Single-Cell Whole Transcriptome Profiling With the SOLiD™ System

Introduction

The ability to study the expression patterns of an individual cell holds great promise to reveal molecular events that dictate normal development as well as tumorigenesis. However, it is frequently impractical to obtain large amounts of biological materials for expression analysis, especially with clinical samples. Furthermore, there is no established method that is capable of scanning the whole transcriptome at the individual cell level. Development of such a method would greatly facilitate advancement in developmental and stem cell biology.

Next-generation sequencing (NGS) technologies have been used in expression analysis as a high-throughput and cost-effective tool. For example, by sequencing and "counting" the cDNA tags derived from small RNAs with great depth and accuracy, thousands of small RNAs and their variants or isoforms can be simultaneously identified. These advances have dramatically increased our understanding of the regulatory networks for gene expression. Successful implementation of next-generation sequencing technology at the single-cell level would undoubtedly push expression analysis to a whole new level.

Previously, Kurimoto et al. [1] reported a single-cell whole transcriptome amplification method that efficiently amplifies double-stranded cDNA (ds cDNA) up to 3 kb with minimal bias. Here, we combined the cDNA amplification method with the Applied Biosystems SOLiD™ System to study

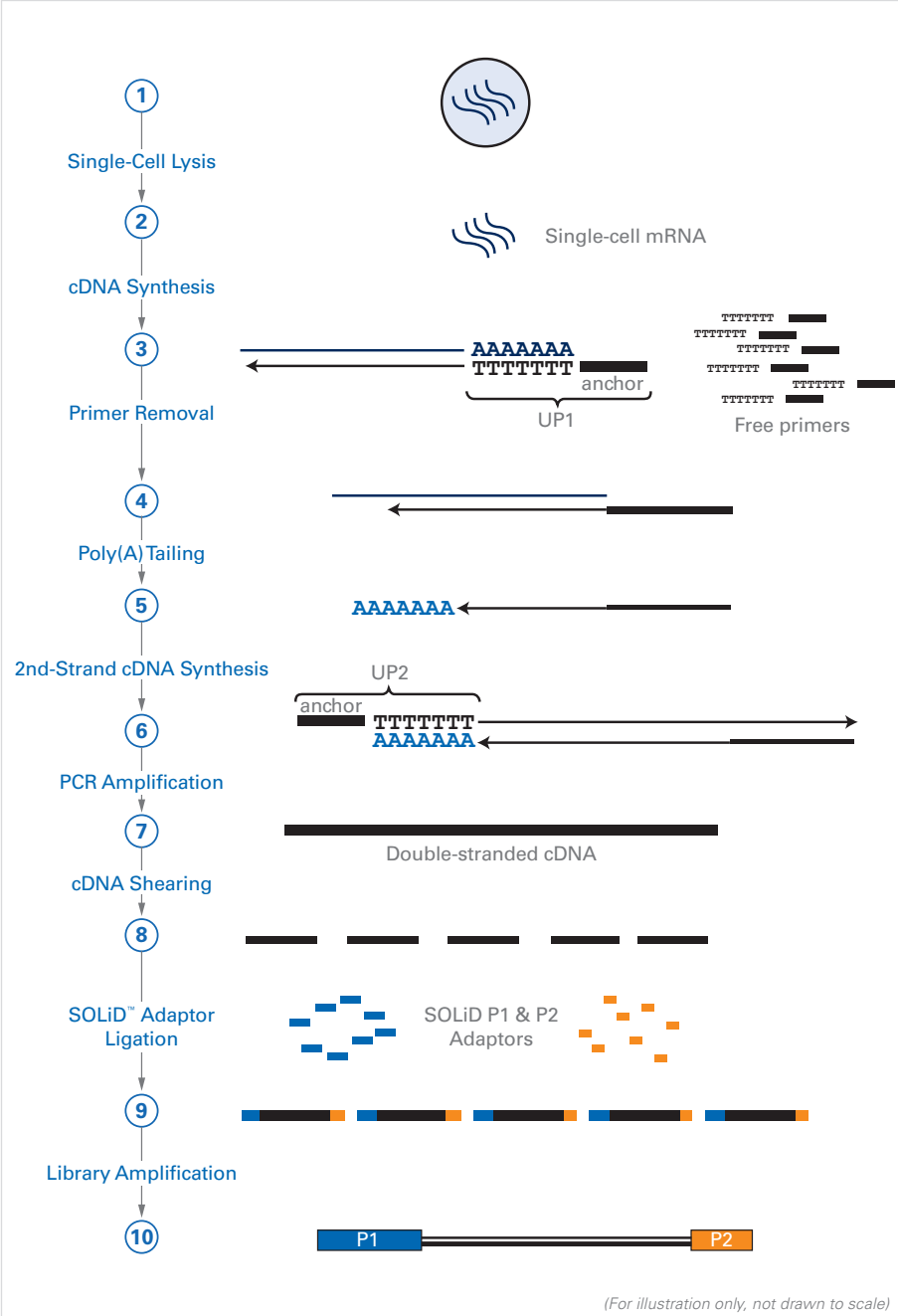

Figure 1. Major Steps of the SOLiD™ System Single-Cell Expression Library Preparation.

the expression profile of a single cell. Our studies indicated that it is feasible to obtain expression profiles at single-cell resolution with the SOLiD™ System. We generated over 168 million sequence reads, or over 80 gigabases of cDNA data. Such an unprecedented throughput allowed us not only to validate the expression patterns of known transcripts, but also to determine the expression of genes missed by microarray analysis. In addition, the method enabled the discovery of novel transcripts, alternative splicing arrangements, and intracellular abundance levels. The consistency and scope of single-cell expression profiling using SOLiD™ technology delivers significant benefits and opens up exciting possibilities in the study of individual stem cells and cancer cells in normal development as well as in tumorigenesis.

