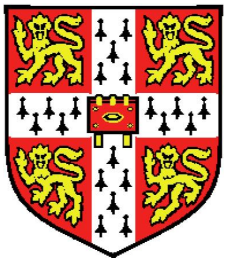
High-throughput Experimental and Computational Studies of Bacterial Evolution



Lars Barquist

Queens' College

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

XX August 2013

Declaration

High-throughput Experimental and Computational Studies of Bacterial Evolution

The work presented in this dissertation was carried out at the Wellcome Trust Sanger Institute between October 2009 and August 2013. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. This thesis does not exceed the limit of 60,000 words as specified by the Faculty of Biology Degree Committee. This thesis has been typeset in 12pt Computer Modern font using LaTeX according to the specifications defined by the Board of Graduate Studies and the Faculty of Biology Degree Committee.

Contents

List of Figures

List of Tables

Chapter 1

Querying bacterial genomes with transposon-insertion sequencing

This chapter is an expansion of the previously published article "Approaches to querying bacterial genomes using transposon-insertion sequencing" (Barquist, Boinett, and Cain, 2013). Amy K. Cain and Christine J. Boinett (Pathogen Genomics, Wellcome Trust Sanger Institute) contributed to the research of the original article. All final language is my own.

1.1 Introduction

A common approach to identifying genomic regions involved in survival under a particular set of conditions is to screen large pools of mutants simultaneously. This can be done with defined mutants(1,2); however, the construction of defined mutant libraries is labor-intensive and requires accurate genomic annotation, which can be particularly difficult to define for non-coding regions. An alternative to defined libraries is the construction and analysis of random transposon-insertion libraries. The original application of this method used DNA hybridization to track uniquely tagged transposon-insertions in Salmonella enterica serovar Typhimurium over the course of BALB/c mouse infection(3). DNA hybridization was eventually superseded by methods that used microarray detection of the genomic DNA flanking insertion sites, variously known as TraSH, MATT, and DeADMAn (reviewed in (4)). However, these methods suffered from many of the problems

microarrays generally suffer from: difficulty detecting low-abundance transcripts, mis-hybridization, probe saturation, and difficulty identifying insertion sites precisely.The application of high-throughput sequencing to the challenge of determining insertion location and prevalence solves many of these problems. Interestingly, the first application of transposon-insertion sequencing, developed by Hutchison et al., actually predates the development of microarray-based methods(5). However, this was applied to libraries of only approximately 1000 transposon mutants in highly reduced Mycoplasma genomes, and the difficulty of sequencing at the time prevented wide spread adoption or high resolution. Modern high-throughput sequencing technology allows the methods discussed in this review to routinely monitor as many as one million mutants simultaneously in virtually any genetically tractable microorganism.

1.2 Protocols

Several methods were developed concurrently for high-throughput sequencing of transposon-insertion sites: TraDIS(6), INSeq(7), HITS(8), and Tn-seq(9) followed by Tn-seq Circle(10) and refinements to the INSeq protocol(11). All of these protocols follow the same basic workflow with minor variations (see Figure 1; Table 1): transposon mutagenesis and construction of pools of single insertion mutants; enrichment of transposon-insertion junctions; and finally, in some protocols a purification step either precedes or follows PCR enrichment before sequencing.

1.2.1 Transposon mutagenesis

Most studies have used either Tn5 or mariner transposon derivatives. Tn5 originated as a bacterial transposon which has been adapted for laboratory use. Large-scale studies have shown that Tn5, while not showing any strong preference for regional GC-content, do have a weak preference for a particular insertion motif(12-14). Transposon-insertion sequencing studies performed with Tn5 transposons in S. enterica serovars have reported a slight bias towards AT-rich sequence regions(6,15). However, this preference does not appear to be a major obstacle to analysis given the extremely high insertion densities obtained with this transposon(6,15,16) (see Table 1). Additionally, Tn5 has been shown to be active in a wide range of bacterial species, though the number of transformants obtained can

vary significantly depending on the transformation efficiency of the host.Mariner/Himar1 transposons on the other hand originate from eukaryotic hosts and have an absolute requirement for TA bases at their integration site(17,18), with no other known bias besides a possible preference for bent DNA(17). This can be a disadvantage in that it limits the number of potential insertion sites, particularly in GC-rich sequence. However, this specificity can also be used in the prediction of gene essentiality in near-saturated libraries: as every potential integration site is known and the probability of integration at any particular site can be assumed to be roughly equal, it is straight-forward to calculate the probability that any particular region lacks insertions by chance. Himar1 transposition can also be conducted in vitro in the absence of any host factors(19), and inserted transposons can then be transferred to the genomes of naturally transformable bacteria through homologous recombination(20). This can be advantageous when working with naturally transformable bacteria with poor electroporation efficiency(8,9). It is worth noting that Tn5 is also capable of transposition in vitro(21), and could potentially be used to increase insertion density and hence the resolution of the assay, particularly in GC-rich genomic regions.

1.2.2 Pool construction

Once mutants have been constructed, they are plated on an appropriate selective media for the transposon chosen, and colonies are counted, picked, and pooled. A disadvantage of this is that the mutants must be recreated for follow up or validation studies. Goodman et al. introduced a clever way around this in the INSeq protocol: by individually archiving mutants, then sequencing combinatorial mutant pools it is possible to uniquely characterize 2n insertion mutants by sequencing only n pools(7). Each mutant is labelled with a unique binary string that indicates which pools it has been added to. These binary strings can then be reconstructed for each insertion observed in these pools by recording their presence or absence in sequencing data, providing a unique pattern relating insertions to archived mutants. The authors control false identifications due to errors in sequencing by requiring that each binary label have a minimum edit distance to every other label, allowing for a robust association of labels with insertions despite sometimes noisy sequencing data. As a proof of concept, the authors were able to identify over 7,000 Bacteroides thetaiotaomicron mutants from only 24 sequenced pools. This effectively

uses methods for the generation of random transposon pools to rapidly generate defined mutant arrays, though it is heavily dependent on liquid-handling robotics.

1.2.3 Enrichment of transposon-insertion junctions

Once pools have been constructed they are grown in either selective or permissive conditions, depending on the experiment, and then genomic DNA is extracted. Fragmentation proceeds either through restriction digestion in the case of transposons modified to contain appropriate sites(7,9,10) or via physical shearing(6,8), then sequencing adapters are ligated to the resulting fragments. PCR is performed on these fragments using a transposon-specific primer and a sequencing adapter-specific primer to enrich for fragments spanning the transposon-genomic DNA junction. Some protocols purify fragments containing transposon insertions using biotinylated primers(10,11) or PAGE(7) before and/or after PCR enrichment. The purification step from the Tn-seq Circle protocol is particularly unusual in that restriction digested fragments containing transposon sequence are circularized before being treated with an exonuclease that digests all fragments without transposon insertions, theoretically completely eliminating background(10). Given the success of protocols that do not include a purification step and the lack of systematic comparisons, it is currently unclear whether including one provides any major advantages.

