FRT-seq: amplification-free, strand-specific transcriptome sequencing

Lira Mamanova[1,3], Robert M Andrews[1,3], Keith D James[1], Elizabeth M Sheridan[1], Peter D Ellis[1], Cordelia F Langford[1], Tobias W B Ost[2], John E Collins[1] & Daniel J Turner[1]

We report an alternative approach to transcriptome sequencing for the Illumina Genome Analyzer, in which the reverse transcription reaction takes place on the flowcell. No amplification is performed during the library preparation, so PCR biases and duplicates are avoided, and because the template is poly(A)$^+$ RNA rather than cDNA, the resulting sequences are necessarily strand-specific. The method is compatible with paired- or single-end sequencing.

Analysis of complementary DNA by next-generation sequencing (RNA-seq) enables us to build an accurate picture of active transcriptional patterns in an organism[1]. The ideal RNA-seq protocol would be accurate, strand-specific and quantitative across a wide dynamic range, compatible with paired-end sequencing, and would detect antisense transcripts unambiguously[2,3]. Some, but not all, of these requirements are met by existing methodologies. Neither polydeoxythymine priming nor random hexamer priming yield the strand-specific information that is essential for comprehensive annotation of the transcriptome[4] and identification of antisense transcription[5,6]. Consequently, several strand-specific approaches to RNA-seq have been developed[3,7–11], and, with the exception of Helicos' 'direct RNA sequencing' approach[3], in each case the cDNA is amplified by PCR, an inherently biased procedure[12]. PCR-amplified libraries can have reduced complexity compared to the total mRNA pool because different fragments tend to amplify with unequal efficiency. This causes drop-out of some RNA species and excessive amplification of others; such PCR duplicates are difficult to distinguish from genuinely abundant RNA species. To overcome these limitations, it is preferable to avoid library amplification altogether[3,13].

Here we report an RNA-seq approach for the Illumina Genome Analyzer in which reverse transcription takes place on the flowcell surface ('FRT-seq'; **Supplementary Fig. 1**, **Supplementary Table 1** and Online Methods). The method is strand-specific, amplification-free, compatible with paired-end sequencing and avoids any ambiguities that might arise from the addition of nontemplated nucleotides by the reverse transcriptase[14]; in our method, the latter will occur at the 3′ end of the adaptor sequence, and these nucleotides are therefore not sequenced.

To evaluate the performance of reverse transcriptase in the flowcell environment, we exploited the ability of this enzyme to use DNA as well as RNA as a template and performed first-strand synthesis on a PCR-amplified PhiX DNA library (Illumina). We then generated clusters and sequenced following the standard protocols. We calculated sequence coverage in 10 base pair (bp) bins and compared it to that obtained from the same library following the standard protocol, in which Taq polymerase performs first-strand synthesis. The two enzymes performed similarly (**Supplementary Fig. 2a**). We then divided the PhiX genome (mean G+C content = 44.7%) into low (<44.7%) and high (>44.7%) G+C content bins and calculated Spearman correlations between sequence coverage and G+C content for both bins using 20–210 bp window sizes at 10 bp intervals (**Supplementary Table 2**). We found a moderate positive correlation for both enzymes with the low G+C content bin, indicating underrepresentation of low G+C content sequences in the mapped sequence data, and a much weaker correlation at high G+C content. The correlation at low G+C content was stronger for Taq polymerase than for reverse transcriptase. Additionally, we found a moderate negative correlation between coverage difference for the two enzymes and G+C content (**Supplementary Table 2** and **Supplementary Fig. 2b**). Together, this confirmed that the reverse transcriptase was no less efficient at seeding clusters than Taq polymerase. There was no discernible difference in the percentage of reads mapping to the PhiX genome or in the read quality of the sequences produced with either enzyme (data not shown).

