BMC Genomics

RESEARCH ARTICLE Open Access

Complete genome sequence and annotation of the laboratory reference strain *Shigella flexneri* serotype 5a M90T and genome-wide transcriptional start site determination

Ramón Cervantes-Rivera[1,2,3], Sophie Tronnet[1,2,3] and Andrea Puhar[1,2,3*]

Abstract

Background: *Shigella* is a Gram-negative facultative intracellular bacterium that causes bacillary dysentery in humans. *Shigella* invades cells of the colonic mucosa owing to its virulence plasmid-encoded Type 3 Secretion System (T3SS), and multiplies in the target cell cytosol. Although the laboratory reference strain *S. flexneri* serotype 5a M90T has been extensively used to understand the molecular mechanisms of pathogenesis, its complete genome sequence is not available, thereby greatly limiting studies employing high-throughput sequencing and systems biology approaches.

Results: We have sequenced, assembled, annotated and manually curated the full genome of *S. flexneri* 5a M90T. This yielded two complete circular contigs, the chromosome and the virulence plasmid (pWR100). To obtain the genome sequence, we have employed long-read PacBio DNA sequencing followed by polishing with Illumina RNA-seq data. This provides a new hybrid strategy to prepare gapless, highly accurate genome sequences, which also cover AT-rich tracks or repetitive sequences that are transcribed. Furthermore, we have performed genome-wide analysis of transcriptional start sites (TSS) and determined the length of 5′ untranslated regions (5′-UTRs) at typical culture conditions for the inoculum of in vitro infection experiments. We identified 6723 primary TSS (pTSS) and 7328 secondary TSS (sTSS). The *S. flexneri* 5a M90T annotated genome sequence and the transcriptional start sites are integrated into RegulonDB (http://regulondb.ccg.unam.mx) and RSAT (http://embnet.ccg.unam.mx/rsat/) databases to use their analysis tools in the *S. flexneri* 5a M90T genome.

Conclusions: We provide the first complete genome for *S. flexneri* serotype 5a, specifically the laboratory reference strain M90T. Our work opens the possibility of employing *S. flexneri* M90T in high-quality systems biology studies such as transcriptomic and differential expression analyses or in genome evolution studies. Moreover, the catalogue of TSS that we report here can be used in molecular pathogenesis studies as a resource to know which genes are transcribed before infection of host cells. The genome sequence, together with the analysis of transcriptional start sites, is also a valuable tool for precise genetic manipulation of *S. flexneri* 5a M90T. Further, we present a new hybrid strategy to prepare gapless, highly accurate genome sequences. Unlike currently used hybrid strategies combining long- and short-read DNA sequencing technologies to maximize accuracy, our workflow using long-read DNA sequencing and short-read RNA sequencing provides the added value of using non-redundant technologies, which yield distinct, exploitable datasets.

Keywords: *Shigella flexneri* serotype 5a M90T, Genome, Transcriptional start sites, TSS, Chromosome, Virulence plasmid, pWR100, Pseudogene, Insertion sequence, RegulonDB, RSAT

* Correspondence: andrea.puhar@umu.se
[1]The Laboratory for Molecular Infection Medicine Sweden (MIMS), 901 87 Umeå, Sweden
[2]Umeå Centre for Microbial Research (UCMR), 901 87 Umeå, Sweden
Full list of author information is available at the end of the article

Background

Shigella is an enteroinvasive Gram-negative bacterium that causes shigellosis or bacillary dysentery in humans. *Shigella* is responsible for significant morbidity and mortality, particularly in young children and immunocompromised adults [1, 2]. In 2010, around 188 million cases of shigellosis occurred globally, including 62.3 million cases in children younger than 5 years [3–5]. A vast majority of the disease burden due to *Shigella* spp. can be attributed to *S. flexneri* in the developing world and to *S. sonnei* in more industrialized regions [1].

S. flexneri has a low infection dose of only 10 to 100 bacteria [6]. *Shigella* causes disease by invading the colonic mucosa, resulting in an intense acute inflammatory response. The bacterium spreads via the fecal-oral route upon ingestion of contaminated food or water and also via person-to-person contact [7].

S. flexneri 5a M90T, along with *S. flexneri 2a,* is one of the most commonly employed laboratory reference strain for *S. flexneri* in many independent research groups across the globe [8–27]. Indeed, much of our knowledge of the molecular mechanisms of *Shigella* pathogenesis has been obtained using *S. flexneri* M90T as a model. The genome of this strain is composed of a circular chromosome and a megaplasmid (virulence plasmid), called pWR100 [25].

The pathogenesis of *Shigella* spp. strictly depends on the virulence plasmid, which encodes several factors that are essential for invasion and subversion of host defenses [28]. So far, chromosomally encoded genes have received little attention, as most *Shigella* research has focused on the plasmid-encoded virulence genes. However, some of the genes encoded on the chromosome may play an important role in *Shigella* pathogenesis. For instance, transfer of chromosomal DNA from *S. flexneri* 5a M90T into commensal *E. coli* followed by phenotyping during infection allowed the identification of the *his*, *purE* and *arg-mtl* loci that are required for the full inflammatory reaction [29, 30]. Similarly, in vivo phenotyping a deletion mutant of *shiA*, a gene encoded within the chromosomal SHI-2 pathogenicity island, was found to attenuate inflammation [31]. Genome comparison in *S. flexneri* 5a M90T had previously revealed the presence of SHI-2, which further encodes genes necessary to virulence such as the aerobactin siderophore system and *colV* necessary to colicin synthesis [32]. The use of In Vitro Expression Technology (IVET) lead to the discovery of several chromosomal genes that are overexpressed intracellularly in *S. flexneri* 5a M90T [33]. Recently, differential expression analysis by RNAseq during anaerobiosis, an important environmental cue encountered by *Shigella* in the gut lumen [34], highlighted several regulated chromosomal genes [35]. Many more chromosomal genes contributing to virulence were reported in other *S. flexneri* strains. For example, a screen of *S. flexneri* 2a SA100 chromosomal fragments fused to promoterless *gfp* revealed a wealth of metabolic genes that are overexpressed intracellularly [36], which were characterized in depth in several follow up studies. A microarray screen performed on intracellular *S. flexneri* 2a 2457 T identified *icgR*, which regulates bacterial growth within the cytosol of epithelial cell [37]. The same strain was found to secrete a protein encoded by the chromosomal gene *pic*, which is necessary for enterotoxin-induced watery diarrhea [38].

Due to its prime importance, the virulence plasmid was one of the first genomic elements to be sequenced, at least partially, in *S. flexneri* 5a M90T; a major breakthrough at the time [28, 39]. The virulence plasmid was later renamed pWR501 [28, 39]. The *S. flexneri* 5a M90T chromosome has also been sequenced and assembled earlier [40], but this sequence is not complete, as it is only reported as a genome scaffold with many gaps. Moreover, the sequence assembly and annotation was based on another *S. flexneri* strain, *S. flexneri* serotype 5b 8401 [41]. Taken together, the currently available hybrid genome is composed of a chromosome sequence draft [40] and the pWR501 sequence [28, 39] that were sequenced independently. To better understand the pathogenic mechanisms and to identify the genetic elements that are involved in pathogenicity and its regulation, it is essential to have a fully sequenced and annotated genome.

Transcriptomic analysis has been increasingly employed to dissect the molecular mechanisms of host-pathogen interactions for a wide range of bacteria [42–45]. However, only few studies employing RNA-seq have been carried out in *Shigella* [23, 35, 46]. The lack of a *S. flexneri* 5a M90T high-quality genome for transcriptome data analysis is a hinderance, leading to poor reads alignment in our experience. Thus, the availability of the annotated full genome of *S. flexneri* serotype 5a strain M90T paves the way to use this model organism for molecular pathogenesis studies by transcriptome analysis. Taken together, in spite of the wealth of molecular pathogenesis data obtained with *S. flexneri* 5a M90T, we are still in need of a complete and high-quality genome sequence for this strain [23].

Genes in prokaryotic cells can have more than one transcriptional start site (TSS). Typically, transcription starts at position − 20/− 40 from the first translatable codon in bacteria [47]. However, it is already known that in many bacteria the TSS is variable, depending on the environment. Further, it is also known that TSS vary depending on how bacteria respond to a specific stimulus [48]. Knowing the operon and gene structure is essential to understand gene expression and regulation. Hence, the determination of the TSS is one of the first steps in understanding the molecular mechanisms that are implicated in gene regulation.

Primary transcripts of prokaryotes carry a triphosphate at their 5′-ends. In contrast, processed or degraded RNAs only carry a monophosphate at their 5′-ends [49]. The differential RNA-seq (dRNA-seq) approach used here exploits the properties of a 5′-monophosphate-dependent exonuclease (TEX) to selectively degrade processed transcripts, thereby enriching for unprocessed RNA species carrying a native 5′-triphosphate [49]. TSS can then be identified by comparing TEX-treated and untreated RNA-seq libraries, where TSS appear as localized maxima in coverage enriched upon TEX-treatment [42].

Here we present the full, high-quality, and annotated genome of *S. flexneri* 5a M90T. Furthermore, we identified the genes that are expressed during mid-exponential growth in TSB, the typical condition used for in vitro infections with *Shigella*. In addition, we determined the active TSS during mid-exponential growth in TSB and the length of 5′-UTR regions.

Results

Complete and gapless genome assembly of *S. flexneri* 5a M90T

To determine the genome sequence of *S. flexneri* serotype 5a strain M90T whole-genome sequencing was conducted with 3-cell sequencing in a PacBio single-molecule real-time (SMRT) sequencing system [50]. This generated a raw output of 93,316 subreads with mean length of 8387 bp and the longest read of 12,275 bp. The sequences totaled 782,710,041 bp, which corresponds to ~ 157-fold genome coverage. This coverage is high enough to avoid any possible sequencing error.

Genome assembly was carried out with Canu/1.7 [51], feeding PacBio raw data. This assembly generated two contigs without any gap and suggested circular replicons. For the larger contig, the output from Canu retained 14, 193 reads of 5938 bp average read length, with a total contig length of 4,596,714 bp (Fig. 1a and Table 1), indicating that this contig corresponds to the chromosomal replicon. For the smaller contig, Canu retained 1491 reads of 5938 bp average read length, with a total length of 232,195 bp (Fig. 1b and Table 2). The small size of this replicon suggested that it corresponds to the virulence plasmid. These two replicons roughly correspond to the expected size for the chromosome and virulence plasmid of *S. flexneri* 5a M90T, in accordance with previous reports [28, 39, 40].

