Whole Transcriptome Profiling Using the SOLiD™ 3 System

Introduction

Sequence-based approaches to the study of gene expression have the advantage of querying known as well as previously unknown RNAs in a sample, also termed "hypothesis-neutral" discovery. cDNA copies of all the RNA present in the sample are made, sequenced, and mapped to a reference genome. Finally the structure is deduced using bioinformatic tools. Here we describe the Whole Transcriptome Library Protocol, which allows the rapid construction of strand-specific libraries from a wide range of RNA species. This easy-to-use, sensitive method uses existing commercial products to clone RNA fragments and sequence the resulting cDNAs using the SOLiD™ 3 System (Figure 1). Applied Biosystems open-source and freely available analysis tools facilitate mapping sequences to a reference, counting each short read mapped to a given site, and identifying exon junctions.

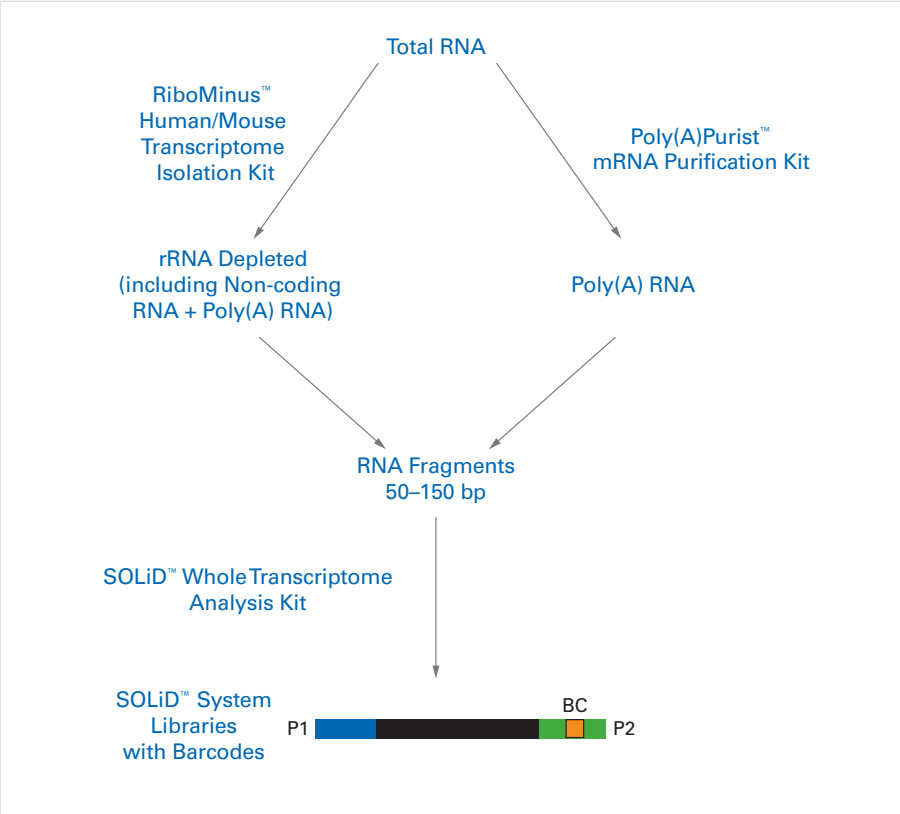

Figure 1. Whole Transcriptome Analysis Workflow Using the SOLiD™ 3 System. Different RNA libraries can be prepared for sequencing on the SOLiD™ 3 System following variations to a common, robust workflow. This enables a wide variety of RNA molecules to be cloned and sequenced for further study by making slight modifications to the basic technique.

Experimental Considerations in Transcriptome Analysis

Successful whole transcriptome analysis depends on RNA quality and efficient conversion to cDNA libraries, as this will dictate what sequences are generated. There are currently two approaches for constructing whole transcriptome libraries. One approach begins with RNA that has been enriched for poly(A) RNA. The other approach starts with total RNA that contains all the different species of RNA molecules found in the cell—poly(A) RNA, noncoding RNA (ncRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), etc. Nontarget RNAs, like structural rRNA and tRNA, represent over 90% of total cellular RNA. Depleting nontarget RNA species allows enrichment for the RNAs of interest. Using rRNA-depleted RNA for whole transcriptome research allows the study of both ncRNAs and coding RNAs (poly(A) RNA). Recent studies have shown an abundance of ncRNAs in the cell, and suggest the role they may have in the control of gene expression [1].

