D694–D697 *Nucleic Acids Research, 2016, Vol. 44, Database issue* *Published online 17 November 2015*
doi: 10.1093/nar/gkv1239

VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on

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

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

Received September 15, 2015; Revised November 1, 2015; Accepted November 2, 2015

ABSTRACT

The virulence factor database (VFDB, http://www.mgc.ac.cn/VFs/) is dedicated to providing up-to-date knowledge of virulence factors (VFs) of various bacterial pathogens. Since its inception the VFDB has served as a comprehensive repository of bacterial VFs for over a decade. The exponential growth in the amount of biological data is challenging to the current database in regard to big data analysis. We recently improved two aspects of the infrastructural dataset of VFDB: (i) removed the redundancy introduced by previous releases and generated two hierarchical datasets – one core dataset of experimentally verified VFs only and another full dataset including all known and predicted VFs and (ii) refined the gene annotation of the core dataset with controlled vocabularies. Our efforts enhanced the data quality of the VFDB and promoted the usability of the database in the big data era for the bioinformatic mining of the explosively growing data regarding bacterial VFs.

INTRODUCTION

Bacterial pathogens threaten public health worldwide (1). The success of pathogenic bacteria depend on their ability to deploy virulence factors (VFs) to establish infections, survive in the hostile host environment and as a result cause disease. Elucidating the molecular mechanisms of VFs can improve our understanding of the molecular and cellular basis of bacterial pathogenesis, which is one of the most important and intriguing topics in microbiology. Moreover, a full understanding of the mechanisms of VFs essential to any one pathogen can provide new avenues for developing potential new approaches to combating or preventing disease (2).

The virulence factor database (VFDB) is an integrated online resource for curating information about VFs of bacterial pathogens (3). Since its inception in 2004, VFDB has been dedicated to providing up-to-date knowledge of VFs from various medically significant bacterial pathogens. During the past decade, VFDB has offered the scientific community a comprehensive repository of the broadest and most up-to-date information regarding various bacterial VFs. Furthermore, previous upgrades of VFDB have equipped the database with platforms for intra-genera as well as inter-genera comparative analyses of bacterial VFs (4,5).

Nowadays, next generation sequencing technologies (NGS) are widely used in genomic, transcriptomic and metagenomic studies related to pathogenic bacteria (6–9). A reliable and nonredundant dataset of known reference VFs will add value to the analysis of this NGS data explosion from the perspective of the study of bacterial pathogenicity. We have recently received increasing numbers of feedbacks from the VFDB database users asking for assistance in screening potential VF-related genes from raw preliminary high-throughput NGS data. Unfortunately, issues of data redundancy in the current database and lagging curation of current gene annotations hamper the efficient use of the VFDB dataset for big data mining. In this communication, we first review the progression of VFDB during the past decade to clarify the current barriers for future application. Then, we present the recent efforts to facilitate further big data analysis, including the major improvements in the infrastructure of the VFDB dataset. Furthermore we describe the inclusion of six additional bacterial pathogens into VFDB and a new alternative JavaScript-rich web interface.

DECADE REVIEW

The first release (R1) of VFDB was announced in 2005. VFDB aimed to provide a comprehensive repository of known VFs of various well-characterized, medically significant bacterial pathogens to facilitate future research (3). By manual collection from original research papers and public databases, VFDB R1 was populated with the most up-to-date knowledge of bacterial pathogens (e.g.

[*]To whom correspondence should be addressed. Tel: +86 106 787 7735; Fax: +86 106 787 7736; Email: yangj@ipbcams.ac.cn
Correspondence may also be addressed to Qi Jin. Tel: +86 106 787 7732; Fax: +86 106 787 7736; Email: zdsys@vip.sina.com
[†]These authors contributed equally to this work as first authors.

nomenclature, classification, characteristics of the virulence mechanism and manifestation of the diseases caused) and all VFs that had been demonstrated by experiment (e.g. virulence-associated genes, protein structural features, functions, mechanisms and key literature) (Figure 1). The initial version of VFDB R1 covered only 16 genera of bacteria pathogens, although it continuously expanded in the follow-up releases and now covers 30 genera (see below for details of recently included pathogens). For brevity, it contains only sequence data from one representative strain for each VF. Thus, VFDB R1 constitutes the core of the database.

