SureSelect™ Solution-Based Enrichment of Genomic Regions for Targeted Resequencing on the SOLiD™ System

Introduction

The identification of genetic variants and mutations associated with human disease requires the development of a robust and cost-effective approach for systematic resequencing of candidate regions in the human genome. When combined with a solution-based hybridization enrichment approach, the industry-leading throughput of the SOLiD™ System facilitates deep sequencing of target genomic regions of interest. The method employed by the Agilent SureSelect™ Target Enrichment System extracts target regions from genomic libraries by hybridization to in-solution biotinylated cRNA probes, or "baits". Post-enrichment material is amplified and used directly for downstream steps, including emulsion PCR (ePCR) and sequencing on the SOLiD™ System (Figure 1). The inherent scalability and flexibility for automation of the SureSelect™ in-solution enrichment system coupled with the ultra-high throughput of the SOLiD™ sequencing platform provides an integrated approach to targeted resequencing.

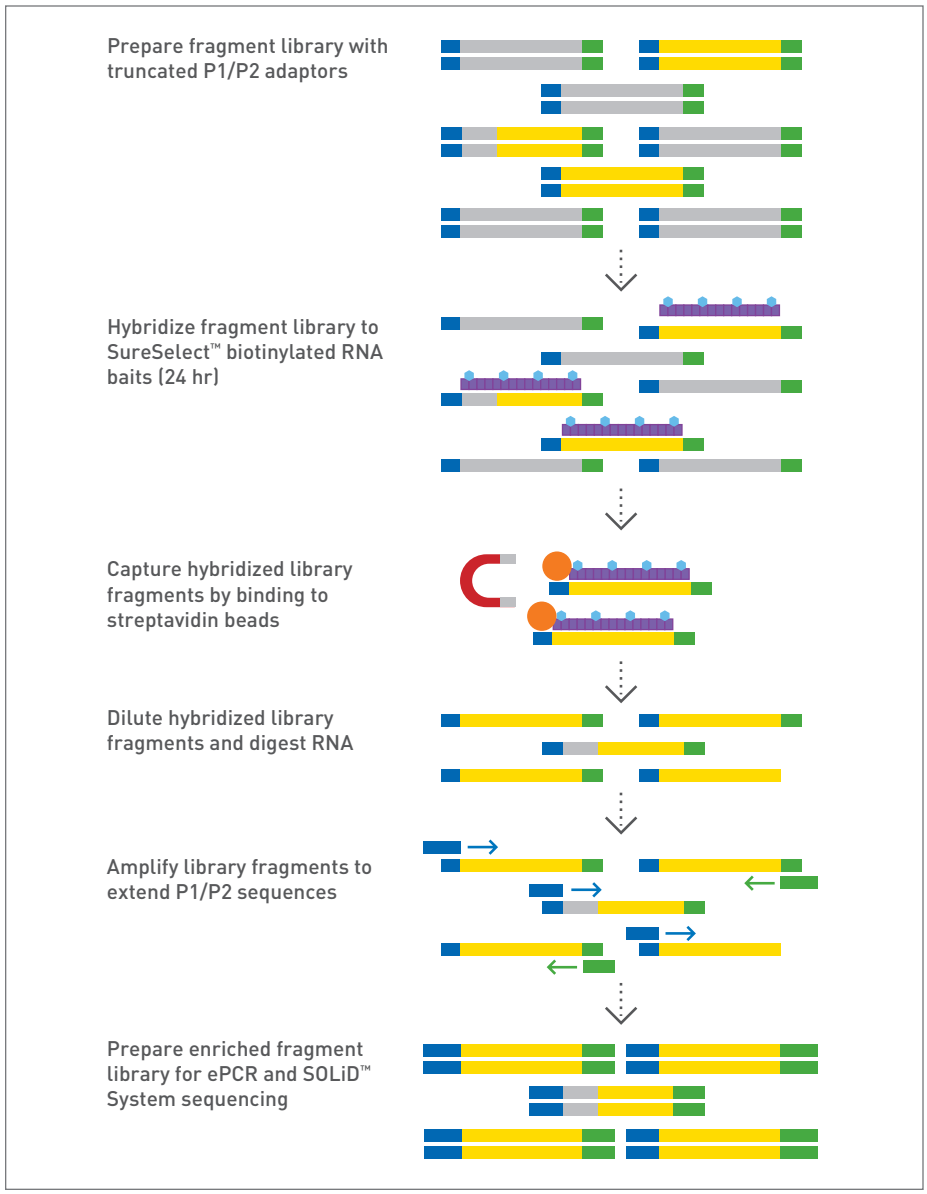

Figure 1. Workflow for SureSelect™ In-Solution Enrichment of Target Genomic Regions Prior to Sequencing on the SOLiD™ System.

