RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase

RYAN M. NOTTINGHAM,[1,2,3] DOUGLAS C. WU,[1,2,3] YIDAN QIN,[1,2] JUN YAO,[1,2] SCOTT HUNICKE-SMITH,[1] and ALAN M. LAMBOWITZ[1,2]

[1]Institute for Cellular and Molecular Biology, [2]Department of Molecular Biosciences, University of Texas at Austin, Austin, Texas, 78712, USA

ABSTRACT

Next-generation RNA sequencing (RNA-seq) has revolutionized our ability to analyze transcriptomes. Current RNA-seq methods are highly reproducible, but each has biases resulting from different modes of RNA sample preparation, reverse transcription, and adapter addition, leading to variability between methods. Moreover, the transcriptome cannot be profiled comprehensively because highly structured RNAs, such as tRNAs and snoRNAs, are refractory to conventional RNA-seq methods. Recently, we developed a new method for strand-specific RNA-seq using thermostable group II intron reverse transcriptases (TGIRTs). TGIRT enzymes have higher processivity and fidelity than conventional retroviral reverse transcriptases plus a novel template-switching activity that enables RNA-seq adapter addition during cDNA synthesis without using RNA ligase. Here, we obtained TGIRT-seq data sets for well-characterized human RNA reference samples and compared them to previous data sets obtained for these RNAs by the Illumina TruSeq v2 and v3 methods. We find that TGIRT-seq recapitulates the relative abundance of human transcripts and RNA spike-ins in ribo-depleted, fragmented RNA samples comparably to non-strand-specific TruSeq v2 and better than strand-specific TruSeq v3. Moreover, TGIRT-seq is more strand specific than TruSeq v3 and eliminates sampling biases from random hexamer priming, which are inherent to TruSeq. The TGIRT-seq data sets also show more uniform 5′ to 3′ gene coverage and identify more splice junctions, particularly near the 5′ ends of mRNAs, than do the TruSeq data sets. Finally, TGIRT-seq enables the simultaneous profiling of mRNAs and lncRNAs in the same RNA-seq experiment as structured small ncRNAs, including tRNAs, which are essentially absent with TruSeq.

Keywords: diagnostics; high-throughput sequencing; small noncoding RNA; transcriptome; tRNA; TruSeq

INTRODUCTION

Next-generation RNA sequencing (RNA-seq) is a powerful tool for analyzing the transcriptomes of cells and tissues, as well as changes in transcriptomes due to environmental stimuli, differentiation, infection, and pathogenesis (Wang et al. 2009; Ozsolak and Milos 2011; Westermann et al. 2012; Vikman et al. 2014). Current RNA-seq methods differ in the protocols used for RNA sample preparation, reverse transcription, addition of RNA-seq adapters, sequencing platforms, and bioinformatic pipelines. Several large-scale studies have addressed the variability between different RNA-seq methods and platforms (Levin et al. 2010; Li et al. 2014; SEQC/MAQC-III Consortium 2014). Overall, these studies found broad agreement in RNA profiles obtained by different methods, but with variations in the quantitation of transcript abundance reflecting different biases inherent in each method.

The overall workflow for transcriptional profiling through RNA-seq is similar among various protocols. First, total RNA from cells or tissues is typically enriched for desired classes of RNA, e.g., by ribo-depletion to remove large and small rRNAs, poly(A)-selection to enrich for mRNAs and other polyadenylated transcripts, or size-selection for small noncoding RNAs (ncRNA). Second, RNA-seq libraries are prepared by converting RNA sequences into cDNAs by using a reverse transcriptase (RT), with adapter sequences for the desired sequencing platform added either before or after cDNA synthesis by using an RNA or DNA ligase. Third, cDNAs are amplified by PCR and then size-selected to remove sequencing adapter dimers and/or enrich for specific RNA size classes. Finally, the cDNA library is sequenced and mapped to a reference genome or transcriptome, enabling analysis of transcriptional profiles including changes in RNA expression levels and alternative splicing under different conditions.

[3]**These authors contributed equally to this work.**

Corresponding author: lambowitz@austin.utexas.edu

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.055558.115.

RNA-seq libraries for long RNAs, such as mRNAs and lncRNAs, are typically prepared separately from those for small ncRNAs, such as miRNAs. In widely used methods, mRNAs and lncRNAs are fragmented in order to increase their 5′ to 3′ coverage and accommodate the shorter read lengths of commonly used sequencing platforms. Such fragmentation can lead to length biases due to increased sampling of longer RNAs (Oshlack and Wakefield 2009). As an alternative to RNA fragmentation, cDNAs of polyadenylated transcripts can be synthesized directly from intact, total RNA samples by using an oligo(dT) primer annealed to their 3′ ends, followed by second-strand synthesis and DNA fragmentation. However, conventional protocols using this method underrepresent 5′ RNA sequences due to the low processivity of the retroviral RTs currently used for cDNA synthesis and do not detect nonpolyadenylated transcripts (Mohr et al. 2013).

RNA-seq libraries of small ncRNAs are prepared from size-selected RNA fractions by ligation of an adapter to the 3′ RNA end, enabling reverse transcription from a DNA primer annealed to that adapter. A second adapter for PCR amplification is added either to the 5′ RNA end prior to reverse transcription or to the 3′ cDNA end after reverse transcription. The Illumina Small RNA Kit uses adapters that target small ncRNAs, such as miRNAs and snoRNAs, with 5′ monophosphate and 3′ OH termini (Illumina product literature) followed by size-selection of cDNAs prior to sequencing. This and similar methods using sequential RNA-seq adapter ligation are typically time consuming, have well-documented sequence and secondary structure biases, and can yield undesirable side products (Hafner et al. 2011; Zhuang et al. 2012; Raabe et al. 2014).

The preservation of strand information in RNA-seq has become increasingly important in light of the growing appreciation of the extent and significance of antisense transcription (Pelechano and Steinmetz 2014). The commonly used non-strand-specific Illumina TruSeq v2 method starts with fragmented RNA (either ribo-depleted or poly(A)-selected) and uses random-hexamer priming for cDNA synthesis by SuperScript II RT, a derivative of the retroviral M-MLV RT. Second-strand synthesis is performed via nick translation by using RNase H and DNA polymerase I, and the resulting double-stranded cDNAs are 3′ A-tailed and ligated to double-stranded DNA adapters at both ends, thus erasing strand information. Widely used strand-specific methods involve the incorporation of dUTP during second-strand synthesis, yielding a dU-containing DNA strand that is either degraded by UDP-N-glycosylase/USER (Parkhomchuk et al. 2009; NEBNext Ultra Directional RNA Library Prep Kit) or excluded by using a PCR polymerase that stalls at dU sites in DNA (Illumina TruSeq v3). In both the NEBNext and Illumina protocols, antisense artifacts from re-copying of the cDNA by the retroviral RT are mitigated by adding actinomycin D, which inhibits DNA-dependent DNA synthesis (Ruprecht et al. 1973; Perocchi et al. 2007). Sequential RNA-seq adapter ligation, which is used for small RNA-seq (see above), also preserves strand information and can be used with intact or fragmented longer transcripts (Illumina Directional mRNA-seq). A comparison of numerous strand-specific methods in profiling the yeast transcriptome concluded that dUTP incorporation had the most advantages but adapter ligation to fragmented RNAs also performed well, albeit with the same issues as for small ncRNA-seq (see above; Levin et al. 2010).

All current RNA-seq methods use retroviral RTs, which have inherently low processivity and fidelity (Harrison et al. 1998; Malboeuf et al. 2001; Mohr et al. 2013). Recently, we harnessed another type of RT for RNA-seq—thermostable group II intron-encoded reverse transcriptases (TGIRTs) from bacterial thermophiles (Mohr et al. 2013). TGIRT enzymes can be expressed and purified from *Escherichia coli* in concentrated, active forms, which have higher thermostability, processivity, and fidelity than retroviral RTs (Mohr et al. 2013). Moreover, TGIRTs possess a novel end-to-end template-switching activity that can be used to directly attach a 3′ RNA-seq adapter to target RNA sequences during cDNA synthesis, obviating the need for RNA ligase. This template-switching reaction is efficient and inherently strand specific. Additionally, the high thermostability, processivity, and strand-displacement activity of TGIRT enzymes (Mohr et al. 2013) enable RNA-seq of highly structured small ncRNAs, such as tRNAs (Katibah et al. 2014; Shen et al. 2015; Zheng et al. 2015). Here, we show that TGIRT-seq of well-characterized human RNA reference samples yields comprehensive transcriptional profiles of whole-cell RNAs with more diversity and less bias than conventional methods.

RESULTS

RNA sample preparation, sequencing, and mapping pipeline

To assess the ability of a TGIRT enzyme to comprehensively profile whole-cell RNAs, we used the commercially available TGIRT-III enzyme (InGex, LLC) to construct RNA-seq libraries from two well-characterized, commercially available human reference RNA sets: the Universal Human Reference RNA (UHR) and the Human Brain Reference RNA (HBR) (Fig. 1A). The samples were prepared to match the study design used by the Association of Biomolecular Resource Facilities (ABRF) NGS study and the Sequencing Quality Control project (Li et al. 2014; SEQC/MAQC-III Consortium 2014). Each human RNA reference sample was doped with a different External RNA Control Consortium (ERCC) spike-in mix (Sample A—UHR plus ERCC Mix 1; Sample B—HBR plus ERCC Mix 2) and then mixed with each other at known ratios (Sample C—3:1 A:B; Sample D—1:3 A:B) to assess the dynamic range and ability of the RNA-seq method to recapitulate the relative abundance of differentially expressed transcripts.

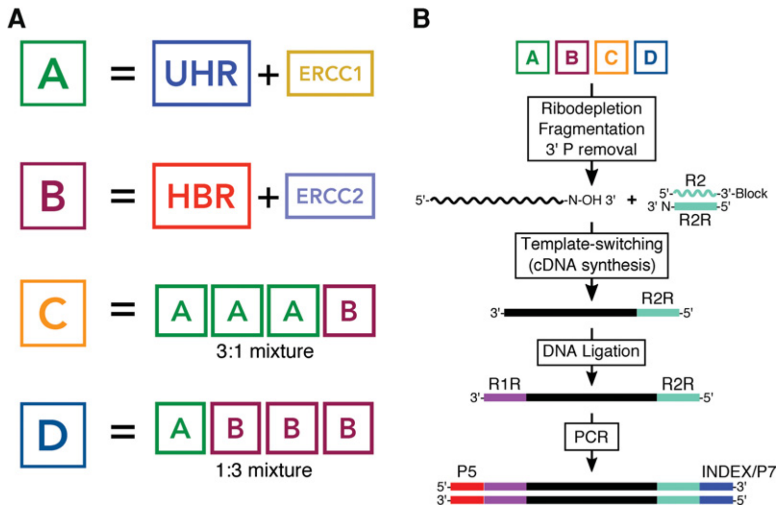

FIGURE 1. RNA sample and TGIRT-seq library preparation. (*A*) Sample A is composed of Universal Human Reference RNA (UHR) mixed with ERCC Spike-in Mix 1, and Sample B is composed of Human Brain Reference RNA (HBR) mixed with ERCC Spike-in Mix 2. Samples A and B were mixed at ratios of 3:1 or 1:3 to constitute Samples C and D, respectively. (*B*) TGIRT-seq library preparation was carried out as previously described (Qin et al. 2016). RNA samples were ribo-depleted to remove cytosolic and mitochondrial rRNAs, fragmented to predominantly 70–100 nucleotides (nt) by incubating with Mg^{2+} at high temperature, and treated enzymatically to remove the resulting 3′ phosphates (–3′ P), which block TGIRT template switching. The fragmented RNA (wavy line) was then used as input for cDNA synthesis by TGIRT template switching, which primes cDNA synthesis from an initial RNA template (R2 RNA)/DNA primer (R2R DNA) substrate that has a single nucleotide 3′ overhang (*N*, an equimolar mix of A, C, G, and T) that can base pair with the 3′ end of the target RNA, seamlessly adding an RNA-seq adapter sequence (R2R) at the start of the cDNA (solid line). This is followed by ligation of a DNA oligonucleotide containing the second RNA-seq adapter sequence (R1R) to the 3′ end of the cDNA. Finally, cDNAs were amplified by PCR to add capture (P5/P7) and index sequences compatible with Illumina sequencing.