With the second release (R2) of VFDB in 2008, comparative genomic approaches were introduced into the database to explore the diversity of bacterial genomes in terms of virulence genes and their organization (4). The availability of complete genome sequences of different species/strains of the same genus enabled comparative genomics, which is a powerful tool to describe genome plasticity of a pathogenic genus and investigate the repertoire of VFs in the gene pool. Common as well as species- or strain-specific VFs may be defined this way. Commonly shared VFs likely indicate universal requirements of members of a genus to cause specific aspects of their infection process. Conversely, strain- or species-specific VFs typically determine more limited and variable phenotypes. Based on the original dataset, VFDB R2 attempted to include all homologs of VF-related genes within fully sequenced bacterial genomes (Figure 1). Therefore, VFDB R2 expanded the content of the VFDB in terms of bacterial genomics but as a consequence also generated considerable redundancy in the sequence data. In addition to the experimentally verified VFs collected in the original dataset, VFDB R2 included predicted VFs, derived from sequenced genomes, for comparison purposes (4). It should be noted that the comparative genomic analyses in VFDB R2 are limited to intra-genera comparisons.

In an attempt to maximize the power of the comparative genomics platform in VFDB, the third release (R3) in 2012 included inter-genera comparative analysis of VFs involved in four major categories: (i) host cell attachment and invasion; (ii) bacterial secretion systems and effectors; (iii) toxins and (iv) iron-acquisition systems (5). Via comparisons of the different composition and organization of VFs from various bacterial pathogens, VFDB R3 was dedicated to identifying the common themes (e.g. core components and phylogenetic clades) in bacterial virulence. This was an attempt to summarize the genetic diversity of bacterial VFs and shed new light on the selection forces that shape the evolutionary history of bacterial pathogenesis. In contrast to the previous two releases, which focused on VFs of medically important bacterial pathogens, VFDB R3 included VF homologs present in animal and plant pathogens to better describe bacterial virulence evolution (5). The VFs included in VFDB R3 are derived from 75 genera of bacteria, which greatly expanded the number of pathogens covered by the database (Figure 1). However, it should be noted that the additional pathogens included in R3 have not yet been formally introduced in VFDB as curation of the full metadata, such as general pathogenesis phenotypes, is still in progress.

In summary, over the past decade, in addition to regular data updates, the VFDB underwent integration of a comparative genomics platform and dual major dataset expansions. These continuous upgrades have increased the coverage of the range of VFs presented in VFDB in terms of both pathogens and genomes (Figure 1). However, the upgrades also resulted in considerable data redundancy and a mixture of experimentally confirmed and predicted VFs in the current dataset, which confounds its application to future big data analysis of bacterial VFs.

RECENT IMPROVEMENTS

We reorganized the VFDB infrastructure with an improved database schema and carefully removed data redundancy to meet the requirements of big data analysis. NGS-based genomic, transcriptomic and metagenomic approaches have revolutionized modern microbial research in recent years (10,11). The power of big data boosted data-centric strategies as an important alternative to traditional hypothesis-driven scientific studies. However, given the vast amount of sequencing data generated from NGS-based studies and the limited computing resources available for regular bioinformatic analysis, the reliable and timely analysis of big data remains a challenging task (12,13). Removal of data redundancy is an essential step in handling and mining big data as minimized datasets dramatically speed up bioinformatic analysis. Recently, we have got many requests asking for assistance in using the VFDB database for VFs screening from preliminary NGS data. Therefore, we combined the sequence data from all previous releases and generated two hierarchical, nonredundant datasets (sets A and B) for free download to meet the demands of different applications. Set A is a core dataset that covers genes associated with experimentally verified VFs, which collects only ∼2400 gene sequences currently. The small dataset size makes set A suitable for quick preliminary screening of bacterial VFs with limited computing resource. Nevertheless, confident additions of trusted homologs could also be beneficial to the computational biologists sometimes. So the full dataset (set B) including >30 000 sequences of all genes related to known and predicted VFs in the database is also available for additional in-depth analysis. But users need to pay special attention to further biological interpretation of the results based on set B.

Additionally, we have refined the gene annotation of the core dataset (set A) using a more consistent and controlled vocabulary to facilitate future big data mining using the VFDB. As the original annotation of the VF-related genes available from GenBank was provided by a large number of different groups without restrictions on the vocabulary (14), it is sometimes difficult to link VF-related genes across different pathogens that share similar mechanisms. Moreover, high-quality annotation with intensive curation is invaluable for extracting key information from big data for further biological interpretation (15). Based on the original literature or recent reviews, we polished the annotation of VF-related genes with standardized, authorized and well-defined terms to make the information readily accessible to both humans and computers. Supplementary Table S1 illustrates a comparison of the original annotations from GenBank and the curated annotations in VFDB for genes related to known VFs in *Shigella*. Only annotations

