4 Deep sequencing of the *S.* Typhi Ty2 transcriptome

4.1 Aims of the work described in this Chapter

To survey the *S.* Typhi transcriptome using the RNA-seq protocol developed in Chapter 3.

4.2 Introduction

DNA sequencing has been exploited to determine the whole genome sequence of hundreds of prokaryotic and eukaryotic species [76,198,199]. The availability of whole genome sequence has facilitated gene identification, transcriptomic studies and underpins experiments to link genotype to phenotype. To date genome-wide analysis of the transcriptome has relied, to a significant degree, on the use of DNA microarrays. However, recent advances in DNA sequencing technologies has facilitated the determination of nucleotide sequence with a genomic read depth several orders of magnitude greater than was previously possible. In this chapter, this technology is applied to characterise the transcriptome of *S.* Typhi by mapping sequence reads of cDNA prepared by reverse transcription of total cellular RNA depleted of ribosomal RNA.

Bacterial genomes are relatively small and have a high density of coding sequence in comparison to most eukaryotes. For example, the genome of *S.* Typhi is ~4.8 Mbp in length, with ~4,700 open reading frames currently defined in available annotation [76,136]. As outlined in detail in Chapter 1, *S.* Typhi is interesting in that, unlike most *Salmonella* serotypes that have a broad host range and cause localised gastroenteritis, this pathogen is highly host adapted and causes systemic typhoid fever only in higher primates (host-restricted). The genome of *S.* Typhi harbours novel features including horizontally acquired genetic islands specific to this serotype and ~220 pseudogenes,

genes that are potentially inactivated in this pathogen but intact in related species such as *S.* Typhimurium [77]. Of particular note, *S.* Typhi also expresses a polysaccharide, known as the Vi capsule, associated with increased virulence.

Illumina-based high throughput sequencing was exploited to characterise the transcriptome of *S.* Typhi and bioinformatics approaches were developed and used to identify key advantages of the approach over microarray analysis. Further, this analysis was performed in a strand-specific manner, allowing overlapping transcripts encoded on opposite strands to be readily identified. In addition to confirming previous transcriptional data, this method has been able to identify novel transcripts, including many potential non-coding small RNAs, putative *cis*-acting RNA elements and previously hypothetical genes. This study has also defined the transcription of pseudogenes under these conditions and annotated a *S.* Typhi Ty2 specific island which maps significant transcript data. By mapping transcripts to the whole genome this approach has identified expressed regions of *S.* Typhi prophages that encode putative cargo genes and potentially identified antisense expression of known coding sequences. This method has putatively mapped a significant amount of sequence data to the 5' untranslated regions of genes.

4.3 Results

4.3.1 Mapping Sequence Reads to the Annotated Ty2 Genome

The protocol devised in Chapter 3 was used in to perform an entire survey of the *S.* Typhi transcriptome. RNA was prepared from three replicates of *S.* Typhi Ty2

BRD948 grown to mid-log phase in LB broth (OD_{600}=0.6). This material was pooled, reverse transcribed and subjected to Illumina sequencing. The sequence reads were then mapped to the Ty2 reference genome (table 4.1). To achieve the transcript plot, each nucleotide of the genome was assigned a value derived by a pileup of ~36bp nucleotide sequence reads generated from the Typhi cDNA (figure 4.1). Sequence reads were then mapped to each strand of the *S.* Typhi Ty2 BRD948 genome and the sequence coverage per base plotted and visualised using Artemis software (figure 4.2). Importantly, no sequence data mapped to *aroC*, *aroD* or *htrA,* which all harbour large deletions in this attenuated strain (figure 4.3).

Table 4.1 Analysis of sequences mapped to the Ty2 genome

	Sequencing Data Table		
Flowcell/Lane	876/2	1104/2	1354/1
Strain	BRD948	BRD948	BRD948
rRNA Depletion	Oligo	Oligo	Oligo
Mass of Total RNA	300	100	100
Read Length	36	36	36
No Reads	56085 89	7183969	5848604
Total Mapped	5438270	6513814	5356994
Total Mapped (%)	96.9	90.7	91.5
Total Uniquely Mapped	1493759	2942477	2326287
Total Mapped Uniquely (%)	0.266334189	0.409589323	0.397750814
Reads Mapped to CDS	1235932	2212650	1937241
Reads Mapped to NC sequences	257827	729827	389046
Reads mapped to pseudogenes	12403	45771	36678
Reads mapped to hypothetical genes	131242	266139	264871
GC content (ALL)	0.519742418	0.41840995	0.453535147
GC content (UNIQUE)	0.50318224	0.44267108	0.471412408
Coding Sense	1124620	1732827	1565156
Antisense Coding	111461	480504	372642

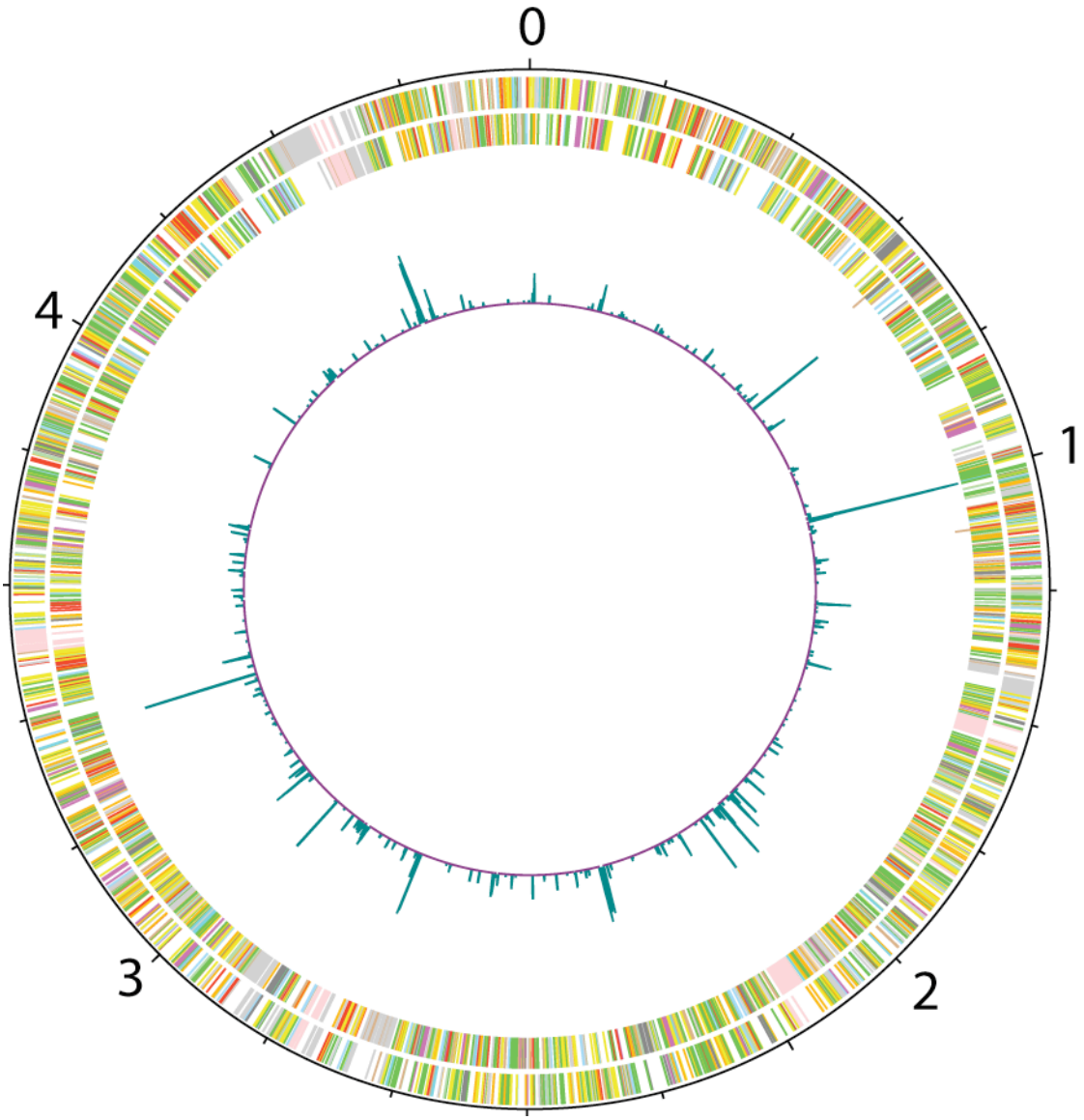

Figure 4.1 Circular plot of mapped sequence data.

Circles are described from outer to inner. The outermost represents base number in megabases, next outermost represents CDS annotated on the forward strand. The circle inside that represents CDS on the reverse strand and the innermost circle represents the plot of sequence data aligning to both strands. Each gene is coloured according to the original CT18 annotation [76] and represent different gene classes.

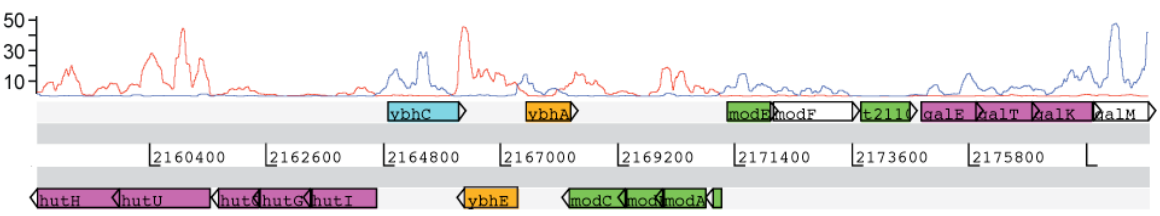

Figure 4.2 Artemis representation of transcriptome plot.

Blue plot represents forward strand and red, reverse strand. y-axis depth coverage of sequence reads.

