From reads to disease variants

Introduction to openCGA and IVA for variant prioritization

Javier Pérez Florido

javier.perez.florido.sspa@juntadeandalucia.es

http://www.clinbioinfosspa.es
Clinical Bioinformatics Area
Fundación Progreso y Salud / Servicio Andaluz de Salud
Seville

Clinical Bioinformatics Area

JUNTA DE ANDALUCIA
Fundación Progreso y Salud
CONSEJERÍA DE SALUD Y FAMILIAS

OpenCB

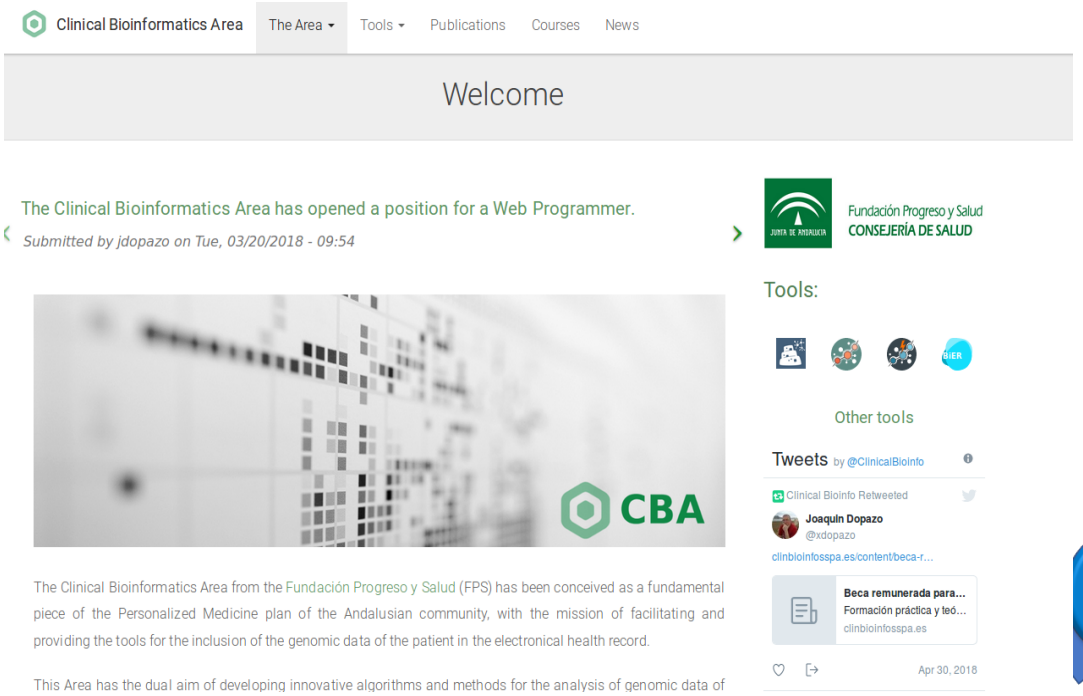
Overview

- Introduction
- *Big Data* in Genomics
- OpenCB: Open source initiative for Computational Biology
- A case study: Personalized Medicine Module (MMP)

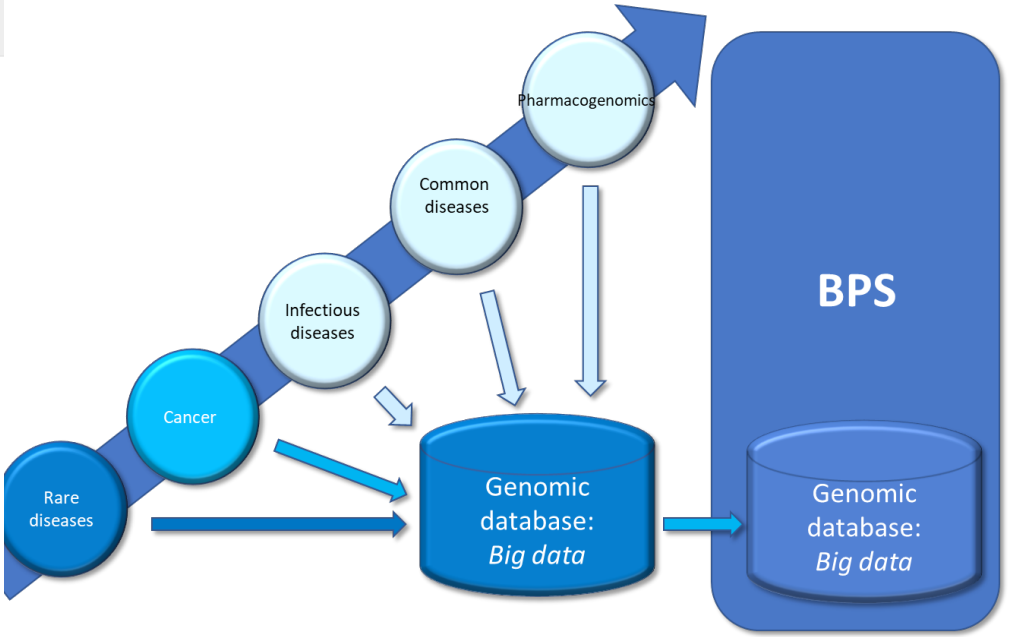
Overview

- **Introduction**
- *Big Data* in Genomics
- OpenCB: Open source initiative for Computational Biology
- A case study: Personalized Medicine Module (MMP)

The Clinical Bioinformatics Area



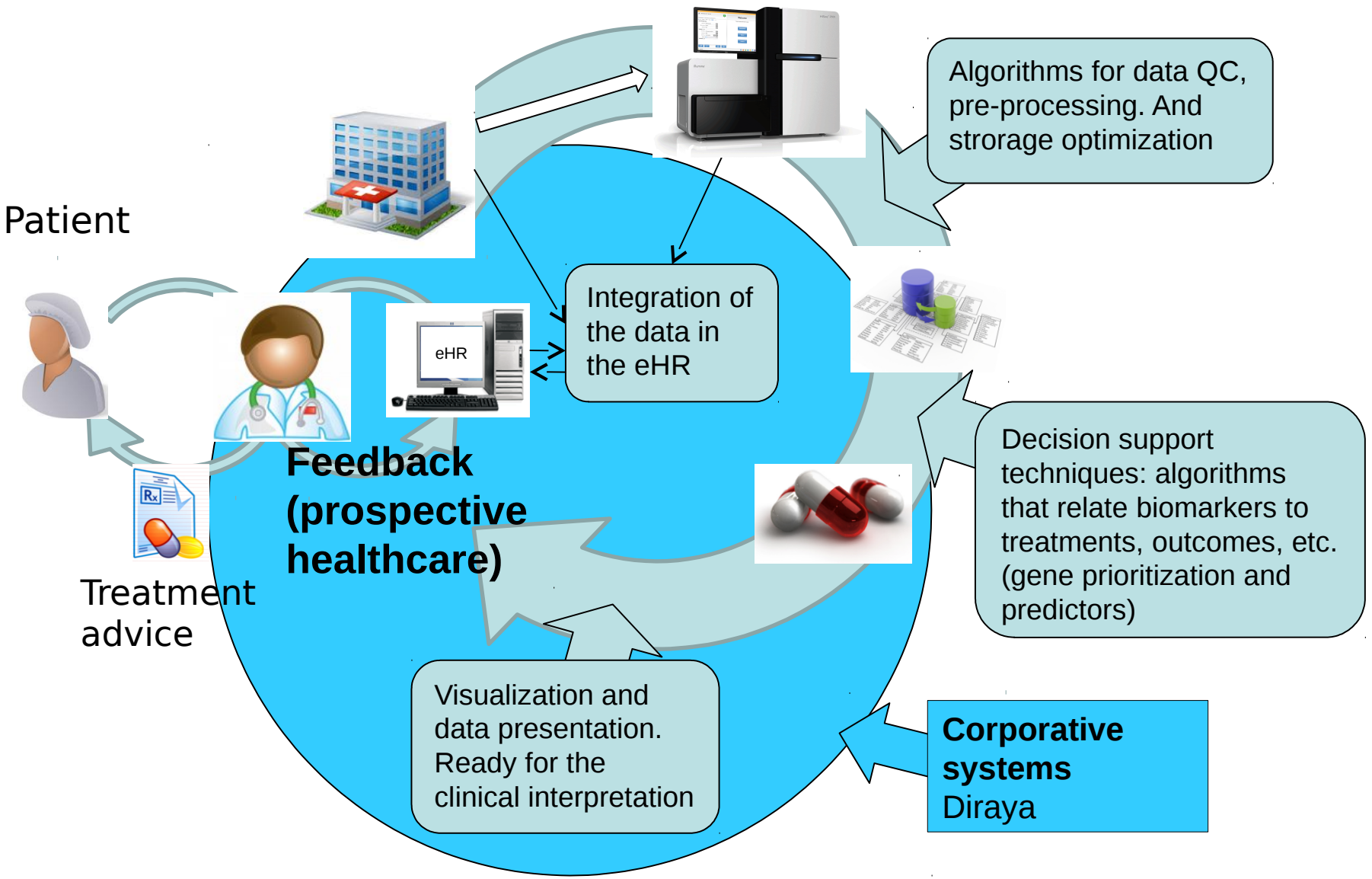
http://www.clinbioinfosspa.es/



The **Bioinformatics Area**, created in June 2016 in the Fundación Progreso y Salud, has as main goal supporting the Program of Personalized Medicine of the Andalusian Community by facilitating the use of genomic data for precision diagnostic and treatment recommendation, implementing a prospective health care functionality in the public health system.

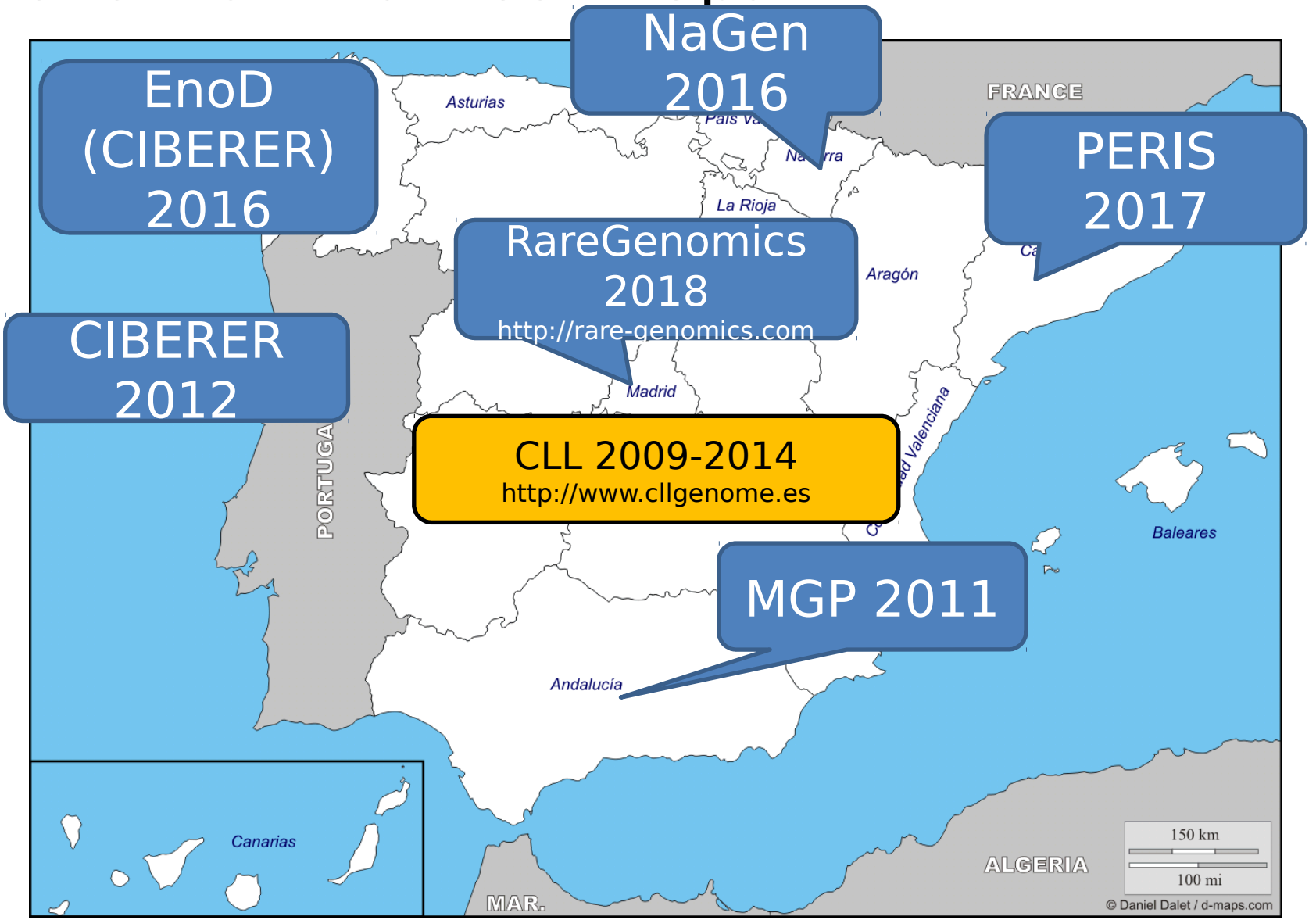
Introduction

Bioinformatics for personalized medicine within the Andalusian health system



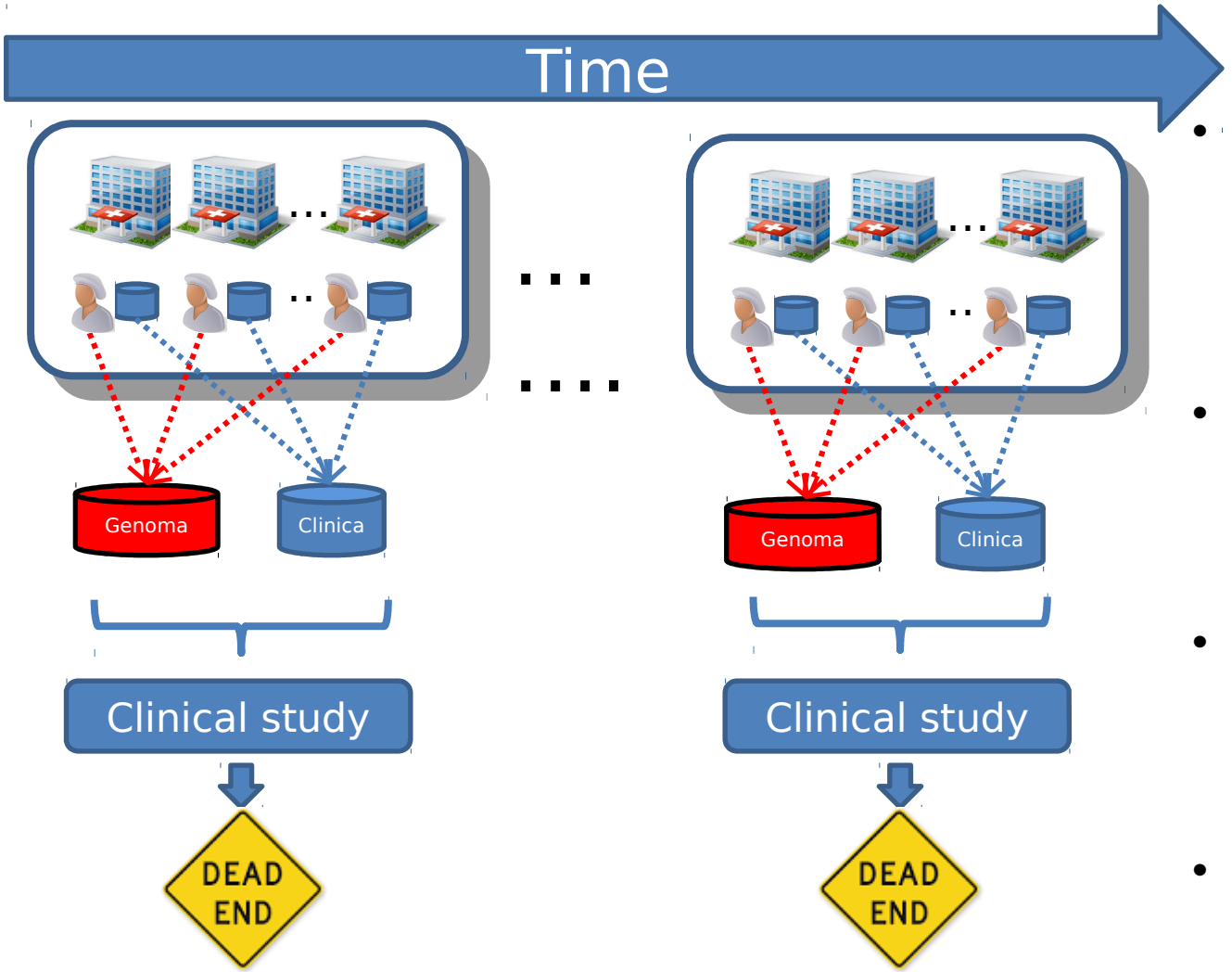
Introduction

Genomic initiatives in Spain



Introduction

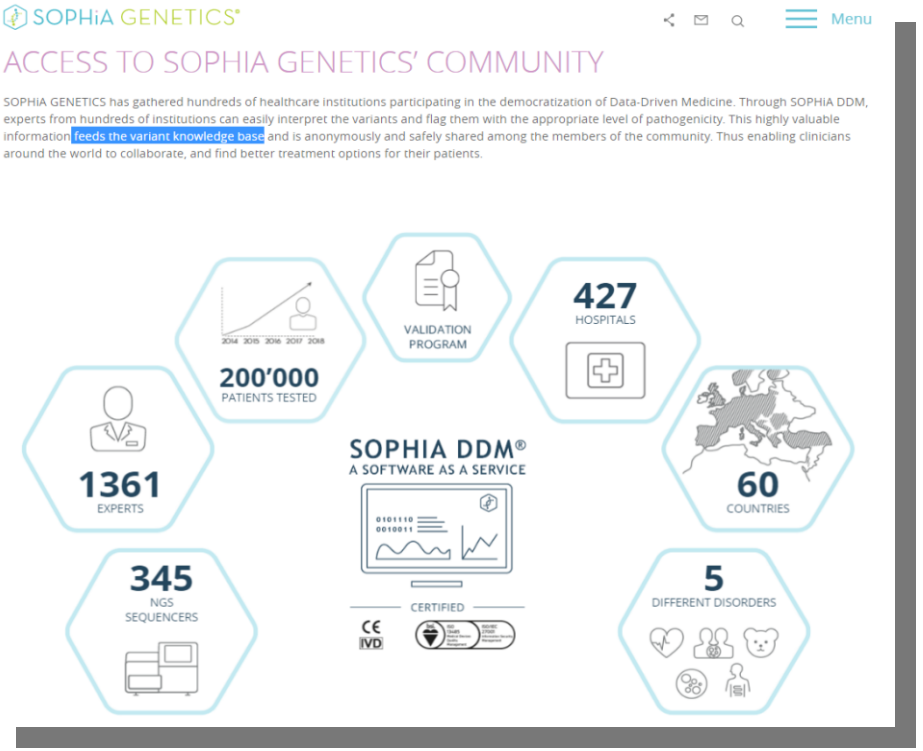
Genomic initiatives without universal eHR are just clinical studies but not Personalized Medicine

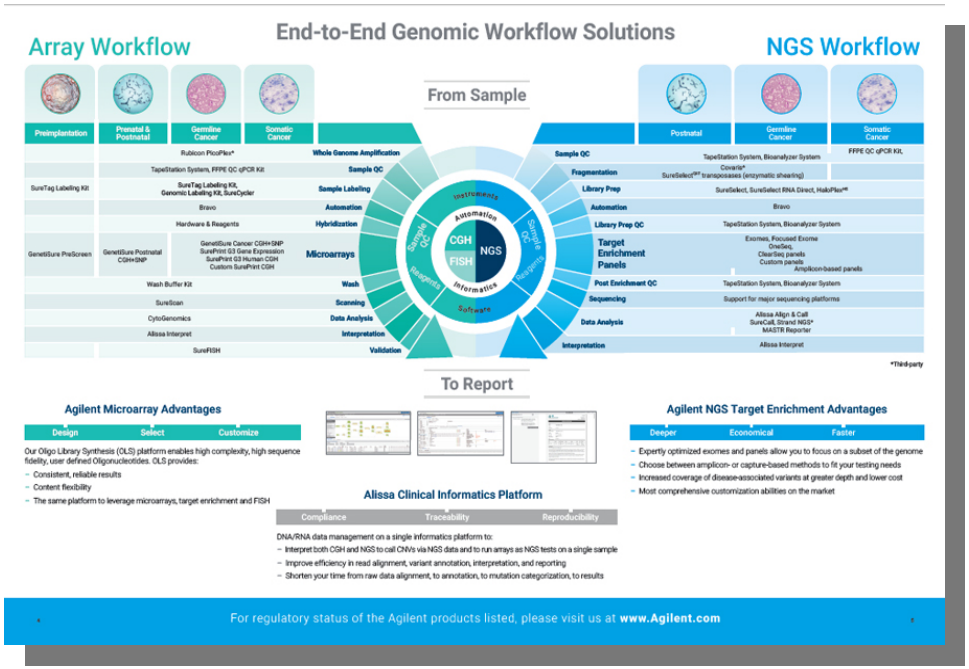


- Each study requires of a specific genomic and clinical data collection into an external database
- Static clinical data (e.g. if a control becomes a case the external DB will not be updated)
- Limited genomic data reuse for purposes different from the original study
- Model of GEL (100,000 genomes) and other initiatives.

Introduction

The value of genomic data for the health system

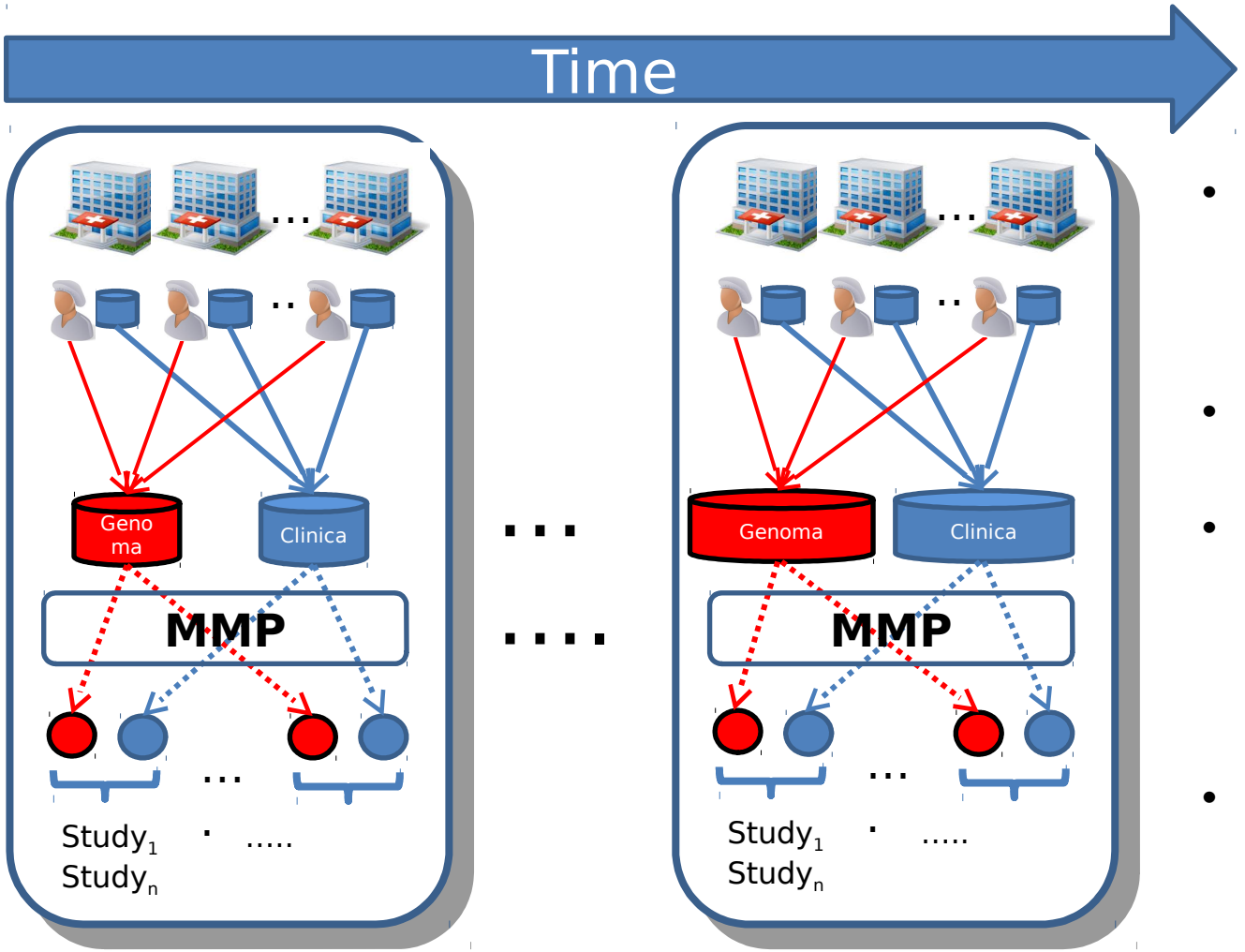




Currently, because of the lack of adequate resources, genomic data analysis is externalized to companies or private software. We lose the control on the data externalized and actually work for external companies for free

Introduction

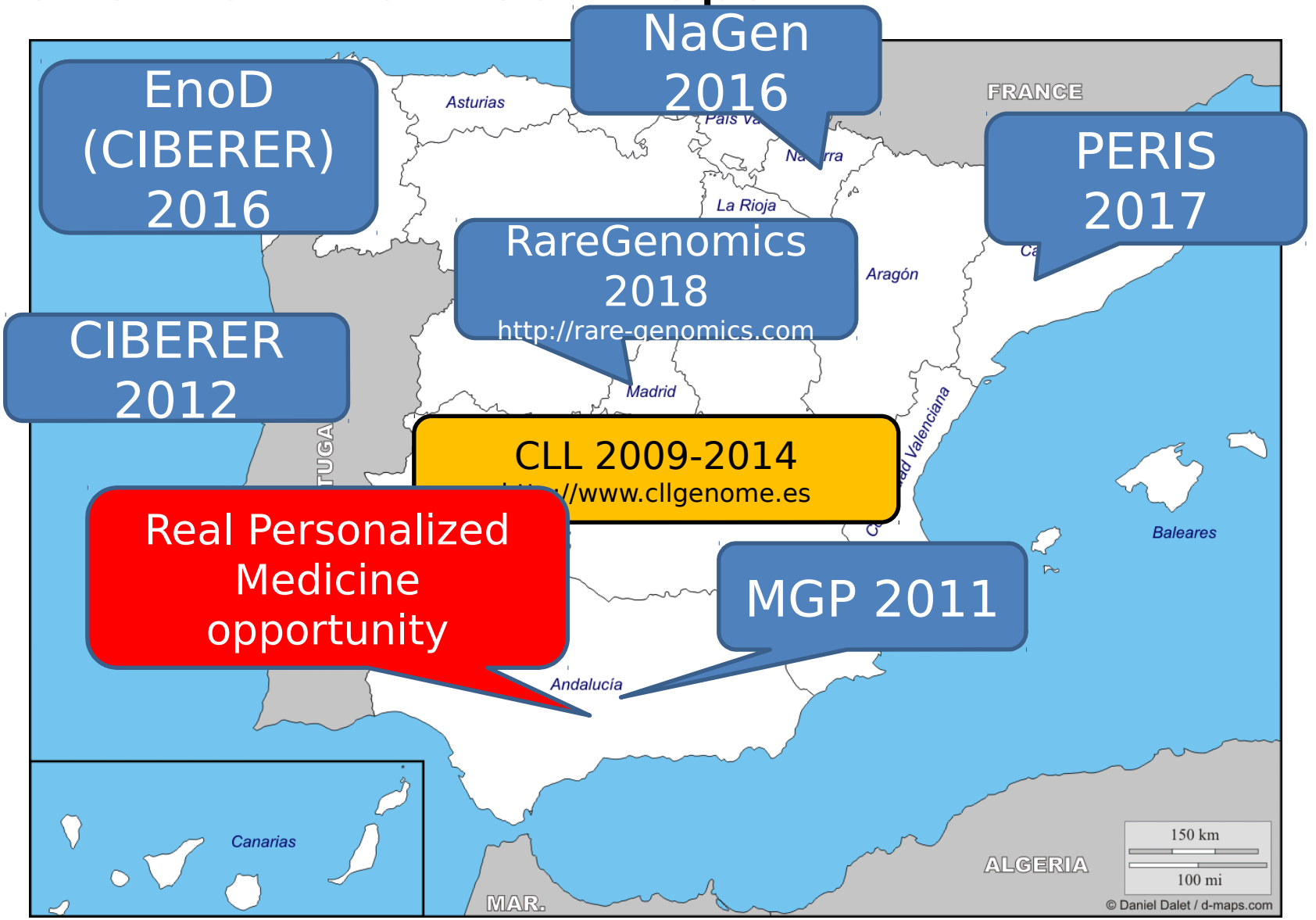
Personalized medicine requires a model that integrates genomic data and universal eHR



- The whole health system becomes a enormous potential <u>prospective study</u>
- Clinical data <u>dynamically</u> associated to patients
- Possibility of many clinical studies by <u>reanalyzing genomic data</u> under diverse perspectives (with no extra investment)
- Growing genomic DB with increasing study possibilities

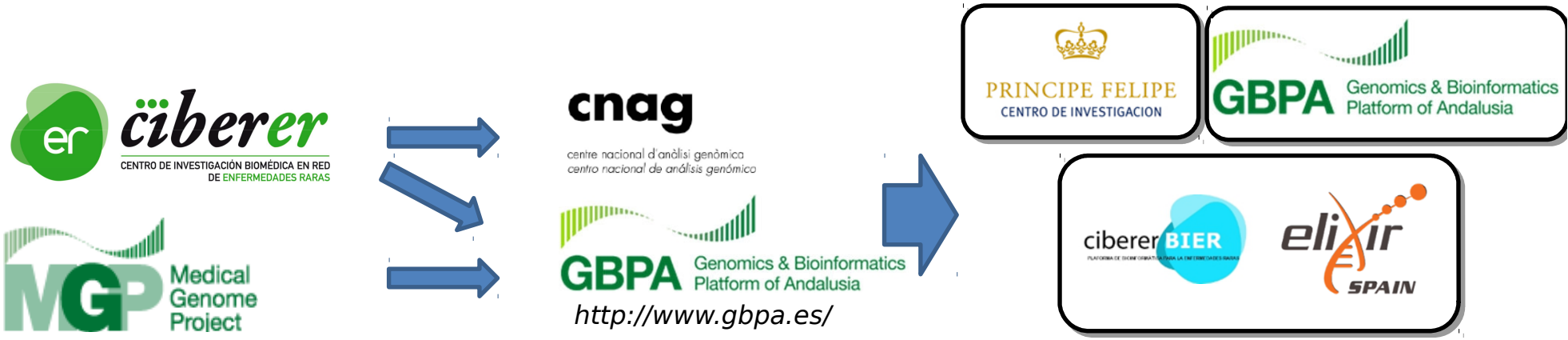
Introduction

Genomic initiatives in Spain



Introduction

Previous initiatives: MGP and CIBERER initiatives to sequence rare disease patient exomes



Sample providers *Sequencing platforms* *Data analysis*

Diseases with

- Unknown causal genes
- No mutations in known genes

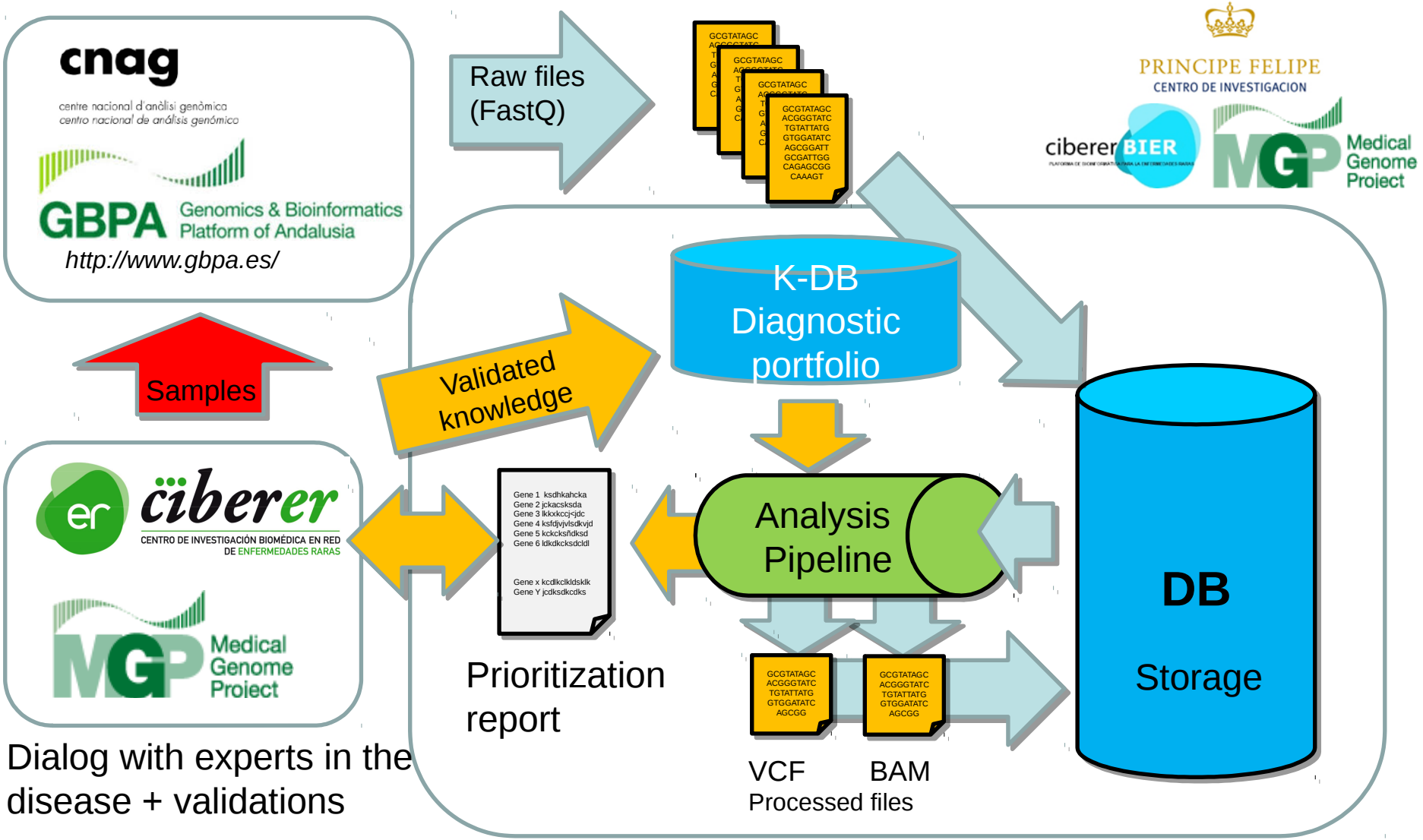
Search for:

- New disease genes
- Susceptibility genes

A total of 1044 exomes of 300 healthy controls and patients of more than 30 diseases were sequenced between 2012 and 2013.

