3 A Method for RNA-seq in prokaryotes

3.1 Aims of the work described in this Chapter

Recent advances in DNA sequencing technology have transformed the amount of sequencing information that can be obtained in single sequencing runs by orders of magnitude. The aim of this chapter is to apply such technology to sequence cDNA made from mRNA, thus developing a novel approach to assess gene expression profiles in bacteria.

3.2 Introduction

The advent of high-throughput sequencing technologies has permitted new approaches to exploring functional genomics, including the direct sequencing of complementary DNA (cDNA), an approach that has been named RNA-seq. Recent publications have described the exploitation of this high-density, high-resolution technology to build an accurate picture of the transcriptional patterns within eukaryotic organisms [188,189]. These studies have demonstrated a number of advantages over microarray-based techniques, including greater sensitivity, increased dynamic range, reduced background noise, single nucleotide resolution and the capacity to map data to the entire genome sequence [190]. Furthermore, the results are not biased by array design as most expression arrays have limited density and features are normally designed on the basis of classical *in silico* genome annotation. Consequently, RNA-seq has led to the discovery of novel genetic features [191,192].

One of the drawbacks of the initial RNA-seq studies, relative to microarray work, was the lack of strand-specificity in the data, as these protocols sequence double stranded cDNA, which masks directionality [188,189]. However, two recent studies have

demonstrated that directionality can be retained through asymmetrically modifying the 5' and 3' ends of the RNA molecules prior to reverse transcription, either by attaching RNA linkers [193] or switch strand PCR [194], allowing the transcripts to be mapped back to the reference genomes in a strand specific manner. This is crucial for resolving overlapping genetic features, detecting antisense transcription and assigning the sense strand for non-coding RNA (ncRNA).

This chapter details the development of a directional RNA-seq protocol that eliminates both the need for second strand cDNA synthesis and the modification of transcripts prior to reverse transcription. This method was applied to the rRNA depleted RNA population of *S.* Typhi allowing capture of an unbiased view of the coding and non-coding RNA transcriptome. In combination with this method, the teams at the WTSI have also developed a computational pipeline based on Artemis (www.sanger.ac.uk/Projects/Pathogen/Transcriptome/) that facilitates mapping of the transcriptome data [195,196]. Together, these methods should greatly enhance the understanding of microbial genome's transcriptional content.

3.3 Results

3.3.1 Directional Sequencing of Single Stranded cDNA

Sequencing using the Illumina platform requires the ligation of adapters onto either end of a DNA molecule (http://icom.illumina.com), which are necessary for PCR amplification, flow cell attachment and sequencing reaction priming. As DNA ligase only works efficiently on DNA duplexes, samples are prepared as double stranded DNA and subjected to an end repair reaction, using a mixture of enzymes to repair 3' overhangs and extend from recessed 3' giving blunt ended products. DNA molecules

are subsequently 3' monoadenylated and ligated to adapter dimers with a 5' monothymidine overhang. During the course of development of a high-throughput RNA-seq method for *S.* Typhi, in cases where the second strand cDNA synthesis failed or was omitted, the Illumina sequencing team were still able to generate data using the standard Illumina sample preparation methods but unlike the double stranded cDNA method, the transcript data retained directional fidelity.

Four hypotheses were possible in explaining the ligation of linkers to single stranded cDNA and subsequent processing to generate sequence data and maintain directionality (figure 3.1). The first hypothesis required the ligation of single stranded DNA adaptors to the single stranded cDNA molecules (figure 3.1a). This is possible because T4 DNA ligase can ligate single stranded DNA molecules, albeit at low efficiency [195] and directionality would be maintained because the second strand is never synthesised (figure 3.1a). The alternate possibilities involved the formation of duplexes during the end repair reaction (figure 3.1b-d). Either annealed RNA fragments (the remains of transcripts that served as templates in the reverse transcription reaction) or inter- or intra-molecular hybridisation of cDNA was suggested to prime complementary strand synthesis, leading to the formation of blunt ended, double stranded constructs that could then function as the substrate for the efficient ligation of adapters. If complementary strand synthesis were primed by annealed RNA fragments, this strand would be composed of both RNA and DNA, which cannot be amplified and sequenced by DNA-dependent DNA polymerases. Consequently only the original single stranded cDNA strand would be sequenced. If complementary strand synthesis were primed by intra- or inter-molecular cDNA annealing, then 3' end processing would produce a reverse complement of the annealed cDNA's 5' end. Hence sequences with sense and antisense orientations

would be segregated into the 3' and 5' regions of the cDNA strands respectively, so by sequencing only the 5' end, all sequence reads maintain the same orientation relative to the original direction of transcription.

