FEMS Microbiology Letters, 362, 2015, fnv011

doi: 10.1093/femsle/fnv011
Advance Access Publication Date: 28 January 2015
Research Letter

RESEARCH LETTER – Pathogens & Pathogenicity

The genome of *Shigella dysenteriae* strain Sd1617 comparison to representative strains in evaluating pathogenesis

Ajchara A. Vongsawan[1,*], Vinayak Kapatral[2], Benjamin Vaisvil[2], Henry Burd[2], Oralak Serichantalergs[1], Malabi M. Venkatesan[3] and Carl J. Mason[1]

[1]Department of Enteric Diseases, Armed Forces Research Institute of Medical Sciences, Bangkok 10400, Thailand, [2]Igenbio Inc., Chicago, IL 60607, USA and [3]Walter Reed Army Institute of Research, Division of Bacterial and Rickettsial Diseases, Silver Spring, MD 20910, USA

*Corresponding author: Department of Enteric Diseases, Armed Forces Research Institute of Medical Sciences, 315/6 Rajvithi Road, Bangkok 10400, Thailand. ajcharataa@gmail.com, carl.mason@afrims.org, vinayak@igenbio.com

One sentence summary: The *Shigella dysenteriae* strain Sd1617 serotype 1 has been sequenced and analyzed. It is widely used as model strain for vaccine design, trials and research. A combination of next-generation sequencing platforms and assembly yielded two contigs representing a chromosome size of 4.34 Mb and the large virulence plasmid of 177 kb.

Editor: Simon Silver

ABSTRACT

We sequenced and analyzed *Shigella dysenteriae* strain Sd1617 serotype 1 that is widely used as model strain for vaccine design, trials and research. A combination of next-generation sequencing platforms and assembly yielded two contigs representing a chromosome size of 4.34 Mb and the large virulence plasmid of 177 kb. This genome sequence is compared with other *Shigella* genomes in order to understand gene complexity and pathogenic factors.

Keywords: *Shigella dysenteriae*; genome; comparison

INTRODUCTION

Shigella dysenteriae strain Sd1617 serotype 1 is the Gram-negative, facultative anaerobe and is an etiological agent of epidemic *Shigella* dysentery. This organism is adapted to colonize, cause disease in the colon and rectal epithelial tissue of humans and primates (Nie *et al.*, 2006). *Shigella* infection is through the fecal–oral route with an infection dose of 10–100 bacterium capable of inducing shigellosis (DuPont *et al.*, 1989). *Shigella dysenteriae* strain Sd1617 was originally isolated from a 1968 outbreak of epidemic dysentery in Guatemala (Mata *et al.*, 1970; Mendizabal-Morris *et al.*, 1971; Venkatesan *et al.*, 2002). Pandemic *S. dysenteriae* type 1 swept through Central America from 1969 through 1972 mounting to over 112 000 cases and 10 000 deaths in Guatemala (Gangarosa *et al.*, 1970; Mendizabal-Morris *et al.*, 1971). Since then, episodes of periodic outbreaks have occurred with antimicrobial resistance patterns and plasmid profiles similar to the 1969–1972 pandemic strain (Reller *et al.*, 1971; Parsonnet *et al.*, 1989).

The genus *Shigella* is classified into *S. dysenteriae* (Group A), *S. flexneri* (Group B), *S. boydii* (Group C) and *S. sonnei* (Group D) which are distinguished by their O-antigen and biochemical properties (Torres 2004; Li *et al.*, 2009; Martinez-Becerra *et al.*, 2012). Except for *S. sonnei*, the other three groups have multiple serotypes and subtypes. *Shigella dysenteriae* serotype 1 is considered the most virulent. This has led to a resurgence in *Shigella*

Received: 12 January 2015; **Accepted:** 15 January 2015

pathogenesis research and vaccine development strategies (Trofa *et al.*, 1999).

Broadly, *Shigella* genomes share a core backbone highly similar to that of *Escherichia coli* K12. In addition, all virulent *Shigella* spp. contain a large virulence plasmid of ∼180–215 kb, which is essential for pathogenesis. A series of gene deletions and acquisitions on the chromosome, in association with the presence of the virulence plasmid, has resulted in an '*E. coli* pathotype' that is believed to have adapted to a unique microenvironment within the host. (Maurelli *et al.*, 1998; Jin *et al.*, 2002). Bacterial genomes within the same species or clinical isolated strains varies in size due to continuous sequence fluctuation (Schmidt and Hensel 2004) and rearrangement in response to environmental factors leading to genome elasticity and species diversity (Juhas *et al.*, 2009). In this work, we have sequenced the chromosome and large virulence plasmid of *S. dysenteriae* strain Sd1617 and performed comparative analysis with other *Shigella* genomes in order to understand pathogenic features with a long-term goal to develop effective vaccine strains.

MATERIALS AND METHODS

DNA preparation

The *S. dysenteriae* strain Sd1617 was provided by Samuel B. Formal, (Walter Reed Army Institute of Research, USA). Genomic DNA was prepared using standard protocols from a single red colony of strain Sd1617 grown on Congo Red agar plate. This was further cultured in a 2 ml LB grown to log phase at 37°C. The cells were centrifuged and the cell pellet was used for genomic DNA extraction following the Puregene genomic DNA extraction kit (Qiagen Inc., MD, USA). The CosMC extraction kit (Agencourt Inc, Boston, MA, USA) and Qiagen large extraction kit were used to extract plasmid DNA. The plasmid DNA was separately isolated and was verified by agarose gel electrophoresis.

