6 A simple method for directional RNA-seq

6.1 Introduction

One of the drawbacks of the initial RNA-seq studies, relative to microarray work, was the lack of information on the direction of transcription. These protocols sequenced ds-cDNA, thereby masking directionality by showing equal signal on both strands (Nagalakshmi *et al*., 2008; Wilhelm *et al*., 2008). This has resulted in a number of techniques being developed in order to retain information on the direction of transcription in the output of such studies. Such data are crucial for resolving overlapping genetic features, detecting antisense transcription and assigning the sense strand for non-coding RNA (ncRNA). These published methods for directional RNA-seq either modify the RNA molecules prior to reverse transcription, through attaching RNA linkers (Lister *et al*., 2008) or by bisulfite-induced cytosine deamination (He *et al*., 2008), or by modifying the first cDNA strand prior to second strand synthesis by adding cytosine residues to the 3' end in a template switching PCR (Cloonan *et al*., 2008). These techniques, all developed for the study of eukaryotic cells, modify the nucleic acid sample prior to the second strand of the duplex being synthesised, thereby allowing reads to be assigned to a specific strand of the genome. However, adding extra steps to any sample preparation protocol increases the risks of sample biases being introduced or exacerbated. Furthermore, the high ribonuclease activity within bacterial cells makes mRNA highly unstable: prokaryotic mRNA typically has a half-life of minutes, whereas in eukaryotic cells such transcripts usually have a half-life on the order of an hour (Rauhut and Klug, 1999). Hence a protocol that minimizes sample manipulation, whilst retaining information on the template strand of transcription, is ideal for studying bacterial gene expression.

6.2 Description and validation of the RNA-seq methodology

6.2.1 Illumina sequencing libraries can be generated from ss-DNA

Sequencing using the Illumina platform requires the ligation of adapters, necessary for PCR amplification, flow cell attachment and sequencing reaction priming, onto

either end of a DNA molecule (Bentley *et al*., 2008). The standard Illumina library preparation protocol requires that samples are prepared in a double-stranded form and subjected to an end repair reaction, using either Klenow to resect 3' overhangs or T4 polymerase to extend from recessed 3' ends to give 'polished' blunt-ended products. These are subsequently 3' monoadenylated and the Illumina adapters, in the form of dimers with a 3' monothymidine overhang, are ligated.

