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

February 19-22

San Antonio, TX

ABRF

DNA Sequencing Research Group

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

Why Perform Region Capture?

- Better suitable for some studies, such as gene testing, GWAS etc.
  - Where information of the whole genome is unnecessary
  - In order to keep costs low

Size of Human genome = 3.4 billion base pairs

On an Illumina HiSeq,

- A 100bp paired-end run will provide 100-200Gb of data which is sufficient to call mutations
- Per sample cost is about $10,000*
- Will take about 10 days per sample

*Only includes Illumina reagent costs.

2009/10 DSRG study

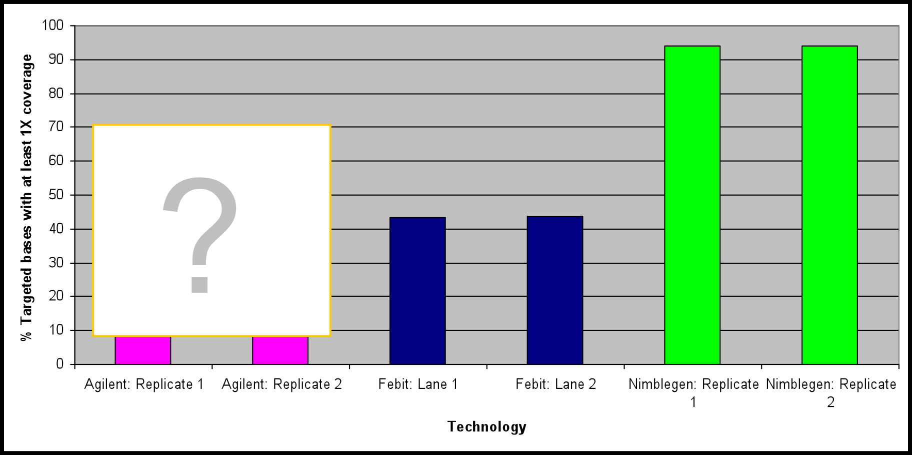
•DNA: ‘The Human Reference Genetic Material Repository DNA Sample’ (Coriell catalog ID: NS12911)   http://huref.jcvi.org/

•Two types of regions selected (total ~3.5Mb):

1. 2Mb continuous region
2. 31 individual genes*

* The genes selected ranged widely in regards to size (2kb to 400kb), exon numbers, GC content, number of transcripts and repetitive nature of the sequences. All companies were provided with ensembl gene IDs and genomic locations.

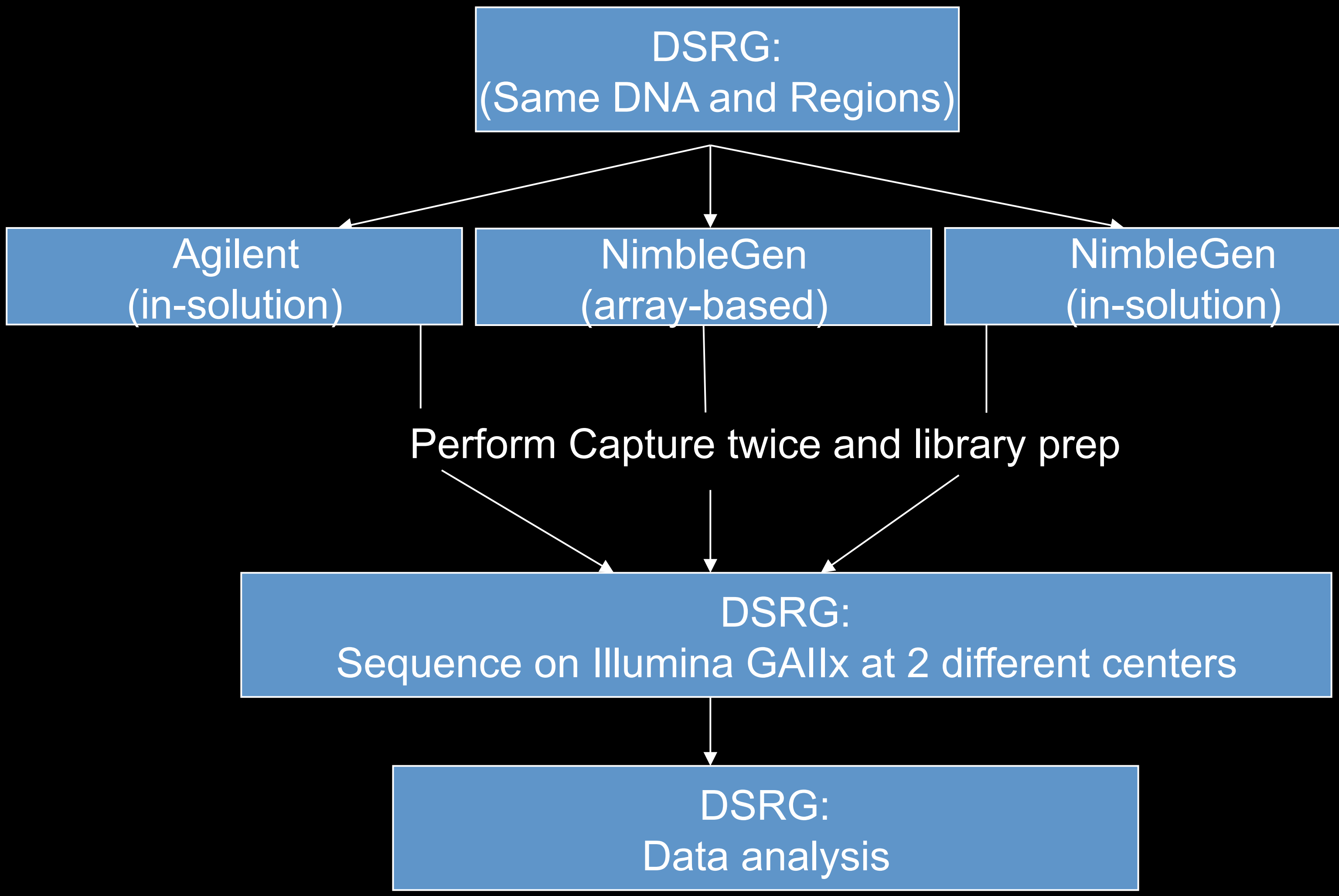
2009/10 DSRG study: Sensitivity



- Agilent accidentally used a different genome build to design the assay
- In addition, NimbleGen introduced a new in-solution capture method

REPEAT STUDY!