Materials and Methods

RNA Isolation

Five micrograms of total RNA from the Human Brain Reference (HBR) RNA (Ambion, P/N AM6050) was used as starting material for the whole transcriptome library construction discussed here. If the library is to be made from poly(A) RNA, Poly(A)Purist™

Kits (Ambion, P/N AM1916 or AM1922) have been shown to give high-quality results. Ribosomal RNA removal can be accomplished using the RiboMinus™ Eukaryote Kit for RNA-Seq (Invitrogen, P/N A10837-08, or A10838-08 for plants). Approximately 1 µg of poly(A) RNA or 0.5 µg of rRNA-depleted RNA is needed for making libraries.

Whole Transcriptome RNA Library Preparation

The RNA is randomly fragmented using RNase III (Ambion, P/N AM2290), and 100–200 bp fragments are isolated after gel electrophoresis. The RNA fragments are then converted to cDNA libraries in a strand-specific manner using the Whole Transcriptome Library Protocol (http://solid.appliedbiosystems.com). DNA "barcodes" can be incorporated into the libraries to allow pooling of multiple samples on a single sequencing run if desired.

SOLiD™ Sequencing

The cDNA libraries are clonally amplified onto beads by emulsion PCR using standard protocols from the SOLiD™ System User Manual (http://solid.appliedbiosystems.com). These beads are enriched and deposited onto the surface of a glass slide for sequencing. Current scientific publications estimate that 40–50 million mappable RNA sequences are needed to detect the maximum number of known transcripts from a library constructed from poly(A) RNA. This number should be considered as the minimum number of sequences needed for a whole transcriptome experiment. The SOLiD™ 3 System is capable of sequencing more than 400 million individual reads in a single run. Thirty-five to fifty bases of sequence are generated from the cloned fragments using the SOLiD™ 3 System and are used for subsequent analysis.

Analysis

The sequences generated by the SOLiD™ 3 System can be analyzed using a number of tools. Applied Biosystems has developed the Applied Biosystems Whole Transcriptome Analysis Pipeline (http://solidsoftwaretools.com/gf/project/transcriptome/), which allows basic analysis such as mapping sequences to a reference, counting the number of sequences mapped to known RNAs, and identifying both known and novel exon junctions. The data output from this pipeline is readily imported to the UC Santa Cruz genome browser (http://genome.ucsc.edu/) for visualization of the results. This tool has been designed to allow additional downstream analysis scripts to be developed for further investigation.

Results

Reproducibility and Dynamic Range

The technical reproducibility of the system was measured by comparing the sequencing results from two independent runs of HBR RNA (Figure 2). As can be seen, a Spearman Rank correlation of >0.98 is obtained and more than 22,000 transcripts are detected. Additionally, the levels of 95% of the transcripts do not vary in the two samples by more than 2.3-fold. Figure 2 shows that the entire system, from library generation to sequencing to data analysis, is highly reproducible. This high reproducibility is critical for comparing different samples and will allow fewer technical replicates to be run. A wide dynamic range is very desirable for any gene expression system. Figure 2 also displays the dynamic range of known transcripts detected in these samples, and in this case it is measured to be between 10^5 and 10^6.

Reliable Mapping of Sequences to the Genome and to Known RefSeqs

To assess the ability of the SOLiD™ System to detect known transcripts, a large number of clones was sequenced

