Complementing Next Generation Sequencing Technologies With Agilent's SureSelect DNA Capture Array

Application Note

Authors

Andy Bhattacharjee[1]
Emily Hodges[2,3]
Ben Gordon[1]
Michelle Rooks[2,3]
Zhenyu Xuan[2]
Leo Brizuela[1]
W. Richard McCombie[2]
Gregory Hannon[2,3]

[1] Agilent Technologies, Inc.
5301 Stevens Creek Boulevard
Santa Clara, CA 95051

[2] Watson School of Biological Sciences
[3] Howard Hughes Medical Institute
Cold Spring Harbor Laboratories
1 Bungtown Road
Cold Spring Harbor, NY 11724

Abstract

Massively parallel DNA sequencing technologies are of pivotal importance in genome biology and medicine, as they can potentially enable comprehensive and systematic evaluation of genetic variation. Currently, these sequencing technologies are geared toward sequencing whole genomes. A broader adoption of these technologies requires a more cost-effective method with higher throughput and greater versatility than PCR—a method such as target enrichment, the targeted resequencing of multiple discrete genomic regions of interest. To validate the use of Agilent DNA microarrays for target enrichment, Agilent collaborated with the laboratory of Dr. Greg Hannon (Cold Spring Harbor Laboratory) to capture exonic regions relevant to a breast cancer sequencing study. We targeted 0.025% of the human genome, using an Agilent 244K array of 60-mer probes to capture approximately 1,287 discrete genomic regions. The captured DNA was then released and sequenced. Various hybridization conditions were tested, ultimately obtaining a 2,700-fold enrichment of sequencing reads within targeted regions. The system was seen to be effective, with sequencing reads covering over 99.8% of the targeted regions and 98% of the targeted bases with at least one read and with a normalized average per-base read depth of 27 per million 32-base reads. These results confirm that Agilent DNA microarrays can provide a rapid and effective solution for targeted sequencing of genomic regions of interest.

Introduction

A growing body of evidence suggests that many complex diseases harbor variations on the genomic scale that range in size from a single-base pair to millions of bases.[1,2] Next-generation, or "high-throughput," sequencing approaches hold promise to aid in dissecting such complex variations, but the cost of whole-genome sequencing remains prohibitive for most researchers. To address this issue, it has been recently demonstrated that sequencing efforts can be directed to user-defined target regions.[3-6] This approach, often called genome partitioning or target enrichment, can greatly expand the potential applications of sequencing technologies by focusing sequencing efforts and costs specifically on regions of interest such that the limitations of next-generation sequencing are no longer prohibitive factors.

PCR has been the dominant technology for targeted amplification for nearly two decades, but its use has been constrained due to difficulties in multiplexing and to limited amplicon sizes. Recently, alternative approaches extend target enrichment to larger numbers of regions (e.g., all 200,000 human exons) or to very large contiguous regions (tens or even hundreds of megabases). These approaches leverage advances in DNA synthesis technology to reduce sample complexity, either by employing large libraries of oligonucleotides[3] or oligonucleotide microarrays for enrichment of target regions.[4-6]

This application note describes the use of Agilent SureSelect DNA Capture Array technology to target genomic regions of interest for high-throughput sequencing. We demonstrate that this method is capable of isolating user-defined genomic regions from complex eukaryotic genomes. For our test case (outlined in **Figure 1**), we selected 1,287 discrete targeted regions from three chromosomal loci suspected of being involved in familial breast cancer. Probes were tiled at 3-bp steps using 60-mer oligonucleotides across targeted regions. We used sonicated HapMap target DNA to carry out several optimizations following a routine of end-repair and adaptor ligation. Target DNA was then hybridized to the DNA oligonucleotide microarray. After hybridization, the bound DNA fraction was eluted from the microarray and sequenced. Data obtained were analyzed to determine enrichment and sequence information.

