A Comparison of Next Generation Sequencing and Microarrays for Whole Transcriptome Expression Profiling

Penn Whitley, Andrew Lemire, Joel Brockman, Sheila Heater, Jeff Schageman, Jian Gu, Kristi Lea, Luming Qu, Charmaine San Jose, Natalie Hernandez
Kelli Bramlett, Diane Ilsley and Robert Setterquist, Life Technologies/Ambion R&D, 2130 Woodward, Austin, TX, USA, 78744

ABSTRACT

Microarray based expression profiling has been remarkably successful at elucidating the spatio-temporal patterns of mRNA transcripts within cells and tissues, however there are a number of shortcomings to the existing technology. Both sensitivity and specificity can be low with microarrays. Accuracy can also be negatively affected by the low dynamic range of existing microarray technology. Perhaps more importantly, microarrays restrict the expression profiling data to specific annotations and content. Digital expression profiling using RNA-Seq and next generation sequencing (NGS) promises to reduce or in some cases eliminate these weaknesses. In order to evaluate the merits of RNA-Seq for expression profiling, we have performed an extensive comparison of data generated with the ABI SOLiD™ NGS platform and the Affymetrix Human Exon 1.0 ST GeneChip® platform. Using the Microarray Quality Control Consortium RNA samples as a model system we have demonstrated increased sensitivity, specificity and accuracy of RNA-Seq data relative to the microarray platform. TaqMan® realtime PCR was used as a third platform technology to assess relative performance of the NGS and array data and to validate the findings for both systems. We also show performance of the NGS data using a panel of 92 synthetic RNA spikes. This model system indicates that NGS offers extremely high sensitivity and accuracy, with no attenuation of signal at the high end of the dynamic range, as has been seen with microarrays. We have also compared the results of RNA-Seq using either polyA RNA fractions or total RNA that has been depleted of rRNA. While the results are highly concordant, the two sample types offer unique advantages and disadvantages. Using total RNA for RNA-Seq gives a fuller picture of the transcriptome, including non-coding RNA and non-polyadenylated transcript expression profiles, but may require more sequencing depth to attain the same level of sensitivity. RNA-Seq with polyA selected RNA results in high sensitivity and accuracy for expression profiling, but does not survey the entire transcriptome sequence space. Applications, such as, novel transcript discovery, splice variant discovery, allele specific expression and traditional gene expression profiling may require the use of one or both RNA sample types.

INTRODUCTION

Recent advances in array design promise the ability to detect alternative splicing as well as differential gene expression. Next generation sequencing (NGS) technologies, such as the Applied Biosystem SOLiD™ System, provide a digital expression profiling readout that is fundamentally different than analog measurement systems like microarrays. The SOLiD™ System has been shown to provide quality alternative isoform detection and differential expression analysis [1,2]. It can also provide data on allele-specific expression, alternative splice variants, expressed SNPs (single nucleotide polymorphisms), translocations and fusion transcripts, and information leading to the discovery of novel transcribed regions. Here we report a comparison of whole transcriptome expression profiling using SOLiD™ System RNA sequencing and Affymetrix Human Exon 1.0 ST GeneChip® analysis. We demonstrate that sequencing provides greater sensitivity, accuracy and dynamic range than exon arrays, however, in general there is good concordance between the platforms. We also use synthetic spike-in transcripts to assess absolute sensitivity and accuracy for the RNA-Seq method, and briefly compare with published spike-in results for GeneChip® arrays. We also discuss how the two technologies can be used together to provide good throughput and highly validated results.

MATERIALS AND METHODS

RNA Samples and Whole Transcriptome Sequencing Sample Preparation

Ambion® FirstChoice® Human Brain Reference RNA (HBRR), Ambion® FirstChoice® HeLa RNA and Stratagene Universal Human Reference RNA (UHRR) were used as starting material for both array analysis and SOLiD™ System sequencing. For SOLiD™ whole transcriptome sequencing, RNA samples were processed using the Ambion® Poly(A)Purist™ Kit to obtain poly(A) RNA or the Invitrogen Ribominus kit to selectively deplete rRNA. 50 ng samples were then used for library preparation using the SOLiD™ Whole Transcriptome Analysis Kit. Six technical replicates were prepared for each of the 2 MAQC PolyA RNA sample types (3 users/2 replicates each). For array analysis, 50 ng of HBRR or UHRR RNA was processed using the Ambion® WT Expression Kit and the Affymetrix GeneChip® WT Terminal Labeling Kit. Again, 6 technical replicates were prepared for each of the 2 RNA sample types (3 reagent lots/2 replicates each). The manufacturers' recommended methods were followed for all sample preparation steps. The synthetic spikes were prepared at Ambion from plasmid stocks provided by the ERCC. Roughly 15ng of the spike pools were added per 500ng polyA.

