



Sequence Capture



Technical Note

Comparison of Enrichment Technologies for Targeted Resequencing of Custom Regions



Daniel Burgess, PhD
Senior Scientist, Sequence Capture
Roche NimbleGen Research and Development

Contributing Authors:

Brian Thomson, Mark D'Ascenzo
Roche NimbleGen Research and Development, Madison, Wisconsin, USA

Introduction

The wide gap between the high-throughput capacity of the next-generation sequencing platforms and our ability to utilize them efficiently for most research applications has been an important driver of technological advances with next-generation sequencers. A hypothetical example of the basic problem is a lab working to identify a novel disease-causing mutation localized to a 5 Mb contiguous region of a chromosome. Without good functional candidate genes in the region, the options are severely limited. Technically, it may be possible to search for mutations in a 5 Mb region using long range PCR, coupled with second generation sequencing, but the high cost of reagents and materials would make it impractical. Whole-genome sequencing would also provide the necessary raw data, but ~99.8% of the resulting sequence would originate outside of the target, and so this approach is also extremely inefficient, time consuming, and expensive. The same challenge applies where the mutation suspect is not a positional candidate gene, but instead a large number of functional candidate genes dispersed throughout the genome.

To address this gap, methods have been developed for enriching targeted portions of the genome as a sample preparation step to next generation sequencing¹⁻⁵. These methods generally rely on hybridization of genomic DNA libraries to large pools of complementary oligonucleotide probes from DNA or RNA that are attached to a microarray surface or another type of retrievable substrate (e.g. biotin-streptavidin beads). The hybridized complexes are washed to remove nonspecific library fragments, then amplified and sequenced, resulting in a much higher proportion of sequence reads originating from the region of interest. Here, we examine the ability of two of these technologies, NimbleGen Sequence Capture and Agilent SureSelect technologies, to enrich a typical set of mutational candidate genes for resequencing: the coding exons of chromosome X.

Results

We compared the performance between the Agilent SureSelect Human Chromosome X Kit (Part Number G4459A) and the NimbleGen Sequence Capture 385K array by performing enrichments for chromosome X exons according to each of the manufacturer's instructions. To facilitate the comparison, all enrichments were performed using the same Illumina single-end-read library constructed from genomic DNA (HapMap NA12878). Two lanes of Illumina Genome Analyzer IIx (40 bp single-read) sequence data were analyzed for each experiment. The purity filtered (PF) sequence reads were randomly subsampled to match the smaller of the two datasets (~3.3 M reads) so that an identical number of reads could be analyzed for each experiment. To account for differences between the platforms in probe number, size, and placement for purposes of comparing target coverage and SNP identification statistics, the defined target for both experiments was the sequence contained within chromosome X coding exons that had at least 1 bp overlap with both NimbleGen and Agilent probes (HG18 coordinates).

NimbleGen Sequence Capture generates more unique sequence aligned to the target

The SureSelect experiment yielded a smaller number of PF reads than the Sequence Capture experiment, however the fraction of reads aligned to genome and the error rates reported by the Genome Analyzer IIx output report were very similar (Table 1), reducing the likelihood that variation in sequence quality would significantly affect the comparison. All available PF reads were filtered for uniqueness by removing reads with identical start coordinates and sequences (i.e. reads not clearly derived from distinct captured molecules). After aligning to Chromosome X and removing duplicates, a much smaller fraction of all available PF reads remained for the SureSelect experiment (14%) compared to the Sequence Capture experiment (26%). Because the same library preparation was divided and used for both experiments, this factor cannot account for the reduced complexity in the SureSelect enriched library. Instead, it is possible that one or more intrinsic differences between the enrichment processes may have contributed to this result.

Sequence, Mapping and Coverage Statistics		
Platform	NimbleGen Sequence Capture 385K Array	Agilent SureSelect
Sequencing Output Summary		
Clusters (PF)	35,847,466	23,692,988
Yield (Gbp)	1.43	0.95
Total Reads Aligned to Genome	29,997,656	19,452,951
Total Reads Aligned to Genome/Clusters (PF)	84%	82%
Error Rate	0.31%	0.32%
Mapping Statistics (using all available PF reads)		
Unique Reads Aligned to Chromosome X	9,483,382	3,333,734
Unique Reads Aligned to Chromosome X/Clusters (PF)	26%	14%
Mapping Statistics (subsample of 3,333,734 unique PF reads aligned to Chromosome X)		
Common Targeted Exons	7,163	7,163
Percent Reads Aligned to Common Targeted Exons	67%	68%
Targeted Bases (from common targeted exons)	2,639,051	2,639,051
Percent Targeted Bases Covered with Read Depth ≥1X	98%	91%
Average Read Depth over Targeted Bases	30.5	31.5
Median Read Depth over Targeted Bases	28	27