All RNA samples were of high quality as evaluated by Bioanalyzer profiles (Supplemental Fig. S1A). For RNA-seq library preparation, each replicate of Samples A–D was ribo-depleted to remove rRNAs and then fragmented with Mg^{2+} at high temperature (NEBNext Magnesium RNA Fragmentation Kit) to yield fragments of ~100 nt, as judged by Bioanalyzer traces (Supplemental Fig. S1B). Following fragmentation, the RNAs were treated with T4 polynucleotide kinase under conditions that remove 3′ phosphates, which impede TGIRT template switching (Mohr et al. 2013). Half of this processed RNA was then used as the input for RNA-seq library preparation via the TGIRT template-switching method developed in our laboratory (Fig. 1B; Qin et al. 2016). The remainder was stored for comparisons with a second TGIRT enzyme denoted TeI4c RT to be done later.

The synthesis of cDNAs by TGIRT template switching involves the direct extension of an initial RNA template/DNA primer substrate comprised of an RNA oligonucleotide containing an Illumina Read 2 sequencing primer-binding site (R2 RNA) annealed to a complementary DNA (R2R DNA). The DNA primer has a single nucleotide 3′ overhang that can base pair with the 3′ end of the target RNA, resulting in a seamless junction between the cDNA and the RNA-seq adapter (Fig. 1B). For the construction of RNA-seq libraries with minimal 3′-end bias, the single nucleotide 3′ overhang is an equimolar mix of A, C, G, and T nucleotides (denoted N; Mohr et al. 2013). The resulting cDNAs are then ligated to a DNA oligonucleotide containing the complement of an Illumina Read 1 sequencing primer-binding site (R1R DNA) followed by 12 cycles of PCR, which synthesizes the second DNA strand and adds Illumina-compatible capture and index sequences.

TGIRT-seq libraries for Samples A–D were constructed in triplicate and sequenced using the Illumina NextSeq 500 platform, with each replicate generating 51.6–88.5 million 75-nt paired-end reads. After trimming, the reads were mapped to the human genome (GRCh38 version 76) in two steps, as described in Materials and Methods (Qin et al. 2016). The first step was end-to-end mapping using HISAT, which maps most of the reads and identifies splice junctions. In the second step, unmapped reads from the first step were remapped by using Bowtie2 local alignment to improve the mapping of RNAs that have post-transcriptionally added nucleotides (e.g., the 3′ CCA of tRNAs).

Overview of TGIRT-seq versus TruSeq v2 and v3 RNA-seq data sets

Mapping statistics for TGIRT-seq libraries of human reference RNA Samples A–D are summarized in Table 1 (for combined replicates of Samples A–D) and Supplemental Table S1 (for individual replicates of Samples A–D) and compared to published ABRF data sets generated from similarly prepared ribo-depleted and fragmented high quality Samples A–D by using either the non-strand-specific TruSeq v2 or the strand-specific TruSeq v3 protocol (see Supplemental Table S2 for sample IDs; Li et al. 2014). In contrast to the TGIRT-seq data sets, the TruSeq data sets were generated from libraries that used the entire yield of fragmented, ribo-depleted RNA for each replicate as input (roughly twice that of the TGIRT-seq libraries) and were amplified by 15 cycles of PCR instead of 12 cycles for TGIRT-seq. The ABRF libraries had been sequenced on the Illumina HiSeq platform and generated 87.3–217.6 million 50-nt paired-end reads for TruSeq v2 libraries and 46.6–92.4 million 50-nt paired-end reads for TruSeq v3. The TruSeq v2 data sets were obtained in triplicate at three different sites (denoted L/R/V), while the TruSeq v3 data sets were obtained at one site (site W) in quadruplicate.

TABLE 1. Mapping statistics for combined human reference RNA data sets

	TGIRT-seq				TruSeq v3				TruSeq v2					
STATISTICS	A	B	C	D	A	B	C	D	L–A	L–B	R–A	R–B	V–A	V–B
Raw read pairs ($\times 10^6$)	57.7–66.4	51.6–88.5	48.6–78.6	54.7–70.1	46.6–78.0	50.5–92.4	61.5–86.0	51.7–73.4	172.8–217.6	96.4–161.6	106.5–127.4	106.2–119.4	87.3–121.3	115.0–126.0
Trimmed pairs ($\times 10^6$)[a]	56.2–65.1	50.3–86.4	47.8–76.6	52.8–66.4	43.9–73.0	47.5–87.3	57.2–80.5	48.3–67.8	164.0–206.4	91.7–153.7	102.0–118.9	101.5–113.3	84.5–117.2	111.2–121.6
End-to-end mapped pairs ($\times 10^6$)[b]	49.4–55.8	42.1–69.6	40.5–66.6	44.1–54.8	41.6–68.3	44.6–81.1	52.7–75.7	44.5–60.8	152.9–191.9	85.1–142.8	91.6–105.4	92.8–102.1	78.8–109.1	103.6–113.0
Local mapped pairs ($\times 10^6$)[c]	5.1–7.1	6.2–12.1	5.0–7.9	6.3–8.4	2.0–4.1	2.5–5.5	3.8–4.6	3.4–6.2	7.5–9.8	4.6–7.5	7.2–11.3	6.8–11.3	4.0–4.7	5.3–5.8
Total mapped pairs ($\times 10^6$)[d]	54.5–62.8	48.3–81.8	45.5–74.5	50.5–63.2	43.6–72.4	47.1–86.6	56.7–79.8	47.8–67.1	60.4–201.7	89.8–150.2	99.9–116.7	99.6–111.1	82.8–114.8	108.8–118.9
Genomic mapping rate (%)[e]	96.5–98.1	94.7–96.1	95.3–97.2	95.2–97.3	99.2–99.3	99.06–99.09	99.1–99.2	98.98–99.0	97.7–98.0	97.7–97.9	97.9–98.2	98.07–98.13	97.9–98.0	97.77–97.83
Genomic unique mapping rate (%)[f]	76.4–79.9	68.3–71.3	73.6–76.5	70.2–73.3	92.1–94.6	91.6–93.0	88.5–92.7	87.9–90.5	86.6–87.4	85.1–86.3	83.8–86.9	82.8–86.9	88.8–89.0	89.9–90.3
Unique pairs mapped to features (%)[g]	87.0–87.7	90.0–90.4	88.6–89.7	89.6–90.3	87.5–88.9	89.8–90.7	88.7–90.4	85.4–88.0	91.8–92.0	92.4–92.5	90.9–91.3	92.1–92.2	90.1–90.2	91.9–92.0

Mapping statistics for combined data sets of human reference RNA samples A–D generated by TGIRT-seq (75-nt paired-end reads, three replicates of each sample), TruSeq v3 (50-nt paired-end reads, four replicates of each sample), and TruSeq v2 (50-nt paired-end reads, three replicates of each sample) at three different sites (L/R/V). The values shown are ranges for the number of replicates.

[a] Number of read pairs after adapter trimming.

[b] Number of read pairs that mapped to the human genome (GRCh38 version 76 with addition of rRNA contigs; see Materials and Methods) using HISAT end-to-end alignment.

[c] Number of unmapped read pairs from HISAT that mapped to the human genome using local alignment with Bowtie2.

[d] Sum of end-to-end and locally aligned read pairs including multiply mapped pairs.

[e] Percentage of read pairs that mapped to the human genome including multiply and uniquely mapped pairs.

[f] Percentage of read pairs that mapped uniquely to the human genome including end-to-end and locally mapped pairs.

[g] Percentage of uniquely mapped read pairs that mapped concordantly in the correct orientation to annotated features in the human reference genome.

For comparisons with TGIRT-seq, the raw sequencing data previously generated in the ABRF-NGS study were reprocessed and mapped in the same manner as for TGIRT-seq data in the current study (see above and Materials and Methods). Plots of the paired-end read span distribution for all data sets obtained using each method showed asymmetric peaks at 72 nt for the TGIRT-seq libraries versus 106 and 138 nt for the TruSeq v2 and v3 libraries, respectively (Supplemental Fig. S1C). These size differences could reflect a higher proportion of reads corresponding to structured small ncRNAs and/or somewhat smaller RNA fragment size in the TGIRT-seq libraries. For all comparisons beyond initial gene mapping, TGIRT-seq reads were clipped to 50 nt in order to match the read length of the TruSeq libraries (see Materials and Methods). After trimming, ~95% of reads were retained across all data sets (Table 1). The TGIRT-seq libraries gave highly reproducible data sets, with Spearman's correlation coefficients comparing normalized read counts of genes between replicate samples ≥0.952 (Supplemental Fig. S2).

Broadly, all libraries displayed high mapping rates (>95% of trimmed reads) to the human genome with ~10% of each library's mapped reads coming from the second local alignment step. Both TruSeq protocols achieved slightly higher overall genomic mapping rates, as well as somewhat higher proportions of uniquely mapped read pairs. The latter likely reflects that the TGIRT-seq data sets include a higher proportion of reads for structured small ncRNAs (e.g., tRNAs and snoRNAs), which map to multiple loci (including pseudogenes), and possibly other unannotated RNAs that map to multiple loci and are not present in the TruSeq data sets (see below). The percentage of uniquely mapped read pairs that mapped concordantly in the correct orientation to annotated genomic features was similar for all libraries (Supplemental Tables S1, S2).

A concern for TGIRT-seq is that multiple end-to-end templates witches might artifactually link sequences of different RNA fragments. However, analysis by TopHat-Fusion (Kim and Salzberg 2011) showed that the frequency of fusion reads was low for all three methods (<0.3%), with TGIRT-seq having a frequency of fusion reads similar to that of TruSeq v2 and lower than that of TruSeq v3, when normalized for read depth (Supplemental Fig. S3A). These findings may reflect that the frequency of multiple end-to-end template switches by the TGIRT enzyme is lower than the frequency of template switches resulting from the retroviral RT used in TruSeq falling off one template and reinitiating on another, which has been shown previously to give artifactual chimeric reads (Ouhammouch and Brody 1992; Mader et al. 2001; Cocquet et al. 2006). The fusion reads detected in each RNA sample by TGIRT-seq showed some correlation between technical replicates (Spearman correlation coefficients ~0.6), suggesting that a significant proportion either represent bona fide RNA fusions or are generated by a nonrandom process (e.g., template switching between complementary sequences during PCR) rather than end-to-end template switching by the TGIRT enzyme, which is expected to be random (Supplemental Fig. S3B)

Diversity of RNAs detected by TGIRT-seq compared to TruSeq

Figure 2 (combined replicates) and Supplemental Table S3 (individual replicates) compare the proportion of concordant read pairs that mapped uniquely in the correct orientation to different annotated genomic features in the various libraries. As expected, all three methods produced libraries with reads mapping mostly to protein-coding genes. In contrast to TruSeq, TGIRT-seq detected a higher proportion of small ncRNAs, while TruSeq v3 was enriched for reads mapping to mitochondrial RNAs. Comparison of the small noncoding RNAs detected by each method showed that TGIRT-seq enabled simultaneous sequencing of tRNAs, which are essentially absent in the TruSeq data sets, and enhanced detection of snoRNAs and snRNAs, two other classes of structured RNAs that are refractory to reverse transcription by retroviral RTs (Fig. 2B). piRNAs were underrepresented in TGIRT-seq libraries compared to TruSeq libraries, likely due to modification of their 3′ ends by a 2′*O*-methyl group (Kirino and Mourelatos 2007), which inhibits TGIRT template switching (Mohr et al. 2013). All library preparation methods gave low levels of reads mapping to miRNA loci, reflecting their low relative abundance and/or underrepresentation in whole-cell RNA libraries because of size-selection steps that remove PCR primers and adapter artifacts using bead-based methods rather than a gel.