SOLiD™ Whole Transcriptome Sequencing and Analysis

Whole transcriptome libraries prepared with the SOLiD™ Whole Transcriptome Analysis Kit were amplified onto beads by emulsion PCR using recommended SOLiD™ System sequencing protocols. Enriched beads were deposited onto glass slides using the octet partition gasket and sequenced using the SOLiD™ 3 System and 50 bp reads. 20–30 million beads were obtained from each octet region for a total of 6.5–10.1 million uniquely mapped reads per sample. genome (hg18) using Applied Biosystems Whole Transcriptome (WT) Analysis Pipeline— a free off-instrument data analysis software package. The resulting uniquely mapped reads were imported as a sorted MAX file into Partek GS v6.5beta software which was then used to count tags mapping to transcripts in the RefSeq database (31,600), normalize the counts with median scaling or quantiles, perform statistical analysis, and visualize sequence tags with an integrated genome browser. The spike data consisted of 2 full slides (1 run) of HeLa resulting in approximately 100 million uniquely mapped reads. The CountTags module of the Applied Biosystems Whole Transcriptome (WT) Analysis Pipeline was used to assign counts to the spike-in data.

Exon Array Analysis

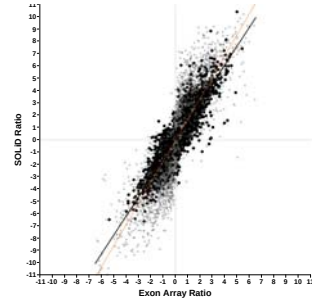
Hybridization and scanning of Affymetrix GeneChip® Human Exon 1.0 ST Arrays were performed according to manufacturer's recommended methods. Again, Partek software was used to import .cel files and pre-process the data using default RMA parameters. Gene-level signal was calculated as the simple average of all relevant core exon-level estimates.

General Analysis Methods

RNA-Seq and exon array data were compared by first merging transcript-level estimates by simple matching of RefSeq identifiers. This merge resulted in 15,065 RefSeq ID's common to both datasets. Differentially expressed genes were identified using a T test and fold-change threshold as recommended by the Microarray Quality Control (MAQC) Consortium (3). TaqMan® real-time PCR data downloaded from the FDA MAQC website was used to provide a third platform comparison. Merging of this dataset with array and SOLiD™ System sequencing data was performed based on RefSeq annotation provided by the MAQC project (n=735 matching assays). Tags per kilobase (TPKB) were calculated by dividing the number of counts mapped to a given transcript by the transcript length and multiplying by 1000. TPKB is used to filter low expressed RNAs.

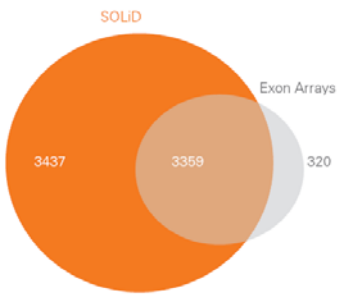
RESULTS

Figure 1. Correlation of RNA-Seq and Exon Array Ratios



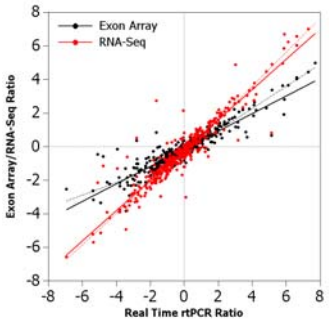
The scatterplot compares the Log2(HBRR/ UHRR) ratios for data from SOLiD™ System sequencing and exon array analysis. Black data points represent genes detected at TPKB≥5. If the gene was detected in both UHRR and HBRR at mean TPKB≥5, the data point was retained (n=9,279). The grey data points are those that did not pass this threshold. The red regression line was calculated with all 15,065 data points mapped between the platforms, and has a Pearson correlation of 0.869. The black regression line was calculated for only those RefSeq transcripts that pass the 5 TPKB filter. The correlation rises to 0.908 after filtering. This is due to removal of lower expressed transcripts that have higher noise levels and transcripts that are not detected at all in at least one sample.

