Output description for Exome pipeline

Exome pipeline is a bioinformatics best-practice variant calling analysis pipeline used for WES-Seq (whole exome sequencing) or target sequencing. The pipeline focused in variant calling and annotation of candidate low frequency variants.

This document describes the output produced by the pipeline and location of output files.

Pipeline overview:

The pipeline is built using Nextflow and processes data using the following steps:

- FastQC v0.11.3 - read quality control
- Trimmomatic v.0.33 - adapter and low quality trimming
- BWA v.0.7.12 - mapping against reference genome
- SAMtools v1.2 - alignment result processing
- Picard v.1.140 - enrichment and alignment metrics
- GATK v.3.4.46 - variant calling.
- KGGSeq v.0.8 - variant annotation.
- MultiQC v1.5 - quality statistics summary

> Each analysis folder contains a log folder with the log files for each process and each sample.

Preprocessing

FastQC

Quality control is performed using FastQC. FastQC gives general quality metrics about your reads. It provides information about the quality score distribution across your reads, the per base sequence content (%T/A/G/C). You get information about adapter contamination and other overrepresented sequences. For further reading and documentation see the FastQC help.

Results directory: ANALYSIS/{ANALYSIS_ID}/QC/fastqc

- There is one folder per sample.
- Files:
  - `{sample_id}/{sample_id}_R[12]_fastqc.html`: html report. This file can be opened in your favourite web browser (Firefox/chrome preferable) and it contains the different graphs that fastqc calculates for QC.
  - `{sample_id}/{sample_id}_R[12]_fastqc` : folder with fastqc output in plain text.
  - `{sample_id}/{sample_id}_R[12]_fastqc.zip`: zip compression of above folder.
  - `{sample_id}/{sample_id}.trimmed_R[12]_fastqc.html`: html report. This file can be opened in your favourite web browser (Firefox/chrome preferable) and it contains the different graphs that fastqc calculates for QC for trimmed reads.
  - `{sample_id}/{sample_id}.trimmed_R[12]_fastqc`: folder with fastqc output in plain text for trimmed reads.
  - `{sample_id}/{sample_id}.trimmed_R[12]_fastqc.zip`: zip compression of above folder for trimmed reads.

Trimming

Table of Contents

- Pipeline overview:
- Preprocessing
  - FastQC
  - Trimming
- Mapping
  - BWA
- Picard
  - MarkDuplicates
  - HsMetrics
- Variant Calling
  - GATK
- Annotation and Filtering
  - KGGSeq
- Quality control report
  - MultiQC
- Annex II
- Annex III

Trimmomatic is used for removal of adapter contamination and trimming of low quality regions. Parameters included for trimming are:

- Nucleotides with phred quality < 10 in 3'end.
- Mean phred quality < 15 in a 4 nucleotide window.
- Read lenght < 70

MultiQC reports the percentage of bases removed by trimming in bar plot showing percentage or reads trimmed in forward and reverse.

Note:The FastQC plots displayed in the MultiQC report shows *untrimmed* reads. They may contain adapter sequence and potentially regions with low quality. To see how your reads look after trimming, look at the FastQC reports in the `ANALYSIS/{ANALYSIS_ID}/QC/fastqc` directory.

Results directory: ANALYSIS/{ANALYSIS_ID}/QC/trimmomatic

- There is one folder per sample.
- Files:
  - `{sample_id}/{sample_id}_R[12].trimmed.fastq.gz`: contains high quality reads with both forward and reverse tags surviving.
  - `{sample_id}/{sample_id}_R[12]_unpaired.fastq.gz`: contains high quality reads with only forward or reverse tags surviving.

NOTE: This results are not delivered to the researcher by default due to disk space issues. If you are interesested in using them, please contact us and we will add them to your delivery.

Mapping

BWA

BWA, or Burrows-Wheeler Aligner, is designed for mapping low-divergent sequence reads against reference genomes. The result alignment files are further processed with SAMtools, sam format is converted to bam, sorted and an index *.bai* is generated.

