D398–D401 *Nucleic Acids Research, 2006, Vol. 34, Database issue*
doi:10.1093/nar/gkj033

*Shi*BASE: an integrated database for comparative genomics of *Shigella*

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

State Key Laboratory for Molecular Virology and Genetic Engineering, Beijing 100176, China and
[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Received August 14, 2005; Revised and Accepted September 26, 2005

ABSTRACT

Among the major enteric bacterial pathogens, *Shigella* is found to display extreme genome diversity and dynamics, which imposes a challenge in comparative genomic studies. To facilitate further studies in this area, we have constructed an integrated online database, *Shi*BASE (http://www.mgc.ac.cn/ShiBASE/), which contains *Shigella* genomic sequences of four species and additional comparative genomic hybridization (CGH) data of 43 serotypes. *Shi*BASE offers online comparative analysis on DNA sequences, gene orders, metabolic pathways and virulence factors. In addition, *Shi*BASE has a newly developed online comparative visualization service, *Shi*-align, which enables the alignment of any query sequence with the reference genome sequences.

INTRODUCTION

Shigella species are the causative agents of bacillary dysentery or shigellosis, which remains a threat to public health worldwide, particularly in developing countries. Estimated shigellosis episodes are 160 million per annum, with 1.1 million deaths, the majority of which are children younger than 5 years of age (1). The clinical manifestation of shigellosis range from mild diarrhoea to severe forms of the disease which includes fever, abdominal cramps and blood, and pus and mucus in the stools (2).

Much of *Shigella* pathogenesis seems to be the result of multiple effects of its plasmid-borne type III secretion system (TTSS) encoded by *mxi* and *spa* genes. The TTSS secretes many effector proteins, including IpaA, IpaB, IpaC and IpgD, which can mediate epithelial signalling events, cytoskeletal rearrangements and other actions (3). The plasmid also encodes a 120 kDa outermembrane protein called IcsA, responsible for actin-based motility leading to intra- and intercellular spread of the bacteria by the binding of N-WASP (4). Furthermore, in addition to the virulence plasmid, many chromosomal loci, such as *Shigella* pathogenicity islands SHI-1 and SHI-2 also contribute to virulence (5,6). For a better understanding of the pathogenesis and epidemiology of *Shigella* infection, comparative genomics has emerged to be an important area of research.

Based on their somatic O-antigen, *Shigella* is divided into four species: *S.dysenteriae*, *S.flexneri*, *S.boydii* and *S.sonnei*. Our group and others have determined the complete genome sequences of five *Shigella* strains that cover all four species (7–9). In addition, comparative genomic hybridization (CGH) microarray analysis has been performed on 43 *Shigella* strains of different serotypes (J. Peng, X. Zhang, J. Yang, J. Wang, E. Yang, W. Bin, C. Wei, M. Sun and Q. Jin, manuscript in preparation), which revealed extensive diversity and dynamics between *Shigella* genomes. Hence, an integrated database capable of dealing with the complex genome comparison is urgently needed in addition to currently available online database, *etal*.,*oli*BASE (10), which is limited within pairwise genome comparison.

We present here *Shi*BASE (http://www.mgc.ac.cn/ShiBASE/), a freely accessible online resource that is focused on the comparative genomics of *Shigella*. *ShiBASE* is able to summarize large volumes of genomic sequence and comparative genomic data in a visually intuitive format. *Shi*BASE also provides a novel *Shi*-align service for visualizing BLAST hits between query sequences and known *Shigella* genomes, thus allowing rapid examination of synteny, genomic rearrangements and putative genomic islands.

DATABASE CONTENT AND CONSTRUCTION

There are five completed *Shigella* genomes available. The *S.flexneri* 2a strain 2457T was excluded from the dataset of *Shi*BASE owing to its genome being highly similar to another serotype 2a strain, 301 (Sf301). Variation between the genomes of 2457T and Sf301 is limited to the quantity of IS elements and some single-nucleotide differences (8). The four *Shigella* genomes included in the current release of *Shi*BASE are *S.dysenteriae* 1 strain 197 (Sd197), *S.boydii*

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

4 strain 227 (Sb227), *S.sonnei* strain 046 (Ss046) and *S.flexneri* 2a strain 301.