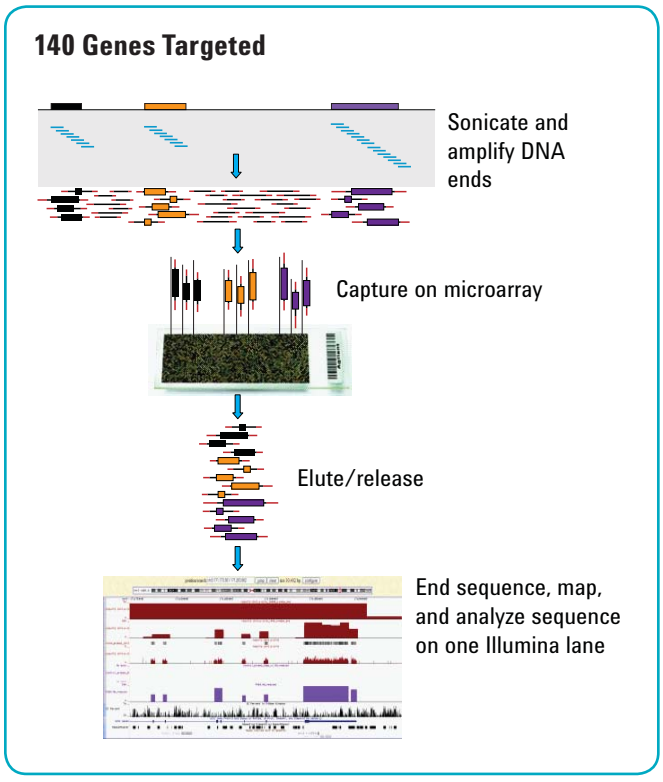

Figure 1. Schema of on-array capture assay followed by DNA sequencing. 1,447 targeted exonic regions (black, orange, and purple) were tiled by 60-mer probes with 3-bp tiling (blue). Fragmented DNA was ligated to adapters (red), size-selected, and amplified. DNA was hybridized, washed in stringent buffer twice, and the bound fraction was eluted by heat treatment. Eluted DNA was further amplified, loaded on Illumina Genome Analyzer (black, orange, and purple), and analyzed as described.

Materials and Methods

Microarray design

We sought to understand the performance of Agilent SureSelect DNA Capture Arrays for target enrichment by targeting 1,447 exons covering 516,006 bp (excluding flanking regions). To target these exons, probes were designed to cover each exon and the flanking 100- to 150-bp regions. Intervals were merged when flanking regions overlapped, resulting in a total of 1,287 targeted intervals. For each interval, probes were designed in a manner previously described by Hodges et al.,[6] except that most probes were tiled at a higher density (3-bp tiling), and probes were fixed in length at 60 bases without T_m adjustments. As a result, approximately 774 kb of the genome was targeted, including exons and their flanks on a single 244K array with 230,014 probes. Probes were designed using 3-bp tiling, meaning that probes were tiled across the target region such that two neighboring probes had start points 3 bp apart and an overlap of 57 bp. To reduce non-specific binding of genomic elements, probes containing highly repetitive elements were excluded by considering the genomic frequency of all 15-mer subsequences.[6] This method results in slightly different coverage of the genome than is produced by more conservative RepeatMasker filtering. An additional set of 100 control regions comprised of 88,435 bp were tiled with probes at 20-bp offsets and using RepeatMasker probe filtering.

DNA library preparation

Individual HapMap purified DNA (ID NA12762) was obtained from the US National Institute of General Medical Sciences Human Genetic Cell Repository. DNA (1–5 µg) was fragmented to a range of 150–800 bp using a Covaris S1 instrument. Nebulizers (Illumina kit) or sonicators (Diagenode) may also be used for fragmentation. Fragmented DNA ends were repaired, phosphorylated, and adenylated according to manufacturer's instruction (Illumina). Adenylated ends were ligated to Illumina-

compatible adaptors. Following ligation, DNA fragments of approximately 150–300 bp were size-selected by 2% agarose gel purification (TAE). After gel purification (eluted in 30-µl elution buffer), at least 8–12 parallel reactions of PCR enrichment were performed with the Illumina adaptor-compatible primers. PCR reaction components for each 50-µl reaction were as follows: 25-µl Phusion HF Master Mix (Finnzymes), 1-µl Forward Primer (50 µM), 1-µl Reverse Primer (50 µM), 1-µl Adaptor Ligated Template, and 22-µl Nuclease-free water. PCR reaction conditions: **Step 1)** 98°C, 30 seconds; **Step 2)** 98°C, 10 seconds; **Step 3)** 65°C, 30 seconds; **Step 4)** 72°C, 30 seconds; repeat steps 2 through 4, 17 times for a total of 18 cycles; **Step 5)** 72°C, 5 minutes; hold at 4°C. Pooled reactions (4 pools each containing 3 separate PCR reactions) were subjected to PCR product purification using a Qiagen (QIAquick or minElute) column. Reactions were eluted in 30-µl elution buffer. Before hybridization, an aliquot of the pre-hyb DNA was archived for downstream qPCR assays.