1.3 Reproducibility, accuracy, and concordance with previous methods

A number of studies have looked at the reproducibility of transposon-insertion sequencing. Multiple studies using different protocol variations have repeatedly shown extremely high reproducibility in the number of insertions per gene (correlations of 90%) in replicates of the same library grown and sequenced independently(7,9,10), and good reproducibility (correlations between 70-90%) in independently constructed non-saturated libraries(9,22). Van Opijnen and Camilli(22) compared traditional 1 X 1 competition experiments between wild-type and mutant Streptococcus pneumoniae to results obtained by transposon-insertion sequencing and showed that there was no significant difference in results over a range of tested conditions.The accuracy of transposon-insertion sequencing in determining library composition has also been assessed. Zhang et al. constructed a library of identified

transposon-insertion mutants in known relative quantities, and then were able to recover the relative mutant prevalence with transposon-insertion sequencing(23). Additionally, by estimating the number of PCR templates prior to enrichment, this study showed that there is a high correlation between enrichment input and sequencing output. Two studies have evaluated concordance between results obtained with transposon-insertion sequencing and microarray monitoring of transposon insertions in order to demonstrate the enhanced accuracy and dynamic range of sequencing over previous methods. In the first, 19 libraries of 95 enterohemorrhagic Escherichia coli (EHEC) transposon mutants that had previously been screened in cattle using signature-tagged mutagenesis (STM) were pooled and re-evaluated using the TraDIS protocol(24). The original STM study had identified 13 insertions in 11 genes attenuating intestinal colonization in a type III secretion system located in the locus of enterocyte effacement (LEE)(25). By applying sequencing to the same samples, an additional 41 mutations in the LEE were identified, spanning a total of 21 genes. Additional loci outside the LEE which have been previously implicated in intestinal colonization but had not been detected by STM were also reported by TraDIS.The second study re-evaluated genes required for optimal growth determined by TraSH in Mycobacterium tuberculosis(26,27). The greater dynamic range of sequencing as compared to microarrays allowed easier discrimination between insertions that were nonviable and those that were only significantly underrepresented. The authors estimate that genes called as required by sequencing in their study are at least 100-fold underrepresented in the pool. In comparison, the threshold in the previous microarray experiment reported genes that had log probe ratios at least 5-fold lower than average between transposon-flanking DNA hybridization and whole genomic DNA hybridization. Additionally, the nucleotide-resolution of insertion sequencing allowed the authors to identify genes which had required regions, likely corresponding to required protein domains(23), but which tolerated insertions in other regions. Altogether the authors increase the set of genes predicted to be required for growth in laboratory conditions in M. tuberculosis by more than 25% (from 614 to 774).

1.4 Gene requirements

The earliest application of transposon-insertion sequencing was to determine the minimal set of genes necessary for the survival of Mycoplasma(5). This essential genome is of

great interest to synthetic and systems biology where it is seen as a foundation for engineering cell metabolism, and in infection biology and medicine where it is seen as a promising target for therapies. However, it is important to remember that essentiality is always relative to growth conditions: a biosynthetic gene that is non-essential in a growth medium supplying a particular nutrient may become essential in a medium that lacks it. Traditionally, gene essentiality has been determined in clonal populations(1,28,29); since the high-throughput transposon sequencing protocols described here necessarily contain a short period of competitive growth before DNA extraction, many of these studies prefer to refer to the required genome for the particular conditions under evaluation.Because of this short period of competitive growth, and because many otherwise required genes tolerate insertions in their terminus(7,27,30) or outside essential domains(23) the determination of required genomic regions is not completely straight-forward and a number of approaches have been taken to counter this. These include only calling genes completely lacking insertions as required(9), determining a cut-off based on the empirical or theoretical distribution of gene-wise insertion densities(6,15,27,30). Additionally, windowed methods have been developed which can be used to identify essential regions in the absence of gene annotation(23,31), and have had success in identifying required protein domains, promoter regions, and non-coding RNAs (ncRNAs). The organisms that have been evaluated for gene requirements under standard laboratory conditions are summarized in Table 1.In agreement with previous studies(1,28), many required genes identified by transposon-insertion sequencing are involved in fundamental biological processes such as cell division, DNA replication, transcription and translation(6,7,15,27), and many of these requirements appear to be conserved between genera and classes(15,16). However, a recent study defining required gene sets in Salmonella serovars has found that phage repressors, necessary for maintaining the lysogenic state of the prophage, are also required(15), even though mobile genetic elements such as phage are usually considered part of the accessory genome. This study also highlights the need for temperance when interpreting the results of high-throughput assays of gene requirements. For example, many genes in Salmonella Pathogenicity Island 2 (SPI-2) did not exhibit transposon-insertions, despite clear evidence from directed knockouts showing that these genes are non-essential for viability or growth. Under laboratory conditions, SPI-2 is silenced by the nucleoid-forming protein H-NS(32,33), which acts by oligermerizing along silenced regions of DNA blocking RNA polymerase access. A previous study has shown that transposon

insertion cold spots can be caused by competition between high-density proteins and transposases for DNA(34). This suggests that H-NS may be restricting transposase access to DNA, though this has not previously been observed in transposon-insertion sequencing data, and will require additional work to confirm.

1.5 Defining conditional gene requirements

One of the most valuable applications of the transposon-insertion sequencing method is the ability to identify genes important in a condition of interest, by comparing differences in the numbers of sequencing reads from input (control) mutant pools to output (test) pools that have been subject to passaging in a certain growth condition. Insertion counts are compared from cells in the input pool and those after passage, thereby indentifying genes that either enhance or detract from survival and/or growth in the given condition, defined by decreased or increased insertion frequency, respectively. A further application of this method involves comparing insertions between biologically linked conditions, such as cellular stresses or different stages of a murine infection, to gain insight into complex systems(22).So far, transposon-insertion sequencing has been used to investigate a number of interesting biological questions: bile tolerance in S. Typhi(6) and S. Typhimurium(35), bacteriophage infection of S. Typhi(36), antibiotic resistance in Pseudomonas aeruginosa (10), cholesterol utilisation in M. tuberculosis(27) and a number of stress and nutrient conditions in S. pneumoniae(22). Transposon-insertion sequencing of populations passed through murine models have been used to assess genes required to establish the gut commensal B. thetaiotaomicron in its niche(7), for Haemophilus influenzae infection(8), as well as S. pneumoniae responses to two in vivo niches - the lung and nasopharynx(22). A further extension of the method examined double mutant libraries, that is transposon mutant libraries generated in a defined deletion background, to tease apart complex networks of regulatory genes(9).Two studies in particular illustrate the power of using transposon-insertion sequencing to identify conditionally required genes. In the first, Goodman et al. set out to determine the genes necessary for the establishment of the commensal B. thetaiotaomicron in a murine model(7). First, the growth requirements of transposon mutant populations in the cecum of germ-free mice was assessed, and genes required for growth in monoassociation with the host were found to be enriched in functions such as energy production and amino acid metabolism. By

further comparing monoassociated transposon mutant libraries with those grown in the presence of three defined communities of human gut-associated bacteria, the authors identified a locus up-regulated by low levels of vitamin B12 that is only required in the absence of other bacteria capable of synthesizing B12. This showed that the gene requirements of any particular bacterium in the gut are at least partially dependent on the metabolic capabilities of the entire community and emphasizes the importance of testing in vivo conditions to complement in vitro study.The second study, conducted by van Opijnen and Camilli, aimed to map the genetic networks involved in a range of cellular stress responses in S. pneumoniae(22). Seventeen in vitro conditions were tested, including: pH, nutrient limitation, temperature, antibiotic, heavy metal, and hydrogen peroxide stress. Approximately 6% of disrupted genes resulted in increased fitness in some condition, suggesting that some genes are maintained despite being detrimental to the organism under particular conditions. These would be interesting candidates for further functional and evolutionary study, as the maintenance of these genes is presumably highly dependent on the conditions the bacteria faces, and may have implications for our understanding of e.g. gene loss in the process of bacterial host adaptation(37). Two additional in vivo experiments were performed in a murine model, where cells were recovered from the lung and nasopharynx. Combining this data, over 1,800 genotype-phenotype genetic interactions were identified. These interactions were mapped and pathways identified. Between the two in vivo niches, certain stress responses pathways were markedly different. For example, temperature stress produced a distinct response in the lung, compared to the nasopharyanx, which is perhaps to be expected as temperature varies greatly between these two sites. By further examining sub-pathways required in the two different niches and comparing them to in vitro requirements, the authors were able to draw conclusions regarding the condition S. pneumoniae faces when establishing an infection. This comprehensive mapping of genotype-phenotype relationships will serve as an important atlas for further studies.

