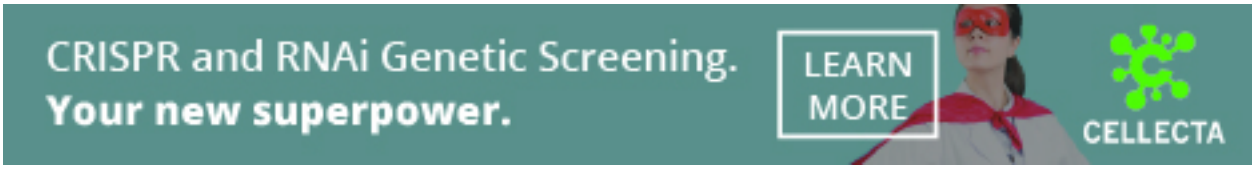
Transposase mediated construction of RNA-seq libraries

Jason Gertz, Katherine E. Varley, Nicholas S. Davis, et al.

Genome Res. 2012 22: 134-141 originally published online November 29, 2011
Access the most recent version at doi:10.1101/gr.127373.111



Method

Transposase mediated construction of RNA-seq libraries

Jason Gertz,[1] Katherine E. Varley,[1] Nicholas S. Davis,[1] Bradley J. Baas,[2] Igor Y. Goryshin,[2] Ramesh Vaidyanathan,[2] Scott Kuersten,[2] and Richard M. Myers[1,3]

[1]*HudsonAlpha Institute for Biotechnology, Huntsville, Alabama 35806, USA;* [2]*Epicentre (An Illumina Company), Madison, Wisconsin 53713, USA*

RNA-seq has been widely adopted as a gene-expression measurement tool due to the detail, resolution, and sensitivity of transcript characterization that the technique provides. Here we present two transposon-based methods that efficiently construct high-quality RNA-seq libraries. We first describe a method that creates RNA-seq libraries for Illumina sequencing from double-stranded cDNA with only two enzymatic reactions. We generated high-quality RNA-seq libraries from as little as 10 pg of mRNA (~1 ng of total RNA) with this approach. We also present a strand-specific RNA-seq library construction protocol that combines transposon-based library construction with uracil DNA glycosylase and endonuclease VIII to specifically degrade the second strand constructed during cDNA synthesis. The directional RNA-seq libraries maintain the same quality as the nondirectional libraries, while showing a high degree of strand specificity, such that 99.5% of reads map to the expected genomic strand. Each transposon-based library construction method performed well when compared with standard RNA-seq library construction methods with regard to complexity of the libraries, correlation between biological replicates, and the percentage of reads that align to the genome as well as exons. Our results show that high-quality RNA-seq libraries can be constructed efficiently and in an automatable fashion using transposition technology.

[Supplemental material is available for this article.]

RNA-seq is a powerful technique that allows for sensitive digital quantification of transcript levels (Mortazavi et al. 2008; Nagalakshmi et al. 2008). It enables the detection of noncanonical transcription start sites (Liu et al. 2011) as well as termination sites (Wang et al. 2008), alternative splice isoforms (Wang et al. 2008; Jiang and Wong 2009), transcript mutations/editing (Rosenberg et al. 2011), and allelic biases in transcript abundance (Pickrell et al. 2010). Methods that preserve the strand from which the transcript originated also allow for the identification of antisense transcription (He et al. 2008; Perkins et al. 2009), which can play a role in post-transcriptional regulation. Because of the power of RNA-seq and the prevalence of aberrant gene-expression patterns in many diseases, there is a growing need to construct libraries efficiently from low starting amounts of RNA in a high-throughput and reproducible fashion.

Ultra-high throughput, "next-generation" DNA sequencing library construction is a time-consuming process that typically has some sample loss at each step. A recent advance in library construction is the use of transposases to randomly integrate sequencing adapters into the DNA of interest (Adey et al. 2010). This approach creates sequencing-ready DNA libraries in a few steps with minimal hands-on time. The resulting libraries exhibit even coverage across the human genome when constructed from low amounts of genomic DNA (Adey et al. 2010). Transposon-based library construction has also been successfully applied to pyrosequencing of the RNA genomes of strains of simian hemorrhagic fever virus (Lauck et al. 2011). The success of transposon-based genomic library construction for genomic analyses suggests that it should be possible to use transposases to construct high-quality RNA-seq libraries.

Recently, several techniques developed for constructing RNA-seq libraries which maintain the transcript strand-of-origin were evaluated (Levin et al. 2010). Each protocol had varying levels of strand specificity, library complexity, and reproducibility. One of the overall best methods tested involved incorporating uracil into the second cDNA strand. The strand is subsequently degraded specifically by treatment with uracil DNA glycosylase and endonuclease VIII, which leaves only sequence reads that map to the strand-of-origin of each transcript (Parkhomchuk et al. 2009). The application of transposases to construct strand-specific RNA-seq libraries is an appealing approach for efficiently creating RNA-seq libraries with maximal information.

Here we describe the development of a transposon-based method for RNA-seq library construction, called Tn–RNA-seq. The method is fast and requires only two steps and two purifications after cDNA is made. The protocol is fully automatable and is compatible with robotics. We also extend and modify the transposase-based RNA-seq method to create directional RNA-seq libraries capable of preserving the strand information from which the transcript originated.

Results

Efficient transposition-based RNA-seq library construction

The strategies of each protocol are outlined in Figure 1. To construct nondirectional standard RNA-seq libraries, we prepared double-stranded (ds) cDNA from fragmented mRNA (Mortazavi et al. 2008). The ds cDNA was end-repaired, A-tailed, ligated to sequencing adapters, and amplified (Fig. 1A). For the nondirectional transposon-based RNA-seq method (Tn–RNA-seq), mRNA was not frag-

[3]**Corresponding author.**
E-mail rmyers@hudsonalpha.org.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.127373.111.

