Admera Health
126 Corporate Boulevard • South Plainfield, NJ 07080
+1-908-222-0533 • custom-services@admerahealth.com • www.admerahealth.com
CLIA ID 31D2038676

Whole Exome Sequence Analysis Report

Platform: NovaSeq X Plus (2x150**)**
Capture Kit: IDT xGen Exome Hyb Panel v2
Project ID: XXXXX-XX
Species: Human
Reference: hg38

Analysis Schema

1. Bioinformatics Pipeline Overview

The diagram presented below offers an insightful overview of the comprehensive bioinformatics pipeline assembled by leveraging the finest components of diverse computational tools, extended by our proprietary in-house computational methods. To ensure transparency and reproducibility, list of the software and R-packages employed throughout these analyses, complete with specific versions, is provided at the end of this report.

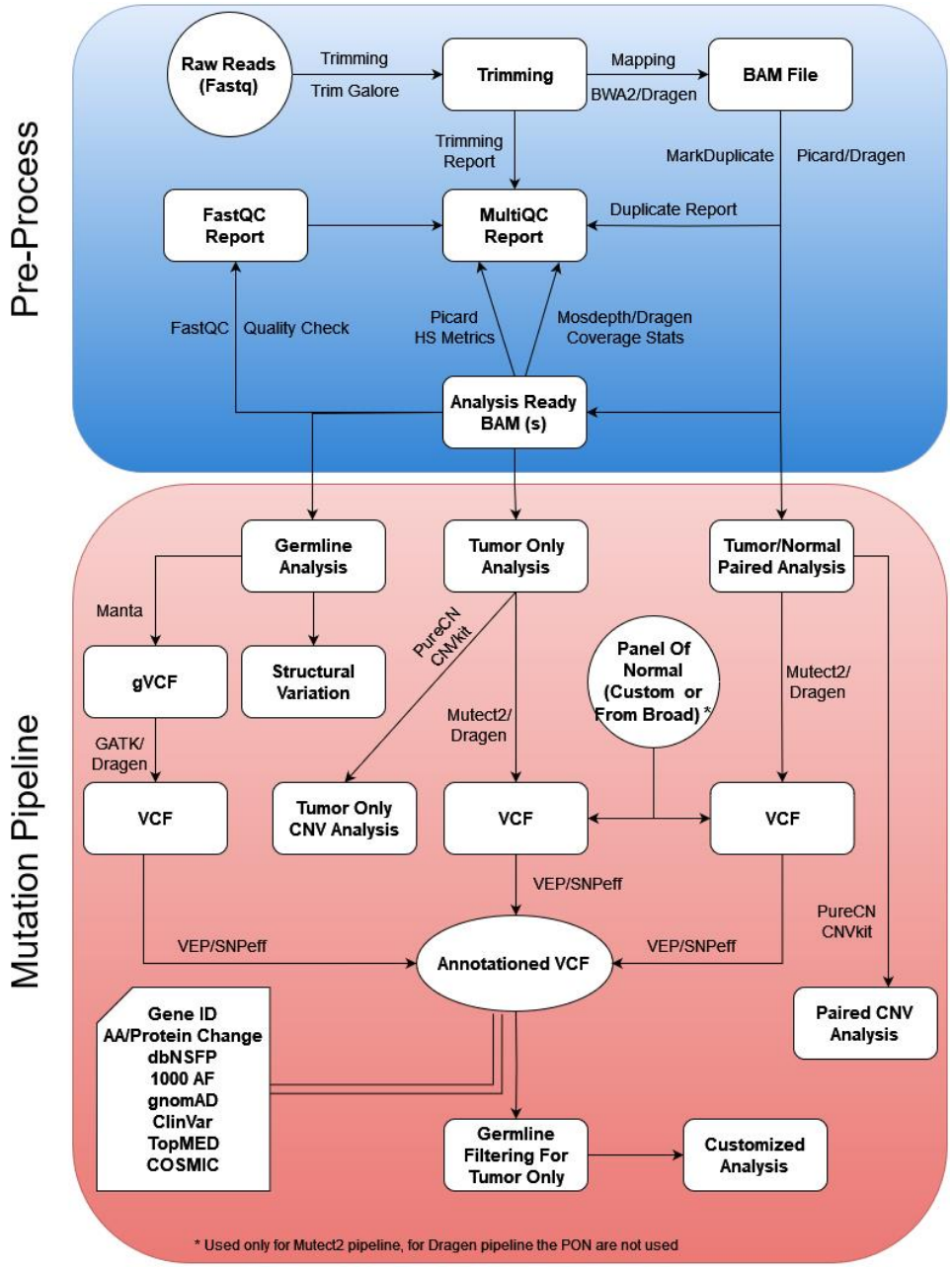

Figure 1 Flowchart of WES Pipeline

2. Trimming, Sequence Alignment and Duplicate Detection

As depicted in Figure 1, our data processing workflow begins with a critical step in which raw sequencing reads undergo trimming. This trimming process is essential to remove potential contamination from sequencing adapters or low-quality bases, which could otherwise impact the accuracy of subsequent analyses. After the trimming process, the remaining high-quality reads are aligned to the appropriate reference genome using the robust bwa-mem2 (or Dragen). Subsequently, duplicate reads are identified using the MarkDuplicates utility from Picard (or with Dragen), and analysis-ready BAM files are generated. Detailed duplication statistics are presented in Figure 2 for visual reference.

