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

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

[1]The Laboratory for Molecular Infection Medicine Sweden (MIMS), Umeå Centre for Microbial Research (UCMR) and Department of Molecular Biology, Umeå University, 90 187 Umeå, Sweden

Abstract

Background

Shigella is a Gram-negative, facultatively 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* serovar 5a M90T has been extensively used to understand the molecular mechanisms of pathogenesis, its complete genome sequence is not available, greatly limiting studies employing high-throughput sequencing and systems biology approaches.

Results

We have sequenced, assembled and annotated the full genome of *S. flexneri* 5a M90T. This yielded two complete contigs, the chromosome and the virulence plasmid. Further, we have performed genome-wide analysis of transcriptional start sites in bacteria grown in tryptic soy broth at 37 °C to mid-exponential phase, corresponding to the typical culture conditions for the inoculum in *in vitro* infection experiments. We have used the results from the transcriptional start site determination to manually curate the gene structure annotation. This analysis identified ~2,000 transcriptional units.

Conclusions

We provide the first complete genome for a *S. flexneri* serovar 5a strain, specifically the laboratory reference strain M90T. This opens the possibility of employing *S. flexneri* M90T in high-quality systems biology studies, for example transcriptomic analyses and differential expression analysis. Moreover, in molecular pathogenesis studies our data can be used as a resource to know which genes are transcribed before infection of host cells, thereby allowing to consider or exclude the possible involvement of a gene of interest.

Keywords (3-10)

Shigella flexneri serovar 5a M90T; Genome; Transcriptional start sites; Chromosome; Virulence plasmid

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 (<5 years old) and immunocompromised adults[1, 2]. In 2010, around 188 million cases of shigellosis occurred globally, including 62.3 million cases in childen 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.

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

Shigella flexneri serovar 5a M90T is the laboratory reference strain for *S. flexneri*. Indeed, the vast majority of our knowledge on the molecular mechanisms of *Shigella* pathogenesis has been obtained using *S. flexneri* M90T as model. The genome of this strain is composed of one chromosome and one giant virulence plasmid, called pWR100[8]. The pathogenesis of *Shigella* spp.

strictly depends on the virulence plasmid, which encodes several factors that are essential to invasion and subversion of host defenses[9].

Owing to its primary importance, the virulence plasmid of *S. flexneri* 5a M90T is the first genomic element that was sequenced in this strain[10]. However, the virulence plasmid (renamed pWR501 in this study, because it was marked with a transposon to select different regions of the plasmid) was sequenced with a now obsolete technology (ABI377 sequencer, Applied Biosystems)[10]. To sequence in ABI377 instruments, it was necessary to nebulize the DNA and size fractionate it by agarose gel electrophoresis to obtain fragments in the range of 0.7 to 2.0 kb, followed by cloning into cosmids for sequencing[10]. Using this protocol, the probability to lose some fragments or to introduce mutations is high in comparison with new technology such as PacBio[11] that is cloning- and PCR-free.

So far, chromosomally encoded genes have received little attention in *Shigella* research, as most of the work in this field has been focused on the plasmid genes. However, some of the genes codified on the chromosome could play an important role in *Shigella* pathogenesis. The *S. flexneri* 5a M90T chromosome has been sequenced and assembled[12], but unfortunately this sequence is not complete as only a genome scaffold was obtained.

Moreover, the assembly was prepared based on another *S. flexneri* strain, namely *S. flexneri* serovar 5b 8401[13].

In conclusion, 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.

Genes in prokaryotic cells can have more than one transcriptional start site (TSS). Typically, transcription starts in position -20 / -40 from the first translatable codon[14]. However, it is already known that in many bacteria the transcriptional start position is variable depending on the environment. Further, it is also known that TSSs vary depending on how bacteria respond to a specific stimulus[15].

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[16]. The differential RNAseq (dRNAseq) approach used here to determine TSSs, exploits the properties of a 5'-monphosphate-dependent exonuclease (TEX) to selectively degrade processed transcripts, thereby enriching for unprocessed RNA species carrying a native 5'-triphosphate[16]. TSSs can then be identified by comparing TEX-treated with untreated RNAseq libraries, as they appear as localized maxima in coverage enriched upon TEX-treatment[17].

