Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

Universität Zürich^UZH

Beaconize this: Databases for Cancer Genomics

and the

Development of Open Data Standards

Michael Baudis

Professor of Bioinformatics
University of Zürich
Swiss Institute of Bioinformatics **SIB**
GA4GH Workstream Co-lead *DISCOVERY*
Co-lead ELIXIR Beacon API Development
Co-lead ELIXIR hCNV Community

SIB
Swiss Institute of
Bioinformatics

elixir

Theoretical Cytogenetics and Oncogenomics

Cancer Genomics | Data Resources | Methods & Standards for Genomics and Personalized Health

Michael Baudis

Curators

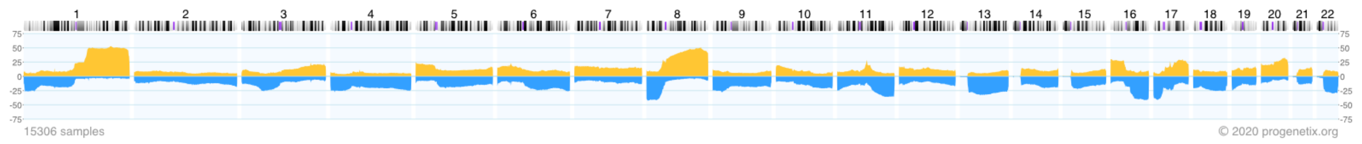
Data ~~Parasites~~

Theoretical Cytogenetics and Oncogenomics

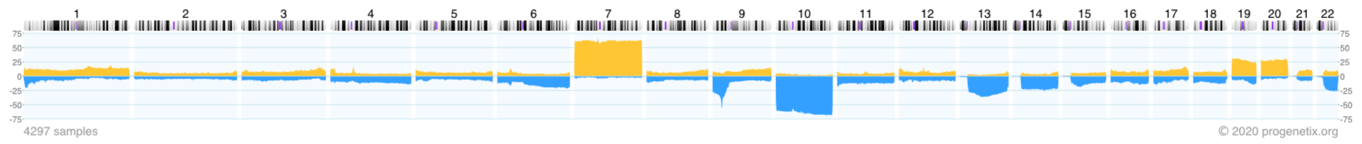
... but what does this entail @baudisgroup?

- patterns & markers in cancer genomics, especially somatic structural genome variants

Malignant Breast Neoplasm (NCIT:C9335)

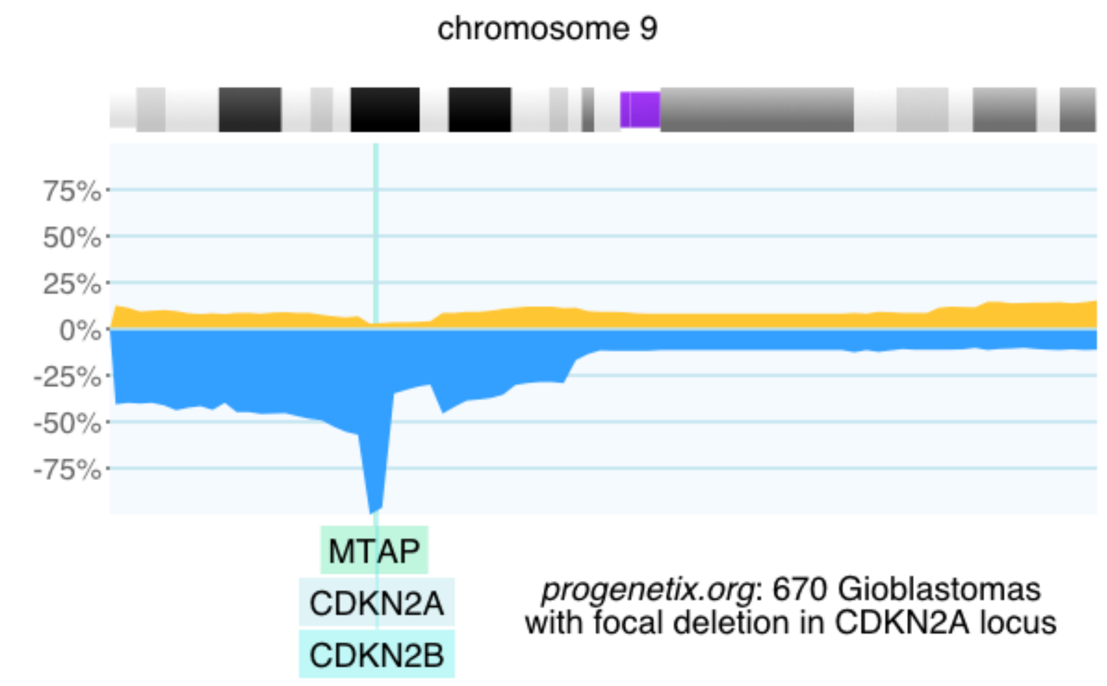


Glioblastoma (NCIT:C3058)



