doi:10.1093/nar/gkp811

A simple method for directional transcriptome sequencing using Illumina technology

Nicholas J. Croucher, Maria C. Fookes, Timothy T. Perkins, Daniel J. Turner, Samuel B. Marguerat, Thomas Keane, Michael A. Quail, Miao He, Sammey Assefa, Jürg Bähler, Robert A. Kingsley, Julian Parkhill, Stephen D. Bentley, Gordon Dougan and Nicholas R. Thomson*

The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, Cambridgeshire CB10 1SA, UK

Received July 10, 2009; Revised September 14, 2009; Accepted September 15, 2009

ABSTRACT

High-throughput sequencing of cDNA has been used to study eukaryotic transcription on a genome-wide scale to single base pair resolution. In order to compensate for the high ribonuclease activity in bacterial cells, we have devised an equivalent technique optimized for studying complete prokaryotic transcriptomes that minimizes the manipulation of the RNA sample. This new approach uses Illumina technology to sequence single-stranded (ss) cDNA, generating information on both the direction and level of transcription throughout the genome. The protocol, and associated data analysis programs, are freely available from http://www.sanger.ac.uk/Projects/Pathogens/Transcriptome/. We have successfully applied this method to the bacterial pathogens *Salmonella bongori* and *Streptococcus pneumoniae* and the yeast *Schizosaccharomyces pombe*. This method enables experimental validation of genetic features predicted *in silico* and allows the easy identification of novel transcripts throughout the genome. We also show that there is a high correlation between the level of gene expression calculated from ss-cDNA and double-stranded-cDNA sequencing, indicting that ss-cDNA sequencing is both robust and appropriate for use in quantitative studies of transcription. Hence, this simple method should prove a useful tool in aiding genome annotation and gene expression studies in both prokaryotes and eukaryotes.

INTRODUCTION

The advent of high-throughput sequencing technologies has permitted new approaches to exploring functional genomics, including the direct sequencing of complementary cDNA generated from messenger and structural RNAs (RNA-seq). Recent publications have exploited this high-resolution technology to study the RNA population in eukaryotes. These studies have demonstrated a number of dramatic advantages over previous microarray-based techniques, including greater sensitivity, increased dynamic range, reduced background noise and improved precision of mapping data to the genome sequence (1). Furthermore, the results are not biased by array design: whilst most expression arrays have a limited oligonucleotide probe density and are designed on the basis of classical *in silico* genome annotation, RNA-seq has already begun to be used to discover novel genetic features (2,3).

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 double-stranded (ds) cDNA, thereby masking directionality by showing equal signal on both strands (4,5). However, three recent studies using the new sequencing technologies have demonstrated that directionality can be retained. This is 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 (6) or by bisulfite-induced cytosine deamination (7), or by modifying the first cDNA strand prior to second strand synthesis by adding cytosine residues to the 3′-end (8).

*To whom correspondence should be addressed. Tel: +44 1223 494740; Fax: +44 1223 494919; Email: nrt@sanger.ac.uk

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

These techniques allow the transcripts to be mapped back to the reference genomes in a strand-specific manner.

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 (9). Hence, a protocol that minimizes sample manipulation, whilst retaining information on the template strand of transcription, is ideal for studying bacterial gene expression. Here we report a directional RNA-seq method that eliminates both the need for second strand cDNA synthesis and modification of transcripts prior to reverse transcription. We have used this method to study transcriptional patterns within the bacterial pathogens *Salmonella bongori* and *Streptococcus pneumoniae* and the yeast *Schizosacchomyces pombe,* allowing us to capture an unbiased view of the transcriptome of these organisms at given points during their growth *in vitro*. We have also developed a computational pipeline that allows the transcriptome data to be mapped and visualized in the context of the genome annotation using the freely available programme, Artemis (10). This method should greatly enhance our understanding of microbial genome content and gene expression.

MATERIALS AND METHODS

Oligonucleotide model

A DNA oligonucleotide with the sequence AACATCTGC AAG$(N)_{19}$CAGCGACGCATC$(N)_5$, either alone or in the presence of an equimolar amount of a 3′ phosphorylated RNA oligonucleotide of sequence GAUGCGUCGCUG, was diluted to a concentration of 120 nM in Tris–EDTA buffer and subjected to standard Illumina library preparation reactions.

Preparation of RNA samples

RNA samples were extracted from *S. bongori* grown to $OD_{600} = 0.6$ in Luria Broth at 37°C. Samples were extracted from *S. pneumoniae* ATCC 700669 grown to $OD_{600} = 0.8$ in Brain–Heart Infusion (Oxoid) at 37°C. Cultures were fixed through mixing with RNAProtect (Qiagen) in a 1:2 ratio. Cells were then pelleted (4600 rpm, 4°C, 25 min) and lysed through incubation in lysozyme (1 mg ml^{-1}, 4 min, room temperature for *S. bongori*; 15 mg ml^{-1}, 10 min, 37°C for *S. pneumoniae*). RNA was extracted from 100 µl aliquots of the resuspended cells using the SV Total RNA Extraction System (Promega) according to manufacturer's instructions. RNA was extracted from *S. pombe* as described by Lyne *et al.* (11).