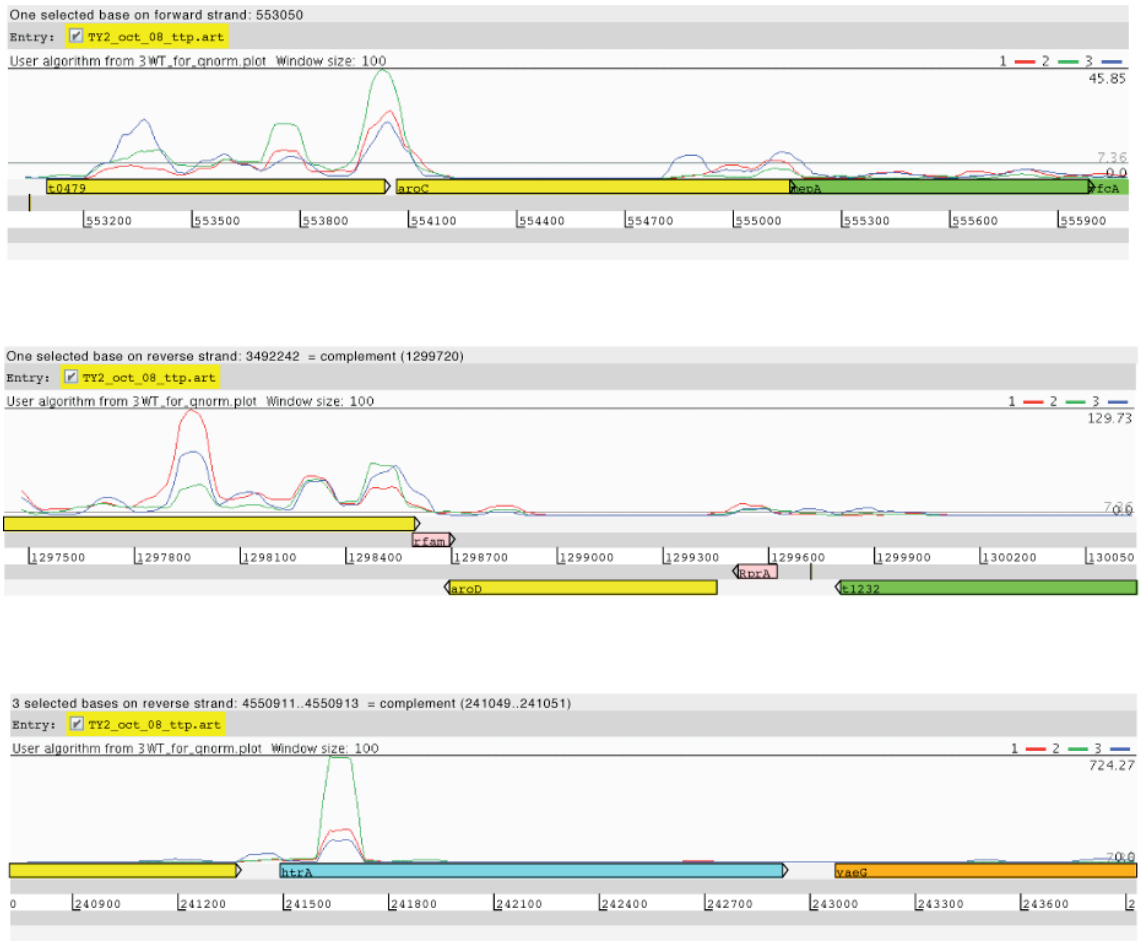

Figure 4.3 Defined deletions in BRD948.

Coding strand traces for three RNA-seq sequencing experiments with no mapping of sequenced cDNA to deleted loci. Coloured traces represent each sequencing run, red, first; green, second; blue, third.

4.3.2 Coding Sequences

4.3.2.1 General Features

The methodology allowed a strand specific alignment of sequence, which has not previously been performed using RNA-seq. To test the genome-wide strand specificity of the sequence information, the arithmetic mean (AM) for mapped

sequence reads was determined for the coding strand of the genes currently annotated on the *S*. Typhi Ty2 genome. This value was then plotted against the AM for the putative non-coding strand (figure 4.4). 91% of the reads mapped to previously annotated *S*. Typhi Ty2 coding strand providing supporting evidence for a successful deconvolution of the strands. Further detailed evaluation of the plots revealed regions with high numbers of mapped reads consistent with annotated coding sequences. Examples of such analysis are shown in figure 4.5(a and b). Although the sequence coverage varied across each coding sequence, indicated by peaks and troughs, the profile was remarkably consistent between experiments (figure 4.6). However, many intergenic regions and 34% of the annotated coding sequences had few (AM<1) or no mapped reads. For sequence data mapping to a region where CDS orientation is highly "mosaic" the plots align to the annotation (figure 4.7), further illustrating the strand-specific nature of the RNA-seq data. Sequence reads that mapped to non-coding strands may represent transcriptionally active but previously unannotated features of the genome. Indeed, these data enabled identification of putative errata in the annotation of a number of, mostly hypothetical genes (figure 4.4) such as the hypothetical locus t2145. This predicted CDS mapped significant sequence data to the opposing strand, which is proximal to the 5' region of the gene *gltA* (figure 4.8). These data are also consistent with mapping of sequenced transcripts upstream of known riboswitch encoding genes such as *btuB* [200,201] and *glmS* [202].

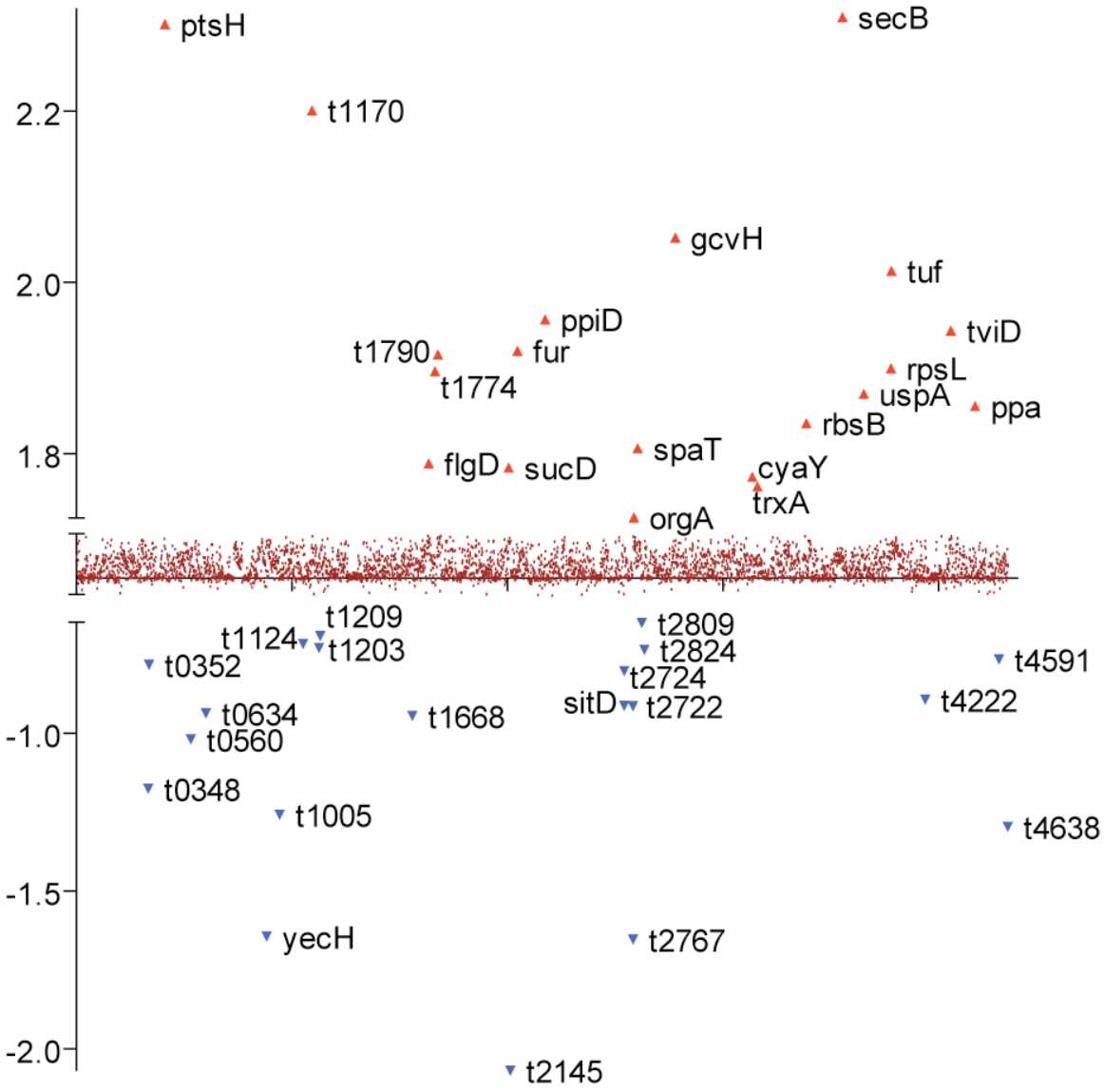

Figure 4.4 Sense and Antisense outliers.

The AM for both the coding strand and non-coding strand was determined for each annotated gene. This plot represent the log(AM+1)sense-log(AM+1)antisense. The highest and lowest 20 genes are plotted on separate y-axis scales and genes in between as plotted as burgundy.

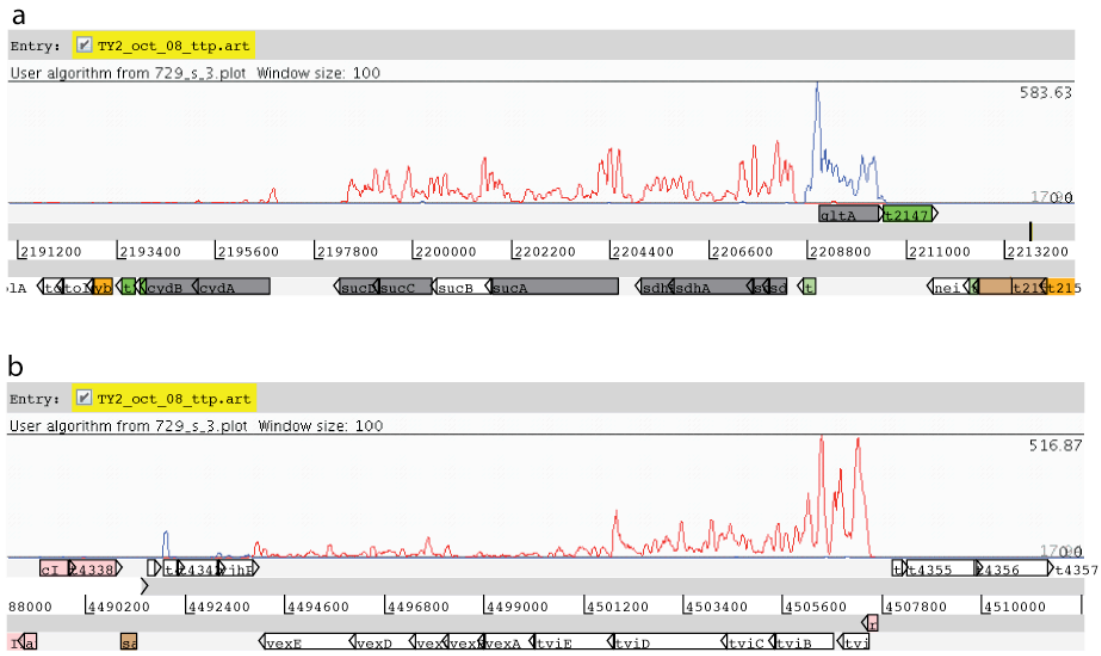

Figure 4.5. Transcripts mapping to well characterised regions.