▲ **Table 1: Sequence, mapping and coverage statistics.** Captured (enriched) single-read genomic DNA libraries were generated using NimbleGen Sequence Capture or Agilent SureSelect methods and sequenced using two lanes of Illumina Genome Analyzer IIx 40 bp single-read sequencing. Data listed under the heading "Sequencing Output Summary" were taken from the Illumina Genome Analyzer II Read Report. The reads were filtered to remove possible duplicates (i.e. reads with identical start coordinates and sequences) and aligned to the genome (HG18) using Bowtie (version 0.10.0)⁶ to generate the remaining statistics.

The uniformity of sequence coverage is higher with NimbleGen Sequence Capture

To remove enriched library complexity as a factor in other performance metrics, mapping statistics were calculated using an identical number of unique, chromosome X aligned, reads (n = 3,333,734) obtained by random subsampling reads from the larger data set (Table 1). The percent of these reads that aligned to the exon target with at least 1 bp overlap were nearly identical for SureSelect (68%) and Sequence Capture (67%) experiments; however, the distribution of their reads was significantly different. The aligned reads from the Sequence Capture experiment generated an average read depth of 30.5 over the target, with 98% of the target bases covered by at least one read. The aligned reads from the SureSelect experiment generated a similar average read depth of 31.5, but only covered 91% of the target bases. For the Sequence Capture experiment

88.4% of target bases were covered to a read depth of at least 10, while only 81.1% were covered to this depth with the SureSelect experiment (Figure 1). Thus, the sequence coverage from the SureSelect experiment was less uniformly distributed than the data from the Sequence Capture experiment, with a greater fraction of target bases covered to depths farther from the mean.

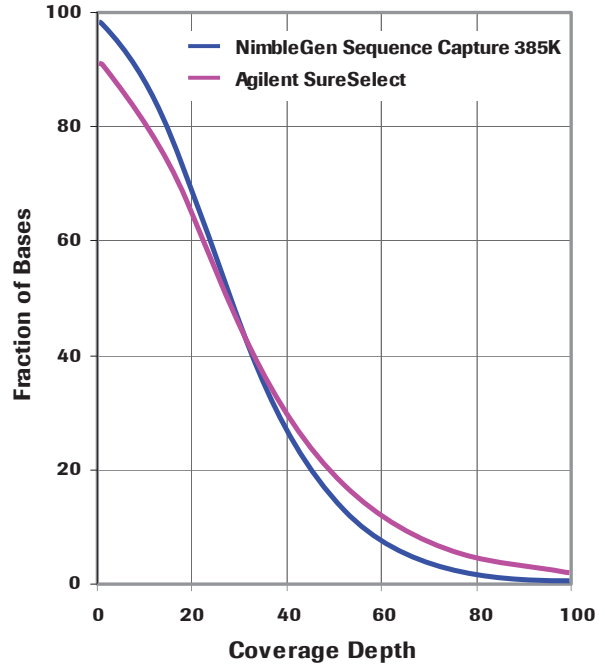
NimbleGen Sequence Capture recovers more known SNPs than Agilent SureSelect

Using the same number of unique, genomically aligned reads, SNPs were called and compared with the list of known HapMap SNPs for this sample to determine concordance rates for both platforms (Table 2). Thirteen of 352 (3.7%) known heterozygous SNPs were missed (i.e. false negative or misclassified) by the NimbleGen Sequence Capture platform, while SureSelect failed to identify 48 (13.6%) of known heterozygous SNPs. NimbleGen Sequence Capture correctly identified 94.9% of known homozygous SNPs in the sample, while SureSelect correctly identified 88.2%.

Conclusions

The NimbleGen Sequence Capture 385K array performed better than the Agilent SureSelect platform for enriching a large set of chromosome X exons prior to next-generation sequencing and SNP discovery. This reason for the difference in performance can be attributed to the greater uniformity of sequence coverage achieved by the Sequence Capture method. Uniformity of sequence coverage is a particularly important metric for enrichment methods because it defines the amount of raw sequence required to achieve a minimum coverage depth over the target. Minimum coverage, in turn, is critical for calling SNPs with high confidence.

