SOLiD™ System Barcoding

Introduction

Genomic studies require large sample sizes to ensure sufficient power to detect specific mutation(s) or alteration(s) associated with a particular trait. Sample multiplexing technologies enable scientists to increase the power of their studies by increasing the number of samples they can cost effectively analyze. The incorporation of sample multiplexing into the SOLiD System protocol provides researchers with the flexibility to efficiently analyze and compare the same targeted regions amongst several biological samples using a simplified workflow. With sample multiplexing, a suite of applications including validation of whole genome association results, rare and somatic mutation detection, expression analysis, and epigenetic studies can be performed on the SOLiD System.

SOLiD™ System Barcodes

SOLiD System barcodes contain unique sequences designed for optimal multiplexing. Sixteen different barcodes were selected based on uniform melting temperature (Tm), low error rate, and orthogonal sequences that are unique in color space. Barcodes are added to the 3′ end of the target sequence using a modified version of the P2 adaptor (Figure 1). SOLiD System barcoding enables the assignment of a unique identifier to templated beads that are made from one individual library. Once these identifiers are assigned, multiple batches of templated beads may be pooled together and sequenced in a single flowcell run (Figure 1). The combination of two sequencing slides

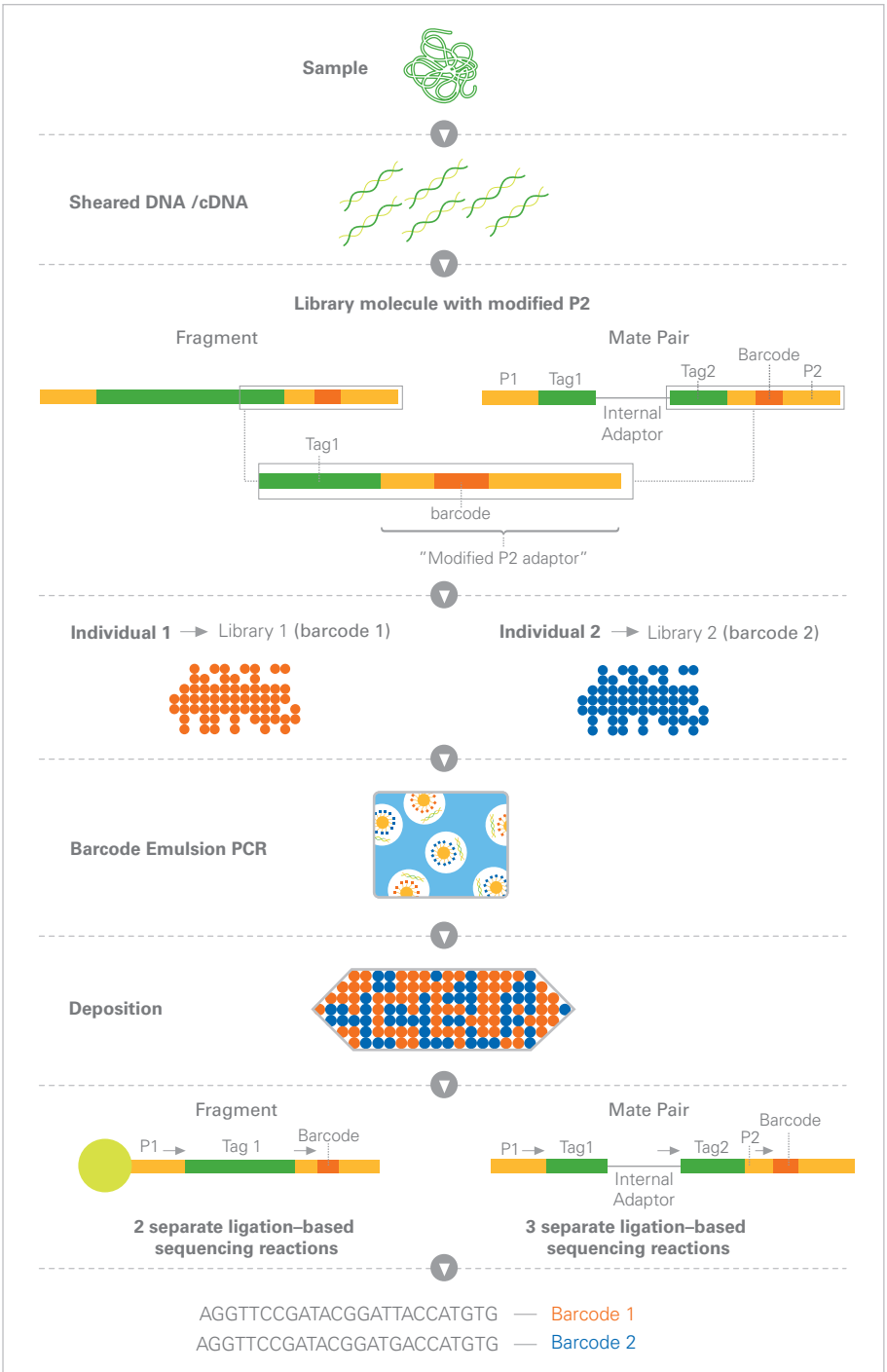

Figure 1. Integration of SOLiD™ System barcodes into the library construction workflow.

with eight segments each and the capability of sixteen different barcodes enables the interrogation of up to 256 samples in a single run (Figure 2). Data analyses can then trace the sequence data back to a specific sample using its respective identifier.

Following sequencing of the target DNA, additional rounds of ligation–based sequencing are performed using primer sets complimentary to the barcode. The resulting reads can then be sorted by the barcode and aligned in groups to the reference sequence.

Robustness and Accuracy of the SOLiD System

SOLiD System barcoding technology utilizes the same ligation-based sequencing methodology as a standard SOLiD sequencing run. Ligation-based sequencing provides a high degree of sequence fidelity. The high accuracy rate achieved reduces the chance that miscalled bases result in the wrong barcode assignment.

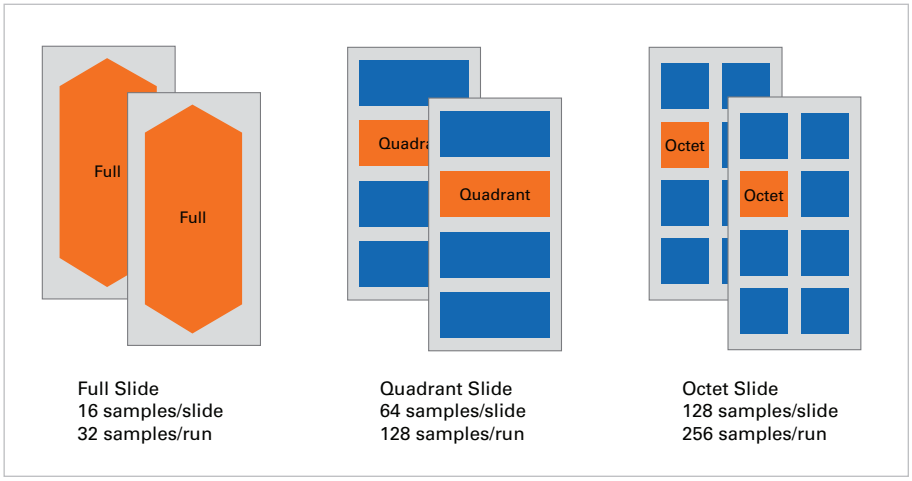

Figure 2. SOLID Barcoding enables the flexibility to process 16- 256 samples per run.