Polishing of genome assembly using RNA-seq reads

We employed reads from RNA-seq experiments performed on an Illumina HiSeq2000 system to polish the assembled genome. For the first round of polishing, we used the BWA software [53, 54] to align with the assembled genome the reads generated from a library in which the rRNA was depleted with RiboZero (RNAseq-RZ). This step allowed us to polish all the transcribed regions, independently of post-transcriptional processing, as with this method of rRNA depletion all other classes of RNAs are retained. The resulting alignment was used to feed Pilon/1.22 [55] for a first round of iterative genome assembly polishing. The second round of polishing was performed with the dataset generated with RNA from which the rRNA was depleted with 5′-phosphate-dependent Exonuclease (RNAseq-TEX). The polishing process was stopped when no further changes were observed in the Pilon output. This reiterative polishing allowed to correct 140 errors in the first round and 59 errors in the second round. Both obtained replicons were gap-free and circular molecules (Fig. 1). The total coverage of the genome with a depth of ≥5 was 98.77% with a mean coverage of 989.9X for the RNA-seq reads, indicating that polishing genome sequences with RNA-seq reads is an approach that can correct mistakes efficiently since there are no major gaps in the coverage (Figure S1). A comparative alignment with previously published DNA small reads obtained by Illumina sequencing of *S. flexneri* 5a M90T [35] showed that the coverage was 99.98% with a depth of ≥5 with mean coverage 126X and a more evenly distributed coverage throughout the genome with respect to RNA-seq reads (Figure S1). Taken together, these analyses show that polishing a genome assembled from long-read DNA sequences with either DNA or RNA Illumina short-read sequences can yield very good results. However, the hybrid workflow presented here provides the added value that it employs non-redundant techniques yielding distinct datasets (genomic sequences and transcriptomic data), which can be further used for other purposes, thereby maximizing the research output.

Genome structure comparison

To examine the genome structure among *S. flexneri* genomes, we performed genome-wide alignments with the Mauve alignment tool [56] of the three available complete chromosome sequences of *S. flexneri* (*S. flexneri* 2a 301: NC_004337, *S. flexneri* 5 8401: NC_008258 and *S. flexneri* 5a M90T: NZ_CP037923), with *S. flexneri* 2a set as reference. Because unfortunately the virulence plasmid sequence of *S. flexneri* 5 8401 is not available, for the virulence plasmid comparison we used only two sequences (pCP301 from *S. flexneri* 2a 301: NC_004851 and pWR100 from *S. flexneri* 5a M90T: NZ_CP037924), with pCP301 set as reference. We identified a high number of homologous genomic regions in the compared chromosome and plasmid sequences, shown as boxes of the same color (Fig. 2a). For the chromosomes, not many major loss or insertions of regions were found, but the alignment showed a high degree of genome reshuffling and several recombination events. In contrast, for the virulence plasmid several non-homologous regions, seen as empty line or boxes, were identified (Fig. 2b).

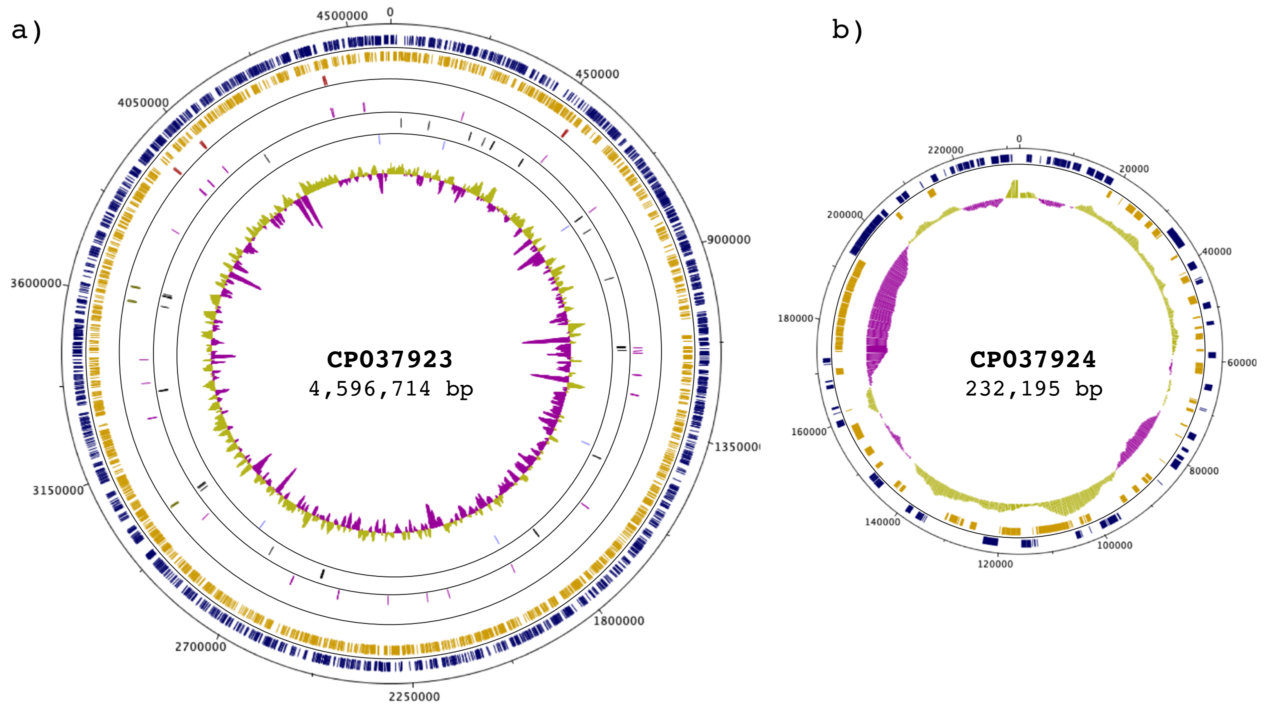

Fig. 1 Circular map of the *S. flexneri* 5a M90T genome. The genome is composed of a chromosome (CP037923) and one plasmid (CP037924). The outermost ring represents the nucleotide position (continuous, black). The two following rings within the scale ring depict coding regions (CDSs) in the forward (blue) and reverse (yellow) strand. Moving towards the center, the next rings depict the rRNA in the forward (red) and reverse (green) strand, followed by rings showing the tRNAs in the forward (purple) and reverse (black) strands. The next ring depicts ncRNAs in both strands (light blue). The following ring shows regulatory elements on both strands (light green). The innermost ring shows the GC content. The figure was generated by Circular-Plot [52]

Gene prediction and functional annotation

Gene prediction and annotation was carried out using three different pipelines: RAST [57], Prokka [58] and Prokaryotic Genome Annotation Pipeline (PGAP)/NCBI [59]. For subsequent analysis, we selected the PGAP/NCBI annotation. However, gene annotations with RAST and Prokka are available as Supplementary Information in the GenBank format (Table S1 and File S1, S1.1, S2 and S2.1). The total number of predicted genes was 4996, of which 769 are pseudogenes (frameshifted = 406, incomplete = 305, internal stop = 166 and multiple problems = 103). From the 769 pseudogenes, 640 were predicted on the chromosome and 129 on the virulence plasmid (Table 1 and Table 2).

Our data showed that *S. flexneri* 5a M90T has a high number of pseudogenes (see Table 1 and Table 2 for the number of pseudogenes and Table S4 for a complete list of pseudogenes) and insertion sequences (IS) (Table 3). In the genome of *S. flexneri* 5a M90T that we are reporting there are 13 different families of IS on the chromosome and 15 families on the virulence plasmid (Table 3). Pseudogenes are defined as fragments of once-functional genes that have been silenced by one or more nonsense, frameshift or missense mutation [60]. Pseudogenes can be the result of errors during the replication process or the effect of IS that shift the open reading frame and modify the DNA sequence. The silencing of the genes can be at two different levels, a) Transcriptional or b) Translational. We verified the expression of the identified pseudogenes, both in the chromosome and in the plasmid, using our RNA-seq data (described later). Our results show that 99% of all identified pseudogenes are transcribed, many highly, indicating that their inactivation did not occur at transcriptional level at least (Fig. 3). The *S. flexneri* 5a M90T annotated genome sequence is integrated into RegulonDB [61](http://regulondb.ccg.

Table 1 General features of the *S. flexneri* 5a M90T chromosome compared with the sequence and annotation of the previous versions

Accession number GenBank	Length (bp)	Genes	CDSs	rRNA	tRNA	ISs	Pseudogenes	Reference
CM001474	4,580,866	4605	4013	22	99	385	197	Onodera, N. T. et al., 2012 [40]
CP037923	4,596,714	4049	4629	22	102	296	640	This work

Table 2 General features of the *S. flexneri* 5a M90T virulence plasmid compared with the sequence and annotation of the previous versions

Accession number GenBank	Length (bp)	Genes	CDSs	ISs	Pseudogenes	Reference
NC_024996	213,494	104	104	22	5	Buchrieser, C. et al., 2000 [28].
AF348706	221,851	294	293	153	0	Venkatesan, M. M. et al., 2001 [39].
CP037924	232,195	307	320	106	129	This work

unam.mx) and RSAT [62] (http://embnet.ccg.unam.mx/rsat/) databases to use their analysis tools in the *S. flexneri* 5a M90T genome.

Whole-genome transcriptional start site determination

To obtain differential RNA-seq (dRNA-seq) data, RNA samples were prepared from triplicate *S. flexneri* 5a M90T cultures grown in TSB at 37 °C and 150 RPM until $OD_{600} = 0.3$. This resulted in a dataset of ~ 120 million reads mapped to the genome of *S. flexneri* 5a M90T presented in this work (GenBank accession no. CP037923 and CP037924). A total of 14,051 TSSs (Fig. 4) were automatically annotated with ReadXplorer [63] based on the dRNA-seq data and evenly distributed on the forward and the reverse strands. Then, these were categorized according to their position in relation to the annotated genes. TSS located ≤300 nt upstream of the start codon and on the sense strand of an annotated gene were designated as primary transcriptional start site (pTSS)(Fig. 4a). TSS within an annotated gene were designated as secondary TSS (sTSS)(Fig. 4a). On the virulence plasmid we annotated 835 TSS, of which 443 were categorized as primary and 392 as secondary TSS (Fig. 4c). For the chromosome we annotated 13,216 TSS, of which 6280 were designated as primary TSSs and 6936 as secondary TSS (Fig. 4b, Table S2 and S3). In total we have annotated 6723 putative pTSS and 7328 putative sTSS. This number corresponds to roughly 2.7 TSS per CDS. The global TSS map of *S. flexneri* 5a M90T and the genome sequence has been integrated into RegulonDB (http://regulondb.ccg.unam.mx/) [61] for easy accessibility and visual display.

