4 Essential genes in Typhi and Typhimurium

4.1 Introduction

Both Typhi and Typhimurium are important human pathogens and multiple strains of both serovars have been sequenced, revealing approximately 99% similarity between orthologous coding sequences at the amino acid level (Deng et al. 2003; Holt et al. 2008; McClelland et al. 2001; Parkhill et al. 2001a). Despite this similarity, Typhi is the causative agent of typhoid fever, exclusively affecting humans while Typhimurium is a leading cause of foodborne gastroenteritis and infects a wide range of mammals and birds. Key areas of investigation have been to find explanations for how these organisms are capable of causing such different disease phenotypes in humans, and why Typhi is human-restricted.

The question of which genes are required for cellular viability is of fundamental importance to biology. Recent developments have seen estimates of the minimal gene set required for bacterial cell viability, via single gene deletion (Baba et al. 2006; de Berardinis et al. 2008), random mutagenesis, resulting in vast libraries where a single gene is represented by multiple mutants (French et al. 2008; Gallagher et al. 2007) or ordered mutagenesis to produce one mutant per gene (Liberati et al. 2006). Such studies performed in the laboratory, by their nature, identify those bacterial genes required for viability under specific laboratory growth conditions. Signature-tagged mutagenesis (STM) (Hensel et al. 1995) and transposon-site hybridisation (TraSH) (Sassetti et al. 2001) are methods that make use of hybridisation to identify genes disrupted by transposon insertions. They are also negative selection methods, whereby transposon mutants lost under selection through a functional screen represent the genes required for that function, and have been used to identify "niche-specific" virulence genes in bacterial pathogens (reviewed in (Andrews-Polymenis et al. 2009)). Most recent bacterial transposon mutant library screens have used a few thousand transposon mutants, which represent, on average, several insertions per gene (Gallagher et al. 2007; Glass et al. 2006; Laia et al. 2009; Liberati et al. 2006; Salama et al. 2004). A more recent method, transposon mediated differential hybridisation (TMDH) (Charles and Maskell 2001), has been used to analyse approximately one million mutants to identify essential genes in *Staphylococcus aureus* (Chaudhuri et al. 2009). However, these approaches are all sub-optimal, due to inaccuracy in the estimation of the transposon insertion site from microarrays, and because some genes, especially those which are smaller, will be missed by chance.

Current knowledge regarding how Typhi and Typhimurium differ is, for the most part, concentrated upon virulence and pathogenicity factors. Prior to genome sequencing, the presence of plasmids was shown to affect *Salmonella* virulence. A 100kb plasmid, encoding the *spv* (*Salmonella* plasmid virulence) genes, is found in some Typhimurium strains and contributes significantly towards systemic infection in animal models (Gulig and Curtiss 3rd 1987; Gulig et al. 1993). Plasmids encoding similar functions have been found in many serovars of *Salmonella*, but never in Typhi, which historically was rarely found to harbour plasmids. However, since the early 1970s, plasmids of incompatibility group (Inc) HI1 have been isolated in Typhi, although these do not harbour the *spv* genes. Initially characterised based on genes encoding resistance to the first generation antibiotics used to treat typhoid (review: (Phan and Wain 2008)), there is also evidence that Typhi strains carrying an IncHI1 plasmid present a higher bacterial load in the blood during human infection (Wain et al. 1998).

The horizontal acquisition of pathogenicity islands during the evolution of the salmonellae is also believed to have impacted upon their disease potential. *Salmonella* pathogenicity islands (SPI)-1 and -2 are common to both serovars, and are required for invasion of epithelial cells (review in (Darwin and Miller 1999)) and survival inside macrophages respectively (Ochman et al. 1996; Shea et al. 1996). SPI-7 and SPI-10, however, are unique to Typhi (with respect to Typhimurium). SPI-7 harbours putative virulence genes encoding the Vi capsular polysaccharide, a SopE effector protein and a type IVB pilus system (Pickard et al. 2003; Seth-Smith 2008). The *sef* chaperone/usher fimbrial operon is encoded on SPI-10 and has been implicated in virulence and host adaptation (Townsend et al. 2001).

However, the acquisition of virulence determinants is not the sole explanation for the differing disease phenotypes displayed in humans by Typhimurium and Typhi. Indeed, genome degradation is a feature of the Typhi genome, in common with other host-restricted serovars such as Paratyphi A (humans) and Gallinarum (chickens). In each of these serovars, pseudogenes account for 4-7% of the genome (Holt et al. 2009b; Thomson et al. 2008). Loss of gene function has occurred in *shdA*, *ratB* and *sivH* of Typhi, genes which have been shown to encode intestinal colonisation and persistence determinants in Typhimurium (Kingsley et al. 2003). Numerous sugar transport and degradation pathways have also been interrupted (Parkhill et al. 2001a), but remain intact in Typhimurium, potentially underlying the restricted host niche occupied by Typhi.

Given the close phylogenetic relationship between Typhi and Typhimurium, a whole genome approach was required to ask whether these organisms are different not only in clinical phenotype but also the base gene set required for survival. Using Illumina-based transposon directed insertion-site sequencing (TraDIS) with large mutant libraries of both Typhimurium and Typhi, we investigated whether these *Salmonella* share the same essential gene set, or whether there are differences which reflect intrinsic differences in the pathogenic niches these bacteria inhabit.

4.2 Methods

4.2.1 Strains

The Typhimurium strain used was SL3261, which contains a deletion relative to the parent strain, SL1344. The 2166bp deletion stretches from 153bp within *aroA* (normally 1284bp) to the last 42bp of *cmk*, hence forming two pseudogenes and deleting the intervening gene SL0916 completely.

The Typhi strain used was WT26 pHCM1, a derivative of the attenuated Ty2-derived strain CVD908-*htrA* which has stable deletion mutations in *aroC*, *aroD* and *htrA* (Tacket et al. 1997). WT26 (Turner et al. 2006) has a point mutation in *gyrA* conferring reduced susceptibility to fluoroquinolone antibiotics and the multiple antibiotic resistance plasmid, pHCM1, has been introduced. These additions were intended to allow the transposon mutant library to be used for fluoroquinolone resistance and plasmid studies.

4.2.2 Transposome preparation

The TraDIS transposon is a derivative of EZ-Tn5 <R6Kγori/KAN-2> (Epicentre Biotechnologies, Wisconsin, USA) with outward oriented T7 and SP6 promoters at each end, and with R6Kγori deleted. The transposon was PCR amplified using Pfu Ultra Fusion II (Stratagene, California, USA) and the following oligonucleotides:

5'-CTGTCTCTTATACACATCTCCCT

5'-CTGTCTCTTATACACATCTCTTC

The resulting amplicon was phosphorylated using polynucleotide kinase (New England Biolabs, Hitchin, UK). 400 ng of this DNA were incubated with EZ-Tn5™ transposase (Epicenter Biotechnologies) at 37^{o}C for 1h then stored at -20 ^{o}C.

4.2.3 Cell transformation and transposon library creation

In the following section, the Typhi mutant library was generated by Keith Turner and Duy Phan, the Typhimurium library by myself.

Bacterial cells (Typhi or Typhimurium) for electrotransformation were grown in 2 x TY broth to an OD_{600} of 0.3 – 0.5, then cells were harvested and washed three times in ½ x vol 10% glycerol. Cells were finally resuspended in 1/1000 x vol 10% glycerol and stored at -80 ^{o}C. 60 µl cells were mixed with 0.2 µl transposomes and electrotransformed in a 2 mm electrode gap cuvette using a BioRad GenePulser II set to 1.4 kV, 25 µF and 200 Ω. Cells were resuspended in 1 mL SOC medium (Invitrogen, Paisley, UK) and incubated at 37 ^{o}C for 2h then spread on L-agar supplemented with kanamycin at 7.5 µg/mL and 20 µg/mL for Typhi and Typhimurium respectively. For Typhi, the L-agar was also supplemented with “aro mix” (40 µg/ml each of L-phenylalanine and L-tryptophan, and 10 µg/ml each of *p*-aminobenzoic acid and 2,3-dihydroxybenzoic acid final concentration). After incubation overnight at 37 ^{o}C, the number of colonies on several plates was estimated by counting a proportion of them, and from this the total number of colonies on all plates was estimated conservatively. Kanamycin resistant colonies were scraped off plates and resuspended in sterilised deionised water using a bacteriological spreader.

Typically, ten or more electrotransformations were performed to generate one batch of mutants. The number of mutants in each batch ranged from 42,000 to 148,000. From the estimated total number of mutants and using the OD_{600} to estimate the cell concentration in each batch, volumes containing approximately similar numbers of mutants from 13 batches were pooled to create the Typhi mutant library mixture, estimated to include 1.1 million mutants. Ten batches were similarly pooled to create the Typhimurium mutant library, estimated to include 930,000 mutants.

4.2.4 DNA manipulation and sequencing

Daniel Turner developed and helped to optimise TraDIS and performed the nucleotide sequencing for Typhi and Typhimurium. Duy Phan and I performed the sequence analysis required to optimise TraDIS.

4.2.4.1 Optimisation using Typhi

Five µg of Typhi genomic DNA from the mutant pool was fragmented to an average size of either 200 or 300 bp by Covaris AFA (Quail et al. 2008) and the Illumina DNA fragment library preparation was performed following the manufacturer’s instructions, but using 1.5x the recommended reagent volumes in each step. DNA fragments from the library, ligated to Illumina adapters, were run in a 12 cm 2 % agarose gel in 1 x TBE buffer, at 6 V cm^{-1} without the preceding column clean up step. After 45 minutes, fragments corresponding to an insert size of 250-350 bp were excised, and DNA was extracted from the gel slice without heating (Quail et al. 2008). The DNA was quantified on an Agilent DNA1000 chip, following the manufacturer's instructions. Template DNA of either 0.1 ng, 1 ng, 25 ng, 65 ng or 100 ng was used in a PCR of 32 or 22 cycles to amplify insertion sites at the 3' end or the 5' end of the transposon, using the transposon-specific forward primer 5F (Sigma-Aldrich, Dorset, UK, HPLC purified):

5'-AATGATACGGCGACCACCGAGATCTACACCTGAATTACCCTGTTATCCCTATTTAGGTGAC

or 3F (Sigma-Aldrich, HPLC purified):

5'- AATGATACGGCGACCACCGAGATCTACACCTGACCTCTAGAGTCGACTGGCAAAC

and a custom Illumina reverse primer V3.3 (Sigma-Aldrich, HPLC purified):

5'-AAGCAGAAGACGGCATACGAGATCGGTACACTCTTTCCCTACACGACGCTCTTCCGATCT

The resultant libraries were sequenced on a paired or single end Illumina flowcell using an Illumina GAII sequencer for 36 or 54 cycles of sequencing, using 2x Hybridization Buffer and the custom sequencing primer 5TMDH2seq (Sigma-Aldrich, HPLC purified):

5'- ATCCCTATTTAGGTGACACTATAGAAGAGATGTGTA

or 3TMDH1seq (Sigma-Aldrich, HPLC purified):

5' - TTATGGGTAATACGACTCACTATAGGGAGATGTGTA

The custom sequencing primers were designed such that the first 10 bp of each read was transposon sequence.