Methods

The major steps of single-cell analysis include single-cell lysis, cDNA synthesis, SOLiD™ library adaptor ligation, and final library amplification (Figure 1). Individual mature mouse oocytes and blastomeres from four-cell-stage fertilized embryos were isolated from wild-type mice; oocytes were also recovered from Dicer knockout mice. Those single cells were picked and lysed as previously described [1]. Subsequently, mRNA in the lysate was reverse-transcribed into first-strand cDNA by incubation at 50°C for 30 minutes with a poly(T) primer (UP1) containing an anchor sequence. Excess nonextended primers were digested with exonuclease I.

After the first-strand cDNA synthesis, a poly(A) tail was added to the 3′ end of first-stand cDNAs using terminal transferase. This allowed the synthesis of second-strand cDNA with a poly(T) primer (UP2) containing another anchor sequence. The newly synthesized ds cDNA molecules were PCR amplified. The vast majority of the ds cDNA molecules fell within the range of 0.5 kb to 3 kb, consistent with the sizes of the bulk messenger RNA. After purification, an aliquot of the ds cDNA was amplified for a few more cycles. The cDNA amplification procedure has been optimized to minimize PCR bias, and the resulting cDNA has been shown to be representative of the expression profiles from single cells [1].

For SOLiD™ System library preparation, 200 ng to 500 ng of the single-cell cDNA (0.5–3 kb) was sheared with the Covaris S2 system to yield short DNA fragments between 80 and 130 bp. The ends of the target DNA were repaired and subsequently ligated to SOLiD™ System P1 and P2 adaptors. The resulting ligated population was resolved on a 6% PAGE gel, and the fraction containing 150 to 200 bp DNA was excised and PCR amplified. To minimize PCR-introduced bias, the number of amplification cycles should be minimized, and this minimum number can be estimated by the ability to visualize amplified products on a standard Lonza FlashGel™ system. Typically, 8 to 10 cycles of PCR are necessary to generate sufficient library material (as judged by gel visualization). After purification of the amplified products, the SOLiD™ System single-cell expression library is ready for emulsion PCR.

