miRNA Discovery and Profiling with the SOLiD™ Small RNA Expression Kit

Figure 1: Workflow for SOLiD™ Small RNA Library Preparation

Introduction

Small non-coding RNA plays a key role in regulating a variety of biological processes, including developmental timing, cellular differentiation, tumor progression, neurogenesis, transposon silencing, and viral defense. Small RNA, typically only ~18–40 nucleotides (nt) in length, encompasses many classes: microRNA (miRNA), short interfering RNA (siRNA), piwi-interacting RNA (piRNA), repeat-associated short interfering RNA (rasiRNA), and other uncharacterized and novel small RNAs. Small RNA functions in gene regulation by binding to its targets and modulating gene expression using mechanisms such as heterochromatin modification, translational inhibition, mRNA decay and nascent peptide turnover. As they were discovered relatively recently, the study of small RNA is a young and rapidly changing field in which novel forms remain and the sequence modifications and activity of mature miRNAs are not well understood[1].

The current tools for studying small RNA are inadequate for whole genome discovery and characterization of novel small RNA. High throughput platforms based on probe-hybridization, such as microarrays, require a priori knowledge of the miRNA sequences, have limited dynamic range, and have poor sensitivity. Moreover, current sequencing-based methods for small RNA library preparation are time-consuming, prone to variation due to complex protocols, and require a significant amount of input RNA. TaqMan® Assays provide a robust, highly sensitive turn key method for profiling miRNA but are limited to analysis of known species.

Here we describe a new robust method for hypothesis-neutral, whole genome analysis of small non-coding RNA in general and miRNA in particular, using a simplified, single day procedure for preparing small RNA. This new approach, coupled with the SOLiD™ System, provides an extremely sensitive method for digital expression that enables the discovery and profiling of novel small RNA from a small quantity of total RNA (10-500 ng). The SOLiD™ System's ultra-high throughput (greater than 200 million mappable sequence reads, or tags, per run), wide dynamic range, and

high sensitivity, make it particularly suited for analyzing low RNA expression levels and measuring accurate fold changes at these levels.

Methods

Total RNA was isolated from samples of human placenta (300 ng RNA) and lung (500 ng RNA). The small RNA (10-40 nt) was purified by flashPAGE™ fractionation and converted to amplified libraries using the SOLiD™ Small RNA Expression Kit (Figure 1). The resulting cDNA libraries, containing the adaptor sequences necessary for SOLID sequencing, were clonally amplified onto beads with emulsion PCR using the SOLiD™ ePCR Kit. Following the standard protocol [SOLiD™ System 2.0 User Guide], the four libraries were deposited onto separate segments of a single, four quadrant slide and sequenced on the SOLiD™ Analyzer.

A combined total of 173 million sequence reads were generated: 40.5, 45.7 and 46.2 million for the three quadrants containing placenta library, and 40.5 million for the single quadrant of lung library, respectively (Figure 2). All reads were mapped sequentially against (1) miRNA Sanger reference sequences –build 10, (2) ribosomal RNA, tRNA and low complexity regions of the human genome, (3) human RefSeq sequences and (4) human genome (NCBI build 36).

The miRNA reference was based on miRBASE sequence data and constructed to enable detection of alternative miRNA forms with variable 5' and 3' ends. The reads mapped to this miRNA reference identified 34.4 million reads from the three placenta library quads and 13.3 million reads from the lung library quad, allowing 0-1 mismatch. A high percentage (92-95%) of these reads matched the miRNA reference sequence in a unique location.

After filtering for rRNA, tRNA, repeated regions of the human genome or human RefSeq sequences, 29.7 million from the three placenta quads and 12.1 million from the lung quad remained and were used for the subsequent analysis/ profiling. The number of reads per miRNA was determined and used as the expression level for that particular miRNA.

Figure 2: Four samples were sequenced on the SOLiD™ Analyzer using a quad slide configuration. A total of 173M reads were generated and a subset (44.2 M) of these reads were used for quantitative analysis.

