Variant Storage Engine

Table of Contents:

- Overview
- Operations
- Study oriented
- VCF types

Overview

OpenCGA Variant Storage Engine provides a centralized solution to storage Genomic Variants. It provides a source of data for analysis and visualization in compatible viewers like GenomeMaps. Allowing a fast reading and filtering for variants will speed up analysis, with fastest and more accurate results.

There are an increasing number of biological formats supported by OpenCGA related with a common NGS pipeline. Within this formats, we focus on Genomic Variants due to the complexity and analysis capabilities

Operations

There is an extensive list of operations that can be executed with the Variant Storage Engine. There operations are:

- Sample Genotype Aggregation
- Variant Annotation
- Secondary Index
- Export / Import

Study oriented

The OpenCGA Variant Storage will create an independent database for each project. This database, same way as the projects in Catalog, is divided by studies. This allows to distribute the data into independent studies. Allows queries across multiple studies. Reduces the disk space consumption and the required time to generate the variant annotation by using the same variant annotation across the same database.

VCF types

- **Aggregated VCFs**: Variant files with no sample specific values. Just aggregated data
- **Merged VCFs**: Variant files with a batch of samples with specific samples data.
- **gVCFs**: Single sample files with information for all the positions.