1.6 Monitoring ncRNA contributions to fitness

To date, four studies have used transposon-insertion sequencing to examine the contribution of non-coding RNAs (ncRNAs) and other non-coding regions to organismal fitness (see Table 1). Two of these examined requirements for non-coding regions in the relatively

under-explored bacterial species Caulobacter crescentus(16) and M. tuberculosis(23). Both utilized analytical techniques that allowed for the identification of putative required regions in the absence of genome annotation. Twenty-seven small RNAs (sRNAs) had previously been detected in C. crescentus(38); 6 were found to be depleted in transposon insertions indicating an important role in basic cellular processes. Additionally, the well-characterized ncRNAs tmRNA and RNaseP, as well as 29 non-redundant tRNAs were found to be required. An additional 90 unannotated non-disruptable regions were identified throughout the genome, implying an abundance of unexplored functional non-coding sequence. While the non-coding transcripts of M. tuberculosis have been explored more thoroughly than those of C. crescentus, most remain functionally uncharacterized, though there are hints that some of these may be involved in pathogenicity(39). Using a mariner transposon-based assay and a windowed statistical analysis that accounted for the distribution of potential TA integration sites, 35 intergenic regions were identified as putatively required in the M. tuberculosis genome(23). In common with the C. crescentus study, the RNA component of RNase P, required for the maturation of tRNAs, and tmRNA, involved in the freeing of stalled ribosomes, were identified as required (Figure 2 A) together with 10 non-redundant tRNAs and potential promoter regions. However, due to the lower overall insertion density and lack of TA sites in some GC-rich regions, there were some regions that could not be assayed and the resolution was limited to 250 bases.A recent study has examined ncRNA requirements in the S. enterica serovars Typhi and Typhimurium(15). Using the tRNAs as a model set of ncRNAs, this study showed that the high transposon insertion density achieved by the TraDIS protocol is capable of assaying the requirement for genomic regions as small as 70 to 80 bases. S. enterica, together with the closely related E. coli, has served as a model organism for the discovery and elucidation of ncRNA function, and extensive annotations of non-coding transcripts are available(40-44). As a result this study was able to assay approximately 300 non-coding regions with evidence for function or transcription. Among the ncRNAs identified as required were RNase P; the RNA component of the signal recognition particle, involved in targeting proteins to the plasma membrane; and a number of known autoregulatory ribosomal protein leader sequences(45), as well as providing evidence for a novel leader sequence, StyR-8(43), that appears to be involved in the autoregulation of the rpmB gene. In total, this study identified 15 confirmed and putative ncRNAs required for robust competitive growth on rich media in both serovars, including a number of known

sRNAs involved in stress response.A particularly exciting study has been conducted in S. pneumoniae TIGR4 combining RNA-seq with transposon-insertion sequencing(46). To identify sRNA loci the authors first sequenced size-select RNA from the wild type and three two-component system knockouts, identifying 89 putative sRNAs, 56 of which were novel. Fifteen of these candidates, selected on the basis of high expression and low predicted folding free energy, were assayed for their ability to establish invasive disease in a murine model. Of these 8 sRNA deletions showed a significant attenuation of disease. To more broadly establish the roles of sRNAs in infecting particular organs, transposon insertion libraries were administered directly to the nasopharnyx, lungs, or blood of mice, and bacteria were harvested following disease progression. Twenty-six, 28, and 18 sRNAs were found to attenuate infection in the nasopharnyx, lung and blood respectively. These results were then validated with targeted deletions of 11 sRNAs (Figure 2 B). In addition to establishing the role of sRNAs in S. pneumoniae virulence, this study illustrated the power of combining RNA-seq and transposon-insertion sequencing to rapidly assign phenotypes to non-coding sequences.

1.7 Limitations

In this review, we have largely focused on the potential of transposon insertion sequencing. However, this technology does have a number of important limitations, which we collect here and summarize in Table 2. As discussed previously, requirements for particular nucleotides at insertion sites, such as the TA required by the Mariner transposon, or preference for certain sequence composition, such as the AT bias exhibited by Tn5, can limit the density of observed insertions in certain genomic regions. This may impact any down-stream analysis, and can potentially bias results, particularly the determination of gene requirements. Even if this bias has been accounted for, transposon-insertion screens will always over-predict gene requirements in comparison to targeted deletion libraries as discussed previously. However, this over-prediction can be controlled either through careful consideration of known insertion biases as in many Mariner-based studies, or by high insertion densities, such as those achieved in several Tn5-based studies (Table 1). Once the library has been created, only regions that have accumulated insertions in the conditions of library creation will be able to be assayed for fitness effects in further conditions. This means that regions that lead to slow growth phenotypes when

disrupted in standard laboratory conditions may be difficult to assay in other conditions. Additionally, the dynamic range of fitness effects detected will depend on the complexity of the input library(s). The absence of insertions may be a particular problem for assaying small genomic elements, such as sRNAs or short ORFs. Finally, the validation of hypotheses derived from transposon-insertion sequencing will require the construction of targeted deletions, as individual mutants cannot be recovered from pools unless specialized protocols have been followed during library construction (as in Goodman, 2009(7)).

1.8 The future of transposon-insertion sequencing

Transposon-insertion sequencing is a robust and powerful technique for the rapid connection of genotype to phenotype in a wide range of bacterial species. Already, a number of studies have demonstrated the effectiveness of this method and the results have been far-reaching: enhancing our understanding of basic gene functions, establishing requirements for colonization and infection, mapping complex metabolic pathways, and exploring non-coding genomic dark matter. Due to the range of potential applications of transposon-insertion sequencing, along with the decreasing cost and growing accessibility of next-generation sequencing, we believe that this method will become increasingly common in the near future. A number of bacterial species have already been subjected to transposon-insertion sequencing (Table 1). Microarray-based approaches to monitoring transposon mutant libraries have even been applied to eukaryotic systems(47), and similarly transposon-insertion sequencing can potentially be applied to any system where the creation of large-scale transposon mutant libraries is technologically feasible. Recently the Genomic Encyclopedia of Bacteria and Archea (GEBA)(48) has been expanding our knowledge of bacterial diversity through targeted genomic sequencing of underexplored branches of the tree of life. Applying transposon-insertion sequencing in a comparative manner(15) across the bacterial phylogeny will provide an unprecedented view of the determinants for survival in diverse environments. While most transposon-insertion sequencing studies to date have focused on pathogenic bacteria, these techniques could also have applications in energy production, bioremediation, and synthetic biology.The combination of transposon-insertion sequencing with other high-throughput and computational methods is already proving to be fertile ground for enhancing our understanding of

bacterial systems. For instance, by using transposon-insertion sequencing in a collection of relatively simple conditions combined with a computational pathway analysis, van Opijnen and Camilli were able to provide a holistic understanding of the genetic subsystems involved in a complex process such as S. pneumoniae pathogenesis(22). In the future, methods to assay phenotype in a high-throughput manner(49,50) may be combined with transposon-insertion sequencing to provide exhaustive simple genotype-phenotype associations with which to understand complex processes in a systems biology framework. We look forward to the adoption of these data sets by the community as an important tool for rapid hypothesis generation.

Chapter 2

A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium

This chapter is a modified version of the previously published article "A comparison of dense transposon insertion libraries in the Salmonella serovars Typhi and Typhimurium" (Barquist, Langridge, et al., 2013). This work is a result of collaboration with Gemma C. Langridge (Pathogen Genomics, Wellcome Trust Sanger Institute), who constructed the transposon mutant library and contributed sections to a draft manuscript.

