**Sequencing QC Report (Somatic)**

# 1. INTRODUCTION

This is the QC report for BUILD\_NAME. 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**

# 2. 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 including multiple QC metrics. Some metrics are listed below with brief description:

**Per Base Sequence Quality**: Aggregated quality score statistics at each position along all reads in each sample/lane. It is normal with all Illumina sequencers for the median quality score to start out lower over the first 5-7 bases and to then rise. The average quality score will steadily drop over the length of the read. With paired end reads the average quality scores for read 1 will almost always be higher than for read 2.

**Per Sequence Quality Scores**: A plot of the total number of reads vs the average quality score over full length of that read. The distribution of average read quality should be fairly tight in the upper range of the plot.

**Per Base Sequence Content**: This plot reports the percent of bases called for each of the four nucleotides at each position across all reads in the file. For whole genome shotgun DNA sequencing the proportion of each of the four bases should remain relatively constant over the length of the read with %A=%T and %G=%C.

**Per Sequence GC Content**: Plot of the number of reads vs. GC% per read. A normal GC content for all reads is assumed to be the theoretical distribution. For whole genome shotgun sequencing the expectation is that the GC content of all reads should form a normal distribution with the peak of the curve at the mean GC content for the organism sequenced. If the observed distribution deviates too far from the theoretical, FastQC will call a Fail. The average GC-content in human genomes ranges from 35% to 60% across 100-Kb fragments, with a mean of 41%.

**Per Base N Content**: Percent of bases at each position or bin with no base call, i.e. ‘N’. In our production pipeline, the adapter bases will all be masked as N.

**Sequence Length Distribution**: Quality trimmed fastq files from our pipeline are shown in this report.

**Duplicate Sequences**: See section 2 in this report.

**Overrepresented Sequences**: List of sequences which appear more than expected in the file. A sequence is considered overrepresented if it accounts for ≥ 0.1% of the total reads. In DNA-Seq data no single sequence should be present at a high enough frequency to be listed, though it is not unusual to see a small percentage of adapter reads.

Please note, some QC cutoffs used in the multiQC are for reference purposes only and should not be followed without careful examination. For example, certain samples may be flagged as QC failures by “Per base sequence content" due to the biased composition of the first several bases of the read, which is a normal phenomenon for many types of libraries. In addition, samples failed on other multiQC checks may be included in the analysis due to 1). New/replacement sample materials are no longer available; 2). Investigators request to include them.

For all fastqc results, please refer to:

**FASTQC\_REPORT**

# 3. 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 or clustering duplicates within the same sequencing lane and PCR duplicates within one sample prep library run on one lane. The secondary duplication rate consists of PCR duplicates coming from the same sample prep library run across two or more lane level BAM files. Total duplication rate is the sum of the two. Figure1 is a violin plot showing total duplication rate grouped by number of lanes per subject.

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

The purpose of this plot is 1, showing the total duplication rate of each sample in the build; 2, showing how many lanes of data we’ve sequenced for each sample; 3, showing how the duplication rate is related to number of lanes data is coming from. A high duplication rate alone might just mean the DNA is over sequenced. The unique reads still provide sufficient information. A high duplication rate with an extremely low coverage is probably a sign of poor DNA quality (e.g. less than average cov 10X and more than 50% duplication rate).

**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)**

# 4. 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**

# 5. 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.

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

# 6. EXOMECQA

Coverage uniformity evaluation was also conducted using ExomeCQA4. The two scores (CCS and UE, defined below) are calculated based on both GENE\_NUM genes and EXON\_NUM target regions of capture kit 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**

# 7. 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 SAMPLE\_NUM samples 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 5a. histogram for all Fold 80 base penalty for samples**

**Figure 5b. boxplot for median and mean insert sizes**

**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**

**8. 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, sex outliers are samples showed ChrY/ChrX ratios lie outside the range of average ratio +/- standard deviation of the opposite sex.

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

# 9. 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).**

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

**LASER\_DIR**

# 10. 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 10a. VAF distribution across all base changes**

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

# 11. 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).