Library construction and sequencing

Both single- (ss) and ds-stranded cDNA samples were prepared identically, according to the manufacturer's recommended protocol (12). Briefly, cDNA samples were first sheared by nebulization (35 psi, 6 min). Duplexes were then blunt ended through an end repair reaction using large Klenow fragment, T4 polynucleotide kinase and T4 polymerase. A single 3′ adenosine moiety was added to the cDNA using Klenow exo^{-} and dATP. Illumina adapters, containing primer sites for flow cell surface annealing, amplification and sequencing, were ligated onto the repaired ends of the cDNA. Gel electrophoresis was used to select for DNA constructs 200–250 bp in size, which were subsequently amplified by 18 cycles of PCR with Phusion polymerase. These libraries were denatured with sodium hydroxide and diluted to 3.5 pM in hybridization buffer for loading onto a single lane of an Illumina GA flow cell. Cluster formation, primer hybridization and sequencing reactions were according to the manufacturer's recommended protocol (12).

RNA sample processing and reverse transcription

The 16S and 23S rRNA were removed from the bacterial RNA samples, at a concentration of 0.83 mg ml^{-1}, either by complementary oligonucleotide hybridization (MicrobExpress, Ambion) according to manufacturer's instructions. The 5.8S, 18S and 28S rRNA were removed from the *S. pombe* RNA sample using the mRNA-ONLY kit (Epicentre Biotechnologies) according to manufacturer's instructions. Quality and quantity of RNA were checked both before and after depletion using Agilent 2100 Bioanalyser RNA Nano Chips. Genomic DNA was removed through treatment with amplification grade DNase I (Invitrogen). DNase I treatment was repeated until DNA could not be detected by a genome-specific PCR. RNA was denatured at 70°C for 10 min in the presence of 3 µg µl^{-1} random hexamer primers, then cooled on ice for 5 min. cDNA was then synthesized through reverse transcription using SuperScript III (Invitrogen) at 42°C for 2 h. When specified, actinomycin D (actD) (Sigma) was added to this reaction at a concentration of 6 µg ml^{-1}. Second strand synthesis was performed by incubating first strand cDNA with DNA polymerase I (Invitrogen) and RNase H (Invitrogen) in second strand buffer at 16°C for 2.5 h.

Read mapping and visualization

We aligned all uniquely mapping reads to the genome of *S. bongori* ATCC 43975 (http://www.sanger.ac.uk/Projects/Salmonella/S_bong.embl), *S. pneumoniae* ATCC 700669 (accession number FM11187) (13) or *S. pombe* (accession numbers CU329670-2 and X54421) (14) using SSAHA2 (15). The cigar2Coverage programme was used to convert the SSAHA2 output into a format that can be displayed directly Artemis via the 'Add User Plot' function. This allows the mapped transcriptome data to be viewed, in a strand-specific manner, as a graph relative to the genome annotation. An analogous output can also be generated by mapping reads using MAQ (16) and producing a graph using maqpilup2depth.pl. All aspects of this computational pipeline are freely available from http://www.sanger.ac.uk/Projects/Pathogens/Transcriptome/. Prior to comparing results from different techniques, plots were standardized according to the number of uniquely mapping reads to account for the varying number of reads output from different Illumina runs and the different levels of rRNA depletion. This was achieved by multiplying plots by a constant factor calculated from the ratio of the number of reads in each sample that could be aligned to a single locus in the genome by SSAHA2. Coding sequences (CDS) expression levels were quantified as mean read fold coverage values, which represent the number of sequence reads mapping to the CDS sequence divided by the length of the CDS. Strand bias values were calculated as the density of reads mapping to the CDS in the expected orientation divided by the total density of reads mapping to the CDS.

RESULTS

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 (12). 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.

Using RNA samples prepared from *S. pneumoniae*, we found that sequencing ss-cDNA retained information on the direction of transcription that generated the template RNA molecule. Four mechanisms by which ss-cDNA might undergo correct processing to generate Illumina libraries were proposed (summarized in Figure 1). The first required the ligation of adapters to the ss-cDNA molecules (Figure 1a). This is possible because T4 DNA ligase can ligate ss-DNA molecules, albeit at low efficiency [(17); Figure 1a (iii)], and directionality would be maintained because the second strand is never synthesized [Figure 1a (iv)]. The alternate possibilities involved the formation of duplexes during the end repair reaction (Figure 1b–d). Either annealed RNA fragments [the remains of transcripts that served as templates in the reverse transcription reaction; Figure 1b (ii)] or inter or intramolecular hybridization of cDNA [Figure 1c (ii) and 1d (ii)], was 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 1b (iii)], which cannot be amplified and sequenced by DNA-dependent DNA polymerases. Consequently, only the original ss-cDNA strand would be sequenced [Figure 1b (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 1c (ii) and 1d (ii)]. Hence, sequences with different orientations relative to the original transcript would

