Targeted Resequencing

The combination of Illumina's Genome Analyzer and Agilent's SureSelect Target Enrichment System offers a flexible and reliable solution for targeted resequencing.

Highlights

- **Streamlined Workflow:**
  Easy, automatable sequencing of hundreds of regions, across thousands of samples
- **Flexible Kits:**
  Customizable for any region of interest
- **Superior Data Quality:**
  Uniform coverage, high reproducibility, and high specificity
- **Unbiased Allele Representation:**
  Supports SNP-calling resequencing applications

Introduction

Despite great efforts, the scientific community still faces tremendous challenges in understanding the genetics behind complex diseases such as cancer. Systematic detection of common and rare variants through targeted resequencing of many samples will likely hold the key to advancing our knowledge of the basis of disease and the development of treatments.

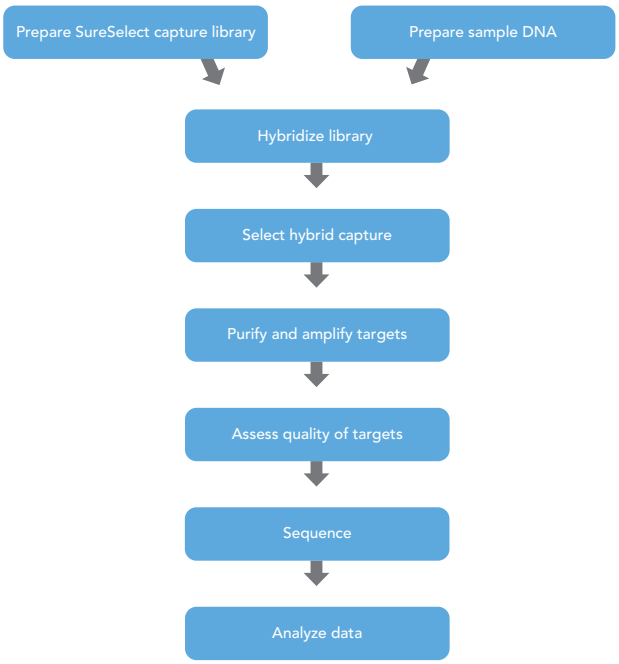
Historically, targeted resequencing experiments have been limited by the high costs of Sanger sequencing and lack of a scalable, efficient method for partitioning genomic regions of interest. The availability of Illumina's cost-effective, high-throughput Genome Analyzer sequencing platform and Agilent Technologies' SureSelect Target Enrichment System brings targeted resequencing applications never considered possible within reach of even small labs.

Efficient Targeting Approach

The Genome Analyzer is the most widely adopted and easiest to use next-generation sequencing platform. Superior data quality and a wide range of read lengths have made it the system of choice for whole-genome *de novo* sequencing and resequencing. In addition to these applications, the ever-increasing output of the Genome Analyzer can be harnessed to analyze a select region of interest in the genome, but in many more samples.

To leverage the advantages of high-throughput sequencing for focused genomic analysis, Agilent Technologies developed the SureSelect Target Enrichment System. Taking advantage of Agilent's probe design expertise and SurePrint technology for synthesizing long, high-fidelity custom oligonucleotides, SureSelect provides a highly flexible, automatable, and reliable hybridization-based method that enables researchers to isolate up to 5 Mb of genomic regions per sample for high-throughput sequencing. Unlike other target enrichment approaches, SureSelect is optimized to work with the long reads produced on the Genome Analyzer.