Results directory: ANALYSIS/{ANALYSIS_ID}/Alignment

- There is one folder per sample.
- This files can be used in IGV for alignment visualization.
- Files:
  - `{sample_id}/{sample_id}_sorted.bam` : sorted aligned bam file.
  - `{sample_id}/{sample_id}_sorted.bam.bai`: index file for soreted aligned bam.

NOTE: This results are not removed due to disk space issues, only last bam processed bam file is retained.. If you are interesested in using them, please contact us and we will try to generate them and add them to your delivery.

Picard

MarkDuplicates

The MarkDuplicates module in the Picard toolkit differentiates the primary and duplicate reads using an algorithm that ranks reads by the sums of their base-quality scores, which helps to identify duplicates that arise during sample preparation e.g. library construction using PCR.

The Picard section of the MultiQC report shows a bar plot with the numbers and proportions of primary reads, duplicate reads and unmapped reads.

Results directory: ANALYSIS/{ANALYSIS_ID}/Alignment

- There is one folder per sample.
- This files can be used in IGV for alignment visualization.
- Files:
  - `{sample_id}/{sample_id}.woduplicates.bam` : sorted aligned bam file.

- `{sample_id}/{sample_id}.woduplicates.bam.bai`: index file for soreted aligned bam.

HsMetrics

Metrics for the analysis of target-capture sequencing experiments are calculated with Picard CollectHsMetrics. The metrics in this class fall broadly into three categories:

- Basic sequencing metrics that are either generated as a baseline against which to evaluate other metrics or because they are used in the calculation of other metrics. This includes things like the genome size, the number of reads, the number of aligned reads etc.
- Metrics that are intended for evaluating the performance of the wet-lab assay that generated the data. This group includes metrics like the number of bases mapping on/off/near baits, %selected, fold 80 base penalty, hs library size and the hs penalty metrics. These metrics are calculated prior to some of the filters are applied (e.g. low mapping quality reads, low base quality bases and bases overlapping in the middle of paired-end reads are all counted).
- Metrics for assessing target coverage as a proxy for how well the data is likely to perform in downstream applications like variant calling. This group includes metrics like mean target coverage, the percentage of bases reaching various coverage levels, and the percentage of bases excluded by various filters. These metrics are computed using the strictest subset of the data, after all filters have been applied.

The Picard section of the MultiQC report shows a table with data as the Bait design efficiency, Bait territory, Fold 80 base penalty, Fold enrichment; near, on and off bait bases; on target bases, Pf (passing filter) reads, Pf unique reads, Pf uq (unique) bases aligned, Pf uq reads aligned, Total reads Mean bait coverage, Mean target coverage, On bait vs selected, Target territory, Zero cvg targets pct.

Results directory: ANALYSIS/{ANALYSIS_ID}/stats/bamstats

- Files:
  - `hsMetrics_all.out` : summary of the some of the most meaningful columns in picard hsmetrics output for all the samples in the project.
  - `{sample_id}_hsMetrics.out`: full picard hsmetrics output per sample.
  - Description of Picard hsMetrics columns in its output can be found in AnnexIII

Variant Calling

GATK

Best Practice GATK protocol has been used for the variant calling of germinal variations in the family (http://www.broadinstitute.org/gatk/guide/best-practices, november 2015). This workflow comprises three different steps:

1. **Data preprocessing:**

    1. **Realignment:** starting from BAM file generated with BWA and Picard MarkDuplicates, realignment around candidate indels is performed in order to improve mapping in complicated zones (low complexity, homopolymers, etc).

       Results directory: ANALYSIS/{ANALYSIS_DIR}/variants/variants_gatk/realignment

       - `{sample_id}.realigned.bam` : realigned bam.
       - `{sample_id}.realigned.bam-bai`: index for realigned bam.
       - `{sample_id}.woduplicates.bam-IndelRealigner.intervals`: file with "problematic" regions

where realignment was needed.

NOTE: This results are not removed due to disk space issues, only last bam processed bam file is retained.. If you are interesested in using them, please contact us and we will try to generate them and add them to your delivery.

2. **Base Recalibrarion**: Next step carries out a phed quality recalibration of bases, using a gold standard set of known SNPS (dbSNP138) using a machine learning approach.