TGIRT-seq recovers relative abundances of spike-ins and differentially expressed genes

RNA-seq can theoretically be used to quantitate the true abundance of RNAs in samples (Mortazavi et al. 2008). Previous large-scale studies comparing RNA-seq protocols across platforms concluded that quantitation is reproducible using a single protocol but not between protocols and that absolute quantitation is typically inaccurate when judged against spike-in transcripts of known concentration (SEQC/MAQC-III Consortium 2014). Therefore, protocols are evaluated on their ability to reproduce differential levels of RNA transcripts between samples. The human reference RNA samples used here and previously have several built-in ground truths that provide convenient measures of relative abundance recovery in libraries generated by different methods.

First, Samples A and B contain ERCC spike-in mixes that are of known sequence and concentration. As shown in Figure 3A (left panel), TGIRT-seq recovered the abundance of ERCC spike-ins in a manner highly correlated with their known values, similar to libraries prepared using TruSeq v2 and v3 (Fig. 3A, middle and right panels). The sensitivity of

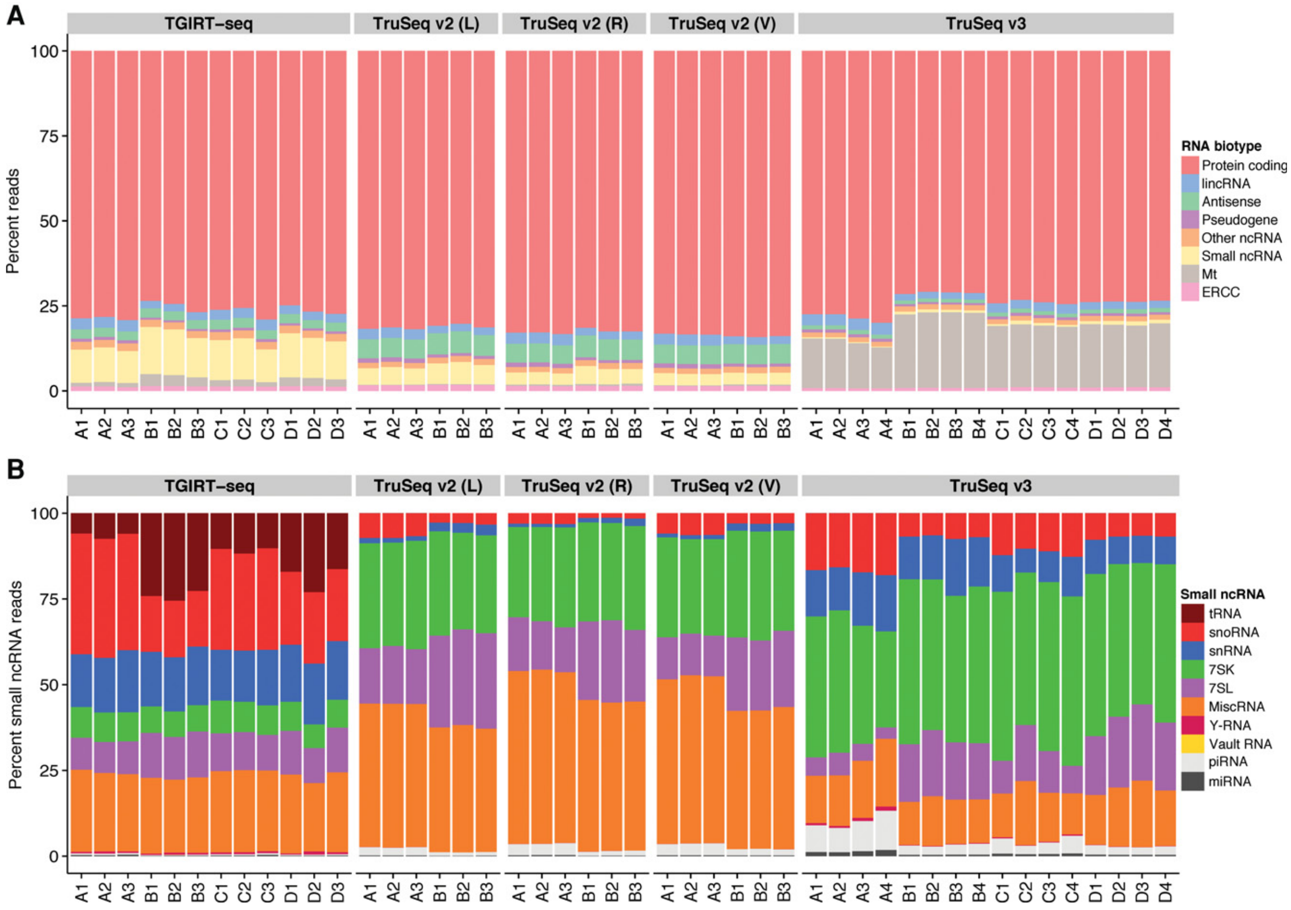

FIGURE 2. TGIRT-seq reads map mostly to protein-coding genes but with greater representation of small ncRNAs than TruSeq libraries. (*A*) Stacked bar graphs showing the percentage of uniquely mapped reads for each class of annotated genomic features in Ensembl GRCh38 release 76, Genomic tRNA Database, and piRNABank (Qin et al. 2016) for different library preparation methods for numbered replicates of Samples A–D. (*B*) Stacked bar graphs showing the percentage of small noncoding RNA reads that map to different classes of small ncRNAs for different library preparation methods for numbered replicates of Samples A–D. MiscRNA includes ribozymes, such as RNase P RNA, imprinted transcripts, such as Xist, and other transcripts that cannot be classified into other RNA annotation categories. (*Left* panels) TGIRT-seq; (*middle* panels) TruSeq v2 (from ABRF at three different sites, L/R/V); (*right* panels) TruSeq v3 (from ABRF at site W). Features and small ncRNA classes are color coded as indicated to the *right* of the bar graphs.

the three methods was similar with TGIRT-seq having a roughly twofold lower limit of detection when compared to the TruSeq libraries at a threshold of 1 FPKM (fragments per kilobase per million mapped reads). TruSeq v2 libraries had a slightly higher number of detected spike-in species, likely due to their greater sequencing depth (Supplemental Table S1).

Second, each of the 92 polyadenylated ERCC spike-in transcripts is grouped into one of four classes (0.5:1, 0.67:1, 1:1, 4:1) according to the relative abundance of the spike-in between Mix 1 (Sample A) and Mix 2 (Sample B). TGIRT-seq recapitulated these differences in abundance better than the strand-specific TruSeq v3 method and almost as well as the non-strand-specific TruSeq v2 method (Fig. 3B). For TGIRT-seq and TruSeq v2, empirical fold-change ratios were more highly correlated with their expected values for abundant spike-ins (those to the right of each panel), as previously observed for TruSeq v2 (SEQC/MAQC-III Consortium 2014), whereas empirical fold-change ratios were poorly correlated with their expected values for TruSeq v3 (Fig. 3B).

Third, the mixing of Samples A and B to constitute Samples C and D defines an expected order of dilution of the human reference set RNAs. For both TGIRT-seq and TruSeq v3 (Samples C and D were not analyzed by TruSeq v2 in the ABRF study), most protein-coding gene transcripts followed a consistent titration order, with those following inconsistent order corresponding to transcripts with small fold changes between Samples A and B (Fig. 4A). For both methods, there was also a slight bias toward inconsistent titration order for transcripts higher in B than in A (tail on right side of the red peak).

More detailed analysis of protein-coding gene transcripts detected by TGIRT-seq and TruSeq v3 in Samples A–D (Fig. 4B) revealed that both protocols performed similarly in recovering the known mixing ratios between samples. The TGIRT-seq libraries had an observed mixing ratio

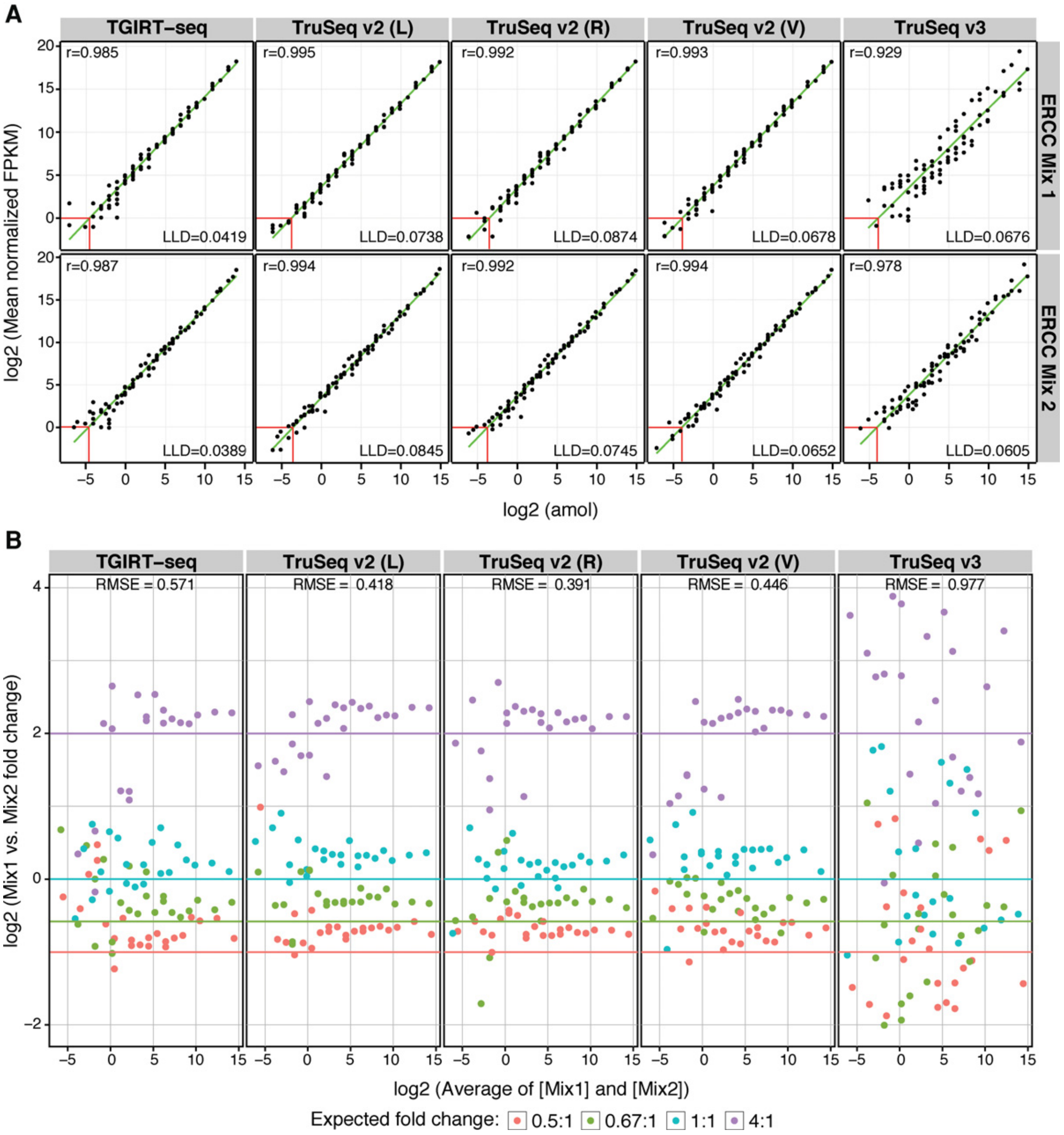

FIGURE 3. TGIRT-seq recapitulates the relative abundance of ERCC spike-ins added to human reference set RNAs. (*A*) Normalized read counts (fragments per kilobase per million mapped; FPKM) of each ERCC spike-in detected by ≥10 reads were plotted against their expected amounts in attomoles for combined data sets for different replicates obtained using different RNA-seq library preparation methods: (*Left* panels) TGIRT-seq; (*middle* panels) TruSeq v2 at sites L, R, and V; (*right* panels) TruSeq v3. The *upper* row displays data for ERCC Spike-in Mix 1 (in Sample A), and the *lower* row displays data for ERCC Spike-in Mix 2 (in Sample B). Each dot represents a particular spike-in RNA. Pearson's correlation coefficients (*r*) are shown on the graphs. Red lines at the *bottom left* indicate the lower limit of detection (LLD) in attomoles with a threshold of FPKM = 1. (*B*) Observed fold differences between Mix 1 and Mix 2 of individual ERCC transcripts were plotted against the average of their expected concentrations in Samples A and B for each library preparation method: (*left* panel) TGIRT-seq; (*middle* panels) TruSeq v2; (*right* panel) TruSeq v3. Each RNA is color-coded to indicate its expected fold change according to the key shown at the *bottom* of the figure. Each colored line represents ideal recovery of the predicted fold change for each class of RNA. Root mean square error (RMSE) values are shown at the *top*.