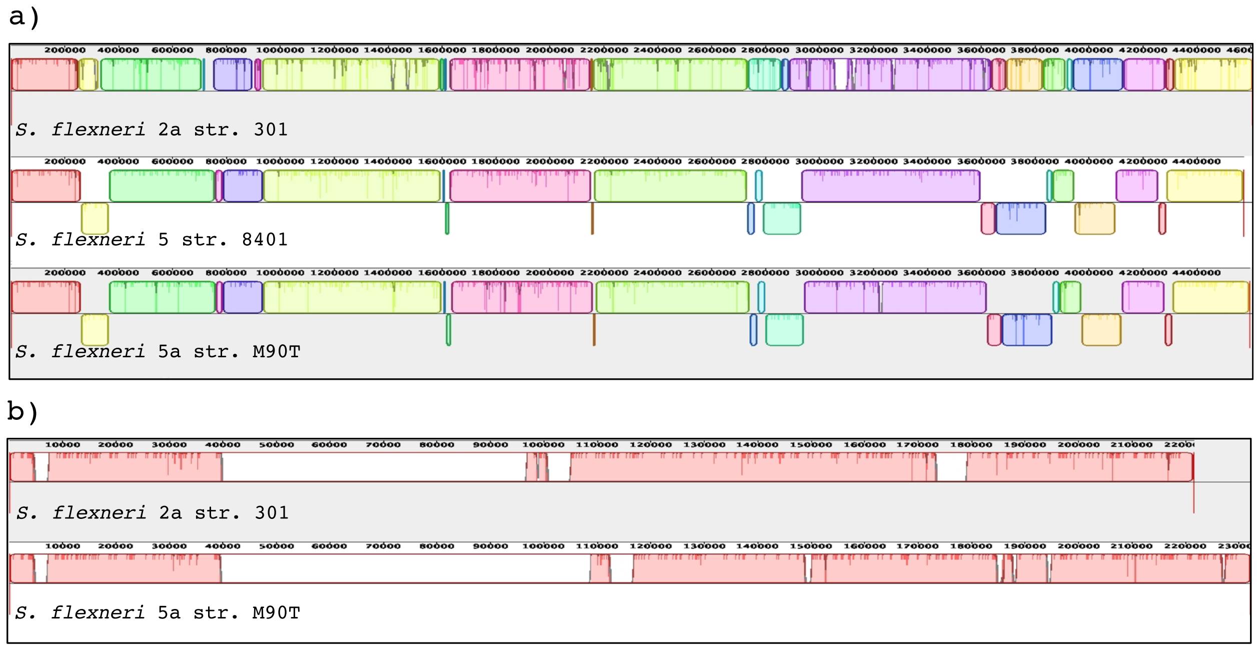

Fig. 2 Comparative genomic map of sequenced *S. flexneri* strains. **a** Chromosome comparison of *S. flexneri* 2a 301 (NC_004337), *S. flexneri* 5 8401 (NC_008258) and *S. flexneri* 5a M90T (NC_CP037923), **b** Virulence plasmid comparison of *S. flexneri* 2a 301(NC_004851) and *S. flexneri* 5a M90T (NZ_CP037924). Genome-wide alignment was performed with Mauve [56] progressive alignments to determine conserved sequence regions. This alignment resulted in many large synteny locally collinear blocks (LCBs). Each syntenical placement of the homologous region of the genome is represented as unique colored block, whilst divergent regions are seen as an empty block or line. Indentations within boxes highlight small mutations. Blocks above and below the center line depict the orientation of the genomic region compared to *S. flexneri* 2a strain 301

Table 3 Insertion sequences (IS) identified in *S. flexneri* 5a M90T

Genomic element	Insertion sequence type	Number of IS
Chromosome	**IS1**	**109**
	IS110	**6**
	IS200/IS605	**3**
	IS3	**73**
	IS3-like	**44**
	IS4	**21**
	IS4-like	**3**
	IS481	**1**
	IS66	**20**
	IS66-like	**6**
	IS91	**7**
	ISC	**1**
	ISNCY	**2**
pWR100	**IS1**	**11**
	IS110	**2**
	IS110-like	**2**
	IS21	**2**
	IS256	**2**
	IS3	**33**
	IS3-like	**4**
	IS4	**5**
	IS4-like	**1**
	IS5	**4**
	IS630	**4**
	IS66	**21**
	IS66-like	**3**
	IS91	**9**
	ISL3	**3**
Total		**402**

Analysis of the length of 5′-UTRs and leaderless transcripts

The TSS analysis shows that the longest 5’UTR in *S. flexneri* 5a M90T is 190 bp on the chromosome and 128 bp on the virulence plasmid (Fig. 5), while the shortest leader in both replicons is only 1 nt long. The average length of leaders on the virulence plasmid is 18 nt and 20 nt on the chromosome. Most primary and secondary TSS have a 5′-UTR of variable length, but we have found 172 TSS without leader region on the chromosome and 6 on the virulence plasmid (Table S2 and S3). The graphical visualization of 5′-UTRs is available at RegulonDB (http://regulondb.ccg.unam.mx/).

Data accessions

The fully sequenced and annotated *S. flexneri* 5a M90T genome is available in GenBank under the accession numbers CP037923 (chromosome) and CP037924 (virulence plasmid). The raw data from PacBio and Illumina sequencing are available in the SRA database under the accession SRR8921221(RNAseq-RiboZero), SRR8921222(dRNA-Seq_TEX_Positive), SRR8921223 (dRNA-Seq_TEX_Negative), SRR8921224(PacBio raw data) and SRR8921225 (RNAseq-TEX). The expression dataset is available in RegulonDB (http://regulondb.ccg.unam.mx/), which allows graphical visualization of the data.

As this is the only full genome of *S. flexneri* 5a M90T, it has been recognized as the reference genome and included in the RefSeq database with the accession numbers NZ_CP037923 (chromosome) and NZ_CP037924 (virulence plasmid). All data that were generated are integrated into RegulonDB for easy accessibility and visualization with JBrowser [64]. The *S. flexneri* 5a M90T genome is integrated in RSAT [62] database to use its analysis tools.

Discussion

The genome sequence that we report here is longer and contains less genes on the chromosome, but more on the virulence plasmid compared to the sequences published earlier for the chromosome scaffold [40] and the virulence plasmid [28, 39, 40]. Minor differences might be due to the fact that the previously published DNA sequences of *S. flexneri* 5a M90T were obtained from a streptomycin-resistant spontaneous mutant (*S. flexneri* 5a M90T Sm), which was derived from the original *S. flexneri* 5a M90T isolate sequenced here by serial culturing on antibiotic-containing plates [40, 65]. Nevertheless, most of the differences can be ascribed to technological developments. On the one hand, the *S. flexneri* 5a M90T chromosome was previously sequenced with a short-read Illumina sequencer [40]. On the other hand, the previously published virulence plasmid sequences were obtained using medium-read ABI377 Sanger technology [28, 39]. Both for the chromosome and the virulence plasmid, repetitive or AT-rich regions make it difficult to prepare a complete genome sequence with technologies that are not long-read [28, 39, 40] owing to the intrinsic assembly problems of this type of sequences. However, these assembly and annotation problems are circumvented with long-read sequencing such as the PacBio technology [50] employed here. Similarly, while Sanger sequencing remains a highly accurate technology for medium-length reads (> 500 nucleotides), the ABI377 sequencer required nebulization and subsequent size fractionation (in the range of 0.7 to 2.0 kb) of DNA by agarose gel electrophoresis and cloning into cosmids for sequencing [28, 39], which increased the risk of introducing mutations or losing sequences in between DNA fragments. NGS technology such as PacBio/SMRT long-read sequencing [50] is cloning- and PCR-free. The

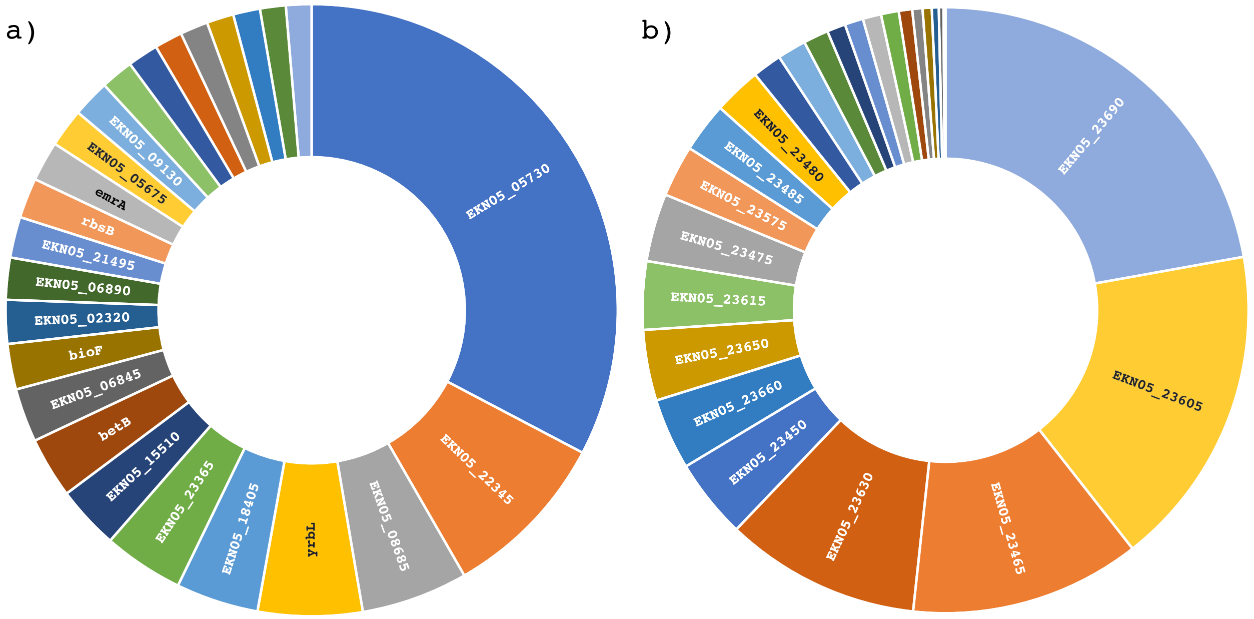

Fig. 3 Sunburst plot of pseudogenes transcript abundance levels in *S. flexneri* 5a M90T, with the top 25 labelled. The size of every box is proportional to the transcript abundance. The total number of reads per pseudogene measured by RNA-seq and counted with htseq/0.9.1 was plotted for **a**) the chromosome and **b**) the virulence plasmid pWR100. Full expression data are available in Table S4