   Results directory:
   ANALYSIS/{ANALYSIS_DIR}/variants/variants_gatk/recalibration

   - `{sample_id}.woduplicates.bam-BQSR.pdf` : pdf file with quality graphics before and after recalibration.
   - `{sample_id}.recalibrated.bam` : recalibrated bam.
   - `{sample_id}.recalibrated.bai` : index for recalibrated bams.
   - `{sample_id}.woduplicates.bam-BQSR.csv`: intermediate file.
   - `{sample_id}.woduplicates.bam-recal2_data.grp`: intermediate file.

   NOTE: BAM files are removed due to disk space issues. If you are interesested in using them, please contact us and we will try to generate them and add them to your delivery.

2. **Variant Calling:** Variant calling is performed for all the samples obtaining a multivcf file. Vcf (variant calling format) describes the variants (positions different from genome reference) present in a group of samples and its genotype in each sample.

   1. **HaplotypeCaller:** this module is capable of calling SNPs and indels simultaneously via local de-novo assembly of haplotypes in an active region. In other words, whenever the program encounters a region showing signs of variation, it discards the existing mapping information and completely reassembles the reads in that region. This allows the HaplotypeCaller to be more accurate when calling regions that are traditionally difficult to call, for example when they contain different types of variants close to each other.
      1. Define active regions: The program determines which regions of the genome it needs to operate on, based on the presence of significant evidence for variation.
      2. Determine haplotypes by assembly of the active regions: For each ActiveRegion, the program builds a De Bruijn-like graph to reassemble the ActiveRegion, and identifies what are the possible haplotypes present in the data. The program then realigns each haplotype against the reference haplotype using the Smith-Waterman algorithm in order to identify potentially variant sites.
      3. Determine likelihoods of the haplotypes given the read data: For each ActiveRegion, the program performs a pairwise alignment of each read against each haplotype using the PairHMM algorithm. This produces a matrix of likelihoods of haplotypes given the read data. These likelihoods are then marginalized to obtain the likelihoods of alleles for each potentially variant site given the read data.
      4. Assign sample genotypes: For each potentially variant site, the program applies Bayes' rule, using the likelihoods of alleles given the read data to calculate the likelihoods of each genotype per sample given the read data observed for that sample. The most likely genotype is then assigned to the samples.
   2. **HardFiltering:** quality filtering of variants following GATK best practices:
      - MQ < 40. RMSMappingQuality. This is the Root Mean Square of the mapping quality of the reads across all

samples.
    - DP <5. LowCoverage
    - QD <2.0. LowQD. This is the variant confidence (from the QUAL field) divided by the unfiltered depth of non-reference samples.
    - FS >60.0. p-value StrandBias. Phred-scaled p-value using Fisher's Exact Test to detect strand bias (the variation being seen on only the forward or only the reverse strand) in the reads. More bias is indicative of false positive calls
    - MQRankSum < -12.5. MappingQualityRankSumTest. This is the u-based z-approximation from the Mann-Whitney Rank Sum Test for mapping qualities (reads with ref bases vs. those with the alternate allele). Note that the mapping quality rank sum test can not be calculated for sites without a mixture of reads showing both the reference and alternate alleles, i.e. this will only be applied to heterozygous calls.
    - ReadPosRankSum < -8.0. VariantReadPosEnd. This is the u-based z-approximation from the Mann-Whitney Rank Sum Test for the distance from the end of the read for reads with the alternate allele. If the alternate allele is only seen near the ends of reads, this is indicative of error. Note that the read position rank sum test cannot be calculated for sites without a mixture of reads showing both the reference and alternate alleles, i.e. this will only be applied to heterozygous calls.
    - SOR > 4.0. StrandOddRank. Strandbias.
3. **Genotype refinement**:
    - PhaseByTransmission: técnica estadística para determinar cuándo es posible qué alelo proviene del padre y cuál de la madre en el niño. Por consenso se pone primero el alelo de la madre y luego el del padre. Madre 1/0 padre 0/0, niño 1(madre) | 0(padre). Sólo se permite el análisis de trios actualmente, por lo que sólo aparecerán las fases calculadas para padre, madre y probando. No se realizará phasing en los hermanos.
    - ReadBackedPhasing: determina la presencia de haplotipos en cada muestra, no entre ellas. Busca grupos de snps que se encuentran en el mismo cromosoma. Por lo que entiendo haplotypeCaller utiliza este mecanismo para determinar los genotipos pero no lo marca en el vcf. Al correr este Walker en el vcf se marcan los genotipos con la | en vez de con / cuando se ha conseguido determinar un haplotipo.