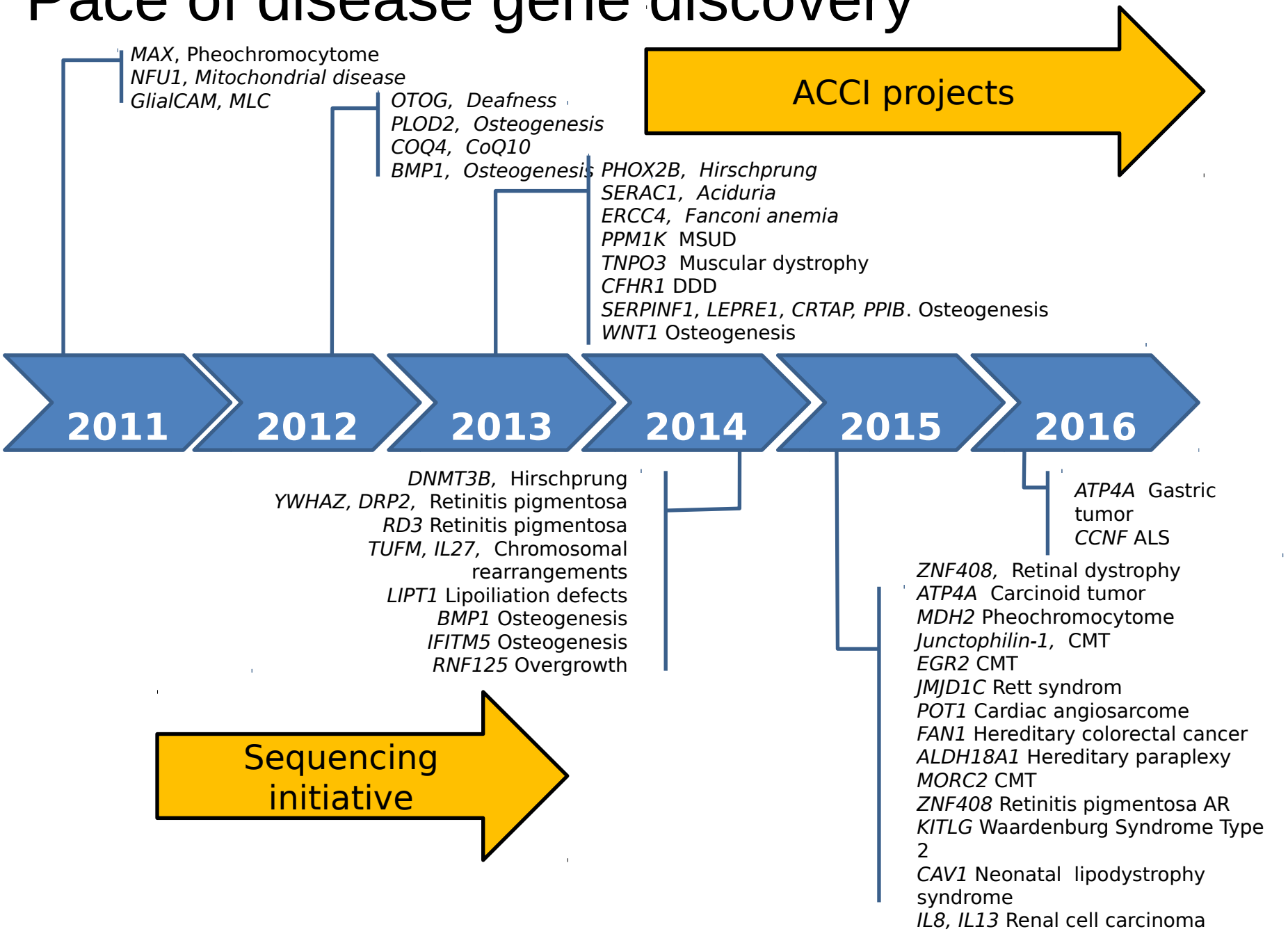
Introduction

Data analysis and the cycle of knowledge generation



Introduction

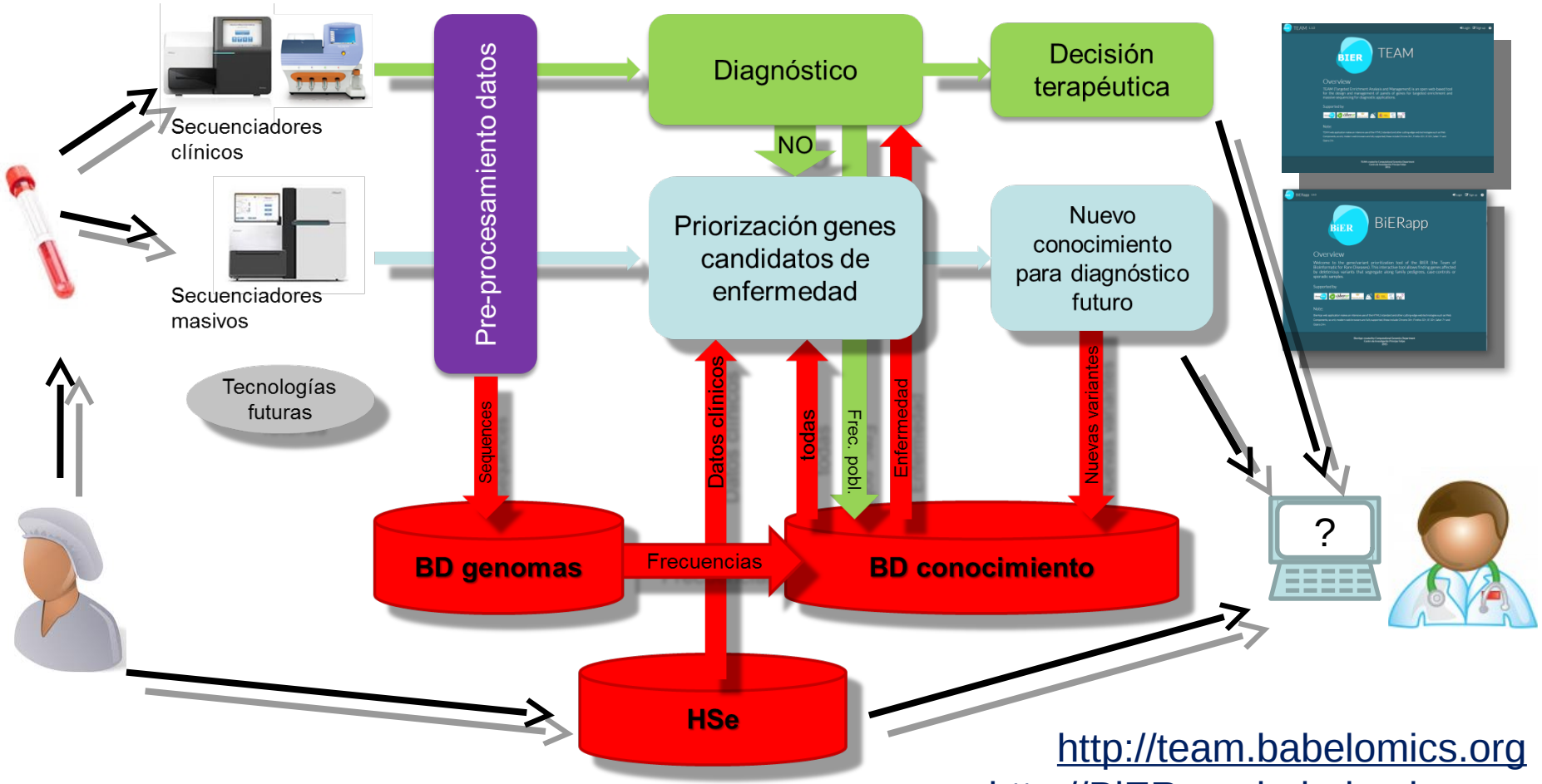
Pace of disease gene discovery



Introduction

CIBERER initiative for diagnosis and biomarker discovery using massive sequencing

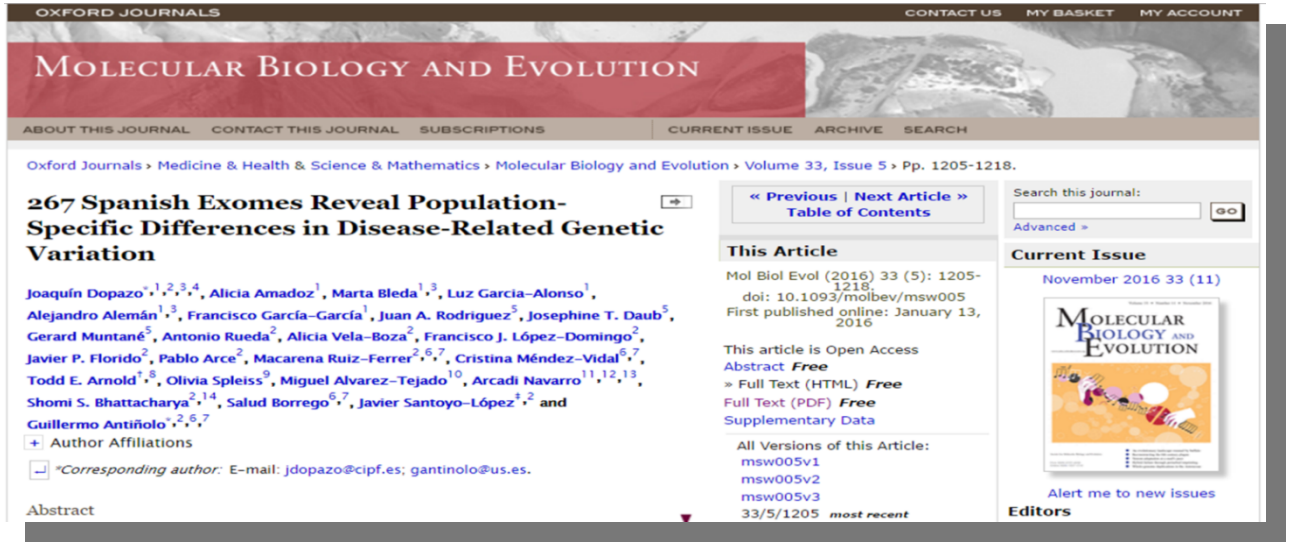
Ongoing (second phase) Project with the collaboration of seven hospitals: La Paz, FJD, Ramón y Cajal, CBM (Madrid), Virgen del Rocío (Sevilla), Hospital del Mar (Barcelona), HU La Fe (Valencia)., **within the context of CIBERER**



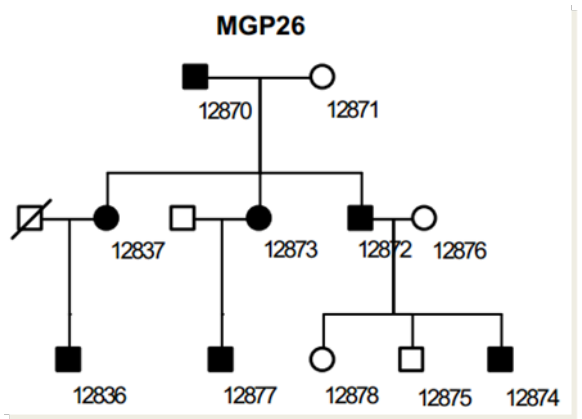
http://team.babelomics.org
http://BiERapp.babelomics.org

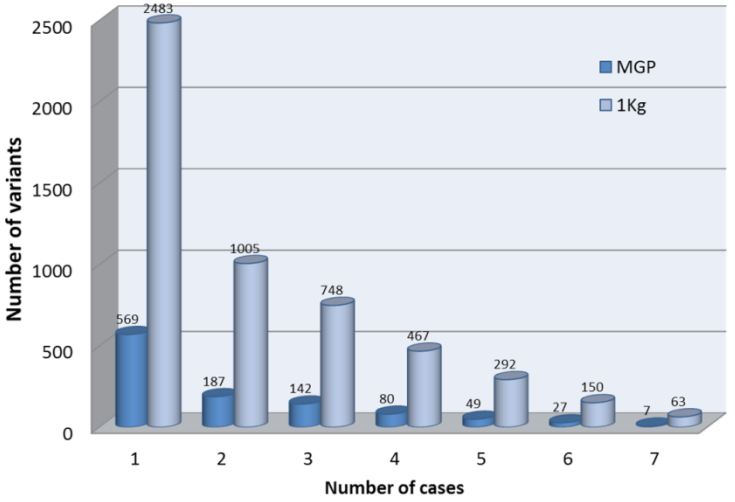
Introduction

Lessons learned: the importance of local variability in the prioritization process



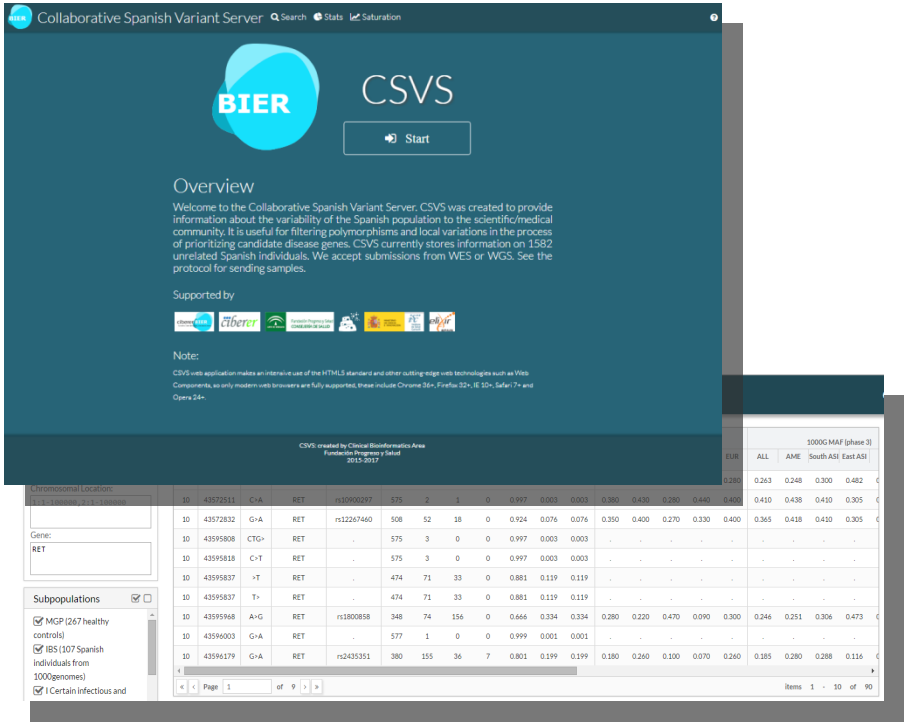
We discovered some 12,000 “spanish” polymorphisms not present in other databases. The filtering efficiency enormously increases using local population data





Introduction

The CSVS is a crowdsourcing project



http://csvs.babelomics.org/

Allelic population frequencies obtained from about 2,000 samples (exomes and genomes) available in CSVS

Scenario: Sequencing projects of healthy population are expensive and funding bodies are reluctant to fund them

CSVS Aim: To offer increasingly accurate information on variant frequencies characteristic of Spanish population.

CSVS Main use: Frequency-based filtering of candidate variants

Main data source: Sequencing projects of individual researchers (CIBERER and others)

Problem: Most of the contributions correspond to patient exomes

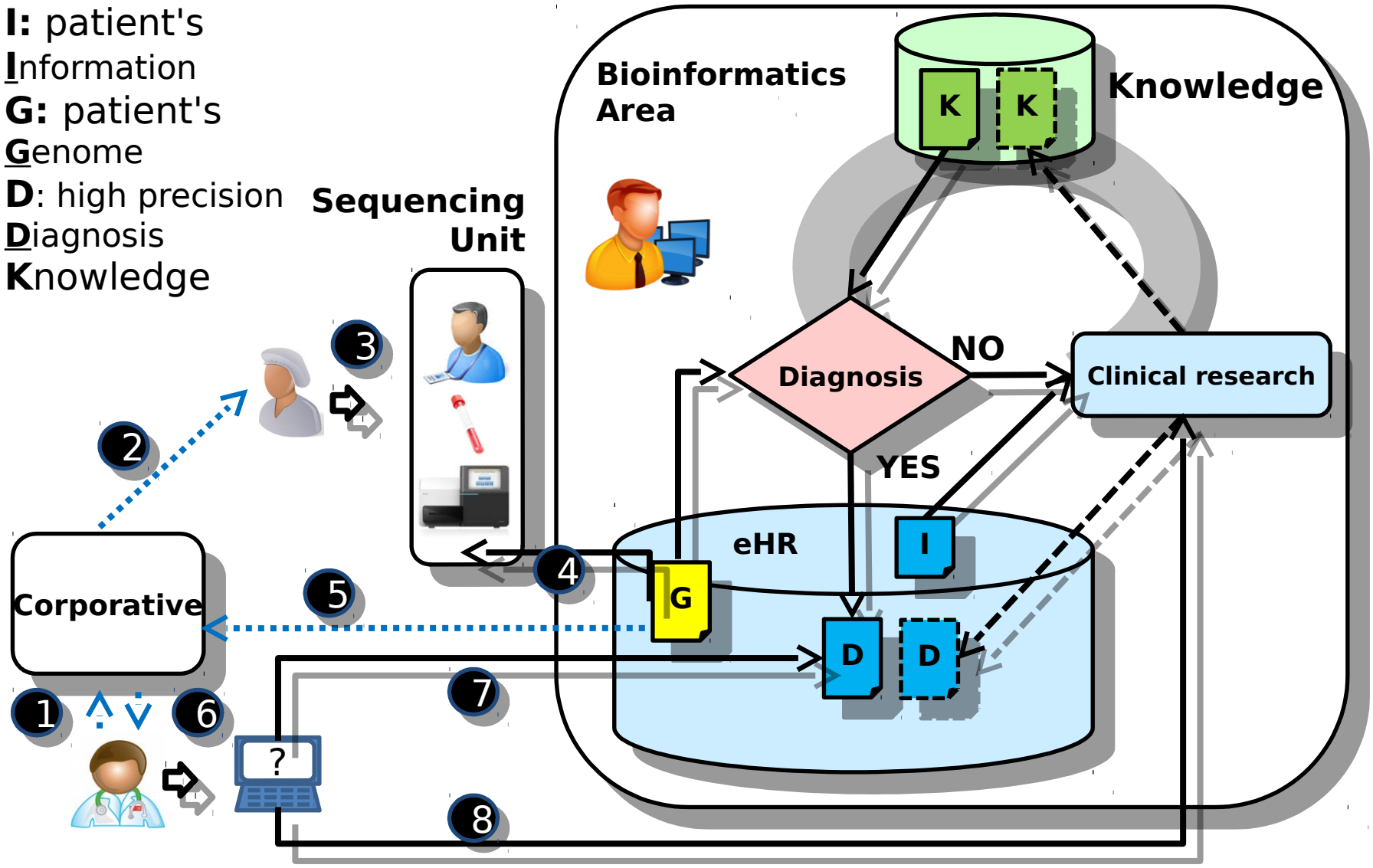
Idea: Patients of disease A can be considered healthy **pseudo-controls** for disease B (providing no common genetic background exist between A and B)

Beacon: CSVS has a Beacon server (Beacon network is search engine across beacons which enables global discovery of genetic mutations)

URL: https://beacon-network.org

Introduction

MMP for diagnosis and clinical research within the Andalusian health system



Overview

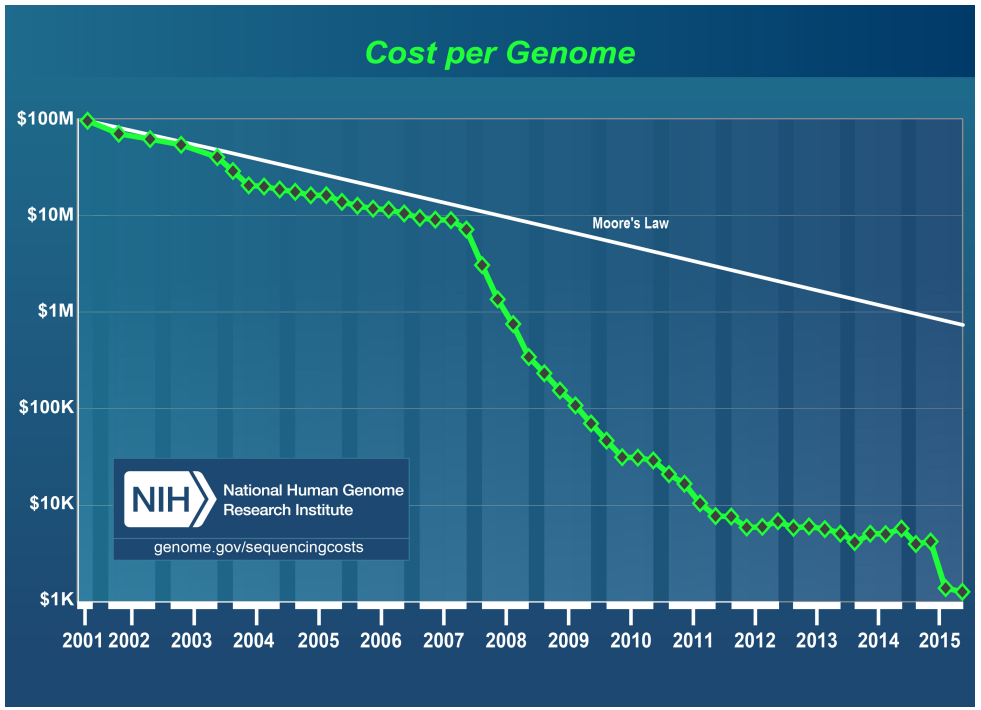
- Introduction
- ***Big Data* in Genomics**
- OpenCB: Open source initiative for Computational Biology
- A case study: Personalized Medicine Module (MMP)

Big data in Genomics
A new scenario in biology

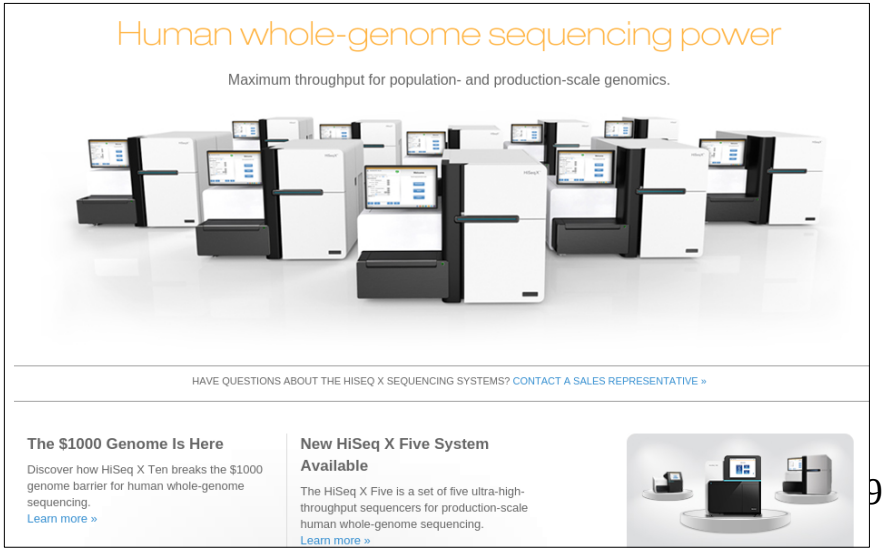
NGS experiments ***have increased data size by more than 5000x*** *compared to microarrays.*

Challenges*:*

- ***Sequencing cost*** keeps falling while throughput keeps increasing. Today a whole genome can be sequenced by less than **$1000**
- ***Variant analysis***: re-sequencing projects such as Whole Genome Sequencing (WGS) aims to find genomic variants and genes involved in phenotypes and diseases



A single HiSeq X Ten System can sequence ~20,000 human genomes a year



Big data in Genomics
A new scenario in biology

NGS experiments ***have increased data size by more than 5000x*** *compared to microarrays.*
Challenges*:*

- ***Sequencing cost*** keeps falling while throughput keeps increasing. Today a whole genome can be sequenced by less than **$1000**
- ***Variant analysis***: re-sequencing projects such as Whole Genome Sequencing (WGS) aims to find genomic variants and genes involved in phenotypes and diseases

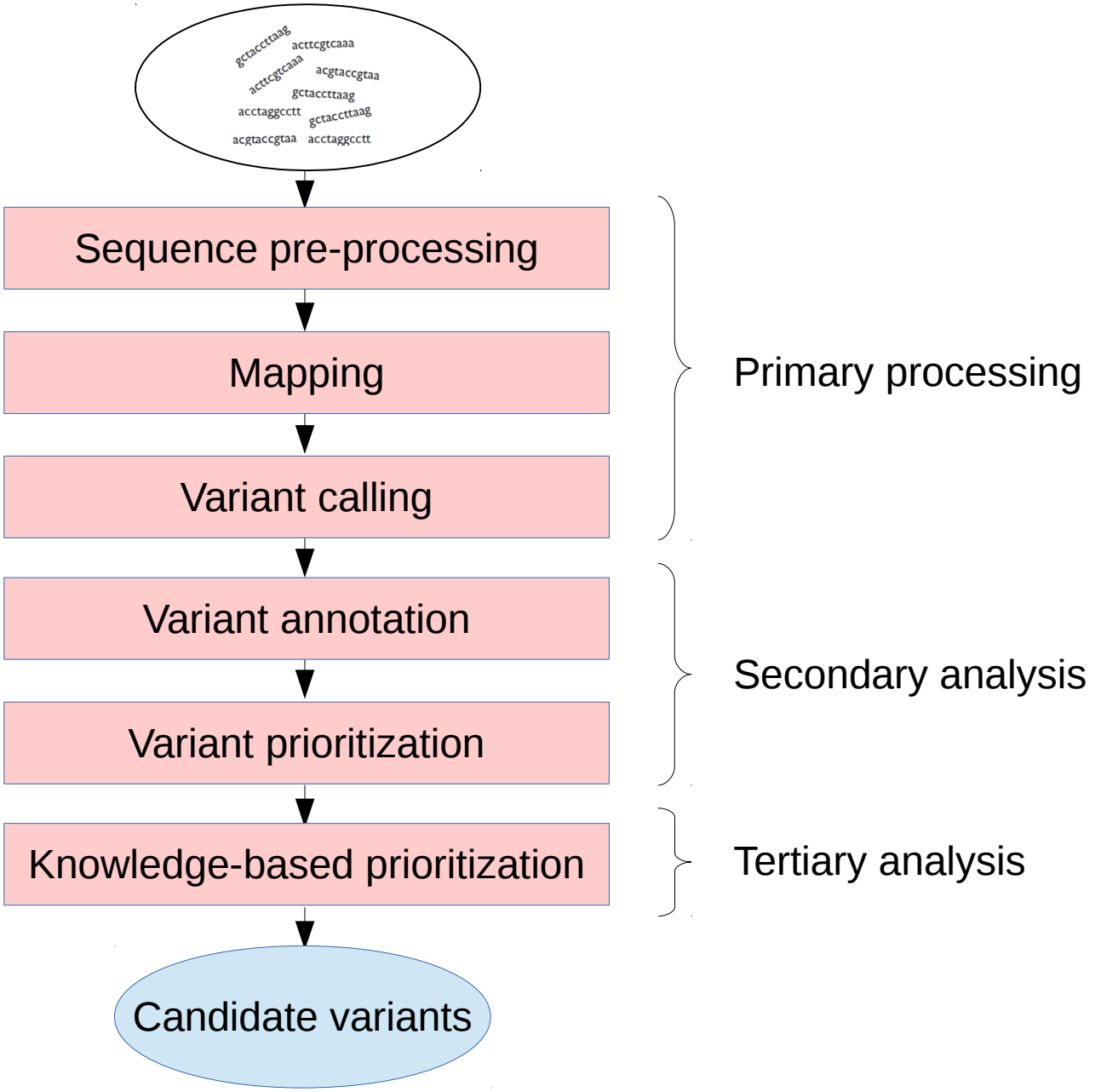
- Illumina NGS sequencer series:
  - **HiSeq 2500** provides high-quality 2x125bp: 50-1000Gb in 1-6 days, 90.2% bases above Q30. One human genome at ~60x coverage
  - **HiSeq 4000** provides high-quality 2x150bp: 125-1500Gb in 1-4 days, >75% bases above Q30. Up to 12 human genomes at ~40x coverage
- ***Each sample*** produces a ***FASTQ*** file ~**1TB** size containing ~**1-2B** reads
- **Illumina X Ten**: Consists if 10 ultra-high-throughput HiSeq X sequencers. First $1000 human genome sequencer, it can sequence up to 20,000 genomes per year

Real flexibility.
Real throughput.
Real data quality.

