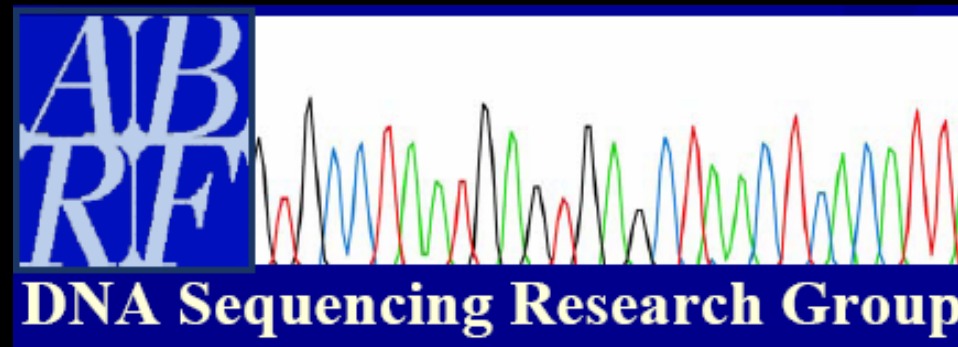


Comparison of Commercially Available Target Enrichment Methods for Next Generation Sequencing with the Illumina Platform

March 20-23, 2010

Sacramento, CA



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

Why do capture?

Using Illumina HiSeq 2000:

1 run = 10 days = 200 Gb => 1 Human Genome at a cost of ~\$10,000



- GWAS
- Exome sequencing
- Candidate gene sequencing



Capture Methods



Many different methods...

1. Multiplex PCR



Some key features:

- High specificity and coverage
- 200-20,000 primer pairs target 10Kb to 10Mb
- Genomic DNA input: 2 μ g
- Requires capital equipment

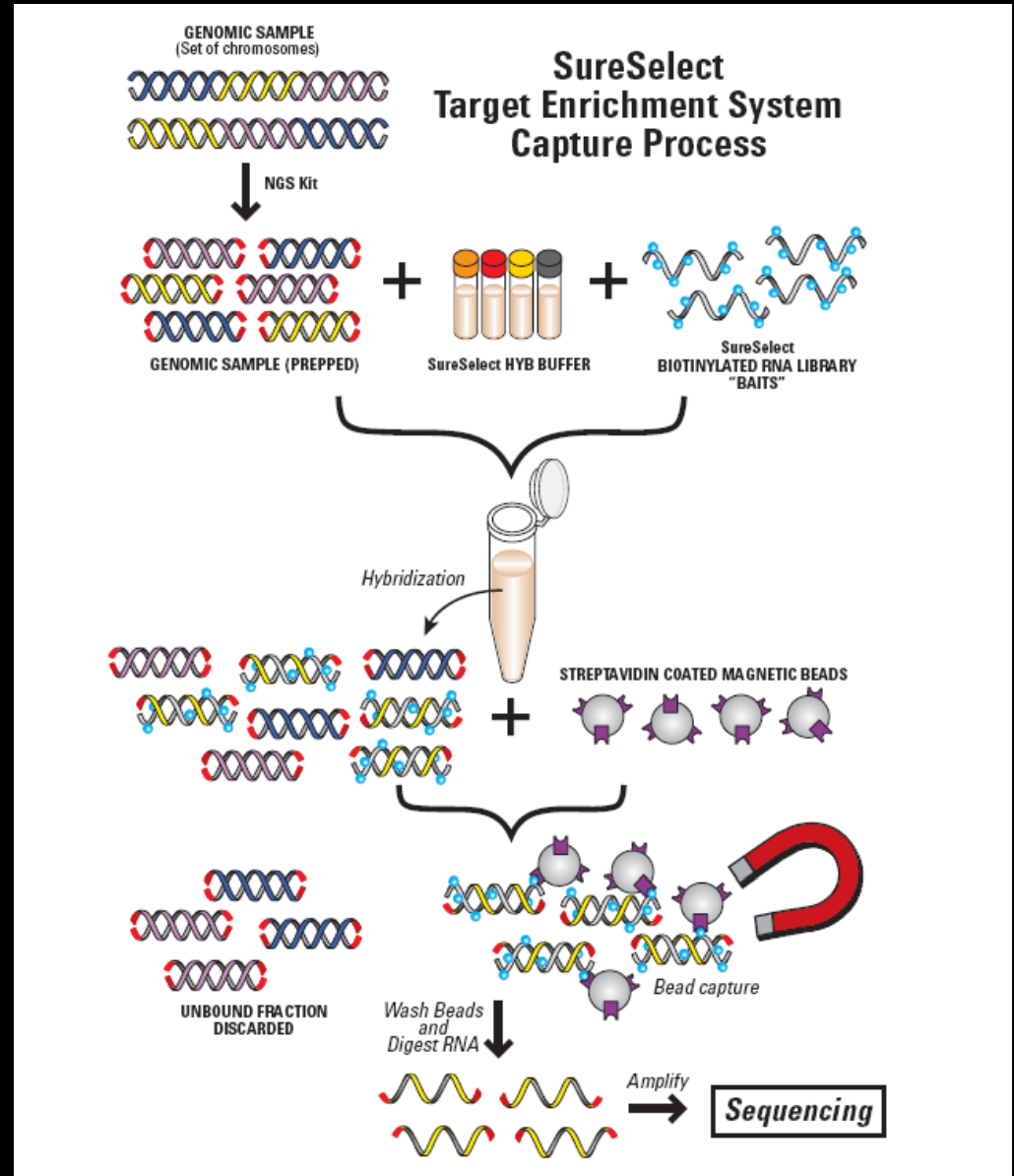
2. Hybridization approaches; in-solution or array based



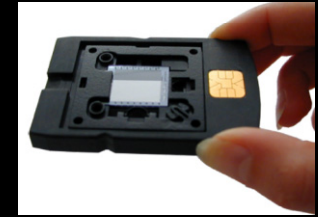
Agilent SureSelect

Some key features:

- eArray: free of charge custom array design tool
- RNA probes: 100 bases
- One array can capture up to 3.3Mb after masking
- Genomic DNA input: 3µg or less
- No capital equipment!
- Sample cost: ~\$1000
- Automation friendly and Scalable
- If done manually many hands-on steps

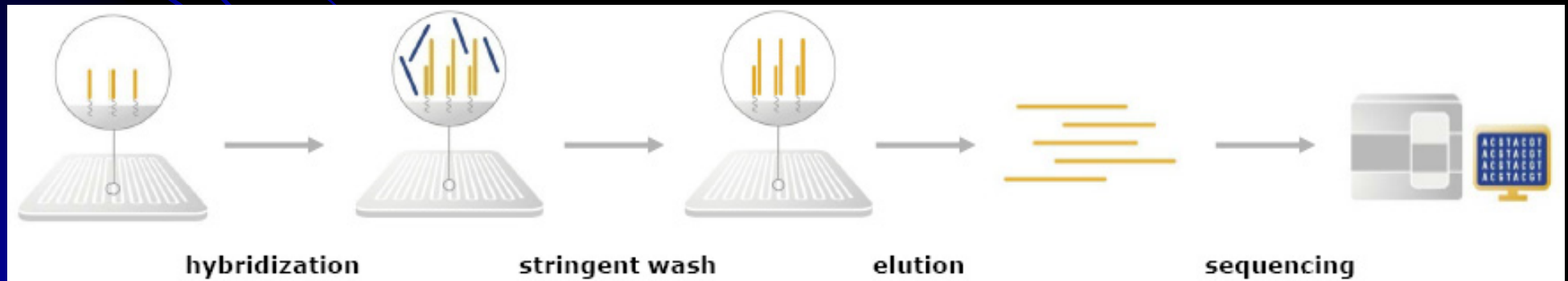
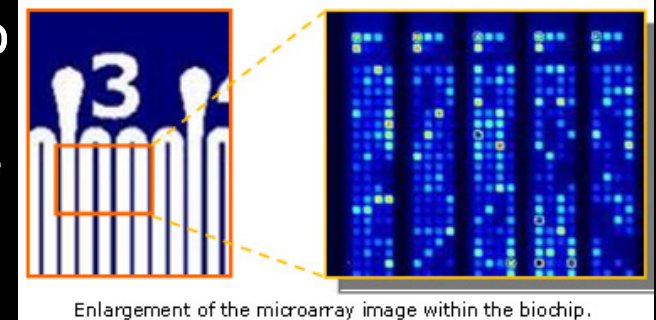


Febit HybSelect



Some key features:

- “Microfluidic Biochip” contains 8 separate micro channels for independent capture
- Requires capital equipment or use of a service provider
- 30 min hands-on time
- Genomic DNA : 1-5 μg
- Oligo length: 60-mer
- One array can capture up to 125Kb after masking
- Sample cost: over \$2000

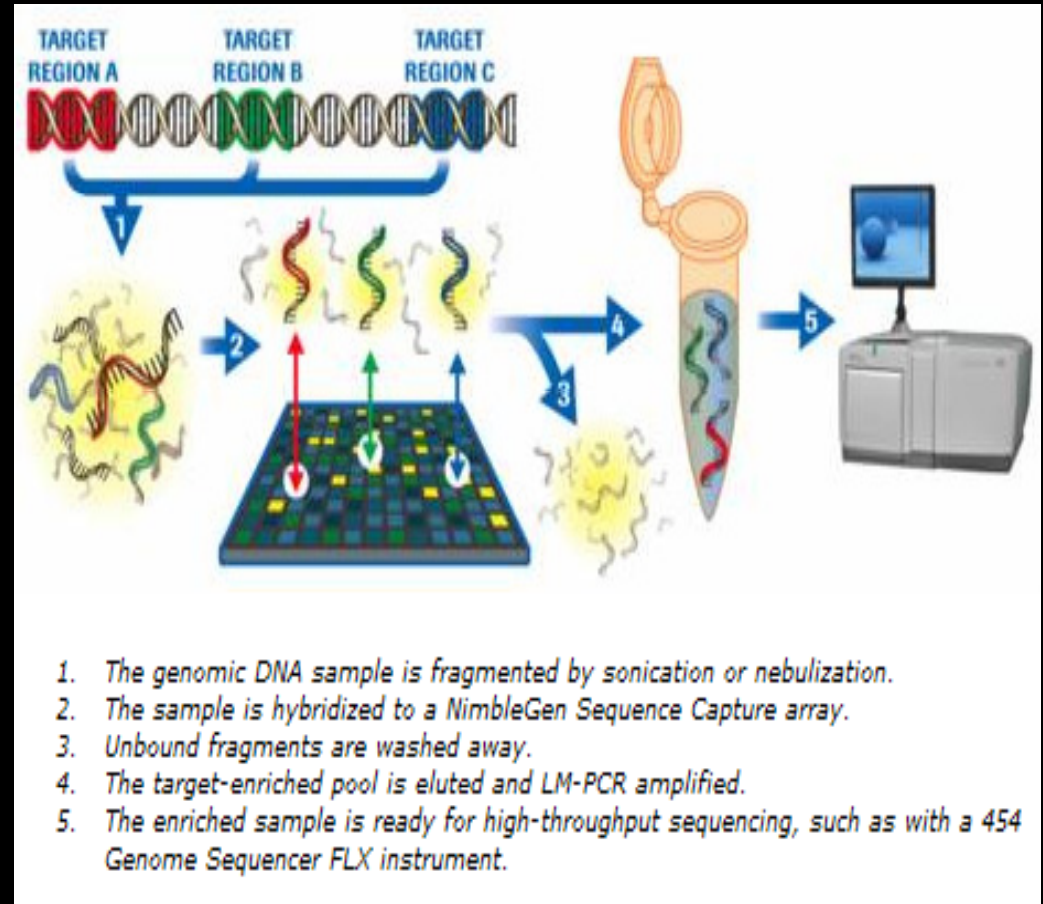


<http://www.febit.com/microarray-sequencing/services/hybselect/>

NimbleGen Array Capture

Some key features:

- Requires a hybridization and elution systems or use of a service provider
- Charged array design (waived if more than 5 arrays)
- Genomic DNA input: 3 μ g or less
- One array capture 5-30 Mb after masking
- Validated arrays for human studies
- Sample cost: ~\$1000
- 2010: In-solution capture is available



2009/10 DSRG study

- Coriell DNA: 'The Human Reference Genetic Material Repository DNA Sample' (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

DSRG:
Select DNA and regions

*** Illumina paired-end
sample prep kit provided
to all participants!

Agilent

Febit

NimbleGen

Illumina library prep
+ In-solution capture

Illumina library prep
+ Array-based capture

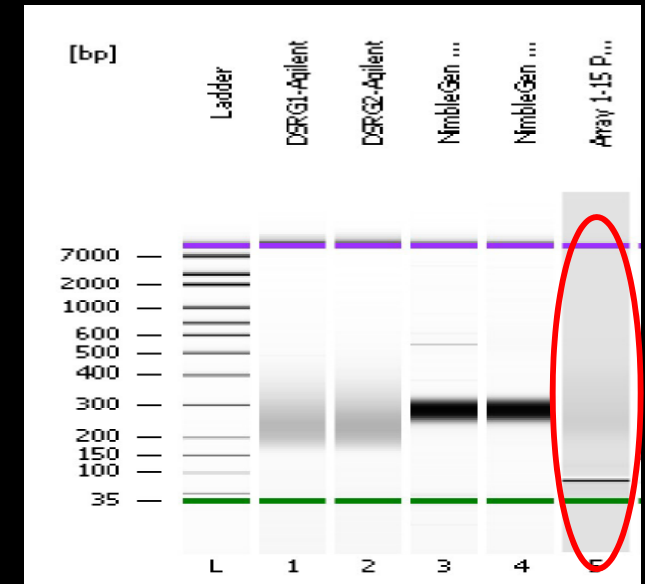
Illumina library prep
+ Array-based capture

DSRG:
Sequence samples on Illumina GAI at 2 different centers

Data Analysis at 2 different centers

QC and Illumina Run Statistics

- Samples were run on the Agilent High Sensitivity chip to assess the quantity and quality
- Samples were loaded in equal nM concentrations for each technology on two Illumina paired-end flowcells
- Two lanes were dedicated for each technology and the flowcells were run at different centers



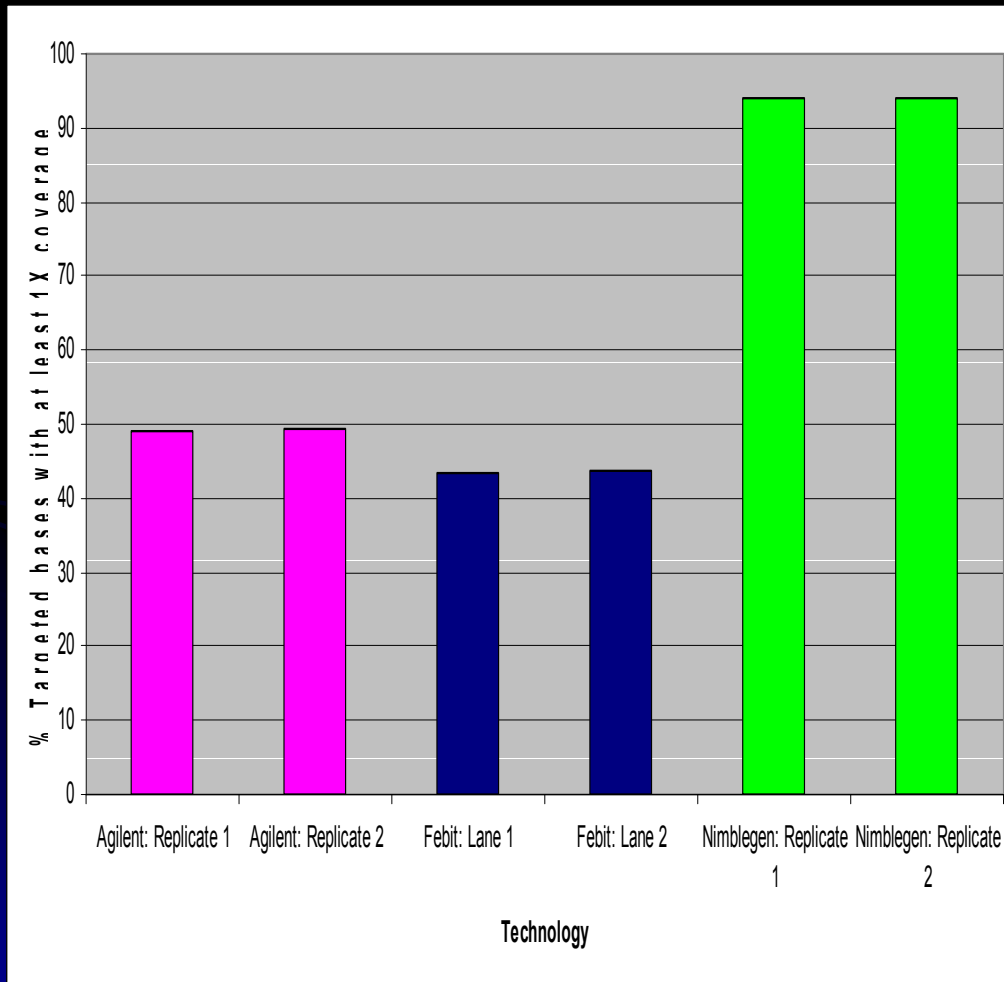
Illumina Primary Analysis

Lane Info		Tile Mean +/- SD for Lane							
Lane	Lane Yield (kbases)	Clusters (raw)	Clusters (PF)	1st Cycle Int (PF)	% intensity after 20 cycles (PF)	% PF Clusters	% Align (PF)	Alignment Score (PF)	% Error Rate (PF)
1	389729	103444 +/- 2114	97432 +/- 1972	183 +/- 10	88.11 +/- 2.05	94.19 +/- 0.26	89.54 +/- 0.11	113.40 +/- 0.30	0.35 +/- 0.07
2	331083	86949 +/- 2321	82771 +/- 2126	179 +/- 9	86.27 +/- 1.89	95.20 +/- 0.19	89.48 +/- 0.14	113.37 +/- 0.23	0.32 +/- 0.01
3	585040	158592 +/- 1581	146260 +/- 1424	163 +/- 6	80.70 +/- 1.57	92.23 +/- 0.68	91.58 +/- 0.10	117.00 +/- 0.25	0.30 +/- 0.03
4	491407	130867 +/- 2549	122852 +/- 2333	174 +/- 6	80.75 +/- 2.31	93.88 +/- 0.42	91.72 +/- 0.09	117.52 +/- 0.18	0.27 +/- 0.02
5	596168	173542 +/- 7350	149042 +/- 6492	157 +/- 6	81.07 +/- 1.75	85.88 +/- 0.39	54.64 +/- 0.35	50.29 +/- 0.33	0.53 +/- 0.03
6	598656	174141 +/- 8715	149664 +/- 7425	157 +/- 7	81.45 +/- 1.80	85.95 +/- 0.54	54.57 +/- 0.39	50.24 +/- 0.36	0.52 +/- 0.03
7	0	13203 +/- 1977	0 +/- 0	0 +/- 0	0.00 +/- 0.00	0.00 +/- 0.00	0	0	0
8	619596	173167 +/- 10933	154899 +/- 7788	148 +/- 18	81.77 +/- 6.22	89.53 +/- 1.52	98.86 +/- 0.08	192.95 +/- 1.21	0.35 +/- 0.09

Data Analysis

1. Combined the datasets from both sequencing centers for each replicate
2. Filtered each data set so that sequences have quality score > 10 for 100% of the bases
3. Mapped reads against the hg19/GRCh37 genome using bowtie 0.12.0
4. Normalized the data sets to equal sizes
5. A series of 'perl' scripts were then used to calculate coverage per position in every targeted region, creating a coverage map
6. Coverage maps were imported into the R statistical computing environment (2.1.0), to find the sensitivity, specificity, and reproducibility for each sample
7. Plots and figures were generated using the "ggplot2" library and MS Excel

Sensitivity: How much was captured?



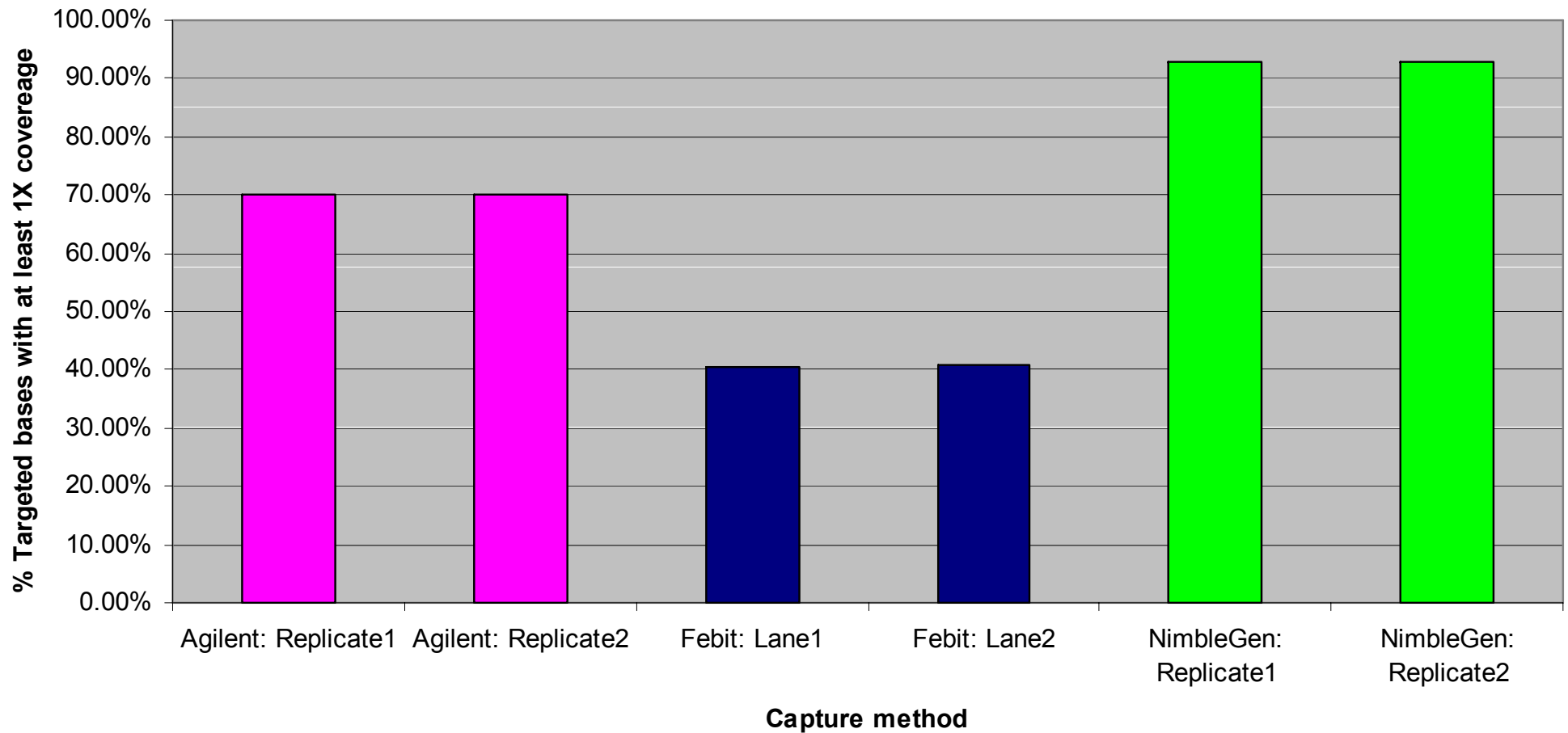
- Converted hg18 genome build coordinates to hg19 co-ordinates
- Determine the overlapping regions between the two different targeted regions
- Reanalyze all the data for only the overlapping regions.

