Published online 4 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D539–D542***
doi:10.1093/nar/gkm951

VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics

Jian Yang[1], Lihong Chen[1,2], Lilian Sun[1], Jun Yu[3] and Qi Jin[1,2,*]

[1]State Key Laboratory for Molecular Virology and Genetic Engineering, National Institute for Viral Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 100176, [2]Institute of Pathogen Biology, Chinese Academy of Medical Sciences, Beijing 100730, China and [3]Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Royal College, Glasgow, G1 1XW, UK

Received September 12, 2007; Revised October 11, 2007; Accepted October 15, 2007

ABSTRACT

Virulence factor database (VFDB) was set up in 2004 dedicated for providing current knowledge of virulence factors (VFs) from various medical significant bacterial pathogens to facilitate pathogenomic research. Nowadays, complete genome sequences of almost all the major pathogenic microbes have been determined, which makes comparative genomics a powerful approach for uncovering novel virulence determinants and hidden aspects of pathogenesis. VFDB was therefore upgraded to present the enormous diversity of bacterial genomes in terms of virulence genes and their organization. The VFDB 2008 release includes the following new features; (i) detailed tabular comparison of virulence composition of a given genome with other genomes of the same genus, (ii) multiple alignments and statistical analysis of homologous VFs and (iii) graphical comparison of genomic organizations of virulence genes. Comparative analysis of the numerous VFs will improve our understanding of the nature and evolution of virulence, as well as the development of new therapeutic and preventive strategies. VFDB 2008 release offers more user-friendly tools for comparative pathogenomics and it is publicly accessible at http://www.mgc.ac.cn/VFs/.

INTRODUCTION

Infectious diseases remain to be one of the biggest threats to public health despite the advance of modern medicine in post-genome era (1). Virulence factors (VFs) refer to the traits encoded by 'virulence genes' that pathogenic microbes are equipped to cause infection. To combat infectious diseases, a better understanding of VFs is absolutely necessary to decipher the mechanisms pathogenic microbes employ. VFDB was built to meet the challenge of providing up-to-date knowledge about VFs from various medically important bacterial pathogens (2).

The term pathogenomics is given to describe genomic approaches in studying microbial pathogens as to how they interact with their hosts, and in other words, pathogenomics is the study of pathogenic microbes and the entities they infect on the genomic level. The availability of complete genome sequences of different microbial species enables comparative studies to identify the common as well as species- or strain-specific VFs. Pathogenic bacteria have acquired various VFs that allow them to colonize diverse niches, cause infection and to survive in the hosts. Commonly shared VFs indicate universal requirement to cause infection by related pathogens, whereas narrowly distributed VFs often determine species- and/or strain-specific characteristics.

As a consequence, comparative genomic approaches were introduced into VFDB to explore VFs within completely deciphered bacterial genomes. The VFDB 2008 release has not only collected up-to-date knowledge about VFs from over 200 complete genomes of pathogenic bacteria, but also has incorporated a set of analytical tools to meet the desire of comparative pathogenomic studies.

DATABASE UPDATES

Data source and construction for comparative analysis

Information of publicly available bacterial genomes was retrieved from the summary page of 'Complete Microbial Genomes' at NCBI (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). RefSeq is a curated non-redundant collection of sequences with uniform format (3).

*To whom correspondence should be addressed. Tel: +86 10 6787 7732; Fax: +86 10 6787 7736; Email: zdsys@sina.com

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

For convenience of later data processing only genomes that are available from RefSeq database were included for further comparative analysis (both pathogenic and non-pathogenic isolates). The complete genome sequences and annotations were batch downloaded from the FTP server of RefSeq (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/).

The VF loci in each genome were obtained from the original literatures and subsequent reviews. Each of the VF genes was verified by sequence-similarity search against the genomes of related bacterial pathogens. Data collected were manually inspected and each homolog group was further validated by multi-alignments (see below).

The NCBI BLAST software was used for local sequence-similarity search (4). A series of BioPerl scripts were designed to extract features for all desired loci from the downloaded genome files in a semi-automated fashion. An enhanced multiple genome map viewer (5) was employed for graphical comparison of the pathogenomic organizations (see below). ClustalW (6) was run in batch by an in-house Perl script to generate multi-alignments for each homolog group.