The HiSeq 2500 is ready for any application, any sample size—**today**.

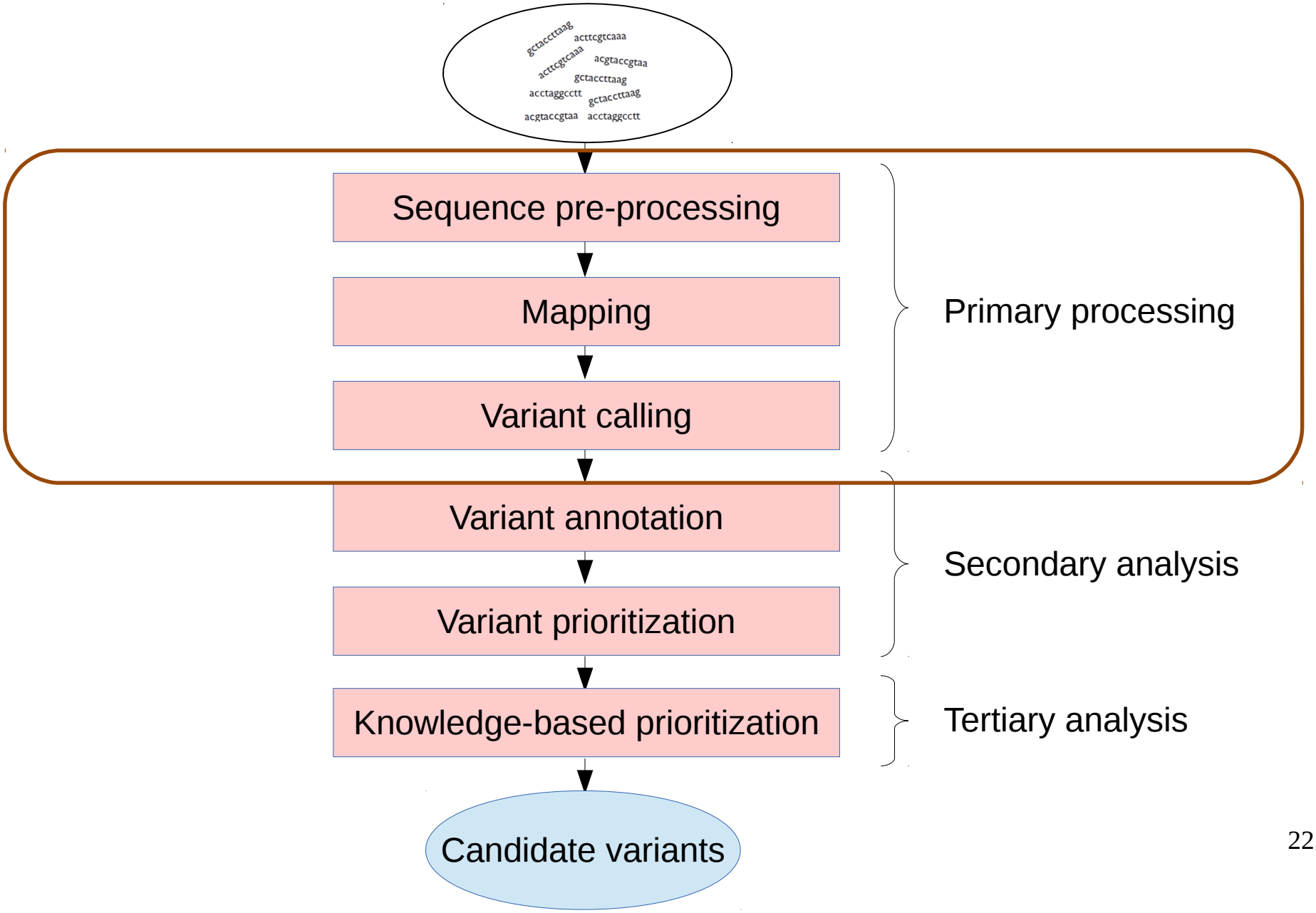
Big data in Genomics

Standard NGS variant data analysis



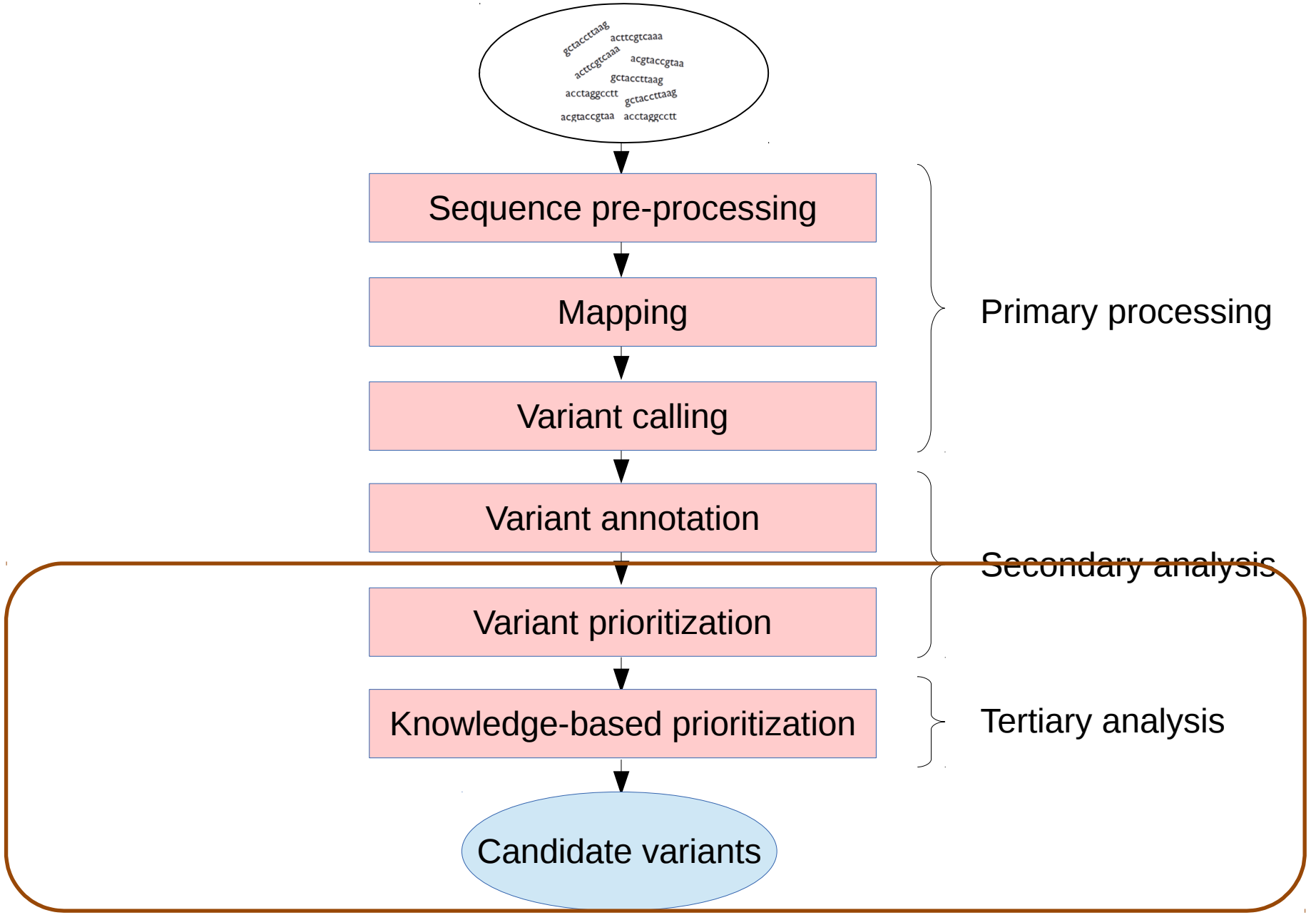
Big data in Genomics

Standard NGS variant data analysis



Big data in Genomics

Standard NGS variant data analysis



Big data in Genomics

Standard NGS variant data analysis

CAUTION!

On average, each *normal* person is found to carry in the **exome**:

~11,000 synonymous variants

~11,000 non-synonymous variants

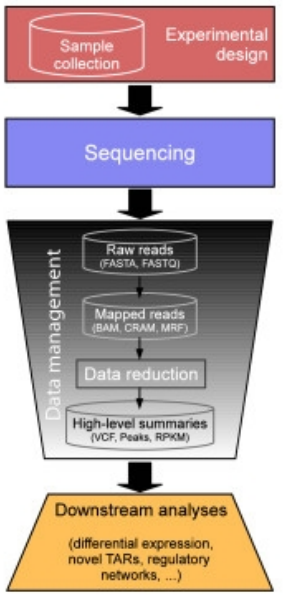
250 to 300 loss-of-function variants in annotated genes

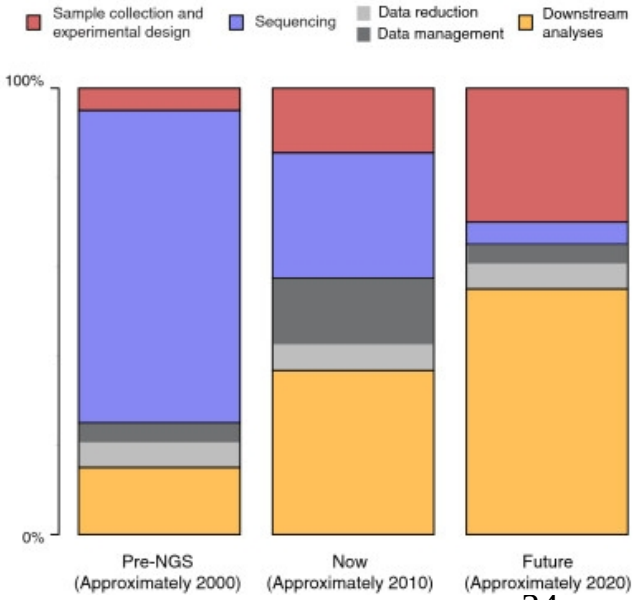
50 to 100 variants previously implicated in **inherited disorders**

1000 Genomes Project Consortium. *A map of human genome variation from population-scale sequencing.* **Nature**. 2010 Oct 28;467(7319):1061-73. PubMed PMID: 20981092

Still a challenge → PRIORITIZATION

- A **whole genome** can carry 3.5 million variants on average
- After annotating there will be **hundreds** of **deleterious** variants

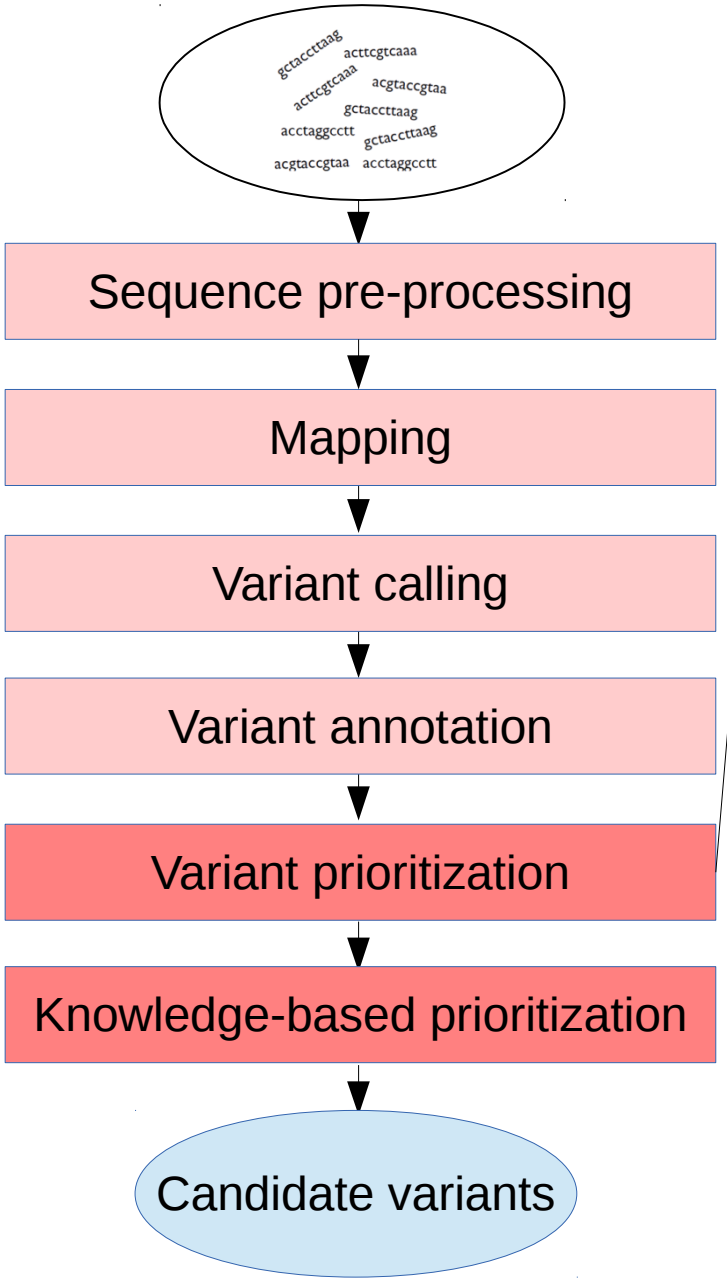




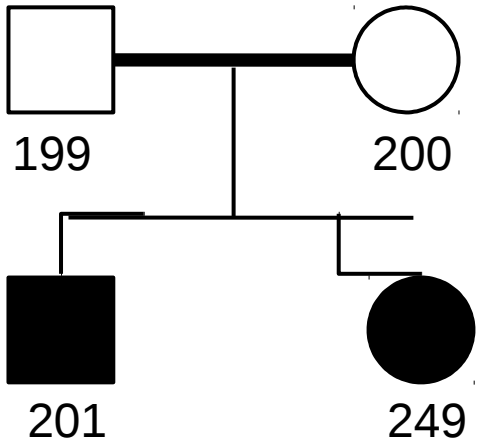
Sboner, Genome Biology 2011

Big data in Genomics

Standard NGS variant data analysis

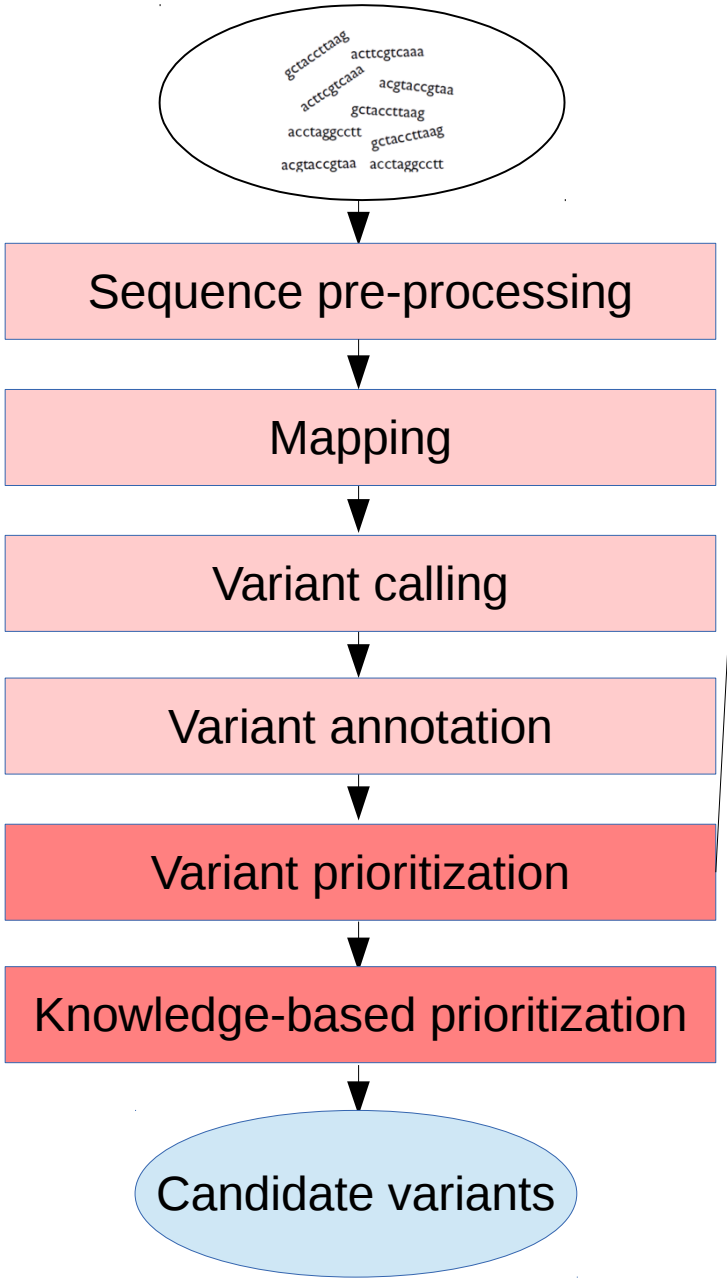


- **Variant level: population frequencies, conservation scores, protein substitution scores....**
- **Experimental design level: case/control, Family, Trios, Disease Panels**

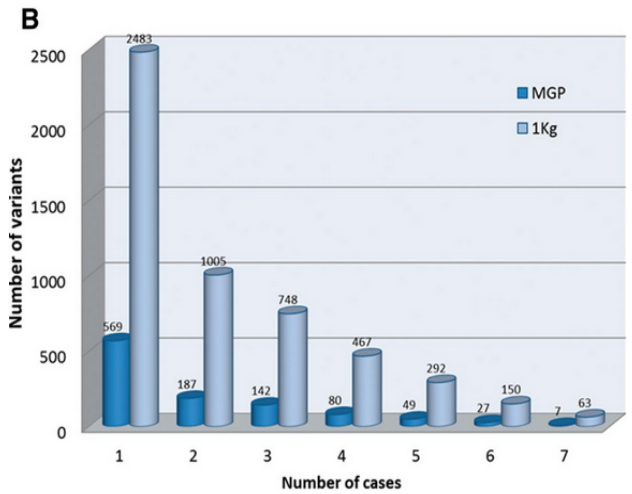


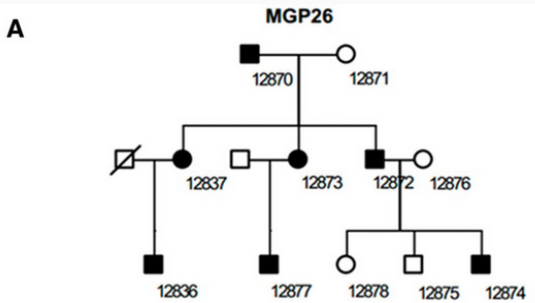
Big data in Genomics

Standard NGS variant data analysis



- **Variant level: population frequencies, conservation scores, protein substitution scores....**
- **Experimental design level: case/control, Family, Trios, Disease Panels**

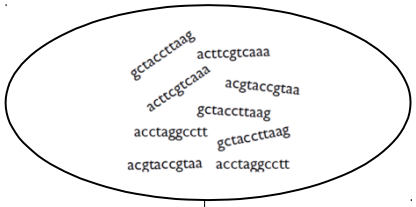




The filtering efficiency of the local population can be between 5 and 10 times those of a general database, such as the 1000 genomes

Big data in Genomics

Standard NGS variant data analysis



Sequence pre-processing

Mapping

Variant calling

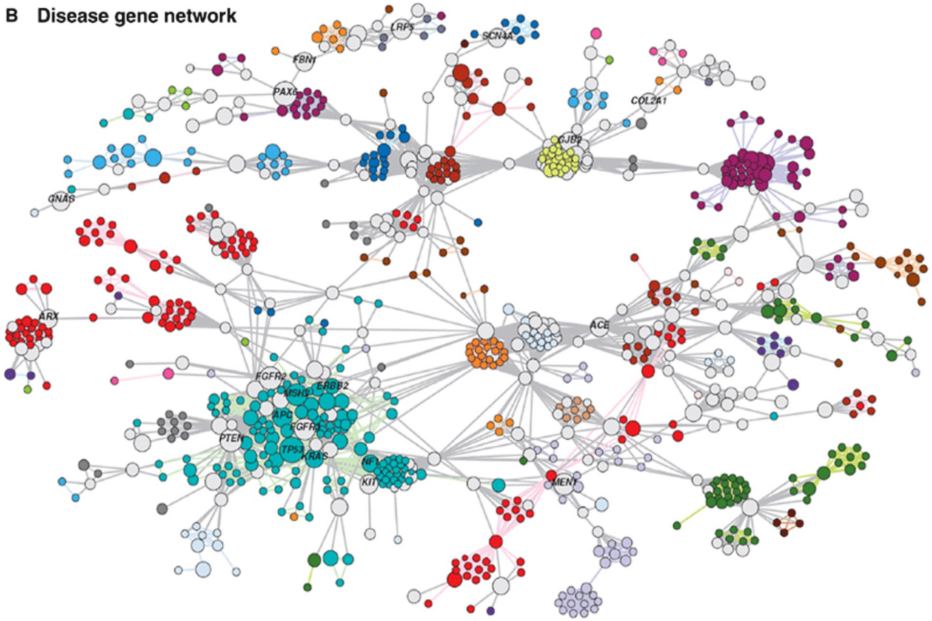
Variant annotation

Variant prioritization

Knowledge-based prioritization

Candidate variants

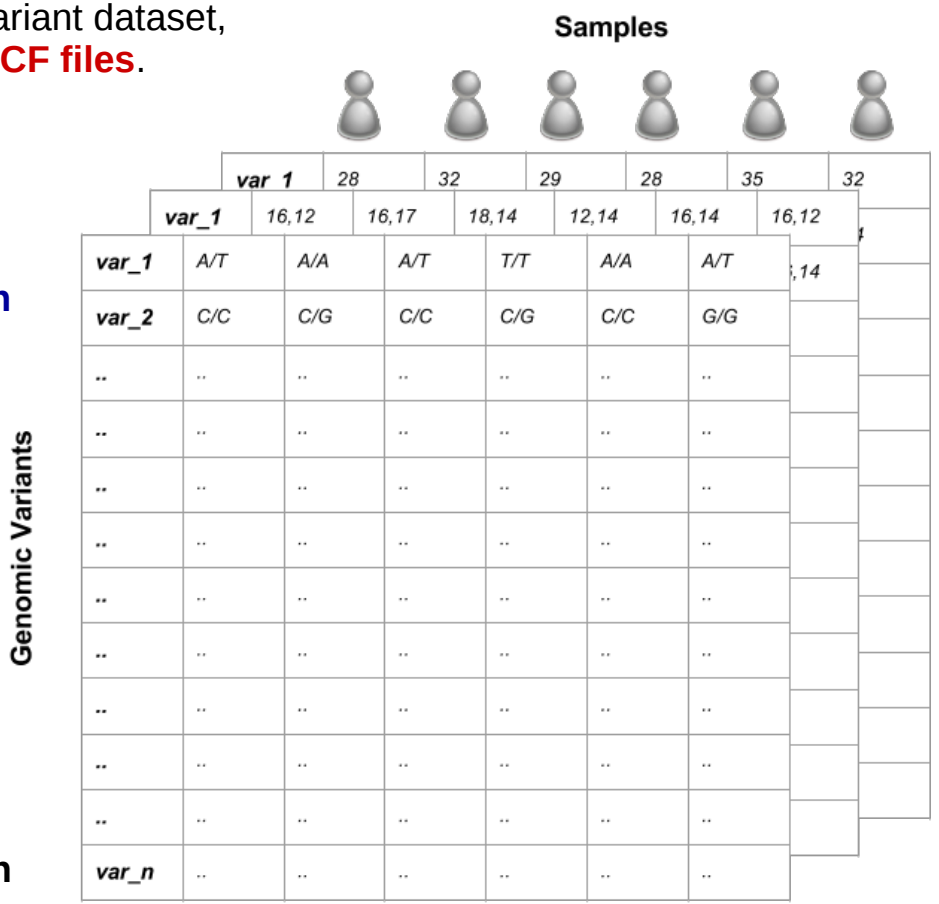
- **Functional (system) level: gene set, network analysis, pathway analysis, burden tests....**



Big data in Genomics

Genomic Variant Dataset, big and complex

Logical view of genomic variant dataset, data come from **different VCF files**.



Hundreds of millions of mutations, some meta data needed: **Variant annotation**

- Clinical info
- Consequence types
- Conservation scores
- Population frequencies
- ...

Genomics England project:

- **250M variants** x **100K samples**, about **25 trillion** points
- With different layers of data, about **100 trillion** points
- A lot of meta data for variants and samples
- About **400TB** to be indexed

Different layers of information:

- Genotype for samples
- Allele counts
- Quality scores
- *Phase* information
- ...

Clinical data: **Sample annotation**

- Phenotype
- Family and population pedigree
- Clinical variables
- ...

Heterogeneous data analysis and algorithms, different technologies and solutions required:

- Search and filter using data and meta data
- Data mining, correlation
- Statistic tests
- Machine learning
- Interactive analysis
- Network-based analysis
- Visualization
- Encryption
- ...

Applications:

- Personalized medicine
- Clinical Genomics
- ...

Big data in Genomics
Some *Big Data current projects*

- ***NIHR BRIDGE***: 10,000 whole genomes from rare diseases, ~1-2PB of data expected
- ***Genomics England (GEL)***: is sequencing **100,000** whole genomes from UK, several rare diseases and cancers being studied, data estimation: ~20PB of BAM and **~400TB of compressed VCF** data are expected! About 100 whole genomes/day, ~5-10TB/day
- ***International Cancer Genome Consortium (ICGC)***: store more than 10,000 sequenced cancers, few PB of data
- **NAGEN:** 1000 whole genomes from rare diseases and cancer
- **France Médicine Génetique 2025**: 10.000 WGS correspondign to 20.000 patients with rare diseases and their families, and 50.000 patients with metastatic or refractory cancers
- **Genome of the Netherlands Consortium (GoNL):** GoNL is interested in genetic variation in the Dutch Population.
- **Andalusian program for Personalized Medicine:** 4500 samples per year (targeted sequencing, rare diseases)
- …..
- Danish National Strategy of Personalized Medicine (2017-2020)
- Qatar Genome Programme (QGP)
- …..

Big data in Genomics

Some *Big Data current projects*

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For A

Archive > Volume 532 > Issue 7600 > News > Article

NATURE | NEWS

AstraZeneca launches project to sequence 2 million genomes

Drug company aims to pool genomic and medical data in hunt for rare genetic sequences associated with disease.

Heidi Ledford

22 April 2016

PDF | Rights & Permissions

One of the world's largest pharmaceutical companies has launched a massive effort to compile genome sequences and health records from two million people over the next decade. In doing so, AstraZeneca and its collaborators hope to unearth rare genetic sequences that are associated with disease and with responses to treatment.

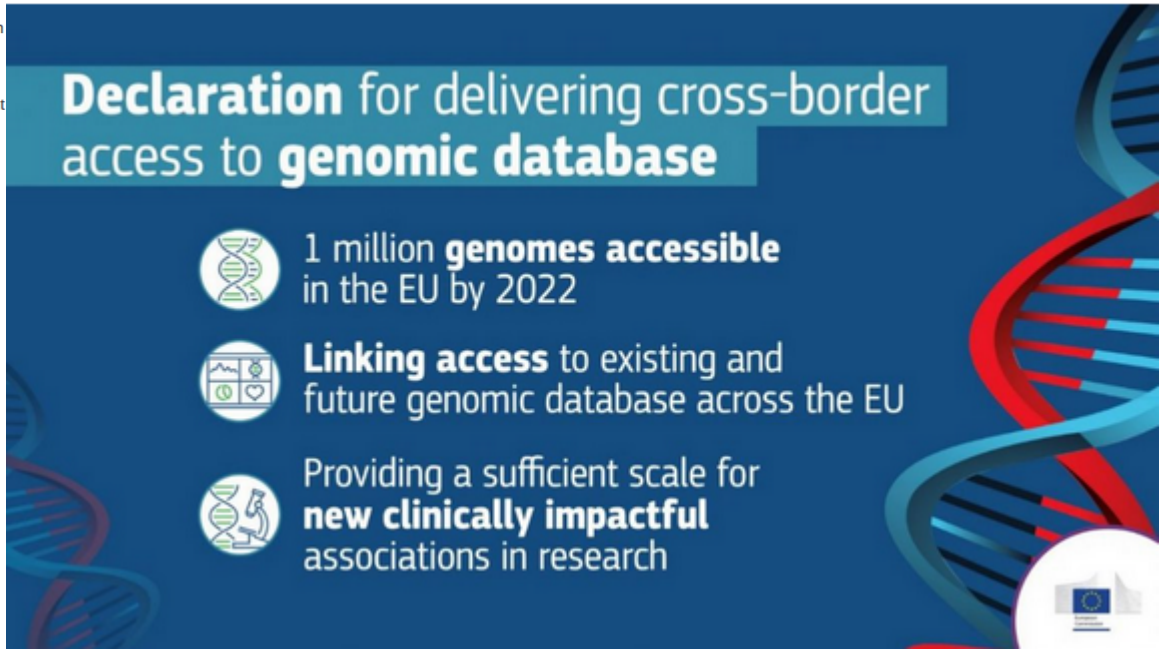
It's an unprecedented number of participants for this type of study, says Ruth March, vice-president and head of personalized health care and biomarkers at AstraZeneca, which is headquartered in London. "That's necessary because we're going to be looking for very rare differences among individuals."

