Array-based Enrichment of Specific Genomic Regions For Applications Using the SOLiD™ System

Introduction

The inherent scalability and ultra high throughput of the SOLiD™ System enables researchers to conduct a wide range of experiments in a more timely and cost effective manner than previously possible. As the cost per sample and time to generate sequence information continues to decrease, whole genome resequencing of complex organisms such as human will soon become routine. Many studies, however, will continue to focus on specific candidate genes or genomic regions and the development of efficient methods for large scale targeted enrichment is critical.

A number of techniques exist to permit the selection of specific target regions including traditional PCR amplification of short regions, long range PCR (LR-PCR) of regions up to 10 KB in length, and more recently array-based methods of enrichment. With the array-based approach, samples are hybridized to custom oligonucleotide arrays with probes designed to match the specific DNA regions of interest[1]. The selection of an appropriate enrichment method is highly dependent on the number of samples and sequence content to be interrogated in a particular project. In studies looking at large numbers of discreet regions across many individuals, array-based approaches are attractive because they are highly multiplexed and enable the interrogation of many

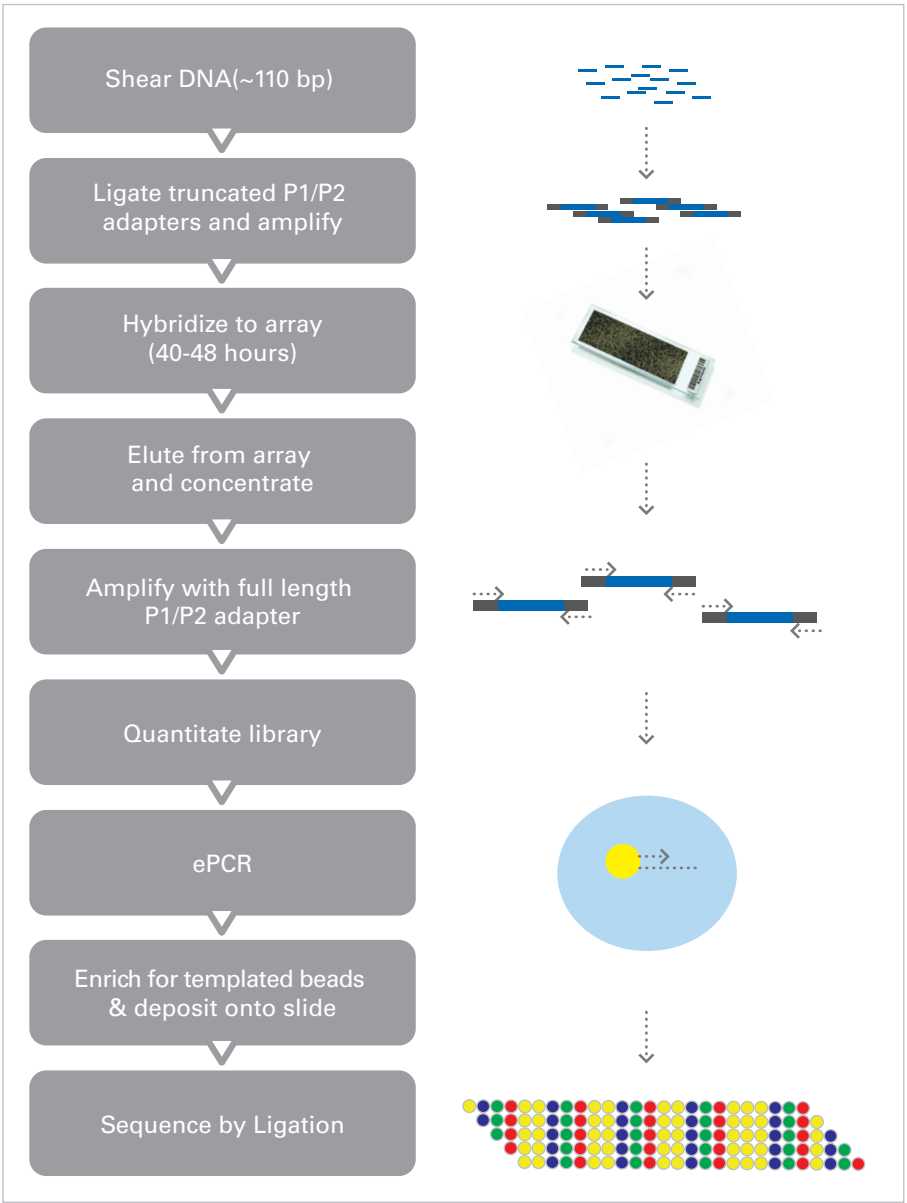

Figure 1: Workflow for array-based enrichment of targeted genomic regions prior to sequencing with the SOLiD™ System

loci. This application note will describe how commercially available custom microarrays have been used to select specific regions of the human genome for sequencing on the SOLiD™ System.

