Genome Sequence of *Shigella flexneri* 2a: Insights Into Pathogenicity Through Comparison with Genomes of *Escherichia coli* K12 and O157

Jin, Q., Yuan, Z., Xu, J., Wang, Y., Shen, Y., Lu, W., … Yu, J. (2002). Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157. Nucleic Acids Research, 30(20), 4432–4441.

Trixie Roque & Jake Woodlee

Departments of Biology and Computer Science
Loyola Marymount University
October 17, 2015

Outline

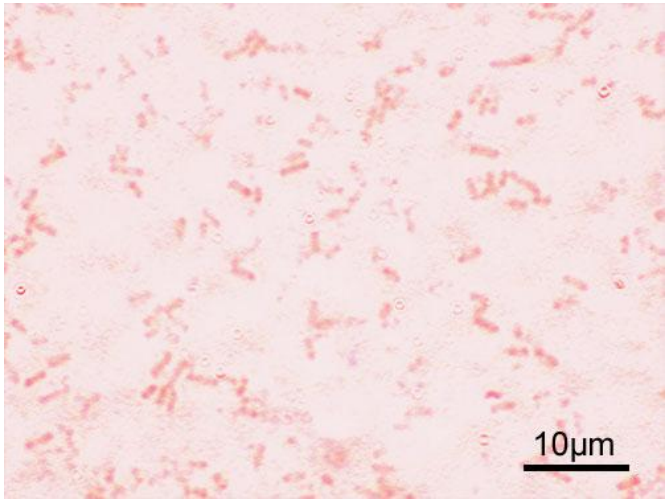
- Shigellosis, caused by *S. flexneri,* is one of the leading causes of death in young children in developing countries.
- Sequencing the genome mainly involved automation to reduce human-induced errors.
- Comparison between *S. flexneri* and its genetic relative, *E. coli.*, revealed distinct and similar characteristics between their chromosomes.
- Viable database for this organism provides a fast and easy way to explore its genome.

Outline

- Shigellosis, caused by *S. flexneri,* is one of the leading causes of death in young children in developing countries.
- Sequencing the genome mainly involved automation to reduce human-induced errors.
- Comparison between *S. flexneri* and its genetic relative, *E. coli.*, revealed distinct and similar characteristics between their chromosomes.
- Viable database for this organism provides a fast and easy way to explore its genome.

S. flexneri is a species of Bacteria that Causes a Worldwide Health Concern to Developing Countries

- Gram-negative, non-sporulating, facultative anaerobes
- Causes bacillary dysentery and shigellosis
- Estimated deaths of 161.1 Million annually
- Invasive in the colon and rectum
- Due to the lack of adequate treatment strategies, the World Health Organization has made an anti-*Shigella* vaccine a priority



Wikipedia, 2015

The Strain Observed was Sf301 serotype 2a

- *S. flexneri* 2a is the most prevalent species and serotype.
- Determining the connection between the chromosome and virulence plasmid required the discovery of the entire genome sequence of *Shigella flexneri*.
- The reference strain was isolated in 1984 from a man in China who carried the disease
- The strain was cultured at 37°C overnight on tryptic soy agar containing 0.01% Congo red. Colonies were inoculated into tryptic soy broth and grown to stationary phase at 37°C for isolating plasmid and chromosomal DNAs

Outline

- Shigellosis, caused by *S. flexneri,* is one of the leading causes of death in young children in developing countries.
- Sequencing the genome mainly involved automation to reduce human-induced errors.
- Comparison between *S. flexneri* and its genetic relative, *E. coli.*, revealed distinct and similar characteristics between their chromosomes.
- Viable database for this organism provides a fast and easy way to explore its genome.

