OpenCGA: a scalable and high-performance platform for big data analysis and visualisation in genomics

Jacobo Coll [1], Pedro Furio [1], Joaquin Tarraga [2], Antonio Altamura [3], Julie Sullivan [3], Daniel Perez-Gil [2], Antonio Rueda [3], M. Bleda [3], David Gomez-Peregrina [4], Jose Miguel Juanes-Tebar [4,5], Pablo Marin-Garcia [4,5], Ignacio Medina-Castello [2]

1. Zetta Genomics, Cambridge, United Kingdom; 2. University of Cambridge, Cambridge, United Kingdom; 3. Genomics England, London, United Kingdom; 4. MGviz.org, Cambridge, UK; 5. Kanteron Systems, Valencia, Spain.

UNIVERSITY OF CAMBRIDGE

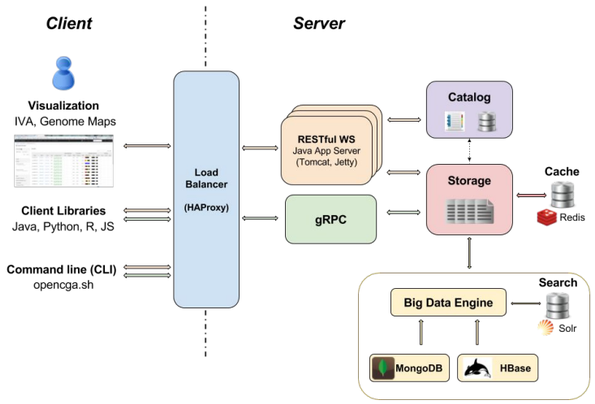
ZETTA GENOMICS
www.zettagenomics.com

kanteron systems

ABSTRACT

Currently, typical large-scale clinical genomics studies consist of ten of thousands of whole genome sequences. Managing these projects requires a platform with the ability to analyze billions of unique variants over hundreds of terabytes of data. **OpenCB** is an open-source project that implements a high-performance, scalable and secure platform for genomic data analysis (**OpenCGA**) and interactive visualisation (**IVA**).

ARCHITECTURE



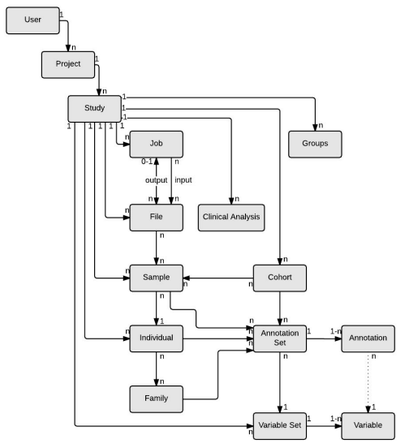
OpenCGA relies on current **big data** technologies such as Hadoop, Spark, MongoDB and Solr to implement an **advanced analytical variant storage engine** that can index and aggregate thousands of whole genomes a day and allows performing real-time queries, genotype aggregations, quality control and genomic analysis such as GWAS and knockout analysis.

The API REST allows access through web clients and CLI. This system is certified for Azure.

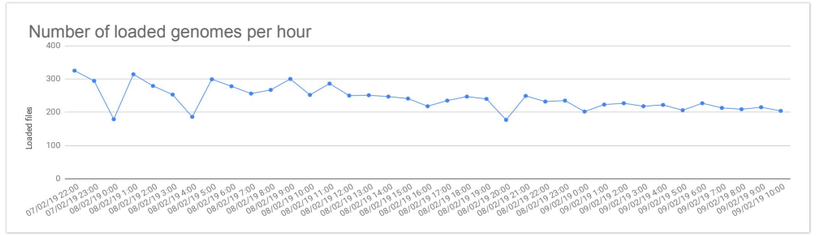
DATA MODEL

A genomic data analysis platform need to keep track of different resources such as metadata of files, sample annotations or jobs. **OpenCGA-Catalog** aims to collect and integrate all the information needed for executing genomic analysis. This information is organized in nine main entities:

- **User**: data related to the user account.
- **Project**: information of a project, covering as many related studies as necessary.
- **Study**: main space set environment. Contain files, samples, individuals, jobs…
- **File**: information of a submitted or generated file.
- **Sample**: information of the sample. Closely related to file entity.
- **Individual**: information of the patient.
- **Cohort**: sets of samples with some common features.
- **Disease panel**: set of variants, genes and/or regions of interest.
- **Job**: job analysis launched using any of the files or samples.



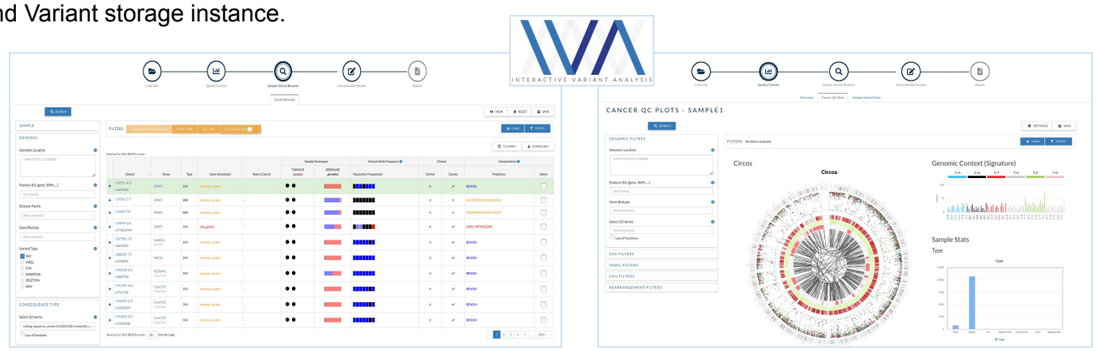
PERFORMANCE



In a Hadoop cluster consisting of 30 nodes and a LSF queue, the loading performance was about 240 genomes per hour or **6,000 samples a day** (Concurrent files loaded: 60, Load time per file: 15 min). **CellBase** was then used to annotate each variant in the database, this annotation include consequence types, population frequencies, conservation scores clinical info... Variant annotation of the 585 million unique variants took about 3 days, **about 200 million variants were annotated per day**.

INTERACTIVE VARIANT ANALYSIS

The Interactive Variant Analysis (**IVA**) tool is an **OpenCB** project implemented to facilitate the filtering, analysis and interpretation of whole genome variant data. IVA is an interactive visualization client for OpenCGA, making it easy to work with **clinical interpretation** and variant information stored in a Catalog and Variant storage instance.



CONCLUSION

OpenCGA has proven to scale and perform very well to nearly 100,000 whole genomes accounting for 584 million aggregated variants or 40TB of data. **OpenCGA** is used as a data platform at GEL and other big genomics institutions, and it is available at Microsoft Azure. In addition, it implements a rich RESTful web service API, a command line interface and multiple client libraries. An Interactive Variant Analysis (**IVA**) browser is provided to analyse and visualise biological information from various data sources. **OpenCGA is open-source and part of OpenCB suite**. http://docs.opencb.org/