Variant Stats

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

Table of Contents:

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

Variant stat file including the following values:

- The total number of alleles (it does not include missing alleles)
- The number of reference alleles found in this variant
- The number of main alternate alleles found in this variant (it does not include secondary alternates)
- The reference allele frequency, i.e., the quotient of the number of reference alleles divided by the total number of alleles.
- The alternate allele frequency, i.e., the quotient of the number of alternate alleles divided by the total number of alleles.
- The number of occurrences for each genotype
- The frequency for each genotype
- The number of missing alleles
- The number of missing genotypes
- The minor allele frequency (maf)
- The minor genotype frequency (mgf)
- The allele with the minor frequency
- The genotype with the minor frequency

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