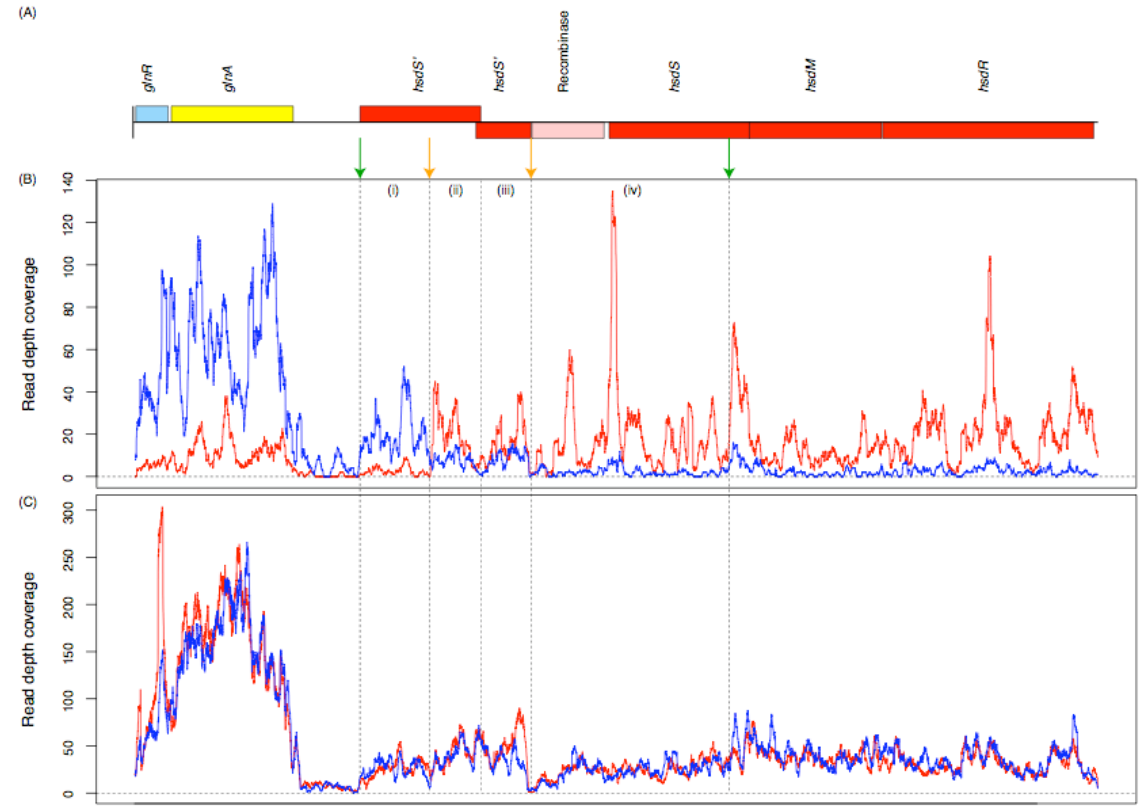

Figure 6.1 Example of the data produced by sequencing ss-cDNA and ds-cDNA. (A) The displayed locus contains the *hsdRMS* operon, which encodes a functional type I restriction system, along with two gene fragments (both labelled *hsdS'*) corresponding to alternative C termini of the specificity subunit, *hsdS*. Also present is a site-specific recombinase, able to invert sequence between two 15 bp repeats, marked with orange arrows, and two 85 bp repeats, marked with green arrows. (B) The data produced by the sequencing of ss-cDNA is shown as Illumina read coverage relative to the annotation. The blue line shows the coverage of reads corresponding to transcription of the top strand of the annotation; the red line shows the data corresponding to transcription of the bottom strand. There appears to be rapid inversion between the 15 bp repeats, hence regions (ii) and (iii) are bidirectionally transcribed. However, inversion between the 85 bp repeats seems to be slower, with no significant antisense transcription observable in regions (i) and (iv). (C) The data produced by ds-cDNA sequencing is shown as read coverage relative to the annotation in the same manner as in (B). In this experiment, the directional information is lost, due to the symmetrical nature of the cDNA being sequenced, but the relative levels of expression of each gene are similar to those in (B).

Unexpectedly, it was found to be possible to produce Illumina libraries from ss-cDNA, generated from *S. pneumoniae* ATCC 700669 RNA. When sequenced, it was

found that such samples retained information on the direction of transcription that generated the template RNA molecule (*e.g.*, Figure 6.1B). Four mechanisms by which ss-cDNA might undergo correct processing to generate Illumina libraries were proposed (summarised in Figure 6.2). The first required the ligation of adapters to the ss-cDNA molecules (Figure 6.2A). This is possible because T4 DNA ligase can ligate ss-DNA molecules, albeit at low efficiency (Kuhn and Frank-Kamenetskii, 2005) (Figure 6.2A, iii), and directionality would be maintained because the second strand is never synthesized (Figure 6.2A, iv). The alternate possibilities involved the formation of duplexes during the end repair reaction (Figure 6.2B–D). Either annealed RNA fragments (the remains of transcripts that served as templates in the reverse transcription reaction; Figure 6.2B, ii) or inter or intramolecular hybridization of cDNA (Figure 6.2C, ii) and Figure 6.2D, ii), were suggested to prime complementary strand synthesis, leading 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 (Figure 6.2B, iii), which cannot be amplified and sequenced by DNA-dependent DNA polymerases. Consequently, only the original ss-cDNA strand would be sequenced (Figure 6.2B, iv). If complementary strand synthesis were primed by intra or intermolecular cDNA annealing, then 3' end processing would produce a reverse complement of the annealed cDNA's 5' end (Figure 6.2C, ii and Figure 6.2D, ii). Hence, sequences with different orientations relative to the original transcript would be segregated into the 3' and 5' regions of the cDNA strands, so by sequencing only the 5' end, all sequence reads maintain the same orientation relative to the original RNA molecule (Figure 6.2C, iv and Figure 6.2D, iv).