4.2.4.2 Typhimurium

Five µg of Typhimurium genomic DNA from the mutant pool was fragmented to an average size of 300 bp by Covaris AFA (Quail et al. 2008) and the Illumina DNA fragment library preparation was performed as for the Typhi libraries above. To amplify the transposon insertion sites, 22 cycles of PCR were performed using 100 ng of DNA fragment library, and the transposon-specific forward primer 5F and custom Illumina reverse primer detailed above.

The amplified library was cleaned up with a QiaQuick PCR product purification column following the manufacturer's instructions, eluted in 30 µl EB, and then quantified by qPCR (Quail et al. 2008). The amplified DNA fragment library was sequenced on a single end Illumina flowcell using an Illumina GAII sequencer, for 36 cycles of sequencing, using 2x Hybridization Buffer and the custom sequencing primer 5TMDH2 detailed above.

4.2.5 Sequence analysis

Using a custom Perl script co-written with Duy Phan (Appendix 8.3.1), the Illumina FASTQ sequence files were parsed for 100% identity at the 5' end to the last 10bp of the transposon (`TAAGAGACAG`). Sequence reads which matched were stripped of the transposon tag and subsequently mapped to the appropriate reference chromosome using Maq version maq-0.6.8 (Li et al. 2008). Ty2 was used as the reference for Typhi and SL1344 for Typhimurium. Precise insertion sites were determined using the output from the Maq mapview command, which gives the first nucleotide position to which each read mapped. The number and frequency of insertions mapping to each nucleotide in the reference genome was then determined. Gene boundaries were defined from the Sanger in-house annotation of the Typhimurium SL1344 sequence, and the publically available Typhi Ty2 sequence (Accession number: AE014613) allowing the number and frequency of transposon insertions to be established for every gene. Genes were grouped into functional classes based on genome annotation (Table 4-1), and the average number of insertions per functional class was calculated by dividing the total number of insertions recovered for the class by the summed total of all gene lengths within the class.

Table 4-1 Functional categories in genome annotation

Function	Colour in genome annotation
Pathogenicity/adaptation	Dark blue*
Energy metabolism	Dark grey/black
Information transfer	Red
Membrane/surface structures	Green
Degradation of small molecules	Purple
Degradation of macromolecules	Cyan
Central/intermediary metabolism	Yellow
Unknown function	Pale green
Regulators	Light blue
Conserved/hypothetical	Orange
Pseudogenes	Brown
Phage/IS elements	Pink

Categories as annotated in the Typhi CT18 genome. *Pathogenicity/adaptation genes are usually coloured white, but for display purposes (e.g. Figure 4-2) have been coloured dark blue.

4.2.6 Statistical analyses

The nature of the statistical analysis was discussed with Leopold Parts who advised on the best way to proceed and subsequently wrote scripts in R to perform the following analyses.

4.2.6.1 High density mutagenesis

We calculated the *P*-value for the distances between insertion sites using F = G/N where G is the number of bases in the genome (4,791,961 / 4,878,012) and N is the number of unique insert sites (371,775 / 549,086) for Typhi and Typhimurium respectively. Across the whole genome, the *P*-value for at least X consecutive bases without an insert site is $e^{(-X/F)}$, giving a 5% cut-off at 39 bp and a 1% cut-off at 60 bp for Typhi. For Typhimurium, there was a 5% cut-off at 27 bp and a 1% cut-off at 41 bp.

To investigate the G+C content insertion site bias, the G+C content and number of insertion sites was calculated for a sliding window of 1 kb (with a 500 bp skip) along the Typhi genome. The average number of insertion sites for a given integer G+C content was determined, and used to calculate the average number of insertion sites in genomic regions with G+C content above and below the genome average (52%). The *P*-value for the distances between insertion sites was again calculated using F = G/N where G is the number of bases in the genomic regions with above/below average G+C content (4,357,000 / 5,225,000) and N is the average number of insert sites (615,547 / 273,576). The *P*-value for at least X consecutive bases without an insert site is again $e^{(-X/F)}$, giving a 5% cut-off at 58 bp and a 1% cut-off at 88 bp.

4.2.6.2 Essential genes per serovar

The number of insertion sites for any gene is dependent upon its length, so the values were made comparable by dividing the number of insertion sites by the gene length, giving an “insertion index” for each gene. The distribution of insertion indices was bimodal, corresponding to the essential (mode at 0) and non-essential models. We fitted gamma distributions for the two modes using the R MASS library (http://www.r-project.org). Log_2-likelihood ratios (LLRs) were calculated between the essential and non-essential models and we called a gene essential if it had an LLR of less than -2, indicating it was at least 4 times more likely according to the essential model than the non-essential model. ‘Non-essential’ genes were assigned for an LLR of greater than 2.

4.2.6.3 Comparison of essential genes between serovars

For every gene g present in both serovars, with $n_{g,A}$ reads observed in Typhi and $n_{g,B}$ reads observed in Typhimurium, we calculated the log_2 fold change ratio $S_{g,A,B} = log_2 \backslash frac(n_{g,A}+100)(n_{g,B}+100)$. The correction of 100 reads smoothes out the high scores for genes with very low numbers of observed reads. We fitted a normal model to the mode +/- 2 sample standard deviations of the distribution of $S_{A,B}$, and calculated p-values for each gene according to the fit. We considered genes to be uniquely essential to one serovar with a P-value of 0.05, according to the normal model.

4.3 Results

4.3.1 Whole genome TraDIS

4.3.1.1 Optimising TraDIS

To determine the optimum sequencing parameters, initially, 32 cycles of PCR were performed on 0.1ng and 1ng of fragmented Typhi DNA, in order to guarantee sufficient template for sequencing. One PCR primer was designed to anneal to the Illumina adapter sequence, and the other was transposon-specific, but tailed to allow cluster amplification later on in the protocol. The very high yield of PCR product (>100nM,), demonstrated that fewer cycles of amplification could be performed in subsequent PCRs. This would reduce amplification bias and potential dropout of rare transposon insertion sites from the resulting sequencing library.

Separate DNA fragment libraries were prepared for sequencing off each end of the transposon, to determine the degree of concurrence, and hence evaluate the completeness of the dataset. Initially these libraries were sequenced with 36 bp reads as paired end, with one end corresponding to the transposon / genomic DNA junction, and the other located 200-300bp away, in the Typhi genomic sequence.

Analysis of the resulting sequences established that read mapping for these sequences was equally robust with single or paired end reads, therefore the extra cost of a paired end run was not justifiable and all subsequent sequencing runs were performed as single end. The analysis also confirmed that more complete coverage was obtained when 1ng template DNA was used in the PCRs, compared to 0.1ng (Table 4-2). With the aim of minimising amplification biases as far as possible, a mixture of both quantitative PCR and regular PCR was used to establish the minimum number of PCR cycles and maximum realistic mass of template DNA that would give a workable yield of amplicon DNA for sequencing. Virtually no difference in amplification yield was observed for 65ng template DNA or above, and 22 cycles were found to be the fewest that gave a reasonable yield of amplified DNA.

Table 4-2 Validating TraDIS

	Sample	Total number of reads	Number of tagged reads (%)	Number of reads mapped to Ty2 (%)	Number of unique insert sites	Average distance between inserts (bp)
a	200bp	9108747	6978334 (77)	4248840 (61)	39103	122.5
	300bp	8678131	6139885 (71)	3686804 (60)	61250	78.2
b	200bp	6188899	4920008 (80)	2049222 (42)	168452	28.4
	300bp	6345624	5357249 (84)	2510720 (47)	225423	21.3
c	100ng	5170743	4019030 (78)	1391070 (35)	268284	17.9
	65ng	5526380	4091846 (74)	1384263 (34)	256431	18.7
	25ng	6368930	4835018 (76)	1837189 (38)	247522	19.4
d	2 lanes	9522310	7219913 (76)	2424118 (34)	312089	15.4
e	4 lanes	21417620	16146777 (75)	5645570 (35)	371775	12.9

A representative lane for each stage of refinement is shown. **(a)** 0.1ng genomic DNA samples, fragmented into 200/300bp lengths, underwent 32 PCR cycles prior to 36 cycles on sequencing. **(b)** 1ng genomic DNA samples, fragmented into 200/300bp lengths, underwent 32 PCR cycles prior to 36 cycles on sequencing. **(c)** Genomic DNA samples underwent 22 PCR cycles and 25-100ng PCR product underwent 54 cycles of sequencing. **(d)** Combined data from 2 sequencing lanes with same conditions as the 100ng sample from (c). **(e)** Combined data from 4 sequencing lanes; the samples from (d) and the 65ng and 25ng samples from (c). All samples were sequenced from the 5’ end of the transposon, with the exception of samples from (a), which were sequenced from the 3’ end.

To verify these results, DNA fragment libraries were prepared from both ends of the transposon, using 22 cycles of PCR and 25, 65 and 100ng of template DNA (fragmented into ~200 bp pieces) in the reaction. These libraries were sequenced in a single end run with a 54bp read length, in order to assess whether the longer reads enhanced read mapping. The greatest number of unique insertion sites were recovered from the 100ng sample (Table 4-2), although the longer sequencing reads did not improve read mapping. Overall, the least biased sequence dataset was obtained from the use of 100ng template DNA and 22 cycles of amplification in the PCR reaction. DNA fragment libraries prepared from each end of the transposon gave highly overlapping results but sequencing two lanes of a library from one end significantly increased the number of unique mutants recovered (Table 4-2). We therefore used 100ng DNA and 22 PCR cycles in all further DNA fragment library preparations from the 5' end of the transposon, and sequenced two lanes per library.

4.3.1.2 Insertion bias

Transposon insertion site bias is often cited as a limitation of transposon mutagenesis techniques. We detected a bias towards insertion in A+T rich regions in both mutant libraries (Typhi shown in Figure 4-1), but the frequency of insertion achieved by the number of transposon mutants ensured that even G+C-rich regions contained numerous insertions. In Typhi, the library with slightly lower coverage, the average insertion rate in G+C regions above 52% (genome average for both these *Salmonella*) was 1 every 19bp, meaning that a region of >88 bp without insertion had a less than 1% probability of occurring by chance. The equivalent values in G+C regions below 52% were 1 insert every 7 bp, with a less than 1% probability that a region >33 bp without insertion occurred by chance. Given that transposon insertion sites successfully delimited essential genes, we have no reason to believe that transposon insertion bias has had any bearing upon our conclusions.

