7 General Discussion

New DNA sequencing technologies have greatly increased the throughput of genome sequencing and have facilitated novel approaches. A combination of Illumina and 454 pyrosequencing recently permitted the re-sequencing of genomes of ~21 *S.* Typhi isolates and the determination of a highly resolved phylogenetic tree [87]. Similar approaches have been been applied to eukaryotic genomes to identify chromosome rearrangements, such as those in carcinoma cells (WTSI, unpublished data). This study has used Illumina sequencing technology in a novel manner to characterise the transcriptome of *S.* Typhi by exploiting the intrinsic secondary structure of RNA species. This secondary structure has enabled ligation of linkers and removed the need for a second strand synthesis step, which normally masks complete identification of the transcribed component. This facile method has allowed us to identify the template strand, something not possible using methods reported in recent related publications in Science [188] and Nature [189]. These RNA-seq articles illustrate the application of deep sequencing technology to further define the transcriptome, however, such techniques are biased to sequence enriched polyadenylated RNA. Prokaryotes do not adenylate mRNA so this method cannot readily be applied to *S.* Typhi. The 16s and 23s rRNA are the most abundant species in the bacterial transcriptome and it was predicted the presence of these RNA molecules would reduce sequencing capacity for less abundant transcripts. Sequencing of the total RNA population in Chapter 3 illustrates the importance of the rRNA depletion step in identifying transcription of less abundant genes. rRNA depletion increased the sequencing capacity available for less abundant transcripts, however, the extra steps involved also increased the manipulation of naturally labile bacterial RNA. Thus, these data are potentially

limited by the level of removal of rRNA, 16s and 23s, the ability of each target cDNA to form a sequencable secondary structure and the half-life of mRNA. Further work should include direct comparison with ds-cDNA Illumina sequencing and application to identifying directionality in eukaryotic transcriptomes.

Importantly, these data contain many endogenous controls such as the presence of high levels of transcription from well-characterised genes such as *fliC* [203], the *viaB* locus [118], *sopE* [103] and the SPI-1 locus [12]. The resolution of the technique is such that it is possible to identify RNA editing features such as attenuation of the threonine leader peptide region [205], recently identified RNA regulatory elements such as the *glmS* riboswitch [202] and the small non-coding RNA, *tke1* [280]. Furthermore, comparison of transcriptomes with the *ompR* mutant *S.* Typhi supports the validity of the sequence data. There was very little sequence data mapping to the *ompC* [139] or *viaB* locus [116] as expected in this mutant. Subsequently, these endogenous controls permitted putative identification of non-coding elements, either as small non-coding RNAs or putative riboswitches. Many had not been previously identified and were supported by further *in silico* analyses. This technique has the capacity to identify the template strand and these data offer the scientific community an important database for further work on prokaryote transcriptomes and the virulence potential of such genes in *Salmonella*. Further work may include confirmation of these binding sites using gel shift assays. A high-throughput epitope tagging programme is currently being devised to study every gene with a predicted DNA binding domain in *Salmonella.*

DNA microarray technology was exploited in Chapter 5 to further define the OmpR regulon. These data indicate OmpR is involved in regulating outer membrane and

virulence associated genes and adjusting the respiratory programme. Many genes that were altered were associated with anaerobic metabolism, an essential component of survival in the anaerobic gut. The chIP-seq method using the Illumina platform was also developed and these data were used to identify binding sites for OmpR. Encoded within many of the chIP-enriched sites was a particular motif that was sometimes encoded as a couplet. This putative consensus sequence was upstream of many of the previously published genes in the OmpR regulon.

As OmpR is a conserved two-component regulator and is present in *E. coli,* why does it control so many horizontally acquired regions of DNA encoded within the *S. enterica* genome acquired since speciation? Some of the OmpR-regulated loci are essential for virulence and these data extend the OmpR regulon to include hypothetical genes, which are not present in *E. coli.* Two novel regions were identified and characterised based on sequence identity to previously described genes. The virulence potential of these genes was assessed using an acute infection murine infection model by competitive challenges but there was no significant difference detected. The gene knockouts were also characterised *in vitro* and growth of the mutant RAK103 was attenuated when the sole carbon source was *N*-acetyl muramic acid, a component of the peptidoglycan cell wall [270]. Persistence is also an important facet of *Salmonella* pathogenicity and mutants in these loci should be analysed in their ability to persist in the murine model for Typhoid fever.

This study encompasses methods developed for exploiting new technology in a novel manner to assess the transcriptome of bacteria. Interrogating a biologically conserved and virulence-essential regulator, the *ompR* gene, has supported the validity of these methods. The regulon was further defined by identifying OmpR bound sequences

using Illumina sequencing technology. Furthermore, newly defined OmpR regulated loci were characterised *in vivo* and *in vitro*. This body of work aims to increase the overall knowledge of the *S.* Typhi organism as a model and to develop methods suitable for interrogating both prokaryotes and eukaryotes and providing a more focussed approach to gene characterisation.