Results directory:
ANALYSIS/{ANALYSIS_ID}/variant_calling/variants_gatk/variants

- Files:
    - `all_samples.vcf`: raw variants outputed by HaplotypeCaller.
    - `all_samples_snps.vcf`: only raw snps.
    - `all_samples_snps_fil.vcf`: snps with marked filters.
    - `all_samples_indels.vcf`: only raw indels.
    - `all_samples_indels_fil.vcf`: indels with marked filters.
    - `all_samples_fil.vcf`: snps and indels filtered combined.
    - `all_samples_PhaseByTransmission.vcf`: variants with phases calculated.
    - `all_samples_ReadBackedPhasing.vcf`: variants with backend phases calculated.
    - `all_samples_gtpos.vcf`: variants with genotype quality fixed.
    - `all_samples_gtpos_fil.vcf`: variants with marked filters.
    - `all_samples_gtpos_fil_annot.vcf`: variants with putative de novo mutations annotated.

Annotation and Filtering

KGGSeq

KGGSeq is used for post-analysis and annotation of variants obtained with GATK (Li, Gui, Kwan, Bao, & Sham, 2012). KGGSeq is used for variant annotation, a tool design for variant priorization in the study of mendelian diseases.

Besides functional annotation some variant filtering is performed:

- Depth < 4
- GQ < 10.0
- PL < 20
- Sequencing quality < 50.0
- Population frequency in ANY database (ESP5400,dbsnp141,1kg201305,exac) > 0.005
- Bed enrichment filter: two set of variants are annotated, all variants called and only variants intersecting the bed enrichment file provided by the commercial library kit.

Moreover variants are prioritise by Genetic inheritance sharing according to the researcher request:

- *De novo*: ONLY include variants at which an offspring has one or two non-inherited alleles AND Exclude variants at which both affected and unaffected subjects have the same heterozygous genotypes
- *Recessive and Compound-heterozygosity*: This function is designed for a disorder suspected to be under compound-heterozygous or recessive inheritance mode, in which both copies of a gene on the two ortholog chromosomes of a patient are damaged by two different mutations or the same mutation. For recessive mode, it simply checks variants with homozygous genotypes in patients. For the compound-heterozygous mode, it can use two different input data, phased genotypes of a patient or unphased genotypes in a trio. Here the trio refers to the two parents and an offspring. When these alleles causing a disease at one locus, it follows the recessive model; and when they are at two loci, it follows the compound-heterozygosity model. In both cases, a gene is hit twice. This is the reason why it has the name 'double-hit gene'.).

Results directory: ANALYSIS/{ANALYSIS_ID}/annotation

- Files:
  - All variants (not filtered by the bed file)
  - `variants_denovo_phased.tab`: FINAL annotation file only for denovo mutations. It contains all paramters from variant calling and all annotation from KGGSeq.
  - `variants_doublehit_phased.tab`: FINAL annotation file only for recessive and doublehit mutations. It contains all paramters from variant calling and all annotation from KGGSeq.
  - `all_samples_gtpos_fil_annot_doublehit.vcf.doublehit.gene.trios.flt.count.txt`: file with double-hit genes.
  - `all_samples_gtpos_fil_annot_doublehit.vcf.doublehit.gene.trios.flt.gty.txt`: file with genotype for double-hit genes.
  - `all_samples_gtpos_fil_annot_doublehit.vcf.flt.txt`: double-hit variants with annotation.
  - `all_samples_gtpos_fil_annot_denovo.vcf.flt.txt`: de novo variants with annotation.
  - Only variants filtered by the bed file. Same as above but in `bedfilter` folder.

Quality control report

MultiQC

MultiQC is a visualisation tool that generates a single HTML report summarising all samples in your project. Most of the pipeline QC results are visualised in the report and further statistics are available in within the report data directory.

