Clontech Laboratories, Inc.
A Takara Bio Company

Strand-Specific Transcriptome Sequencing for Challenging Samples

Craig Betts, Magnolia Bostick, Tommy Duong, & Andrew Farmer[1]

1 Corresponding Author

Clontech Laboratories, Inc., 1290 Terra Bella Ave., Mountain View, CA 94043

Abstract

Next Generation Sequencing (NGS) has empowered a deeper understanding of biology by enabling RNA expression analysis over the entire transcriptome with high sensitivity and dynamic range. A powerful application within this field is stranded RNA-seq, which is necessary to distinguish closely-related genes and non-coding RNAs (e.g. lincRNA) or to define genes in poorly annotated, coding-rich genomes, including many bacteria.

Commonly-used methods to generate strand-specific RNA-seq libraries are plagued by protocols that require several rounds of enzymatic treatments and cleanup steps, making them time intensive and insensitive, and making it challenging to process several samples simultaneously. Here we present a novel, single-tube method, based on Clontech's patented SMART™ technology, which is able to generate strand-specific RNA-seq libraries from minute samples in under four hours. This approach eliminates the multitude of labor-intensive enzymatic steps required by other stranded RNA-seq methods, while maintaining the sensitivity and reproducibility characteristic of SMART. We have successfully tested our technology with input levels from 100 pg to 100 ng of poly(A)-selected RNA, as well as with ribosomally-depleted RNA from FFPE samples, with outstanding reproducibility within and across input levels. Spike in of ERCC controls showed linear detection over six orders of magnitude and strand specificity of over 99%.

The increased sensitivity achieved with SMART requires a very sensitive rRNA removal method (most methods require microgram amounts of total RNA). With this in mind, we have developed a method of rRNA depletion which effectively removes *28S, 18S, 5.8S, 5S,* and *12S* transcripts from mammalian samples down to 10 ng. The remaining, rRNA-depleted RNA can easily be used in downstream sequencing applications with fewer than 5% of reads mapping back to rRNA.

With these tools, researchers can more confidently apply NGS to challenging samples.

Introduction

NGS' short reads can make it difficult to determine which gene or noncoding sequence a particular read comes from. In the SMARTer® Stranded RNA-Seq Kit, we use a directional template-switching reaction to preserve the strand orientation of the RNA and obtain strand-specific sequencing data from the synthesized cDNA. The SMARTer Stranded RNA-Seq Kit generates indexed cDNA libraries that are suitable for RNA-seq on any Illumina® platform. The protocol has been designed for ease of use and direct addition of adapters, and can be completed in less than 4 hr. Importantly, the SMARTer Stranded RNA-Seq Kit produces whole-transcriptome coverage without 5′ or 3′ biases, and yields excellent reproducibility and sensitivity, mappability, and ERCC and MAQC correlation.

Conclusions

The SMARTer Stranded RNA-Seq Kit provides a simple and efficient solution for generating indexed cDNA libraries suitable for NGS on any Illumina platform in less than 4 hr, starting from as little as 100 pg of poly(A)-purified or rRNA-depleted RNA.

- **Robust performance and wide dynamic range:** Single-tube protocol creates sequencing-ready libraries from low-input samples of poly(A)-purified or rRNA-depleted RNA (Figure 1). Data is highly reproducible across a wide range, extending to as little as 100 pg of input RNA (Figure 2).
- **Highly accurate and reproducible results:** All 92 ERCC spike-in control transcripts were detected with expression levels consistent with the quantity spiked in. Results were highly reproducible across replicates and over a thousand fold range of input RNA levels (Figures 2 and 3).
- **Ability to distinguish overlapping and antisense transcripts:** Sequencing reads are assigned to the correct gene in the case of overlapping and antisense transcripts (Figure 4).
- **rRNA depletion from small samples:** RiboGone™ - Mammalian treatment removes rRNA from intact and degraded total RNA, and retains noncoding transcripts for analysis (Figure 5).
- **Highly correlated with other methods of measuring expression:** Differential expression data obtained with the SMARTer Stranded RNA-Seq Kit is highly correlated with MAQC qPCR data (R=0.860; Figure 6).

References

1. Jiang, L. *et al.* (2011) *Genome Res.* **21**(9):1543–1551.
2. MAQC Consortium (2006) *Nat. Biotechnol.* **24**(98):1151–1161.
3. Chenchik, A. *et al.* (1998) In RT-PCR Methods for Gene Cloning and Analysis. Eds. Siebert, P. & Larrick, J. (BioTechniques Books, MA), pp. 305–319.
4. Hansen, T. B. *et al.* (2011) *EMBO J.* **30**(21):4414–4422.