The genome sequences and annotation files from GenBank (11) were converted to tabular files by using the Bioperl toolkit (12). Genes from *Escherichia coli* K12 strain MG1655 (accession no. U00096) were used as references for defining orthologue groups. Pairwise orthologues were identified by INPARANOID program with default parameters (13), and were further inspected manually. Multi-copy genes related to mobile DNA, such as IS elements and bacteriophage were excluded. Protein sequences from each of the orthologue groups were processed in ClustalW (14) allowing for alignments and guide trees. Whole genome sequence comparisons were performed by BLASTN (15) and the information for each maximal segment pair (MSP) was deposited into the database for web illustration. The enzyme commission information of each genome as well as metabolic pathway maps were retrieved from the KEGG ftp server (16), and then revised by manual curation for dynamic web presentation.

*Shi*BASE was built on a RedHat Linux 9.0 operation system, and MySQL was used to construct the database for storage of information, including genome annotations, orthologue groups, sequence comparisons and CGH results. The Perl programming language and several modules (DBI, GD and CGI) were used to generate dynamic web pages. Client-side JavaScript and Java applet were also applied in some cases for data presentation purposes.

COMPARATIVE GENOMICS IN *Shi*BASE

The interspecies comparison of *Shigella* genomes may be carried out at several different levels within *Shi*BASE. One may compare basic genome features, such as genome size, IS elements, genome structure (inversions and deletions) and order of orthologous genes. One can also compare metabolic pathways and virulence factors among *Shigella* genomes. Moreover, the CGH data produced from analyses on 43 different serotypes of *Shigella* strains are also integrated within *Shi*BASE. By using the *Shi*-align visualization platform one can also perform comparative analysis on any given query sequences.

Sequence-based comparison

All *Shigella* and *E.coli* strains sequenced to date are closely related and share an essentially collinear common 'backbone' genome sequence. However, there are unique DNA sequences present in each of the *Shigella* and *E.coli* chromosomes (7). For graphically viewing a selected region across all genomes, we have set a clickable web page in the *Shi*BASE interface. Figure 1A shows a comparison of a 20 kb segment involved in flagella biogenesis from all *Shigella* genomes. The reference genome (from Sb227 in this case) is presented on the top and the query genome sequences are aligned below. The pre-computed MSPs from BLASTN are indicated by red bars between each pair, and additionally are equipped with popup messages that provide details of the sequence alignment. Genome annotations, including genes, stable RNAs and IS elements are also graphically illustrated for each region of synteny. The comparison range can be increased in size (up to 100 kb) and the window can slide along each genome for simple browsing. If necessary users can reverse and display the complementary strand (see Sf301 in Figure 1A).

As shown in Figure 1A, there are different mutations in the *fli* genes across all *Shigella* strains, which explain genetically why members of the *Shigella* species are non-motile. An additional surface structure, namely fimbriae, is also eliminated from the members of the *Shigella* species by a different gene inactivation (data not shown). These are both examples of the genome interrogation that one may perform using *Shi*BASE, thus adding insight into the biology and the evolutional history of the *Shigella* species.

*Shi*BASE also allows investigation of variation in genome structure at the amino acid level by gene order comparison, which is particularly useful for detecting orthologous genes. Orthologues are genes that are evolved vertically from a single ancestral gene in different genomes, and they often retain similar biological functions in the present-day organisms (13). Generally, protein sequence comparison is more sensitive because codon usage often leads to DNA sequence variations. For example, although the coding sequences of *fliC* show little similarity among *Shigella* genomes (Figure 1A) their protein sequences bare 46–62% in overall identity. In order to provide a uniform interface, the comparison of orthologues order was presented in a similar style as the comparison of DNA sequence described above, except that red bars were used here to link orthologue pairs. Multi-alignment and guide tree created by ClustalW for each orthologue group are directly linked from the individual graphic comparison pages.

Function-based comparison

Analysis of metabolic pathways is useful in illuminating the complexity of the *Shigella* metabolic patterns for the explanation of biochemical characteristics that distinguish *Shigella* from other enteric bacteria. By adopting the KEGG pathway maps, which offer graphic representations for most metabolic pathways (16), we have generated a comparative pathway mapviewer. It combines individual pathways of *E.coli* K12 MG1655 and *Shigella* strains and displays the results in one map. For instance, *Shigella* strains do not synthesize lysine decarboxylase since the absence of lysine decarboxylase activity is important to their virulent lifestyle (17). These data can be easily accessed and browsed on our web page.