In order to determine which of these mechanisms described above occurs during library preparation, a 48 nt DNA oligonucleotide composed of a defined 5' sequence tag and RNA oligonucleotide binding site separated by two stretches of random sequence (Figure 6.3A) was designed. Solutions containing either this DNA oligonucleotide alone, or in the presence of a 12 nt RNA oligonucleotide complementary to the binding site, were subjected to standard Illumina sample preparation and sequencing reactions (see Materials and Methods).

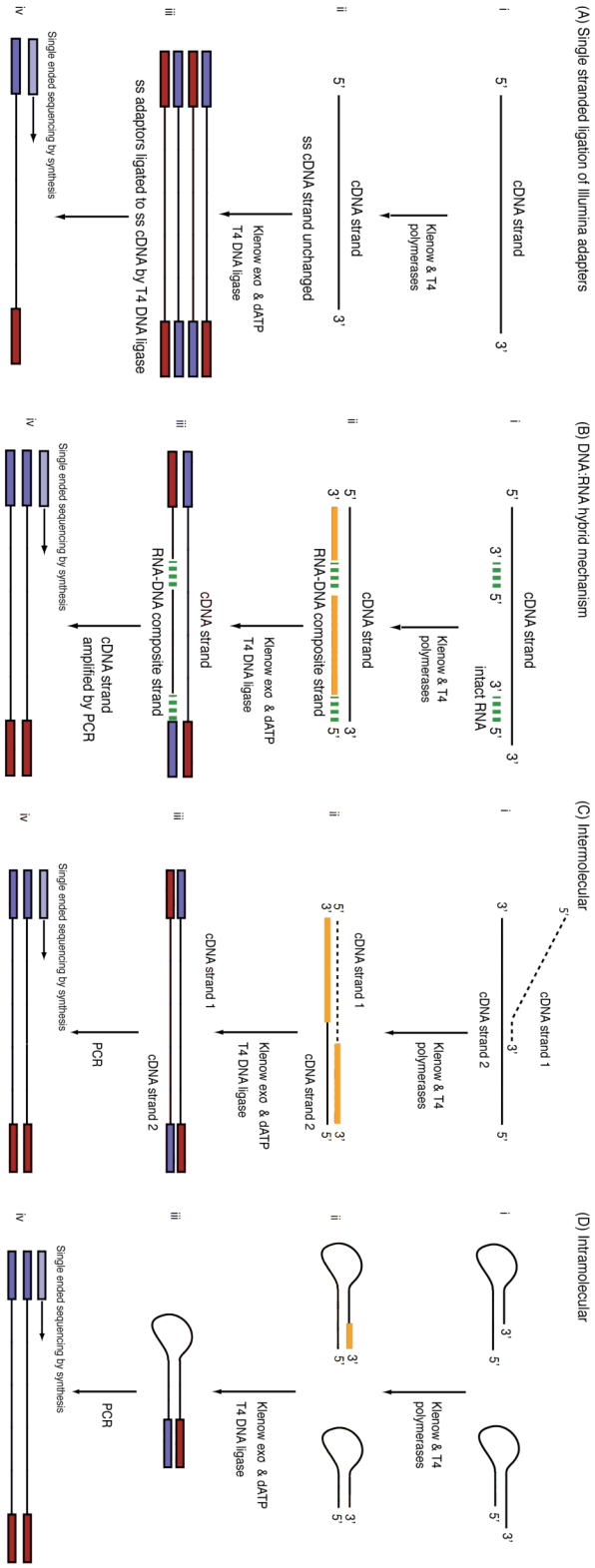

Figure 6.2 The hypotheses proposed to account for the attachment of Illumina adapted dimers to ss-cDNA: (A) attachment of adapters to ss-cDNA, and priming of second strand synthesis by (B) RNA fragments, (C) intermolecular cDNA annealing and (D) intramolecular cDNA annealing.