Methods

Array Design

An Agilent custom oligonucleotide array was designed with probes specific to a 4.3 MB region of interest. Repetitive elements, as defined by RepeatMasker at default settings for hg17 genome, were removed from the target sequence and not included in the array design. Probes were designed to maximize specificity to the target sequence. For more information on this method, please contact your Applied Biosystems Applications Specialist.

Library Construction

A fragment library was constructed from human DNA (NA18507) using the protocol in the Applied Biosystems SOLiD™ System 2.0 User Guide, with minor modifications. Protocol modifications included truncating the library adaptors and PCR primers used to achieve melting temperatures (T_ms) compatible with the hybridization conditions suggested by the array manufacturer and lowering the annealing temperature used during amplification. The resulting library was subjected to a limited number of cycles of PCR amplification to provide 30 ug of DNA.

Array Enrichment

A portion of the fragment library was hybridized to the array using a modified protocol. Hybridized fragments were eluted from the array, concentrated by precipitation, and PCR-amplified to incorporate the standard full length PCR primers used in the SOLiD™ System protocol. The final amount of DNA was determined using an Agilent 2100 Bioanalyzer.

SOLiD Sequencing

The final library was attached to beads via hybridization to the P1 adaptor. After emulsion PCR, the sample was enriched for templated beads and deposited onto a glass slide. One full slide was deposited with beads from the enriched library and a separate slide was deposited with beads made from an un-enriched library as a control. Sequencing by ligation was carried out on the SOLiD Analyzer™ according to standard protocols described in the SOLiD™ System 2.0 User Guide.

Data Analysis

The 35 base pair reads produced from each library were aligned against the human genome sequence (NCBI, b36, hg18) with up to 3 mismatches. These reads were also aligned to the 4.3 MB target sequence with up to 3 mismatches and *de novo* SNP detection was performed against this alignment. The positions of the target sequence, aligned reads and identified SNPs were then mapped to positions in the whole human genome (hg18).

Results

The coverage and matching statistics for the enriched sample are summarized in Table 1. Approximately 50% of the reads which mapped to the human genome were specific to the target sequence. Excellent coverage was achieved with 91% of the region covered by at least one read. Target DNA was 391 fold higher in the enriched sample compared to the un-enriched control.

TABLE 1: COVERAGE STATISTICS FOR ENRICHED SAMPLE

Reads Uniquely Mapping to Human Genome	1.083 GB
Reads Uniquely Mapping to Target Regions	0.585 GB
Enrichment	391
Percent Target Covered	91%
Average Target Coverage	138

Reads uniquely mapping to the Human Genome = Reads were aligned to the human genome sequence (NCBI, b36, hg18) with up to 3 mismatches.
Reads uniquely mapping to target regions = Reads were aligned to the target sequence with up to 3 mismatches.
Enrichment = (Sequence Reads Uniquely Mapping to Target Regions / Sequence Reads Uniquely Mapping to Human Genome) * Maximum Enrichment, where Maximum Enrichment is defined as a ratio of genome length vs. target length and is equal to 722 for this 4.3 MB enrichment target.
Percent Target Covered = Percent of target sequence containing at least one sequence read.
Average Target Coverage = Number of 35bp reads matching target * 35 / target length.

Successful target enrichment by array hybridization is shown in Figure 2. The placement and sequence coverage of uniquely matching reads generated for the enriched sample (Panels A and C) and control sample (Panels B and D) are shown for a region of Human Chromosome 22. Comparison of the number of uniquely mapped reads in the enriched and control samples clearly indicates significant enrichment of specific regions of the genome (Panels A and B). The per base sequence coverage for the enriched sample (Panel C) exceeds 100x (reaching 1500x) while staying under 15x in the control sample (Panel D). Comparison of the chromosomal location of the enriched reads to the probe targets (Panel E) clearly establishes that the vast majority of reads map to regions which overlap array hybridization probes.

It is also important to ensure that the enriched sample maintains an accurate representation of the SNP profile for that individual. The SNPs detected in this study were compared to genotypes in the HapMap database for the same sample (NA18507), 99.9% of the homozygous and 96% of the heterozygous SNPs were correctly identified (data not shown).