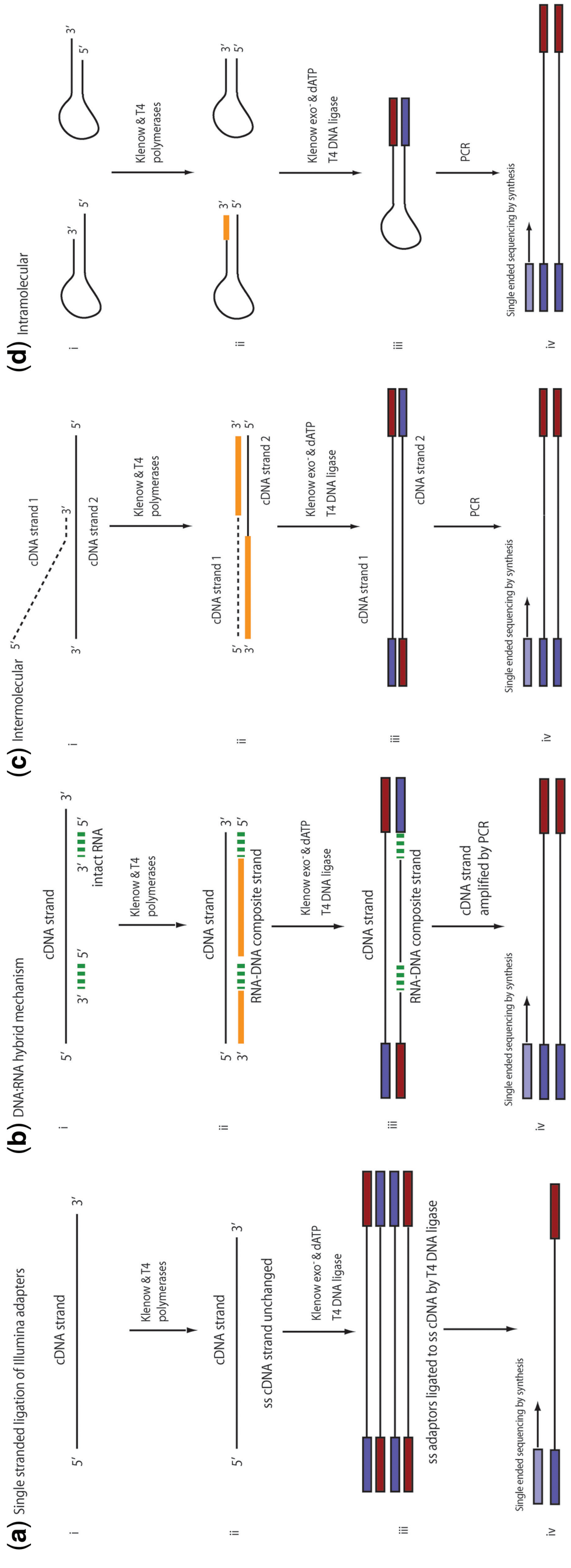

Figure 1. The hypotheses proposed to account for the attachment of Illumina adapted dimers to ss-cDNA: (**a**) attachment of adapters to ss-cDNA, and priming of second strand synthesis by (**b**) RNA fragments, (**c**) intermolecular cDNA annealing and (**d**) intramolecular cDNA annealing.

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 1c (iv) and 1d (iv)].

In order to determine which of these mechanisms described above occurs during library preparation, we designed 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 2a). 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' section).

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. Furthermore, using 54 nt sequence reads, no cases in which the adapters were directly ligated to the unaltered 48 nt oligonucleotide could be identified.

Analysis of 2 162 655 paired 36 nt sequence reads generated from these libraries 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 2c). 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 2b.

In cases where >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 >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 ss DNA 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.

ss-cDNA sequencing retains information on the direction of transcription

An ss-cDNA sequencing protocol was developed and applied to bacterial RNA samples extracted from *S. bongori* and *S. pneumoniae*. Mapping the sequence reads to the reference genomes, and displaying these data as a coverage graph in Artemis, showed that directionality was retained throughout the datasets (Figure 3). For CDS expressed under the conditions tested (in this instance, defined as those having a mean read fold coverage >1), the median strand bias was 97% for both *S. bongori* and *S. pneumoniae*. Hence the results of this method correspond well with the *in silico* genome annotation and provides excellent resolution regarding the direction of transcription throughout the genome.

To investigate whether the same mechanisms were occurring in the mix of bacterial transcripts as in the model oligonucleotide system, an ss-cDNA sample from *S. pneumoniae* was subjected to 54 nt read paired end Illumina sequencing. We identified a dataset of ~3 million reads that existed in pairs where both members could be uniquely mapped to the reference genome. In 60% of cases, these data could be mapped as conventional paired end sequences: the degree of 3′-end processing was sufficiently small not to interfere with alignment to the DNA sequence. Hence, this method can generate data that can be mapped across repetitive regions using read pair information.

In just over half of the remaining cases, the forward and reverse reads aligned to the same strand of the genome. Such chimeric molecules are a result of intramolecular annealing or intermolecular annealing of RNA transcribed from the same DNA strand. In the remaining instances, the reads map to the genome in opposite orientations as excepted, but the distance between them is considerably greater than the insert size. This is a result of intermolecular annealing between RNA strands transcribed in opposite directions. The resulting chimeric cDNA molecule retains the same orientation relative to the direction of transcription throughout its length, but different segments correspond to separate RNA molecules. Hence, this shows that the model system accurately represents the processes occurring in complex transcriptome samples.