We prepared two FRT-seq libraries using a human placental poly(A)$^+$ RNA sample (Clontech) and prepared one paired-end flowcell for each library. We sequenced each library for 2 × 37 cycles on an Illumina Genome Analyzer, generating 3.3 and 3.5 Gb of sequence. For comparison, we prepared two standard RNA-seq libraries from the same sample, using Illumina's random priming protocol and generated 1.6 Gb and 0.5 Gb of sequence. We mapped all reads to annotated genes from the Ensembl database[15], normalized read counts and calculated Pearson correlations between libraries and between all lanes (**Supplementary Table 3**). FRT-seq was highly reproducible, with a Pearson correlation of 0.993 between the datasets obtained from separate libraries (**Fig. 1**). Correlations between individual lanes from the same FRT-seq library were also very strong (0.998–1.000), indicating that the slight discrepancy that exists is due to sampling bias rather than stochastic systematic biases in the library preparation and reverse transcriptase reactions. The correlation between standard RNA-seq libraries was very strong between lanes from

[1]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [2]Illumina Inc., Chesterford Research Park, Little Chesterford, Essex, UK. [3]These authors contributed equally to this work. Correspondence should be addressed to D.J.T. (djt@sanger.ac.uk).

RECEIVED 17 AUGUST 2009; ACCEPTED 14 DECEMBER 2009; PUBLISHED ONLINE 17 JANUARY 2010; DOI:10.1038/NMETH.1417

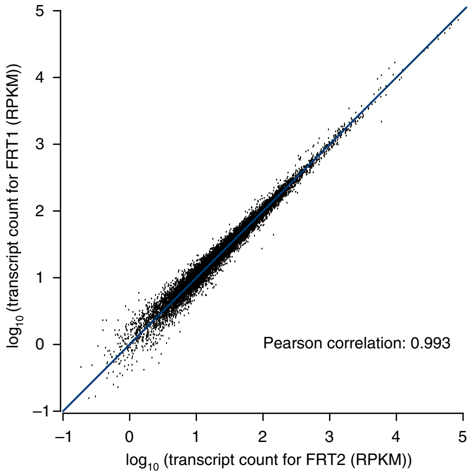

Figure 1 | Correlation plots for FRT-seq libraries. We plotted sequence data obtained from two FRT libraries, FRT1 and FRT2, prepared from the same poly(A)$^+$ RNA sample. All reads were mapped to annotated genes from the Ensembl database. Shown are normalized read counts. RKPM, reads per kilobase of sequence per million reads.

the same library (~1.000) but weaker between libraries (0.866), presumably reflecting stochastic amplification biases incurred during the library-preparation PCR (**Supplementary Fig. 3a–f**). The comparatively poor technical reproducibility is not necessarily representative of the Illumina standard RNA-seq library preparation method *per se* but indicates that care must be taken to ensure consistent results throughout the library preparation. Alternative approaches to RNA-seq have been reported[8,11], in which very good technical reproducibility has been demonstrated (Pearson correlations = 0.98–0.99), but to which our FRT-seq method still compares favorably.

The percentage of duplicate reads was low for the two FRT-seq libraries (6.1% and 7.2% for libraries FRT1 and FRT2, respectively; **Supplementary Table 4**) but was higher and varied appreciably between standard libraries (94.1% and 39.7% for libraries STD1 and STD2, respectively). Regardless of the cause, duplicate sequences will be more prevalent for more abundant transcripts. The frequency of positions at which we observed one or more duplicate sequences was 2.2% for each FRT-seq library, and 74.2% and 13.9% for standard RNA-seq libraries. The fragmentation methods were identical between standard and FRT-seq libraries, indicating that the observed difference in duplication frequency between library types was largely due to PCR bias.

To evaluate the influence of template G+C content on read depth, we divided sequences obtained by both methods into bins of G+C content for the entire mapped fragment. Sequences generated by the PCR-based standard method appeared to be biased away from lower G+C content toward a more neutral G+C content, compared to the FRT-seq data (**Supplementary Fig. 4a**,**b**). This mirrored the effect of PCR on genomic DNA[12].

