Welcome to OpenCGA

OpenCGA is an open-source project that implements a high-performance, scalable and secure platform for Genomic data analysis and visualisation.

OpenCGA provides the most advanced and complete genomic data platform. The performance, scalability and huge number features makes of OpenCGA an unique full-stack solution today. OpenCGA takes care of security and implements a high-performance query engine and analysis frameworks for *Big Data* analysis and visualisation in current genomics. OpenCGA uses the most modern and advanced technologies, and has been designed and implemented to scale to hundreds of thousands if genomes accounting for petabytes of variant data. It is built on top of three main components: *Catalog Metadata Database*, *Variant Storage Engine* and *Analysis Framework*.

Latest news:

OpenCGA v1.2.0 Released

Nacho Medina posted on Jul 21, 2017

We are pleased to announce new version 1.2.0!

Main Features

- **Authenticated** and **secure** platform to query and visualise data. An advanced **permission** system has been implemented to ensure data privacy.
- A **metadata database** to keep track of registered users, projects, studies, files, samples, families, jobs, ...
- **Advanced Clinical Data** database implemented, users can define their data models for samples, patients or families.
- **Alignment storage** allows to index BAM/CRAM, calculate index and query data and coverage
- The most advanced, high-performance and scalable **Variant Storage Engine** solution today. Variant Storage Engine can normalise, load, index, aggregate, annotate and precompute variant stats for hundreds of thousands of whole genomes.
- **Analysis Framework** implemented on top of variant and alignment storage engines. OpenCGA comes with many analysis already implemented such as GWAS. Users can easily extend OpenCGA functionality by implementing a plugin or connecting to a external binary.
- Real **Big Data Analytics** supported, you can use different computing frameworks such as MapReduce or Spark on top HBase or Parquet files.
- Full **Clinical Analysis Solution** implemented, you can create the cases and run different clinical interpretations algorithms from your scripts or from a web application.
- Rich and comprehensive **RESTful Web Services API** with more than 160 endpoints to manage, query and analyse metadata, variants, alignments and clinical data.
- Easy **programmatic access** and **pipeline integration** thanks to the four different **client libraries** developed in **Java, Python, R** and **Javascript**
- Interactive **web-based application** to query, analyse and visualise variants, alignments and clinical data
- **Zetta Genomics** start-up is being launched in 2019 to offer official support and customisation. **OpenCB Enterprise** will be launched in 2020 with many new features!

Metadata and Security

Metadata Database

OpenCGA Catalog implements a high-performance metadata database to track all files metadata, samples, families, ...

Security

OpenCGA implements **authentication** to control what data can be seen by users. Data such as Files, Samples, Families, .. can be shared in different way.

Variant and Alignment Storage

Variant Database

OpenCGA implements a **high-performance and scalable** variant NoSQL database to store and index thousands of whole genome VCF files. Performance observed show more than 2,000 whole genomes indexed a day.

Many variant operations have been implemented such as variant aggregation, stats calculation, variant annotation, export, ...

We have implemented the most advanced query engine and aggregation framework to query variants.

Alignment Storage

Indexing BAM files and calculating coverage is supported. You can efficiently query all these data through REST web services.

Easy to Use

REST API and Clients

We have implemented a comprehensive REST API to work with Catalog and query Variants and Alignment data in a secure way. To facilitate using REST we have developed four client libraries developed in Java, Python, R and Javascript.

Command Line Interface

OpenCGA implements two different command lines, one for the users and one for the admin. Users can fully operate OpenCGA from the command line.

Analysis Framework

Clinical Analysis

Big Data Analysis

Native Analysis and Plugins

OpenCGA implkements most common analysis such as stats or GWAS among many other ones. We will keep adding more common analysis in each version.

Users can implement their own native analysis for OpenCGA by developing a ***plugin***. These plugins can easily be installed and executed in OpenCGA.

Wrapped Analysis

OpenCGA can also execute any other external binary (*C++, Python R, ...*) by creating a simple wrapper that connect OpenCGA storage engine with the binary. We also provide some official external binaries supported such as *Plink*

Clinical Data and Disease Panels

You can store all you clinical data in our free data model solution in Catalog. You can define your clinical variables and annotate files, samples, individuals, families or cohort. Clinical Data is indexed automatically to provide a real-time queries and aggregations analysis.

Disease Panels are fully supported and versioned.

Clinical Interpretation Analysis

You can define different types of Clinical Analysis. We have implemented some automatic clinical interpretation algorithms for Rare Diseases (families) and Cancer. A Decision Support System has also been implemented in IVA.

Rich Data Models

OpenCGA takes advantage of the rich data models developed in OpenCB. We make an extensive use of **Variant** and **Variant Annotation** data models.

Spark Analysis

OpenCGA implements several analysis top of the Variant storage. These analysis can use different programming models – such as MapReduce – or different technologies such as Spark.

A Spark-based library has developed to provide extra analysis capabilities.

Cloud

Cloud Architecture

OpenCGA architecture was designed to be fully compatible with modern cloud architectures, this makes of OpenCGA extremely efficient and performance in cloud environments.

Microsoft Azure

OpenCGA and Microsoft collaborated to test and validate HDInsight security and analysis performance.

Visualisation

Source Code

Web based on IVA project at https://github.com/opencb/iva/tree/app/hgva

Server based on OpenCGA at https://github.com/opencb/opencga

Contributing

IVA is a collaborative project that aims to integrate as many reference human studies as possible, you can contact us for feature request. If you want to contribute to the code you are more than welcome to contribute to IVA and OpenCGA

Zetta Genomics

Start-up

A University of Cambridge start-up is being launched during 2019. Zetta will provide official support and a number of different services.

This is will be officially announced in later 2019, if you want to know more about this please contact im411@cam.ac.uk

Development

Contributors

Ignacio Medina (HPCS, University of Cambridge)

Source Code

Web based on IVA project at https://github.com/opencb/iva/tree/app/hgva

Server based on OpenCGA at https://github.com/opencb/opencga

Contributing

IVA is a collaborative project that aims to integrate as many reference human studies as possible, you can contact us for feature request. If you want to contribute to the code you are more than welcome to contribute to IVA and OpenCGA

Recent space activity

Pedro Furio
AnnotationSets 1.4.0 updated Nov 10, 2020 view change

Joaquín Tárraga Giménez
Clinical Interpretation Analysis updated Aug 27, 2020 view change

Pedro Furio
Command Line updated Jun 10, 2020 view change

Joaquín Tárraga Giménez

Release Notes updated Jun 04, 2020 view change

Will Spooner
Roadmap updated Jun 04, 2020 view change