- bioinformatics support in collaborative studies
- reference resources for curated cancer genome variations
- bioinformatics tools & methods
- standards and reference implementations for data sharing in genomics and personalized health
- open research data "ambassadoring"

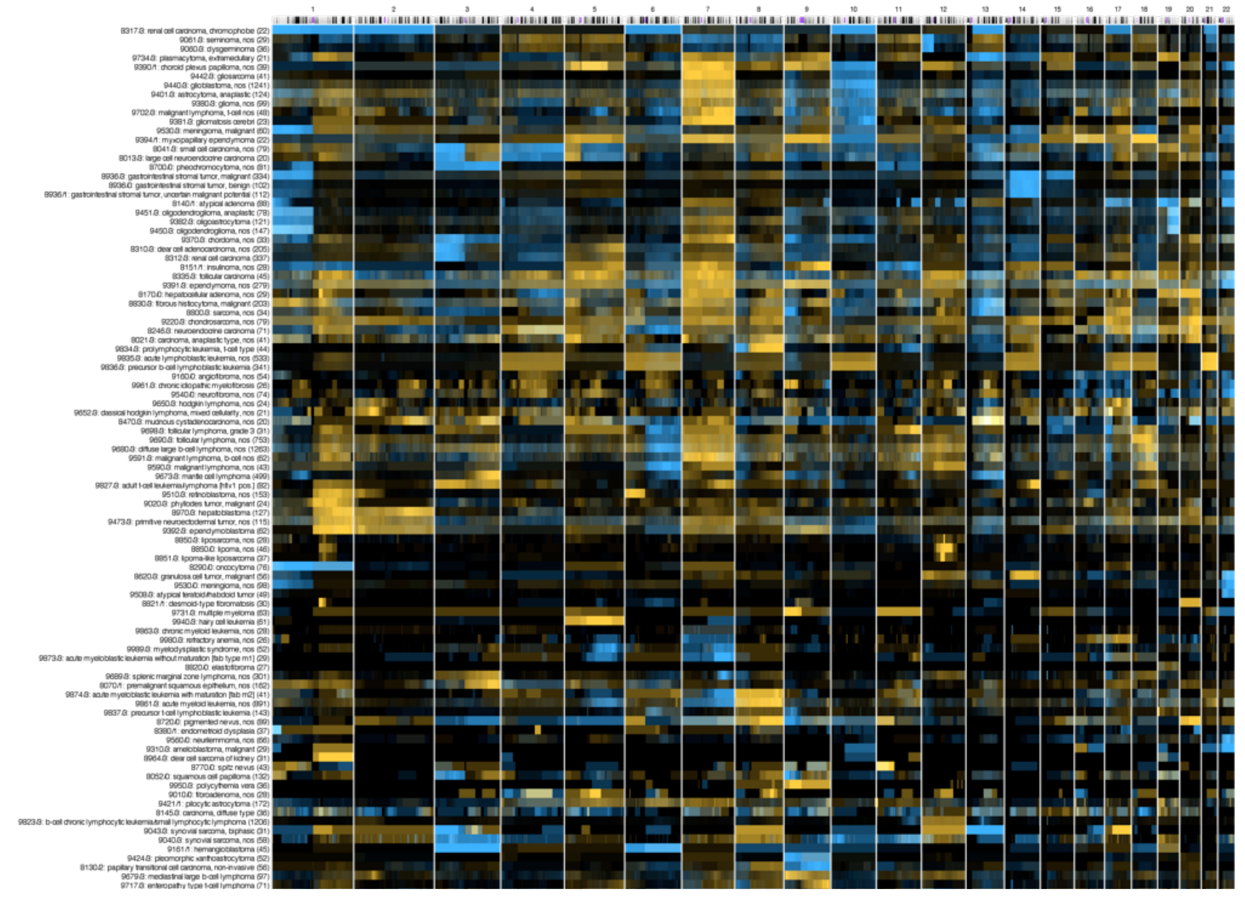
Theoretical Cytogenetics and Oncogenomics Research | Methods | Standards

