

## Common Infrastructure for National Cohorts in Europe, Canada, and Africa - CINECA -

### Deliverable D5.2 - Universal FAIR Data Compliant Federated Biomarker Discovery Service

Work Package:	WP5 - Healthcare Interoperability and Clinical Applications
Lead Beneficiary:	European Molecular Biology Laboratory
WP Leader(s):	Andrew Stubbs (EMC), Patrick Ruch (HESSO)
Contributing Partner(s):	
Contractual Delivery Date:	30th April 2023
Actual Delivery Date:	3rd May 2023
Authors of this Deliverable:	Saskia Hiltmann, Jasper Ouwerkerk, Andrew Stubbs
Contributors:	Saskia Hiltmann, Jasper Ouwerkerk, Helena Rasche, Willem de Koning
Reviewed by:	Kaur Alasoo
Approved by:	Thomas Keane
Dissemination Level:	Public
Type of Deliverable:	Demonstrator
Grant agreement:	No. 825775 Horizon 2020 (H2020-SC1-BHC-2018-2020)
Type of action:	RIA
Start Date:	1 Jan 2019
Duration:	54 months

## Table of contents:

1. Executive Summary	3
2. Project objectives	3
3. Detailed report on the deliverable	4
3.1 Background	4
3.2 Description of Work	5
3.2.1 Data import: Galaxy EGA HTSget Download client	5
3.2.1.1 Use case: Trio Analysis for RD-Connect synthetic dataset Beyond the 1 Million Genomes (synB1MG)	8
3.2.2 Data analysis: tools for biomarker discovery in Galaxy	11
3.2.3 Data sharing: Beacon Integration in Galaxy	13
3.2.4. Future steps and Beyond CINECA	15
4. References	16
5. Abbreviations	16
6. Delivery and schedule	16
7. Adjustments made	17
8. Appendices	17



## 1. Executive Summary

This deliverable (D.5.2) allows clinical researchers to access and reuse existing data from public omics databases, to integrate these data with their own data and to enable access to a suite of bioinformatics tools, via the Galaxy workflow platform, for advanced cancer biomarker discovery and for patient stratification.

This deliverable (D5.2) provides FAIR data access from GA4GH public genomics data resources (e.g. European Genome Archive (EGA) using the GA4GH htsget protocol directly into Galaxy. We developed an online tutorial to explain our workflow in detail, which is associated with the Galaxy Training Network (GTN). This study shows it is feasible to adopt *end-to-end* (i.e. data to results) scalable FAIR analysis of clinical data, and ultimately for any future analysis on data available at the EGA. Additionally we have developed a unified resource, Cancer Galaxy, as a landing page for Cancer related workflows and tools<sup>1</sup> which includes tools implemented/developed by CINECA and other bioinformatics groups from the EU and worldwide.

## 2. Project objectives

The main objective of WP5 within the CINECA project is to develop federated clinical applications that leverage services developed in the CINECA project. Data discovery provided by WP1,2,3 to identify relevant cohorts for further retrieval and analysis using the FAIR Cancer Galaxy platform.

This report outlines the outcome from **Task 5.2: FAIR data analysis for cancer biomarker discovery**, deliverable 5.2 (D5.2) which aimed to utilise federated data (T5.1) from the Catalog of FAIR data (T1.2) available via the interactive query portals developed in T1.4. Secure access to these data will be managed with the ELIXIR AAI service delivered in T2.3.

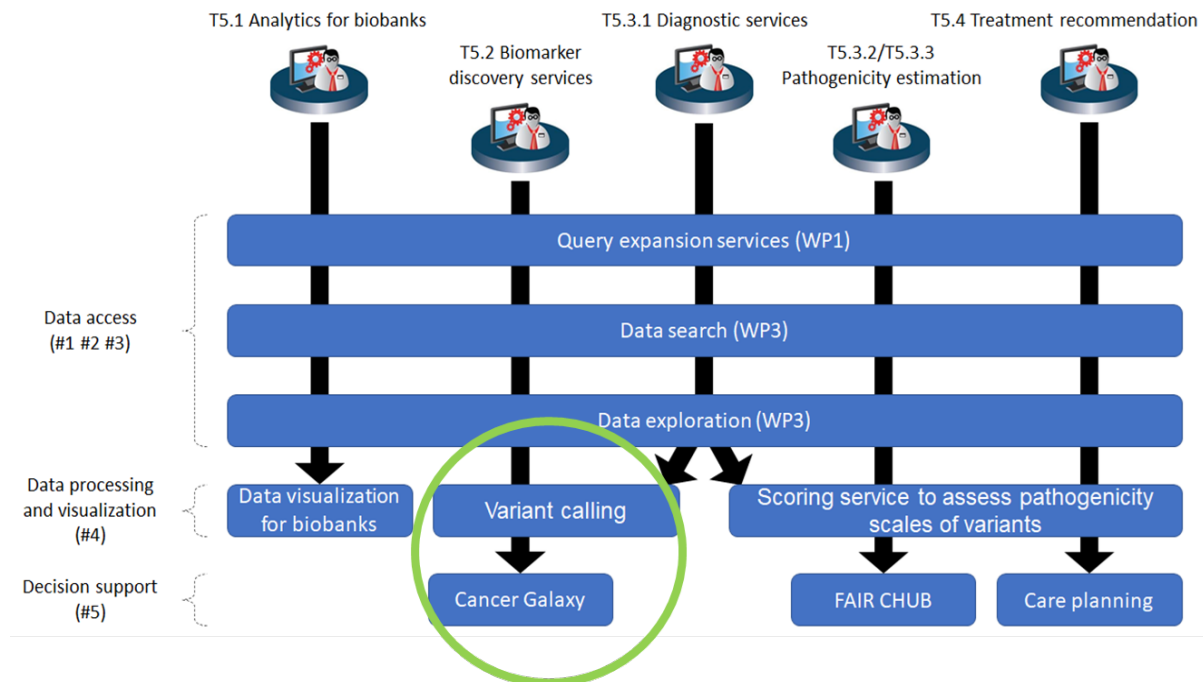
Within CINECA the completion of D5.2 has contributed to the following objectives:

- a) To Develop a GA4GH compliant service to access and retrieve omic data which supports FAIR data principles
- b) To implement a suite of tools for multimodal cancer analysis (Cancer Galaxy)

---

<sup>1</sup> <https://cancer.usegalaxy.eu/>





Interoperability of the services developed by WP5.

### 3. Detailed report on the deliverable

#### 3.1 Background

A biomarker is a measured indicator of normal biological processes, pathogenic processes or responses to an exposure, including therapeutic interventions and can be derived from molecular, histologic, radiographic, or physiologic characteristics. Since the purpose of a biomarker is to aid in the early diagnosis of disease as well as risk prediction, and to be useful, it should influence clinical decision-making while minimizing harm and expense, and ideally lead to decreases in the number of cancer-related deaths.