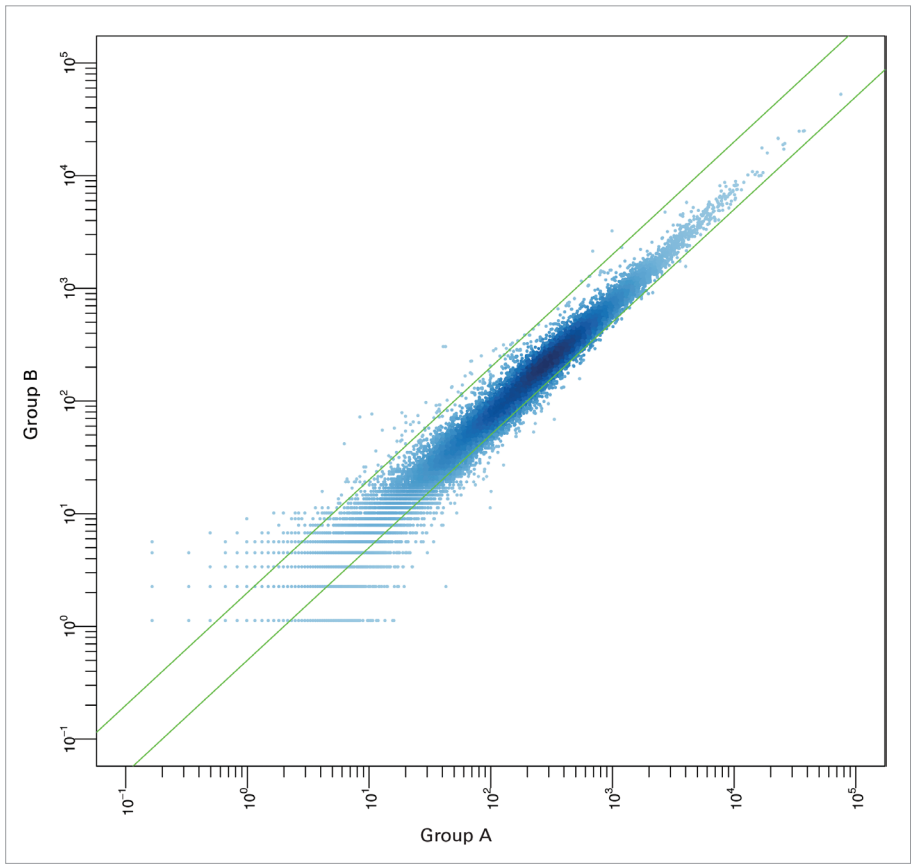

Figure 2. Poly(A)+ Transcript Level Count Reproducibility Scatterplot. The number of known transcripts identified from HBR RNA was obtained from a library that was sequenced independently by two different groups. Transcripts were counted as present calls if one or more reads mapped in both data sets. The dynamic range of known transcripts detected in these samples is between 10^5 and 10^6. Green lines indicate 2-fold change, and the blue markers darken with higher density.

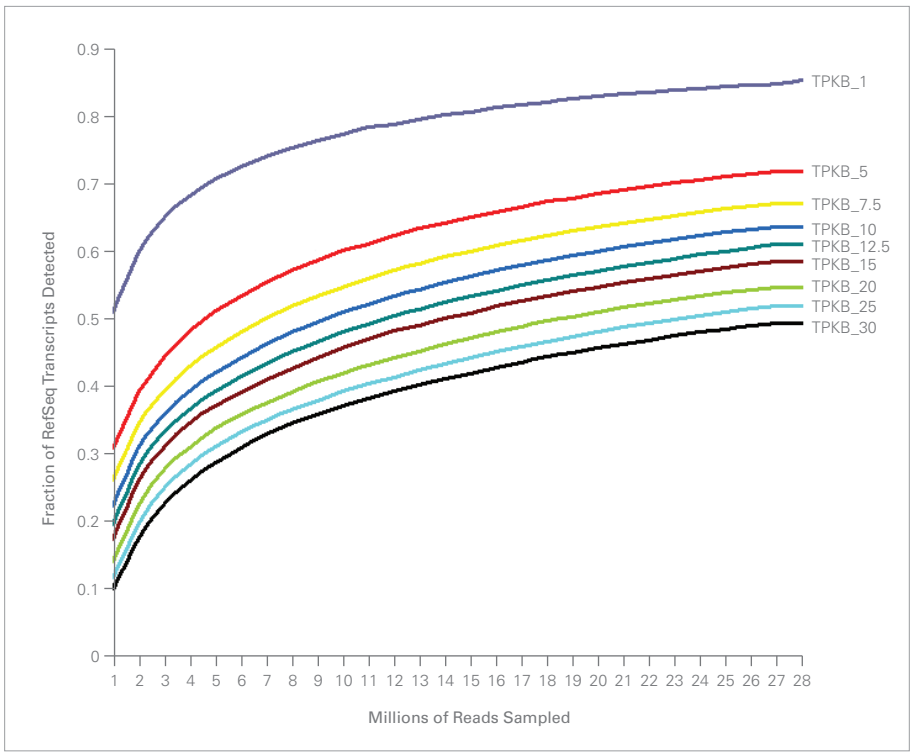

Figure 3. Fraction of Known RefSeqs Detected in a Library Prepared from the MAQC UHR RNA Sample. The fraction of known RefSeqs begins to plateau as 30 million mappable tags are detected, when 40–45 million mappable sequences are used for analysis. Colored lines indicate thresholds required to be counted as a present call where TPKB represents tags per KB of transcript length. It is observed in the graph that as the number of sequence tags mapped per kilobase of RefSeq increases, a smaller fraction of total RefSeqs are detected.

