Published online 8 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D641–D645***
doi:10.1093/nar/gkr989

VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors

Lihong Chen, Zhaohui Xiong, Lilian Sun, Jian Yang* and Qi Jin*

State Key Laboratory for Molecular Virology and Genetic Engineering, Institute of Pathogen Biology, Chinese Academy Medical Sciences and Peking Union Medical College, Beijing 100176, China

Received September 15, 2011; Accepted October 17, 2011

ABSTRACT

The virulence factor database (VFDB, http://www.mgc.ac.cn/VFs/) has served as a comprehensive repository of bacterial virulence factors (VFs) for >7 years. Bacterial virulence is an exciting and dynamic field, due to the availability of complete sequences of bacterial genomes and increasing sophisticated technologies for manipulating bacteria and bacterial genomes. The intricacy of virulence mechanisms offers a challenge, and there exists a clear need to decipher the 'language' used by VFs more effectively. In this article, we present the recent major updates of VFDB in an attempt to summarize some of the most important virulence mechanisms by comparing different compositions and organizations of VFs from various bacterial pathogens, identifying core components and phylogenetic clades and shedding new light on the forces that shape the evolutionary history of bacterial pathogenesis. In addition, the 2012 release of VFDB provides an improved user interface.

INTRODUCTION

Bacterial virulence factors (VFs) are fascinating for a number of reasons. First, the ability of successful pathogens to establish infections, produce disease and survive in a hostile environment is provided by a large armamentarium of virulence mechanisms. Elucidating the molecular mechanisms of VFs can improve understanding of the cellular and molecular basis of pathogenesis. Second, many important virulence factors interact with host cells and modulate their functions. Investigating the complex and finely balanced interactions between hosts and pathogens can uncover useful tools for studying normal host cellular processes. Third, a much deeper understanding of the mechanisms of action of VFs will inform new avenues for identifying promising approaches to disease prevention and therapy. Fueled by recent technological innovations in the life sciences, the field of microbial virulence has expanded rapidly over the past decade.

Since its inception in 2004, the virulence factor database (VFDB, http://www.mgc.ac.cn/VFs/) has provided the broadest and most comprehensive up-to-date information regarding experimentally validated bacterial virulence factors (e.g. extracellular products, such as enzymes and toxins and secreted effectors or cell-associated products, such as capsular polysaccharides and outer membrane proteins), and has further explored plasticity in the repertoire of VFs on an intra-genera level since its second release (1,2). To summarize the common themes in bacterial virulence and to reflect the diversity of genomic encoding, structural architecture and functional originality, we recently updated VFDB with an enhanced user interface and new contents dedicated to inter-genera comparative analysis of VFs involved in host cell attachment and invasion, bacterial secretion systems and effectors, toxins, and iron-acquisition systems (Table 1).

DATABASE UPDATES

Data sources and processing

The core dataset of VFDB only covers experimentally demonstrated VFs from 24 genera of medically important bacterial pathogens. Several predicted VFs from complete genomes were also included for comparative analyses in the second release (2), but this information is still far from sufficient for a comprehensive study of the genetic diversity and molecular evolution of VFs. Many VFs found in human pathogens have homologues present in animal or plant pathogens, and, sometimes, even in non-pathogens. Additionally, the genomic sequences encoding most functionally validated VFs are fragmentary, rather than complete genomes in the public domain. Therefore, via exhaustive literature screening and expert review, the

*To whom correspondence should be addressed. Tel: +86 10 6787 7732; Fax: +86 10 6787 7736; Email: zdsys@vip.sina.com
Correspondence may also be addressed to Jian Yang. Tel: +86 10 6787 7735; Fax: +86 10 6787 7736; Email: yangj@ipbcams.ac.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Table 1. Data summary of newly released contents for the diversity and evolution analyses of VFs (as of September 2011)

VF super-family	Number of subclasses	Number of VFs	Number of genera involved	Number of VFs-related genes	Number of related references
Adhesion and invasion	10	429	47	2016	387
Secretion systems	6	217	48	2879	483
Toxin	12	487	42	564	218
Iron acquisition	3	72	30	551	69
Total	31	1205	75	5955	1133

basic information on >1200 VFs was collected from over 1100 original research papers (Table 1). These collected VFs, derived from 75 genera of bacteria, were organized into four super-families and 31 subclasses in VFDB (Table 1). Nevertheless, we do not intend to discuss the biological diversity of certain VFs; therefore, only experimentally verified VFs were included. In addition, if more than one sequence was available for an individual species, only the representative one was collected into the database for the sake of brevity.

