.

Pathogens recently included in VFDB

Since the last release in 2016, two additional genera of pathogens, *Francisella* and *Klebsiella*, have been formally introduced into VFDB. *F. tularensis* is the causative agent of the life-threatening zoonotic disease tularemia (rabbit fever) and is a tier I priority pathogen due to its extreme virulence and ease of aerosol dissemination (17). *Klebsiella* species are ubiquitous in nature and routinely found in the human mouth, skin and intestines as normal microbiota. However, *K. pneumonia* is an important opportunistic human pathogen and a leading cause of both community and nosocomial infections (18). To date, 32 genera of pathogens with medical importance are formally included in VFDB (with full information available), and 42 additional genera are also partially included in the database for the intergenus comparison of VFs, as previously described (12) (a full list of pathogens is available from the main page of the interactive interface at http://www.mgc.ac.cn/cgi-bin/VFs/v5/main.cgi). VFDB is dedicated to providing comprehensive information on bacterial VFs and will continue to cover additional genera of medically important pathogens in future updates.

DISCUSSION

The rapid expansion of second-generation sequencing technology (such as Illumina) has dramatically decreased the cost and time for determining a bacterial genome. However, it usually produces only a draft rather than a complete bacterial genome. In contrast, the recently available third-generation sequencing technologies (such as Pacific Biosciences and Oxford Nanopore), which can produce very

long sequencing reads, are capable of generating complete or nearly complete bacterial genomes without manual assistance. Therefore, *de novo* sequencing instead of resequencing is now readily feasible for the scientific community to decode potential novel or variant pathogens not only in cases of outbreaks but also in routine clinical practice. However, it remains a challenge for microbiologists or physicians with limited bioinformatics skills to efficiently define and extract biologically relevant information from volumes of genomic data. Based on the VFDB database, VFanalyzer was therefore developed to meet this demand by providing an automatic and comprehensive platform for accurate bacterial VF identification.

Several other online resources for bacterial genomic analysis also provide VF screening services, such as PATRIC (19), VRprofile (20) and VirulenceFinder (21). Although generally based on VFDB datasets, the majority of them depend solely on BLAST searches. Owning to the inherent complexity of bacterial VFs and the genetic diversity of bacterial genomes, accurate *in silico* identification of VFs is a challenging task. VFanalyzer circumvents this difficulty by combining whole-genome ortholog grouping, hierarchical and iterative homolog searching and context-based data refinement. Based on our own testing datasets, after careful configuration optimization, VFanalyzer achieved a relatively high overall accuracy that is comparable to human curation (data not shown). In addition, all of the current tools produce a list of predicted VFs in the query genome only, whereas none of them provide comparative results across query and reference genomes for further in-depth analysis. Indeed, the comparative pathogenomic presentation of VFanalyzer results is particularly useful for highlighting the divergence and conservation of bacterial VFs encoded by different bacterial genomes. Nevertheless, users should always pay attention to the biological interpretation of the VFanalyzer outputs, as any *in silico* result requires further experimental validation for verification.

The current VFanalyzer platform has some limitations. First, VFanalyzer is designed to analyze only genomes from the aforementioned 32 genera of pathogens, since the internal pipeline requires the comparative pathogenomic datasets of VFDB. The accumulation of bacterial VFs depends on in-depth molecular microbiological studies on bacterial pathogenesis, which are generally focused on human pathogens with medical importance, such as those currently covered by VFDB. Although it is possible for genomes of other genera to force a VFanalyzer run using other closely related, well-studied pathogens as references, this abuse of VFanalyzer is unsupported and discouraged, as it may yield unexpected results outside the original schema. Second, VFanalyzer accepts only complete or nearly complete draft genomes to produce reliable results, since it will try to integrate the genomic context for final data refinement. Severely fragmented genomes will obfuscate the result improvement process and, as a consequence, decrease the overall accuracy of VFanalyzer. For raw WGS reads or metagenomics data, an additional barrier is the efficient network transfer and server handling of the big data (including both storage and analysis). Further investigations are required to overcome these problems in the future. Finally, VFanalyzer only systematically screens the query genome to identify potential VFs based primarily on homologies with known VFs in our database. Therefore, the pipeline reports only homologs of known VFs (whether closely related or divergent) rather than the *de novo* prediction of any novel VFs. Some other tools are already available to recognize possible new VFs using machine-learning approaches, such as VICMpred (22), VirulentPred (23), EffectiveT3 (24) and T4EffPred (25). However, the current methods are usually limited to only one or several subtype(s) of VFs or subclass(es) of pathogens. Therefore, the development of a comprehensive, homology-independent bacterial VF prediction method will be an interesting focus for future studies.

FUNDING

National Key Research and Development Program from the Ministry of Science and Technology of China [2016YFC1202404 to J.Y.]; National Basic Research Program from the Ministry of Science and Technology of China [2015CB554204 to L.C.]; CAMS Innovation Fund for Medical Sciences [2017-I2M-3-017 to J.Y.]; National Scientific Data Sharing Platform for Population and Health. Funding for open access charge: National Key Research and Development Program.
Conflict of interest statement. None declared.