Genome sequencing and annotations

The complete strain 1617 was sequenced using multiple next-generation sequencing strategies. First, a random library was constructed and the 454 sequencing method was used to generate and assemble into 604 contigs. In addition, two rounds of Ion Torrent (316 chip) and Illumina Mi Seq paired-end runs were performed to generate 562 contigs. The genome of *S. dysenteriae* strain 197 genome (PRJNA58213) was used as a reference for genome assembly. Based on the above assemblies, two contigs each representing the *S. dysenteriae* strain 1617 chromosome and plasmid were obtained. The ORFs in *S. dysenteriae* strain 1617 were identified using a combination of Glimmer (v 2.1), CRITICA and Prokpeg (a protein sequence similarity-based ORF caller) as described in Kapatral *et al.* (2002). The reconciled predicted ORFs were integrated into the ERGO annotation environment for computing protein similarities and function identification (Overbeek *et al.*, 2003). The genome features in *S. dysenteriae* strain Sd1617 (SLG) were compared to other draft genomes such as *S. dysenteriae*1617 (WIV: PRJNA59463 is in 67 contigs), *S. dysenteriae* M131649 (SDY: PRJNA346 is in 239 contigs) and fully sequenced *S. dysenteriae* Sd197 (SDS: PRJNA58213 is in 2 contigs) genome.

Genome accession

The genome sequence of *S. dysenteriae* strain Sd1617 and the large virulence plasmid are deposited in GenBank with the accession numbers CP006736 (4.3 Mb) and CP006737 (177 kb), respectively.

RESULTS

Genome analysis

The *S. dysenteriae* strain Sd1617 genome is in two contigs with a total size of ∼4.3 Mb with average GC content of 50%, where as the plasmid varied between 40 and 53% GC. The genome features are given in Table 1. The size of the chromosome is

Table 1. Genome-wide statistics of strain Sd1617.

Sl	Features	No	Percent
1	Contigs	2	–
2	Total nucleotide bases	4 480 198	–
3	ORF coding bases	3 826 144	–
4	RNAs, total	96	–
5	RNAs, transfer	93	–
6	RNAs, ribosomal	3	–
7	ORFs, total	6409	100
8	ORFs, with assigned function	5313	82
9	ORFs, with assigned function with no similarities	42	1
10	ORFs, without assigned function	1204	18
11	ORFs, without assigned function with similarity	1113	17
12	ORFs, in asserted pathways	2421	38
13	ORFs, not in asserted pathways	4090	64
14	ORFs, not in asserted pathways with assigned function	2886	45
15	ORFs, without assigned function with sims	1113	17
16	ORFs, in paralog clusters	3026	47
17	ORFs, in COGs	4358	68
18	ORFs, in Pfam domains	3810	59
19	ORFs, in operons/chromosomal clusters	5830	91
20	ORFs, in possible fusion events	5403	84
21	ORFs in possible fusion events as composite	1820	28
22	ORFs in possible fusion events as component	4815	75

4.3 Mb and the virulence plasmid is 177 kb. A total of 6505 ORFs have been identified including rRNA and tRNA operons on the chromosome and 394 ORFs on the large virulence plasmid. Using the ERGO annotation procedures as described in Overbeek *et al.* (2003), over 82% of ORFs were identified with a functional annotation. However, 67% of ORFs belonged to COG categories (Tatusov *et al.*, 2000) and 59% of the ORFs have pfam domain (Finn *et al.*, 2014) signifying potential functions. A significant number of ORFs (84%) were identified as fusions proteins or frameshifts. Fusion proteins as composite (28%) consisting of two or more fused ORFs and nearly 75% ORFs representing individual components were identified.

Pseudogenes

Pathogenic bacterial genomes in particular carry higher numbers of pseudogenes as identified in *S. flexneri* 2a (Jin *et al.*, 2002; Wei *et al.*, 2003). These pseudogenes are believed to be non-functional genes or may have been generated due to errors in ORF calling (Hansen-Wester and Hensel 2001; Overbeek *et al.*, 2003; Pieper *et al.*, 2009; Figueira *et al.*, 2013). Pseudogenes are also believed to have arisen from mutations due to in-frame stop codons, frameshifts or inaccurate start codon (Gohmann *et al.*, 1994; Pieper *et al.*, 2009). In *S. dysenteriae* strain 1617 genome, 1041 pseudogene ORFs were identified that contained a classical start and stop bacterial codons. 573 pseudogene ORFs do not have any assigned functions and 1113 pseudogene ORFs that have similarities to sequences in the database (conserved hypotheticals). A total of 42 unique pseudogene ORFs (19 on the chromosome and 23 on the virulence plasmid) have no similarities to any sequences in the database. A total of 568 pseudogenes have sequence similarities to other genes with recognizable functional domains. Of these, some are conserved among other bacterial ORFs including *Shigella* spp., *E. coli* and *Salmonella* spp. Among them, hypothetical proteins with no domains (396 ORFs), hypothetical cytosolic proteins (52 ORFs), hypothetical exported proteins (12 ORFs), hypothetical membrane-associated proteins (23 ORFs), hypothetical membrane-spanning proteins (45 ORFs) and a single ORF for hypothetical ATP binding protein were identified. Other hypothetical pseudogenes ORFs with domains such as nucleoside transporter, ParB-nuclease domain, NmrA family protein, STAS domain, RNA-binding protein, pyrimidine permease RutG, YdhiL, YdiH, YacL, etc. were found as well.

