hCNV Community

Standards and protocols for CNV discovery and data exchange

Andrew Stubbs
Antonio Rausell
Krzysztof Poterlowicz
Michael Baudis

www.elixir-europe.org

2025-05-15

hCNV products

- Resource (https://refcnv.org/) : Progenetix cancer CNV reference resource and upcoming refCNV for reference copy number data accessible via Beacon API - Michael
- Resource (htsget2Galaxy) : FAIR access to GA4GH databases - Andrew
- Resource (CNV 2 Galaxy Beacon) : integrate rare disease workflow with external data (e.g hCNV from refcnv) and Cancer data in Genomic Beacons. – Khaled / Krzys
- Resource: - bioinformatics and ML tools for CNV assessment, including a couple of slides opening to non-coding SNVs at the very end (which can be of interest for the rare-disease group even if is beyond the CNV scope). - Antonio
- Standards: Beacon, VRS and VCF support for CNV representation and querying

progenetix.org

Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **150'000 cancer CNV profiles**
- more than 9**00 diagnostic types**
- runs on a **Beacon API**
- inclusion of reference datasets (e.g. TCGA)
- support for SNV data (TCGA, cell lines...)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services

Universität Zürich[UZH]

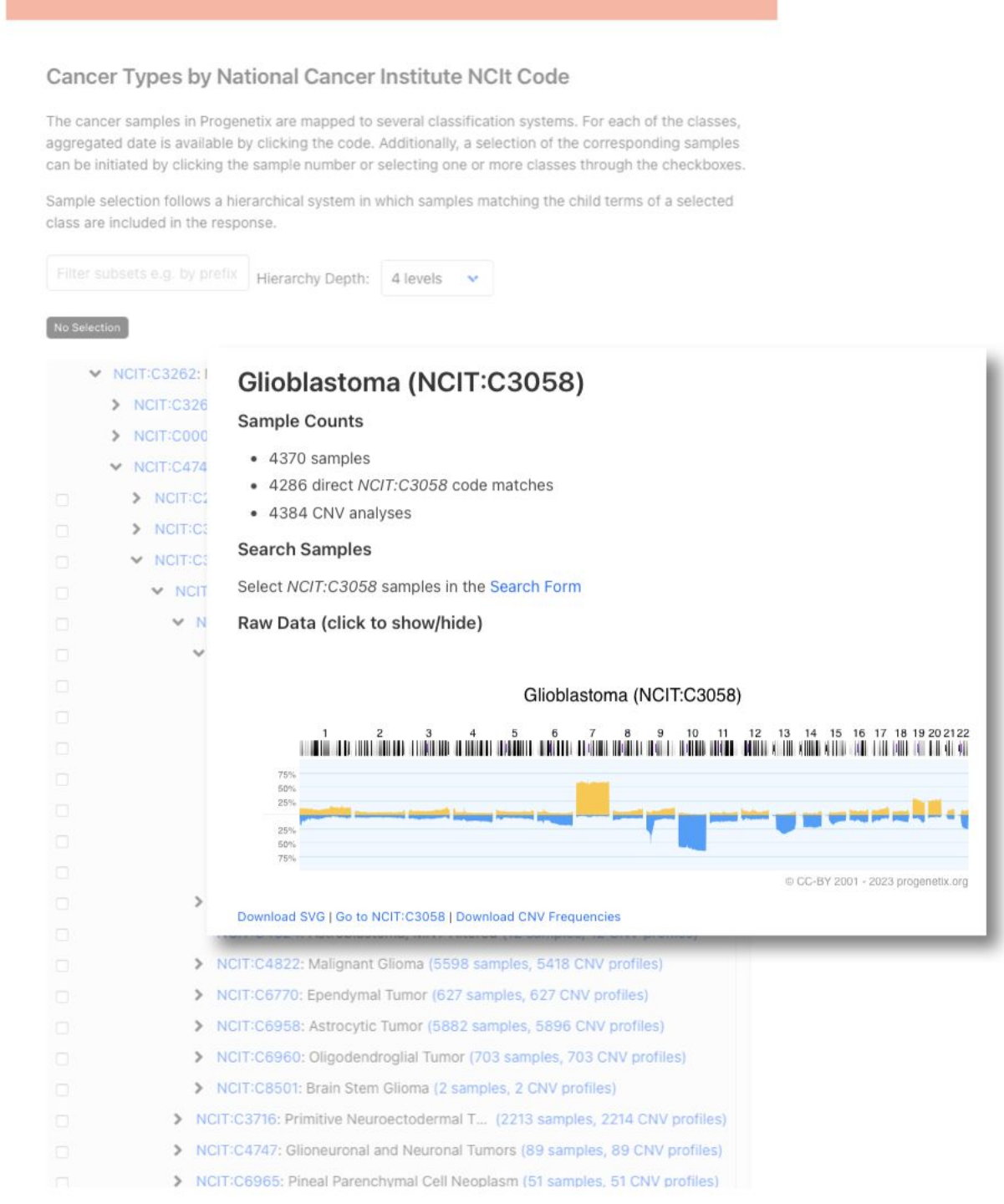
progenetix

SIB Swiss Institute of Bioinformatics

progenetix.org

Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **150'000 cancer CNV profiles**
- more than 9**00 diagnostic types**
- runs on a **Beacon API**
- inclusion of reference datasets (e.g. TCGA)
- support for SNV data (TCGA, cell lines...)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services



Universität Zürich[UZH]

progenetix

SIB Swiss Institute of Bioinformatics

progenetix.org

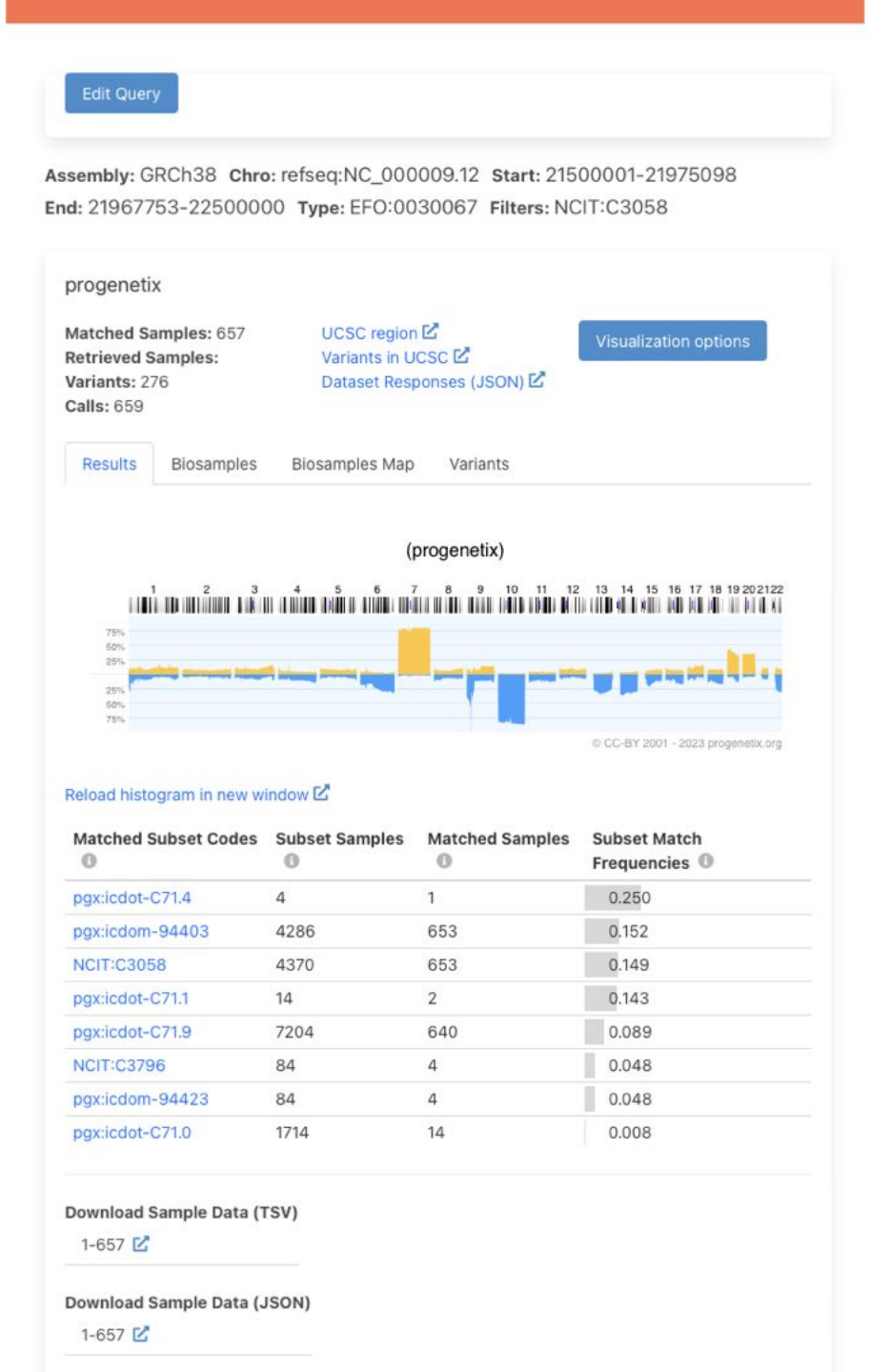
Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **150'000 cancer CNV profiles**
- more than 9**00 diagnostic types**
- runs on a **Beacon API**
- inclusion of reference datasets (e.g. TCGA)
- support for SNV data (TCGA, cell lines...)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services