Hybridization, washes, and elution on Agilent DNA microarrays

The arrays were hybridized as outlined in the Agilent aCGH* manual with minor modifications. Briefly: 10–20 µg sample DNA (in 138-µl volume), 5-µl Blocking oligo 1 (200 µM) Forward Primer, 5-µl Blocking oligo 2 (200 µM) Reverse Primer, 5-µl Blocking oligo 3 (200 µM) reverse complement of Forward Primer, 5-µl Blocking oligo 4 (200 µM) reverse complement of Reverse Primer, 50-µl Human Cot-1 DNA (1 mg/ml), 52 µl 10x Blocking Buffer, and 260-µl 2x Hyb buffer were mixed and denatured at 95°C for 3 minutes followed by transfer to 37°C for 30 minutes. The mixture was centrifuged at 17,800g for 1 minute. A volume of 490-µl hybridization mixture was dispensed onto the center of the gasket slide with the Agilent label facing up and the gasket slide in the hybridization chamber. The DNA array was placed with the active side contacting the hybridization mixture. The resulting slide-gasket sandwich with hybridization solution was incubated at 65°C for 65 hours on a rotisserie in an Agilent hybridization oven. The aCGH Wash Buffer #2 was preheated at 37°C in a hybridization oven overnight before washing. Following hybridization, the slide-gasket sandwich was disassembled in aCGH Wash Buffer #1 at room temperature. The slide was washed for 10 minutes in Wash Buffer #1 and transferred to Wash Buffer #2 at 37°C and washed for 5 minutes. In one hybridization experiment, Cot-1 was omitted, and in a second experiment, the conditions of hybridization and wash were as described in Hodges et al.[6] In some experiments, blocking oligos 1–4 were not used. In these cases, the DNA was prepared in a volume of 158 µl.

Slides were dried by centrifugation at 600 rpm for 30 seconds. Approximately 490 µl nuclease-free water was added to a new gasket in the hybridization chamber and the dried slide was placed atop as described earlier, creating a fresh slide-gasket sandwich. The hybridization chambers containing slide-gasket sandwiches were placed in a separate 95ºC Scigene 700 series oven for 10 minutes. Following the heat denaturation, the assemblies were quickly and carefully removed from the oven. For each chamber, the chamber screw was loosened a quarter turn, while grasping the slide sandwich and chamber base with a paper towel due to the heat. The chamber was tilted so that the narrow end (no label) pointed upward. A 1-ml 30G syringe was inserted into the narrow end of the slide sandwich, through the rubber ring between the gasket and the slide. As much of the eluate as possible was removed with the syringe.

Lyophilization and PCR enrichment

Following elution, the eluted DNA was lyophilized down to 50 µl in a speed vac set on "high" for approximately 2.5–4 hours. An aliquot of the re-suspended eluate was used as a template for PCR enrichment using the following PCR conditions. Five independent 50-µl PCR reactions were performed using 5-µl aliquots of eluted DNA (template). The PCR reaction components used were as follows: 25-µl Phusion HF Master Mix, 1-µl Forward Primer (50 µM), 1-µl Reverse Primer (50 µM), 5-µl Eluted Template, 18-µl nuclease-free water. The PCR reaction conditions were: **Step 1)** 98°C, 30 seconds; **Step 2)** 98°C, 10 seconds; **Step 3)** 65°C, 30 seconds; **Step 4)** 72°C, 30 seconds; repeated step 2-4 for 17 times for a total of 18 cycles; **Step 5)** 72°C, 5 minutes. Following PCR, two and a half reactions (125 µl) were pooled and purified on a QIAGEN column using the PCR purification kit. The DNA was quantified by a Thermo Scientific Nanodrop 7500 spectrophotometer and diluted to a working concentration of 10 nM. Cluster generation was performed in each Illumina flow cell lane. A standard Illumina sequencing primer was used for 36 cycles of base incorporation.