3.5Mb

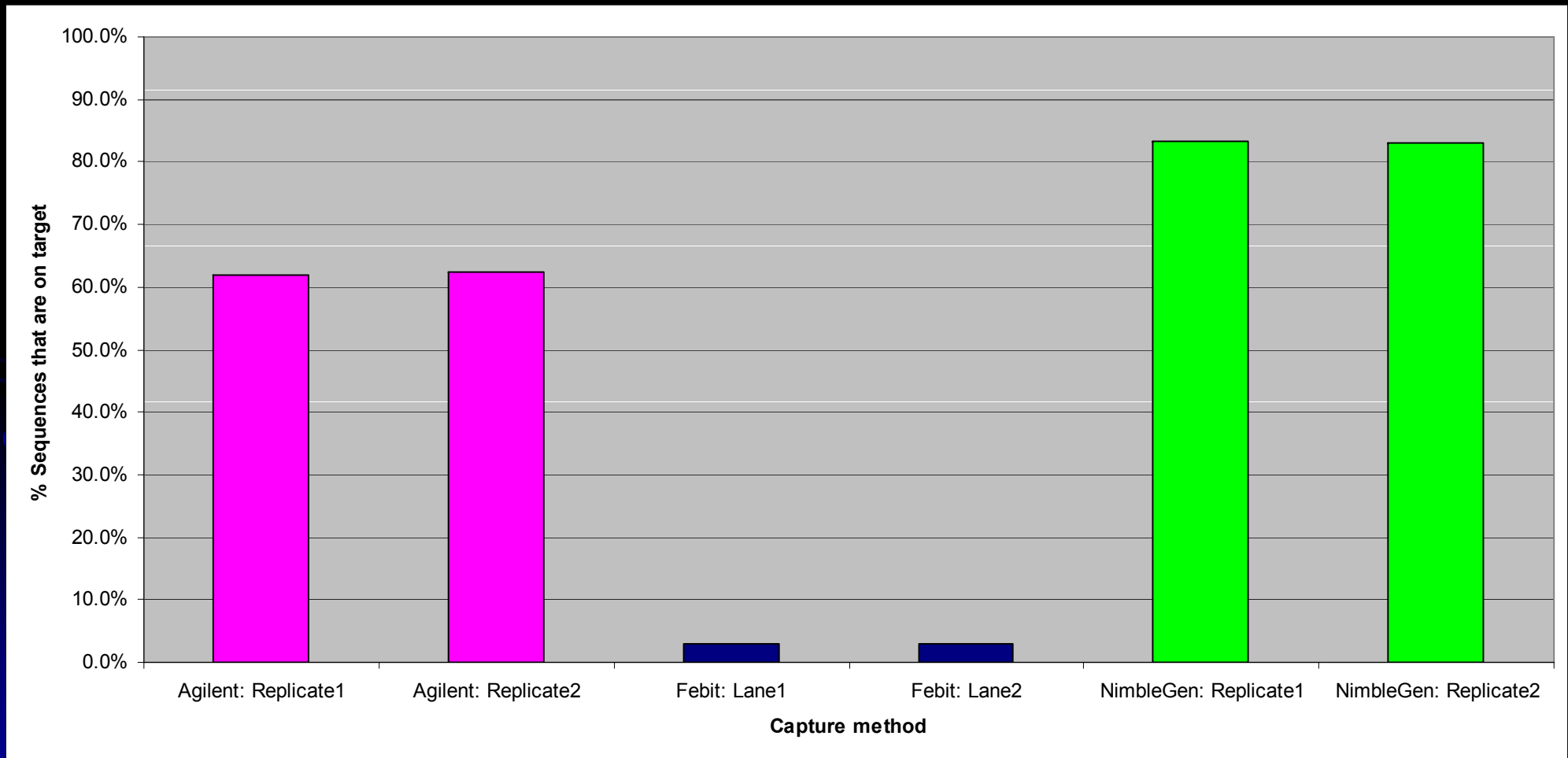


2.1Mb

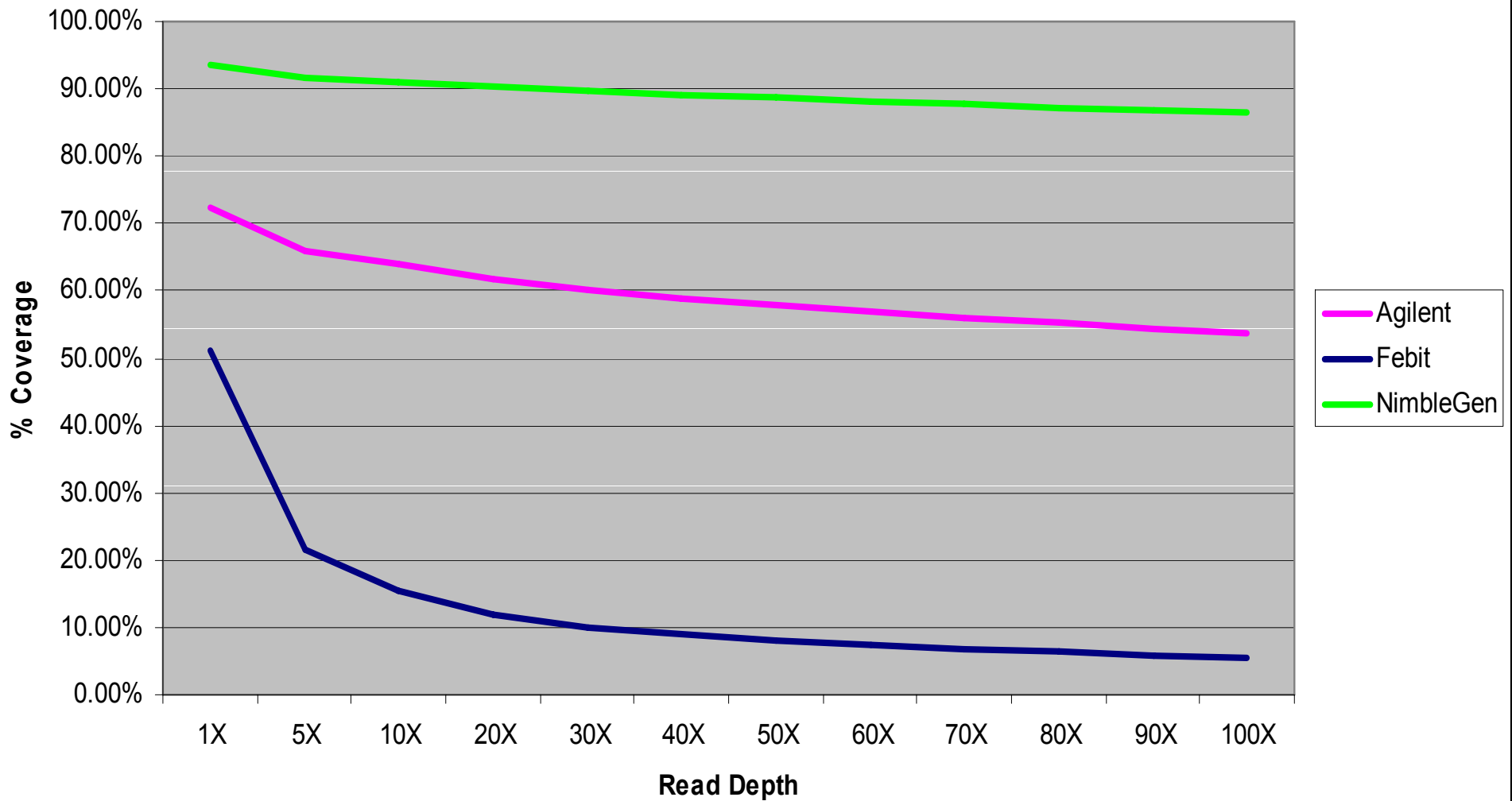
Sensitivity: How much of the 2.1Mb was captured by at least 1 read?