Universität Zürich[UZH]

progenetix

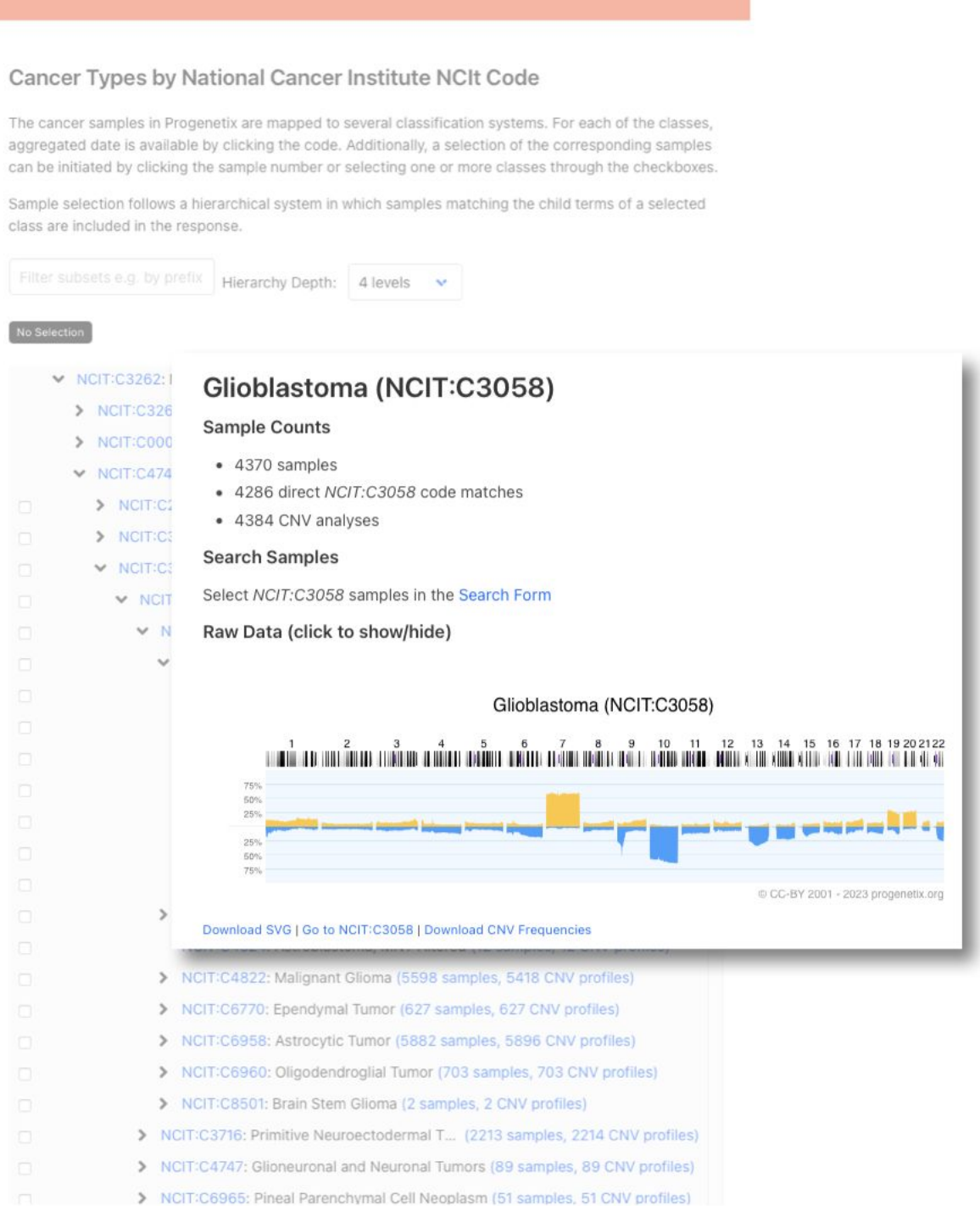
SIB Swiss Institute of Bioinformatics



progenetix.org

Cancer Genomics Reference Resource

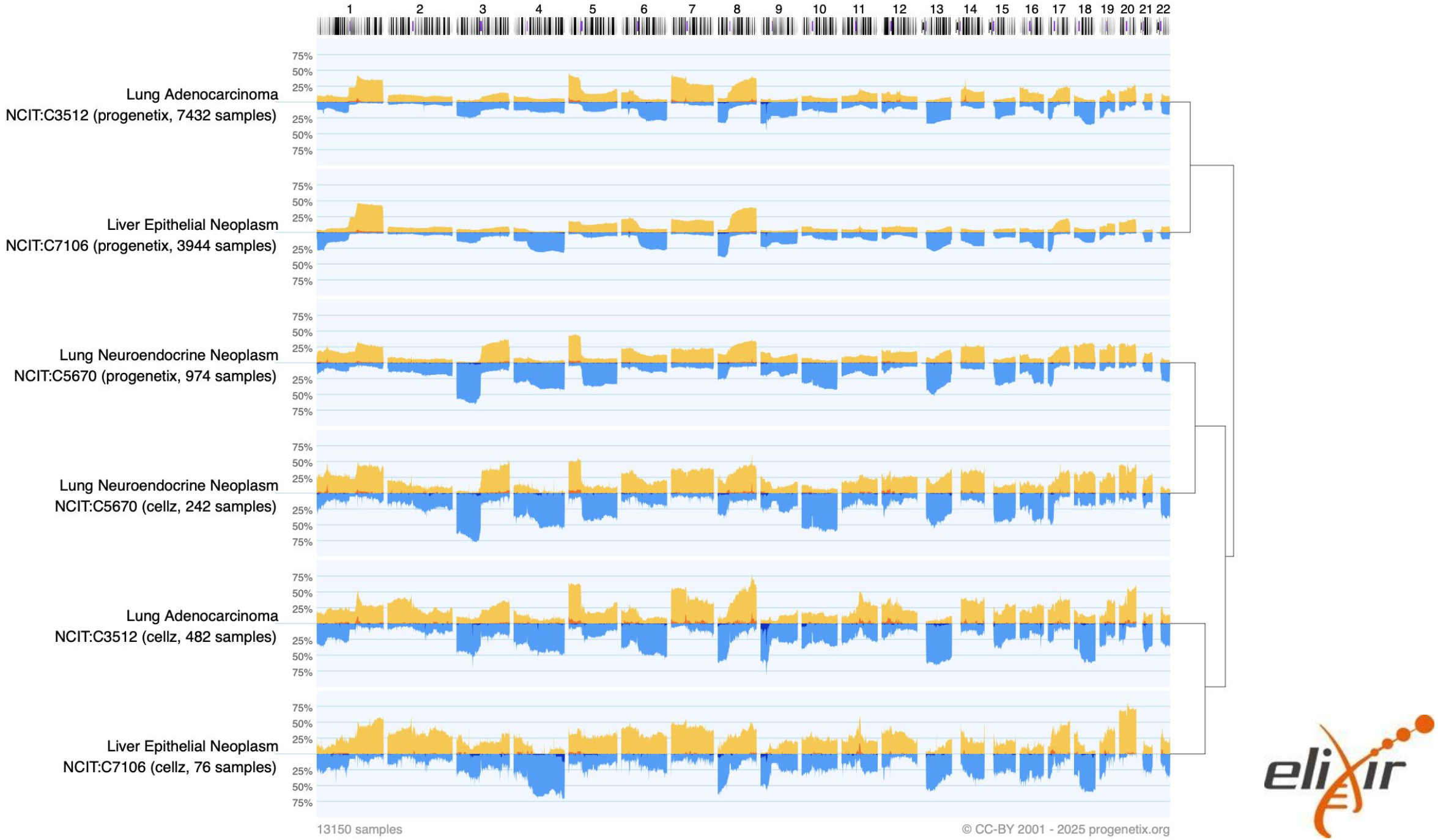
- ***open*** resource for oncogenomic profiles
- over **150'000 cancer CNV profiles**
- more than 9**00 diagnostic types**
- runs on a **Beacon API**
- inclusion of reference datasets (e.g. TCGA)
- support for SNV data (TCGA, cell lines...)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services