In the cases where intermolecular annealing appeared to have occurred, there was no evidence for sequence-specific interactions. The reverse reads were observed to predominately map to highly expressed loci, such as the rRNA operons. Hence, annealing appears mainly to be a function of concentration rather than sequence, suggesting the method should not be biased by sequence-specific interactions.

ss-cDNA sequencing retains information on the level of transcription

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

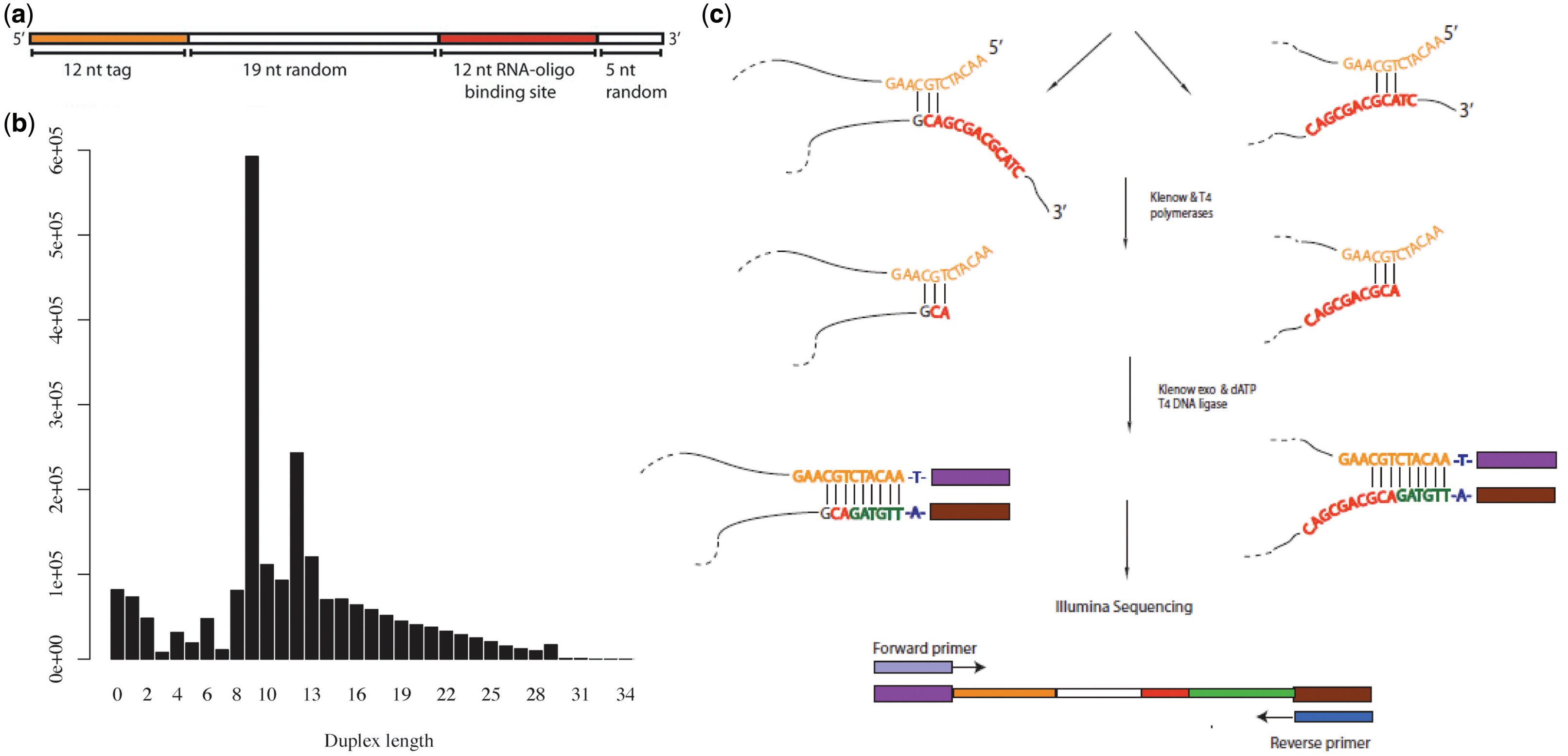

Figure 2. (**a**) Schematic representation of the DNA oligonucleotides from which Illumina libraries were generated. (**b**) Distribution of duplex lengths amongst a sequenced sample of single-stranded DNA oligonucleotides. We extracted reads corresponding to the oligonucleotide by searching the output 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 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.