The nucleotide and amino-acid sequences of VF-encoding genes and related annotation information were extracted from individual GenBank (3) records using *ad hoc* BioPerl scripts. The conserved domain(s) of each protein were recognized by local Pfam (4) search using the HMMER3 program (http://hmmer.org/), and the related protein structure information was available from the PDB database (5) via batch BLAST search followed by manual curation. Homologue groups were determined by reciprocal BLAST on individual datasets of each subclass, and the results were further curated based on conserved synteny. Next, the MatGAT program (6) and DaliLite server (7) were used to calculate pairwise sequence and structure (if it exists) similarities, respectively, among each group. The T–coffee package (8) was employed to generate multiple alignment for each homologue group. For highly divergent proteins, the segments of respective conserved domain(s) were used instead of full sequences for producing reliable alignments. The ESPript web server (9) was used to render structure information on multiple alignments. The MEGA software (10) was used to build phylogenetic trees based on the multiple alignment of the core component/domain of each subclass of VFs. The overall data-processing procedure is shown in Figure 1.

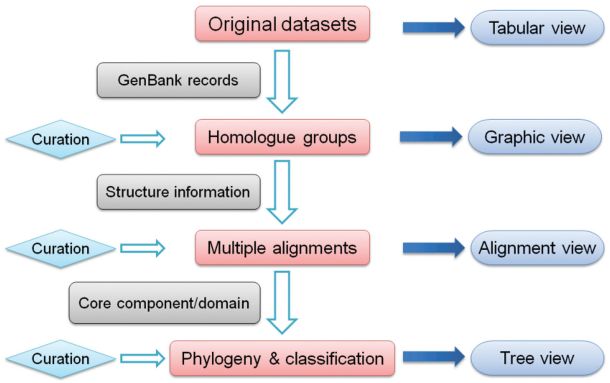

Figure 1. The overall method of data processing and presentation.

Data presentation and web interface

The effective presentation of data is one of the key criteria for any good database to provide users with the most intuitive and easy-to-understand results. The VFDB offers four main styles to visualize the comparative results of each subclass of VFs during different analysis stages (Figure 1). The information gleaned from literature is presented in a concise table (exemplified in the right panel of Figure 2A), which covers the basic data for each VF, such as organism name, taxonomic class, VF name/family, known/proposed function, key component(s) and a direct link to the original literature available in PubMed (11). The linear graphic view is used to display unambiguously the diversity of VFs in terms of genetic composition or genomic organization (Figure 2B). The manually curated multiple alignments (Figure 2C) and phylogenetic tree built from the core component/domain (Figure 2D) are also available to enable users to further analyze the sequence/structure diversity and molecular evolutionary relationships of homologous VFs from various pathogens.

For the 2012 release of the VFDB, we built a more responsive and intuitive user interface with high-performance grids, expandable trees, collapsible menus and tabbed panels using ExtJS (http://www.sencha.com/), which is a cross-browser JavaScript library for building rich internet applications. This library provides users with the look and feel of a desktop application rather than a traditional web page. For example, the aforementioned tables are fully sortable and filterable by a single click on the column title, and each column is also movable and scalable (or hidden) by dragging and dropping on the title. These features that were previously available only in standalone applications will undoubtedly provide the database users with better experiences than before.

The main web interface is vertically divided into two panels: a collapsible menu panel on the left and a tabbed content panel on the right (for example, see Figure 2A). The menu panel provides a tree-like organization of all subclasses of VFs with direct links to each individual page for easy navigation. To maximize the visible region of the content panel, the menu is collapsed into a clickable vertical bar automatically upon page load (for example see Figure 2D). The bottom tool bar of the content panel provides several convenient functional buttons on the left side for easy manipulation of the