Full tabular comparison of the pathogenomic composition

Tabular style was commonly used in scientific literatures for comparative analysis. For each genus a full comparison of pathogenomic composition is given as a spreadsheet to integrate information about VFs and genomes (see Figure 1A for example). The far left column organizes all known VFs in functional groups (toxins, lipase, etc.) and the next column lists all VF genes, and each row gives gene IDs (i.e. 'locus_tag' in annotation files) of the respective genomes, and pseudogenes are highlighted by star marks. Each gene ID in the table is a hyperlink that connects its individual page for full DNA and protein sequences. All tables can be downloaded as Excel files by terminal users.

For a quick glance of overall information, the above full-detailed tables can be converted to simplified tables, where each VF occupies a single row and gene details are replaced by symbols; '+' for presence, '−' for absence, and '±' for partial or non-functional genes. The simplified table can be viewed in text mode, symbolic mode or schematic mode.

In some genus there are many complete genome sequences, and an extreme case is *Streptococcus* with 25 complete genomes up to date. Taking the consideration that a selective subset of genomes might be more interesting to certain studies, a special filter is designed to allow generating tables that contain only selected genomes. For example, a table can be generated by expressing data of only 12 pyogenes genomes within the genus of *Streptococcus*.

Multiple alignments of homologous virulence genes

Analysis of homologous genes is a powerful approach for elucidating gene structure, function and evolution. The diversity of nucleotide sequences of bacterial genes often reflects particular niches a microbe colonizes in vivo and in the environment (7). From the full-comparison table described above, a single click on gene name will return a page with multi-alignment of both nucleotide and amino acid sequences if homologous gene(s) existing. A summary table on top of the alignment gives statistics about unmatched overhang, length of the alignment and percentage of polymorphic sites. A configurable phylogenetic tree constructed by ClustalW is displayed beneath the alignment when more than two sequences are involved.

A filter is also designed to perform multi-alignment on selected sequences only. In this case however there is no pre-computed result by default, and an online ClustalW must be run which constructs an alignment in a few seconds.

Concise graphical comparison of pathogenomic organization

Bacterial genome evolution has been driven by nucleotide substitutions and indels, as well as the changes of the genome architecture by genetic rearrangements including translocations and inversions (8). Recent comparative genomic studies have revealed that the dynamic changes of genome structures contribute greatly to the adaptive evolution of certain bacterial pathogens, such as *Shigella* (9). To unambiguously display the dynamic features of the genomes and to compare VFs' genomic organization among related pathogens, an enhanced multiple genome map viewer was implemented, which depicts all VF genes as clickable arrows (or bars) and color-coded by functional classifications. Since details about genes unrelated to virulence are hidden, the map becomes concise, although not to scale, and suitable for quick examinations of the genome organization of VFs among related genomes.

The viewer page provides three different representing styles: (i) complete mode which exhibits full scale pathogenomic map that is informative but usually large in size; (ii) compact mode that provides details of all virulence loci but omits flanking genes/regions; (iii) overview mode that scales the map to fit full screen without giving details. To facilitate interpretation of pathogenomic synteny under the overview mode, there are lines to connect homologous VF genes of the adjacent genomes when only one replicon is available (or selected by the users) for each genome. Terminal users can also run the viewer with select genomes of their interest. The usefulness of displaying synteny is highlighted in the case of *Listeria* species (Figure 1B); their genomes exhibit a high synteny in virulence gene organization. It is in agreement with the recent listerial pangenome studies, which revealed the lack of inversions or shifting of large genome segments in the sequenced *Listeria* genomes. The possible reason may be the low occurrence of transposons and insertion sequence elements in those genomes (10).

