SOLiD™ System High-throughput Analysis of Differential Gene Expression

Introduction

Analysis of gene expression patterns provides valuable insight into the role of differential expression in normal biological and disease processes. TaqMan® Gene Expression Assays are considered the 'gold standard' when a defined number of genes are being studied due to the wide dynamic range (8 to 9 logs) and robustness of this system. High density microarrays have been used to globally assay mRNA expression levels, however, microarrays are limited in their dynamic range and can be relatively ineffective at measuring low copy genes. Furthermore, traditional microarray approaches depend upon 3' biased sample preparation and hypothesis driven probe design, limiting their ability to detect novel exons or differentiate between splice variants. Sequence tag-based expression techniques such as SAGE (Serial Analysis of Gene Expression), SuperSAGE, CAGE (Cap Analysis Gene Expression), and 5' SAGE, provide extremely sensitive, hypothesis-neutral sample preparation methods. To date, however, these assays have been limited by the throughput of traditional sequencing technologies.

The SOLiD™ System overcomes the limitations of both microarray and SAGE technologies by providing a highly sensitive, hypothesis-neutral, method for the detection of gene expression on a genome-wide scale. Collaborators at the University of Tokyo have modified their 5' SAGE protocol[1] for use with the SOLiD System (termed 5' SOLID). They used this high throughput method to generate millions of 5' biased tags and to study the expression profiles of a colon cancer cell line treated with various anticancer compounds.

Method

HT-29 colon cancer cells were treated with 5-Azacytidine, Tricostatin A or a combination of the two drugs. Fragment libraries were generated from approximately 5 ug of polyA+RNA isolated from each of the three treatment cases plus one untreated control sample using the 5' SOLiD method described here (Figure 1). The 5' SOLiD library protocol, based off a traditional 5' SAGE method, resulted in a random collection of unique sequence tags from the 5' UTR of mRNA molecules. The 5' SOLiD tags were mapped to the NCBI, b36, hg18 reference sequence and these tags were then mapped to RefSeqs. The tags mapping to each RefSeq were then counted. The association between the results from two independent runs of the same library was calculated using the Pearson's statistical test. Details of the protocol, analysis tools and results are described in an upcoming publication.

Results

Reproducibility and Dynamic Range

A combined total of 40 x 10^6 tags (8.7 x 10^6, 11.7 x 10^6, 9.7 x 10^6 and 10.4 x 10^6) were generated from the four libraries and mapped to the human genome. These tags were then mapped against
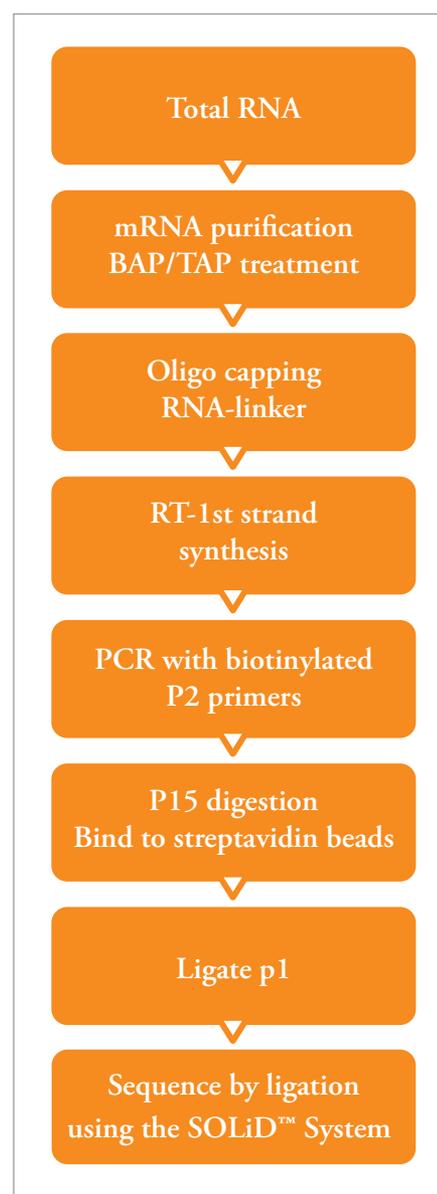