Universität Zürich[UZH]

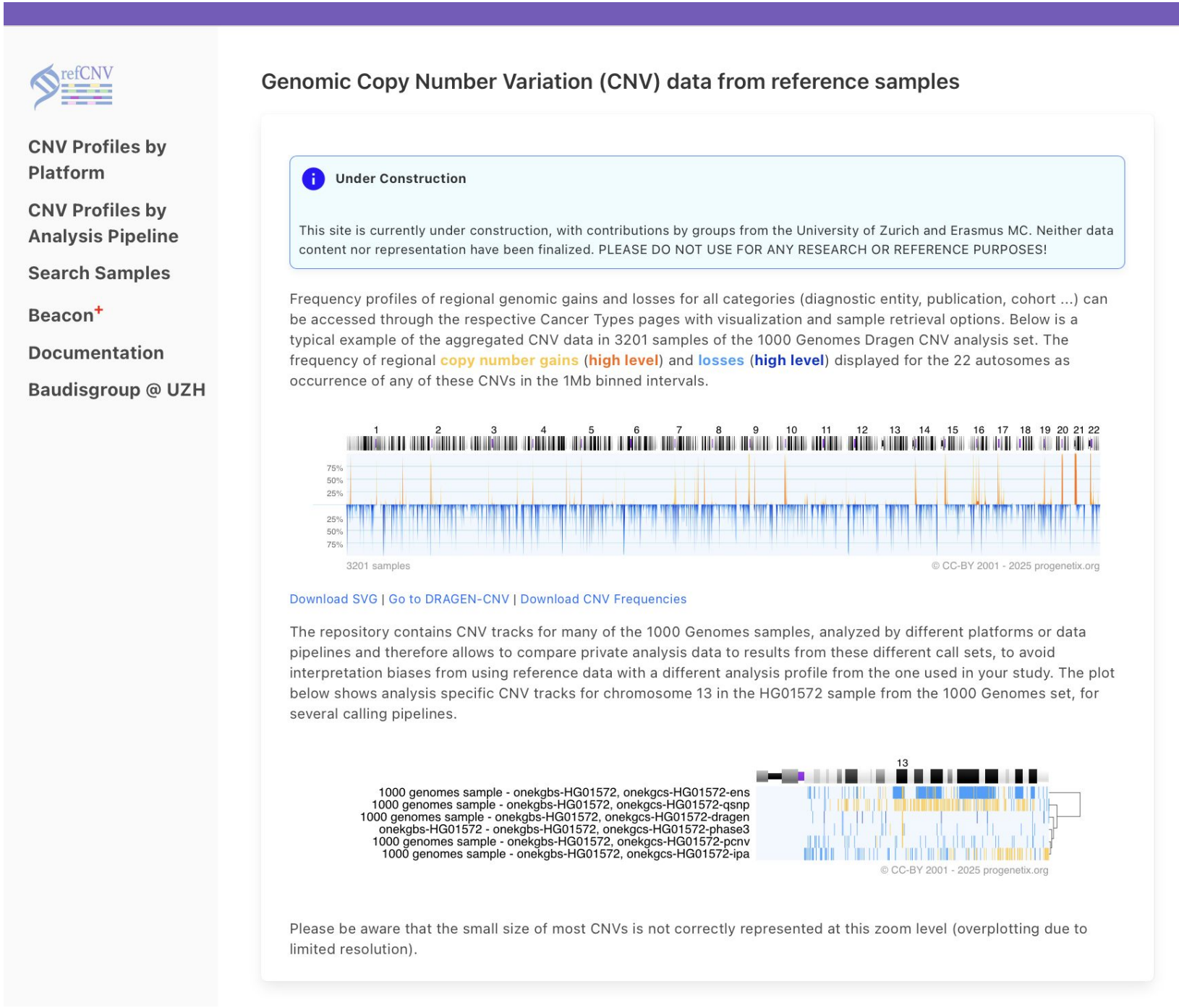
progenetix

SIB Swiss Institute of Bioinformatics


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Lung Adenocarcinoma
NCIT:C3512 (progenetix, 7432 samples)
Liver Epithelial Neoplasm
NCIT:C7106 (progenetix, 3944 samples)
Lung Neuroendocrine Neoplasm
NCIT:C5670 (progenetix, 974 samples)
Lung Neuroendocrine Neoplasm
NCIT:C5670 (cellz, 242 samples)
Lung Adenocarcinoma
NCIT:C3512 (cellz, 482 samples)
Liver Epithelial Neoplasm
NCIT:C7106 (cellz, 76 samples)
75%
50%
25%
13150 samples
© CC-BY 2001 - 2025 progenetix.org

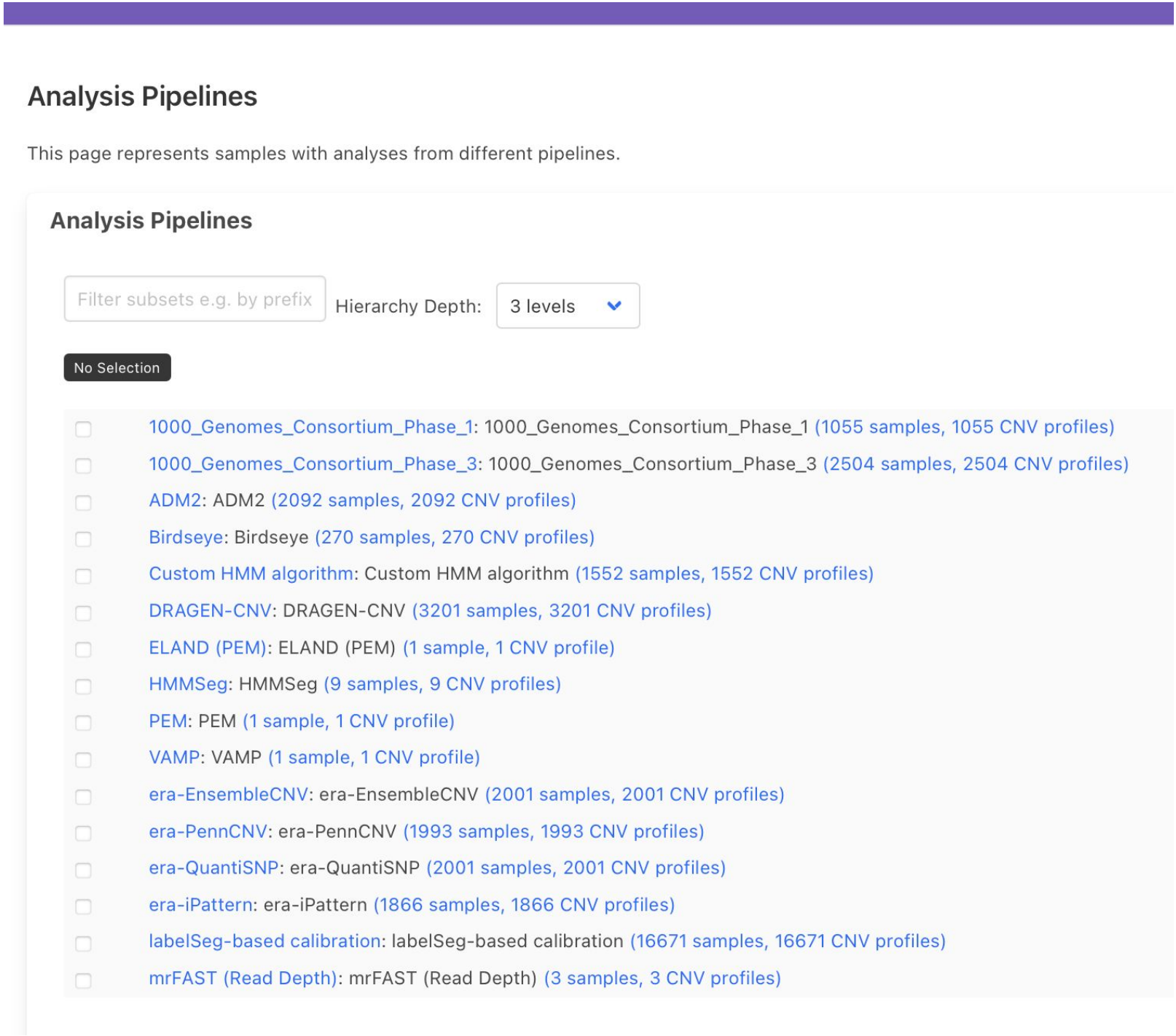
refCNV
Germline CNVs…

- germline CNVs vary widely by genomic background but importantly also by technical assessment
  - sequencing & array
  - bioinformatics workflows
- frequency based information can be misleading
- starting from a benchmarking project (A. Stubbs et al.) we started to do a "non-judgemental" collection w/ technical annotations
- Beacon API, obviously …