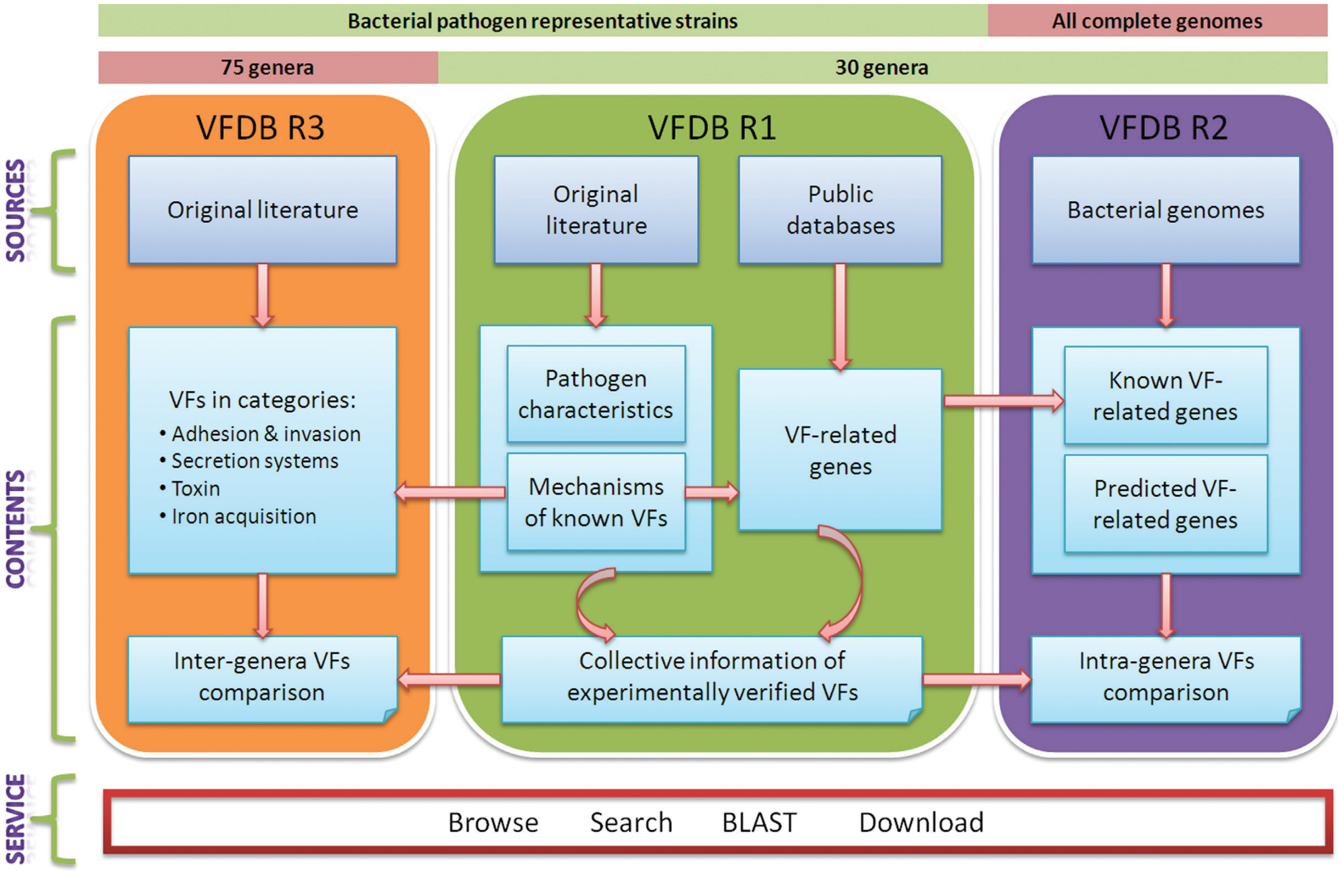

Figure 1. Schematic diagram of the relationship between previous releases of virulence factor database (VFDB). R1 forms the base and core of the database, whereas R2 and R3 provided a comparative genomics platform and expanded the data contents in terms of genomes and pathogens (illustrated by the horizontal bars on the top), respectively.

of genes associated with experimentally verified VFs were involved in this curation process because the true molecular functions and roles in pathogenesis of predicted VF-related genes remain unclear. Nevertheless, curation of the annotation of the core dataset will be included in future routine updates of the VFDB.

Since the latest release in 2012, VFDB has expanded with additional six genera: *Acinetobacter*, *Aeromonas*, *Anaplasma*, *Burkholderia*, *Coxiella* and *Rickettsia*. The most up-to-date statistics of the database (daily update) is available at the status page of VFDB (http://www.mgc.ac.cn/VFs/status.htm). In addition, we recently reconstructed the web page interface of the VFDB with a highly responsive and intuitive user interface using advanced JavaScript programming. The new interface presents collapsible menus, expandable trees and sortable tables, which provide an improved user experience with a look and feel similar to desktop applications. As the new web interface requires JavaScript support from the users' web browser, we will continue to maintain the previous web pages to make the database easily accessible by all users worldwide. The dramatic increase of bacterial genome sequences in the public domain in recent years makes it difficult to handle all complete genomes in the VFDB R2 for VFs comparative analysis. Therefore, we have changed our strategy to collect only selected representative genomes by default. The reference and representative genomes selected by NCBI (http://www.ncbi.nlm.nih.gov/genome/browse/reference/) will be covered in priority (16). Nevertheless, users are welcome to send requests for the inclusion of any publicly available complete genome of specific interest for inclusion in VFDB comparative pathogenomic platform.

