Portfolio: Germline
Product: Exome v 6.0
Published: June 2019

BROAD INSTITUTE

GENOMIC SERVICES

Implementation and Performance Assessment of Broad Genomic's Germline Exome v 6.0

BACKGROUND

The Broad Institute's Genomics Platform evaluated several new commercially available germline exome sample preparation kits. Our goal was to identify an off-the-shelf solution that met the following criteria:

1. Meets or exceeds the inter- and intra-target coverage of our v5.0 exome
2. Compatible/scalable with existing sample preparation (library construction) workflows
3. Meets or exceeds the end-to-end process time of our v5.0 exome
4. Provides a cost competitive product
5. Includes novel, recently curated content

Considerations 1-4 were readily identified in the Twist Biosciences - Human Core Exome. However, many new and novel targets of interest as identified by members of the Broad Institute's somatic and germline research community, were not adequately covered.

We collaborated to augment the Human Core Exome, adding hundreds of targets to the design, resulting in the development and implementation of a new standard exome for the Broad Institute's Genomic Services, designated as Germline Whole Exome Sequencing (WES) v6.0.

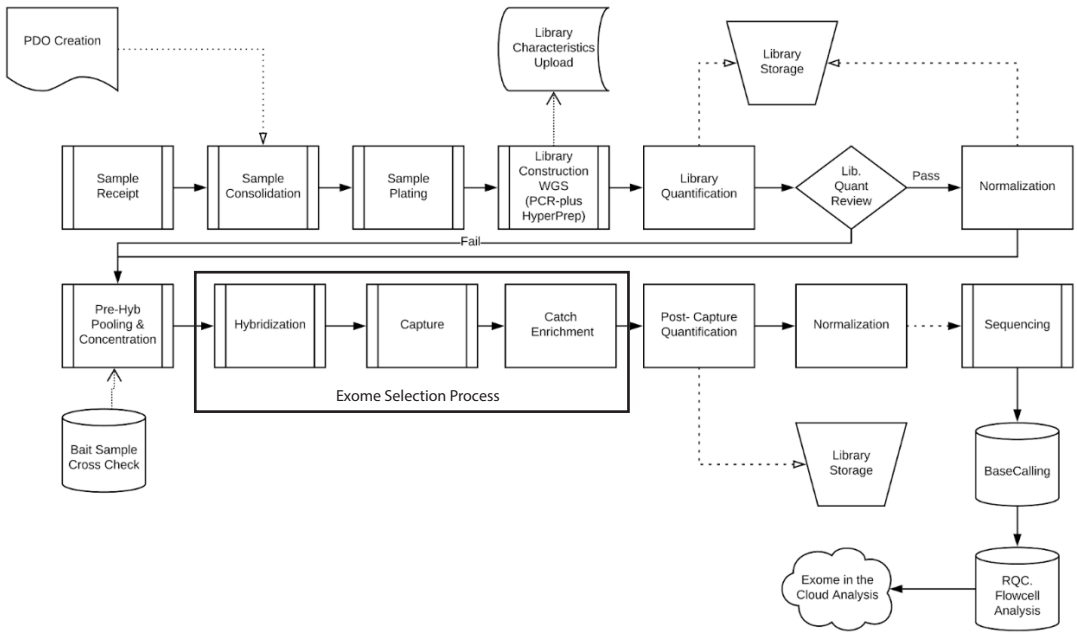
PRE-IMPLEMENTATION EVALUATION DESIGN

Ninety-six germline samples were processed according to the germline exome v5.0 workflow. This workflow is characterized by a tagmentation process which combines enzymatic shearing and adapter addition into one step. The resulting library then undergoes two hybridization/capture events.

The same ninety-six samples were then processed according to v6.0 workflow. This protocol utilizes mechanical shearing and ligation of adapters - process steps that are shared with our Whole Genome Sequencing (WGS), allowing us to unify both workflows. This process also features an updated probe set, and undergoes only one hybridization/capture event, reducing overall process time.

After exome capture, libraries were sequenced on the NovaSeq using the S4 flowcell with XP setup. Data analysis was performed in the cloud pipeline using the bait and target interval files based on the hg38 reference build.

See figure below for the v6.0 workflow schema.



PRE-ANALYTICAL CONSIDERATIONS

Benchmarking of the performance of the germline exome v6.0 protocol includes a comparison of product specification metrics generated between the v5.0 and v6.0 protocol. An equivalent amount of sequencing data was generated for each condition, measured as bases sequenced. It should be noted that differences in these metrics may not only be the result of the performance of the assays but also a function of the panel designs and the algorithm used in that design processes. As such this review's focuses on assay performance and not necessarily on panel design.