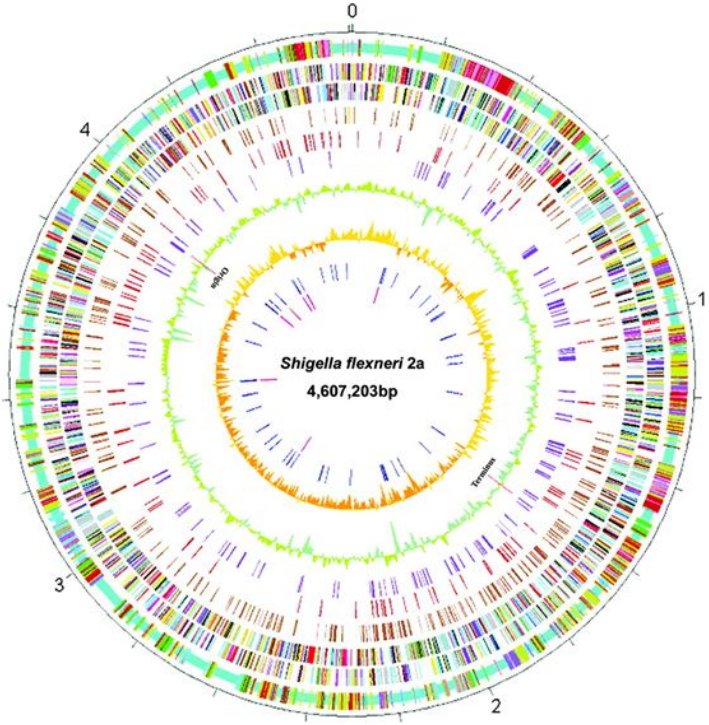
Automated Shotgun Sequencing was Used to Create the Genome

- Initially involved the employment of a highly accurate base-calling software, called *phred*
- Significantly reduced human interaction with the DNA sequences, thus also reducing the errors that would have resulted from human involvement
- After reaching 318 overlapping regions in the species' genome, a program called *consed* was then used for sequence finishing.
- Identifying open reading frames involved the Glimmer 2.0 software, but some manual inspection was still employed for overlapping ORFs.
- The databases BLASTP and COGs were used to identify families of related proteins. Genomic comparison with E. coli K12 was then executed using the GenomeComp software.

Outline

- Shigellosis caused by *S. flexneri* is one of the leading causes of death in young children in developing countries.
- Sequencing the genome mainly involved automation to reduce human-induced errors.
- Comparison between *S. flexneri* and its genetic relative, *E. coli.*, revealed distinct and similar characteristics between their chromosomes.
- Viable database for this organism provides a fast and easy way to explore its genome.

Circular Genome Map of Sf301 Chromosome Compared with those of *E.coli* K12 MG1655 and 0157 EDL933



Qi Jin et al. Nucl. Acids Res. 2002;30:4432-4441

Nucleic Acids Research

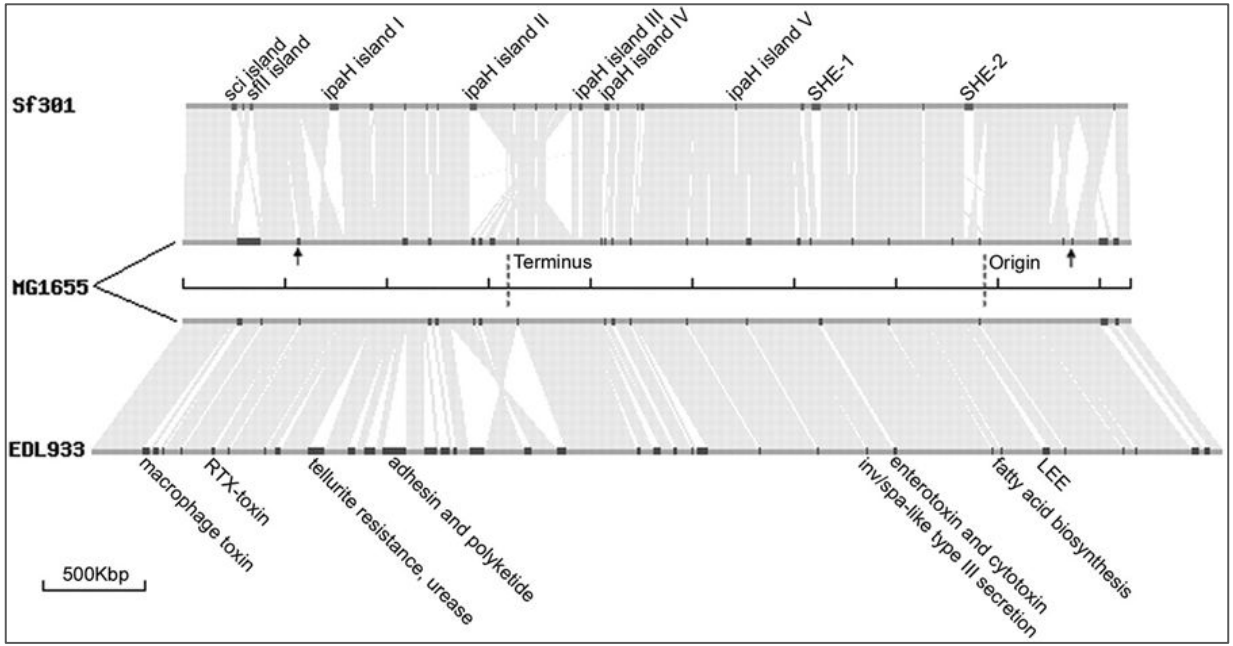
Comparing the Chromosomes of *S. flexneri* and *E. coli* Shows their Similarities