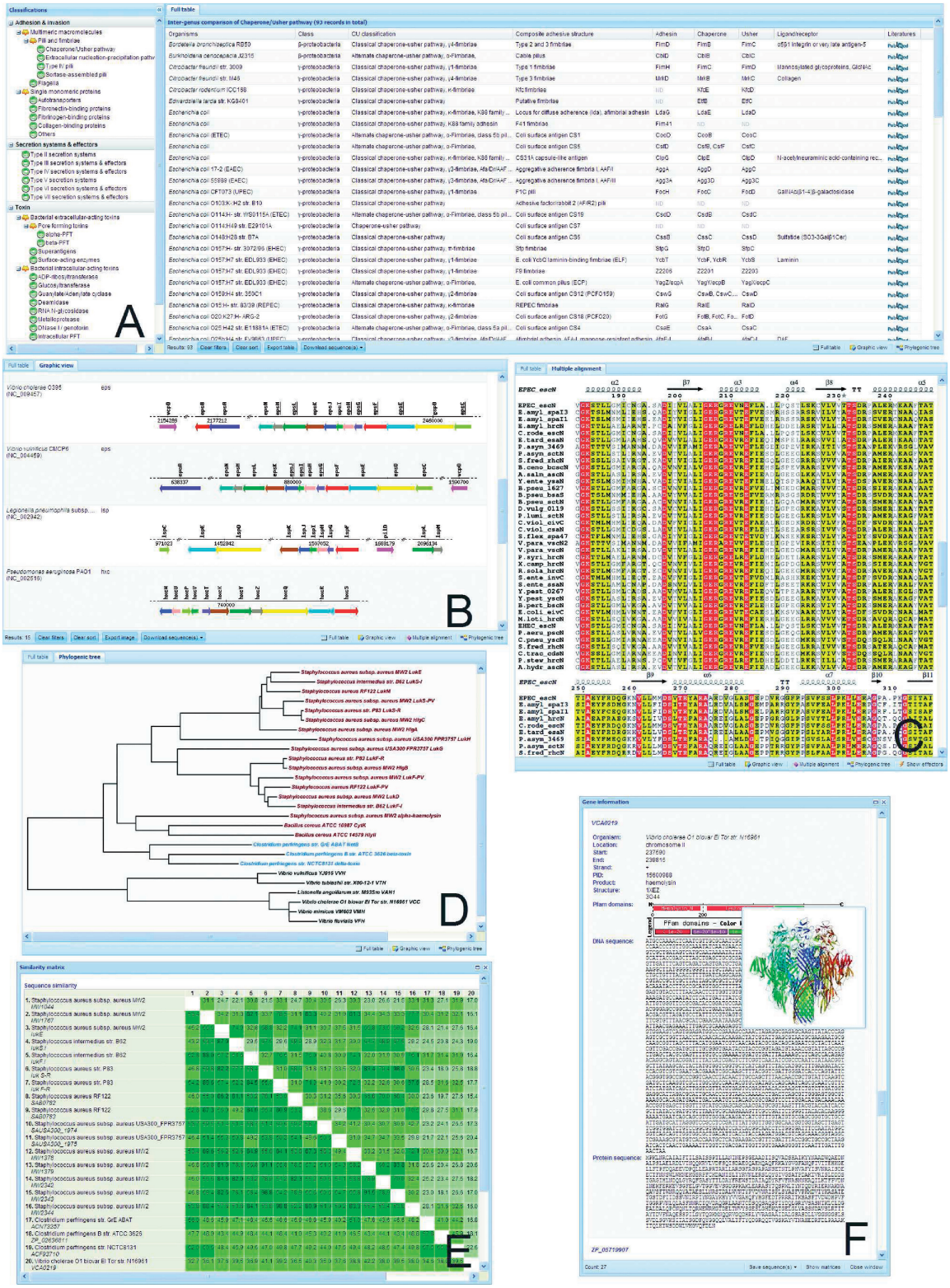

Figure 2. The updated VFDB web interface. (**A**) Menu panel (left) and tabular view of basic data sets (right). (**B**) Graphic view for multiple-component VFs, color-coded by homologue groups. (**C**) Structure-based multiple alignment of homologous VFs. (**D**) Deduced phylogenetic tree based on key component/domain (the menu panel is collapsed as a vertical bar in the left). (**E**) Color-shaded matrix of pairwise sequence similarities. (**F**) Popup window with detailed gene information along with a graphic illustration of conserved domains and a 3D structure preview.

tables and for saving the web contents as a local file (Excel table, PNG figure or FASTA sequences). In addition, there are icon buttons on the right side of the tool bar for rapid switching between the aforementioned different data presentation styles.

Genetic diversity of VFs

The diversity of genetic composition or genomic organization of homologous VFs from different pathogens may reflect the evolutionary relationships of the bacteria in terms of virulence. To facilitate future studies on the diversity of VFs, the composition and organization of VFs are highlighted in the graphic view for easy comparison. For single-gene-encoded VFs, domain architectures are shown as colored bars with a direct link to the respective protein family information. As for multiple-component VFs, all genes are depicted as clickable arrows in the linear map and are color-coded by homologue groups (Figure 2B). Therefore, it becomes straightforward to find out whether those VF-related genes are clustered or scattered on the genomes of various pathogens. For example, the genes encoding synthesis of type IVa pili are generally dispersed throughout the bacterial genome while those of type IVb pili are arranged in a contiguous cluster. We endeavor herein to provide a framework for further investigations into whether these genes were acquired separately or whether all genes were previously in a single cluster that was disrupted by genomic rearrangements.