2.1 Introduction

Salmonella enterica subspecies enterica serovars Typhi (S. Typhi) and Typhimurium (S. Typhimurium) are important human pathogens with distinctly different lifestyles. S. Typhi is host-restricted to humans and causes typhoid fever. This potentially fatal systemic illness affects at least 21 million people annually, primarily in developing countries (1-3) and is capable of colonizing the gall bladder creating asymptomatic carriers; such individuals are the primary source of this human restricted infection, exemplified by the case of Typhoid Mary (4). S. Typhimurium, conversely, is a generalist, infecting a wide range of mammals and birds in addition to being a leading cause of foodborne gastroenteritis in human populations. Control of S. Typhimurium infection in livestock

destined for the human food chain is of great economic importance, particularly in swine and cattle (5,6). Additionally, S. Typhimurium causes an invasive disease in mice, which has been used extensively as a model for pathogenicity in general and human typhoid fever specifically (7).

Despite this long history of investigation, the genomic factors that contribute to these differences in lifestyle remain unclear. Over 85% of predicted coding sequences are conserved between the two serovars in sequenced genomes of multiple strains (8-11). The horizontal acquisition of both plasmids and pathogenicity islands during the evolution of the salmonellae is believed to have impacted upon their disease potential. A 100kb plasmid, encoding the spv (Salmonella plasmid virulence) genes, is found in some S. Typhimurium strains and contributes significantly towards systemic infection in animal models (12,13). S. Typhi is known to have harbored IncHI1 plasmids conferring antibiotic resistance since the 1970s (14), and there is evidence that these strains present a higher bacterial load in the blood during human infection (15). Similar plasmids have been isolated from S. Typhimurium (16-18). Salmonella pathogenicity islands (SPI)-1 and -2 are common to both serovars, and are required for invasion of epithelial cells (reviewed in (19)) and survival inside macrophages respectively (20,21). S. Typhi additionally incorporates SPI-7 and SPI-10, which contain the Vi surface antigen and a number of other putative virulence factors (22-24).

Acquisition of virulence determinants is not the sole explanation for the differing disease phenotypes displayed in humans by S. Typhimurium and S. Typhi; genome degradation is an important feature of the S. Typhi genome, in common with other host-restricted serovars such as S. Paratyphi A (humans) and S. Gallinarum (chickens). In each of these serovars, pseudogenes account for 4-7% of the genome (9,25-27). Loss of function has occurred in a number of S. Typhi genes that have been shown to encode intestinal colonisation and persistence determinants in S. Typhimurium (28). Numerous sugar transport and degradation pathways have also been interrupted (9), but remain intact in S. Typhimurium, potentially underlying the restricted host niche occupied by S. Typhi.

In addition to its history as a model organisms for pathogenicity, S. Typhimurium has recently served as a model organism for the elucidation of non-coding RNA (ncRNA) function (29). These include cis-acting switches, such as RNA-based temperature and magnesium ion sensors (30,31), together with a host of predicted metabolite-sensing

riboswitches. Additionally, a large number of trans-acting small RNAs (sRNAs) have been identified within the S. Typhimurium genome (32), some with known roles in virulence (33). These sRNAs generally control a regulon of mRNA transcripts through an antisense binding mechanism mediated by the protein Hfq in response to stress. The functions of these molecules have generally been explored in either S. Typhimurium or E. coli, and it is unknown how stable these functions and regulons are over evolutionary time (34).

Transposon mutagenesis has previously been used to assess the requirement of particular genes for cellular viability. The advent of next-generation sequencing has allowed simultaneous identification of all transposon insertion sites within libraries of up to 1 million independent mutants (35-38), enabling us to answer the basic question of which genes are required for in vitro growth with extremely fine resolution. By using transposon mutant libraries of this density, which in S. Typhi represents on average ¿ 80 unique insertions per gene (35), shorter regions of the genome can be interrogated, including ncRNAs (38). In addition, once these libraries exist, they can be screened through various selective conditions to further reveal which functions are required for growth/survival.

Using Illumina-based transposon directed insertion-site sequencing (TraDIS (35)) with large mutant libraries of both S. Typhimurium and S. Typhi, we investigated whether these Salmonellae require the same protein-coding and non-coding RNA (ncRNA) gene sets for competitive growth under laboratory conditions, and whether there are differences which reflect intrinsic differences in the pathogenic niches these bacteria inhabit.

2.2 Methods

2.2.1 Strains

S. Typhimurium strain SL3261 contains a deletion relative to the parent strain, SL1344, was used to generate the large transposon mutant library. The 2166bp deletion ranges from 153bp within aroA (normally 1284bp) to the last 42bp of cmk, forming two pseudogenes and deleting the intervening gene SL0916 completely. For comparison, we utilized our previously generated S. Typhi Ty2 transposon library (35).

2.2.2 Annotation

For S. Typhimurium strain SL3261, we used feature annotations drawn from the SL1344 genome (EMBL-Bank accession FQ312003.1), ignoring the deleted aroA, ycaL,and cmk genes. We re-analyzed our S. Typhi Ty2 transposon library with features drawn from an updated genome annotation (EMBL-Bank accession AE014613.1.) We supplemented the EMBL-Bank annotations with non-coding RNA annotations drawn from Rfam 10.1 (39), Sittka et al. (40), Chinni et al. (41), Raghavan et al. (42), and Krger et al. (32). Selected protein-coding gene annotations were supplemented using the HMMER webserver (43) and Pfam (44).

2.2.3 Creation of S. Typhimurium transposon mutant library

S. Typhimurium was mutagenized using a Tn5-derived transposon as described previously (35). Briefly, the transposon was combined with the EZ-Tn5 transposase (Epicenter, Madison, USA) and electroporated into S. Typhimurium. Transformants were selected by plating on LB agar containing 15 ?g/mL kanamycin and harvested directly from the plates following overnight incubation. A typical electroporation experiment generated a batch of between 50,000 and 150,000 individual mutants. 10 batches were pooled together to create a mutant library comprising approximately 930,000 transposon mutants.

2.2.4 DNA manipulations and sequencing

Genomic DNA was extracted from the library pool samples using tip-100g columns and the genomic DNA buffer set from Qiagen (Crawley, UK). DNA was prepared for nucleotide sequencing as described previously (35). Prior to sequencing, a 22 cycle PCR was performed as previously described (35). Sequencing took place on a single end Illumina flowcell using an Illumina GAII sequencer, for 36 cycles of sequencing, using a custom sequencing primer and 2x Hybridization Buffer (35). The custom primer was designed such that the first 10 bp of each read was transposon sequence.

2.2.5 Sequence analysis

The Illumina FASTQ sequence files were parsed for 100% identity to the 5 10bp of the transposon (TAAGAGACAG). Sequence reads which matched were stripped of the

transposon tag and subsequently mapped to the S. Typhimurium SL1344 or S. Typhi Ty2 chromosomes using Maq version maq-0.6.8 (45). Approximately 12 million sequence reads were generated from the sequencing run which used two lanes on the Illumina flowcell. 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 appropriate genome was then determined.

2.2.6 Statistical analysis of required genes

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. As before (35) the distribution of insertion indices was bimodal, corresponding to the required (mode at 0) and non-required models. We fitted gamma distributions for the two modes using the R MASS library (http://www.r-project.org). Log2-likelihood ratios (LLR) were calculated between the required and non-required models and we called a gene required if it had an LLR of less than -2, indicating it was at least 4 times more likely according to the required model than the non-required model. Non-required genes were assigned for an LLR of greater than 2. Genes falling between the two thresholds were considered ambiguous for the purpose of this analysis. This procedure lead to genes being called as required in S. Typhimurium when their insertion index was less than 0.020, and ambiguous between 0.020 and 0.027. The equivalent cut-offs for the S. Typhi library are 0.0147 and 0.0186, respectively.