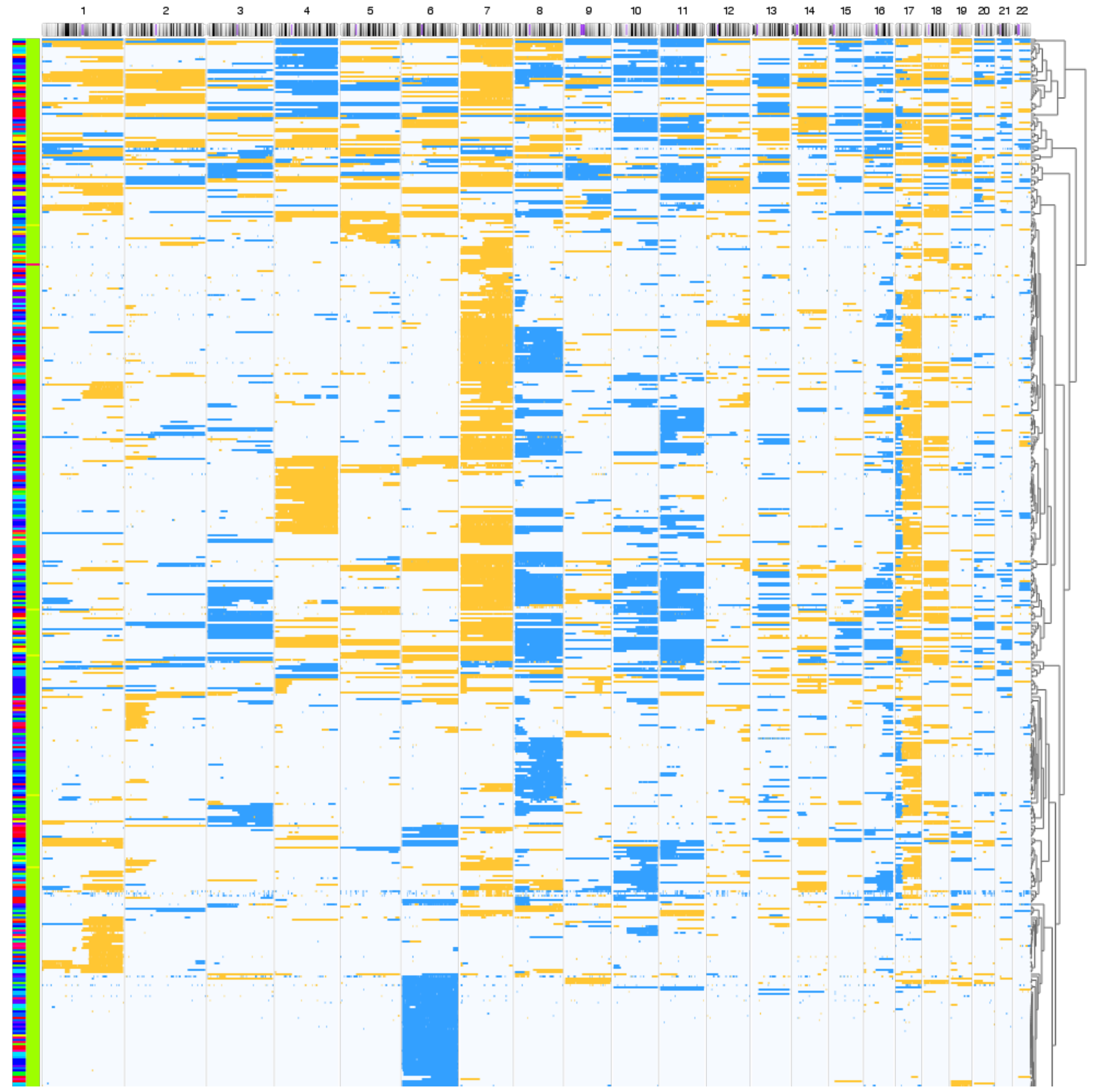


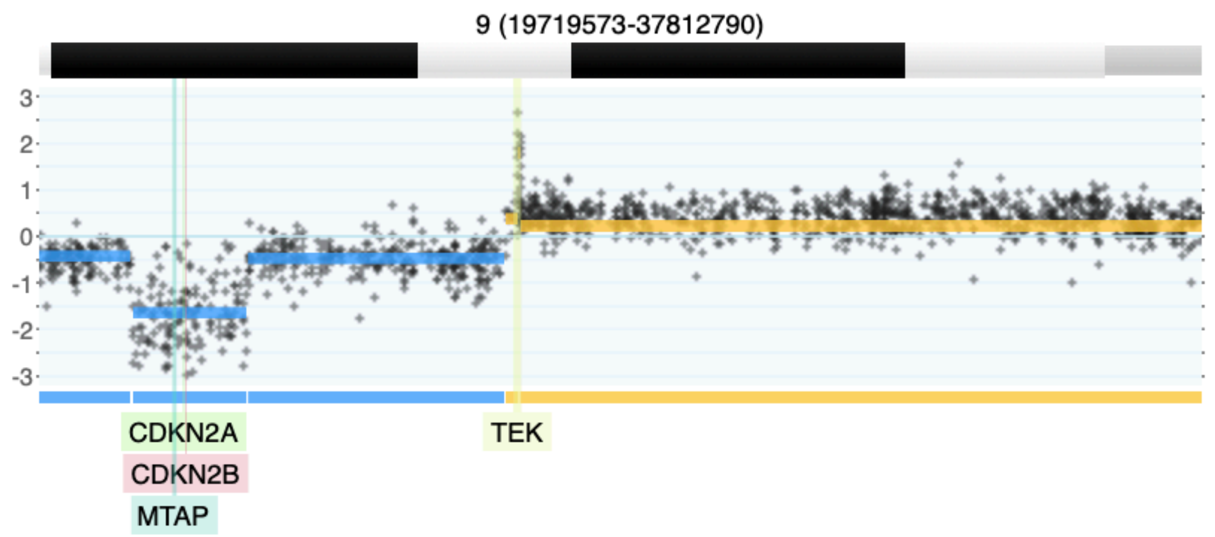
progenetix.org: 670 Gioblastomas with focal deletion in CDKN2A locus

Genomic Imbalances in Cancer - Copy Number Variations (CNV)

- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)



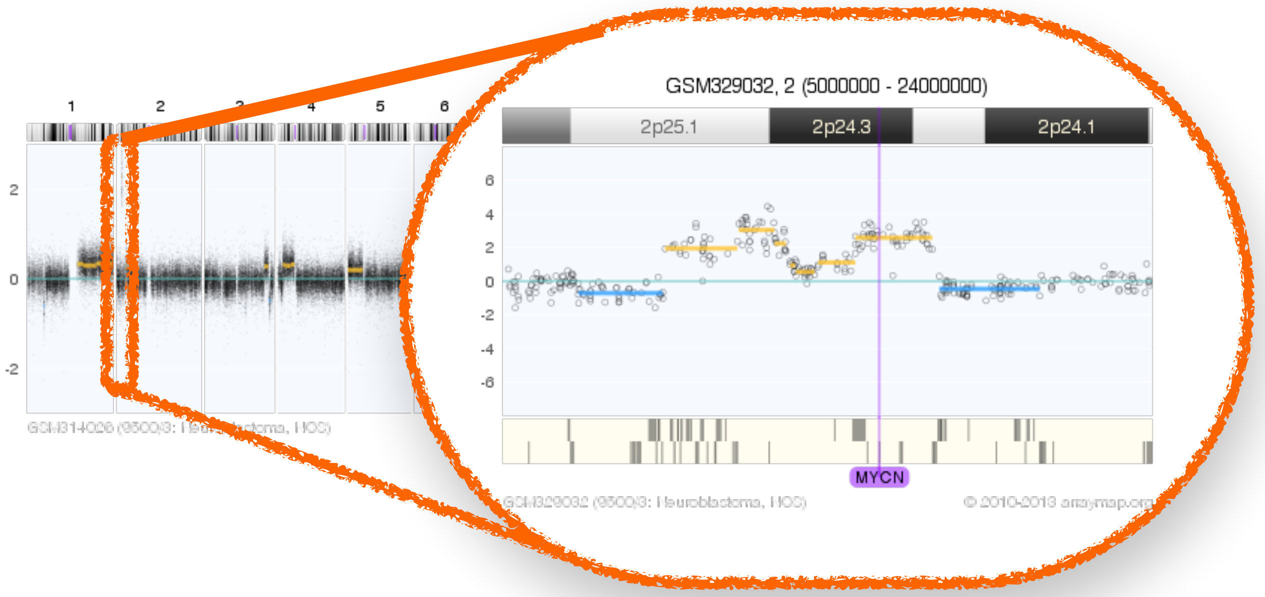




pgxcs::GSE19612::GSM491153

© 2019 progenetix.org

2-event, homozygous deletion in a Glioblastoma



MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

progenetix.net: storage and visualization of genomic aberration data in human malignancies