slightly lower than expected, while the TruSeq v3 libraries had an observed mixing ratio slightly higher than expected, possibly reflecting small differences in hand mixing during sample preparation. As expected, the mixing ratio was recovered better for highly abundant transcripts than for lower abundance transcripts (compare top 1% of transcripts [red] versus bottom 75% of transcripts [gray]). Importantly, the difference in relative abundance of protein-coding gene transcripts between Sample A and B determined by TGIRT-seq was highly correlated with that determined by TaqMan RT-qPCR of similarly prepared reference samples in the MAQC study (MAQC Consortium 2006; Spearman

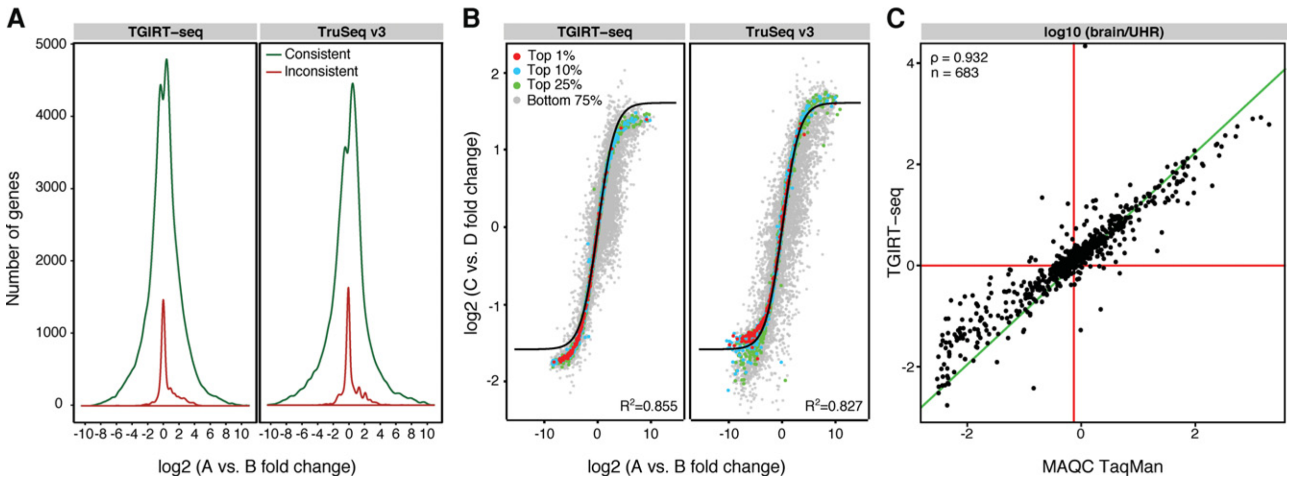

FIGURE 4. TGIRT-seq reproduces the titration order and relative abundances of human reference RNA transcripts. (*A*) The number of protein-coding gene RNAs whose normalized read counts were consistent or inconsistent with their relative abundance in Samples A–D (data sets for combined replicates) were plotted against the observed fold change for TGIRT-seq (*left* panel) and TruSeq v3 (*right* panel) libraries. The majority of protein-coding genes followed the titration order (i.e., A > C > D > B in Sample A or B > D > C > A in Sample *B*) in both methods. (*B*) Log_2 fold changes of protein-coding gene RNAs observed between Samples C and D were plotted against the $\log_2$ fold changes of these protein-coding gene RNAs between Samples A and B for TGIRT-seq (*left* panel) and TruSeq v3 (*right* panel). Each RNA is color-coded to indicate its expression level relative to all other RNAs according to the key at the *upper left* of the panel: top 1% (red); top 10%—RNAs in the top 10% but not the top 1% (blue); top 25%—RNAs in the top 25% but not in the top 10% (green); bottom 75% (gray). The black curve indicates ideal recovery of expected fold difference ratios. Correlation values are displayed at the *bottom right* of each graph. (*C*) The ability of TGIRT-seq to reproduce differences in RNA expression levels between Sample A (UHR) and Sample B (brain) was correlated with the relative abundance of RNAs between these two samples determined by TaqMan RT-qPCR in the MAQC study (MAQC Consortium 2006). The plot shows the $\log_{10}$ values for the relative abundance of 683 protein-coding transcripts between Samples A and B for combined TGIRT-seq data sets versus the TaqMan RT-qPCR values. Number of transcripts (*n*) and the Spearman correlation coefficient (ρ) are displayed on the graph.

$\rho = 0.932$; Fig. 4C, compared to $\rho = 0.92$–0.93 and 0.90 for TruSeq v2 and v3, respectively; Supplemental Fig. S4).

Strand specificity and biases

Because TGIRT-seq libraries are generated via a template-switching reaction that links an RNA-seq adapter to the beginning of the cDNA sequence and TGIRT enzymes minimally recopy the initial cDNA (Qin et al. 2016), they are inherently strand specific without relying on dUTP incorporation or the inclusion of actinomycin D as in TruSeq libraries. Comparison of the percent of reads mapping to the annotated strand of protein-coding RNAs indicated that the TGIRT-seq libraries have significantly better strand specificity than TruSeq v3 libraries (*P*-value $< 1.56 \times 10^{-6}$), while the TruSeq v2 libraries showed no strand specificity as expected (Fig. 5A). Analysis of reads mapping to the ERCC spike-in sequences confirmed that the TGIRT enzyme minimally recopies the initial cDNA, while TruSeq v3 libraries (generated by SuperScript II) contained a significantly higher proportion of antisense spike-in sequences (*P*-value $< 2.45 \times 10^{-6}$), presumably reflecting recopying of initial ERCC spike in cDNA by the retroviral RT (Fig. 5B). The lower strand specificity of TruSeq v3 was additionally confirmed by the relatively high frequency of the exact complement of annotated splice junctions in the TruSeq v3 data sets, as described below.

We also compared nucleotide frequency biases at the 5′ and 3′ ends of reads generated by the three methods (Fig. 5C). Such biases are well documented for random hexamer priming used for cDNA synthesis in both TruSeq methods (Hansen et al. 2010), but could also arise from A-addition to the 3′ ends of dsDNAs prior to adaptor ligation or the ligation itself in TruSeq, or from the template-switching reaction used for RNA-seq adapter attachment in TGIRT-seq. In addition, 5′ and 3′ biases for all methods could arise from other steps, including nontemplated nucleotide addition at the 3′ ends of cDNAs after reaching the 5′ end of the RNA template, RNA fragmentation, dephosphorylation, or DNA or RNA ligases used to attach RNA-seq adapters (Hafner et al. 2011; Kwok et al. 2013; Lee et al. 2013; Wery et al. 2013; Zajac et al. 2013; Jackson et al. 2014). As the TruSeq libraries were generated by using random hexamer priming for cDNA synthesis, biases at the both the 5′ and 3′ ends of the RNA reads reflect where the primer anneals and may include sequence errors due to mispriming. In contrast, TGIRT template switching yields full-length reads of RNA fragments, beginning directly at their 3′ ends.

Plots of nucleotide frequency as a function of position from the 5′ and 3′ ends of the RNA read show that TGIRT-seq and TruSeq libraries have similar nucleotide frequencies at the 3′ end of the RNA sequence with guanosine comprising >25% of the 3′ terminal nucleotides (position −1) and other nucleotides correspondingly underrepresented (Fig. 5E, right

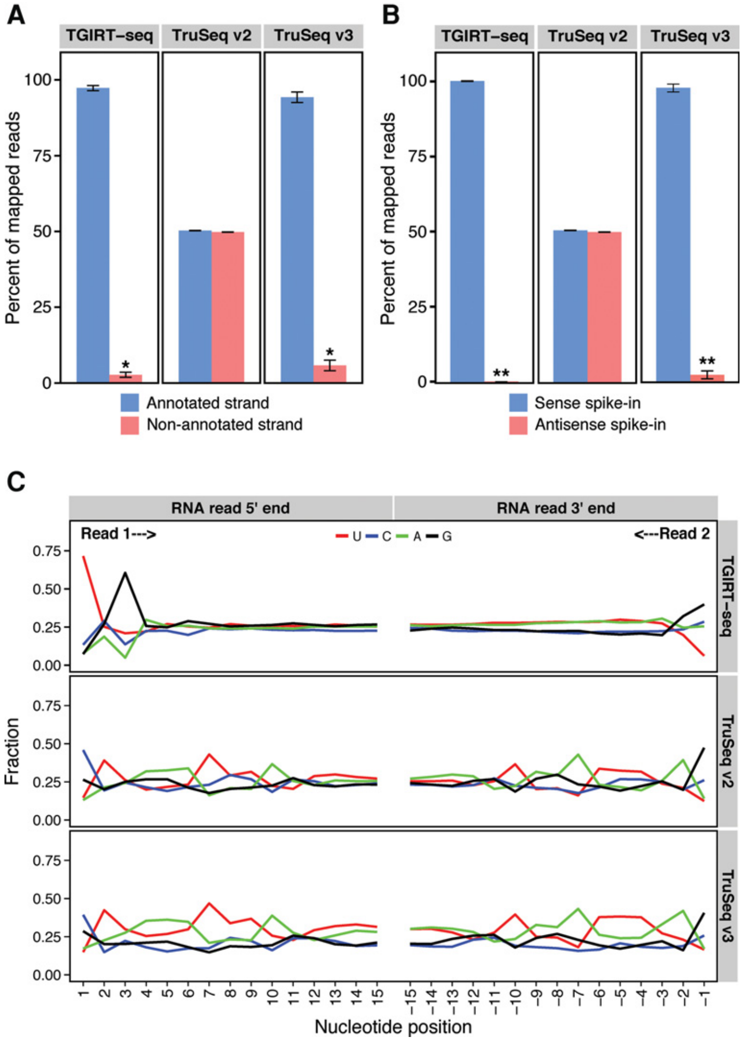

FIGURE 5. TGIRT-seq is more strand specific than TruSeq v3 and eliminates biases due to random hexamer priming. (*A*) The mean percentage of reads mapping to the correct annotated strand for all protein-coding gene RNAs in combined data sets for different replicates was plotted for each library preparation method: (*left* panel) TGIRT-seq; (*middle* panel) TruSeq v2; (*right* panel) TruSeq v3. Error bars represent standard deviation. Asterisk indicates *t*-test $P < 1.56 \times 10^{-6}$ for the difference in values between TGIRT-seq and TruSeq v3. (*B*) The mean percentage of reads mapping to the correct strand for ERCC spike-in RNAs was plotted for each library preparation method: (*left* panel) TGIRT-seq; (*middle* panels) TruSeq v2; (*right* panel) TruSeq v3. Error bars represent standard deviation. Double asterisk indicates *t*-test $P < 2.45 \times 10^{-6}$ for the difference in values between TGIRT-seq and TruSeq v3. TruSeq v2 data sets from all three sites (L/R/V) were combined for this analysis. (*C*) 5′ and 3′ end biases of different library preparation methods. The plots show the fraction of each nucleotide at each of the first 15 bases of Read 1 (equivalent to the 5′ end of the RNA sequence) and the first 15 complementary bases of Read 2 (equivalent to the 3′ end of the RNA sequence) after adapter trimming. (*Upper* panel) TGIRT-seq; (*middle* panel) TruSeq v2; (*lower* panel) TruSeq v3. TGIRT-seq shows similar 3′-nuclotide biases to TruSeq but different 5′-nucleotide biases. The latter may reflect different patterns of nontemplated nucleotide addition at the 5′ end of trimmed reads, preferred termination sites for the RTs, or biases introduced by DNA ligase for addition of the second RNA-seq adapter at the 3′ end of the cDNA. TGIRT-seq shows less bias than TruSeq at nucleotide positions 4–12 from the 5′ or 3′ ends, which are likely introduced by random hexamer priming in the TruSeq methods (Hansen et al. 2010). TruSeq v2 data sets from all three sites (L/R/V) were combined for this analysis.