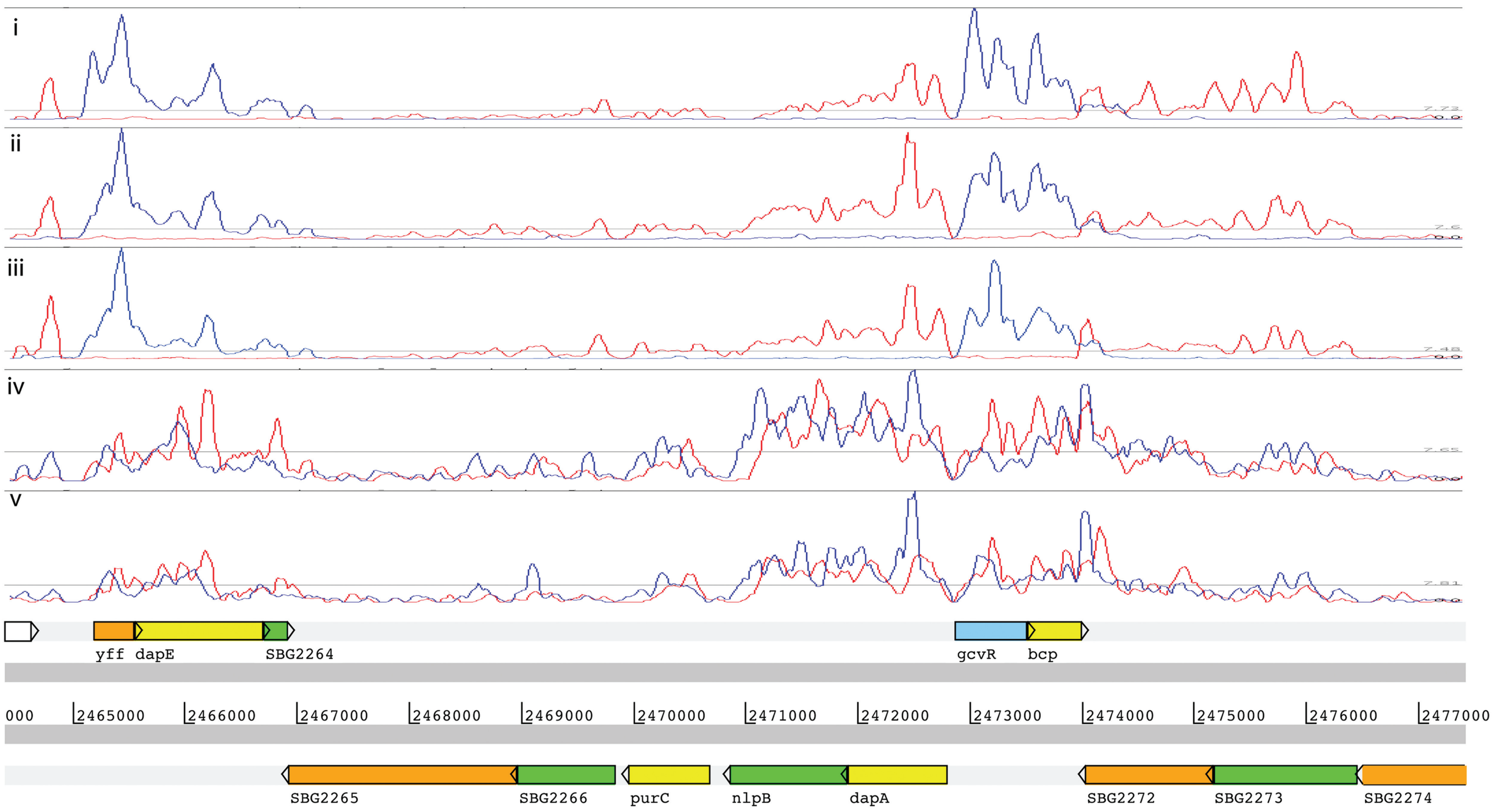

Figure 3. RNA-seq data displayed in Artemis. Mapped RNA-seq data is displayed as a plot showing sequence depth for the forward (blue) and reverse strand (red). The *S. bongori* genome annotation is also shown. The graphs, from the top downwards, represent the result of sequencing (**i**) undepleted ss-cDNA (**ii**) depleted ss-cDNA (**iii**) depleted ss-cDNA with actD present in the reverse transcription reaction (**iv**) ds-cDNA and (v) ds-cDNA with actD present in the reverse transcription reaction.