22:134–141 © 2012 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/12; www.genome.org

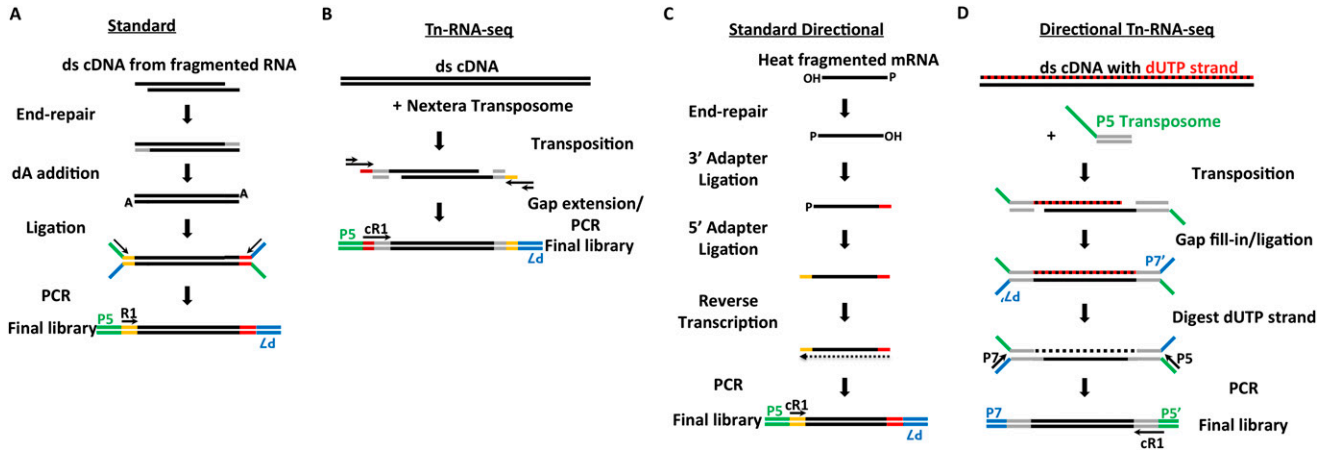


Figure 1. RNA-seq methods overview. (*A*) In the standard adapter ligation RNA-seq library construction protocol, double-stranded cDNA made from fragmented mRNA is subjected to end repair, dATP addition, adapter ligation, size selection, and PCR. (*B*) For the Tn–RNA-seq method described here, double-stranded cDNA is incubated with transposome (transposase complexed with transposon) and then undergoes PCR. (*C*) In the directional RNA ligation approach (standard directional), poly(A)-selected mRNA is fragmented with heat and end repaired. 3′ and then 5′ adapters are ligated onto single-stranded RNA before reverse transcription, followed by PCR. (*D*) The directional Tn–RNA-seq library construction described here starts with double-stranded cDNA, in which the second strand synthesized contains uracils instead of thymines. cDNA synthesis is followed by transposition of sequencing adapters, gap fill-in/ligation, USER digestion, and PCR. P5 and P7 correspond to Illumina cluster generation primers. R1 identifies the sequencing primer and cR1 indicates custom sequencing primer. (′) Reverse complement.

mented before cDNA synthesis. Instead, we incubated the ds cDNA with a transposome (hyperactive Tn5 transposase bound to synthetic 19-bp mosaic end-recognition sequences appended to Illumina sequencing adapters) (Adey et al. 2010) to simultaneously fragment and attach adapters (Fig. 1B). Because the transposome is a mixture containing two different sequences (shown in red and yellow in Fig. 1B), it can insert in either orientation, resulting in a nondirectional library. During the transposition process, only the transferred strand of each transposon end is covalently linked to the target DNA. Due to the staggered fashion of the transposition, a 9-bp gap between the nontransferred strand and the target DNA is created. Extension synthesis from the target DNA into this gap, followed by copying of the attached transposon end by strand displacement, creates the 3′ adapter sequence. Suppression PCR (Rand et al. 2005) is then used to select for templates with heterologous adapters. During PCR, index barcodes can be added to allow for the mixing of multiple samples in one sequencing lane. Following purification, PCR products are ready for single-end or paired-end sequencing with custom sequencing primers.

The entire process is automatable and feasible in 96-well plates, making large-scale Tn–RNA-seq library construction with robotics an appealing combination. Standard RNA-seq library construction (Mortazavi et al. 2008) requires multiple enzymatic reactions between cDNA synthesis and the final PCR step, compared with the one reaction with the Tn–RNA-seq method we describe here (Fig. 1B). The Tn–RNA-seq protocol cuts down significantly on sample preparation time and could yield higher quality RNA-seq libraries by minimizing sample loss during multiple reactions and purifications.

We constructed RNA-seq libraries with the standard method and the Tn–RNA-seq method to compare library quality. We extracted high-quality total RNA (RNA integrity number of 9.5 on an Agilent Bioanalyzer) from the human endometrial adenocarcinoma cell line ECC-1 (Mo et al. 2006), and purified ECC-1 mRNA by using poly(A) selection on magnetic beads. Two biological replicates were used for each method with a starting amount of 50 ng of ECC-1 mRNA. Each library was sequenced on one lane of an Illumina GAIIx using a custom sequencing primer specific to the transposon end to generate an average of 28 million pass-filter 36-bp reads. We used multiple metrics to assess quality of each library. The results are shown in Table 1 (see Methods for details on calculations) and some typical examples are presented in Figure 2 and Supplemental Figure S5. The libraries were evaluated on how well they aligned to exons, their complexity, and their biological reproducibility. In each of these categories, the Tn–RNA-seq protocol showed similar performance to the standard protocol, indicating that high-quality RNA-seq libraries can be constructed using transposition technology.