Methods

Library Construction

Fragment libraries were constructed from 3 µg of human genomic DNA (NA15510 and NA18507) using a modified protocol of the Applied Biosystems® SOLiD™ 3 System Library Preparation Guide. Truncated adaptors for library construction were utilized to minimize nonspecific capture during SureSelect™ in-solution hybridization. PCR primers and cycling conditions for genomic library amplification were correspondingly adjusted.

Solution-Based Enrichment

With the Agilent SureSelect™ Target Enrichment System, sets of in-solution baits were custom-designed using the eArray tool (http://www.opengenomics.com/earray) for capture of noncontiguous target regions ranging from a total of 0.2 to 3.1 Mb in length. Genomic libraries were hybridized to SureSelect™ biotinylated RNA baits for 24 hours. Hybridized library fragments were isolated by magnetic capture using streptavidin beads and purified according to the Agilent SureSelect™ protocol. The enriched fragments were subjected to another round of PCR amplification to incorporate standard full-length SOLiD™ adaptor sequences. Unenriched control libraries were also generated to determine enrichment efficiency.

SOLiD™ Sequencing

Templated beads derived from either SureSelect™-enriched libraries or unenriched libraries were prepared according to the Applied Biosystems® SOLiD™ 3 System Templated Bead Preparation Guide. Each sample was deposited on a quadrant of the slide at bead densities ranging from 130K to 200K beads per panel. Sequencing by ligation was carried out on the SOLiD™ 3 Analyzer in accordance with the Applied Biosystems® SOLiD™ 3 System Instrument Operation Guide.

Data Analysis

50 bp reads from each library were analyzed using the SOLiD™ Accuracy Enhancer Tool (http://info.appliedbiosystems.com/solidsoftwarecommunity) and then aligned against the hg18 human genome reference sequence, allowing up to 4 mismatches per read. For the enriched and unenriched samples, coverage across the genome was visualized using the UCSC Genome Browser. SNP detection was performed against the reads aligning to targeted sequences using the SOLiD™ BioScope™ v1.0 Resequencing Pipeline with modified parameters. Enrichment performance was evaluated by calculating the proportion of uniquely mapped reads aligning to the targeted sequences for each sample. To assess enrichment performance of the unenriched samples, noncontiguous target regions of a total of either 2.7 Mb or 3.1 Mb in length were used for calculations.

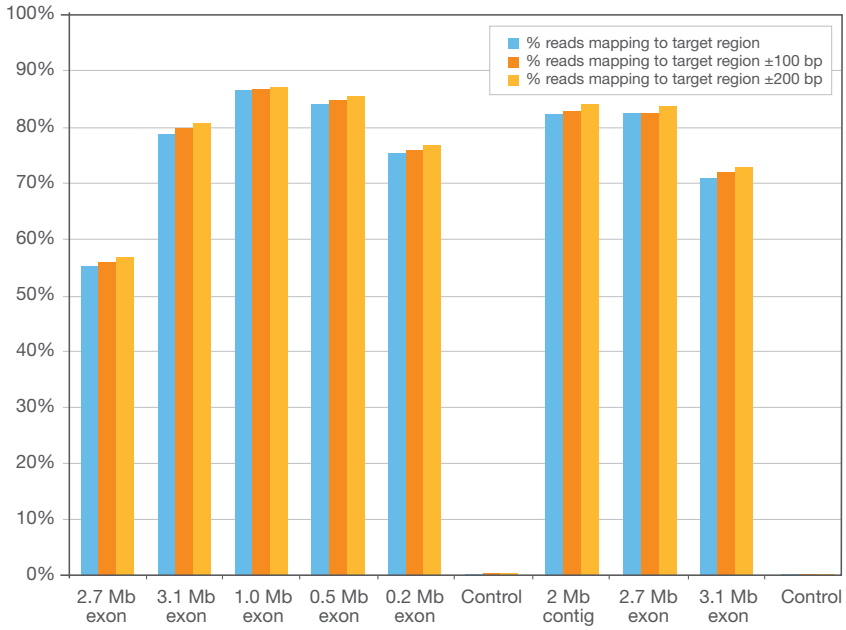

Figure 2. Enrichment Performance Among a Diverse Set of SureSelect™ Target Enrichment Designs. The blue bars indicate the percentage of uniquely mapped SOLiD™ reads aligning to the target region. The other bars indicate the percentage of SOLiD™ reads aligning to the target region, as well as any region within 100 bases (orange) or 200 bases (yellow) of the target region. The designated target region for analysis of the unenriched control was the 3.1 Mb exon design.

