Roadmap

In this section you can find only the **main top-level features** planned for major releases. For a more detailed list you can go to GitHub Issues at https://github.com/opencb/opencga/issues.

From OpenCGA version 2.0.0 we follow **time-based releases**, two minor releases a year will be scheduled in April and October.

Table of Contents:

- OpenCGA 2.x Releases
  - 2.1.0 (Nov 2020)
    - General
    - Catalog
    - Storage Engines
      - Variant
    - FIHR
  - 2.0.0 (June 2020)
    - General
    - Catalog
    - Storage Engines
      - Alignment
      - Variant
    - Analysis
      - Framework
      - Variant
      - Clinical Interpretation
    - Clinical
    - Cloud
- OpenCGA 1.x Releases
  - 1.4.0 (March 2019)
    - General
    - Catalog
    - Variant Storage
  - 1.3.0 (November 2017)
    - General
    - Catalog
    - Variant Storage
    - Alignment Storage
- Unscheduled features

OpenCGA 2.x Releases

2.1.0 (Nov 2020)

You can track GitHub issues at GitHub Issues 2.1.0. You can follow the development at GitHub Projects.

General

- Main feature in this release is **Federation**
- Implement a **Centralised Log** analytic solution, we are planning to use Kibana

Catalog

- Implement a new **Notification** system, Catalog will notify to a message queue *(RabbitMQ, Apache Kafka),* this will allow other applications to know what's going on
- Improve **RESTful** web services by adding standardise **error codes** to the response, this will improve debugging

Storage Engines

Variant

- Implement a new **Cache** functionality, some sample and family-based variant queries and anlaysis can take up to few seconds, since this data is read-only this could be easily cached

FIHR

- Initial support of **FIHR**, in this release we will extend Catalog data models and we will implement FIHR import/export functionality
- Implement **FIHR Genomics** API, this will allow FIHR applications to query genomic variants in OpenCGA

2.0.0 (June 2020)

You can track GitHub issues at GitHub Issues 2.0.0. You can follow the development at GitHub Projects.

General

- Improve **Docker** images, now stable versions with the different variant storage are pushed to Docker Hub
- Upgrade **dependencies**: MongoDB 4.2, Solr 8.1.1, JUnit 5.5.1, ...
- **Clean ups** and **remove** deprecated code and APIs

Catalog

- Add **ACID Transactions** to all database operations
- Improve **Audit**, extend audit data model and ensure all actions are now audited. Also, make audit *queryable*.
- Implement a new **Task** system, this will be used internally by OpenCGA to schedule some jobs, this new functionality can be also used by external applications
- Improve **RESTful** web services response and **warning/error** notifications
- Prepare OpenCGA for supporting **Federation** in next releases
- Improve **performance** and **test coverage**

Storage Engines

Alignment

- Support CRAM file

Variant

- Implement **structural variant imprecise** queries
- Implement new **Variant Score** to store results from analysis such as GWAS, this can be used when filtering
- Remove any **blocking variant operation**, any variant operation should be able to run at any time in a consistent way
- Improve **HBase sample index**, this will improve the **performance** of some **queries and analysis**
- Implement HBase-based **aggregations**
- Support new **HBase 2.0** version
- Improve **testing** and **benchmark** module

Analysis

Framework

- Develop an **Analysis Framework**, this will allow users to extend and customise OpenCGA with their own analysis
- Implement a **WrappedAnalysis** functionality in this framework to make easy to use any external tool such as Plink (see below in *Varlant Analysis* section)

Variant

- Implement on-demand **Variant Stats** and **Variant Sample Stats**
- Add GWAS **variant analysis**, this can optionally be stored and indexed in the new **Variant Score** object
- Add *Plink* as **wrapped analysis**

Clinical Interpretation

- Implement **Cancer Tiering** interpretation analysis algorithm
- Network-based clinical interpretation algorithm *(experimental)*
- Implement **Secondary Findings** analysis

Clinical

- Network-based clinical interpretation algorithm *(experimental)*

Cloud

- Full support for **Microsoft Azure and HDInsight 4.0,** this also includes **Azure AD, Azure Blob** and **Azure Batch**. We would like to **thank very much Microsoft Azure** for their amazing support and help here.
- Add **Kubernetes** for deployment and orchestration

Note: some of these features might be released in the Enterprise version coming soon

OpenCGA 1.x Releases

1.4.0 (March 2019)

General

- Implement the new **HTSGET 1.0** protocol
- **IVA 0.9.0** will implement a full study and clinical analysis among many other features
- Add many more negative and variant **functional tests**
- **Documentation** improvements with new diagrams and tutorials