Total RNA

↓

mRNA purification BAP/TAP treatment

↓

Oligo capping RNA-linker

↓

RT-1st strand synthesis

↓

PCR with biotinylated P2 primers

↓

P15 digestion Bind to streptavidin beads

↓

Ligate p1

↓

Sequence by ligation using the SOLiD™ System

Figure 1. 5' SOLiD™ Workflow

known RefSeqs resulting in 5.3×10^6, 7.5×10^6, 5.8×10^6 and 6.5×10^6 unique tags mapping respectively. In order to assess the reproducibility of the system, unique tags mapping to RefSeqs isolated from two independent runs of one library were compared (Figure 2). The number of unique tags representing each RefSeq was plotted for both runs. A Pearson coefficient (r^2) of >.99 was calculated, illustrating a strong correlation between the results of the two independent runs. This high degree of reproducibility between sequencing runs is essential for accurate detection of subtle changes in gene expression.

This data also demonstrates the sensitivity to detect low levels of gene expression with the SOLiD System. Transcripts were detected over a range in expression levels from <1 copy per cell to over 100,000 copies per cell, corresponding to a dynamic range of $>10^5$.

In addition, the dynamic range of the system was calculated by normalizing the number of unique RefSeq tags from each sample to 6×10^6. Estimating the number of RNA molecules present in a cell at 300,000, the copy number of a gene-specific RNA can be calculated using the following formula: Copy Number = # of gene tags/6,000,000 X 300,000. Using this formula, the copy number of genes identified from the four libraries was found to vary between 0.02 – 4000 copies/cell, or a dynamic range of 10^5 for all runs. In comparison to other gene expression platforms, the SOLiD System provides a dynamic range that is several orders of magnitude greater than standard microarrays (Table 1). It is also

TABLE 1. Dynamic range of various gene expression detection platforms [2,3]

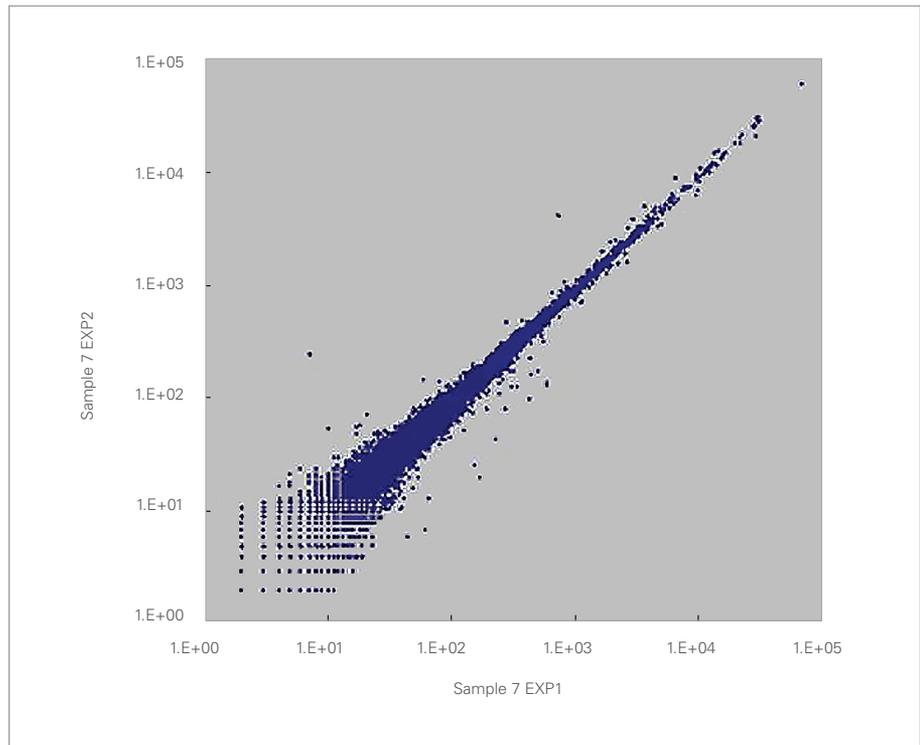|	
Arrays	2-3 logs
SOLiD	5 logs
TaqMan	7-8 logs