Table 1. General features of the Sf301 genome compared with genomes of *E.coli* K12 and 0157, and the virulence plasmid, pWR501, from *S.flexneri* M90T 5a

Chromosome	Sf301	MG1655[a]	EDL933[b]
Total length (bp)	4 607 203	4 639 221	5 528 445
No. of total ORFs	4434	4289	5349
Average length of ORFs (bp)	891	954	905
Percentage of coding sequence (%)	80.4	87.8	87.1
G + C content			
Total genome (%)	50.89	50.79	50.40
Protein coding regions (%)	51.95	51.85	51.51
RNA genes (%)	54.79	54.84	54.88
Intergenic regions (%)	46.07	42.28	42.76
Ribosomal RNA			
No. of 16S	7	7	7
No. of 23S	7	7	7
No. of 5S	8	8	8
No. of transfer RNA	97	92	93
No. of tmRNA	1	1	1
No. of non-classical RNA	9	5	5
Translocations and inversions[c]	13	–	1
IS elements	314	39	40
Of which partial copies	67	7	19
Plasmid	pCP301	pWR501[d]	
Total length (bp)	221 618	221 851	
No. of total ORFs	267	293	
Average length of ORFs (bp)	658	636	
Percentage of coding sequence	76.24	82.09	
G + C content			
Total (%)	45.77	46.36	
Coding regions (%)	46.13	46.95	
Intergenic regions (%)	44.59	43.69	
IS elements	88	92	
Of which partial copies	62	69	

[a]Data are from Blattner *et al.* (10).
[b]Data are from Perna *et al.* (11).
[c]Only those with DNA segments >5 kb are listed.
[d]Data are from Venkatesan *et al.* (8).

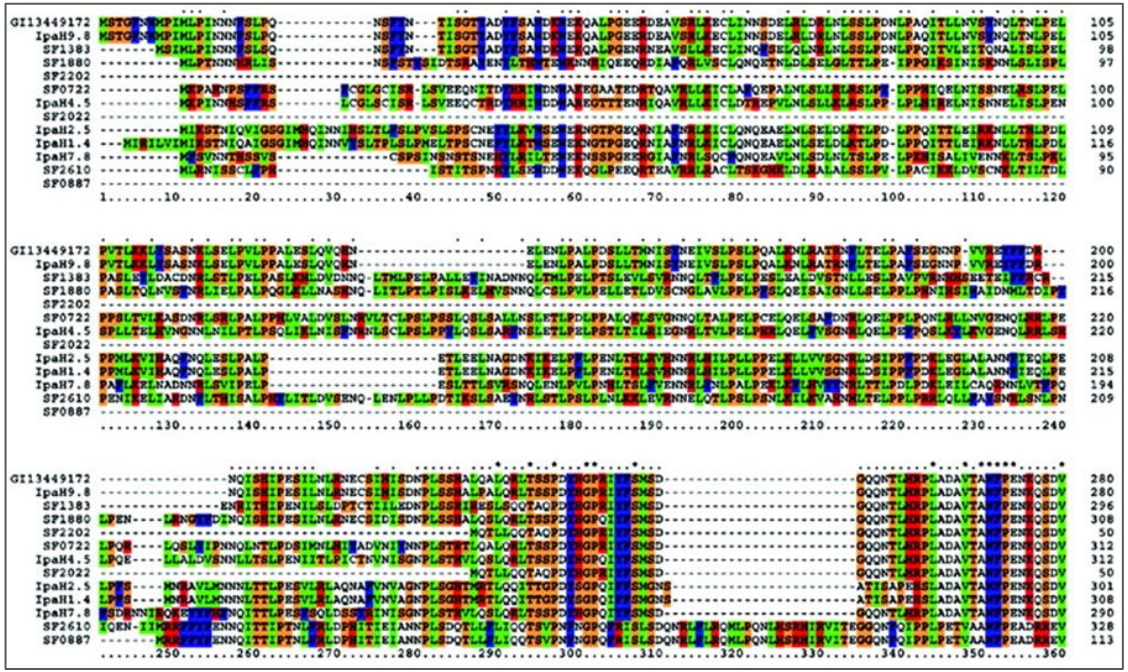
Schematic Representation of Translocations and Inversions, and Strain-Specific Islands