Directional transcripts generally map to the coding strand for (a) succinate dehydrogenase operon (b) *viaB* locus. Red plot represents the reverse strand and blue, forward, window size=100bp.

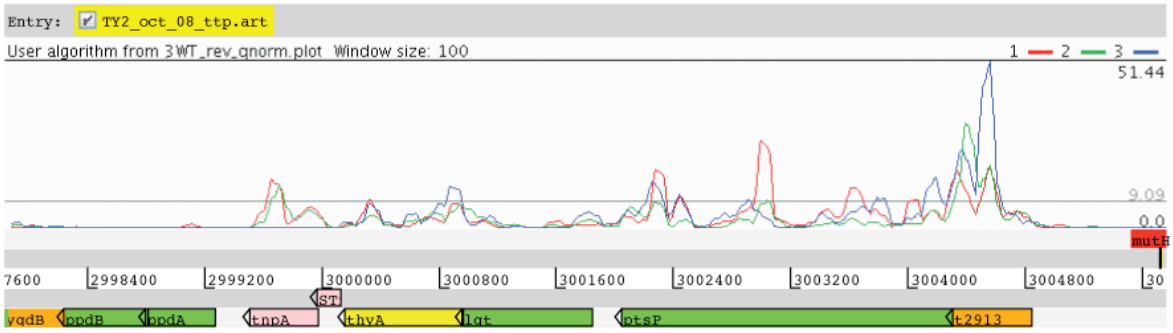

Figure 4.6 Consistency between sequencing experiments.

Colours represent sequencing run samples mapped to the reverse strand with each plot coloured, red represents the first sequencing run; green, second; blue, third.

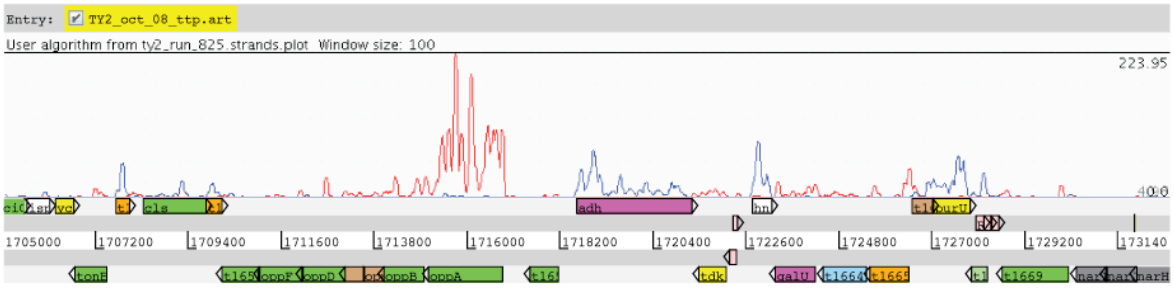

Figure 4.7 “Mosaic coding”.

Region where coding orientation is mosaic the directional sequence data is consistent with the annotation. Reverse strand, red; forward strand, blue.

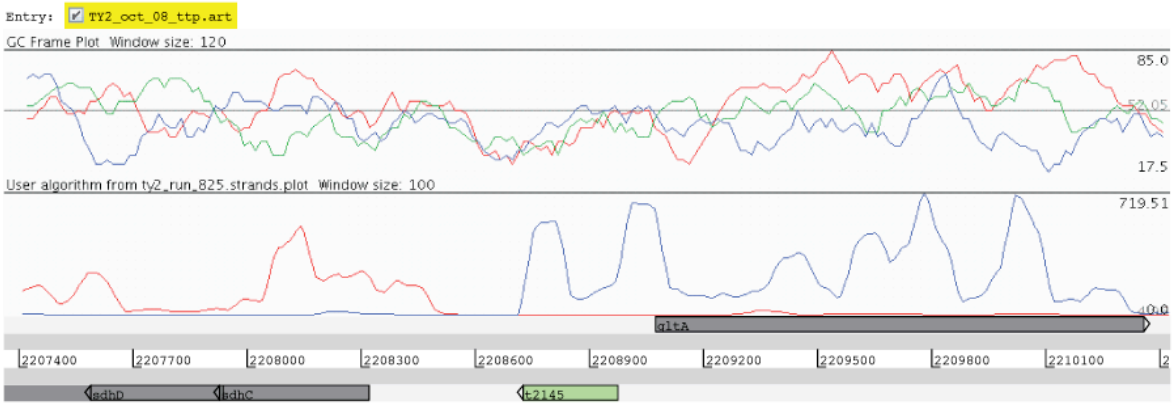

Figure 4.8. Putative error in the annotation.

Directional RNA-seq data highlights putative errata in the previous annotation. GC frame plot indicates ORF annotated as t2145 is not coding and sequence data maps to the 5’ region of *gltA*, indicating a putative riboswitch.

In order to provide an overview of the gene classes identified in the genome-wide transcriptome, the AM for each CDS was determined and compared to the previously assigned functional group classification [76]. Each functional class, represented as a percentage of the total genome-wide predicted number of CDSs, was compared with proportion of each in the entire transcriptome (figure 4.9). This approach effectively identified transcriptionally active classes expressed in the mRNA populations. A ratio of >1 represents a highly transcriptionally active class. The ratio for outer membrane/surface structures, regulators, conserved hypotheticals and central

intermediary metabolism were approximately 1. Interestingly, energy metabolism, pathogenicity/adapation/chaperones and information transfer were "over-represented" in the transcriptome ranging from 1.57 to 2.23. As may be expected, transcriptionally silent prophage elements (ratio ~ 0.75) are under represented, as are genes predicted to encode proteins required for degradation of both macromolecules and small molecules. Interestingly, pseudogenes represent 4.6% of the "coding sequences" yet a ratio of only 0.15 was observed for this gene class in the transcriptome.

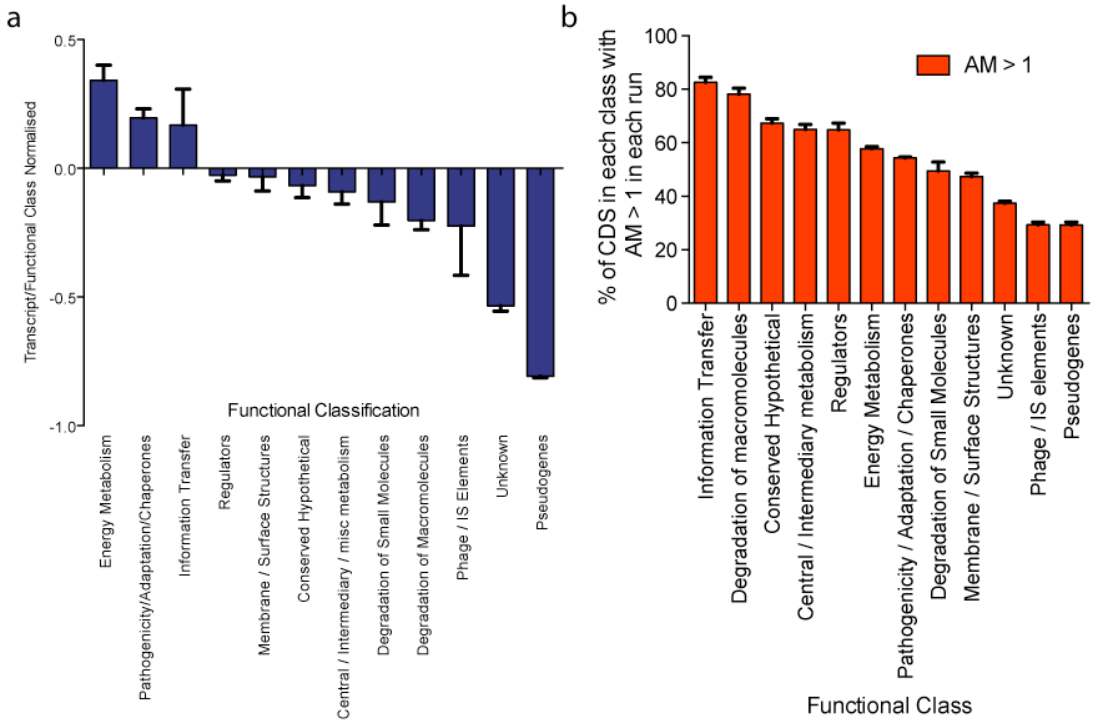

Figure 4.9. Functional classification of sequenced transcripts.
(a) The total number of reads/bp mapped to each CDS are assigned to functional class described previously [76]. These data were then normalised by the number of CDS for each function encoded within the entire genome. A ratio of 1 represents transcription of functional class on par with its genome content. A ratio of more than one represents a transcriptionally over-active class, and less than one, under-active. (b) Overview of *S.* Typhi Ty2 transcriptome assigned to functional class. The percentage of CDS in each functional class with an AM ≥ 1.

The sequence coverage for the five most highly transcribed genes ranged from an AM of 996 to 596 reads/base on the sense strand. The reads mapping to large operons with

high levels of transcription generally tended to map more 5' than 3' but the previously published CDS annotation is generally supported by these data sets. As expected high coverage was observed for abundant proteins such as flagellin (*fliC*) [203] and outer membrane porin C (*ompC*) [138] as well as for genes in the TCA cycle, such as the succinate dehydrogenase operon (*sdhCDAB, sucABCD*) [204]. The resolution of the data is striking as even features such as transcriptional attenuators could be readily identified. For example, analysis of transcripts upstream of the threonine leader peptide, *thrL* (figure 4.10), provides a classic example of an attenuator [205]. The *thrL* peptide sequence contains eight threonine residues stimulating attenuation when threonine is abundant, as in the LB broth cultures employed in these studies.