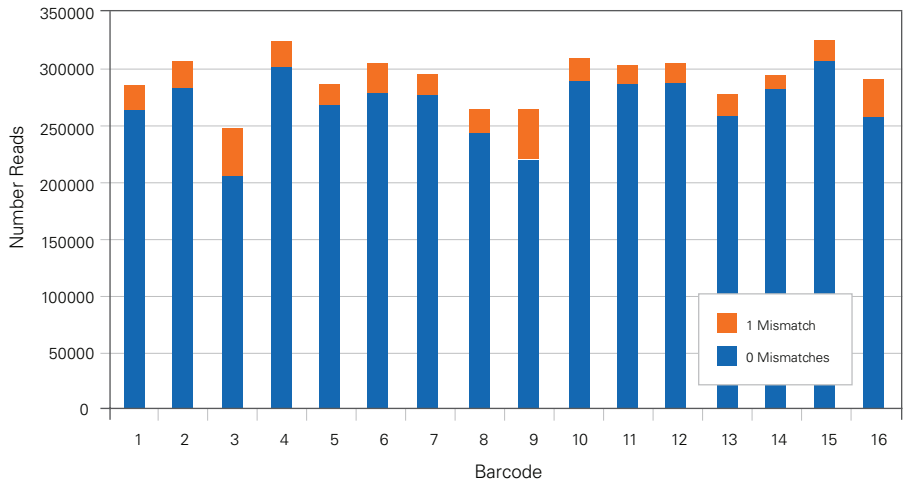

Figure 3. Accurate mapping of high quality reads to barcodes with 0 or 1 mismatches. The proportion of high quality 25 bp reads coupled with barcode sequences containing 0 mismatches (blue bars) or 1 mismatch (orange bars) is noted.

A recent report (Parameswaran et. al, 2007) has suggested that the addition of barcodes in other sequencing technologies may bias the relative representation of amplicons in a pool, resulting in non-uniform coverage. To assess whether this phenomenon occurs in the SOLiD barcoding system, sixteen barcodes were pooled together and used to create a single *E. coli* DH10B library. Multiple rounds of ligation-based sequencing were performed on the library for read lengths of 25 bp for the target DNA and 5 bp for the tagging code. Results demonstrate the uniform representation of each barcode in the pool and that over 99% of all high-quality 25 bp reads were coupled with an accurate barcode (Figure 3). The correlation between high-quality reads and barcodes permits corrections of barcodes with single mismatches, further increasing the rate of association between target and barcode sequences (orange bars, Figure 3).

Barcode Performance

Implementation of barcodes in a small RNA discovery project was performed to demonstrate the robustness and utility of barcodes in a real world application. Two RNA samples (lung and placenta tissue) were processed under various conditions using the SOLiD™ Small RNA Expression Kit to yield fifteen libraries. The resulting libraries were assigned a specific barcode, pooled and sequenced using the SOLiD system. The reads from each of the pooled libraries were then separated based on specific barcode sequences and mapped against the miRNA sequences available in the Sanger miRNA database. The number of unique reads per miRNA was determined and used as the expression level for that particular miRNA. The relatively even distribution of detected miRNAs in each library is an indicator of the success of the barcoding approach in multiplexing several samples along with the ability

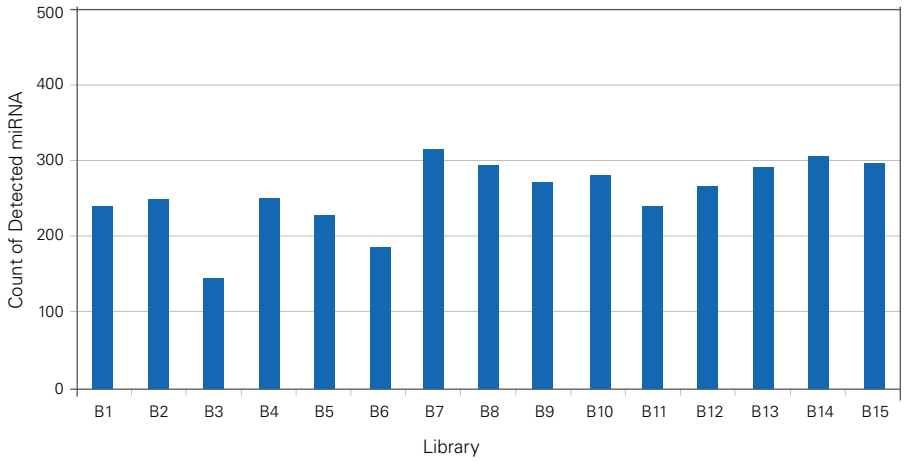

Figure 4. Fifteen samples were processed via the SOLiD™ Small RNA Expression Kit, assigned a specific barcode, pooled and sequenced using the SOLiD system. The reads from each of the pooled libraries were then separated based on their barcode sequence and mapped against the RefSeq collection (NCBI)