Figure 2. Comparison of unique tags, mapping to NCBI Reference Sequences, from two independent runs of one library of mRNA.
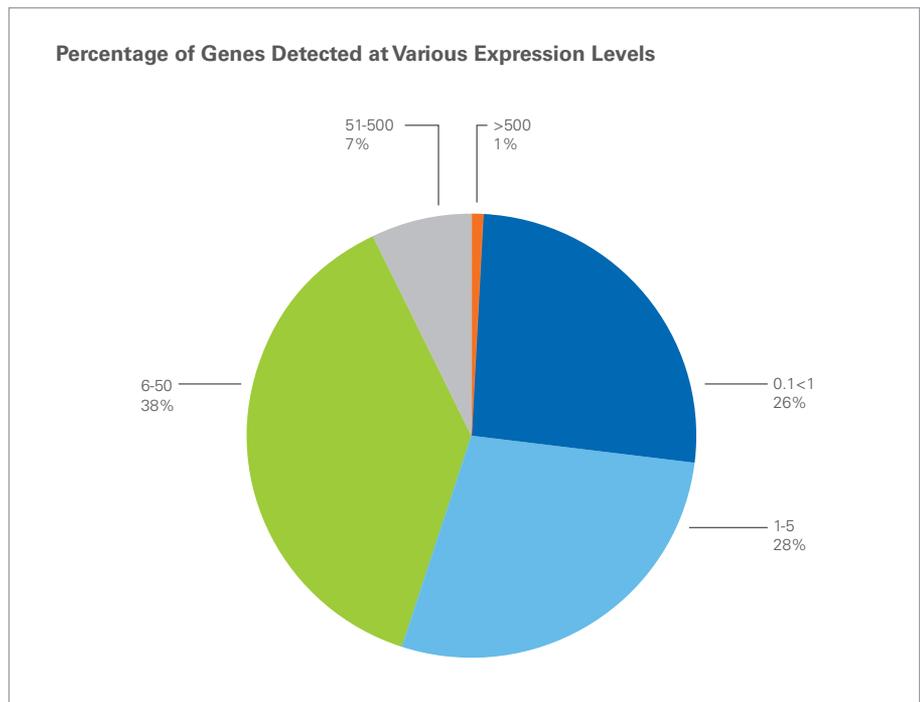

Figure 3. Percentage of genes detected by 5' SOLID as a function of copy per cell. The number of transcripts mapping to both the human genome and RefSeqs were first normalized to 6 million tags/library. The number of copies/cell for each transcript was computed using the formula cited in the text and grouped as shown in the figure. The percentage of the total number of transcripts in each group is shown.
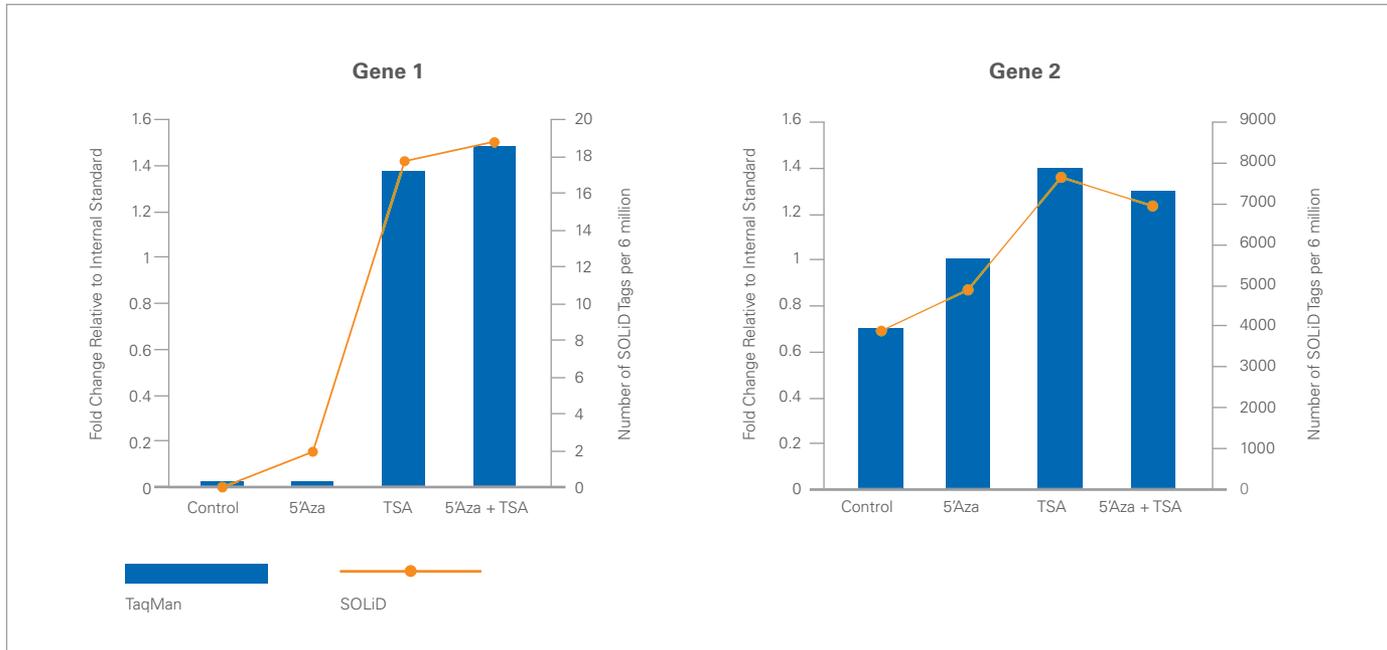
Figure 4. Differential gene expression in a colon cancer cell line in response to various anticancer compounds. The vertical bars represent the fold change of the target gene relative to the internal control gene with different treatments and the scale is shown on the left Y axis. The number of 5' SOLiD tags for the same genes and treatments are represented by the lines and the scale is shown on the right Y axis.