Output directory: ANALYSIS/{ANALYSIS_ID}/stats

- `multiqc_report.html`: MultiQC report - a standalone HTML file that can

be viewed in your web browser
- `multiqc_data/`: Directory containing parsed statistics from the different tools used in the pipeline

For more information about how to use MultiQC reports, see http://multiqc.info

Annex II

Column	Meaning
Chromosome	chromosome number
StartPosition	Human genome reference position
ReferenceAlternativeAllele	reference/alternative allele
rsID	SNP rs ID
MostImportantFeatureGene	Gene Symbol
MostImportantGeneFeature	Gene feature {missense,intronic, ncRNA, etc}
RefGeneFeatures	Gene Features {codons,transcripts,etc}
SLR	Sitewise Likelihood-ratio (SLR) test statistic for testing natural selection on codons. A negative value indicates negative selection, and a positive value indicates positive selection. Larger magnitude of the value suggests stronger evidence.
SIFT_score	SIFT uses the 'Sorting Tolerant From Intolerant' (SIFT) algorithm to predict whether a single amino acid substitution affects protein function or not, based on the assumption that important amino acids in a protein sequence should be conserved throughout evolution and substitutions at highly conserved sites are expected to affect protein function.A small scoreindicates a high chance for a substitutionto damage the protein function.
Polyphen2_HDIV_score	"Polyphen2 score based on HumDiv, i.e. hdiv_prob. The score ranges from 0 to 1, and the corresponding prediction is ""probably damaging"" if it is in [0.957,1]; ""possibly damaging"" if it is in [0.453,0.956]; ""benign"" if it is in [0,0.452]. Score cutoff for binary classification is 0.5, i.e. the prediction is ""neutral"" if the score is smaller than 0.5 and ""deleterious"" if the score is larger than 0.5. Multiple entries separated by "";"""
Polyphen2_HVAR_score	Polyphen2 predicts the possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations by an iterative greedy algorithm. In the present study, we use the original scores generated by the HumVar (instead ofHumDiv) trained model as it is preferred for the diagnosis of Mendelian diseases. The scores range from 0 to 1. A substitution with larger score has a higher possibility to damage the protein function.

Column	Meaning
LRT_score	LRT employed a likelihood ratio test to assess variant deleteriousnessbased on a comparative genomics data set of 32 vertebrate species. The identified deleterious mutations could disrupt highly conserved amino acids within protein-coding sequences, which are likely to be unconditionally deleterious.The scores range from 0 to 1. A larger score indicates a larger deleterious effect.
MutationTaster_score	MutationTaster assesses the impact of the disease-causing potential of a sequence variant by a naive Bayes classifier using multiple resources such as evolutionary conservation, splice-site changes, loss of protein features and changes that might affect mRNA level. The scores range from 0 to 1. The larger score suggests a higher probability to cause a human disease.
MutationAssessor_score	"MutationAssessor ""functional impact of a variant : predicted functional (high, medium), predicted non-functional (low, neutral)"" Please refer to Reva et al. Nucl. Acids Res. (2011) 39(17):e118 for details"
FATHMM_score	"FATHMM default score (weighted for human inherited-disease mutations with Disease Ontology); If a score is smaller than -1.5 the corresponding NS is predicted as ""D(AMAGING)""; otherwise it is predicted as ""T(OLERATED)"". If there's more than one scores associated with the same NS due to isoforms, the smallest score (most damaging) was used. Please refer to Shihab et al Hum. Mut. (2013) 34(1):57-65 for details"
VEST3	VEST 3.0 score. Score ranges from 0 to 1. The larger the score the more likely the mutation may cause functional change. In case there are multiple scores for the same variant, the largest score (most damaging) is presented. Please refer to Carter et al., (2013) BMC Genomics. 14(3) 1-16 for details. Please note this score is free for non-commercial use. For more details please refer to http://wiki.chasmsoftware.org/index.php/SoftwareLicense. Commercial users should contact the Johns Hopkins Technology Transfer office.
CADD_score	Combined Annotation Dependent Depletion (CADD) score for funtional prediction of a SNP. Please refer to Kircher et al. (2014) Nature Genetics 46(3):310-5 for details. The larger the score the more likely the SNP has damaging effect.
GERP++_NR	Neutral rate
GERP++_RS	RS score, the larger the score, the more conserved the site
phyloP	PhyloP estimates the evolutional conservation at each variant from multiple alignments of placental mammal genomes to the human genome based on a phylogenetic hidden Markov model.
29way_logOdds	SiPhy score based on 29 mammals genomes. The larger the score, the more conserved the site.