For both methods, we assessed the evenness of sequence coverage along the length of genes, both in their entirety and across individual exons (**Supplementary Fig. 5**). Representation was more even in the FRT-seq libraries compared to standard libraries.

To determine how closely the FRT-seq data correlated with microarray-derived expression data, we ran the poly(A)$^+$ RNA sample on Human Expression BeadChips (Illumina) in triplicate and compared the results to transcript counts obtained from FRT-seq and standard RNA-seq libraries (**Supplementary Fig. 6**). The Pearson correlation between transcription levels derived from array data and those obtained from FRT-seq (0.676) was substantially better than between array data and standard RNA-seq library (0.482), indicating that FRT-seq is the more quantitative approach. Correlations between individual RNA-seq libraries and array data differed slightly, reflecting differences in library quality (0.423 and 0.493 for libraries STD1 and STD2, respectively), whereas those between libraries FRT1 and FRT2 were in close agreement (0.676 and 0.674, respectively). These correlations are lower than has been reported previously for standard libraries[16]. The arrays used in our study, Illumina HumanWG-6 v3 Expression BeadChips, have been designed to detect mainly the 3′ end of transcripts, whereas the FRT-seq data represent entire transcripts, making the two types of data difficult to normalize and hindering direct comparison. Additionally, the background signal of arrays may contribute to the failure of sequence and array data to correlate perfectly[16]. Nevertheless, our results revealed that PCR amplification bias is a major cause of discordance between array and sequence data. Lists of called genes and read counts from both FRT-seq and standard libraries are available at ftp://ftp.sanger.ac.uk/pub/transseq/.

Sequences obtained using FRT-seq are necessarily strand-specific. To demonstrate this, we mapped all reads to the US

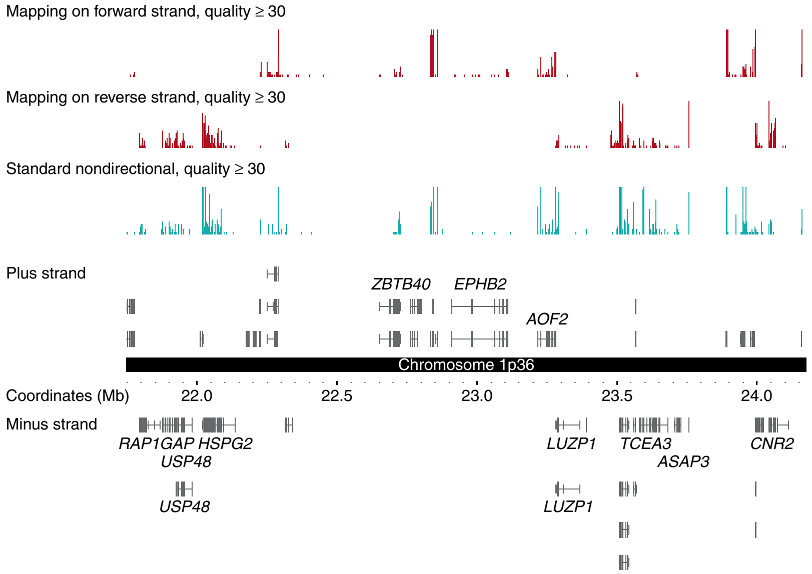

Figure 2 | Strand-specificity of FRT-seq. Sequences generated by FRT-seq were mapped to the human genome. The .wig files are displayed in modified Integrated Genome Browser format (red). For comparison, sequences made using the standard RNA-seq library preparation protocols and flowcell amplification are shown (blue). Below is a representation of the region of human chromosome 1p36 with genes shown in Ensembl together with the strands from which the transcript was produced.

National Center for Biotechnology Information (NCBI) build 36 version of the human genome and created forward and reverse strand .wig files for viewing in the Integrated Genome Browser (**Fig. 2**). The majority of reads produced by FRT-seq mapped with the first read corresponding to the sense strand and the second read corresponding to the antisense strand. For the standard, nondirectional libraries, reads mapped to both strands with similar frequency (**Supplementary Figs. 7a**,**b** and **Supplementary Table 5**).