Abbreviations

ERCC—External RNA Controls Consortium (1)

FPKM—Fragments per kilobase of transcript per million mapped reads

MAQC—MicroArray Quality Control project (2)

RPKM—Reads per kilobase of exon per million reads

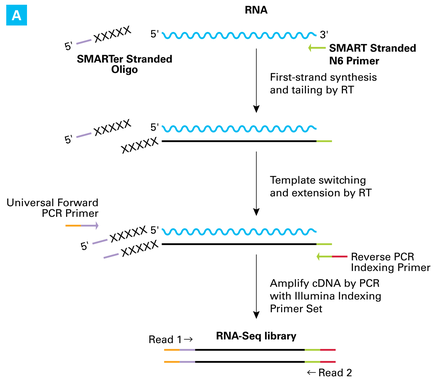
rRNA—Ribosomal RNA

RT—Reverse transcriptase

SMART—**S**witching **M**echanism **a**t 5′ End of **R**NA **T**emplate

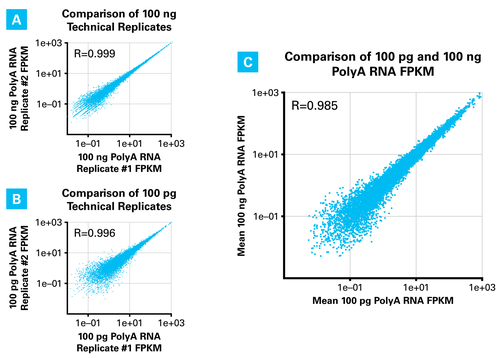
633685

1 SMARTer Stranded RNA-Seq Library Generation



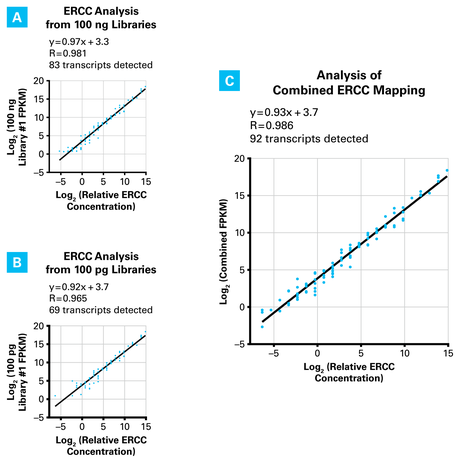
The SMARTer Stranded RNA-Seq Kit produces high-yield, high-quality cDNA irrespective of input RNA concentrations. Panel A. Flowchart of SMARTer Stranded RNA-Seq library generation. The SMARTer Stranded RNA-Seq Kit utilizes the proprietary SMART Stranded N6 Primer and SMARTScribe™ Reverse Transcriptase to perform first strand synthesis and tailing. When SMARTScribe RT reaches the 5′ end of the RNA fragment, its terminal transferase activity adds a few additional nucleotides to the 3′ end of the cDNA. The SMARTer Stranded Oligo base-pairs with this non-templated nucleotide stretch, creating an extended template to enable SMARTScribe RT to continue replicating to the end of the oligonucleotide (3). **Panel B.** cDNA libraries produced with the SMARTer Stranded RNA-Seq Kit. Human Brain Poly A$^+$ RNA (Clontech) was spiked with ERCC control RNA and serially diluted to prepare RNA samples containing between 100 ng–100 pg Human Brain Poly A$^+$ RNA. cDNA libraries were successfully prepared in triplicate according to the SMARTer Stranded protocol with SeqAmp™ DNA Polymerase and twelve different Illumina indices, and visualized as an Agilent 2100 Bioanalyzer gel-like image using a High Sensitivity DNA Chip. Libraries had comparable yields and purity irrespective of input RNA concentrations. Libraries were sequenced on an Illumina HiSeq® 2000 instrument, with ~300M 2 x 100 bp paired end reads.

2 Robust Performance and Wide Dynamic Range



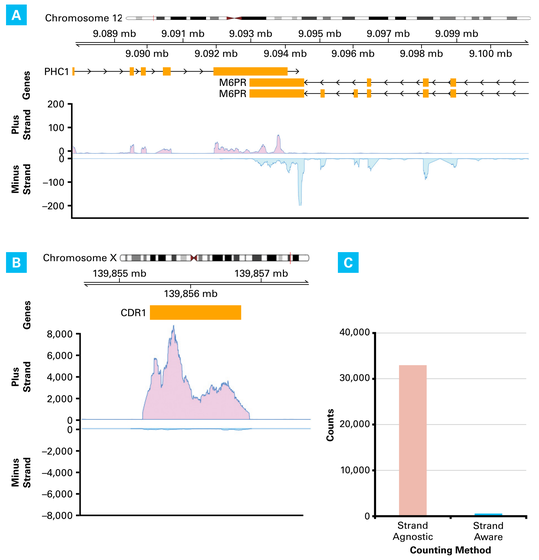
Obtain highly reproducible and sensitive directional RNA-seq data across a wide range of input RNA with the SMARTer Stranded RNA-Seq Kit. Scatter plots of expression (FPKM) comparing pairs of cDNA library replicas created from 100 ng or 100 pg of input RNA (**Panels A** and **B** respectively) show high reproducibility across a wide range of input levels. **Panel C.** cDNA libraries prepared from 100 ng and 100 pg of human brain polyA$^+$ RNA show a high correlation, suggesting consistency across input levels. Axes are plotted on a $\log_{10}$ scale. Insets indicate the Pearson coefficient of correlation between replicates (R).