worth noting that the dynamic range obtained with this platform can be simply extended by additional sequencing of the beads from the same libraries. With additional coverage we expect the range to be comparable to that of TaqMan assays.

Several publications have estimated that microarrays are capable of detecting transcripts present at 3-5 copies/cell[3]. The data from the 5' SOLiD runs indicate that this system can detect transcripts present at levels 100X lower than microarrays. The copy number of all transcripts detected was calculated and the results showed that approximately 55% of the transcripts identified were present at levels less than 3-5 copies/cell or at levels below what can be quantitatively detected using microarrays (Figure 3).

Quantitation of Gene Expression Levels

A number of genes were identified by the 5' SOLiD method as having differential expression in response to drug treatment. These results were compared to results obtained using Applied Biosystems TaqMan® Assays. The data for two of these genes is shown above (Figure 4). Gene 1 expressed at very low levels in the control sample and after 5-Azacytidine treatment. Its expression level increased dramatically upon treatment with Tricostatin A and remained at a comparable level when the two drugs were combined. On the other hand, Gene 2 was expressed at modest levels in the control sample and increased significantly with both 5 Azacytidine and Tricostatin A treatment remaining elevated when both drugs were present. These results correlate closely with those previously obtained using Taqman® Gene Expression Assays.

Conclusion

A robust 5' SAGE workflow has been developed for the SOLiD™ System (5' SOLID). This workflow demonstrates significant advantages in sensitivity and dynamic range when compared to traditional approaches for studying whole genome gene expression. The expression levels of mRNA molecules were readily determined after sequencing unique sequence tags isolated from the 5' UTR of full length mRNAs. The flexible slide format used in the SOLiD System allows for the analysis of 1-16 samples per run. This enables the simultaneous analysis of both case and control samples on one slide. This approach was shown to be highly reproducible and demonstrated a dynamic range which is orders of magnitude greater than microarrays. Transcript levels for several genes identified in the 5' SOLiD analysis were confirmed by experiments using TaqMan® Gene Expression Assays, hence the SOLiD System provides a highly sensitive, hypothesis-neutral method for the detection of transcripts on a genome-wide scale.

References

[1]Hashimoto, SI., et al, *Nature Biotechnology*. 22:1146-1149 (2004).

[2]Design and Performance of the GeneChip® Human Genome U133 Plus 2.0 and Human Genome U133A 2.0 Arrays – Technical Note. Affymetrix Inc.

[3]Jun, L., et al., *Genomics*. 84:631-636 (2004).

Applied Biosystems