michael baudis, md

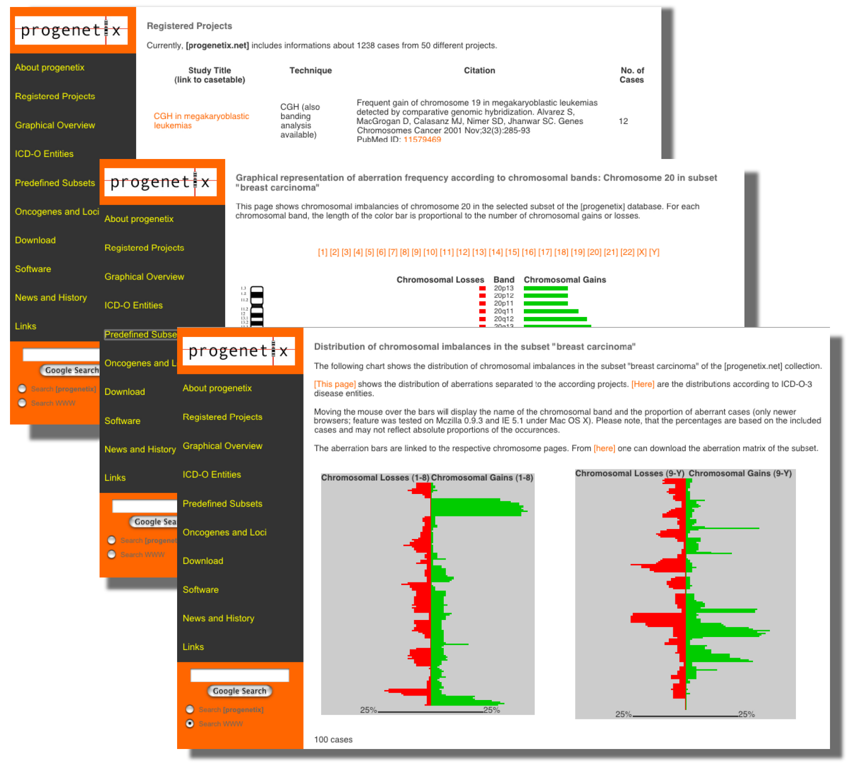
progenetix.net

Over the last decade, techniques for the genome wide scanning for genomic imbalances in malignant neoplasia have been developed, e.g. Comparative Genomic Hybridization (CGH).

Currently, no comprehensive online source for CGH data with a standardized format suitable for data mining procedures has been made available for public access. Such a data repository could be valuable in identifying genetic aberration patterns with linkage to specific disease entities, and provide additional information for validating data from large scale expression array experiments.

A case and band specific aberration matrix was selected as most suitable format for the mining of CGH data. The [progenetix.net] data repository was developed to provide the according data to the research community for a growing number of human malignancies.

In the current implementation, two main purposes are being served. First, access to the band specific pattern of chromosomal imbalances allows the instantaneous identification of genomic “hotspots”. Second, the band specific aberration matrices can be included in data mining efforts. As an example, the clustering off all informative cases from the current (September 2001) dataset is shown here (online source under www.progenetix.net/bcats/clustered.png).



Registered Projects

Currently, **[progenetix.net]** includes informations about 1238 cases from 50 different projects.

Study Title (link to casetable)	Technique	Citation	No. of Cases
CGH in megakaryoblastic leukemias	CGH (also banding analysis available)	Frequent gain of chromosome 19 in megakaryoblastic leukemias detected by comparative genomic hybridization. Alvarez S, MacGrogan D, Calasanz MJ, Nimer SD, Jhanwar SC. Genes Chromosomes Cancer 2001 Nov;32(3):285-93 PubMed ID: 11579469	12

Graphical representation of aberration frequency according to chromosomal bands: Chromosome 20 in subset "breast carcinoma"

This page shows chromosomal imbalancies of chromosome 20 in the selected subset of the [progenetix] database. For each chromosomal band, the length of the color bar is proportional to the number of chromosomal gains or losses.

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [22] [X] [Y]

Distribution of chromosomal imbalances in the subset "breast carcinoma"

The following chart shows the distribution of chromosomal imbalances in the subset "breast carcinoma" of the [progenetix.net] collection.