Comparative secretion systems

Secretion of proteins across *Shigella* spp. inner and outer membrane is dependent on a variety of secretions systems. Seven bacterial secretion systems have been found that transport proteins, toxins or DNA into other either prokaryotes or eukaryotes (Tseng, Tyler and Setubal 2009; Hayes, Aoki and Low 2010). Six of seven protein secretion pathways have been detected in Gram-negative bacteria and the seventh being usually seen in Gram-positive bacteria (Tseng, Tyler and Setubal 2009). Secretion systems are categorized into contact dependent (T3SS, T4SS, T5SS and T6SS) (Hayes, Aoki and Low 2010) and contact independent (T1SS and T2SS). The comparative secretory systems found in the chromosome (Table S1, Supplementary Information) and plasmid (Table S2, Supplementary Information) are summarized.

The hemolysin toxin type I secretion system (T1SS) encoded by the operon *tolC-hlyD-hlyB* complex in *E. coli* and is partially found in *S. dysenteriae* strain Sd1617 genome. ORFs for hemolysin (HlyA) have not been found; however, sequences for hemolysin type ATP-binding protein, transmembrane subunit, adaptor protein (HylD) and the outer membrane proteins have been identified. Other proteins belonging to this system such as proteases, toxins, metalloproteases and uncoupler proteins have also been identified. The translocase-mediated uncoupler family, which includes multidrug-resistance protein A, B and the outer membrane proteins, is also present. A dedicated T2SS (type II secretion system) has been identified on the chromosome. A total of 11 ORFs dedicated to general secretion proteins except for GspA, GspB, GspN and prepilin peptidase were identified in this genome. Among the secreted proteins via T2SS, enzymes such as alkaline phosphatase, acid phosphatase, phosphatidylglycerophosphatases and others have been identified (Table S1, Supporting Information). In *S. dysenteriae* strain 1617, all the three stages of Sec-dependent system (T2SS) including protein targeting, translocation and maturation proteins are present.

Similarly, the components of twin-arginine translocation (TAT) pathway proteins (TatA, TatB, TatC and TatE) that are necessary for twin-arginine motif-containing proteins such as ABC transporter substrate binding proteins, peroxidase family protein, oxidoreductases, trimethylamine N-oxide reductase are present in all the *Shigella* genomes compared. The virulence plasmid of *S. dysenteriae* strain 1617 contains a contact-dependent secretion system for transferring effectors and toxins into the eukaryotic cells (T3SS). Most of the T3SS genes encode as an operon on a 31 kb invasion-associated region of the virulence plasmid. Other T3SS effectors that are encoded outside this 31 kb entry region include ORFs for SenA and SenB enterotoxins. The *Shigella* spp. T3SS is different compared to other Gram-negative bacteria such as *Salmonella* spp. and *Yersinia* spp., where the pathogenicity islands (PAI) are typically located on the chromosome. For instance, *S. enterica* PAI, SPI1 and SPI2, each encoding a T3SS, are present on the chromosome (Hansen-Wester and Hensel, 2001). In *Shigella* spp., the T3SS encodes the invasion plasmid antigens (Ipa) IpaA, IpaB, IpaC and IpaD as well as more than 20 accessory proteins (*mxi-spa* genes) that make up the needle-like type III secretion apparatus that injects the Ipa proteins and effector proteins into the host cells (Table S3, Supporting Information). The expression of the Ipa and the *mxi-spa* operon is under the control of regulatory proteins VirB (or IpaR) and an AraC-like activator protein VirF which is encoded on the virulence plasmid, which in turn is regulated by the HN-S protein found on the chromosome (Table S1, Supporting Information). Results from proteomic studies with *S. dysenteriae* strain 1617 in gnotobiotic piglets have indicated that there is an increased expression *in vivo* of proteins belonging to the T3SS as compared to *in vitro* growth.

The Osp families of proteins that are known to be involved in the manipulation of the host innate and adaptive immune response are an example of such increased *in vivo* expression (Phalipon and Sansonetti 2007; Pieper *et al.*, 2009; Zurawski *et al.*, 2009). The *S. dysenteriae* strain 1617 genome does not have ORFs for type IV secretion system, but has ORFs for type V secretion system. The type V secretion system uses the Sec system for translocating proteins to cross the inner membrane into the periplasm. In addition, three additional proteins DsbA, DsbD and disulfide bond formation protein B are necessary for specific protein translocation through the outer membrane. The type V autotransporter protein is reminiscent of the IgA1 protease. We have identified in *Shigella* specific adhesins aidA-1, extracellular protease and pertactin precursor proteins (Table S1, Supporting Information). The VirG (IcsA protein) protein on the virulence plasmid is another example of a type V secretion system that is present on the virulence plasmids of all

Shigella spp. and is required for intercellular spreading of the bacteria.