Here we present the full, high-quality, annotated genome of *S. flexneri* serovar 5a M90T. Further, we identify all the genes that are expressed during mid-exponential growth in tryptic soy broth (TSB), the typical condition used for *in vitro* infections with *Shigella*. Moreover, we determine all the active transcriptional start sites during mid-exponential growth in TSB and also detect some RNA regulatory elements that are localized in the 5′-UTR regions.

Results

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

To determine the genome sequence of *S. flexneri* serovar 5a strain M90T whole-genome sequencing was conducted with 3 cells sequencing in a PacBio single-molecule real-time (SMRT) sequencing system. This generated a raw output of 93,316 subreads with mean length of 8,387 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 correct any possible sequencing error.

Genome assembly was carried out with Canu/1.7, feeding PacBio raw data. This analysis generated two contigs without any gaps and suggested circular replicons. For the larger contig, the output from Canu retained 14,193 reads of 5,938 bp average read length, with a total contig length of 4,596,712 bp, suggesting 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 232,191 bp. The size of the smaller replicon strongly suggests that it correspond to the virulence plasmid. These two replicons are around the expected size for the chromosome and virulence plasmid of *S. flexneri* 5a M90T, according to previous results[10, 12].

Genome assembly polishing using RNAseq reads

We employed RNAseq results to polish the assembled genome, using reads from RNAseq experiments performed on an Illumina HiSeq system. For the first round of polishing, we used the assembled genome as a reference to align the reads generated with RNA from which the rRNA was depleted with RiboZero (RNAseq-RZ) using the BWA software. The RNAseq-RZ polish step allowed us to polish all transcribed regions independently of posttranscriptional processing, as with this method of rRNA depletion all other classes of RNAs are retained. The resulting alignment was used to feed Pilon for a second round of iterative genome assembly polishing. This second round of polishing was performed with the data set 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 Pilon. The final polished genome assembly yielded one chromosome of 4,596,714 bp length and the virulence plasmid pWR100 with 235,195 bp of length. Both replicons were gap-free and circular molecules. The full genome sequence was deposited in GenBank with the accession numbers: CP037923 (chromosome) and CP037924 (pWR100).

The genome sequence that we report here contains some substantial differences compared to the previously sequenced genome[10, 12]. The chromosome of *Shigella flexneri* 5a M90T has 43,018 pb more than the previous version of *S. flexneri* M90T Sm, a streptomycin resistant derivative of M90T[12]. This can be explained by technological advances. In fact, here genome sequence was carried out with long read sequencing and the final genome assembly resulted devoid of gaps. In the older study[12] genome sequencing was performed using Illumina technology, which yields shorter reads. This can result in gaps during genome assembly, especially in repetitive sequences. The virulence plasmid pWR100 alone has been previously sequenced two further times by different groups. The first version[9] has 213,494 pb of length, and contains many gaps. The second version[10] has 221,851 bp of length, 8357 bp more than the first version. The sequence of pWR100 that we are reporting here is 232,190 bp of total length, 10,339 bp more than in [10]. Again, technological advances can explain these incongruences. Both previously sequenced versions were obtained using ABI377 technology, which required the construction of a cosmid library for sequencing. This protocol is prone to loss of DNA fragments, which could be the reason for creating gaps in the sequence.

Another important characteristic of the *S. flexneri* 5a M90T genome is its high content of insertion sequences and repeated regions that further complicate the genome assembly process. The PacBio sequencing protocol employed here is very well-suited to avoid poor assembly especially with a genome with a high content of repetitive or insertion sequences. In conclusion, the approach used here allowed us to report the *S. flexneri* 5a M90T genome sequence with much less errors.