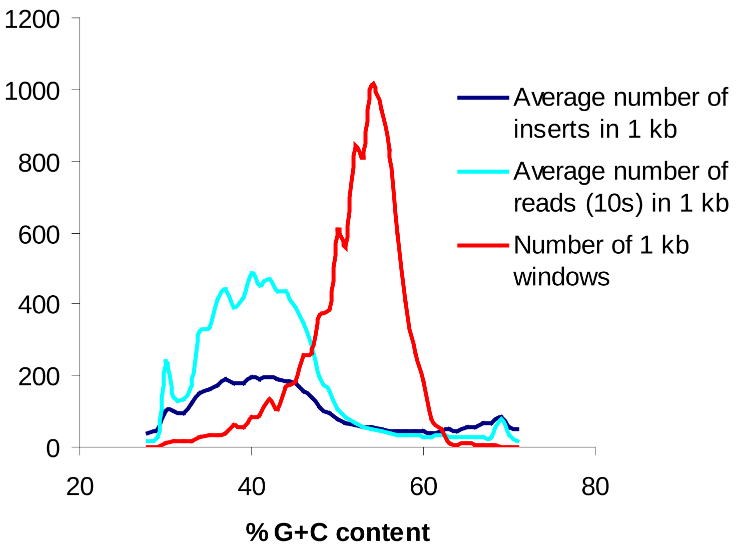

Figure 4-1 Transposon insertion site bias in Typhi

The %G+C content and number of insertions and sequence reads (in tens) was determined for every 1 kb window (with a 500 bp skip between windows) across the genome. Red line: the number of 1 kb windows with a particular %G+C content; dark blue line: the average number of transposon insertions found in 1 kb windows of a particular %G+C content; light blue line: the average number of mapped reads found in windows of a particular %G+C content, in tens. The average G+C content in *S.* Typhi is 52% but the highest number of insertion sites and mapped reads occur in windows with 40% G+C content.

4.3.1.3 Typhi

Genomic DNA was extracted from the ~1 million mutant pool for nucleotide sequencing from the transposon into the adjacent sequences of the insertion sites. Since Typhi had been used for optimisation of TraDIS, all four sequencing lanes using the optimised parameters were combined. These four lanes from the Illumina sequencing flow cell generated over 21 million nucleotide sequence reads, of which 75% included an identical match to the 10 base transposon nucleotide sequence tag (Table 4-2). Of these tagged sequence reads, 5.6 million were mapped to the Typhi Ty2 chromosome sequence. This allowed the identification of 371,775 individual transposon insertion sites; an average of 1 insertion site for every 13 bp. This represents an average of more than 80 inserts per gene, which is far in excess of the number of insertions achieved previously for bacterial transposon mutant libraries which have reported an average of 5 to 17 inserts per gene (Gallagher et al. 2007; Laia et al. 2009; Salama et al. 2004; Sassetti et al. 2001), and makes possible the assay of every gene in the genome. The distribution of mapped sequence reads across the whole genome is shown in Figure 4-2A.

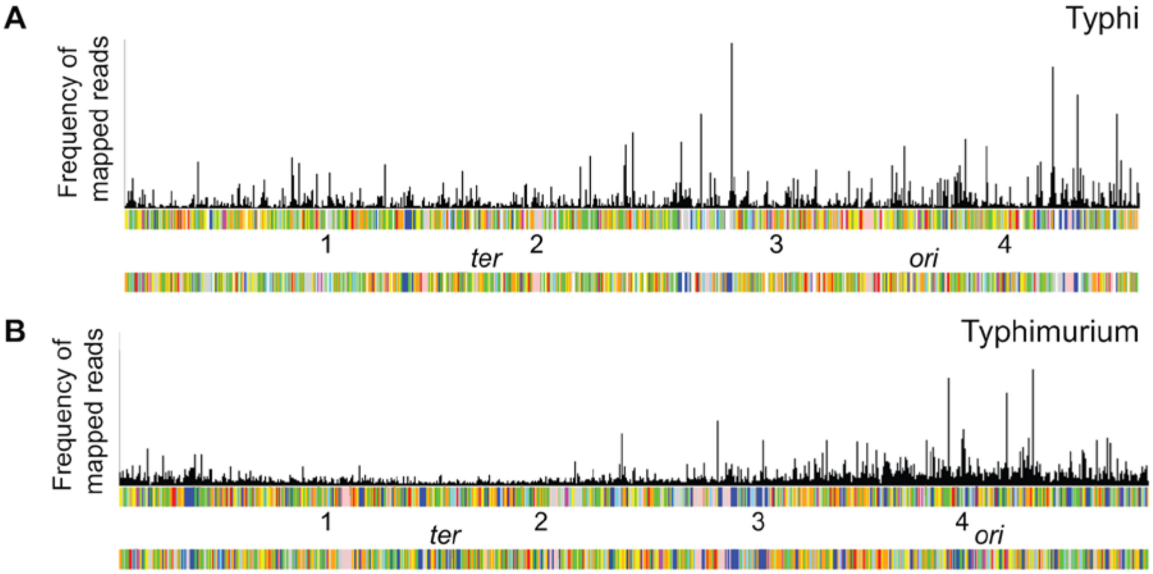

Figure 4-2 Whole genome view of transposon insertion sites

A) Frequency and distribution of transposon directed insertion-site sequence reads across A) the entire Typhi Ty2 genome and B) the entire Typhimurium SL1344 genome, scaled to heights of 2500 and 800, respectively. Numbers represent megabases; the y-axes show the number of mapped sequence reads within a window size of 3; *ori* and *ter* indicate the approximate positions of the replication origin and terminus respectively. Genes are colour-coded according to function: dark blue, pathogenicity/adaptation; black, energy metabolism; red, information transfer; dark green, membranes/surface structures; cyan, degradation of macromolecules; purple, degradation of small molecules; yellow, central/intermediary metabolism; light blue, regulators; pink, phage/IS elements; orange, conserved hypothetical; pale green, unknown function; brown, pseudogenes.

4.3.1.4 Typhimurium

Approximately 12 million sequence reads were generated from the Typhimurium sequencing run which used two lanes on the Illumina flowcell. Almost 90% of the reads contained a 100% identical match to the transposon tag and 49% of these could be mapped to the SL1344 reference sequence (N.B. 35% mapped to three plasmids present in this strain and a further 16% contained Illumina adapter sequence). Combining the two sequencing lanes meant that, in total, 549,086 unique insertion sites were recovered from the ~1 million mutant library, an average of one insertion every 9bp, or over 100 unique inserts per gene (Figure 4-2B). There is an apparent bias in the frequency of transposon insertion towards the origin of replication. This likely occurred as the bacteria were in exponential growth phase immediately prior to transformation with the transposon. In this phase of growth, multiple replication forks would have been initiated, meaning genes closer to the origin were in greater copy number and hence more likely to be a target for insertion. This is not evident in the Typhi library, but we believe this is disguised partly by a few over-represented mutants (high peaks in Figure 4-2A) and partly because Typhi has a slower growth rate than Typhimurium.

4.3.2 Essential genes in Typhi

Three hundred and fifty-six genes had an LLR of less than −2; thus these genes were considered essential for growth under standard laboratory conditions, based on the criteria given in the methods. Conversely, 4162 genes had an LLR greater than 2 and were thus non-essential. Nineteen genes had LLRs between the two cut-off values and so it was not possible to assign these as essential or non-essential with the same degree of confidence. In addition, the density of insertions across the genome was such that a 60bp region without insertion had only a 1% chance of occurring randomly. We could not therefore make conclusions with confidence for very short genes (less than 60bp) with no insertions which may have been missed by chance. However, there were only 2 annotated genes less than 60bp long that had no mapped insertion sites. Thus, we effectively assayed all but 2 very short annotated genes, and were able to draw conclusions with statistical confidence for 4518 of 4537 (99.6%) annotated genes in the genome (Figure 4-3, Appendix 8.3.2).

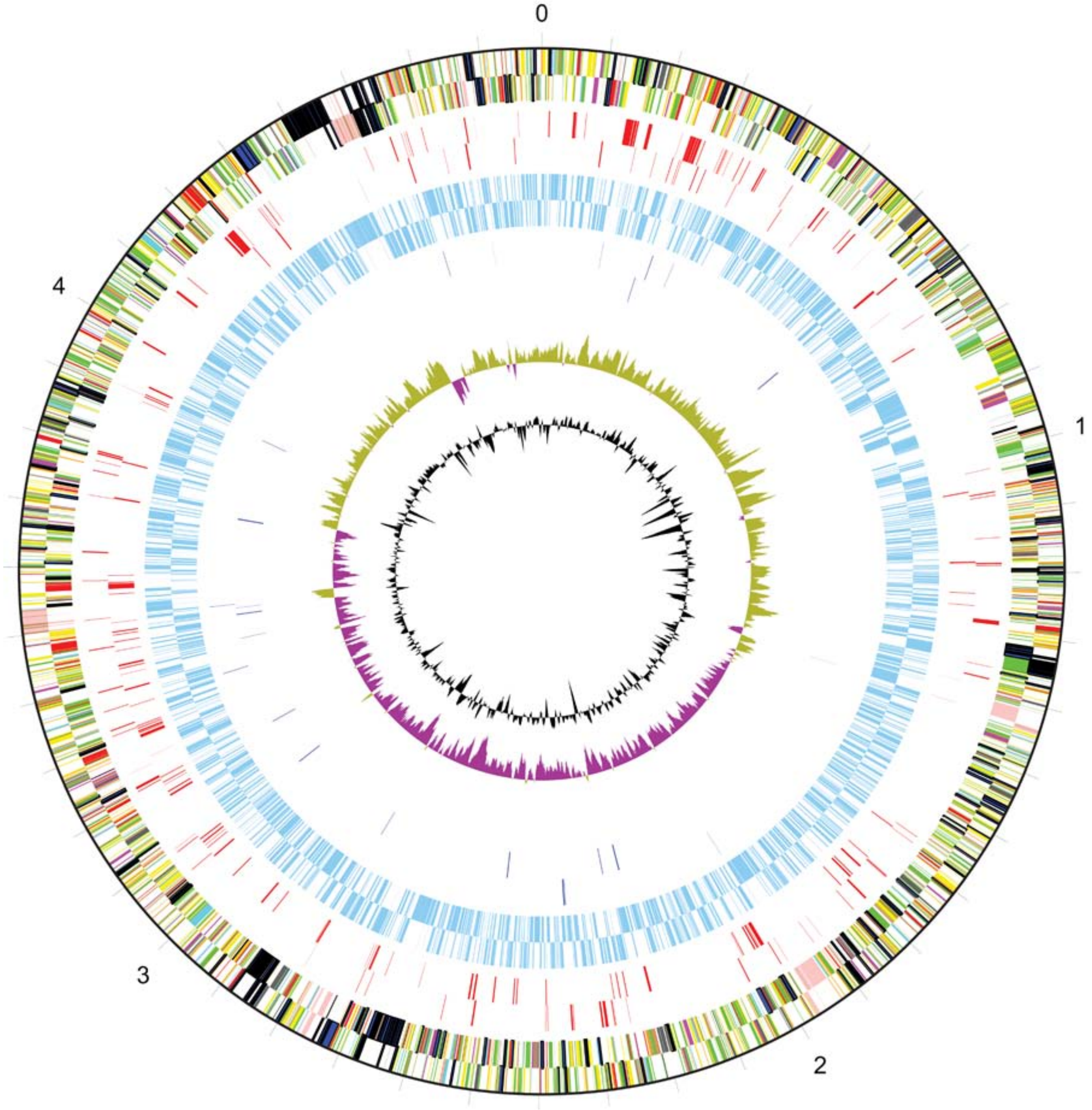

Figure 4-3 Whole genome assay of Typhi

The outer scale is marked in megabases. Circles range from 1 (outer circle) to 6 (inner circle) and represent genes on both forward and reverse strands. Circle 1, all genes; circle 2, Typhi essential genes (red); circle 3, Typhi non-essential genes (light blue); circle 4, 26 genes essential in Typhi only (dark blue); circle 5, GC bias ((G-C)/(G+C)), khaki indicates values >1; purple <1; circle 6 %G+C content. Genes in outer circle are colour-coded according to function: dark blue, pathogenicity/adaptation; black, energy metabolism; red, information transfer; dark green, membranes/surface structures; cyan, degradation of macromolecules; purple, degradation of small molecules; yellow, central/intermediary metabolism; light blue, regulators; pink, phage/IS elements; orange, conserved hypothetical; pale green, unknown function; brown, pseudogenes.

