**Sequencing QC report scheme**

**This is the whole exome/targeted 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) 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 (G. Jun, M. Flickinger, et,al. Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data, American journal of human genetics doi:10.1016/j.ajhg.2012.09.004) to cross-individual contamination. Note that verifyBamID performs well in mostly diploid samples (e.g. germline data), but there are caveats to usage with tumor samples: those with substantial allelic imbalance can be misinterpreted as contamination. Our contamination rate flag is 5%. No samples excluded from variant calling for the build.**

**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 samples is based on samtools mpileup function. Coverage is sorted from lowest to highest. Mosdepth coverage plots reflect the coverage distribution for each sample. One curve represents one sample. Due to large numbers of disease groups, 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**

**ExomeCQA(Qingyu Wang, Cooduvalli S. Shashikant, Matthew Jensen, et al. Novel metrics to measure coverage in whole exome sequencing datasets reveal local and global non-uniformity, Scientific Reports, 7-885, 2017) was run as the coverage qc tool for cohort-based WES datasets.**

**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.**

**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 in the capture region is.**

**Both tables and figures are stats based on gene level. For more details of target exon capture region level stats, refer to**

**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 dceg control sample (V3+UTR capturekit)**

**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 for all 5300 samples in familial\_build\_2019\_23806 in this summary. For more details and metrics, refer to:**

**PRECALLING\_QC\_REPORT**

**Fold80 is defined as the fold over-coverage necessary to raise 80% of bases in non-zero-coverage targets to the mean coverage level in those targets). The higher Fold80 penalty is, the less even coverage is.**

**CollectOxoGMetrics calculates the Phred-scaled probability that an alternate base call results from an oxidation artifact. This probability score is based on base context, sequencing read orientation, and the characteristic low allelic frequency. (Maura Costello, Trevor J. Pugh, Timothy J. Fennell, 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 Research, Volume 41, Issue 6, 2013).**

**CollectSequencingArtifactMetrics examines two sources of sequencing errors associated with hybrid 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 nucleic acid sample including DNA already damaged before library prep (such as FFPE). Hence named pre-adapter. 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. A lower Qscore means a higher probability that a REF\_BASE:ALT\_BASE observation randomly picked from the data will be due to this preadapter/baitbias artifact, rather than a true variant for the lowest base change.**

**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 (OPTIONAL)**

**Sex concordance between reads mapped to ChrY/ChrX ratio from sequence data and IDENTIFILER detected gender.**

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

**8. ANCESTRY CHECK**

**Laser tool (C Wang, X Zhan, L Liang, GR Abecasis, X Lin (2015) Improved Ancestry estimation for both genotyping and sequencing data using projection Procrustes analysis and genotype imputation. American Journal of Human Genetics, 96: 926-937.) 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: filtered variant total counts on case and control samples; ti/tv(transition/transversion) ratio on case and control samples; base change counts across different levels of phred-scale quality for each disease group.**

**In substitution mutations, transitions are defined as the interchange of the purine-based A↔G or pryimidine-based C↔T. Transversions are defined as the interchange between two-ring purine nucleobases and one-ring pyrimidine bases. The possible transversions are A↔C, A↔T, C↔G, G↔T. (Jing Wang, Leon Raskin, David C. Samuels, et al. Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics. 31(3): 318–323. 2015)**

**Figure 10a. Ensemble filtered variant count by sample (samples are sorted alphabetically). Variants are filtered by ensemble majority voting, capturekit intersection region filter.**

**For the elevated variant count of BL and PLCO group in the post filter plot, it is mainly due to ethical group (AFR vs EUR) difference.**

**Table 6a. 10 samples with lowest variant count**

**Figure 10b. Ensemble filtered variant ti/tv ratio by sample (samples are sorted alphabetically). Variants are filtered by ensemble majority voting, capturekit intersection region filter.**

**Table 6b. 10 samples with lowest ti/tv ratio**

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