Gene prediction and functional annotation

Gene prediction was carried out with three different pipelines; a) RAST, b) PGAP/NCBI and c) PROKKA. The number of predicted genes, RNAs and CDS were different for all three analyses (table). This can at least in part be explained by the fact that all the pipelines used for gene prediction and annotation employ different database for homology searches. The RAST pipeline used the Taxonomy ID: 1086030 from NCBI, which corresponds to *S. flexneri* serotype 5a strain M90T. RAST was able to predict and annotate 5,299 genes, but was unable to predict any ncRNA. The pipeline PGAP/NCBI was less efficient for gene prediction and annotation. The PGAP pipeline was able predict and annotate 4,077 genes plus 784 pseudogenes (frameshifted=406, incomplete=305, internal stop=166 and multiple problems=103). Gene prediction and annotation with PROKKA was the most efficient pipeline. We were able to predict and annotate 5,021 genes plus 220 ncRNAs with this pipeline. The most relevant difference of PROKKA in comparison with the other two pipelines is that PROKKA uses other and multiple databases, namely Rfam, Aragon, RNAmmmer and Prodigal, to find sequence homologies.

In other members of the Enterobacteriaceae family that have been previously sequenced it has been shown that they have a high number of pseudogenes (reference). Pseudogenes are generally classified as a reminiscence of an evolutive process, but if these genes are actively transcribed, it means that they can play a role in the bacterial gene expression network under specific condition, for example some of them could work as regulatory RNAs.

TSS annotation

To obtain differential RNAseq (dRNAseq) data, RNA samples were obtained from triplicate *S. flexneri* 5a M90T cultures grown in Tryptic Soy Broth (TSB) at 37 °C and 180 rpm until OD_{600}=0.3. This resulted in a dataset of ~120 million reads mapped to the previously completed reference genome of *S. flexneri* 5a M90T. A total of XXXXX TSS were automatically annotated based on the dRNAseq data, evenly distributed on forward and reverse strands. These were then categorized according to their position in relation to annotated genes: TSS in intergenic regions, located ≤300 nt upstream of the start codon and on the sense strand of an annotated gene, were assigned as primary TSS (pTSS). TSS within an annotated genes were assigned as internal sense (isTSS) or antisense (asTSS) when they were found on the sense or antisense strand, respectively. TSS in intergenic regions and not associated with any gene were assigned as "orphan" (oTSS). When TSS were positioned within 100 nt of a primary or orphan TSS and on the same strand, they were designated as secondary (sTSS).

Length of 5′ UTRs and leaderless transcripts

The average length of the 5′UTR in *S. flexneri* 5a M90T is XXXX nt (figure), with a distribution peak between XXX and XXX nt; XXX % of 5′UTRs were between XXXX and XXXX nt long. This is in the same range as other bacteria such as *Salmonella enterica*[18], *Helicobacter pylori*[17] and *Streptomyces coelicolor*[19].

The length of a 5′UTR can provide insight into the regulation of gene expression. Long 5′UTRs may contain riboswitches or provide binding sites for small regulatory RNAs (reference). Leaderless genes are translated by a different mechanism than genes with a leader sequence, and have been shown to be differentially regulated under stress conditions compared to leader-lead genes.

Promoter prediction and annotation based on dRNAseq

After completion of TSS annotation, we identified and annotated promoters that were associated to transcribed genes or operons based on dRNAseq results. For this purpose, specific promoter motifs were mapped to sequences upstream of previously identified TSSs. We were able to identify XXXXX active promoters during growth in TSB.

Material and methods

Bacterial strain and culture condition

The *Shigella* strain that was used for sequencing was obtained from Dr. Philippe Sansonetti, Institut Pasteur, Paris, France [8]. *Shigella flexneri* serovar 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, subcultured 1:100, and grown at 37°C in a shaking incubator at 150 RPM.

DNA purification and genome sequencing

Genomic DNA was isolated from overnight cultures of *S. flexneri* 5a M90T according to the kit manufacturer's instructions (Wizar^{R} Genomic DNA purification kit, Promega, Inc.). Isolated DNA was cleaned up as many times as necessary with phenol-chloroform (until no white interphase between the water and organic phase was forming)[20] to obtain a high quality and quantity of genomic DNA (20 μg) for PacBio library preparation[11]. Library preparation was carried out by Novogene. Sequencing was performed using a PacBio RSII sequencer at Novogene Inc., Hong Kong, China.