QC by qPCR

qPCR was performed according to the guidelines in the SYBR® Green PCR Master Mix product insert (Applied Biosystems). In this set of experiments, qPCRs were performed for a selected

*A portion of Agilent's aCGH protocol and buffers, specifically the hybridization and wash, were employed in the development of the DNA Capture technology.

set of five targeted regions in order to confirm successful "capture" prior to sequencing **(Figure 2)**. A non-targeted region (GAPDH) was also included as a negative control that was not represented on the array. The difference between pre-selection and post-selection CT values indicates the level of enrichment. The increase in CT value for GAPDH suggests that it was diluted rather than enriched, while the reductions in CT measured for selected targeted exons indicate that they were in fact enriched.

Read mapping and coverage analysis

The ELAND program output provided by Illumina was used to map all reads to the human genome, with a 25- or 32-base seed for mapping, allowing at most 2 mismatches. Only uniquely mapped reads were retained for further analysis. In order to get a comprehensive view of the enrichment coverage and read depth, we utilized a data analysis tool to compute statistics. For these calculations, we considered all genomic bases covered by at least one probe as targets.

Results and Discussion

We attempted to determine the set of conditions that provide optimal performance for on-array capture. To evaluate the performance of each condition, we measured the enrichment (fraction of reads in targeted regions divided by the fraction of genome targeted) and per-base read depth (average number of sequencing reads obtained for each targeted base). Data for the five conditions that we evaluated are summarized in **Table 1** (see page 7). An illustration of typical base coverage for four representative exons is shown in **Figure 3**.

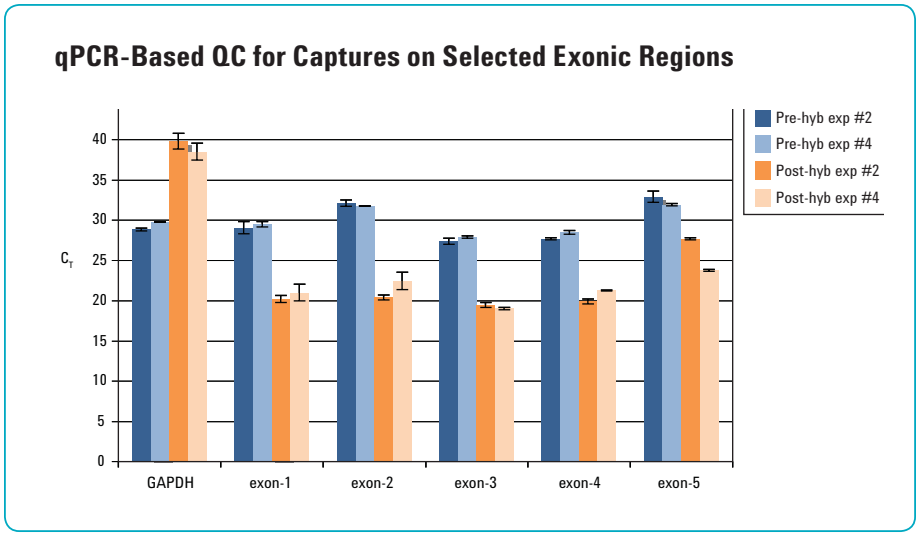

Figure 2. qPCR-based QC for capture of selected exons. CT versus samples across different experiments.