Many of the 356 essential genes are required for fundamental biological processes, including cell division, DNA replication, transcription and translation (Table 4-3). The full list is available in Appendix 8.3.3. A few are worthy of note, including DNA polymerase III, a multimeric enzyme encoded by eight subunit genes, six of which are identified as essential. The remaining two genes are *holE*, (LLR = 4.83), and *holC* (LLR = 5.36) which are unlikely to be essential. All the aminoacyl-tRNA synthetase genes were identified as candidate essential genes except for *trpS* (t4024) and *trpS* (t4557) which are both tryptophanyl-tRNA synthetases and therefore mutually redundant. Similarly, of the 11 genes that are involved in peptidoglycan biosynthesis, 9 were assigned as essential, while *ddlA* and *ddlB* were assigned as non-essential; both these genes perform the same function (Figure 4-4). Of the 356 Typhi candidate essential genes identified by TraDIS, 256 (~70%) are also essential in *Escherichia coli* (Baba et al. 2006), including 110 of the genes in Table 4-3. Of the 100 genes essential in Typhi but not *E. coli*, almost half are involved in energy metabolism or regulation of gene expression.

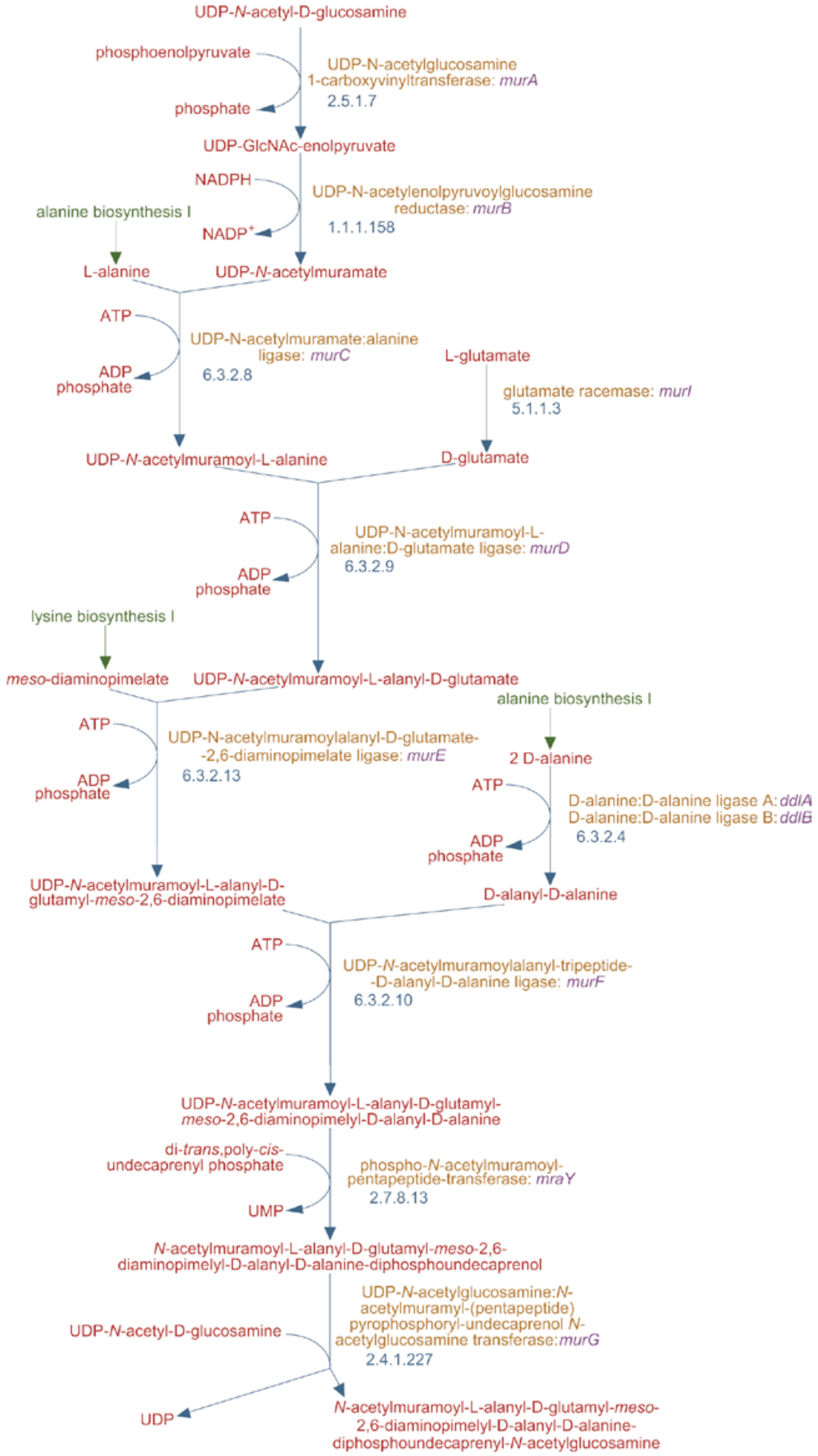
UDP-N-acetyl-D-glucosamine
phosphoenolpyruvate
phosphate
UDP-N-acetylglucosamine 1-carboxyvinyltransferase: murA
2.5.1.7
UDP-GlcNAc-enolpyruvate
NADPH
NADP+
alanine biosynthesis I
UDP-N-acetylenolpyruvoylglucosamine reductase: murB
1.1.1.158
L-alanine
UDP-N-acetylmuramate
ATP
ADP
phosphate
UDP-N-acetylmuramate:alanine ligase: murC
6.3.2.8
L-glutamate
glutamate racemase: murI
5.1.1.3
UDP-N-acetylmuramoyl-L-alanine
D-glutamate
ATP
ADP
phosphate
UDP-N-acetylmuramoyl-L-alanine:D-glutamate ligase: murD
6.3.2.9
lysine biosynthesis I
meso-diaminopimelate
UDP-N-acetylmuramoyl-L-alanyl-D-glutamate
alanine biosynthesis I
ATP
ADP
phosphate
UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase: murE
6.3.2.13
2 D-alanine
ATP
ADP
phosphate
D-alanine:D-alanine ligase A: ddlA
D-alanine:D-alanine ligase B: ddlB
6.3.2.4
UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminopimelate
D-alanyl-D-alanine
ATP
ADP
phosphate
UDP-N-acetylmuramoylalanyl-tripeptide--D-alanyl-D-alanine ligase: murF
6.3.2.10
UDP-N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminopimelyl-D-alanyl-D-alanine
di-trans,poly-cis-undecaprenyl phosphate
UMP
phospho-N-acetylmuramoyl-pentapeptide-transferase: mraY
2.7.8.13
N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminopimelyl-D-alanyl-D-alanine-diphosphoundecaprenol
UDP-N-acetyl-D-glucosamine
UDP-N-acetylglucosamine:N-acetylmuramyl-(pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase: murG
2.4.1.227
UDP
N-acetylmuramoyl-L-alanyl-D-glutamyl-meso-2,6-diaminopimelyl-D-alanyl-D-alanine-diphosphoundecaprenyl-N-acetylglucosamine

Figure 4-4 Peptidoglycan biosynthesis

All genes are essential in this pathway, except for *ddlA* and *ddlB*, which perform the same enzymatic function. Reaction direction indicated by arrows, blue numbers represent enzyme commission (E.C.) numbers, green words and arrows represent incoming substrates from other metabolic pathways.

Table 4-3 Known genes coding for fundamental biological processes in Typhi

Biological process	**Sub-process (total number of genes)**		**Essential genes**	**Non-essential genes**
Cell division		20	***ftsAH**J**LQWXYZ***, ***mukB***, **t0429**	***ftsNK***, *minC**DE***, *sdiA, cedA, sulA,* t3932
DNA replication	DNA Polymerase I	1	*polA*	
	DNA Polymerase II	1		*polB*
	DNA Polymerase III	8	***dnaEN**Q**X***, ***holAB**D*	*holC*
	Supercoiling	4	***gyrAB***, ***par**CE*	
	Primosome-associated	10	***dnaBC**GT*, *priAB, rep,* ***ssb*(t4161)**	*priC, ssb*(t4237)
Transcription	RNA polymerase	3	***rpoABC***	
	Sigma, elongation, anti- and termination factors	9	***rpoDE**H*, ***nusA**B**G***, *rho*	*rpoNS*
Translation	tRNA-synthetases	23	***glyQ**S*, ***hisS*** , *lysS*, ***metG***, ***pheST***, ***proS***, ***serS***, ***thrS***, ***tyrS***, ***aspS***, ***asnS***, *alaS*, ***valS***, ***leuS***, *ileS*, ***gltX***, ***glnS***, ***cysS***, ***argS***	***trpS*(t4024)**, *trpS*(t4557)
	Ribosome components	56	***rplBCDEFJ**K**LMNOPQRSTUVWX**Y*, ***rpmABCDH**IJ*(t4086), *r**psABCDE**F**GHIJKLMN**O**PQRS**U*	*rplAI*, *rpmE*(t3522), *rpmE*(t2391), *rpmFGJ*(t2390), *rpsT*
	Initiation, elongation and peptide chain release factors	13	***fusA, infABC, prfA**B,* ***tsf***	*efp, prfCH, selB, tufAB*

Gene names in bold are also essential in *E. coli* (Baba et al. 2006).

The high density of insertions across the Typhi genome allows a clear demarcation between many candidate essential and non-essential genes. As an example, topoisomerase IV, an essential enzyme for maintaining DNA supercoiling, is encoded by *parC* and *parE* and almost no insertion sites were identified for these genes, or for *plsC*, a lipid biosynthesis gene (Figure 4-5).

