Basic statistics

OpenCGA provides a set of analysis o compute basic statistics given a variant dataset. In order to get richer statistics, the variant data should comprise annotation and pedigree (samples, phenotypes,...).

OpenCGA computes three types of statistics:

- Variant stats
- Sample stats
- Cohort stats
- Family stats

Next sections describe these statistics.

Table of Contents:

Variant stats

Variant stats are calculated for each variant, in addition, you may specify a set of samples (aka, cohort) in order to take into account only those samples.

Variant stats include 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

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

Sample stats

Sample stats are calculated for each sample. It includes the following information:

- The total number of variants.
- The number of variants per chromosome.
- The number of variants per consequence type.
- The number of variants per biotype.
- The number of variants per type (SNV, INDEL,...)
- The number of variants per genotype.
- The transition-to-transversion ratio (ti/tv ratio).
- A heterozigosity score.
- A missingness score.
- A list of the most affected genes.
- The number of variants per indel length
- A list of HPO and genes for loss of function (LoF) variants.
- A list of the most frequent variant traits.

Summary statistics are stored in a JSON format file.