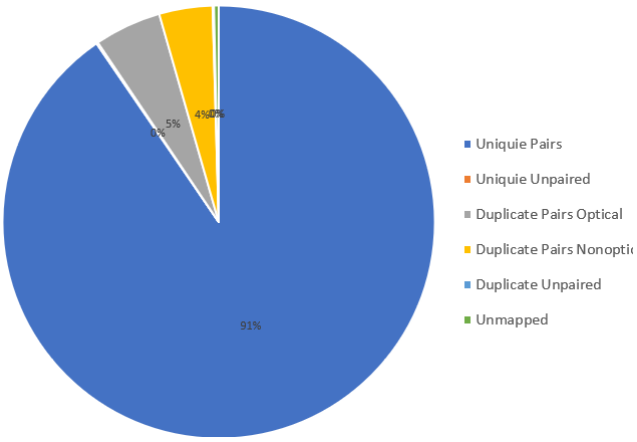

Figure 2 : Duplication Statistics

3. Comprehensive Quality Assessment

We initiated a comprehensive quality control (QC) analysis using the analysis-ready BAM file as our foundation. Initially, we employed the FastQC tool, a widely recognized tool for evaluating the quality of high-throughput sequencing data. The FastQC report supplied us with valuable insights into multiple aspects of the sequencing data, aiding in the detection of potential issues or biases that could impact downstream analyses. Subsequently, utilizing both the probes and target bed files, we extracted hybrid-selection (HS) metrics through the Picard tool. These metrics provided us with essential information regarding the efficiency and accuracy of the hybrid selection process. Table 1 presents a selection of significant measurements extracted from these metrics for reference. Furthermore, we calculated coverage statistics using Mosdepth (or Dragen), offering crucial insights into the depth and uniformity of coverage across the genome. To streamline our QC assessment, we aggregated various QC metrics into a user-friendly MultiQC HTML report. This comprehensive report, available as "multiqc_report.html," encompasses statistics related to trimming, the FastQC report, MarkDuplicate matrices, hybrid capture metrics, and coverage statistics.

Table 1 Key Hybrid Selection (HS) Metrics for Sequencing Experiment

Metric	Description	Value
PF_UQ_BASES_ALIGNED	The number of bases aligned to the target regions after quality filtering	4.64E+09
PCT_SELECTED_BASES	The fraction of PF_BASES_ALIGNED located on or near a baited region (ON_BAIT_BASES + NEAR_BAIT_BASES)/PF_BASES_ALIGNED.	0.925272
PCT_OFF_BAIT	The fraction of PF_BASES_ALIGNED that are mapped away from any baited region, OFF_BAIT_BASES/PF_BASES_ALIGNED.	0.074728
PCT_TARGET_BASES_1X	The fraction of all target bases achieving 1X or greater coverage.	0.98697
PCT_TARGET_BASES_2X	The fraction of all target bases achieving 2X or greater coverage.	0.985505
PCT_TARGET_BASES_10X	The fraction of all target bases achieving 10X or greater coverage.	0.977559
MEAN_BAIT_COVERAGE	The mean coverage of all baits in the experiment.	71.79495
MEAN_TARGET_COVERAGE	The mean coverage of a target region.	53.61135
PCT_USABLE_BASES_ON_BAIT	The number of aligned, de-duped, on-bait bases out of the PF bases available.	0.581808
PCT_USABLE_BASES_ON_TARGET	The number of aligned, de-duped, on-target bases out of all of the PF bases available.	0.350533
ZERO_CVG_TARGETS_PCT	The fraction of targets that did not reach coverage=1 over any base.	0.009739
PCT_EXC_DUPE	The fraction of aligned bases that were filtered out because they were in reads marked as duplicates.	0.113319
PCT_EXC_ADAPTER	The fraction of aligned bases that were filtered out because they were in reads with mapping quality 0 and the looked like adapter reads.	0
PCT_EXC_MAPQ	The fraction of aligned bases that were filtered out because they were in reads with low mapping quality.	0.040738
PCT_EXC_BASEQ	The fraction of aligned bases that were filtered out because they were of low base quality.	0.017257
PCT_EXC_OVERLAP	The fraction of aligned bases that were filtered out because they were the second observation from an insert with overlapping reads.	0.129501

4. Variant Detection and Annotations

Based on experiment type (Gernline vs Somatic) we employ different strategies to accurately detect and if available annotate mutations.

4.1. Germline Analysis

The germline analysis is the examination of an individual's genetic makeup as it is inherited from their parents. This analysis primarily involves non-tumor samples, such as those derived from blood or saliva. To detect germline SNPs and Indels, we utilized GATK4 (or Dragen). Subsequently, the identified genetic variants are annotated using available resources. Please review Annotations section for further details. If requested, we perform Structural Variation analysis using Manta tool.

4.2. Somatic Analysis

We employ MuTect2 workflow to identify potential somatic mutations within the genomic data. Initially, candidate somatic mutations are detected using base utility Mutect2 (or Dragen), when available matched normal sample is supplied to reduce potential germline variants. To further reduce false positive, particularly for Tumor only samples, we utilize Broad's Panel of Normals (PoN) using '-pon' flag (applicable to human samples only analysis using Mutect2). The step generates a list of potential mutations.

Next, to enhance the precision of the analysis using Mutect2 we obtained orientation model using LearnReadOrientationModel utility. The orientation models helps distinguishing genuine mutations from sequencing artifacts. Subsequently, to detect cross-sample contamination (human samples only), pileup summaries are calculated for SNPs common in ExAC03 and contamination table is created. Finally, from list of raw mutation, potential somatic variants are detected using the orientation bias model and contamination table (when available).

The Orthogonal Dragen somatic pipeline utilizes a distinct approach, employing a probability model that accounts for somatic, germline, and noise artifacts. It assigns "somatic quality" (SQ) scores to somatic variants and produces a VCF file containing filtered variants, ensuring reliable detection of somatic mutations in both tumor-normal and tumor-only samples. Please review here for detailed information https://support-docs.illumina.com/SW/dragen_v42/Content/SW/DRAGEN/SomaticMode.htm

The resulting somatic variants from either of the pipeline (Mutect2 or Dragen) are annotated using available resources. Please review Annotations section for further details. If requested, for paired samples, we perform somatic Structural variation analysis using Manta tool.

4.3. Annotations

For both Germline and Somatic mutations, we include essential functional annotations. These annotations cover a diverse array of species and encompass:

- Details regarding genes and transcripts impacted by mutations.
- The precise mutation locations, including classifications like upstream, downstream, exonic, intergenic, and more.
- Insights into the consequences of the mutations, ranging from stop gain and missense to frameshift mutations. Figure 3 and Table 2 below show the pie chart and counts

In the case of human samples, we conduct thorough annotations using prominent databases such as dbNSFP, COSMIC, and ClinVar. Additionally, we supply population allele frequency information sourced from resources like 1000 Genomes (1000G) and gnomAD.

Table 2 Coding Consequences Table

Count	Type
stop_gained	74
frameshift_variant	241
stop_lost	18
start_lost	14
inframe_insertion	182
inframe_deletion	193
missense_variant	10729
protein_altering_variant	2
start_retained_variant	4
stop_retained_variant	10
synonymous_variant	11707
coding_sequence_variant	22

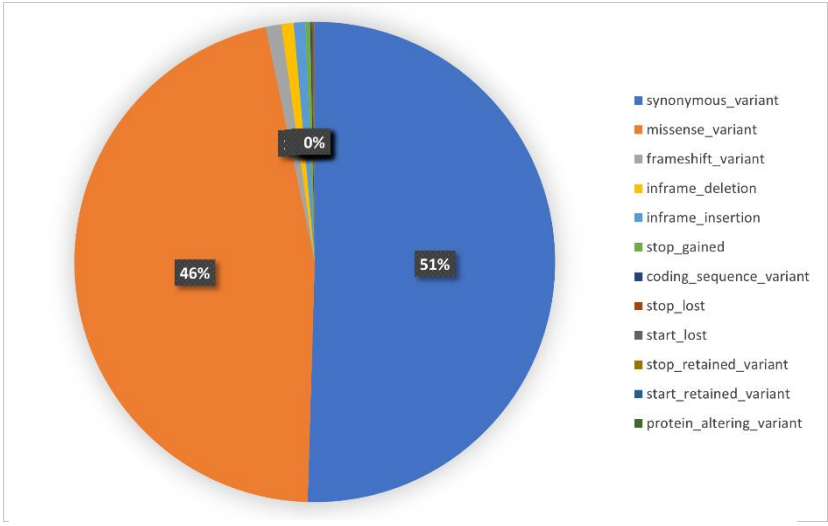

Figure 3 Coding Consequences Pie Chart

5. Contact Us

Address: 126 Corporate Boulevard, South Plainfield, New Jersey 07080
Email: custom-services@admerahealth.com
Phone: 908-222-0533