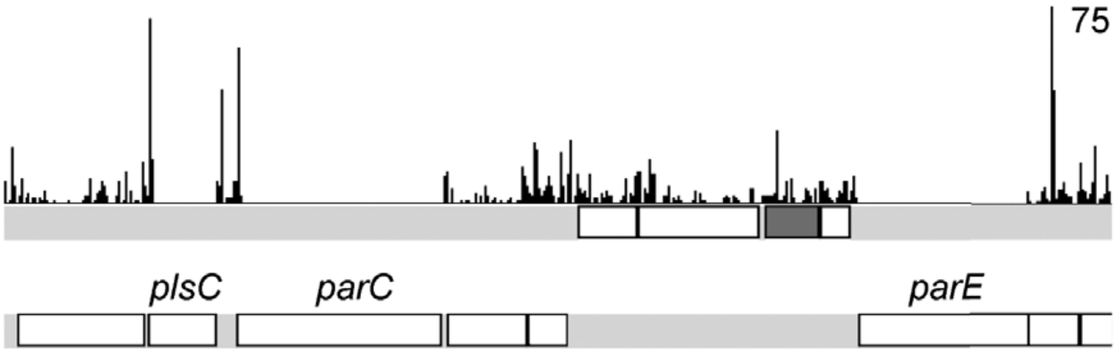

Figure 4-5 Essential genes in Typhi

Detailed plot generated using Artemis (Rutherford et al. 2000). The essential *plsC* gene and topoisomerase IV genes, *parC* and *parE*, showing the absence of transposon insertions. The maximum number of sequence reads within this plot is 75; white boxes represent genes, and grey boxes pseudogenes.

Inportantly, the genome coverage of the Typhi million mutant library is so great that insertions into small intergenic regions between essential genes such as *pyrH*, *frr* and *dxr* can also be seen clearly (Figure 4-6). This demonstrates that the insertion of this transposon is unlikely to have polar effects within operons. Elsewhere, the intergenic region between essential genes *leuS* and *rlpB* is only 14 bp but we observed 6 sequence reads mapping to 1 insertion site here without any insertions into the adjacent coding sequence.

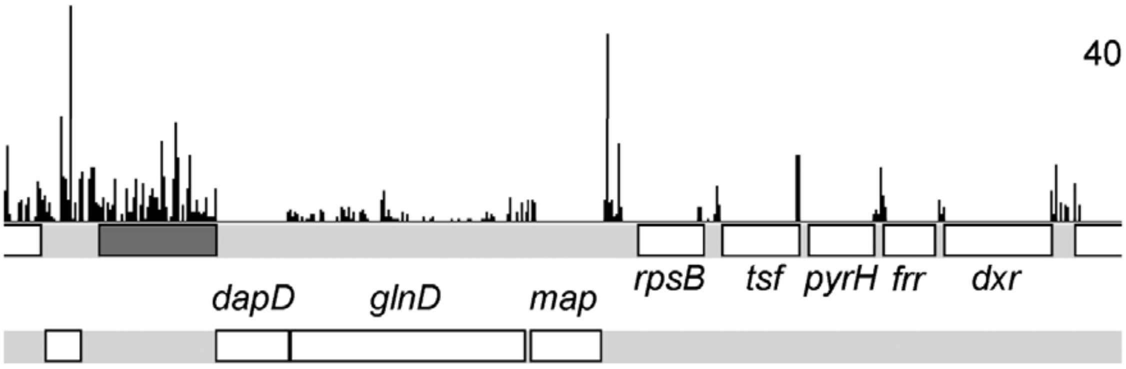

Figure 4-6 Insertions between genes

Detailed plot generated using Artemis (Rutherford et al. 2000). Sequence reads mapping to regions between essential genes. The maximum number of sequence reads within this plot is 40; white boxes represent genes, and grey boxes pseudogenes.

4.3.3 Essential genes in Typhimurium

The TraDIS analysis of the Typhimurium mutant library allowed us to identify 318 essential genes, and 4,135 non essential genes (Appendix 8.3.4 and 8.3.5). We were unable to assign 39 genes which had LLRs between -2 and 2, but all other genes contained enough insertions or were of sufficient length (>41bp) to generate credible LLR scores. Thus, every gene was assayed and we were able to draw conclusions with statistical confidence for 99.1% (4453/4492 genes) of the coding genome in a single sequencing run (Figure 4-7).

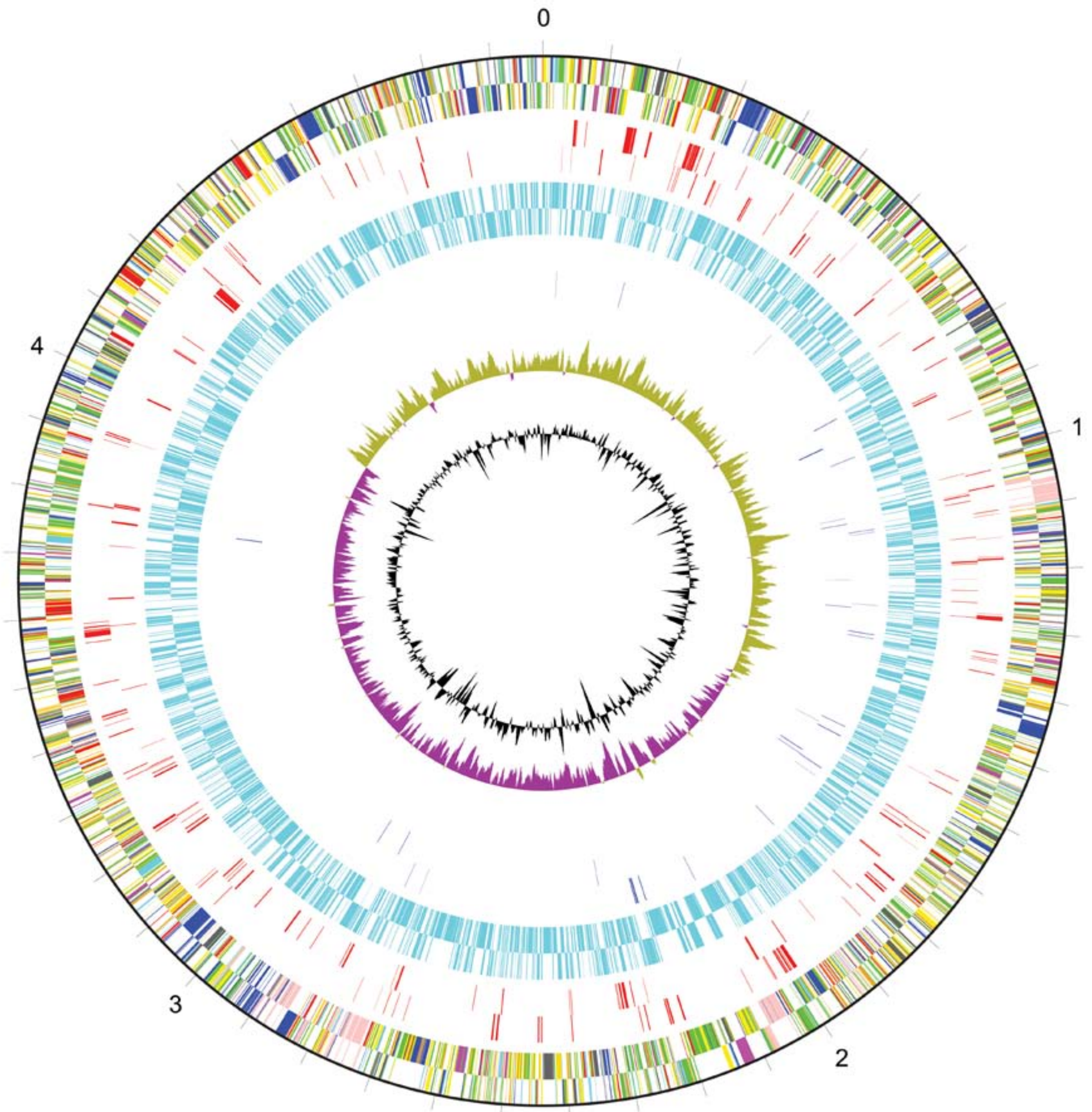

Figure 4-7 Whole genome assay of Typhimurium

The outer scale is marked in megabases. Circles range from 1 (outer circle) to 6 (inner circle) and represent genes on both forward and reverse strands. Circle 1, all genes; circle 2, Typhimurium essential genes (red); circle 3, Typhimurium non-essential genes (light blue) circle 4, 37 genes essential in Typhimurium only (dark blue); circle 5, GC bias ((G-C)/(G+C)), khaki indicates values >1; purple <1; circle 6 %G+C content. Genes in outer circle are colour-coded according to function: dark blue, pathogenicity/adaptation; black, energy metabolism; red, information transfer; dark green, membranes/surface structures; cyan, degradation of macromolecules; purple, degradation of small molecules; yellow, central/intermediary metabolism; light blue, regulators; pink, phage/IS elements; orange, conserved hypothetical; pale green, unknown function; brown, pseudogenes.

In some cases (for longer genes) we were also able to assay below the gene level to identify parts of genes that were essential. One example is *polA* where transposon insertions occurred across the majority of the 3' part of the gene but the 5' end contained no insertions (Figure 4-8). This 'essential' region corresponds to two protein domains (PF02739 and PF01367) which are involved in 5' to 3' exonuclease activity and have a role in DNA binding. Another example is *ams* (*rne*) which again only contained insertions through the 3' part of the gene. A Pfam search (Finn et al. 2007) based on the amino acid sequence of this gene predicts the presence of two domains in the preserved 5' end that are responsible for RNA-binding and cleavage. Our data suggest that it is the activity of these domains and not the whole gene that is essential for growth under laboratory conditions. It should be noted that genes with essential domains at the 3' end are unlikely to contain transposon insertions along the whole length of the gene, as insertions prior to the essential domain would interrupt translation of the required region.

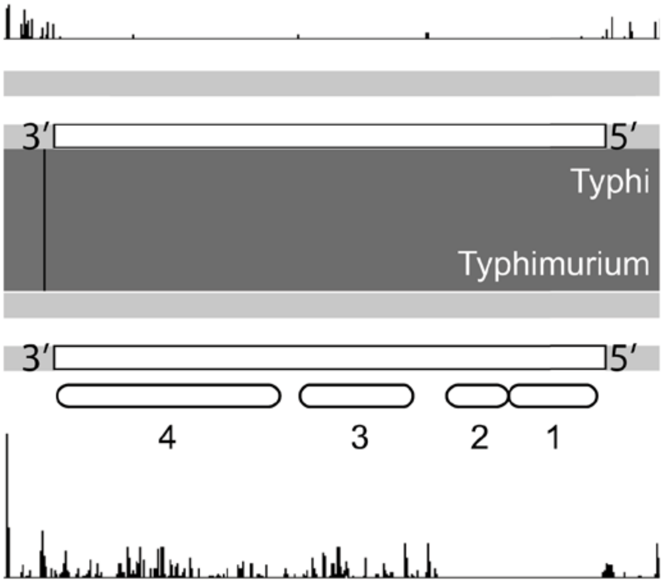

Figure 4-8 Domains in *polA* are essential

Image generated using Artemis Comparison Tool (ACT (Carver et al. 2005)) for mapped insertion sites across *polA*, scaled to a height of 50 with a window size of 3. Dark grey indicates sequence similarity. Direction of transcription is given and Pfam domains are shown in rounded boxes marked 1-4: 1, PF02739; 2, PF01367; 3, PF01612; 4, PF00476. Domains 1 and 2 do not tolerate insertion in either Typhi or Typhimurium.