RNA purification and sequencing

S. flexneri 5a M90T was subcultured 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)[20]. 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 13000

RPM at 4°C using a table-top centrifuge. 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 5 min at 65°C. Afterwards, total RNA was extracted using the hot-phenol method[16]. Contaminating DNA was digested by DNase I (Roche; 1 U/□g RNA, 60 min, 37°C) in the presence of RNase inhibitor (RNaseOUT, ThermoFisher Scientific; 0.1 U/□l) followed by clean up of RNA by phenol/chloroform/isoamyl alcohol and precipitation of RNA by 2.5 volumes of ethanol containing 0.1 M sodium acetate pH 5.5 and 20 □g of glycogen (Roche)[20]. 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 309 pb long product of the *hfq* gene from *S. flexneri* 5a M90T.

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 (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 Illumina Genome Analyzer 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[21, 22]. These RNAseq results obtained from Novogene were used to polish the genome assembly.

Genome assembly and annotation

De novo genome assembly was performed with the script Canu/1.7[23] 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[24]. Polishing of genome assembly was done in two rounds: the first one was carried out using the RNAseq output files from the samples in which the rRNA was depleted with RiboZero (RNAseq-RZ); the second one was carried out with the RNAseq results from the samples in which the rRNA was depleted with TEX (RNAseq-

TEX). Genome annotation and polishing was ran at Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) of SciLifeLab at Uppsala University, Sweden.

Annotation of assembled contigs was done using three different pipelines: 1) NCBI Prokaryotic Genome Annotation Pipeline (PGAP)[25], 2) prokka/1.12-12547ca[26] and 3) Rapid Annotation using System Technology version 2.0 (RAST)[27]. Genome annotation with Prokka was ran on the UPPMAX server facilities at Uppsala University, Sweden.

The assembled and annotated genome was manually curated using Artemis[28] for visualizing and editing the genome files. The genome was deposited in GenBank with accession numbers CP037923 (chromosome) and CP037924 (virulence plasmid).

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[16], while RNAs containing a 5′ mono-phosphate (5′P) are selectively degraded[16, 17]. The rRNA depleted RNA was split into two aliquots. One aliquot was treated with Terminator 5′-Phosphate-Dependent Exonuxlease (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/0.1M sodium acetate (pH 5.5) and 20 μg of glycogen (Roche) mixture. After TEX treatment both samples (TEX+ and TEX-) were treated with 5′ Pyro phosphohydrolase (RppH) (NewEngland BioLabs, Inc.) to generate 5′-mono-phosphates for linker ligation, and again purified by organic extraction and ethanol precipitation. RppH[29] treatment was carried out for 60 min at 37°C. An RNA adaptor (5′-GACCUUGGCUGUCACUCA-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.

A separate library was constructed for TEX- and TEX+ samples. The libraries were constructed using Illumina Genome Analyzer 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[21, 22].

Reads mapping of TSS library

Reads in the FASTQ format were cleaned up with trimmomatic/0.36[30] to remove sequences originating from Illumina adaptors and low quality reads. The files were aligned with the previously assembled reference genome of *S. flexneri* 5a M90T (accession numbers CP037923 and CP037924) with bowtie2/2.3.4.3[31] using –X 1000 such that only mate pairs were reported if separated by less than 1000 bp. All the other setting were implemented with the default option. After alignment was completed, samtools/1.9[32] was used to remove duplicates and select for reads that were aligned in proper pairs. The number of reads aligned to the reference genome was summarized using coverageBed (part of the bedtools software package)[33]. Reads per Kb/million reads (RPKM) was calculated as a measure of expression of all genes individually[34] using the formula: RPKM = number of mapped reads / total number of reads / gene length x 1000,000,000.

Transcriptional start sites annotation and classification

To map RNAseq output reads, reads were split by replicon, converted to BAM format and sorted by position with Sam tools/1.9[28]). These BAM files were used as input for TSSAR (TSS annotation regime for dRNAseq data)[35] for automatic *de novo* TSS annotation. Default parameters (p-value threshold 0.0001, noise threshold 2, merge range 5) were used for automatic TSS annotation. For the analysis, the results from the three biological replicates were pooled and TSS within 5 nt of each other were clustered into one. Genome regions with read start distribution that do not conform to Poisson distribution are omitted from TSSAR analysis[35]. Such regions were then manually annotated by scanning the respective wiggle files for nucleotides with an abrupt increase in coverage. Transcriptional start sites 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 internal. Peaks within gene boundaries and on the opposite strand from the gene were classified as antisense. 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 position start with -1.