In CINECA we have focused on genomic or transcriptomic features which will serve as biomarkers in both genetic disease progression or in T5.2 for cancer progression and patient stratification. However, experimental replication for validation is essential to provide robust biomarkers that can progress from preclinical into clinical utility. Replication studies by researchers not involved with the original study require access to both the data and the methodology used in the original discovery study which is known to be low; for example in The Reproducibility Project: Cancer Biology only managed to replicate 5 of 17 highly cited articles, in Science and Nature the success was 57 to 67%.

Ensuring data are findable, accessible, interoperable, and reusable (FAIR) will allow researchers to efficiently leverage the wealth of genomic and clinical data currently generated in heterogeneous formats and by diverse organisations around the world. Within and beyond CINECA, large genomic repositories (e.g. EGA; Freeberg et al. 2022) are important sources for secure storage and access of these international multi-omic resources. The GA4GH consortium has developed the BEACON

technology<sup>2</sup> to deliver universal, standardised FAIR and GDPR controlled access to these large repositories. The BEACON service returns a summarised view of the underlying genomic data whilst GA4GH-compliant HTSGET protocol provides a service to access the underlying genomic content for those researchers granted authorised access from the relevant Data Access Committee (DAC).

supporting requests over genomic ranges. PyEGA3, a python tool, implements the HTSGET protocol as a service is available for the EGA<sup>3</sup> for securely streaming genomic data (Keller, *et al.* 2019).

Access to methodology for reproducing scientific studies can be achieved using code repositories such as GitHub, but the researcher is then responsible to replicate the dependencies of both the programming environment and the associated packages or libraries. To alleviate this challenge there are several workflow applications (e.g. Nextflow, Galaxy) whereby Galaxy is a FAIR and user-friendly web-based data analysis platform. It aims to make bioinformatics more accessible for non-bioinformaticians by reducing the technical burden of installing tools, managing resources, and running analyses by way of the command line. Instead, users interact with web-based forms to configure and run their analyses. Galaxy is open source, and can be set up and run on private infrastructure, or users can make use of one of many public (and free) Galaxy servers. Furthermore, the Galaxy training network<sup>4</sup> (GTN) hosts a large repository of training materials covering a wide range of scientific data analysis use cases, where scientists can use self-paced training materials for learning new techniques and analysis tools.

Federated workflows used to analyse these data are part of existing Galaxy services for cancer genomics and include fusion gene detection (Teles Alves, *et al.* 2015, Hiltmann, *et al.* 2015, Albuquerque, *et al.* 2017). To include all possible tools and existing workflows to analyse cancer omics data we have implemented Cancer Galaxy<sup>5</sup>, a service at usegalaxy.eu. This service will provide clinical research scientists with a secure electronic notebook of their bioinformatics analysis with complete provenance for the data, the methods and the parameters. In summary Cancer Galaxy is a unique application for FAIR data analysis used by researchers involved with clinical and translational research projects.

## 3.2 Description of Work

This deliverable (D5.2) is outlined in three categories, (i) data import (3.2.1), (ii) data analysis (3.2.2) with various tools and workflows and (iii) data sharing via beacons (3.2.3)

### 3.2.1 Data import: Galaxy EGA HTSget Download client

The European Genome-Phenome Archive (EGA) (<https://ega-archive.org/>, Freeberg *et al.* 2022) offers a variety of secure data access and download services based on GA4GH standards (Figure 1). The EGA supports the HTSget protocol (Kelleher *et al.* 2019), which was designed to allow secure, efficient and reliable access to sequencing read and variation data.

---

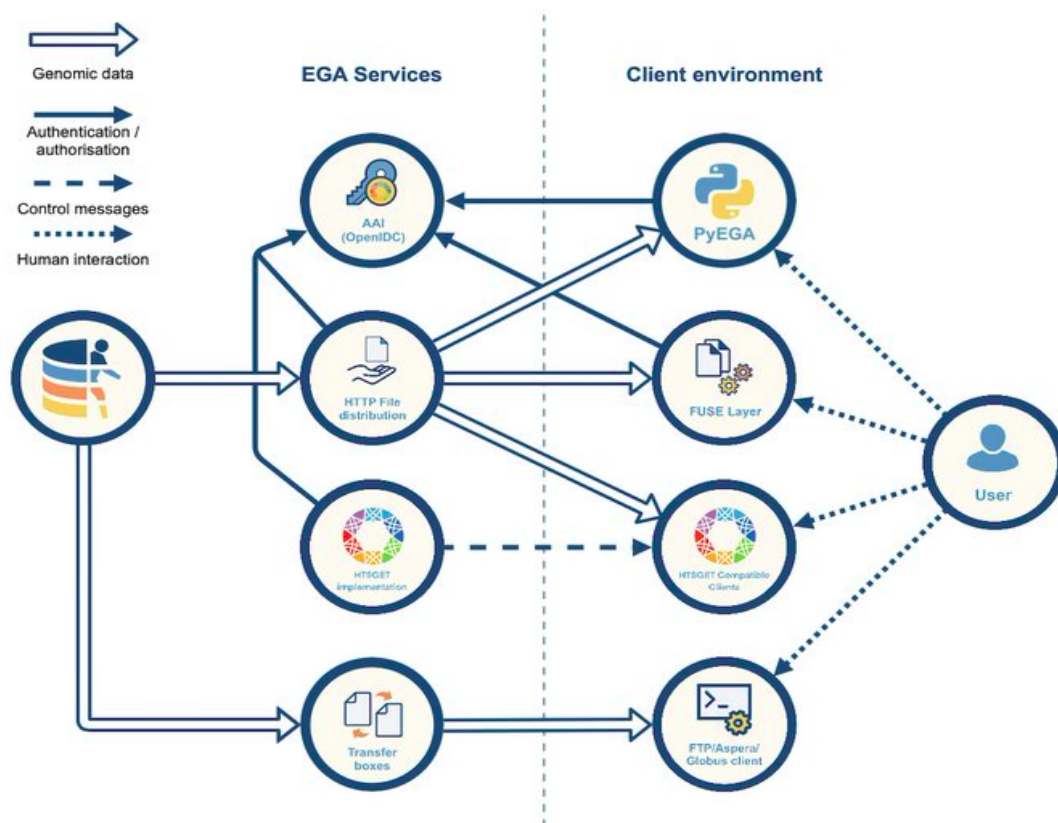
<sup>2</sup> <https://elixir-europe.org/internal-projects/commissioned-services/beacon-network-service>

<sup>3</sup> <https://github.com/EGA-archive/ega-download-client>

<sup>4</sup> <https://training.galaxyproject.org>

<sup>5</sup> <https://cancer.usegalaxy.eu/>





**Figure 1:** EGA data access infrastructure. FUSE: Filesystem in Userspace. AAI: Authentication and Authorization Infrastructure. OpenIDC: OpenID Connect, an open standard and decentralised authentication protocol. Clients may be implemented to connect to this infrastructure. Our Galaxy clients utilise PyEGA which supports AAI and the HTSget protocol.

