Agilent's SureSelect Target Enrichment System:

Bringing Cost and Process Efficiency to Next-Generation Sequencing

Product Note

Authors

Fred P. Ernani, Ph.D.
Agilent Technologies, Inc.
Santa Clara, USA

Emily M. LeProust, Ph.D.
Agilent Technologies, Inc.
Santa Clara, USA

Abstract

While next-generation sequencing has revolutionized the way genomes are sequenced, this technology possesses a fundamental weakness—the inability to easily target specific regions of a genome. To address this, Agilent has released the highly powerful SureSelect Target Enrichment System, developed in collaboration with the Broad Institute.[1] This system uses an extremely efficient hybrid selection technique which significantly improves the cost- and process efficiency of the sequencing workflow, allowing for a larger number of samples per study. With sample input requirements at or below 3µg of genomic DNA, even the most precious of samples can be utilized for massively-parallel sequencing without risk of depletion. The system leverages well-known Agilent strengths: 1) proprietary SurePrint oligonucleotide synthesis of complex libraries consisting of ultra-long oligonucleotides greater than 100 bases in length; 2) custom library design fully integrated with eArray, Agilent's unique design tool; and 3) quality manufacturing processes that ensure the greatest reliability and consistency. The system can also easily be incorporated into an automated environment, further increasing process efficiencies, while minimizing total sample costs.

Agilent Technologies

Introduction

Next-generation sequencing technology has brought a high level of efficiency to the process of genome sequencing, but these workflows tend to be complex, time-consuming, costly to perform, and generate enormous amounts of data that need to be analyzed, moved, and stored. And despite the increased efficiency of these systems, they all lack the ability to target specific areas of interest for sequence interrogation. With the SureSelect Target Enrichment System, only the genomic areas of interest can be sequenced, creating process efficiencies that reduce costs and allow more samples to be analyzed per study. Reducing the amount of DNA being interrogated allows investigators to perform the experiments they want to with more statistically relevant numbers of samples.

Easily automatable, the SureSelect protocol is labor and process efficient. Using the protocol with an automation solution such as the Agilent Bravo Automated Liquid Handling Platform can reduce the amount of labor per target enrichment experiment and minimize the variability of sample processing — a capability unique to the SureSelect system.

As shown in **Figure 1**, the SureSelect Target Enrichment System workflow is solution-based and is performed in microcentrifuge tubes or microtiter plates. This format is more amenable to automation, and it can be scaled to meet the needs of larger sequencing projects, a limitation inherent in other commercially available methods of target enrichment.

Design a Customized Kit Specific To Your Studies

Custom SureSelect kits are created on Agilent's web-based design tool, eArray. This turnkey bioinformatics tool is provided free of charge, allowing biologists to design customized experiments without the need to invest in specialized bioinformatics tools. The eArray portal assists in the development of microarrays for a wide range of applications, from gene expression, CGH, and CNV analysis, to ChIP-on-Chip assays, and has now been further developed to enable design of custom solution-based kits for target enrichment. Whether the goal is to capture a specific set of exons on the X chromosome or a number of defined regions of interest based upon genome-wide association studies, eArray allows researchers to easily build a kit specifically for the study at hand. Utilizing an intuitive user interface with design wizards, eArray walks researchers through the creation of their target enrichment kit. The kit design is then sent to Agilent for processing and manufacturing. Agilent enables researchers to quickly and easily do custom science through this unique, flexible manufacturing process.

RNA-Driven DNA Capture

Each target enrichment kit comes packaged with a mixture of custom SureSelect RNA oligonucleotides, or "baits," that are biotinylated for easy capture onto streptavidin-labeled magnetic beads, as well as buffers and blocking agents necessary for performing the capture process **(Figure 1)**. To perform the capture, genomic DNA is sheared and assembled into a library format specific to the sequencing instrument utilized downstream. Size selection is performed on the library prior to capture and confirmed by a method such as electrophoresis on the Agilent Bioanalyzer. Size-selected libraries are then incubated with SureSelect baits for 24 hours. RNA bait-DNA hybrids are then "fished" out of the complex mixture by incubation with streptavidin-labeled magnetic beads and captured onto a strong magnet. After the beads have been washed, the RNA bait is then digested so that the only remaining nucleotide is the targeted DNA of interest. A few cycles of DNA amplification are performed at the end of the capture, and the targeted sample is then loaded onto the sequencing instrument.