Column	Meaning
LRT_Omega	Estimated nonsynonymous-to-synonymous-rate ratio (reported by LRT)
AffectedRefHomGtyNum	Number of affected individuals with reference homozygote at this variant;
AffectedHetGtyNum	Number of affected individuals with heterozygote at this variant;
AffectedAltHomGtyNum	Number of affected individuals with non-ref homozygote;
UnaffectedRefHomGtyNum	Number of unaffected individuals with reference homozygote at this variant;
UnaffectedHetGtyNum	Number of unaffected individuals with heterozygote at this variant;
UnaffectedAltHomGtyNum	Number of unaffected individuals with non-ref homozygote;
DenovoMutationEvent	"n the main output file, there is a column named DenovoMutationEvent to record the genotypes of a child and his or her parents. Example: N140_0:0/1:46,59&N140_1:0/0:57,0&N140_2:0/0:68,0. The child N140_0 has genotype 0/1 with 46 and 59 reads carrying reference alleles and alternative alleles respectively. The father N140_1 and mother N140_2 are homozygous 0/0."
UniProtFeatureForRefGene	Annotate a variant of coding gene using the UniProt protein annotations.
GeneDescription	Gene description
Pseudogenes	Pseudogenes listed in http://tables.pseudogene.org/set.py?id=Human61
DiseaseName(s)MIMid	"Disorder, () Phenotype mapping method :1 - the disorder is placed on the map based on its association with a gene, but the underlying defect is not known.2 - the disorder has been placed on the map by linkage; no mutation has been found. 3 - the molecular basis for the disorder is known; a mutation has been found in the gene.4 - a contiguous gene deletion or duplication syndrome, multiple genes are deleted or duplicated causing the phenotype."
GeneMIMid	GeneMIMid : Gene/locus MIM no.
SIFT_pred	SIFT prediction filter
Polyphen2_HDIV_pred	"Polyphen2 prediction based on HumDiv, ""D"" (""porobably damaging""), ""P"" (""possibly damaging"") and ""B"" (""benign""). Multiple entries separated by "";"""

Column	Meaning
Polyphen2_HVAR_pred	"Polyphen2 prediction based on HumVar, ""D"" (""porobably damaging""), ""P"" (""possibly damaging"") and ""B"" (""benign""). Multiple entries separated by "";"""
LRT_pred	Classification using LRT (D = deleterious, N = neutral, or U = unknown)
MutationTaster_pred	Classification using MutationTaster (A = disease_causing_automatic, D = disease_causing, N = polymorphism, or P = polymorphism_automatic)
MutationAssessor_pred	"MutationAssessor ""functional impact of a variant : predicted functional (high, medium), predicted non-functional (low, neutral)"""
FATHMM_pred	FATHMM prediction filter.
DiseaseCausalProb_ExoVarTrainedModel	Conditional probability of being Mendelian disease-causing given the above prediction scores under a logistic regression model trained by our dataset ExoVar.
IsRareDiseaseCausal_ExoVarTrainedModel	Classification using the logistic regression model (Y = disease-causing or N = neutral)
BestCombinedTools:OptimalCutoff:TP:TN	The subset of original prediction tools (out of the 13 tools) used for the combined prediction by our Logistic Regression model which have the largest posterior probability among all possible combinatorial subsets: the cutoff leads to the maximal Matthews correlation coefficient (MCC): the corresponding true positive and true negative at the maximal MCC.
TFBSconsSite[tfbsName:rawScore:zScore]	Conserved TFBSs in the UCSC genome browser
vistaEnhancer[enhancerName:positive/negative]	Known enhancers in the VISTA enhancer browser
PubMedIDIdeogram	PubMed ID of articles in which the term and the cytogeneic position of the variant are co-mentioned
PubMedIDGene	PubMed ID of articles in which the term and the gene containing the variant are co-mentioned

Annex III

BAIT_SET	The name of the bait set used in the hybrid selection.
GENOME_SIZE	The number of bases in the reference genome used for alignment.
BAIT_TERRITORY	The number of bases which are localized to one or more baits.
TARGET_TERRITORY	The unique number of target bases in the experiment, where the target sequence is usually exons etc.