A

Virulence factors	Related genes	*L. innocua* Clip11262 (serovar 6a) chromosome NC_003212	*L. monocytogenes* EGD-e (serovar 1/2a) chromosome NC_003210	*L. monocytogenes* F2365 (serovar 4b) chromosome NC_002973	*L. welshimeri* SLCC5334 (serovar 6b) chromosome NC_008555
Adherence					
Fibronectin-binding protein	*fbpA*	lin1943	lmo1829	LMOf2365_1857	lwe1848
GW autolysin	*ami*	lin2703	lmo2558	LMOf2365_2530	lwe2508
Internalin J (LPXTG protein)	*inlJ*		lmo2821	LMOf2365_2812	
Listeria adhesion protein	*lap*	lin1675	lmo1634	LMOf2365_1656	lwe1650
Bile resistance					
Bile-salt hydrolase	*bsh*		lmo2067	LMOf2365_2098	
Enzyme					
Metalloprotease	*mpl*		lmo0203	LMOf2365_0214	
PC-PLC	*plcB*		lmo0205	LMOf2365_0216	
PI-PLC	*plcA*		lmo0201	LMOf2365_0212	
Serine-threonine phosphatase	*stp*	lin1935	lmo1821	LMOf2365_1849	lwe1840
Intracellular Survival					
Lipoate protein ligase A1	*lplA1*	lin0931	lmo0931	LMOf2365_0952	lwe0911
Oligopeptide-binding protein	*oppA*	lin2300	lmo2196	LMOf2365_2229	lwe2213
Sugar-uptake system	*hpt*		lmo0838	LMOf2365_0855	
Surface-virulence associated protein	*svpA*	lin2289	lmo2185	LMOf2365_2218	lwe2202
Invasion					
Autolysin (GW protein)	*aut*		lmo1076		
Cell wall hydrolase	*iap/cwhA*	lin0591	lmo0582	LMOf2365_0611	lwe0549
Internalin A (LPXTG protein)	*inlA*		lmo0433	LMOf2365_0471	
Internalin B (GW protein)	*inlB*		lmo0434	LMOf2365_0472*	
Lipoprotein promoting entry protein	*lpeA*	lin1961	lmo1847	LMOf2365_1875	lwe1866
Virulence protein (LPXTG protein)	*vip*		lmo0320	LMOf2365_0338	
Nucleation-promoting factor					
ActA	*actA*		lmo0204	LMOf2365_0215	
Regulation					
AgrA/AgrC	*agrA*	lin0044	lmo0051	LMOf2365_0060	lwe0042
	agrC	lin0043	lmo0050	LMOf2365_0059	lwe0041
CheA/CheY	*cheA*	lin0700	lmo0692	LMOf2365_0728	lwe0661
	cheY	lin0699	lmo0691	LMOf2365_0727	lwe0660
LisR/LisK	*lisR*	lin1414	lmo1377	LMOf2365_1396	lwe1393
	lisK	lin1415	lmo1378	LMOf2365_1397	lwe1394
Positive regulatory factor	*prfA*		lmo0200	LMOf2365_0211	
VirR/VirS	*virR*	lin1856	lmo1745	LMOf2365_1770	lwe1762
	virS	lin1852	lmo1741	LMOf2365_1766	lwe1758
Surface protein anchoring					
Lipoprotein diacylglyceryl transferase	*lgt*	lin2625	lmo2482	LMOf2365_2455	lwe2430
Lipoprotein-specific signal peptidase II	*lspA*	lin1958	lmo1844	LMOf2365_1872	lwe1863
Sortase A	*srtA*	lin0929	lmo0929	LMOf2365_0950	lwe0907
Sortase B	*srtB*	lin2285	lmo2181	LMOf2365_2214	lwe2198
Toxin					
Listeriolysin O	*hly*		lmo0202	LMOf2365_0213	