COMPARING THE TARGETS: GENCODE V29

The v6.0 targets cover slightly fewer bases overall than the previous targets, at 35,086,188 bases, compared to 38,692,858 bases for v5.0. However, a greater proportion of the Gencode v29 coding region is now covered.

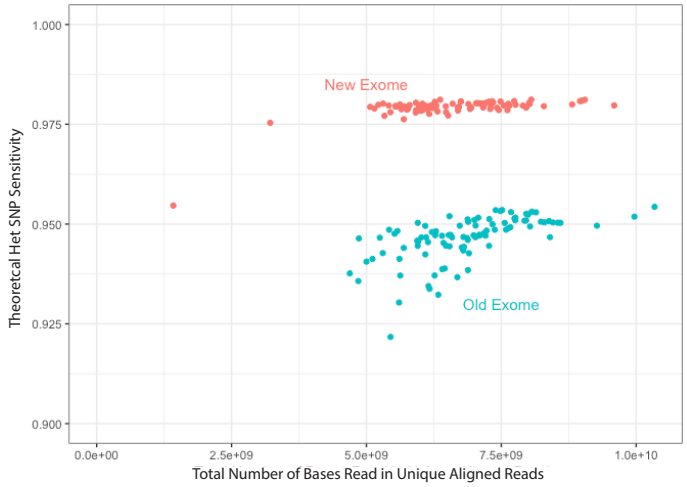
Overall, 322 genes have at least half of their coding regions added to the target region when moving to the new exome (based on GencodeV29). The new targets overlap 34,149,260 of the 34,491,558 GencodeV29 bases in coding regions validated at level 2 or better - 99.01% of bases in those regions.

	Total Target Bases	Exclusive bases	Exclusive to Gencode v29	Exclusive @ level 2 or better
WES v6.0	35,086,188	755,214	376,167	368,924
WES v5.0	38,692,858	4,361,884	274,051	136,326

THEORETICAL HET SENSITIVITY

The theoretical HET Sensitivity metric is the predicted sensitivity for heterozygous SNPs in the target region, and is calculated based on the distribution of base coverages over the targets.

The new exome performs significantly better than the old exome. This is due to more consistent and even coverage by the new exome selection.



SENSITIVITY AND SPECIFICITY

In addition to theoretical performance, we also evaluate performance by comparing actual calls to a known truth set. Our in-house control samples (NA12878), processed with the new exome selection and previous exome selection were run through base HaplotypeCaller from GATK4.0.11.0. Each selection was called over its own target interval. No variant filtering was performed for this evaluation. The calls were evaluated against the GIAB3.3.2 truth set using the Benchmarking pipeline from the GA4GH consortium. The new exome provides both improved SNP - sensitivity, specificity, false and positive rates, and INDEL - sensitivity, specificity, and false positive rates.

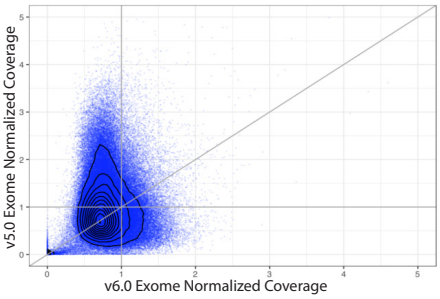
	SNP Sensitivity	SNP Specificity	SNP False Positive Rate (per Mb)	INDEL Sensitivity	INDEL Specificity	INDEL False Positive Rate (per Mb)
WES v6.0	98.5%	98.6%	8.8	85.6%	77.8%	3.37
WES v5.0	96.3%	97.7%	15.6	78.4%	66.6%	8.93

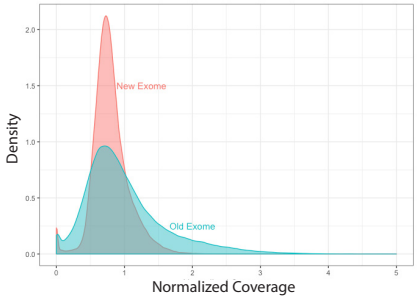
COVERAGE

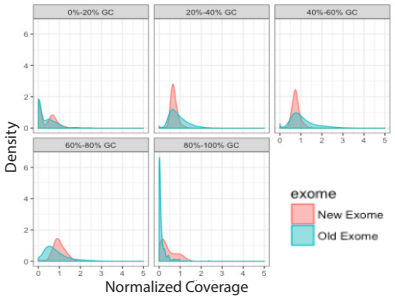
INTER TARGET COVERAGE

We calculated the mean normalized coverage of targets for each version of the exome workflow. The new v6.0 exome shows much more consistent normalized coverage between targets. Conversely there are numerous targets that are poorly covered by the old exome which are well covered by the new exome. Defining poor coverage as less than 0.2, and good coverage as 0.8, then 2,060, targets covering 325,008 bases are well covered by the new exome and poorly covered by the old exome. This corresponds to 0.95% of the intersection of the two exomes. Meanwhile, only 18 targets covering 1918 bases are poorly covered by the new exome and well covered by the old exome. This corresponds to 0.01% of the intersection of the two exomes.