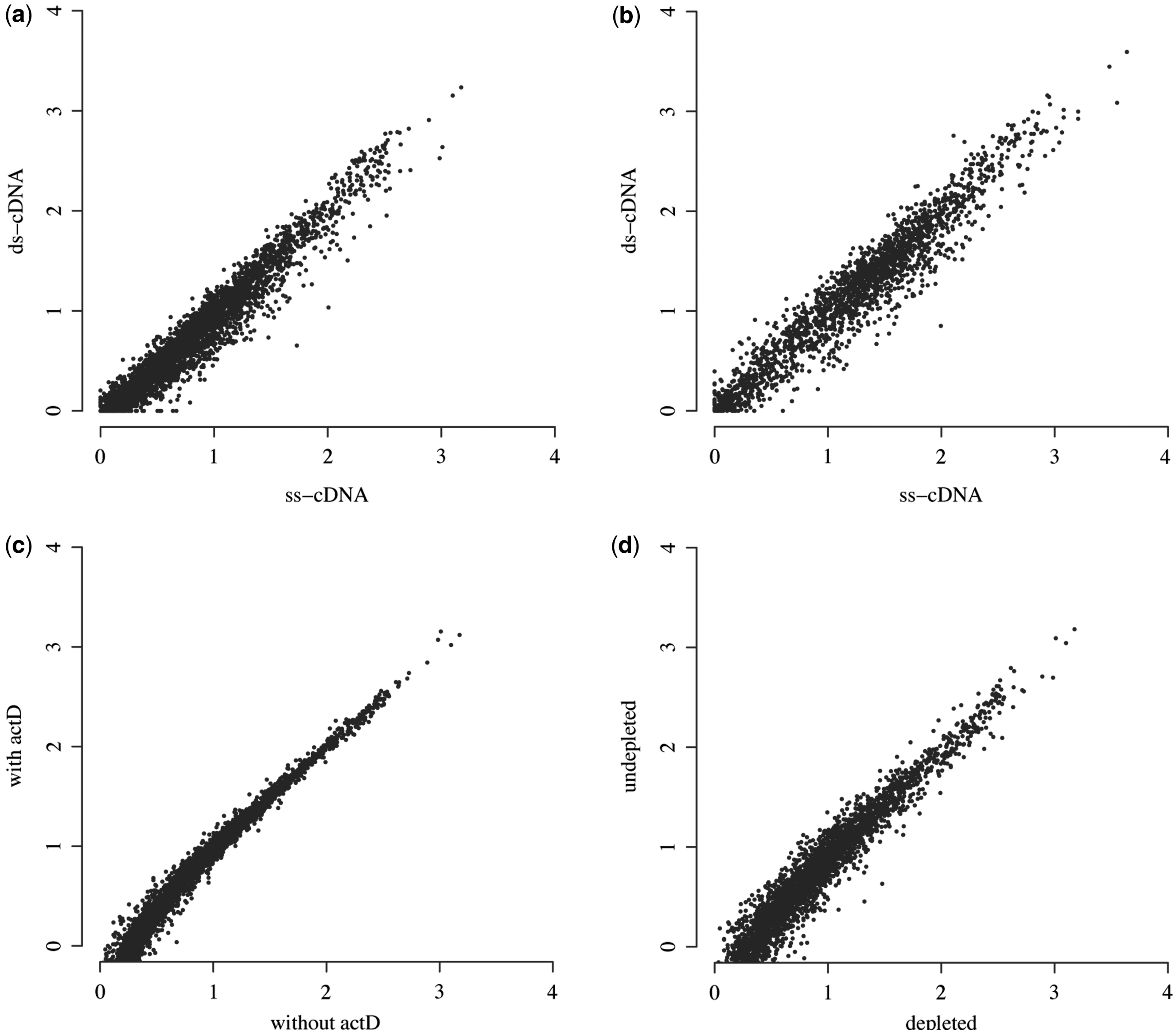

Figure 4. Scatter plots showing the correlation between the genome-wide levels of CDS expression in RNA-seq datasets. Each data point represents the standardized mean fold coverage of a CDS, plotted as log (mean + 1). The top two plots show the correlation between the measured level of CDS expression between technical replicates sequencing ss-cDNA and ds-cDNA for (**a**) *S. bongori* and (**b**) *S. pneumoniae*. The bottom two plots show the impact of modifications to the methodolgy, when applied to *S. bongori*, on the resulting dataset. (**c**) shows that the addition of actD has little impact on the calculated level of transcription across the genome. Similarly, (**d**) shows that depletion of rRNA causes little alteration in the results obtained.

studies of gene expression through comparisons against microarray data (2,4,5,18). In order to validate ss-cDNA sequencing against the results of ds-cDNA sequencing, ss-cDNA and ds-cDNA technical replicates were produced from RNA extracted from both *S. bongori* and *S. pneumoniae* by synthesizing the second DNA strand for only half the sample. As expected, the presence of the complementary strand abrogated the directionality of the data [Figure 3 (iv) and (v)]. However, there was a high degree of correlation in the mean number of reads mapping to each CDS using the two techniques for both species ($R = 0.93$ for *S. bongori* and $R = 0.91$ for *S. pneumoniae*; Figure 4a and b), suggesting that the mechanism by which adapters are attached to ss-cDNA does not cause sample bias. The number of sequence reads per lane for the single- and double-stranded samples was similar for both species, as was the proportion of reads mapping to the genome. This suggests that libraries constructed from ss- or ds-cDNA yield the same quantity and quality of sequence data. 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.

The impact of other variations on the basic RNA-seq technique was also studied using samples extracted from *S. bongori*. One such alteration is the addition of actD to the reverse transcription reaction. This antibiotic has been found to reduce the level of antisense artefacts produced during RNA-dependent DNA synthesis (19). However, no change in the observed level of antisense transcription relative to sense transcription was observed when actD was added. Correspondingly, there was little alteration in the pattern of peaks and troughs observed when the mapped data was viewed in Artemis [Figure 3 (iii)]. These observations suggest that the antisense transcription observed in our datasets may represent a genuine biological phenomenon. Furthermore, the mean expression levels for CDS were strongly correlated between the datasets produced with and without actD ($R = 0.99$; Figure 4c), demonstrating the highly reproducible nature of technical replicates using this technique. Hence, this method should have the sensitivity and precision to detect small changes in gene expression.