DISCUSSION

The exponential growth in the amount of biological data provides both opportunities and challenges for current biomedical research. Recent developments have further improved the data quality of VFDB and enhanced the usability of the database for the bioinformatic mining of the explosively growing data regarding bacterial VFs. Biocuration in the big data era plays an essential role in biological discoveries and specifically in biomedical research. The scientific community needs both robust analytical tools and well-curated reference datasets to gain novel insights and find meaningful information from the ever increasing wealth of sequence data (15). VFDB is dedicated to providing comprehensive and up-to-date knowledge of bacterial VFs to the community to unpick the sophisticated virulence strategies implemented by bacterial pathogens. The development of methods or tools to circumvent current technical hurdles for the identification of potential VFs from large datasets remains the future goal of the VFDB.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Basic Research Program from the Ministry of Science and Technology of China [2015CB554204]; National Science and Technology Major Projects for Major New Drugs Innovation and Development [2013ZX09304101]; Program for Changjiang Scholars and Innovative Research Team in University [IRT13007]; PUMC Youth Fund and the Fundamental Research Funds for the Central Universities [3332013118]. Funding for open access charge: Biologic Medicine Information Center, National Scientific Data Sharing Platform for Population and Health.
Conflict of interest statement. None declared.

REFERENCES

1. van Oosten,M., Hahn,M., Crane,L.M., Pleijhuis,R.G., Francis,K.P., van Dijl,J.M. and van Dam,G.M. (2015) Targeted imaging of bacterial infections: advances, hurdles and hopes. *FEMS Microbiol. Rev.*, **39**, 892–916.
2. Unala,C.M. and Steinert,M. (2014) Microbial peptidyl-prolyl cis/trans isomerases (PPIases): virulence factors and potential alternative drug targets. *Microbiol. Mol. Biol. Rev.*, **78**, 544–571.
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. 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.
6. Bucci,V. and Xavier,J.B. (2014) Towards predictive models of the human gut microbiome. *J. Mol. Biol.*, **426**, 3907–3916.
7. Kay,G.L., Sergeant,M.J., Giuffra,V., Bandiera,P., Milanese,M., Bramanti,B., Bianucci,R. and Pallen,M.J. (2014) Recovery of a medieval Brucella melitensis genome using shotgun metagenomics. *MBio*, **5**, e01337–e01314.
8. Lax,S., Smith,D.P., Hampton-Marcell,J., Owens,S.M., Handley,K.M., Scott,N.M., Gibbons,S.M., Larsen,P., Shogan,B.D., Weiss,S. *et al.* (2014) Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science*, **345**, 1048–1052.
9. Lu,X., Zhang,X.X., Wang,Z., Huang,K., Wang,Y., Liang,W., Tan,Y., Liu,B. and Tang,J. (2015) Bacterial pathogens and community composition in advanced sewage treatment systems revealed by metagenomics analysis based on high-throughput sequencing. *PLoS One*, **10**, e0125549.
10. Buchan,B.W. and Ledeboer,N.A. (2014) Emerging technologies for the clinical microbiology laboratory. *Clin. Microbiol. Rev.*, **27**, 783–822.
11. Falony,G., Vieira-Silva,S. and Raes,J. (2015) Microbiology meets big data: the case of gut microbiota-derived trimethylamine. *Annu. Rev. Microbiol.*, **69**, 305–321.
12. Dunne,W.M. Jr, Westblade,L.F. and Ford,B. (2012) Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. *Eur. J. Clin. Microbiol. Infect. Dis.*, **31**, 1719–1726.
13. Naccache,S.N., Federman,S., Veeraraghavan,N., Zaharia,M., Lee,D., Samayoa,E., Bouquet,J., Greninger,A.L., Luk,K.C., Enge,B. *et al.* (2014) A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome Res.*, **24**, 1180–1192.
14. 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.
15. Howe,D., Costanzo,M., Fey,P., Gojobori,T., Hannick,L., Hide,W., Hill,D.P., Kania,R., Schaeffer,M., St Pierre,S. *et al.* (2008) Big data: the future of biocuration. *Nature*, **455**, 47–50.
16. Tatusova,T., Ciufo,S., Federhen,S., Fedorov,B., McVeigh,R., O'Neill,K., Tolstoy,I. and Zaslavsky,L. (2015) Update on RefSeq microbial genomes resources. *Nucleic Acids Res.*, **43**, D599–D605.