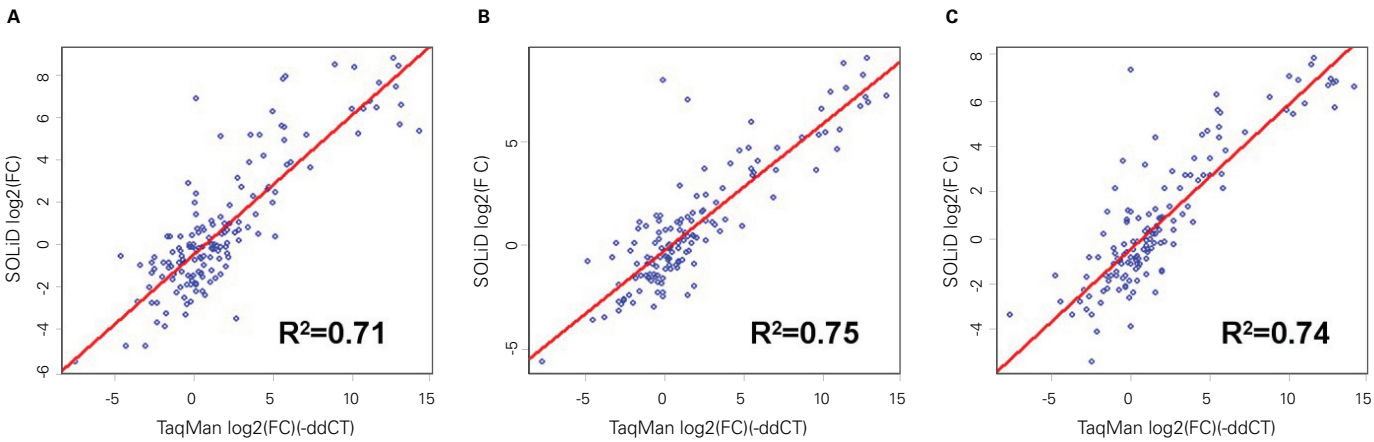

Figure 5. Fold change (Placenta vs. Lung) correlation of three small RNA expression libraries analyzed with the SOLiD™ System and TaqMan® Gene Expression Assays.

to segregate the data resulting from each individual library (Figure 4).

In addition, fold change correlations for three of the libraries (placenta vs lung), were done with 210 TaqMan miRNA assays. The high level of correlation with TaqMan assays and the degree of reproducibility of the correlations between the three libraries is a strong indication of the ability to mix several bar coded samples and extract expression data for each of the pooled samples (Fig. 5; A, B, and C).

Versatility of SOLiD Barcoding

The SOLiD System barcoding technology maximizes throughput for a wide range of resequencing and tag-based applications (Table 1). Not only is sample multiplexing applicable to rare-variant detection amongst hundreds of samples, but also to time course-based or drug-induced differential gene expression analyses. In addition, for complex genome sequencing or for structural variation studies, multiplexing of different sized mate-paired libraries allows complete assembly and uniform coverage of the highly complex genomes. Furthermore, the SOLiD sample multiplexing technology offers a controlled experimental environment for concurrent processing and sequencing of a diverse set of samples.

Conclusion

The SOLiD barcoding technology dramatically increases the throughput and scalability of the SOLiD System. Moreover, this key attribute considerably simplifies the sequencing workflow and drastically reduces costs by pooling multiple library preparations into a single emulsion PCR and sequencing reaction. Future advances of this feature include expansion of the number of barcodes, resulting in an increase in the number of libraries sequenced, as well as application of barcodes to mate-paired libraries.

Reference

1 Parameswaran, P, et al. *Nucleic Acid Res* (2007) 35:e130.

TABLE 1. Flexibility of barcoding to high-throughput sequencing applications.

Application	Functionality of Barcoding
Resequencing of highly complex genomes	Analysis of different sized mate-paired libraries of the same genomic sample
Structural variation analysis	Analysis of different sized mate-paired libraries of the same genomic sample
Targeted resequencing for SNP detection	Assignment of rare SNPs to an individual in a sample population(s)
Mutation analysis for complex or Mendelian disorders	Screening of multiple exons within genes for a single patient sample or screening of a genomic region in multiple patient samples
Gene expression	Examination of differentially expressed genes within a time course series or between drug treatments

Printed in the USA, 04/2008, Publication 139AP05-01

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