Libraries were successfully generated both in the presence and absence of the RNA oligonucleotide, demonstrating that adapter ligation did not require RNA-primed complementary strand synthesis. Analysis of 2,162,655 paired 36 nt sequence reads generated from libraries produced in the absence of RNA revealed that in 88% of the DNA molecules, the RNA binding site had been partially replaced by sequence representing the reverse and complement of the known 5' end tag of the 48-mer DNA oligonucleotide (as shown in Figure 6.3C). This indicated that duplexes had been formed through intra or intermolecular annealing followed by processing of the 3' end. The most common species (29% of the sequenced population) had 9 nt of reverse complement of the 5' tag at the 3' end (equivalent to a 9 bp 'duplex length'), which is likely to have arisen from the scenarios outlined in Figure 6.3C.

In cases where more than 12 nt of sequence is generated at the 3' end, the calculated duplex length depends on whether annealing occurs intra or intermolecularly. If annealing is intramolecular, then the reverse complement of the 5' end of the random sequence region is found near the 3' end, resulting in a duplex length greater than 12 nt. This is observed in around a third of cases. However, if intermolecular hybridization occurs, then the reverse complement of the annealed molecule's 5' region is synthesized at the 3' end of the sequenced molecule. In such a case, a duplex length of 12 nt will usually be observed. This is because only the 12 nt 5' tag, common to all molecules, can be identified as having its reverse complement at the 3' end; 3' end processing otherwise replaces random sequence with the reverse complement of another molecule's random sequence. Such a scenario is likely to account for much of the 12% of the sequenced population with a 12 bp duplex length. Similar results are observed when libraries are constructed from the ssDNA in the presence of the RNA oligonucleotide (data not shown). Hence, this shows that Illumina libraries can be constructed from ss-cDNA using standard protocols, with both intra and intermolecular annealing occurring to a comparable extent and contributing to the formation of duplexes during the end repair reaction.