When mean normalized coverage is stratified by GC content we observe particular improvements in regions with GC content below 20% or above 60%.

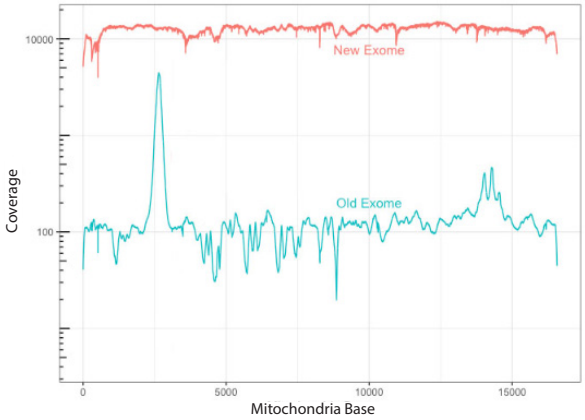






MITOCHONDRIA

Additionally the v6.0 exome has much better coverage of the mitochondria, with average coverage increasing from ~100x to 10,000x.



Y CHROMOSOME

The new exome selection leads to more consistent coverage of the Y chromosome than the previous method, which had more variability.

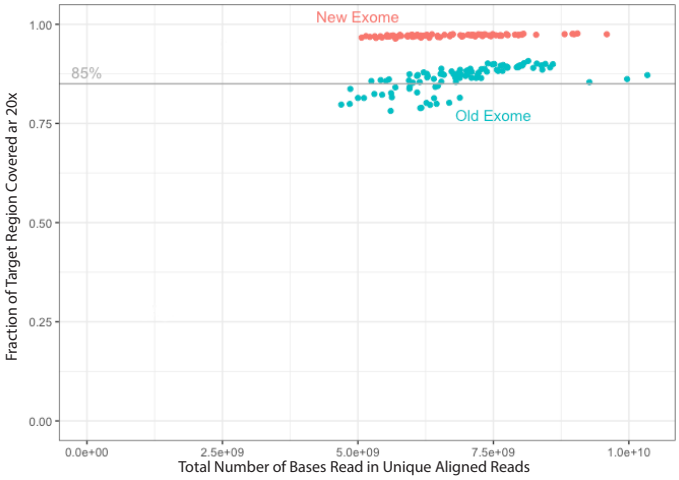
ACMG59

We also considered coverage over the ACMG59 genes. In general, coverage is more consistent, and drops below 20x are reduced in the new exome. There are eight regions which have an average coverage below 20x, only two of which are coding regions included in the target list, but not in the bait design.

- Coding, included in target, NOT included in baits
  - ◊ 161 base region in PKP2: This region is the most concerning, as it was well covered by the old exome. It is a repetitive region (soft-masked in the reference).
  - ◊ 112 base region in KCNH2: This region is a bit better covered in the old exome than in the new exome.
- Coding, included in targets and baits, low coverage (all of these regions have high GC content)
  - ◊ 188 base region in KCNQ1
  - ◊ 177 base region in RYR1
  - ◊ 126 base region in TGFBR1
- Coding, NOT in targest or baits
  - ◊ 34 base region in KCNQ1: This is a very small 5 base coding region and its surrounding padding.
- Non-Coding, NOT in targets or baits
  - ◊ 58 base region in GLA
  - ◊ 702 base region in SDHD

ABSOLUTE COVERAGE

The deliverable for both the germline exome v5.0 and v6.0 products is defined as a minimum of 85% of target bases achieving 20X or greater coverage (roughly equivalent to 60x MTC). Due to the plexing strategy employed in the lab, samples typically receive between 8-10 Gigabases ±2. The v5 exome exhibits coverage levels with 95% of samples reaching between 85 and 100% of targets at 20x coverage. The v6 exome significantly outperforms with 95% of samples receiving between 95 and 100% of targets at 20x coverage.



SUMMARY

The new v6.0 exome was compared to our previous v5.0 exome. The new exome performs significantly better across a wide range of metrics. We observe greater coverage of the coding regions of Gencode v29 targets, the Y chromosome, mitochondrial targets, and the ACMG59. More broadly we saw improved coverage between intervals, as well as between bases within particular intervals. This leads to improved sensitivity and specificity for known SNPs and INDELs, which was also reflected in the significantly better heterozygous SNP sensitivity.