Catalog

- Complete and test all **delete** operations and implement *delete by queries* to make easier to delete batches of resources, with this the **REST API** can be considered complete
- Implement a new **admin** REST API, this will allow OpenCGA administrator to execute administrative tasks remotely
- New **PermissionRule** feature, you can define rules for assigning permissions automatically when new data is created, e.g. *set VIEW permission to USER to all samples where HOSPITAL = 'X'*

- New implementation of how **clinical data** (*annotation sets*) are store in the database, this new physical schema significantly improves querying annotations (even with nested objects or arrays), *group by* aggregations, *include/exclude* filtering and allow to *flatten* the annotations
- Complete ***ClinicalAnalysis and ClinicalInterpretation*** data models and functionality
- Add **DiseasePanel** entity to manage panels

Variant Storage

- Final **HBase variant storage** implementation. New architecture should scale to few million of genomes and billion of variants.
- Support the last pending structural variant: **Translocation**. With this all structural variants are properly represented and stored
- Improve **variant stats** and add **simple variant analysis** such as association or Hardy-Weinberg test, this will be stored and indexed in the new ***VariantScore*** object
- Add INDEL **left-alignment** normalisation to *VariantNormaliser*
- **Variant Benchmark suite** to study scalability and performance
- Add a native implementation of Genomics England Tiering analysis

1.3.0 (November 2017)

General

- CLI **autocompletion** implemented
- New single CLI for execute **migrations** automatically
- New and fully functional **R client library** for REST web services, with this the four client libraries are completed
- New **IVA 0.9.0** is developed coordinately to exploit all the new features, they will be released together
- Many more **functional tests** added to test all new functionality described below
- Review and improve **Swagger** documentation and descriptions
- **Documentation** improvements with new diagrams and tutorials

Catalog

- New ***Family*** data model finished, now it is production ready, this completes and integrates three related data models: *Sample, Individual* and *Family*
- New ***Versioning*** feature implemented for *Sample, Individual* and *Family*. Now you can track any change in those data models, users can query o review any *version* of those documents
- New ***Export*** functionality implemented, this allows to export a *Project* as it was at any specific release, this can then imported in a new OpenCGA server
- New Study administrative group called ***admins***, all users in this group will be granted some special permissions at Study level such as *create groups* or *share* data, this will make Study administration much easier
- New ***Confidential*** permission for Variable Sets, now you can make some clinical data private for some users
- New ***ClinicalAnalysis*** data model added, this allows to define and stored different clinical interpretation analysis, this is still experimental and it should not be used in production
- Improvements in ***Group By*** queries, now you can pass a ***count*** parameter and aggregations only use data you can view, this can be useful for summarising data. Also, this has been added to *Individual* and *Family*
- Ensure that all query **GET** REST web services accept **comma-separated list of IDs**, at the moment only few of them accept ID lists, this will reduce the number of REST calls needed improving the performance
- New REST web service to **execute remote scripts** for Catalog, for instance "*move samples from Study*"
- **Performance improvements** when checking permissions (ACL) in *create* and *update* methods, now on average 50% less database queries are needed

Variant Storage

- Improve support for **Structural Variants**, in this release we will fully support *Insertion, Deletion* and *Copy Number* variants
- New ***VariantMetadata*** implemented, this is *exported* together with the variant data to be further analysed with other OpenCB projects using Spark
- New ***VariantScore*** object added to Variant data model, this will allow to store variant scores from cohort-related analysis such as association or Hardy-Weinberg tests in the next release
- Implement some **HBase** physical schema improvements and a better integration with Solr
- Support ***Amazon EMR*** Hadoop cluster
- **Performance improvements** when querying variants from samples, this will have a big impact in clinical interpretation analysis

Alignment Storage

- Major improvements in **BAM query** engine. New **server-side** filters added, this is a more efficient implementation since the data sent through the network is reduced. The available filters now are: *region, minMapQ, maxNumberMismatches, maxNumberHits, properlyPaired, maxInsertSize, unmmapped* and *duplicated.*
- New **coverage** calculator using **BigWig**. Now coverage is calculated and stored in BigWig format, the *windowSize* is configurable. Also, coverage can now be queried for a *region* and optionally a *windowSize,* the server will **aggregate and compute the average** in *windowSizes.*
- New **REST** and **gRPC** APIs implementing the new query filters and coverage functionality. When using **REST** a JSON string is returned using GA4GH data model. When **gRPC** is used a binary stream is obtained. Note that in both protocols the filters are applied in the server.

Unscheduled features

The following features have been accepted but no release version has been assigned:

- Add test for the CLI
- Support Slurm
- Add **Reactive Programming** (RxJava) and **Events**, this will allow to be easily integrated into other custom Java-based applications
- New **Gene Expression** database, this will include a Gene Annotation based on CellBase

You can find detailed information for some of them at https://github.com/opencb/opencga/milestone/10