2010/11 DSRG study



DSRG:
(Same DNA and Regions)

Agilent
(in-solution)

NimbleGen
(array-based)

NimbleGen
(in-solution)

Perform Capture twice and library prep

DSRG:
Sequence on Illumina GAIIx at 2 different centers

DSRG:
Data analysis

*** Illumina paired-end library prep kits provided to all participants!

Agilent vs. NimbleGen Custom Kits

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Assay design	Customer	Vendor	Vendor

Agilent vs. NimbleGen Custom Kits

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Assay design	Customer	Vendor	Vendor
Design charge?	None	Yes*	None

* Waived if more than 5 assays

Agilent vs. NimbleGen Custom Kits

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Assay design	Customer	Vendor	Vendor
Design charge?	None	Yes*	None
Probe type	RNA	DNA	DNA

* Waived if more than 5 assays

Agilent vs. NimbleGen Custom Kits

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Assay design	Customer	Vendor	Vendor
Design charge?	None	Yes*	None
Probe type	RNA	DNA	DNA
Total size captured¨	<200Kb-6.8 Mb	5-30Mb	100Kb to 50Mb

* Waived if more than 5 assays

¨ Size captured varies per kit

Agilent vs. NimbleGen Custom Kits

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Assay design	Customer	Vendor	Vendor
Design charge?	None	Yes*	None
Probe type	RNA	DNA	DNA
Total size captured¨	<200Kb-6.8 Mb	5-30Mb	100Kb to 50Mb
DNA input amount	1µg	1-3µg	1µg

* Waived if more than 5 assays

¨ Size captured varies per kit

Agilent vs. NimbleGen Custom Kits

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Assay design	Customer	Vendor	Vendor
Design charge?	None	Yes*	None
Probe type	RNA	DNA	DNA
Total size captured¨	<200Kb-6.8 Mb	5-30Mb	100Kb to 50Mb
DNA input amount	1µg	1-3µg	1µg
Cost per sample	~$1000**	~$1000**	~$1000**

* Waived if more than 5 assays

¨ Size captured varies per kit

** Cost varies depending on the size of kit

Agilent vs. NimbleGen Custom Kits

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Assay design	Customer	Vendor	Vendor
Design charge?	None	Yes*	None
Probe type	RNA	DNA	DNA
Total size captured¨	<200Kb-6.8 Mb	5-30Mb	100Kb to 50Mb
DNA input amount	1µg	1-3µg	1µg
Cost per sample	~$1000**	~$1000**	~$1000**
Multiplexing?	Yes	No	2Q

* Waived if more than 5 assays

¨ Size captured varies per kit

** Cost varies depending on the size of kit

Agilent vs. NimbleGen Custom Kits

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Assay design	Customer	Vendor	Vendor
Design charge?	None	Yes*	None
Probe type	RNA	DNA	DNA
Total size captured¨	<200Kb-6.8 Mb	5-30Mb	100Kb to 50Mb
DNA input amount	1µg	1-3µg	1µg
Cost per sample	~$1000**	~$1000**	~$1000**
Multiplexing?	Yes	No	2Q
Automation friendly?	Yes	No	Yes

* Waived if more than 5 assays

¨ Size captured varies per kit

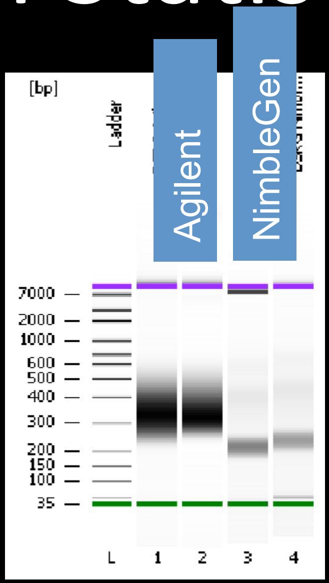
** Cost varies depending on the size of kit

Kits Used in the DSRG Study

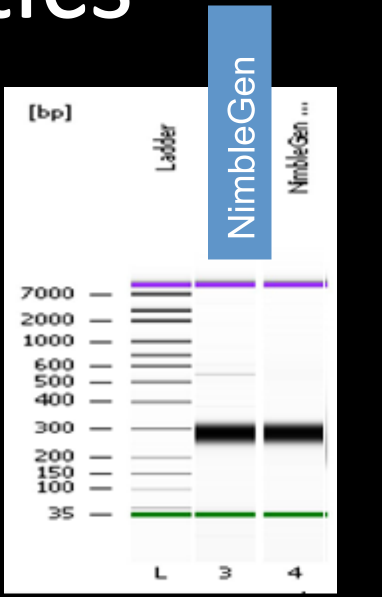
- Agilent SureSelectXT Custom MP0 (3.0Mb-6.8Mb) Kit (in-solution)
- NimbleGen SeqCap EZ Choice (in-solution)
- NimbleGen Sequence Capture Arrays

QC and Illumina Run Statistics

- Each company was asked to perform the capture twice so we can look at reproducibility
- Ran all libraries on the Agilent High Sensitivity chip
- Samples were loaded in equal nM concentrations on an Illumina paired-end flowcell at two different centers
- Two lanes were loaded per technology