Only γ proteobacteria are known to harbor a type VI secretion system. Orthologs of this protein have been detected in some virulent *Shigella* genomes such as *S. sonnei*, *S. flexneri* and *S. dysenteriae* which is considered to be important in pathogenesis, interesting these are absent in non-virulent bacteria (Dudley *et al.*, 2006; Shrivastava and Mande 2008; Chow and Mazmanian 2010). However, there are no ORFs for proteins associated with type VI secretion systems in *S. dysenteriae* (Strain 1012, 155–74) except for VgrG, ClpV protein and serine-thereonine kinase PrkA homologs. Other sequenced strains such as *S. flexneri* 2a Str 2457T and Str 8401 also have PrkA homologs. There are no ORFs in the type VI secretion system in the any *Shigella* genomes compared.

Motility

Although *Shigella* spp. like other Gram-negative bacteria has flagellar genes for motility and chemotaxis, it is non-motile. ORFs for flagellar-specific transcriptional activators such as FlhD, FlhC, regulatory ORFs for flagellin-specific sigma factor FliA and anti-sigma factor FlgM are present. We have identified ORFs for flagellar basal body rod proteins such as FlgB, FlgC and three ORFs for FlgF, FlgG, FlhA had frameshifts. ORFs for basal body P-ring protein (FlgA), flagellar biosynthetic protein FlhE, FliZ, flagellar L-ring protein FlgH, flagellar P-ring protein FlgI, flagellar protein FliS and FliT were identified. Flagella brake protein YcgR, flagella assembly protein Flk and flagella synthesis protein FlgN (frameshift) are present. Among the hook-associated proteins, FlgE, FlgK (frameshift) and FlgL have also been identified. Hook-basal body complex proteins such as FlhP, FlhO, FliE are absent. Other ORFs for proteins such as FleN, FliL, FliJ or flagellar biosynthetic proteins such as FlhA, FlhB, FlhF, FliP and FliQ have not been identified. None of the ORFs for the three flagellar motor switch proteins such as FliG, FliM and FliN proteins were detected in this genome. Among the Class III flagellar operons, motor proteins (MotA, MotB) and four methyl-accepting chemotaxis proteins have been identified. There are three ORFs for flagellin protein (similar to FliC protein of *Salmonella* spp.) one of them is embedded in an IS element (SLGSEL1 type) and an ORF for flagellar capping protein FliD. It appears that this organism is non-motile despite having several flagellar genes.

Shiga toxins

The ORFs for two-shiga toxins subunit A (StxA) were identified in all four genomes compared in this study whereas the ORF for shiga toxin B chain precursor (StxB) was found in all except in the draft genome *S. dysenteriae* strain Sd 1617 and are flanked by prophages. Other key ORFs for virulence proteins such as cell surface protein (YadA), protease ATP-binding subunit CipA and ATP-dependent endopeptidase subunit ClpP (EC 3.4.21.92), apyrase (EC 3.6.1.5) are present in all the genomes compared. All the genomes have three ORFs for enterotoxins in strain Sd1617 (SLG) draft strain Sd1617 (WIV) and strain Sd197 (SDS); however, only two ORFs were identified in the draft strain M121649 genome. Three ORFs for *N*-acetylmuromoyl-L-alanine amidase (EC 3.5.1.28), involved in hydrolysis of murein, were also found in all the genomes except in the strain Sd1617 draft genome. In addition, three copies on plasmid borne enterotoxin, two copies each of IpaH, MxiG, type III secreted protein SctT and SctV are present on the virulence plasmid (Table S3, Supporting Information).

Lipopolysaccharide (LPS)

LPS is a major component of the proteobacterial outer membrane and is a major virulence factor in all Gram-negative pathogens. Typically, the Gram-negative bacterial LPS consists of three covalently linked moieties; the lipid A region, a conserved core oligosaccharide region and a serotype-specific O-antigen composition of polysaccharide side chains (Hong and Payne 1997). The O-antigen determines host immunological specificity to each serotype and contributes to the design of protective vaccine complexity at the species level (Passwell *et al.*, 2001). Vaccine designs for *S. dysenteriae* targeted at the O-antigens are still under progress and protective vaccines are still unavailable (Venkatesan and Ranallo 2006; Levine *et al.*, 2007; Phalipon, Mulard and Sansonetti 2008). The ORFs for phospholipid-LPS ABC membrane transporter (MsbA) that is involved in outer membrane translocation has been identified in all four *S. dysenteriae* genomes. Similarly, ORFs for inner core biosynthesis proteins of LPS and homologs of KdtA, HtrB and MsbB are also present. A second ORF for MsbB (MsbB2), that is normally present on the virulence plasmid, is embedded within an IS element on the *S. dysenteriae* strain Sd1617 plasmid. Other important ORFs for proteins involved in the LPS such as RfaC, RfaF, RfaG, WaaP, WaaQ and RfaI (LPS 3-alpha-galactosytransferase) were identified in all the *Shigella* genomes compared in this study (Table S1, Supporting Information). However, ORFs for LPS-specific glucosyltransferase I (EC 2.4.1.73) or glucosyltransferase II (EC 2.4.1.58) are absent in this *S. dysenteriae* strain Sd1617, although they are present in strain Sd197 and strain Sd1617 draft genomes. Chromosomal-encoded operons such as *galETKM* and *rfbBDACX* that code for functional enzymes in the biosynthesis of the O-antigens are present in *S. dysenteriae* strain Sd1617 and other *Shigella* strains (Yao and Valvano 1994; Kuntumalla *et al.*, 2011). Five PAIs have been identified in *Shigella* genomes: SHI-0, SHI-1 (she PAI), SHI-2, SHI-3 and SRL and the addition of a PAI-like cluster encoding the *ipa-mix-spa* genes on the virulence plasmid (Torres 2004). The integration and excision of PAIs from the chromosome is integrase dependent (Turner *et al.*, 2001; Luck *et al.*, 2004; Turner *et al.*, 2004) while in *S. flexneri*, strain PAI and the excision from the chromosome is considered to be spontaneous (Rajakumar, Sasakawa and Adler 1997; Turner *et al.*, 2001; Luck *et al.*, 2004). LPS genes are necessary for bacterial resistance to acidic host conditions (Nie *et al.*, 2006). The biosynthesis of the O-antigen is carried out by genes located on the chromosome at the O-antigen gene cluster *wbbP* (*rfpB*) and also on the 9 kb plasmid pHW400 (Gohmann *et al.*, 1994; Feng *et al.* 2007).

