Whole Transcriptome Analysis of Total Human RNAs by Massively Parallel Sequencing on the SOLiD™ System

Poster Number P11.119

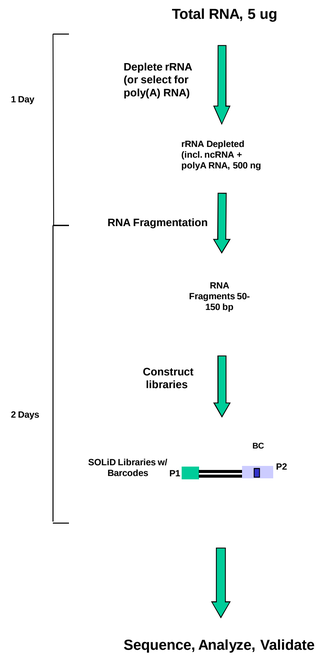
Bob Nutter[1], Diane Ilsely[2], Scott Kuersten[2], Joel Brockman[2], Jeff Schagemen[2], Catalin Barbarciou[1], Brian Tuch[1], Kelli Bramlett[2], Jian Gu[2], Helen Chen[2], Sheila Heater[2], Tom Bittick[2], Bob Setterquist[2] and Asim Siddiqui[1] Applied Biosystems 1850 Lincoln Centre Dr. Foster City, CA and 22130 Woodward St., Austin, TX

ABSTRACT

Detailed analysis of the entire transcriptome of higher organisms is for the first time demonstrating the complexity of the structure of RNA and providing a better understanding of the role different types of RNA play in the control of gene expression. We report here the results obtained using a prototype version of the SOLiD™ Whole Transcriptome Analysis system being developed by Applied Biosystems. Using RNA from HeLa, Human Brain (HBR) and Universal Human Reference (UHR), we constructed libraries from rRNA-depleted and poly A RNA to show the performance of this integrated system in terms of maintaining sample representation, strandedness, and reproducibility. We demonstrate representation by use of a synthetic RNA spike-in mixture, in which six heterologous transcripts were added to total RNA at different concentrations. Sequencing shows the expected dose dependent response curve and uniform coverage across each transcript. We highlight reproducibility of the system from data generated from independent field confirmation sites. Feedback from customers indicates that the protocol was robust and easy to use, while sequencing data across testing sites showed similar detection of unique mappable sequences, presence of RefSeq transcripts, and coverage across genes. Libraries representing the transcriptome were constructed from as little as 0.4 ug of rRNA-depleted RNA without appreciable loss of coverage. The correlation among the sites for RefSeq transcripts was > 0.90. The successful testing of this approach will result in the release of a SOLiD™ Whole Transcriptome Analysis Kit in June 2009.

MATERIALS AND METHODS

Figure 1 Workflow used to construct Whole Transcriptome libraries for sequencing



Whole Transcriptome (WT) Analysis can be performed using RNA from a variety of sources. In the experiments described here, 5-10 ug of total RNA was used as starting material. Total RNA was first fractionated to allow the study of either rRNA-depleted or polyA RNA, depending on the nature of the study. In these studies, typically 1 ug of fractionated RNA was used, although as little as 0.4 ug of RNA was used in some instances. Briefly, RNA was first fragmented by digesting with RNase III, and then size fractionated using electrophoresis. Fragments 50-150 bp were collected and purified. Using the Small RNA Expression Kit (SREK, PN 4397682), cDNA libraries were constructed by ligating the RNA fragments to adaptors in a strand-specific manner and converting them into double stranded cDNA libraries using reverse transcription followed by PCR amplification. The entire process could be completed in 2-3 days, depending on the type of RNA used as starting material. Library fragments were sequenced on a SOLiD™ System Analyzer using standard run conditions. Resulting sequences were analyzed using the freely available Whole Transcriptome Analysis Pipeline (http://solidsoftwaretools.com) from Applied Biosystems. Subsequent analysis was conducted using publicly available software.