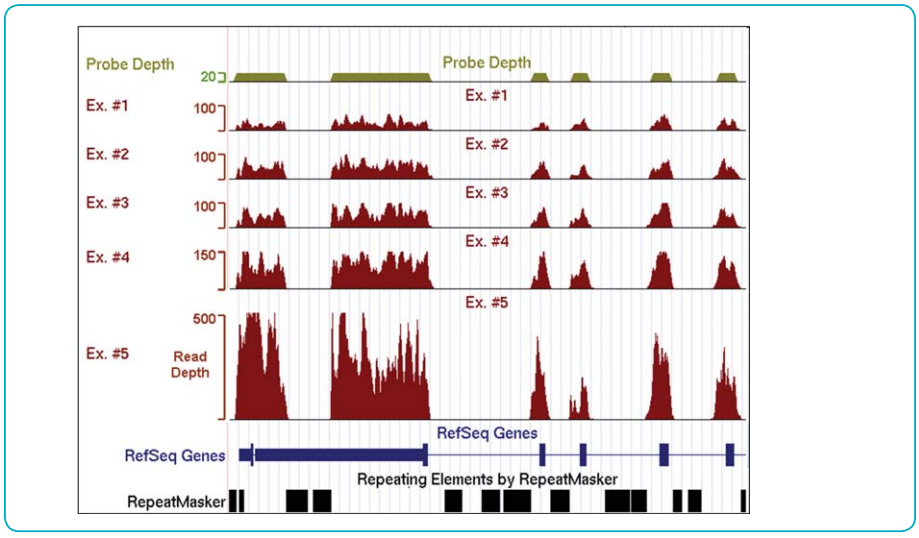

Figure 3. Illustration of typical read depths across targeted intervals. Read depths were computed by counting the number of reads covering each individual base in the genome. Visualization was performed using the UCSC genome browser.

First, we observed that addition of Cot-1 DNA significantly improves the enrichment of targeted DNA regions. Using 10 µg of amplified HapMap DNA, we performed the assay both in the absence of Cot-1 and with 50 µl of Cot-1 DNA (Experiments #1 and #2), as described in Agilent's aCGH protocol. Cot-1 improved the enrichment from 672-fold to 1,296-fold, and increased the average per-base sequence read depth from 13 to 24 **(Table 1)**. We believe that the Cot-1 DNA scavenges undesired repetitive DNA in solution, reducing its non-specific binding to the microarray.

Holding the amount of target DNA constant at 10 µg, we tested hybridization at 42°C and formamide in Experiment #3 (conditions used in Hodges et al.).[6] We observed a 1,071-fold enrichment and average per-base read depth of 26, suggesting that the hybridization conditions were comparable to the Agilent 65°C hybridization conditions.

To assess the importance of the input amount of target DNA, we tested the effect of doubling the DNA amount to 20 µg (Experiment # 4). This resulted in an increase of enrichment to 1,500-fold from the 1,296-fold increase obtained with 10-µg input. More significantly, we observed dramatically improved per-base read depth, with the largest fraction of bases covered by at least 20 reads (84.75%), and an average read depth of 67.46 reads per base **(Table 1)**. We believe that increasing the amount of input DNA shifts the equilibrium to favor more hybridization. While this shift might also increase the amount of non-specific hybridization, the increased enrichment suggests that the shift is greater for hybridization of desired targets. An additional benefit of this behavior is that it enables pooling of samples, for which reads could be assigned back to individual samples using a barcoding strategy.

To assess the importance of probe tiling density, we investigated the behavior of probes with 20-bp spacing (resulting in a 40-bp overlap) as compared to probes with 3-bp spacing (and 57-bp overlap) by examining the behavior of 100 additional control exons targeted at 20 bp on the same array. The average per-base read depth of 20-bp tiled control intervals in Experiment #4 was 28.2-fold, compared to 67.5-fold observed on the 3-bp tiling regions **(Table 1)**. Across all experiments, we observed drops in average per-base read depth of between 2.6- to 3.2-fold for the 20-bp tiling compared to 3-bp tiling, despite the more extensive 7-fold decrease in probe density. This suggested that, by using 20-bp spacing, a single 244K Agilent microarray can be used to target 4.8 Mb of genomic sequence in exonic regions, albeit with lowered read depths.

We tested this concept under an entirely different array design targeting different regions of the genome, including non-exonic and promoter regions (data not shown). One design tested increased spacing (15 bp) to cover larger portions of the genome on chromosome 7 (3.7 Mb), while another design targeted a 787-kb target region within this locus at 3-bp spacing. By increasing the targeting from 787 kb to 3.7 Mb (factor of 4.8-fold) we only observed a reduction of normalized read depth from10 to 2.76 reads per base per million reads of 32 bases (3.6-fold). Taken together, this suggests that larger targeting regions can be captured on the Agilent 244K array without significantly sacrificing read depth.

Although the capture performance was good, we sought to understand other factors that constrained it from being closer to ideal. We tested the hypothesis that adaptor sequences used in generating the libraries were causing non-specific hybridization via interactions between adapters. We tested this by varying the concentration of Cot-1 and using blocking oligos corresponding to the forward and reverse primers (Experiment # 5). We observed significant improvement in enrichment from 1,500- to 2,682-fold and targeted reads to up to 66.85% under conditions described **(Table 1)**. These results suggest the importance of including the blocking oligos for optimal target capture performance.