Qi Jin et al. Nucl. Acids Res. 2002;30:4432-4441

Nucleic Acids Research

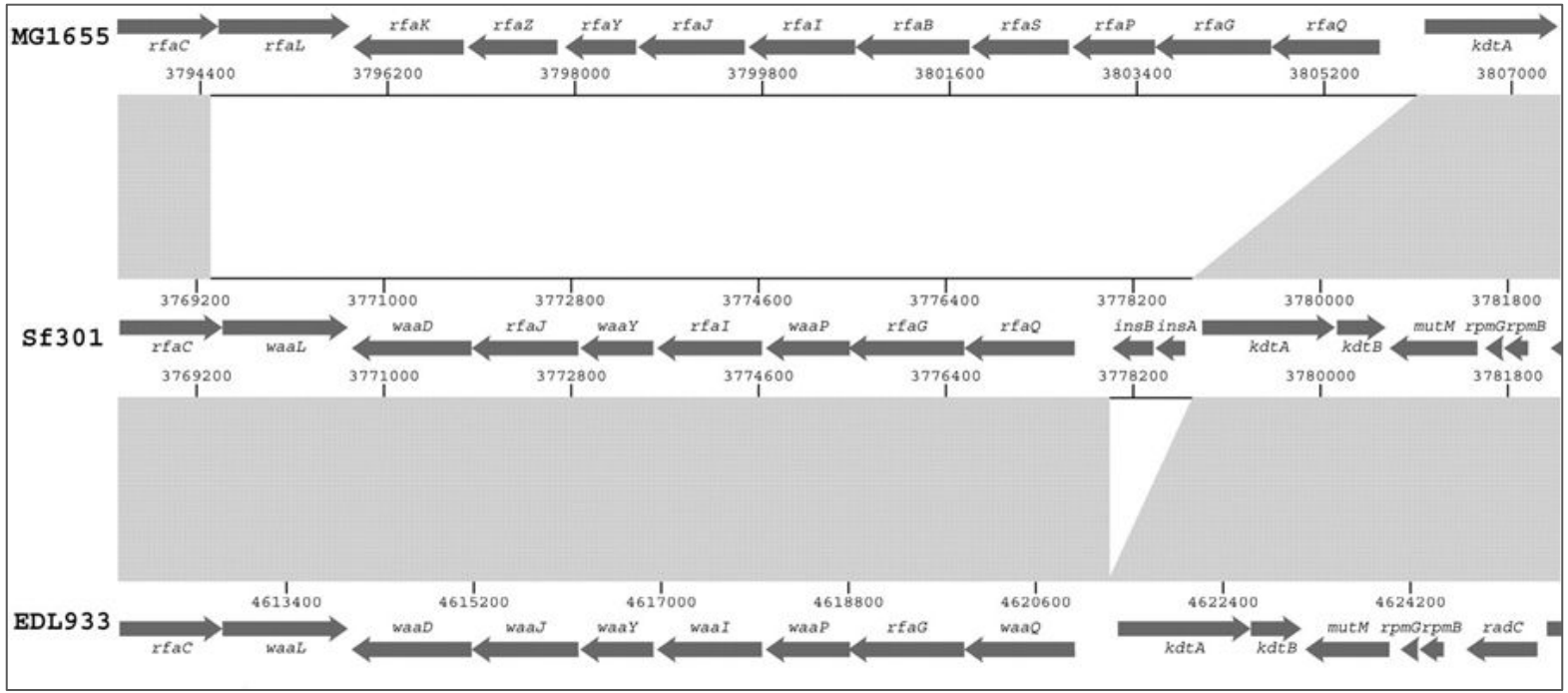
Amino Acid Sequence Alignment of N-terminal Halves of IpaH Proteins were Identified in Sf301