The virulence genes and toxins identified in *S. dysenteriae* strain 1617 were compared with other serotypes (Table S3, Supporting Information). Strain Sd1617 shows variation in the expression of its virulence genes in the Osp family. In Sd1617, both closed and draft genomes contain only the OspF, OspG and OspG where OspB and OspE family proteins are absent in strain Sd1617 but present in strain Sd197. Both proteins IpgB1 and IpgB2 which are a part of type III secretion system effector were found present in the closed *S. dysenteriae* strain Sd1617 and absent in the draft genomes. This emphasizes the importance of closing draft strains for *Shigella* genomes. It is interesting to find that the presence of ORF for IpgB2 is in strain Sd197 genome and not complete strain Sd1617 genome. Several ORFs with similarity to invasion proteins necessary for host entry have been

identified, along with other adhesion belonging to NIpC/P60 family proteins in all the *Shigella* genomes.

Outer membrane proteins

ORFs determining outer membrane proteins that play a role both in secretion and pathogenesis have been identified. These outer membrane proteins serve as antigens for vaccine production and are categorized based on domains into adherence, outer membrane adherence protein, outer membrane assembly protein, outer membrane autotransporter domain protein, outer membrane channel protein and porins. ORFs coding for lipoproteins such as outer membrane lipoprotein (Lpp), Blc, Pcp, SmpA and lipoprotein carried protein LolA and LolB have been identified in the fully sequenced strain Sd1617. Determinants of other outer membrane proteins present in strain Sd1617 include those for OmpA, OmpC, OmpF, OmpN, OmpH, OmpW, YopM, Slp precursors, assembly factor family protein YaeT and siderophore receptor protein. Several members of the outer membrane usher protein family such as FimD, PapC and SefC have been identified. The orthologs of these proteins have been found in other genomes as well.

The ORF for outer membrane protein OmpA in strain Sd1617 genome is located on the chromosome downstream of an IS element (SLGSEL1014). The translational start site of OmpA protein in other *S. dysenteriae* is 72 nt downstream in WIV and WOP genomes compared to SLG genome. However, the *ompA* gene in SLG genome is identical to other enteric orthologs such as *E. coli* genome. Similarly, the ORF for OmpC protein is found in all the closed and draft *Shigella* genomes compared; however, in strain Sd1617, there is an IS element (SLGSEL1003) upstream of this ORF (Fig. 1). Other ORFs coding for OmpF, OmpH and OmpN proteins do not have insertion elements. It is not clear whether the IS elements located adjacent to any ORF gene can relocate it to a different location in strain Sd1617 for ORFs coding OmpA or OmpC or if the insertion elements located upstream can disrupt the gene function altogether. In strain Sd1617, IS upstream in the promoter region of ORF *ompA* could be detrimental to the gene function (Fig. 1). The Omps have homology to each other and to other outer membrane proteins such as porins. The OmpA protein is an ~34 kd in size with versatile function. One function is the pathogen-associated molecular patterns through the TLR-2 receptor-mediated signaling to induce the host innate immune signaling cascade of inflammatory responses via the NF-kappa B release of proinflammatory cytokines which in turn recruit the host cell mediated or adaptive immune response as seen in *S. flexneri* (Pore *et al.*, 2010). We also have identified four ORFs for uncharacterized cell surface proteins in tandem and perhaps play a role in immunogenicity, which is present in all *Shigella* genomes compared.

Iron acquisition

Iron is key for survival for *Shigella* infection and it encounters limited iron availability especially in the human host environment. Multiple dedicated systems for iron acquisition Iro (IroB, IroC, IroD and IroE proteins) and a dedicated iron ABC transporter system SitA, SitB, SitC, SitD and SitE proteins (Table S3, Supporting Information). ORFs for ferrous iron transport uptake protein A and protein B are found in all the *Shigella* genomes. In addition, other iron uptake systems such as heme-binding proteins (ShuS and ShuT), iron di-citrate transport system proteins (FecA, FecB, FecC, FecD and FecC) are also present for additional iron uptake under iron-limiting conditions (Runyen *et al.*, 2003).

