Gene Expression in human samples using SOLiD™ Next Generation Sequencing

Catalin Barbacioru, Melissa Barker, Raymond Samaha, Jingwei Ni, Yongming Sun, Jian Gu, Scott Kuersten, Bob Setterquist, Roland Wicki and Eugene Spier
Applied Biosystems, 850 Lincoln Center Dr., Foster City, CA 94404

ABSTRACT

Analysis of gene expression patterns provides valuable insight into the role of differential expression in biological and disease processes. High density microarrays -- the standard for global gene expression analysis -- are limited in their dynamic range and can be ineffective at measuring genes expressed at a low level. Additionally, hybridization based platforms require an a priori knowledge of the mRNA sequences and are therefore unsuited for hypothesis free RNA discovery type of studies. The SOLiD™ System overcomes the limitations of microarray technologies by providing an ultra high throughput sequence-based platform for quantitative measurement of expression of RNA molecules. SOLiD produces 100's of millions of short reads (25-50bp) in a single run, requiring low sample input. This allowed us to explore gene expression profiles at the whole genome scale, without prior RNA sequence knowledge enabling an entirely new scale of biological experimentation, with large dynamic range, a tunable depth of coverage for rare transcript discovery and quantification. We measured TaqMan ddCt values for 667 genes between UHR and brain tissues and observed good correlation (r^2=0.89) with log2(fold change count) for two samples.

INTRODUCTION

Using a newly developed library protocol which requires low sample input and results in sequence ready samples in less than a day, we explored the mRNA transcripts expression profiles in two samples, brain and universal human reference (UHR).

Total RNA undergoes a sequence of specific sample preparation steps including fragmentation, adaptor hybridization, ligation, reverse transcription, PCR, gel purification. The resulting libraries were sequenced using the SOLID™. 50-base long (color) sequence information is then used to identify the corresponding transcripts and genes, and determine their abundance.

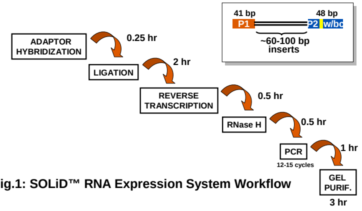


Fig.1: SOLiD™ RNA Expression System Workflow

RESULTS

Mapping. Reads are mapped to a color coded reference sequence obtained by concatenating all 32,661 human RefSeq entries (NCBI RefSeq release 26). Reads matching uniquely the RefSeqs are used for consequent analysis. The reads that did not map to the NCBI database were also mapped sequentially to several other references including tRNAs, rRNAs, and the human genome:

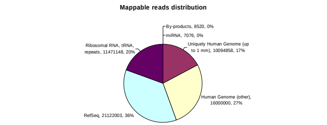


Fig 2: Mapping stats for 59 Mln reads in UHR

Coverage.
Mapped reads distribution is shown above, e.g., 21 Mln reads (36% of total) map to RefSeqs including NRs (non-coding). The number of reads per starting point ranging from 1 to 24,000. As expected, the number of starting points corresponding to one transcript is linearly correlated to the size of the transcript (Fig. 3 A, r^2 = 0.93). The location of starting points is uniformly distributed across transcripts (Fig. 3 B) with dips at the ends.

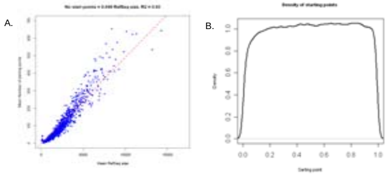


Fig. 3: Length of the transcript and number of starting points. A) Number of starting points as a function of transcript length; B) All transcripts coverage

Detection. We defined RefSeq as "detected" if at least 3 SOLiD reads uniquely mapped to this RefSeq. Based on this – we saturated detection at ~6 and 8 Mln reads for Brain and UHR, respectively detecting ~80% of all RefSeqs.