panels). By the third nucleotide, however, the TGIRT-seq reads reverted to a relatively even distribution of all four nucleotides, while both TruSeq libraries showed nucleotide biases characteristic of random hexamer priming (Hansen et al. 2010; Li et al. 2014). At RNA 5′ ends, TGIRT-seq libraries showed a high proportion of U at the 5′-most nucleotide (position +1), which may reflect nontemplated addition of an A residue to the 3′ end of the cDNA, and G at position +3, which may reflect a third position bias of the Thermostable RNA/DNA Ligase used for adaptor ligation to the cDNAs (Jackson et al. 2014), before reverting to relatively uniform nucleotide frequencies (Fig. 5C, upper left panel). In contrast, the TruSeq libraries have a high proportion of Cs at their 5′ ends (middle and lower left panels) and have biases up to position +12, which are complementary to the biases at the 3′ end of the reads and presumably result from random hexamer priming of second-strand synthesis (Hansen et al. 2010). We conclude that TGIRT template switching does not introduce substantial 3′-end bias and eliminates biases due to random hexamer priming that are inherent in TruSeq.

Gene coverage

Previous methods for constructing RNA-seq libraries using RNA fragmentation have been thought to give fairly uniform coverage across protein-coding genes (Mortazavi et al. 2008). However, in both TGIRT-Seq and TruSeq v2-generated data sets, a higher proportion of bases mapped to coding regions than in the TruSeq v3 data sets, while TruSeq v3 reads had a higher proportion of bases mapping to introns. All three library types gave similar proportions of bases mapping to mRNA 5′- and 3′-untranslated regions (UTRs) and very low proportions of bases mapping to intergenic regions, indicating little contamination by genomic DNA (Fig. 6A; Supplemental Table S1).

Further, comparison of base coverage from the 5′ to 3′ end of the one thousand most abundant protein-coding gene RNAs in the combined data sets for each method showed that TGIRT-seq libraries gave better coverage of 5′-end sequences than did libraries prepared by either TruSeq protocol, while the strand-specific TruSeq v3 libraries overrepresented sequences at the 3′ ends of

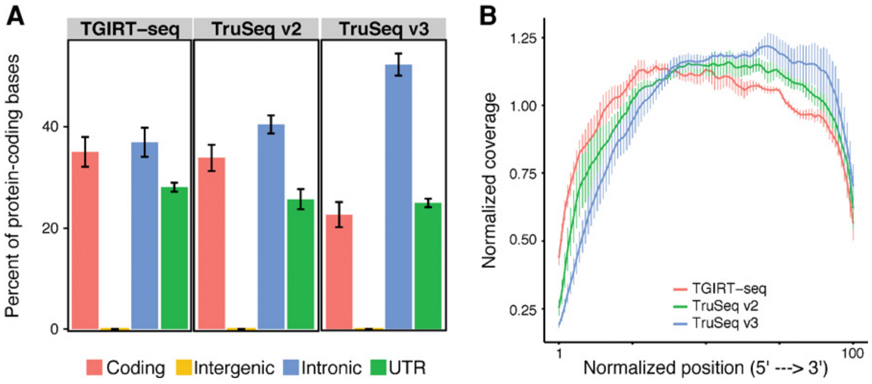

FIGURE 6. TGIRT-seq data sets show more uniform gene coverage than TruSeq data sets. (*A*) The percentage of bases of protein-coding genes classified as coding, intergenic, intronic, and untranslated region (UTR) was plotted for combined replicate data sets for each library preparation method: (*left* panel) TGIRT-seq; (*middle* panel) TruSeq v2; (*right* panel) TruSeq v3. Error bars represent the standard deviation. (*B*) The normalized coverage of the 1000 most abundant protein-coding gene transcripts is plotted against normalized gene position from the 5′ (*left*) to the 3′ (*right*) end of the RNA transcript for combined replicate data sets for each library preparation method. Error bars represent standard deviation.

these genes as reported previously (Fig. 6B; Li et al. 2014). Although TGIRT enzymes were shown previously to give better representation of 5′-proximal sequences than a retroviral RT in RNA-seq libraries prepared by oligo(dT) priming (Mohr et al. 2013), their ability to do so in RNA-seq libraries prepared from fragmented RNA was surprising. These differences may reflect a combination of somewhat shorter RNA fragment size in the TGIRT-seq libraries, which results in more frequent sampling across RNA molecules (Mortazavi et al. 2008; Oshlack and Wakefield 2009), and the high processivity and strand displacement of the TGIRT enzyme operating at 60°C, which enables it to reverse transcribe continuously and through structured regions to the 5′ ends of even short RNA fragments more frequently than can retroviral RTs operating at lower temperature (Mohr et al. 2013).

Detection of different mRNA species and splice junctions

In both TGIRT-seq and TruSeq data sets, the large majority of unique read pairs mapped to protein-coding genes (Fig. 2A; Supplemental Table S3). When normalized for read depth and with reads clipped to the same length, TGIRT-seq and TruSeq v2 detected more protein-coding gene transcripts and more annotated splice junctions than did TruSeq v3 (Fig. 7A,B, respectively). Closer analysis revealed that the TGIRT enzyme detected more splice junctions near the 5′ end of mRNAs than either TruSeq method (Fig. 7C), consistent with the better representation of 5′ RNA ends in the TGIRT-seq libraries.

Comparison of Sashimi plots generated by the Integrative Genomics Viewer (IGV) for abundant protein-coding gene RNAs in TGIRT-seq and TruSeq v3 data sets (Sample A—UHR) shows substantial increases in reads detecting the 5′-most splice junctions, as well as additional junctions detected throughout the genes in the TGIRT-seq data sets (Fig. 8). Strikingly, TGIRT-seq also detected abundant reads corresponding to snoRNAs encoded within introns in ribosomal protein genes, whereas these embedded snoRNAs were not readily detected by TruSeq v3 (Fig. 8, boxed regions for *RPS8* and *RPL17*).

Analysis of detected splice sites showed that TGIRT-seq and TruSeq v3 detected annotated and unannotated splice junctions, which were overwhelmingly canonical GU–AG junctions (>90%), with relatively few U12-type splice junctions (AU-AC, <0.1%; Supplemental Fig. S5A,B). The TruSeq v3 data sets showed a higher proportion of junctions that were the antisense to annotated junctions, consistent with a higher proportion of cDNA recopying by the retroviral RT, even in the presence of actinomycin D (Fig. 7D, green). Examination of these junctions showed that they are the exact complement of the annotated junctions and part of longer antisense reads that are not present in the TGIRT-seq data sets (Supplemental Fig. S6A,B). Both methods identified a low proportion of novel junctions with noncanonical splice sites of unknown significance, with the proportion of such reads appearing somewhat higher for TGIRT-seq (Fig. 7D, purple).

Profiles of small ncRNAs in TGIRT-seq libraries

A major difference between TGIRT-seq and TruSeq v2 and v3 libraries is the greater coverage of structured small ncRNAs by TGIRT-seq (Fig. 2B; Supplemental Table S3). To assess the ability of TGIRT-seq to quantitate small ncRNAs, we analyzed the titration order of small ncRNAs that are differentially expressed between Samples A and B, as done previously for protein-coding gene transcripts (see Fig. 4A). As shown in Supplemental Figure S7, we found that the majority of all classes of small ncRNA analyzed (tRNAs, snoRNAs, snRNAs, miRNAs, Y-RNAs, and Vault RNAs) followed the expected titration order between samples, albeit to different degrees for different RNA classes.

Figure 9 shows profiles of small ncRNAs detected by TGIRT-seq in Sample A and Sample B rank-ordered by relative abundance. Each class of RNA shows a skewed distribution, with a few predominant species followed by a longer tail of less abundant species. Glu tRNAs were the most abundant tRNA in both the UHR and HBR sets with >40% and 30% of tRNA reads mapping to Glu tRNAs with anticodon CTC (Fig. 9A). Read alignments via IGV showed that most of the tRNAs are fragmented by the library preparation

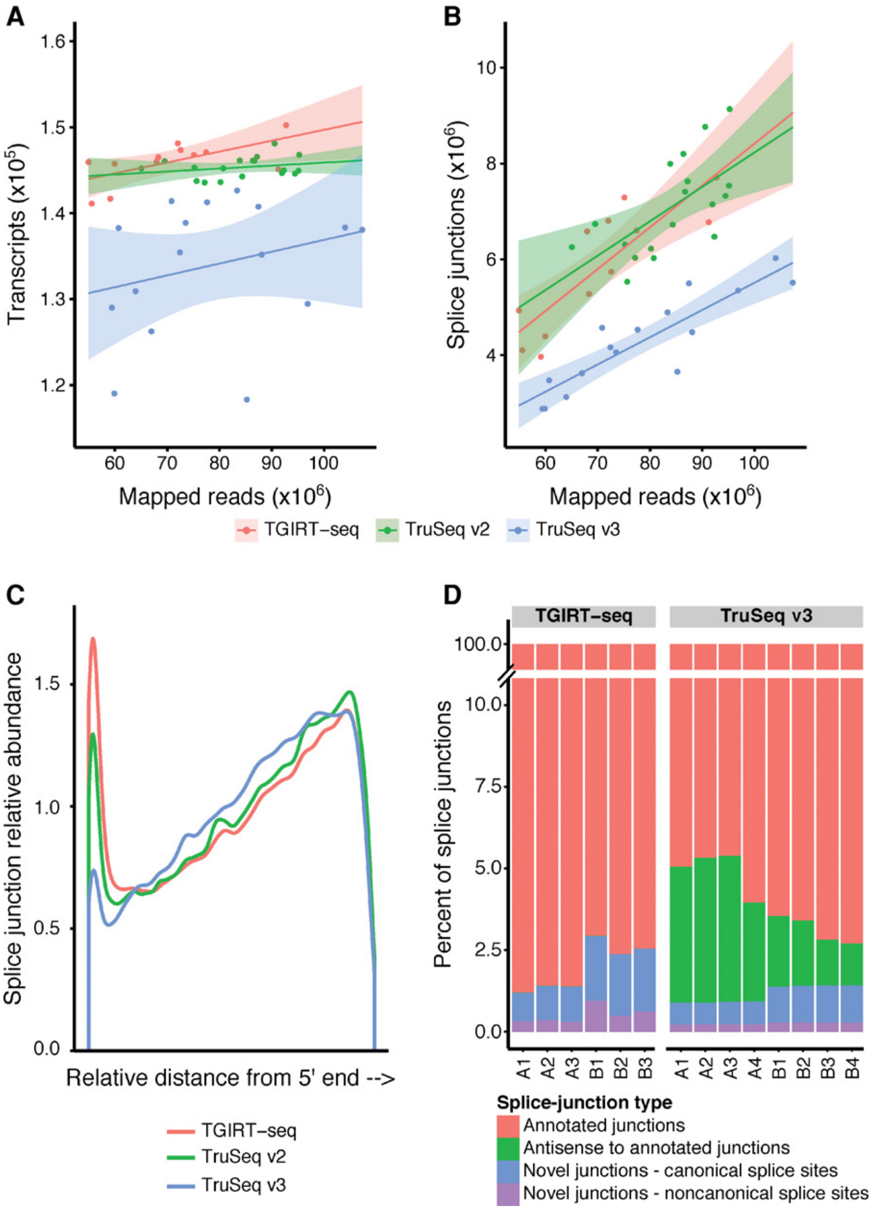

FIGURE 7. TGIRT-seq detects more transcripts and splice junctions than does TruSeq. (*A*,*B*) The number of protein-coding gene transcripts and annotated splice junctions detected as a function of mapped reads for combined data sets for Samples A–D for each library preparation method. The TruSeq v2 libraries were down-sampled to match the sequencing depth of TGIRT-seq and TruSeq v3 libraries. Shaded areas represent 95% confidence intervals for the predicted fit. (*C*) Splice junction density plotted versus relative distance from the 5′ end of protein-coding gene RNAs for combined data sets for Samples A–D for each library preparation method. (*D*) Distribution of splice junction types for replicates of Samples A and B for TGIRT-seq and TruSeq v3. Antisense to annotated junctions are exact complements on the opposite strand of the annotated junctions. Novel junctions are those for which no annotation currently exists in the reference genome used.