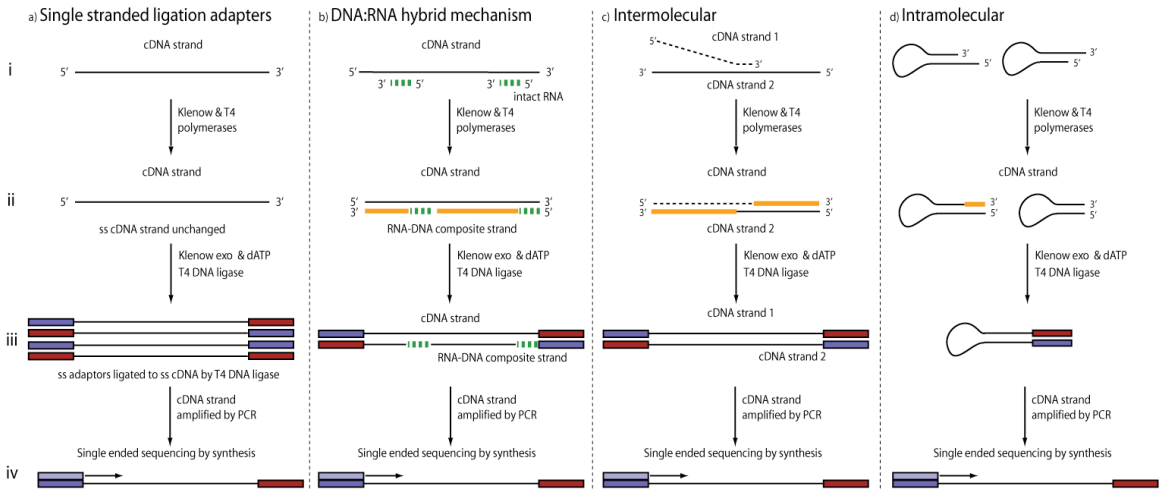

Figure 3.1 The hypotheses proposed to account for the attachment of Illumina adapted dimers to sscDNA

a) attachment of adapters to ss cDNA, and priming of second strand synthesis by b) RNA fragments c) intermolecular cDNA annealing d) intramolecular cDNA annealing. Dark blue rectangles represent linker sequence and red, the reverse complement of the linker sequence. Light blue rectangles represent sequencing primer used on the Illumina slide during sequencing.

To test these hypotheses, a 48-mer DNA oligonucleotide was designed, consisting of a defined sequence tag, an RNA oligonucleotide-binding site and two stretches of random sequence (figure 3.2a). Solutions containing either this DNA oligonucleotide alone, or in the presence of a 12 nucleotide RNA oligonucleotide complementary to the binding site, were subjected to standard Illumina sample preparation and sequencing reactions. If RNA primed the second strand synthesis, then resection of the overhanging 5 nucleotide random sequence at the 3' end would be observed and libraries would not be generated in the absence of the RNA oligonucleotide (figure 3.1b). Conversely, if inter- or intra-molecular cDNA annealing were the dominant mechanism, then the reverse complement of the known 5' sequence tag would be

observed at the 3' end of the oligo (figure 3.1c and d). However, if the mechanism was simple annealing of linkers directly to single stranded-cDNA, the DNA oligonucleotide would remain unaltered (figure 3.1a).