Conclusions

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 (London, England)* 2013, **382**(9888):209-222.
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.** *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2014, **59**(7):933-941.
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 (London, England)* 2012, **380**(9859):2095-2128.
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.** *Bulletin of the World Health Organization* 1999, **77**(8):651-666.
5. Kotloff KL, Riddle MS, Platts-Mills JA, Pavlinac P, Zaidi AKM: **Shigellosis.** *Lancet (London, England)* 2018, **391**(10122):801-812.
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):1126-1128.
7. The HC, Thanh DP, Holt KE, Thomson NR, Baker S: **The genomic signatures of *Shigella* evolution, adaptation and geographical spread.** *Nature reviews Microbiology* 2016, **14**(4):235-250.
8. Sansonetti PJ, Kopecko DJ, Formal SB: **Involvement of a plasmid in the invasive ability of *Shigella flexneri*.** *Infection and immunity* 1982, **35**(3):852-860.
9. 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):760-771.
10. 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*.** *Infection and immunity* 2001, **69**(5):3271-3285.

11. Rhoads A, Au KF: **PacBio Sequencing and Its Applications.** *Genomics, proteomics & bioinformatics* 2015, **13**(5):278-289.
12. Onodera NT, Ryu J, Durbic T, Nislow C, Archibald JM, Rohde JR: **Genome sequence of *Shigella flexneri* serotype 5a strain M90T Sm.** *Journal of bacteriology* 2012, **194**(11):3022.
13. 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.
14. 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.
15. 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):683-695.
16. Schoenberg DR: **The end defines the means in bacterial mRNA decay.** *Nature chemical biology* 2007, **3**(9):535-536.
17. 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):250-255.
18. Kroger C, Dillon SC, Cameron AD, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hebrard M, Handler K *et al*: **The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium.** *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**(20):E1277-1286.
19. 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 biology* 2011, **8**(3):468-477.
20. 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.** *The EMBO journal* 1990, **9**(7):2331-2340.
21. Wingett SW, Andrews S: **FastQ Screen: A tool for multi-genome mapping and quality control.** *F1000Research* 2018, **7**:1338.
22. Ewels P, Magnusson M, Lundin S, Kaller M: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics (Oxford, England)* 2016, **32**(19):3047-3048.

23. 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 research* 2017, **27**(5):722-736.

24. 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.

25. 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):6614-6624.

26. Seemann T: **Prokka: rapid prokaryotic genome annotation.** *Bioinformatics (Oxford, England)* 2014, **30**(14):2068-2069.

27. 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.

28. 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 (Oxford, England)* 2012, **28**(4):464-469.

29. Deana A, Celesnik H, Belasco JG: **The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal.** *Nature* 2008, **451**(7176):355-358.

30. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics (Oxford, England)* 2014, **30**(15):2114-2120.

31. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nature methods* 2012, **9**(4):357-359.

32. 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 (Oxford, England)* 2009, **25**(16):2078-2079.

33. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics (Oxford, England)* 2010, **26**(6):841-842.

34. Li P, Piao Y, Shon HS, Ryu KH: **Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data.** *BMC bioinformatics* 2015, **16**:347.

35. Amman F, Wolfinger MT, Lorenz R, Hofacker IL, Stadler PF, Findeiss S: **TSSAR: TSS annotation regime for dRNA-seq data.** *BMC bioinformatics* 2014, **15**:89.

Table 1. Summary of PacBio raw data.

Sample ID	Number of Reads	Number of Bases(bp)	Mean Read Length(bp)	N50 Read Length(bp)
SF5aM90T	93,316	782,710,041	8,387	12,275

Table 2. Summary of features annotated by PROKKA, RAST and PGAP/NCBI automatic annotation pipeline.

Feature	PROKKA	RAST	PGAP/NCBI
Contigs	2	2	2
Bases	4,828,902	4,828,902	4,828,902
rRNA	22	22	22
mRNA	5,367	5,299	4,077
Gene	5,367	5,299	4,077
CDS	5,021	5,299	4,861
tRNA	103	123	102
ncRNA	220	0	7
Pseudogenes	0	0	784

Figure 1. Circular map of the chromosome and the virulence plasmid pWR100. The outer ring (grey) depicts the length of the replicon, the inner rings (red) show the ORFs on both strands.