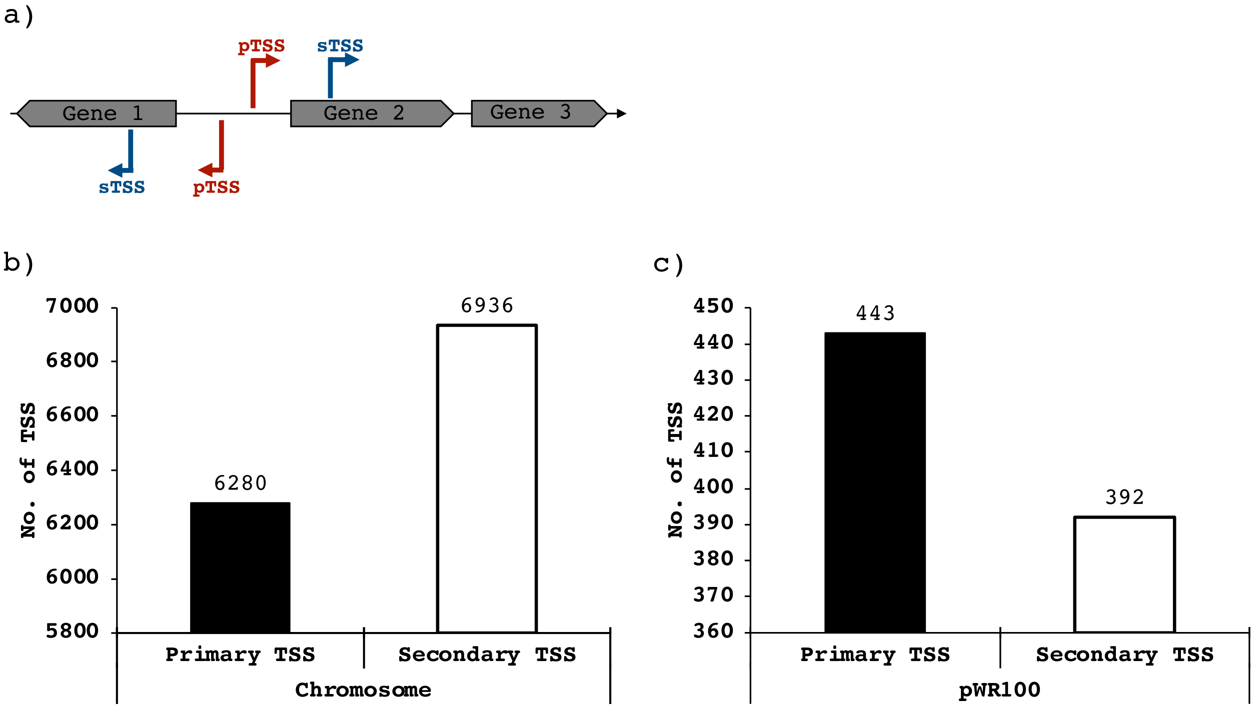

Fig. 4 Number of identified Transcriptional Start Sites (TSS) in *S. flexneri* 5a M90T grown in TSB to $OD_{600} = 0.3$. **a** Schematic representation of primary TSS (pTSS) and secondary (sTSS), **b** Plot of identified TSS on the chromosome and **c** pWR100

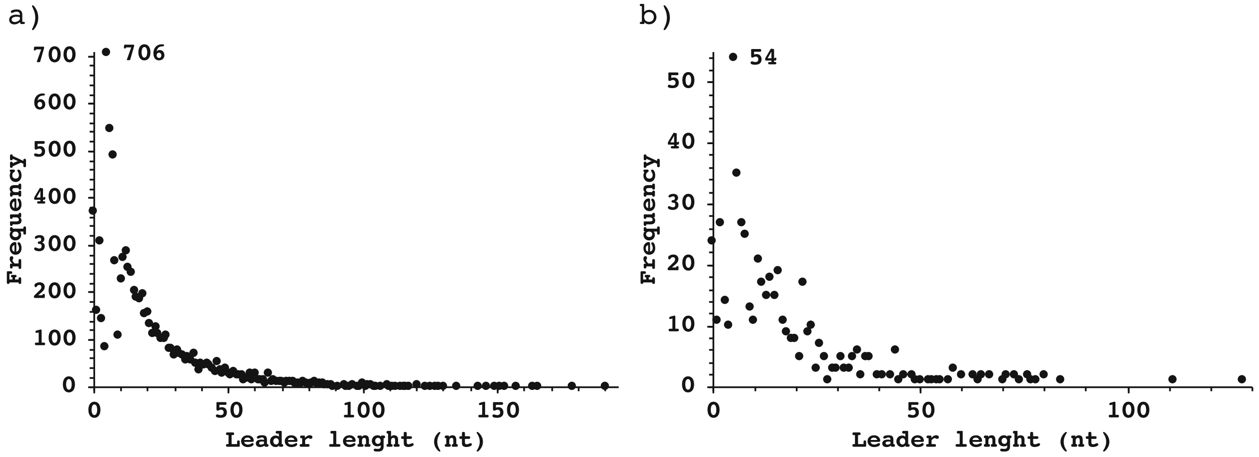

Fig. 5 Histogram of 5′-UTR lengths in *S. flexneri* 5a M90T. The distribution of 5′-UTR lengths ranges from 0 (leaderless) to maximum 190 nt. Transcripts with a 5′-UTR of 5 nt are the most abundant. **a** 5′-UTR lengths in the chromosome, **b** 5′-UTR lengths in the virulence plasmid

virulence plasmid described previously [28] is 213,194 bp long (RefSeq accession number NC_024996) [28], 8357 bp shorter than the sequence published 1 year later (GenBank accession number AF348706, 39]. The virulence plasmid sequence that we report here is 10,344 bp longer than the one with accession number AF348706 [39] (Table 2). It is already known that the genomic structure of the virulence plasmid is a mosaic with many repeated regions and AT-rich tracks [39]. The chromosome that was sequenced previously is 4,589,866 bp in length (GenBank accession number CM001474) including many regions with gaps that are represented in the sequence with "N". The total number of "N" in the genome scaffold is 11,901 bp [40]. A random check of 20 of these regions showed that they are repeated sequences or AT-rich tracks. The chromosome sequence reported here is gapless, it includes the missed 11,901 bp in the previously reported chromosome sequence [40] plus an extra 15,848 bp that were not present in the previously sequenced scaffold chromosome sequence. All together, these new regions in the genome sequence are summing up to a total 27,749 bp extra on the chromosome.

The number of genes, including putative pseudogenes, that have been annotated in *S. flexneri* 5a M90T earlier is smaller [40] because of the technical advances in sequencing and computer power for genome analysis and annotation (Table 1 and Table 2). The genome reported here has 4949 CDSs, of which 4629 are on the chromosome, in contrast to 4013 CDSs annotated on the draft chromosome previously [40] (Table 1), resulting in 616 new CDSs. This discrepancy in the numbers of CDSs could be a consequence of gaps in the previously reported chromosome sequence. In the case of the virulence plasmid, we report 27 new CDSs (Table 2) compared to the last sequence [39].

Of all the annotated genes, 769 are putative pseudogenes (Table 1 and Table 2), i.e. 563 more than previously reported based on the available hybrid genome composed of the chromosome scaffold and the plasmid assemblies [28, 39, 40]. This number is within the range of 229–858 reported before for *Escherichia* and *Shigella* genera members [60, 66]. When bacteria evolve from free-living to intracellular, the genome undergoes adaptive evolution [66–69]. Pseudogenes can be considered as a genomic record of the proteins, enzymes or pathways that are no longer necessary as the bacterium has adapted to a new environment [69–71]. Pseudogenes are continually created in bacterial genomes from ongoing mutational processes and are subject to degradation and eventual removal by further accumulation of mutations [71–74]. For example, intracellular pathogenic bacteria, such as *Mycobacterium leprae*, may accumulate a large number of pseudogenes [74, 75]. Pseudogenes are particularly prevalent in bacterial species that have recently become associated with or are dependent on eukaryotic hosts, as is the case of *Salmonella* and *Shigella* [68, 70]. Besides *Salmonella*, in other members of the *Enterobacteriaceae* such as *E. coli* and various *Shigella* species or strains, many CDSs have been annotated as pseudogenes [41, 68, 70, 76–78]. The numbers reported here are in the same range as found in other bacteria such as *Salmonella enterica* [68], *Helicobacter pylori* [42] and *Streptomyces coelicolor* [79]. The process of gene conversion into pseudogenes or gene decay is much faster in *Shigella* than in *E. coli* [80], possibly reflecting adaptive microevolution resulting in a transition of *Shigella* from a commensal to an intracellular pathogen. It has been shown that *Shigella* diverged from *E. coli* in multiple independent events [81–83]. Previous studies reported the inactivation of genes that hamper intracellular life in *S.*

flexneri 5a M90T by various types of mutations, for example in the *nadA*, *nadB* locus encoding the capacity to synthetize cadaverine [84, 85] or the *fim* cluster encoding fimbriae [86], pointing towards an adaptative process driving the intracellular lifestyle of *S. flexneri*.

In this study, we show that the pseudogenes that are present in the genome are transcribed, some of which highly, under laboratory growth conditions (Fig. 3 and Table S4). While almost all pseudogenes identified in this study are transcribed, it is not known whether these pseudogene transcripts have a role in *S. flexneri* fitness. Due to the compact genome architecture in bacteria, it is unclear why prokaryotic genomes would contain reduced coding-capacities. It is tempting to speculate that pseudogenes maintain some function, resulting in positive selective pressure to maintain their presence. One possible explanation to maintain a high number of transcribed pseudogenes is that they keep some residual function as regulatory elements or confer genome plasticity [13]. An alternative explanation is that the high number of transcribed pseudogenes reflects ongoing evolutionary processes [68, 70, 80]. The persistence of pseudogenes and their impact on *S. flexneri* fitness need to be studied further case by case.

Pseudogenes and IS appear to drive the bacterial genome remodeling. The presence of a large number of pseudogenes in a genome usually correlates with a high number of IS [87]. The IS play a very important role, particularly in genome evolution of pathogenic bacteria [88]. IS transposition can have different outcomes, from simple gene inactivation to constitutive expression or repression of adjacently located genes by delivering IS-specified promoter or terminator sequences, respectively. Multiple copies of IS elements promote various genomic rearrangements such as inversions, deletions, duplications and fusions between replicons. IS are determining factors for the efficiency of gene transfer between different bacterial strains or species [88]. Here, we show that the number of IS in the virulence plasmid is much higher than in the chromosome, indicating that the plasmid undergoes more active genome remodeling (Table 3).