European Commission > Strategy > Digital Single Market > Policies >

Digital Single Market

POLICY

European '1+ Million Genomes' Initiative



Declaration for delivering cross-border access to **genomic database**

1 million **genomes accessible** in the EU by 2022

Linking access to existing and future genomic database across the EU

Providing a sufficient scale for **new clinically impactful** associations in research

Big data in Genomics

Big data analysis challenges

- ***Data Analysis and visualization***: Real-time and Interactive graphical data analysis and visualization is needed.
- ***Data mining***: Complex queries, aggregations, correlations, ...
- ***Security***: sometimes data access require authentication, authorization, *encryption*, ...
- ***Performance and scalability***: software must be high-performance and scalable
- ***Data Integration***: different types of data such as variation, expression, ChIP, ...
- ***Share and collaboration***: data models to ease the collaboration among different groups. Avoid moving data.
- ***Knowledge base and sample annotations***: many of the visual analytic tools need genome and sample annotations

Do current bioinformatic tools solve these problems?

Big data in Genomics

Current status of *big data* tools in bioinformatics

- Many bioinformatic tools are great but, in general, not designed and implemented for processing and analyzing *big data*.
- Tools usually don't exploit the parallelism of modern hardware and current high-performance and scalable technologies. Poor performance and scalability.
- We need to develop new generation of software and methodologies to:
  - Improve performance and scalability of analysis
  - Store data efficiently and secured to be queried and visualized
- **Bioinformaticians have new challenges and requirements, given the volume, complexity, heterogeneity and nature of data.**

OXFORD

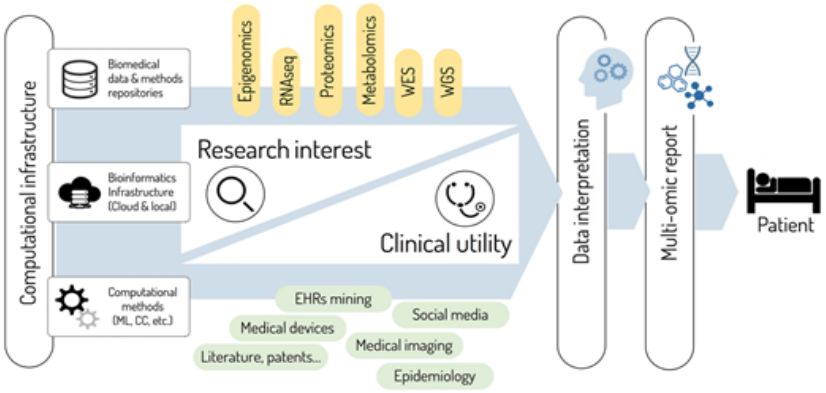
Briefings in Bioinformatics, 2017, 1–15

doi: 10.1093/bib/bbx144
Paper

Precision medicine needs pioneering clinical bioinformaticians

Gonzalo Gómez-López, Joaquín Dopazo, Juan C. Cigudosa, Alfonso Valencia and Fátima Al-Shahrour

Corresponding author. Fátima Al-Shahrour, Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernández Almagro, 3. ES-28029 Madrid, Spain. Tel.: +34 91 732 80 00; E-mail: falshahrour@cnio.es



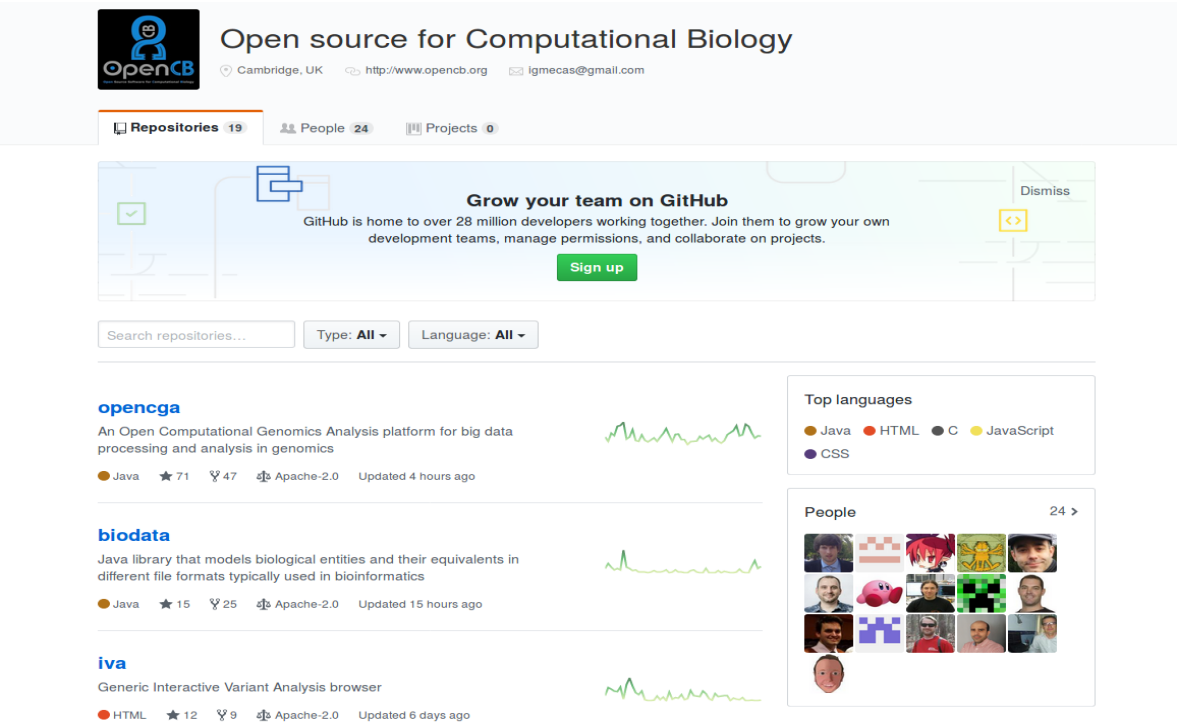
Overview

- Introduction
- Big Data in Genomics
- **OpenCB: Open source initiative for Computational Biology**
- A case study: Personalized Medicine Module (MMP)

OpenCB

Open source initiative for Computational Biology

- OpenCB is a collaborative project that **aims** to design and develop high-performance and scalable solutions for genomic big data analysis using **most modern computing technologies**.
- OpenCB is a collaborative project with more than 15 actives developers and data analysts and more than 12 repositories (**http://www.opencb.org**)
- No one computing programming language oriented: BioPerl, BioPython, Bioconductor, ... Good software solutions may use different languages and technologies to solve different problems and use cases
- So far, is where all the software we develop is being released. About 15 active committers. Available as open-source at GitHub **https://github.com/opencb**



OpenCB

Some relevant projects

- ***biodata*** (https://github.com/opencb/biodata) and ***ga4gh*** (https://github.com/opencb/ga4gh)
  - Contain all data models (Variant, Alignment...) parsers and converters (*avro, protobuf*) for all OpenCB projects
- ***CellBase*** (https://github.com/opencb/cellbase)
  - A NoSQL database the integrates the most relevant biological information about genomic features and proteins, gene expression regulation, etc. **The biological knowledge-base for OpenCB project** . A ***Variant Annotation*** tool implemented.
  - A high-performance NoSQL implementation, a CLI and web services implemented
- ***HPG BigData*** (https://github.com/opencb/hpg-bigdata)
  - Hadoop-based implementation of data converters (avro, parquet) and bioinformatic tools (ie. samtools)
  - Simple indexing for HBase, Hive and Impala developed
  - C code embedded using JNI to speed-up processing
- ***OpenCGA*** (https://github.com/opencb/opencga)
  - Integrates most of the OpenCB projects to provide a scalable and high-performance platform for the analysis of HT genomic data
  - **OpenCGA Catalog** provides an authenticated environment, files and sample annotations, system audit, ...
  - **OpenCGA Storage** is a plugin-oriented framework that allows to index hundreds of millions of variants for thousands of samples in different storage engines. Stats and annotation implemented..
- ***Interactive Variant analysis (IVA)***
  - Web-based variant analysis tool
- ***Genome Maps*** (https://github.com/opencb/genome-maps)
  - A web-based NGS and genome browser: http://genomemaps.org/
- Many other related projects for big data analysis and visualization, check https://github.com/opencb

OpenCB: CellBase

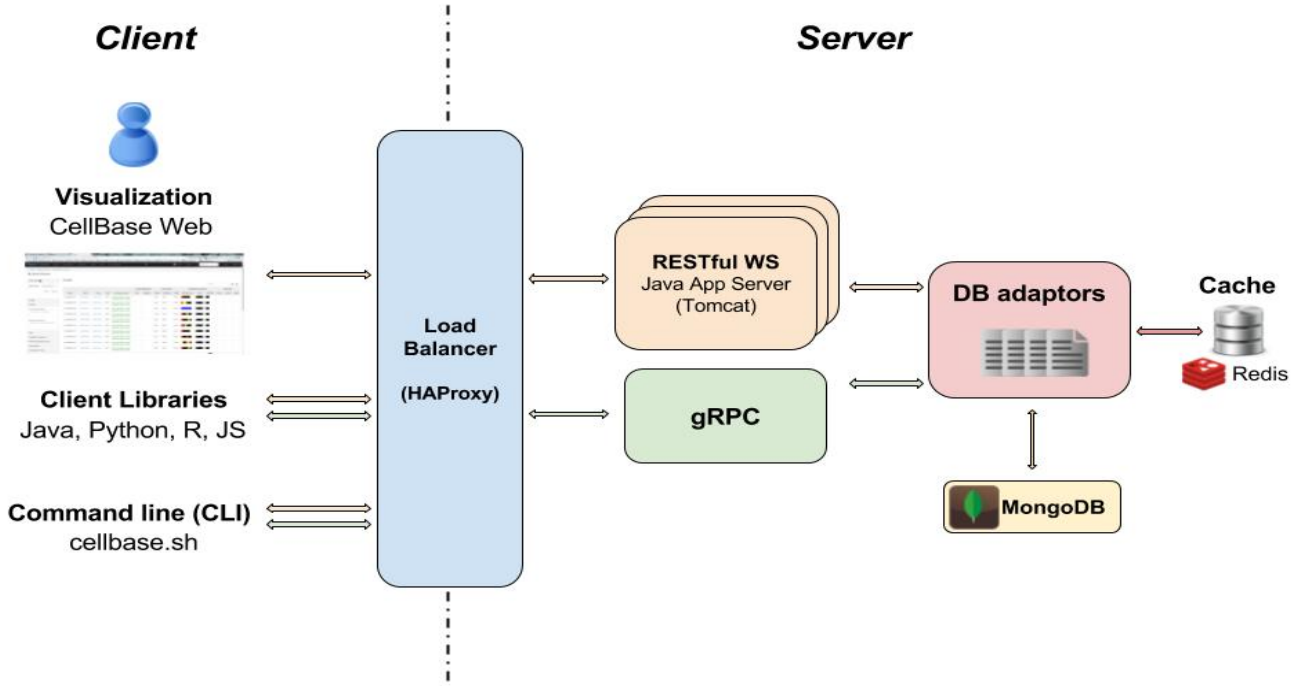
An integrative database and RESTful Web Service API

- ***CellBase*** is a comprehensive integrative NoSQL database and a *RESTful Web Service API*, designed to provide a ***high-performance and scalable*** solution.
- It integrates the most relevant biological information about genomic features and proteins, gene expression regulation, functional annotation, genomic variation...Currently contains more than 2TB of data
- Used by EMBL-EBI, ICGC, GEL, MMP among others

Category	Data source	Version/Date	
		CellBase v3 (*March 2015*)	CellBase v4 (*June 2016*)
Core	Ensembl Core	v79	v82
Protein	UniProt	March 2015	Release 2015_10
	InterPro	v50	v54
	Polyphen2/Sift	Ensembl 79	Ensembl v82
Variation	Ensembl Variation	v79	v82
	1000 genomes project		Phase 3 2016-05
	ExAC		0.3.1
	GoNL		Release 5
	UK10K		2016-05
	ESP		2016-05
Regulatory	Ensembl Regulatory	v79	v82
Conservation	PhastCons		June 2016
	PhyloP		June 2016
	GERP++		June 2016
Clinical	ClinVar	March 2015	2016-12
	COSMIC	v71	v79
	HPO		2015-11
	DisGeNET		Version 3.0
Biological Networks	Reactome	v51	June 2016
	IntAct	March 2015	June 2016
Others	DGIdb		2.0
	Gene Expression Atlas		June 2016
	CADD		v1.3

- **Official Domain and swagger: http://bioinfo.hpc.cam.ac.uk/cellbase/webservices/**
- **Current version:** v4.7.1
- **Data sources for homo sapiens:** http://bioinfo.hpc.cam.ac.uk/cellbase/webservices/rest/v4/meta/hsapiens/versions?assembly=grch37&limit=-1&skip=-1&skipCount=false&count=false&Output%20format=json
- **Project: https://github.com/opencb/cellbase Wiki: http://docs.opencb.org/display/cellbase/**

OpenCB: CellBase
Arquitecture

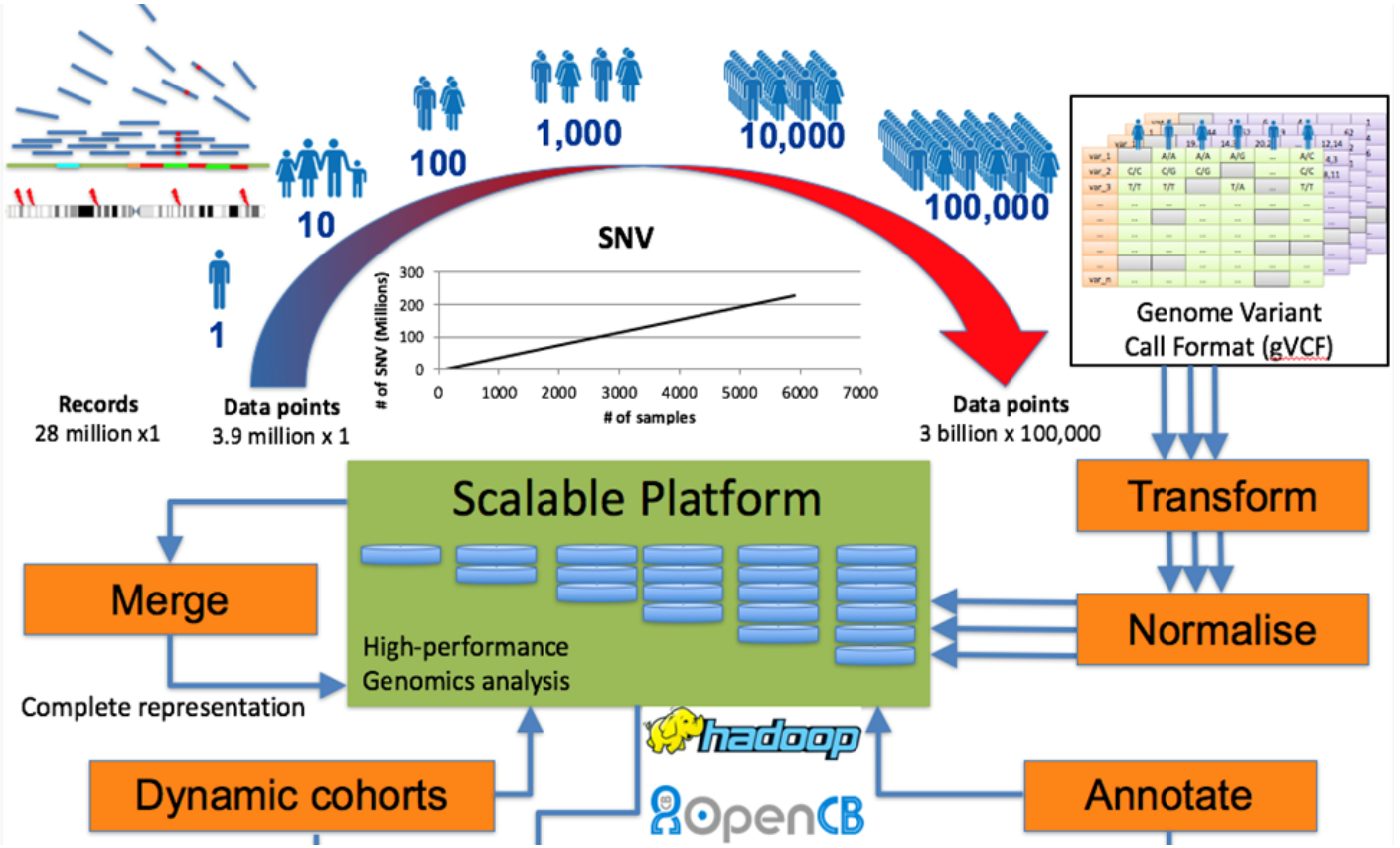


- **Example of variant annotation**:
  - http://bioinfo.hpc.cam.ac.uk/cellbase/webservices/rest/v4/hsapiens/genomic/variant/19:45411941:T:C/annotation

OpenCB: OpenCGA

Overview and goals

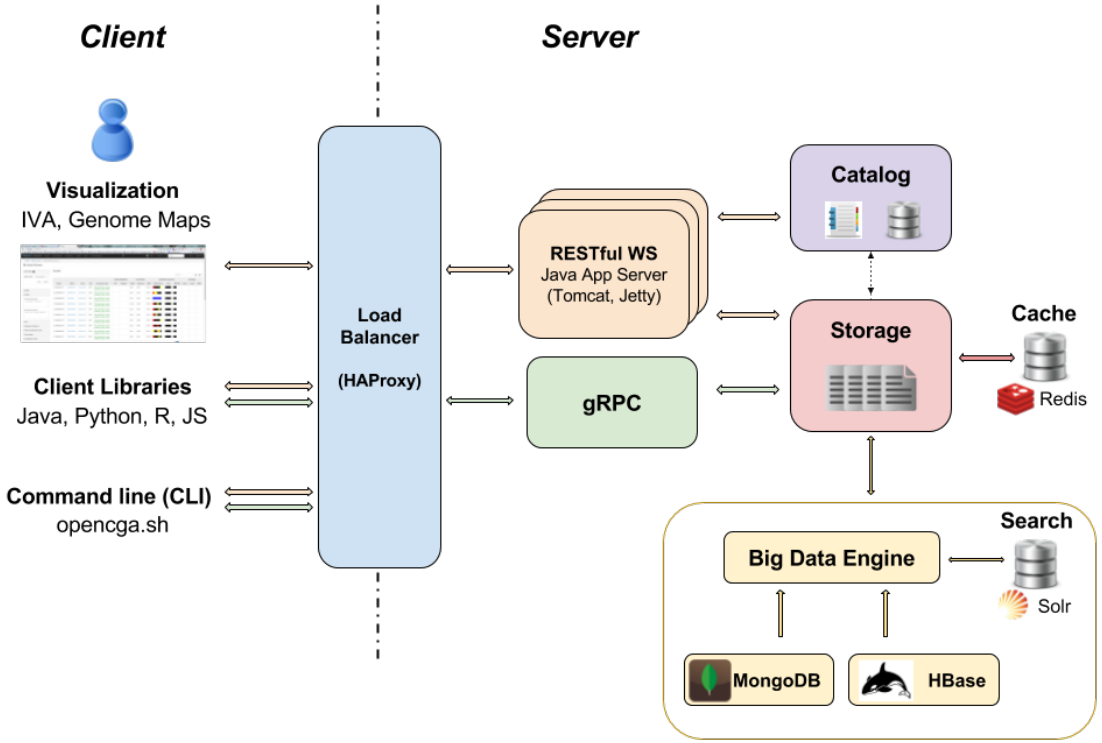
- Open-source Computational Genomics Analysis (**OpenCGA**) aims to provide to researchers and clinicians a ***high performance and scalable solution*** for genomic big data processing and analysis
- **OpenCGA** is built on OpenCB
- Project: https://github.com/opencb/opencga ; Doc: http://docs.opencb.org/display/opencga
- Currently, the fastest and more powerful genomic engine in the world. Used in Genomics England (GEL) and MMP

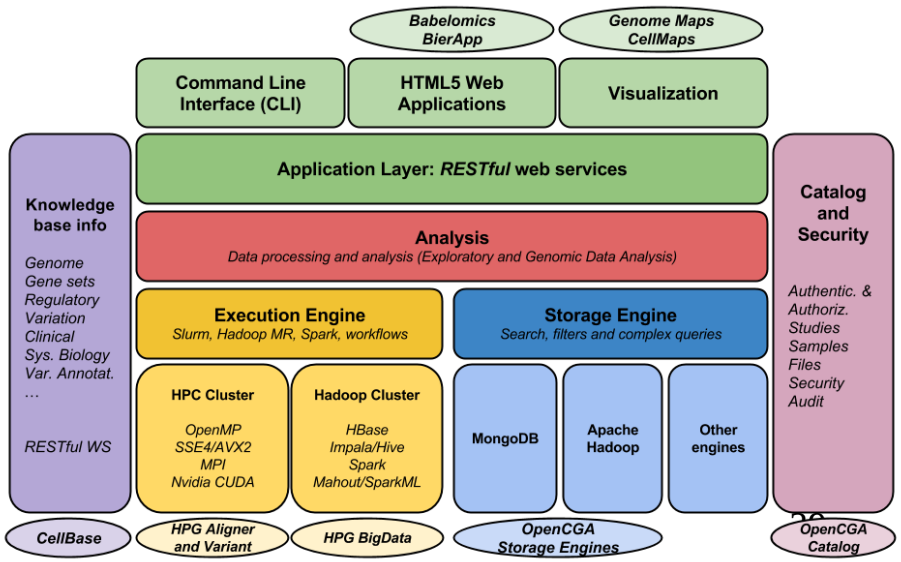


OpenCB: OpenCGA

Overview and goals

- Main features:
  - High-performance and scalable variant storage and index that allow to load and merge VCF/gVCF files
  - Annotate and calculate statistics for all the variants
  - Client libraries developed in Java, Python, R and Javascript
  - Clinical interpretation analysis of samples and families
  - Integrated Catalog keeps track of users, files, jobs, clinical data...
  - Interactive web-based data mining tool based on IVA

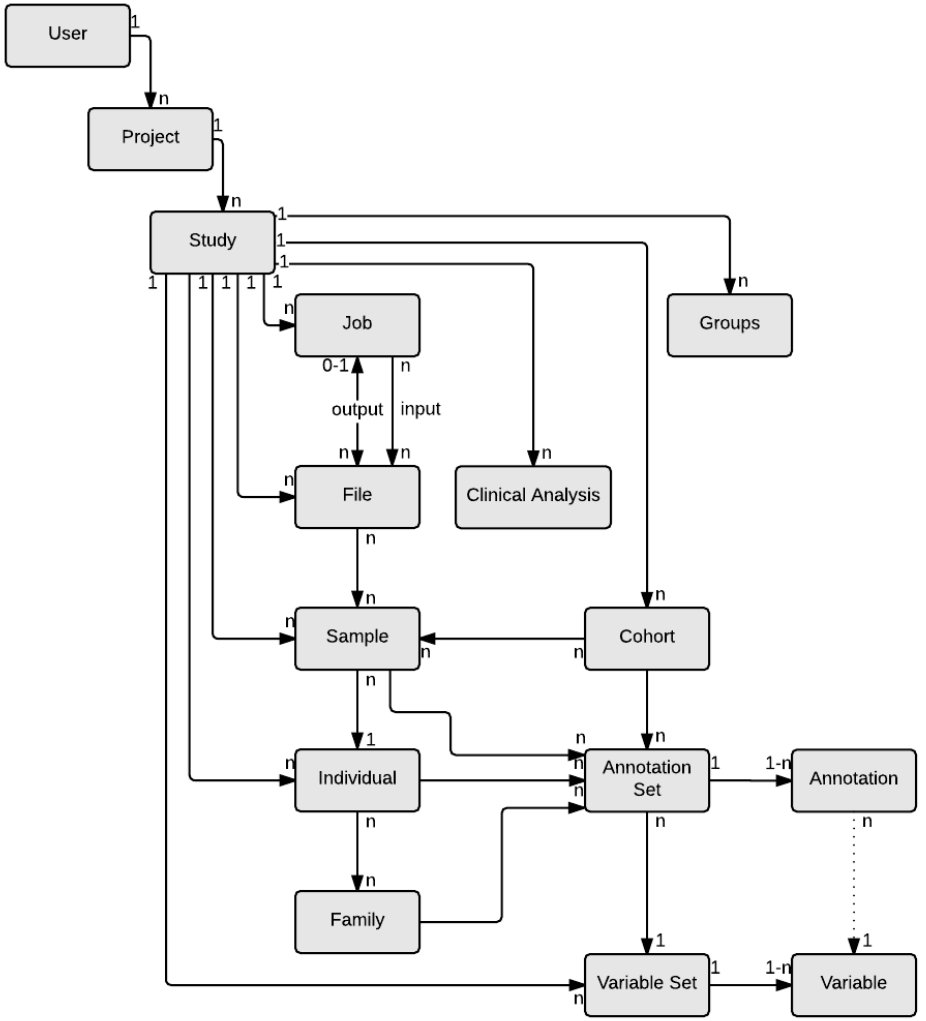




OpenCB: OpenCGA

Catalog, a *metadata* database

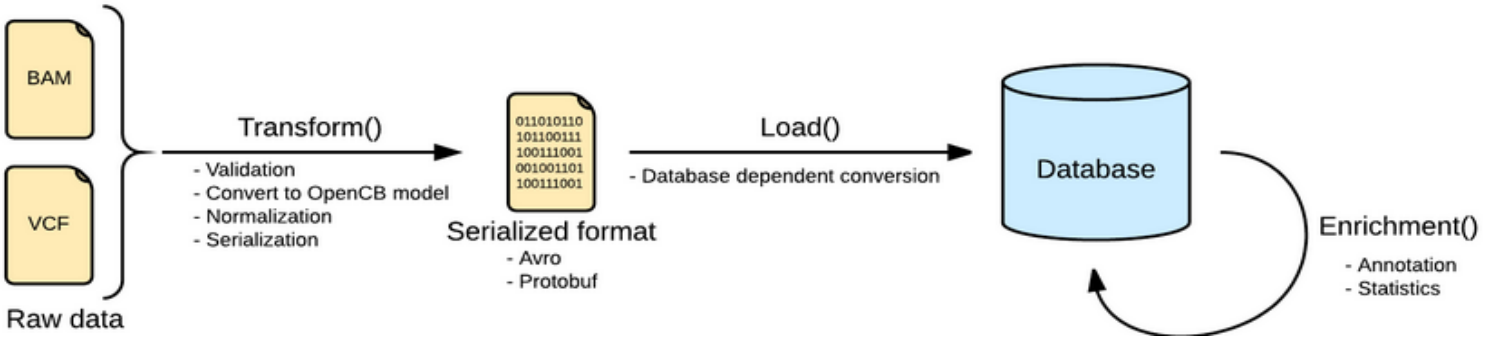
- **OpenCGA Catalog** provides a user authentication and authorization environment, implements *sample annotation* database, tracks all files and job, audit, …
- ***Authentication & authorization***: *roles* definition, detailed ACL system implemented to share sample, file, jobs, ...
- ***File and jobs***: file and jobs are tracked, stats calculated
- ***Sample annotations*** is one of the main features:
  - Allow complex queries and aggregations
  - Allow to detect bias and other problems with the data
  - Cohort definitions
- ***Audit***: all actions (login, data indexing, ...) and queries are audited
- Allow to the different big data storage engines to perform optimizations
- Implements a job launcher and a execution monitor
- Data models at https://github.com/opencb/opencga/tree/develop/opencga-catalog/src/main/java/org/opencb/opencga/catalog/models



OpenCB: OpenCGA
Storage Engines

OpenCGA Storage provides a *pluggable* Java framework for storing and querying alignment and variant data

- Two default implementations: **MongoDB** and **Hadoop** for huge performance and scalability ~ hundreds of thousands of samples



Variant Index Pipeline

1. **Transform**
   - Validation: VCF files are read using the library HTSJDK → validation
   - Files are converted to Biodata models. The metadata is stored into a file serializing in json a single instance of the biodata model “VariantSource” (header and some general stats). The real variants data is serialized in file (avro) with a set of variant records described as the biodata model “Variant”
   - Normalization → unify the variants representation, since the VCF specification allows multiple ways of referring to a variant and some ambiguities.
2. **Load**
   - Variants are loaded into the database and merged with the existent ones.
3. **Enrichment**
   - Variant annotation (from Cellbase or read from local files provided by the user)
   - Stats calculation useful for filtering variants

OpenCB: OpenCGA

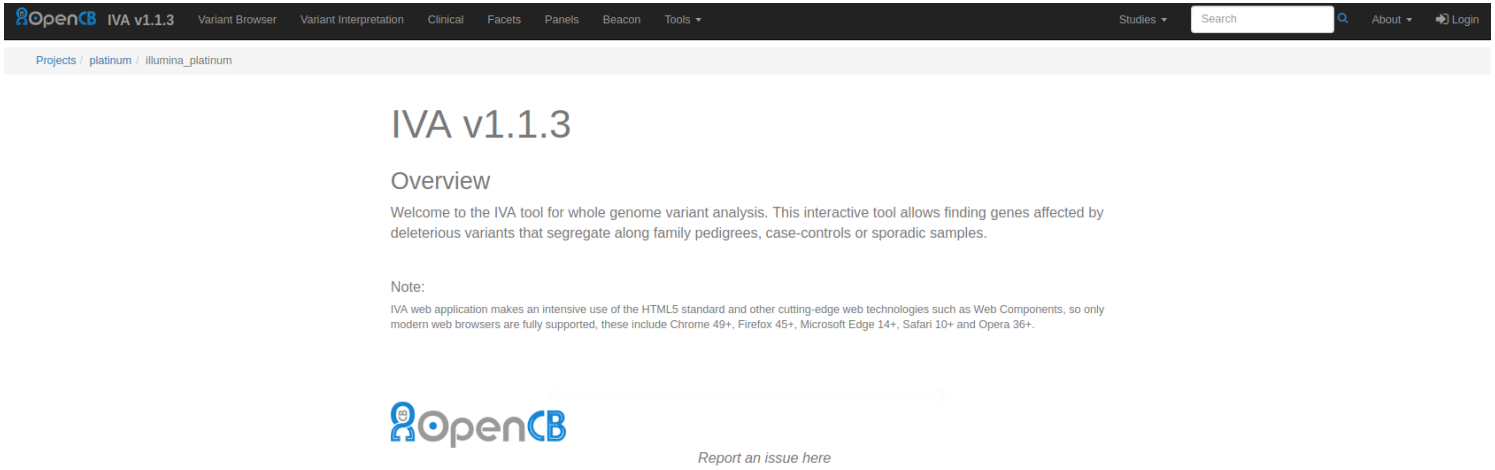
User interfaces: REST, command lines and client libs