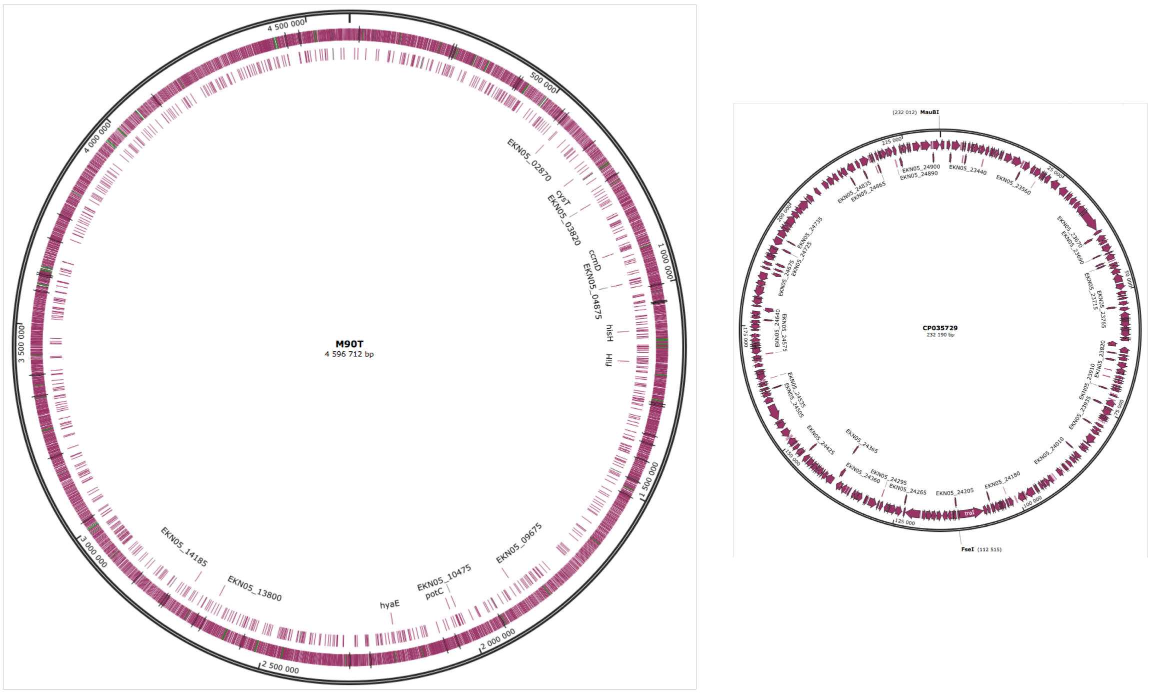
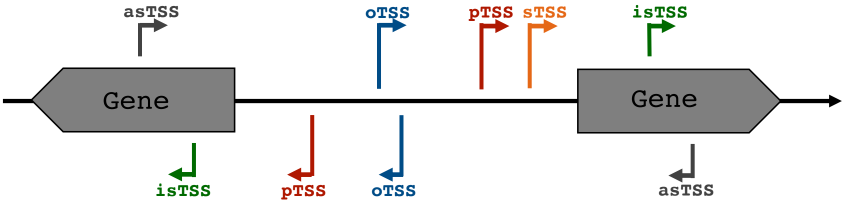


Figure 2. Schematics of categorisation of transcription start sites into primary (pTSS), secondary (sTSS), internal sense (isTSS), internal antisense (asTSS) and orphan (oTSS) based on differential RNA sequencing.



Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Accesion numbers: CP035729 and CP035729

Data are deposited at xxx under the number yyy.

- The datasets generated and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS]

Request for material should be directed to and will be fulfilled by Andrea Puhar (andrea.puhar@umu.se).

Competing interests

The authors declare no competing interests.

Funding

We gratefully acknowledge funding from the Kempe Foundation, the Knut and Alice Wallenberg Foundation, Umeå University and The Laboratory for Molecular Infection Medicine Sweden (MIMS).

Author's contribution

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.

Acknowledgements

We are grateful for travel support and courses organized by The National Doctoral Programme 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.