BAIT_SET	**The name of the bait set used in the hybrid selection.**
BAIT_DESIGN_EFFICIENCY	The ratio of TARGET_TERRITORY/BAIT_TERRITORY. A value of 1 indicates a perfect design efficiency, while a valud of 0.5 indicates that half of bases within the bait region are not within the target region.
TOTAL_READS	The total number of reads in the SAM or BAM file examined.
PF_READS	The total number of reads that pass the vendor's filter.
PF_UNIQUE_READS	The number of PF reads that are not marked as duplicates.
PCT_PF_READS	The fraction of reads passing the vendor's filter, PF_READS/TOTAL_READS.
PCT_PF_UQ_READS	The fraction of PF_UNIQUE_READS from the TOTAL_READS, PF_UNIQUE_READS/TOTAL_READS.
PF_UQ_READS_ALIGNED	The number of PF_UNIQUE_READS that aligned to the reference genome with a mapping score > 0.
PCT_PF_UQ_READS_ALIGNED	The fraction of PF_UQ_READS_ALIGNED from the total number of PF reads.
PF_BASES_ALIGNED	The number of PF unique bases that are aligned to the reference genome with mapping scores > 0.
PF_UQ_BASES_ALIGNED	The number of bases in the PF_UQ_READS_ALIGNED reads. Accounts for clipping and gaps.
ON_BAIT_BASES	The number of PF_BASES_ALIGNED that are mapped to the baited regions of the genome.
NEAR_BAIT_BASES	The number of PF_BASES_ALIGNED that are mapped to within a fixed interval containing a baited region, but not within the baited section per se.
OFF_BAIT_BASES	The number of PF_BASES_ALIGNED that are mapped away from any baited region.
ON_TARGET_BASES	The number of PF_BASES_ALIGNED that are mapped to a targeted region of the genome.
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.

BAIT_SET	**The name of the bait set used in the hybrid selection.**
PCT_OFF_BAIT	The fraction of PF_BASES_ALIGNED that are mapped away from any baited region, OFF_BAIT_BASES/PF_BASES_ALIGNED.
ON_BAIT_VS_SELECTED	The fraction of bases on or near baits that are covered by baits, ON_BAIT_BASES/(ON_BAIT_BASES + NEAR_BAIT_BASES).
MEAN_BAIT_COVERAGE	The mean coverage of all baits in the experiment.
MEAN_TARGET_COVERAGE	The mean coverage of a target region.
MEDIAN_TARGET_COVERAGE	The median coverage of a target region.
MAX_TARGET_COVERAGE	The maximum coverage of reads that mapped to target regions of an experiment.
PCT_USABLE_BASES_ON_BAIT	The number of aligned, de-duped, on-bait bases out of the PF bases available.
PCT_USABLE_BASES_ON_TARGET	The number of aligned, de-duped, on-target bases out of all of the PF bases available.
FOLD_ENRICHMENT	The fold by which the baited region has been amplified above genomic background.
ZERO_CVG_TARGETS_PCT	The fraction of targets that did not reach coverage=1 over any base.
PCT_EXC_DUPE	The fraction of aligned bases that were filtered out because they were in reads marked as duplicates.
PCT_EXC_MAPQ	The fraction of aligned bases that were filtered out because they were in reads with low mapping quality.
PCT_EXC_BASEQ	The fraction of aligned bases that were filtered out because they were of low base quality.
PCT_EXC_OVERLAP	The fraction of aligned bases that were filtered out because they were the second observation from an insert with overlapping reads.
PCT_EXC_OFF_TARGET	The fraction of aligned bases that were filtered out because they did not align over a target base.
FOLD_80_BASE_PENALTY	The fold over-coverage necessary to raise 80% of bases in "non-zero-cvg" targets to the mean coverage level in those targets.
PCT_TARGET_BASES_1X	The fraction of all target bases achieving 1X or greater coverage.