Figure 1: Streamlined Targeted resequencing Workflow



Low Input Requirements

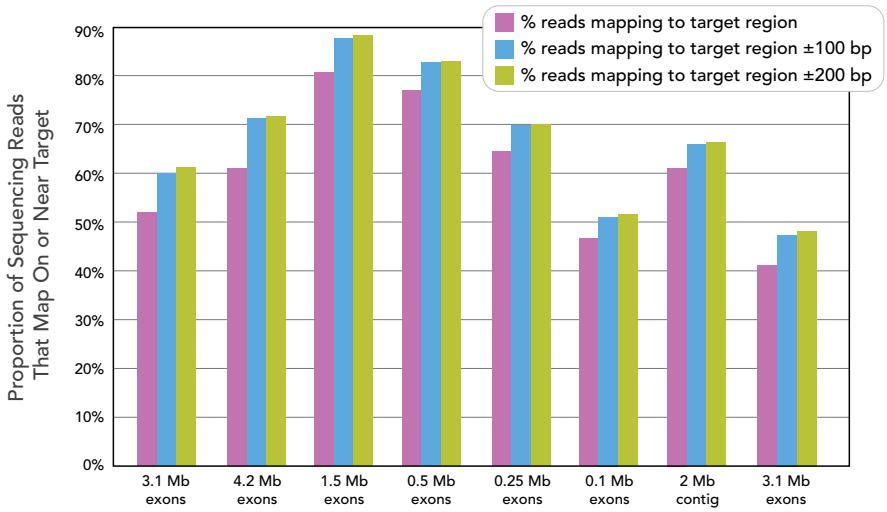
Efficient sample use is essential to the success of many projects, particularly those involving material available in limited supply such as cancer studies. Extending the use of these minimal sample amounts, the Genome Analyzer and SureSelect Target Enrichment System both require very little input DNA. This is accomplished through SureSelect's solution-phase capture technique that takes place in a very small reaction volume. Using long RNA oligos, SureSelect uniquely harnesses attributes important for higher and more efficient capture capacity as compared to other methods, driving the initial sample input requirement to a minimum. Likewise, the Genome Analyzer sample preparation protocol requires DNA input amounts as low as 100 ng.

Streamlined Workflow

From probe design to data generation, targeted resequencing is easily accomplished through a simple, streamlined workflow using the SureSelect Target Enrichment System and the Genome Analyzer (Figure 1).

Custom SureSelect Kits are produced using Agilent's established web-based design tool, eArray (www.opengenomics.com/earray). An intuitive user interface with design wizards walks researchers through the creation of a target enrichment kit. eArray's design algorithms

Figure 2: High Target Specificity



Optimized design algorithms and protocol limit the capture of near target sequences. This graph shows the percentage of sequence reads that map to targeted regions. Each set of three values represents the percent of reads that are: on target (purple), on target plus within 100 bp of target (blue), and on target plus within 200 bp of target (green).

work over a diverse set of genomic locations, including small and large exons, short and long contiguous genomic targets, genome targets within repeats, and non-coding DNA.

Each custom target enrichment kit contains a mixture of custom SureSelect RNA oligonucleotides, or baits, that are biotinylated for easy capture using streptavidin-conjugated magnetic beads. RNA baits are hybridized with a DNA library prepared using Illumina's Genomic DNA Sample Prep Kit. RNA-bait-DNA hybrids are then isolated from the complex mixture using streptavidin-conjugated magnetic beads. The RNA bait is then digested, leaving the targeted DNA to be sequenced. The solution-phase protocol enables automation and minimizes experimental variability.

After a few cycles of DNA amplification, the enriched library is loaded onto an Illumina flow cell and undergoes standard clonal amplification on the Illumina Cluster Station. After clonal clusters are generated, the flow cell is transferred to the Genome Analyzer for massively parallel sequencing by synthesis. The Genome Analyzer system's simple workflow takes approximately four days to go from sample prep to data output for 36 bp reads.

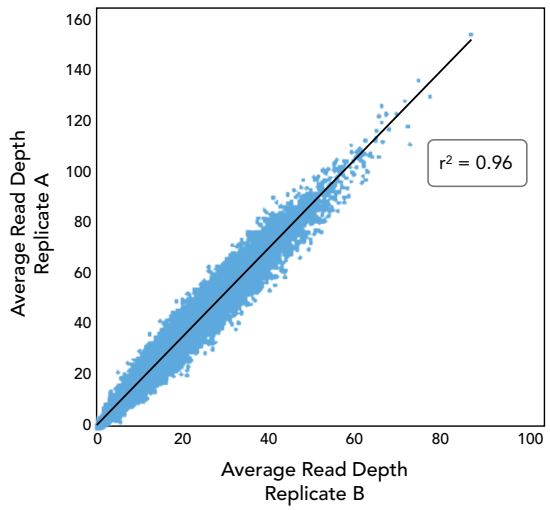
High-Quality Data

Internal and external validation ex-periments have demonstrated the uniform coverage, high reproducibility, and little or no allele bias provided by the SureSelect enrichment method. Combined with the Genome Analyzer system's high sequence accuracy and 75+ bp read lengths, researchers can generate superior data quality for a wide range of targeted resequencing applications.