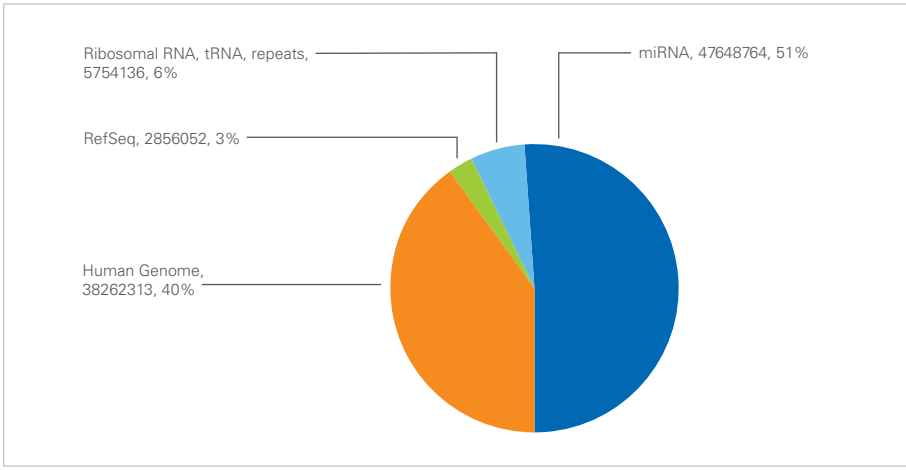

Figure 3: Distribution of 94M unique tags mapping to respective databases. Sequence tags were mapped with 0-1 mismatch against miRNA, ribosomal RNA, tRNA and low complexity regions in the human genome, known RefSeq sequences and human genome (NCBI build 36). The figure shows the numbers and percent of total tags that map to each reference.

Results

Detection of putative novel small RNA.

Distribution of the 94 million tags is described in Figure 3. Approximately 51% of the tags mapped to known miRNAs. Greater than 40% of the tags mapped to genomic regions but not to any known RNA species. The small percentage (9%) of the tags that mapped to known rRNA, tRNA, repeat regions, and RefSeq indicate that the library preparation method successfully enriched for the small RNA fraction. A subset of the tags that mapped to the human genome but not to miRNA could represent previously uncharacterized miRNAs. Confirmation of these species as novel small RNA is currently being validated with TaqMan® Assays.

Reproducibility and Dynamic Range

To assess the reproducibility of the instrument system for miRNA detection, unique sequence tags mapping to

miRBASE and isolated from two independent runs of the placenta library (2 quad segments), were compared (Figure 4). A linear regression between the two sets of counts produced a coefficient of determination R^2 of >.9996, illustrating an excellent correlation between the results of the two quadrants over the entire six logs (base 10) of dynamic range. This high degree of reproducibility is essential for accurate detection of subtle changes in small RNA expression between two samples run on the same slide.

miRNA Expression Analysis—Correlation to TaqMan results

A major limitation of microarrays is the observed "ratio compression" of expression levels. The wide dynamic range observed on the SOLiD™ Analyzer suggested that expression levels derived from the SOLiD™ System and TaqMan® platforms would show strong correlation. Of the 423 miRNAs identified in placenta tissue, TaqMan® Assays for 244 microRNA targets were readily available. Comparison of expression levels between the two platforms for these miRNAs demonstrated a high correlation coefficient of 0.87 (Figure 5, Panel A). This value is considerably greater than the typical correlation observed between microarrays and TaqMan® results. Restricting the analysis to miRNAs that show significant differential expression increased the correlation coefficient to 0.90 (Figure 5, Panel B). These data demonstrate that the SOLiD™ System generates miRNA gene expression profiles that correlate well with TaqMan® data (R = 0.90), suggesting that the SOLiD™ platform is a valid profiling tool for gene expression analysis.

Variability in miRNA start points

Analysis of several previously characterized miRNA revealed a significant number of molecules with different 5′ start points than previously described on the Sanger database (Figure 6). This variability may be due to alternative or permissive processing and may provide new insight into the regulation of miRNA. Further studies are required to understand the biological significance of these miRNA isomiRs.