We also evaluated 5′ and 3′ bias in each library by calculating the relative coverage across transcripts. Figure 3 shows that the Tn–RNA-seq libraries show a subtle depletion in coverage at the 10%-most 5′ ends of transcripts. This depletion is due to the nature of the transposon-based library construction. To sequence the ends of transcripts, a transposase would have to integrate one transposon near the very end of the transcript and that transposon would have to be in the correct orientation with the sequencing primer facing toward the 3′ end of the transcript. Even in this case, only sequencing reads generated from one strand would map to the most 5′ end of the transcript. Depletion is not seen on the 3′ end of the transcript, which is most likely due to the presence of the poly(A) tail, which gives the transposase extra substrate to integrate the transposon. The depletion of 5′ ends is seen across all sizes of transcripts (Supplemental Fig. S2), and when analyzed on a base-pair scale, corresponds to a more than twofold depletion in coverage of the first 50 nt of the transcript relative to the standard method (Supplemental Fig. S3). We observed a subtle reduction in the number of short transcripts, <200 nt, which were detectable with the Tn–RNA-seq approach. We found that 396 short transcripts were detectable with the standard approach and 348 short transcripts were detectable with the Tn–RNA-seq method, which represents a 12% reduction in the number of short transcripts detected. It is important to point out that the higher alignment percentage to exons of the Tn–RNA-seq method, compared with the standard protocol, may be due to the depletion of 5′ ends,

Table 1. RNA-seq library construction comparison in ECC-1

Protocol	Reads aligned to genome	Percent of aligned reads that map to RefSeq transcripts	Complexity	Biological Replicate Correlation (Pearson)	Biological Replicate Correlation (Spearman)
Standard Rep1	18,347,613	73.7%	85.77%	0.983	0.986
Standard Rep2	9,218,462	74.3%	85.79%		
Tn–RNA-seq Rep1	22,945,616	81.3%	89.08%	0.955	0.986
Tn–RNA-seq Rep2	18,513,940	85.1%	85.20%		
Directional Tn–RNA-seq Rep1	20,474,907	74.5%	81.60%	0.981	0.988
Directional Tn–RNA-seq Rep2	25,848,260	72.5%	81.43%		

because observed 5′ ends of transcripts may differ from RefSeq annotations. While there is a slight depletion in coverage at the 5′ end of transcripts, the impact on library quality and expression measurements (discussed below) is negligible.

To determine whether the Tn–RNA-seq protocol produces data indicating the same gene-expression levels as does the standard protocol, we calculated RPKM (reads per kilobase per million aligned reads) (Mortazavi et al. 2008) values for each RefSeq gene (see Methods). The results are shown in Figure 4. The Pearson correlation (r) between log base 2 of RPKM values from the standard and Tn–RNA-seq protocols is 0.959. The Spearman rank correlation (ρ), which is more appropriate given the overall distribution of RPKM values, is 0.979. The high correlation in expression values indicates that the Tn–RNA-seq protocol allows for efficient construction of high-quality RNA-seq libraries while maintaining the integrity of transcript measurements.

Consistent Tn–RNA-seq libraries constructed from low amounts of input mRNA

We next sought to establish whether the Tn–RNA-seq method is robust to differing amounts of starting material and determine the amount of mRNA required to construct a reliable RNA-seq library. To test library construction with lower starting amounts of mRNA, we constructed seven Tn–RNA-seq libraries with between 10 ng and 1 pg of mRNA. The yield of Tn–RNA-seq library construction was dependent on the amount of mRNA that was used. Libraries made with between 10 and 0.5 ng of mRNA yielded ~600 ng of DNA, while libraries constructed with <100 pg of mRNA yielded between 30 and 100 ng of DNA.

The quality metrics for each Tn–RNA-seq library are displayed in Supplemental Table S1. All six libraries made from between 10 ng and 10 pg of mRNA had at least 72% of aligned reads map to known transcripts, while the library made from 1 pg of mRNA had 62% of aligned reads map to known transcripts. Library complexity also remained high for all libraries except for the library constructed with 1 pg of mRNA (Fig. 5A). In general, Tn–RNA-seq libraries made with 10 pg or more of mRNA exhibited consistent quality measures, showing that high-quality RNA-seq libraries can be constructed with the transposon-based method from as little as 10 pg of mRNA, which represents ~1 ng of total RNA or ~200 cell equivalents.

We also examined whether expression measurements were consistent across different amounts of starting materials (Supplemental Fig. S4). We found that all libraries made from at least 10 pg of mRNA were very consistent with the libraries constructed from 50 ng of mRNA. For all libraries except for the library made with 1 pg of mRNA, the rank correlation of expression measurements with the 50 ng of mRNA library exceeded 0.96 (Fig. 5B).

Library insert size is influenced by the amount of mRNA used; smaller amounts of starting material result in smaller insert sizes (Supplemental Fig. S1). This is due to the relative ratio of target DNA to transposome, since the transposase does not enzymatically turn over in these reactions. Based on these observations, it may be possible to alter the insert size by changing the concentration of transposase relative to the amount of mRNA. Our results indicate that transposon-based library construction can be used on limiting amounts of mRNA as low as 10 pg.

Strand-specific transposon-based RNA-seq library construction

A limitation of the above-mentioned transposon-based approach is that the transposition reaction is inherently nondirectional. This means that the resulting cDNA is captured without regard to the original transcript strand information. To create libraries that preserve strand information we adapted a previously described approach to specifically mark one strand of cDNA by incorporating

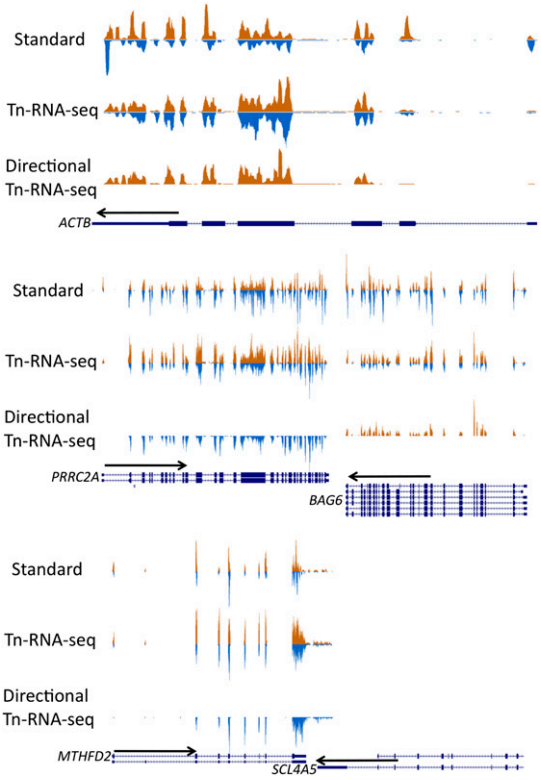


Figure 2. The directional Tn–RNA-seq method exhibits complete strand specificity. Aligned reads for the standard, Tn–RNA-seq, and directional Tn–RNA-seq method are displayed for three genomic loci. Reads mapping to the positive strand are shown in orange and reads mapping to the negative strand are shown in blue. Arrows indicate the direction of transcription for each RefSeq gene.