4.3.4 Comparing Typhi and Typhimurium

4.3.4.1 Shared essential genes

The essential gene set of Typhi (Langridge et al. 2009b), was compared with the list generated from Typhimurium. We considered that the two serovars shared an essential gene if the gene had an LLR of < -2 in both organisms. Using this criterion, a total of 267 essential genes were shared, which represents 75% of the Typhi set (Figure 4-9A, Appendix 8.3.6). These sets were also compared with one generated in *E. coli* by systematic gene knockouts (Baba et al. 2006), which revealed a total of 226 essential genes shared between all three, validating TraDIS as an approach for the identification of candidate essential genes (Figure 4-9B). As expected, the *Salmonella* serovars share slightly more essential genes with each other (267) than either does with *E. coli* (~258).

The majority of shared essential genes between all three bacteria are responsible for fundamental cell processes, including cell division, transcription and translation. A number of key metabolic pathways are also represented, such as fatty acid and peptidoglycan biosynthesis (Appendix 8.3.7). Interestingly, 16 genes annotated as conserved hypothetical are essential in both *Salmonella* serovars (11 shared with *E. coli*), indicating the presence of important genes whose functions in cell survival have yet to be elucidated.

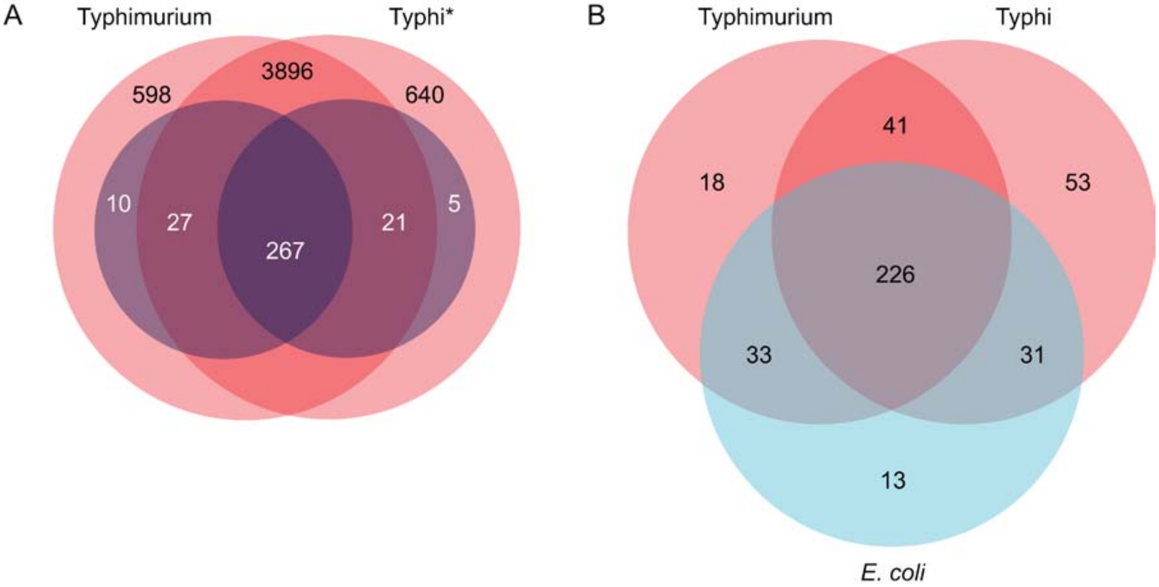

Figure 4-9 Comparison of essential genes

Venn diagrams showing (A) the overlap of all genes (red) and essential genes (blue) between Typhimurium and Typhi. Black numbers refer to all genes, white numbers to essential genes. *, details of the five Typhi essential genes omitted from the comparison are given in Appendix 8.3.8. (B) the overlap of essential genes between Typhimurium, Typhi and *E. coli*.

4.3.4.2 Essential genes only present in one serovar

Ten genes essential in Typhimurium were absent from Typhi (Table 4-4). Five of these were encoded on phage, of which three are repressors. One of the remaining phage essential genes encodes a PhoP/PhoQ regulated protein and the other is involved in natural bacterial transformation. These warrant further investigation as they are genes that have been acquired and then become essential for survival in rich media. The non-phage related essential genes included one encoding the antitoxin element of a chromosomally encoded toxin/antitoxin system. Others encoded a lipoprotein, a cation transporter and an electron transfer flavoprotein and are likely to have been lost from the Typhi genome, since they are found in many other *Salmonella* serovars.

Table 4-4 Genes uniquely essential in Typhimurium

	Ty inserts	Ty reads	SL inserts	SL reads	SL ID	SL gene length	Ty ID	Ty gene length	Name	Function
No orthologue in Typhi	-	-	18	123	SL0742	1250	-	-	-	putative cation transporter
	-	-	4	21	SL0831	836	-	-	-	putative electron transfer flavoprotein (beta subunit)
	-	-	0	0	SL0950	323	-	-	-	putative prophage protein
	-	-	11	75	SL1179	770	-	-	envF	lipoprotein
	-	-	2	4	SL1480	230	-	-	-	putative cytoplasmic protein
	-	-	1	3	SL1560	698	-	-	-	putative membrane protein
	-	-	3	27	SL1967	677	-	-	-	putative prophage protein
	-	-	3	34	SL2549	209	-	-	-	endodeoxyribonuclease
	-	-	10	146	SL2633	977	-	-	-	putative repressor protein
	-	-	0	0	SL2695	959	-	-	-	putative competence protein
Present in Typhi but essential only in Typhimurium†	22	156	16	174	SL1561	1208	t1534‡	122	sseJ	*Salmonella* translocated effector protein (SseJ)
	-	-	4	149	SL2593	449	STY2066*	-	-	putative DNA-binding protein
	33	463	5	26	SL0032	422	t0033	287	-	putative transcriptional regulator
	68	325	9	36	SL0623	623	t2232	557	lipB	lipoate-protein ligase B
	147	3451	10	64	SL0702	878	t2156	875	-	putative glycosyl transferase
	188	2959	9	61	SL0703	1115	t2155	1115	-	galactosyltransferase
	230	3478	15	67	SL0706	1760	t2152	1761	-	putative glycosyltransferase. From NCBI STM gene
	84	1041	2	4	SL0707	815	t2151	815	-	conserved hypothetical protein
	46	361	13	69	SL0722	1550	t2136	1550	cydA	cytochrome d ubiquinol oxidase subunit I
	73	1604	5	22	SL1069	674	t1789	674	-	putative secreted protein
	17	182	1	1	SL1203	131	t1146	137	-	hypothetical protein
	18	286	1	5	SL1264	296	t1209	296	-	putative membrane protein
	35	305	2	5	SL1341	209	t1275	209	ssaH	putative pathogenicity island protein
	44	387	1	3	SL1342	230	t1276	230	ssaI	putative pathogenicity island protein
	142	3178	5	14	SL1343	731	t1277	731	ssaJ	putative pathogenicity island lipoprotein
	70	747	4	44	SL1355	761	t1289	761	ssaT	putative type III secretion protein
	81	708	6	35	SL1532	932	t1511	932	sifB	putative virulence effector protein
	118	1635	10	44	SL1563	743	t1536	743	-	putative periplasmic amino acid-binding protein
	107	2440	5	44	SL1564	629	t1537	629	-	putative ABC amino acid transporter permease
	181	1562	19	92	SL1628	1355	t1612	1364	-	hypothetical protein
	23	177	1	5	SL1659	164	t1640	164	-	conserved hypothetical protein
	35	269	3	9	SL1785	377	t1022	377	-	conserved hypothetical protein
	164	2808	9	27	SL1793	896	t1016	896	pagO	inner membrane protein
	23	155	1	4	SL1823	953	t0988	953	msbB	lipid A acyltransferase
	55	338	10	57	SL2064	983	t0786	983	rfbV	putative glycosyl transferase
	82	483	6	58	SL2065	1274	t0785	1280	rfbX	putative O-antigen transporter
	40	195	4	11	SL3828	1811	t3658	1811	glmS	glucosamine-fructose-6-phosphate aminotransferase

SL, Typhimurium; Ty, Typhi; †*P*-value (associated with log2 read ratio) < 0.05; ‡ *sseJ* is a pseudogene in Typhi; *This gene is not present in Typhi Ty2 but is in CT18 so the STY identifier is given.

Five essential genes were only present in Typhi (Table 4-5) of which four were phage-related, including two phage repressors. The other two phage-related genes encode a glycosyl transferase and a putative DNA repair protein. One essential Typhimurium glycosyl transferase is orthologous to a pseudogene in Typhi, suggesting that the Typhi phage glycosyl transferase is acting as a functional replacement. Typhi also contains a pseudogene for *priC*, whose gene product normally interacts with RecA. The essential phage DNA repair protein is predicted to interact with RecA also, again suggesting some overlap of function. The remaining essential gene present only in Typhi is predicted to encode a secreted protein and is of interest as genomic comparisons with other *Salmonella* serovars indicate that only Paratyphi A, another human-restricted serovar, contains this gene.

Table 4-5 Genes uniquely essential in Typhi

	SL inserts	SL reads	Ty inserts	Ty reads	Ty ID	Ty gene length	SL ID	SL gene length	Name	Function
No orthologue in Tm	-	-	0	0	t1378	212	-	-	-	hypothetical protein
	-	-	1	2	t1920	386	-	-	-	putative DNA-binding protein
	-	-	3	70	t3402	551	-	-	cI	repressor protein
	-	-	3	45	t3415	722	-	-	-	hypothetical protein
	-	-	1	6	t4531	131	-	-	-	hypothetical secreted protein
Present in Typhimurium but essential only in Typhi†	43	493	3	22	t0123	440	SL0119	440	yabB	conserved hypothetical protein
	117	571	11	32	t0203	1262	SL0203	1262	hemL	glutamate-1-semialdehyde 2,1-aminomutase
	122	965	1	1	t0224	1334	SL0224	1334	yaeL	putative membrane protein
	65	446	1	12	t0270	557	SL2604	557	rpoE	RNA polymerase sigma-E factor
	139	757	0	0	t0587	2267	SL2246	2267	nrdA	ribonucleoside-diphosphate reductase 1 alpha chain
	113	641	14	38	t2140	2783	SL0718	2783	sucA	2-oxoglutarate dehydrogenase E1 component
	112	711	10	16	t2177	1622	SL0680	1622	pgm	phosphoglucomutase
	75	694	10	36	t2274	938	SL0582	938	fepB	ferrienterobactin-binding periplasmic protein precursor
	80	542	8	13	t2276	989	SL0580	989	fepD	ferric enterobactin transport protein FepD
	93	591	2	2	t2277	971	SL0579	971	fepG	ferric enterobactin transport protein FepG
	64	508	4	4	t2278	776	SL0578	776	fepC	ferric enterobactin transport ATP-binding protein FepC
	198	1116	12	116	t2410	2336	SL0444	2336	lon	Lon protease
	94	504	7	16	t2730	1043	SL2809	1043	recA	recA protein
	131	699	13	29	t2996	1973	SL3052	1928	tktA	transketolase
	76	358	3	9	t3120	1415	SL3173	1415	rfaE	ADP-heptose synthase
	211	1928	6	50	t3265	1052	SL3321	1052	degS	serine protease
	41	405	3	10	t3326	587	SL3925	587	yigP	conserved hypothetical protein
	121	557	16	34	t3384	2006	SL3872	2006	rep	ATP-dependent DNA helicase
	172	1194	5	18	t3621	2768	SL3947	2768	polA	DNA polymerase I
	116	775	9	13	t3808	1028	SL3677	1028	waaF	ADP-heptose-LPS heptosyltransferase II
	138	1082	8	29	t4411	932	SL4294	932	miaA	tRNA delta-2-isopentenylpyrophosphate transferase

SL/ Tm, Typhimurium; Ty, Typhi; †*P*-value (associated with $\log_2$ read ratio) < 0.05.