We calculated a p-value for the observed number of insertion sites per gene using a Poisson approximation with rate R = N/G where N is the number of unique insert sites (549,086) and G is the number of bases in the genome (4,878,012). The P-value for at least X consecutive bases without an insert site is e(-RX), giving a 5% cut-off at 27 bp and a 1% cut-off at 41 bp.

For every gene g with ng,A reads observed in S. Typhi and ng,B reads observed in S. Typhimurium, we calculated the log2 fold change ratio Sg,A,B = log2 ((ng,A+100)/(ng,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 SA,B, and calculated p-values for each gene according

to the fit. We considered genes with a P-value of 0.05 or less under the normal model to be uniquely required by one serovar.

2.3 Results and Discussion

2.3.1 TraDIS assay of every Salmonella Typhimurium protein-coding gene

Approximately 930,000 mutants of S. Typhimurium were generated using a Tn5-derived transposon. 549,086 unique insertion sites were recovered from the mutant library using short-read sequencing with transposon-specific primers. This is a substantially higher density than the 371,775 insertions recovered from S. Typhi previously (35). The S. Typhimurium library contains an average of one insertion every 9bp, or over 100 unique inserts per gene (Figure 1). The large number of unique insertion sites allowed every gene to be assayed; assuming random insertion across the genome, a region of 41bp without an insertion was statistically significant (P ¡ 0.01). As previously noted in S. Typhi, the distribution of length-normalized insertions per gene is bimodal (see supplementary figure 1), with one mode at 0. We interpret genes falling in to the distribution around this mode as being required for competitive growth within a mixed population under laboratory conditions (hereafter required). Of these, 57 contained no insertions whatsoever and were mostly involved in core cellular processes (see Table 1, Supplementary Dataset).

There was a 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. We also observed a bias for transposon insertions in A+T rich regions, as was previously observed in the construction of an S. Typhi mutant library (35). However, the insertion density achieved is sufficient to discriminate between required and non-required genes easily. As was first seen in S. Typhi (35), we observed transposon insertions into genes upstream of required genes in the same operon, suggesting that most insertions do not have polar effects leading to the inactivation of downstream genes.

Analysis of the S. Typhimurium mutant library allowed us to identify 353 coding

sequences required for growth under laboratory conditions, and 4,112 non-required coding sequences (see Supplementary Dataset). We were unable to assign 65 genes to either the required or non-required category. 60 of these genes, which we will refer to as ambiguous, had log-likelihood ratios (LLRs) between -2 and 2. The final 5 unassigned genes had lengths less than 60 bases, and they were removed from the analysis. All other genes contained enough insertions or were of sufficient length to generate credible LLR scores. Thus, every gene was assayed and we were able to draw conclusions for 98.7

2.3.2 Cross-species comparison of genes required for growth

Gene essentiality has previously been assayed in Salmonella using insertion-duplication mutagenesis (46). Knuth et al. estimated 490 genes are essential to growth in clonal populations, though 36 of these have subsequently been successfully deleted (47). While TraDIS assays gene requirements after a brief period of competitive growth on rich media, we identify a smaller required set than Knuth et al. of approximately 350 genes in each serovar, closer to current estimates of approximately 300 essential genes in E. coli (48).

To demonstrate that TraDIS does identify genes known to have strong effects on growth, as well as to test our predictive power for determining gene essentiality, we compared our required gene sets in S. Typhimurium and S. Typhi to essential genes determined by systematic single-gene knockouts in the Escherichia coli K-12 Keio collection(48). We identified orthologous genes in the three data sets by best reciprocal FASTA hits exhibiting over 30% sequence identity for the amino acid sequences. Required orthologous genes identified in this manner share a significantly higher average percent sequence identity with their E. coli counterparts than expected for a random set of orthologs, at 94% identity as compared to 87% for all orthologous genes. In 100,000 randomly chosen gene sets of the same size as our required set we did not find a single set where the average shared identity exceeded 90%, indicating that required genes identified by TraDIS are more highly conserved at the nucleotide level than other orthologous protein coding sequences.

Baba et al.(48) have defined an essentiality score for each gene in E. coli based on evidence from four experimental techniques for determining gene essentiality: targeted knock-outs using ?-red mediated homologous recombination(48), genetic footprinting (49,50), large-scale chromosomal deletions (51), and transposon mutagenesis (52). Scores

range from -4 to 3, with negative scores indicating evidence for non-essentiality and positive scores indicating evidence for essentiality. Comparing the overlap between essential gene sets in E. coli, S. Typhi, and S. Typhimurium, we find a set of 228 E. coli genes which have a Keio essentiality score of at least 0.5 (i.e. there is evidence for gene essentiality; See Figure 2.) that have TraDIS-predicted required orthologs in both S. Typhi and S. Typhimurium, constituting 85% of E. coli genes with evidence for essentiality indicating that gene requirements are largely conserved between these genera. Including orthologous genes that are only predicted to be essential by TraDIS in S. Typhi or S. Typhimurium raises this figure to nearly 93%. The majority of shared required 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 (Table 1). A recent study in the alphaproteobacteria Caulobacter crescentus reported 210 shared essential genes with E. coli, despite C. crescentus sharing less than a third as many orthologous genes with E. coli as Salmonella serovars (38). This suggests the existence of a shared core of approximately 200 essential proteobacterial genes, with the comparatively rapid turnover of 150 to 250 non-core lineage-specific essential genes.

If we make the simplistic assumption that gene essentiality should be conserved between E. coli and Salmonella, we can use the overlap of our predictions with the Keio essential genes to provide an estimate of our TraDIS libraries accuracy for predicting that a gene will be required in a clonal population. Of the 2632 orthologous E. coli genes which have a Keio essentiality score of less than -0.5 (i.e. there is evidence for gene non-essentiality), only 33 are predicted to be required by TraDIS in both Salmonella serovars. S. Typhi contains the largest number of genes predicted by TraDIS to be required with E. coli orthologs with negative Keio essentiality scores. However, even if we assume these are all incorrect predictions of gene essentiality, this still gives a gene-wise false positive rate (FPR) of 2.7% (81 out of 2981 orthologs) and a positive predictive value (PPV) of 75% (247 with essentiality scores greater than or equal to 0.5 out of 328 predictions with some Keio essentiality score.) Under these same criteria the S. Typhimurium data set has a lower gene-wise FPR of 1.6% (51 out of 3122 orthologs) and a higher PPV of 82% (234 out of 285 predictions as before), as we would expect given the librarys higher insertion density. In reality these FPRs and PPVs are only estimates; genes which are not essential in E. coli may become essential in the different

genomic context of Salmonella serovars and vice versa, particularly in the case of S. Typhi where wide-spread pseudogene formation has eliminated potentially redundant pathways (26,27). Additionally, TraDIS will naturally over-predict essentiality in comparison to targeted knockouts, as our library creation protocol necessarily contains a short period of competitive growth between mutants during the recovery from electro-transformation and selection. As a consequence, genes which cause major growth defects, but not necessarily a complete lack of viability in clonal populations, may be reported as required.

2.3.3 Serovar-specific genes required for growth

Many of the required genes present in only one serovar encoded phage repressors, for instance the cI proteins of Fels-2/SopE and ST35 (see Supplementary Tables 2 and 3). Repressors maintain the lysogenic state of prophage, preventing transcription of early lytic genes (53). 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 broadens the definition of required genes; such repressors may not be required for cellular viability in the traditional sense, but once present in these particular genomes, their maintenance is required for continued viability, as long as the rest of the phage remains intact.