An appreciable percentage of reads mapped with the second read corresponding to the sense strand and the first read corresponding to the antisense strand (2.55%), compared to the gene annotation. This is the least likely combination to arise from chimerism but would be expected for antisense transcripts. The value is highly consistent between the different libraries and between different lanes within the same library. Approximately 40% of sequences mapping within the 1 kb upstream regions were in the antisense orientation, compared to <3% overall, indicating significant enrichment of antisense reads in the promoter regions (two-tailed $P < 0.0001$, Fisher's exact test), consistent with them being genuine antisense transcripts[6] (**Supplementary Table 6**).

A reasonably high proportion of sequences mapped to intergenic regions, both for FRT-seq and standard RNA-seq libraries. When we performed FRT-seq on zebrafish ovary poly(A)$^+$ RNA, mapping to the zebrafish reference genome version Zv8, very few intergenic sequences were evident (**Supplementary Fig. 8**). It is possible that the commercial human placental poly(A)$^+$ RNA sample may have been contaminated with DNA or unspliced RNA, or that the human gene annotations in the Ensembl database are incomplete[16].

In conclusion, FRT-seq enables amplification-free RNA-seq and generates sequences that are strand-specific and compatible with paired-end sequencing, and presents no opportunity for the formation of intermolecular priming artifacts. We anticipate that this method will be the method of choice for transcriptome sequencing in the future.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Accession codes. European Nucleotide Archive: ERA000183 (sequence data).

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank N. Bason and M. Quail for preparing the standard RNA-seq library, J. Burton for coordinating the sequencing of standard RNA-seq libraries, and J. Ule and M. Wollerton for advice on RNA ligation. This work was supported by the Wellcome Trust grant (WT079643).

AUTHOR CONTRIBUTIONS

D.J.T. and T.W.B.O. devised the project; L.M. and D.J.T. devised the experimental protocols; L.M. and E.M.S. planned and carried out laboratory work; R.M.A. and K.D.J. performed data analysis; P.D.E. performed microarray work; and C.F.L. oversaw analysis and array work. D.J.T., J.E.C. and L.M. wrote the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Wang, Z., Gerstein, M. & Snyder, M. *Nat. Rev. Genet.* **10**, 57–63 (2009).
2. Wu, J.Q. *et al. Genome Biol.* **9**, R3 (2008).
3. Ozsolak, F. *et al. Nature* **461**, 814–818 (2009).
4. David, L. *et al. Proc. Natl. Acad. Sci. USA* **103**, 5320–5325 (2006).
5. Carninci, P. *et al. Science* **309**, 1559–1563 (2005).
6. Katayama, S. *et al. Science* **309**, 1564–1566 (2005).
7. Lister, R. *et al. Cell* **133**, 523–536 (2008).
8. Cloonan, N. *et al. Nat. Methods* **5**, 613–619 (2008).
9. Croucher, N.J. *et al. Nucleic Acids Res.* **37**, e148 (2009).
10. He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N. & Kinzler, K.W. *Science* **322**, 1855–1857 (2008).
11. Parkhomchuk, D. *et al. Nucleic Acids Res.* **37**, e123 (2009).
12. Kozarewa, I. *et al. Nat. Methods* **6**, 291–295 (2009).
13. Lipson, D. *et al. Nat. Biotechnol.* **27**, 652–658 (2009).
14. Chen, D. & Patton, J.T. *Biotechniques* **30**, 574–580, 582 (2001).
15. Hubbard, T.J. *et al. Nucleic Acids Res.* **37**, D690–D697 (2009).
16. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. & Gilad, Y. *Genome Res.* **18**, 1509–1517 (2008).

ONLINE METHODS

Fragmentation. We fragmented 250 ng of a human placental poly(A)$^+$ RNA (Clontech) by metal ion hydrolysis (Ambion), ethanol precipitated it and dephosphorylated it nonspecifically using Antarctic phosphatase (New England Biolabs).