REFERENCES

1. Rasko,D.A., Webster,D.R., Sahl,J.W., Bashir,A., Boisen,N., Scheutz,F., Paxinos,E.E., Sebra,R., Chin,C.S., Iliopoulos,D. *et al.* (2011) Origins of the E. coli strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N. Engl. J. Med.*, **365**, 709–717.
2. Rohde,H., Qin,J., Cui,Y., Li,D., Loman,N.J., Hentschke,M., Chen,W., Pu,F., Peng,Y., Li,J. *et al.* (2011) Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *N. Engl. J. Med.*, **365**, 718–724.
3. Chen,L., Yang,J., Yu,J., Yao,Z., Sun,L., Shen,Y. and Jin,Q. (2005) VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.*, **33**, D325–D328.
4. Yang,J., Chen,L., Sun,L., Yu,J. and Jin,Q. (2008) VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. *Nucleic Acids Res.*, **36**, D539–D542.
5. Haft,D.H. (2015) Using comparative genomics to drive new discoveries in microbiology. *Curr. Opin. Microbiol.*, **23**, 189–196.
6. Fitch,W.M. (2000) Homology a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.
7. Delcher,A.L., Bratke,K.A., Powers,E.C. and Salzberg,S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.
8. Li,L., Stoeckert,C.J. Jr and Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
9. Masignani,V., Balducci,E., Di Marcello,F., Savino,S., Serruto,D., Veggi,D., Bambini,S., Scarselli,M., Arico,B., Comanducci,M. *et al.* (2003) NarE: a novel ADP-ribosyltransferase from Neisseria meningitidis. *Mol. Microbiol.*, **50**, 1055–1067.
10. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
11. Mistry,J., Finn,R.D., Eddy,S.R., Bateman,A. and Punta,M. (2013) Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.*, **41**, e121.
12. Chen,L., Xiong,Z., Sun,L., Yang,J. and Jin,Q. (2012) VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. *Nucleic Acids Res.*, **40**, D641–D645.

13. Chen,L., Zheng,D., Liu,B., Yang,J. and Jin,Q. (2016) VFDB 2016: hierarchical and refined dataset for big data analysis–10 years on. *Nucleic Acids Res.*, **44**, D694–D697.
14. Chen,L., Liu,B., Yang,J. and Jin,Q. (2014) DBatVir: the database of bat-associated viruses. *Database*, **2014**, bau021.
15. Chen,L., Liu,B., Wu,Z., Jin,Q. and Yang,J. (2017) DRodVir: a resource for exploring the virome diversity in rodents. *J. Genet. Genomics*, **44**, 259–264.
16. Chen,Y., Shi,M., Cheng,Y., Zhang,W., Tang,Q. and Xia,X.Q. (2018) FVD: The fish-associated virus database. *Infect. Genet. Evol.*, **58**, 23–26.
17. Steiner,D.J., Furuya,Y. and Metzger,D.W. (2014) Host-pathogen interactions and immune evasion strategies in Francisella tularensis pathogenicity. *Infect. Drug Resistance*, **7**, 239–251.
18. Podschun,R. and Ullmann,U. (1998) Klebsiella spp. as nosocomial pathogens: epidemiology, taxonomy, typing methods, and pathogenicity factors. *Clin. Microbiol. Rev.*, **11**, 589–603.
19. Wattam,A.R., Davis,J.J., Assaf,R., Boisvert,S., Brettin,T., Bun,C., Conrad,N., Dietrich,E.M., Disz,T., Gabbard,J.L. *et al.* (2017) Improvements to patric, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Res.*, **45**, D535–D542.
20. Li,J., Tai,C., Deng,Z., Zhong,W., He,Y. and Ou,H.Y. (2018) VRprofile: gene-cluster-detection-based profiling of virulence and antibiotic resistance traits encoded within genome sequences of pathogenic bacteria. *Brief. Bioinform.*, **19**, 566–574.
21. Joensen,K.G., Scheutz,F., Lund,O., Hasman,H., Kaas,R.S., Nielsen,E.M. and Aarestrup,F.M. (2014) Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. *J. Clin. Microbiol.*, **52**, 1501–1510.
22. Saha,S. and Raghava,G.P. (2006) VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria using amino acid patterns and composition. *Genomics Proteomics Bioinformatics*, **4**, 42–47.
23. Garg,A. and Gupta,D. (2008) VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. *BMC Bioinformatics*, **9**, 62.
24. Arnold,R., Brandmaier,S., Kleine,F., Tischler,P., Heinz,E., Behrens,S., Niinikoski,A., Mewes,H.W., Horn,M. and Rattei,T. (2009) Sequence-based prediction of type III secreted proteins. *PLoS Pathog.*, **5**, e1000376.
25. Zou,L., Nan,C. and Hu,F. (2013) Accurate prediction of bacterial type IV secreted effectors using amino acid composition and PSSM profiles. *Bioinformatics*, **29**, 3135–3142.