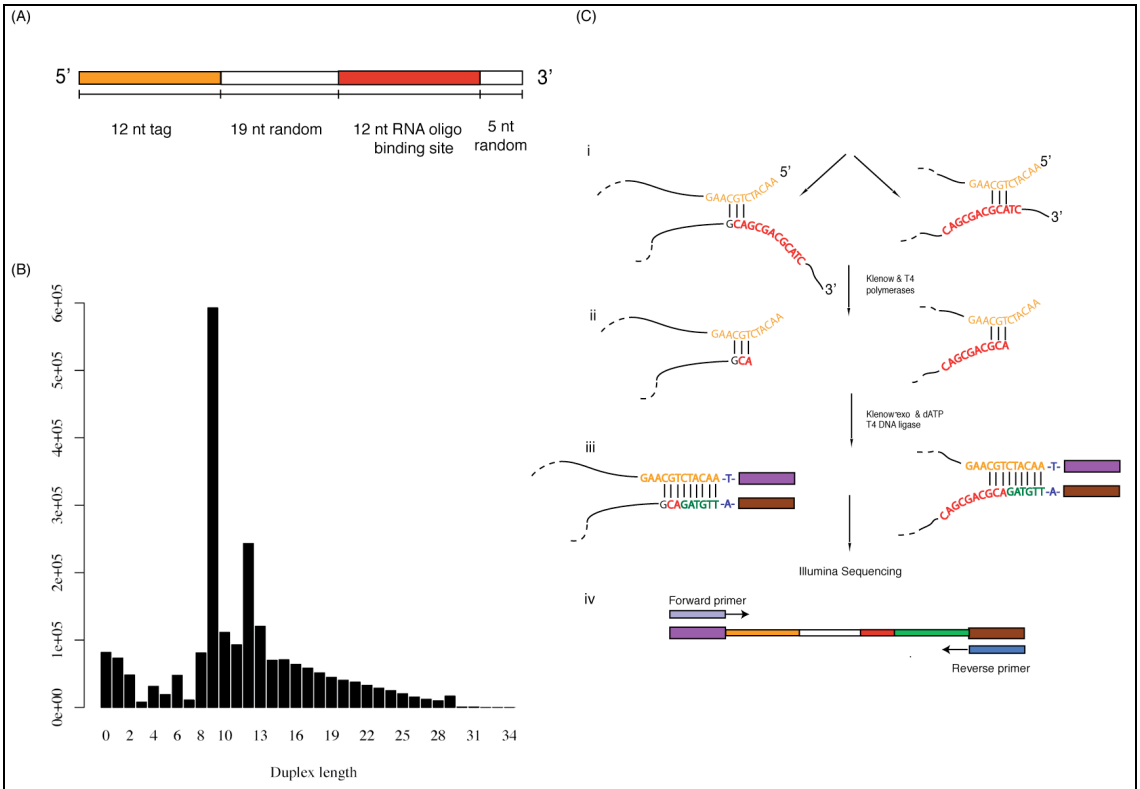

Figure 6.3 (A) Schematic representation of the DNA oligonucleotides from which Illumina libraries were generated. (B) Distribution of duplex lengths amongst a sequenced sample of single-stranded DNA oligonucleotides. Reads corresponding to the oligonucleotide were extracted by searching the output data 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 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.

6.2.2 ss-cDNA sequencing retains data directionality

An RNA sample from an exponentially growing pneumococcal culture was processed to produce an ss-cDNA sample, half of which underwent second strand synthesis to produce an equivalent ds-cDNA sample. Both of these were then sequenced on the Illumina platform to give 54 nt paired end reads, revealing only the ss-cDNA sample retained information on the direction of transcription (Figure 6.1). The proportion of reads mapping to CDSs in the sense direction varied with the level of putative antisense transcription, the degree to which it overlapped with convergently

transcribed CDSs and the amount of experimental noise, with a median percentage of 87%. This compares to a value of 50% for the ds-cDNA sample.

A high proportion of the cDNA aligns to rDNA sequences (68% in the case of ss-cDNA, 48% in the case of ds-cDNA), which map redundantly to the four almost identical pneumococcal rRNA operons. Hence for the ss-cDNA and ds-cDNA samples, about 24% and 46%, respectively, of the mapping reads could be aligned as pairs in which both read mates could be assigned to unique locations on the chromosome. Of these, 97% of the pairs in the ds-cDNA sample correspond to 'proper pairs', which map as such data are expected to, with the two reads aligning to complementary strands of the genome at sites separated by an insert size congruent with that expected from library construction. The comparable figure for the ss-cDNA sample is 62%; these represent cases where the level of 3' end processing of the cDNA is not sufficient to interfere with the mapping of the reverse read of the pair. The majority of the remainder (36% of the uniquely aligned pairs) mapped to the genome with an insert size greater than 1 kb, with the two reads either mapping to the same strand or complementary strands. This is indicative of intermolecular annealing: the sequence of the reverse read originates from a different cDNA molecule to that of the forward read, resulting in a chimeric molecule that maps to widely separated regions of the genome. The relative orientations of the two reads within the pair is determined by whether the two annealing cDNA molecules were produced from RNA transcribed from the same strand of the genome, or not; correspondingly, half (51%) align to the complementary strand, and the remainder to the same strand. Hence, in accordance with the results of the model oligonucleotide system, intermolecular annealing is observed to occur between cDNA strands during library generation.

Similar fractions of both the ss-cDNA and ds-cDNA samples (1.9%) map to the same strand of the genome to sites within 1 kb of each other. Such an arrangement would be expected as a consequence of intramolecular annealing, with the reverse read sequence originating through reverse complementation of the forward read. That there is no detectable excess of such read pairs when processing ss-cDNA rather than ds-cDNA indicates that intramolecular annealing is not a significant contributory process to library formation in complex transcriptome samples.