3′ adaptor ligation. We ligated an adaptor onto the 3′ end of the RNA using T4 RNA ligase 1 (New England Biolabs). This adaptor matched the Illumina reverse PCR primer[17] in sequence but consisted of 20 RNA nucleotides at the 5′ terminus, and the remaining bases were DNA nucleotides. The adaptor was phosphorylated at the 5′ end and blocked by dideoxy cytosine at the 3′ end.

To remove excess adaptor, we ran ligation products in a denaturing acrylamide gel, retaining the portion of the gel containing oligonucleotides larger than the adaptor and discarding the portion containing oligonucleotides equal to or smaller than the adaptor. We then extracted oligonucleotides from the gel.

5′ adaptor ligation and cleanup. We phosphorylated templates at the 5′ (unligated) ends using polynucleotide kinase (New England Biolabs) and ligated to them a 5′ adaptor. This adaptor matched the Illumina forward PCR primer[17] in sequence and had a similar composition to the 3′ adaptor described above. We cleaned the fully ligated product using SPRI beads (Agencourt Bioscience Corporation).

Product validation and quantification. We quantified the library using a Bioanalyzer 2100 RNA chip (Agilent Technologies Inc), following the manufacturer's protocol.

Reverse transcription and sequencing. We reverse-transcribed ligated RNA libraries on an Illumina flowcell, in a manner that is analogous to the first-strand synthesis of DNA templates, but using reverse transcriptase (Invitrogen), and performed the cluster amplification and sequencing reactions following the manufacturer's recommended protocol for standard templates. We prepared a reverse transcriptase buffer mix in a total volume of 2.0 ml (1× SuperScript II RT buffer (Invitrogen), 0.017 M DTT and 2 M betaine (Sigma)).

We filtered this buffer through a 0.2-μm filter and collected it in a 15 ml Falcon tube. We then added 187.5 μl of 40 U μl^{-1} RNaseOut (Invitrogen) and mixed thoroughly. We pipetted 90 μl of this mixture into each tube in an 8-tube strip labeled D and stored it at 4 °C until needed.

Next, we prepared 1,120 μl of reverse transcriptase enzyme mix: 1,058.4 μl reverse transcriptase buffer mix from preceding step, 1 U μl^{-1} SuperScript II reverse transcriptase (Invitrogen) and 500 μM dNTP mix (Illumina). We pipetted 140 μl of this mixture into each tube of an 8-tube strip labeled E and stored at 4 °C until needed.

Cluster Station steps. We pipetted 140 μl of Illumina Hybridization Buffer (HT1) into each tube of an 8-tube strip labeled A and pumped through a paired end flowcell with the following Illumina Cluster Station parameters: aspiration rate = 60 and volume = 120.

We diluted RNA libraries to 500 pM (based on Agilent Bioanalyzer 2100 RNA chip quantification) using Illumina Hybridization Buffer (HT1) and pipetted 90 μl to each tube of an 8-tube strip labeled B. We pumped this through the flowcell with following conditions: (i) TempRamp temperature = 96, rate = 1; (ii) pump reagent with aspiration rate = 15, volume = 75; (iii) pump reagent with aspiration rate = 100, volume = 10; (iv) wait duration = 30,000; and (v) TempRamp temperature = 40, rate = 0.05.

We pipetted 140 μl of Illumina Wash Buffer (HT2) into each tube of an 8-tube strip labeled C and pumped through the flowcell with following conditions: (i) pump reagent with aspiration rate = 15, volume = 75; and (ii) TempRamp temperature = 42, rate = 1.

We pumped the reverse transcriptase buffer mix (prepared above; labeled D) through the flowcell with aspiration rate = 15 and volume = 70.

