Datasets and Studies

Table of Contents:

- Data organisation
- Datasets
- Variant Anotation

Data organisation

OpenCGA uses a two-level structure to organise datasets, these are *Projects* and *Studies* and are used to organise HGVA data and metadata:

- ***Projects*** is the top-level and can contain one or more *studies*. *Projects* are specific for one species and assembly, all *studies* in a *project* are stored and indexed together in the same database and, therefore, they share the variant annotation.
- ***Study***, in turn, represents a particular dataset which can contain *samples metadata* and *cohorts*, and obviously all the genomic variants. For example, the *1000 Genomes Project* is defined as a *study* in OpenCGA and belongs to *Reference GRCh37* project. You can also define *cohorts* in the *studies*, they are just a set of samples defined within a study. For example, populations and super-populations within The 1000 Genomes Project are defined as *cohorts*, so EUR, AMR or GBR are examples of *cohorts.*

You can get more information about data organisation at OpenCGA Catalog Data Management. *Projects* and *Studies* have a unique ***alias*** to ease their usage from the command-line and REST API, you can find more information about how to query data programmatically at RESTful Web Services and Clients. Please, see next section the full list and organisation of the currently available *Projects* and *Studies* (datasets) in HVGVA.

Datasets

In this sections you can find all datasets loaded in HGVA and how they are organised in *Projects* and *Stu dies* (see previous section).

Project name *(alias)*	**Studies**		**HGVA Version (date)**	
	Name	**Alias**	**v1 (*Dec. 2016*)**	**v2 (*Jan. 2018*)**
Reference GRCh37 *(reference_grch 37)*	1000 Genomes Project GRCh37	*1kG_phase3*	Phase 3 2016-05	Phase 3 2016-05
	Exome Sequencing Project (ESP6500)	*ESP6500*	2016-05	2016-05
	Exome Aggregation Consortium (ExAC)	*EXAC*	0.3.1 2016-05	0.3.1 2016-05
	Genome of the Netherlands (GoNL)	*GONL*	Release 5 2016-05	Release 5 2016-05
	UK10K Project	*UK10k*	2016-05	2016-05
	DiscovEHR	*DISCOVEHR*	-	
	Genome Aggregation Database (gnomAD Exomes)	*GNOMAD_EXOM ES*	-	
	Genome Aggregation Database (gnomAD Genomes)	*GNOMAD_GENO MES*	-	
	Spanish Medical Genome Project (MGP)	*MGP*	2016-12	2016-12
Reference GRCh38 *(reference_grch 38)*	1000 Genomes Project GRCh38	*1kG_phase3*	Phase 3 2016-10	Phase 3 2016-10
	ESP6500	*ESP6500*	-	
	UK10K Project (*****)	*UK10K*	-	
	DiscovEHR (*****)	*DISCOVEHR*	-	
	Genome Aggregation Database (gnomAD Exomes) (*****)	*GNOMAD_EXOM ES*	-	
	Genome Aggregation Database (gnomAD Genomes) (*****)	*GNOMAD_GENO MES*	-	
Cancer GRCh37 *(cancer_grch37)*	QIMR Berghofer Melanoma	*QIMR_Berghofer_ Melanoma*	2016-12	2016-12
	Chronic Myeloid Leukemia - Russian Academy of Medical Sciences	*RAMS_CML*	2016-12	2016-12
Platinum *(platinum)*	Illumina Platinum	*illumina_platinum*	2015-08	2015-08

(*) Liftover carried out by Genomics England (GEL)

Variant Anotation

Variant annotation was carried out by the CellBase project. Please, check CellBase documentation for details on additional data sources: Data sources and species