Detection of Novel Exons and Alternate Transcripts

Another important feature for whole transcriptome analysis is the ability to analyze all currently annotated exons, as well as novel exons arising from previously unrecognized splice sequences [3]. To achieve this, it is necessary to have uniform coverage across all transcripts, and sufficient coverage to have confidence that the exon has been correctly identified. The large number of unique tags generated by the SOLiD™ 3 System—greater than 400 million per run—and the Whole Transcriptome Analysis Pipeline tool provide uniform coverage across the transcriptome and high confidence mapping to allow detection of novel exons (Figure 4) and discrimination between alternative transcripts (Figure 5).

Genomic DNA Strand Specificity

Recent publications using high-throughput sequencing have shown that as many as 6000 known transcripts are also synthesized from the "antisense"

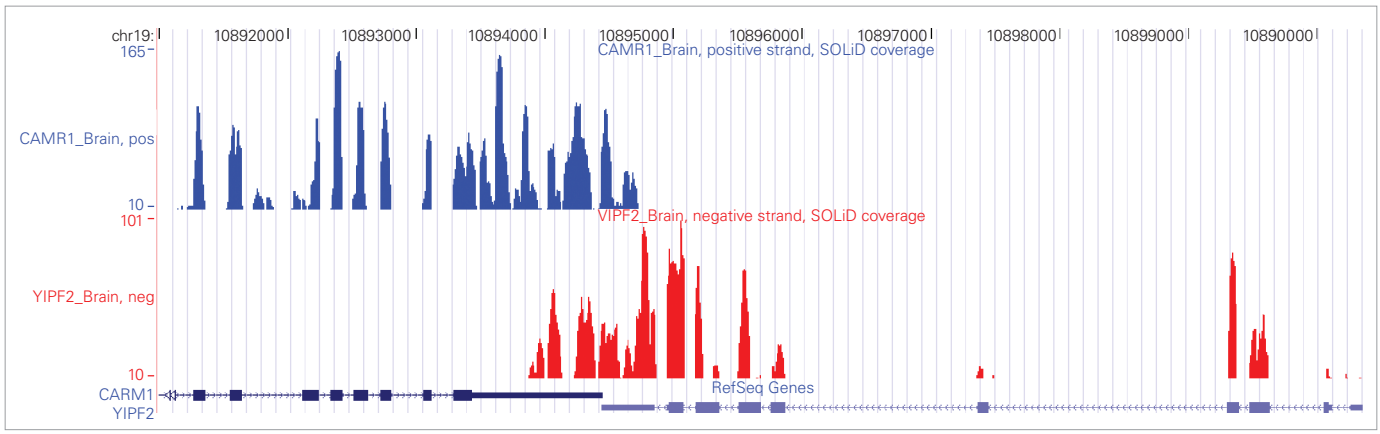

Figure 4. Strand-specific Read Distribution. UCSC genome browser showing sequences generated from HBR RNA mapping to specific strands (positive and negative) in a region of the human genome where the 3′ ends of two genes overlap. If individual RNA molecules are mapped uniquely to the strand of the DNA from which it was synthesized, the strandedness of the RNA is said to be preserved. It is clear that this is the case with the sequences generated from the two genes. The data also suggest that the current annotation for these genes is not complete, as there are sequences mapping beyond the current exons.

and mapped back to the RefSeq database. The number of known transcripts detected was graphed as a function of the number of sequence reads required in Figure 3. The number of sequence tags per kilobase of transcript length (TPKB) required to call a RefSeq transcript "present" dictates the fraction of RefSeqs that will be detected at any given number of sequences mapped [2] (Figure 3). The higher the TPKB threshold to detect a transcript, the higher the confidence of accurately measuring the amount of that transcript present in the sample.

strand of the same DNA region [4]. This work suggests antisense transcription is not the exception, but is common in humans and most likely other higher organisms. Additionally, other types of ncRNAs have been shown to represent a significant fraction of the genome. These transcripts are synthesized from