Qi Jin et al. Nucl. Acids Res. 2002;30:4432-4441

Nucleic Acids Research

Comparison of the rfa/waa Region



Qi Jin et al. Nucl. Acids Res. 2002;30:4432-4441

Nucleic Acids Research

Insertion sequence elements found in the three strains were also compared with those of 5a

Table 2. IS elements identified in genomes of Sf301, MG1655 and EDL933, the virulence plasmid, and pWR501, from *S.flexneri* 5a

Name	Length (bp)	No. of ORFs	No. of intact elements					No. of partial elements				
			Sf301	K12	0157	pCP301	pW501	Sf301	K12	0157	pCP301	pWR501
IS1	768	2	108	6	2	2	3	9	0	0	1	1
iso-IS1	803	2	0	0	0	0	0	1	0	3	5	5
IS2	1331	2	30	6	1	1	2	5	1	0	2	2
IS3	1258	2	5	5	0	0	0	3	0	2	7	8
IS4	1428	2	18	1	0	1	1	3	0	0	1	2
IS5	1198	1	0	10	0	0	0	0	1	0	0	0
iso-IS10R[a]	1329	1	13	0	0	2	0	0	0	0	0	0
IS21	2131	2	0	0	0	0	0	0	0	0	3	3
IS91	1830	1	3	0	0	0	0	2	0	0	6	6
IS100	1963	2	0	0	0	0	0	0	0	0	7	6
IS150	1443	3	0	1	0	0	0	5	0	0	2	2
IS186	1372	1	0	3	0	0	0	0	0	0	0	0
IS600	1264	2	35	0	0	3	2	17	1	6	10	13
IS629	1310	2	10	0	18	8	5	11	0	3	3	9
IS630	1164	1	0	0	0	1	1	0	0	4	2	2
IS911	1250	2	16	0	0	1	1	0	4	0	0	0
IS1294	1714	1	0	0	0	1	2	3	0	0	7	4
IS*Sfl1*	929	1	0	0	0	1	1	0	0	0	2	3
IS*Sfl2*	1374	1	6	0	0	2	2	0	0	0	1	0
IS*Sfl3*	1302	1	0	0	0	1	1	1	0	0	1	1
IS*Sfl4*	2754	3	3	0	0	2	2	7	0	1	2	2
Total			247	32	21	26	23	67	7	19	62	69

[a]iso-IS10R is a homolog of IS10R identified in Sf301 in this study.