S. Typhimurium and S. Typhi both contains 8 apparent large phage-derived genomic regions (54,55). We were able to identify required repressors in all the intact lambdoid, P2-like, and P22-like prophage in both genomes, including Gifsy-1, Gifsy-2, and Fels-2/SopE (see supplementary tables 2 and 3). With the exception of the SLP203 P22-like prophage in S. Typhimurium, all of these repressors lack the peptidase domain of the classical lambda repressor gene cI. This implies that the default anti-repression mechanism of Salmonella prophage may be more similar to a trans-acting mechanism recently discovered in Gifsy phage (56) than to the phage lambda repressors RecA-induced self-cleavage mechanism. We are also able to confirm that most phage remnants and fusions contained no active repressors, with the exception of the SLP281 degenerate P2-like prophage in S. Typhimurium. This degenerate prophage contains both intact replication and integration genes, but appears to lack tail and head proteins, suggesting it may depend on another phage for production of viral particles. Both genomes also encode P4-like satellite prophage, which rely on helper phage for lytic functions and utilize a complex

antisense-RNA based regulation mechanism for decision pathways regarding cell fate (57) using structural homologs of the IsrK (58) and C4 ncRNAs (59), known as seqA and CI RNA in the P4 literature, respectively. While the mechanism of P4 lysogenic maintenance is not known, the IsrK-like ncRNAs of two potentially active P4-like prophage in S. Typhi are required under TraDIS. This sequence element has previously been shown to be essential for the establishment of the P4 lysogenic state (60), and we predict based on our observations that it may be necessary for lysogenic maintenance as well. The fact that some lambdoid prophage in S. Typhimurium encode non-coding genes structurally similar to the IsrK-C4 immunity system of P4 raises the possibility that these systems may be acting as a defense mechanism of sorts, protecting the prophage from predatory satellite phage capable of co-opting its lytic genes.

In addition to repressors, 4 prophage cargo genes in S. Typhimurium and one in S. Typhi are required (See Tables 2 and 3; Supplementary Tables 2 and 3). The S. Typhimurium prophage cargo genes encode a PhoPQ regulated protein, a protein predicted to be involved in natural transformation, an endodeoxyribonuclease, and a hypothetical protein. The S. Typhi prophage cargo gene encodes a protein containing the DNA-binding HIRAN domain (61), believed to be involved in the repair of damaged DNA. These warrant further investigation, as they are genes that have been recently acquired and become necessary for survival in rich media.

To compare differences between requirements for orthologous genes in both serovars, we calculated log-fold read ratios to eliminate genes which were classified differently in S. Typhi and S. Typhimurium but did not have significantly different read densities (see Methods.) Even after this correction, 36 S. Typhimurium genes had a significantly lower frequency of transposon insertion compared to the equivalent genes in S. Typhi (P ¡ 0.05), including four encoding hypothetical proteins (Table 2). This indicates that these gene products play a vital role in S. Typhimurium but not in S. Typhi when grown under laboratory conditions.

A major difference between the two serovars is in the requirement for genes involved in cell wall biosynthesis (see Figure 3). A set of four genes (SL0702, SL0703, SL0706, and SL0707) in an operonic structure putatively involved in cell wall biogenesis is required in S. Typhimurium but not in S. Typhi. The protein encoded by SL0706 is a pseudogene in S. Typhi (Ty2 unique ID: t2152) due to a 1bp deletion at codon 62 that causes a frameshift (Figure 4a). This operon contains an additional two pseudogenes in S. Typhi

(t2154 and t2150), as well as a single different pseudogene (SL0700) in S. Typhimurium, indicating that this difference in gene requirements reflects the evolutionary adaptation of these serovars to their respective niches. Similarly, four genes (rfbV, rfbX, rfbJ and rfbF) within an O-antigen biosynthetic operon are required by S. Typhimurium but not S. Typhi. There appears to have been a shuffling of O-antigen biosynthetic genes since the divergence between the two serovars, and rfbJ, encoding a CDP-abequose synthase, has been lost from S. Typhi altogether. These broader requirements for cell wall-associated biosynthetic and transporter genes suggest that surface structure biogenesis is of greater importance in S. Typhimurium.

We also identified seven genes from the shared pathogenicity island SPI-2 that appear to contain few or no transposon insertions only in S. Typhimurium under laboratory conditions. These genes (spiC, sseA, and ssaHIJT) are thought to encode components of the SPI-2 type III secretion system apparatus (T3SS)(62). In addition, the effector genes sseJ and sifB, whose products are secreted through the SPI-2-encoded type 3 secretion system (T3SS) (63,64), also fell into the required category in S. Typhimurium alone. All of these genes display high A+T nucleotide sequence and have been previously shown (in S. Typhimurium) to be strongly bound by the nucleoid associated protein H-NS, encoded by hns (65,66). Therefore, rather than being required, it is instead possible that access for the transposon was sufficiently restricted that very few insertions occurred at these sites. In further support of this hypothesis, a comparison of the binding pattern of H-NS detected in studies using S. Typhimurium LT2 with the TraDIS results from the SPI-2 locus indicated that high regions of H-NS enrichment correlated well with both the ssa genes described here and with sseJ (65,66) (see Supplementary Figure 1). An earlier study also suggests that high-density DNA binding proteins can block Mu, Tn5, and Tn10 insertion (67); however, a genome-wide study of the effects of H-NS binding on transposition would be necessary to confirm this effect.

Indeed, the generation of null S. Typhimurium mutants in sseJ and sifB, as well as many others generated at the SPI-2 locus suggest that these genes are not truly a requirement for growth in this serovar (64,68-70). While this is a reminder that the interpretation of gene requirement needs to be made with care, the effect of H-NS 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 H-NS binding), which is not what we observed. In total, only 21 required genes fall into the

hns-repressed category described in Navarre, et al. (66)(see Supplementary Table 1); the remainder (almost 400) contained sufficient transposon insertions to conclude they were non-required. In addition, we noted that all SPI-1 genes that encode another Type III secretion system and are of high A+T content were also found to be non-required. This phenomenon was not observed in S. Typhi, possibly because the strain used harbors the pHCM1 plasmid, which encodes the H-NS-like protein sfh and has been shown to affect H-NS binding (71,72).

Twenty-two S. Typhi genes had a significantly lower frequency of transposon insertion compared to orthologs in S. Typhimurium (P ¡ 0.05), indicating that they are required only in S. Typhi for growth under laboratory conditions (Table 3), including the fepBDGC operon. This indicates a requirement for ferric (Fe(III)) rather than ferrous (Fe(II)) iron. This can be explained by the presence of Fe(III) in the bloodstream, where S. Typhi can be found during typhoid fever (15). These genes function to recover the ferric chelator enterobactin from the periplasm, acting with a number of proteins known to aid the passage of this siderophore through the outer membrane (73). It has long been noted that aroA mutants of S. Typhi, deficient in their ability to synthesize enterobactin, exhibit severe growth defects on complex media, while similar mutants of S. Typhimurium grow normally under the same conditions (74), though the mechanism has not been clear. Our results suggest that this difference in growth of aroA mutants is caused by a requirement for iron uptake through the fep system in S. Typhi. During host adaptation, S. Typhi has accumulated pseudogenes in many iron transport and response systems (27), presumably because they are not necessary for survival in the niche S. Typhi occupies in the human host, which may have led to this dependence on fep genes. In contrast, S. Typhimurium generally causes intestinal rather than systemic infection and is able to utilize a wider range of iron sources, including Fe(II), a soluble form of iron present under anaerobic conditions such as those found in the intestine (75).