Jacob and Monod proposed that the transcriptional architecture of bacteria is driven by three elements: one activator, one repressor and the polymerase binding site [89–92]. According to this model, transcription starts exclusively at one specific nucleotide. Further, the classical model of an operon comprises a group of genes under the control of a regulatory protein, where transcription results in a polycistronic mRNA with a single TSS and a single transcriptional terminator site (TTS). This classical model of operon may be valid for a specific group of genes under specific growth or environmental conditions. However, more recent evidence indicates that the transcriptional architecture of bacteria is far more complex than originally proposed. Many examples have established that an operon may encode alternative transcriptional units which are active under varying environmental condition [48, 93–96]. It has been suggested that an alternative model of transcriptional architecture called "noncontiguous operon" occurs in bacteria [97]. For example, for *E. coli* MG1655, which encodes roughly 4600 genes, > 14,000 TSS were documented [98, 99]. Our results for *S. flexneri* 5a M90T in terms of TSS number – about 14,051 TSS for a genome of 4.8 Mb - are close to what was found in *E. coli* MG1655.

To achieve a complex landscape of alternative transcriptional units, transcriptional regulation occurs at multiple levels [100]. Different lengths of the 5′-UTR play a very important role in translational regulation [94, 101, 102]. Indeed, the length of a 5′-UTR can provide insight into the regulation of gene expression [101]. For example, long 5'UTRs may contain riboswitches or provide binding sites for small regulatory RNAs [103]. Leaderless genes are translated by a different mechanism than genes with a leader sequence, as is the case for *virF* in *Shigella* spp. [104].

Conclusions

The genome sequence reported here is the first complete, gapless genome sequence for *S. flexneri* 5a M90T. Automatic annotation combined with manual curation allowed us to provide a high-quality reference genome that will be extremely useful to several types of studies, for example transcriptomics, differential expression analyses, or genome evolution. Moreover, in molecular pathogenesis projects, our results can be used as a resource to know which genes are transcribed before infection of host cells. The genome sequence together with the analysis of transcriptional start sites is also a valuable tool for precise genetic manipulation of *S. flexneri* 5a M90T.

In the present study we describe a hybrid cutting-edge workflow for genome sequencing with long reads and polishing with RNA-seq data and produced a high-quality, gapless reference genome. As input, many genome assemblies currently only use short-read DNA sequences, which are highly accurate but typically lead to poor or no coverage of repetitive and AT-rich regions. An alternative approach is to use long-read DNA sequences, which provide outstanding scaffolding power but lower fidelity. In contrast to currently used hybrid approaches combining redundant long- and short-read DNA sequencing technologies, our hybrid workflow exploiting the strengths of non-redundant long-read DNA sequencing and short-read RNA sequencing has the added value of yielding distinct datasets (genomic data and transcriptomic data), which can be further used for other purposes. This workflow proved as a powerful strategy for genome assembly, polishing and annotation that could be implemented for any type of organism.

S. flexneri serotype 5a strain M90T is a very important model to study the molecular pathogenesis mechanisms. The availability of a full genome opens the door to discovering new genomic elements and gene regulatory networks that are involved in *Shigella* pathogenicity.

Methods

Bacterial strain and culture conditions

The *S. flexneri* serotype 5a strain M90T that was used in this study was obtained from Dr. Philippe Sansonetti, Institut Pasteur, Paris, France. This strain was collected by Dr. Samuel Formal for the Walter Reed Army Institute of Research collection and first described in 1982 [25]. This strain is not streptomycin resistant, unlike a derivative of M90T that was later obtained by serial passaging on antibiotic containing plates [65] and sequenced in 2012 [40]. *S. flexneri* 5a M90T was cultured on tryptic soy broth agar plates with 0.01% (w/v) Congo red (TSBA-CR). Red colonies were selected to ensure the presence of the virulence plasmid (pWR100). Overnight bacterial cultures were grown at 37 °C in tryptic soy broth medium, sub-cultured 1:100, and grown at 37 °C in a shaking incubator at 150 RPM, until OD_{600} = 0.3.

DNA purification and genome sequencing

Genomic DNA was isolated using the Wizard^{R} Genomic DNA purification kit (Promega, Inc.) from overnight cultures of *S. flexneri* 5a M90T according to the manufacturer's instructions. Isolated DNA was cleaned as many times as necessary with phenol-chloroform (until no white interphase between the watery and organic phase was forming) [105] to obtain a high quality-quantity of genomic DNA (20 μg) for PacBio library preparation [50]. Library preparation was carried out by Novogene Inc. Sequencing was performed using a PacBio RSII sequencer at Novogene Inc., Hong Kong, China.

Total RNA purification and sequencing

S. flexneri 5a M90T was sub-cultured until $OD_{600} = 0.3$ and the culture was mixed with 0.2 volumes of stop solution (95% EtOH and 5% phenol pH 4, v/v) [105]. Samples were allowed to incubate on ice for at least 30 min, but not longer than 2 h, to stabilize the RNA and prevent degradation. After the incubation on ice, the cells were harvested by centrifugation for 5 min at 13, 000 RPM at 4 °C. Cell pellets were frozen with liquid nitrogen and stored at -80 °C until RNA extraction.

Frozen cell pellets were thawed on ice and resuspended in lysis solution (0.5% SDS, 20 mM sodium acetate pH 4.8, 10 mM EDTA pH 8). Bacterial cells were lysed by incubating the samples for 5 min at 65 °C. Afterwards, total RNA was extracted using the hot-phenol method [105]. Contaminating DNA was digested by DNase I (Roche; 1 U/μg RNA, 60 min, 37 °C) in the presence of RNase inhibitor (RNaseOUT, ThermoFisher Scientific; 0.1 U/μl) followed by clean-up of RNA by phenol/chloroform/isoamyl alcohol and precipitation of RNA with 2.5 volumes of ethanol containing 0.1 M sodium acetate pH 5.5 and 20 μg of glycogen (Roche) [105]. Removal of residual DNA was subsequently verified by control PCR using the oligos SF-Hfq-F 5′-ACGATGAAATGGTTTATCGAG-3′ and SF-Hfq-R 5′-ACTGCTTTACCTTCACCTACA-3′, which amplify a 909 bp long product of the *hfq* gene from *S. flexneri* 5a M90T including 300 pb upstream and 300 bp downstream of the *hfq* gene.

The RNA concentration was measured using a NanoDrop ND-1000 spectrophotometer (Saveen & Werner AB, Limhamn, Sweden). Thereafter, the integrity of the 16S and 23S rRNA was checked by agarose gel electrophoresis, using 1% agarose in 1X TAE buffer (40 mM Tris acetate, 1 mM EDTA at pH 8.3 ± 0.1).

The rRNA was depleted from three biological replicates of total RNA with RiboZero according to the manufacturer's instructions (Illumina, Inc.). Library preparation and sequencing were performed at the EMBL Genomics Core Facility (Heidelberg, Germany).

The rRNA from another set of three biological replicates was depleted with Terminator-5′-Phosphate-Dependent Exonuclease (TEX) (Lucigene, Inc.). Library preparation and sequencing was performed at Novogene, Inc. The libraries were constructed using the TruSeq Stranded kit and were sequenced on an Illumina HiSeq2000 platform (Illumina, Inc.) with a paired-end protocol and read length of 150 nt (PE150), resulting in a total output of roughly 20 million (M) per sample. All reads outputs were checked for passage of Illumina quality standards [106, 107]. These RNA-seq results obtained from EMBL and Novogene were used to polish the genome assembly.

Genome assembly and annotation

De novo genome assembly was performed with the script Canu/1.7 [51] implementing the pacbio-raw option using all its default parameters. Output files from Canu assembly were used as input to polish the genome assembled with Pilon/1.22 [55]. Polishing of the genome assembly was done in two rounds: the first one was carried out using the RNA-seq output files from the samples in which the rRNA was depleted with RiboZero (RNAseq-RZ) and the second one was carried out with the RNA-seq results from the samples in which the rRNA was depleted with TEX (RNAseq-TEX). Genome annotation and polishing was performed at Uppsala Multidisciplinary Center for Advanced Computational

Science (UPPMAX) of SciLifeLab at Uppsala University, Sweden.

Gene prediction and annotation was carried out with three different pipelines: a) RAST [57], b) Prokka [58] and c) NCBI Prokaryotic Genome Annotation Pipeline (PGAP) [59]. The RAST pipeline used the Taxonomy ID: 1086030 from NCBI, which corresponds to *S. flexneri* serotype 5a strain M90T. The most striking feature of Prokka, which is distinct from other pipelines, is the use of multiple databases to find sequence homologies. For subsequent analysis, we selected the PGAP/NCBI annotation.

The assembled and annotated genome was manually curated using Artemis [52] for visualizing and editing the genome files. The genome was deposited in GenBank with accession numbers CP037923 (chromosome) and CP037924 (virulence plasmid).

The genome-scale alignments were performed using Mauve alignment tools [56]. For the chromosome, the sequences from *S. flexneri* 2a strain 301 (NC_004337), *S. flexneri* 5 strain 8401 (NC_008258) and *S. flexneri* 5a strain M90T (NZ_CP037923) were used. For the virulence plasmid, the sequences from *S. flexneri* 2a strain 301 (NC_004851) and *S. flexneri* 5a strain M90T (NZ_CP037924) were used. To ease the comparison, the strand direction of the *S. flexneri* 5a strain M90T sequence was shifted using the script emboss/6.6.0 [108]. The start nucleotide in the *S. flexneri* 5a strain M90T sequence was selected manually using *S. flexneri* 2a strain 301 as reference.

RNA treatment for transcriptional start site (TSS) determination and sequencing

To determine transcriptional start sites, the RNA of three biological replicates in which the rRNA had been depleted with RiboZero (Illumina, Inc.) was used. To enrich for primary transcripts, we exploited the property that primary bacterial transcripts are protected from exonucleolytic degradation by their triphosphate (5’PPP) RNA ends [49], while RNAs containing a 5′ monophosphate (5’P) are selectively degraded [42, 49]. The rRNA-depleted RNA was split into two aliquots. One aliquot was treated with Terminator 5′-Phosphate-Dependent Exonuclease (TEX+), the other aliquot was incubated only with TEX buffer (TEX-) as a control. TEX treatment was carried out for 60 min at 30 °C. One unit of TEX was used per 1 μg of rRNA-depleted RNA. Following organic extraction (25:24:1 v/v phenol/chloroform/isoamyl alcohol), RNA was precipitated overnight with 2.5 volumes of ethanol/0.1 M sodium acetate (pH 5.5) and 20 μg of glycogen (Roche) mixture. After TEX treatment, both samples (TEX+ and TEX-) were treated with 5′ Pyrophosphohydrolase (RppH) (NewEngland BioLabs, Inc.) to generate 5′-mono-phosphates for linker ligation [109], and again purified by organic extraction and ethanol precipitation. RppH treatment was carried out for 60 min at 37 °C. An RNA adaptor (5′-GACCUUGGCU GUCACUCA-3′) was ligated to the 5′-monophosphate of the RNA end by incubation with T4 RNA ligase (NewEngland BioLabs, Inc.), at 25 °C for 16 h. As last step, the RNA adaptor that had been ligated to the RNA was phosphorylated with T4 PNK (NewEngland BioLabs, Inc.) at 37 °C for 60 min.