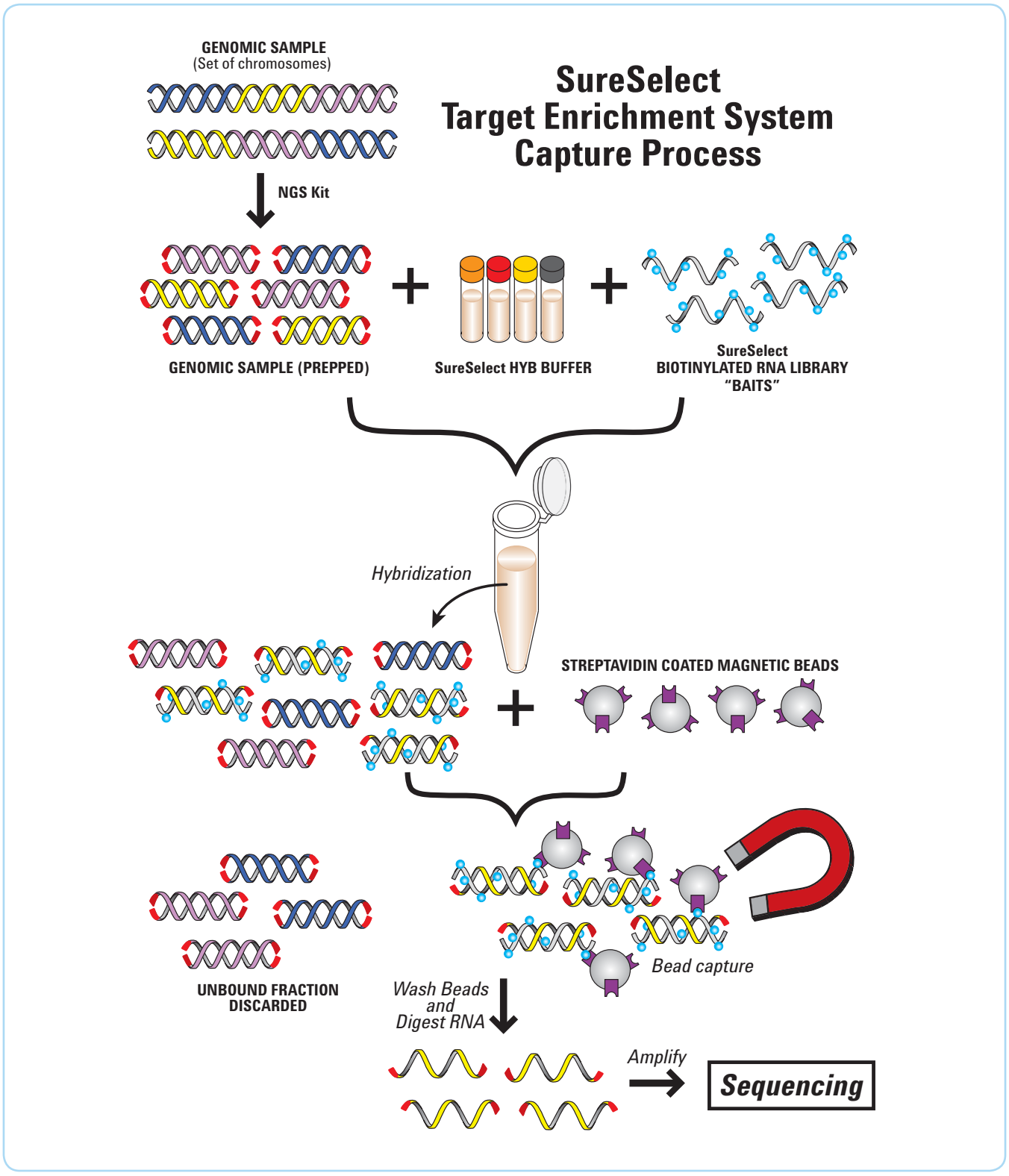

Figure 1. SureSelect Target Enrichment System Workflow

Enrichment Design Algorithms Optimized for Target Region Specificity

Validation of a custom kit may be of concern as each new kit can be composed of a unique set of probes that may not be optimized for the method of choice—in this case, target enrichment. eArray's design algorithms have been proven to work over a diverse set of genomic locations, such as small and large exons, short and long contiguous genomic targets, genome targets within repeat areas, and non-coding DNA. To assess the performance of these algorithms and SureSelect Target Enrichment kits, a diverse set of enrichment designs were created and tested in parallel against multiple samples **(Figure 2)**. To measure performance, one of the most direct measures of enrichment efficiency was used — analyzing the sequencing reads for percent in the targeted regions. As shown in **Figure 2**, the percent on target for seven designs is within approximately 40-80%, demonstrating efficiency in focusing on only regions of interest across several diverse kit designs. This translates into enrichments of 300-7400 fold; fold enrichment is largely dependent upon the size of the targeted region as well as the number of reads per sequencing run. Small capture designs typically yield a higher level of enrichment due to the kit's efficiency in focusing sequencing reads on a smaller subset of the targeted genome.