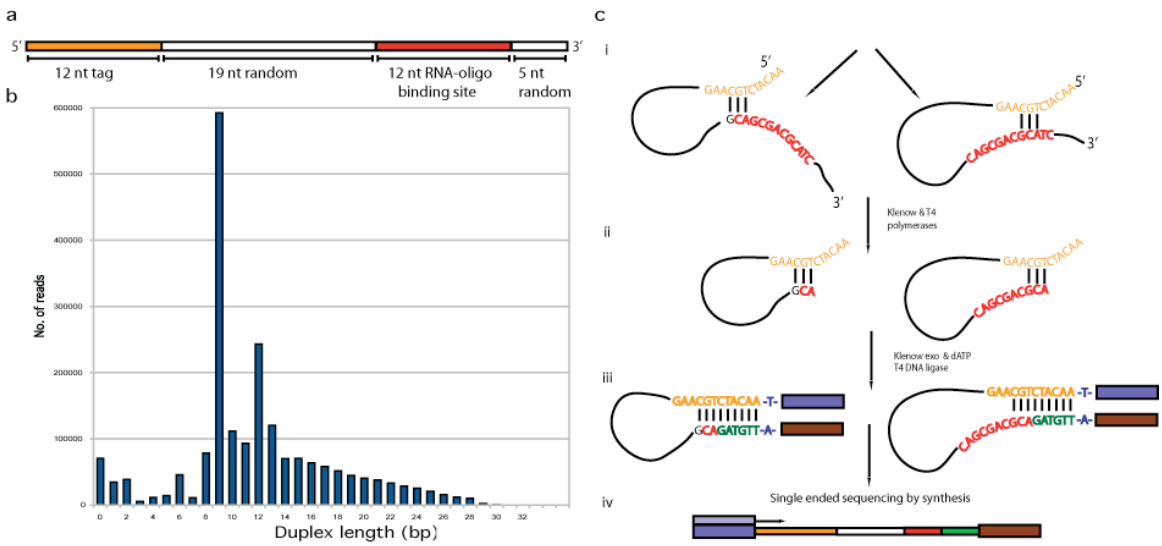

Figure 3.2 Single stranded sequencing

a) Schematic representation of the DNA oligonucleotides from which Illumina libraries were generated. b) Distribution of duplex lengths amongst a sequenced sample of single stranded DNA oligonucleotides. Reads were extracted corresponding to the oligonucleotide by searching the output data for the 12 nt known sequence tag. Duplex lengths were then calculated by counting the number of bases at the 3' end found to be the reverse complement of those at the 5' end. This revealed a smooth distribution of values over a range of sizes, with large peaks at 12 bp (likely resulting from intermolecular annealing) and 9 bp (probably the consequence of a 3 bp duplex that can form between the known sequence tag and RNA binding site). c) Two proposed mechanisms for the formation of species with a 9 nt of reverse complementarity between the 5' and 3' ends, the most common duplex length observed. The vast majority of these were found to have 3 bp "seed duplexes" formed by base pairing between –CGT- in the sequence tag and either the –GCA- in the 3' half of the RNA oligonucleotide binding site, or the CA- dinucleotide at the start of the binding site when the preceding nucleotide (the last of the 19 nt random sequence) was G.

Following the standard Illumina preparation protocol, libraries were successfully generated both from mixtures of the DNA and RNA oligonucleotides, as well as the DNA oligonucleotide alone (but not the RNA oligonucleotide alone), suggesting RNA molecules were not required for priming of second strand synthesis. Illumina

sequence reads were obtained from a library constructed from a solution of single stranded DNA oligonucleotides. In 88% of the sequences, the RNA binding site sequence had been at least partially altered to the reverse complement of the known sequence tag, implying intra- or inter-molecular annealing had occurred and primed second-strand synthesis to give blunt-ended duplexes. The most common species (29% of the sequenced population, figure 3.2b) had 9 nucleotides of reverse complement of the 5' tag at the 3' end (a 9 bp "duplex length"), which is likely to have arisen from the scenarios outlined in Figure. 3.2c. The second most common species (12% of the sequenced population) have a 12 bp duplex length. This is likely an artefact of inter-strand annealing: because the 19 nucleotide random sequences will be different in the two annealed strands, only the 12 nucleotide tags will match between the two ends of the repaired duplex. However, a third of the duplex lengths are longer than 12 bp, indicating that intra-strand annealing also occurs to a detectable extent (because the complementary nature of the two ends extends beyond the known tags into the random sequence unique to each oligonucleotide molecule). Hence, this has demonstrated that Illumina libraries can be constructed from single stranded DNA using standard techniques, where both inter- and intra-strand annealing occur to a comparable extent and make a significant contribution to the formation of double stranded cDNA during end repair.

3.3.2 Sequencing of the *S.* Typhi transcriptome

The total RNA population was extracted from *S.* Typhi BRD948 [177], single stranded cDNA generated by a reverse transcription reaction and the cDNA was used as the substrate for standard library construction reactions and Illumina sequencing. This project generated a dataset of 5.4 million 36 nucleotide reads that was mapped

back to the genome in a strand-specific manner. However, as the total RNA population was sampled, the majority of reads mapped to the rRNA operons, suggesting removal of these molecules prior to reverse transcription would greatly increase the sensitivity of the technique.