- **RESTful API**:
  - more than 160 web services developed for Catalog (e.g. search *samples* or *files*) and Analysis (e.g. query *variants* or fetch *alignment reads*)
  - Other features: authentication and authorization, Swagger documentation (http://bioinfo.hpc.cam.ac.uk/hgva/webservices/), ...
- ***opencga.sh*** command line: you can query remote REST or gRPC services after authentication
- **Client libs**: allow to easily query REST web services. At the moment Java, Python, R and Javascript are fully developed and with similar level of functionality (e.g. authentication, parallel fetches, error support, …).
- **Current version**: v1.4.1
  - Roadmap: http://docs.opencb.org/display/opencga/Roadmap

OpenCB: Interactive Variant Analysis (IVA)

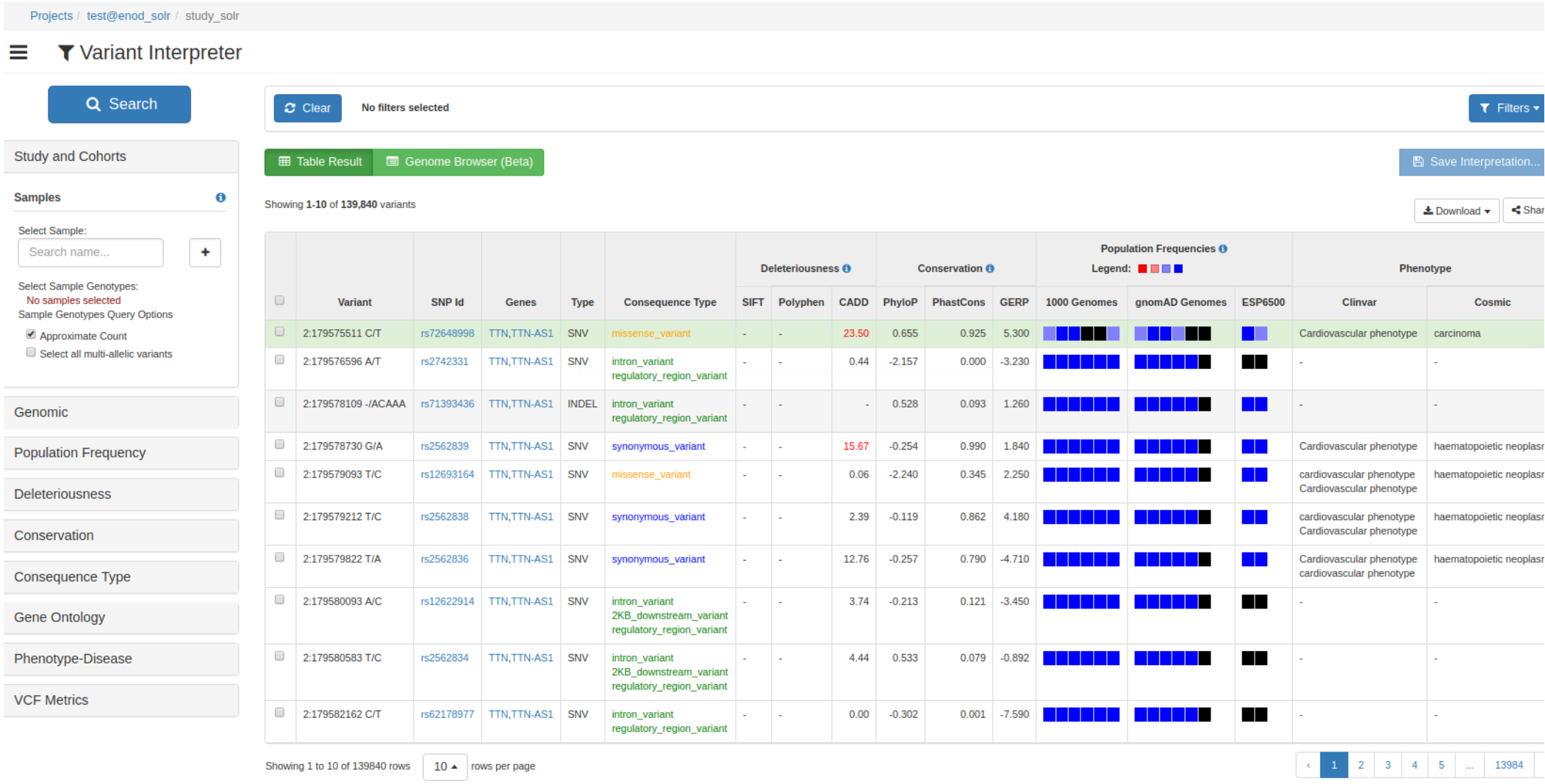
An interactive web-based variant analysis suite

- A collaborative project to implement a rich and interactive web-based variant analysis tool on top of **OpenCGA**, currently supports:
  - Variant Browser
  - Prioritization
  - Clinical Analysis
  - Panels
  - Tools
  - ...



OpenCB: Interactive Variant Analysis (IVA)

An interactive web-based variant analysis suite



Other OpenCB projects: HGVA

The Human Genomic Variation Archive

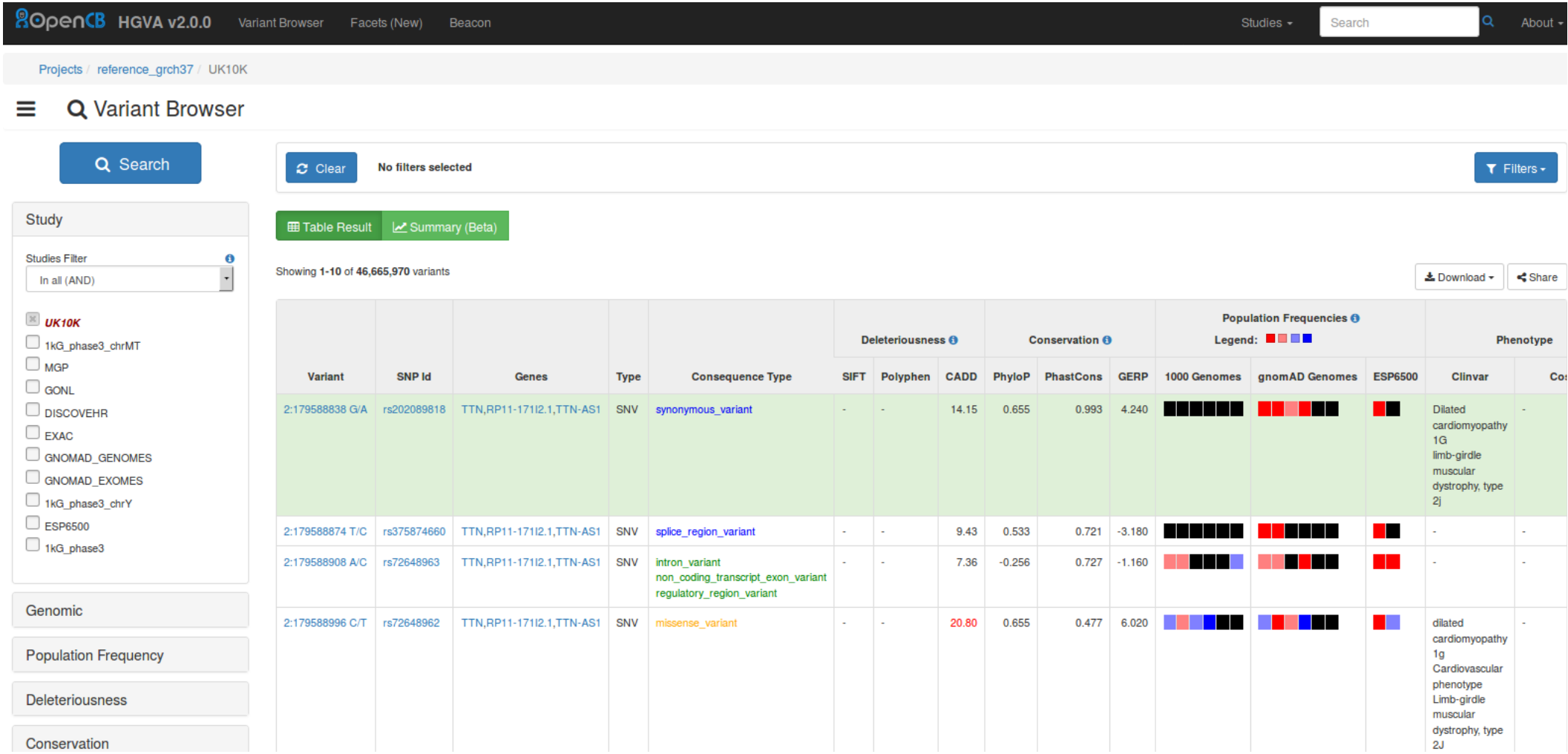
- Integrates variants from the main reference human projects
  - Adds valuable information such as variant annotation: consequence types, population frequencies, protein effect predictions, variant-associated phenotype, etc
- Main features:
  - Annotation performed using CellBase
  - Population frequencies calculated, including populations and super-populations
  - Data is indexed using OpenCGA
  - Rich interactive web-based data mining tool based on IVA

Project name *(alias)*	Studies		HGVA Version (date)	
	Name	Alias	v1 (*Dec. 2016*)	v2 (*Jan. 2018*)
Reference GRCh37 *(reference_grch37)*	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_EXOMES*	-	
	Genome Aggregation Database (gnomAD Genomes)	*GNOMAD_GENOMES*	-	
	Spanish Medical Genome Project (MGP)	*MGP*	2016-12	2016-12
Reference GRCh38 *(reference_grch38)*	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_EXOMES*	-	
	Genome Aggregation Database (gnomAD Genomes) (*)	*GNOMAD_GENOMES*	-	
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)

Other OpenCB projects: HGVA

The Human Genomic Variation Archive



- Published at NAR, 2017
- http://hgva.opencb.org

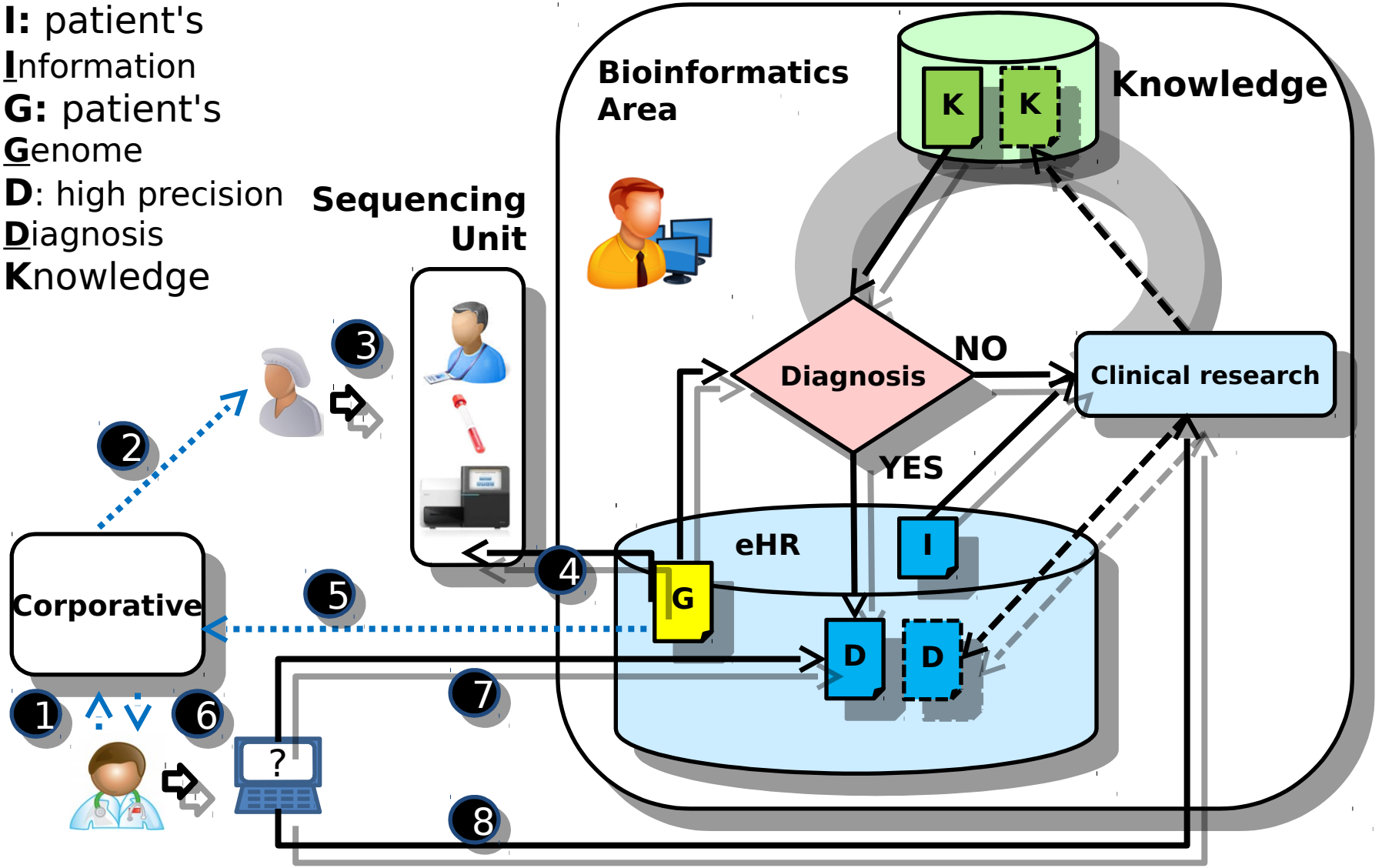
Overview

- Introduction
- Big Data in Genomics
- OpenCB: Open source initiative for Computational Biology
- **A case study: Personalized Medicine Module (MMP)**

Big data in Genomics

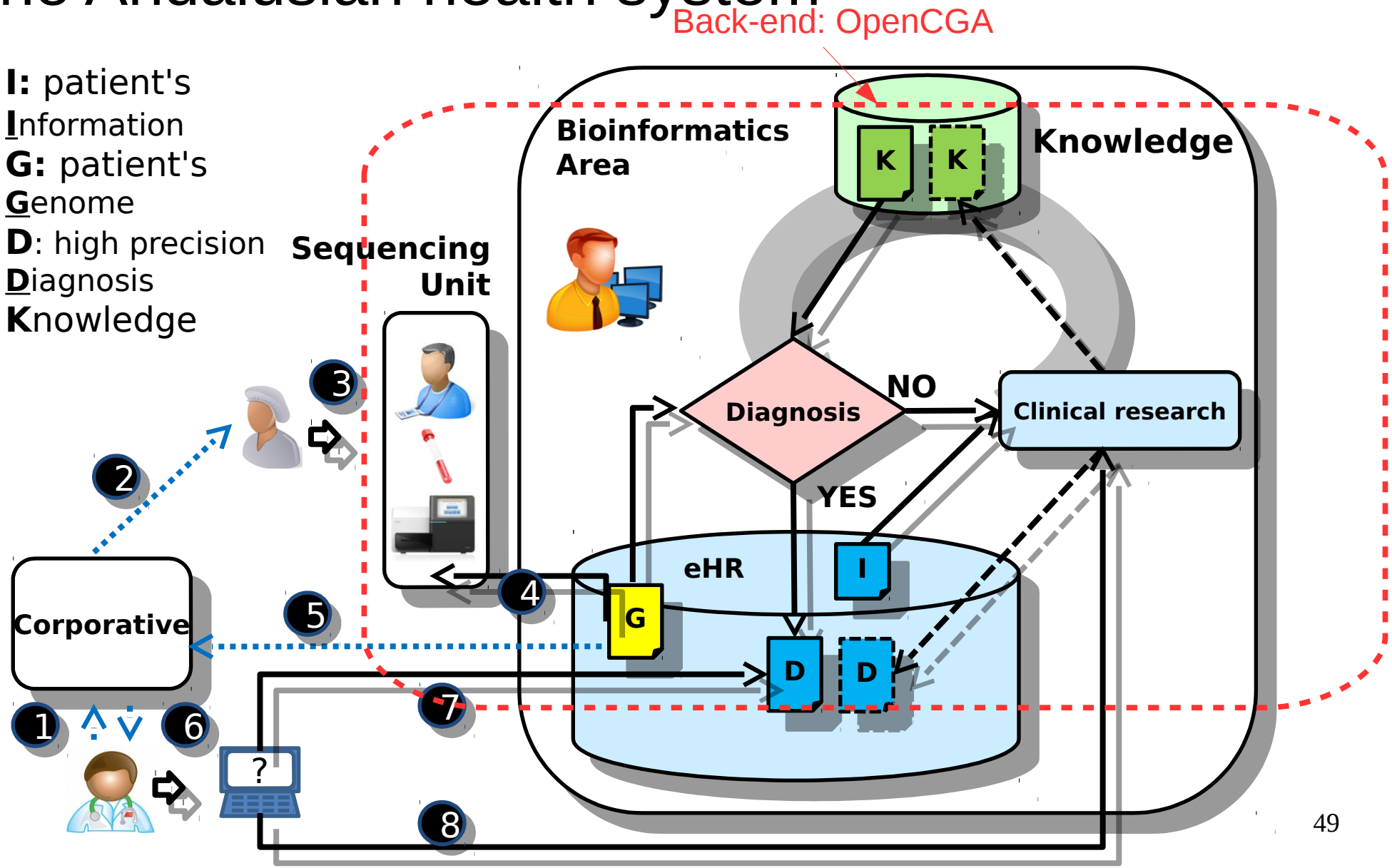
MMP for diagnosis and clinical research within the Andalusian health system

I: patient's **I**nformation
G: patient's **G**enome
D: high precision **D**iagnosis
Knowledge



Big data in Genomics

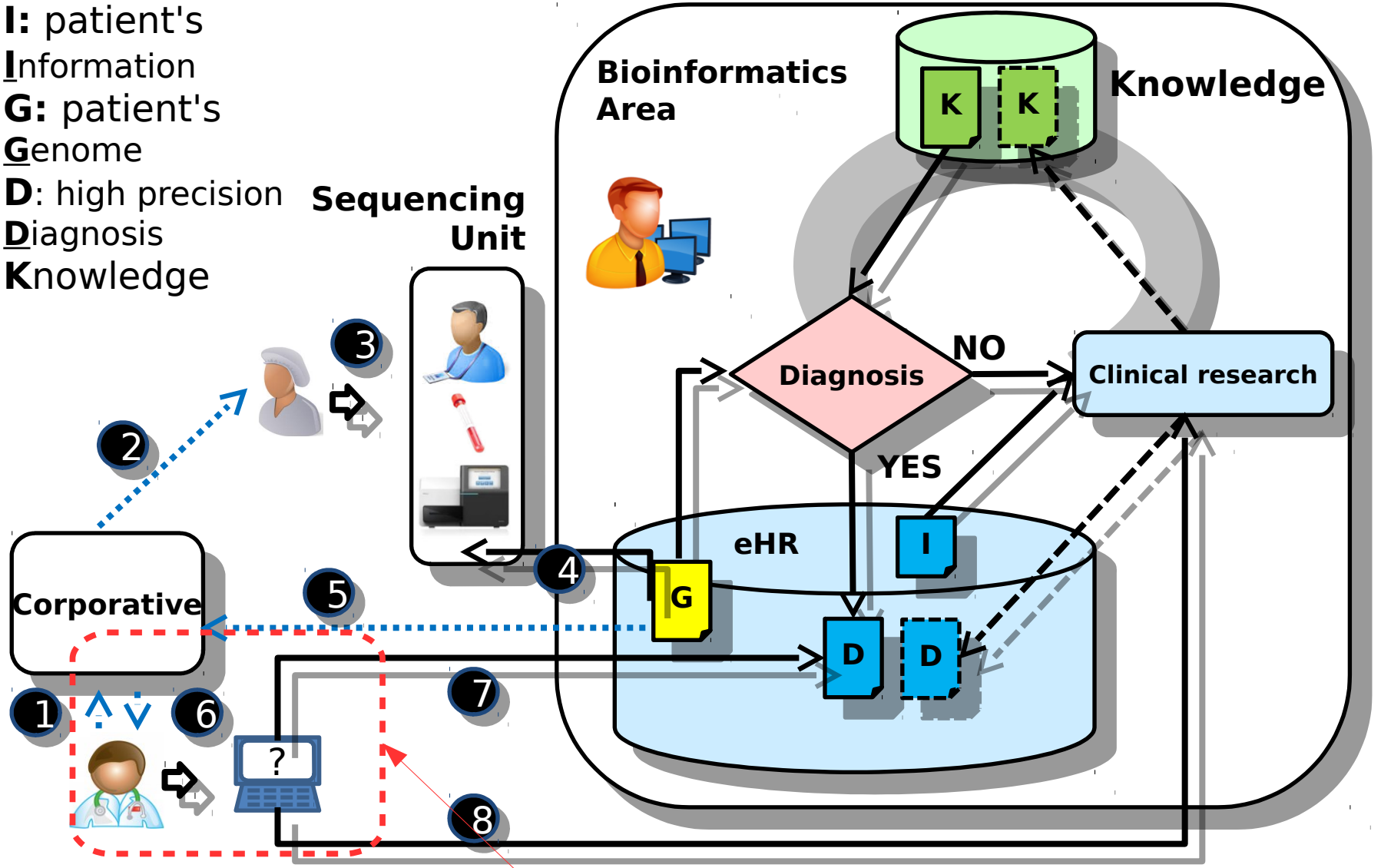
MMP for diagnosis and clinical research within the Andalusian health system



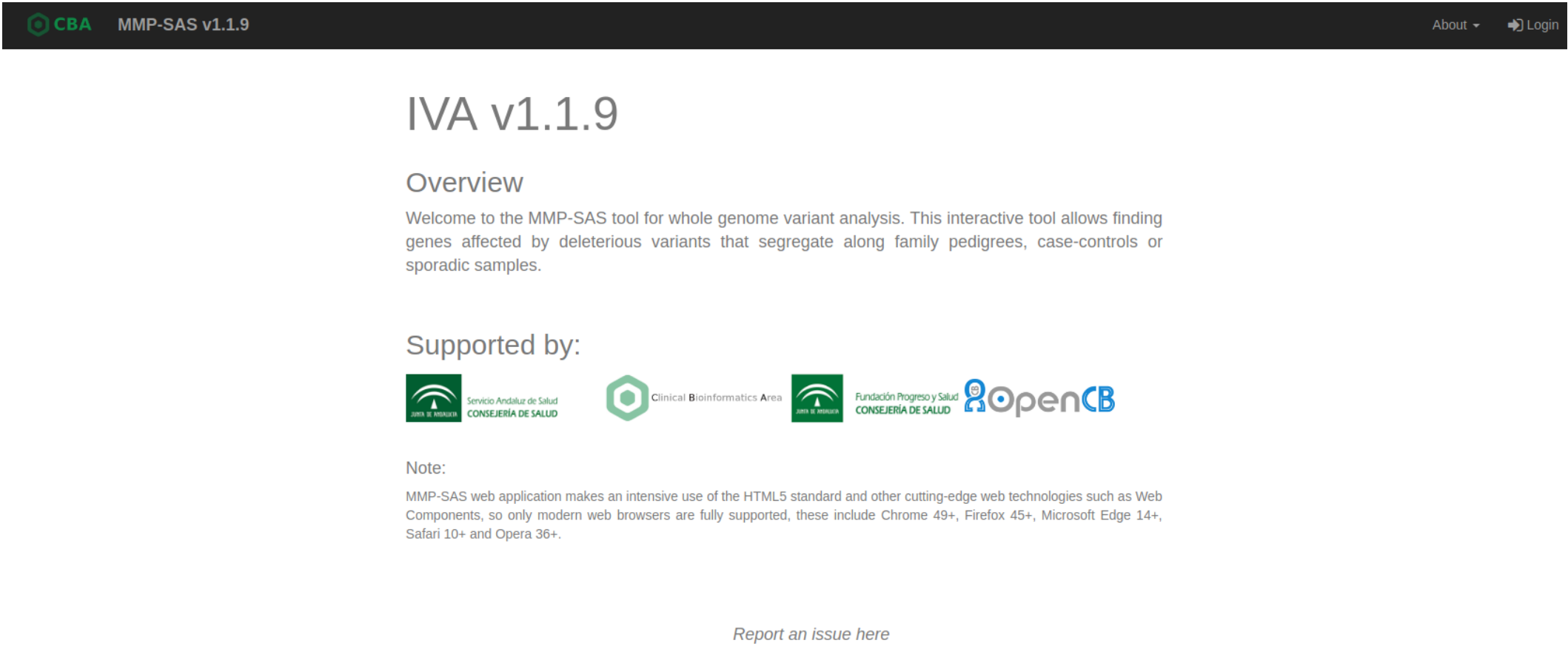
Big data in Genomics

MMP for diagnosis and clinical research within the Andalusian health system

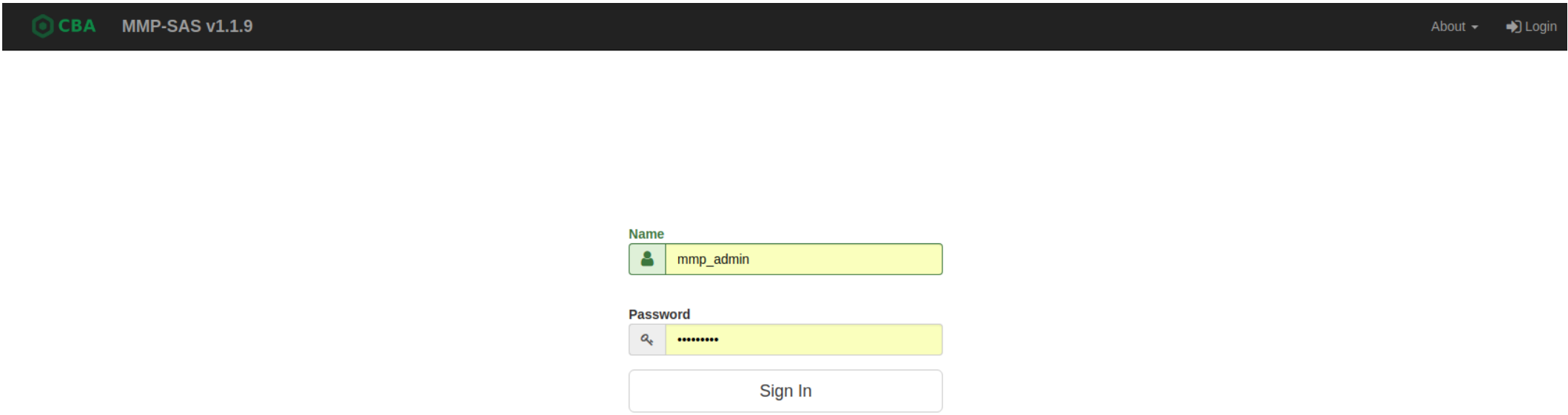
I: patient's **I**nformation
G: patient's **G**enome
D: high precision **D**iagnosis
Knowledge