As part of our federated biomarker discovery service, we have implemented a download client in Galaxy ([source code](#)<sup>6</sup>, [Galaxy tool shed](#)<sup>7</sup>) that can securely connect to the EGA data server and download authorised datasets directly from the EGA into Galaxy’s computational environment. This tool was submitted to Galaxy’s Intergalactic Utilities Commission (IUC); all tools submitted here undergo a strict peer review process to ensure adherence to the latest best coding practices, security standards, and FAIR principles.

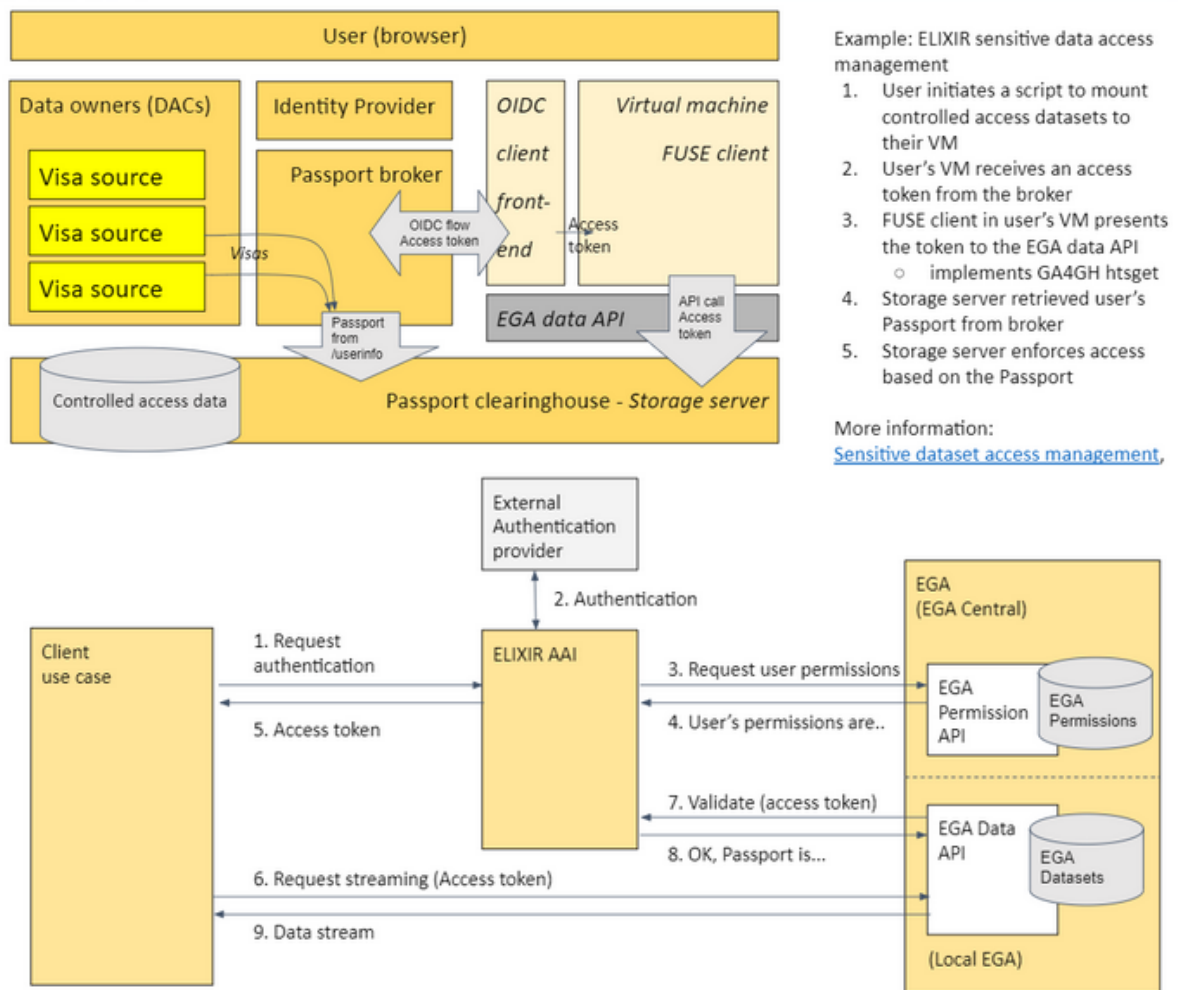
Users must separately obtain authorization to datasets via the relevant Data Access Committee, but once obtained, Galaxy will take care of user authentication behind the scenes. In the pilot version of the tool, EGA credentials must be manually connected to the Galaxy user’s account via a one-time operation in the user preferences menu, but in future releases of Galaxy, when a user is logged in with

<sup>6</sup> <https://github.com/galaxyproject/tools-iuc/tree/main/tools/pyega3>

<sup>7</sup>

[https://toolshed.g2.bx.psu.edu/repository/view\\_repository?id=47eef4cb4009abe4&changeset\\_revision=9564758e8638](https://toolshed.g2.bx.psu.edu/repository/view_repository?id=47eef4cb4009abe4&changeset_revision=9564758e8638)

an AAI-connected account, this step can be omitted and Galaxy will automatically determine data authorization based on the GA4GH passport and visa authentication framework (Figure 2).



**Figure 2:** The GA4GH passport and visa implementation and how it interacts with ELIXIR AAI. Galaxy allows logging via ELIXIR AAI, and this can be leveraged to provide EGA data access and authorization credentials in the Galaxy download client automatically.



The Galaxy download client allows several operations; it can list the full set of datasets the user is currently authorised for in EGA, it can list the files present in each of these datasets, it can download selected files in their entirety and place them in the user's Galaxy analysis environment, and lastly it also supports specifying genomic ranges to fetch when the full datasets are not required. This saves valuable processing time and storage space. The user interface (Figure 3) is a simple form requesting the required information from the user (e.g. operation type, dataset IDs and optionally a genomic range). Galaxy automatically stores the full provenance (parameter settings, software versions, input datasets) for any analysis performed.