Multiple emulsion PCR reactions were carried out according to the SOLiD™ System User Guide to yield a sufficient quantity of templated beads. The single-cell expression libraries were sequenced to either 35- or 50-base read length on the SOLiD™ System.

Reads were initially mapped against known RefSeq transcripts. However, such an approach limits the mapping of the reads to known exons and splicing events. In order to identify and align the reads that map to novel exons and alternative splicing junctions, we

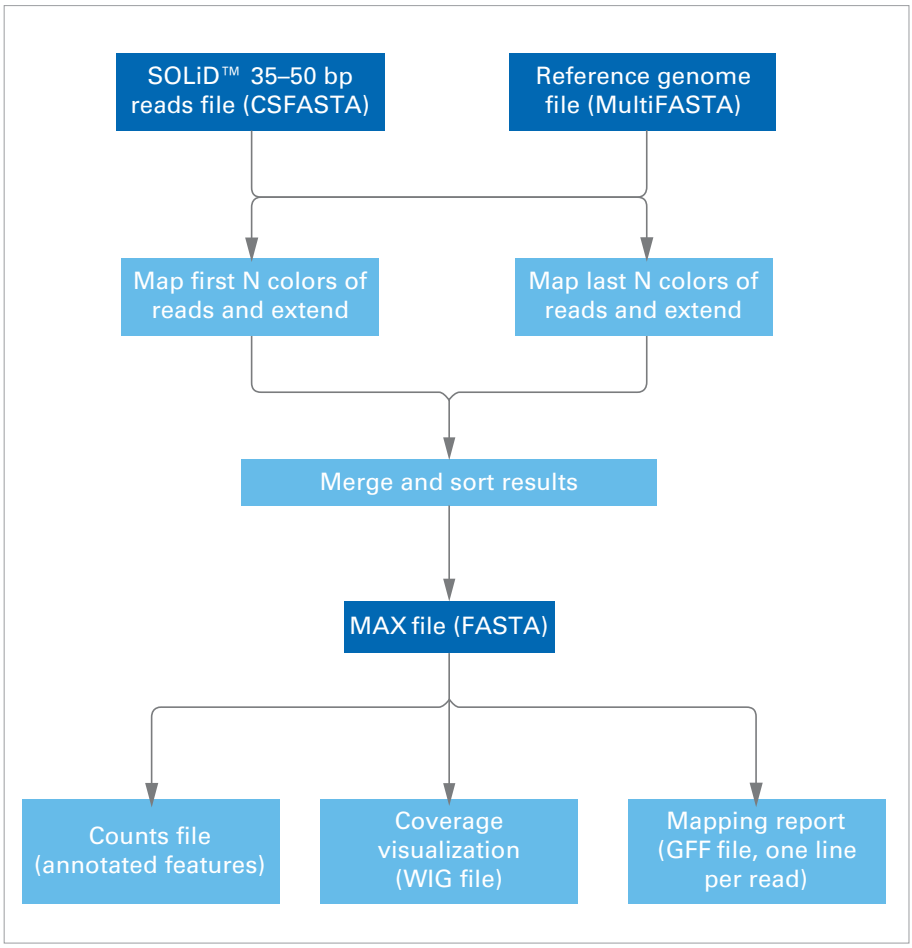

Figure 2. Overview of the SOLiD™ System Single-Cell Expression Sequencing Data Analysis Pipeline. In this study, N = 25.

developed new mapping tools that used the mouse genome sequence directly as a reference instead of RefSeq. As outlined in Figure 2, the software split up the reads into two parts: 35-base reads were split into 25- and 10-base reads from both ends, and 50-base reads were split into two 25-base parts from both ends. After the split, 25-base fragments were mapped to the genome using Applied Biosystems MapReads software. The assumption is that at least one of the split reads should come from a contiguous region with length greater than or equal to half of the total length of the unsplit read. The mapped split reads were extended according to a scoring function. This approach enables the discovery of novel splice junctions and exons. All analysis was performed in color space, and a MAX file in FASTA format was then generated and sorted. The MAX file was subsequently used to generate: 1) a Counts File, which records the sequence read counts of annotated features; 2) a WIG file, which is used for visualization of coverage with the generic UCSC Genome Viewer; and 3) a Mapping Report, which is a GFF file.