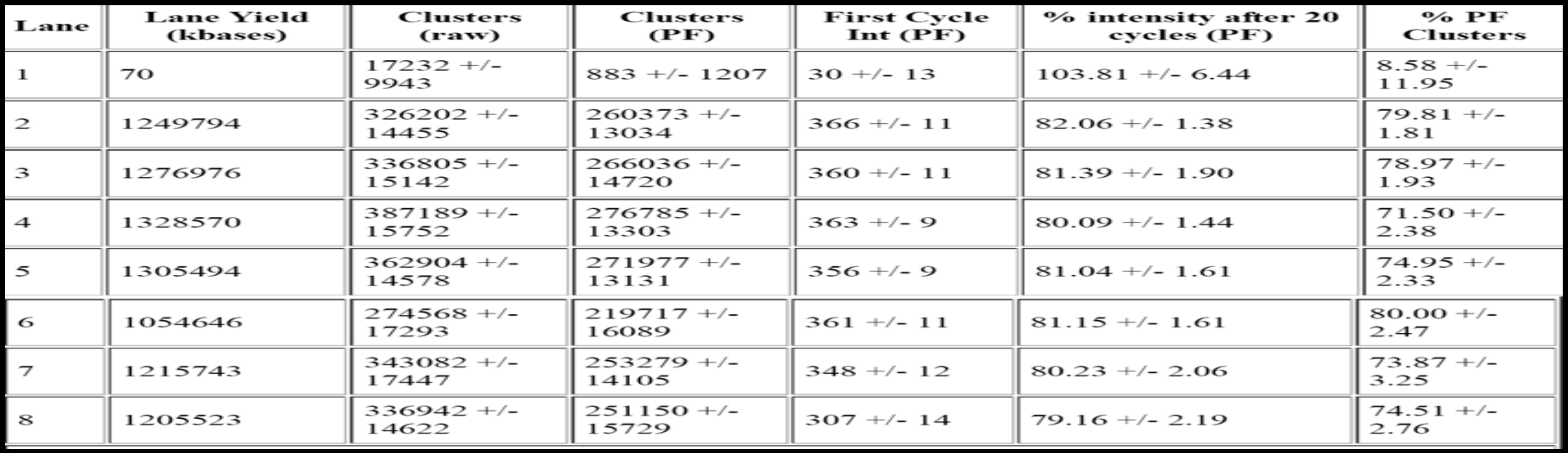


In-solution



Array-based

Illumina Primary Analysis



Lane	Lane Yield (kbases)	Clusters (raw)	Clusters (PF)	First Cycle Int (PF)	% intensity after 20 cycles (PF)	% PF Clusters
1	70	17232 +/- 9943	883 +/- 1207	30 +/- 13	103.81 +/- 6.44	8.58 +/- 11.95
2	1249794	326202 +/- 14455	260373 +/- 13034	366 +/- 11	82.06 +/- 1.38	79.81 +/- 1.81
3	1276976	336805 +/- 15142	266036 +/- 14720	360 +/- 11	81.39 +/- 1.90	78.97 +/- 1.93
4	1328570	387189 +/- 15752	276785 +/- 13303	363 +/- 9	80.09 +/- 1.44	71.50 +/- 2.38
5	1305494	362904 +/- 14578	271977 +/- 13131	356 +/- 9	81.04 +/- 1.61	74.95 +/- 2.33
6	1054646	274568 +/- 17293	219717 +/- 16089	361 +/- 11	81.15 +/- 1.61	80.00 +/- 2.47
7	1215743	343082 +/- 17447	253279 +/- 14105	348 +/- 12	80.23 +/- 2.06	73.87 +/- 3.25
8	1205523	336942 +/- 14622	251150 +/- 15729	307 +/- 14	79.16 +/- 2.19	74.51 +/- 2.76

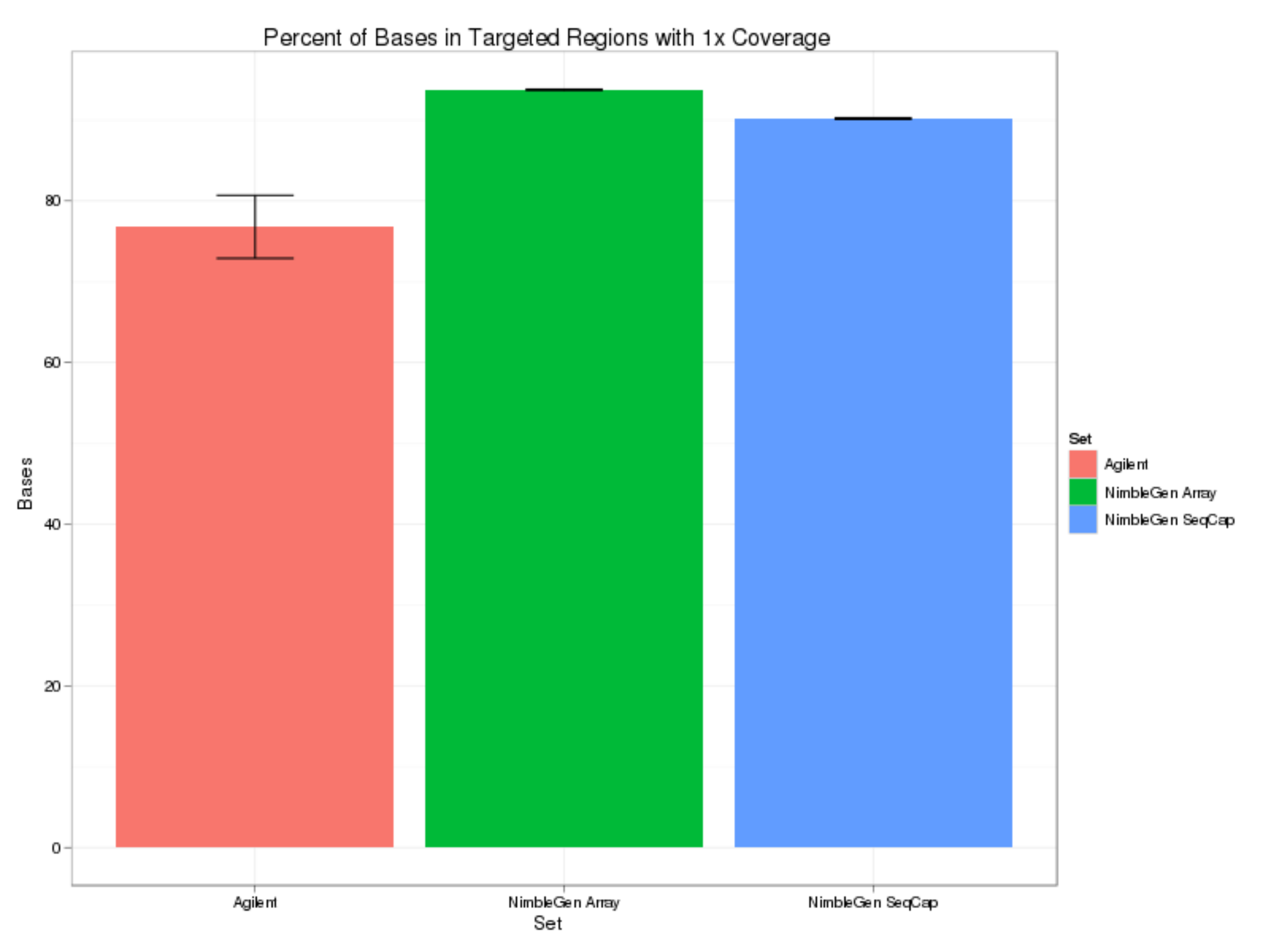
Elizabeth Ketterer, Kendra Walton

Data Analysis

1. Filtered each data set so that sequence quality score > 10 for 100% of the bases
2. Mapped reads against the hg19/GRCh37 genome using “bowtie 0.12.7”
3. Normalized the data sets to equal sizes
4. ‘perl’ scripts used to calculate coverage per position in every targeted region, creating a coverage map
5. Coverage maps imported into the “R statistical computing environment (2.1.0)”, to find the sensitivity, specificity, and reproducibility for each sample
6. Plots and figures generated using the "ggplot2" library and MS Excel