Insertion sequence

Seventeen types of IS elements based on nature of sequences and length were identified in strain Sd1617 and compared various types of IS elements with other *Shigella*-sequenced genomes. Strain Sd1617 contains the highest number of IS elements compared to other *Shigella* genomes strain (Yang *et al.*, 2005; Nie *et al.*, 2006). Within the interspecies, it is clear that Sd1617 is significantly different in IS composition than Sd197 and SdM131649 (Table 2). In strain Sd1617, the unknown types of IS elements are 2-fold higher than other genomes. Interestingly, the average length of IS1 family is smaller in strain Sd1617 compared to strain Sd197 or strain SdM131649. In strain Sd1617, the majority of known IS elements are skewed towards IS1 and IS3 analogous to the IS distribution seen in strain Sd197 serotype. IS605 is present in all compared *Shigella* spp. genomes in one copy of 1206 bp long where IS3/IS481 is also present, show a minor variation in their length with the strain Sd1617 and strain Sd197 chromosomes containing 405 bp compared to the draft strain SdM131649 and Sd1617 both contain 368 bp in length (Table 2). The frequency of shorter IS lengths in strain Sd1617 complete genome is perhaps due to recombination/excision of selective IS element families but not other IS elements.

Plasmid maintenance

Several toxin–anti-toxin pairs necessary for plasmid stabilization and plasmid/chromosome segregation are present on the *Shigella* spp. chromosome. A post-segregation toxin and anti-toxin and two pairs Yee-toxin and Yee anti-toxin have been identified in this genome. In addition, a second pair of post-segregation toxin and anti-toxin is present on the plasmid. It is not clear what the roles of these genes are other than plasmid maintenance, stabilization and segregation. On the plasmid contig, both partition proteins ParA and ParB have been identified. For plasmid post-segregation processes, three ORFs encoding microcin process peptidase 1, and endopeptidase including

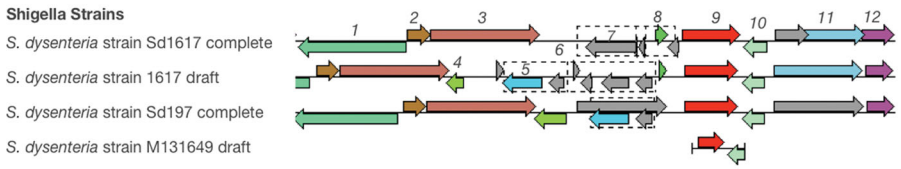

Figure 1. The ORF OmpA is located in an IS element region in all *Shigella* genomes. Similar colored ORFs have identical functions. (**1**) The IS elements are shown in dashed boxes. The ORFS are (1) helicase, (**2**) membrane protein, (**3**) hypothetical protein, (**5–7**) transposase, (**8**) hypothetical protein, (**9**) OmpA, (**10**) YcgB protein, (**11**) Lon protease (frameshift in strain Sd 1917) and (**12**) FabA protein.

Table 2. Occurrence and types of various IS elements types categorized in *Shigella* spp. The highest number of IS elements were identified in closed strain Sd1917 signifying genome plasticity.

		Strain Sd1617		Strain Sd197		Strain SdM131649		Strain Sd1617	
Sl	Types of IS elements	No	Length	No	Length	No	Length	No	Length
1	IS1	831	221	625	334	688	337	741	307
2	IS110	9	490	8	868	4	609	7	819
3	IS2 repressor TnpA	38	276	33	373	12	386	35	361
4	IS2 transposase TnpB	47	458	33	780	34	746	34	810
5	IS200	2	433	2	427	2	403	1	525
6	IS21	9	489	5	708	4	725	2	319
7	IS256	4	552	5	527	5	513	3	562
8	IS3	146	556	196	477	170	469	186	500
9	IS4	16	647	2	585	1	567	4	501
10	IS5	5	577	20	899	18	971	16	785
11	IS605	1	1206	1	1206	1	1206	1	1206
12	IS66	11	544	10	583	10	579	11	464
13	IS91	4	1089	6	951	5	898	4	1277
14	ISL3	3	347	2	375	1	330	0	0
15	Unknown transposase	663	171	107	268	32	353	331	217
16	IS3/IS481	1	405	1	405	1	369	1	369
17	ISTn3	0	0	0	0	1	2964	0	0
	Total	1790		1056		989		1377	

independent two pairs of post segregation toxin/anti toxins have been identified. These pairs are found in other *Shigella* genomes as well.

DISCUSSION

Whole-genome sequencing and comparative analysis of *Shigella* strains provide an understanding of strain-specific genes and their possible role in physiology and pathogenesis. Using multiple next-generation sequencing methods, we have fully sequenced a pandemic *S. dysenteriae* strain Sd 1617. This genome appears to have a reduced genome size compared to other *Shigella* serotypes; however, it has retained all the essential genes necessary for survival and the virulence plasmid necessary for pathogenesis. Even though many virulence genes are allocated to the large virulence plasmid, strains Sd1617 as well as other virulent serotypes appear to reserve their set of chromosomal genes from the shared core genes that induce shigellosis. The mobility of insertion sequence elements across bacterial genomes and within the same bacterial genome can progressively facilitate genome plasticity, survival and virulence traits. Previously identified IS families are distributed throughout the chromosome and large virulence plasmid of which several transposable elements are of unknown origin.