Two alternative methods were successful in reducing the amount of highly abundant 16S and 23S rRNA molecules, thus enriching for mRNA and non-coding RNA (ncRNA) transcripts: Reduction through hybridisation of capture oligonucleotides and subsequent annealing to biotinylated oligonucleotides and removal by magnetic streptavidin coated beads or degradation using a terminator exonuclease specific for 5' monophosphorylated transcripts [197]. Both approaches depleted the rRNA and increased the number of uniquely mapping sequences (table 3.1). Plotting the arithmetic mean coverage per base of each annotated coding sequence in the *S.* Typhi genome in biological replicates using the two different techniques yielded a correlation coefficient of R=0.90, suggesting the two different methods give comparable results (figure 3.3). No clear outliers were evident in the correlation plots, suggesting there is no bias for a subset of transcripts specific to either method; however, depletion through hybridisation to oligonucleotides is likely to be the more specific technique, as any cleaved transcripts or RNA species generated through endonucleolytic cleavage of a larger precursor would be removed by the exonuclease treatment.

Table 3.1 Read mapping for different depletion methods

Depletion Method	Undepleted	Oligonucleotide hybridisation	Terminator exonuclease
Amount RNA (μg)	100	100	100
Amount cDNA (μg)	20	20	20
Size fraction (bp)	200-250	200-250	200-250
Read Length (bp)	36	36	36
No. Reads	5372456	7183969	7043338
Proportion of Reads Mapped	94%	91%	94%
Proportion of Reads Mapped Uniquely	14%	41%	51%

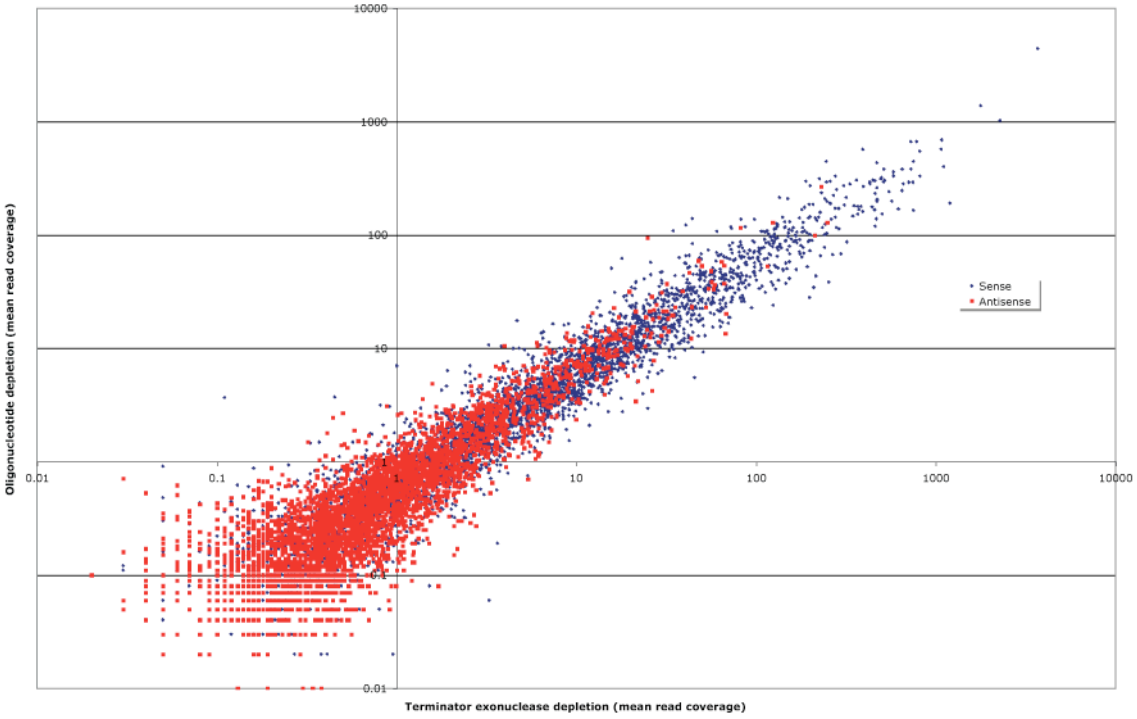

Figure 3.3 Quantitative comparison of terminator exonuclease and oligonucleotide hybridisation rRNA depletion techniques.

This log-log plot shows the mean read coverage in the sense and antisense directions for all annotated CDS features in the *S.* Typhi Ty2 genome in datasets depleted by the two different techniques. The Pearson R_2 and Spearmann correlation coefficients of the results of the two methods are >0.80 in both directions, indicating that these results are robust and reproducible. The absence of large numbers of outliers demonstrates that the terminator exonuclease does not remove a large number of non-rRNA transcripts.