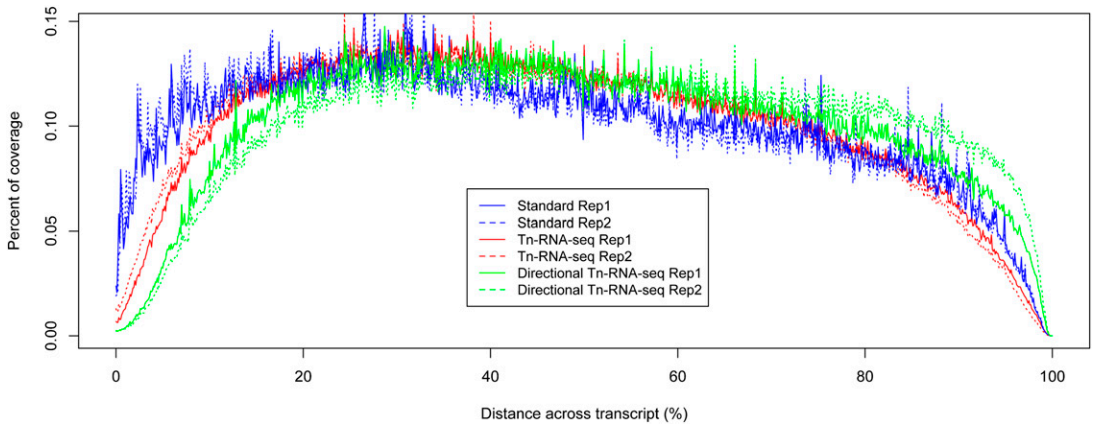


Figure 3. Transposon-based libraries show expected depletion of coverage at 5′ ends of transcripts. The percentage of coverage (*y*-axis), averaged across all transcripts is plotted as a function of distance across the transcripts (*x*-axis). 0% corresponds to the 5′ ends and 100% corresponds to the 3′ ends of transcripts.

dUTP during the second-strand synthesis (Parkhomchuk et al. 2009). We modified the Tn–RNA-seq method to accommodate uracil-containing cDNA and preserve the stranded information content of the samples. After adapters were attached, the combination of uracil DNA glycosylase (UDG) and endonuclease VIII (Endo VIII) degraded the second strand, leaving only the first strand of cDNA, which is the reverse complement of the original transcript.

After first-strand cDNA synthesis from 50 ng of ECC-1 mRNA, we treated the reaction with a nucleotide phosphatase to remove nucleotides, since free nucleotide contamination in the second-strand reaction would result in a decrease in strand specificity. The reaction was then column purified and used for second-strand cDNA synthesis in the presence of a nucleotide mix containing dUTP instead of dTTP.

Purified uracil-containing double-stranded cDNA was then incubated with a single transposome containing a unique sequence (P5), which is appended to the 5′ end of the transferred strand (Fig. 1D, shown in green). After transposition, DNA fragments contain the P5 sequence at the 5′ ends of both strands of cDNA. The nontransferred strand is replaced with a modified oligonucleotide containing a different sequence (P7) appended to the 5′ end (Fig. 1D, shown in blue) and the 9-bp gap is filled in and ligated to the template. Because the cDNA is marked, the uracil-containing second strand can be removed prior to PCR by treating the cDNA library with UDG and Endo VIII. The surviving fragments are then amplified and enriched using Phusion DNA polymerase, which is very inefficient at extending templates that contain uracils, providing an additional level of strand specificity (Greagg et al. 1999). We sequenced 36 bases of these final libraries from a single end on an Illumina GAIIx using a custom sequencing primer specific to the P5-containing transposon end. This directional library method (directional Tn–RNA-seq) is designed to produce all sequencing tags oriented 3′–5′ relative to the original RNA transcripts.

We observed striking strand specificity in the genomic alignments produced from these libraries (Fig. 2). *ACTB*, one of the highest expressed genes in ECC-1, is shown in Figure 2, top, and we found that all reads map to the expected strand. Determining strand specificity can help to disambiguate some genes; for example,