Studies comparing vaccine strains to the parental strains could help elucidate key mechanisms utilized for survival and thus in turn could greatly assist in vaccine design strategies. Humans are the natural host for *Shigella* infection. Several animal models have revealed important proteins involved in a host–pathogen interaction; however, differences in pathogenesis are largely attributed to the differences in host immune cell population. Comparative gene expression studies pre- and post-host exposure could help in understanding critical genes necessary for survival within host and further facilitate gene selection for knock-out strains as attenuated vaccine candidates. The complete genome sequence of *S. dysenteriae* strain Sd 1617 further facilitates gene and protein expression studies, which aid in efficient vaccine development.

SUPPLEMENTARY DATA

Supplementary data is available at FEMSLE online.

ACKNOWLEDGEMENTS

The authors thank Stephen Baker for critical reading of the manuscript, Ryan Ranallo and Melanie Melendrez for their constructive comments and Akamol Suvarnapunya for suggestions on DNA extraction kits. They also thank Piyarat Pootong, Orntipa Sethabutr, and Ladaporn Bodhidatta for their support.

FUNDING

This work was supported by the National Institute of Allergy and Infectious Diseases (NIAID), under IAA # Y1-AI-4906–03, and by the US Army Medical Research and Materiel Command (USAMRMC).

Conflict of interest statement. None declared.

REFERENCES

Chow J, Mazmanian SK. A pathobiont of the microbiota balances host colonization and intestinal inflammation. *Cell Host Microbe* 2010;**7**:265–76.

Dudley EG, Thomson NR, Parkhill J, *et al*. Proteomic and microarray characterization of the AggR regulon identifies a pheU pathogenicity island in enteroaggregative *Escherichia coli*. *Mol Microbiol* 2006;**61**:1267–82.

DuPont HL, Levine MM, Hornick RB, *et al*. Inoculum size in shigellosis and implications for expected mode of transmission. *J Infect Dis* 1989;**159**:1126–8.

Feng L, Perepelov AV, Zhao G, *et al*. Structural and genetic evidence that the *Escherichia coli* O148 O antigen is the precursor of the *Shigella dysenteriae* type 1 O antigen and identification of a glucosyltransferase gene. *Microbiology* 2007;**153**:139–47.

Figueira R, Watson KG, Holden DW, *et al*. Identification of *Salmonella* pathogenicity island-2 type III secretion system

effectors involved in intramacrophage replication of *Salmonella enterica* serovar typhimurium: implications for rational vaccine design. *MBio* 2013;**4**:e00065.

Finn RD, Bateman JC, Eberhardt RY, *et al*. The pfam protein families' database. *Nucleic Acids Res* 2014;**42**:222–30.

Gangarosa EJ, Perera DR, Mata LJ, *et al*. Epidemic Shiga bacillus dysentery in Central America. II. Epidemiologic studies in 1969. *J Infect Dis* 1970;**122**:181–90.

Gohmann S, Manning PA, Alpert CA, *et al*. Lipopolysaccharide O-antigen biosynthesis in *Shigella dysenteriae* serotype 1: analysis of the plasmid-carried rfp determinant. *Microb Pathog* 1994;**16**:53–64.

Hansen-Wester I, Hensel M. *Salmonella* pathogenicity islands encoding type III secretion systems. *Microbes Infect* 2001;**3**: 549–59.

Hayes CS, Aoki SK, Low DA. Bacterial contact-dependent delivery systems. *Annu Rev Genet* 2010;**44**:71–90.

Hong M, Payne SM. Effect of mutations in *Shigella flexneri* chromosomal and plasmid-encoded lipopolysaccharide genes on invasion and serum resistance. *Mol Microbiol* 1997;**24**:779–91.

Jin Q, Yuan Z, Xu J, *et al*. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Res* 2002;**30**:4432–41.

Juhas M, van der Meer JR, Gaillard M, *et al*. Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol Rev* 2009;**33**:376–93.

Kapatral V, Anderson I, Ivanova N, *et al*. Genome sequence and analysis of the oral bacterium *Fusobacterium nucleatum* strain ATCC 25586. *J Bacteriol* 2002;**184**:2005–18.

Kuntumalla S, Zhang Q, Braisted JC, *et al*. In vivo versus in vitro protein abundance analysis of *Shigella dysenteriae* type 1 reveals changes in the expression of proteins involved in virulence, stress and energy metabolism. *BMC Microbiol* 2011;**11**:147.

Levine MM, Kotloff KL, Barry EM, *et al*. Clinical trials of *Shigella* vaccines: two steps forward and one step back on a long, hard road. *Nat Rev Microbiol* 2007;**5**:540–53.

Li Y, Cao B, Liu B, *et al*. Molecular detection of all 34 distinct O-antigen forms of *Shigella*. *J Med Microbiol* 2009;**58**:69–81.

Luck SN, Turner SA, Rajakumar K, *et al*. Excision of the Shigella resistance locus pathogenicity island in *Shigella flexneri* is stimulated by a member of a new subgroup of recombination directionality factors. *J Bacteriol* 2004;**186**:5551–4.

Martinez-Becerra FJ, Kissmann JM, Diaz-McNair J, *et al*. Broadly protective *Shigella* vaccine based on type III secretion apparatus proteins. *Infect Immun* 2012;**80**:1222–31.

Mata LJ, Gangarosa EJ, Caceres A, *et al*. Epidemic Shiga bacillus dysentery in Central America. I. Etiologic investigations in Guatemala, 1969. *J Infect Dis* 1970;**122**:170–80.