Separate libraries were constructed for TEX- and TEX+ samples. The libraries were constructed using TruSeq Stranded Kit (Illumina, Inc.) and sequenced on an Illumina HiSeq2000 platform (Novogene, Inc.) with a paired-end protocol and read length of 150 nt (PE150), resulting in a total output of roughly 20 million (M) per sample/library sequenced. All reads were checked for passage of Illumina quality standards [106, 107].

Reads mapping of TSS library

Reads in the FASTQ format were cleaned up with trimmomatic/0.36 [110] to remove sequences originating from Illumina adaptors and low quality reads. The files were aligned to the genome of *S. flexneri* 5a M90T prepared in the present work (GenBank accession numbers CP037923 and CP037924) with bowtie2/2.3.4.3 [111] using –X 1000 such that only mate pairs were reported if separated by less than 1000 bp. All the other settings were implemented with the default option. After the alignment was completed, samtools/1.9 [53] was used to remove duplicates and select for reads that were aligned in proper pairs. Reads aligned to the reference genome was converted to BAM format with samtools/1.9. The final analysis for identification and annotation of TSSs into pTSS and sTSS was done with ReadXplorer [63, 112].

Reads mapping of transcribed pseudogenes

To count the number of reads aligned per pseudogene we used the alignment results for TSS mapping. After sorting the alignment with samtools/1.9 [53], the reads counting per pseudogene was performed with htseq/0.9.1 [113] using stranded mode and pseudogene as a feature type. The total expression of pseudogenes is the average of the six libraries used for TSS determination.

Transcriptional start sites annotation and classification

To map dRNA-seq outputs, reads were split by replicon, converted to BAM format and sorted by position with samtools/1.9 [28]. These BAM files were used as input for ReadXplorer [63, 112] for automatic de novo TSS annotation. For the analysis, results from three biological replicates were pooled and TSS within 10 nt of each other were clustered into one. Such regions were then manually annotated by scanning the respective wiggle files for nucleotides with an abrupt increase in coverage. TSS were classified according to their genomic context. Peaks in an intergenic region and on the same strand as the closest downstream gene were classified as primary. Peaks within gene boundaries and on the same strand as the gene were qualified as secondary. All TSS positions were assigned relative to the start of the associated annotated gene. With the first base of the gene being positive + 1, all upstream positions start with − 1.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-020-6565-5.

Additional file 1: Figure S1. Genome coverage plot of *S. flexneri* 5a M90T with aligned DNA or RNA short-read sequences. a) Genome coverage of small DNA Illumina sequences. The percentage of genome covered with a depth of ≥5 is 99.98% with a mean coverage of 126X. b) Genome coverage of small RNA sequences. The percentage of genome covered with a depth of of ≥5 is 98.77% with a mean coverage of 989.9X. Coverage calculation was performed with Samtools depth [54, 55] using sorted alignments in BAM format as input. The figure was generated with ggplot2.3.2.1 in the R. 3.6.1 environment.

Additional file 2: Table S1. Comparison of general features of the *S. flexneri* 5a M90T predicted with three different pipelines: Prokka [59], RAST [58] and PGAP/NCBI [60].

Additional file 3: Table S2. Transcriptional start sites determined in *S. flexneri* 5a M90T chromosome with ReadExplorer [64, 113].

Additional file 4: Table S3. Transcriptional start sites determined in *S. flexneri* 5a M90T virulence plasmid (pWR100) with ReadExplorer [64, 113].

Additional file 5: Table S4. Pseudogenes transcription abundance level. Reads counting per pseudogene was performed with htseq/0.9.1 [113] using stranded mode and pseudogene as a feature type. The total expression of pseudogenes is the average of the six libraries used for TSS determination.

Additional file 6: File S1. General features of the *S. flexneri* 5a M90T chromosome annotated with Prokka [59].

Additional file 7: File S1.1. General features of the *S. flexneri* 5a M90T virulence plasmid (pWR100) annotated with Prokka [59].

Additional file 8: File: S2. General features of the *S. flexneri* 5a M90T chromosome annotated with RAST [58].

Additional file 9: File: S2.1. General features of the *S. flexneri* 5a M90T virulence plasmid (pWR100) annotated with RAST [58].

Abbreviations

CDS: Coding sequence; dRNA-seq: Diffential RNA sequencing; IS: Insertion sequence; ncRNA: Non-coding RNA; pTSS: Primary transcriptional start site; SMRT: Single molecule real-time; sTSS: Secondary transcriptional start site; TEX: 5′-monophosphate-dependent exonuclease; TSB: Tryptic soy broth; TSBA-CR: Tryptic soy broth agar with 0.01% (w/v) Congo red; TSS: Transcriptional start site; TTS: Transcriptional terminator site

Acknowledgements

We are grateful for travel support and courses organized by The National Doctoral Program in Infections and Antibiotics (NDPIA).
The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project SNIC 2017-7-258.
We acknowledge the EMBL Genomics Core Facility (Heidelberg, Germany) for RiboZero library preparation and sequencing.
The authors thank RegulonDB (CCG-UNAM) staff, especially Jaír S. García-Sotelo (LIIGH-UNAM), for their support in data managing and display in RegulonDB, and also RSAT staff for integrating the *S. flexneri* 5a M90T genome (CP037923, CP037924) in its analysis tools. We further thank Atin Sharma, Dorothee Langenbach, David A. Cisneros (Umeå University, Sweden) and Edgardo Sepúlveda (CICESE, Baja California, Mexico) for critical reading of the manuscript.

Authors' contributions

RCR and ST performed experiments. RCR analyzed all data. RCR and AP designed research. RCR and AP wrote the manuscript. All authors corrected and approved the manuscript.

Funding

We gratefully acknowledge funding from the Kempe Foundation (JCK-1528), the Knut and Alice Wallenberg Foundation (KAW 2015.0225), Centre for Microbial Research (UCMR), Umeå University and The Laboratory for Molecular Infection Medicine Sweden (MIMS).
The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data nor in writing the manuscript. Open access funding provided by Umea University.

Availability of data and materials

Data are deposited at GenBank database under the accession number: CP037923 and CP037924.
The raw data are deposited in SRA under the accession number SRR8921221(RNAseq-RiboZero), SRR8921222(dRNA-Seq_TEX_Positive), SRR8921223 (dRNA-Seq_TEX_Negative), SRR8921224(PabBio raw data) and SRR8921225 (RNAseq-TEX). https://dataview.ncbi.nlm.nih.gov/
Request for material should be directed to and will be fulfilled by Andrea Puhar (andrea.puhar@umu.se).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]The Laboratory for Molecular Infection Medicine Sweden (MIMS), 901 87 Umeå, Sweden. [2]Umeå Centre for Microbial Research (UCMR), 901 87 Umeå, Sweden. [3]Department of Molecular Biology, Umeå University, 901 87 Umeå, Sweden.

Received: 24 June 2019 Accepted: 7 February 2020
Published online: 06 April 2020

References

1. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, Wu Y, Sow SO, Sur D, Breiman RF, et al. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. Lancet. 2013;382(9888):209–22.
2. Livio S, Strockbine NA, Panchalingam S, Tennant SM, Barry EM, Marohn ME, Antonio M, Hossain A, Mandomando I, Ochieng JB, et al. *Shigella* isolates from the global enteric multicenter study inform vaccine development. Clin Infect Dis. 2014;59(7):933–41.
3. Lozano R, Naghavi M, Foreman K, Lim S, Shibuya K, Aboyans V, Abraham J, Adair T, Aggarwal R, Ahn SY, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the global burden of disease study 2010. Lancet. 2012;380(9859):2095–128.
4. Kotloff KL, Winickoff JP, Ivanoff B, Clemens JD, Swerdlow DL, Sansonetti PJ, Adak GK, Levine MM. Global burden of *Shigella* infections: implications for vaccine development and implementation of control strategies. Bull World Health Organ. 1999;77(8):651–66.
5. Kotloff KL, Riddle MS, Platts-Mills JA, Pavlinac P, Zaidi AKM. Shigellosis. Lancet. 2018;391(10122):801–12.
6. DuPont HL, Levine MM, Hornick RB, Formal SB. Inoculum size in shigellosis and implications for expected mode of transmission. J Infect Dis. 1989;159(6):1126–8.