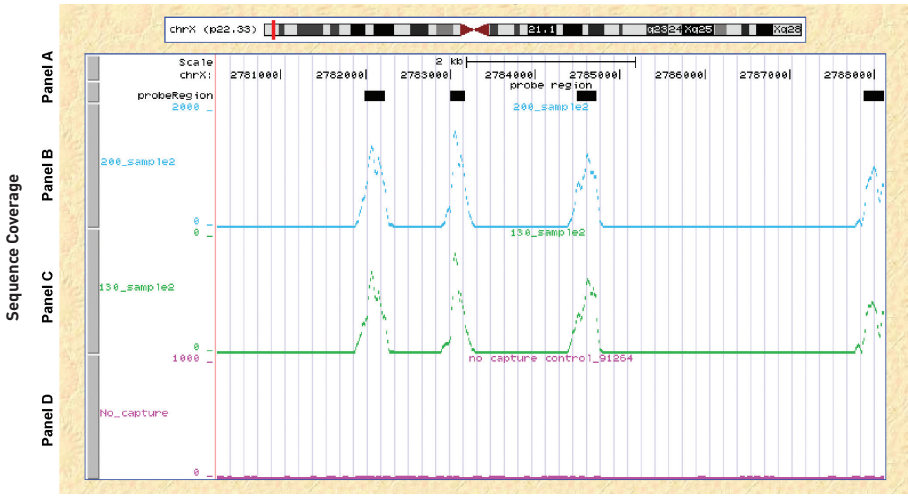

Figure 3. Sequence Coverage Generated by the SOLiD™ System for a Sample Region of the Human X Chromosome. Panel A highlights the target chromosomal location of the probes contained in the SureSelect™ Oligo Capture Library. The number of uniquely mapped tags across the sample chromosomal region for enriched samples (Panels B and C) and the unenriched sample (Panel D) is shown. The difference in coverage levels between the two enriched samples in the targeted regions corresponds to differences in bead deposition density (200K beads/panel for Panel B and 130K beads/panel for Panel C).

Results

Highly specific enrichment of the target genomic regions was demonstrated with an average of over 2,500-fold enrichment among the enriched samples. Optimal enrichment performance was achieved by using the truncated library adaptors during library construction, as shorter adaptor sequences decreased the melting temperature of the library molecules to well below the hybridization incubation temperature, thus

preventing the cross-hybridization of single-stranded library strands and nonspecific capture.

Of all the reads mapping uniquely to the human genome, the percentage of "on-target" reads ranged from 50 to 85% (Figure 2). For the unenriched controls, fewer than 0.1% of the total number of mapped reads were specific to the target sequences. Coverage profile characteristics such as peak amplitude are governed by probe performance and sequence content for a single cohort of targeted loci, but are reproducible when the same cohort of targeted loci is independently interrogated with the same probe set.

To assess the reproducibility of the SureSelect™ solution-based system, we compared the results from independent technical replicates targeting regions of the human X chromosome (Figure 3). Specifically, two independent hybrid selections, prepared from the same source DNA and using the same SureSelect™ Oligo Capture Library, were subjected to SOLiD™ System sequencing. Base-by-base coverage profiles for the two samples (Figure 3, Panels B and C) demonstrated remarkable similarity. Comparison of these two profiles (Figure 3, Panels B and C) with that of the unenriched sample (Figure 3, Panel D) clearly indicates specific enrichment of the targeted regions (Figure 3, Panel A) using SureSelect™ methodology. Differences in the amplitude of sequence coverage between the replicates for the targeted regions reflect the differences in bead deposition density between the two samples rather than in enrichment efficiency. As shown in Figure 3, fold coverage of the target region increases linearly as a function of bead density. The nearly identical coverage profiles of two independently enriched samples run on two different instruments illustrate the propensity of the integrated workflow for robust, reproducible, and specific enrichment.