Because the shearing of DNA prior to creation of prepared libraries is random, no matter how specific the capture methodology, both the targeted region and near DNA targets will be captured. Agilent has optimized the design algorithms and the protocol to prepare libraries prior to capture so as to limit the near target sequences captured. These near-target sequences are of little to no use to researchers and may also entrain repeat regions close to targeted regions, in particular, exons. To this end, the specificity of the SureSelect System was measured by analyzing sequencing reads exactly on target, within 100 bp of the target, and within 200 bp of the target. **Figure 2** shows the off-target capture rate is not substantially increased by including sequence reads within up to 200 bp of the target.

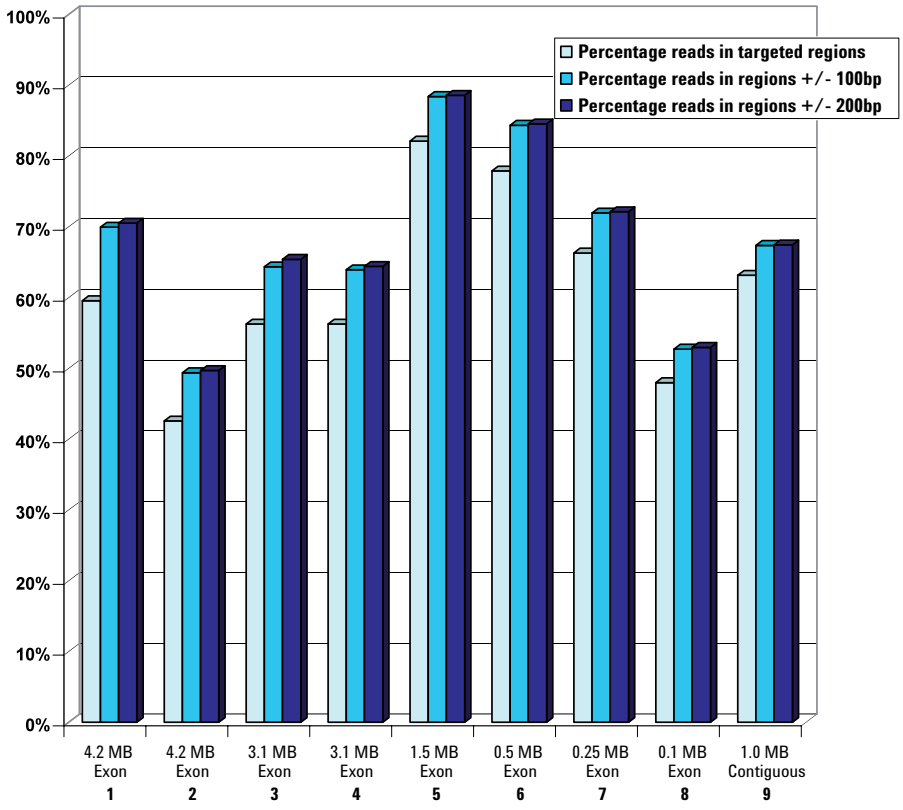

Figure 2. SureSelect Performance over a Diverse Set of Target Enrichment Designs

Percentage of sequence reads that map to targeted regions for different library designs. Each set of three values represents the percent of reads that are: on target (light blue), on target plus within 100bp of target (med blue), and on target plus within 200 bp of target (dark blue). Samples 1 and 2) 4.2 MB Exon Design, different samples; 3 and 4) 3.1 MB Exon Design, X Chromosome Demo Kit, different samples; 5) 1.5 MB Exon Design 6) 0.5 MB Exon Design; 7) 0.25 MB Exon Design; 8) 0.1 MB Exon Design; 9) 1.0 MB Contiguous Region Design.