Specificity: How many of the reads mapped to the intended targets?

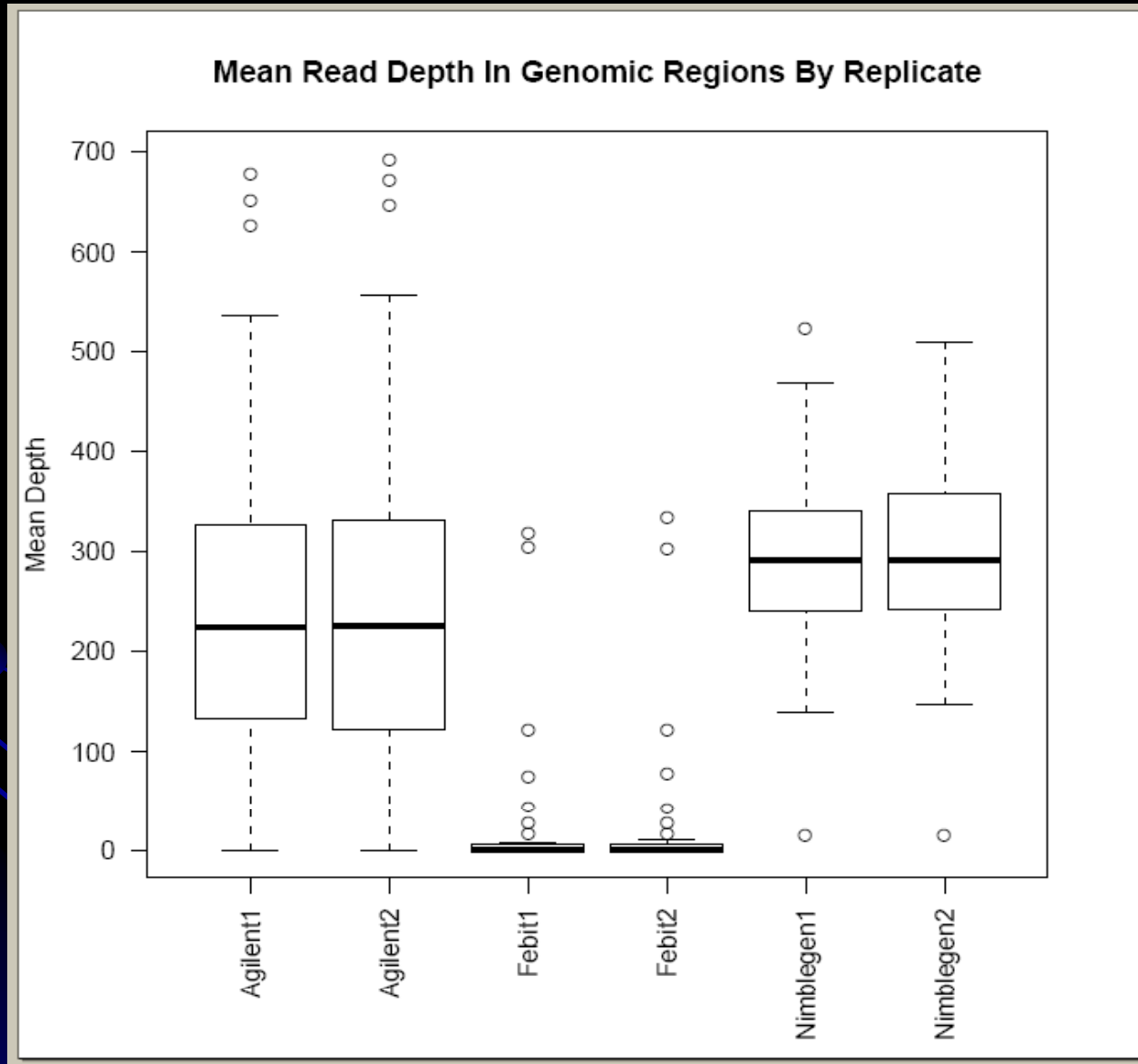


% Coverage vs. Read Depth



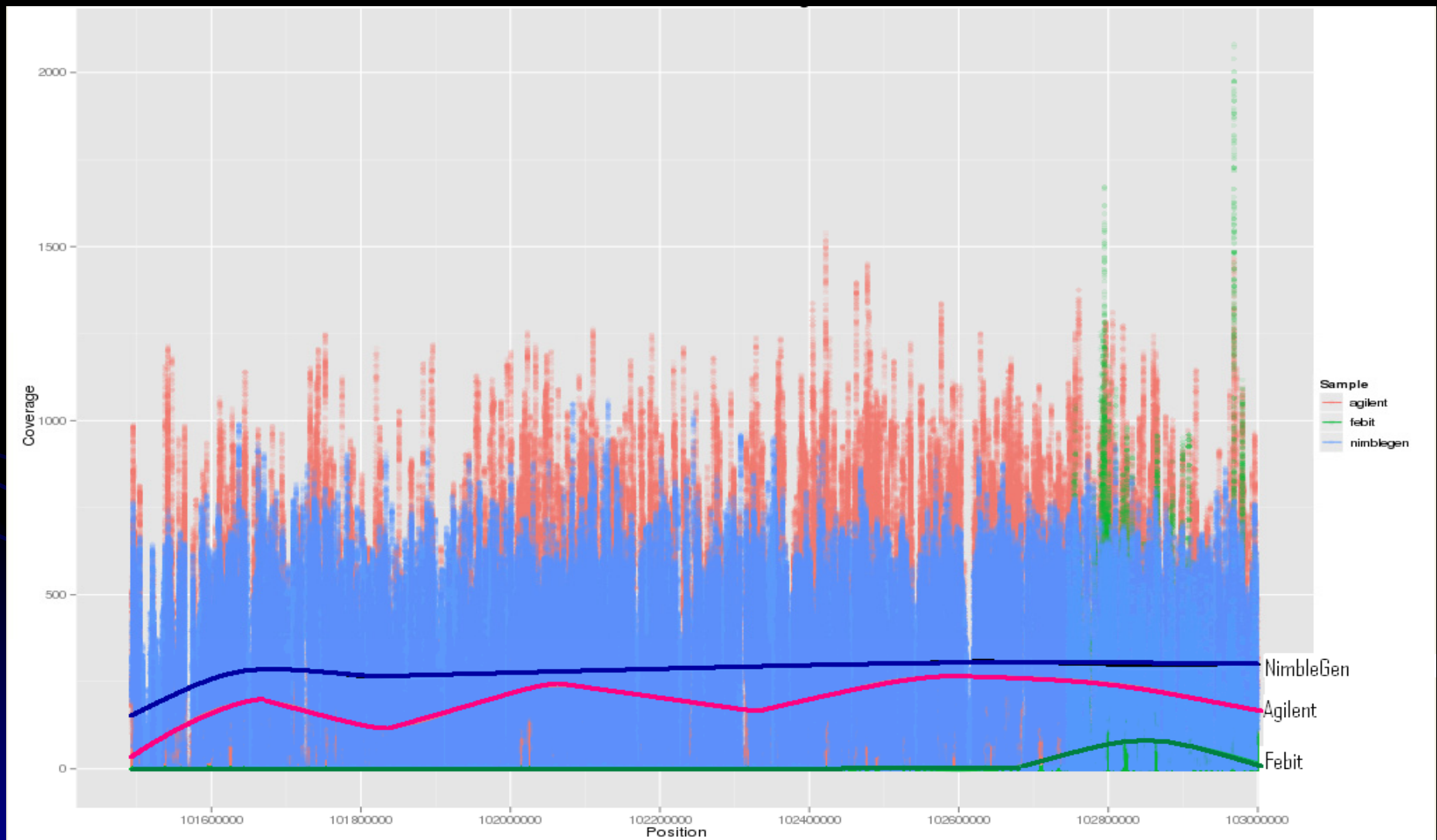
Kip Bodi, Aaron Noll

Coverage of Overlapping Genes



Aaron Noll

Coverage of the Continuous Region



In Summary...

Description	Agilent In-solution	Febit Array	NimbleGen Array
Cost per sample	✓		
Sample input requirement	✓	✓	✓
Quality of captured sample			✓
Reproducibility	✓		✓
Sensitivity			✓
Specificity			✓
Coverage			✓
Scalability	✓		

Future Directions...

- Repeat Agilent capture with probes designed to the 3.5Mb region of hg19
- Febit capture: What went wrong? Repeat?
- Determine reasons for low coverage areas that are common to all three technologies
- Perform SNP analysis
 - check how well each technology can detect SNPs in a given region
 - look for false positives and negatives

Many Thanks!!!

Agilent/Vector Biotech

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

Febit

Han Lee
Jack Leonard
Natalie Oliveira

NimbleGen

Daniel Burgess
Lance Brown
Michael Frawley
Xinmin Zhang

Illumina

Jonathan Pinter

Data Analysis:

Aaron Noll
Kip Bodi

Others:

Christine Brennan
Constance Esposito
Elizabeth Ketterer
Karen Staehling
Kendra Young

DSRG

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



The Association
of Biomolecular
Resource Facilities