7. The HC, Thanh DP, Holt KE, Thomson NR, Baker S. The genomic signatures of *Shigella* evolution, adaptation and geographical spread. Nat Rev Microbiol. 2016;14(4):235–50.
8. Sorbara MT, Foerster EG, Tsalikis J, Abdel-Nour M, Mangiapane J, Sirluck-Schroeder I, Tattoli I, van Dalen R, Isenman DE, Rohde JR, et al. Complement C3 Drives Autophagy-Dependent Restriction of Cyto-invasive Bacteria. Cell Host Microbe. 2018;23(5):644–52 e645.
9. Abdel-Nour M, Carneiro LAM, Downey J, Tsalikis J, Outlioua A, Prescott D, Da Costa LS, Hovingh ES, Farahvash A, Gaudet RG, et al. The heme-regulated inhibitor is a cytosolic sensor of protein misfolding that controls innate immune signaling. Science. 2019;365:eaaw4144. https://doi.org/10.1126/science.aaw4144.
10. Scribano D, Damico R, Ambrosi C, Superti F, Marazzato M, Conte MP, Longhi C, Palamara AT, Zagaglia C, Nicoletti M. The *Shigella flexneri* OmpA amino acid residues 188EVQ190 are essential for the interaction with the virulence factor PhoN2. Biochem Biophys Rep. 2016;8:168–73.
11. Brotcke Zumsteg A, Goosmann C, Brinkmann V, Morona R, Zychlinsky A. IcsA is a *Shigella flexneri* adhesin regulated by the type III secretion system and required for pathogenesis. Cell Host Microbe. 2014;15(4):435–45.
12. Liu Z, Mar KB, Hanners NW, Perelman SS, Kanchwala M, Xing C, Schoggins JW, Alto NM. A NIK-SIX signalling axis controls inflammation by targeted silencing of non-canonical NF-kappaB. Nature. 2019;568(7751):249–53.
13. Pilla G, McVicker G, Tang CM. Genetic plasticity of the *Shigella* virulence plasmid is mediated by intra- and inter-molecular events between insertion sequences. PLoS Genet. 2017;13(9):e1007014.
14. Krokowski S, Lobato-Marquez D, Chastanet A, Pereira PM, Angelis D, Galea D, Larrouy-Maumus G, Henriques R, Spiliotis ET, Carballido-Lopez R, et al. Septins Recognize and Entrap Dividing Bacterial Cells for Delivery to Lysosomes. Cell Host Microbe. 2018;24(6):866–74 e864.
15. Pasqua M, Grossi M, Scinicariello S, Aussel L, Barras F, Colonna B, Prosseda G. The MFS efflux pump EmrKY contributes to the survival of *Shigella* within macrophages. Sci Rep. 2019;9(1):2906.
16. Weiner A, Mellouk N, Lopez-Montero N, Chang YY, Souque C, Schmitt C, Enninga J. Macropinosomes are key players in early *Shigella* invasion and vacuolar escape in epithelial cells. PLoS Pathog. 2016;12(5):e1005602.
17. Sidik SM, Salsman J, Dellaire G, Rohde JR. *Shigella* infection interferes with SUMOylation and increases PML-NB number. PLoS One. 2015;10(4):e0122585.
18. Wandel MP, Pathe C, Werner EI, Ellison CJ, Boyle KB, von der Malsburg A, Rohde J, Randow F. GBPs Inhibit Motility of *Shigella flexneri* but Are Targeted for Degradation by the Bacterial Ubiquitin Ligase IpaH9.8. Cell Host Microbe. 2017;22(4):507–18 e505.
19. Ciancarella V, Lembo-Fazio L, Paciello I, Bruno AK, Jaillon S, Berardi S, Barbagallo M, Meron-Sudai S, Cohen D, Molinaro A, et al. Role of a fluid-phase PRR in fighting an intracellular pathogen: PTX3 in *Shigella* infection. PLoS Pathog. 2018;14(12):e1007469.
20. Meghraoui A, Schiavolin L, Allaoui A. Single amino acid substitutions on the needle tip protein IpaD increased *Shigella* virulence. Microbes Infect. 2014;16(7):532–9.
21. Kim HN, Seok SH, Lee YS, Won HS, Seo MD. Crystal structure and functional characterization of SF216 from *Shigella flexneri*. FEBS Lett. 2017;591(21):3692–703.
22. Roehrich AD, Bordignon E, Mode S, Shen DK, Liu X, Pain M, Murillo I, Martinez-Argudo I, Sessions RB, Blocker AJ. Steps for *Shigella* gatekeeper protein MxiC function in hierarchical type III secretion regulation. J Biol Chem. 2017;292(5):1705–23.
23. Silue N, Marcantonio E, Campbell-Valois FX. RNA-Seq analysis of the T3SA regulon in *Shigella flexneri* reveals two new chromosomal genes upregulated in the on-state. Methods. 2019. https://doi.org/10.1016/j.ymeth.2019.03.017.
24. Valencia-Gallardo C, Bou-Nader C, Aguilar-Salvador DI, Carayol N, Quenech'Du N, Pecqueur L, Park H, Fontecave M, Izard T, Tran Van Nhieu G. *Shigella* IpaA Binding to Talin Stimulates Filopodial Capture and Cell Adhesion. Cell Rep. 2019;26(4):921–32 e926.
25. Sansonetti PJ, Kopecko DJ, Formal SB. Involvement of a plasmid in the invasive ability of *Shigella flexneri*. Infect Immun. 1982;35(3):852–60.
26. Ferrari ML, Malarde V, Grassart A, Salavessa L, Nigro G, Descorps-Declere S, Rohde JR, Schnupf P, Masson V, Arras G, et al. *Shigella* promotes major alteration of gut epithelial physiology and tissue invasion by shutting off host intracellular transport. Proc Natl Acad Sci U S A. 2019;116(27):13582–91.
27. Mirza N, Sowa AS, Lautz K, Kufer TA. NLRP10 Affects the Stability of Abin-1 To Control Inflammatory Responses. J Immunol. 2019;202(1):218–27.
28. Buchrieser C, Glaser P, Rusniok C, Nedjari H, D'Hauteville H, Kunst F, Sansonetti P, Parsot C. The virulence plasmid pWR100 and the repertoire of proteins secreted by the type III secretion apparatus of *Shigella flexneri*. Mol Microbiol. 2000;38(4):760–71.
29. Sansonetti PJ, Hale TL, Dammin GJ, Kapfer C, Collins HH Jr, Formal SB. Alterations in the pathogenicity of *Escherichia coli* K-12 after transfer of plasmid and chromosomal genes from *Shigella flexneri*. Infect Immun. 1983;39(3):1392–402.
30. Falkow S, Schneider H, Baron LS, Formal SB. Virulence of *Escherichia-Shigella* genetic hybrids for the Guinea pig. J Bacteriol. 1963;86:1251–8.
31. Ingersoll MA, Moss JE, Weinrauch Y, Fisher PE, Groisman EA, Zychlinsky A. The ShiA protein encoded by the *Shigella flexneri* SHI-2 pathogenicity island attenuates inflammation. Cell Microbiol. 2003;5(11):797–807.
32. Moss JE, Cardozo TJ, Zychlinsky A, Groisman EA. The *selC*-associated SHI-2 pathogenicity island of *Shigella flexneri*. Mol Microbiol. 1999;33(1):74–83.
33. Bartoleschi C, Pardini MC, Scaringi C, Martino MC, Pazzani C, Bernardini ML. Selection of *Shigella flexneri* candidate virulence genes specifically induced in bacteria resident in host cell cytoplasm. Cell Microbiol. 2002;4(9):613–26.
34. Marteyn B, West NP, Browning DF, Cole JA, Shaw JG, Palm F, Mounier J, Prevost MC, Sansonetti P, Tang CM. Modulation of *Shigella* virulence in response to available oxygen in vivo. Nature. 2010;465(7296):355–8.
35. Vergara-Irigaray M, Fookes MC, Thomson NR, Tang CM. RNA-seq analysis of the influence of anaerobiosis and FNR on *Shigella flexneri*. BMC Genomics. 2014;15:438.
36. Runyen-Janecky LJ, Payne SM. Identification of chromosomal *Shigella flexneri* genes induced by the eukaryotic intracellular environment. Infect Immun. 2002;70(8):4379–88.
37. Morris CR, Grassel CL, Redman JC, Sahl JW, Barry EM, Rasko DA. Characterization of intracellular growth regulator *icgR* by utilizing transcriptomics to identify mediators of pathogenesis in *Shigella flexneri*. Infect Immun. 2013;81(9):3068–76.
38. Faherty CS, Harper JM, Shea-Donohue T, Barry EM, Kaper JB, Fasano A, Nataro JP. Chromosomal and plasmid-encoded factors of *Shigella flexneri* induce secretogenic activity *ex vivo*. PLoS One. 2012;7(11):e49980.
39. Venkatesan MM, Goldberg MB, Rose DJ, Grotbeck EJ, Burland V, Blattner FR. Complete DNA sequence and analysis of the large virulence plasmid of *Shigella flexneri*. Infect Immun. 2001;69(5):3271–85.
40. Onodera NT, Ryu J, Durbic T, Nislow C, Archibald JM, Rohde JR. Genome sequence of *Shigella flexneri* serotype 5a strain M90T Sm. J Bacteriol. 2012;194(11):3022.
41. Nie H, Yang F, Zhang X, Yang J, Chen L, Wang J, Xiong Z, Peng J, Sun L, Dong J, et al. Complete genome sequence of *Shigella flexneri* 5b and comparison with *Shigella flexneri* 2a. BMC Genomics. 2006;7:173.
42. Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R, et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature. 2010;464(7286):250–5.
43. Hannemann S, Gao B, Galan JE. *Salmonella* modulation of host cell gene expression promotes its intracellular growth. PLoS Pathog. 2013;9(10):e1003668.
44. Hannemann S, Galan JE. *Salmonella enterica* serovar-specific transcriptional reprogramming of infected cells. PLoS Pathog. 2017;13(7):e1006532.
45. Albrecht M, Sharma CM, Reinhardt R, Vogel J, Rudel T. Deep sequencing-based discovery of the *Chlamydia trachomatis* transcriptome. Nucleic Acids Res. 2010;38(3):868–77.
46. Nickerson KP, Chanin RB, Sistrunk JR, Rasko DA, Fink PJ, Barry EM, Nataro JP, Faherty CS. Analysis of *Shigella flexneri* Resistance, Biofilm Formation, and Transcriptional Profile in Response to Bile Salts. Infect Immun. 2017;85(6):e01067–16. https://doi.org/10.1128/IAI.01067-16.
47. Mendoza-Vargas A, Olvera L, Olvera M, Grande R, Vega-Alvarado L, Taboada B, Jimenez-Jacinto V, Salgado H, Juarez K, Contreras-Moreira B, et al. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. PLoS One. 2009;4(10):e7526.
48. Kroger C, Colgan A, Srikumar S, Handler K, Sivasankaran SK, Hammarlof DL, Canals R, Grissom JE, Conway T, Hokamp K, et al. An infection-relevant transcriptomic compendium for *Salmonella enterica* Serovar Typhimurium. Cell Host Microbe. 2013;14(6):683–95.
49. Schoenberg DR. The end defines the means in bacterial mRNA decay. Nat Chem Biol. 2007;3(9):535–6.
50. Rhoads A, Au KF. PacBio sequencing and its applications. Genomics Proteomics Bioinformatics. 2015;13(5):278–89.
51. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36.

52. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 2012;28(4):464–9.
53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
54. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
55. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963.
56. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004; 14(7):1394–403.
57. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, et al. The RAST server: rapid annotations using subsystems technology. BMC Genomics. 2008;9:75.
58. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.
59. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016;44(14):6614–24.
60. Goodhead I, Darby AC. Taking the pseudo out of pseudogenes. Curr Opin Microbiol. 2015;23:102–9.
61. Santos-Zavaleta A, Salgado H, Gama-Castro S, Sanchez-Perez M, Gomez-Romero L, Ledezma-Tejeida D, Garcia-Sotelo JS, Alquicira-Hernandez K, Muniz-Rascado LJ, Pena-Loredo P, et al. RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. Nucleic Acids Res. 2019;47(D1):D212–d220.
62. Nguyen NTT, Contreras-Moreira B, Castro-Mondragon JA, Santana-Garcia W, Ossio R, Robles-Espinoza CD, Bahin M, Collombet S, Vincens P, Thieffry D, et al. RSAT 2018: regulatory sequence analysis tools 20th anniversary. Nucleic Acids Res. 2018;46(W1):W209–w214.
63. Hilker R, Stadermann KB, Doppmeier D, Kalinowski J, Stoye J, Straube J, Winnebald J, Goesmann A. ReadXplorer--visualization and analysis of mapped sequences. Bioinformatics. 2014;30(16):2247–54.
64. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, Goodstein DM, Elsik CG, Lewis SE, Stein L, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol. 2016;17:66.
65. Allaoui A, Mounier J, Prevost MC, Sansonetti PJ, Parsot C. *icsB*: a *Shigella flexneri* virulence gene necessary for the lysis of protrusions during intercellular spread. Mol Microbiol. 1992;6(12):1605–16.
66. Liu Y, Harrison PM, Kunin V, Gerstein M. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. Genome Biol. 2004;5(9):R64.
67. Kuo CH, Ochman H. The extinction dynamics of bacterial pseudogenes. PLoS Genet. 2010;6(8):e1001050.
68. Holt KE, Thomson NR, Wain J, Langridge GC, Hasan R, Bhutta ZA, Quail MA, Norbertczak H, Walker D, Simmonds M, et al. Pseudogene accumulation in the evolutionary histories of *Salmonella enterica* serovars Paratyphi a and Typhi. BMC Genomics. 2009;10:36.
69. Feavers IM, Maiden MC. A gonococcal *porA* pseudogene: implications for understanding the evolution and pathogenicity of *Neisseria gonorrhoeae*. Mol Microbiol. 1998;30(3):647–56.
70. Feng Y, Chen Z, Liu SL. Gene decay in *Shigella* as an incipient stage of host-adaptation. PLoS One. 2011;6(11):e27754.
71. Suzuki K, Nakata N, Bang PD, Ishii N, Makino M. High-level expression of pseudogenes in *Mycobacterium leprae*. FEMS Microbiol Lett. 2006;259(2): 208–14.
72. Toh H, Weiss BL, Perkin SA, Yamashita A, Oshima K, Hattori M, Aksoy S. Massive genome erosion and functional adaptations provide insights into the symbiotic lifestyle of *Sodalis glossinidius* in the tsetse host. Genome Res. 2006;16(2):149–56.
73. Belda E, Moya A, Bentley S, Silva FJ. Mobile genetic element proliferation and gene inactivation impact over the genome structure and metabolic capabilities of *Sodalis glossinidius*, the secondary endosymbiont of tsetse flies. BMC Genomics. 2010;11:449.
74. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, et al. Massive gene decay in the leprosy bacillus. Nature. 2001;409(6823):1007–11.
75. Williams DL, Slayden RA, Amin A, Martinez AN, Pittman TL, Mira A, Mitra A, Nagaraja V, Morrison NE, Moraes M, et al. Implications of high level pseudogene transcription in *Mycobacterium leprae*. BMC Genomics. 2009;10:397.
76. Wei J, Goldberg MB, Burland V, Venkatesan MM, Deng W, Fournier G, Mayhew GF, Plunkett G 3rd, Rose DJ, Darling A, et al. Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. Infect Immun. 2003;71(5):2775–86.
77. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, Tang X, Wang J, Xiong Z, Dong J, et al. Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. Nucleic Acids Res. 2005;33(19):6445–58.
78. Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, Wang J, Liu H, Yang J, Yang F, 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(20):4432–41.
79. Vockenhuber MP, Sharma CM, Statt MG, Schmidt D, Xu Z, Dietrich S, Liesegang H, Mathews DH, Suess B. Deep sequencing-based identification of small non-coding RNAs in *Streptomyces coelicolor*. RNA Biol. 2011;8(3): 468–77.
80. Hershberg R, Tang H, Petrov DA. Reduced selection leads to accelerated gene loss in *Shigella*. Genome Biol. 2007;8(8):R164.
81. Peng J, Yang J, Jin Q. The molecular evolutionary history of *Shigella* spp. and enteroinvasive *Escherichia coli*. Infect Genet Evol. 2009;9(1):147–52.
82. Pupo GM, Lan R, Reeves PR. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc Natl Acad Sci U S A. 2000;97(19):10567–72.
83. Yang J, Nie H, Chen L, Zhang X, Yang F, Xu X, Zhu Y, Yu J, Jin Q. Revisiting the molecular evolutionary history of *Shigella* spp. J Mol Evol. 2007;64(1):71–9.
84. Fernandez IM, Silva M, Schuch R, Walker WA, Siber AM, Maurelli AT, McCormick BA. Cadaverine prevents the escape of *Shigella flexneri* from the phagolysosome: a connection between bacterial dissemination and neutrophil transepithelial signaling. J Infect Dis. 2001;184(6):743–53.
85. Prunier AL, Schuch R, Fernandez RE, Maurelli AT. Genetic structure of the *nadA* and *nadB* antivirulence loci in *Shigella* spp. J Bacteriol. 2007;189(17): 6482–6.
86. Bravo V, Puhar A, Sansonetti P, Parsot C, Toro CS. Distinct mutations led to inactivation of type 1 fimbriae expression in *Shigella* spp. PLoS One. 2015; 10(3):e0121785.
87. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev. 2014;38(5):865–91.
88. Vandecraen J, Chandler M, Aertsen A, Van Houdt R. The impact of insertion sequences on bacterial genome plasticity and adaptability. Crit Rev Microbiol. 2017;43(6):709–30.
89. Mao X, Ma Q, Liu B, Chen X, Zhang H, Xu Y. Revisiting operons: an analysis of the landscape of transcriptional units in *E coli*. BMC Bioinformatics. 2015;16:356.
90. Yamamoto M, Nomura M. Organization of genes for transcription and translation in the *rif* region of the *Escherichia coli* chromosome. J Bacteriol. 1979;137(1):584–94.
91. Horii T, Ogawa T, Ogawa H. Organization of the *recA* gene of *Escherichia coli*. Proc Natl Acad Sci U S A. 1980;77(1):313–7.
92. Jacob F, Perrin D, Sanchez C, Monod J. Operon: a group of genes with the expression coordinated by an operator. C R Hebd Seances Acad Sci. 1960; 250:1727–9.
93. Krasny L, Tiserova H, Jonak J, Rejman D, Sanderova H. The identity of the transcription +1 position is crucial for changes in gene expression in response to amino acid starvation in *Bacillus subtilis*. Mol Microbiol. 2008; 69(1):42–54.
94. Warrier I, Ram-Mohan N, Zhu Z, Hazery A, Echlin H, Rosch J, Meyer MM, van Opijnen T. The transcriptional landscape of *Streptococcus pneumoniae* TIGR4 reveals a complex operon architecture and abundant riboregulation critical for growth and virulence. PLoS Pathog. 2018;14(12):e1007461.
95. Albrecht M, Sharma CM, Dittrich MT, Muller T, Reinhardt R, Vogel J, Rudel T. The transcriptional landscape of *Chlamydia pneumoniae*. Genome Biol. 2011; 12(10):R98.
96. Schluter JP, Reinkensmeier J, Daschkey S, Evguenieva-Hackenberg E, Janssen S, Janicke S, Becker JD, Giegerich R, Becker A. A genome-wide survey of sRNAs in the symbiotic nitrogen-fixing alpha-proteobacterium *Sinorhizobium meliloti*. BMC Genomics. 2010;11:245.
97. Saenz-Lahoya S, Bitarte N, Garcia B, Burgui S, Vergara-Irigaray M, Valle J, Solano C, Toledo-Arana A, Lasa I. Noncontiguous operon is a genetic

organization for coordinating bacterial gene expression. Proc Natl Acad Sci U S A. 2019;116(5):1733–8.

98. Thomason MK, Bischler T, Eisenbart SK, Forstner KU, Zhang A, Herbig A, Nieselt K, Sharma CM, Storz G. Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in *Escherichia coli*. J Bacteriol. 2015;197(1):18–28.
99. Conway T, Creecy JP, Maddox SM, Grissom JE, Conkle TL, Shadid TM, Teramoto J, San Miguel P, Shimada T, Ishihama A, et al. Unprecedented high-resolution view of bacterial operon architecture revealed by RNA sequencing. mBio. 2014;5(4):e01442–14.
100. Miravet-Verde S, Llorens-Rico V, Serrano L. Alternative transcriptional regulation in genome-reduced bacteria. Curr Opin Microbiol. 2017;39:89–95.
101. Wen J, Harp JR, Fozo EM. The 5 UTR of the type I toxin ZorO can both inhibit and enhance translation. Nucleic Acids Res. 2017;45(7):4006–20.
102. Dong F, Xia L, Lu R, Zhao X, Zhang Y, Zhang Y, Huang X. The *malS*-5'UTR weakens the ability of *Salmonella enterica* serovar Typhi to survive in macrophages by increasing intracellular ATP levels. Microb Pathog. 2018; 115:321–31.
103. Heroven AK, Nuss AM, Dersch P. RNA-based mechanisms of virulence control in *Enterobacteriaceae*. RNA Biol. 2017;14(5):471–87. https://doi.org/10.1080/15476286.2016.1201617.
104. Di Martino ML, Romilly C, Wagner EG, Colonna B, Prosseda G. One Gene and Two Proteins: a Leaderless mRNA Supports the Translation of a Shorter Form of the *Shigella* VirF Regulator. mBio. 2016;7(6):e01860–16. https://doi.org/10.1128/mBio.01860-16.
105. Blomberg P, Wagner EG, Nordstrom K. Control of replication of plasmid R1: the duplex between the antisense RNA, CopA, and its target, CopT, is processed specifically in vivo and in vitro by RNase III. EMBO J. 1990;9(7): 2331–40.
106. Wingett SW, Andrews S. FastQ Screen: A tool for multi-genome mapping and quality control. F1000Research. 2018;7:1338.
107. Ewels P, Magnusson M, Lundin S, Kaller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047–8.
108. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16(6):276–7.
109. Deana A, Celesnik H, Belasco JG. The bacterial enzyme RppH triggers messenger RNA degradation by 5′ pyrophosphate removal. Nature. 2008; 451(7176):355–8.
110. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.
111. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357–9.
112. Hilker R, Stadermann KB, Schwengers O, Anisiforov E, Jaenicke S, Weisshaar B, Zimmermann T, Goesmann A. ReadXplorer 2-detailed read mapping analysis and visualization from one single source. Bioinformatics. 2016; 32(24):3702–8.
113. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.