Detailed information on each gene, including genomic location, coding strand, scientific name, product and sequences, as well as a graphic illustration of conserved domains and a preview of 3D structure(s) (if they exist) are available from a popup window upon clicking on the linear map (Figure 2F). By default, the linear maps are ordered on the basis of the phylogenetic tree (see below) to emphasize potential correlations between genetic variations and molecular evolution of VFs. In addition, the linear maps of each VF are also organized in a highly scalable grid, which enables users to sort and filter the VFs easily to construct customized graphic comparisons.

Sequence/structure variations and phylogenetic analysis

We explore the sequence and structure similarity of each homologue group in order to provide insight into how VFs may have evolved from common ancestors or may have exploited different mechanisms to arrive at similar biological activities. The multiple-alignment of homologous VFs is displayed by superimposing the crystal structure of the representative protein, and secondary structural elements are highlighted on top of the alignment (Figure 2C). It will be helpful to disclose possible similar structures deduced from homologous sequences. Color-shaded matrices summarizing the pairwise sequence/structure similarities among each homologue group are also provided in an attempt to illustrate sequence variations within each group and reveal potential protein pairs that share low sequence similarities but produce highly similar 3D structures. For example, within the α-hemolysin subfamily of β-barrel pore-forming toxins, the overall sequence identities of the core leukocidin domain from *Vibrio cholerae* cytolysin (VCC) and most of other members are <30% (Figure 2E), but their structure comparison scores are notably high, indicating clear similarities at the structural level. However, it should be noted that protein pairs displaying significant sequence homology and similar enzymatic activities might still differ in host cell targets, thereby playing different roles in bacterial pathogenesis, such as *Escherichia coli* SopE and SopE2, *Pseudomonas* ExoS and ExoT, and the *Shigella* IpaH proteins.

The growing diversity of VFs has prompted numerous efforts to develop classification schemas and unravel the evolutionary origins of VFs. For example, six major fimbrial clades of chaperone/usher systems and seven different families of T3SS are already well-established (12,13). Therefore, we performed extensive phylogenetic analysis of subclass or subfamily in the VFDB. Phylogenetic trees are labeled by species and color-coded by bacterial taxonomy (for example, see Figure 2D). This analysis may not only provide insights into the evolutionary history of VFs but may also facilitate future classification of newly identified VFs using the existing schemas. As a preliminary result, we found aerolysin-like toxin family and α-hemolysin family each might be further divided into two groups (Figure 2D), though additional investigations are needed.

DISCUSSION

Bacterial pathogenicity is one of the most important subjects in microbiology. The pathogenicity of bacteria depends on the ability to employ virulence factors, which are localized to the cell surface, released into the extracellular milieu or injected directly into host cells. Obviously, much is yet to be learned from the sophisticated virulence strategies posed by bacterial pathogens. There is an increasing need to review the entire field and perform bioinformatic mining of the explosively growing data regarding bacterial VFs. VFDB is dedicated to meeting these demands by providing up-to-date, thought-provoking information and various analytical tools. Nevertheless, we acknowledge that our work represents only a preliminary characterization of VFs. The increasingly rapid expansion of knowledge concerning the multifaceted aspects of VFs will continue to challenge our capacity to compile the latest and most relevant information for the scientific community.

FUNDING

National Basic Research Program from the Ministry of Science and Technology of China (grants 2009CB522603 and 2011CB504904 to J.Y. and Q.J., respectively); Beijing Nova Program (grant 2009A67 to J.Y.). Funding for open access charge: Beijing Nova Program.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. 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.
3. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2011) GenBank. *Nucleic Acids Res.*, **39**, D32–D37.
4. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
5. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
6. Campanella,J.J., Bitincka,L. and Smalley,J. (2003) MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences. *BMC Bioinformatics*, **4**, 29.
7. Holm,L. and Park,J. (2000) DaliLite workbench for protein structure comparison. *Bioinformatics*, **16**, 566–567.
8. Di Tommaso,P., Moretti,S., Xenarios,I., Orobitg,M., Montanyola,A., Chang,J.M., Taly,J.F. and Notredame,C. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.*, **39**, W13–W17.
9. Gouet,P., Robert,X. and Courcelle,E. (2003) ESPript/ENDscript: extracting and rendering sequence and 3D information from atomic structures of proteins. *Nucleic Acids Res.*, **31**, 3320–3323.
10. Tamura,K., Dudley,J., Nei,M. and Kumar,S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.*, **24**, 1596–1599.
11. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Federhen,S. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.
12. Nuccio,S.P. and Baumler,A.J. (2007) Evolution of the chaperone/usher assembly pathway: fimbrial classification goes Greek. *Microbiol. Mol. Biol. Rev.*, **71**, 551–575.
13. Troisfontaines,P. and Cornelis,G.R. (2005) Type III secretion: more systems than you think. *Physiology*, **20**, 326–339.