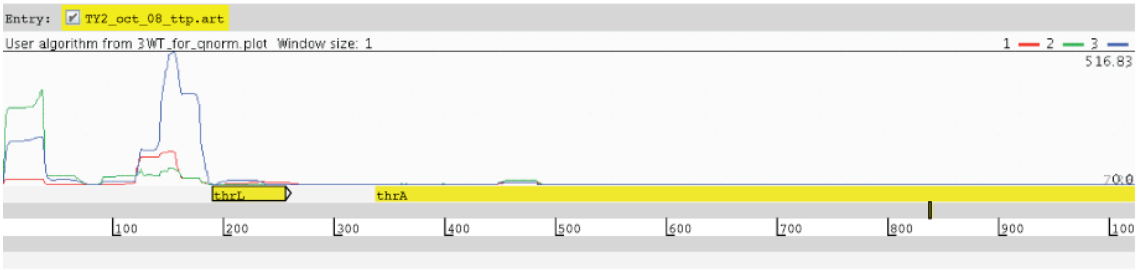

Figure 4.10. Attenuation.

RNA polymerase activity is inhibited after translation of the threonine rich leader peptide (thrL). LB is nutrient rich, thus reducing the requirement of threonine biosynthesis genes to be expressed. This feedback loop attenuates transcription. Run 1, red; run 2, green; run 3, blue. NB. Only data mapped to the forward strand is shown.

4.3.2.2 Virulence Genes

Genes encoding pathogenicity/adaptation/chaperone, outer membrane and central metabolism proteins were generally highly expressed (figure 4.9). Under the growth conditions employed, the most significant region of highly transcribed classical virulence associated genes clustered with SPI-1. The regulation of SPI-1 expression is

complex and involves the interaction of several regulators including HilA, HilD, InvF, SprAB and OmpR [206]. Transcripts for each of these regulators were highly represented but not all transcripts of the SPI-1 needle complex translocon were equally represented (figure 4.11). For example, *spaN* (t2794) and *spaM* (t2795) had an AM of 40 nc/bp and 62 nc/bp respectively, while *spaI* (t2796) and *spaO* (t2793) that flank the latter, and are transcribed in the same operon had an AM of 12.9 nc/bp and 6.08 nc/bp respectively in one experiment. These data suggest that mRNA stability may play an important role in control of translocon expression. Furthermore, transcripts originating from genes encoding SPI-1 effector proteins encoded outside of SPI-1 were also highly expressed. For example, transcript mapping to the *sigE* (t1829, AM=32.7) and *sigD* (t1828, AM=53.5) of SPI-5 both had high sequence coverage. In sharp contrast to the relatively high coverage of SPI-1 transcripts, SPI-2 transcripts had an arithmetic-mean coverage of just 1.8/CDS (figure 4.12). This disparity was expected since it is known that SPI-2 encoded genes are up-regulated in response to environmental conditions in the *Salmonella*-containing vacuole, under the control of the *phoPQ* and *ssrAB* two component regulators (Deiwick et al, 1999). Perhaps surprisingly however, the transcripts from two genes, *sscB* (t1271, AM=16.3) and *sseF* (t1272, AM=12.5), have considerable sequence intensity.

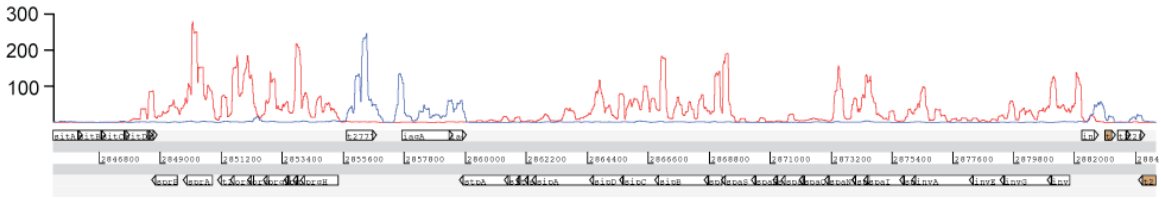

Figure 4.11 Coverage for SPI-1.

Plots represent fold-depth of sequence data mapped to each strand, forward strand, blue, red strand, reverse.

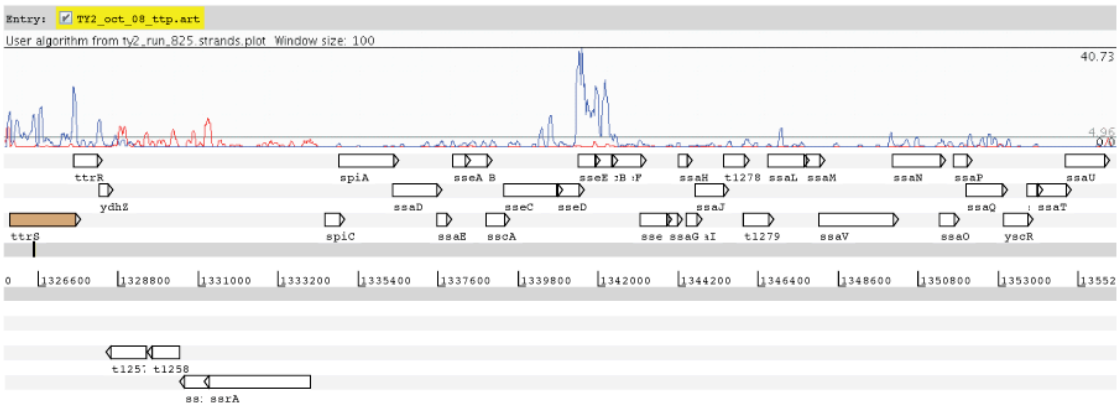

Figure 4.12 Coverage across SPI-2.

Sequence data mapped to two genes within SPI-2, suggesting an alternative regulator controls expression of genes *sseE* and *sscB*, Forward strand, blue; red strand, reverse.

Pili are a family of proteinaceous hair-like structures that can function as adhesins. In *S.* Typhi Ty2 there are a total of 13 pili determinants including a type IV pilus encoded by the *pil* operon, and 6 members of the usher chaperone family [76,81,123,207,208]. Interestingly, little to no transcripts mapping to pili operons were detected, suggests that complex and currently unknown signals are required for transcription and elaboration of these surface appendages.

Some of the most highly expressed genes in the *S.* Typhi genome were genes associated with flagella biogenesis and structural components and chemotaxis. The flagella major subunit, *fliC*, had an AM of 514. This high level of transcription underlines the role of this protein as the major subunit of the flagella structure. SPI-7 encodes *tvi* and the *vex* loci responsible for Vi capsule biosynthesis [118] and export. The *viaB* locus is known to be up-regulated in low osmotic potential environments under the control of the *ompR*/*envZ* two component regulator. Sequence coverage mapping to both the biosynthesis and export associated genes were high (figure 4.5b), with the entire locus having an average AM of 71.8 reads/base with a range of 8.48

(*vexE*) to 179 reads/base (*tviA*), Ty2 is known to be highly Vi positive when grown in LB [116].

4.3.2.3 Phage and putative cargo genes

S. Typhi harbours a number of distinct prophage, whose content can vary between the different evolutionary lineages [89,209]. Such prophages are regarded as being predominantly transcriptionally silent in the genome and can encode horizontally acquired 'cargo' genes potentially encoding factors that modify the virulence potential of the host bacteria. This analysis confirms that most of the resident prophage are indeed predominantly transcriptionally inactive (figure 4.13) but it is worth noting that the mapping was sufficiently sensitive to highlight low level transcription across phage regions involved in maintaining lysogeny. However, many of the prophage harbour transcriptionally active regions and some of these mapped over well known cargo or moron genes such as *sopE* (t4303, AM=186) encoded by the sopE phage (figure 4.13(a)). Similar analysis of this phage and others within the *S*. Typhi Ty2 genome highlights several transcriptionally active regions, which may encode novel cargo genes. Informatics analysis of these regions in some cases supports this hypothesis in that the genes do not encode known phage proteins but have functional protein similarities with genes in other pathogens such as *E. coli* 0157H7 (figure 4.13(b)), *Vibrio cholerae* (figure 4.13(c)) or the eukaryotic signalling enzymes phospholipase and putative threonine/serine kinases (figure 4.13(d)). Thus, these methodologies may provide a novel approach to identifying virulence genes expressed during the lysogenic phase.

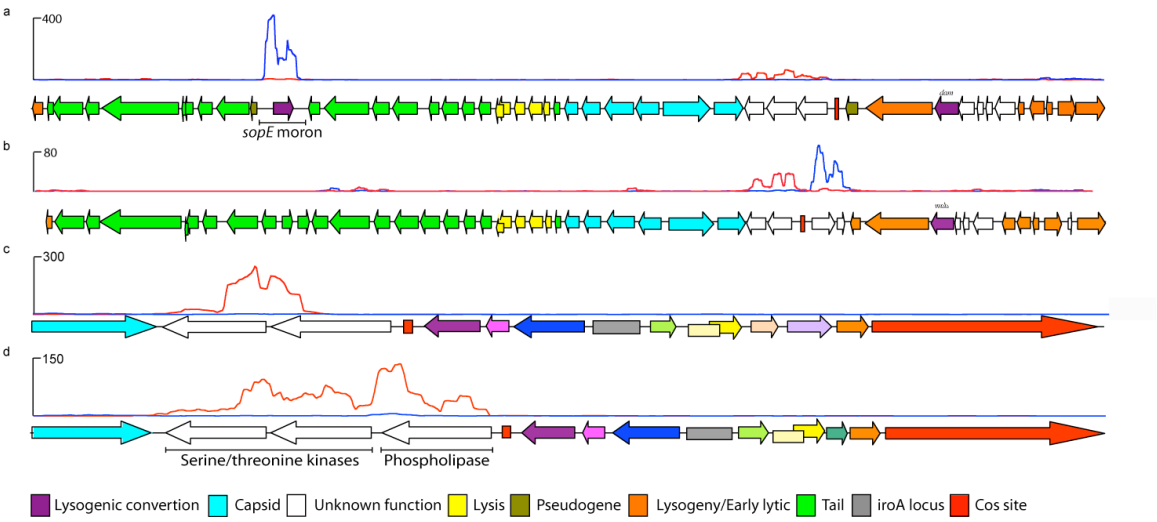

Figure 4.13. Sequenced transcripts identifying cargo genes within *Salmonella* prophage.

(a) Genetic organisation of the SopE prophage aligned with mapped sequence reads (blue forward strand, red reverse) illustrates "expression" of the sopE moron and a putative cargo region. (b) Genetic organisation of the ST35 prophage. The low-GC region maps significant sequence coverage compared with the prophage machinery putatively identifying it as cargo. (c) Putative ST2-27 Prophage Cargo. Low GC region maps significant sequence intensity. (d) Putative ST46 Prophage Cargo. Low GC region maps significant sequence intensity.

4.3.2.4 Pseudogenes

S. Typhi, in common with other host-adapted pathogens, harbours a large number (~220) of putatively inactivated pseudogenes. Genome degradation may contribute to host restriction by inactivating pathways essential for infections in the non-permissive host. Theoretically, putative pseudogenes can still express a functional truncated protein domain, as for example has been demonstrated for the CTL gene encoding a toxin in *Chlamydia trachomatis* [210,211]. Based on the previous annotation there were nine pseudogenes in *S.* Typhi Ty2 that exhibited high levels of transcription, suggesting that they may be expressed as functional proteins. Interestingly, peptides that corresponded to the open reading frame upstream of the inactivating stop coding of one of the transcribed pseudogenes, *hdsM* (t4575) where detected using Mass

Spectrometric analysis of *S.* Typhi Ty2 extracts (our unpublished data). This represents the only evidence in this study of translated pseudogenes. This combined with the shear lack of transcriptional abundance of other *S.* Typhi pseudogenes further supports the current interpretation that these genes are no longer active.

4.3.2.5 Hypothetical Genes

Many genes in the *S.* Typhi genome were initially annotated as hypothetical coding sequences in the absence of any direct evidence for transcription or translation into a protein product. A considerable number of these hypothetical genes were also identified as orthologues in the genomes of *E. coli,* and other bacterial species. The annotation derived by Parkhill *et al* (2001) assigned each predicted coding sequence to a functional class according to *in silico* analyses. Thus, all genes that were annotated as putative, probable, predicted, hypothetical and possible were included in this analysis (appendix 9.5). This analysis encompassed ~2900 genes. 72 genes mapped no sequences for all three experiments and 677 genes had an average AM ranging from 0.01 to 0.10. 1751 genes had an average AM from 0.11 to 1.0o and 293 genes mapped an average AM ranging from 1.01 to 10.0. 75 genes ranged from an average AM 10.0 to 730.

4.3.2.6 A *S.* Typhi Ty2 Specific Insertion

Annotation performed following the sequencing of the *S.* Typhi Ty2 genome failed to identify a variable region that mapped significant transcriptome sequence data [136]. This region, now referred to as ST20, shows variation between *S.* Typhi isolates and is composed of a mosaic of up to three known inserts, a, b and c [87]. *S.* Typhi strains CT18 and Ty2 encode forms a and b respectively. The synteny of this island is

specific to Ty2 and maps significant transcript sequence data (figure 4.14). This region, which encodes 27 putative CDSs in Ty2 mapped significant numbers of sequence reads. The highly transcribed genes are functionally homologous to two restriction enzymes, a negative regulator of *N*-acylhomoserine and protein with no significant sequence homology.

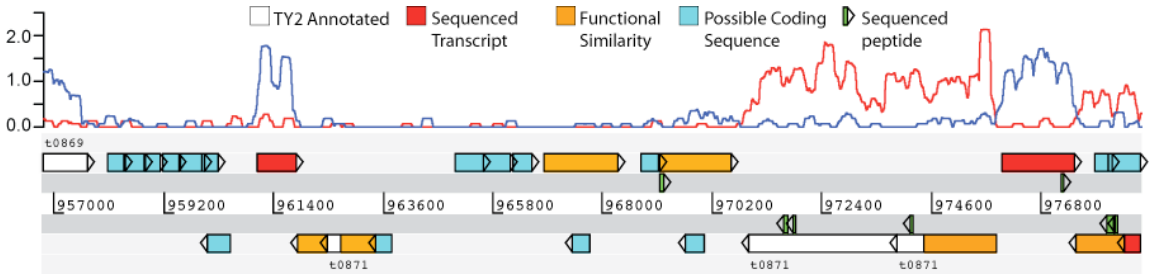

Figure 4.14. Third Party Annotation of the Ty2-specific Insertion.

Alignment of transcriptome sequence data to this previously poorly-annotated region of Ty2 represents further evidence for putative CDS. All possible CDS's are coloured blue, those with predicted functional protein homology are mustard, white are previously annotated and red annotated by transcriptome sequence data. Green features are mapped peptide sequences (FDR<0.076), which map to the same translation frame as the aligned

4.3.3 Non-Coding Sequences

The AM for each currently annotated sncRNA was determined. 67 of the known 151 have an average AM > 1, ranging from 1.17 to 1004 reads/base-pair (figure 4.15a). Furthermore, many of the transcripts identified by sequencing mapped back to regions of the *S.* Typhi Ty2 genome that were previously unannotated, predicting a further 40 regions as expressed non-coding sequences. Furthermore, many of these were not unique within the Ty2 genome and similar sequences were annotated as paralogues to include in this analysis. Furthermore, 127 CDS were identified that were preceded by putative 5'UTR transcripts, 31 of which were more than 150bps in length (appendix 9.6). There were also two novel putative 3'UTRs adjacent to *sprB* and *ramA*

respectively. Subsequently, we determined the AM for each prediction (figure 4.15b) and 85 of the 239 elements had an average AM > 1. Taken together, these sequence data suggests that there may be many previously unidentified functional non-coding RNAs present in *S.* Typhi, and potentially in other bacteria. Consequently, bioinformatics analyses were used to further interrogate these data [212].

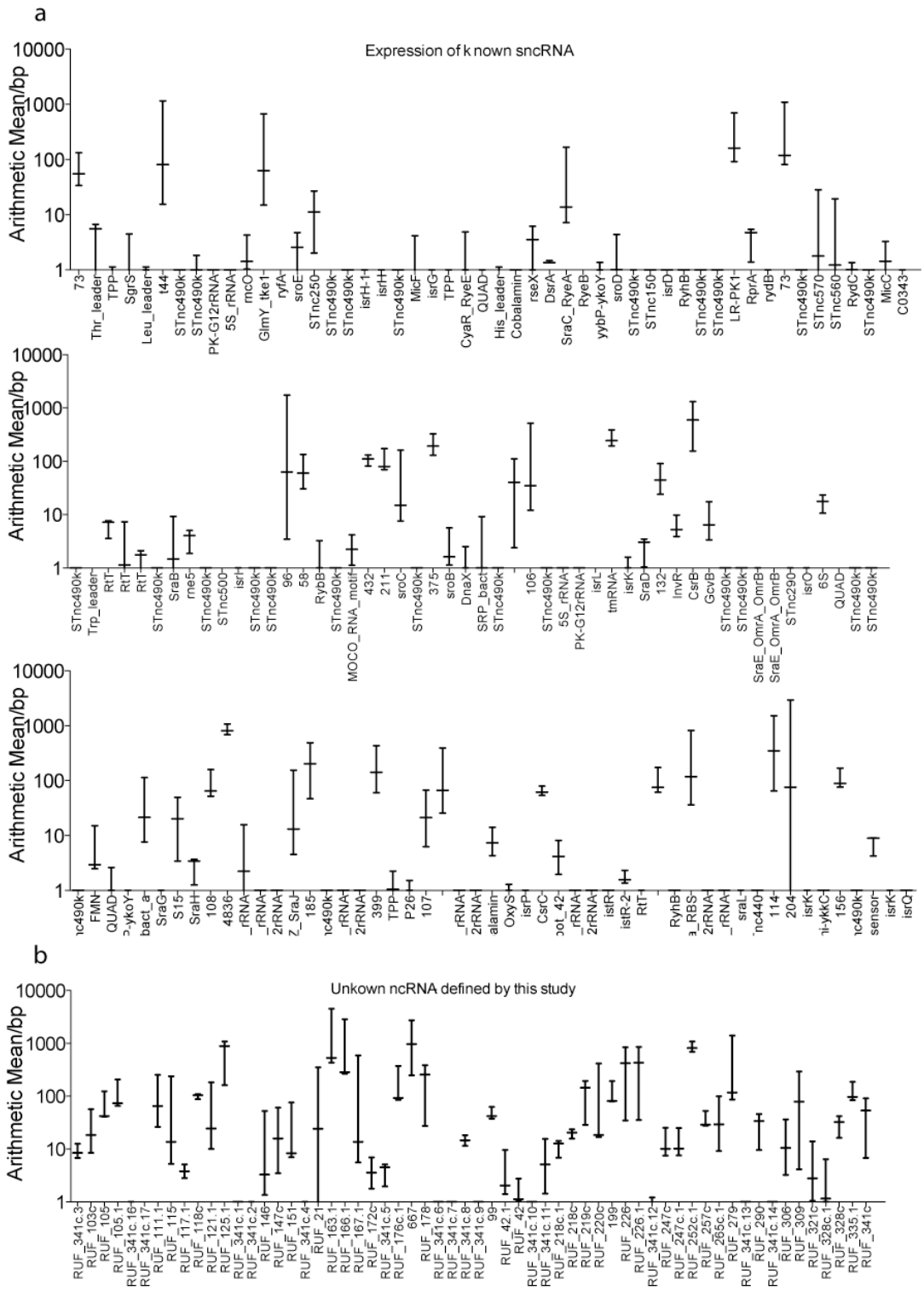

Figure 4.15 Coverage of putative non-coding RNAs.

(a) Sequence coverage (AM) of ncRNAs in each sequencing run, derived from previous annotation and publication. (b) Predicted ncRNA based on sequenced transcripts mapping to previously unannotated intergenic region or paralogous regions of ncRNA derived by this study.