For cross-platform comparison between the SOLiD™ System and microarray techniques, previously published data on the expression profiles of 80 pooled mouse four-cell-stage embryos (320 blastomeres in total) were used. For comparison with real-time PCR, duplicate sets of diluted cDNA samples from single oocytes and blastomeres were analyzed by TaqMan® real-time PCR analysis with 384-well plates on the Applied Biosystems AB7900 Real-Time PCR System.

Results

Quality of cDNA Samples—The original cDNA synthesis procedure for single cells was optimized to produce cDNAs within the size range of 0.5 to 3.0 kb. The procedure is very consistent in terms of the cDNA size range, indicating that the majority of the expressed genes were converted into long, if not full-length, cDNAs.

Summary of SOLiD™ System Sequencing Results—After SOLiD™ System sequencing, we obtained a total of more than 168 million 35- or 50-base reads from single blastomeres and oocytes. As mentioned in the Methods section, we initially performed the direct mapping of all reads against the RefSeq database and mapped over 27 million reads for blastomeres, 9 million for the wild-type oocytes, and 11 million for the Dicer knockout oocytes (Table 1, Part I).

Since a significant portion of the reads did not map to the RefSeq sequences, we performed another round of mapping against the mouse genome using the read-split method outlined in Figure 2. We mapped over 64 million reads to the mouse genome with this method (Table 1, Part II) for blastomeres, 14.6 million reads for wild-type oocytes, and 21.3 million for Dicer knockout oocytes. The SOLiD™ System sequence reads were mapped to 189,620 known exons of the 9 mm mouse genome (Build 37, released in July 2007, assembly by NCBI), and those reads were assigned to all known mouse RefSeq transcripts. In total, 60% of all unfiltered reads mapped to either RefSeq genes or the mouse genome.

SOLiD™ System Robustness—To determine the consistency of the SOLiD™ System sequencing chemistry, we determined the concordance among the SOLiD™ System reads that were mapped to the positive and negative strands of the reference. Since both strands of the cDNA are derived from a single strand of mRNA, the positive and negative cDNA strands should maintain a high concordance ratio throughout the SOLiD™ System sample preparation. As shown in Figure 3A, we saw very high concordance (R=0.995) between the positive and negative strands for sequence reads from all single blastomeres. Other experiments with the wild-type and Dicer knockout oocytes also showed the same high concordance (data not shown). This indicates the robustness and high consistency of the SOLiD™ System single-cell profiling workflow.

In addition to strand comparison, we also examined the consistency of performance among different SOLiD™ instruments, which is crucial for meaningful data comparison between different experiments. In Figure 3B, the same mouse blastomere library was

TABLE 1. SUMMARY OF SINGLE-CELL SEQUENCING EXPERIMENTS USING BLASTOMERES, WILD-TYPE OOCYTES, AND DICER KNOCKOUT MUTANT OOCYTES.

Mapping Summary	Blastomere	Wild-Type Oocyte	Dicer Knockout Oocyte
Total reads processed: 168,883,617	110,232,318	20,998,366	37,652,933
Part I: Mapping to 21438 RefSeq	27,211,437	9,334,408	11,121,519
0 mm to 21438 RefSeq	13,638,318	5,071,973	6,558,271
1 mm to 21438 RefSeq	7,474,352	2,382,288	2,658,557
2 mm to 21438 RefSeq	6,098,767	1,880,147	1,874,691
Part II: Mapping to genome	64,807,549	14,623,581	21,299,168
Unmatched	44,997,355	6,003,943	14,472,985
Filtered sequences (primers, rRNAs)	427,414	370,842	1,880,780
Junction reads	4,215,238	922,656	1,170,273

Table 1. Matching of SOLiD™ System Reads Performed in Two Parts. Part I was against the RefSeq for known transcripts, and Part II against the mouse genome. Matching against both references enables the discovery of both known and novel transcripts.