Read Distribution and Sequence Coverage

Another metric of great importance to DNA sequencing research is read distribution, as it affects the ability to adequately cover genomes to identify sequence variation. However, none of the commercially available next-generation sequencers can output a perfectly even read distribution across a genome of interest due to the nature of massively parallel sequencing. Statistically there will always be some stretches of DNA that are read out more than others. The challenge for any sample preparation method aimed at specifically selecting targets for sequencing is to not introduce any further sequence bias. There are two ways of looking at this metric: 1) plot the distribution of the actual percent of bases with a certain number of reads; and 2) plot the cumulative number of bases with at least a certain depth. **Figure 3** shows a plot of a representative sample run on an Illumina Genome Analyzer. This plot shows that roughly 80% of all bases have at least a 20X read depth, indicating the suitability of this method for identifying SNPs with targeted re-sequencing. Plots like this will vary considerably based upon the capture design and run parameters set on the sequencer, but despite this variability, the sequence coverage seen with the SureSelect Target Enrichment System is very even, showing negligible bias. For example, if a capture design is made to target 3.3 MB of genomic locations, and 10 million reads are obtained from the sequencing run, the distribution of read depth is typically centered around 30X with 70% of the reads within 1.7 logs. Additionally, in this type of scenario one could expect approximately 95%$^{+}$ of the targeted bases to have at least one read, 80%$^{+}$ to have at least 5 reads, and 50%$^{+}$ to have at least 20 reads, making the SureSelect Target Enrichment System ideally suited for interrogating genomes for mutations.

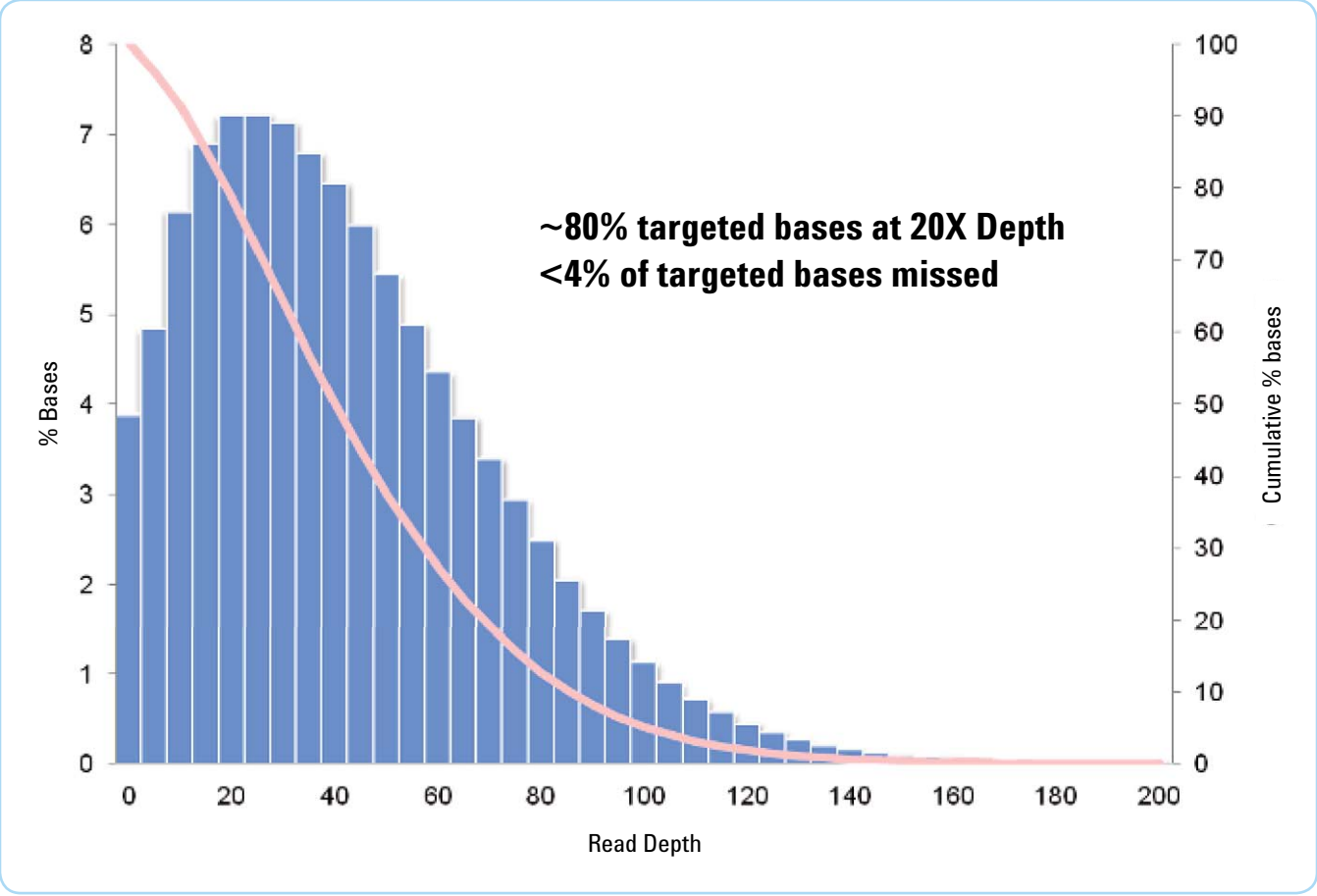

Figure 3. SureSelect Target Enrichment Sequence Coverage

Sequencing read depth across target intervals for a genomic DNA sample. Sample prepared for the Illumina Genome Analyzer and captured with an Agilent SureSelect Target Enrichment library covering exonic regions totaling 3.3Mb with 50% probe overlap. The columns show the distribution of sequencing read depths per base pair. Read depth is shown on the X-axis, binned percentage of reads at each read depth on the Y-axis (left). The pink line and right Y-axis show the cumulative read depth as a percent of total bases.

Robust and Reproducible Results

To demonstrate that the SureSelect Target Enrichment System process is both robust and reproducible, the same sample was captured with the same kit twice and run on two separate lanes of an Illumina Genome Analyzer. Sequence coverage results were then compared between replicates and plotted **(Figure 4)**. The read depth for corresponding genomic locations between experiments shows strong correlation, indicating experiment-to-experiment reproducibility. Additionally, the technical replicates gave results that were very consistent for the following metrics: actual bases represented, % on target bases, and % of reads with 30X read depth. Note that the differences in read depth from one replicate to the other is driven by the difference in total reads from the sequencer— the sample represented on the Y axis generated more reads than the sample represented on the X axis. However, the R^2 value of 0.9616 shows strong correlation between sample runs, indicating high reproducibility of the SureSelect Target Enrichment System.

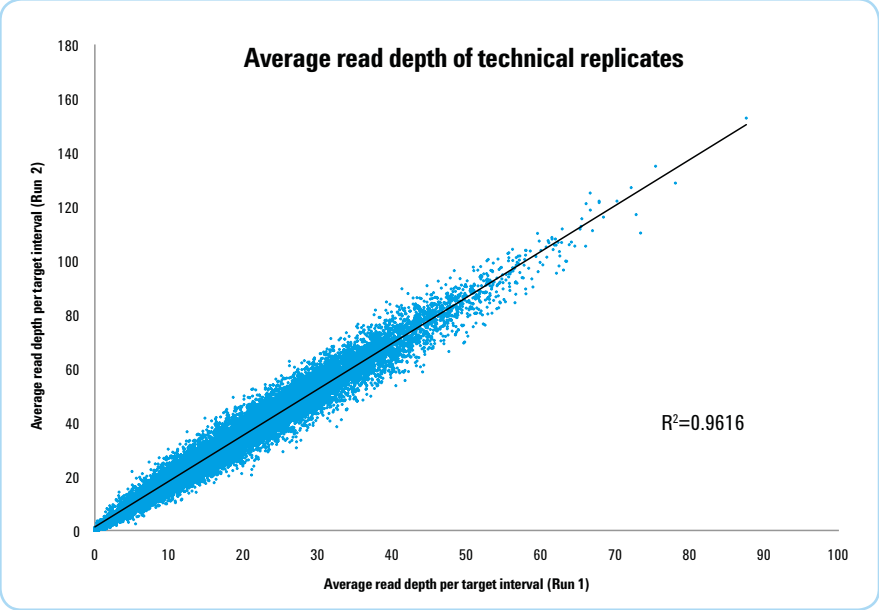

Figure 4. Technical Replicates Show Strong Correlation
Correlation of read depth across target intervals for technical replicate captures. Replicate captures were performed on a SureSelect Target Enrichment library designed for exonic targets covering 3.3Mb with 50% probe overlap.

Little or No Allele Bias

Any target enrichment process performed prior to next-generation sequencing has the potential to bias allele representation, complicating the identification of sequence variation. To demonstrate that bias is limited in the SureSelect Target Enrichment System process, sequence data were analyzed for allele balance. **Figure 5** shows data derived from a control sample of known heterozygosity for >20000 SNP reads. This information was used to deconvolute the allelic origin of each associated read. The results show little to no bias as the coverage across both alleles for several SNPs is quite similar. Thus, the Agilent target enrichment approach is efficient at capturing DNA regardless of whether the DNA strand targeted is "wild type" or contains mutations, an indicator of the utility of ultra-long oligonucleotides for capture.

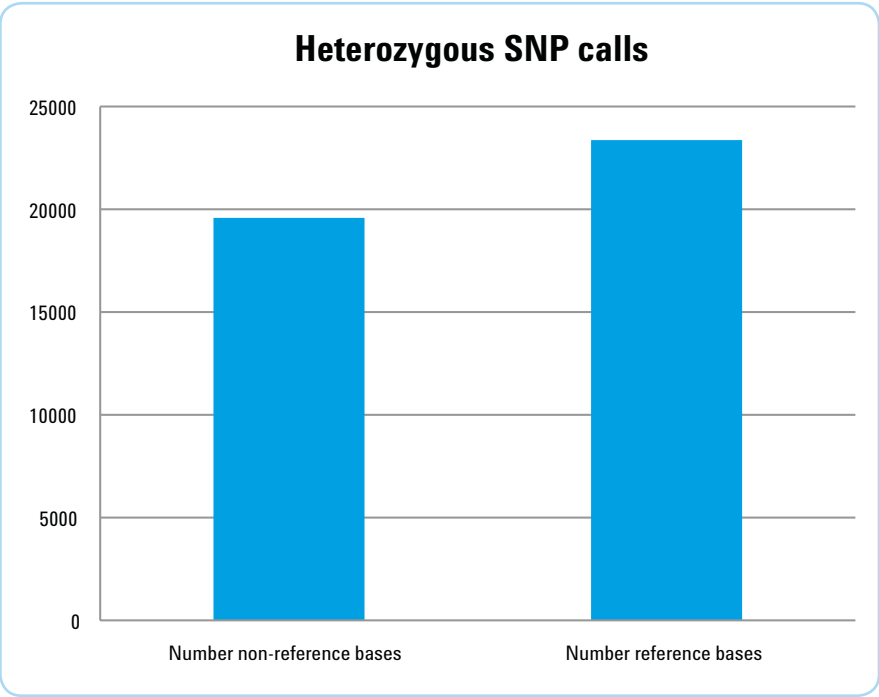

Figure 5. Allele Balance
The number of reference bases vs. non-reference bases in heterozygous SNP calls.

Robust and Reliable SNP Identification and Confirmation

One of the main applications of target enrichment prior to next-generation sequencing is the follow up to whole genome association studies to: 1) confirm SNPs; 2) define new SNPs; and 3) correlate these mutations to disease states. We wanted to verify that the SureSelect system could reliably capture known and unknown SNPs and that newly identified SNPs are genuine. Using the SureSelect method, we selected both known SNPs and a novel SNP identified after a SureSelect enrichment experiment (confirmed by PCR and Sanger sequencing). As shown in **Figure 6**, both the novel SNP and the known SNPs were confirmed by Sanger sequencing indicating that the SureSelect System is robust and reliable for the identification of SNPs.

Capture Method for Genomes with Large Deletions

One of the biggest challenges for target enrichment sequence capture methods is the ability to capture regions of genomes that possess large deletions. Methods that utilize shorter oligonucleotides capture these regions less efficiently than systems utilizing complex mixtures of longer (>100 bp) oligonucleotides, simply due to the less favorable hybridization kinetics. To improve capture performance across regions with insertions and deletions, the SureSelect Target Enrichment System utilizes complex mixtures of RNA-based oligonucleotides of 120 bp. Capture efficiency around stretches of deletions is improved by both the longer oligonucleotides and the chemical nature of the capture oligonucleotide. RNA's stronger affinity for DNA improves the SureSelect Target Enrichment System's efficiency in binding targeted regions that possess mutations such as insertions and deletions. The long capture probes are also more tolerant of mutations, as shown in **Figure 7** where a deletion of five

SNP ID	Chr	Basepair	Reference	SNP
rs11204545	1	246126076	47	42
rs11204546	1	246126085	52	40
rs831043	2	169811332	44	52
rs6886628	5	112927351	52	41
Novel	9	103381434	36	44
rs944682	1	150959212	0	91

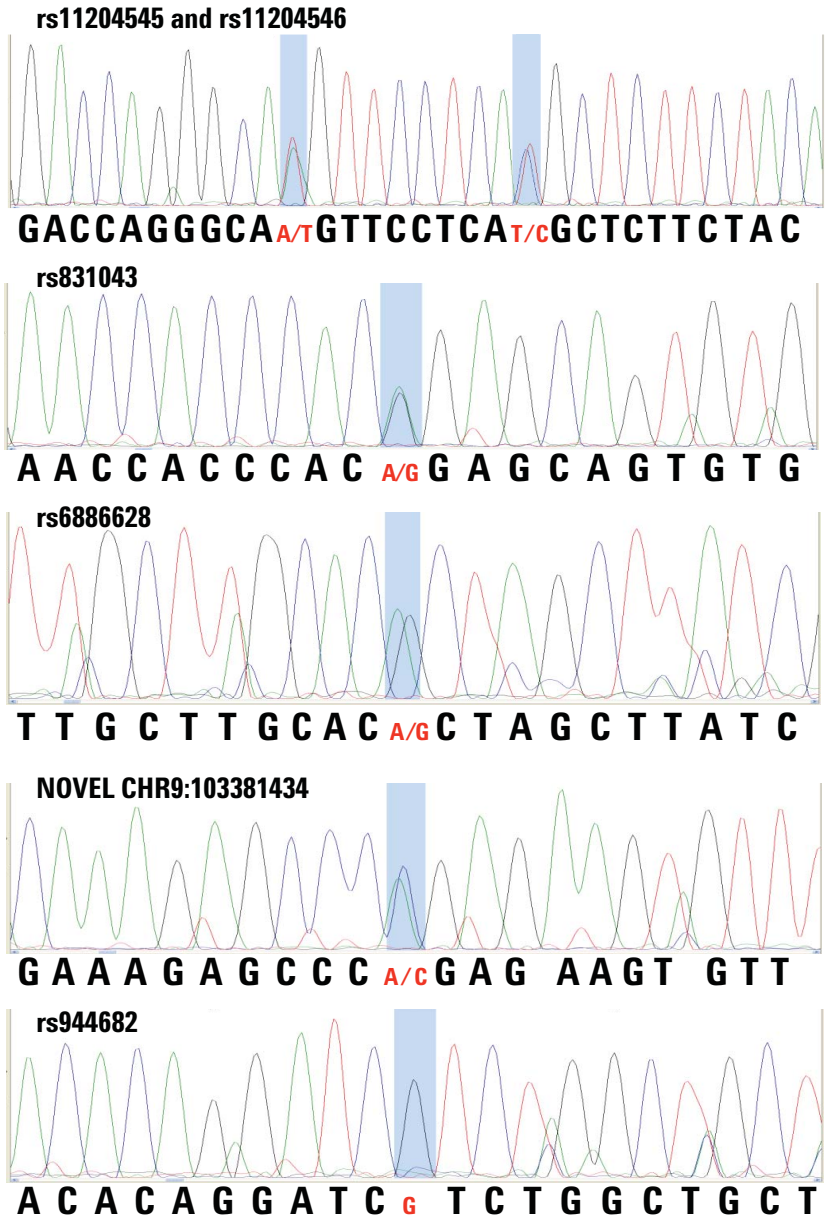

Figure 6. SureSelect Target Enrichment SNP Identification (Sanger confirmation)
Sanger sequencing confirmation of example single nucleotide polymorphisms found in a SureSelect Target Enrichment System captured genomic DNA sample.

nucleotides on human chromosome X from a patient with Menkes Disease did not prevent this region from being represented with 10X depth on an Illumina Genome Analyzer.

Summary

The data presented here demonstrate the robustness and reliability of the SureSelect Target Enrichment System. A natural extension of Agilent's ability to manufacture custom-designed high quality ultra-long oligonucleotides with SurePrint inkjet technology, the SureSelect system offers unparalleled efficiency. By requiring 10-fold less input DNA than other commercial systems and by focusing DNA sequencing on genomic areas of interest, the SureSelect System enables studies that were previously unfeasible due to the rarity of DNA sample and/or the overall cost of a sequencing study. The first truly scalable product for target enrichment, the SureSelect Target Enrichment System will enable researchers to process anywhere from a handful to thousands of samples, using the same approach. The SureSelect line of custom target enrichment products represent a revolutionary approach to DNA re-sequencing using next-generation sequencing systems.

References

1. Gnirke, A., et al. (2009). **Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing.** *Nat Biotechnology* Feb;27(2):182-9.

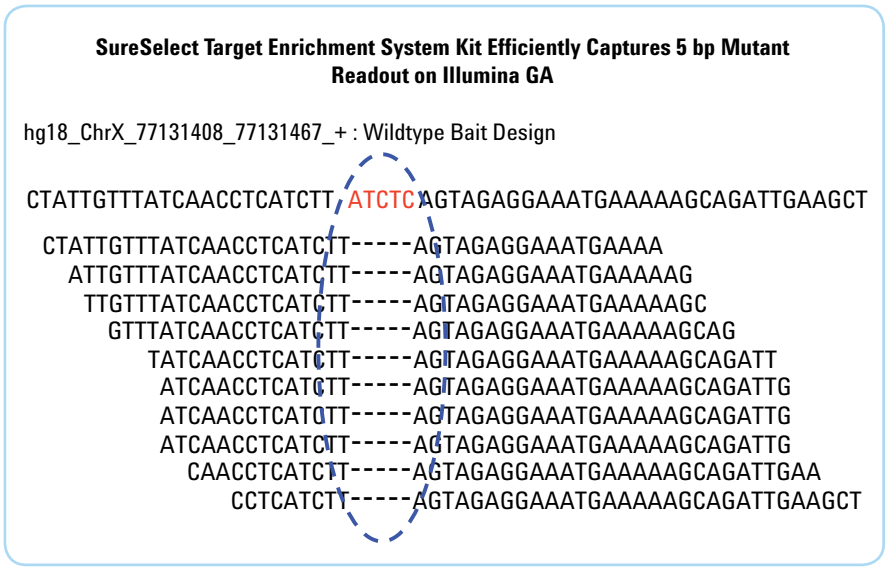

Figure 7. SureSelect Target Enrichment Captures Region with Large Deletion
Sequencing results from captures of a 5BP deletion on the X-Chromosome: Menke's Syndrome.

SureSelect Target Enrichment System Kits	
Part Number	**# Rxns**
G3360A	10
G3360B	25
G3360C	50
G3360D	100
G3360E	250
G3360F	500
G3360G	1000
G3360H	2000
G3360J	5000
SureSelect Human Chromosome X Exome Kit	
Part Number	**# Rxns**
G4459A	5

Learn more:
www.opengenomics.com/sureselect

For technical support:
Email sureselect.support@agilent.com

Find an Agilent customer center in your country:
www.agilent.com/chem/contactus

U.S. and Canada
1-800-227-9770
agilent_inquiries@agilent.com

Asia Pacific
adinquiry_aplsca@agilent.com

Europe
info_agilent@agilent.com

Published in U.S.A, March 16, 2009
Publication Number 5990-3532EN

Agilent Technologies