Transcription and translation in prokaryotes is commonly regulated by changes to the conformational structure of *cis*-acting non-coding RNAs called riboswitches. These RNA's generally bind metabolites related to the function of each downstream gene [213,214] and have been identified bioinformatically based on sequence conservation of the 5' UTR. Several known riboswitches, such as *btuB* [200], *glmS* [202] and TPP [215] were highly represented in these data. Possibly one of the most interesting regions encoding novel non-coding RNAs was part of SPI-1 (figure. 4.16). Two of these SPI-1 associated transcripts were identified by the programme RNAz as candidate riboswitches, here designated SPIS1 and SPIS2 (RUF220c and RUF219c) (4.16b and c). SPIS1 and SPIS2 are located directly upstream of the AraC-like regulators *sprA* (t2988) and *sprB* (t2987), respectively. The third candidate element which is predicted to be a 3'UTR, named SPIS3 (RUF218c) (figure 4.16), is antisense to the *sitD* gene, an iron transport protein [216] and a hypothetical protein O30622 (t2767), which may have been acquired independently of the rest of SPI-1 [12]. The sequence of RUF218c is conserved across cyanobacteria, firmicutes and proteobacteria. The *sitA* gene maps sequenced transcripts (average AM=1.27), whereas *sitB*, *sitC* and *sitD* have slightly lower levels of expression (AM = 0.37, 0.38 and 0.61, respectively). It is possible that RUF218c is an antisense repressor of these proteins as it is predicted to form a moderately stable MFE secondary structure compared to a shuffled ensemble of sequences that have the same di-nucleotide composition (p=0.0090). The fourth candidate element, named SPIS4 (RUF221) maps to the 5' UTR of *iagA* (t2999) (figure 4.16), an invasion protein regulator. The structure of this RNA (34% G+C) is not predicted to be significant (RNAz p=0.0037 and shuffling p=0.2627) (figure 4.16c).

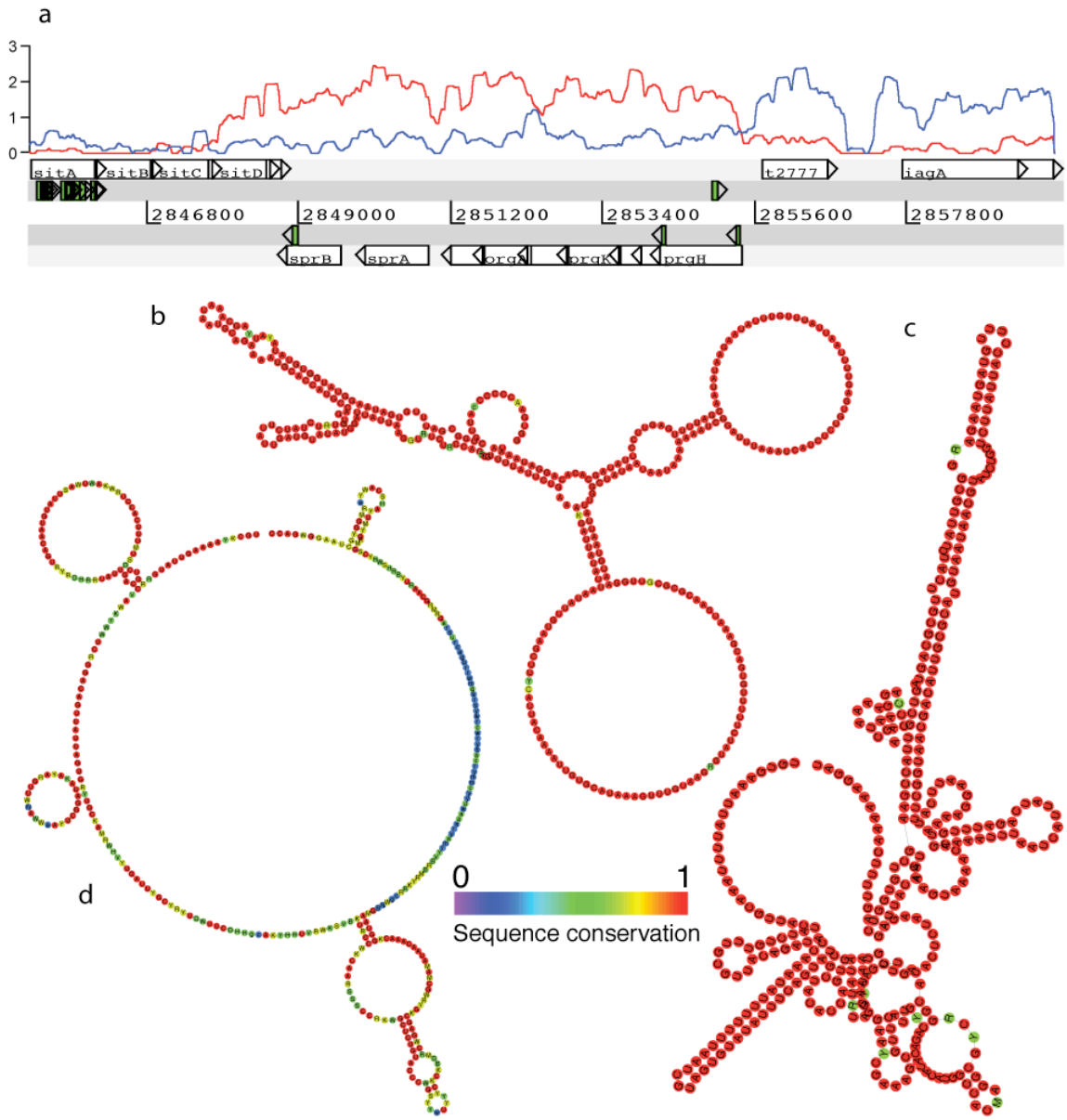

Figure 4.16. SPI-1 Sequence coverage

(a) Artemis representation of part sequenced transcript mapping to the highly transcribed region SPI-1. Log(10) +1 of Forward strand plot, blue; Log(10)+1 of reverse strand; red. SPIS1 is region upstream of *sprA*; SPIS2 is region upstream of *sprB*; SPIS3 is region downstream of *sprB*; SPIS4 is region upstream of *iagA*. Dark grey horizontal line represent translation of each CDS and green ORFs represent sequenced peptides mapping to those CDS. (b) Predicted secondary structure of data mapped to region 5' of *sprA* (p(RNAz)= 0.933655). (c) Predicted secondary structure of region 5' to *sprB* (p(RNAz)= 0.809987). (d) Predicted secondary structure of region 5' to *iagA* (p(RNAz)=0.003772).

A further non-coding feature of note is RUF107c, which is highly expressed in these

S. Typhi samples. This element, predicted to be highly structured by RNAz

(p=0.9396), has approximately 115 paralogues in *Salmonella* (figure 4.17). Further, it is conserved across ~82 bacterial species but is chiefly restricted to Enterobacteriaceae. The genomic context of RUF107c is not consistent with a *cis*-regulatory or a transposable element, as the sequence does not consistently co-occur with either CDSs or near transposases, respectively.

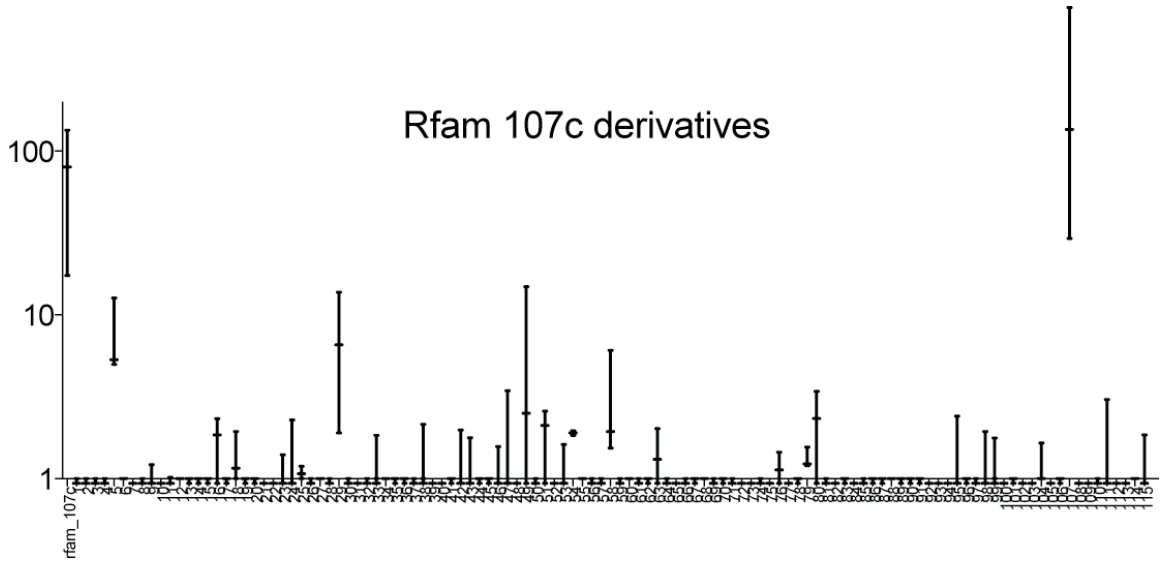

Figure 4.17. Paralogues of RUF107c.

AM of sequence coverage for predicted RUF107c paralogues in *S*. Typhi.

4.3.4 Transcriptome comparison with *ompR*-null mutant

4.3.4.1 Quality control

The validation of the first RNA-Seq experiments using *S*. Typhi Ty2 BRD948 prompted further experiments utilising similar RNA prepared from an *ompR* mutant of the same strain. The OmpR/EnvZ system is a key two component regulator controlling the expression of virulence phenotypes by *S*. Typhi. Hence, RNA-seq analysis could potentially reveal some novel aspects of OmpR mediated expression regulation.

The data described here are the product of a number of Illumina sequencing runs carried out during the development of protocols designed to deplete RNA preparations of 16s and 23s and provide high-throughput strand specific sequence information. Mapping statistics for each run and starting material are detailed in table 4.2.