2.3.4 TraDIS provides resolution sufficient to evaluate ncRNA contributions to fitness

Under a Poisson approximation to the transposon insertion process, a region of 41 (in S. Typhimurium) or 60 bases (in S. Typhi) has only a 1% probability of not containing an insertion by chance. NcRNAs tend to be considerably shorter than their protein-coding

counterparts, but this gives us sufficient resolution to assay most of the non-coding complement of the Salmonella genome. As a proof of principle, we performed an analysis of the best-understood class of small ncRNAs, the tRNAs. Francis Crick hypothesized that a single tRNA could recognize more than one codon through wobble recognition (76), where a non-canonical G-U base pair is formed between the first (wobble) position of the anticodon and the third nucleotide in the codon. As a result, some codons are covered by multiple tRNAs, while others are covered non-redundantly by a single tRNA. We expect that singleton wobble-capable tRNAs, that is wobble tRNAs which recognize a codon uniquely, will be required. In addition, we inferred the requirement for other tRNAs through the non-redundant coverage of their codons and used this to benchmark our ability to use TraDIS to reliably interrogate short genomic intervals.

The S. Typhi and S. Typhimurium genomes encode 78 and 85 (plus one pseudogene) tRNAs respectively with 40 anticodons, as identified by tRNAscan-SE (77). In S. Typhi, 10 out of 11 singleton wobble tRNAs are predicted to be required or ambiguous, compared to 16 tRNAs below the ambiguous LLR cut-off overall (significant enrichment at the 0.05 level, two-tailed Fishers exact test p-value: 6.4e-08.) Similarly in S. Typhimurium, 9 of 11 singleton wobble tRNAs are required or ambiguous compared to 15 required or ambiguous tRNAs overall, again showing a significant enrichment of required tRNAs in this subset (Fishers exact test p-value: 5.2e-07.) The one singleton wobble tRNA which is consistently not required in both serovars is the tRNA-Pro(GGG), which occurs within a 4-member codon family. It has previously been shown in S. Typhimurium that tRNA-Pro(UGG) can read all four proline codons in vivo due to a cmo5U34 modification to the anticodon, obviating the need for a functional tRNA-Pro(GGG) (78) and making this tRNA non-required. The other non-required singleton wobble tRNA in S. Typhimurium, tRNA-Leu(GAG), is similarly a member of a 4-member codon family. We predict tRNA-Leu(TAG) may also be capable of recognizing all 4 leucine codons in this serovar; Such a leucine "four-way wobble" has been previously inferred in at least one other bacterial species (79,80).

Of the 6 required non-wobble tRNAs in each serovar, four are shared. These include two non-wobble singleton tRNAs covering codons uniquely, as well as a tRNA with the ATG anticodon which is post-transcriptionally modified by the required protein mesJ/tilS to recognize the isoleucine codon ATA (80). An additional two required tRNAs in both serovars, one shared and one with a differing anticodon, contain Gln anticodons and are

part of a polycistronic tRNA operon containing other required tRNAs. This operon is conserved in E. coli with the exception of an additional tRNA-Gln at the 3' end that has been lost in the Salmonella lineage. It is possible that transposon insertions early in the operon may interfere with processing of the polycistronic transcript in to mature tRNAs. Finally, we do not observe insertions in a tRNA-Met and a tRNA-Val in S. Typhi and S. Typhimurium, respectively.

Using this analysis of the tRNAs we estimate a worst-case PPV for these short molecules (76 bases) at 81%, in line with our previous estimates for conserved protein-coding genes, and a FPR of ¡4%, higher than for protein-coding genes but still well within the typical tolerance of high-throughput experiments. This assumes that the required operonic tRNA-Glns and the serovar-specific tRNA-Met and tRNA-Val are all false positives; it is not clear that this is in fact the case.

Surveying the shared required ncRNA content of both serovars (see Table 4), we find that the RNA components of the signal recognition particle (SRP) and RNaseP, two universally conserved ncRNAs, are required as expected. The SRP is an essential component of the cellular secretion machinery, while RnaseP is necessary for the maturation of tRNAs. We also find a number of required known and potential cis-regulatory molecules associated with genes required for growth under laboratory conditions in both serovars. The RFN riboswitch controls ribB, a 3,4-dihydroxy-2-butanone 4-phosphate synthase involved in riboflavin biosynthesis, in response to flavin mononucleotide concentrations (81). Additionally, we are able to assign putative functions to a number of previously uncharacterized required non-coding transcripts through their 5' association with required genes. SroE, a 90 nucleotide molecule discovered in an early sRNA screen (82), is consistently located at the 5' end of the required hisS gene across its phylogenic distribution in the Enterobacteriaceae. Given this consistent association and the function of HisS as a histidyl-tRNA synthetase, we hypothesize that this region may act in a manner similar to a T-box leader, inducing or repressing expression in response to tRNA-His levels. The thrU leader sequence, recently discovered in a deep-sequencing screen of E. coli (42), appears to regulate a polycistronic operon of required singleton wobble tRNAs. Three additional required cis-regulatory elements, t44, S15, and StyR-8, are associated with required ribosomal proteins, highlighting the central role ncRNA elements play in regulating fundamental cellular processes.

2.3.5 sRNAs required for competitive growth

Inferring functions for potential trans-acting ncRNA molecules, such as anti-sense binding small RNAs (sRNAs), from requirement patterns alone is more difficult than for cis-acting elements, as we cannot rely on adjacent genes to provide any information. It is also important to keep in mind that TraDIS assays requirements after a brief competition within a large library of mutants on permissive media. This may be particularly important when surveying the bacterial sRNAs, which are known to participate in responses to stress (29).

This is demonstrated by two sRNAs involved in the ?E-mediated extracytoplamic stress response, RybB and RseX, both of which can be successfully knocked out in S. Typhimurium (83). In S. Typhi, rpoE is required, as it also is in E. coli (48,84). However, in S. Typhimurium, rpoE tolerates a heavy insertion load, implying that ?E mutants are not disadvantaged in competitive growth. In S. Typhimurium, the sRNA RseX is required. Overexpression of RseX has previously been shown to compensate for ?E essentiality in E. coli by degrading ompA and ompC transcripts (85). This suggests that RseX may also be short-circuiting the ?E stress response network in S. Typhimurium (Figure 4). To our knowledge, this is the first evidence of a native (i.e. not experimentally induced) activity of RseX.

S. Typhi on the other hand requires ?E along with its activating proteases RseP and DegS and anchoring protein RseA, as well as the ?E-dependent sRNA RybB, which also regulates OmpA and OmpC in S. Typhimurium, along with a host of other OMPs (86). It is unclear why the ?E response is required in S. Typhi but not S. Typhimurium, though it may partially be due to the major differences in the cell wall and outer membrane between the two serovars. In addition, there are significant differences in the OMP content of the S. Typhi and S. Typhimurium membranes that may be driving alternative mechanisms for coping with membrane stress. For instance, S. Typhi completely lacks OmpD, a major component of the S. Typhimurium outer membrane (87) and a known target of RybB (29).

Two additional sRNAs involved in stress response are also required by both S. Typhi and S. Typhimurium. The first, MicA, is known to regulate ompA and the lamB porin-coding gene in S. Typhimurium (88), contributing to the extracytoplasmic stress response. The second, DsrA, has been shown to negatively regulate the nucleoid-forming

protein H-NS and enhance translation of the stationary-phase alternative sigma factor ?S in E. coli (89), though its regulation of ?S does not appear to be conserved in S. Typhimurium (90). Both have been previously deleted in S. Typhimurium, and so are not essential. H-NS knockouts have previously been shown to have severe growth defects in S. Typhimurium that can be rescued by compensatory mutations in either the phoPQ two-component system or rpoS, implying that the lack of H-NS is allowing normally silenced detrimental regions to be transcribed (66). As MicA has recently been shown to negatively regulate phoPQ expression in E. coli (91), it is tempting to speculate that MicA may be moderating the effects of DsrA-induced H-NS repression; however, it is currently unclear whether sRNA regulons are sufficiently conserved between E. coli and S. enterica to justify this hypothesis.