Enrichment is highly specific for target regions. As shown by several experiments in Figure 2, most reads map directly to target regions. Some additional reads map within 100 bp of the target, but the off-target capture rate is not substantially increased by including reads within 200 bp of the target.

An experiment targeting 3.3 Mb of exonic regions resulted in approximately 80% of the bases having at least 20× coverage and more than 96% with 6× coverage or greater, demonstrating uniform coverage. It is important to note that these results were from a seven million

Figure 3: High Reproducibility



Read depth across target intervals correlate strongly between technical replicate captures. Replicate captures were performed on a SureSelect Target Enrichment library designed for exonic targets covering 3.3 Mb with 50% probe overlap.

read experiment. With the current throughput of the Illumina Genome Analyzer, the percent of bases with 20× coverage would be substantially higher.

Technical replicates show that the SureSelect method is robust and reproducible. The read depth for corresponding genomic locations between experiments shows strong correlation (Figure 3).

The ability to confidently call heterozygous or mutant SNPs is essential for many targeted resequencing applications, such as the characterization of cancerous mutations in regions of interest. With oligo baits of 120 bp, this capture approach is robust enough to detect variations in target regions, supporting SNP calling applications. Data from a sample with more than 20,000 known heterozygous SNPs show that coverage across both alleles is similar, demonstrating that Agilent's target method has little or no allele bias (data not shown). Coupled with the industry-leading accuracy of the Genome Analyzer, researchers can be assured of generating the highest data quality. With 2×75 bp paired-end sequencing, the Genome Analyzer consistently generates 12–15 Gb of mappable data, with more than 70% of base calls having Q30 or greater quality scores*.

Data Analysis

Sequence data generated from samples partitioned by SureSelect samples are analyzed using Illumina's standard Genome Analyzer Analysis Pipeline, CASAVA (Consensus Assessment of Sequence and Variation), and GenomeStudio™ software applications. Data from several sequencing runs can be aggregated to generate consensus variation calls with CASAVA. GenomeStudio software's Integrated Genome Browser allows users to visually examine the on-target and off-target coverage in a sample. In addition, GenomeStudio software provides the ability to export a table of SNPs for further investigation.

Summary

Whether the goal is to sequence genes or regions in very large populations, follow up on genomic regions of interest identified in genome-wide association studies (GWAS), focus on genes involved in certain pathways, identify signatures associated with disease prevalence, or discover rare variants, combining Agilent's front-end SureSelect Target Enrichment System with Illumina's Genome Analyzer offers a flexible and efficient solution for targeted resequencing.

Ordering Information

Products Available From Illumina†	Quantity	Catalog No.
Genome Analyzer	1	SY-301-1201
Cluster Station	1	SY-301-2001
Standard Genomic DNA Sample Prep Kit	10 Samples	FC-102-1001
Standard Cluster Generation Kit (for GA$_{II}$)	1 Flow Cell	GD-203-2001
Standard Sequencing Kit (36 cycle)	1 Flow Cell	FC-104-3002

†Representative kits shown here. Additional options are listed in the Illumina Product Guide.

Products Available From Agilent Technologies	Reactions	Catalog No.
SureSelect Target Enrichment System Kits	10	G3360A
	25	G3360B
	50	G3360C
	100	G3360D
	250	G3360E
	500	G3360F
	1000	G3360G
	2000	G3360H
	5000	G3360J
SureSelect Human Chromosome X Exome Kit	5	G4459A

Illumina, Inc. • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2010 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, GoldenGate Indexing, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, CSPro, GenomeStudio, Genetic Energy, HiSeq, and HiScan are registered trademarks or trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners. Pub. No. 770-2009-014 Current as of 17 April 2009

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA