A Comprehensive Assay for Single Nucleotide Polymorphism, Copy Number Variants and Loss of Heterozygosity Using SureSelect Target Enrichment

Kyeong Soo Jeong[1], Arjun Vadapalli[1], Ashutosh Ashutosh[1], Paula Costa[1], Brian Peter[1], Stephanie Fulmer-Smentek[1], Magnus Isaksson[1], Jayati Ghosh[1], Douglas Roberts[1], and Holly Hogrefe[2]

Agilent Technologies, Diagnostics and Genomics Group [1]Santa Clara, California, USA [2]La Jolla California, USA

Agilent Technologies

Abstract

Genetic variation in the mammalian genome spans a size extreme that includes cytogenetically recognizable elements and single-nucleotide polymorphisms. Copy number variations (CNVs) are an important intermediate size class of structural variations that involve unbalanced arrangements that increase or decrease the DNA content in mammalian genomes. CNVs are responsible for a continuous spectrum of phenotypic effects ranging from adaptive traits to embryonic lethality. The technology platforms available to identify CNVs include fluorescence in situ hybridization (FISH), array comparative genomic hybridization (aCGH) and more recently next generation sequencing. More mature platforms such as FISH and aCGH suffer from low resolution of genomic regions. The rapid development of low cost short-read sequencing technologies has paved the way to detect mutations and high resolution structural variation detection in a single experiment. Here we describe a comprehensive assay that enables researchers to identify SNP, INDEL, CNV, and LOH using SureSelect target enrichment. This design can be employed as a standalone entity or in concert with other bait designs for SNP and INDEL detection. We also describe methods for data analysis and visualization.

Motivation

Using next generation sequencing technologies the user can either sequence the entire genome or sequence regions captured with target enrichment assays such as SureSelect™. The choice between whole genome sequencing (WGS) and target enrichment based sequencing depends on balancing cost and sequencing output. The current cost of whole genome sequencing to can cost over a thousand dollars and increase the computational time to analyze the data. A more economical alternative to WGS is sequencing of small gene panels or an entire exome that represents a highly enriched subset of the human genome. Since current exome panels will not provide users with a uniform probe spacing across the genome we are presenting a SureSelect design with functional resolution of about 300kb in CNV detection. CNV in clinically relevant regions from Clingen (formerly ICCG) can be identified at 25~50 kb resolution. Targets also include SNP regions with high minor allele frequencies, allowing detection of LOH as low as ~2Mb resolution.

Bait design

A majority of the baits locations were selected based on known SNP positions taken from dbSNP 138 database. An empirical selection of the final set of SNPs was done after screening available candidates for their performance with the SureSelect XT assay. The probe design accounted for GC content, the likelihood of mapping uniquely to the genome, and nearby sequence complexity or randomness in order to maximize bait specificity.

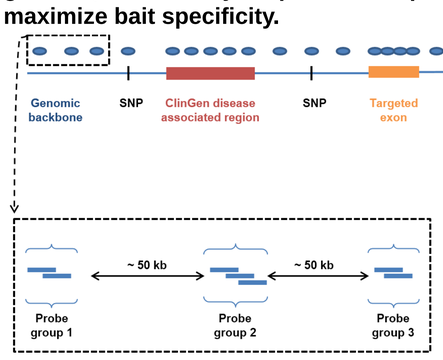

Figure 1. OneSeq backbone design layout

OneSeq Workflow

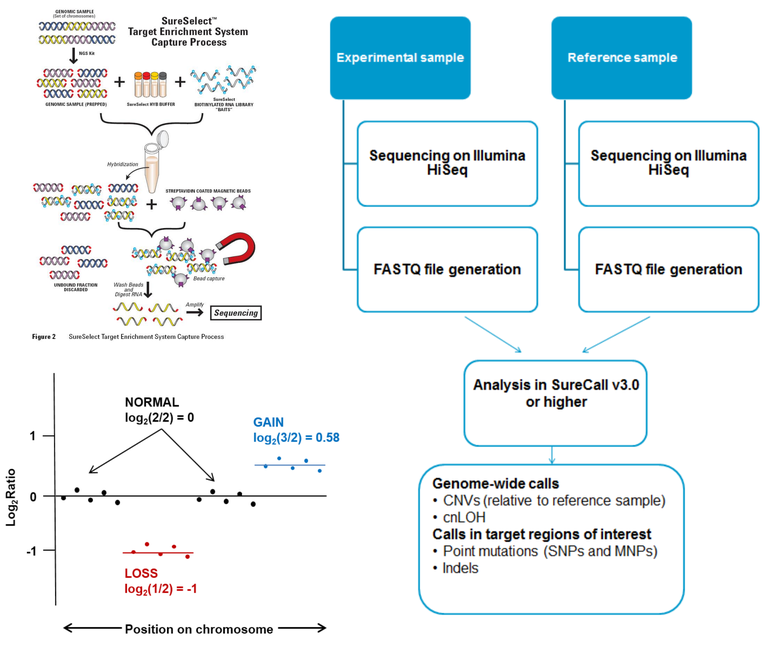

Figure 2. OneSeq workflow showing SureSelect XT library preparation, capture and sequencing. The log2 ratio of sample over reference read depths is used to detect aberrations in SureCall 3.0.

Capture Performance

OneSeq design shows excellent capture and uniformity.

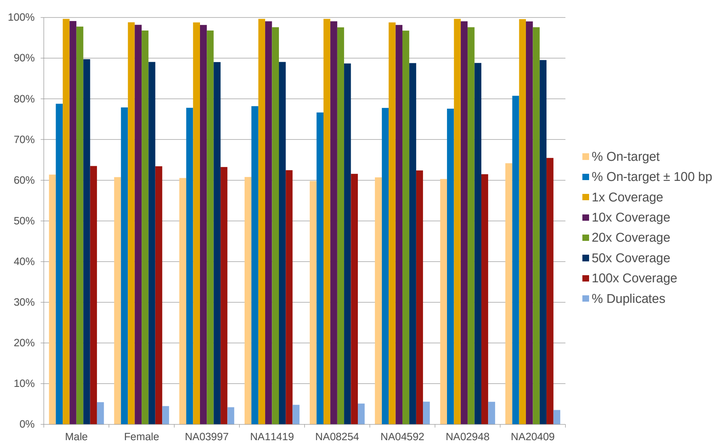

Figure 3. Summary of typical capture performance. OneSeq was performed with Coriell samples and OneSeq constitutional research panel. The vertical axis shows the percentage of on-target, coverage or duplicates. Each sample includes 7Gb of sequencing.

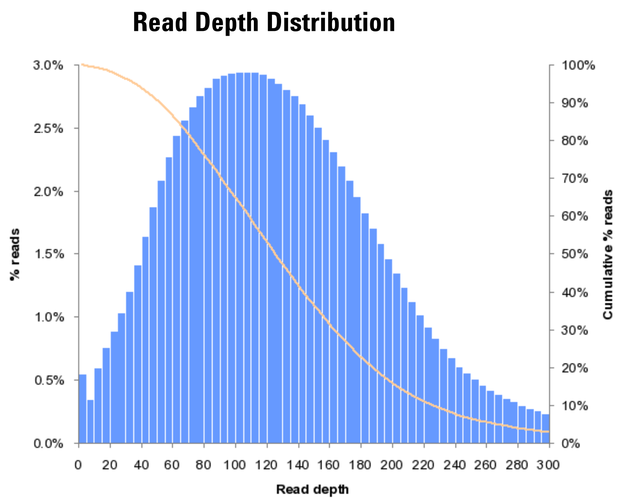

Figure 4. Read depth distribution in OneSeq constitutional research panel. The cumulative percent of reads is shown by an orange line.

Trisomy 13

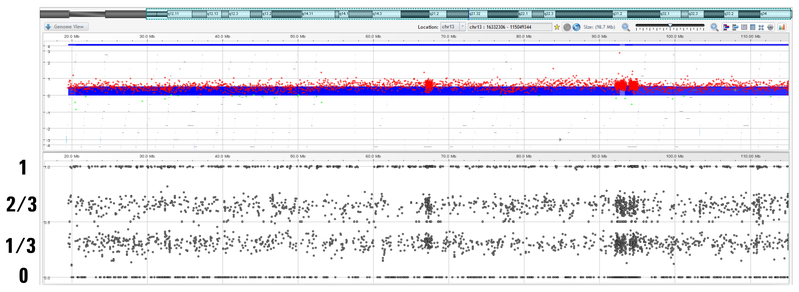

Figure 5. Detection of trisomy 13 in Coriell sample NA02948 with karyotype 47, XY, +13. The B-Allele frequency plot is shown below the log2 ratio data.

18q Deletion

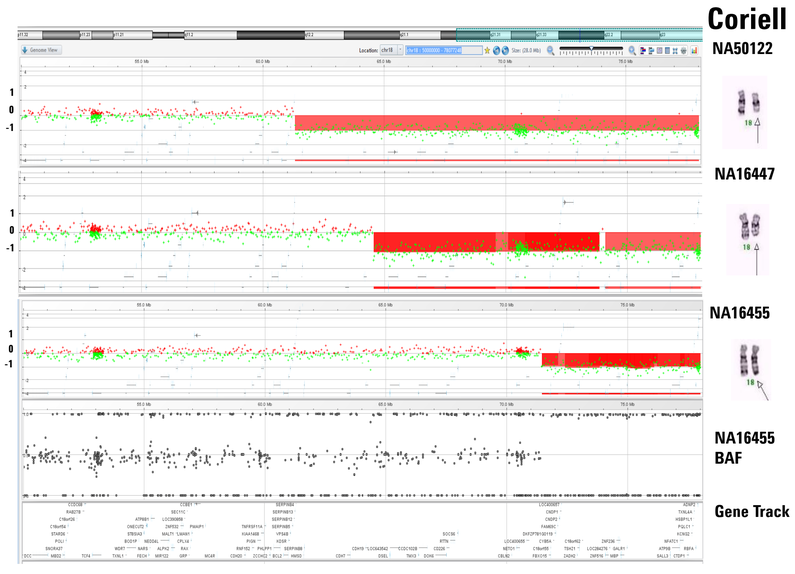

Figure 6. Chromosome 18q deletion break point detection. Samples include NA50122 showing deletion in 18q21.33q23, NA16447 showing deletion at 18q22.1q23, and NA16455 with deletion present at 18q22.3q23.

Short CNVs

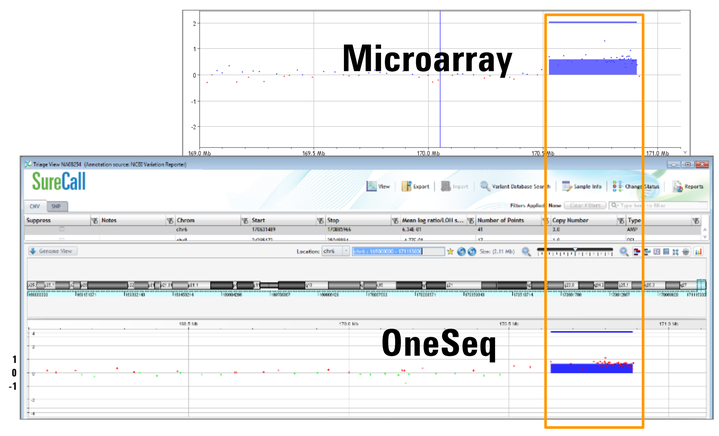

Figure 7. Comparison of ~370kb amplification found with OneSeq Target Enrichment versus CGH+SNP microarray on Coriell sample NA08254

Indels or SNPs

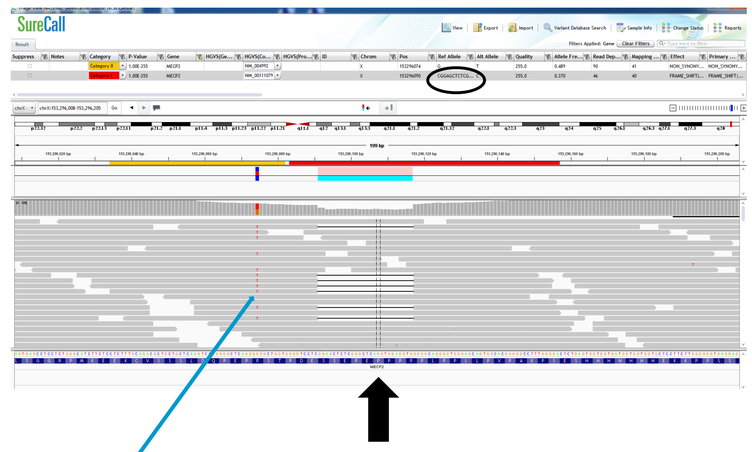

Figure 8. Detection of heterozygous SNP and indel in the same assay using NA16382

Copy-neutral LOH

Figure 9. NA20408 shows maternal uniparental disomy (UPD) in chromosome 15 which is associated with Prader-Willi syndrome.

Conclusions

- Agilent's OneSeq provides a comprehensive, efficient, robust, and cost-effective means to assess SNPs, INDELs, CNVs, and LOHs in one assay.
- Different capture sizes show comparable high performance regardless of various custom targeted regions.
- High reproducibility of enrichment, depth distribution, and sequence coverage from multiplexed sequencing.
- The OneSeq workflow from DNA samples to analysis provides a complete solution to determine DNA aberrations of around 300kb.

For Research Use Only. Not for Use in Diagnostic Procedures.