3.3.3 Mapping and visualising data

In order to sample a wide range of different transcriptional patterns, *S.* Typhi cultures were grown shaking in LB to three different time points (OD_{600}=0.3, 0.45 and 0.6) in 50ml falcon tubes and in 15ml volumes. Illumina sequencing of the pooled samples generated 45 million ~36 nucleotide reads from 7 independent experiments. 4.8 million reads mapped to annotated coding sequences (CDS). Strand specific mapping allowed the unambiguous assignment of the 'sense' strand for transcriptional units,

including ncRNA, and permitted deconvolution of overlapping genetic features on opposite strands.

These data were use to assist development of a computational pipeline at the WTSI (composed of tools freely available from http://www.sanger.ac.uk/Projects/Pathogens/Transcriptome/) that allow the visualisation of sequence read depth mapped to the annotated genome sequence within the freeware program Artemis (figure 3.4). The programme MAQ [185], an algorithm that retains non-uniquely mapping reads within the dataset, was used to calculate a fold coverage value for every nucleotide in the genome in both the forward and reverse directions. A second script, maqpileuptodepth.pl, is able to produce an analogous output from data mapped with MAQ. Using these tools, patterns of transcription were observed within *S.* Typhi to single nucleotide resolution.

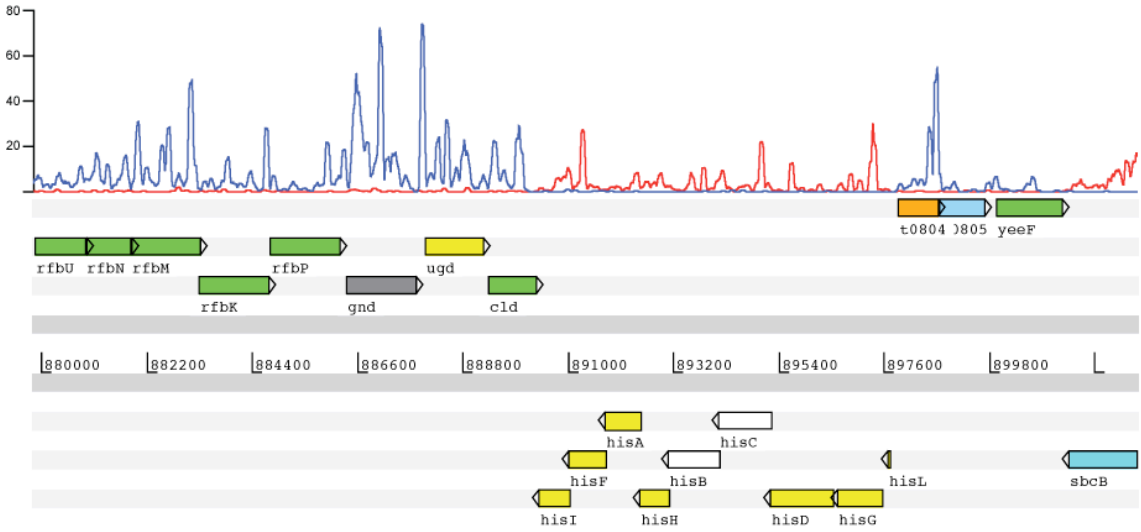

Figure. 3.4 Representation of transcriptomic sequence read coverage plot relative to genome annotation, as displayed in Artemis

A histidine biosynthesis gene cluster in *S.* Typhi Ty2. The graph lines represent the mean number of sequence reads (averaged over a 100 nt sliding window) mapping to the forward (blue) or reverse (red) strand.

3.4 Discussion and conclusion

This chapter summarises the development of a high-throughput RNA-seq method for RNA isolated from *S.* Typhi. This method identifies the template strand for the transcriptome effectively deconvoluting the sequence data. This approach includes a novel method for the isolation of a total RNA sample, depleted of rRNA, suitable for high-throughput sequencing, and a new approach for retaining directional fidelity in transcriptomic data by sequencing single-stranded cDNA, a method that is actually simpler than the original RNA-seq protocols as it abrogates the need for second strand cDNA synthesis. These sequence data have assisted in the development of the Illumina sequencing and analysis pipeline, which includes custom-built programmes for mapping and visualising the output data in context with the reference genome.

This method has the capacity to further our understanding of the *S.* Typhi genome expression.

Datasets produced in this manner allow the detection of ncRNA, operon structures and 5' and 3' untranslated regions, features crucial for gene regulation that are difficult to predict from genome sequences *de novo*, across the entire chromosome. Hence as well as measuring transcriptional activity it is clear that this approach will prove to be of great value for complementing and refining current genome annotations in prokaryotes.