**Figure 3:** The EGA download client interface in Galaxy.

### 3.2.1.1 Use case: Trio Analysis for RD-Connect synthetic dataset Beyond the 1 Million Genomes (synB1MG)

The synthetic data (synB1MG) represent the genomes from six sets parent-offspring trios from the Beyond 1 Million Genomes (B1MG) project<sup>8</sup> that have been *spiked* to include the causal rare variant for each of the pathologies. Our implementation of the PyEGA3 tool allows users to filter datasets, available on the EGA, based on their metadata and scale up analysis. We have implemented a Galaxy workflow for trio analysis on parent-offspring trios B1MG. Moreover, we added the [clin.iobio tool](https://clin.iobio.io/)<sup>9</sup> for variant analysis (Ward, *et al.* 2022), including trio analysis with the complete workflow, including data retrieval with PyEGA3, implemented within Galaxy and available from WorkflowHub for

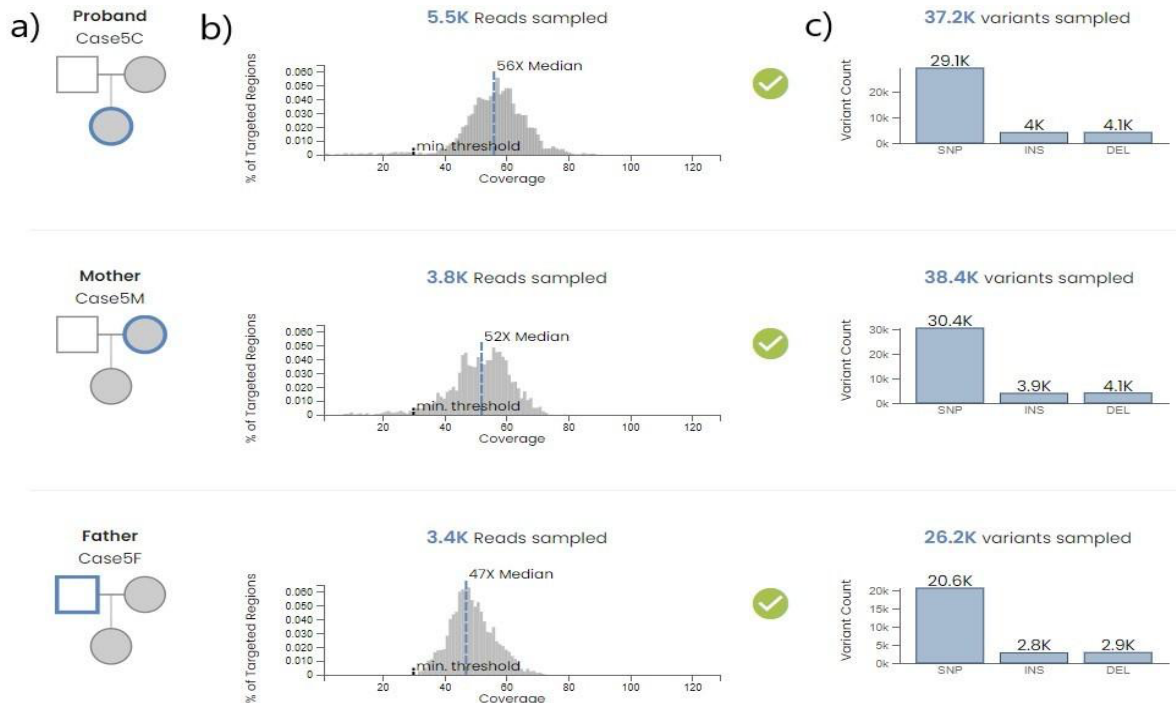
<sup>8</sup> <https://b1mg-project.eu>

<sup>9</sup> <https://clin.iobio.io/>

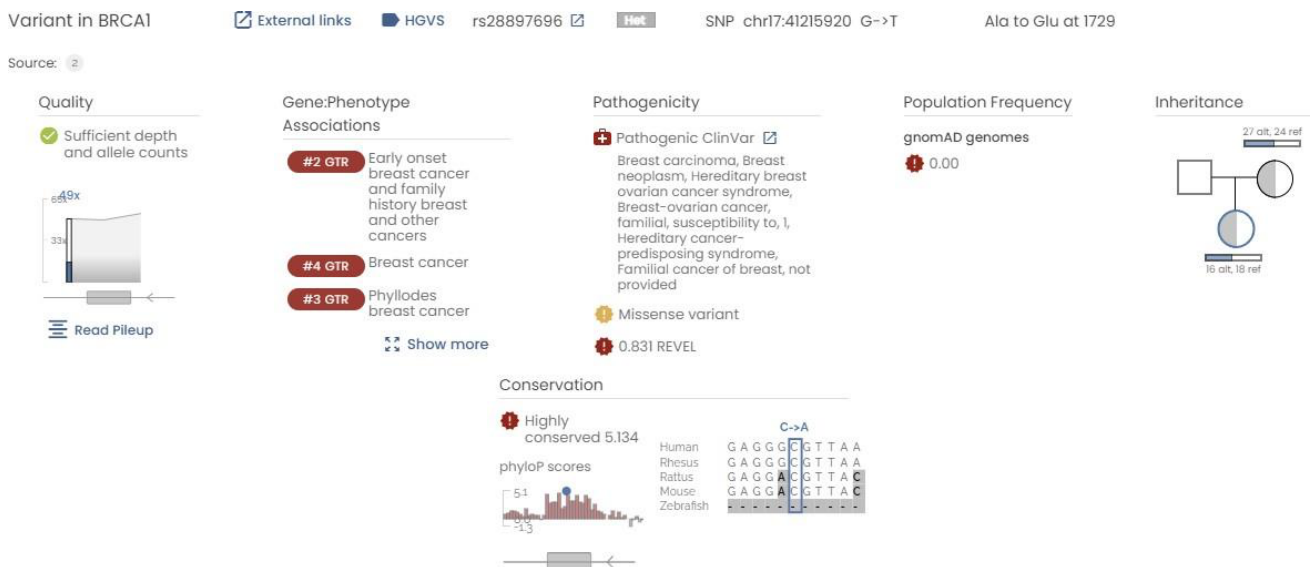




discoverability<sup>10</sup>. Pathogenic variant detection is achieved with either GEMINI [Integrative Exploration of Genetic Variation and Genome Annotations] (Paila et al, 2013) or using [clin.iobio](https://clin.iobio). An example of the output generated using [clin.iobio](https://clin.iobio) for the breast cancer proband Case5, from which the causative variant is BRCA1, can be observed in Figures 4 and 5.



**Figure 4.** Quality control figures created by [clin.iobio](https://clin.iobio). Here a) shows a visual representation of the family pedigree, b) shows the distribution of coverage across different regions, and c) shows the variant type distribution.



**Figure 5.** Overview of [clin.iobio](https://clin.iobio) results for the spiked-in variant for use case #5. The figure shows statistics on quality of the variant, phenotype associations, pathogenicity, population frequency, inheritance, and conservation.

<sup>10</sup> <https://workflowhub.eu/workflows/363>

The results of this workflow using either GEMINI or clin.iobio to identify the causal variants in the probands of the synB1MG datasets demonstrate that clin.iobio correctly identifies the variants whilst GEMINI suffers from too many false positive results (Table 1).

Case	Disease	Pattern	GEMINI	Clin.iobio
Case 1	Congenital myasthenic syndrome	Autosomal Dominant De Novo	0	1
Case 2	Macular dystrophy	Autosomal Dominant	77	1
Case 3	Muscular dystrophy	Autosomal Recessive Compound Heterozygous	0	2
Case 4	Mitochondrial disorder	Autosomal Recessive Homozygous Consanguineous	26	1
Case 5	Breast cancer	Autosomal Dominant	142	1
Case 6	Congenital myasthenic syndrome	Autosomal Dominant De Novo	0	1

**Table 1:** causative variants identified using the TrioAnalysis workflow with either GEMINI or clin.iobio used for the proband variant selection.