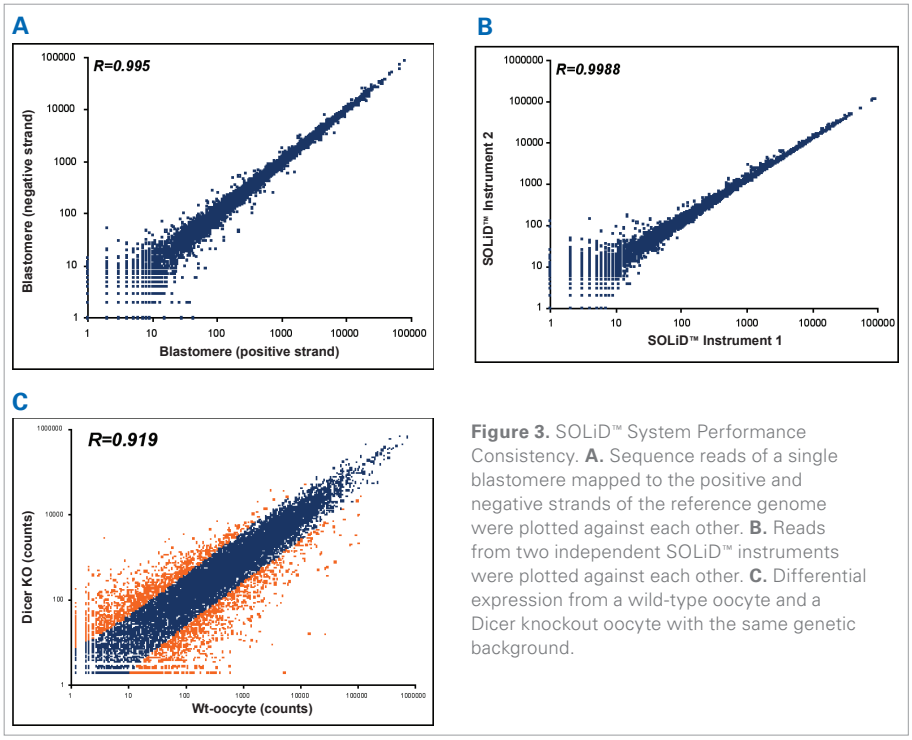

Figure 3. SOLiD™ System Performance Consistency. **A.** Sequence reads of a single blastomere mapped to the positive and negative strands of the reference genome were plotted against each other. **B.** Reads from two independent SOLiD™ instruments were plotted against each other. **C.** Differential expression from a wild-type oocyte and a Dicer knockout oocyte with the same genetic background.

used to prepare templated beads and sequenced on two independent SOLiD™ instruments by different operators. The reads for each transcript generated from two instruments were plotted against each other. The results show that the concordance ratio was remarkably high (R=0.9988), further validating the consistency of the instruments.

Furthermore, we determined the correlation of the SOLiD™ System sequencing results for different oocytes at the same developmental stage. As shown in Figure 3C, a wild-type mouse oocyte and a Dicer knockout oocyte with the same genetic background were plotted against each other. The result indicates that while sequence reads were overall well correlated with those cells (R=0.943), correlation was not as tight as those seen in Figures 3A and 3B. The result is expected, since the Dicer knockout is known to alter the expression landscape of the oocyte despite the otherwise consistent genetic background. Overall, our results clearly demonstrated the consistency and robustness of the SOLiD™ System sequencing and mapping algorithms.