Forty one essential genes in Typhimurium were present as orthologues in both serovars; this number was 79 for Typhi. The cutoff for essentiality was placed at an LLR < -2, meaning there was the possibility that a gene with an LLR of < -2 in one serovar had an LLR just above this threshold in the other. Hence, we also calculated $\log_2$ read ratios for each essential gene to establish whether the number of mapped reads (frequency of transposon insertions) per gene was significantly different between serovars. The number of mapped reads acts as a proxy for the frequency of transposon insertion in a gene, so the $\log_2$ read ratio indicated whether a gene had greater or fewer insertions in Typhimurium or Typhi.

Using these ratios, we found genes that appeared essential in one serovar (i.e. LLR < -2) and as unassigned/non-essential in the other (LLR > -2) but did not have a significantly different frequencies of transposon insertion, according to our cutoff ($P < 0.05$). We termed these genes "putative" essential genes, of which there are 14 and 58 for Typhimurium and Typhi respectively (Appendix 8.3.9).

4.3.4.3 Genes essential in Typhi only

Twenty-one essential Typhi genes had a significantly lower frequency of transposon insertion compared to orthologues in Typhimurium ($P < 0.05$), including two encoding conserved hypothetical proteins. The *fepBDGC* operon (Figure 4-10) was essential only in Typhi, indicating an apparent requirement for ferric (Fe(III)) rather than ferrous (Fe(II)) iron. These genes function to recover ferric enterobactin from the periplasm by acting with two other proteins known to aid the passage of this siderophore through the outer membrane. FepA is the outer membrane receptor for ferric enterobactin and provides a gated pore which is activated in the presence of TonB. While neither *fepA* (LLR = 27.7) nor *tonB* (LLR = 7) were found to be essential using TraDIS, it is probable that when Typhi enters the bloodstream, these two genes are then required for uptake of ferric enterobactin.

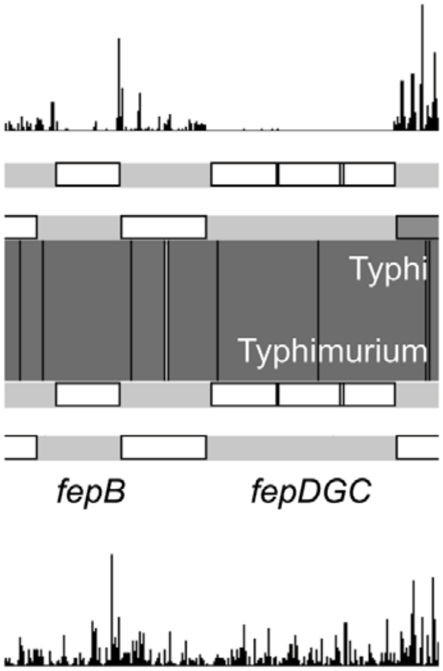

Figure 4-10 Genes uniquely essential in Typhi

ACT view of frequency and distribution of mapped insertion sites across *fepBDCG*. Dark grey blocks indicate sequence similarity. Scaled to a height of 50 with a window size of 3. White boxes indicate genes, grey boxes pseudogenes.

4.3.4.4 Genes essential in Typhimurium only

Twenty seven essential Typhimurium genes had a significantly lower frequency of transposon insertion compared to the equivalent genes in Typhi ($P < 0.05$), including five encoding hypothetical proteins (Table 4-4). This indicates that these gene products play a vital role in Typhimurium but not in Typhi when grown under laboratory conditions. One protein of note is encoded by SL0706, which is a pseudogene in Typhi (Ty2 unique ID: t2152) due to a 1bp deletion at codon 62 that causes a frameshift (Figure 4-11). SL0706 is predicted to encode a glycosyl transferase and in total, three glycosyl transferases (SL0702, *wbaV*) and one galactosyl transferase (SL0703) are essential in Typhimurium. However, only one enzyme of this type is essential in Typhi, suggesting that surface structure biogenesis is of greater importance in Typhimurium. This is possibly because Typhi expresses the cell-surface Vi antigen and so selection of surface structures may be less intense.

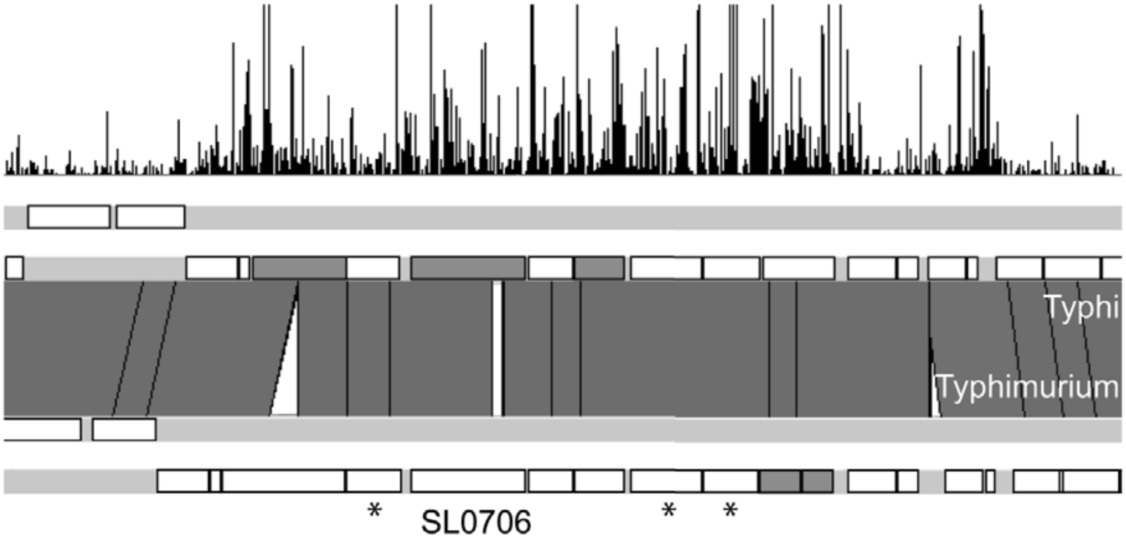

Figure 4-11 Gene uniquely essential in Typhimurium

ACT view of frequency and distribution of mapped insertion sites across the genomic region surrounding SL0706. Scaled to a height of 50 with a window size of 3. White boxes indicate genes, grey boxes pseudogenes. * These genes are also uniquely essential in Typhimurium.

We also identified four genes from SPI-2 that appear uniquely essential in Typhimurium under laboratory conditions (Figure 4-12). These genes (*ssaHIJT*) are thought to encode structural components of the SPI-2 type III secretion system apparatus (T3SS) (Kuhle and Hensel 2004). In addition, the effector genes *sseJ* and *sifB*, whose products are secreted through the SPI-2 type 3 secretion system (T3SS) (Freeman et al. 2003; Miao and Miller 2000), were also found to be uniquely essential in Typhimurium.

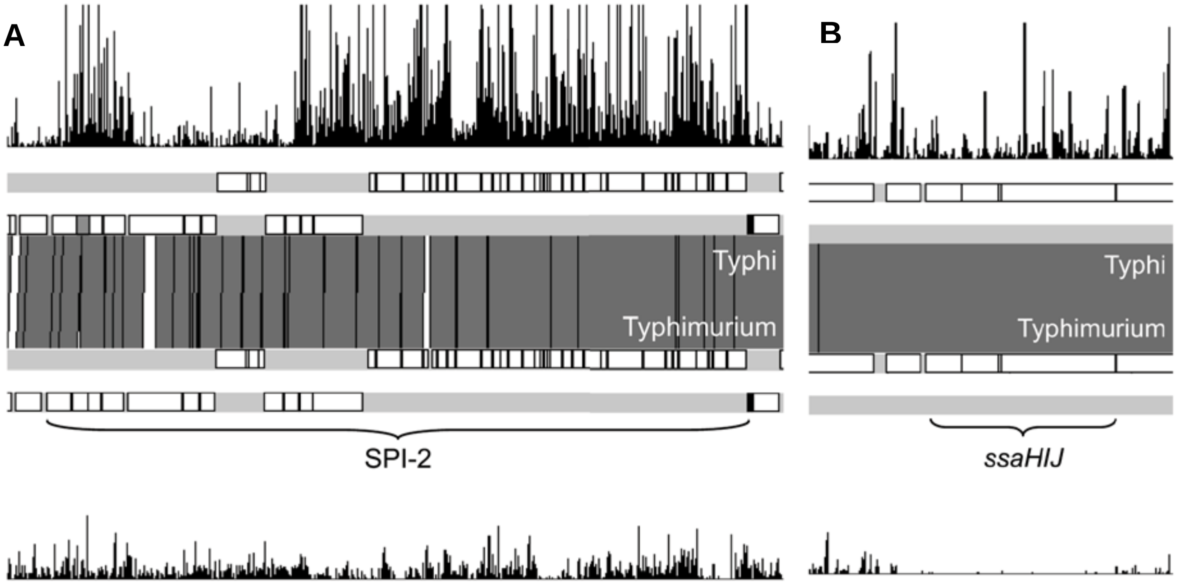

Figure 4-12 SPI-2 genes uniquely essential in Typhimurium

ACT view of frequency and distribution of mapped insertion sites across A) SPI-2 and B) *ssaHIJ*. Scaled to a height of 50 with a window size of 3. White boxes indicate genes, grey boxes pseudogenes.

All of these genes display high A+T nucleotide sequence and have been previously shown (in Typhimurium) to be strongly bound by the nucleoid associated protein HN-S, encoded by *hns* (Lucchini et al. 2006; Navarre et al. 2006). Therefore, rather than being essential, it is instead possible that access for the transposon was sufficiently restricted that very few insertions occurred at these sites. Indeed, the generation of null Typhimurium mutants in *sseJ* and *sifB*, as well as many others generated at the SPI-2 locus suggest that these genes are not truly essential in this serovar (Freeman et al. 2003; Hensel et al. 1997b; Hensel et al. 1998; Ohlson et al. 2005). While this is a reminder that the interpretation of why a gene is essential needs to made with care, the effect of HN-S upon transposon insertion is not genome-wide. If this were the case, there would be an under-representation of transposon mutants in high A+T regions (known for HN-S binding), which is not what was observed. In total, only 15 candidate essential genes fall into the '*hns*-repressed' category described by Navarre and colleagues (Navarre et al. 2006), the remainder (almost 400) contained sufficient transposon insertions to conclude they were non-essential. In addition, we noted that all SPI-1 genes which encode another Type III secretion system and are of high A+T content were found to be non-essential.