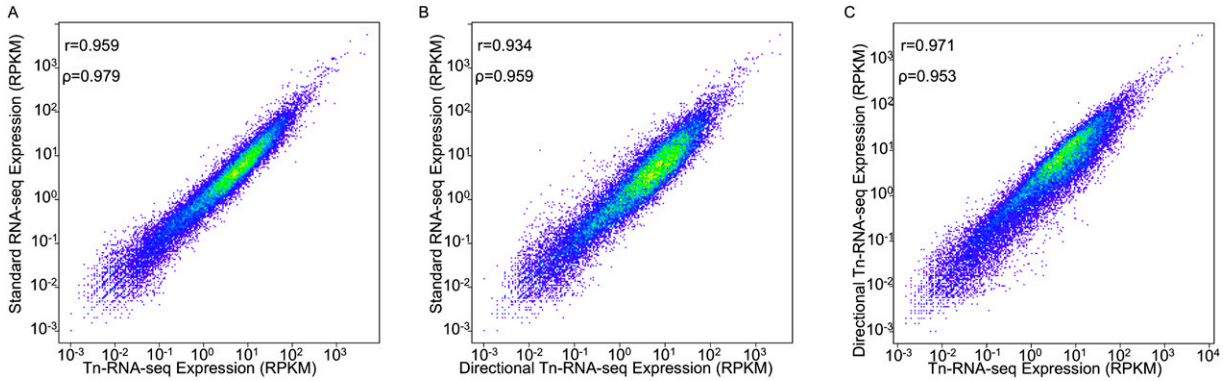


Figure 4. Expression values are consistent between standard RNA-seq library construction and transposon-based RNA-seq library construction in ECC-1. (*A*) Scatterplot showing expression values for standard RNA-seq library construction (*y*-axis) and the Tn–RNA-seq library construction (*x*-axis). The Pearson correlation between the standard and Tn–RNA-seq protocols is 0.959, and the Spearman rank correlation is 0.979. (*B*) Scatterplot displaying expression values for standard RNA-seq library construction (*y*-axis) and the directional Tn–RNA-seq library construction (*x*-axis). The Pearson correlation between the standard and directional Tn–RNA-seq protocols is 0.934, and the Spearman rank correlation is 0.959. (*C*) Scatterplot displaying expression values for Tn–RNA-seq library construction (*x*-axis) and the directional Tn–RNA-seq library construction (*y*-axis). The Pearson correlation between the Tn–RNA-seq and directional Tn–RNA-seq protocols is 0.971, and the Spearman rank correlation is 0.953.

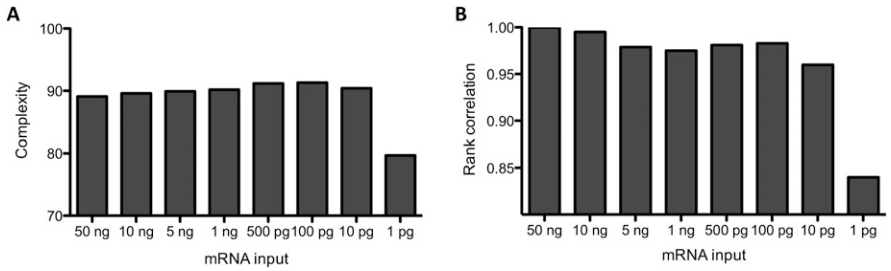


Figure 5. Tn–RNA-seq libraries constructed from as low as 10 pg of mRNA are high quality and show highly correlated expression levels. (*A*) Library complexity, calculated as the number of different alignment positions in a random set of 1 million aligned reads divided by 1 million, is shown for libraries made with between 50 ng and 1 pg of mRNA. (*B*) Rank correlations of expression measurements between the library constructed with 50 ng of mRNA and every other Tn–RNA-seq library are displayed.

SLC4A5 (Fig. 2, bottom). When measured by nondirectional RNA-seq, *SLC4A5* appears expressed because of reads mapping to the 3′ end of the gene. However, strand-specific RNA-seq shows that these reads originate from the antisense strand, not the coding strand, and represent either antisense transcription or read-through of the 3′ end of *MTHFD2*, a gene 3′ and oriented opposite to the *SLC4A5* gene.

To determine the overall strand specificity of the directional Tn–RNA-seq method, we calculated the percentage of reads mapping to the expected strand of RefSeq genes. We found that in both replicates >99.4% of reads map to the expected strand (99.5% and 99.4% for individual replicates). This is likely an underestimate of the strand specificity of the method, as there is expected to be some antisense transcription as well as alternative 5′ and 3′ UTR boundaries (as may be the case with *SLC4A5*) that are not represented in the RefSeq annotations. This level of strand specificity is in the same range as the most strand-specific methods analyzed by Levin et al. (2010) in yeast, indicating that the directional Tn–RNA-seq method exhibits a degree of strand specificity that is comparable to the most specific methods available.

The coverage across the length of transcripts for the directional Tn–RNA-seq RNA-seq libraries yields an interesting pattern. We observed substantial depletion at the 5′ end of transcripts and increased coverage at the 3′ end of transcripts compared with the standard RNA-seq library construction protocol. This pattern can be explained by the strand specificity of the directional Tn–RNA-seq method. Strand-specific reads in these libraries should always sequence from the 3′ end toward the 5′ end of the transcript. This would cause depletion in coverage at the 5′ end, because two transposition events near the 5′ end would be required to sequence the 5′−most portion of the transcript. The 3′ end harbors an overabundance of reads compared with the standard method because, regardless of where the transposon is integrated, the 3′ most transposon will be the sequencing primer that generates the sequence read.

While the directional Tn–RNA-seq method yields highly strand-specific libraries, we also wanted to assess the quality of the libraries using the same metrics discussed above. Table 1 shows that the directional Tn–RNA-seq libraries have similar levels of reproducibility, complexity, and alignability. The complexity of the directional Tn–RNA-seq libraries is lower compared with the standard and nondirectional Tn–RNA-seq libraries. We believe that this is due in part to a reduction in the number of possible alignment locations. The number of possible unique genome mapping locations, which includes the strand that the read matches, is cut by half in the directional Tn–RNA-seq libraries due to the strand specificity. Overall, these results show that our directional Tn–RNA-seq protocol results in high-quality strand-specific RNA-seq libraries that preserve transcript measurements (Fig. 4B,C).

To compare the directional Tn–RNA-seq with a standard directional RNA-seq protocol, we created strand-specific RNA-seq libraries using single-stranded RNA ligation (Lister et al. 2008), similar to the Illumina TruSeq small RNA protocol (Fig. 1C; see Methods). We created strand-specific libraries from universal human reference RNA (Novoradovskaya et al. 2004) using both methods to compare performance. For each library, we calculated the percentage of reads mapping to the expected strand of RefSeq genes. Similar strand specificity was observed with each protocol. The library constructed with the standard directional approach exhibited 99.46% of reads mapping to the expected strand, and the library constructed with the directional Tn–RNA-seq method had 99.51% of reads mapping to the expected strand. We next analyzed expression measurements from the directional libraries and found a high correlation (rank correlation: 0.96) between the two methods (Supplemental Fig. S6). These results indicate that the directional Tn–RNA-seq method maintains the same strand specificity of a standard method, and that expression measurements are also consistent between directional approaches.

