Variant Stats

Variant Stats contain a basic information for each variant in a different cohort.

Table of Contents:

- Implementation
- Input
  - Parameters
- Output
  - Files
    - Variant Stats Data Model
  - Index
- Useful Links

Implementation

Variant Stats is implemented using Hadoop MapReduce over HBase.

Input

Parameters

OpenCGA support different input parameters:

- Variant Query
- Sample list, cohort or query

Output

Files

If the stats are not indexed, the analysis produces a Variant stats file in json format with the following model schema:

Variant Stats Data Model

cohortId *String*	Unique cohort identifier within the study.
sampleCount *int*	Count of samples with non-missing genotypes in this variant from the cohort. This value is used as denominator for genotypeFreq.
fileCount *int*	Count of files with samples from the cohort that reported this variant. This value is used as denominator for filterFreq.
alleleCount *int*	Total number of alleles in called genotypeCounters. It does not include missing alleles. This value is used as denominator for refAlleleFreq and altAlleleFreq.
refAlleleCount *int*	Number of reference alleles found in this variant.
refAlleleFreq *float*	Reference allele frequency calculated from refAlleleCount and alleleCount, in the range [0,1]
altAlleleCount *int*	Number of main alternate alleles found in this variants. It does not include secondary alternates.
altAlleleFreq *float*	Alternate allele frequency calculated from altAlleleCount and alleleCount, in the range [0,1]
missingAlleleCount *int*	Number of missing alleles.
missingGenotypeCount *int*	Number of genotypes with all alleles missing (e.g. ./.). It does not count partially missing genotypes like "./0" or "./1".
genotypeCount *Map<String, int>*	Number of occurrences for each genotype. This does not include genotype with all alleles missing (e.g. ./.), but it includes partially missing genotypes like "./0" or "./1". Total sum of counts should be equal to the count of samples.

genotypeFreq *Map<String, float>*	Genotype frequency for each genotype found calculated from the genotypeCount and samplesCount, in the range [0,1]
maf *float*	Minor allele frequency. Frequency of the less common allele between the reference and the main alternate alleles. This value does not take into acconunt secondary alternates.
mafAllele *String*	Allele with minor frequency.
mgf *float*	Minor genotype frequency. Frequency of the less common genotype seen in this variant. This value takes into account all values from the genotypeFreq map.
mgfGenotype *String*	Genotype with minor frequency.
filterCount *Map<String, int>*	The number of occurrences for each FILTER value in files from samples in this cohort reporting this variant. As each file can contain more than one filter value (usually separated by ';'), the total sum of counts could be greater than to the count of files.
filterFreq *Map<String, float>*	Frequency of each filter calculated from the filterCount and filesCount, in the range [0,1]
qualityCount *int*	The number of files from samples in this cohort reporting this variant with valid QUAL values. This value is used as denominator to obtain the qualityAvg
qualityAvg *float*	The average Quality value for files with valid QUAL values from samples in this cohort reporting this variant. Some files may not have defined the QUAL value, so the sampling could be less than the filesCount.

Index

Pre-computed stats are useful for filtering variants. This stats are intra-study, calculated within a given cohort.

Useful Links

- https://en.wikipedia.org/wiki/Genetic_association
- https://en.wikipedia.org/wiki/Genome-wide_association_study
- https://www.cog-genomics.org/plink/1.9/assoc