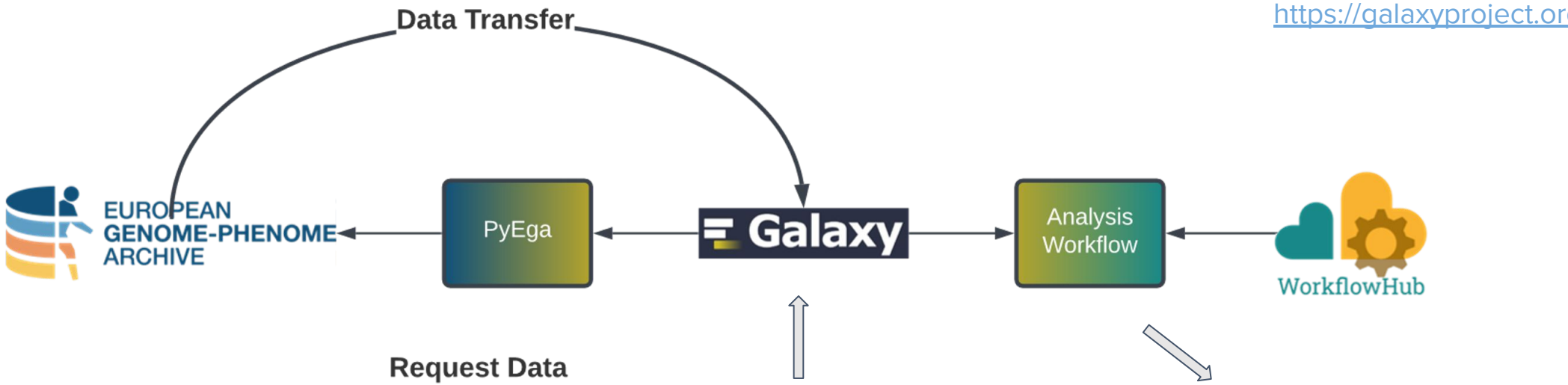
refCNV
Germline CNVs…

- germline CNVs vary widely by genomic background but importantly also by technical assessment
  - sequencing & array
  - bioinformatics workflows
- frequency based information can be misleading
- starting from a benchmarking project (A. Stubbs et al.) we started to do a "non-judgemental" collection w/ technical annotations
- Beacon API, obviously …



HTSgetGalaxy :FAIR CLOUD Analysis

Galaxy & GA4GH alignment:
https://galaxyproject.org/ga4gh/



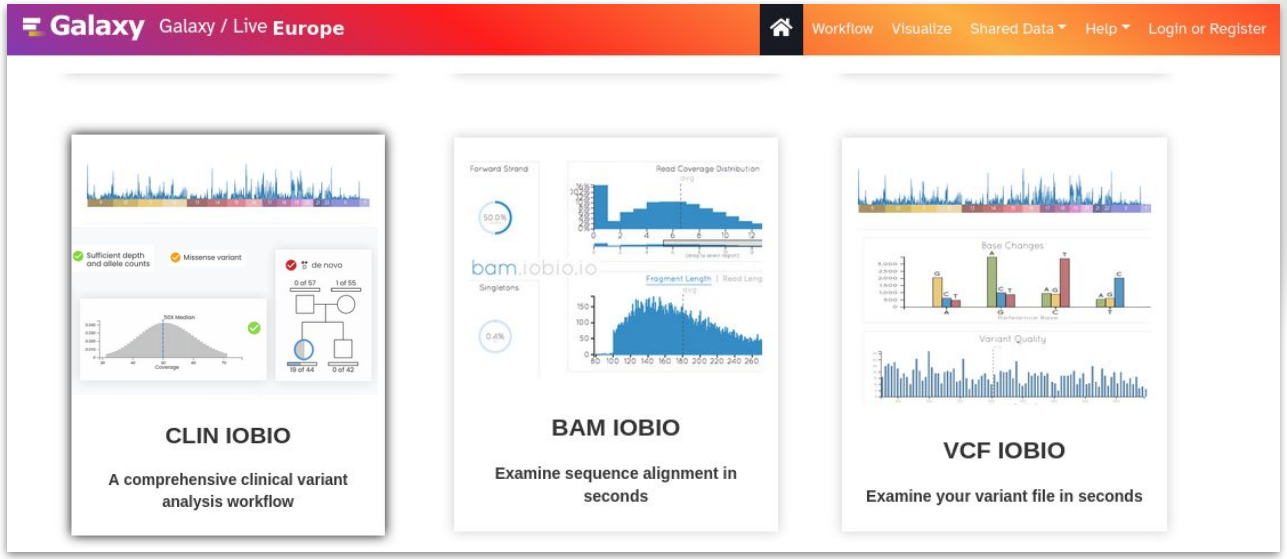
Log in with ELIXIR AAI

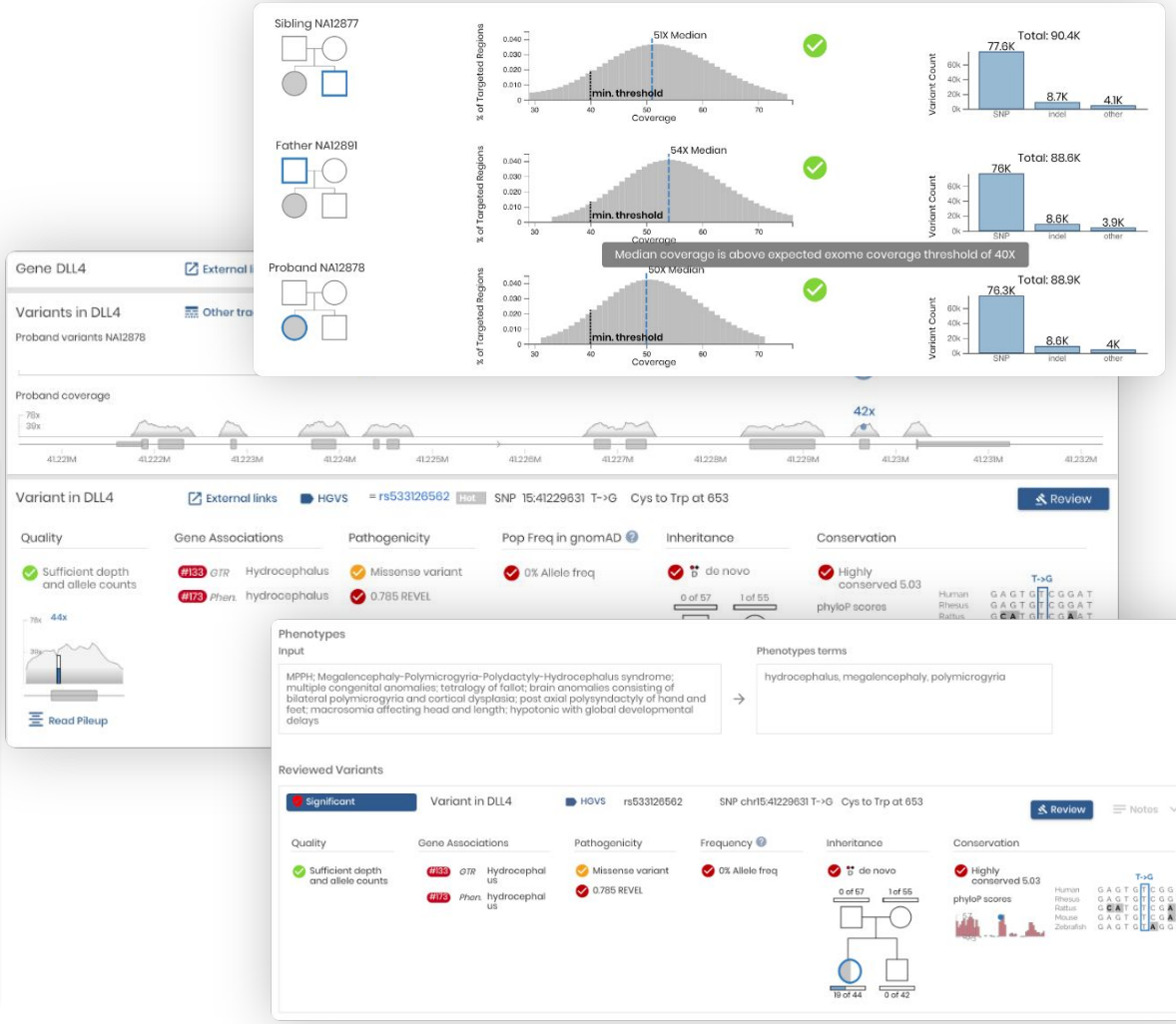
Authentication (passports/visa)

