# A complete workflow from sample preparation to analysis using SureSelect target enrichment system for Ion Proton semiconductor sequencing

Christian Le Cocq[1] ,Kyeong Soo Jeong[1], Arjun Vadapalli [1], Joseph Ong[2], Elin Agne[3], Filip Karlsson[3], Ashutosh Ashutosh[1], Francisco Useche[1], Jayati Ghosh[1], Henrik Johansson[3], Scott Happe[2], Douglas Roberts[1], and Holly Hogrefe[4]
Agilent Technologies, Diagnostics and Genomics Group [1]Santa Clara, California 95051 [2]Cedar Creek, Texas 78612 [3]Uppsala, Sweden [4]La Jolla, California 92037

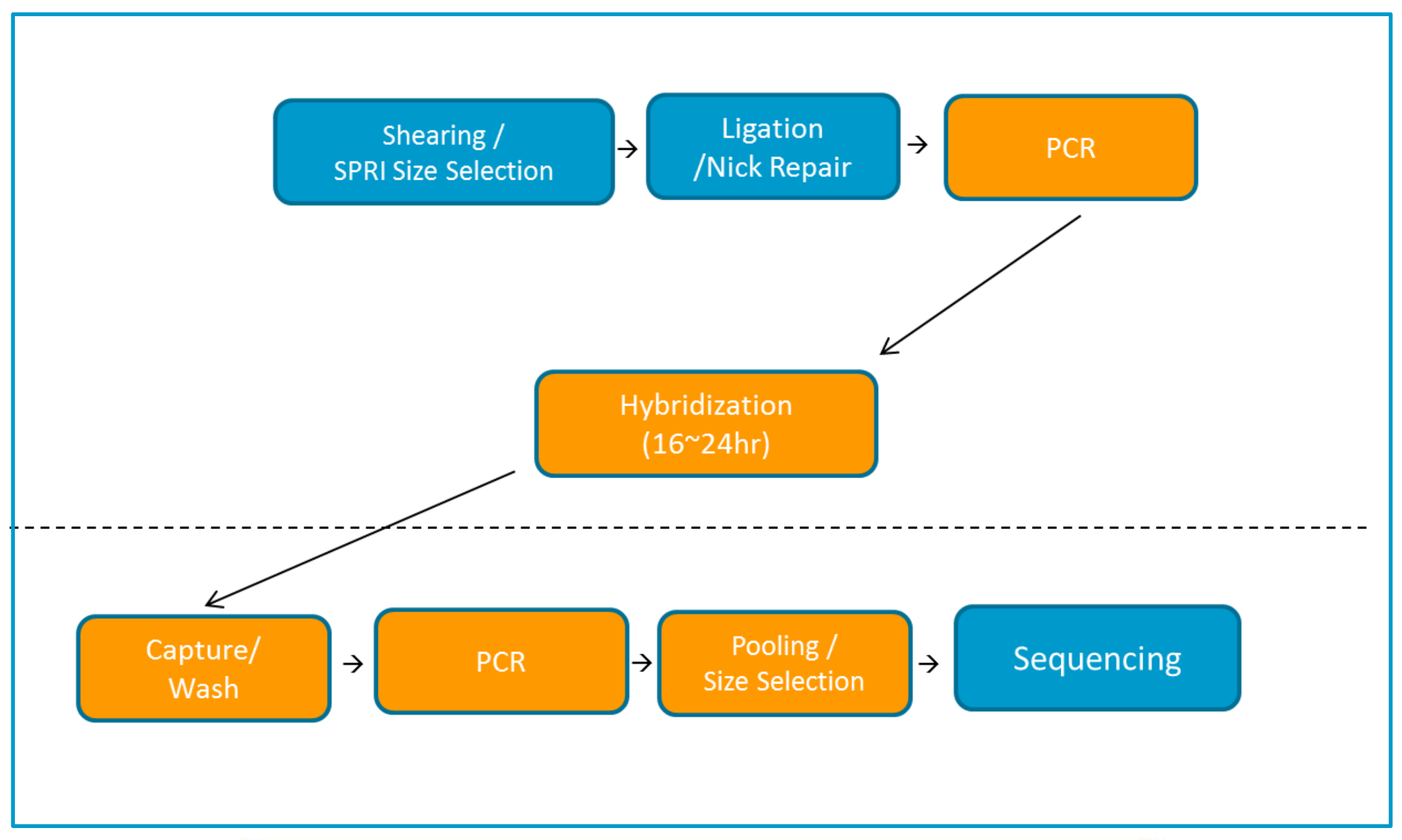
Agilent Technologies

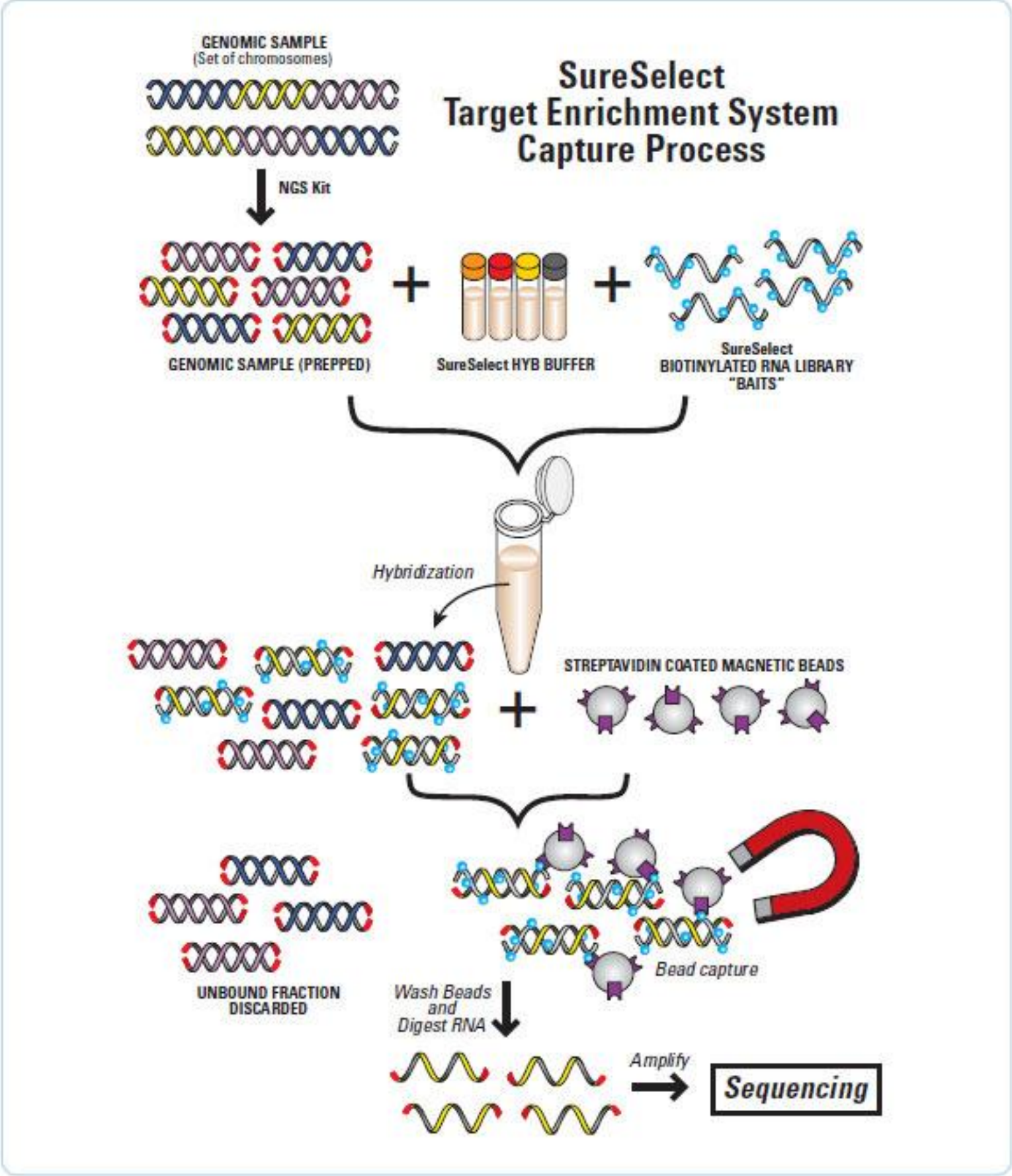
## Abstract

Whole exome or targeted sequencing for protein-coding regions has provided a cost effective way to identify common and rare polymorphisms that are associated with Mendelian disorders and complex diseases. With increased capacity of semiconductor sequencing, highly multiplexed samples can be studied in a single sequencing run. However, a complete workflow processing raw DNA samples to identify DNA variants in target regions is not easily accessible. Here we describe an analysis workflow to study multiplexed samples in semiconductor sequencing for several target sizes: 50Mb (Human All Exon), 3.2Mb (Human Kinome) and a 1Mb custom design. The workflow includes library preparation, SureSelect target enrichment, semiconductor sequencing, and variant calling with SureCall software (beta version). Improved and simplified steps for library preparation and target enrichment maximize multiplexing and produce consistent results in the Ion Proton sequencer. Sequencing output can be easily analyzed, visualized and summarized in a report with SureCall which is optimized for use with Agilent's target enrichment system. We demonstrate high capture efficiency, uniformity, and reproducibility of enrichment. The results from different capture sizes show comparable high performance regardless of various targeted regions. The combination of efficient target enrichment system, semiconductor sequencing, and SureCall software provides a fast and convenient tool to assess DNA variants in genomic regions of interest.