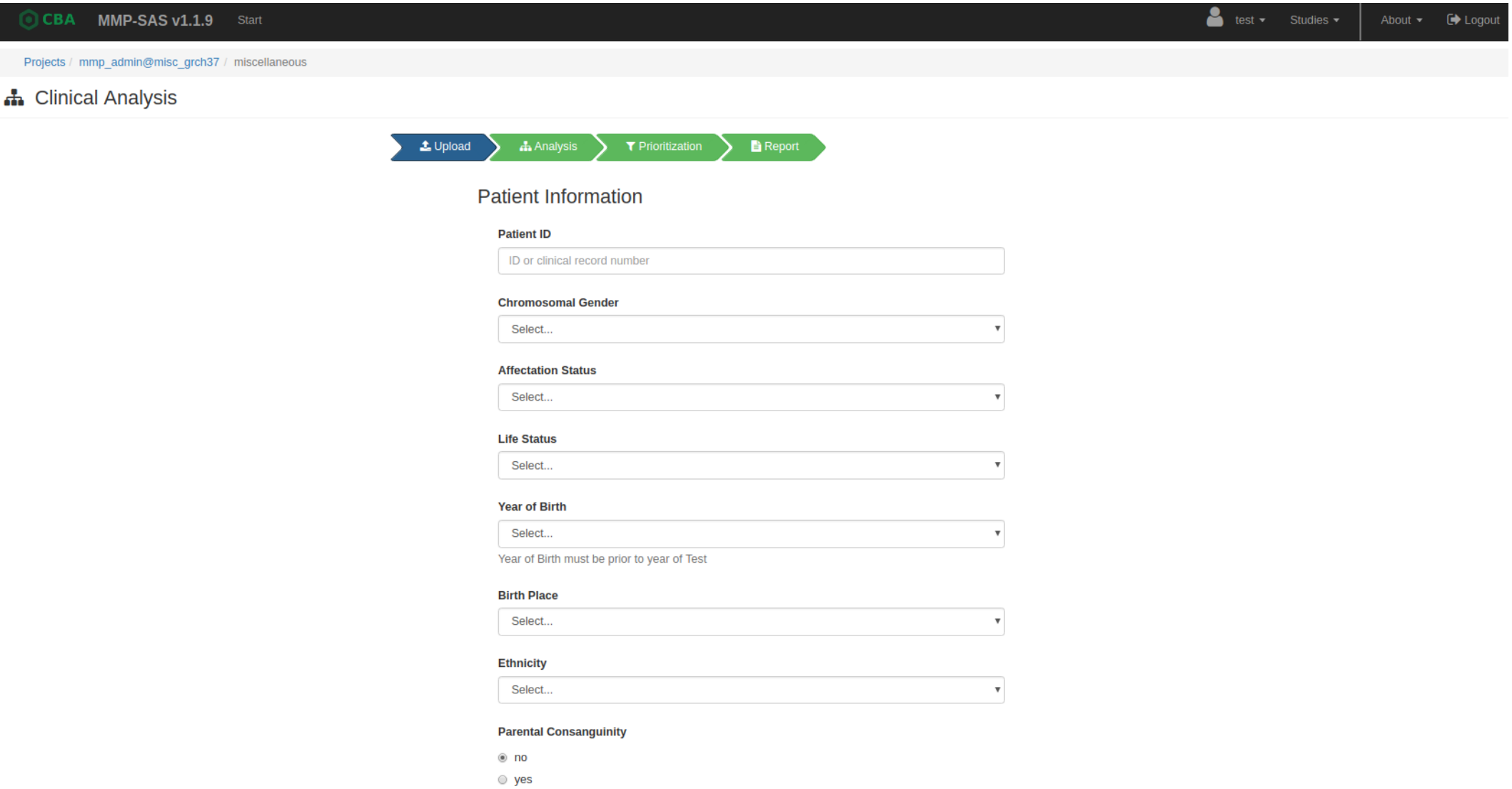
Front-end: MMP-SAS



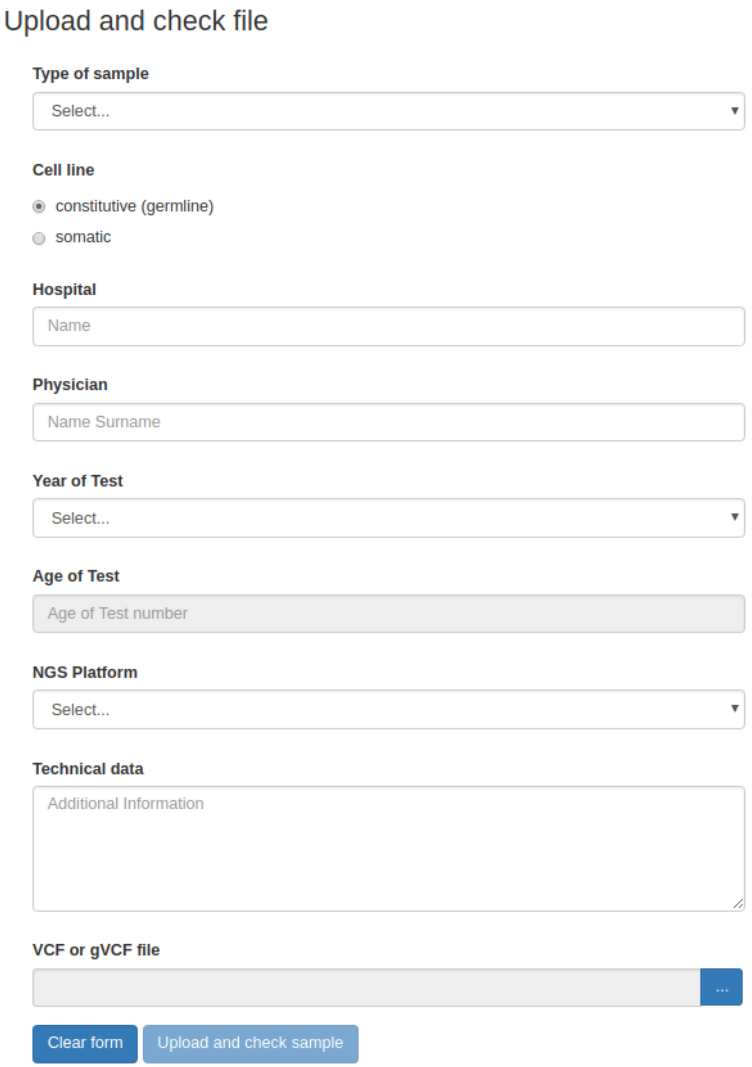
Front-end: MMP-SAS



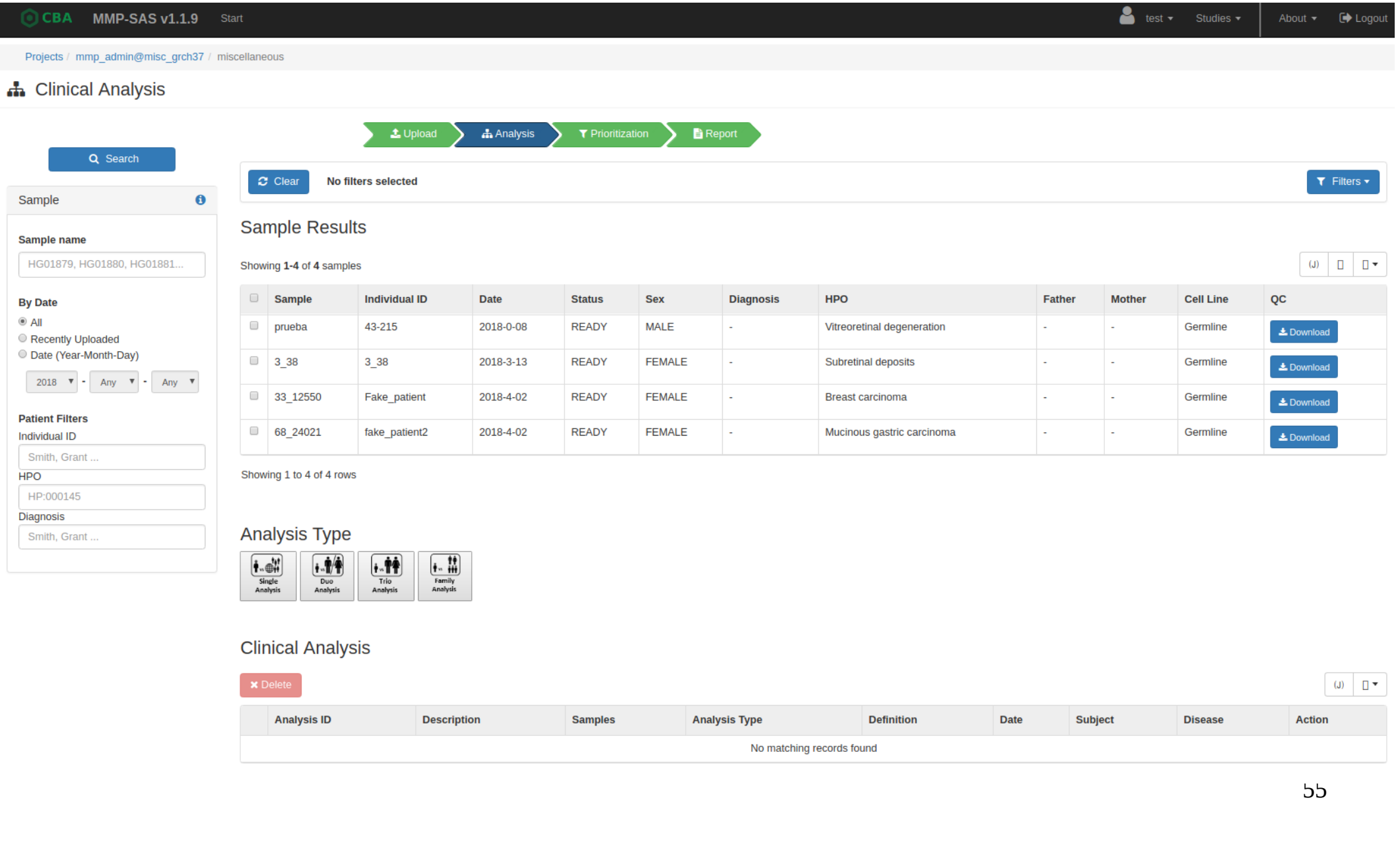
Front-end: MMP-SAS



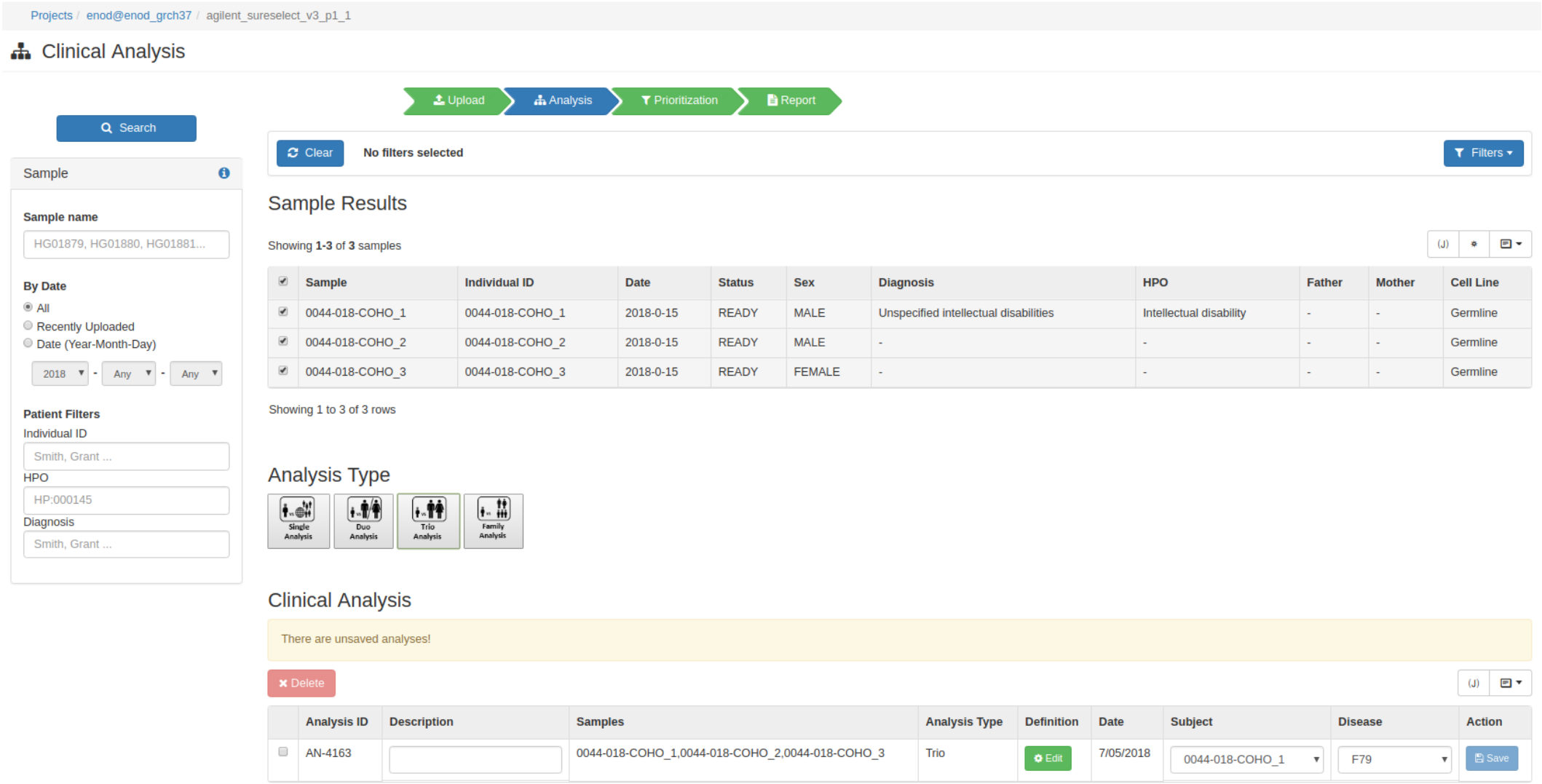
Front-end: MMP-SAS



Front-end: MMP-SAS



Front-end: MMP-SAS



Front-end: MMP-SAS

Family Editor

Name: FA-6031

Description:

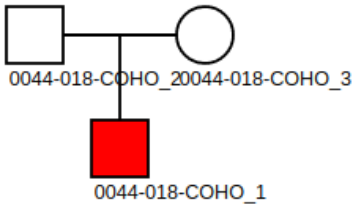
Diseases

⊞ Add custom disease

Unspecified intellectual disabilities(F79)

Intellectual disability(HP:0001249)

Sample ID	Father	Mother	Phenotypes	Deceased	Parental Consanguinity	Sex
0044-018-COHO_1 (0044-018-COHO_1)	0044-018-COHO_2	0044-018-COHO_3	F79 HP:0001249			MALE
0044-018-COHO_2 (0044-018-COHO_2)	Select a sample...	Select a sample...	F79 HP:0001249			MALE
0044-018-COHO_3 (0044-018-COHO_3)	Select a sample...	Select a sample...	F79 HP:0001249			FEMALE



Cancel OK

Front-end: MMP-SAS

Projects / enod@enod_grch37 / agilent_sureselect_v3_p1_1

Clinical Analysis

Upload | Analysis | Prioritization | Report

Clinical Analysis AN-3622

Select Recent Analysis:

AN-3622

Select analysis: Browse...

AN-3622 Clinical Analysis Summary

Clinical Analysis

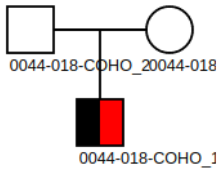
Name:	AN-3622
Disease:	Unspecified intellectual disabilities (F79)
Type:	TRIO
Date:	15 Jan 18

Sample

Name:	0044-018-COHO_1
Somatic:	false
Phenotypes:	Unspecified intellectual disabilities (F79) Intellectual disability (HP:0001249)

Subject

Name:	0044-018-COHO_1
Sex (karyotype):	MALE (XY)
Date of Birth:	1920 (UNKNOWN)
Pedigree:	



Phenotypes:	Unspecified intellectual disabilities (F79) Intellectual disability (HP:0001249)

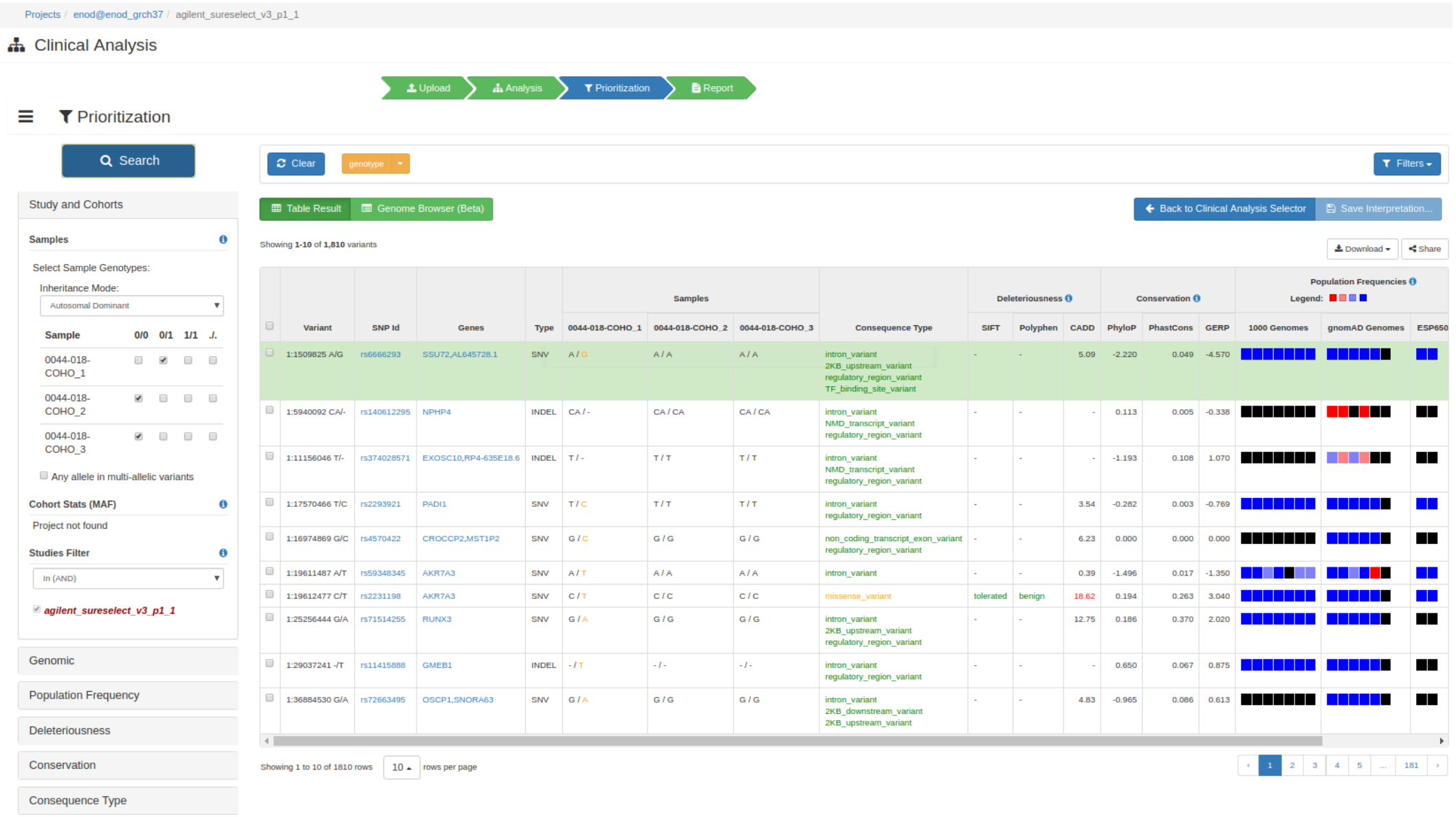
Interpretation Algorithms

Several prioritization and interpretation algorithms are available, you can choose an **interactive** tool:

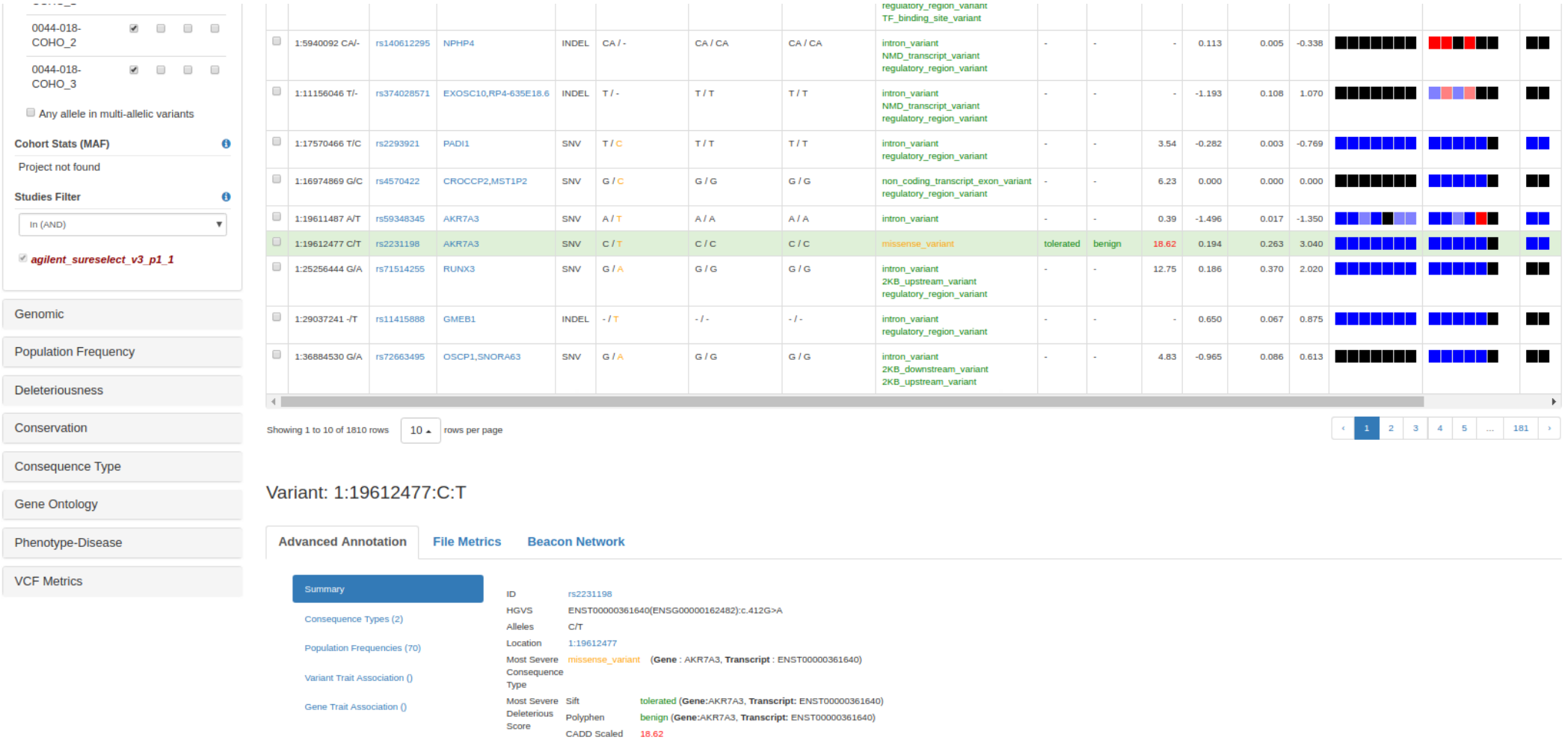
Interactive Prioritization (based on **TEAM** paper)

OK

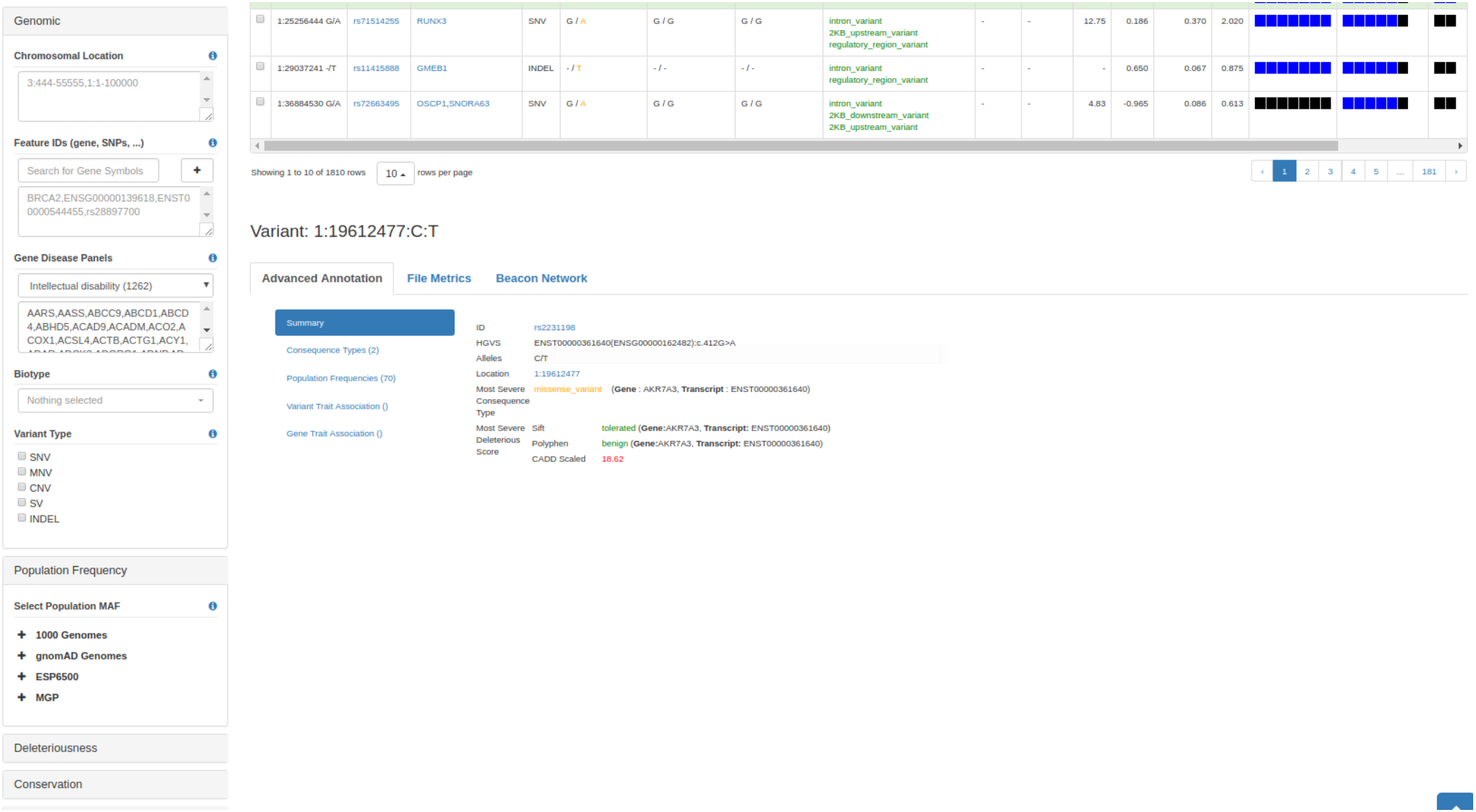
Front-end: MMP-SAS



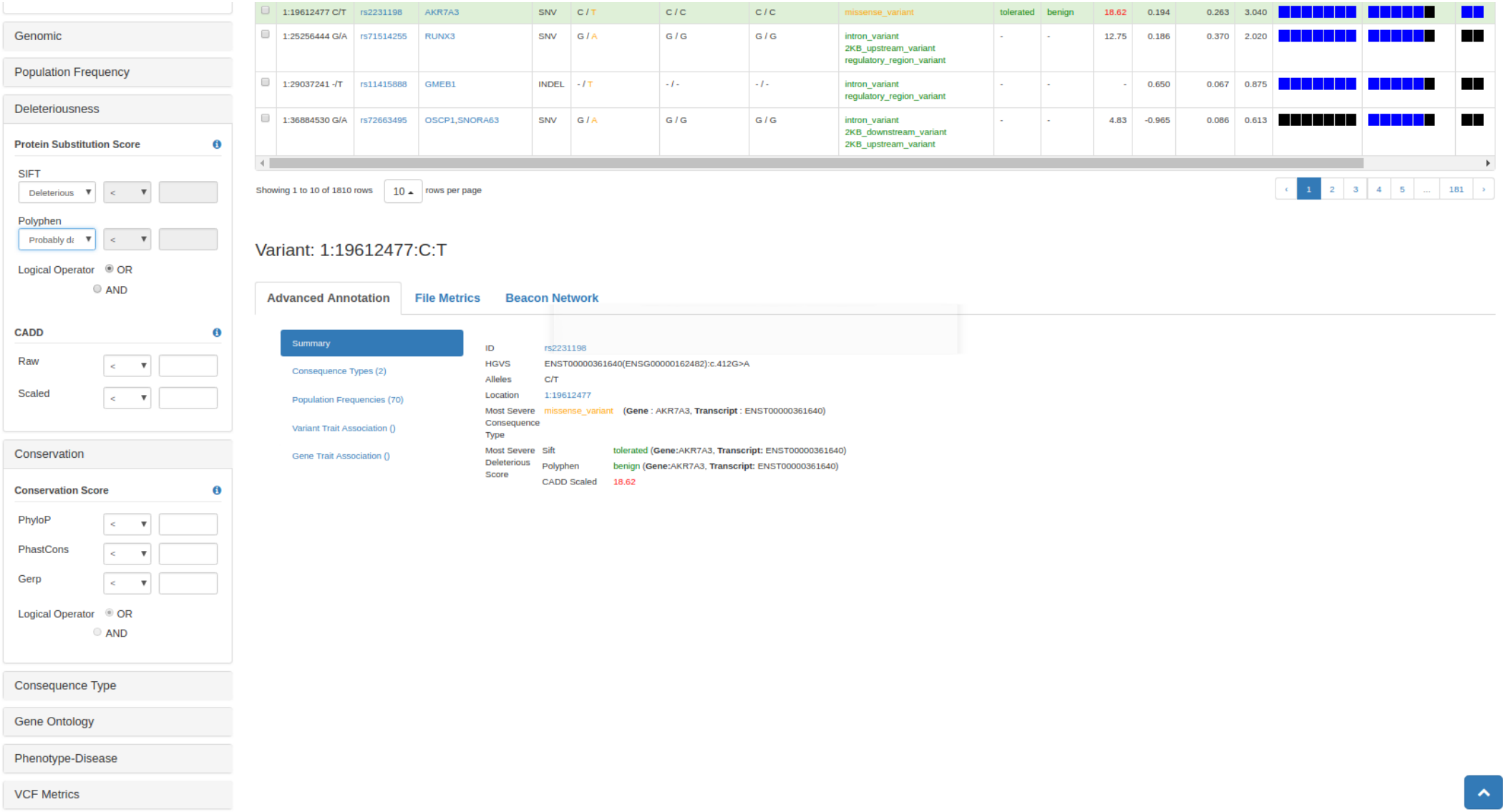
Front-end: MMP-SAS



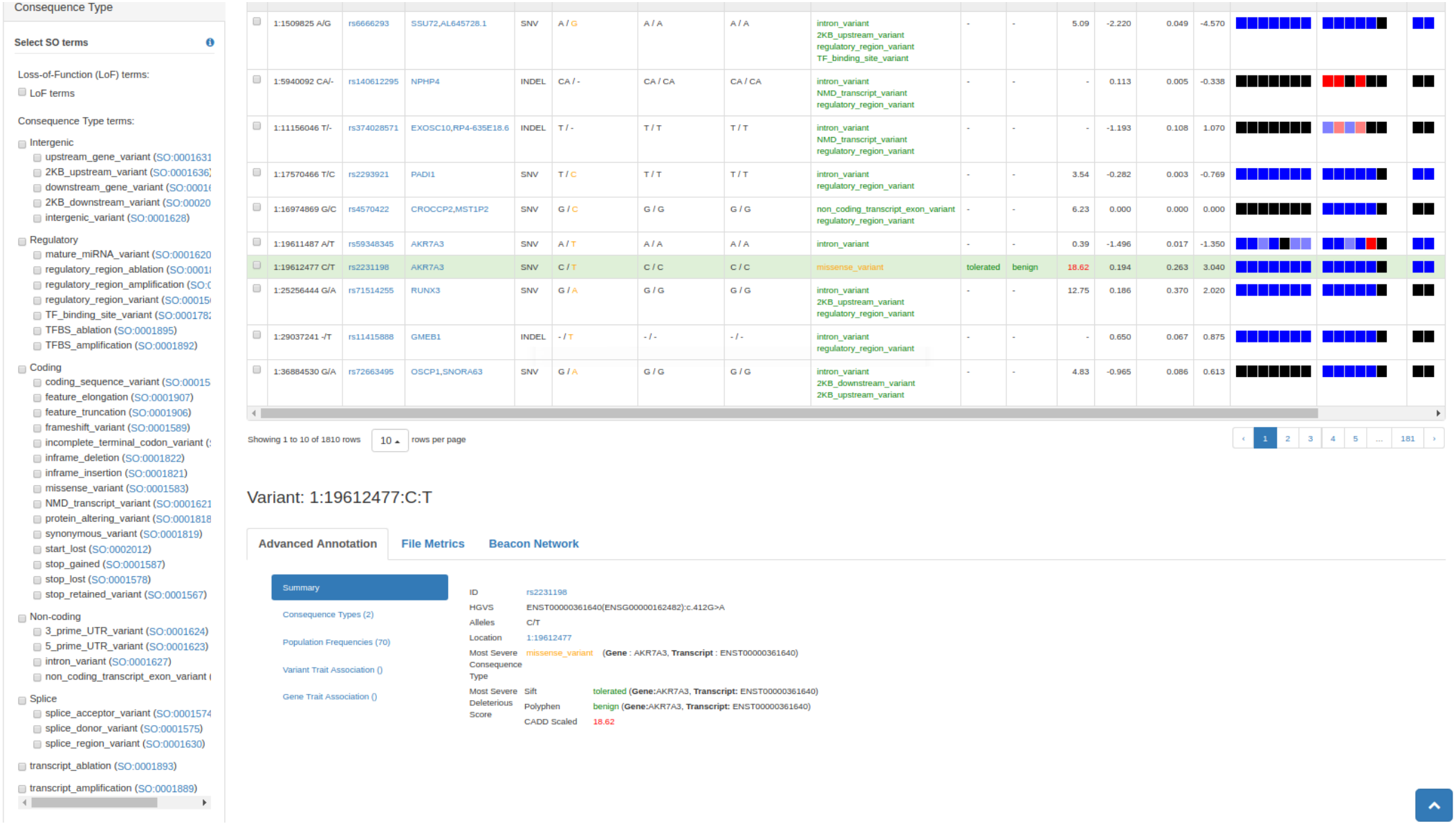
Front-end: MMP-SAS



Front-end: MMP-SAS

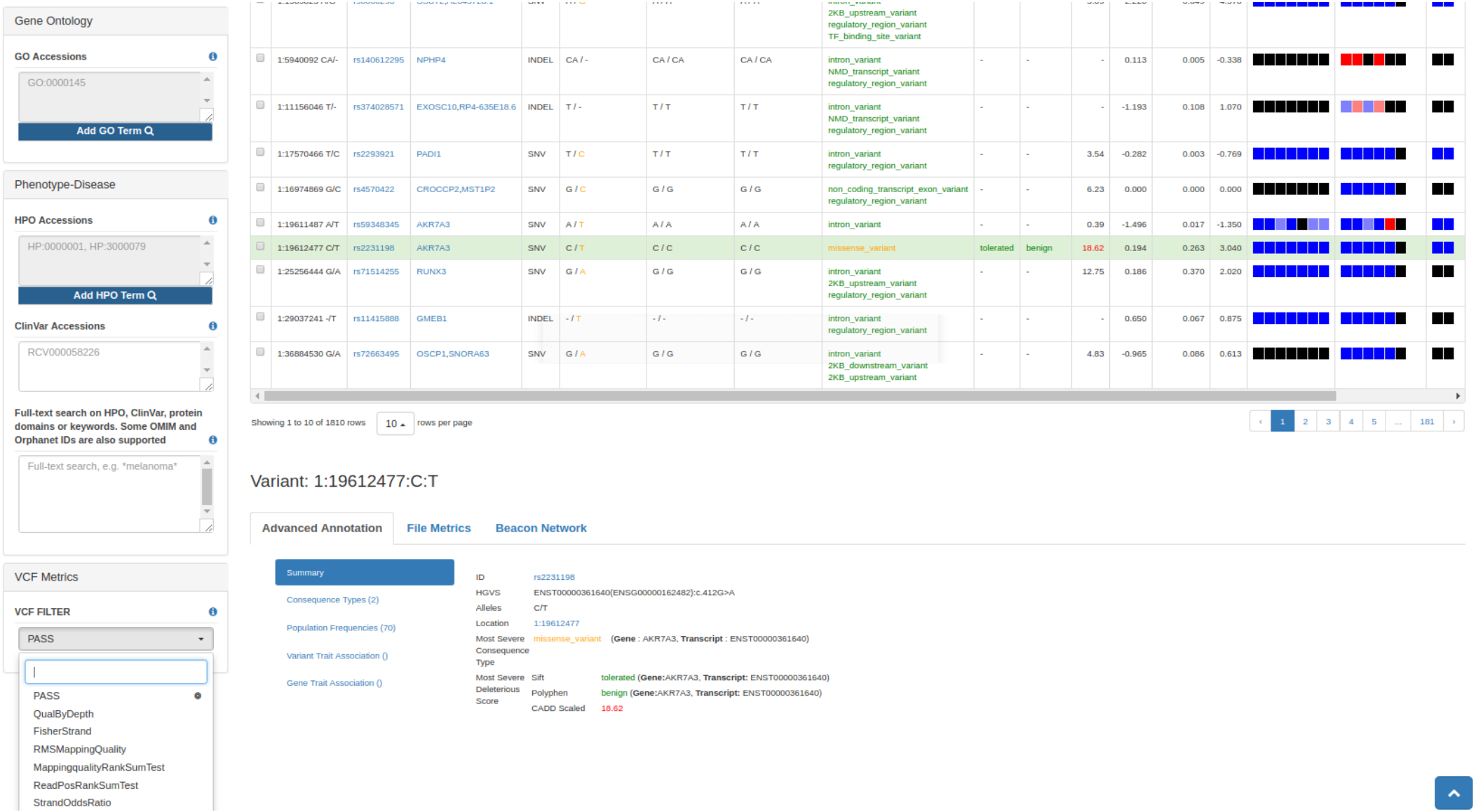


Front-end: MMP-SAS

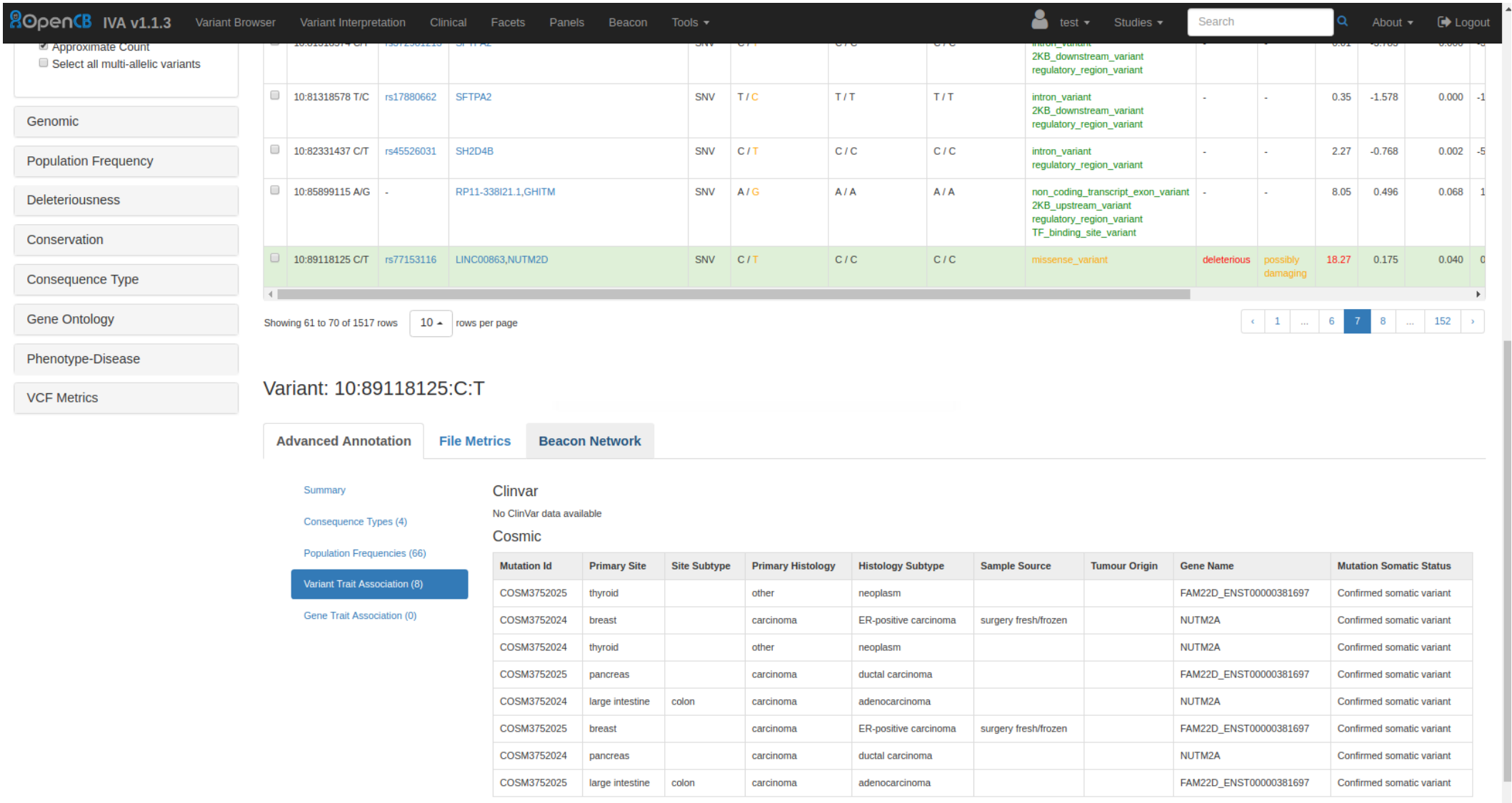


Consequence type: the location or effect of a sequence variation on a transcript

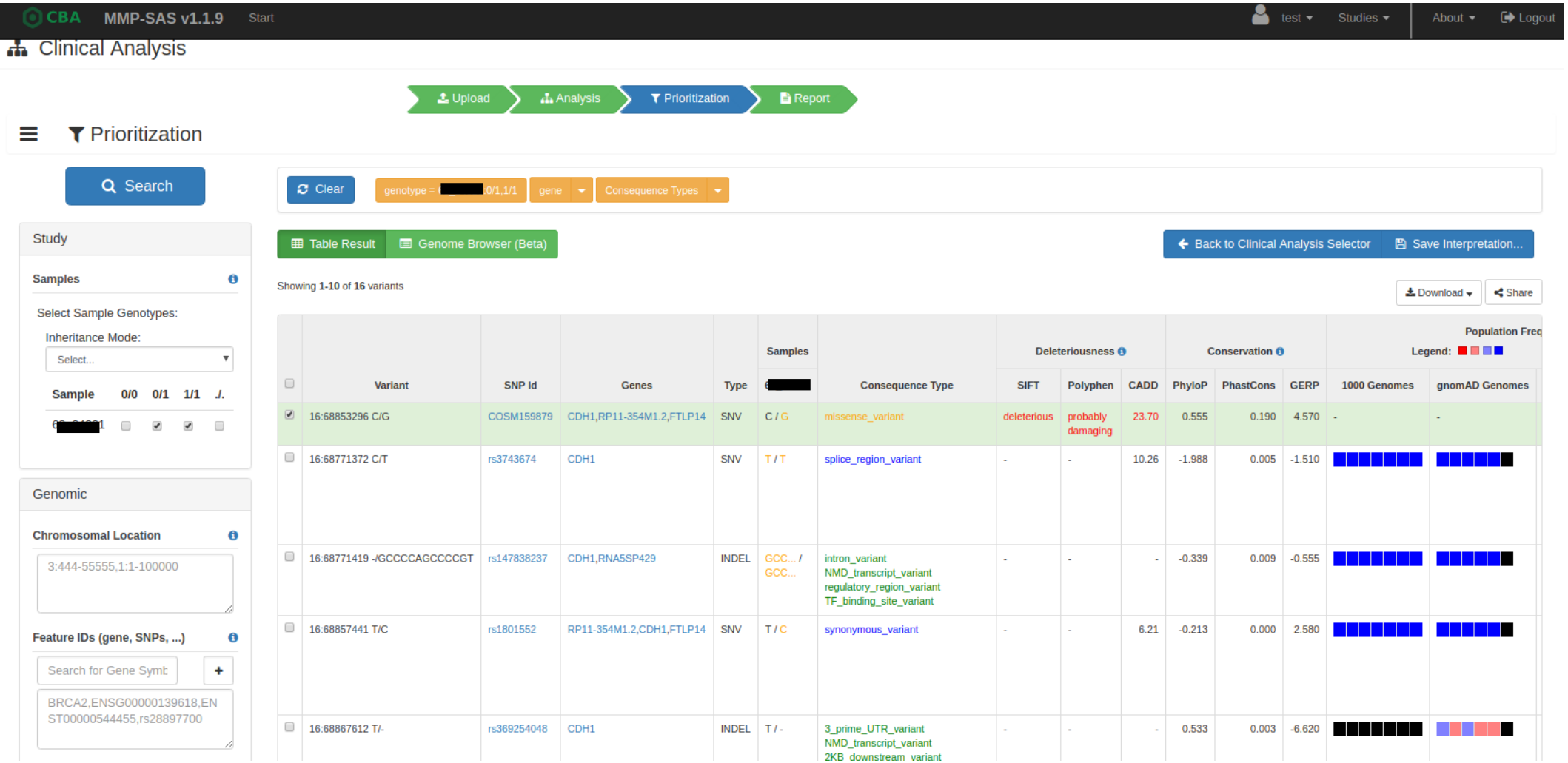
Front-end: MMP-SAS



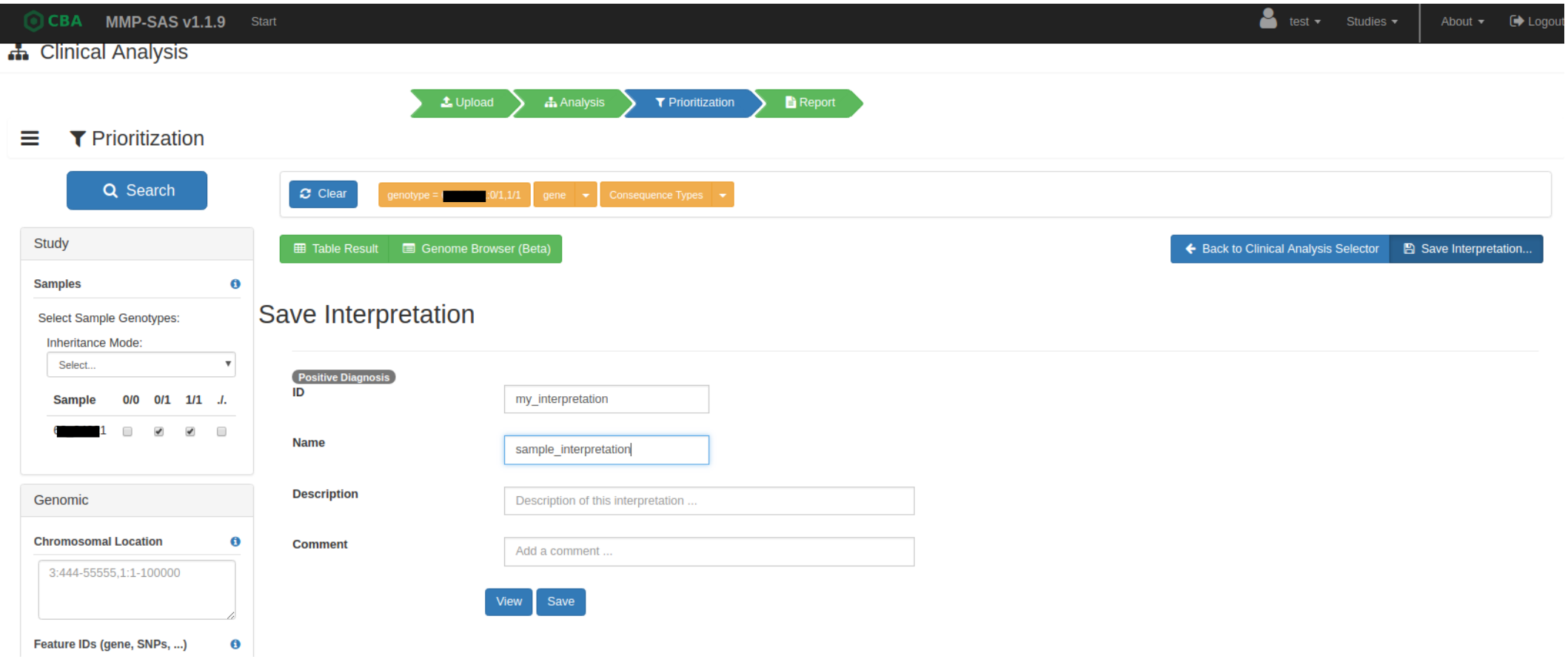
Front-end: MMP-SAS



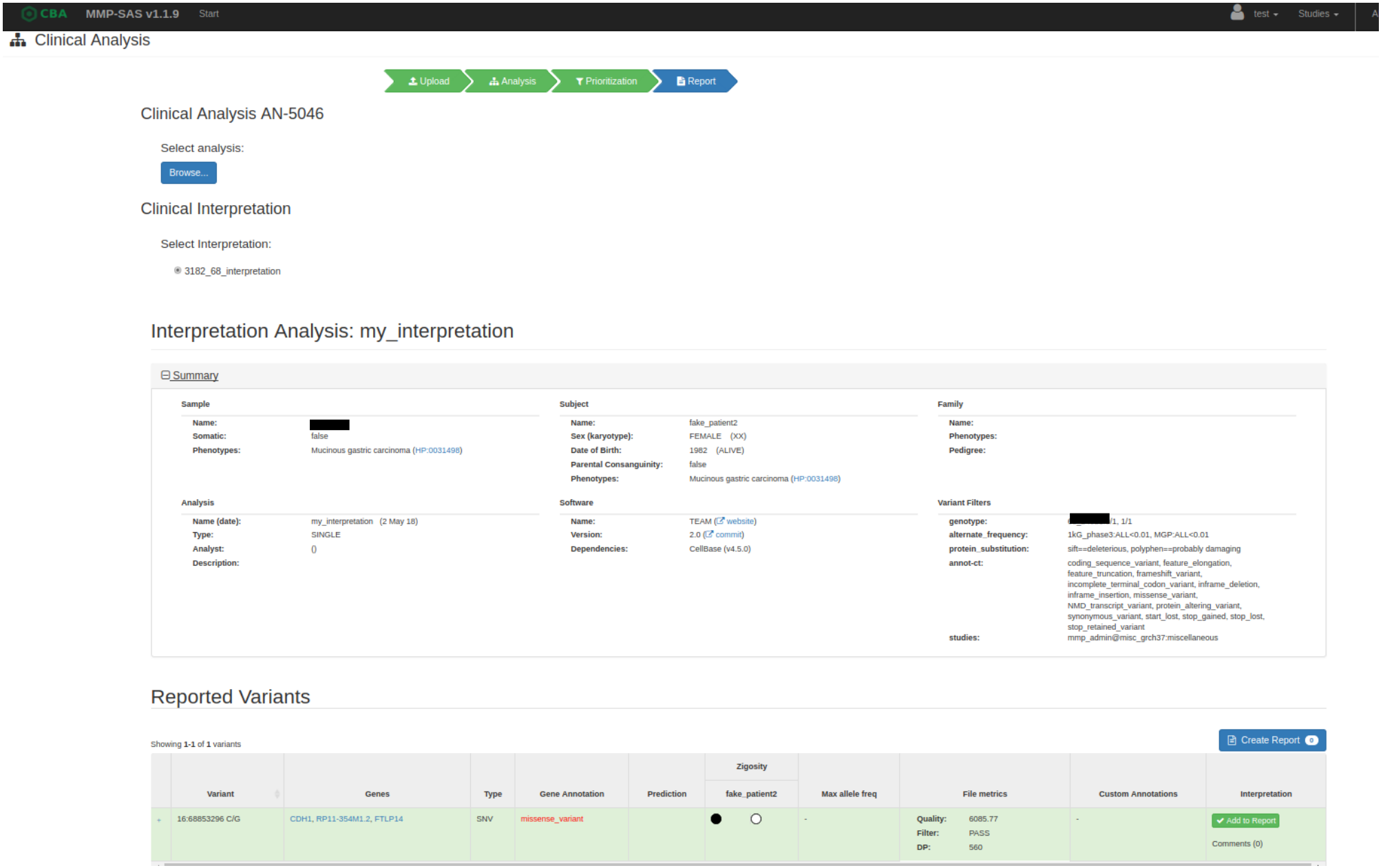
Front-end: MMP-SAS



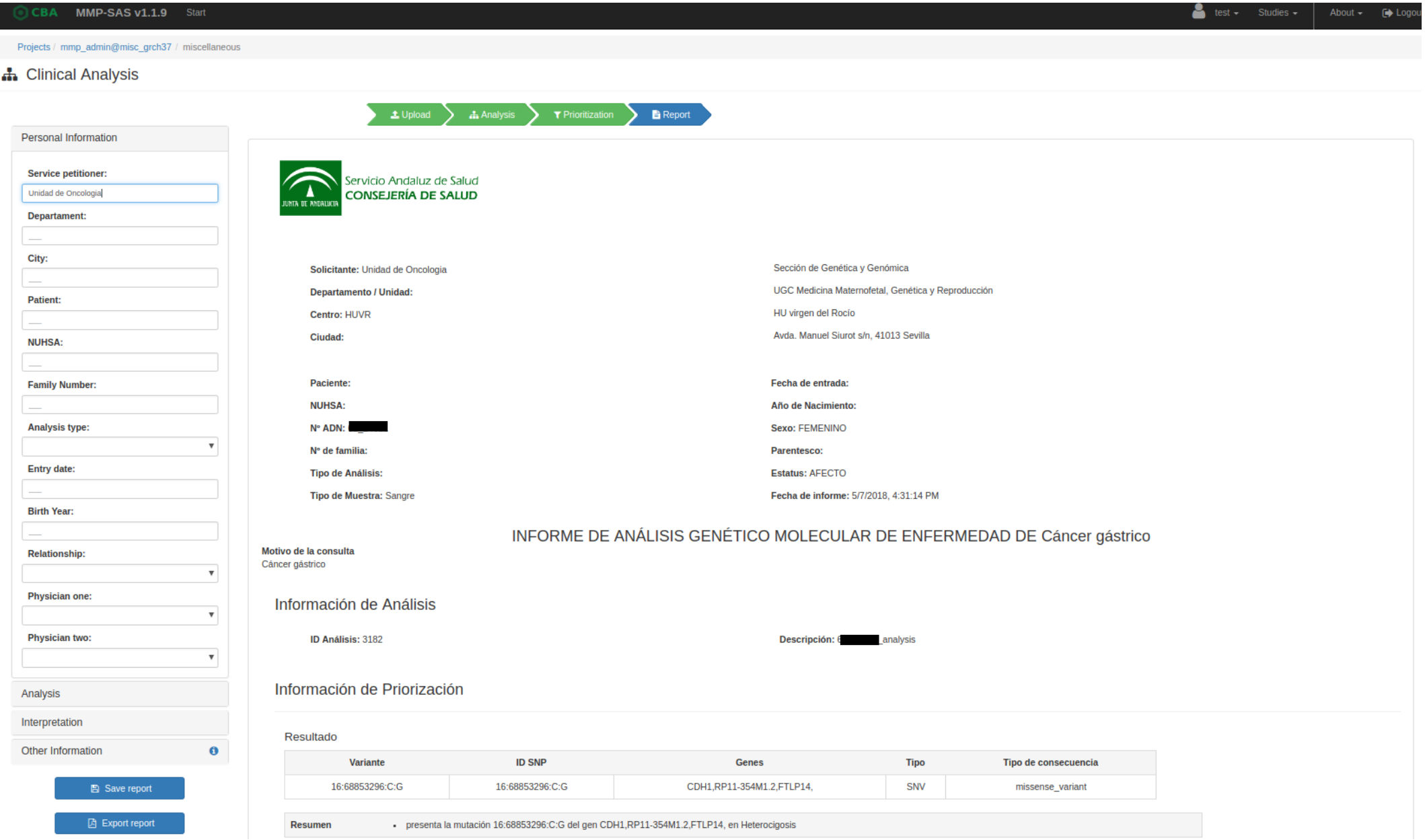
Front-end: MMP-SAS



Front-end: MMP-SAS

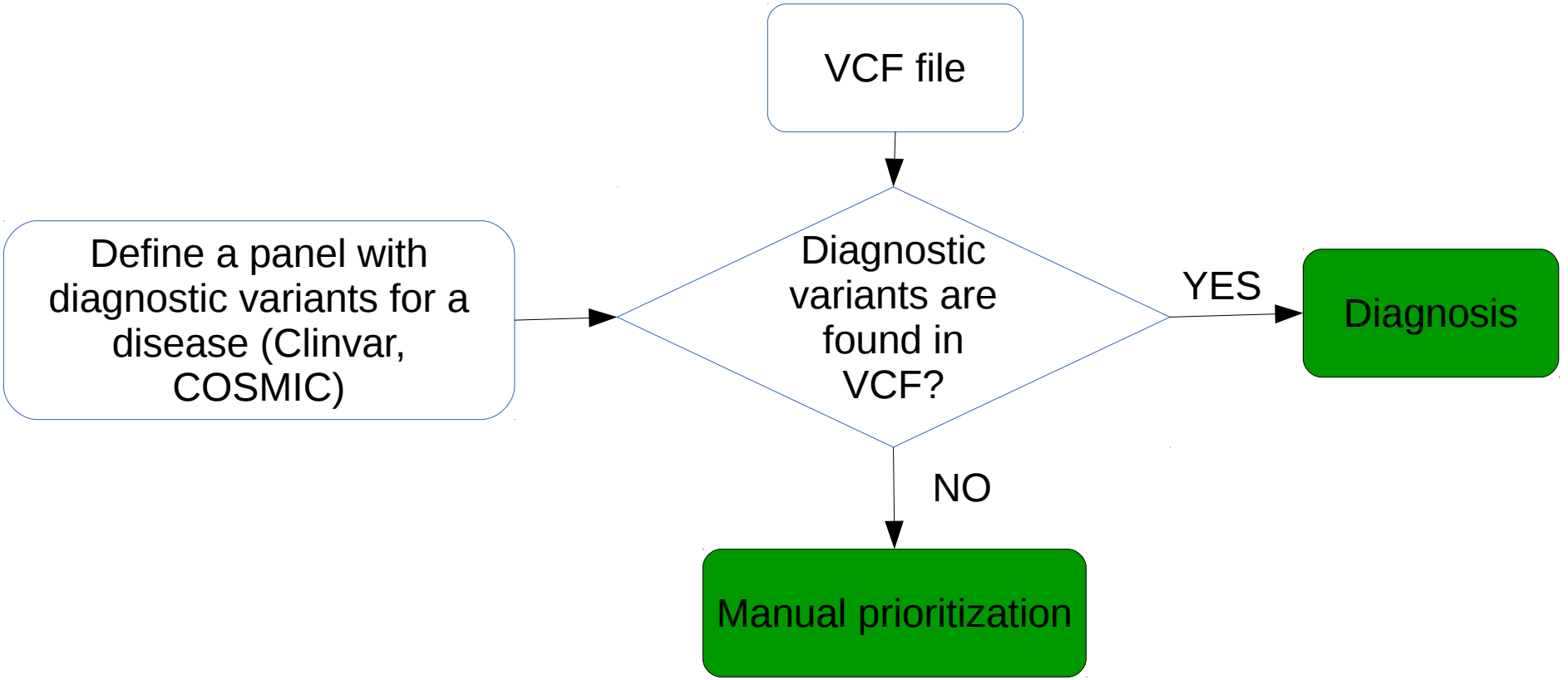


Front-end: MMP-SAS



IVA dev version (available soon)

- Gene panel component
  - Design and management of panels of genes
  - Goal: Diagnostic purposes



IVA dev version (available soon)

- Gene panel component
  - Essential for defining (informed consent is necessary):
    - **Pharmacogenetic variants**: variability in genes implicated in drug response can modulate treatment efficacy or predispose to adverse drug reactions
    - **Secondary findings:** reporting of known pathogenic or expected pathogenic variants in the 56 ACMG genes even when unrelated to the primary medical reason for testing

ACMG STATEMENT | Genetics inMedicine

Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics

Sarah S. Kalia, ScM[1], Kathy Adelman[2],
Christine Eng, MD[6], James P. Evans, MD, PhD
Teri E. Klein, PhD[10], Bruce R. Korf, MD, PhD
C. Sue Richards, PhD[14], Christopher N. Vlango
David T. Miller, MD, PhD[18]; on behalf of the A

Genetics inMedicine | LETTER TO THE EDITOR

ACMG secondary findings 2.0

To the Editor: The authors of the recent American College of Medical Genetics and Genomics secondary findings

leads to an inconsistency in that many EP novel predicted loss of function variants would score as "likely pathogenic" in the new pathogenicity criteria, while it is unclear whether expected pathogenic missense variants should be reported as secondary findings. The second issue is that an explicit decision should be made as to the pathogenicity threshold

IVA dev version (available soon)

- Variant classification according to ACMG Standards:
  - Pathogenic, likely pathogenic, benign, likely benign, uncertain significance

ACMG STANDARDS AND GUIDELINES | Genetics inMedicine

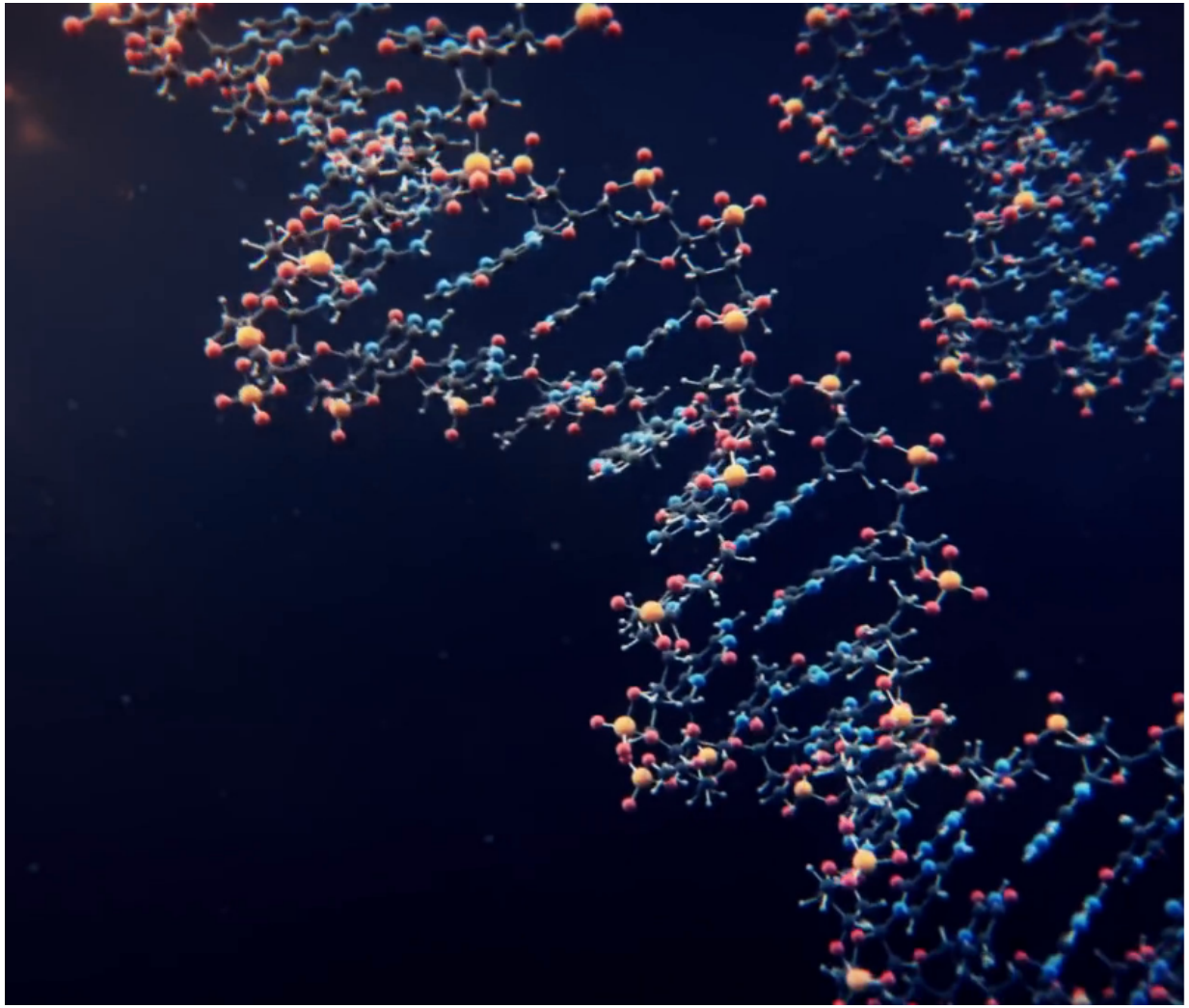
Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