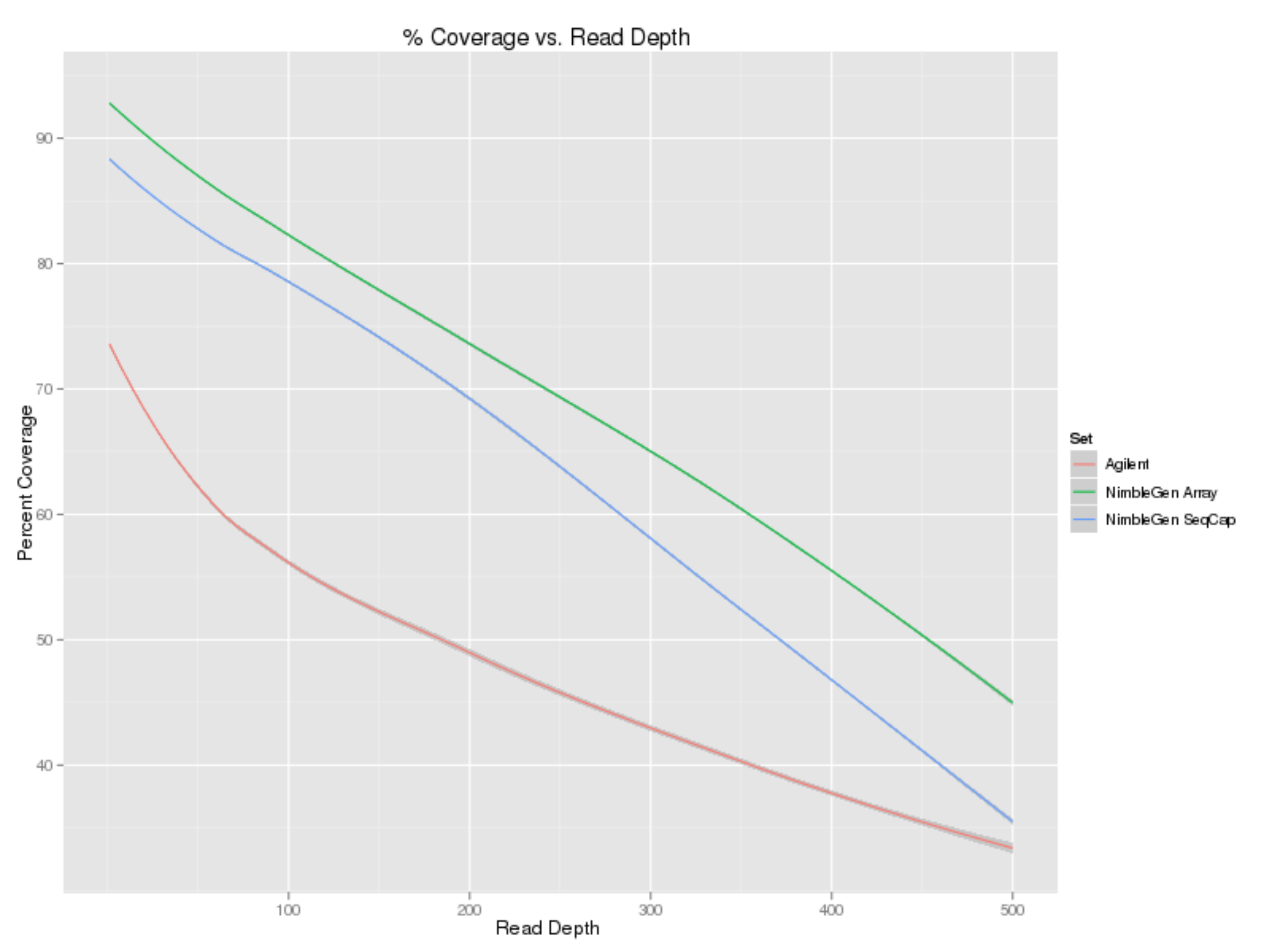
Kip Bodi

% Coverage of 3.5 Mb region by at least 1x



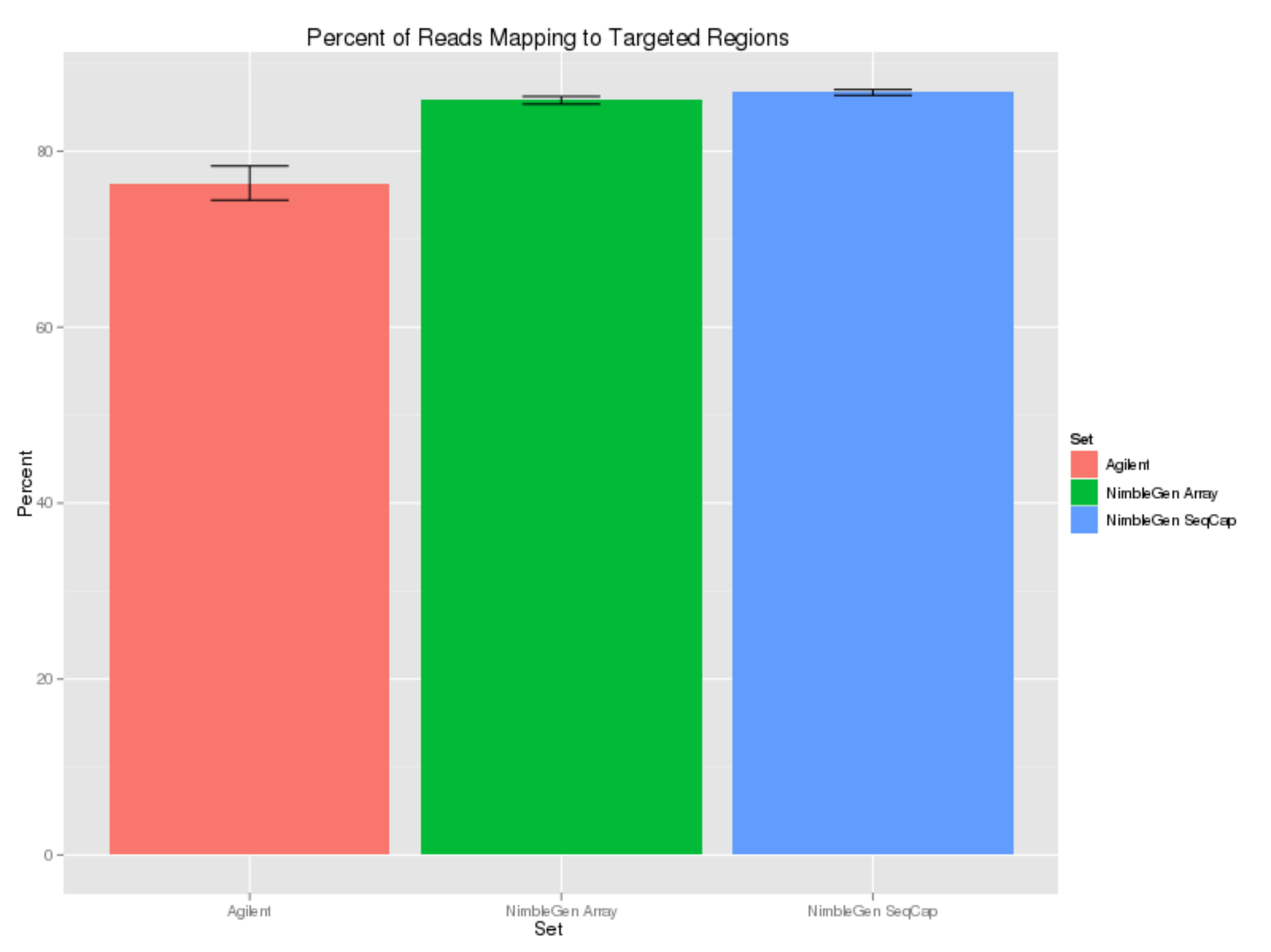
Kip Bodi

% Coverage vs. Read Depth



Kip Bodi

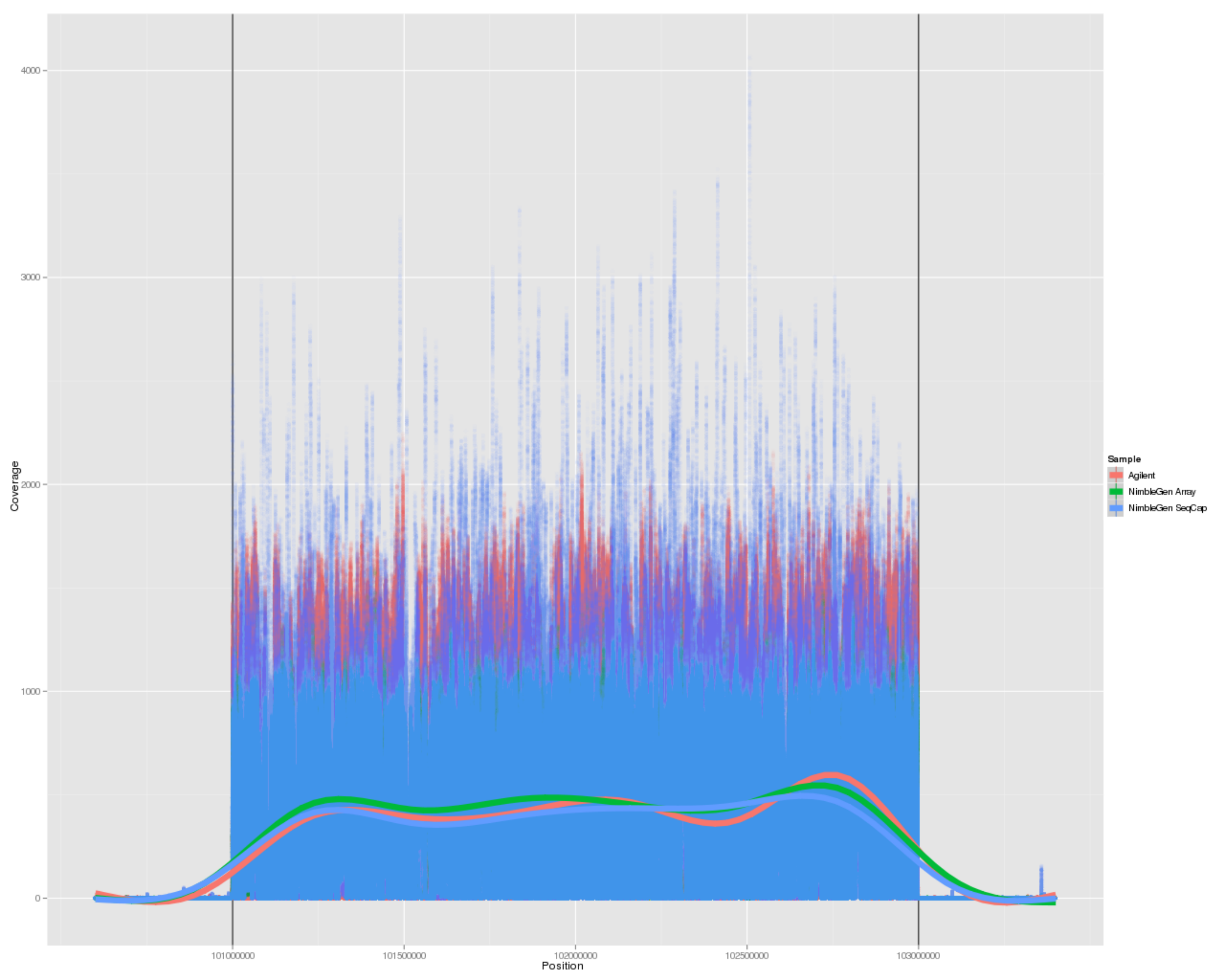
% Reads Mapping to Target (On Target)