We observed few reads further than 100 bp from the targeted intervals, highlighting the specificity of the target capture **(Figure 4)**. We further observed that on average, maximum enrichment is achieved inside the intervals at 100–200 bp from the boundaries. Note that although a maximum of 20 probes target each base when using 3-bp tiling, this coverage was not enforced at interval boundaries in this design. Therefore, bases at interval edges had fewer probes for capture, an effect that likely contributed to the observed edge phenomenon. Duplicating probes near boundaries may reduce this behavior.

Unlike the large number of different regions targeted in this study, some studies target fewer, larger regions of the genome.

For such studies, there are also fewer boundaries, so more of the probes are utilized to their full potential. Additionally, the larger the fraction of the genome targeted, the greater the number of random reads expected to fall in targeted regions. Together, these observations may partly explain the improved performance seen by another group when targeting larger regions of the genome.[5] Interestingly, the fraction of reads in their study that mapped to targeted regions ranged from 15% for 1-bp tiled 200-kb regions to 35% for 500-kb regions. In our study, regions of approximately 500 kb provided equivalent enrichment with 3-bp tiling. This suggests a potential several-fold improvement on performance by Agilent. We suspect that the increased number of molecules per Agilent feature may partially explain the improved performance, though further experiments are necessary to understand these differences.

Although the enrichment achieved in Experiment #5 was more than adequate for most applications (including barcoding strategies), we sought to understand what noise components constrained it from being closer to ideal. To do this, we studied the characteristics of poorly enriched targets and of undesired reads found outside of targeted regions. Within target regions, we observed that low read depths in certain regions could often be explained by non-optimal GC content (data not shown) or by artifacts in ELAND mapping. Regarding the latter effect, we found that some regions of targeted areas have sub-sequences of 30 bp (or longer) that are not unique. We examined 60 of these by hand and found that reads for most were indeed produced by the sequencer, but that the reads had been rejected because they had multiple genomic alignments. Outside of targeted regions, we observed a tendency for undesired reads to occur in areas of high GC content. We also observed a disproportionate number of undesired reads in exons of untargeted genes, suggesting cross-hybridization, possibly from targeted regions with homologous domains.

Conclusion

In summary, we have demonstrated the effective use of Agilent microarrays for targeted sequence enrichment of exonic regions in the human genome. We observed enrichments that resulted in an approximate 2,700-fold increase in representation of target regions. Agilent's SureSelect DNA Capture Array (G4458A), in conjunction with customer array design through eArray, serves as a powerful tool for target enrichment prior to next-gen sequencing. This is ideal for researchers who may only want to sequence a small number of samples with various capture designs, while reducing overall cost and efforts for sequencing projects.