procedure as expected, but have 3′ termini corresponding to those expected for mature tRNAs with post-transcriptionally added CCA tails (Supplemental Fig. S8). Certain base modifications were detected by distinctive patterns of nucleotide misincorporation, as described previously (Auffinger and Westhof 1998; Katibah et al. 2014), including the highly conserved 1-methyladenosine (m^{1}A) at A58 and 1-methylguanosine (m^{1}G) at G9 in iMet-CAT and at G37 in Leu-CAG.

Similar analysis of snoRNAs (Fig 9B) and snRNAs (Fig. 9C) showed a broad distribution of snoRNA species in each sample, while the RNU2-2P accounted for 30%–40% of reads mapping to snRNAs in both Samples A and B. Read alignments of snoRNAs and snRNAs revealed that these transcripts were fragmented by the Mg^{2+} treatment of the RNAs but had coverage profiles extending the full-length of the mature RNA (Supplemental Figs. S9, S10). Many of the miRNA detected when mapping to the genome were annotated as predicted but not validated miRNAs (not shown). However, mapping of the TGIRT-seq reads to miRBase detected validated miRNAs, with miR-125b-5p being one of the most abundant in both Samples A and B (Fig. 9D). Read alignments of the miRBase-mapped miRNAs showed mainly full-length mature miRNAs, with some (e.g., let7a-5p) showing post-transcriptionally added 3′ A and U, which may influence miRNA function or stability (Supplemental Fig. S11A; Burroughs et al. 2010; Wyman et al. 2011; Koppers-Lalic et al. 2014). Finally, TGIRT-seq detected both Y-RNA and Vault RNAs (Fig. 9E,F). Read alignment plots of Y-RNAs showed that these species mostly shared common 3′ ends (Supplemental Fig. S11B), while Vault RNAs showed a mixture of mature 3′ ends and 3′-end extensions, particularly for VTRNA1-2 and VTRNA1-3 (Supplemental Fig. S11C).

DISCUSSION

The TGIRT-seq data sets obtained here for well-characterized human RNA reference samples validate this method and indicate that it has several advantages compared to the widely used TruSeq methods for transcriptome profiling. These advantages include simplicity, rapid processing time, efficacy with very small amounts of RNA (previously demonstrated for human plasma RNAs; Qin et al. 2016), and significantly higher strand specificity than TruSeq v3. The higher strand specificity of TGIRT-seq is shown by the higher proportion of reads mapping to the correct strand

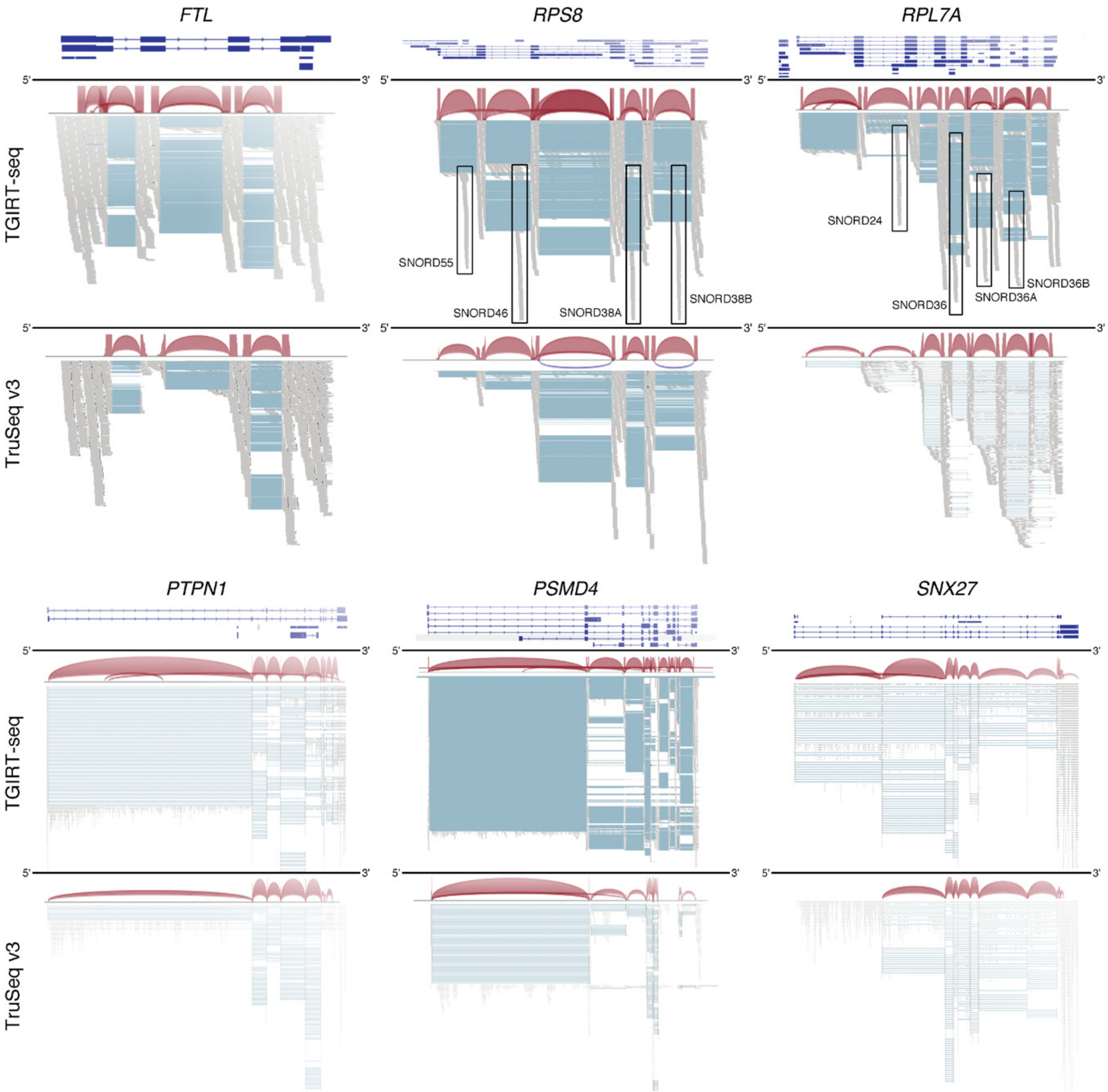

FIGURE 8. Enrichment for 5′ proximal splice junctions by TGIRT-seq. IGV screen captures of Sashimi plots for protein-coding gene transcripts from Sample A2 for both methods (Universal Human Reference: *FTL*, *RPS8*, *RP17A*, *PTPN1*, *PSMD4*, and *SNX27*). (*Upper* plot for each gene) TGIRT-seq; (*lower* plot for each gene) TruSeq v3. Genes are identified by name *above* their gene maps, which are displayed at *top* in blue, with exons depicted as rectangular boxes and introns depicted as lines. Splice junctions are depicted in Sashimi plots as arcs between exons. The height and thickness of each arc depict the depth of coverage for that particular junction. Splice junctions in the sense strand are displayed in maroon, and those in the antisense strand are displayed in blue (e.g., TruSeq *RPS8* gene). Read alignments are shown *below* with reads from each method colored gray and splice junctions depicted as thin blue bars between gray reads. Reads corresponding to snoRNAs encoded in introns in ribosomal protein genes are shown in boxes along with the snoRNA name.

of protein-coding genes (Fig. 5A) and RNA spike-ins (Fig. 5B), and by the lower frequency of antisense reads of annotated mRNA splice junctions, which are surprisingly abundant in the TruSeq v3 data sets (2.97 ± 1.58% of splice-junction reads for all replicates; Fig. 7D). We also find that the TGIRT-seq data sets obtained here show more uniform 5′ to 3′ gene coverage and detect more splice junctions, particularly near the 5′ ends of mRNAs, than do the published TruSeq data sets. And finally, a major advantage of TGIRT-seq is the ability to profile protein-coding and lncRNAs together with small ncRNAs in the same sequencing run.