This workflow is accompanied by a [comprehensive tutorial](#)<sup>11</sup>, submitted to the [Galaxy training network](#)<sup>12</sup> (GTN), entitled Trio Analysis using Synthetic Datasets from RD-Connect GPAP (Figure 6). This tutorial provides a step-by-step guide through the process of obtaining data access to EGA datasets, connecting Galaxy with EGA user credentials, and using the download client to import datasets directly from EGA to the Galaxy analysis environment. The tutorial continues with a downstream analysis of the data, performing a trio analysis on a Beyond 1 Million Genomes (B1MG) breast cancer dataset from the RD-Connect GPAP (Genome-Phenome Analysis Platform) project. This tutorial was developed as part of the CINECA end-to-end federated analysis learning pathway<sup>13</sup>. The full workflow of this use-case is available both from the GTN, as well as from the ELIXIR workflow hub<sup>14</sup>.

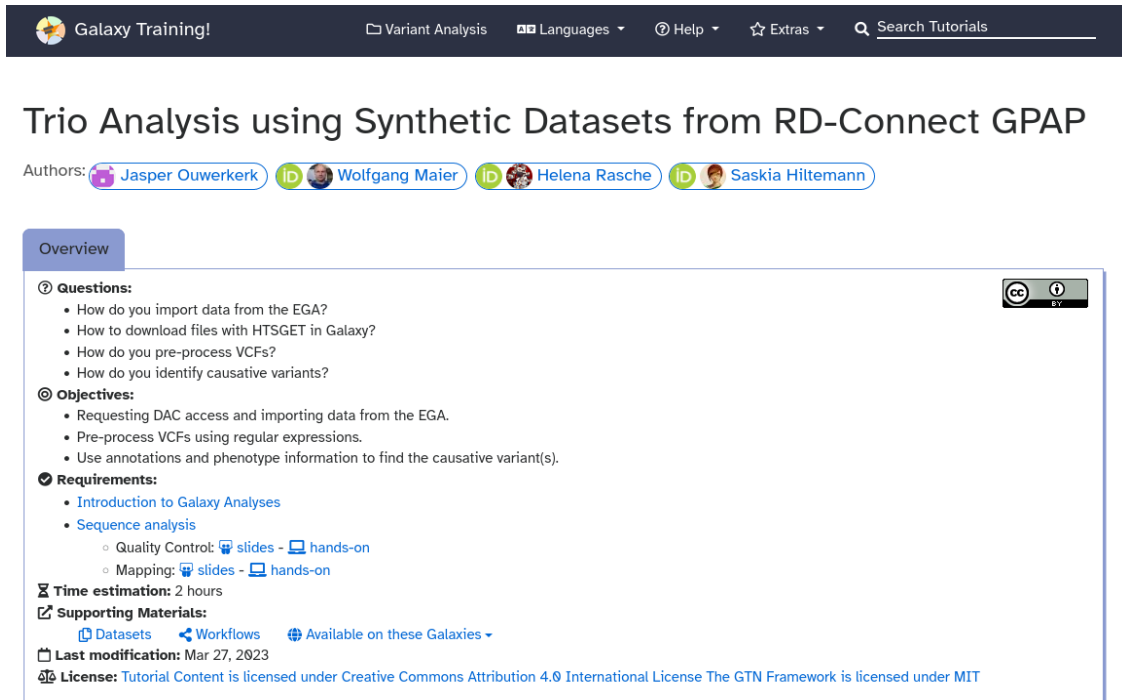
<sup>11</sup> <https://training.galaxyproject.org/training-material/topics/variant-analysis/tutorials/trio-analysis/tutorial.html>

<sup>12</sup> <https://training.galaxyproject.org/>

<sup>13</sup> <https://www.ebi.ac.uk/training/materials/cineca-federated-data-analysis/introduction/>

<sup>14</sup> <https://workflowhub.eu/>





Galaxy Training! Variant Analysis Languages Help Extras Search Tutorials

## Trio Analysis using Synthetic Datasets from RD-Connect GPAP

Authors: [Jasper Ouwerkerk](#) [Wolfgang Maier](#) [Helena Rasche](#) [Saskia Hiltmann](#)

**Overview**

**Questions:**

- How do you import data from the EGA?
- How to download files with HTSGET in Galaxy?
- How do you pre-process VCFs?
- How do you identify causative variants?

**Objectives:**

- Requesting DAC access and importing data from the EGA.
- Pre-process VCFs using regular expressions.
- Use annotations and phenotype information to find the causative variant(s).

**Requirements:**

- Introduction to Galaxy Analyses
- Sequence analysis
  - Quality Control: [slides](#) - [hands-on](#)
  - Mapping: [slides](#) - [hands-on](#)

**Time estimation:** 2 hours

**Supporting Materials:**

[Datasets](#) [Workflows](#) [Available on these Galaxies](#)

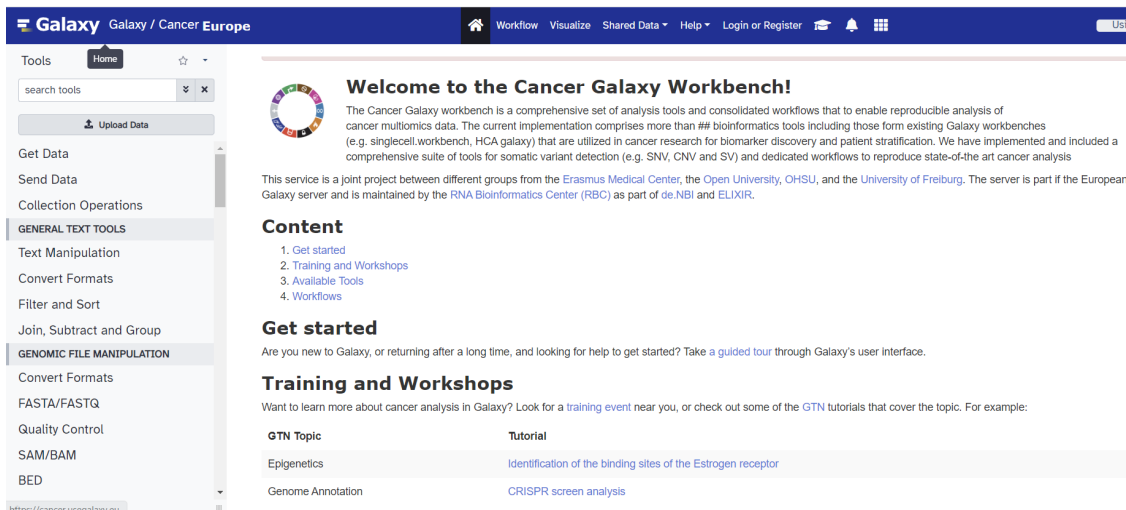
**Last modification:** Mar 27, 2023

**License:** Tutorial Content is licensed under Creative Commons Attribution 4.0 International License The GTN Framework is licensed under MIT

**Figure 6:** Screenshot of the tutorial demonstrating the HTSget EGA download client and downstream data analysis. This tutorial is one of the modules in the CINECA end-to-end federated data analysis learning pathway.