NOTE: Pilot implementation uses username/password for authentication, future release of Galaxy will support use of passports & visas

Data analysis & reporting: Galaxy clin.iobio trio analysis

- Galaxy Interactive Tool
- Docker based
- Interactive variant exploration

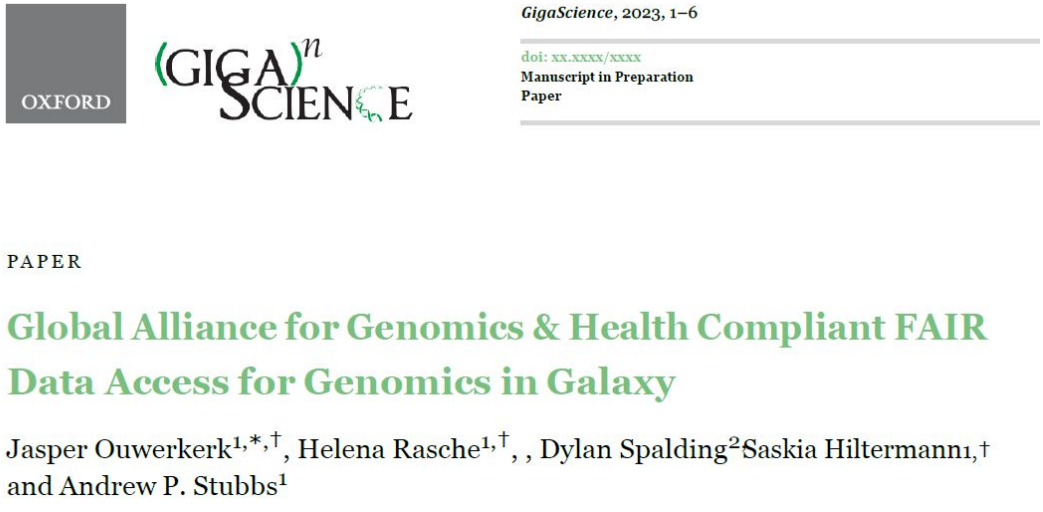


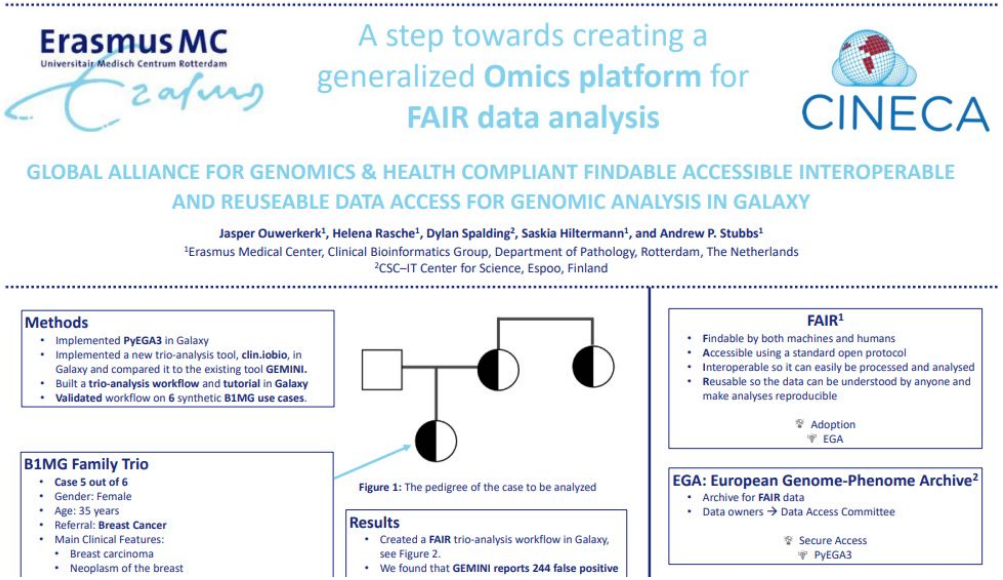


HTSgetGalaxy : Summary

We have implemented:

- Secure access to GA4GH EGA service using PyEGA3 Galaxy service
- Standard analysis for B1MG synthetic data (Filtering etc...)
- Interactive gene variant detection for trio-analysis with clin.iobio in Galaxy
- Related Galaxy Training Network (GTN)
- Single source of Cancer analysis tooling for scientific community – Cancer Galaxy (https://cancer.usegalaxy.eu/)





hCNV Galaxy and Nextflow

Galaxy

- Structural genomic variant calling tools with user-friendly interface for running tools and managing data.
- Galaxy is also a powerful platform for developing and sharing bioinformatics workflows deposited in the WorkflowHub.
- Galaxy training network tutorial (slides, hands on, workflows)

Nextflow

- Nextflow workflows/modules available to be used on a variety of compute platforms, including local machines, clusters, and clouds.
- Nextflow can be used to run hCNV on multiple samples in parallel, which can significantly reduce the amount of time required to analyze a large dataset.

Galaxy Training Network

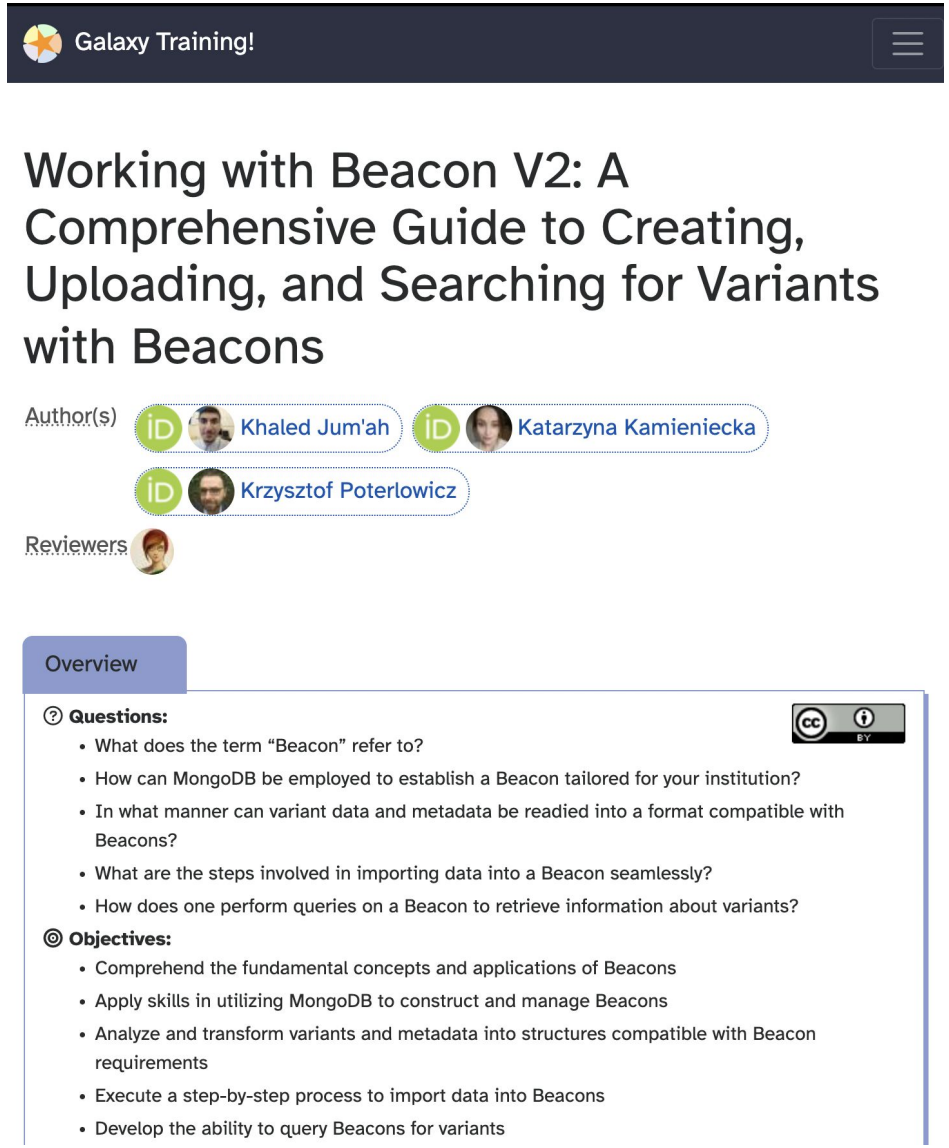
UNIVERSITY of BRADFORD

Control-FREEC Training material

- Define the specific challenges in locating human Copy Number Variances (hCNVs).
- Show how to pre-process the sequenced reads for hCNVs detection.
- Show an example on using Control-FREEC to detect the hCNVs from cancer dataset.
- Visualise the hCNVs' findings.