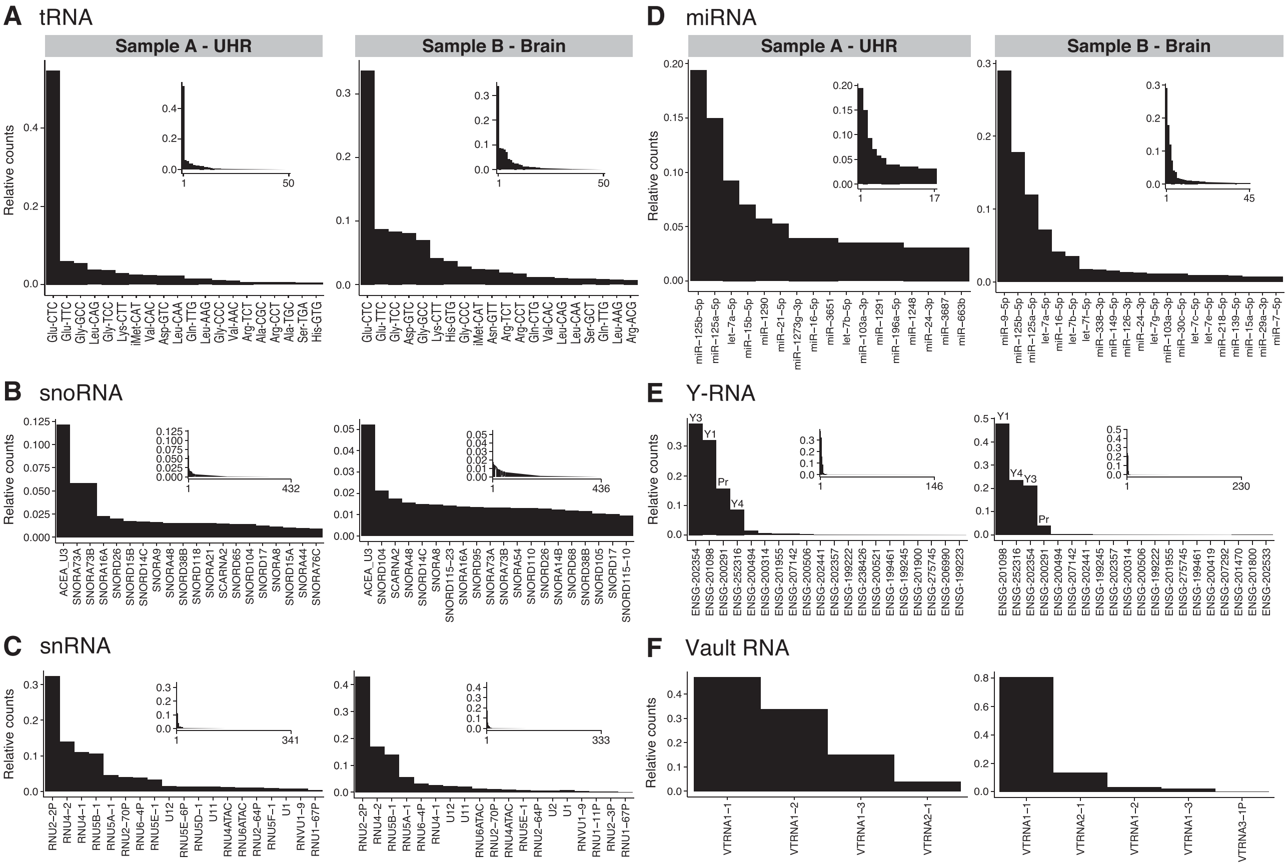

FIGURE 9. TGIRT-seq detects tRNAs and other small noncoding RNAs. (*A*–*F*) Detected species of various small ncRNAs in combined replicates of Sample A (Universal Human Reference, UHR, *left*) and Sample B (Human Brain Reference, Brain, *right*). Detected species of tRNAs (grouped by anticodon), snoRNAs, snRNAs, miRNAs, Y RNAs, and Vault RNAs with ≥10 reads in combined TGIRT-seq data sets for Samples A and B were rank ordered by their relative abundance. The bar graphs show the most abundant small ncRNA species in each category. Highly abundant Y RNAs are identified both by Ensembl ID *below* and abbreviated gene symbol *above* (Pr, predicted). *Inset* graphs show the full profile for each RNA class with the total number of species detected (grouped by anticodon for tRNA), except for Vault RNAs where all detected species could be displayed in the bar graph.

To validate the TGIRT-seq method, we tested its ability to recapitulate the relative abundance of transcripts and RNA spikes-ins in ribo-depleted, fragmented human reference RNAs. We find that TGIRT-seq does so comparably to the non-strand-specific TruSeq v2 method and better than the strand-specific TruSeq v3 method (Figs. 3, 4). Additionally, we show that changes in relative abundance of human mRNAs quantitated by TGIRT-seq in the UHR and HBR reference RNA samples is highly correlated (ρ = 0.932) with TaqMan RT-qPCR data previously published for these samples (Fig. 4C; MAQC Consortium 2006). Addressing possible liabilities, we find that the TGIRT-template-switching activity used for RNA-seq library construction does not introduce substantial 3′-RNA-end biases in human reference RNA samples compared to TruSeq (Fig. 6C), nor does it produce large numbers of fusion reads due to multiple, sequential template switches (Supplemental Fig. S3). Indeed, the frequency of fusion reads in the TGIRT-seq data sets was similar to that for TruSeq v2 and lower than that for TruSeq v3 (Supplemental Fig. S3A).

The virtues of TGIRT-seq for transcriptome profiling include more transcripts detected per million reads than TruSeq v3 (Fig. 7A), more uniform 5′ to 3 gene coverage (Fig. 6B), and detection of more splice junctions, particularly near the 5′ end of mRNAs than either TruSeq method (Fig. 7B,C). TGIRT enzymes were previously shown to give more uniform 5′ to 3′ coverage on intact mRNAs using an oligo(dT) primer (Mohr et al. 2013). Their ability to do so even for fragmented RNAs may reflect that the higher processivity, strand-displacement activity, and operating temperature (60°C) of TGIRT enzymes enable more complete copying of even short RNA fragments than can be done by retroviral RTs used in TruSeq. The latter enzymes have relatively low processivity and dissociate readily from RNA templates, particularly at structured regions (Ouhammouch and Brody 1992; Wu et al. 1996; Mader et al. 2001; Cocquet et al. 2006; Mohr et al. 2013), and these deficiencies may be exacerbated in TruSeq by annealed random hexamers, which constitute potential barriers to cDNA synthesis. Although differences in RNA fragment length between the TGIRT-seq and

TruSeq libraries are relatively small (Supplemental Fig. S1C), it is possible that the somewhat shorter RNA fragment sizes in the TGIRT-seq libraries also contribute to better sampling of 5′ ends of mRNAs. However, shorter inserts are thought to be less advantageous for splice variant analysis by paired-end sequencing (Katz et al. 2010) and thus unlikely to be a factor in the ability of TGIRT-seq to detect substantially more splice junctions than TruSeq v3 (Figs. 7, 8).

The template-switching activity of TGIRT enzymes provides a novel, simple, and efficient method for adding RNA-seq adapters at RNA 3′ ends that obviates the need for an RNA ligase or random hexamer priming, both of which introduce biases (Hansen et al. 2010; Hafner et al. 2011). As indicated above, concerns that TGIRT template switching might introduce additional 3′ RNA-end biases or produce large numbers of fusion reads proved unfounded, and the method eliminates sampling biases due to random hexamer priming that are inherent to TruSeq (Fig. 6C; Supplemental Fig. S3A; Hansen et al. 2010). This template-switching activity together with the high processivity and strand-displacement activities of TGIRT enzymes makes it possible to obtain full-length coverage of highly structured and modified RNAs, such as tRNAs, snoRNAs, and Y RNAs, in the same RNA-seq data sets as protein-coding RNAs and lncRNAs. We note that miRNAs may be underrepresented in whole-cell RNA-seq libraries due to loss during size-selective bead clean-up of RNA adapters, and that TGIRT-seq poorly detects mature piRNAs due to 2′*O*-methyl groups at their 3′ end, which inhibit template switching. The greater representation of small ncRNAs in TGIRT-seq libraries comes at the cost of a lower proportion of mRNA-mapped reads, but could be compensated for by greater read depth, particularly as sequencing costs continue to decrease and throughput continues to increase.

Finally, we note that the same TGIRT-seq method can be used for transcriptional profiling of a variety of different types of RNA samples, including unfragmented cellular RNA preparations where it was shown previously to give full-length reads of small ncRNAs (Katibah et al. 2014; Shen et al. 2015; Zheng et al. 2015) and for the analysis of protein-bound RNA fragments in procedures like HITS-CLIP or RIP-seq, where the TGIRT template-switching activity enables more efficient and rapid RNA-seq library construction from small amounts of RNA sample than conventional methods. TGIRT-seq should also be readily adaptable for different sequencing platforms.

MATERIALS AND METHODS

Human reference RNA samples

Human reference RNA samples were prepared in the same format used by the SEQC/MAQC-III Consortium and ABRF Initiative using ERCC spike-ins (Jiang et al. 2011; Li et al. 2014; SEQC/MAQC-III Consortium 2014). Briefly, Sample A was prepared by doping 50 µL of Universal Human Reference RNA at 1 µg/µL (Agilent) with 1 µL of ERCC ExFold Mix 1 (Thermo Fisher Scientific), and Sample B was prepared by doping 50 µL of Human Brain Reference RNA at 1 µg/µL (Life Technologies) with 1 µL ERCC ExFold Mix 2. Samples A and B were then mixed in 3:1 or 1:3 ratios to constitute Samples C and D, respectively. Each sample was evaluated for integrity by using an Agilent 2100 Bioanalyzer (RNA 6000 Nano Chip Total Eukaryote RNA Assay, RIN ≥8.2) and then aliquoted and stored at −80°C.

Construction and sequencing of RNA-seq libraries

For RNA-seq library sample replicates, a fresh aliquot of 2 µg of each RNA sample was ribo-depleted by using a RiboZero Gold (Human/Mouse/Rat) kit (Illumina). The RNAs were then fragmented to predominantly 70–100 nucleotide fragments by using an NEBNext Magnesium RNA Fragmentation Module (New England Biolabs) at 94°C for 7 min and treated with T4 polynucleotide kinase (Epicentre) to remove 3′ phosphates and 2′, 3′ cyclic monophosphates, which impede TGIRT template switching (Mohr et al. 2013). RNAs were purified after ribo-depletion, fragmentation, and dephosphorylation by using a modified version of the Zymo RNA Clean & Concentrator protocol (addition of eight sample volumes of ethanol to increase retention of very small RNA species). Half of the recovered RNA was used for cDNA synthesis via TGIRT template switching with 1 µM TGIRT-III RT (InGex, LLC) for 15 min at 60°C, as previously described (Qin et al. 2016). The template-switching reaction seamlessly links the complement of an Illumina Read 2 sequencing primer-binding site (R2R DNA) to the 5′ end of the cDNA during cDNA synthesis, after which a DNA oligonucleotide containing the complement of an Illumina Read 1 sequencing primer-binding site (R1R DNA) is ligated to the cDNA 3′ end by using Thermostable 5′ AppDNA/RNA Ligase (New England Biolabs). Ligated cDNAs with R1R and R2R sequencing adapters on either end were then amplified for 12 cycles of PCR (Qin et al. 2016; initial denaturation at 98°C for 5 sec, followed by cycles of 98°C for 5 sec, 65°C for 10 sec, 72°C for 10 sec), during which Illumina capture and index sequences were added. Libraries were size-selected by using Ampure XP beads (Beckman-Coulter) and evaluated on an Agilent 2100 Bioanalyzer. TGIRT-seq libraries were sequenced using an Illumina NextSeq 500 instrument (75-nt paired-end reads) at the Genome Sequencing and Analysis Facility (GSAF) at the University of Texas at Austin.

Data processing and mapping

Data from TGIRT-Seq libraries were compared to data generated by ABRF on the Illumina HiSeq platform from similarly prepared ribo-depleted and fragmented RNA constructed using Illumina TruSeq v2 and TruSeq v3 protocols downloaded from NCBI Gene Expression Omnibus accession number GSE48035 (sample IDs in Supplemental Table S2). For all data sets, raw reads were trimmed using Trimmomatic (Bolger et al. 2014) with the following options: ILLUMINACLIP:adapters.fa:2:10:10:1:true, LEADING:10, TRAILING:10, SLIDINGWINDOW:4:8, MINLEN:18, AVGQUAL:20. TGIRT-seq data sets were trimmed referencing a customized adapter sequence file. ABRF TruSeq libraries were trimmed by using the TruSeq2-PE.fa file. For comparisons between TGIRT-seq and TruSeq quantitation of relative RNA abundance (Figs. 3, 4; Supplemental Fig. S4), fusion reads (Supplemental Fig. S3), strand specificity (Fig. 5), 5′ to 3′ gene coverage (Fig. 6), and transcript and splice junction detection (Figs. 7, 8; Supplemental Figs. S5, S6), the TGIRT-seq reads were clipped to 50 nt to match the read length of the TruSeq libraries before adapter trimming.