The consequences of depleting the 16S and 23S rRNA were also examined. 79.9% of the undepleted sample sequence reads map to the rRNA operon, compared to 65.1% of the depleted sample. This depletion greatly increases the proportion of the sample aligning to the chromosome outside of the rRNA operons, thereby significantly improving the quantification of mRNA and ncRNA expression. Comparison of the calculated levels of expression for each CDS in the genome revealed that depletion had little overall effect ($R = 0.96$; Figure 4d), although there is a more pronounced deviation at low expression levels. Calculations of transcription rates at such low expression levels will always be subject to greater variation due to the difficulties of measuring a continuous variable, such as expression, using discrete data, such as sequence reads. However, this also shows there does not appear to be a problem with saturation of sequencing capacity by the high levels of rRNA, as transcription of genes expressed at a low level can be detected even in undepleted samples. Eliminating the depletion step further reduces the amount of RNA sample manipulation, thereby preserving transcript integrity. The need for depletion is likely to become less as the depth and read length of Illumina sequencing improves.

ss-cDNA sequencing can be used to study spliced eukaryotic transcriptomes

In order to test whether this method would work well for eukaryotes, we sequenced ss- and ds-cDNA generated from *S. pombe* grown to mid-log phase in yeast extract medium. This retained the directionality observed in the bacterial datasets (Figure 5a), with a median strand bias of 99% for genes with a mean fold coverage >1. This improvement relative to the prokaryotic results is probably due to the lower gene density of the eukaryote (0.397 genes kb^{-1} for *S. pombe*, compared with 0.867 genes kb^{-1} for *S. bongori* and 0.899 genes kb^{-1} for *S. pneumoniae*) reducing the level of overlapping transcription. Hence, the strand specificity of this technique appears to reflect the genuine balance of sense and antisense transcription across CDSs, as opposed to any inherent limitation of the method.

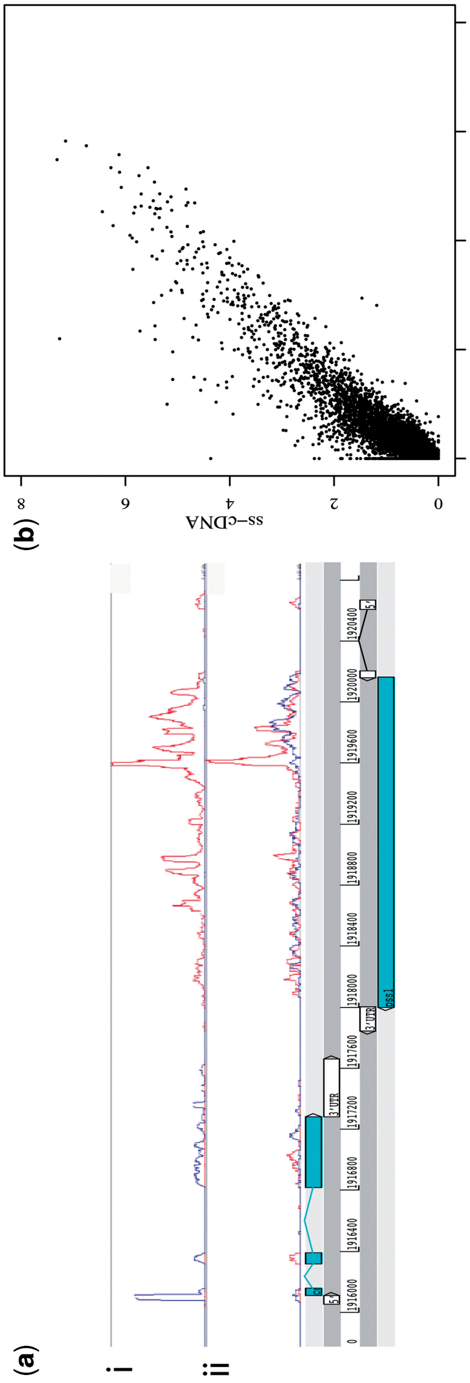

Figure 5. Application of ss-cDNA sequencing to the eukaryote *S. pombe*. **(a)** Sequence reads aligned to the genomic locus on chromosome I containing the *cdc42* and *pss1* genes. The mapped sequence read depth is displayed as in Figure 3, with graphs representing the result of sequencing (i) ss-cDNA and (ii) ds-cDNA. This shows that this technique is adept at delimiting intron–exon boundaries and detecting 5′ and 3′ untranslated regions (UTR; the UTR in the figure are annotated in the publicly available version of the *S. pombe* genome). **(b)** Scatter plot showing the correlation between the measured CDS expression levels obtained by sequencing ss- and ds-cDNA, displayed as in Figure 4.