Control-FREEC workflow

- Publish the workflow on Galaxy and Workflow hub to be accessible by the community



Beacon v2 tools

UNIVERSITY of BRADFORD

Beacon2 Import

Run Tool

Import JSON formatted datasets to beacon database
(Galaxy Version 1.0.7+galaxy0)

Tool Parameters

DATABASE HOST - optional

127.0.0.1

Hostname/IP of the beacon database (--db-host)

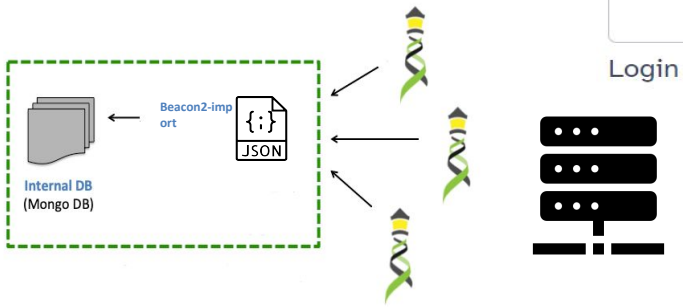
DATABASE PORT *

27017

Port of the beacon database (--db-port)

DATABASE PASSWORD - optional

Login password for the beacon database (--



Beacon2 Bracket

Run Tool

Specifies a sequence ranges for both the start and end positions of a genomic variation
(Galaxy Version 1.0.7+galaxy0)

Tool Parameters

DATABASE HOST - optional

127.0.0.1

Hostname/IP of the beacon database (--db-host)

Beacon2 Gene

Run Tool

Queries the beacon database and retrieve the genomic variants matching gene symbol
(Galaxy Version 1.0.7+galaxy0)

Tool Parameters

DATABASE HOST - optional

127.0.0.1

Hostname/IP of the beacon database (--db-host)

Beacon2 Sequence

Run Tool

Query for the existence of a specified sequence at a given genomic position
(Galaxy Version 1.0.7+galaxy0)

Tool Parameters

DATABASE HOST - optional

127.0.0.1

Hostname/IP of the beacon database (--db-host)

Beacon2 Range

Run Tool

Retrieve the genomic variants from the beacon database by specifying start and end positions
(Galaxy Version 1.0.7+galaxy0)

Tool Parameters

DATABASE HOST - optional

127.0.0.1

Hostname/IP of the beacon database (--db-host)

hCNV data analyses with Nextflow modules

UNIVERSITY of BRADFORD

modules Public

Forked from nf-core/modules

Repository to host tool-specific module files for the Nextflow DSL2 community!

Star

biohackeu23 {B}{H} nf-core

Nextflow MIT License Updated last month

new module: beacon2-ri-tools #4264

Open **kkamieniecka** opened this issue last month · 0 comments

kkamieniecka commented last month Member

beacon2-ri-tools transform genomic variations data (VCF) to queryable data (MongoDB)
See extended documentation here.

+ Add tasklist

kkamieniecka added the new module label last month

elixir

bioinformatics and ML tools for CNV assessment

…. Antonio

CNV annotation & Query Standards Development

GA4GH VRS & Beacon Scouts

- unambiguous representation and recovery of genomic variation data still represent major problems in large scale genomic studies (meta-analyses and data federated data discovery)
- structural genome variants present additional challenges both on epistemic and technical levels
- GA4GH has emerged as the major driver for standards development and harmonization
- ELIXIR hCNV members contribute to the development and improvement of CNV representation and query options as well as associated vocabularies
  - VCF 4.4+ D and J flags (e.g. indicating dosage w/o locus assignment)
  - EFO relative copy number count term tree
  - VRS v1.3/v2 representation of copyNumberChange and copyNumberCount
  - Beacon query types specifically suited for CNV or "fuzzy location" events

CNV annotation & Query Standards Development

GA4GH VRS & Beacon Scouts: CNV Classes

GA4GH VRS1.3+	Beacon	VCF v4.4+	SO
EFO:0030070 gain	DUP or **EFO:0030070**	DUP SVCLAIM=D	SO:0001742 copy_number_gain
EFO:0030071 low-level gain	DUP or **EFO:0030071**	DUP SVCLAIM=D	SO:0001742 copy_number_gain
EFO:0030072 high-level gain	DUP or **EFO:0030072**	DUP SVCLAIM=D	SO:0001742 copy_number_gain
EFO:0030072 high-level gain	DUP or EFO:0030073	DUP SVCLAIM=D	SO:0001742 copy_number_gain
EFO:0030067 loss	DEL or **EFO:0030067**	DEL SVCLAIM=D	SO:0001743 copy_number_loss
EFO:0030068 low-level loss	DEL or **EFO:0030068**	DEL SVCLAIM=D	SO:0001743 copy_number_loss
EFO:0020073 high-level loss	DEL or **EFO:0020073**	DEL SVCLAIM=D	SO:0001743 copy_number_loss
EFO:0030069 complete genomic loss	DEL or **EFO:0030069**	DEL SVCLAIM=D	SO:0001743 copy_number_loss

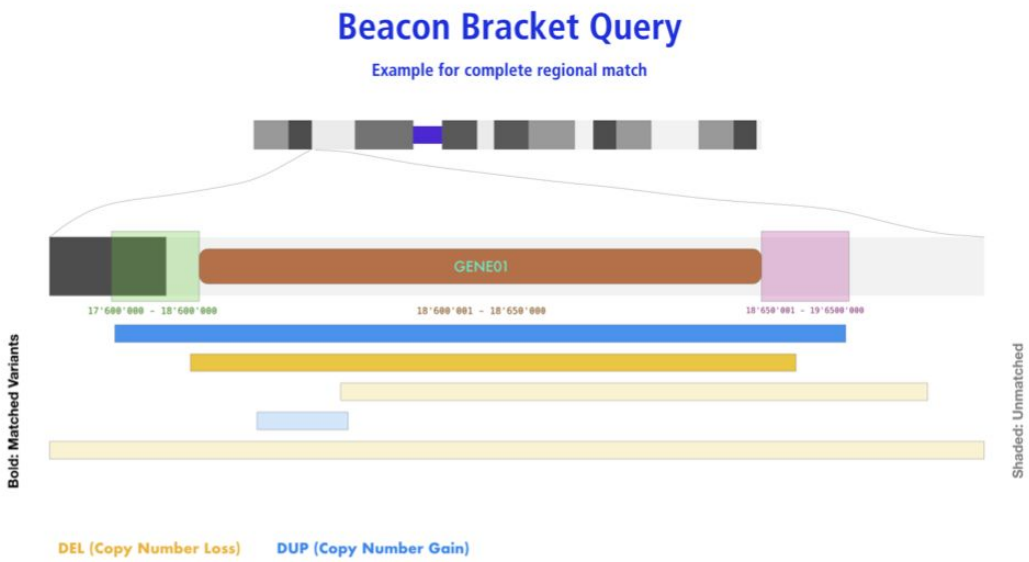
CNV annotation & Query Standards Development