Another example is given in Supplementary Figure 1, which is the comparative view of the tyrosine metabolism pathway. It shows that Sf301, Sb227 and Ss046 possess the *hpa* operon, which offers them the ability to catabolize the aromatic compound 4-hydroxyphenylacetate (4-HPA), whereas Sd197 is lacking the *hpa* operon (18).

Although the four species of *Shigella* were originally divided on the basis of O-specific polysaccharide of the LPS, they additionally demonstrate distinctive differences in their pathological and epidemiological features. Furthermore, *S.dysenteriae* serotype 1 solely possesses the cytotoxic Shiga toxin and causes disease with neurological and renal complications (19). To highlight the diversity of virulence factors presented in each genome, we collated all known and putative virulence factors in a table, which links each virulence gene to detailed pages, allowing for further analysis or DNA/protein sequence retrieval.

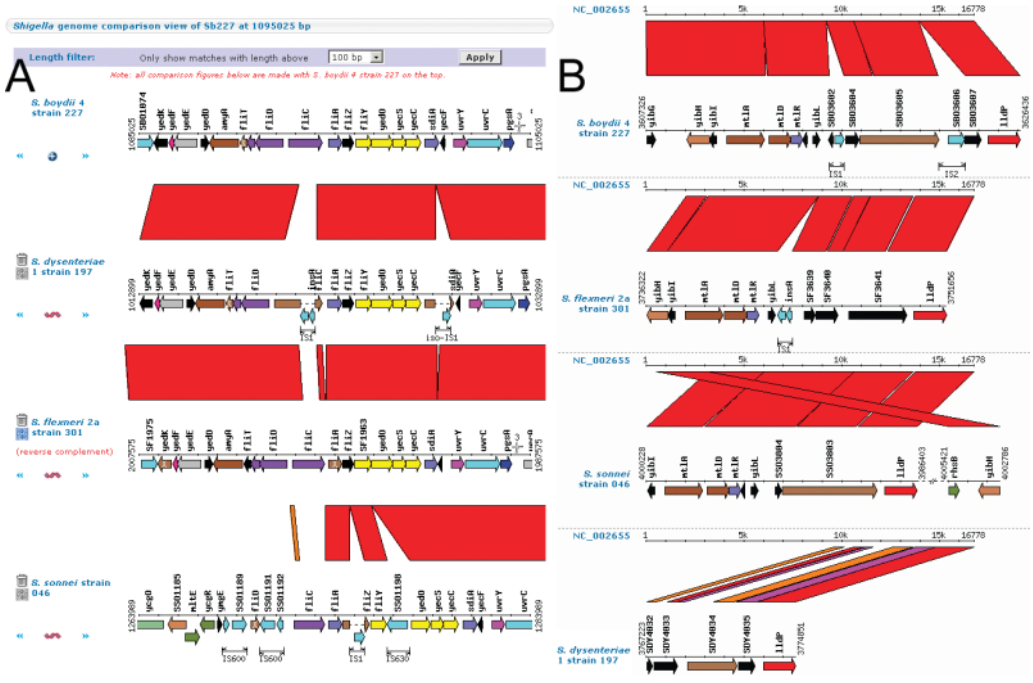

Figure 1. Sequence-based comparison in *Shi*BASE. (**A**) *Shigella* genome comparison view. Genome sequence of Sb227 was set as reference on the top. The current comparison window size is 20 kb and it can be increased by clicking the circular plus ('+') icon on the top left. Note that the reverse complementary strand of Sf301 was used in this comparison figure. Arrows represent genes, of which those with write cross or with dashed lines that connect separate portions are pseudogenes owing to mutations or insertions, respectively. The compared region contains a set of flagellar genes (*fli*) and they were inactivated by different pseudogenes in all four genomes. The left part of the corresponding region in Ss046 (on the bottom) was translocated elsewhere. (**B**) Result of *Shi*-align by querying with a ~17 kb segment from EHEC. Coloured bars between each pair represent MSPs from BLASTN of different identity values: red, >95%; pink, 90–95%; and orange, <90%.