For accurate detection of genetic variants, the extent of coverage for the target regions was assessed for all of the enrichment samples. For these samples, 92% or more of the target bases were covered by at least 1 read, while 89% or more of the target bases were covered by at least 20 reads (Figure 4). Thus, the specificity and sensitivity of the SureSelect™ System and the throughput of the SOLiD™ System are ideal for detection of genetic variants.

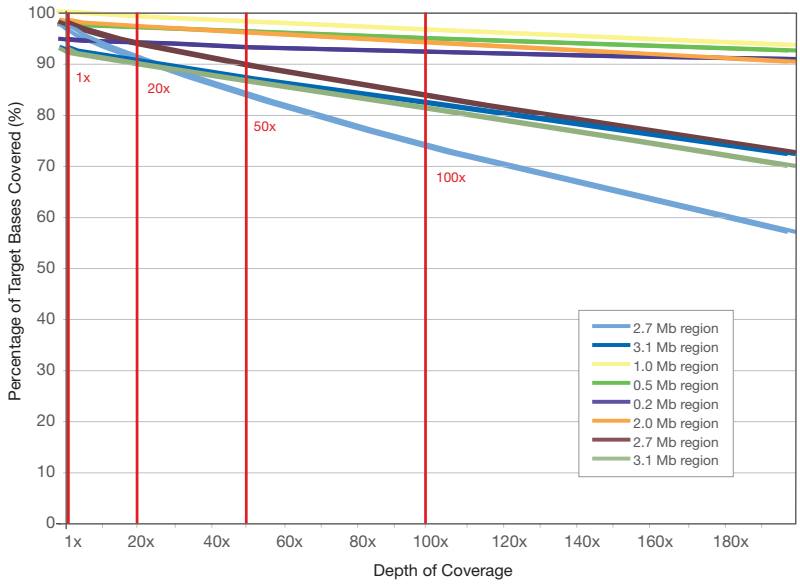

Figure 4. Percentage of Bases in the Targeted Regions Versus Depth of Coverage for All Enrichment Samples. Red vertical lines indicate read depth of targeted regions at 1x, 20x, 50x, and 100x coverage.

Table 1. SNP Calls for Enriched Samples Sequenced by the SOLiD™ System.

SNP Classification	Number of Calls
Heterozygous true positives	1,038
Homozygous true positives	5,500
Heterozygous false positives	15
Homozygous false positives	15
Heterozygous undercalls	55
Heterozygous uncalled	13
Homozygous uncalled	69

Table 2. SNP Discovery Statistics in Enriched Samples.

	Homozygous SNPs	Heterozygous SNPs
Total number of SNPs identified in enriched regions	1,650	2,465
Novel SNPs	134	785
Specificity*◊ (including undercalls)	93.8%	99.7%
Specificity*◊ (excluding undercalls)	93.7%	99.7%
Sensitivity*‡ (including undercalls)	99.7%	93.9%
Sensitivity*‡ (excluding undercalls)	99.7%	95.0%
Concordance with dbSNP** in enriched regions	98.7%	98.9%

*** Reported values are based on comparisons to the following version of HapMap: http://hapmap.ncbi.nlm.nih.gov/genotypes/latest/forward/non-redundant/**
‡ Sensitivity = (True Positives)/(True Positives + False Negatives)
◊ Specificity = (True Negatives)/(True Negatives + False Positives)
**** Values based on comparison to dbSNP 129**

In order to determine sensitivity, specificity, and concordance of SNP detection, SNPs identified in this study were compared to genotypes in the HapMap database for the same sample. Classification of SNPs is outlined in Table 1. A total number of 1,650 homozygous SNPs and 2,465 heterozygous SNPs were identified in the targeted regions by the SOLiD™ System. Reported sensitivity and specificity values (Table 2) take into account the number of undercalls, or true heterozygous SNPs called as homozygous SNPs. In general, specificity and sensitivity values increase with additional sequence coverage.

Conclusion

The SOLiD™ System and the SureSelect™ Target Enrichment System provide a powerful targeted resequencing solution. The scalability, convenience, and reproducibility of the SureSelect™ solution-phase enrichment method, combined with the throughput and accuracy of the SOLiD™ System, enable researchers to perform deep sequencing of specific regions of interest for rare variant discovery, to better understand areas such as tumorigenesis, population diversity, microbial resistance, and disease susceptibility.

Printed in the USA. 02/2010 Publication 139AP19-01. CO11539 0210

applied biosystems™
part of life technologies™

Headquarters
850 Lincoln Centre Drive | Foster City, CA 94404 USA
Phone 650.638.5800 | Toll Free 800.327.3002
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