### 3.2.2 Data analysis: tools for biomarker discovery in Galaxy

Galaxy offers many tools and workflows for biomarker discovery. We have collected these resources in a single view on the European Galaxy server<sup>15</sup>. This view filters the Galaxy tool panel to those tools relevant to Cancer analysis and biomarker discovery, as well as providing a dedicated home page listing relevant resources such as tools, workflows and tutorials. This Galaxy environment can be accessed via <https://cancer.usegalaxy.eu> (Figure 7).



Galaxy / Cancer Europe Workflow Visualize Shared Data Help Login or Register

Tools Home search tools Upload Data

Get Data Send Data Collection Operations

**GENERAL TEXT TOOLS**

Text Manipulation Convert Formats Filter and Sort Join, Subtract and Group

**GENOMIC FILE MANIPULATION**

Convert Formats FASTA/FASTQ Quality Control SAM/BAM BED

**Welcome to the Cancer Galaxy Workbench!**

The Cancer Galaxy workbench is a comprehensive set of analysis tools and consolidated workflows that to enable reproducible analysis of cancer multomics data. The current implementation comprises more than ## bioinformatics tools including those form existing Galaxy workbenches (e.g. singlecell.workbench, HCA galaxy) that are utilized in cancer research for biomarker discovery and patient stratification. We have implemented and included a comprehensive suite of tools for somatic variant detection (e.g. SNV, CNV and SV) and dedicated workflows to reproduce state-of-the-art cancer analysis

This service is a joint project between different groups from the Erasmus Medical Center, the Open University, OHSU, and the University of Freiburg. The server is part of the European Galaxy server and is maintained by the RNA Bioinformatics Center (RBC) as part of de.NBI and ELIXIR.

**Content**

- Get started
- Training and Workshops
- Available Tools
- Workflows

**Get started**

Are you new to Galaxy, or returning after a long time, and looking for help to get started? Take a [guided tour](#) through Galaxy's user interface.

**Training and Workshops**

Want to learn more about cancer analysis in Galaxy? Look for a [training event](#) near you, or check out some of the [GTN tutorials](#) that cover the topic. For example:

GTN Topic	Tutorial
Epigenetics	Identification of the binding sites of the Estrogen receptor
Genome Annotation	CRISPR screen analysis

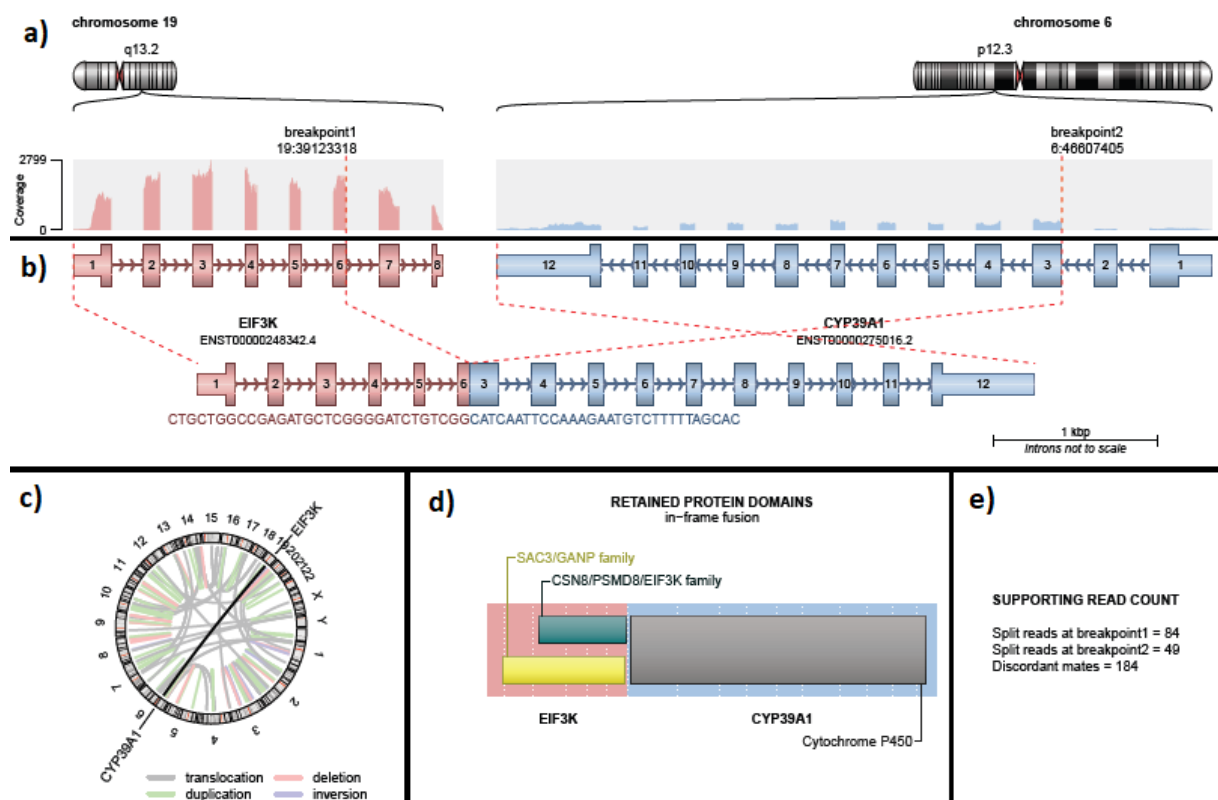
**Figure 7:** Screenshot of Cancer Galaxy workbench.

<sup>15</sup> <https://usegalaxy.eu>



### 3.2.2.1 Use case: Fusion gene detection

Fusion genes are a genetic mutation caused by structural rearrangement of the genome resulting in the read-through in-frame combination of two previously distinct genes into one (Lei et al. 2016). Fusion genes are essential biomarkers for cancer since they contribute to around 20% of human cancer morbidity (Mitelman et al. 2007). Galaxy has implemented two state-of-the-art fusion gene detection tools: STAR-Fusion (Haas et al. 2019), which uses only RNA-Seq data, and Arriba (Uhrig et al. 2021), which can use both Whole Genome Sequencing (WGS) and RNA-Seq data. Arriba improves the detection of fusion gene detection by calling structural variants, large genomic mutations, from WGS data. We have implemented the Arriba tool in Galaxy by exposing the best practices parameters in the upstream sequence alignment tool STAR (Dobin et al. 2013). In addition, STAR-Fusion and Arriba are implemented in a Galaxy workflow to compare the output of both tools and to simplify fusion gene calling using the best practices. The workflow has been run on RNA-Seq ([SRR2532336](https://www.ncbi.nlm.nih.gov/sra/?term=SRR2532336))<sup>16</sup> and WGS ([SRR892296](https://www.ncbi.nlm.nih.gov/sra/?term=SRR892296))<sup>17</sup> data of the HCC1395 breast cancer cell line. This workflow produces multiple outputs, namely the detected fusion genes (Figure 8) and plots comparing the fusion gene detection tools (Figure 9) using only RNA-Seq (STAR and Arriba) or a combination of WGS and RNA-Seq data (arriba\_sv).

