Sample Stats

Sample Stats contain a basic information for each indexed Sample in a different cohort.

Table of Contents:

Implementation

Sample Stats is implemented using Hadoop MapReduce over HBase.

Input

Parameters

OpenCGA support different input parameters:

- Variant Query
- Sample list, cohort or query

Output

Files

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.

Index

Pre-computed stats are useful for filtering Samples and can be indexed in OpenCGA Catalog using a predefined variable set

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