Figure 2. Concordance of Differentially Expressed Transcripts



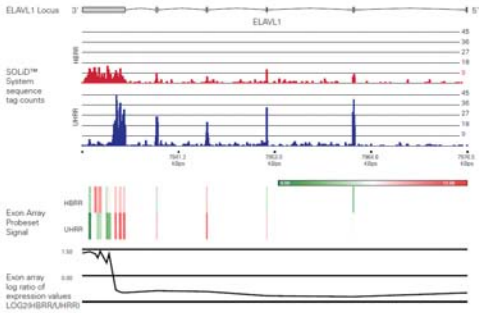
The concordance of differentially expressed gene (DEG) lists for the 2 platforms is shown here. Of 15,065 transcripts that could be mapped to both platforms, 7116 were found to be differentially expressed by at least one technology. Differential expression was determined by using a T test and fold-change threshold (p<.001 and >2-fold change between RNA samples). Only 320 out of 3679 DEGs found by exon array were not found by SOLiD™ System sequencing (~9%), indicating very high agreement. 3,437 out of 6,796 DEGs found by SOLiD™ were not detected by arrays. The apparent increase in DEG detection by SOLiD™ System sequencing is thought to be due in large part to the increase in dynamic range and accuracy (see figures 3 and 7).

Figure 3. Correlation of RNA-Seq and Exon Array Ratios with Real-Time rtPCR



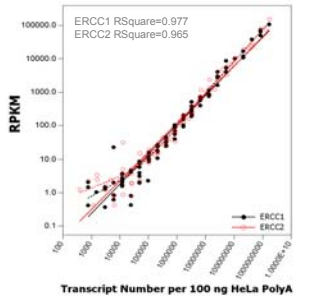
The scatterplot compares the Log2(HBRR/UHRR) ratios of data obtained via SOLiD™ System sequencing and exon arrays with TaqMan® realtime PCR data (4). Only transcripts that could be mapped at >5 TBKP in both samples were included in the analysis (n=406). Pearson correlations of data from TaqMan® real-time PCR with that from the SOLiD™ System (r=0934) or exon arrays (r=0.922) is similar. The slope (m) of the regression fits however, indicate that SOLiD™ System sequencing (m=0.901) shows a much greater dynamic range than exon arrays (m=0.521). This difference indicates significantly greater accuracy, relative to a "Gold Standard" method. Solid lines indicate linear regression fits and dashed lines are Lowess smoothing fits.

Figure 4. Alternative Isoform Detection



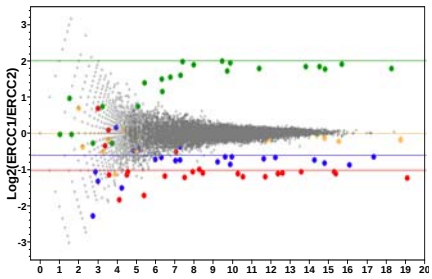
The ELAVL1 gene shows alternative isoform usage in HBRR vs. UHRR. In HBRR, relatively constant low level expression of ELAVL1 across the predicted RefSeq transcripts is seen, whereas in UHRR, slightly higher levels of ELAVL1 transcripts are evident. UHRR appears to preferentially use a shortened 3' exon that is likely the result of alternative polyadenylation site usage. Although the RefSeq database does not contain more than one isoform for this gene, there are 2 Ensembl transcripts which precisely predict this behavior (not shown). The heatmap and ratio profile plot show exon-level array analysis data that is concordant with this interpretation. There are 9 probesets on the array that interrogate the 3' Exon. The heatmap and ratio plot clearly show the 3' most 6 probesets decreasing in UHRR, relative to HBRR. These probesets show perfectly concordant data with RNA-Seq.