Trimmed sequences from TGIRT-seq and ABRF data sets were aligned to the human reference genome (GRCh38 version 76) supplemented with additional contigs containing 5S rRNA genes (2.2-kb 5S rRNA repeats from the cluster on chromosome 1 [1q42]; GeneBank: X12811) and 45S rRNA genes (43-kb 45S rRNA repeats containing 5.8S, 18S, and 28S rRNA sequences from clusters on chromosomes 13, 14, 15, 21, and 22; GeneBank: U13369) plus ERCC spike-in reference sequences. The alignment was done by using the read mapping pipeline developed previously for TGIRT-seq (Qin et al. 2016) with TopHat replaced by HISAT (Kim et al. 2015) with the following parameters: very-sensitive, –no-mixed, –no-discordant, and –known-splicesite-infile splicesite-file. The splicesite-file was generated by combining the .gtf file of GRCh38 version 76 and the .gtf file of ERCC spike-ins provided by the vendor (Thermo Fisher Scientific). Unmapped and nonconcordantly mapped sequences were then extracted and remapped to the reference sequence by using Bowtie2 – local (Langmead and Salzberg 2012). Concordant reads from the alignment file were extracted and merged with the output from HISAT. The merged .bam files were then filtered by using SAMtools (Li et al. 2009) with option -q 15 to extract uniquely mapped reads. Uniquely mapped read pairs were then converted to .bed files and intersected with the annotations in GRCh38 (version 76) and piRNABank (Sai Lakshmi and Agrawal 2008) using BEDtools (Quinlan and Hall 2010) with intersect options: -f 0.5 -wb. Additionally, reads intersecting tRNA genes were extracted for further alignments to the Genomic tRNA Database (Lowe and Eddy 1997) to improve read assignments. To ensure that overlapping features are limited to the same strand, additional options: -s/-S were used for TGIRT-seq and TruSeq v3, respectively; these options take into account that TGIRT-seq Read 1 corresponds to the RNA strand, while TruSeq Read 1 corresponds to the cDNA strand. TGIRT-seq data were additionally mapped to the miRBase reference (Kozomara and Griffiths-Jones 2014) to identify reads mapping to miRNAs.

Sequence analysis

The results of the intersected .bed files were counted using awk, sort, and uniq commands in a UNIX environment on the Lonestar server from the Texas Advanced Computing Center (TACC). Tables of counts were generated by merging individual count files with customized R scripts. Pairwise differential expression data was generated by DESeq2 (Love et al. 2014). Base distribution was produced by CollectRNASeqMetrics from Picard tools (Broad Institute) using alignments to protein-coding gene RNAs only. Library strandedness and transcript and splice junction detection were determined by RNA-SeQC (Broad Institute; Deluca et al. 2012). Coverage plots, read alignments, and Sashimi plots were generated using the Integrative Genomics Viewer (Robinson et al. 2011; Thorvaldsdóttir et al. 2013). For analyses of splice junction sequences, read containing spliced junctions (annotated as N in the cigar string) were extracted in .bed format with BEDtools bamtobed with options: -cigar and converted to a junction annotation .bed file with a customized program. The .bed file was then parsed by customized python scripts using the pyfaidx package (Shirley et al. 2015) in order to extract stranded dinucleotides from the intron 5′ and 3′ ends.

DATA DEPOSITION

The TGIRT-seq data sets described in this manuscript have been deposited in the National Center for Biotechnology Information Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under accession number SRP066009.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

COMPETING INTEREST STATEMENT

Thermostable group II intron reverse transcriptase (TGIRT) enzymes and methods for their use are the subject of patents and patent applications that have been licensed by the University of Texas and East Tennessee State University to InGex, LLC. A.M.L. and the University of Texas are minority equity holders in InGex, LLC. A.M.L. and some present and former Lambowitz laboratory members receive royalty payments from sales of TGIRT enzymes and licensing of intellectual property.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants GM37949 and GM37951 and Welch Foundation Grant F-1607. We thank Vishwanath Iyer (University of Texas at Austin) for comments on the manuscript and the staff of the Genome Sequencing and Analysis Facility at the University of Texas at Austin for their technical support and assistance.

Received November 30, 2015; accepted December 30, 2015.

REFERENCES

Auffinger P, Westhof E. 1998. Appendix 5: location and distribution of modified nucleotides in tRNA. In *Modification and editing of RNA* (ed. Grosjean H, Benne R), pp. 569–576. American Society of Microbiology Press, Washington, DC.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30:** 2114–2120.

Burroughs AM, Ando Y, de Hoon MJ, Tomaru Y, Nishibu T, Ukekawa R, Funakoshi T, Kurokawa T, Suzuki H, Hayashizaki Y, et al. 2010. A comprehensive survey of 3′ animal miRNA modification events and a possible role for 3′ adenylation in modulating miRNA targeting effectiveness. *Genome Res* **20:** 1398–1410.

Cocquet J, Chong A, Zhang G, Veitia RA. 2006. Reverse transcriptase template switching and false alternative transcripts. *Genomics* **88:** 127–131.

DeLuca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G. 2012. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28:** 1530–1532.

Hafner M, Renwick N, Brown M, Mihailović A, Holoch D, Lin C, Pena JTG, Nusbaum JD, Morozov P, Ludwig J, et al. 2011. RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA* **17:** 1697–1712.

Hansen KD, Brenner SE, Dudoit S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38:** e131.

Harrison GP, Mayo MS, Hunter E, Lever AM. 1998. Pausing of reverse transcriptase on retroviral RNA templates is influenced by secondary structures both 5′ and 3′ of the catalytic site. *Nucleic Acids Res* **26:** 3433–3442.

Jackson TJ, Spriggs RV, Burgoyne NJ, Jones C, Willis AE. 2014. Evaluating bias-reducing protocols for RNA sequencing library preparation. *BMC Genomics* **15:** 569.

Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. 2011. Synthetic spike-in standards for RNA-seq experiments. *Genome Res* **21:** 1543–1551.

Katibah GE, Qin Y, Sidote DJ, Yao J, Lambowitz AM, Collins K. 2014. Broad and adaptable RNA structure recognition by the human interferon-induced tetratricopeptide repeat protein IFIT5. *Proc Natl Acad Sci* **111:** 12025–12030.

Katz Y, Wang ET, Airoldi EM, Burge CB. 2010. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* **7:** 1009–1015.

Kim D, Salzberg SL. 2011. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol* **12:** R72.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12:** 357–360.

Kirino Y, Mourelatos Z. 2007. Mouse Piwi-interacting RNAs are 2′-O-methylated at their 3′ termini. *Nat Struct Mol Biol* **14:** 347–348.

Koppers-Lalic D, Hackenberg M, Bjinsdorp IV, van Eijndhoven MAJ, Sadek P, Sie D, Zini N, Middeldorp JM, Ylstra B, de Menezes RX, et al. 2014. Nontemplated nucleotide additions distinguish the small RNA composition in cells from exosomes. *Cell Rep* **8:** 1649–1658.

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* **42:** D68–D73.

Kwok CK, Ding Y, Sherlock ME, Assmann SM, Bevilacqua PC. 2013. A hybridization-based approach for quantitative and low-bias single-stranded DNA ligation. *Anal Biochem* **435:** 181–186.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359.

Lee C, Harris RA, Wall JK, Mayfield RD, Wilke CO. 2013. RNase III and T4 polynucleotide kinase sequence bias and solutions during RNA-seq library construction. *Biol Direct* **8:** 16.

Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. 2010. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* **7:** 709–715.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25:** 2078–2079.

Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, Viale A, Wright C, Schweitzer PA, Gao Y, et al. 2014. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol* **32:** 915–925.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25:** 955–964.

Mader RM, Schmidt WM, Sedivy R, Rizovski B, Braun J, Kalipciyan M, Exner M, Steger GG, Mueller MW. 2001. Reverse transcriptase template switching during reverse transcriptase-polymerase chain reaction: artificial generation of deletions in ribonucleotide reductase mRNA. *J Lab Clin Med* **137:** 422–428.

Malboeuf CM, Isaacs SJ, Tran NH, Kim B. 2001. Thermal effects on reverse transcription: improvement of accuracy and processivity in cDNA synthesis. *Biotechniques* **30:** 1074–1078.

MAQC Consortium, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, et al. 2006. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24:** 1151–1161.

Mohr S, Ghanem E, Smith W, Sheeter D, Qin Y, King O, Polioudakis D, Iyer VR, Hunicke-Smith S, Swamy S, et al. 2013. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* **19:** 958–970.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat Methods* **5:** 621–628.

Oshlack A, Wakefield MJ. 2009. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* **4:** 14.

Ouhammouch M, Brody EN. 1992. Temperature-dependent template switching during in vitro cDNA synthesis by the AMV-reverse transcriptase. *Nucleic Acids Res* **20:** 5443–5450.

Ozsolak F, Milos PM. 2011. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* **12:** 87–98.

Parkhomchuk D, Borodina T, Amstislavskly V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A. 2009. Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res* **37:** e123.

Pelechano V, Steinmetz LM. 2014. Gene regulation by antisense transcription. *Nat Rev Genet* **14:** 880–893.

Perocchi F, Xu Z, Clauder-Münster S, Steinmetz LM. 2007. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res* **35:** e128.

Qin Y, Yao J, Wu DC, Nottingham RM, Mohr S, Hunicke-Smith S, Lambowitz AM. 2016. High-throughput sequencing of human plasma RNA by using thermostable group II intron reverse transcriptases. *RNA* **22:** 111–128.

Quinlan AR, Hall IM. 2010. BEDtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842.

Raabe CA, Tang T-H, Brosius J, Rozhdestvensky TS. 2014. Biases in small RNA deep sequencing data. *Nucleic Acids Res* **42:** 1414–1426.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29:** 24–26.

Ruprecht RM, Goodman NC, Spiegelman S. 1973. Conditions for the selective synthesis of DNA complementary to template RNA. *Biochim Biophys Acta* **294:** 192–203.

Sai Lakshmi S, Agrawal S. 2008. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res* **36:** D173–D177.

SEQC/MAQC-III Consortium, Su Z, Łabaj PP, Li S, Thierry-Mieg J, Thierry-Mieg D, Shi W, Wang C, Schroth GP, Setterquist RA, et al. 2014. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol* **32:** 903–914.

Shen PS, Park J, Qin Y, Li X, Parsawar K, Larson MH, Cox J, Cheng Y, Lambowitz AM, Weissman JS, et al. 2015. Protein synthesis. Rqc2p and 60S ribosomal subunits mediate mRNA-independent elongation of nascent chains. *Science* **347:** 75–78.

Shirley MD, Ma Z, Pedersen BS, Wheelan SJ. 2015. Efficient "pythonic" access to FASTA files using pyfaidx. *PeerJ PrePrints* **3:** e1196.

Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14:** 178–192.

Vikman P, Fadista J, Oskolkov N. 2014. RNA sequencing: current and prospective uses in metabolic research. *J Mol Endocrinol* **53:** R93–R101.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10:** 57–63.

Wery M, Descrimes M, Thermes C, Gautheret D, Morillon A. 2013. Zinc-mediated RNA fragmentation allows robust transcript reassembly upon whole transcriptome RNA-Seq. *Methods* **63:** 25–31.

Westermann AJ, Gorski SA, Vogel J. 2012. Dual RNA-seq of pathogen and host. *Nat Rev Microbiol* **10:** 618–630.

Wu W, Henderson LE, Copeland TD, Gorelick RJ, Bosche WJ, Rein A, Levin JG. 1996. Human immunodeficiency virus type 1 nucleocapsid protein reduces reverse transcriptase pausing at a secondary structure near the murine leukemia virus polypurine tract. *J Virol* **70:** 7132–7142.

Wyman SK, Knouf EC, Parkin RK, Fritz BR, Lin DW, Dennis LM, Krouse MA, Webster PJ, Tewari M. 2011. Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res* **21:** 1450–1461.

Zajac P, Islam S, Hochgerner H, Lönnerberg P, Linnarsson S. 2013. Base preferences in non-templated nucleotide incorporation by MMLV-derived reverse transcriptases. *PLOS One* **8:** e85270.

Zheng G, Qin Y, Clark WC, Dai Q, Yi C, He C, Lambowitz AM, Pan T. 2015. Efficient and quantitative high-throughput tRNA sequencing. *Nat Methods* **12:** 835–837.

Zhuang F, Fuchs RT, Sun Z, Zheng Y, Robb GB. 2012. Structural bias in T4 RNA ligase-mediated 3′ adapter ligation. *Nucleic Acids Res* **40:** e54.

RNA
A PUBLICATION OF THE RNA SOCIETY

RNA-seq of human reference RNA samples using a thermostable group II intron reverse transcriptase

Ryan M. Nottingham, Douglas C. Wu, Yidan Qin, et al.

RNA published online January 29, 2016