**Sequencing QC Report (Somatic)**

**This is the somatic exome sequencing build BUILD\_NAME qc summary report. Total number of samples sequenced for this build is SAMPLE\_NUM. The capture kit used is BED\_FILE.**

**Table 1. Subject count breakdown before QC**

**1. MULTIQC**

FastQC(consolidated in MultiQC format 1) is a widely used high through put sequencing data quality control tool. It provides a modular set of analyses which includes: Basic Statistics, Per Base Sequence Quality, Per Sequence Quality Scores, PerBase Sequence Content, Per Sequence GC Content, Per Base N Content, Sequence Length Distribution, Duplicate Sequences,

Overrepresented Sequences, Adapter Content, Kmer Content, Per Tile Sequence Quality. For all fastqc results, please refer to

**FASTQC\_REPORT**

**2. SEQUENCING DUPLICATION RATE CHECK**

Duplicate reads are sequencing reads that start at the exact same position and contain the same CIGAR string (Compact Idiosyncratic Gapped Alignment Report). Duplicates can arise from three sources: true biological duplication, PCR duplication, and optical or clustering duplication. In biological duplication, two individual sequenced molecules are coincidentally identical; this is unlikely with standard-depth sequencing but may occur in deep sequencing. PCR duplication is the result of PCR amplification of a library molecule, and multiple copies of that amplified molecule being sequenced. Optical duplication refers to large clusters of molecules being erroneously called as two separate clusters (on non-patterned flowcells), or duplicates generated when a template molecule on a flowcell seeds more than 1 cluster on a flowcell (on patterned flowcells). We use Picard MarkDuplicates to remove duplicate reads from both lane level BAM files and subject (merged lane) level BAM files. The lane level duplication rate consists of optical duplicates within the same sequencing lane and PCR duplicates within one sample prep library run on one lane. The subject level duplication rate consists of PCR duplicates coming from the same sample prep library run across two or more lane level BAM files. Below is a beeswarm plot showing total duplication rate distribution across number of lanes a subject is consisted of.

**Figure 1. subject level duplication rate by number of lanes**

**Table 2. 10 samples with highest duplication rate (LANE\_NUM: the number of lanes the data comes from; TOTAL\_DUP: aggregated duplication rate of the subject; SECONDARY\_DUP: residual duplication rate after lane dup removed)**

**3. CONTAMINATION CHECK**

We use verifyBamID 2 to cross-individual contamination. Note that verifyBamID performs well in germline samples. However, tumor samples with substantial allelic imbalance can be a challenge for this tool. Our recommended contamination QC cutoff is 5%. Samples with contamination rate greater than 5% may be included in the analysis due to 1). New/replacement sample materials are no longer available; 2). Investigators request to include them. It is recommended to pay extra attention to those samples with higher than 5% contamination rate during your downstream analysis. No samples are excluded from variant calling for the build. Figure 2 shows the sample contamination rate. Samples with highest contamination rates (n = 10) are listed in Table 3.

**Figure 2. sorted contamination rate by subjects**

**Table 3. 10 samples with highest contamination rate**

**4. COVERAGE CHECK**

Average coverage across all case and control subjects is shown in Figure 3. It was calculated based on capturekit region using samtools mpileup function and sorted from lowest to highest. The targeted coverage for germline exome sequencing is on average 40X and 80% of capture region above 15X.The coverage distribution of each sample was also plotted using mosdepth software3. Please refer to build directory for mosdepth plots for each group:

**MOSDEPTH\_DIR**

**Figure 3. coverage scatter plot for all exome samples, with cases and controls separated**

**5. EXOMECQA**

Coverage uniformity evaluation was also conducted using ExomeCQA4. The two scores are calculated based on both GENE\_NUM genes and EXON\_NUM target regions of capturekit bed file across the all samples.

The CCS (Cohort Coverage Sparseness) score measures the percentage of base pairs with low coverage in a specific gene/target exon capture region across all samples. It is defined as the percentage of low coverage (<10X) bases within a given capture region in multiple WES samples. It’s basically the median number of samples with low coverage base percentage at a particular capture region. Thus, it may vary between 0 and 1, with high CCS scores indicating low sequence coverage. In general, a CCS score below 0.2 is considered as good coverage.

The UE (Unevenness) score measures non-uniformity of the coverage in a specific gene/target capture region across all samples. It is calculated based on the number and structural features (height, width, base) of the coverage peaks. The UE score increases with an increase in the number and relative height of peaks within a given capture region, where a score of 1 indicates uniformly distributed coverage, and the higher the score, the less uniform the coverage is in the capture region.

CCS and UE scores provide a cohort level view of the coverage depth and uniformity at given genomic regions. They help users to identify coverage inconsistency across large numbers of samples. This is particularly helpful while assessing the mutations detected through joint analysis.