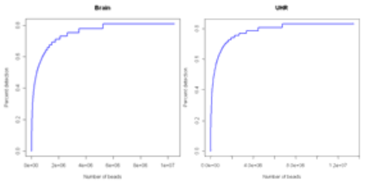


Fig. 4: Percentage of RefSeqs detected in Brain and UHR samples depending on the number of reads.

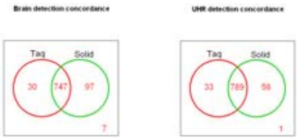


Fig. 5: Detection comparison. 874 RefSeq's have been measured using TaqMan® assays in brain. Transcript detection defined for SOLiD™ system as above, while for TaqMan® assays if Ct measurements are below Ct 35. For 85% and 90% of the interrogated RefSeq's the two platforms are in concordance, some differences being noticed at the low level of expression.

Reproducibility. Both Brain and UHR were run twice: Lib. 1: Oct 26, instrument # 21; Lib. 2: Apr 2, # 73 (Fig. 6). In both samples the r^2=0.99 and most of the outliers are belong to genes with "pile-ups" in start points – see Fig. 6 B-D examples.

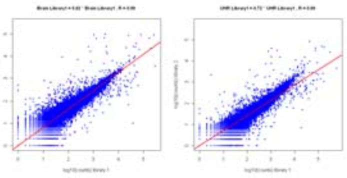


Figure 6: Reproducibility. System reproducibility has been tested using two technical replicates run on two different instruments 5 months apart

Examples. For majority of the transcripts tag coverage follows a Binomial distribution (Fig. 7 A). There are transcripts where the coverage is not uniform showing pile-ups (Fig. 7 B-D).

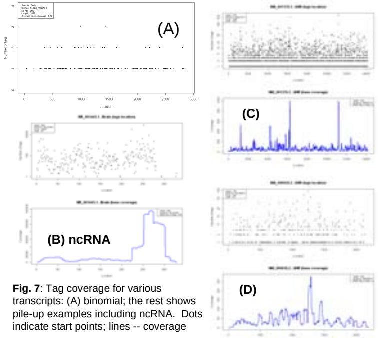
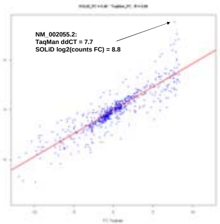


Fig. 7: Tag coverage for various transcripts: (A) binomial; the rest shows pile-up examples including ncRNA. Dots indicate start points; lines -- coverage

Fig. 8: Correlation between TaqMan delta-deltaCt for Brain and UHR and SOLiD log2(fold change count). 677 genes that are present in both samples and platforms are shown; R^2=0.89



The pipe-ups in SOLiD coverage are likely to indicate stable 5' P-ends in complex mixtures of RNA species in the samples.
We found an excellent overall correlation between TaqMan deltaCts and SOLiD log2 (fold change count) when comparing two tissues (Fig. 8)

CONCLUSIONS

The SOLiD™ RNA Expression Kit provides a streamlined workflow that greatly reduces the time, cost, and experimental variability associated with library preparation. Researchers can now generate whole transcriptome libraries in a single day with a simple, easy to use protocol. Using a newly developed library approach we have profiled the expression levels of total RNA for two samples. This protocol is shown to randomly fragment transcripts, except stable 5'-phosphorylated ends that show as pile-ups. The system is capable of detecting low abundant transcripts and running multiple sample on the same slide using multiplexing (AKA bar-codes). SOLiD™ system empowered by the RNA expression kit offers a highly reproducible and sensitive alternative to hybridization array platforms, suitable for both expression profiling as well as discovery of novel transcripts. It provides incredible rich details about a multitude of RNA species expressed in the genome including ncRNA.

REFERENCES

The Microarray Quality Control (MAQC) project shows inter- and intra-platform reproducibility of gene expression measurements. Nat Biotechnol. 2006 Sep;24(9):1151-61

TRADEMARKS/LICENSING

Research Use Only. Not for use in diagnostic procedures.