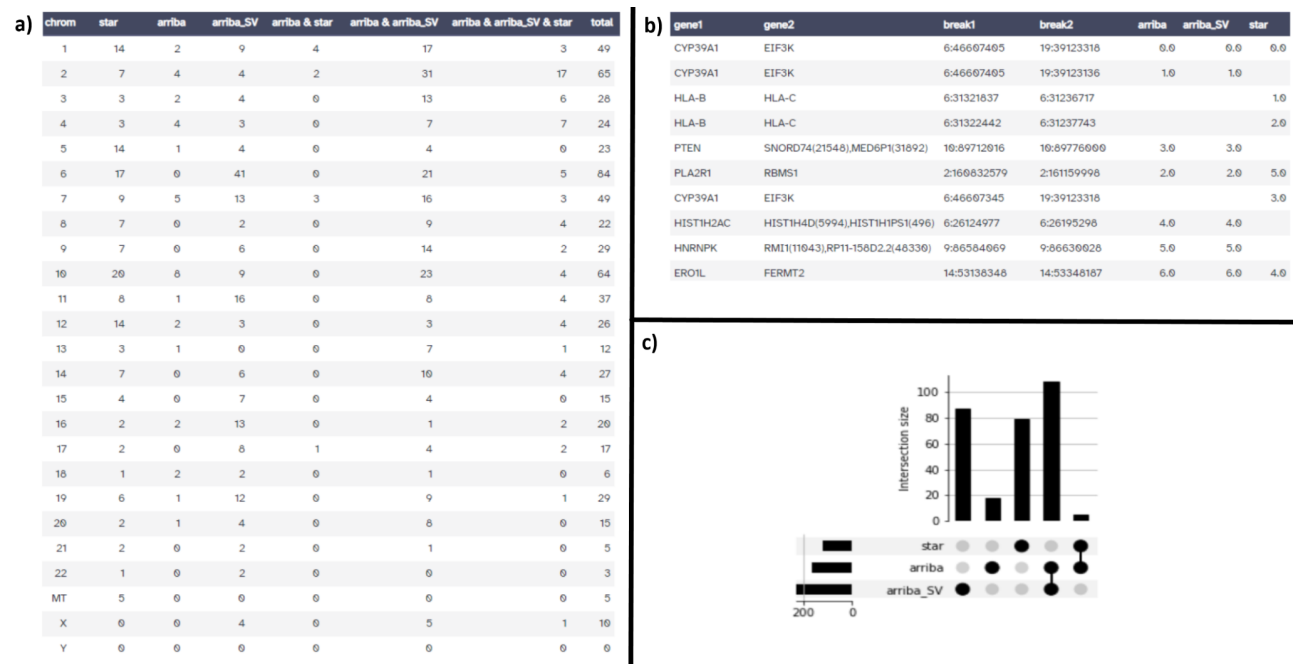


**Figure 8:** The fusion gene visualisations generated by Arriba. It shows a) the fusion gene location and read coverage, b) the fused genes and their associated exons, c) a plot of all the fusion genes across all chromosomes and their types coded in colour, d) the retained protein domains in both genes, and e) the supporting read counts for each breakpoint and discordant mates.

<sup>16</sup> <https://www.ncbi.nlm.nih.gov/sra/?term=SRR2532336>

<sup>17</sup> <https://www.ncbi.nlm.nih.gov/sra/?term=SRR892296>





**Figure 9:** The fusion gene visualisations generated by the workflow. It shows a) a table with the locations of all the detected fusion genes per tool and combinations of them, b) a table with the top 10 detected fusions genes across the tools, c) an upset plot showing the shared and uniquely detected fusion genes per tool.

### 3.2.3 Data sharing: Beacon Integration in Galaxy

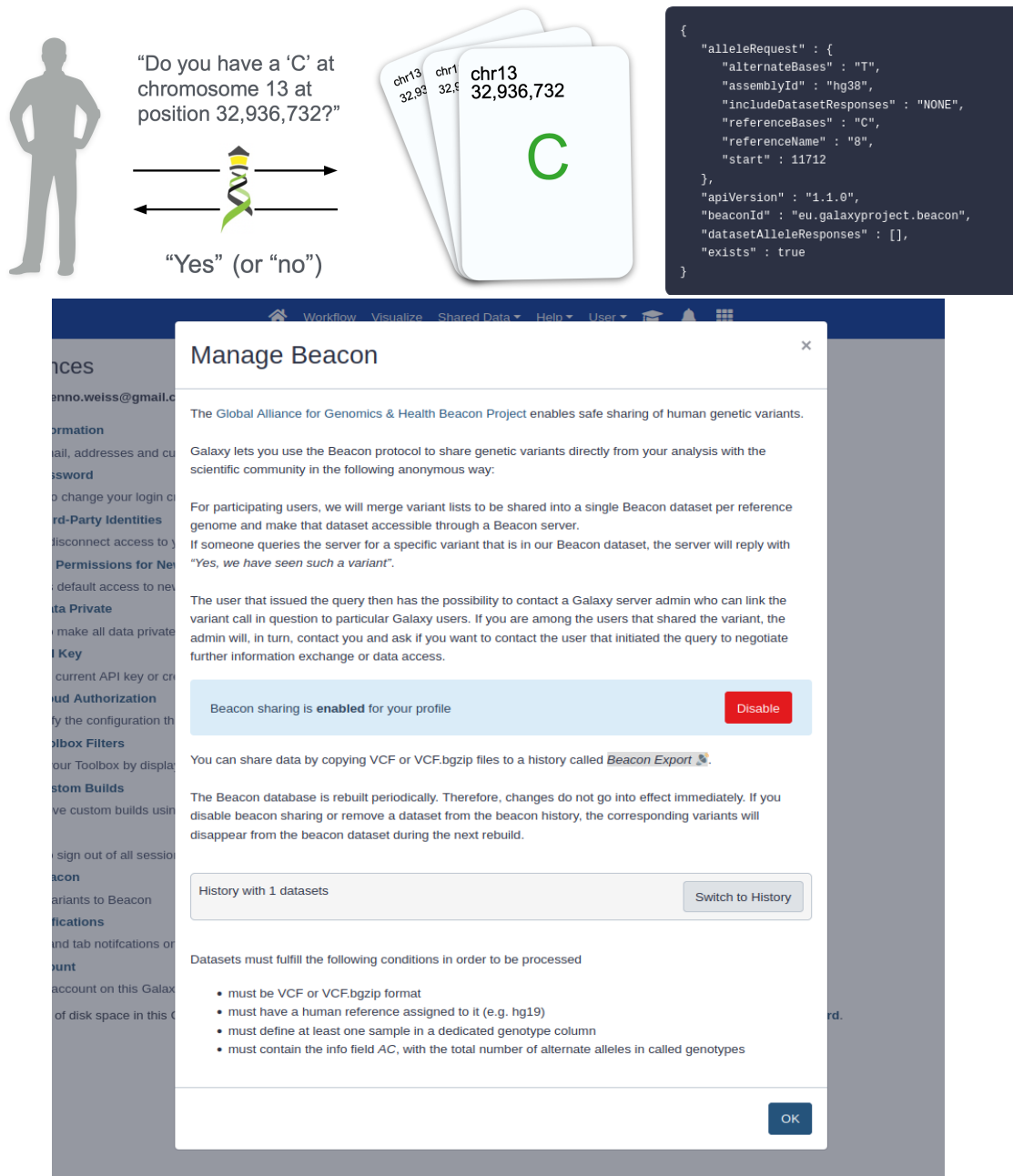
For this section of the deliverable, we had planned to integrate beacons into Galaxy. This has been completed by the larger Galaxy community<sup>18</sup>.