RESULTS

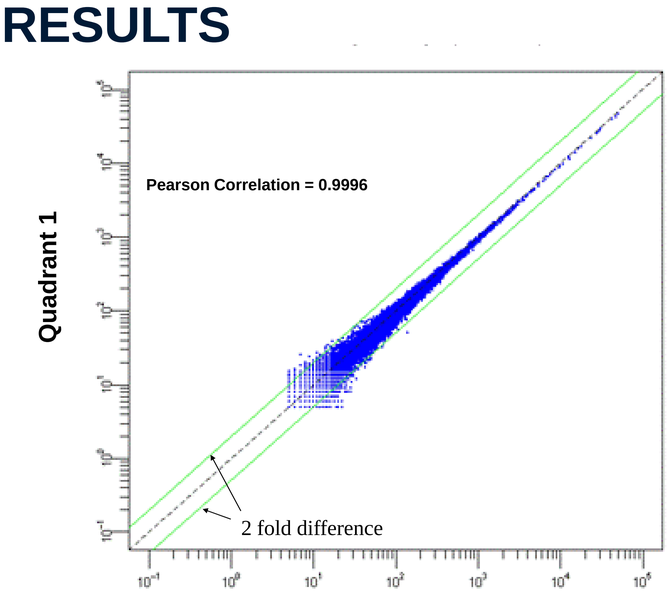


Figure 2: Correlation of SOLID™ Sequencing of WT libraries from two slide quadrants. The X and Y axes represent RefSeq counts from the same library after ePCR and sequenced in two different quadrants on the same slide. The library was constructed from 1 ug of rRNA-depleted HeLa total RNA. The data from each quadrant was analyzed and the number of reads mapping to RefSeq transcripts were determined. The high Pearson correlation shows the reproducibility of the system.

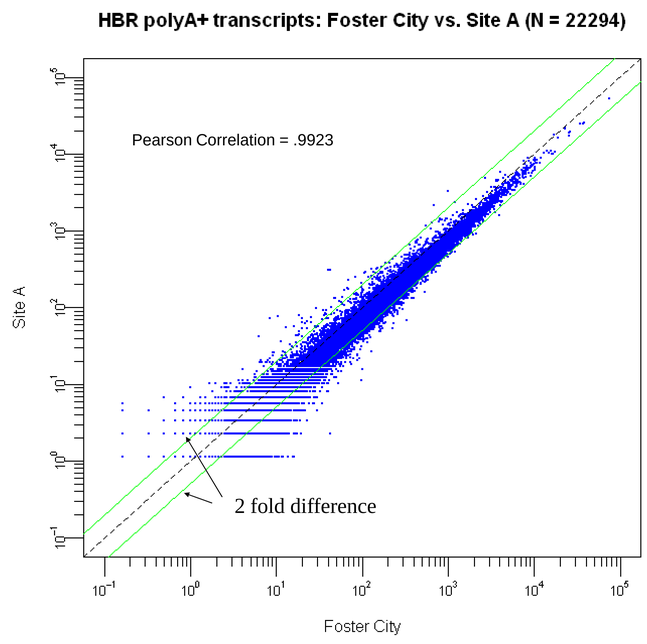


Figure 3: Correlation of cDNA library construction and sequencing between two independent sites. The X-axis shows the RefSeq counts for data obtained in Foster City while the Y-axis shows the RefSeq counts for site A. Each site independently made a library from 1ug of Human Brain Reference polyA RNA (Ambion), followed by SOLiD™ sequencing. The number of reads mapped to known RefSeq transcripts were determined for each library. The data shows high reproducibility between different sites. The average correlation measured among all test sites (n=60) was > 0.8.

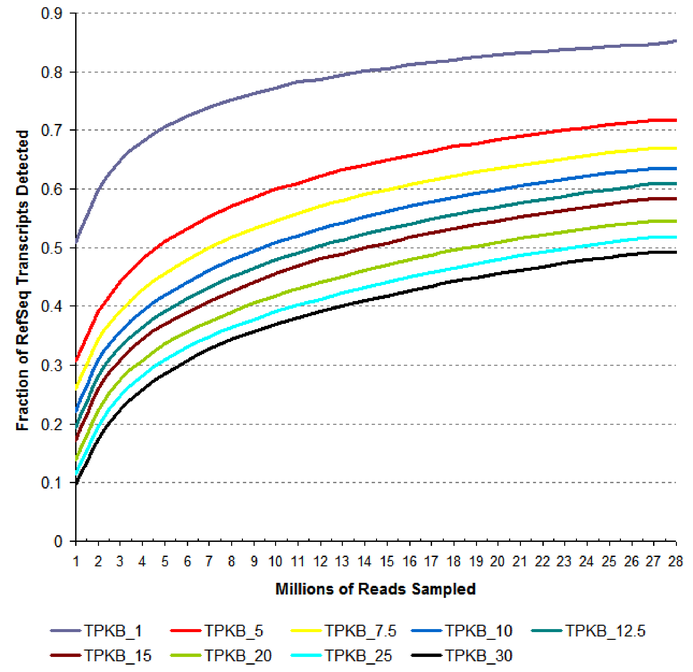


Figure 4: Number of sequences needed to detect RefSeq transcripts. The graph shows the fraction of RefSeq transcripts detected as a function of mapped reads sampled (millions). The colored lines represent various thresholds for the minimum number of sequence tags per kilobase (TPKB) needed for the transcript to be called 'present'. The graph shows a high percentage of RefSeq transcripts are detected with increasing numbers of sequences generated. It is also shown that as more sequences are required to call a RefSeq as 'present', the number of sequences needed to reach saturation also increases. The WT libraries were prepared from 1 ug of rRNA-depleted UHR RNA (Stratagene) as described in the methods.