Cross-Platform Comparison—
To benchmark SOLiD™ System performance against existing platforms, we compared SOLiD™ System single-cell results with those previously published using Affymetrix GeneChip® microarrays. We further validated our findings with TaqMan® real-time PCR analysis. As mentioned, we obtained over 168 million 35- and 50-base reads from our sets of experiments. The reads were mapped to 189,620 known exons of the mouse genome 9 mm and were subsequently assigned to known mouse RefSeq transcripts.

For comparison with microarray data, we used SOLiD™ System data from a single blastomere against a published data set obtained from 80 pooled four-cell-stage embryos (or 320 blastomeres) [2]. We found that 95.5% (6733 genes) of the genes detected by Affymetrix microarray using pools of blastomeres can be unambiguously detected by single-cell sequence reads (Figure 4A). The comparison is based on the 15,776 RefSeq transcripts that have probes on the array from 21,436 known RefSeq transcripts.

The small number of transcripts (317 genes) detected only by microarray were mostly very low expressors that appeared only because hundreds of cells were used in the assay. As shown in Figure 4B, the majority of the genes have low fluorescence intensity that is measured below 100 a.u. It is well known that at the single-cell level, low-abundance genes can be stochastically on or off. Therefore, it is quite possible that genes exhibiting low abundance when pools of cells are analyzed might not be expressed in the particular individual cell analyzed by the SOLiD™ instrument.

To further validate SOLiD™ System results, we chose 380 genes known

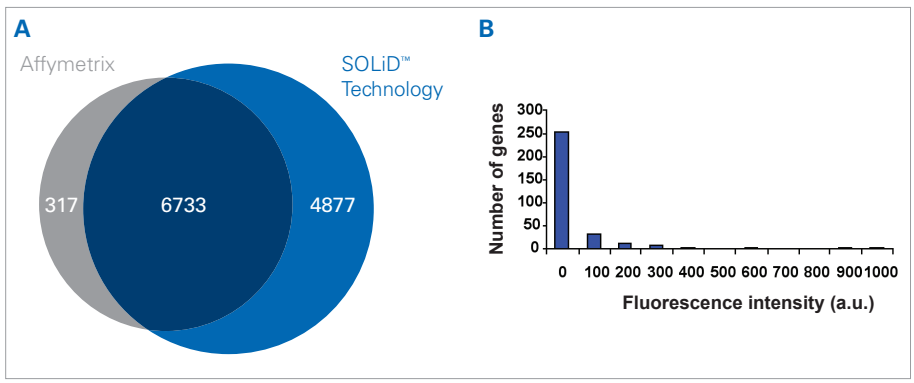

Figure 4. Cross-Comparison Between SOLiD™ System and Microarray Results. **A.** Venn diagram of the number of RefSeq genes with probes on the Affymetrix array that were detected by array and by SOLiD™ System reads. For the 15,776 transcripts with probes on the array, 317 transcripts were detected by array only, 6733 transcripts were detected by both array and SOLiD™ technologies, and 4877 transcripts were detected by SOLiD™ technology only. In total, SOLiD™ technology detected 11,610 transcripts in the RefSeq. **B.** The histogram of the frequencies and fluorescence intensities for the 317 transcripts detected only by microarray.

to play a role in early embryonic development and performed real-time PCR analysis. After looking at the fold changes in gene expression for both Dicer knockout and wild-type oocytes, we found that the concordance ratio between TaqMan® analysis and SOLiD™ instrument data was very high (R=0.926) (Figure 5). This indicates that SOLiD™ System data are highly reliable and in agreement with the most sensitive technology currently on the market.

In contrast, SOLiD™ System analysis detected the expression of 4877 additional known transcripts that were missed by microarray, despite the presence of probes for those genes on the array. This represents 69% more target genes that were detected by the SOLiD™ System. Furthermore, based upon direct genome mapping, an additional 1217 transcripts that do not have probes on the arrays were detected by the SOLiD™ System. In total, 12,827 known genes were detected by SOLiD™ System expression profiling with a single cell, demonstrating the power of the SOLiD™ System to identify existing and new transcripts without prior assumptions.