Kip Bodi

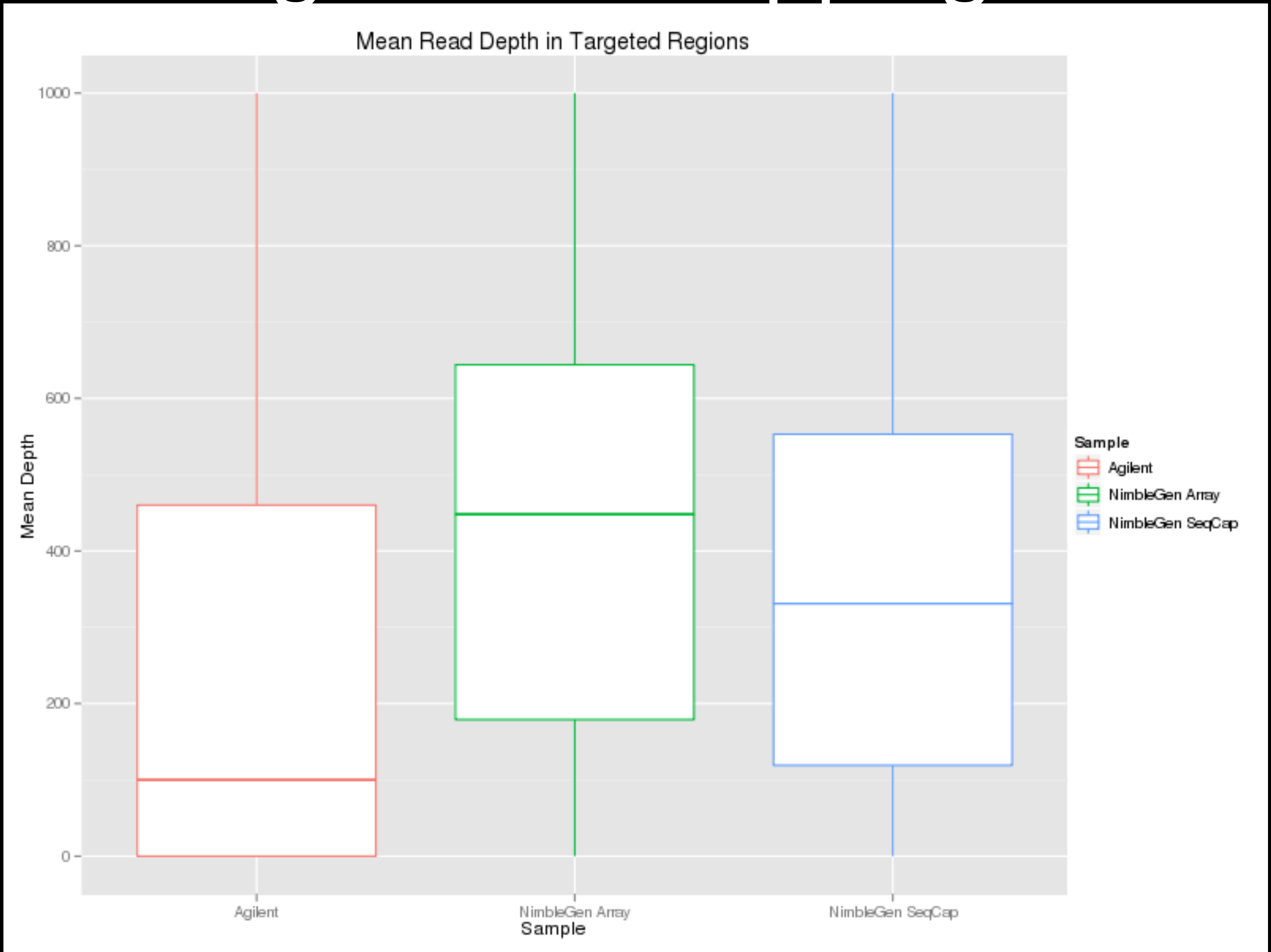
* Adding 100bp to the co-ordinates did not significantly change the % of on target reads