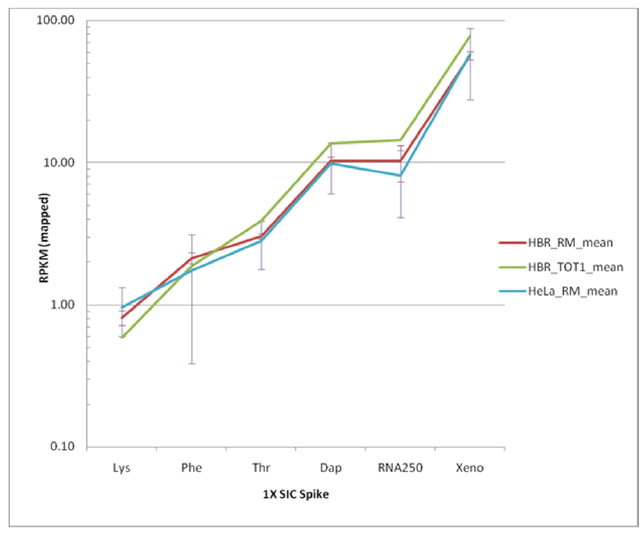


Figure 5: Dose response of curve for RNA spike-in controls. The Affymetrix™ spike-in control transcripts (lys, dap, thr, phe) plus Xeno RNA™ and RNA250 (Ambion) were added to 1ug of total human brain RNA (HBR), rRNA-depleted HeLa RNA, and rRNA-depleted HBR RNA prior to library generation. The resulting libraries were sequenced. The number of sequences mapping to the controls were plotted at each level and were normalized by transcript length and total number of mapped reads (RPKM). The data shows the expected dose response as a function of increasing transcript amount. The relative ratios are Lys (1X), Phe (2X), Thr (4X), Dap and RNA 250 (15X) and Xeno (21X). The data represents triplicate measurements

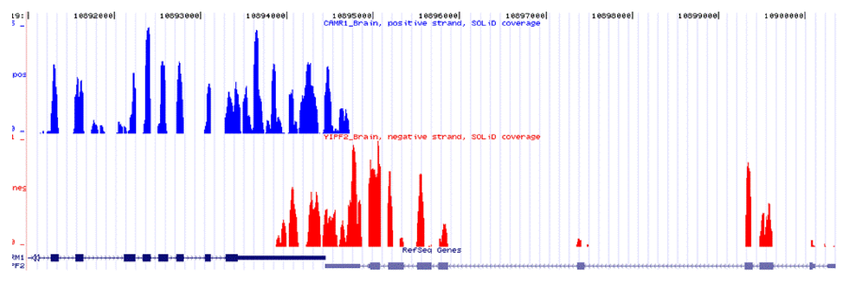


Figure 6: Strand-Specific mapping to the reference genome. Data generated from sequencing a WT library from HBR RNA (Ambion) with the system was visualized using the UC Santa Cruz genome browser. The figure shows the two genes (CARM1, blue; YIPF2, red) are both encoded from the same region of the genome, but on different strands. The figure shows the RNA fragments corresponding to each gene are unambiguously mapped to the correct strand of the DNA. The figure also suggests both genes have exons extending past what is contained in the current annotation. Being able to map RNA to the genome in a strand-specific manner will allow scientists to discover novel exons and develop better annotation of complex genomes.

CONCLUSIONS

- A WT Protocol is available using commercial products that permits rapid, strand-specific conversion of RNA to cDNA for high throughput sequencing.
- The technical reproducibility of the system is extremely high.
- Testing of the system at multiple sites shows good reproducibility in the detection of known RefSeq transcripts.
- The ability to detect known RefSeq transcripts is directly correlated to the number of sequences generated and the threshold set for calling a transcript as 'present'.
- Spike-in controls are detected in a dose-dependant manner and can be used to normalize data and quantitate differences between samples.
- A fully supported WT kit based on this work will be commercially available in June 2009.

TRADEMARKS/LICENSING