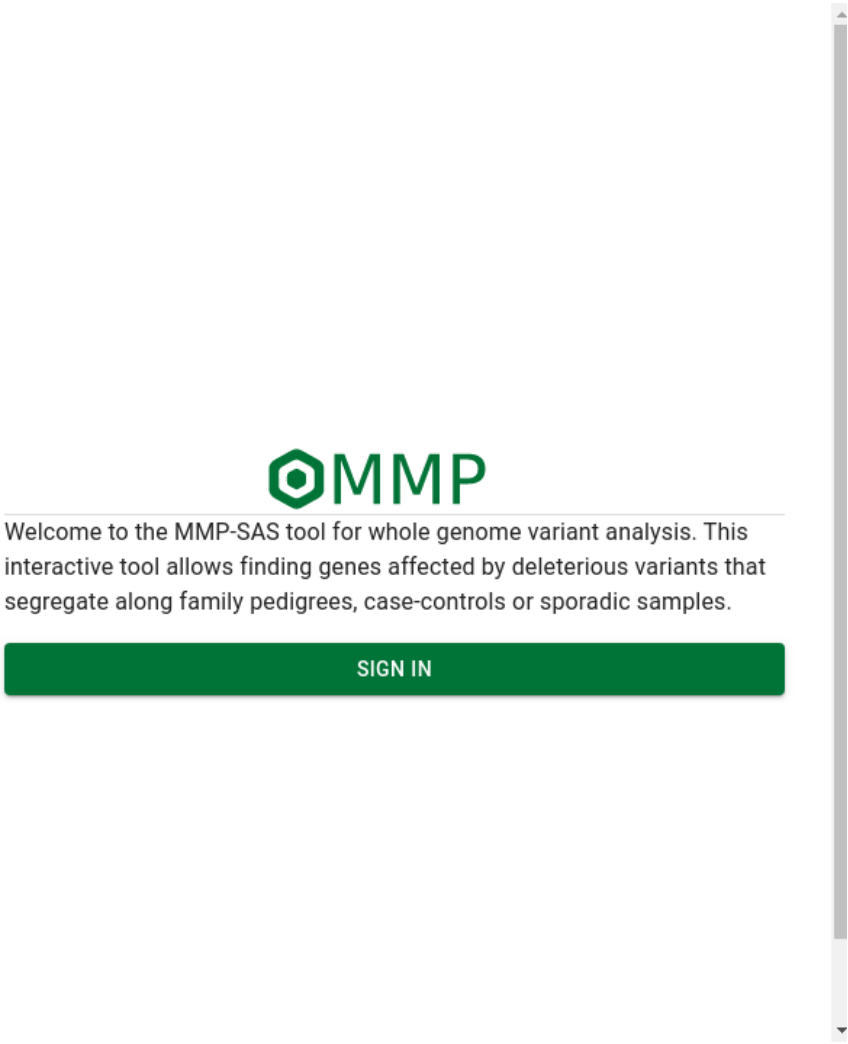
Sue Richards, PhD[1], Nazneen Aziz, PhD[2,16], Sherri Bale, PhD[3], David Bick, MD[4], Soma Das, PhD[5], Julie Gastier-Foster, PhD[6,7,8], Wayne W. Grody, MD, PhD[9,10,11], Madhuri Hegde, PhD[12], Elaine Lyon, PhD[13], Elaine Spector, PhD[14], Karl Voelkerding, MD[13] and Heidi L. Rehm, PhD[15]; on behalf of the ACMG Laboratory Quality Assurance Committee

IVA dev version (available soon)

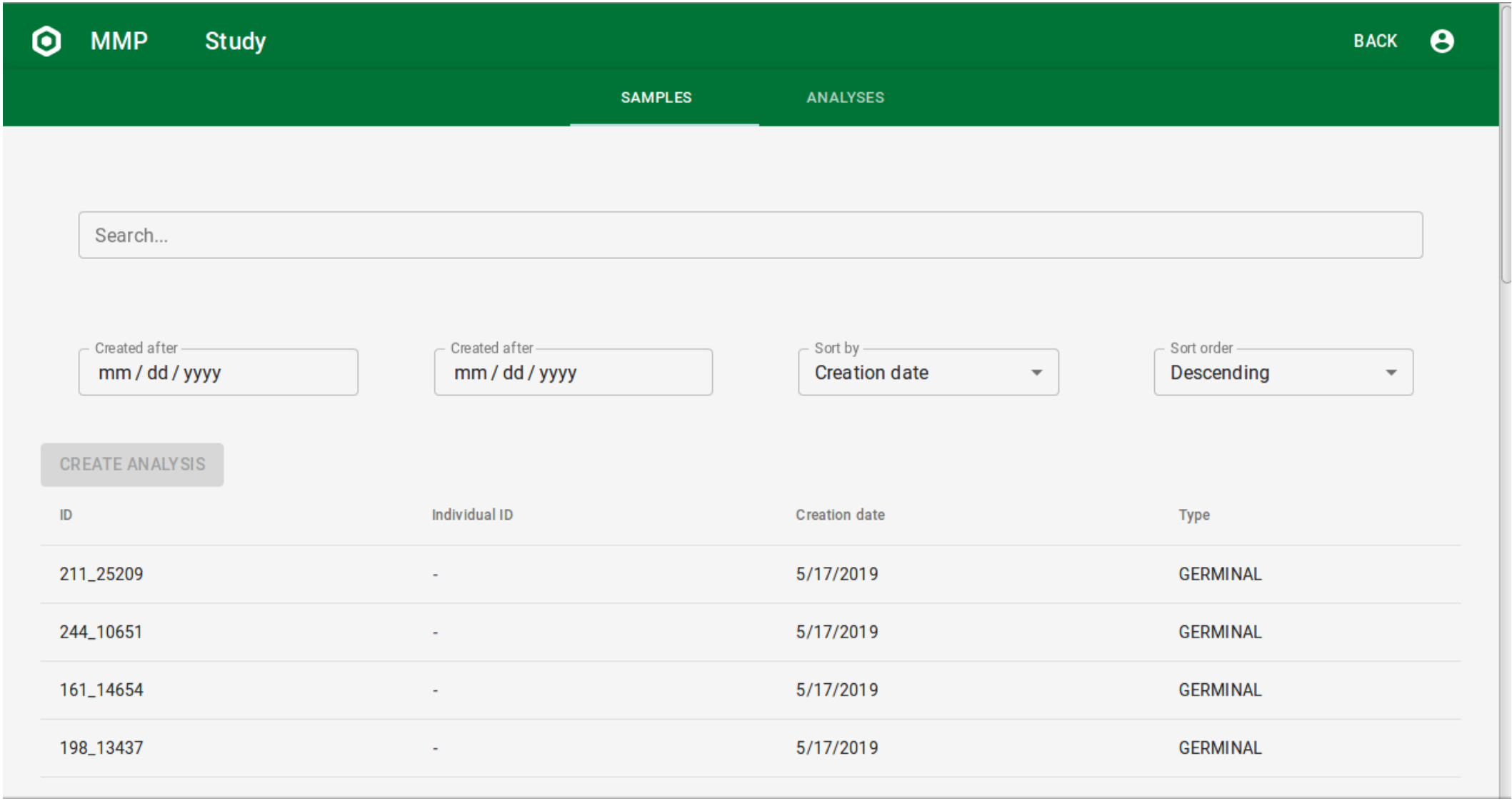
- Cohort information
  - Existence of a given variant in a cohort
- Federation of applications
  - A single installation can query other installations for variants, additional information, stats, etc (more powerful than Beacon)
- User annotation of variants:
    - Technological artifacts
    - Novel and diagnostic variants
    - …..
    -
    -

IVA dev version (available soon)

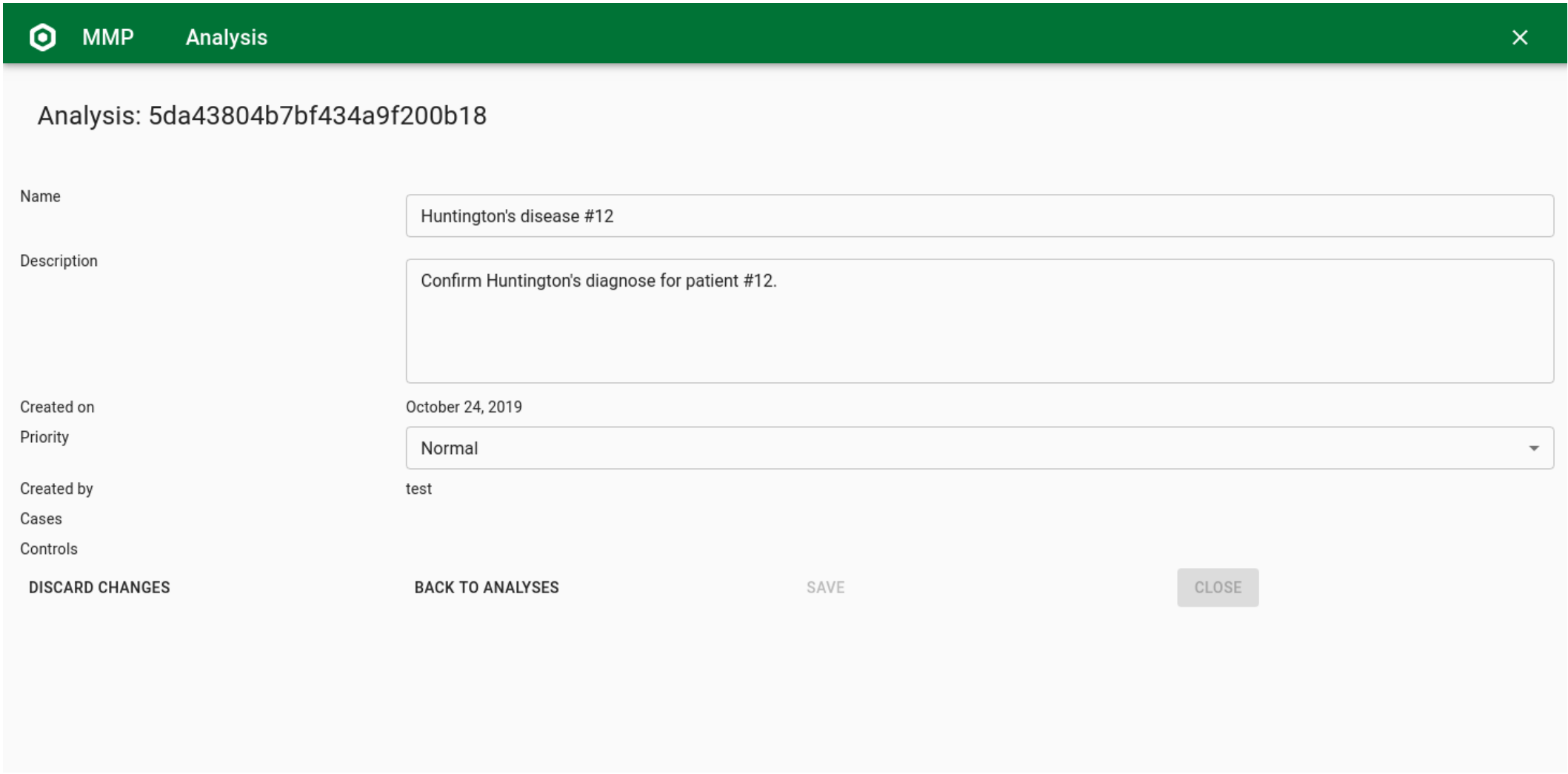




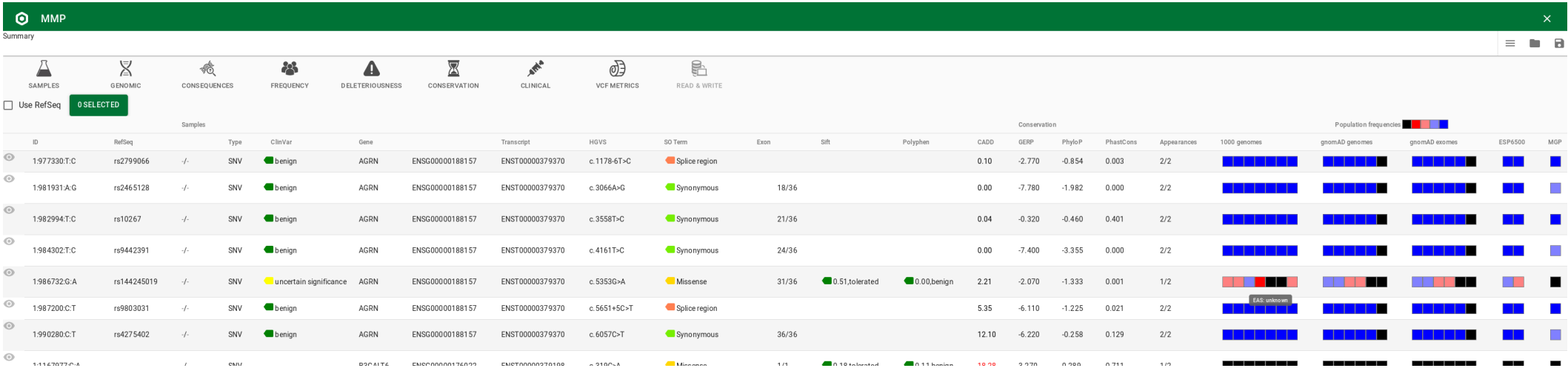
IVA dev version (available soon)



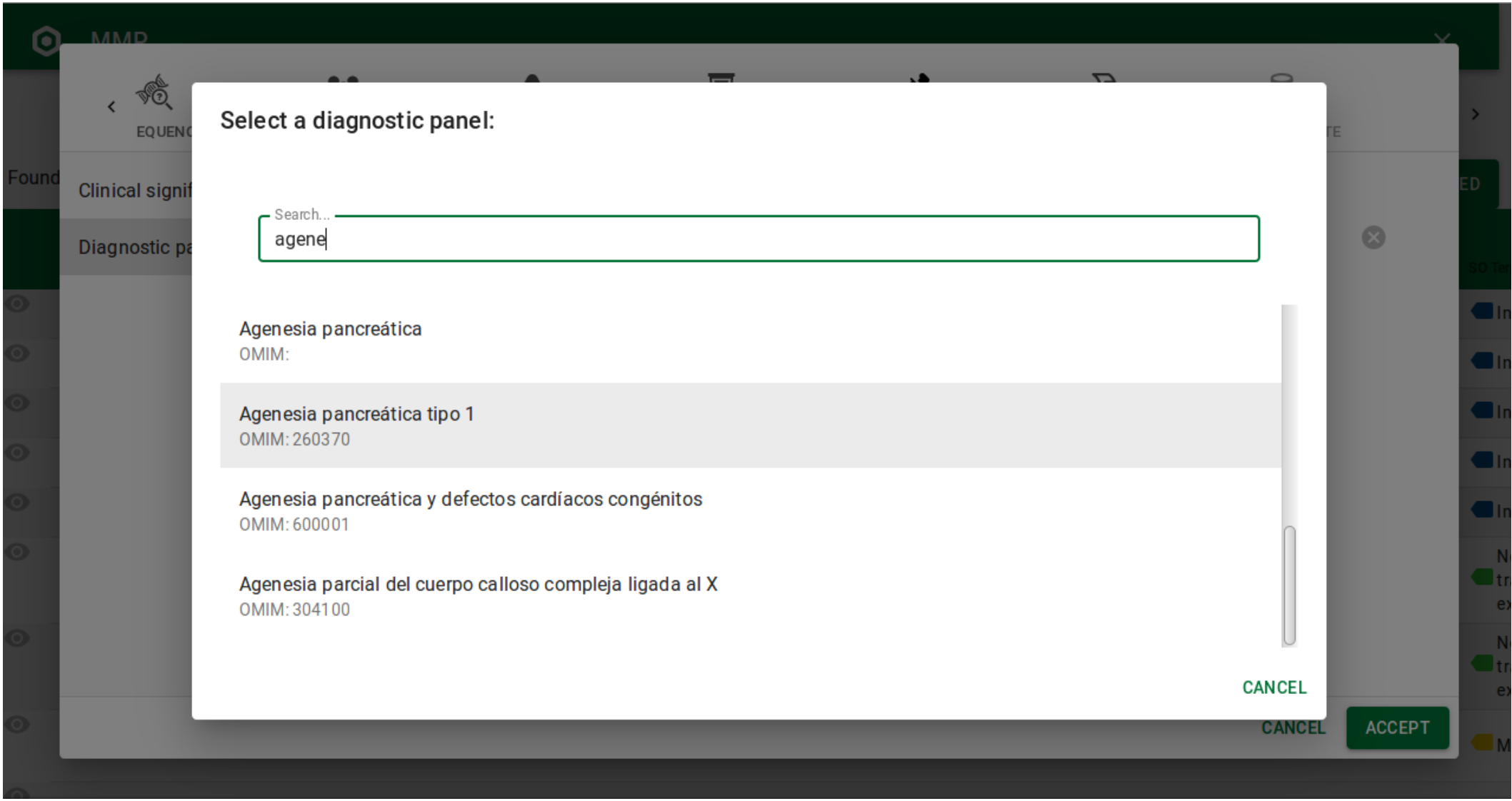
IVA dev version (available soon)



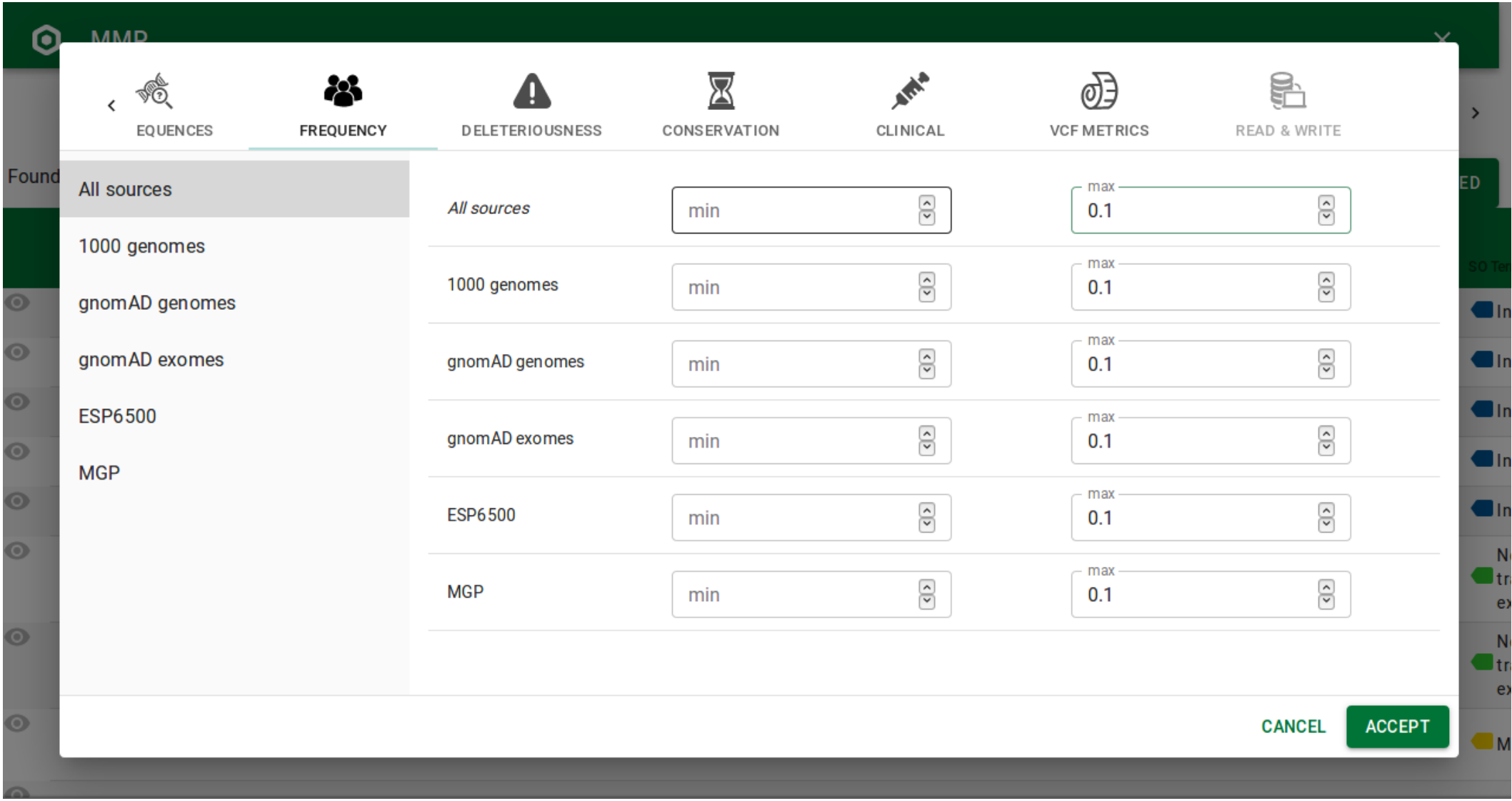
IVA dev version (available soon)



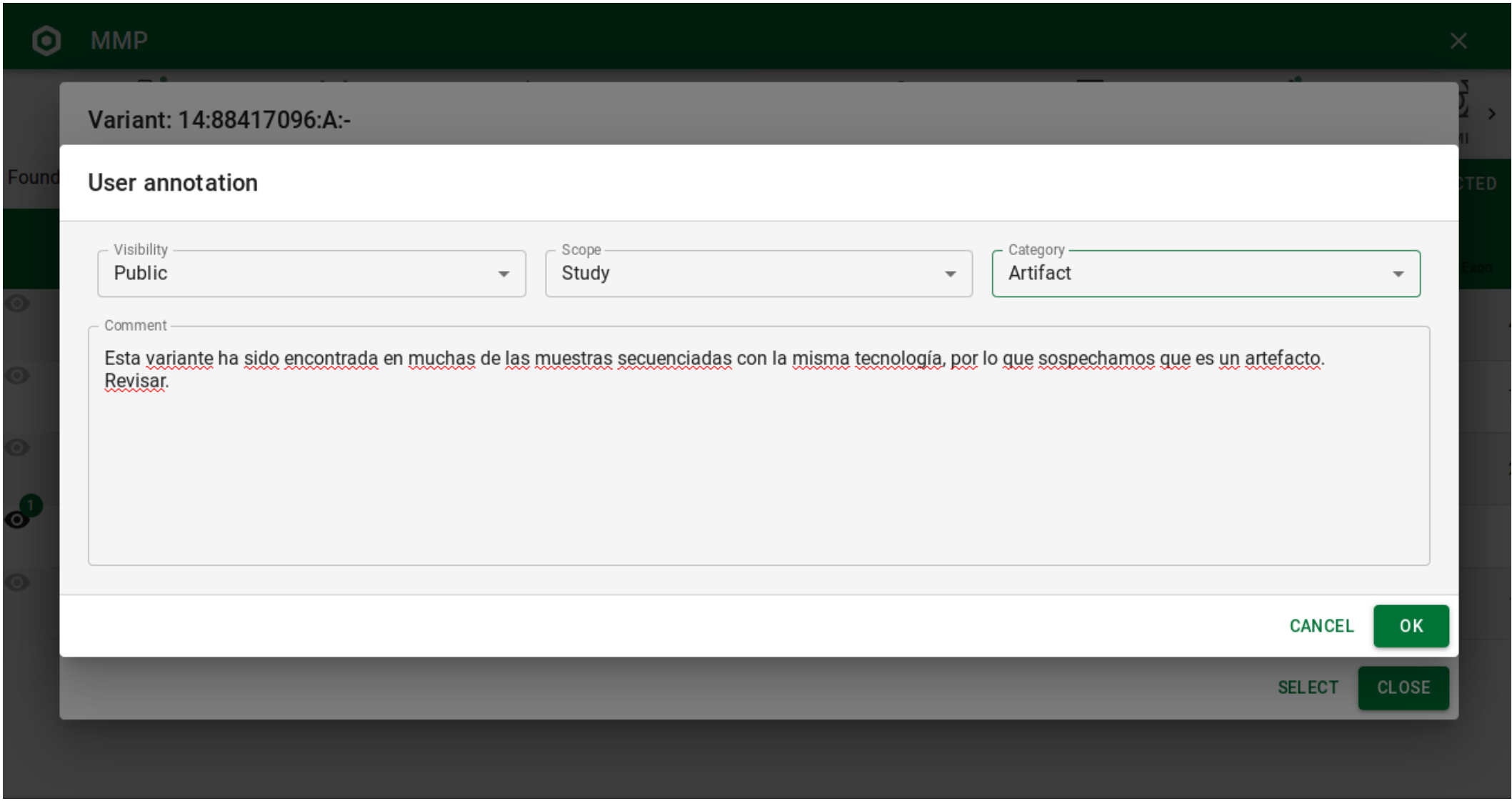
IVA dev version (available soon)



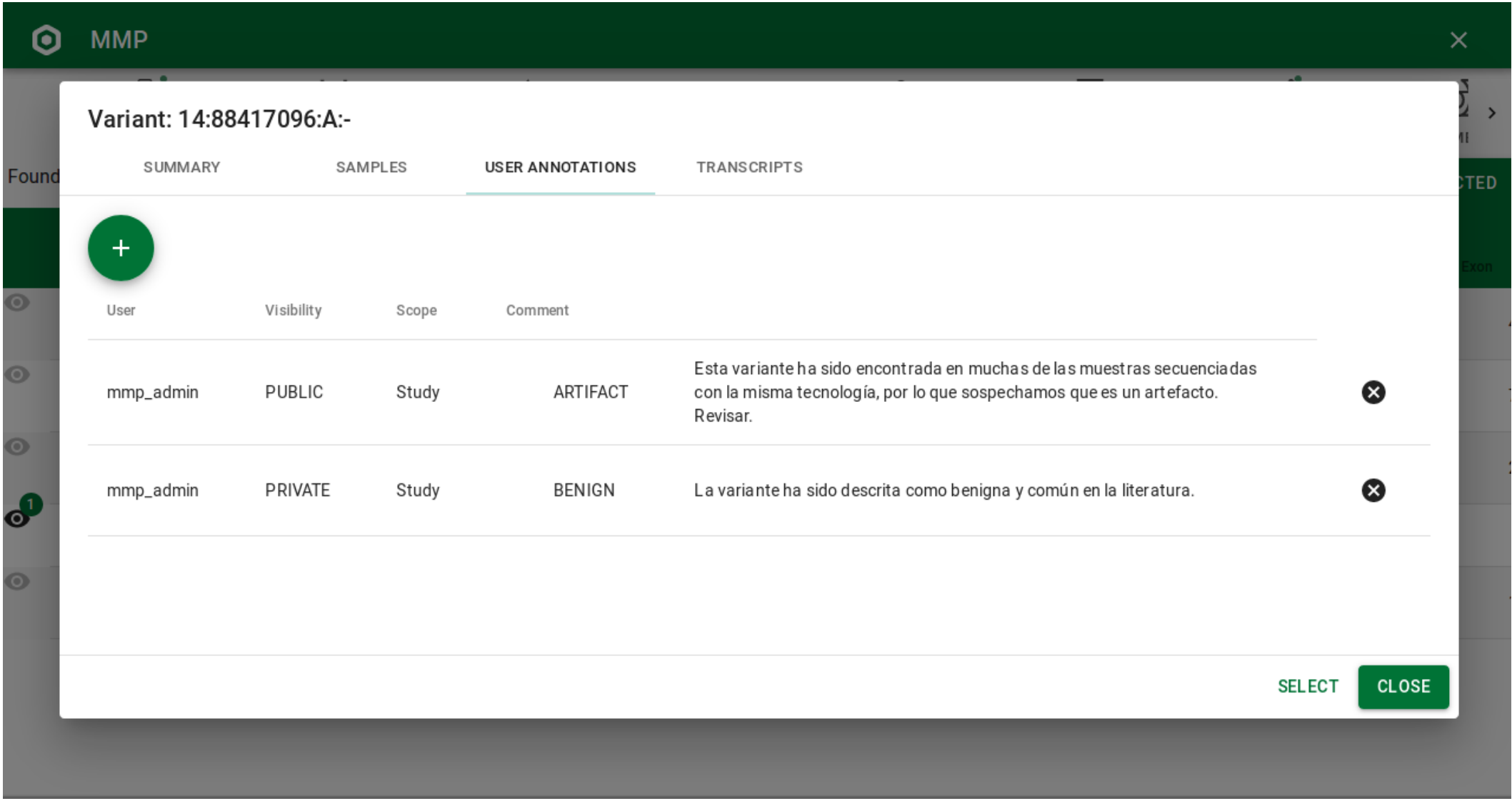
IVA dev version (available soon)



IVA dev version (available soon)



IVA dev version (available soon)



Acknowledgements

- **Ignacio Medina** (*Genomics England, University of Cambridge, OpenCB Team leader)*
- Jacobo Coll (*Genomics England*)
- Javier Lopez (*Genomics England*)
- Pedro Furio (*Genomics England*)
- Joaquin Tarraga (*University of Cambridge*)
- Antonio Rueda (*Genomics England*)

Clinical Bioinformatics Area
Fundación Progreso y Salud, Sevilla, Spain, and…

…the INB-ELIXIR-ES, National Institute of Bioinformatics and the BiER (CIBERER Network of Centers for Research in Rare Diseases)

Follow us on twitter

@xdopazo

@ClinicalBioinfo

https://www.slideshare.net/xdopazo/

GOBIERNO DE ESPAÑA · MINISTERIO DE ECONOMÍA, INDUSTRIA Y COMPETITIVIDAD · Instituto de Salud Carlos III
UNIÓN EUROPEA · Fondo Europeo de Desarrollo Regional
"Una manera de hacer Europa"
JUNTA DE ANDALUCIA · CONSEJERÍA DE SALUD · MGP Medical Genome Project · ciberer BIER · ciberer · INB · Machine Learning for Personalized Medicine · elixir SPAIN · MARIE CURIE ACTIONS