Pseudogenes were identified

Table 3. Pseudogenes with known functions identified in Sf301 genome

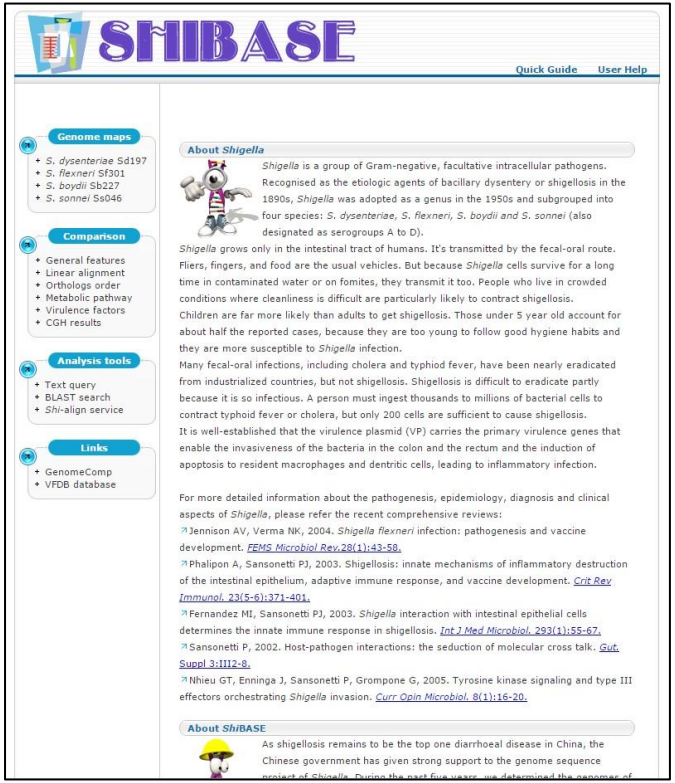
Pathway	Mutation	Description
Carbohydrate metabolism		
araA	Stop codon	L-Arabinose isomerase; arabinose catabolism
ugd	Stop codon	UDP-glucose 6-dehydrogenase; colanic acid synthesis
fucK	Stop codon	L-Funulokinase, fucose catabolism
glcD	Stop codon	Glycolate oxidase subunit D
xylA	Stop codon	D-Xylose isomerase; D-xylose catabolism and D-glucose conversion
aceB	Stop codon	Malate synthetase A; glyoxylate bypass
dgoA	Stop codon	D-Galactonate hydro-lyase; galactonate catabolism
fdhF[a]	Stop codon	Formate dehydrogenase-H; anaerobic respiration
zwf	Stop codon	G6PD; oxidative branch of pentose phosphate pathway
Energy metabolism		
cyoB	Stop codon	Cytochrome o ubiquinol oxidase subunit I; active under high oxygen growth conditions
cyoA	Truncation	Cytochrome o ubiquinol oxidase subunit II; as *cuoB*
acs	Stop codon	Acetyl-CoA synthetase; scavenging acetate
hyfB	Stop codon	Hydrogenase 4 subunit; anaerobic respiration
narZ	Stop codon	NRZ; anaerobic terminal electron acceptor
torA	Stop codon	Trimethylamine N-oxide reductase subunit; electron acceptor (anaerobic respiration)
torD	Insertion	Chaperone of TorA; preventing TorA degradation
Lipid metabolism		
hcaD	Stop codon	Ferredoxin reductase; utilization of aromatic acids
Amino acid metabolism		
speF	Stop codon	Ornithine decarboxylase isozyme; putrescine synthesis
speG	Frame shift	Spermidine acetyltransferase; polyamine synthesis
nadB	Stop codon	Quinolinate thynthetase B; pyridine synthesis
gabD	Stop codon	Succinate-semialdehyde dehydrogenase; aminobutyrate catabolism
mtgA	Frame shift	Peptidoglycan enzyme; cell wall formation
metA	Truncation	Homoserine transsuccinylase; methionine synthesis
cstC	Stop codon	Acetylornithine transaminase; arginine catabolism
Cofactors and vitamins		
nfnB	Insertion	Dihydropteridine reductase; recycling the quinoid dihydrobiopterin cofactor by reducing it
lhr	Stop codon	ATP-dependent helicase, dispensable
lplA	Frame shift	Lipoate-protein ligase A; ligation of lipoyl to apoprotein
Complex lipids		
gldA	Stop codon	Glycerol dehydrogenase; glycerol dissimilation
Complex carbohydrates		
ycjM	Insertion	Putative polysaccharide hydrolase
otsA	Truncation	Trehalose-6-phosphate synthase; response to high osmolarity
aceK	Stop codon	Isocitrate dehydrogenase kinase/phosphatase; control flux between the TCA cycle and the glyoxylate bypass
Translation		
prfB	Stop codon	Peptide chain release factor RF-2
Transport		
araF	Stop codon	L-Arabinose-binding periplasmic protein
cysW	Stop codon	Sulfate transport system permease W protein
yhdX	Truncation	Permease; putative amino acid ABC transporter
ugpC	Insertion	ATP-transporter; glycerol-3-phosphate uptake
rbsA	Insertion	ATP-biding component; D-ribose transport
rbsB	Stop codon	ABC transporter; D-ribose periplasmic binding protein
glvG	Frame shift	6-Phospho-β-glucosidase; arbutin fermentation
ptsA	Stop codon	PEP-protein phosphotransferase system enzyme I
yphF	Stop codon	ABC transporter; periplasmic binding
Signal transduction		
citB	Truncation	Regulator (paired with *citR*); citrate fermentation
kdpE	Stop codon	Regulator of the *kdp* operon; potassium transport
kdpD	Stop codon	Sensor of the *kdpDE* system; potassium transport
narQ	Stop codon	Nitrate/nitrite sensor protein; acts on NarL/NarP
arp	Stop codon	Regulator of acetyl CoA synthetase
malT	Stop codon	Positive regulator of *mal* operon
Cell motility		
fliA	Frame shift	σ^{28} for flagellar operons
flgF	Stop codon	Cell-proximal portion of basal-body rod
flgK	Stop codon	Hook-filament junction protein 1
flgL	Stop codon	Hook-filament junction protein
fliF	Stop codon	Basal-body MS-ring and collar protein
fliJ	Truncation	FliJ protein
flhA	Stop codon	Export of flagellar proteins
Unassigned enzymes		
tesA	Stop codon	Acyl-CoA thioesterase I; hydrolyzes long chain acyl thioesters
pphA	Stop codon	Protein phosphatase 1; modulates phosphoproteins signaling protein misfolding