Extended genome content comparison

In order to extend genomic comparison beyond genome sequences, we have integrated the newly available CGH data into *Shi*BASE. The CGH analysis includes 43 *Shigella* strains of different serotypes. The DNA microarray used to generate these data contains 5051 non-redundant genes from *E.coli* K12 MG1655 and all known *Shigella* genomes. Users may browse or search the CGH data and display the results in serological or clustering order.

Recent phylogenic analyses on eight house keeping genes have grouped all *Shigella* strains into three main clusters with five outliers (20). Since use of the CGH results has generated a similar clustering order, the information is beneficial in demonstrating a correlation between genome structure and phylogenetic relationship. In order to facilitate further comparative analysis on the CGH results, an auto filter was set up to offer users the potential to rapidly examine commonly shared or lost genes within each cluster or serogroup. For example, by screening the CGH results, a total of 228 genes present in the other *S.flexneri* strains are absent from *S.flexneri* 6. Most of the absent genes are related to LPS biosynthesis (such as *rfbEFGIJ*), outer membrane proteins (such as *ompG*, *fhuA* and *nmpC*) or enzymes (such as *cai* operon necessary for carnitine metabolism). Whereas *S.flexneri* 6 solely possesses the *gsp* genes encoding a novel type II secretion system. These findings may offer some reasoning to why *S.flexneri* 6 behaves differently from other *S.flexneri* strains and lies within a different cluster.

Online sequence comparison visualization service *Shi*-align

BLAST is a very powerful tool for finding sequence similarities and it has been widely used in database interrogation and sequence comparisons. However, when performing long sequence alignments, the voluminous and complex BLAST textual output is often difficult for biologists to read and analysis. Hence, several stand-alone programs, such as ACT (http://www.sanger.ac.uk/Software/ACT/) and GenomeComp (21), have been generated to tackle these issues.

In line with these programs, we have developed an online tool named *Shi*-align. *Shi*-align allows the visualization of sequence comparisons, and is incorporated within *Shi*BASE. *Shi*-align permits users to compare any sequence fragments with known *Shigella* genomes and in turn generate a graphic alignment view (Figure 1B) as an alternative to the textual output of BLAST. This tool is particularly useful for viewing the overall organization between query sequences and known *Shigella* genome sequences. Figure 1B shows a ~17 kb segment from enterohaemorrhagic *E.coli* (EHEC) encoding a putative adhesion aligned with corresponding sequences from various *Shigella* genomes in *Shi*-align. This graphical alignment demonstrates that sequences from Sb227 and Sf301 with the exception of IS insertions are essentially collinear with that of EHEC. Whereas the alignment sequences from Ss046 and Sd197 have obvious DNA rearrangements with respect to the EHEC sequence.

CONCLUSIONS AND FUTURE DIRECTIONS

As the developments of sequencing techniques and ubiquity of fast computers have led microbial genomics to the genus scale (22), resources at genus scale are also on demand. *Shi*BASE as well as other recently developed databases, such as *coli*BASE (10) and MolliGen (23), are all examples that aim to fulfil the requirement.

The current release of *Shi*BASE is dedicated to be a comprehensive online resource of *Shigella* and offers a platform for further comparative genomics studies on this important human pathogen. Nevertheless, the close genetic relationship between *Shigella* and *E.coli* is already well established. Moreover, an *E.coli* pathovar, named enteroinvasive *E.coli* (EIEC), is biochemically, genetically and pathogenically closely related to *Shigella* spp. EIEC is considered as an intermediate type in evolution between *E.coli* and *Shigella*. So, in the future, *Shi*BASE will not only include more data of *Shigella* (such as proteomics data) but also collect genomes of all types of pathogenic *E.coli*, especially EIEC.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We thank Dr Stephen Baker for critical reading of the manuscript. This project is supported by the National Basic Research Priorities Program (grant no. 2005CB522904) and the High Technology Research and Development Program (grant no. 2001AA223011) from the Ministry of Science and Technology of China. Funding to pay the Open Access publication charges for this article was provided by MSTC.

Conflict of interest statement. None declared.