High-Resolution Mapping—Millions of SOLiD™ System sequence reads from a single cell can be directly visualized and studied using existing tools, such as the UCSC Genome Browser. In our Dicer knockout oocytes, exon 23 of the Dicer gene has been deleted through LoxP site-directed recombination with Cre recombinase. By mapping SOLiD™ System reads, we discovered that reads were absent at exon 23 in the Dicer knockout, whereas the adjacent exons were covered by SOLiD™ System sequencing reads (Figure 6). Such resolution with SOLiD™ System sequencing allows a direct look at the nucleotide level, facilitating the investigation of splicing mechanisms of the transcription machinery.

New Discovery—The ability to simultaneously study the structure (sequence and splicing variation) and

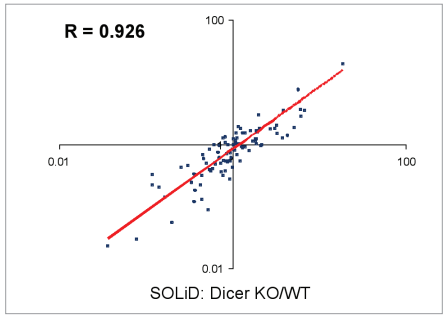

Figure 5. Concordance Plot of the Genes Detected by Both SOLiD™ Technology and TaqMan® Real-Time PCR Analysis.

abundance of transcripts in a single cell opens up exciting opportunities for new discoveries with the SOLiD™ System. As an example, we studied the alternative splicing events by generating all possible combinations of exon–exon junctions as an 84-base long sequence with 42 bases in each exon (~2 million junctions in total). Then, we removed the known exon junctions (~200,000). The remaining junctions were used as a new reference, and reads already aligned to the genome (but not matching known junctions) were mapped against this reference. As shown in Table 2, there were 1753 novel junctions covered by at least 5 sequence reads for one blastomere. For wild-type and Dicer knockout oocytes, the novel junctions detected with at least five SOLiD™ System reads were 2070 and 2411, respectively. Overall, these data show that different isoforms can be expressed in the same cell, and that SOLiD™ technology can readily identify novel splicing junctions and alternative splicing events. The ability to detect individual isoforms of a given gene would not be possible in assays using multiple cells, such as in microarray analysis.

To illustrate that SOLiD™ System sequencing can be successfully used to evaluate expression levels, we looked at the developmental pluripotency associated gene 5 (Dppa5), important in embryonic development. As shown in Figure 7, reads map to exons with a sharp boundary at the exon–intron junction, indicating the correct classification of the exons. Furthermore, the elevated expression of Dppa5 in Dicer mutants indicates the up-regulation of Dppa5, consistent with the suppressive role of the Dicer gene. In total, 1924 transcripts were upregulated in the Dicer knockout oocyte compared to wild-type control. Thus, the

TABLE 2. NOVEL JUNCTIONS DETECTED BY SOLID™ TECHNOLOGY IN BLASTOMERES, AND WILD-TYPE AND DICER KNOCKOUT OOCYTES.

Novel Junctions	4-Cell Blastomere	WT Oocyte	Dicer KO Oocyte
> 2 Hits	6701	9012	11,322
> 5 Hits	1753	2070	2411