Coverage of the 2Mb Continuous Region



Kip Bodi

Coverage of Overlapping Genes



Kip Bodi

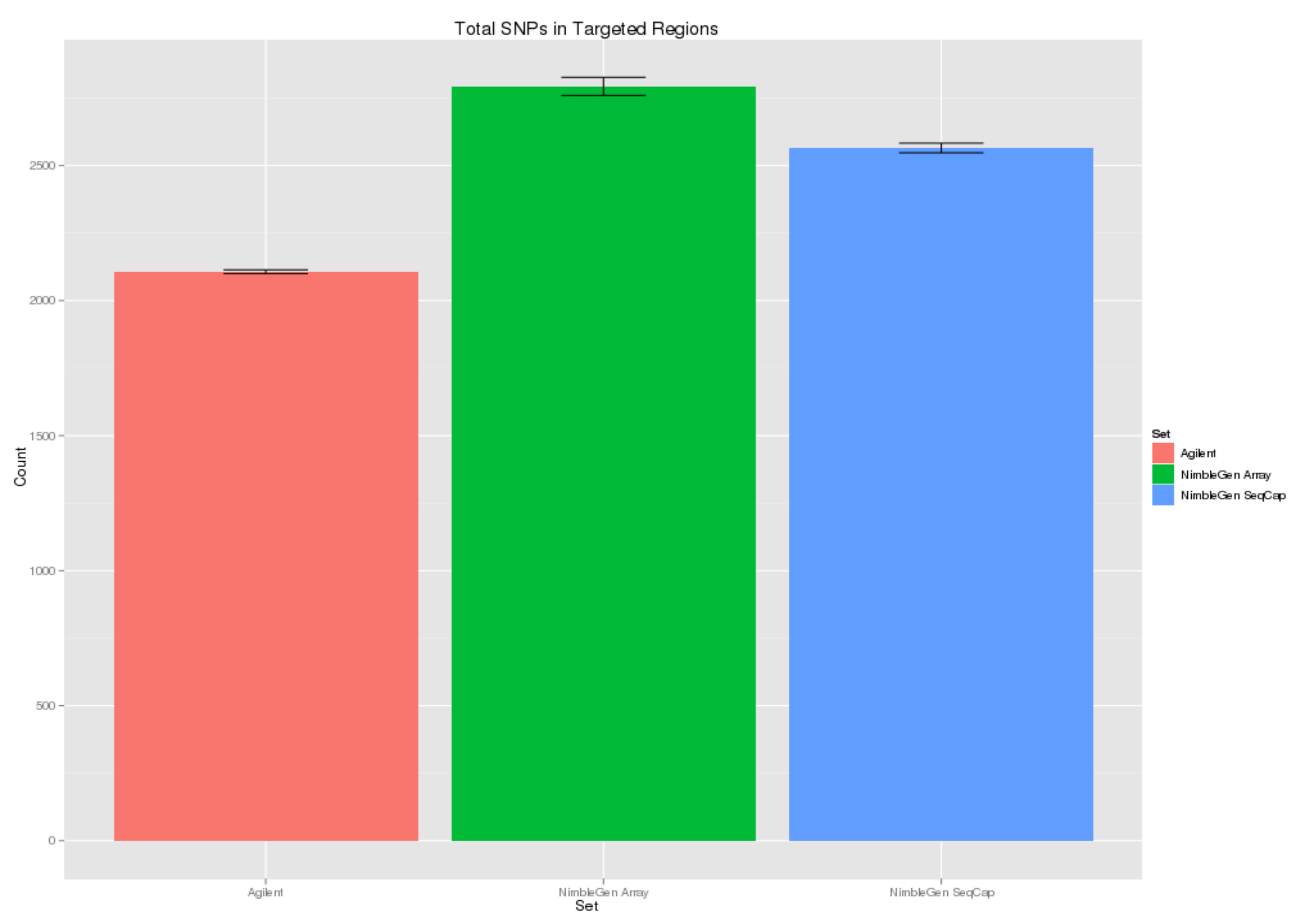
SNP Detection

1. SNP detection was performed by downloading dbSNPs (NCBI) for the regions of interest

2. “samtools” and “bcftools” to generate a list of high quality SNPs (depth >= 5, Q>=20)

3. Then every SNP was compared to the dbSNP list to see if the position and mutation was present in our region of interest (ROI)