Table 4.2. Sequence data mapped for each experiment

Sequencing Data Table						
Flowcell/Lane	876/2	1104/2	1354/1	876/5	1354/2	1104/3
Strain	BRD948	BRD948	BRD948	BRD948DompR	BRD948DompR	BRD948DompR
rRNA Depletion	Oligo	Oligo	Oligo	Oligo	Oligo	Oligo
Mass of Total RNA	300	100	100	300	100	100
Read Length	36	36	36	36	36	36
No Reads	5608589	7183969	5848604	5375891	3098524	7399877
Total Mapped	5438270	6513814	5356994	5249193	2774513	6454861
Total Mapped (%)	96.9	90.7	91.5	97.6	89.5	87.2
Total Uniquely Mapped	1493759	2942477	2326287	1749240	610817	2119151
Total Mapped Uniquely (%)	0.266334189	0.409589323	0.397750814	0.325386062	0.197131602	0.286376517
Reads Mapped to CDS	1235932	2212650	1937241	1500449	491028	1617543
Reads Mapped to NC sequences	257827	729827	389046	248791	119789	501608
Reads mapped to pseudogenes	12403	45771	36678	14126	17185	50531
Reads mapped to hypothetical genes	131242	266139	264871	248791	78606	275976
GC content (ALL)	0.519742418	0.41840995	0.453535147	0.53065418	0.496200834	0.392659886
GC content (UNIQUE)	0.50318224	0.44267108	0.471412408	0.517455104	0.479285167	0.418839576
Coding Sense	1124620	1732827	1565156	1311837	295552	1056574
Antisense Coding	111461	480504	372642	188739	195794	561813

Genomic DNA contamination is evident in samples 876_s_1.wt and 876_s_3.mut when the data are plotted in Artemis (not shown). The histogram (figure 4.18(a)) or box-plot (figure 4.18(b)) also illustrates the differences between the runs with contaminating genomic DNA and cDNA. Further, samples were then prepared on a total of three occasions and the total RNA isolated ranged from 300µg to 100µg as the protocol was optimised. After mapping of read data with a quality score of 30 using MAQ, the AM for each CDS was determined and the data normalised by quantile normalization [187]. Due to inconsistencies in the box-plot and DNA contamination discovered during Illumina sequencing a subset was chosen for further analysis (figure 4.18(c)). Clustering of these sequence data intensities revealed each sample is

more closely related to the sample on the same flowcell rather than the strain it was isolated from. However, the comparative analysis was performed in order to assess the power of these experimental data and methods.

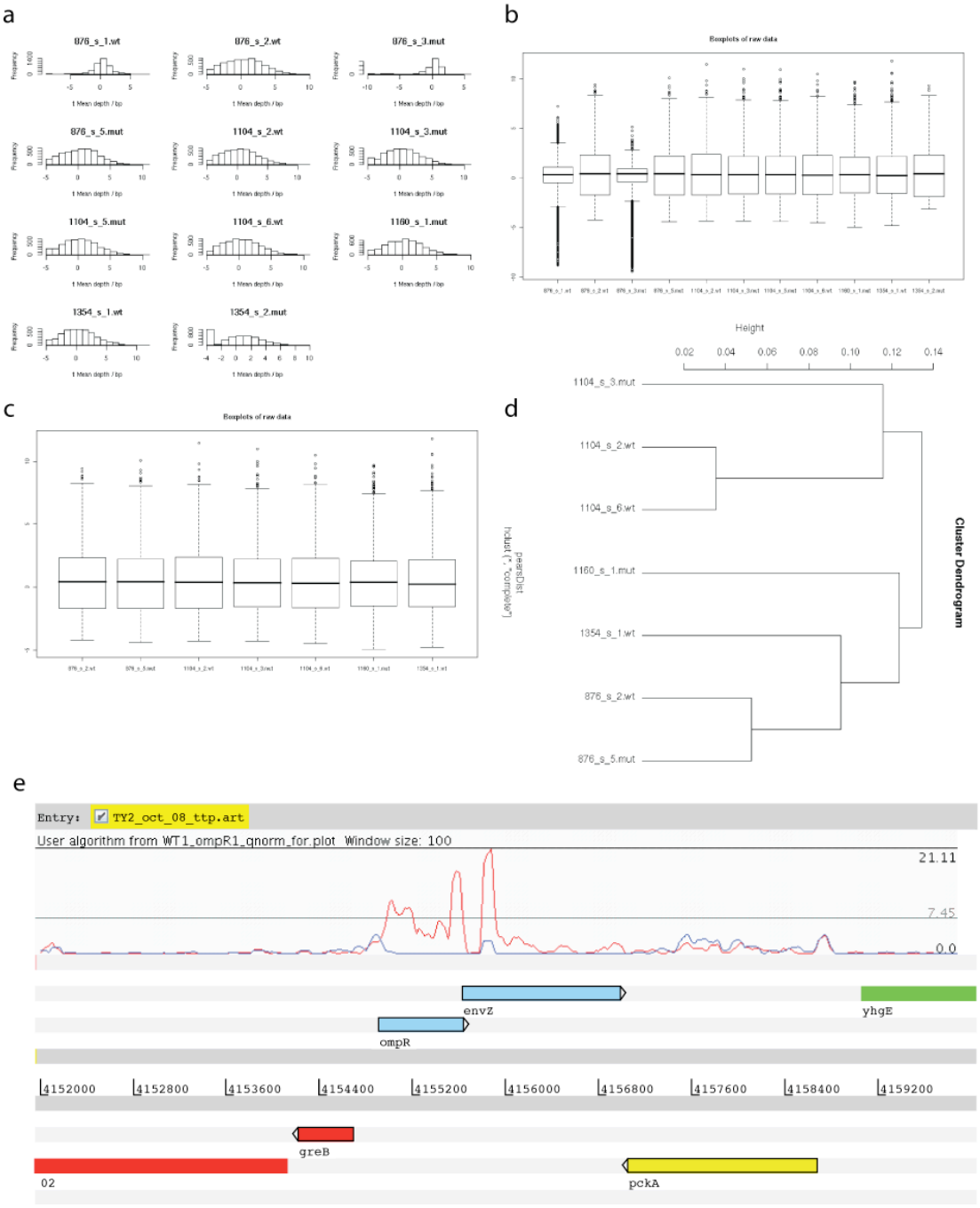

Figure 4.18 Quality control.

(a) Histogram of sequence data mapping to coding sequences (log2) (b) Box plots of depth over coding sequences (c) A subset chosen for further comparative analysis after quantile normalisation of the sequence data mapping to coding sequences (d) Pearson correlation clustering of subset. (e) OmpR deletion. No sequence data are mapped to the region deleted in *ompR* (blue), WT (red).

4.3.4.2 General features of the BRD948 *ompR* transcriptome

Initially the data derived from the transcriptome of the BRD948 *ompR* mutant were mapped to each of the functional gene classes described by Parkhill *et al* (2001) (figure 4.19). Overall, as expected these data were similar between the wild-type and *ompR* mutant derivatives. OmpR is known to strongly activate expression of the *viaB* locus [118] and this region maps a very much reduced sequence data in the null-mutant (figure 4.20). Expression of *ompC* is also strongly activated by the presence of OmpR, and very few transcript reads mapped to this region in RNA seq of the *ompR* strain (figure 4.21).

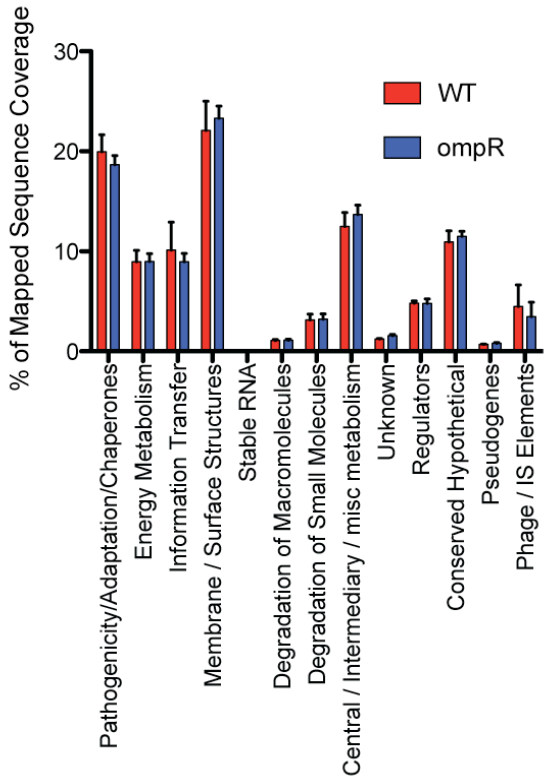

Figure 4.19 Assigning transcriptome to functional class.

Data mapped to the functional class derived by Parkhill et al. (2001) for CT18 to determine the overall AM/bp for each class.

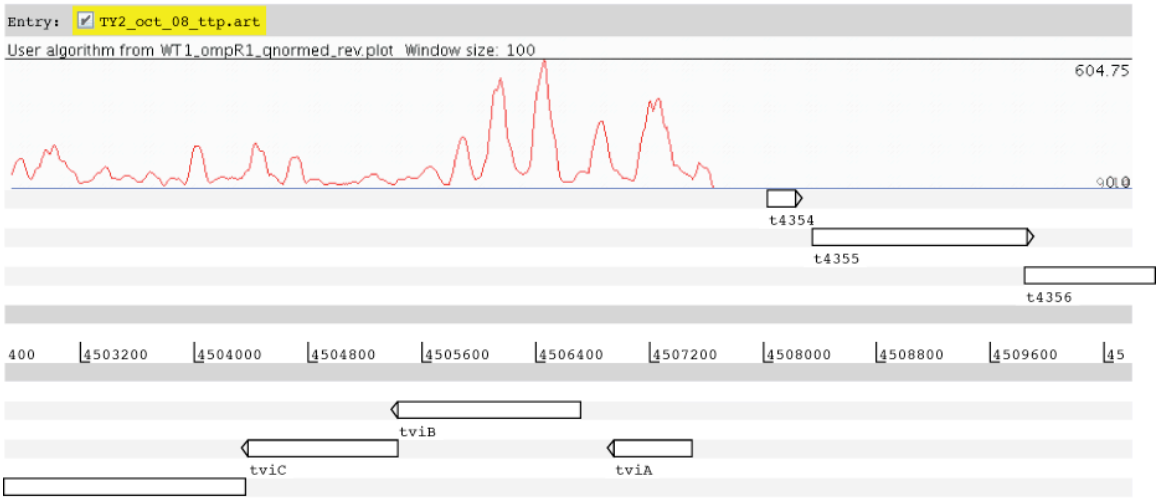

Figure 4.20. Start of the *viaB* locus.

Sequence data mapping to the coding strand of the *viaB* locus is considerably less in the *ompR* mutant (blue) compared with the WT (red).

Figure 4.21 Outer membrane porin C.

The transcript sequenced and mapped to the coding strand in the *ompR* mutant (blue) is clearly less abundance than WT (red).

4.3.4.3 Quantified differences

The AM was determined for all annotated CDSs and these values were taken as intensity values much like values derived by microarray scanning. Using the AM and

the LIMMA package for microarray analysis the data were normalized by quantile normalization [187]. Quantile normalization attempts to match each percentile of each data set. This analysis identified 15 genes (adj $p<0.05$) that were differentially transcribed in the *ompR*-null mutant, none of which were significantly increased (table 4.3). These data were consistent with OmpR acting as a transcriptional activator [217].