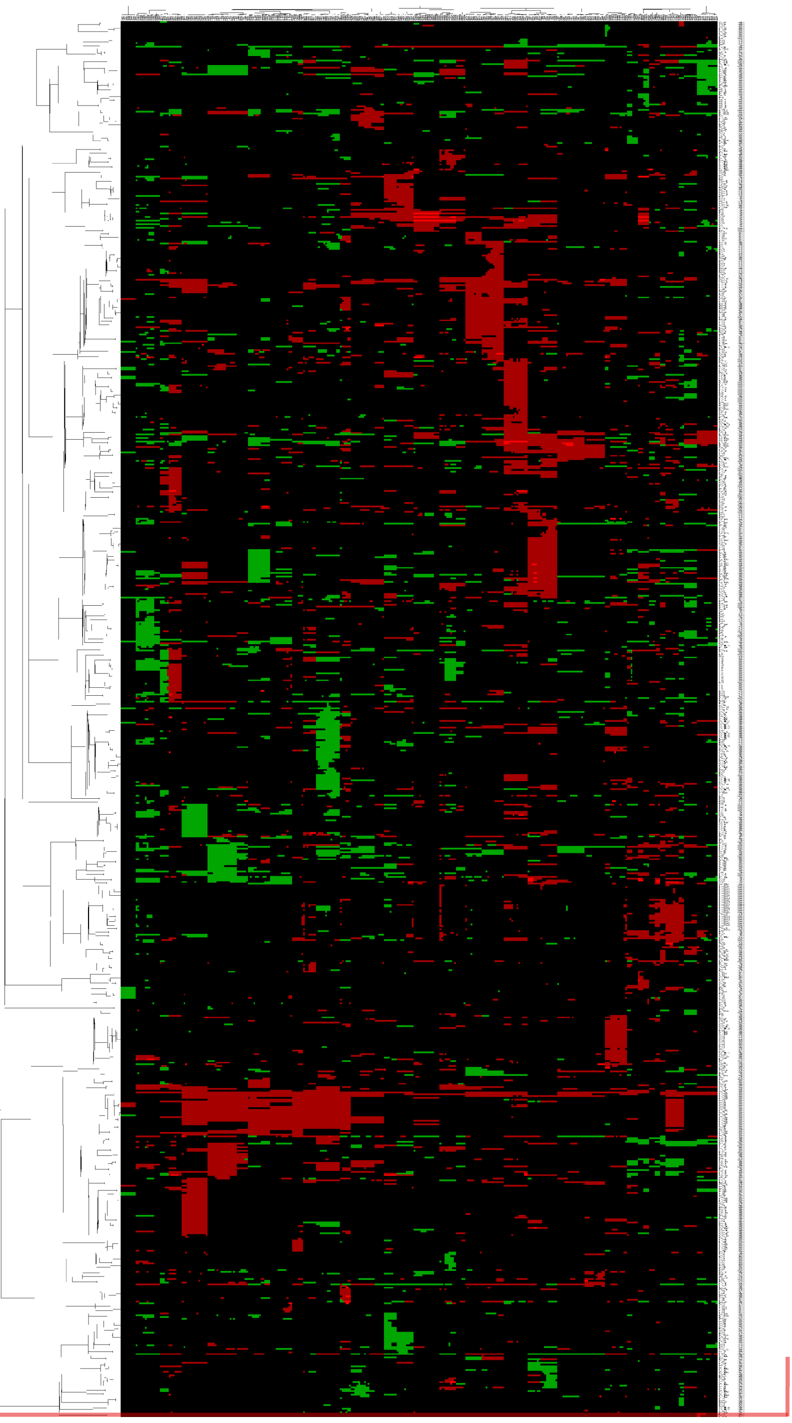
[This page] shows the distribution of aberrations separated to the according projects. [Here] are the distributions according to ICD-O-3 disease entities.

Moving the mouse over the bars will display the name of the chromosomal band and the proportion of aberrant cases (only newer browsers; feature was tested on Mozilla 0.9.3 and IE 5.1 under Mac OS X). Please note, that the percentages are based on the included cases and may not reflect absolute proportions of the occurences.

The aberration bars are linked to the respective chromosome pages. From [here] one can download the aberration matrix of the subset.

Chromosomal Losses (1-8) Chromosomal Gains (1-8)
25%
25%
Chromosomal Losses (9-Y) Chromosomal Gains (9-Y)
25%
25%

100 cases



Hierarchical clustering of band specific chromosomal imbalances from 999 human neoplasias, contained in the [progenetix.net] collection. Cases without aberrations were excluded.

Data selection

PubMed is searched for publications applying CGH to the analysis of malignant tumors. Articles are selected according to their online availability and the description of genomic imbalances on a per case basis.

Transformation of input data

Chromosomal aberration data is transformed via customized parsing commands to a common format adherent to ISCN 1995 recommendations. In some cases, aberration data was transcribed from graphical representations or provided by the authors.

Data storage

Currently, the primary data is stored in a dedicated “off-line” database. Besides case identifier and ISCN adapted chromosomal imbalance data, tumor classification and source information including the PubMed identifier is recorded. Disease entities are reclassified to ICD-O-3 codes.

Text parsing and generation of aberration matrix

For the generation of the case and band specific aberration matrix, a dedicated text pattern comparison model was developed using Perl. Briefly, for each chromosomal band, the aberration field of each case is searched for a variety of patterns containing aberration information applying to that band. A matrix with currently 324 band resolution is generated, annotating chromosomal gains with “1” and losses with “–1”; localized high-level gains are designated “2“.