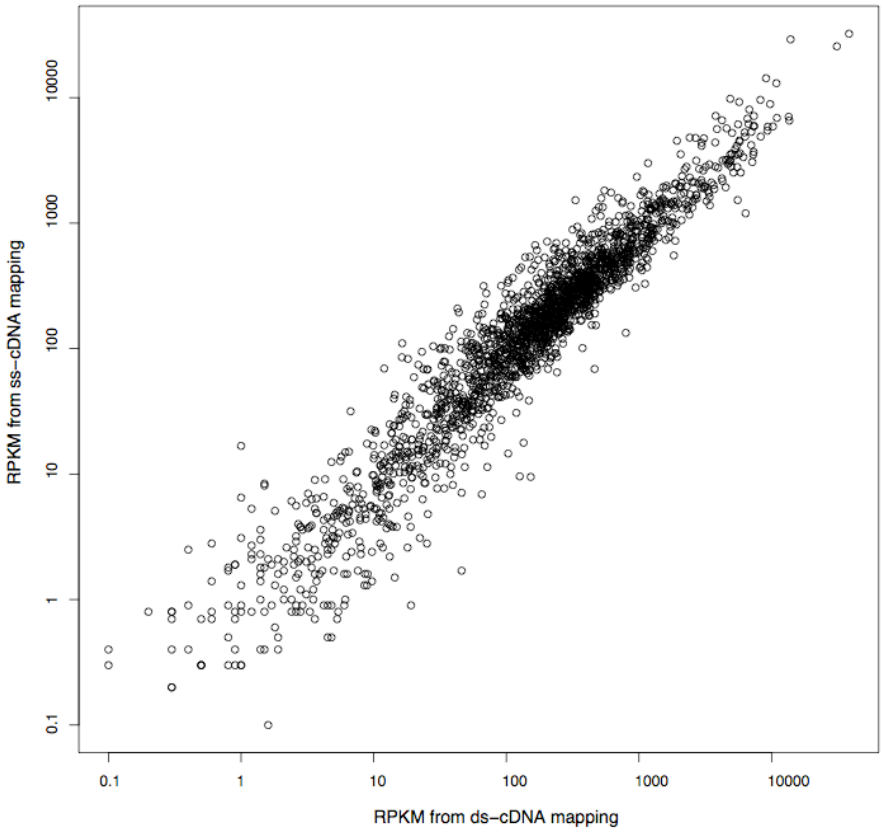

Figure 6.4 Correlation between RPKM values calculated from mapping of ss-cDNA and ds-cDNA to the genome of *S. pneumoniae* ATCC 700669. The RPKM values for each CDS in the genome. with an RPKM greater than zero in both samples, are shown plotted against one another, revealing a tight correlation ($R^2 = 0.84$, $p < 2.2\text{x}10^{-16}$).

6.2.3 ss-cDNA sequencing is quantitative

In order for this technique to be used for quantitative studies of gene expression, the number of reads mapping to a CDS should ideally be directly proportional to its level of transcription. Previous studies have shown that ds-cDNA sequencing is appropriate for quantitative studies of gene expression through comparisons against microarray data (Marioni *et al*., 2008; Mortazavi *et al*., 2008; Nagalakshmi *et al*., 2008; Wilhelm *et al*., 2008). In order to validate ss-cDNA sequencing against the results of ds-cDNA sequencing, the levels of gene expression inferred from the ss-cDNA and ds-cDNA samples were compared. For each annotated CDS, the level of transcription was quantified as the number of reads per kilobase of gene length per million mapped reads (RPKM) (Mortazavi *et al*., 2008). Across the genome, the RPKMs calculated from the two datasets correlated tightly (Pearson correlation, $R^2 = 0.84$, $p < 2.2\text{x}10^{-16}$; Figure 6.4), suggesting that the mechanism by which adapters are attached to ss-cDNA do not distort the proportion of sequence reads originating from each gene.

Hence, simply by not synthesizing the second cDNA strand, information regarding the direction of transcription is retained, without affecting the quantitative nature of the data.

6.3 Discussion

This method represents a novel approach for retaining directional fidelity in transcriptomic data. Sequencing ss-cDNA, a technique simpler than the original RNA-seq protocols as it eliminates the need for second strand cDNA synthesis, minimises the number of steps required to process bacterial RNA samples, which are typically more fragmentary than those of eukaryotes. Evaluation of this technique reveals that it maintains the quantitative aspect of sequencing ds-cDNA, crucial for use in gene expression studies. Hence, despite the requirement that cDNA strand anneal into a duplex to allow adapter to be attached, there does not appear to be an appreciable distortion of the relationship between a gene's level of transcription and the number of sequence reads mapping to it. This seems to be because there is little or no sequence dependence in the annealing of cDNA, which is likely to result from the high concentration of DNA in the end repair reaction and the low temperature at which it is conducted (~23°C). Therefore this approach is a simple way to accurately quantify the level, and direction, of expression across the pneumococcal genome.