

# Comparison of Custom Target Enrichment Methods for Next Generation Sequencing with the Illumina Platform

Anoja Perera, Scottie Adams, David Bintzler, Kip Bodi, Ken Dewar, Deborah Grove, Jan Kieleczawa, Robert Lyons, Tom Neubert, Aaron Noll, Sushmita Singh, Robert Steen, Michael Zianni

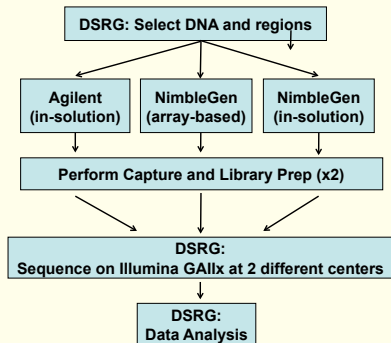
## OVERVIEW

Targeted sequence capture followed by Illumina sequencing is a cost-effective method of finding variations in regions of interest without sequencing an entire genomic sample. We compared three targeted sequence capture technologies: Agilent SureSelect XT, NimbleGen Sequence Capture Arrays, and NimbleGen SeqCap EZ Choice. Sequence data was obtained and analyzed for capture sensitivity, specificity, coverage per region, and capacity for SNP detection.

## INTRODUCTION

The size of the human genome is 3.4 billion base pairs. Using an Illumina HiSeq 2000, one run will generate 200 gigabases of data at a cost of \$10,000, equivalent to one human genome. However, for many applications, only certain regions are of interest, such as coding regions, the exome, candidate genes, or for GWAS. Several vendors currently offer whole exome capture, but currently only NimbleGen and Agilent offer custom targeted sequencing solutions. We designed a custom targeted capture for the hg19 genome and submitted the design to both NimbleGen and Agilent.

## STUDY DESIGN

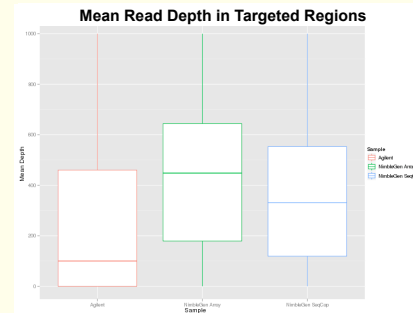
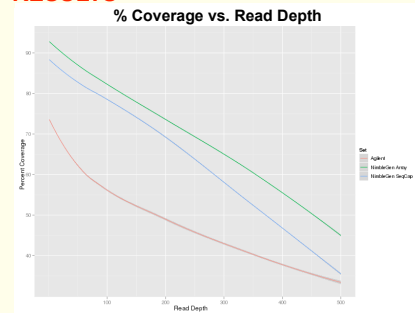


Each company performed two separate capture experiments for each technology. Libraries were checked for quality on an Agilent High Sensitivity chip, and then loaded in equimolar concentrations on an Illumina GAIIx paired-end flowcell at two sequencing centers. Two lanes were loaded per technology.

## ANALYSIS

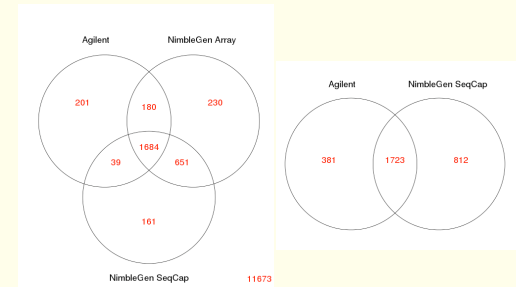
The resulting sequence data from each sample was filtered so that 100% of the bases had a quality score of more than 10. Reads were mapped against the hg19/GRCh37 genome using "bowtie 0.12.7" and the sets were normalized to the total number of mapped reads. Sequences aligning to more than one region and duplicate read pairs were discarded. A series of perl scripts were written to calculate the coverage per position for every targeted region, creating a coverage map. Coverage maps were then imported into the "R statistical computing environment 2.12.1" to find the sensitivity, specificity, and reproducibility for each sample. Plots and figures were generated using the "ggplot2" library.

## RESULTS

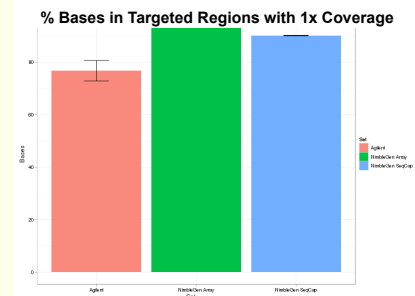


## SNP DETECTION

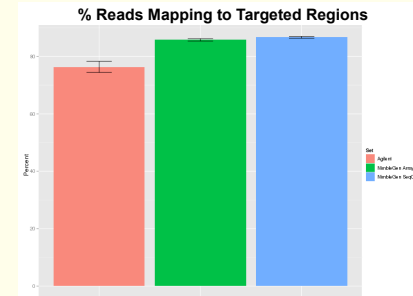
We used "samtools" and "bcftools" to generate a list of high-quality (depth >= 5, Q >=20) SNPs for each sample. Found SNPs were compared with the dbSNP human SNP database and each sample had ~98% agreement. The number of on-target and off-target SNPs for each sample was then compared.



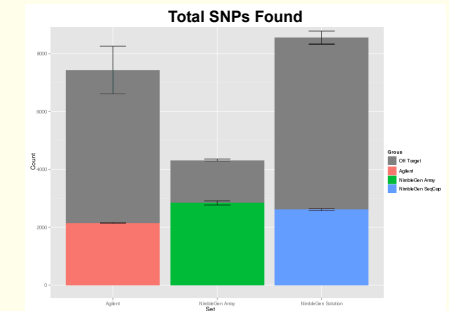
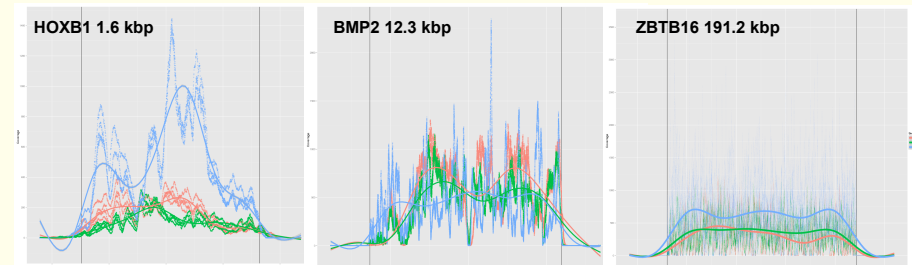
## SENSITIVITY



## SPECIFICITY



## COVERAGE PER REGION



## CONCLUSION

NimbleGen Methods performed best overall. For SNP detection, NimbleGen array-based method performed better than both in-solution methods. However, if experiments involve large sample numbers, in-solution methods are automation friendly and hence less tedious.

## ACKNOWLEDGMENTS

The DSRG thanks Alexander Wong, Fed Ermani, Garick Peters, Katie Weaver, Ken Olinger, and Nick Mapara from Agilent; Dan Burgess, Lance Brown, Michael Frawley, and Xinmin Zhang from NimbleGen; Jonathan Pinter from Illumina; Christine Brennan, Elizabeth Ketterer, Kendra Walton, Madelaine Gogol, Constance Esposito, and Karen Staehling.