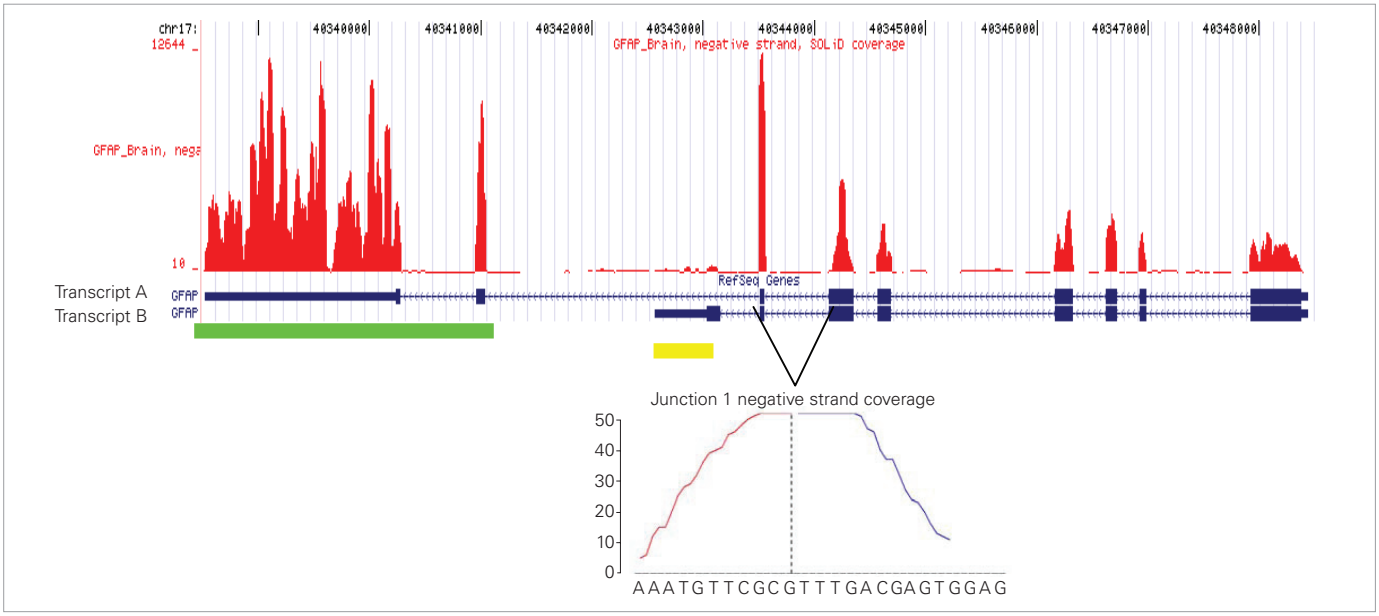

Figure 5. Exon-specific Coverage. UCSC genome browser showing the sequences generated from HBR RNA mapping to a region of the human genome where there are two different transcripts (A and B). The presences of sequences in the green region and absence of sequences in the yellow region identify that only transcript A is detected in this RNA sample. The blow up shows the 50 bp sequences that specifically map across the junction of two exons.

both strands of DNA. Because of the large number of antisense transcripts present, and the need to accurately map ncRNAs, it is important to know which DNA strand each RNA transcript maps to [4,5]. The whole transcriptome system used with the SOLiD™ System preserves the "strandedness" of the RNA by specific ligation of adapters to either the 5' or 3' ends of the RNA molecules before conversion to double-stranded cDNA (Figure 4). This differs from other methods in which the adapters are ligated to the double-stranded cDNA molecules, or methods that utilize random cDNA priming; with such methods it is not possible to know which strand of the DNA the sequences are mapping to, and therefore it is not possible to know which transcript the sequences were derived from [2].

Conclusion

High-throughput sequencing allows scientists to study the complexity of the RNA synthesized in complex genomes. The massively parallel short-read sequencing technology achieved by the SOLiD™ 3 System is ideally suited for whole transcriptome analysis. The addition of the Applied Biosystems Whole Transcriptome Analysis Kit enables the detection of known and novel RNA molecules as well as the resolution of strand specificity. The Whole Transcriptome Analysis Pipeline allows the data generated to be mapped and viewed easily. This complete system provides a powerful solution for the study of complex transcriptomes.

References

1. Huang A (2008) 'Noncoding' RNA Shown to Silence Cancer Suppressor Gene. *The Gazette, Johns Hopkins University*, Vol. **37** No. 19.
2. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, **5(7)**:621–628.
3. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, **40**:1413–1415.
4. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW (2008) The antisense transcriptomes of human cells. *Science*, 19; **322**(5909):1855–1857.
5. Cloonan N, Grimmond SM (2008) Transcriptome content and dynamics at single-nucleotide resolution. *Genome Biology*, **9**(9):234.

Printed in the USA. 05/2009 Publication 139AP15-01 **B-084000 0509**

Applied Biosystems

Headquarters
850 Lincoln Centre Drive | Foster City, CA 94404 USA
Phone 650.638.5800 | Toll Free 800.345.5224
www.appliedbiosystems.com

International Sales
For our office locations please call the division headquarters or refer to our Web site at
www.appliedbiosystems.com/about/offices.cfm