B

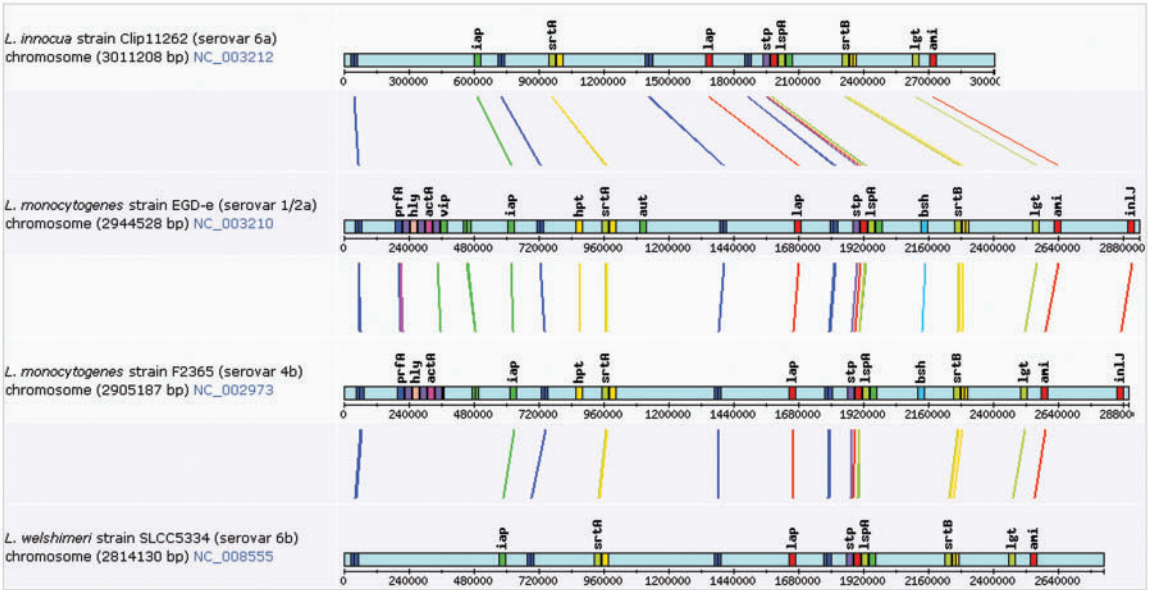

Figure 1. Comparative pathogenomic results of four sequenced *Listeria* genomes. (**A**) Full tabular comparison of pathogenomic composition. (**B**) Graphical overview for the comparison of pathogenomic organization. VF genes are color-coded by their functional classifications. Homologues between each adjacent pair of genomes are indicated by connecting lines for convenience of further interpretation.

DISCUSSION

Virulence involves a wide spectrum of biological activities, which is reflected by the diverse VFs employed by pathogenic microbes to colonize the particular niches in the hosts. A fuller investigation of VFs is highly desirable for pathogenomic research. VFDB 2008 release attempts to meet such a challenge by providing all creditable information up to date and by providing more analytical tools to the terminal users. The comparative pathogenomic results indicate that most pathogens have a flexible gene pool encoding VFs. Different combinations of VFs or organizations on microbial genomes or different expression patterns of VFs may in consequence be responsible for the diverse clinical signs of pathogen infections.

VFDB 2008 release has expanded with additional eight pathogens, which are *Brucella*, *Bartonella*, *Campylobacter*, *Clostridium*, *Corynebacterium* and *Enterococcus*, as well as *Chlamydia* and *Mycoplasma*. VFDB will continue to expand by including more medical significant pathogens, and provide up-to-date information by regular updates. For the convenience of local use, full dataset of VFDB is available for batch download in several forms, including FASTA sequences and tabular (Excel) files. Furthermore, new features and analytical tools are under development which we anticipate to make VFDB a useful pathogenomic resource to the scientific community.

ACKNOWLEDGEMENTS

This work is supported by the National Basic Research Program from the Ministry of Science and Technology of China (grant No 2005CB522904), and National Natural Science Fund from the National Natural Science Foundation of China (grant No 30600022). Funding to pay the Open Access publication charges for this article was provided by the National Natural Science Foundation of China after partially waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Becker,K., Hu,Y. and Biller-Andorno,N. (2006) Infectious diseases-a global challenge. *Int. J. Med. Microbiol.*, **296**, 179–185.
2. 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.
3. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
4. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
5. Yang,J., Chen,L., Yu,J., Sun,L. and Jin,Q. (2006) ShiBASE: an integrated database for comparative genomics of Shigella. *Nucleic Acids Res.*, **34**, D398–D401.
6. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
7. Gilsdorf,J.R., Marrs,C.F. and Foxman,B. (2004) Haemophilus influenzae: genetic variability and natural selection to identify virulence factors. *Infect. Immun.*, **72**, 2457–2461.
8. Mira,A., Klasson,L. and Andersson,S.G. (2002) Microbial genome evolution: sources of variability. *Curr. Opin. Microbiol.*, **5**, 506–512.
9. Yang,F., Yang,J., Zhang,X., Chen,L., Jiang,Y., Yan,Y., Tang,X., Wang,J., Xiong,Z. *et al.* (2005) Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res.*, **33**, 6445–6458.
10. Hain,T., Chatterjee,S.S., Ghai,R., Kuenne,C.T., Billion,A., Steinweg,C., Domann,E., Karst,U., Jansch,L. *et al.* (2007) Pathogenomics of Listeria spp. *Int. J. Med. Microbiol.*, **297**, 541–557.