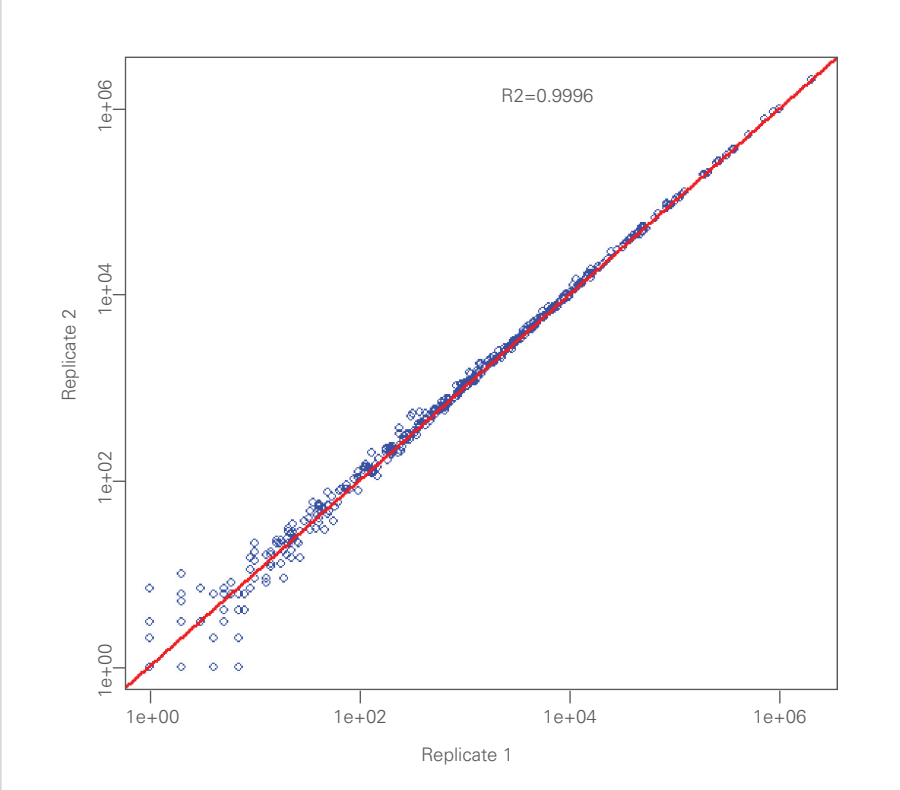

Figure 4: Reproducibility of SOLiD™ Small RNA Expression Method. Two samples from a single placenta library were deposited onto two quadrants of a single slide and analyzed using the SOLiD™ Analyzer. Reproducibility between the two replicate samples was 99.96% for all 423 known miRNAs and their isoforms identified in the tissue. This high degree of reproducibility combined with wide dynamic range allows for accurate detection of subtle changes in expression levels.

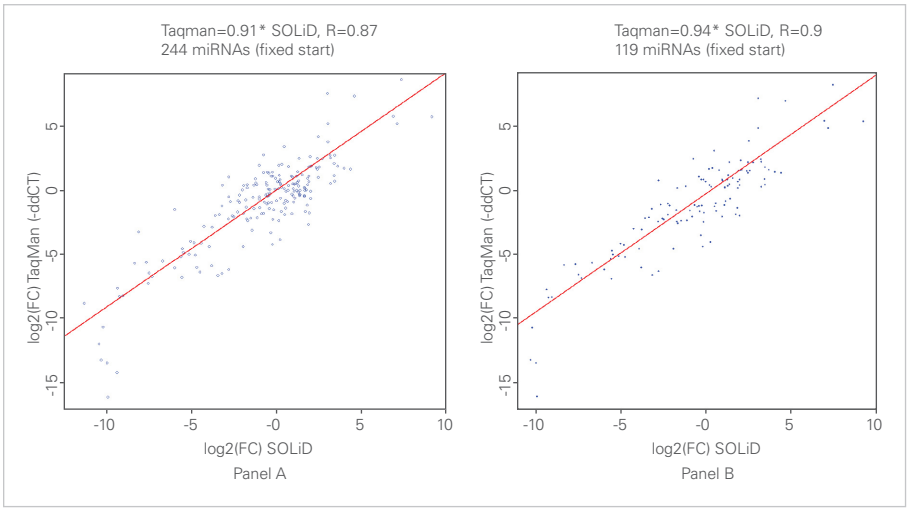

Figure 5: Correlation of SOLID and TaqMan® Results: Data from 244 miRNA, with valid CT measurements and targeted by at least one tag on the SOLiD™ System, were compared (Panel A). A subset of 119 miRNAs demonstrating significant differential expression (p-value of the t-test less than 0.01) between lung and placenta according to TaqMan measurements are represented (Panel B).