Discussion

We have described two techniques for constructing RNA-seq libraries that are based on the introduction of sequencing adapters by transposition into double-stranded cDNA. The first method described is an efficient method for creating strand ambiguous libraries that requires only one enzymatic step to go from double-stranded cDNA to fragments ready to be amplified before sequencing. We found that libraries constructed in this manner performed as well as libraries made using the more laborious standard adapter ligation-based approach. We also found that that the transposon-based approach yielded high-quality RNA-seq libraries that preserved transcript measurements with as low as 10 pg of mRNA. This reduction in required starting material for RNA-seq library construction provides the opportunity to create reproducible RNA-seq libraries from rare cell types or small samples. The transposon-based RNA-seq approach is an attractive option for RNA-seq library construction because of the protocol's efficiency and efficacy. This is especially true for labs preparing a large number of samples, with or without robotics, because the library preparation starting from poly(A) selection takes <8 h to complete.

We also present a method that preserves the strand-of-origin for each transcript sequenced. Knowing the strand orientation of the transcripts can lead to interesting findings about transcript structure (Core et al. 2008; He et al. 2008; Seila et al. 2008). The strand specificity of the directional Tn–RNA-seq method comes from specific digestion of the second cDNA strand combined with novel transposome modifications to control the attachment of specific sequences to the template cDNA. The directional method is more time consuming than the nondirectional transposon-based method, but it provides additional information while maintaining a high level of complexity and reproducibility. The strand specificity is near complete at 99.5% and similar to the best-

published methods in yeast (Levin et al. 2010) and to our results using an RNA ligation approach.

While both transposon-based RNA-seq library construction techniques exhibit high-quality sequencing results, there is a subtle depletion of sequence near the 5′ ends of transcripts that is more pronounced with the directional method. This depletion is expected and unavoidable for directional Tn–RNA-seq due to the nature of the transposition and strand specificity. This depletion at the 5′ ends of transcripts could be lessened by modifying the protocol to sequence toward the 3′ end of the transcript as opposed to toward the 5′ end of the transcript in the protocol presented.

The nondirectional Tn–RNA-seq library construction is amenable to large-scale library construction and automation. Because every enzymatic step is followed by magnetic bead purification (Hawkins et al. 1994), the full library construction protocol can easily be applied to a 96-well plate format, where steps can be completed with robotics. The protocol also allows for multiplex sequencing of samples (Smith et al. 2010). Molecular barcodes can be added during the final PCR step by using different primers, which can result in a significant cost savings. Since the transposase binds to a particular sequence, the sequencing adapters introduced are different from the standard Illumina adapters. Therefore, the Tn–RNA-seq libraries need to be mixed with a custom primer to be sequenced, but otherwise require no special experimental or computational accommodations. Both transposase-based approaches to constructing RNA-seq libraries that are described in this work provide an efficient and streamlined workflow to achieve high-quality characterization of the transcriptome comparable to the current more laborious methods.

Methods

Cell culture and mRNA isolation

The human endometrial cancer cell line ECC-1 was grown in RPMI-1640 (Invitrogen) supplemented with 10% fetal bovine serum (Hyclone) and 1% penicillin-streptomycin (Invitrogen). Two separate growth replicates were used to assess biological replication. To isolate total RNA, we used the Animal Tissue RNA Isolation kit (Norgen) with ~5 million cells scraped from a 100-mm cell culture dish. The samples are DNase treated during the purification, which is important because genomic DNA contamination can be efficiently made into sequencer-ready molecules during the transposition step. Universal human reference RNA was purchased from Agilent. After total RNA was purified, mRNA was enriched using the Dynabeads mRNA Purificaiton Kit (Invitrogen) with the following modifications. The beads were washed twice, instead of once, with Wash Buffer B before each elution. Each sample went through two rounds of binding, washing, and elution. Samples were eluted in 20 μL of Tris-HCL elution buffer during the final elution. All RNA and DNA concentrations were measured with a Qubit fluorometer (Invitrogen).

Standard RNA-seq library construction

Standard library construction was performed as previously described (Mortazavi et al. 2008). For each biological replicate, 50 ng of mRNA was used for each library.

Tn–RNA-seq library construction

Primer and adapter sequences for both transposon-based protocols can be found in Supplemental Table S2. To make cDNA, 1 μL (3 μg) of random hexamers (Invitrogen) was added to poly(A)-selected mRNA in a volume of 20 μL of Tris-HCL elution buffer and incubated at 65°C for 5 min, then placed on ice. First-strand cDNA synthesis was performed by adding 4 μL of First Strand Buffer (Invitrogen), 2 μL of 100 mM DTT (Invitrogen), 0.5 μL of RNaseOUT (Invitrogen), and 1 μL of Superscript II (200 U/μL, Invitrogen), and incubating the mix at 25°C for 12 min, 42°C for 50 min, then 70°C for 15 min. The second strand of cDNA was filled-in by adding 16 μL of water, 5 μL of 10× second Strand Buffer (500 mM Tris-HCl at pH 7.8, 50 mM MgCl2, 10 mM DTT), 3 μL of 10 mM dNTPs (New England Biolabs—NEB), 1 μL of RNase H (10 U/μL, Invitrogen), and 5 μL of DNA Polymerase I (10 U/μL, Invitrogen) to the first-strand reaction on ice, and then incubating at 16°C for 2.5 h. The cDNA was purified with AMPure beads (Beckman Coulter) according to the manufacturer's instructions and eluted in 15 μL of EB (Qiagen).