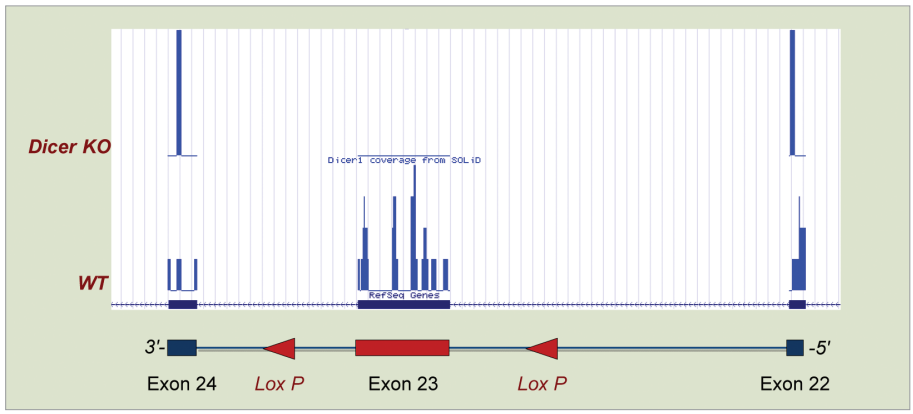

Figure 6. SOLiD™ System Reads Mapping to Exon 23 Regions of the Dicer Gene in Dicer Knockout and Wild-Type Oocytes.

entire transcriptome can be scanned by the SOLiD™ System for changes in expression, facilitating the functional characterization of important genes, such as Dicer, during development.

Discussion

Next-generation sequencing technology is transforming life science research. Its throughput and accuracy enable scientists to address fundamental questions with scope and depth previously unimaginable. Moreover, adoption of the next-generation sequencing technology in new application areas such as transcriptome analysis further demonstrates the versatility of systems such as the SOLiD™ System.

In this study, we demonstrated that the SOLiD™ System generated over 168 million reads from cDNA samples derived from single mouse cells. The ability to scan the whole transcriptome with mRNA material from a single cell opens up exciting possibilities in important areas such as stem cell research, cancer biology, and developmental biology. Contrast that with the limited sensitivity of existing platforms, particularly microarrays, that fail to detect low-abundance genes, despite the use of hundreds of cells. By directly interrogating the sequences instead of relying on reading fluorescence hybridization signals, SOLiD™ technology provides much higher resolution. With the SOLiD™ System and its streamlined sample prep workflow and data analysis capabilities, we demonstrate that the SOLiD™ System can provide quantitative estimates of messenger abundance as well as identify novel alternative splicing events. Overall, the SOLiD™ System delivers clear advantages over existing platforms in terms of sample usage, sensitivity, depth of coverage, and hypothesis-neutral target identification.

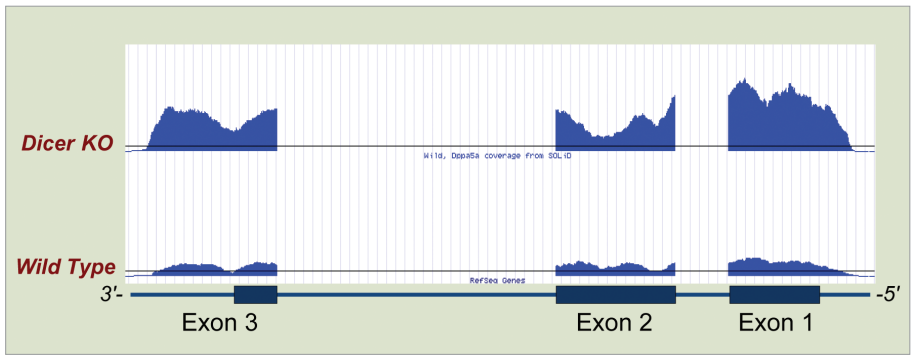

Figure 7. Difference in Expression of Dppa5a Gene in Wild-Type and Dicer Knockout Oocytes.

Reference

1. Kurimoto Ket al. (2006) An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res.* Mar 17;34(5):e42.
2. Maekawa M et al. (2007) Requirement for ERK MAP kinase in mouse preimplantation development. *Development.* Aug;134(15):2751–2759.
3. Tang F et al. (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods.* Advance online publication April 6, 2009. DOI:10.1038/nmeth.1315

Printed in the USA. 04/2009 Publication 139AP16-01

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