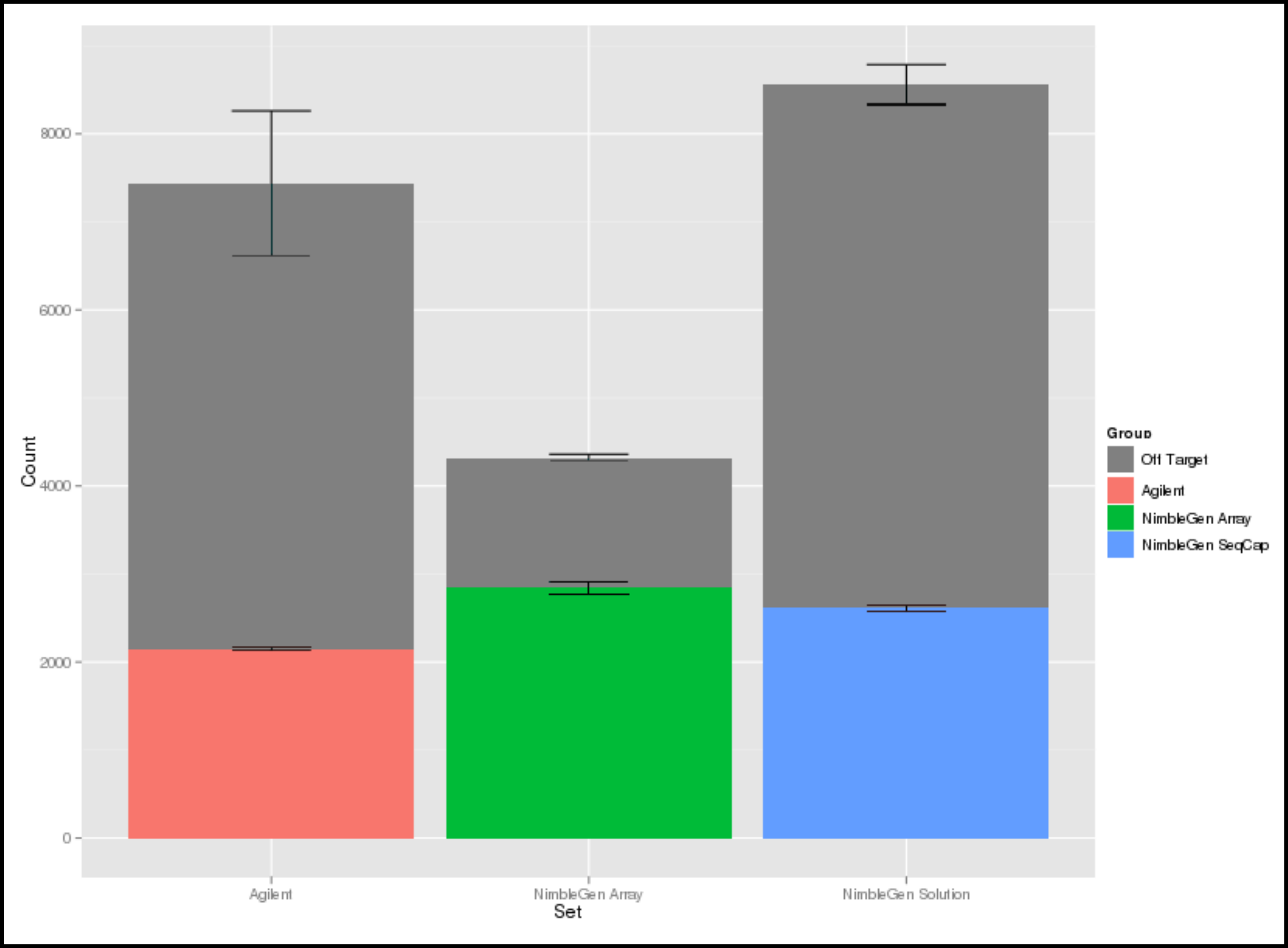
Total SNPs in the Targeted Regions



Of the SNPs found, ~ 98% matched to dbSNPs for each technology

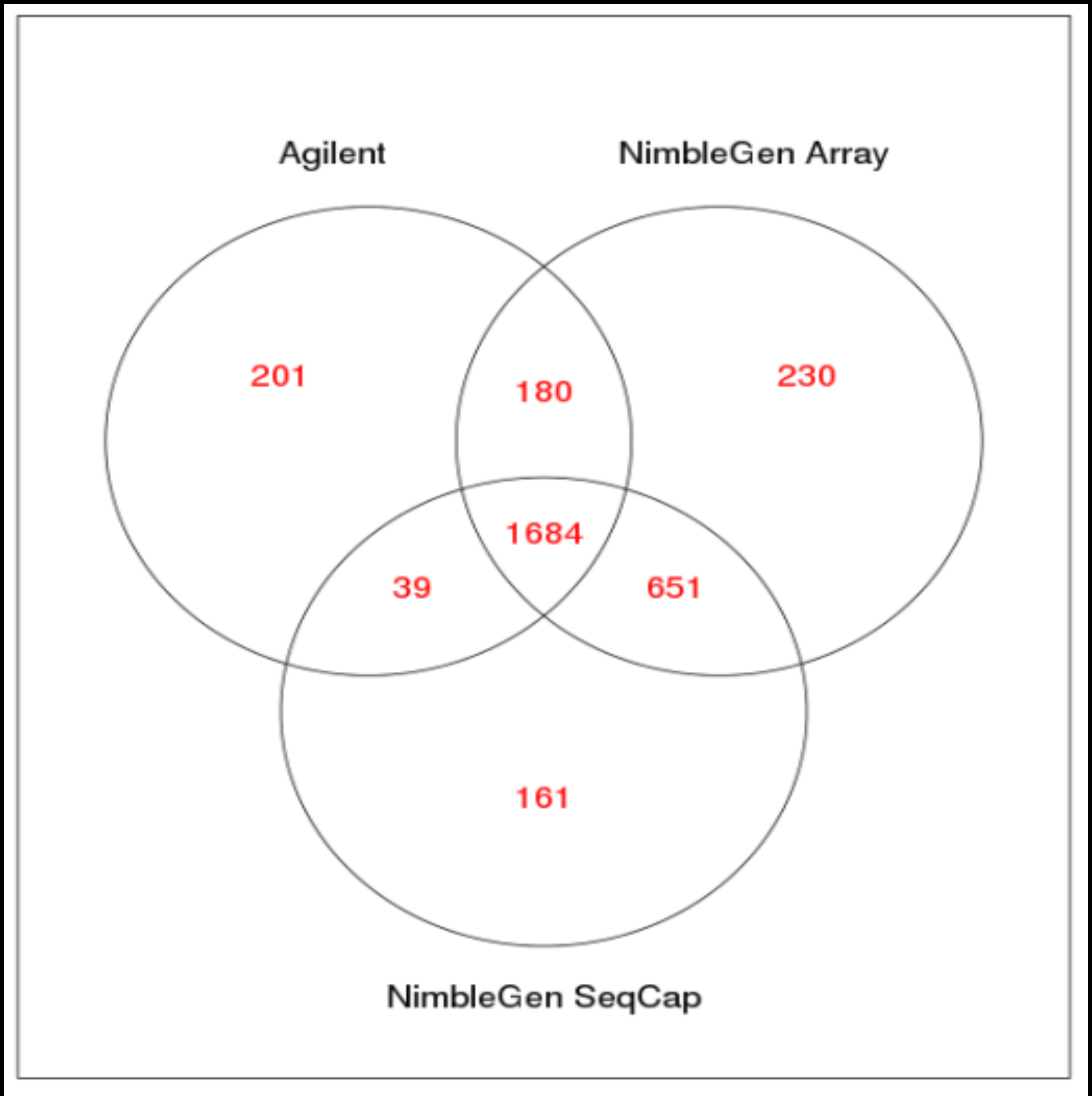
Kip Bodi

% on Target SNPs Compared to % all SNPs in the Data Set



Kip Bodi

Overlap of SNP counts



In Summary…

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Cost per sample	X		X

In Summary…

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Cost per sample	X		X
Sample input requirement	X	X	X

In Summary…

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Cost per sample	X		X
Sample input requirement	X	X	X
Quality of captured sample	X	X	X

In Summary…

Description	Agilent	NimbleGen	NimbleGen
	In-solution	Array	In-solution
Cost per sample	X		X
Sample input requirement	X	X	X
Quality of captured sample	X	X	X
Reproducibility*	X	X	X

* Data not shown due to time restraints

In Summary…

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Cost per sample	X		X
Sample input requirement	X	X	X
Quality of captured sample	X	X	X
Reproducibility*	X	X	X
% Coverage		X	X

* Data not shown due to time restraints

In Summary…

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Cost per sample	X		X
Sample input requirement	X	X	X
Quality of captured sample	X	X	X
Reproducibility*	X	X	X
% Coverage		X	X
On target			X

* Data not shown due to time restraints

In Summary…

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Cost per sample	X		X
Sample input requirement	X	X	X
Quality of captured sample	X	X	X
Reproducibility*	X	X	X
% Coverage		X	X
On target			X
SNP detection		X	

* Data not shown due to time restraints

In Summary…

Description	Agilent In-solution	NimbleGen Array	NimbleGen In-solution
Cost per sample	X		X
Sample input requirement	X	X	X
Quality of captured sample	X	X	X
Reproducibility*	X	X	X
% Coverage		X	X
On target			X
SNP detection		X	
Scalability	X		X

* Data not shown due to time restraints

In Summary

- NimbleGen methods performed best in this study.
- 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.

Future Directions

Examine in detail:

- Where the off-target reads are mapping to
- Why the SNP counts are higher for the in-solution methods but lower for on-target regions
- Whether there are allelic biases
- Ability to call indels and CNVs with each product

Many Thanks!!!

Agilent/Vector Biotech

Alexander Wong
Fred Ernani
Garrick Peters
Katie Weaver
Ken Olinger
Nick Mapara

NimbleGen

Daniel Burgess
Lance Brown
Michael Frawley
Xinmin Zhang

Illumina

Jonathan Pinter

Data Analysis:

Kip Bodi

Others:

Christine Brennan
Elizabeth Ketterer
Kendra Walton
Madelaine Gogol
Constance Esposito
Karen Staehling

DSRG

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

ABRF The Association of Biomolecular Resource Facilities