To incorporate sequencing adapters, we combined the purified cDNA with 4 μL of TA buffer (33 mM Tris-acetate at pH 7.5, 66 mM potassium acetate, 10 mM magnesium acetate, and 0.5 mM DTT) and 0.2 μL of Nextera Enzyme (Epicentre) on ice and incubated at 55°C for 5 min, and then placed the sample on ice. We added 30 μL of QG buffer (Qiagen) to stop the transposase reaction and purified the DNA with 90 μL of AMPure beads, eluting in 22 μL of EB. To PCR amplify the fragments, we added 25 μL of Nextera PCR buffer, 1 μL of 50× Nextera Primer Cocktail, 1 μL of Nextera Adapter 2, and 1 μL of Nextera PCR enzyme (Epicentre) to the purified fragments for a total volume of 50 μL. The reaction was incubated at 72°C for 3 min, then at 95°C for 30 sec, followed by 15 cycles of 95°C for 10 sec, 62°C for 30 sec, and 72°C for 3 min. We purified the PCR amplicons with 90 μL of AMPure beads per the manufacturer's instructions and eluted in 32 μL of EB. We sequenced libraries at a concentration of 6 pM on an illumina GAIIx sequencer with a custom sequencing primer designed to anneal to the transposon sequence. Libraries constructed with <10 ng of mRNA were barcoded and sequenced on an Illumina HiSeq 2000 with a custom sequencing primer and custom index primer. For optimal sequencing results, we found that using 75 ng or less of double-stranded cDNA in the transposition reaction is important. Using larger amounts of cDNA can lead to insert sizes of >1000 bp that do not sequence well.

Directional Tn–RNA-seq library construction

To construct the first strand of cDNA, 50 ng of mRNA in a volume of 20 μL were added to 1 μL (3 μg) of random hexamers (Invitrogen), heated to 65°C for 5 min, and placed directly on ice. Then, 6 μL of 5x first strand buffer, 1 μL of 100 mM DTT, 1 μL of 10 mM dNTPs [NEB], 0.5 μL of RNaseOUT [Invitrogen], 0.5 μL of Actinomycin D [120 ng/μL, Sigma], and 1 μL of Superscript III [200 U/μL, Invitrogen] were added to mRNA/primer mix at room temperature. The reaction was then incubated at 40°C for 90 min and heat inactivated at 70°C for 15 min. The reaction was cooled to 37°C and 1 μL of RNase H (10 U/μL, Invitrogen) and 1 μL of NTPhos (20 U/μL, Epicentre) were added. The reaction was incubated at 37°C for 30 min, then heat inactivated at 75°C for 15 min. The first-strand cDNA was purified with the QIAquick PCR Purification kit (Qiagen) per the manufacturer's instructions and eluted in 25 μL of EB (Qiagen). It was then purified further with a prepared G50 Sephadex column (USA Scientific) to ensure removal of unincorporated dNTPs.

The second strand of cDNA was created by mixing 25 μL of the previously purified single-stranded cDNA, 13 μL of water, 5 μL of NEBuffer 2, 2 μL of 25 mM dNTPs (with dUTP instead of dTTP), 1 μL of random hexamers, and 4 μL of Klenow exo- (5 U/μL, NEB). The reaction was incubated at 37°C for 30 min. The second-strand synthesis product was purified using the MinElute PCR

Purification kit (Qiagen) per the manufacturer's instructions and eluted in 15 μL of EB.

To add sequencing adapters to the double-stranded cDNA, 15 μL of cDNA product, 4 μL of TA buffer (33 mM Tris-acetate at pH 7.5, 66 mM potassium acetate, 10 mM magnesium acetate, and 0.5 mM DTT), and 1 μL of directional Tn–RNA-seq Enzyme mix (Epicentre, beta test material) were mixed together on ice. They were gently vortexed and incubated at 55°C for 5 min. A 30-μL aliquot of QG buffer (Qiagen) was added immediately after the reaction finished. The reaction was cleaned up with 90 μL of AMPure Beads according to the manufacturer's instructions (Beckman Coulter) and eluted in 11 μL of EB. The 11 μL of purified DNA was mixed with 4 μL of Replacement Oligo (Epicentre, beta test material) and 4 μl of Fill-in Reaction Buffer (Epicentre, beta test material) and incubated at 45°C for 1 min, then 37°C for 30 min. A 1-μL aliquot of Gap Filling Enzyme (Epicentre, beta test material) was added and incubated at 37°C for an additional 30 min. The reaction was cleaned up with 36 μL of AMPure beads according to the manufacturer's instructions and eluted in 26 μL of EB.

To digest the second strand of cDNA, the 26-μL purified DNA was added to 3 μL of T4 DNA Ligase buffer (NEB) and 1 μL of USER enzyme mix (1 U/μL, NEB). The reaction was incubated at 37°C for 30 min and the cDNA was purified with 54 μL of AMPure beads according to the manufacturer's instructions and eluted in 25 μL of EB. To amplify the fragments, 25 μL of USER-treated DNA was added to 1 μL of directional Tn–RNA-seq PCR Primer 1 and 1 μL of directional Tn–RNA-seq Primer 2 (Epicentre, beta test material) and 27 μL of Phusion PCR mix (NEB). The mix was incubated at 95°C for 2 min, then 18 cycles of 94°C for 10 sec, 60°C for 30 sec, and 72°C for 3 min were performed. The PCR amplicons were purified with 97 μL of AMPure beads according to the manufacturer's instructions and eluted in 32 μL of EB. Libraries were sequenced on an Illumina GAIIx at a concentration of 6 pM.

Standard directional RNA-seq library construction

Poly(A)-selected Universal Human Reference RNA (Agilent) was heated for 8 min at 94°C in 1x fragmentation buffer (40 mM Tris-Acetate at pH 8.1, 100 mM KOAc, 30 mM MgOAc) and purified using an RNeasy MinElute Kit (Qiagen). The purified and fragmented RNA was treated with 1 μL of Antarctic Phosphatase (5U/μL; NEB) for 30 min at 37°C, and then heat killed for 5 min at 65°C. The samples were then incubated with 2 μL of T4 Polynucleotide Kinase (2U/μL; Epicentre) and 0.7 mM ATP for 30 min at 37°C, followed by purification using an RNeasy MinElute kit (Qiagen). The end-repaired RNA was then ligated and amplified using the Illumina TruSeq small RNA kit. This involves sequential ligation of 3′ and 5′ adapters, followed by reverse transcription and PCR to amplify the completed libraries. Material of the expected size was purified from free adapter product using AMPure Beads.