Table 4.3 Statistically different genes in the *ompR* mutant

Gene Name	Start Co-ord	Stop Co-ord	logFC	t	P-value	adj P-value	B	Product
tviC	4504378	4505424	-8.17	-14.41	5.10E-06	4.24E-03	4.52	Vi polysaccharide biosynthesis protein, epimerase
tviA	4506949	4507488	-8.05	-8.72	1.02E-04	3.15E-02	2.01	Vi polysaccharide biosynthesis protein
tviB	4505427	4506704	-8.03	-13.21	8.63E-06	4.66E-03	4.13	Vi polysaccharide biosynthesis protein, UDP- glucose/GDP-mannose dehydrogenase
tviD	4501859	4504354	-7.09	-16.64	2.12E-06	4.24E-03	5.12	Vi polysaccharide biosynthesis protein
vexB	4498207	4499001	-6.98	-13.36	8.07E-06	4.66E-03	4.18	Vi polysaccharide export inner-membrane protein
vexC	4497486	4498181	-6.89	-9.52	6.11E-05	2.03E-02	2.49	Vi polysaccharide export ATP-binding protein
vexA	4499011	4500078	-5.73	-14.28	5.38E-06	4.24E-03	4.48	Vi polysaccharide export protein
tviE	4500123	4501859	-5.71	-14.07	5.89E-06	4.24E-03	4.41	Vi polysaccharide biosynthesis protein TviE
vexE	4494169	4496139	-5.62	-14.96	4.05E-06	4.24E-03	4.68	Vi polysaccharide export protein
vexD	4496159	4497463	-5.50	-11.28	2.23E-05	8.77E-03	3.36	Vi polysaccharide export inner-membrane protein
hyaA	1500712	1501815	-5.02	-14.36	5.20E-06	4.24E-03	4.50	uptake hydrogenase small subunit
NA	1502243	1503334	-4.82	-12.29	1.33E-05	6.39E-03	3.79	putative secreted hydrolase
envZ	4155629	4156981	-4.17	-11.09	2.47E-05	8.87E-03	3.28	osmolarity sensor protein
slsA	3873860	3874540	-3.63	-8.42	1.25E-04	3.60E-02	1.82	hypothetical protein
NA	1701305	1701808	-3.39	-11.50	1.98E-05	8.57E-03	3.46	hypothetical protein

The entire *viaB* locus is represented in these data, which agrees with Pickard *et al.* [116]. As expected, *envZ* is also significantly decreased. This is an important endogenous control as it is downstream of the defined mutation in *ompR*.. The *ompR* gene is the first gene in the *ompB* locus and is not represented in these data as it was filtered out by the false discovery rate Benjamini-Hochberg correction. Benjamini-Hochberg correction is used as a post-analysis estimation to allow for false discovery

based on multiple analyses and variation between experimental testing [218]. However, this sort of correction also includes a level of false non-discovery rate (FNR) and the shear lack of *ompR* reads in these data suggests such an estimation may be too stringent for this experiment.

Interestingly, there are genes in this dataset that were not previously described as *ompR* regulated. The *slsA* gene (12.3 fold down, adj p<0.036), encoded within SPI-3, is conserved throughout *Salmonella* and sequence identity suggests it is an inner membrane protein with homology to the isochorismatase hydrolase family of enzymes. Isochorismatase hydrolase has been characterised in the phenazine biosynthesis pathway in *Pseudomonas aeruginosa*. Phenazines are a group of 70 related compounds, some of which are antimicrobial and others induce neutrophil cell death [219]. The function of the *slsA* protein is not known so the consequence from its reduced expression remains to be elucidated.

The hydrogenase uptake gene, *hyaA2* is decreased (32.4 fold down, adj p<0.004). *Salmonella* encodes three predicted hydrogenase operons, two hydrogenase 1 operons (*hyaACDEFt1048* and *hyaA2B2C2D2E2F2t1454*) and a hydrogenase 2 operon (*hybOABCDEFG*) that are important factors in respiration. Interestingly, both *hyaA* and *hyaB2* are pseudogenes in *S.* Typhi Ty2 and CT18. The hydrogenase protein complex consists of Hya or Hyb subunits and enzymatically splits H_2 to release electrons to reduce downstream components in the respiratory chain and all three of these operons contribute to virulence in the murine model [220]. Deletion in all three severely attenuates *S.* Typhimurium. Furthermore, the gene divergently transcribed from *hyaA2,* a putative secreted choloylglycine hydrolase, t1459 is also significantly decreased (28.0 fold down, adj p<0.006). The family of choloylglycine hydrolases

cleave carbon-nitrogen binds, exclusive of peptide bonds and includes conjugates bile acid hydrolase and penicillin acylase [221].

Expression of the conserved hypothetical gene, t1641, that exhibits no significant identity to any characterised gene but is conserved in *E. coli* K12, (annotated as yciF) is down (10.4 fold down, adj p<0.009).

4.3.4.4 Quantified differences pre-Benjamini Hochberg correction.

There were 305 genes with 2-fold differences (p<0.05)(appendix 9.7) in these data prior to Benjamini-Hochberg false discovery rate estimation and correction. Ninety-nine of these genes were decreased in expression and 49 genes had one or more contiguous genes differentially regulated, suggesting they are encoded by an operon structure. These data set may provide useful information if corroborated by a different gene expression method as it contains genes reported previously to be *ompR* regulated in *Salmonella*, such as *ssrAB*, *ompC* and o*mpS*. Furthermore, many of the genes predicted to be in an operon structure were differentially expressed with consistent fold-change direction.

4.4 Discussion

4.4.1 General transcriptome results

Overall, application of this Illumina sequencing technology has provided a detailed insight into the entire transcriptome of *S.* Typhi at one particular phase of growth. This technique has proven to be reproducible between experiments and has identified

many previously described features. However, the most interesting part of any study such as this, is the identification of novel expression features. This experiment has mapped sequence data to hypothetical open reading frames and has been able to identify putative errata in the existing annotation. Such information may be crucial for rationalising further work on *Salmonella* and possibly *E. coli.*

Identification of novel features specific to *Salmonella* has in the past permitted identification of virulence mediators. A global survey of the transcriptome on this scale *in vitro,* will unfortunately not identify such candidates without further *in vivo* experimentation, however, it will provide a rationale for further work. Combined with the comparison of *S.* Typhi with the *S.* Typhimurium and/or *E. coli* genomes*,* it is possible to further define variation and consequently highlight determinants potentially involved in pathogenicity.

Possibly the most interesting and exciting data are the transcripts that mapped to four intergenic regions of SPI-1. SPI-1 is predicted to be one of the oldest horizontally acquired regions of the *Salmonella* genome and this region does not exhibit any CDS degradation or intergenic region degradation within *Salmonella*. Such conservation suggests these intergenic regions are as functionally important as the coding sequences and the recent discovery of riboswitches underpins this logic. Investigation of possible ligands that bind to these 5'UTRs should further elucidate mechanisms of SPI-1 regulation. Currently, defined riboswitches are limited to non-coding regions and are generally predicted bioinformatically based on a significant gap between CDS and conservation. It is difficult to determine conservation for *Salmonella* specific islands, however, the identification of such transcribed regions in SPI-1 may permit

rational identification of such regions present in SPI-2, which also contains a significant number of intergenic regions.

Interesting work on riboswitches and the ligands that bind to alter their conformation has identified an antimicrobial compound pyrithiamine pyrophosphate, which inhibits binding of thiamine pyrophosphate and induces cellular dysfunction [222]. Identification of a natural ligand for *Salmonella* specific riboswitches may then facilitate the identification of synthetic ligands with microbicidal activity, thus permitting selective killing that would potentially not affect the overall microbiota. Indirect disruption of the normal flora through antimicrobial therapy has been shown, in many cases, to cause more harm than the targeted infection [223].

4.4.2 OmpR comparison

Due to the lack of any precedent for the use of RNA-seq technology in bacteria, the methods and protocol were continuously modified during the course of these experiments in an attempt to optimize the approach. Cost was also a factor driving some of the changes.

Nagalakshmi *et al.* [188] and Wilhelm *et al.* [189] published the first method for transcriptional analysis by RNA-seq approximately three months after this experiment was designed. These techniques were used to sequence the more stable poly-adenylated RNA of eukaryotes after enrichment using oligo(dT). Furthermore, Marioni *et al.* [224] found that during an analysis of technical reproducibility the cDNA concentration increased the variability of results. The method has undergone multiple variations and iteration throughout development and relies on a considerable

amount of RNA manipulation. This may limit the scope for definitive quantitative analysis.

The parent strain, BRD948, is deficient in part of the stress response (*htrA*) mechanism. The serine protease, HtrA, has been associated with virulence [72] and is homologous to a heat shock protein in *E.* coli involved in protein degradation and fidelity [225]. The aromatic amino acid biosynthesis pathway in *S.* Typhi BRD948 has also been disrupted promoting a need to scavenge *para*-aminobenzoic acid and dihydroxybenzoate [226], which is supplemented in all growth media. All comparisons are between isogenic strains (plus or minus the defined mutation in *ompR*), thus nullifying any effect these modifications may have on the global transcript.

This study was able to identify genes currently reported as being under the control of OmpR as well as extending current information on this regulon. Pickard *et al* (1994) demonstrated the *viaB* locus of *S.* Typhi is OmpR regulated and a deleting mutation arrested Vi expression. Expression of the Vi genes was found to be significantly decreased in the *ompR* mutant using RNA-seq analysis. Interestingly, *ompC* (figure 4.21) is highly activated by the presence of OmpR but is not found in the significantly different gene list. Under these conditions the OmpR regulon apparently includes the SPI-3 gene *slsA*, a putative hydrogenase gene, a putative choloylglycine hydrolases and a conserved hypothetical gene.

Improvements could be made to the experimental approach, based on the variation identified in these experiments and the observations published by Marioni *et al.* [224]. For example, ideally similar concentrations of cDNA should be prepared for a comparative experiment submitted to the same flowcell for sequencing.

4.5 Conclusion

The development of an RNA-seq approach to the analysis of the *S.* Typhi transcriptome provided an enormous amount of novel data. However, to some degree this technique should be regarded as 'a work in progress' and further developments of the experimental protocol and analysis tools will be beneficial. It has taken much iteration to get to this point and we are currently sequencing the transcriptome of *Clostridium difficile, Streptoccocus pneumoniae, Haemophilus influenzae, Campylobacter jejuni* and *Mycobacterium tuberculosis* using this refined protocol. Ultimately, this technique may provide an important quantitative analysis in both eukaryotes and prokaryotes.