4.3.4.5 Functional classification of essential genes

A functional breakdown of the shared and the serovar-specific essential genes is shown in Figure 4-13. The pattern of function essential in both serovars is heavily skewed towards the fundamental biological processes, described above, of energy metabolism, regulation and synthesis of membrane/surface structures. Nonetheless, the distribution of function required individually by Typhi and Typhimurium, while similar in some respects, also reveals some stark differences. Proportionally, regulatory genes and again those for synthesis of membrane/surface structures are equally represented. However, Typhi requires relatively more energy metabolism genes, and in fact has representatives in every functional category. This may partly be a reflection of the loss of overlapping functions in the Typhi genome due to the presence of over 200 pseudogenes. For example, if two genes originally shared an essential function and one of them became a pseudogene, the other would then become essential. Conversely, almost 20% of the genes essential only in Typhimurium are of ‘unknown’ function indicating the presence of extremely important genes whose precise role in cellular viability remains to be elucidated.

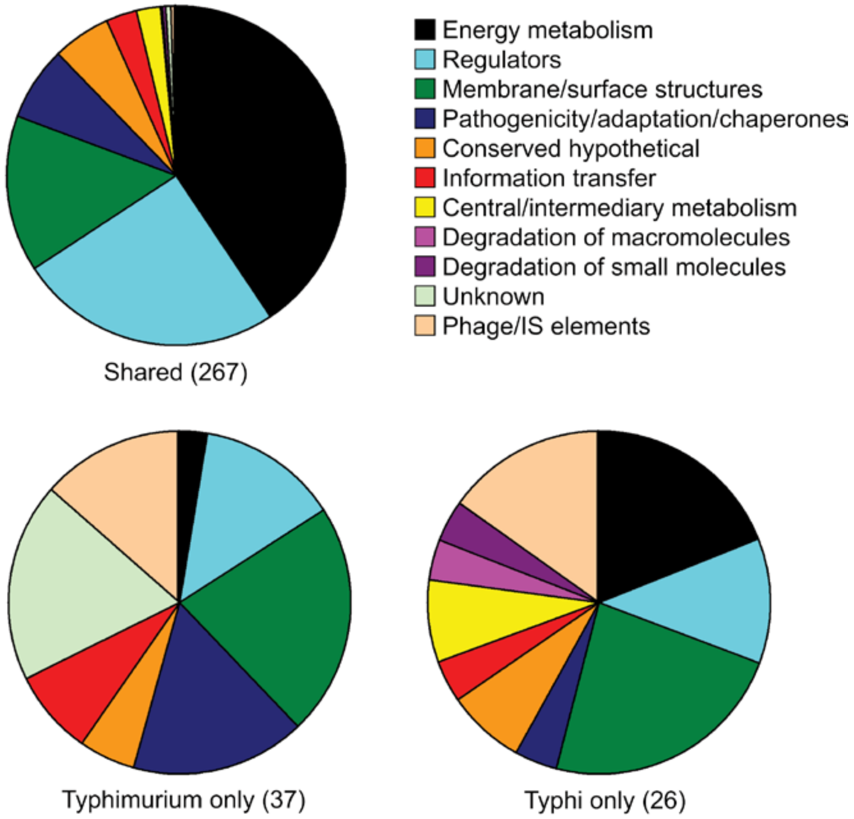

Figure 4-13 Essential genes classified by function

Functional classification for shared (267), Typhimurium only (37) and Typhi only (26) essential genes.

4.4 Discussion

4.4.1 Improvement over microarrays

A variety of previous methods has identified a number of essential and niche-specific genes, but to do this effectively on a genome-wide scale has required the use of microarrays to indirectly assay the sites of transposon insertion. Microarrays have their drawbacks: resolution is limited, and distinguishing a positive from a negative signal for some microarray features can be difficult. With sequencing, the signal is of a “digital” nature; any sequence read that has the 10bp transposon tag with adjacent genomic sequence is almost certainly an indication of the exact position of a transposon insertion site.

The combination of extremely large transposon mutant pools and high-throughput Illumina sequencing from the transposon insertion sites has brought an unparalleled degree of resolution to a transposon mutagenesis screen. Indeed, the number of insertions in the Typhimurium library was sufficiently great that gaps between insertion sites of 27bp had a less than 5% probability of occurring by chance, indicating the resolution available from this approach. This has allowed us to distinguish between essential and non-essential genomic regions to within a few base pairs and to confidently assign over 99% of the genes in both *Salmonella* genomes as essential or non-essential. In addition, there are sufficient insertions to allow the assay of nearly every gene in the genome for a particular growth condition; only small genes, with few or no transposon insertions, cannot be assayed. Thus, TraDIS can be used for the accurate estimation of minimal gene sets, and as a very effective negative selection method.

4.4.2 Assaying short regions

The ability to assay over 99% of the coding genome has implications beyond determining the minimal gene set. When a 60 bp region of the genome that does not contain insertions is statistically significant, it becomes possible to assay some functional domains encoded within genes for their contribution to cell survival. Here we have demonstrated an example of two protein domains in Typhimurium *polA* that produce no viable insertions, while insertions are found in two other domains.

The average level of transposon insertion in the Typhi and Typhimurium libraries of 1 every 10-20 bp also has implications for small RNAs. Initially, a non-coding 60 bp region without insertion is significant, both statistically and because it may identify an essential small RNA. In addition, given that there is a set of known small RNAs in *Salmonella* (Perkins et al. 2009), when these transposon libraries are used in biological screens, many of these will have been assayed for their response to that screen, which may contribute towards understanding of small RNA function. While this is not within the scope of this work, this kind of analysis is part of future efforts aimed at making full use of the transposon libraries as scientific resources.

4.4.3 Essential prophage genes

Many of the essential genes present in only one serovar encoded phage repressors. Repressors maintain the lysogenic state of the prophage, preventing transcription of early lytic genes (Echols and Green 1971). Transposon insertions into these genes will relieve this repression and trigger the lytic cycle, resulting in cell death, and consequently mutants are not represented in the sequenced library. This again questions the definition of ‘essential’ genes; such repressors may not be required for cellular viability in the traditional sense, but once present in these genomes, their maintenance is required for continued survival.

4.4.4 A+T-rich islands protected in Typhimurium

In *Salmonella*, high A+T content is a hallmark of horizontal acquisition, and Tn5 inserts preferentially into such DNA. However, there are sites in Typhimurium where horizontally acquired A+T rich regions showed no increase in insertion frequency. The SPI-2 pathogenicity island and the region surrounding SL0702-3 and SL0706-7 both have average A+T contents of ~53% (compared to a genome average of 48%), and showed an increased frequency of transposon insertion in Typhi. In contrast, the average insertion frequency for these sites was similar to that of the surrounding chromosome in Typhimurium. This indicates a potential Typhimurium-specific mechanism that partially protects some A+T-rich regions from frequent transposon insertion.

Assuming a single sequencing read per transposon mutant, the average frequency of insertion in a single gene residing in the 15kb regions surrounding SPI-2 was 250-350 in both serovars. For the 44 genes in SPI-2, the average was increased over 5-fold in Typhi to ~1900 per gene, as expected in an A+T-rich region. However, this average remained at ~250 per gene in Typhimurium. The nucleoid-associated protein HN-S has been implicated in binding A+T-rich DNA in Gram negative bacteria, and virulence loci in particular have been demonstrated to be repressed by HN-S in Typhimurium (Navarre et al. 2006). It is possible that the presence of HN-S affects the ability of the transposon to integrate into chromosomal DNA, and that *ssaHIJT* represents an extreme case as the number of insertion sites mapped to all other Typhimurium SPI-2 genes was sufficient to assign them as non-essential.

4.4.5 Specific genes required by Typhi

Our data indicate that in Typhi, *recA* is a candidate essential gene (LLR $=-11.5$). Mutants of *recA* exist in *E. coli*, suggesting that it is not an essential gene in this bacterium (Baba et al. 2006). However, in support of the TraDIS data, multiple attempts in our laboratory to generate a *recA* mutant in Typhi, using the suicide vector allelic-exchange method (Turner et al. 2006), have failed (Appendix 8.3.10). During bacterial growth, RecA is involved in DNA replication and the re-activation of stalled replication forks. This occurs via the ‘restart’ primosome, a multimeric enzyme complex made up of 7 proteins encoded by *dnaTBCG* and *priABC* (Sandler and Marians 2000). In *E. coli*, *priC* mutants have little phenotypic effect on growth (Sandler et al. 1999) and in Typhi *priC* is a pseudogene (Parkhill et al. 2001a). However, without *priC*, there is only a *priA*-dependent pathway for replication fork restart and our results suggest that in this background, a *recA* mutant is not viable.

The *fepBDGC* operon, responsible for Fe(III) uptake, is also essential only in Typhi. Fe(III) is present in the mammalian bloodstream, where Typhi can be found during systemic human infection. Both Typhi and Typhimurium encode four transport systems for the uptake of Fe(III), chelated to different siderophores; the Fep system uses a self-encoded siderophore, enterobactin, synthesised by the *entAF* operon (Earhart 1996;

Hantke et al. 2003; Zhou et al. 1999). Fe(III) is transported into the bacterial cell in the form of ferric enterobactin by the TonB-dependent Fep system. During host adaptation, Typhi has accumulated pseudogenes in other iron chelating systems, presumably because they are not necessary for survival in the niche Typhi occupies in the human host.

In contrast, Typhimurium causes intestinal rather than systemic infection, suggesting that a mechanism for obtaining Fe(III), the only form of iron present in the blood, is not a requirement. Instead, FeoAB is more advantageous for Typhimurium (transposon insertion frequency across the *feoAB* operon is much reduced compared to the flanking genes), which encodes a high affinity system for the uptake of Fe(II), a soluble form of iron present under anaerobic conditions such as those found in the intestine (Tsolis et al. 1996).

4.5 Conclusions

Essential genes have been recognised previously using methods screening up to 17 mutants per gene (Laia et al. 2009; Sassetti et al. 2001). The extremely high resolution of TraDIS has allowed us to assay every gene for essentiality in two very closely related salmonellae with different host ranges. High density transposon mutagenesis screens such as ours produce gene lists that must be interpreted in the context under which they were performed. We found, under laboratory conditions, that 48 genes present in both serovars were essential in only one, suggesting that identical gene products do not necessarily have the same phenotypic effects in the two different serovar backgrounds. Predicting the phenotype associated with a gene must therefore be serovar-specific.

The generation of two large-scale transposon mutant libraries has created a valuable biological resource. These libraries are suited to use in high-throughput functional studies and are currently being used in a number of biological screens, including antibiotic resistance, quorum sensing and serum killing. The ability to assay over 99% of each genome also lends itself well to the possibility of screening these libraries through eukaryotic cell infection, a vital stage in the infective process of both serovars.