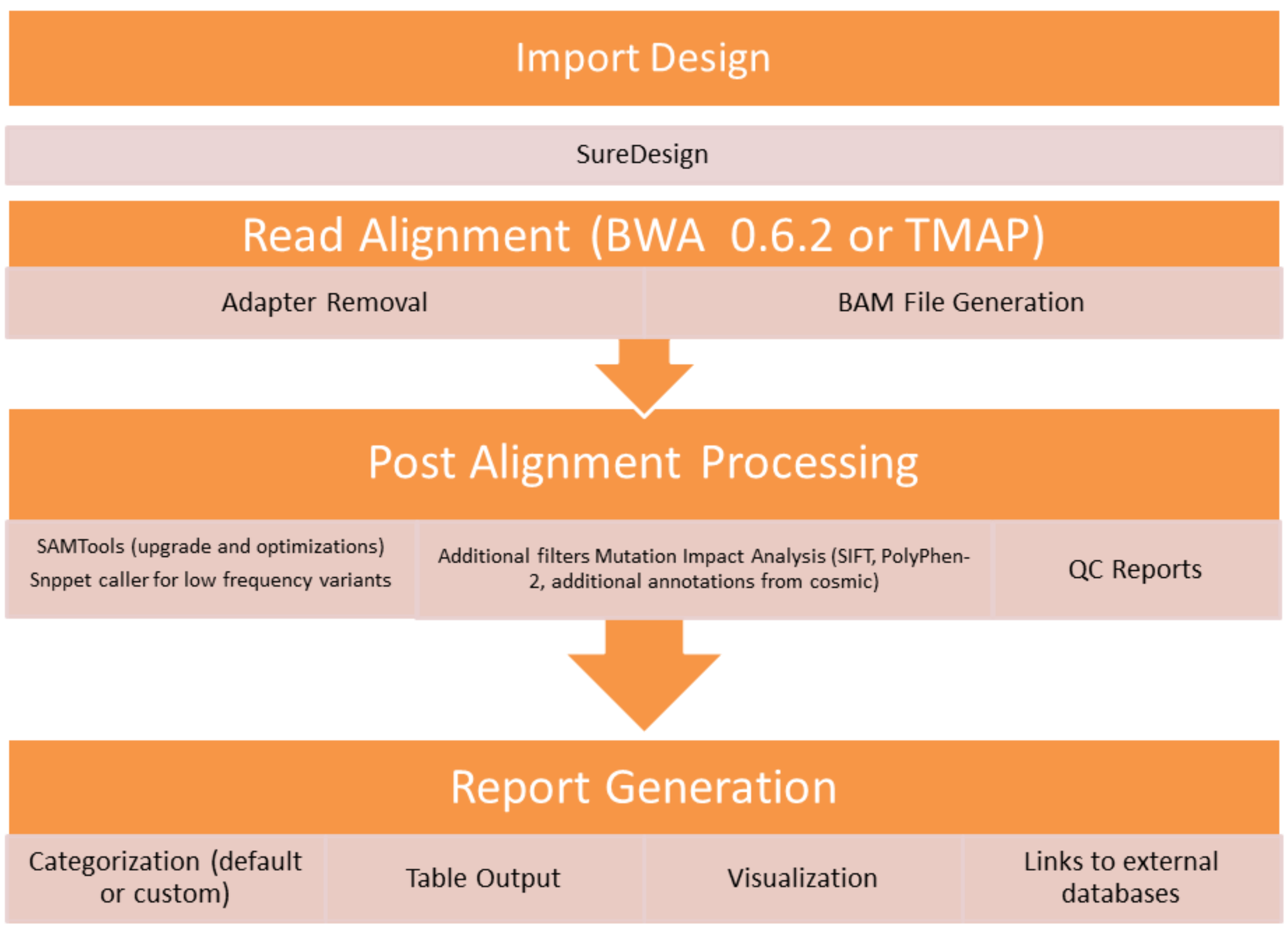
## Materials & Methods

### SureSelect Target Enrichment Workflow


Shearing / SPRI Size Selection → Ligation /Nick Repair → PCR → Hybridization (16~24hr) → Capture/ Wash → PCR → Pooling / Size Selection → Sequencing



## SureCall Workflow


Import Design

SureDesign

Read Alignment (BWA 0.6.2 or TMAP)

Adapter Removal | BAM File Generation

Post Alignment Processing

SAMTools (upgrade and optimizations)
Snppet caller for low frequency variants

Additional filters Mutation Impact Analysis (SIFT, PolyPhen-2, additional annotations from cosmic)

QC Reports

Report Generation

Categorization (default or custom) | Table Output | Visualization | Links to external databases