3 Highly Reproducible Results



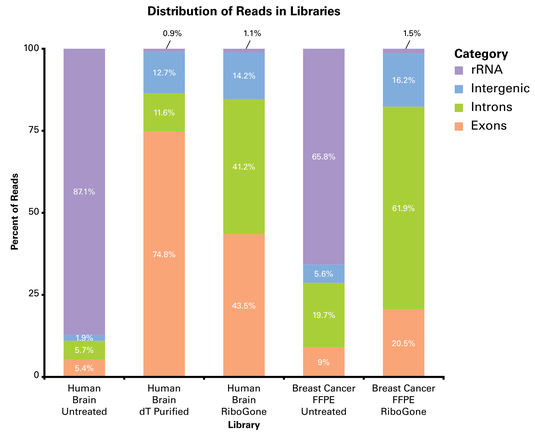
High reproducibility confirmed by ERCC analysis. FPKMs were plotted against the relative concentrations of the ERCC spike-in control RNAs. **Panel A.** Reads mapping to the ERCC data set from one 100 ng input library. **Panel B.** Reads mapping to the ERCC data set from one 100 pg library. **Panel C.** Reads mapping to the ERCC data set from all libraries combined. Similar results are obtained from individual libraries, regardless of input amount (Panels A and B), and the complete set of 92 ERCC transcripts was detected linearly over the entire range of concentrations in the pooled mapping (Panel C). Axes are plotted on a $\log_2$ scale. Slope, Pearson coefficient of correlation (R), and number of transcripts detected are indicated above each graph.

4 Strand Specificity Distinguishes Overlapping and Antisense Transcripts



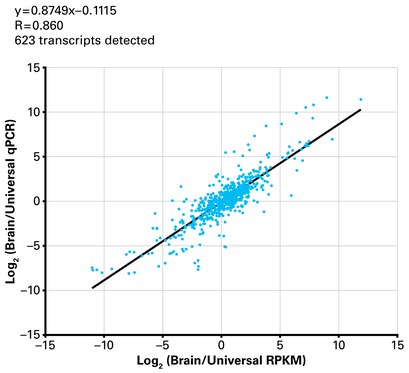
Distinguishing overlapping and antisense transcripts with the SMARTer Stranded RNA-Seq Kit. Panel A. RNA-seq reads from the Human Brain Poly A$^+$ RNA cDNA library were mapped against the human genome. The SMARTer stranded method allowed assignment of sequencing reads to the correct gene in the case of overlapping *PHC1* and *M6PR* transcripts. **Panel B.** Strand-specific coverage of the *CDR1* locus. Nearly all reads are antisense to the annotated transcript, a finding independently reported elsewhere (4). **Panel C.** Comparison of *CDR1* gene counts obtained using either a strand-agnostic or strand-aware method.

5 Efficient rRNA Removal and Non-coding Transcript Analysis from Intact and Degraded RNA



RiboGone treatment removes rRNA efficiently from intact and degraded RNA, while retaining noncoding transcripts for analysis. RNA-Seq libraries were generated from Human Brain Total RNA (Clontech) or Breast Cancer FFPE RNA (Cureline, extracted using a NucleoSpin totalRNA FFPE kit) and treated with the indicated rRNA removal method. Reads were mapped to the hg19 genome and read distributions were determined using Picard RNA-Seq Metrics. Libraries generated from RiboGone-treated RNA had comparably low rRNA reads to oligo(dT) enriched RNA (Takara), while retaining more noncoding reads.

6 RNA-Seq and qPCR Data are Highly Correlated



High correlation between SMARTer RNA-seq data and qPCR data from the MAQC project. Differential expression data was obtained for Human Brain Reference RNA (Ambion) and Human Universal Reference RNA (Agilent) using the SMARTer Stranded kit (after RiboGone rRNA depletion) and compared with qPCR data for the same RNAs obtained through the MAQC study (2). A scatter plot was used to compare the differential expression data. The slope (0.875) and correlation (0.860) for the comparison line of expression ratio (in RPKM) and qPCR ratio (in C_t) are plotted for Human Brain and Human Universal Reference RNAs, on a $\log_2$ scale. The transcripts used in this analysis were the 623 of ~900 transcripts present in the MAQC data set that were also detected in both the Human Brain and Human Universal RNA-seq data sets.