REFERENCES

1. Kotloff,K.L., Winickoff,J.P., Ivanoff,B., Clemens,J.D., Swerdlow,D.L., Sansonetti,P.J., Adak,G.K. and Levine,M.M. (1999) Global burden of *Shigella* infections: implications for vaccine development and implementation of control strategies. *Bull. World Health Organ.*, **77**, 651–666.
2. Jennison,A.V. and Verma,N.K. (2004) *Shigella flexneri* infection: pathogenesis and vaccine development. *FEMS Microbiol. Rev.*, **28**, 43–58.
3. Tran Van Nhieu,G., Bourdet-Sicard,R., Dumenil,G., Blocker,A. and Sansonetti,P.J. (2000) Bacterial signals and cell responses during *Shigella* entry into epithelial cells. *Cell Microbiol.*, **2**, 187–193.
4. Egile,C., Loisel,T.P., Laurent,V., Li,R., Pantaloni,D., Sansonetti,P.J. and Carlier,M.F. (1999) Activation of the CDC42 effector N-WASP by the *Shigella flexneri* IcsA protein promotes actin nucleation by Arp2/3 complex and bacterial actin-based motility. *J. Cell Biol.*, **146**, 1319–1332.
5. Moss,J.E., Cardozo,T.J., Zychlinsky,A. and Groisman,E.A. (1999) The selC-associated SHI-2 pathogenicity island of *Shigella flexneri*. *Mol. Microbiol.*, **33**, 74–83.
6. Rajakumar,K., Sasakawa,C. and Adler,B. (1997) Use of a novel approach, termed island probing, identifies the *Shigella flexneri* she pathogenicity island which encodes a homolog of the immunoglobulin A protease-like family of proteins. *Infect. Immun.*, **65**, 4606–4614.
7. Jin,Q., Yuan,Z., Xu,J., Wang,Y., Shen,Y., Lu,W., Wang,J., Liu,H., Yang,J., Yang,F. *et al.* (2002) Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res.*, **30**, 4432–4441.
8. Wei,J., Goldberg,M.B., Burland,V., Venkatesan,M.M., Deng,W., Fournier,G., Mayhew,G.F., Plunkett,G.,III, Rose,D.J., Darling,A. *et al.* (2003) Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infect. Immun.*, **71**, 2775–2786.
9. Yang,F., Yang,J., Zhang,X., Chen,L., Jiang,Y., Yan,Y., Tang,X., Wang,J., Xiong,Z., Dong,J. *et al.* (2005) Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res.*, **33**, 6445–6458.
10. Chaudhuri,R.R., Khan,A.M. and Pallen,M.J. (2004) coliBASE: an online database for *Escherichia coli*, *Shigella* and *Salmonella* comparative genomics. *Nucleic Acids Res.*, **32**, D296–D299.
11. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2005) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.
12. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
13. Remm,M., Storm,C.E. and Sonnhammer,E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.
14. 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.
15. 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.
16. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
17. Maurelli,A.T., Fernandez,R.E., Bloch,C.A., Rode,C.K. and Fasano,A. (1998) 'Black holes' and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **95**, 3943–3948.
18. Prieto,M.A., Diaz,E. and Garcia,J.L. (1996) Molecular characterization of the 4-hydroxyphenylacetate catabolic pathway of *Escherichia coli* W: engineering a mobile aromatic degradative cluster. *J. Bacteriol.*, **178**, 111–120.
19. Sansonetti,P.J. (2001) Microbes and microbial toxins: paradigms for microbial-mucosal interactions III. Shigellosis: from symptoms to molecular pathogenesis. *Am. J. Physiol. Gastrointest. Liver Physiol.*, **280**, G319–G323.
20. Pupo,G.M., Lan,R. and Reeves,P.R. (2000) Multiple independent origins of Shigella clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc. Natl Acad. Sci. USA*, **97**, 10567–10572.
21. Yang,J., Wang,J., Yao,Z.J., Jin,Q., Shen,Y. and Chen,R. (2003) GenomeComp: a visualization tool for microbial genome comparison. *J. Microbiol. Methods*, **54**, 423–426.
22. Ravel,J. and Fraser,C.M. (2005) Genomics at the genus scale. *Trends Microbiol.*, **13**, 95–97.
23. Barre,A., de Daruvar,A. and Blanchard,A. (2004) MolliGen, a database dedicated to the comparative genomics of Mollicutes. *Nucleic Acids Res.*, **32**, D307–D310.