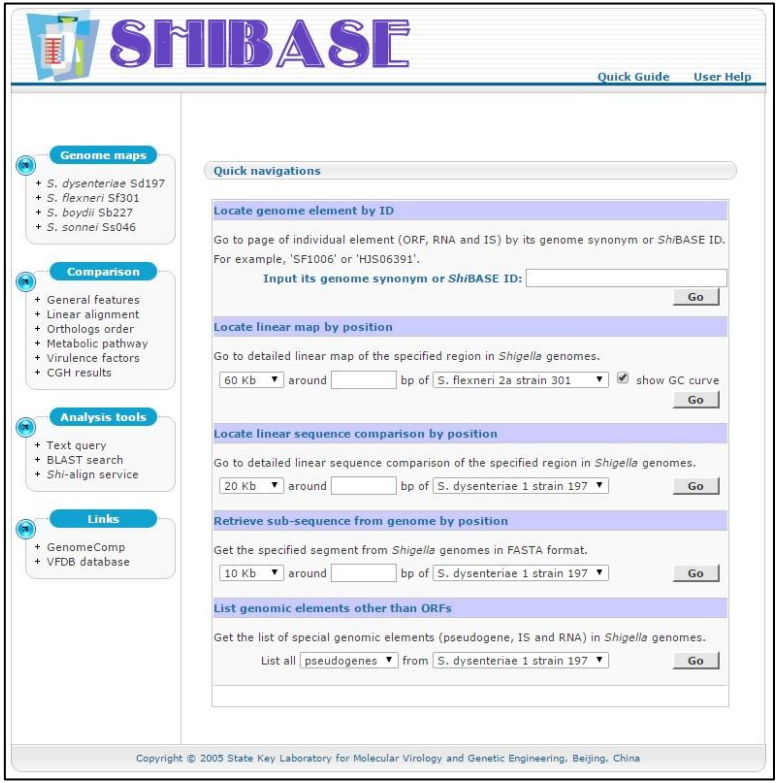
Table 3. *Continued*

Pathway	Mutation	Description
pphB	Stop codon	Removal of a phosphate group attached to serine or threonine residue; signaling protein misfolding through cpxRA system
Unassigned non-enzymes		
yaaJ	Stop codon	Transport protein; sodium/alanine symporter
nfrA	Stop codon	Omp; bacteriophage N4 receptor
csgG	Stop codon	Transporter; curli assembly
csgA	Insertion	Curlin major subunit; coiled surface structures
fepE	Stop codon	Transporter; ferric enterobactin (enterochelin)
fhuE	Stop codon	Omp; receptor for ferric iron uptake
entC	Stop codon	Isochorismate synthase; enterobactin biosynthesis
hlyE	Stop codon	Hemolysin E; hemolytic to sheep blood
hslJ	Truncation	Heat shock protein HslJ
uidB	Truncation	Transporter; specific to α- and β-glucuronides
celD	Insertion	Negative regulator of *cel* operon (cryptic); ferment cellobiose, arbutin and salicin
molR	Insertion	Molybdate metabolism regulator, first fragment
molR_2	Stop codon	Molybdate metabolism regulator, fragment 2
cirA	Stop codon	Porin and receptor; colicin I uptake
focB	Frame shift	Formate transporter (formate channel 2)
emrA	Stop codon	Multidrug resistance secretion protein
ppdA	Frame shift	Prepilin peptidase dependent protein A
glcF	Frame shift	Glycolate oxidase iron–sulfur subunit; ferridoxin related
aer	Stop codon	Aerotaxis sensor receptor; transducing signals for aerotaxis
ompG	Truncation	Outer membrane protein; forms large channels
yaeG	Stop codon	Regulator of D-galactarate, D-glucarate and D-glycerate metabolism
nagD	Stop codon	*N*-Acetylglucosamine metabolism
fimD	Insertion	Export and assembly of type 1 fimbriae