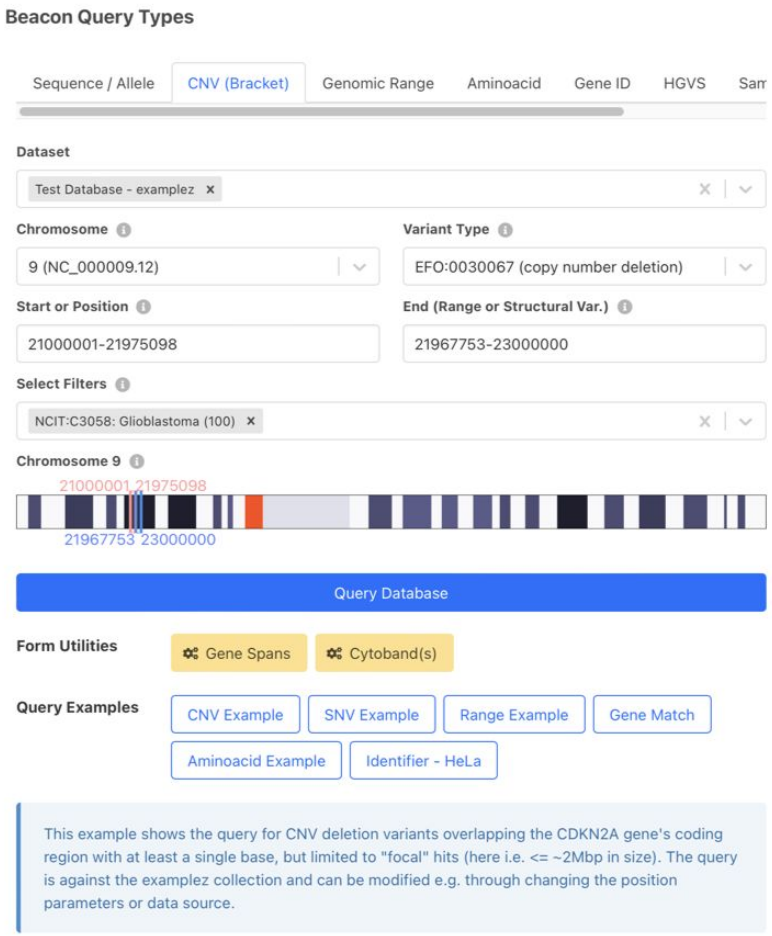
Beacon Scouts: CNV Queries

Variation Queries

Bracket ("CNV") Query

- defined through the use of 2 start, 2 end
- any contiguous variant...





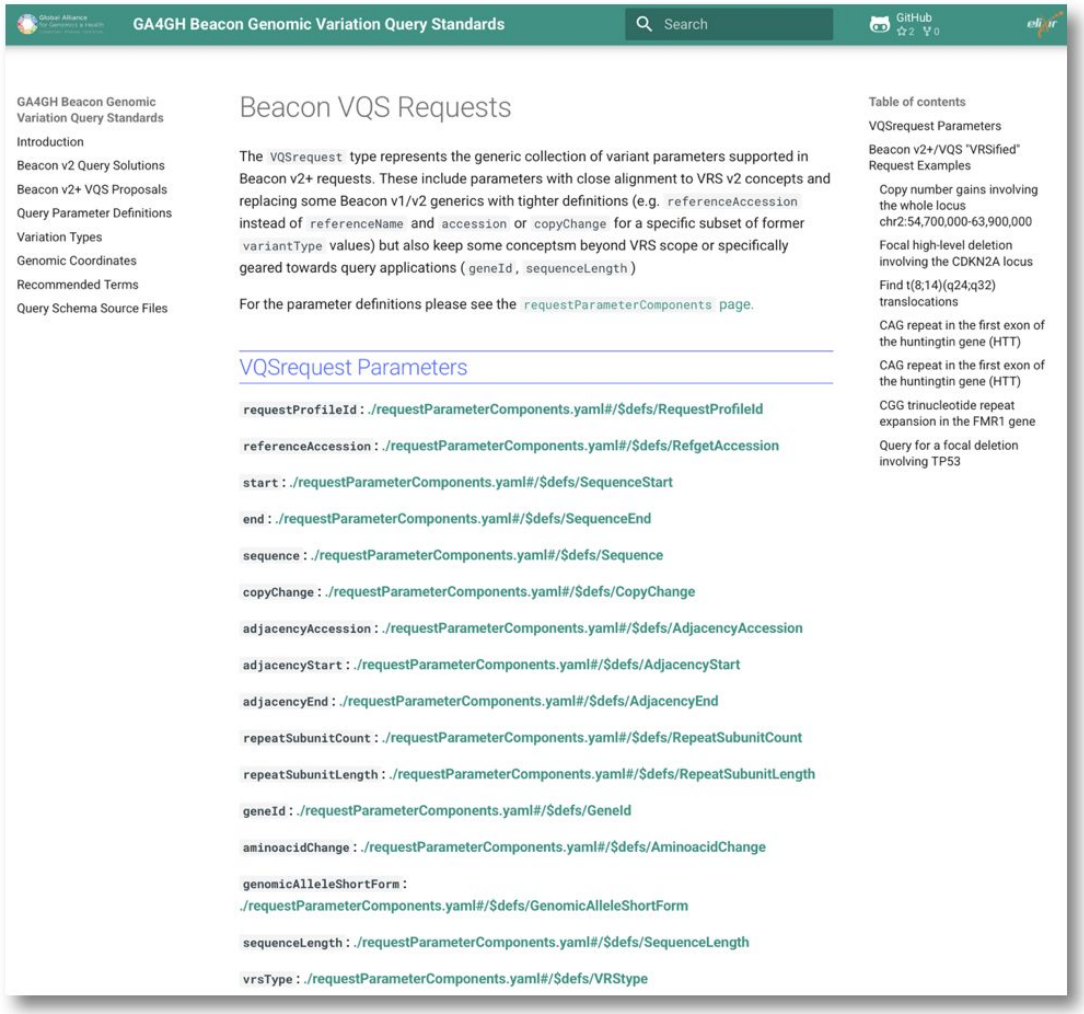
CNV annotation & Query Standards Development

Beacon Scouts: Moving Ahead

Beacon Scouts

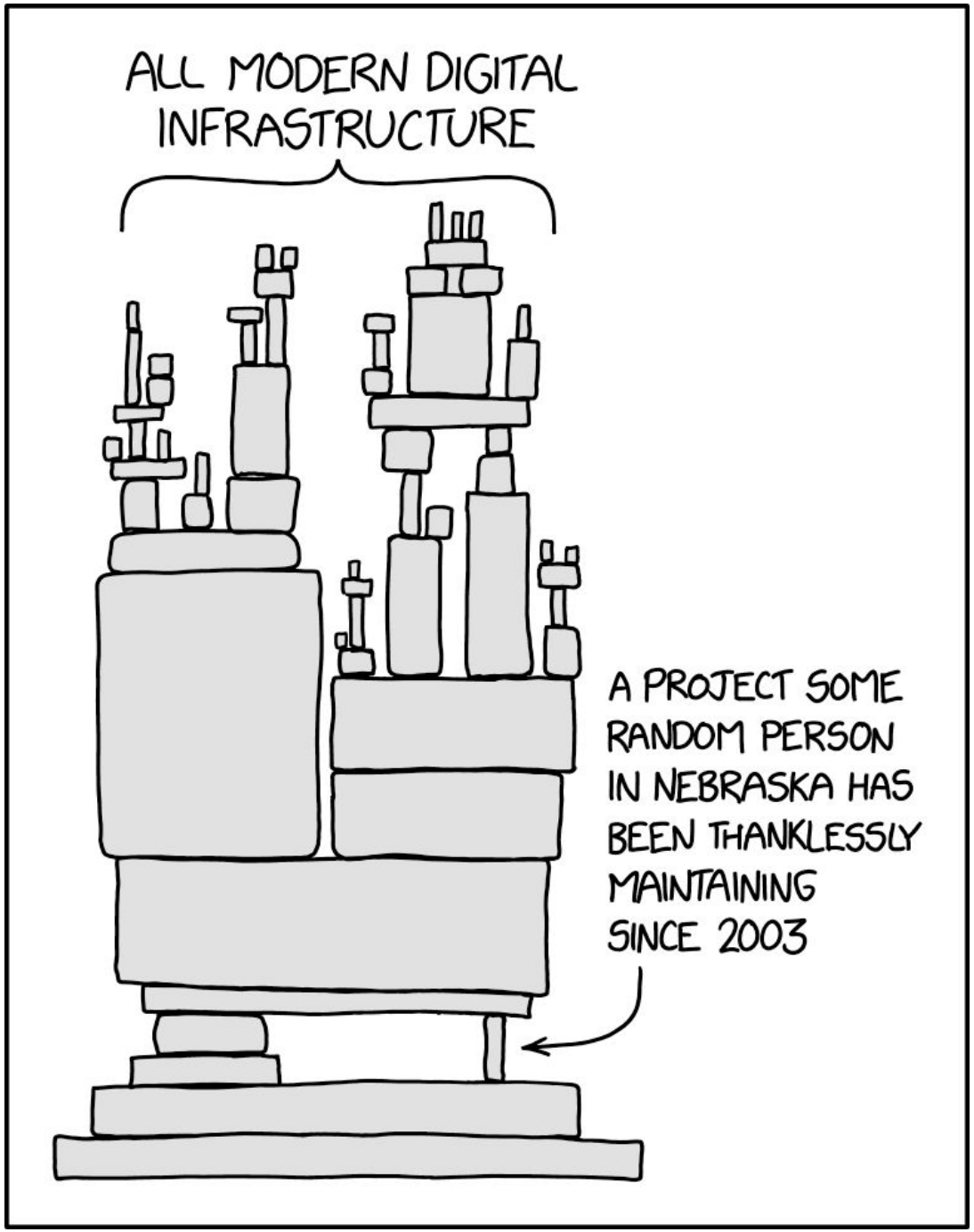
Finding the Paths to Beacon's Future

- Genomic Variation Scouts
  - ➡ extension to the query model based on assessed needs
    - fusions/breakpoints, cytogenetic annotations, repeats, categorical variants...
  - ➡ adoption of evolving VRS... standards for variant representation
    - adjacency, repeats…
    - re-use of parameters where clear (e.g. **`sequenceLength`** instead of **`variantMinLength`** + **`variantMaxLength`**)