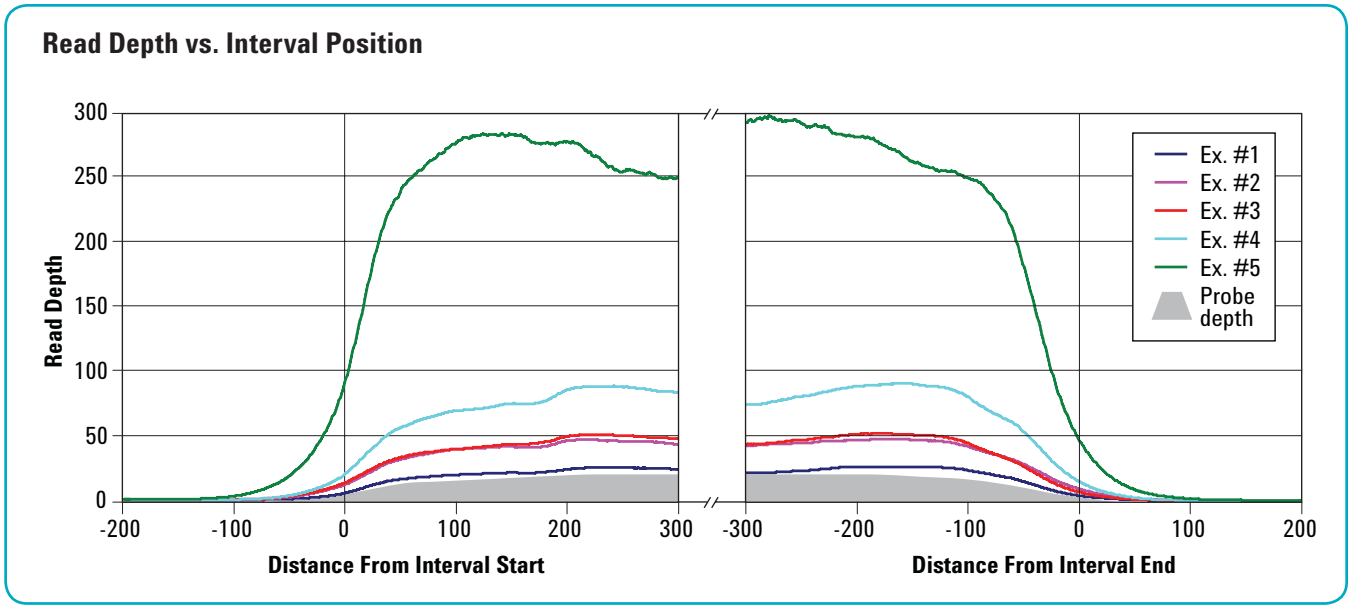

Figure 4. Per-base sequence depth across exonic boundaries. Read depth for bases upstream and downstream of the interval ends was computed and is summarized here, showing high read depths within intervals that drop off near interval boundaries. Additionally, probe coverage is indicated in gray.

	Experiment #1	Experiment #2	Experiment #3	Experiment #4	Experiment #5
Experimental conditions					
Hybridization method	Agilent	Agilent	Hodges et al.5	Agilent	Agilent
Hybridization temperature	65°C Agilent hyb	65°C Agilent hyb	42°C formamide Ng	65°C Agilent hyb	65°C Agilent hyb
Amplified DNA library	10 µg	10 µg	10 µg	20 µg	20 µg
Wash temperature	37°C	37°C	37°C	37°C	37°C
Cot -1 DNA enrichment	–	+	+	+	+
Blocking oligo	no	no	no	no	yes
Total reads with high-quality unique mapping	2,593,531	2,444,807	3,193,797	3,887,525	7,269,039
Number of reads in targeted regions	435,050	790,237	852,740	1,453,922	4,859,621
Percentage reads in targeted regions	16.77%	32.32%	26.70%	37.40%	66.85%
Percent of genome targeted	0.02%	0.02%	0.02%	0.02%	0.02%
Enrichment in targeted regions	672.97	1,296.76	1,071.16	1,500.43	2,682.08
Average read depth	20.19	36.67	39.57	67.46	225.38
Average read depth of controls at 20 bp	7.68	13.32	12.45	28.15	72.87
Normalized depth (per million 32-base reads)	6.93	13.35	11.03	15.45	27.62
Percentage of 1,287 targeted regions with at least:					
1 read	99.69%	100.00%	99.69%	99.92%	99.84%
5 reads	98.76%	99.38%	99.15%	99.30%	99.69%
10 reads	98.21%	98.83%	98.68%	98.99%	99.38%
20 reads	97.44%	97.98%	98.14%	98.37%	98.83%
Percentage of 774,621 targeted bases covered by at least:					
1 read	96.22%	97.52%	97.36%	97.83%	98.37%
5 reads	86.07%	93.39%	92.54%	95.27%	97.25%
10 reads	72.03%	87.89%	85.85%	92.06%	96.07%
20 reads	45.62%	74.38%	71.77%	84.75%	94.02%
30 reads	24.87%	59.13%	57.99%	76.63%	92.00%
40 reads	11.03%	43.20%	44.48%	68.66%	89.92%

Table 1. Experimental Results

References

1. Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661-678.
2. Korbel, J.O. et al. 2007. Paired-End Mapping Reveals Extensive Structural Variation in the Human Genome Science 318 (5849): 420-6.
3. Porecca, G.J. et al. 2007. Multiplex amplification of large sets of human exons. Nat. Methods 4:931-6.
4. Okou, D.T. et al. 2007. Microarray-based genomic selection for high-throughput resequencing. Nat. Methods 4:907-9.
5. Albert, T.J. et al. 2007. Direct selection of human genomic loci by microarray hybridization. Nat. Methods 4:903-905.
6. Hodges, E. et al. 2007. Genome-wide in situ exon capture for selective resequencing Nat. Genet. 39:1522-1527.

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

Printed in the USA, July 13, 2009

5989-8700EN

Agilent Technologies