Website generation

For graphical representation of chromosomal imbalances, HTML pages containing different views of the underlying aberration matrices are generated using Perl. Graphics are implemented using HTML syntax. Besides band specific, whole genomic overviews, chromosome specific pages with links to all involved cases are generated for each ICD-O-3 entity as well as for each registered project. Additionally, those representations are available for several subsets combining related data (e.g. all lymphoid neoplasias, breast carcinoma cases). For each of the groups, the according aberration matrix is linked for download.

Michael Baudis | BCATS Biocomputing at Stanford I | Stanford Nov 2001

BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 12 2001
Pages 1228–1229

Progenetix.net: an online repository for molecular cytogenetic aberration data

Michael Baudis [1, 2,*] *and Michael L. Cleary* [2]

[1] *Medizinische Klinik und Poliklinik V der Universität Heidelberg, Germany and*
[2] *Department of Pathology, Stanford University Medical Center, Stanford, CA 94305, USA*

Received on July 5, 2001; revised on July 9, 2001; accepted on July 16, 2001

ABSTRACT

Summary: Through sequencing projects and, more recently, array-based expression analysis experiments, a wealth of genetic data has become accessible via online resources. In contrast, few of the (molecular-) cytogenetic aberration data collected in the last decades are available in a format suitable for data mining procedures. www.progenetix.net is a new online repository for previously published chromosomal aberration data, allowing the addition of band-specific information about chromosomal imbalances to oncologic data analysis efforts.
Availability: http://www.progenetix.net
Contact: mbaudis@stanford.edu

Neoplastic transformation and progression is the result of genetic defects arising in normal cells and giving rise to a malignant clone. During the process of oncogenesis, some of the usually multiple steps required for acquisition of the full neoplastic phenotype may represent themselves as numerical or structural abnormalities in the chromosomes of the transformed cells.

Over the last decades, the analysis of chromosomal abnormalities in malignant cells has gained importance in oncologic research as well as in clinical practice. A vast number of genetic abnormalities has been identified in the virtually complete range of human neoplasias. Several attempts have been undertaken for collection and classification of those abnormalities, the most widely recognized being the catalog by Mitelman and co-workers (Mitelman, 1994; online access through http://cgap.nci.nih.gov/Chromosomes/Mitelman).

In addition to metaphase analysis of short-term cultivated tumor cells or tumor cell lines, molecular cytogenetic techniques have recently been applied to the analysis of chromosomal abnormalities in primary tumor tissues. One of the more widely used screening techniques is Comparative Genomic Hybridization (CGH; Kallioniemi *et al.*, 1992; du Manoir *et al.*, 1993). Briefly, this method is based on the competitive *in-situ* hybridization of differentially labeled tumor versus normal genomic DNA to normal human metaphase spreads. The calculation of the intensity ratios of the two fluorochromes gives an overview about relative gains and losses of DNA in the tumor genome with mapping to the respective chromosomal bands. The identification of frequently imbalanced regions in tumor entities may point towards tumor suppressor gene or proto-oncogenes mapping to the respective chromosomal bands. Usually, the result of those experiments is communicated either in text format according to the International System for Cytogenetic Nomenclature (Mitelman, 1995) or graphically, with aberration bars next to chromosomal ideograms for the representation of chromosomal gains and losses.

Because in each experiment CGH analysis covers the whole number of chromosomes, the comparision of data sets from related malignancies could lead to the delineation of common as well as divergent genetic pathways defining the respective malignant phenotypes. Although an extremely large number of malignant tumors has been analyzed using this technique, no comprehensive CGH database with band-specific chromosomal aberration information is publicly available[†].

A minimal requirement for such a database would be the conversion of the text or graphical information used in publications to data tables, representing the information about the aberration status of single chromosomal bands for each case. For the site discussed here, this process includes: (1) the transformation of the published results in a format adapted from the ISCN, and (2) the automatic generation of the band specific aberration table.

Due to format variations of the published data, step 1 consists of the manual conversion of the text data or evaluation and conversion of the graphical representations, respectively. Due to the (in computational terms) odd

[*] To whom correspondence should be addressed.