Maurelli AT, Fernandez RE, Bloch CA, et al. "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella spp*. and enteroinvasive *Escherichia coli*. *P Natl Acad Sci USA* 1998;**95**:3943–8.

Mendizabal-Morris CA, Mata LJ, Gangarosa EJ, *et al*. Epidemic Shiga-bacillus dysentery in Central America. Derivation of the epidemic and its progression in Guatemala, 1968–69. *Am J Trop Med Hyg* 1971;**20**:927–33.

Nie H, Yang F, Zhang X, *et al*. Complete genome sequence of *Shigella flexneri* 5b and comparison with Shigella flexneri 2a. *BMC Genomics* 2006;**7**:173.

Overbeek R, Larsen N, Walunas T, *et al*. The ERGO genome analysis and discovery system. *Nucleic Acids Res* 2003;**31**: 164–71.

Parsonnet J, Greene KD, Gerber AR, *et al*., *Shigella dysenteriae* type 1 infections in US travellers to Mexico, 1988. *Lancet* 1989;**2**:543–5.

Passwell JH, Harlev E, Ashkenazi S, *et al*. Safety and immunogenicity of improved *Shigella* O-specific polysaccharide-protein conjugate vaccines in adults in Israel. *Infect Immun* 2001;**69**:1351–7.

Phalipon A, Mulard LA, Sansonetti PJ. Vaccination against shigellosis: is it the path that is difficult or is it the difficult that is the path? *Microbes Infect* 2008;**10**:1057–62.

Phalipon A, Sansonetti PJ. *Shigella's* ways of manipulating the host intestinal innate and adaptive immune system: a tool box for survival? *Immunol Cell Biol* 2007;**85**:119–29.

Pieper R, Zhang Q, Parmar PP, *et al*. The *Shigella dysenteriae serotype 1* proteome, profiled in the host intestinal environment, reveals major metabolic modifications and increased expression of invasive proteins. *Proteomics* 2009;**9**:5029–45.

Pore D, Mahata N, Pal A, *et al*. 34 kDa MOMP of Shigella flexneri promotes TLR2 mediated macrophage activation with the engagement of NF-kappaB and p38 MAP kinase signaling. *Mol Immunol* 2010;**47**:1739–46.

Rajakumar K, Sasakawa C, Adler B. 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* 1997;**65**:4606–14.

Reller LB, Rivas EN, Masferrer R, *et al*. Epidemic shiga-bacillus dysentery in Central America. Evolution of the outbreak in El Salvador, 1969–70. *Am J Trop Med Hyg* 1971;**20**:934–40.

Runyen LJ, Reeves SA, Gonzales EG, *et al*. Contribution of the *Shigella flexneri* Sit, Iuc, and Feo iron acquisition systems to iron acquisition in vitro and in cultured cells. *Infect Immun*.2003;**71**:1919–28.

Schmidt H, Hensel M. Pathogenicity islands in bacterial pathogenesis. *Clin Microbiol Rev* 2004;**17**:14–56.

Shrivastava S, Mande SS. Identification and functional characterization of gene components of Type VI Secretion system in bacterial genomes. *PLoS One* 2008;**3**:e2955.

Tatusov RL, Galperin MY, Natale DA, *et al*. The COG database:a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 2000;**28**:33–6.

Torres AG. Current aspects of Shigella pathogenesis. *Rev Latinoam Microbiol* 2004;**46**:89–97.

Trofa AF, Ueno-Olsen H, Oiwa R, *et al*. Dr. Kiyoshi Shiga: discoverer of the dysentery bacillus. *Clin Infect Dis* 1999;**29**: 1303–6.

Tseng TT, Tyler BM, Setubal JC. Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology. *BMC Microbiol* 2009;**9** Suppl 1:S2.

Turner SA, Luck SN, Sakellaris H, *et al*. Nested deletions of the SRL pathogenicity island of *Shigella flexneri* 2a. *J Bacteriol* 2001;**183**:5535–43.

Turner SA, Luck SN, Sakellaris H, *et al*. Role of attP in integrase-mediated integration of the Shigella resistance locus pathogenicity island of Shigella flexneri. *Antimicrob Agents Ch* 2004;**48**:1028–31.

Venkatesan MM, Hartman AB, Newland JW, *et al*. Construction, characterization, and animal testing of WRSd1, a *Shigella dysenteriae 1* vaccine. *Infect Immun* 2002;**70**:2950–8.

Venkatesan MM, Ranallo RT. Live-attenuated Shigella vaccines. *Expert Rev Vaccines* 2006;**5**:669–86.

Wei J, Goldberg MB, Burland V, *et al*. Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infect Immun* 2003;**71**:2775–86.

Yang F, Yang J, Zhang X, *et al*. Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res* 2005;**33**: 6445–58.

Yao Z, Valvano MA. Genetic analysis of the O-specific lipopolysaccharide biosynthesis region (rfb) of Escherichia coli K-12 W3110: identification of genes that confer group 6 specificity to *Shigella flexneri* serotypes Y and 4a. *J Bacteriol* 1994;**176**:4133–43.

Zurawski DV, Mumy KL, Faherty CS, *et al*., *Shigella flexneri* type III secretion system effectors OspB and OspF target the nucleus to downregulate the host inflammatory response via interactions with retinoblastoma protein. *Mol Microbiol* 2009;**71**: 350–68.