BAIT_SET	The name of the bait set used in the hybrid selection.
PCT_TARGET_BASES_2X	The fraction of all target bases achieving 2X or greater coverage.
PCT_TARGET_BASES_10X	The fraction of all target bases achieving 10X or greater coverage.
PCT_TARGET_BASES_20X	The fraction of all target bases achieving 20X or greater coverage.
PCT_TARGET_BASES_30X	The fraction of all target bases achieving 30X or greater coverage.
PCT_TARGET_BASES_40X	The fraction of all target bases achieving 40X or greater coverage.
PCT_TARGET_BASES_50X	The fraction of all target bases achieving 50X or greater coverage.
PCT_TARGET_BASES_100X	The fraction of all target bases achieving 100X or greater coverage.
HS_LIBRARY_SIZE	The estimated number of unique molecules in the selected part of the library.
HS_PENALTY_10X	The "hybrid selection penalty" incurred to get 80% of target bases to 10X. This metric should be interpreted as: if I have a design with 10 megabases of target, and want to get 10X coverage I need to sequence until PF_ALIGNED_BASES = 10^7 * 10 * HS_PENALTY_10X.
HS_PENALTY_20X	The "hybrid selection penalty" incurred to get 80% of target bases to 20X. This metric should be interpreted as: if I have a design with 10 megabases of target, and want to get 20X coverage I need to sequence until PF_ALIGNED_BASES = 10^7 * 20 * HS_PENALTY_20X.
HS_PENALTY_30X	The "hybrid selection penalty" incurred to get 80% of target bases to 30X. This metric should be interpreted as: if I have a design with 10 megabases of target, and want to get 30X coverage I need to sequence until PF_ALIGNED_BASES = 10^7 * 30 * HS_PENALTY_30X.
HS_PENALTY_40X	The "hybrid selection penalty" incurred to get 80% of target bases to 40X. This metric should be interpreted as: if I have a design with 10 megabases of target, and want to get 40X coverage I need to sequence until PF_ALIGNED_BASES = 10^7 * 40 * HS_PENALTY_40X.

BAIT_SET	The name of the bait set used in the hybrid selection.
HS_PENALTY_50X	The "hybrid selection penalty" incurred to get 80% of target bases to 50X. This metric should be interpreted as: if I have a design with 10 megabases of target, and want to get 50X coverage I need to sequence until PF_ALIGNED_BASES = 10^7 * 50 * HS_PENALTY_50X.
HS_PENALTY_100X	The "hybrid selection penalty" incurred to get 80% of target bases to 100X. This metric should be interpreted as: if I have a design with 10 megabases of target, and want to get 100X coverage I need to sequence until PF_ALIGNED_BASES = 10^7 * 100 * HS_PENALTY_100X.
AT_DROPOUT	A measure of how undercovered <= 50% GC regions are relative to the mean. For each GC bin [0..50] we calculate a = % of target territory, and b = % of aligned reads aligned to these targets. AT DROPOUT is then abs(sum(a-b when a-b < 0)). E.g. if the value is 5% this implies that 5% of total reads that should have mapped to GC<=50% regions mapped elsewhere.
GC_DROPOUT	A measure of how undercovered >= 50% GC regions are relative to the mean. For each GC bin [50..100] we calculate a = % of target territory, and b = % of aligned reads aligned to these targets. GC DROPOUT is then abs(sum(a-b when a-b < 0)). E.g. if the value is 5% this implies that 5% of total reads that should have mapped to GC>=50% regions mapped elsewhere.
HET_SNP_SENSITIVITY	The theoretical HET SNP sensitivity.
HET_SNP_Q	The Phred Scaled Q Score of the theoretical HET SNP sensitivity.

Bibliography 1. *McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. a. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9), 1297–1303. doi:10.1101/gr.107524.110.20* 2. *Li, M.-X., Gui, H.-S., Kwan, J. S. H., Bao, S.-Y., & Sham, P. C. (2012). A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic acids research, 40(7), e53. doi:10.1093/nar/gkr1257*