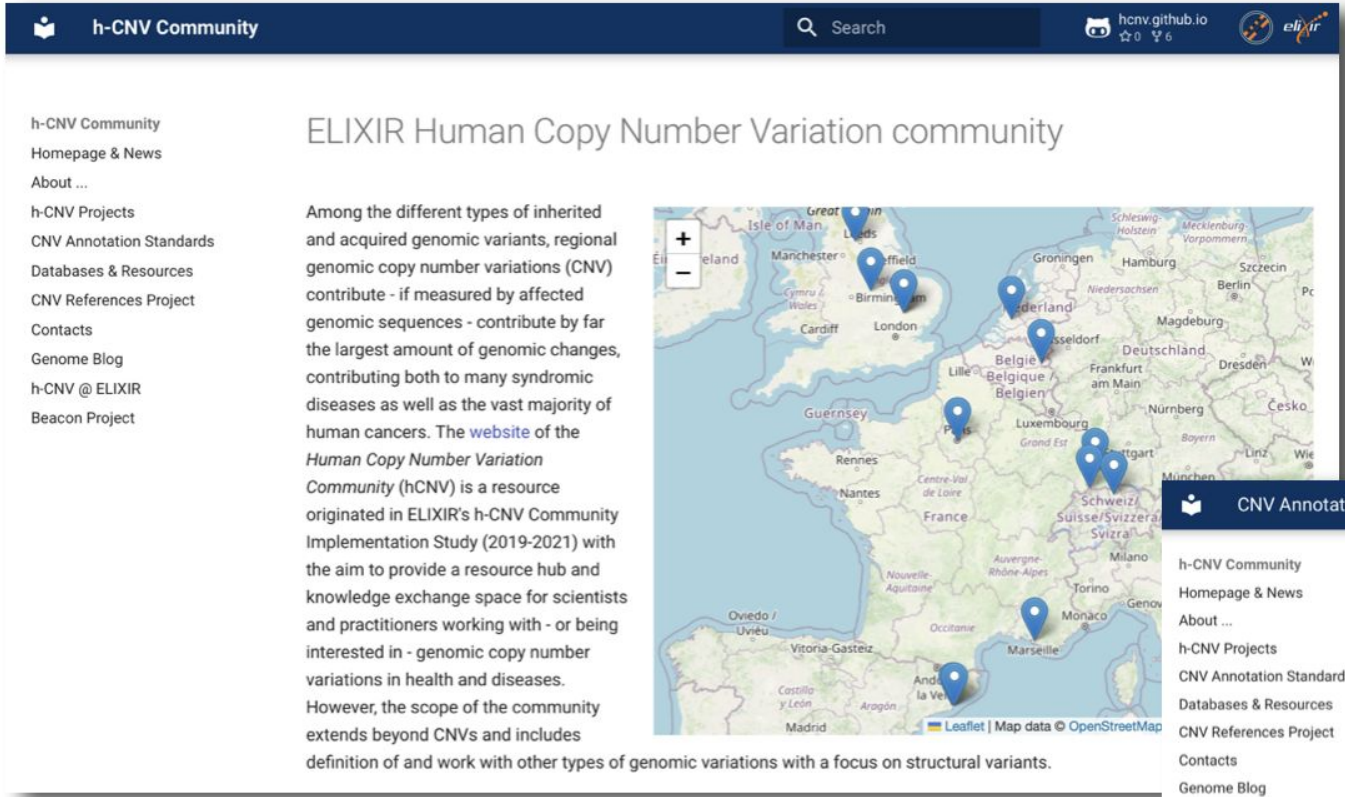
https://genomebeacons.org/variant-query-types/variant-scouts-home/

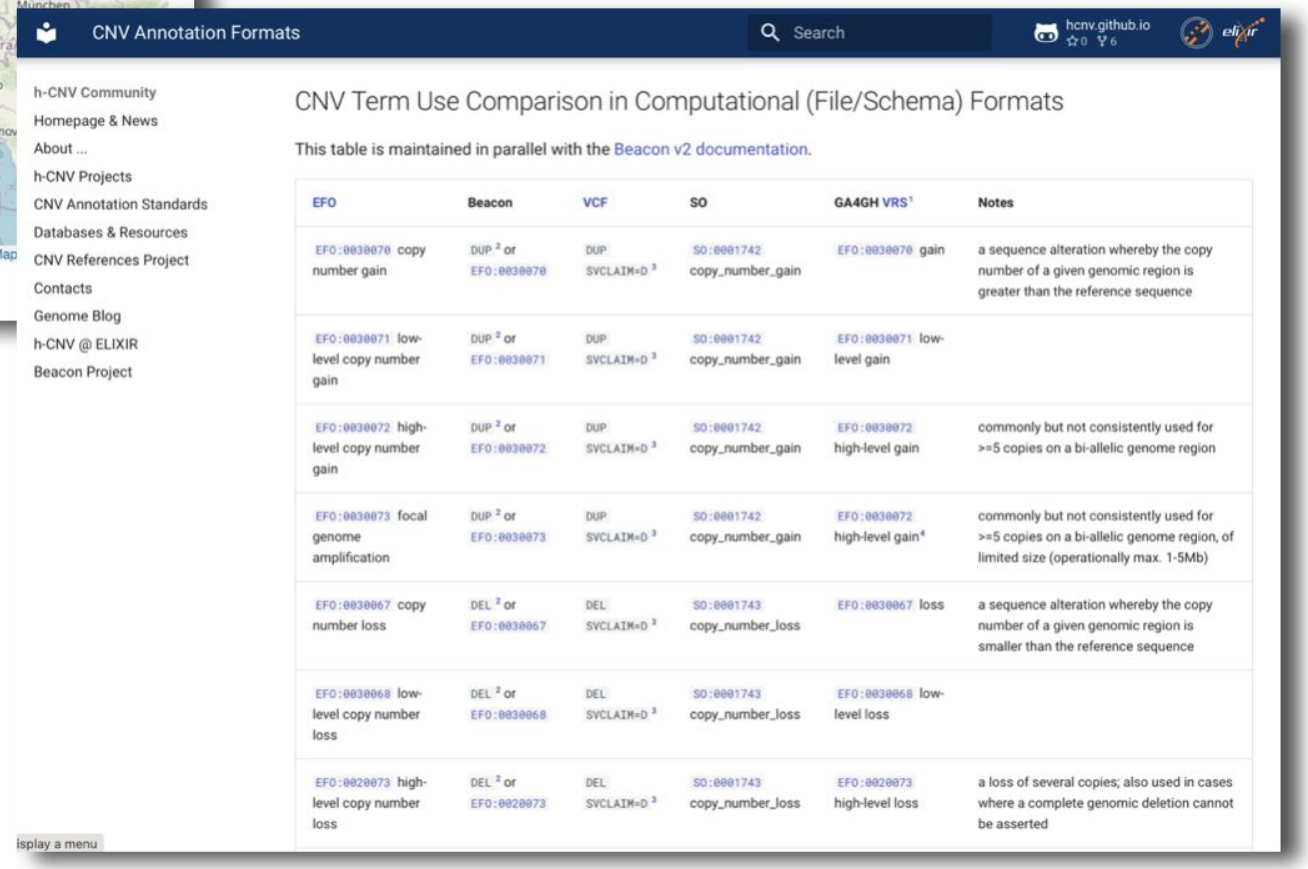
Contribute
Participate
Exchange
Maintain



ELIXIR hCNV Community

https://cnvar.org/





Thank you

www.elixir-europe.org

@ELIXIREurope

/company/elixir-europe