As shown in Figure 5a, the method provides valuable information on the position and orientation of small exons, which should aid the annotation and validation of complex eukaryotic gene models. Furthermore, the comparison between the single- and double-stranded sample results demonstrates that the technique remains quantitative ($R = 0.83$; Figure 5b). This correlation is weaker than that for the bacterial samples, likely a result of the lower density of mapped reads across the larger genome resulting in greater chance variation in the coverage across a CDS. However, this technique would still appear to be appropriate for studying the changes in levels of transcripts within eukaryotic organisms.

DISCUSSION

Here we have presented the development of a basic toolkit for studying gene expression in a strand-specific manner using Illumina platform sequencing. This relies upon a novel approach for retaining directional fidelity in transcriptomic data by sequencing single-stranded cDNA, a method that is simpler than the original RNA-seq protocols as it abrogates the need for second strand cDNA synthesis. We have also adapted the genome annotation and analysis tool to be able to view this data relative to the genome annotation which allows for easy data interpretation and identification of new genetic features.

Extensive evaluation of this technique reveals that it maintains the quantitative aspect of sequencing ds-cDNA, crucial for use in gene expression studies. Although the attachment of adapters is dependent upon the formation of duplexes through annealing of cDNA strands, this does not appear to distort 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). Whilst a large proportion of the interactions involve cDNA generated from rRNA, due to the abundance of these transcripts, the removal of a significant proportion of these molecules through depletion had very little effect on measured gene expression. This demonstrates that quantification of transcript level is robust to changes in the composition of the cDNA mix, highlighting the absence of any requirement for specific interactions for ss-DNA sequencing library preparation.

The success of this technique in both prokaryotes and eukaryotes, with genomes of differing chromosomal GC content (52.1% for *S. bongori*, 39.5% for *S. pneumoniae,* 36.1% for *S. pombe*), demonstrates that it should be applicable to a wide range of species with little alteration. The only limitation in using this method for the study of organisms with larger genomes is available sequencing capacity, which is continually increasing. 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 also prove to be of great value for complementing and refining current genome annotations in both prokaryotes and eukaryotes.

FUNDING

The Wellcome Trust. Funding for open access charge: The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Graveley,B.R. (2008) Molecular biology: power sequencing. *Nature*, **453**, 1197–1198.
2. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
3. Sultan,M., Schulz,M.H., Richard,H., Magen,A., Klingenhoff,A., Scherf,M., Seifert,M., Borodina,T., Soldatov,A., Parkhomchuk,D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.
4. Nagalakshmi,U., Wang,Z., Waern,K., Shou,C., Raha,D., Gerstein,M. and Snyder,M. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.
5. Wilhelm,B.T., Marguerat,S., Watt,S., Schubert,F., Wood,V., Goodhead,I., Penkett,C.J., Rogers,J. and Bahler,J. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.
6. Lister,R., O'Malley,R.C., Tonti-Filippini,J., Gregory,B.D., Berry,C.C., Millar,A.H. and Ecker,J.R. (2008) Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, **133**, 523–536.
7. He,Y., Vogelstein,B., Velculescu,V.E., Papadopoulos,N. and Kinzler,K.W. (2008) The antisense transcriptomes of human cells. *Science*, **322**, 1855–1857.
8. Cloonan,N., Forrest,A.R., Kolle,G., Gardiner,B.B., Faulkner,G.J., Brown,M.K., Taylor,D.F., Steptoe,A.L., Wani,S., Bethel,G. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*, **5**, 613–619.
9. Rauhut,R. and Klug,G. (1999) mRNA degradation in bacteria. *FEMS Microbiol. Rev.*, **23**, 353–370.
10. Carver,T., Berriman,M., Tivey,A., Patel,C., Bohme,U., Barrell,B.G., Parkhill,J. and Rajandream,M.A. (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, **24**, 2672–2676.
11. Lyne,R., Burns,G., Mata,J., Penkett,C.J., Rustici,G., Chen,D., Langford,C., Vetrie,D. and Bahler,J. (2003) Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data. *BMC Genomics*, **4**, 27.
12. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
13. Croucher,N.J., Walker,D., Romero,P., Lennard,N., Paterson,G.K., Bason,N.C., Mitchell,A.M., Quail,M.A., Andrew,P.W., Parkhill,J. *et al.* (2009) Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae*$^{\mathrm{Spain23F}}$ ST81. *J. Bacteriol*, **191**, 1480–1489.
14. Wood,V., Gwilliam,R., Rajandream,M.A., Lyne,M., Lyne,R., Stewart,A., Sgouros,J., Peat,N., Hayles,J., Baker,S. *et al.* (2002) The genome sequence of *Schizosaccharomyces pombe*. *Nature*, **415**, 871–880.

15. Ning,Z., Cox,A.J. and Mullikin,J.C. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.
16. Li,H., Ruan,J. and Durbin,R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.
17. Kuhn,H. and Frank-Kamenetskii,M.D. (2005) Template-independent ligation of single-stranded DNA by T4 DNA ligase. *FEBS J.*, **272**, 5991–6000.
18. Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
19. Perocchi,F., Xu,Z., Clauder-Munster,S. and Steinmetz,L.M. (2007) Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res.*, **35**, e128.