Figure 6. Dose Response of 92 Spike-ins



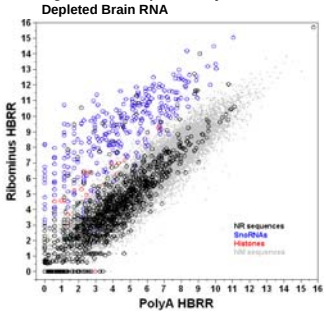
The scatterplot shows dose response data for 2 independent pools of 92 synthetic transcripts designed by the ERCC (6). The 2 pools were spiked into polyA HeLa RNA with each spike at the levels indicated on the X axis. ~50 million uniquely mapped reads were generated for each ERCC pool (ERCC1 and ERCC2). High linearity extends through >5 logs of dynamic range with no attenuation at the high end as is seen for analog microarray measurements where there is only 2-3 logs of linear range. Based on this depth of sequencing we see detection of less than 10,000 copies in 100 ng polyA at 1RPKM (read per Kb per million uniquely mapped reads). Solid lines indicate linear regression fit and dashed lines are Lowess smoothing fits.

Figure 7. MA Plot Comparing Spike Pool Ratios



The MA Plot shows ratio-metric performance of the 2 ERCC pools. The pools are divided equally into 4 sub-pools of 23 spikes each. These sub-pools are designed to evaluate 1.5, 2 and 4 fold change, as well as, no change between the 2 ERCC pools. As can be seen the different ratios are performing as predicted throughout most of the dynamic range. The spikes are designed to extend through 20 log2 units (~6.6 logs). The spikes are very closely approximating expectation with at least 18 log2 units of dynamic range. As was seen in fig. 6, no high end saturation was observed for the ratios, as with array data (7). Green=4 fold, red=2 fold, blue=1.5 fold, orange=no change, grey=RefSeq Genes. Corresponding colored lines indicate expected Log2 values for the 2 pools. This data was not normalized but only log2 transformed with an offset of 1 to remove 0 values.

Figure 5. Scatterplot of PolyA and rRNA Depleted Brain RNA



The scatterplot shows RefSeq transcript profiles for polyA and rRNA depleted HBRR. Although there is fairly high correlation between these sample types, rRNA depleted samples show "up-regulation" of a large number of transcripts. Both snoRNAs and histones are dis higher as expected because of the absence of a polyA tail in these species. A number of other NR refseq transcripts (non-coding) are also up-regulated. We are currently investigating presence of polyA tails in other transcript types that are up-regulated in rRNA depleted samples. This data was quantile normalized because of non-linearity induced by different sensitivity levels for these sample types.

CONCLUSIONS

Both SOLiD™ System whole transcriptome sequencing and exon array analysis are useful tools for the analysis of differential gene expression and alternative splicing. SOLiD™ System transcriptome sequencing, or RNAseq, delivers high sensitivity, accuracy, and broad dynamic range. Furthermore, it is a hypothesis-neutral approach that can be used to discover and annotate novel transcripts and isoforms. On the other hand, exon array technology is relatively inexpensive and easy to use. It is already in use and validated in many labs around the world and offers the ability to rapidly evaluate many samples for alternative splicing events. The two technologies complement each other and can be used where they are strongest. The SOLiD™ System is the platform of choice for discovery, offering hypothesis-neutral analysis, scaleable sensitivity, high accuracy and wide dynamic range. Exon arrays offer cost and time advantages that are attractive for larger scale follow-up or validation studies.

REFERENCES

1. Tang F, Barbacioru C, Wang Y et al. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* 6:377–382.
2. Wang T, Fan L, Watanabe Y et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456:470–476.
3. MAQC Consortium, Shi L, Reid LH et al. (2006) The Microarray Quality Control (MAQC) project shows inter and intra platform reproducibility of gene expression measurements. *Nat. Biotechnology* (9):1151–1161.
4. Canales R, Luo Y, Willey JC et al. (2006) Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol.* (9): 1115–1122.
5. Shrinivasro P.M., et al. (2009) Transcriptome sequencing of the Microarray Quality Control (MAQC) RNA reference samples using next generation sequencing. *BMC Genomics* 10:264.
6. External RNA Controls Consortium (2005) Proposed methods for testing and selecting the ERCC external RNA controls. BMC Genomics 6:150.
7. Cope LM et al. (2003) A Benchmark for Affymetrix GeneChip Expression Measures. Bioinformatics 1(1):1–10.

TRADEMARKS/LICENSING