Conclusion

A simple and robust workflow for small RNA analysis has been developed for the SOLiD™ System, which

demonstrates significant advantages in sensitivity and dynamic range over traditional approaches to studying whole genome RNA expression. The SOLiD System generates greater than 240M sequence tags per run. The single day procedure to prepare small RNA libraries using the SOLiD™ Small RNA Expression Kit represents a significant improvement over the 4 days required by other published methods, saving researchers time and labor. The flexible slide format used in the SOLiD™ System allows for the deposition of 1-16 samples, enabling the analysis of multiple samples and matched controls in a single run. The expression levels of miRNA molecules were readily determined after sequencing unique reads from miRNAs. This approach was shown to be highly reproducible and demonstrated a dynamic range that is orders of magnitude greater than microarrays. miRNA levels analyzed by the SOLiD™ Small RNA Expression System were confirmed by experiments using TaqMan® MicroRNA Assays. The SOLiD System therefore provides a highly sensitive, hypothesis-neutral method for the detection of novel small transcripts on a genome-wide scale.

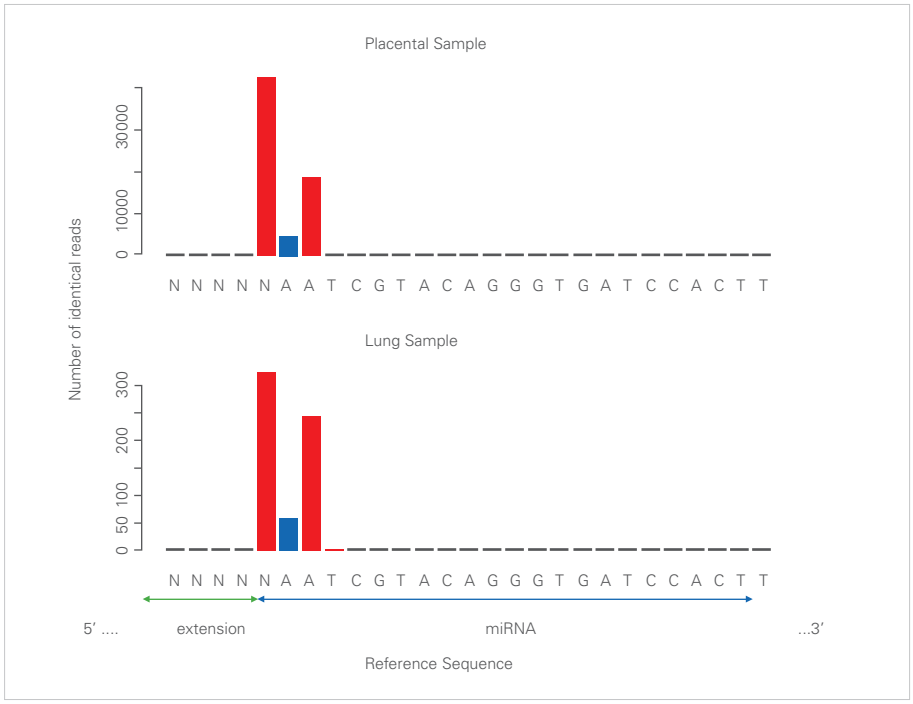

Figure 6: Sequence analysis of known miRNA indicates significant variation at the 5′ end. Reads are plotted according to the position of the 5′ start point. Red bars indicate the number of reads with a 5′ start point that differ from the position previously characterized in the Sanger database.

References

Stephani, G., and Slack, F. J., *Nat Rev Mol Cell Bio.* 3:219-230 (2008).

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