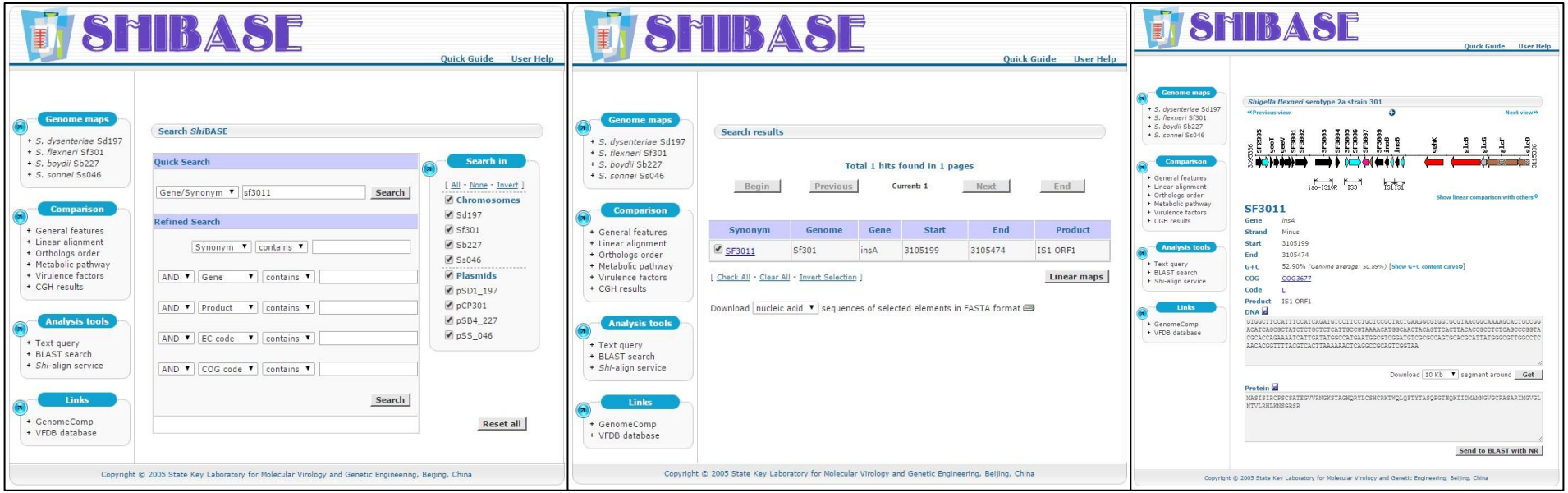
[a]*fdhF* has a stop codon (UAA) in addition to the stop codon UGA used for introducing selenocysteine.

Chinese-operated Database Presents a Viable Source of Genome Information on *Shigella*





*Shi*BASE Provides More Query Options for Finding Specific Genes



Summary

- Combatting Shigellosis is of great concern in developing countries like China.
- Discovering the pathogenicity of *S. flexneri* required the sequencing of its entire genome.
- The research group used softwares in order to efficiently conduct the shotgun sequencing process of *S. flexneri*.
- The results revealed the extreme similarities between 5a and 2a serotypes of *S. flexneri* and between Sf301 and *E. coli* K12.
- A database, called *Shi*BASE, developed by Chinese researchers presents these compiled information to other potential scientists.

Acknowledgments

- Dr. Dahlquist
- Dr. Dionisio
- Biological Database students

Questions?