This work allows users to turn their genomic variants (VCF file) into a Beacon in a user-friendly manner. Users must configure the Beacon in Galaxy, provide the variant data to the dedicated Beacon analysis history, and opt-in to the Beacon being made accessible. Subsequently, this Beacon can be queried like any other Beacons (Figure 10).

To enable this Beacon integration, the administrator of the Galaxy server has to enable it on the server side. To this end, we have developed a tutorial for Galaxy admins on how to configure Galaxy to support this Beacon integration. We are testing this tutorial in an in-person Galaxy administrators workshop in Ghent on April 17-21, 2023.

<sup>18</sup> <https://galaxyproject.org/news/2023-01-beacon-integration/>





**Figure 10:** Integration of a Beacon inside Galaxy. A user query (top left) and a response from the Galaxy beacon in JSON format (top right). Setting up a beacon is a matter of configuring the Beacon settings in the Galaxy interface (bottom) and providing the source VCF file in a Galaxy history.

### 3.2.4. Future steps and Beyond CINECA

We are currently submitting a publication, “FAIR Data Retrieval for Sensitive Clinical Analysis in Galaxy” by Ouwerkerk et al. 2023 in which the content from section 3.2.1 has been drafted ready for submission to Gigascience in April 2023. All products and applications (Table 2) that were generated in the CINECA project that relate to Task 5.2 and D5.2 will be maintained by Erasmus MC for the foreseeable future.

Application Name	Federated Analysis link	Information link	Repository link	Used in
htsget-Galaxy	<a href="https://www.ebi.ac.uk/training/materials/cineca-federated-data-analysis/5-federated-data-analysis/">https://www.ebi.ac.uk/training/materials/cineca-federated-data-analysis/5-federated-data-analysis/</a>	<a href="https://training.galaxyproject.org/training-material/topics/variant-analysis/tutorials/trio-analysis/tutorial.html">https://training.galaxyproject.org/training-material/topics/variant-analysis/tutorials/trio-analysis/tutorial.html</a>	<a href="https://github.com/ErasmusMC-Bioinformatics">https://github.com/ErasmusMC-Bioinformatics</a>	RD-connect synthetic data B1MG
Cancer Galaxy		<a href="https://cancer.usegalaxy.eu/">https://cancer.usegalaxy.eu/</a>		Ongoing multi-omics data analysis use cases

**Table 2:** Application and services that have been delivered as part of D5.2.

This work has been included as outreach for other EU and global projects where federated FAIR CLOUD services would be invaluable for cancer research and for clinical decision support. These include presentations for:

1. GA4GH connect<sup>19</sup> entitled “CINECA WP5: Federated FAIR data biomarker discovery in cancer using GA4GH standard”.
2. EOSC4Cancer<sup>20</sup> entitled “Federated clinical applications in the cancer area at CINECA”.

<sup>19</sup> <https://broadinstitute.swoogo.com/ga4ghaprilconnect23/>

<sup>20</sup> <https://eosc4cancer.eu/>





## 4. References

- Albuquerque et al. Enhancing knowledge discovery from cancer genomics data with Galaxy. *Gigascience*. 2017 May 1;6(5):1-13.
- Dobin et al.. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013 Jan; 29(1): 15–21. <https://doi.org/10.1093/bioinformatics/bts635>
- Freeberg et al. The European Genome-phenome Archive in 2021. *Nucleic Acids Res*. 2022 Jan 7; 50(D1): D980–D987. <https://doi.org/10.1093/nar/gkab1059>
- Haas et al. Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol* 20, 213 (2019). <https://doi.org/10.1186/s13059-019-1842-9>
- Hiltmann et al. Discriminating somatic and germline mutations in tumor DNA samples without matching normals. *Genome Res*. 2015 Sep;25(9):1382-90. <https://doi.org/10.1101/2Fgr.183053.114>
- Kelleher et al. 2019. *Bioinformatics*. 2019 Jan 1; 35(1): 119–121. <https://doi.org/10.1093/bioinformatics/bty492>
- Kelleher et al. 2019. *Bioinformatics*. 2019 Jan 1; 35(1): 119–121. <https://doi.org/10.1093/bioinformatics/bty492>
- Lei et al. Evolutionary Insights into RNA trans-Splicing in Vertebrates. *Genome Biology and Evolution*. 2016 Mar 10; 8(3): 562-577. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4824033/>
- Mitelman et al. The impact of translocations and gene fusions on cancer causation. 2007 Mar 15; 7: 233-245. <https://www.nature.com/articles/nrc2091>
- Paila U, Chapman BA, Kirchner R, Quinlan AR. GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLOS Computational Biology* 2013;9:1–8. <https://doi.org/10.1371/journal.pcbi.1003153>
- Teles-Alves et al.. Next-generation sequencing reveals novel rare fusion events with functional implication in prostate cancer. *Oncogene*. 2015 Jan 29;34(5):568-77. <https://doi.org/10.1038/onc.2013.591>
- Uhrig et al. Accurate and efficient detection of gene fusions from RNA sequencing data. *Genome Res*. 2021 Mar;31(3):448-460. <https://doi.org/10.1101/gr.257246.119>.
- Ward A, et al.. Clin.iobio: A Collaborative Diagnostic Work- flow to Enable Team-Based Precision Genomics. *Journal of Personalized Medicine* 2022;12. <https://doi.org/10.3390/jpm12010073>

## 5. Abbreviations

AAI - Authentication and authorization infrastructure

EGA - European Genome-Phenome Archive

GTN - Galaxy Training Network

SV - Structural Variation/Variant

VCF - Variant Calling Format

WGS - Whole Genome Sequencing

## 6. Delivery and schedule

The delivery is delayed by 6 months due to the effect of the covid pandemic



## 7. Adjustments made

No adjustments to the deliverable were made or required just the time to complete was extended

## 8. Appendices

Appendix 1