## All Exon V5 Capture Performance

**SureSelect with Ion Proton shows excellent capture and uniformity.**

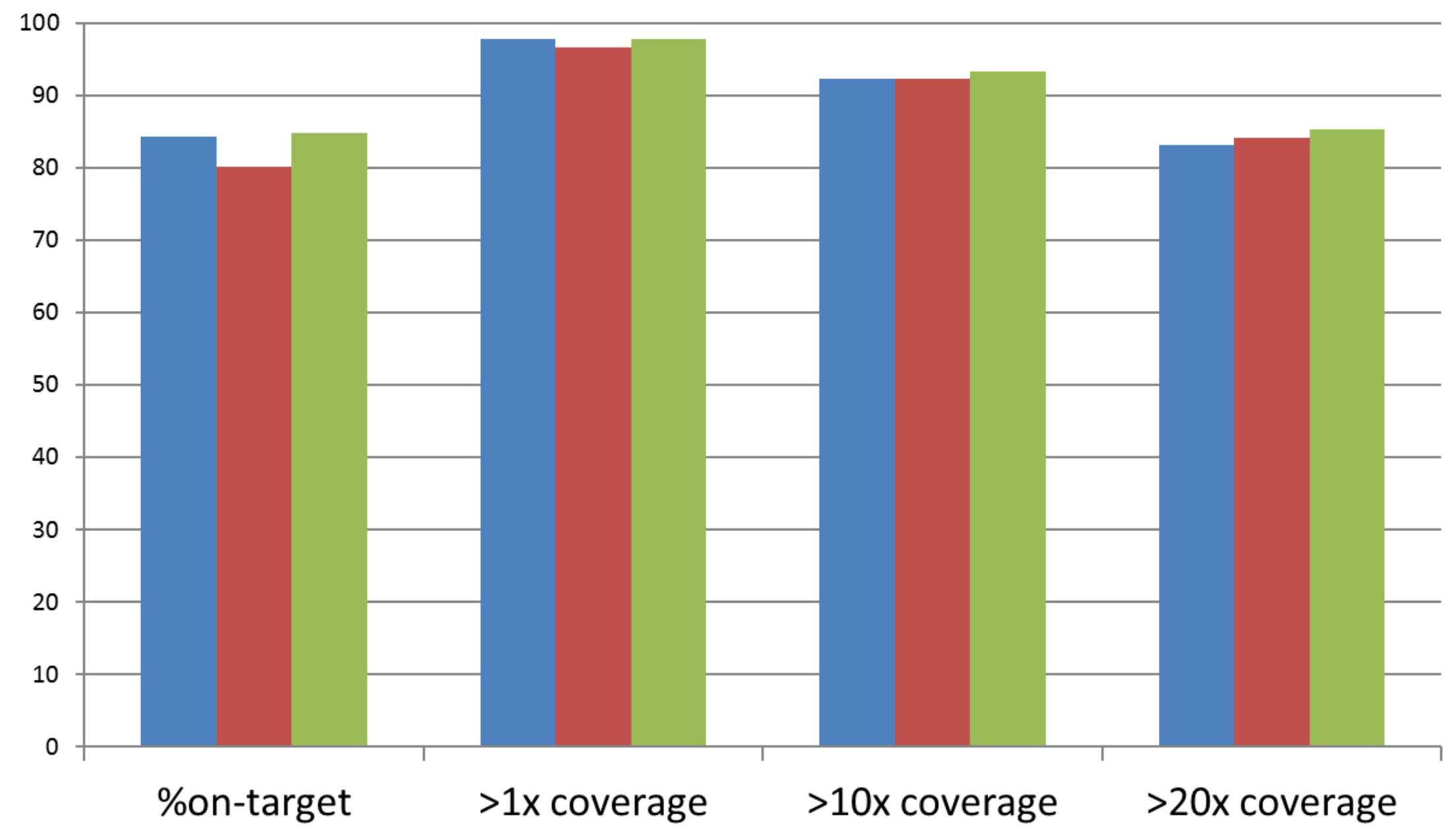


Figure 1. Summary of typical capture performance. SureSelect was performed with Hapmap DNA (NA12878) and Exon V5 baits. The vertical axis shows the percentage of on-target or coverage. Each sample includes 4Gb of sequencing. Coverage shows the percentage of targeted bases with triplicates.

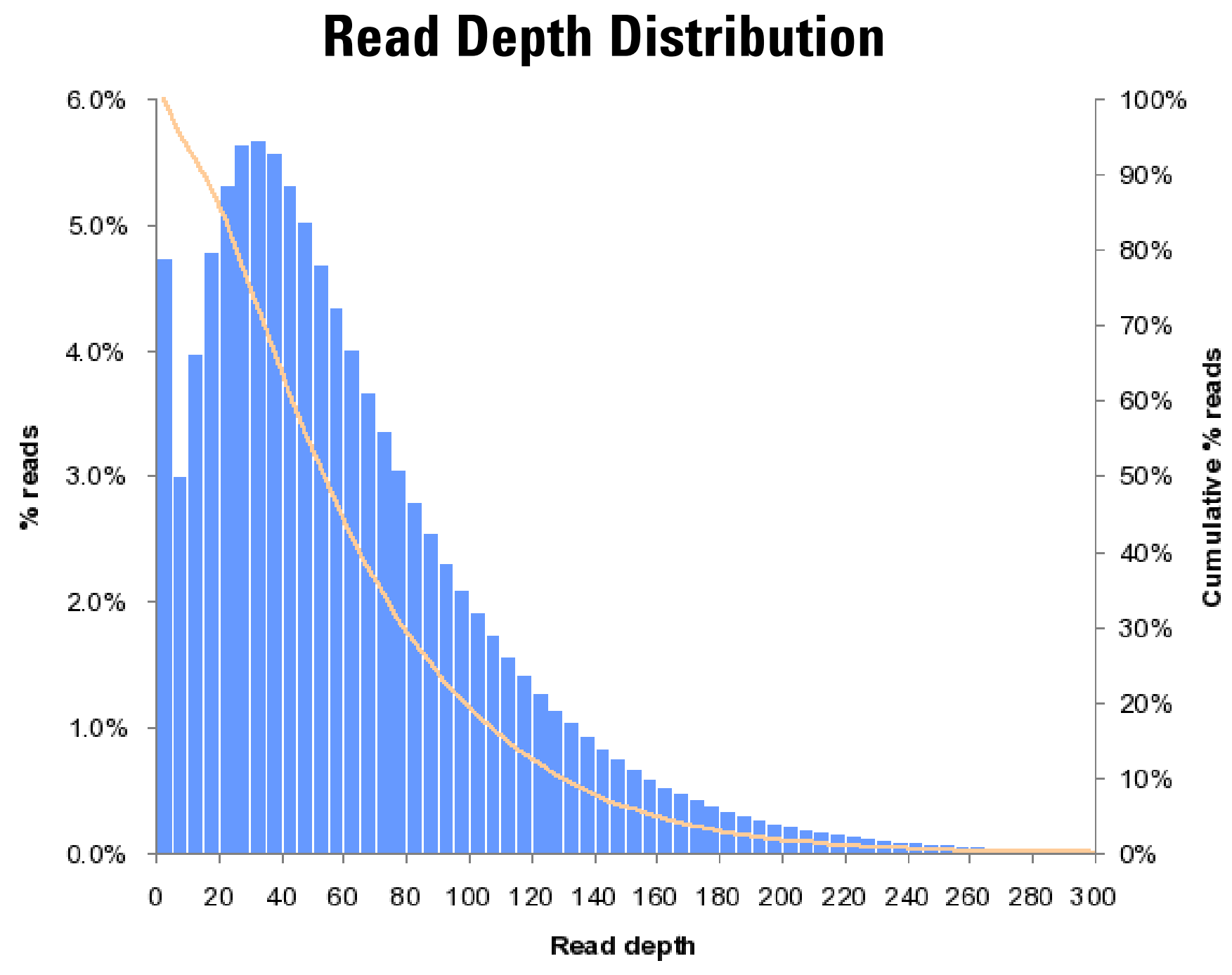


Figure 2. Read depth distribution. The uniquely mapped reads show a uniform read depth distribution. The cumulative percent of reads is shown by an orange line.

## SNP concordance (All Exon V5)

| | |
|---|---|
| Evaluation sites | 49113 |
| Overlapping sites with db135 | 39608 |
| Concordant sites | 39435 |
| Novel sites | 9433 |
| Concordant rates (%) | 99.4 |

Table 1. SNP concordance with dbSNP135. The exon data show high concordance with previously reported SNPs. The amount of sequencing used for comparison is 4 Gb. SNP calling was performed with Genome Analysis Toolkit.

## Kinome Capture Performance

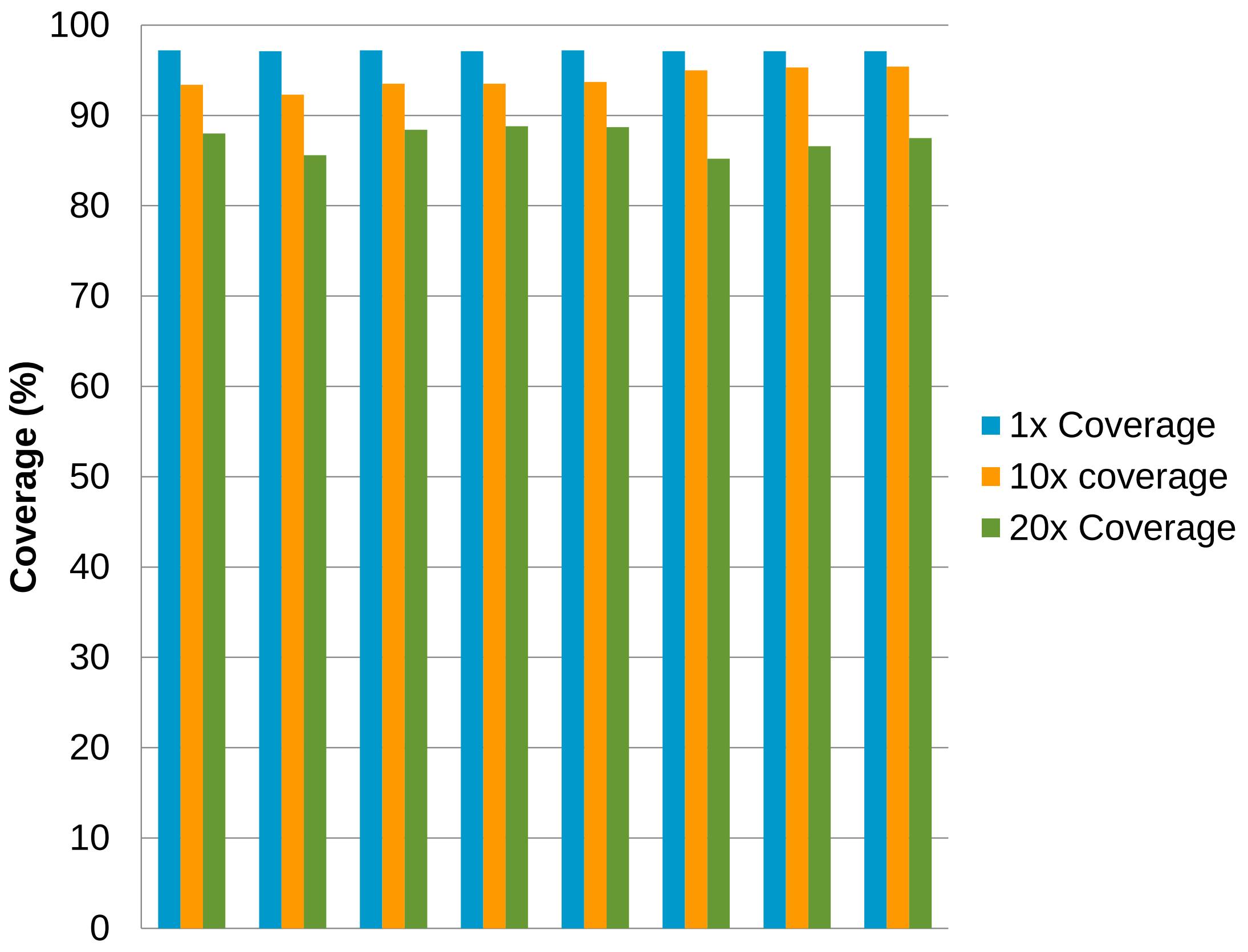


Figure 3. Capture performance with SureSelect Kinome baits of 3.2 Mb. 8 Hapmap DNA (NA12878) samples were analyzed in one Ion Proton run with multiplexing. Each sample was normalized with 0.32 Gb of sequencing. Coverage represents the percentage of targeted bases with at least 1 (blue), 10 (orange) and 20 (green) reads.

## Custom Capture Performance

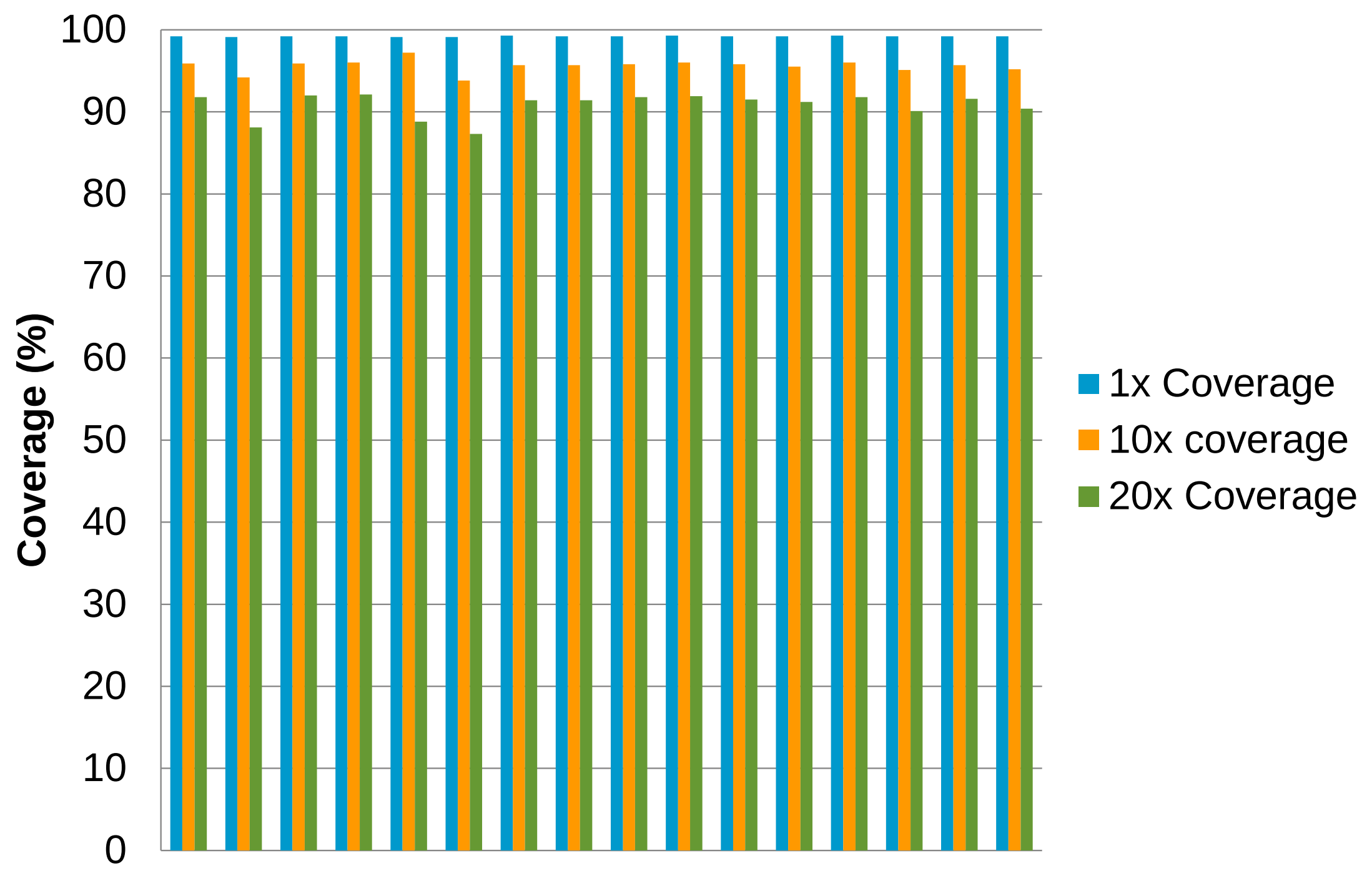
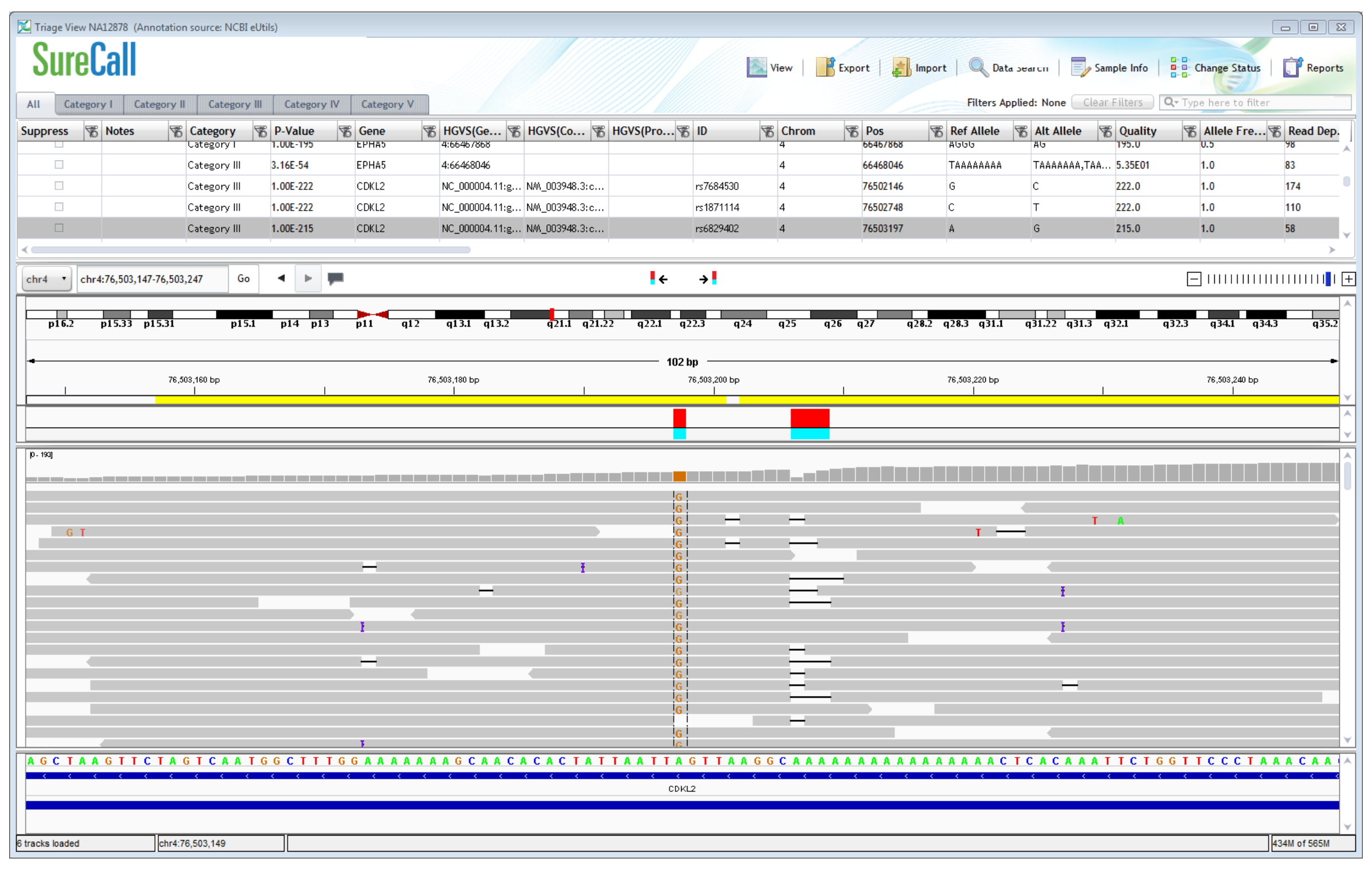


Figure 4. Capture performance with SureSelect Custom baits (1 Mb). 16 Hapmap DNA (NA12878) samples were analyzed with multiplexing. The results are shown with 0.1 Gb of sequencing. The vertical axis shows the percentage of targeted bases with at least 1 (blue), 10 (orange) and 20 (green) reads.

## SureCall results


Triage View NA12878 (Annotation source: NCBI eUtils)

SureCall

Filters Applied: None

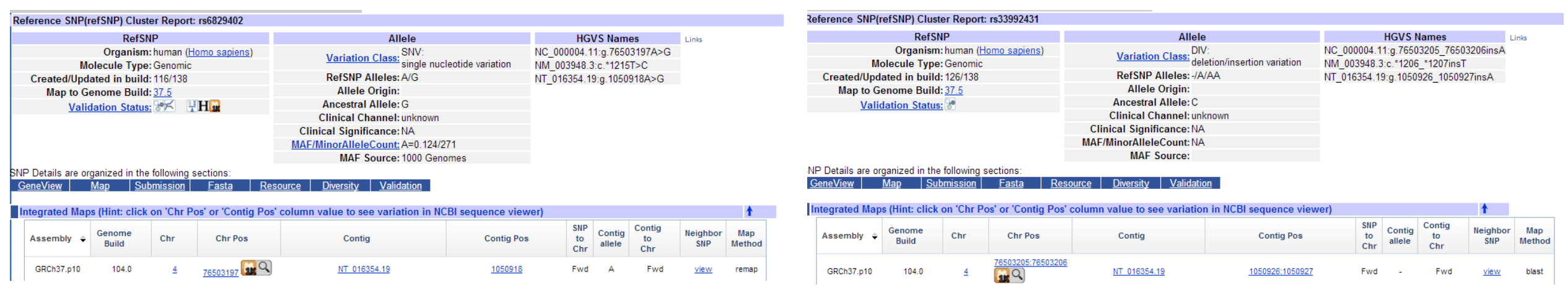
| Suppress | Notes | Category | P-Value | Gene | HGVS(Ge... | HGVS(Co... | HGVS(Pro... | ID | Chrom | Pos | Ref Allele | Alt Allele | Quality | Allele Fre... | Read Dep. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| □ | | Category I | 1.00E-195 | EPHA5 | 4:66467868 | | | | 4 | 66467868 | AGGG | AG | 195.0 | 0.5 | 98 |
| □ | | Category III | 3.16E-54 | EPHA5 | 4:66468046 | | | | 4 | 66468046 | TAAAAAAAA | TAAAAAAA,TAA... | 5.35E01 | 1.0 | 83 |
| □ | | Category III | 1.00E-222 | CDKL2 | NC_000004.11:g... | NM_003948.3:c... | | rs7684530 | 4 | 76502146 | G | C | 222.0 | 1.0 | 174 |
| □ | | Category III | 1.00E-222 | CDKL2 | NC_000004.11:g... | NM_003948.3:c... | | rs1871114 | 4 | 76502748 | C | T | 222.0 | 1.0 | 110 |
| □ | | Category III | 1.00E-215 | CDKL2 | NC_000004.11:g... | NM_003948.3:c... | | rs6829402 | 4 | 76503197 | A | G | 215.0 | 1.0 | 58 |

chr4 | chr4:76,503,147-76,503,247 | Go

AGCTAAGTTCTAGTCAATGGCTTTGGAAAAAAAGCAACACACTATTAATTAGTTAAGGCAAAAAAAAAAAAAAAAAACTCACAAATTCTGGTTCCCTAAACAA

6 tracks loaded | chr4:76,503,149 | 434M of 565M

Figure 5. Triage view in SureCall v1.1 highlighting dbSNP record for SNP rs6829402 on the left and indel rs33992431 on the right for NA12878. The Triage view allows users to view variants and the reads supporting the call in a single view along with annotations from public databases.


Reference SNP(refSNP) Cluster Report: rs6829402

| RefSNP | |
|---|---|
| Organism: | human (Homo sapiens) |
| Molecule Type: | Genomic |
| Created/Updated in build: | 116/138 |
| Map to Genome Build: | 37.5 |
| Validation Status: | |

| Allele | |
|---|---|
| Variation Class: | SNV: single nucleotide variation |
| RefSNP Alleles: | A/G |
| Allele Origin: | |
| Ancestral Allele: | G |
| Clinical Channel: | unknown |
| Clinical Significance: | NA |
| MAF/MinorAlleleCount: | A=0.124/271 |
| MAF Source: | 1000 Genomes |

HGVS Names
NC_000004.11:g.76503197A>G
NM_003948.3:c.*1215T>C
NT_016354.19:g.1050918A>G

SNP Details are organized in the following sections:
GeneView | Map | Submission | Fasta | Resource | Diversity | Validation

Integrated Maps (Hint: click on 'Chr Pos' or 'Contig Pos' column value to see variation in NCBI sequence viewer)

| Assembly | Genome Build | Chr | Chr Pos | Contig | Contig Pos | SNP to Chr | Contig allele | Contig to Chr | Neighbor SNP | Map Method |
|---|---|---|---|---|---|---|---|---|---|---|
| GRCh37.p10 | 104.0 | 4 | 76503197 | NT_016354.19 | 1050918 | Fwd | A | Fwd | view | remap |

Reference SNP(refSNP) Cluster Report: rs33992431

| RefSNP | |
|---|---|
| Organism: | human (Homo sapiens) |
| Molecule Type: | Genomic |
| Created/Updated in build: | 126/138 |
| Map to Genome Build: | 37.5 |
| Validation Status: | |

| Allele | |
|---|---|
| Variation Class: | DIV: deletion/insertion variation |
| RefSNP Alleles: | -/A/AA |
| Allele Origin: | |
| Ancestral Allele: | C |
| Clinical Channel: | unknown |
| Clinical Significance: | NA |
| MAF/MinorAlleleCount: | NA |
| MAF Source: | |

HGVS Names
NC_000004.11:g.76503205_76503206insA
NM_003948.3:c.*1206_*1207insT
NT_016354.19:g.1050926_1050927insA

SNP Details are organized in the following sections:
GeneView | Map | Submission | Fasta | Resource | Diversity | Validation

Integrated Maps (Hint: click on 'Chr Pos' or 'Contig Pos' column value to see variation in NCBI sequence viewer)

| Assembly | Genome Build | Chr | Chr Pos | Contig | Contig Pos | SNP to Chr | Contig allele | Contig to Chr | Neighbor SNP | Map Method |
|---|---|---|---|---|---|---|---|---|---|---|
| GRCh37.p10 | 104.0 | 4 | 76503205:76503206 | NT_016354.19 | 1050926:1050927 | Fwd | - | Fwd | view | blast |

Figure 6. Corresponding dbSNP records shown for SNP (left) and indel (right).

| Coefficient of Variation | Percentage of supporting reads | Total Number of Variants | Percentage of reported Indels |
|---|---|---|---|
| no filter | no filter | 10751 | 87 |
| 3 | 75 | 4326 | 66 |
| 1 | 75 | 3718 | 61 |
| 0.5 | 75 | 1512 | 4 |

Table 2. Indel Filter performance in SureCall v1.1 on Ion Proton SureSelect Kinome data. The higher rate of indel calls on the Ion platform is primarily due to inaccurate flow calls over homopolymer regions. The number of indels reported in the Triage view is controlled by the parameters coefficient of variation and percentage of supporting reads. The coefficient of variation parameter is defined as the standard deviation over the mean indel size. The percentage of supporting reads is defined as the percentage of reads supporting the indel call.

## Conclusions

- **Agilent's SureSelect Target Enrichment for the Ion Proton Platform provides a comprehensive, efficient, robust, and cost-effective means to sequence subsets of the human genome.**
- **Different capture sizes show comparable high performance regardless of various targeted regions.**
- **High reproducibility of enrichment, depth distribution, and sequence coverage from multiplexed sequencing.**
- **Excellent concordance with known dbSNP.**