Table 4 and Figure 4 are stats based on gene level. For more details of target exon capture region level stats, please refer to the following files: **EXOMCQA\_GENE\_REPORT**

**EXOMCQA\_EXON\_REPORT**

**Table 4a. Min, 1st quantile, Median, Mean, 3rd quantile and Max score for US score**

**Table 4b. Min, 1st quantile, Median, Mean, 3rd quantile and Max score for CCS score**

**Figure 4. the histogram of UE score and CCS Score for all sample**

**Table 5. 10 regions with highest Unevenness Score for all sample**

**6. PRE-CALLING QC CHECK**

From the subject level pre-calling qc report, we collect fold 80 base penalty from picard

CollectHsMetrics, average oxidation quality from picard CollectOxoGMetrics and lowest preadapter/baitbias total score from picard CollectSequencingArtifactMetrics in this summary. For more details and metrics, refer to the PRECALLING\_QC\_REPORT:

**PRECALLING\_QC\_REPORT**

Fold 80 penalty score is defined as the fold over-coverage necessary to raise 80% of bases in "non-zero-cvg" targets to the mean coverage level in those targets. Lower Fold 80 score is indicative of more even coverage across targets.

CollectOxoGMetrics calculates the Phred-scaled probability that an alternate base call results from an oxidation artifact. This probability score 5 is based on base context, sequencing read orientation, and the characteristic low allelic frequency. Higher scores indicate lower probability of artifactual calls. In Figure 6, average oxidation Qscore and lowest Qscore from all contexts are boxploted to see the distribution among all subjects.

CollectSequencingArtifactMetrics examines two sources of sequencing errors associated with hybridization selection protocols. These errors are divided into two broad categories, pre-adapter and bait-bias. Pre-adapter errors can arise from laboratory manipulations of a sample damaged before library prep/adapter ligation (examples include sample degradation, or fixative treatments such as formalin in FFPE samples). Bait-bias artifacts occur during or after the target selection step, and correlate with substitution rates that are 'biased', or higher for sites having one base on the reference/positive strand relative to sites having the complementary base on that strand. Most previous problematic samples have oxoG score and preadapter qscore at the 20-30 range. In Figure 7, pre-adapter artifact Qscore and bait-bias artifact Qscore are boxploted to see the distribution among all subjects.

**Figure 5. boxplot for all Fold 80 base penalty for samples**

**Figure 6. boxplot of average oxidation q score and lowest oxidation q score**

**Figure 7. boxplot of Lowest Preadapter and Baitbias Total Score for all samples**

**7. SEX CHECK**

Sample Sex was calculated based on the ratio of mapped reads to ChrY and ChrX. Sex concordance was then checked between ChrY/ChrX mapping ratio from sequence data and IDENTIFILER detected sex. In Figure 8 and table 6, SEX\_OUTLIER samples showed ChrY/ChrX values outlie the opposite se

**Figure 8. ChrY/ChrX reads ratio for all subjects**

**8. ANCESTRY CHECK**

LASER tool6 is used here to estimate individual ancestry by shotgun sequence reads without calling genotypes. LASER uses principal components analysis (PCA) and Procrustes analysis to analyze sequence reads of each sample and place the sample into a reference PCA space constructed using genotypes of a set of reference individuals. Here HGDP(Human Genome Diversity Project) data, including 632,958 autosomal SNPs loci for 938 unrelated individuals is used as reference panel.

**Figure 9. Laser ancestry estimation for all subjects: top 2 plots showing reference only, bottom 2 plots showing study subjects, black projected to reference samples, grey.**

For visualization of each disease group or controls, please refer to build directory:

**LASER\_DIR**

**9. POST-CALLING QC CHECK**

After Ensemble variant calling pipeline, we perform a standard post variant calling qc check on ensemble variants. Items checked include: all base change and indel counts based on capture region and ensemble PASS flag filtered pair call results. PASS flag means..

**Figure 10c. base change counts per disease group**

**10. BAM-MATCHER CHECK**

BAM-matcher 7 is one of the standard method to detect mislabeled pair samples for tumor normal pair somatic calls in high throughput sequencing data. Matching samples(same individual) will have at common fraction close to 1. Siblings and parent-child generated have common fraction between 0.5-0.7. Unrelated samples will have common fraction below 0.5.

**Figure 11. scatter plot of testing site count against concordant rate**

References

1 Ewels, P., Magnusson, M., Lundin, S. & Kaller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047-3048, doi:10.1093/bioinformatics/btw354 (2016).

2 Jun, G. et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet 91, 839-848, doi:10.1016/j.ajhg.2012.09.004 (2012).

3 Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics 34, 867-868, doi:10.1093/bioinformatics/btx699 (2018).

4 Wang, Q., Shashikant, C. S., Jensen, M., Altman, N. S. & Girirajan, S. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity. Sci Rep 7, 885, doi:10.1038/s41598-017-01005-x (2017).

5 Costello, M. et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res 41, e67, doi:10.1093/nar/gks1443 (2013).

6 Wang, C., Zhan, X., Liang, L., Abecasis, G. R. & Lin, X. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. Am J Hum Genet 96, 926-937, doi:10.1016/j.ajhg.2015.04.018 (2015).

7 Paul P.S. Wang, Wendy T. Parker, Susan Branford, Andreas W. Schreiber. BAM-matcher: a tool for rapid NGS sample matching. Bioinformatics 32, 2699–2701, https://doi.org/10.1093/bioinformatics/btw239 (2016).