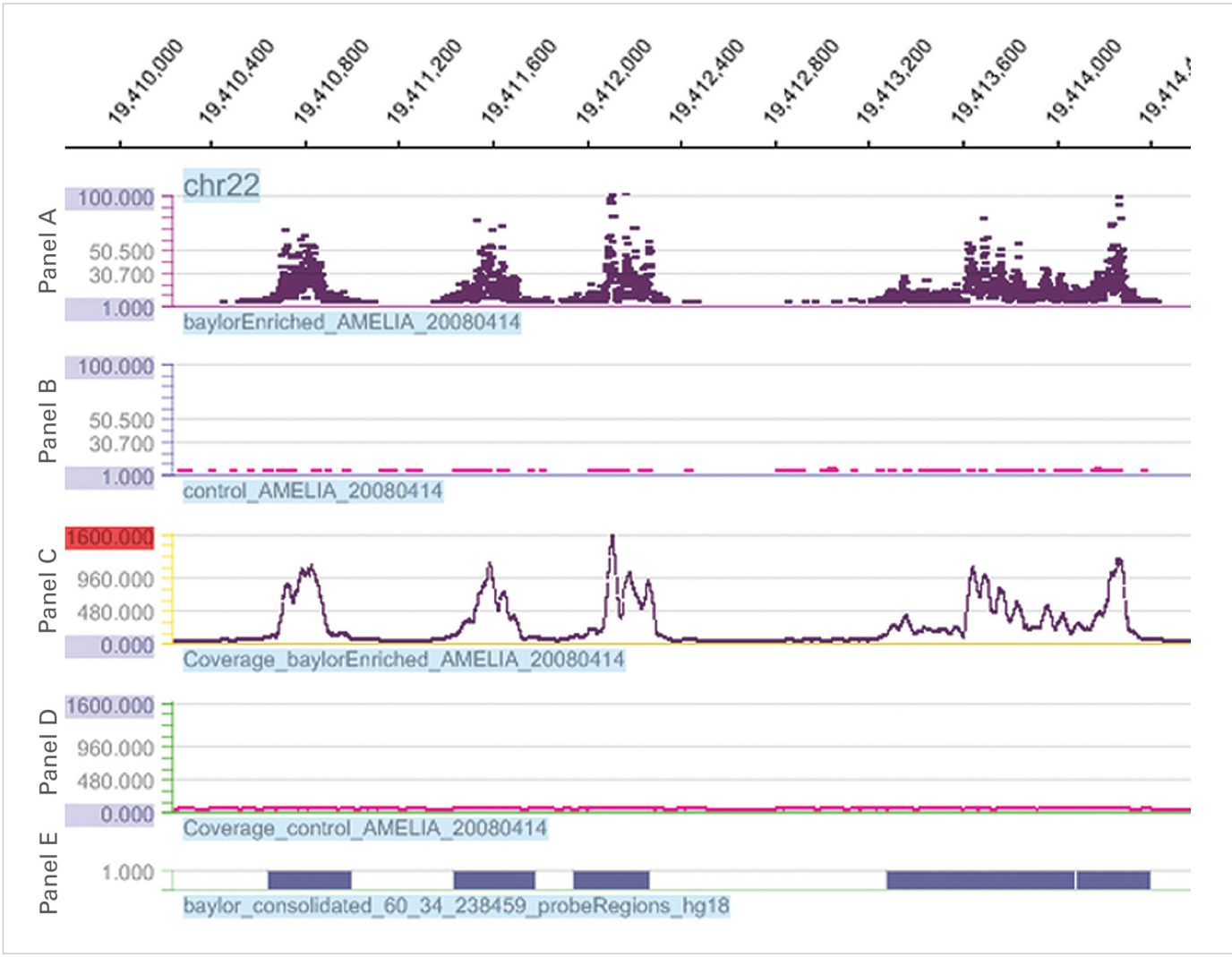

Figure 2: Sequence coverage and placement of uniquely matching sequence tags generated by the SOLiD™ System for a sample region of Human Chromosome 22. Panels A (enriched sample) and B (control sample) show the read placement with the X axis indicating the chromosomal position of mapped reads and the Y axis representing how many times the particular read is sequenced. Panels C (enriched sample) and D (control sample) indicate the coverage per base position across the same chromosomal region. Panel E indicates the location of the probes contained on the oligonucleotide array. Panel A and Panel E can be used to compare the read placement for the enriched sample to the probe regions.

Conclusion

Array hybridization is a viable approach for enrichment of genomic regions for deep sequencing studies on the SOLiD™ System. Preliminary studies, using a commercially available array, demonstrated approximately 390X enrichment of target fragments relative to un-enriched controls. Enrichment was specific to regions targeted by the array probes and the SNP profile was maintained.

Discussion

While microarrays for this application are commercially available, the arrays and the protocols for using them were developed for other applications such as SNP detection, copy number variation or gene expression. These existing protocols need to be modified to use the arrays as DNA capture platforms.

Specific modifications used in this study included:

- Truncating library adapters and primers
- Lowering annealing temperature in library PCR
- Using post-array amplification to full length adaptors before ePCR
- Using the MAUI® Hybridization System instead of Manufacturer's recommended Tecan System
- Increasing DNA input for array hybridization — original input DNA optimized for expression and CGH applications
- Adding an elution protocol to capture hybridized DNA
- Designing arrays using internal probe design criteria. Please contact your Applied Biosystems Applications Specialist for details

References

1. Okou, D.T., et al., *Nat Methods*. 2007 Nov;4(11):907-9. E-pub 2007 Oct 14.

2. Agilent Oligonucleotide Array-Based CGH for Genomic DNA Analysis CGH protocol v4.0 – Agilent Technologies, June 2006

Printed in the USA, 06/2008 Publication 139AP12-01

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