2.4 Conclusions

The extremely high resolution of TraDIS has allowed us to assay gene requirements in two very closely related Salmonellae with different host ranges. We found, under laboratory conditions, that 58 genes present in both serovars were required in only one, suggesting that identical gene products do not necessarily have the same phenotypic effects in the two different serovar backgrounds. Many of these genes occur in genomic regions or metabolic systems which contain pseudogenes and/or have undergone reorganization since the divergence of S. Typhi and S. Typhimurium, demonstrating the complementarity of TraDIS and phylogenetic analysis. These changes may in part explain differences observed in the pathogenicity and host specificity of these two serovars. In particular, S. Typhimurium showed a requirement for cell surface structure biosynthesis genes; this may be partially explained by the fact that S. Typhi expresses the Vi-antigen which masks the cell surface, though these genes are not required for survival in our assay. S. Typhi on the other hand has a requirement for iron uptake through the fep system, which enables ferric enterobactin transport. This dependence on enterobactin suggests that S. Typhi is highly adapted to the iron-scarce environments it encounters during systemic infections. Furthermore, this appears to represent a single point of failure in the S. Typhi iron utilization pathways, and may present an attractive target for narrow-spectrum antibiotics.

Of the approximately 4500 protein coding genes present in each serovar, only about 350

were sufficiently depleted in transposon insertions to be classified as required for growth in rich media. This means that over 92% of the coding genome has sufficient insertion density to be queried in future assays. Dense transposon mutagenesis libraries have been used to assay gene requirements under conditions relevant for infection, including S. Typhi survival in bile (35), Mycobacterium tuberculosis catabolism of cholesterol (92), drug resistance in Pseudomonas aeruginosa (93), and Haemophilus influenzae survival in the lung (94). We expect that parallel experiments querying gene requirements under the same conditions in both serovars examined in this study will yield further insights in to the differences in the infective process between Typhi and Typhimurium, and ultimately the pathways that underlie host-adaptation.

Both serovars possess substantial complements of horizontally-acquired DNA. We have been able to use TraDIS to assay these recently acquired sequences. In particular, weve been able to identify, on a chromosome wide scale, active prophage through the requirement for their repressors. The P4 phage utilizes an RNA-based system to make decisions regarding cell fate, and structurally similar systems are used by P1, P7, and N15 phage (95,96). C4-like transcripts have been regarded as the primary repressor of lytic functions, though the IsrK-like sequence is known to be essential to the establishment of lysogeny in P4 and is transcribed in at least two phage types (60,96). Our observations in S. Typhi suggest an important role for the IsrK-like sequence in maintenance of the lysogenic state in P4-like phage, though the mechanism remains unclear.

Recent advances in high-throughput sequencing have greatly enhanced our ability to detect novel transcripts, such as ncRNAs and short open reading frames (sORFs). In fact, our ability to identify these transcripts now far out-strips our ability to experimentally characterize these sequences. There have been previous efforts at high-throughput characterization of bacterial sRNAs and sORFs in enteric bacteria; however, these have relied on labor-intensive directed knockout libraries (47,97). Here we have demonstrated that TraDIS has sufficient resolution to reliably query genomic regions as short as 60 bases, in agreement with a recent high-throughput transposon mutagenesis study in the alphaproteobacteria Caulobacter crescentus (38). Our method has the major advantage that library construction does not rely upon genome annotation, and newly discovered elements can be surveyed with no further laboratory work.

We have been able to assign putative functions to a number of ncRNAs using TraDIS though consideration of their genomic and experimental context. In addition, ncRNA

characterization generally is done in model organisms like E. coli or S. Typhimurium, and it is unclear how stable ncRNA regulatory networks are over evolutionary time. By assaying two serovars of Salmonella with the same method under the same conditions, we have seen hints that there may be differences in sRNA regulatory networks between S. Typhi and S. Typhimurium. In particular, we have found that under the same experimental conditions, S. Typhi appears to rely on the ?E stress response pathway while S. Typhimurium does not; it is tempting to speculate that this difference in stress response is mediated by the observed difference in requirement for two sRNAs, RybB and RseX. We believe that this combination of high-throughput transposon mutagenesis with a careful consideration of the systems context of individual genes provides a powerful tool for the generation of functional hypotheses. We anticipate that the construction of TraDIS libraries in additional organisms, as well as the passing of these libraries through relevant experimental conditions, will provide further insights into the function of bacterial ncRNAs in addition to the protein-coding gene complement.

Published Works

Publications produced during the course of this thesis:

- Martin M.J., Clare S., Goulding D., Faulds-Pain A., Barquist L., Browne H., Pettit L., Lawley T.D., Dougan G., Wren B.W. **The *agr* locus regulates virulence and colonization genes in *Clostridium difficile* 027.** Manuscript under review, 2013.

- Reuter S., Conner T.R., Barquist L., Walker D., Feltwell T., Harris S.R., Fookes M., Hall M.E., Fuchs T.M., Corander J., Dufour M., Ringwood T., Savin C., Bouchier C., Martin L., Miettinen M., Shubin M., Laukkanen-Ninios R., Sihvonen L.M., Siitonen A., Skurnik M., Falcão J.P., Fukushima H., Scholz H.C., Prentice M., Wren B.W., Parkhill J., Carniel E., Achtman M., McNally A., Thomson N.R. **Parallel independent evolution of pathogenicity within the genus *Yersinia*.** Manuscript under review, 2013.

- Croucher N.J., Mitchell A.M., Gould K.A., Inverarity D., Barquist L., Feltwell T., Fookes M.C., Harris S.R., Dordel J., Salter S.J., Browall S., Zemlickova H., Parkhill J., Normark S., Henriques-Normark B., Hinds J., Mitchell T.J., Bentley S.D. **Within-patient diversification and genomic stasis within a pneumococcal lineage.** Manuscript under review, 2013.

- Barquist L., Boinett C.J., Cain A.K.. **Approaches to querying bacterial genomes with transposon-insertion sequencing.** *RNA Biology*, 10(7), 2013.

- Barquist L., Langridge G.C., Turner D.J., Phan M.D., Turner A.K., Bateman A., Parkhill J., Wain J., Gardner P.P. **A comparison of dense transposon insertion libraries in the *Salmonella* serovars Typhi and Typhimurium.** *Nucleic Acids Research*, 41(8):4549-4564, 2013.

- Burge S.W., Daub J., Eberhardt R., Tate J., Barquist L., Nawrocki E.P., Eddy S.R., Gardner P.P., Bateman A. **Rfam 11.0: 10 years of RNA families.** *Nucleic Acids Research*. 41(D1):D226-D232, 2013

- Croucher N.J., Harris S.R., Barquist L., Parkhill J., Bentley S.D. **A high-resolution view of genome- wide pneumococcal transformation.** *PLoS Pathogens*, 8(6), 2012

- Westesson O., Barquist L., Holmes I. **HandAlign: Bayesian multiple sequence alignment, phylogeny, and ancestral reconstruction.** *Bioinformatics*, 28(8):1170-1171, 2012

- Gardner P.P., Barquist L., Bateman A., Nawrocki E.P., Weinberg Z. **RNIE: genome-wide prediction of bacterial intrinsic terminators.** *Nucleic Acids Research*, 39(14):5845-5852, 2011

References

Barquist, L., C. J. Boinett, and A. K. Cain (2013). "Approaches to querying bacterial genomes with high-throughput transposon insertion sequencing". *RNA Biology* (2013).

Barquist, L., G. C. Langridge, D. J. Turner, M. D. Phan, A. K. Turner, A. Bateman, J. Parkhill, J. Wain, and P. P. Gardner (2013). "A comparison of dense transposon insertion libraries in the Salmonella serovars Typhi and Typhimurium". *Nucleic acids research* (2013).