The performance of NimbleGen Sequence Capture array in this experiment could be due to differences in: 1. The stability or hybridization kinetics of NimbleGen long DNA oligonucleotide probes compared with Agilent RNA probes; 2. Optimized probe selection and design algorithms; 3. The greater number of unique probe sequences available on NimbleGen Sequence Capture 385K arrays; 4. Other differences between the protocols, or some combination of these. These factors will be important topics for future investigation, as targeted enrichment methods become an increasingly important tools in genomic research.





▲ **Figure 1: Sequence coverage uniformity.** A plot of sequence coverage over target bases shows that the NimbleGen Sequence Capture 385K array produced more uniform coverage over a greater fraction of target bases than the Agilent SureSelect platform. The plotted data begins at a coverage depth of 1 and is arbitrarily truncated at a coverage depth of 100 to facilitate display.

SNP Identification Statistics		
Platform	NimbleGen Sequence Capture 385K Array	Agilent SureSelect
Known SNPs:	589	589
Known Heterozygote SNPs:	352	352
Known Homozygote SNPs:	237	237
Heterozygote True Positive:	339	304
Heterozygote False Positive:	1,454	1,551
Heterozygote False Negative:	9	45
Heterozygote Misclassified:	4	3
Homozygote True Positive:	225	209
Homozygote False Positive:	2,007	1,756
Homozygote False Negative:	10	23
Homozygote Misclassified:	2	5
Heterozygote True Positive Rate:	96.3%	86.4%
Homozygote True Positive Rate:	94.9%	88.2%

▲ **Table 2: SNP discovery rates.** SNPs were called from the aligned sequence data using SOApsnp (Release 1.03)⁷ and compared to the known SNPs for sample NA12878 (HapMap database). SNPs were called heterozygotes if they exhibited allele frequencies between 0.2-0.8. The SNP data are not filtered for base call quality scores or read depth.

References

1. Albert, TJ et al. "Direct selection of human genomic loci by microarray hybridization," *Nat Methods*: 4(11):903-5, 2007.
2. Gnirke, A et al. "Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing," *Nat Biotechnol*.:27(2):182-9, 2009.
3. Hodges, E et al. "Genome-wide in situ exon capture for selective resequencing," *Nat Genet*.:39(12):1522-7, 2007.
4. Okou, DT et al. "Microarray-based genomic selection for high-throughput resequencing," *Nat Methods*:4(11):907-9, 2007.
5. D'Ascenzo, M et al. "Mutation discovery in the mouse using genetically guided array capture and resequencing," *Mamm Genome*:20(7):424-36, 2009.
6. Langmead, B et al. "Ultrafast and memory efficient alignment of short DNA sequences to the human genome," *Genome Biol*.:10(3):R25, 2009.
7. Li, R et al. "SNP detection for massively parallel whole-genome resequencing," *Genome Res*.: 19(6):1124-32, 2009.

Array Specs	2.1M Array	385K Array
		
Total features	2.1 million	385,000

Ordering Information

Product	D Delivery Cat. No.	S Service Cat. No.
NimbleGen Custom Sequence Capture 385K Array	05 394 538 001	05 394 546 001
NimbleGen Custom Sequence Capture 2.1M Array	05 329 841 001	—

D = Delivery Array (Array is delivered to you. You run the experiment.)

S = Service Array (Data are delivered to you. Roche NimbleGen runs the experiment as service.)

Microarray Processing Accessories

Reagents	Cat. No.
NimbleGen Sequence Capture Hybridization Kit	05 340 721 001
NimbleGen Sequence Capture Wash and Elution Kit	05 340 730 001
Equipment	Cat. No.
NimbleGen Hybridization System 4 (110V)	05 223 652 001
NimbleGen Hybridization System 12 (110V)	05 223 679 001
NimbleGen Hybridization System 4 (220V)	05 223 687 001
NimbleGen Hybridization System 12 (220V)	05 223 695 001
NimbleGen Sequence Capture Elution System	05 329 752 001

Roche Microarray Technical Support:
www.nimblegen.com/arraysupport

Published by:
 Roche Diagnostics GmbH
 Roche Applied Science
 Werk Penzberg
 82372 Penzberg
 Germany

www.roche-applied-science.com

© 2010 Roche Diagnostics GmbH
 All rights reserved.

For life science research only. Not for use in diagnostic procedures

NIMBLEGEN is a trademark of Roche.
 Other brands or product names are trademarks of their respective holders.