Data analysis

Every sequence library was trimmed to only analyze the first 36 bases in order to make the results comparable. To assess library quality, the sequence reads from each library were aligned to the GRCh37/hg19 build of the human genome using bowtie (Langmead et al. 2009) with the –m 1 option, to guarantee unique mapping. Complexity was measured for each library by taking a random set of 1 million aligned reads and determining how many different alignment start positions (including strand information) were represented. The number of different alignment start positions was then divided by 1 million to calculate complexity. The percentage of aligned reads that map to RefSeq transcripts was determined by comparing bowtie alignments against the human genome to RefSeq transcript coordinates. The strand specificity of directional Tn–RNA-seq libraries was calculated by determining what percentage of reads that aligned to RefSeq transcripts map to the expected strand in the RefSeq annotation. Note that based on the library construction strategy, transcripts originating from the positive strand should generate sequencing reads that map to the negative strand with the directional Tn–RNA-seq method. In the standard directional approach, transcripts originating from the positive strand should generate reads that map to the positive strand.

To calculate expression levels in each library, sequence reads were aligned to a sequence database of all spliced RefSeq transcripts using Bowtie (Langmead et al. 2009) with the following parameters: -n 2 -k 1 -m 10, which allow reads to align to multiple transcripts in order to capture different isoforms. The number of reads aligning to each transcript was multiplied by 1 million, then divided by the length of the transcript in kilobases times the total number of aligned reads to calculate RPKM values. All correlation analysis was performed in R. All Pearson correlations were measured between log base 2 of RPKM values. Coverage across transcripts was calculated by counting the number of reads that align at each position in a RefSeq transcript and dividing the position by the number of base pairs of the full-length transcript.

Data access

The sequencing data and expression measurements from this study have been submitted to the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) under accession no. GSE32307.

Competing interest statement

B.J.B., I.Y.G., R.V., and S.K. are employees of Epicentre (An Illumina Company).

Acknowledgments

We thank Mark Maffitt, Barbara Wold and members of her lab, as well as members of the Myers lab for valuable discussions and contributions. Portions of this work were funded by USAMRMC/TATRC Contract W81XWH-10-1-0790 (to R.M.M.) and by NHGRI ENCODE Grant 5U54HG004576 (to R.M.M.).

References

Adey A, Morrison HG, Asan , Xun X, Kitzman JO, Turner EH, Stackhouse B, MacKenzie AP, Caruccio NC, Zhang X, et al. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* **11:** R119. doi: 10.1186/gb-2010-11-12-r119.

Core LJ, Waterfall JJ, Lis JT. 2008. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322:** 1845–1848.

Greagg MA, Fogg MJ, Panayotou G, Evans SJ, Connolly BA, Pearl LH. 1999. A read-ahead function in archaeal DNA polymerases detects promutagenic template-strand uracil. *Proc Natl Acad Sci* **96:** 9045–9050.

Hawkins TL, O'Connor-Morin T, Roy A, Santillan C. 1994. DNA purification and isolation using a solid-phase. *Nucleic Acids Res* **22:** 4543–4544.

He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. 2008. The antisense transcriptomes of human cells. *Science* **322:** 1855–1857.

Jiang H, Wong WH. 2009. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* **25:** 1026–1032.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.

Lauck M, Hyeroba D, Tumukunde A, Weny G, Lank SM, Chapman CA, O'Connor DH, Friedrich TC, Goldberg TL. 2011. Novel, divergent simian hemorrhagic Fever viruses in a wild ugandan red colobus

monkey discovered using direct pyrosequencing. *PLoS ONE* **6:** e19056. doi: 10.1371/journal.pone.0019056.
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.
Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell* **133:** 523–536.
Liu Y, Han D, Han Y, Yan Z, Xie B, Li J, Qiao N, Hu H, Khaitovich P, Gao Y, et al. 2011. Ab initio identification of transcription start sites in the Rhesus macaque genome by histone modification and RNA-Seq. *Nucleic Acids Res* **39:** 1408–1418.
Mo B, Vendrov AE, Palomino WA, DuPont BR, Apparao KB, Lessey BA. 2006. ECC-1 cells: a well-differentiated steroid-responsive endometrial cell line with characteristics of luminal epithelium. *Biol Reprod* **75:** 387–394.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621–628.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320:** 1344–1349.
Novoradovskaya N, Whitfield ML, Basehore LS, Novoradovsky A, Pesich R, Usary J, Karaca M, Wong WK, Aprelikova O, Fero M, et al. 2004. Universal Reference RNA as a standard for microarray experiments. *BMC Genomics* **5:** 20. doi: 10.1186/1471-2164-5-20.
Parkhomchuk D, Borodina T, Amstislavskiy 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. doi: 10.1093/nar/gkp596.
Perkins TT, Kingsley RA, Fookes MC, Gardner PP, James KD, Yu L, Assefa SA, He M, Croucher NJ, Pickard DJ, et al. 2009. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella typhi*. *PLoS Genet* **5:** e1000569. doi: 10.1371/journal.pgen.1000569.
Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK. 2010. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464:** 768–772.
Rand KN, Ho T, Qu W, Mitchell SM, White R, Clark SJ, Molloy PL. 2005. Headloop suppression PCR and its application to selective amplification of methylated DNA sequences. *Nucleic Acids Res* **33:** e127. doi: 10.1093/nar/gni120.
Rosenberg BR, Dewell S, Papavasiliou FN. 2011. Identifying mRNA editing deaminase targets by RNA-Seq. *Methods Mol Biol* **718:** 103–119.
Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA. 2008. Divergent transcription from active promoters. *Science* **322:** 1849–1851.
Smith AM, Heisler LE, St Onge RP, Farias-Hesson E, Wallace IM, Bodeau J, Harris AN, Perry KM, Giaever G, Pourmand N, et al. 2010. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res* **38:** e142. doi: 10.1093/nar/gkg368.
Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456:** 470–476.

Received June 21, 2011; accepted in revised form October 17, 2011.