We pumped the reverse transcriptase enzyme mix (prepared above; labeled E) through the flowcell with: (i) pump reagent with aspiration rate = 60, volume = 95; (ii) TempRamp temperature = 42, rate = 1; (iii) wait duration = 500,000; (iv) pump reagent with aspriation rate = 15, volume = 10; (v) wait duration = 440,000; (vi) pump reagent with aspiration rate = 15, volume = 10; (vii) wait duration = 440,000; (vii) pump reagent with aspiration rate = 15, volume = 10; (viii) wait duration = 440,000; (ix) pump reagent with aspiration rate = 15, volume = 10; (x) wait duration = 440,000; (xi) pump reagent with aspiration rate = 15, volume = 10; (xii) wait duration = 440,000; (xiii) TempRamp temperature = 70, rate = 1; (xiv) wait duration = 900,000; (xv) TempRamp temperature = 37, rate = 1.

We pipetted 150 μl of 0.1 N NaOH into each tube of an 8-tube strip labeled F and pumped through the flowcell with following conditions: (i) pump reagent with aspiration rate = 15, volume = 120; and (ii) TempRamp temperature = 37, rate = 1.

We pipetted 150 μl of TE (pH 8.0) into each tube of an 8-tube strip labeled G and pumped through the flowcell with following conditions: (i) pump reagent with aspiration rate = 15, volume = 120; and (ii) TempRamp temperature = 37, rate = 1.

We then removed the hybridization manifold, connected an amplification manifold and followed the rest of the standard Illumina amplification recipe without changes. All other procedures were performed following Illumina's recommended protocols.

Read mapping. We filtered read pairs for poly(N) and poly(A) sequences, and mapped to both the human genome sequence (assembly NCBI36) and a nonredundant set of Ensembl gene sequences with corresponding RefSeq entries (downloaded from BioMART[18]). We mapped read pairs using MAQ[19] and removed those for which either or both reads in the pair did not map. We imposed a mapping score cutoff of 30.

Expression arrays. We amplified, in triplicate, 12.5 ng of human placental poly(A) RNA using the Illumina TotalPrep-96 RNA Amplification kit (Applied Biosystems) according to the manufacturer's instructions. We applied 1,500 ng of biotinylated cRNA to an Illumina HumanWG-6 v3 Expression BeadChip (Illumina) for each replicate and hybridized overnight at 58 °C. We washed, detected and scanned chips according to the manufacturer's instructions. We imported scanner output files into BeadStudio software (Illumina) and output nonnormalized, probe-level data text files for subsequent analysis.

doi:10.1038/nmeth.1417

We vst-transformed and quantile-normalized data[20] in Bioconductor (http://www.bioconductor.org/) using the Lumi (http://www.bioconductor.org/packages/2.0/bioc/html/lumi.html) and Limma[21] packages. Expressed genes were called when the microarray probe reported a brightness above background in all three replicates (detection threshold $P = 0.05$). We united probes with the companion gene annotation using the Annotate package (http://www.bioconductor.org/packages/release/bioc/html/annotate.html).

Standard RNA-seq libraries. We produced standard libraries in accordance with Illumina's RNA-seq V3.5 protocol.

Sequenced transcriptome analysis. We normalized the number of read pairs mapping per gene by gene length and number of reads in the run, yielding a value of reads per kilobase of coding sequence per million mapped reads (RPKM). Additionally, we transformed and normalized data a second time by the same method used to generate the microarray data, for the microarray correlation analyses.

Correlation analysis. We computed lane-to-lane Pearson correlations from lane RPKM values and lane-to-microarray Pearson correlations from vsn-transformed[22], quantile-normalized values for both datasets.

17. Bentley, D.R. *et al. Nature* **456**, 53–59 (2008).
18. Smedley, D. *et al. BMC Genomics* **10**, 22 (2009).
19. Li, H., Ruan, J. & Durbin, R. *Genome Res.* **18**, 1851–1858 (2008).
20. Yang, Y.H. *et al. Nucleic Acids Res.* **30**, e15 (2002).
21. Smyth, G.K. *Stat. Appl. Genet. Mol. Biol.* **3**, 3 (2004).
22. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. *Bioinformatics* **18** (Suppl. 1), S96–S104 (2002).

doi:10.1038/nmeth.1417