[†] Links to a number of online CGH resources with different scopes can be found at www.progenetix.net.

progenetix.org

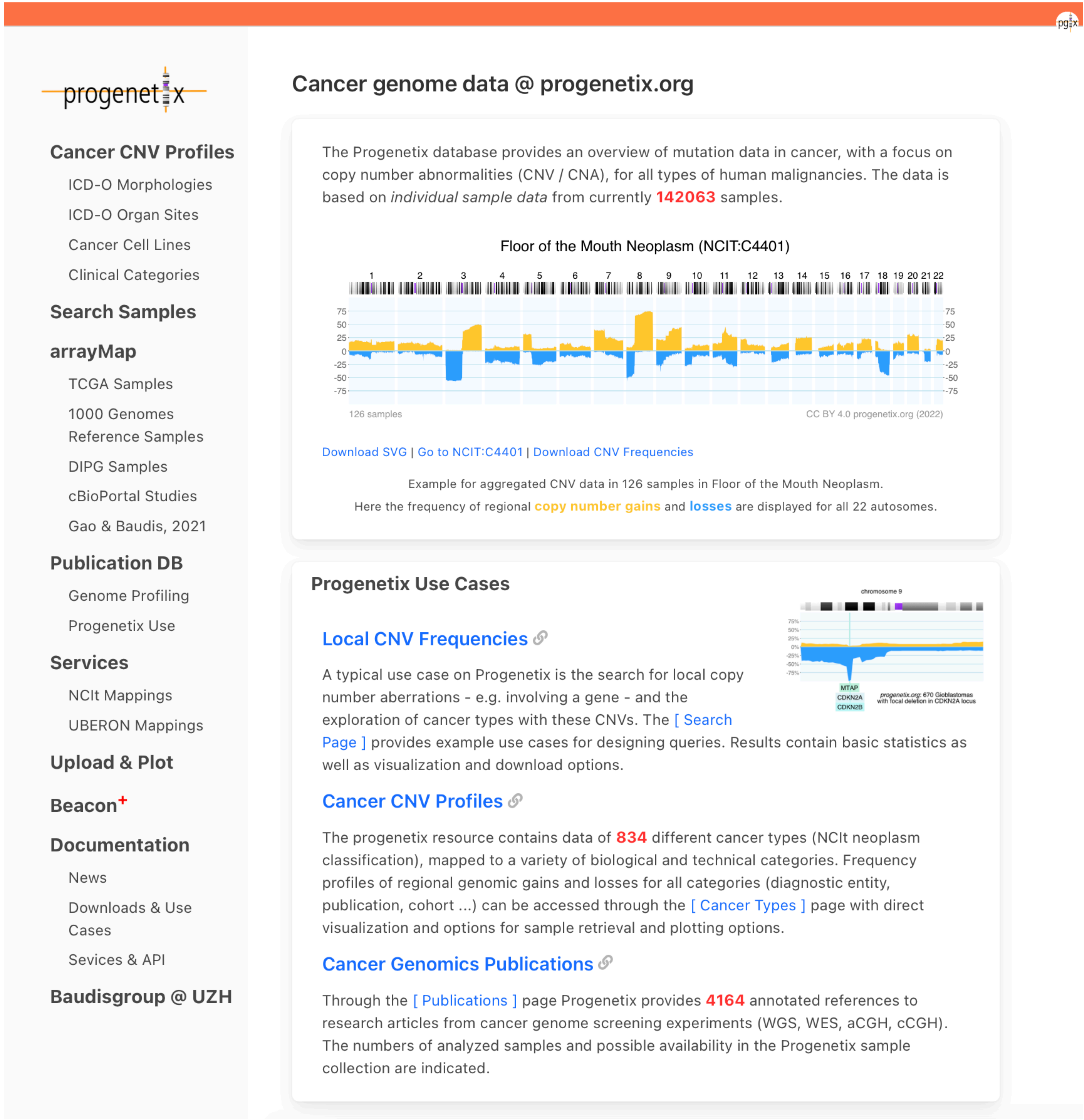
Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **116'000 cancer CNV profiles**
- more than **800 diagnostic types**
- inclusion of reference datasets (e.g. TCGA)
- 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
- recent addition of SNV data for some series

Universität Zürich^{UZH}

progenetix

SIB Swiss Institute of Bioinformatics



Cancer genome data @ progenetix.org

The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* from currently **142063** samples.

Example for aggregated CNV data in 126 samples in Floor of the Mouth Neoplasm.
Here the frequency of regional **copy number gains** and **losses** are displayed for all 22 autosomes.

Progenetix Use Cases

Local CNV Frequencies

A typical use case on Progenetix is the search for local copy number aberrations - e.g. involving a gene - and the exploration of cancer types with these CNVs. The [Search Page] provides example use cases for designing queries. Results contain basic statistics as well as visualization and download options.

Cancer CNV Profiles

The progenetix resource contains data of **834** different cancer types (NCIt neoplasm classification), mapped to a variety of biological and technical categories. Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the [Cancer Types] page with direct visualization and options for sample retrieval and plotting options.

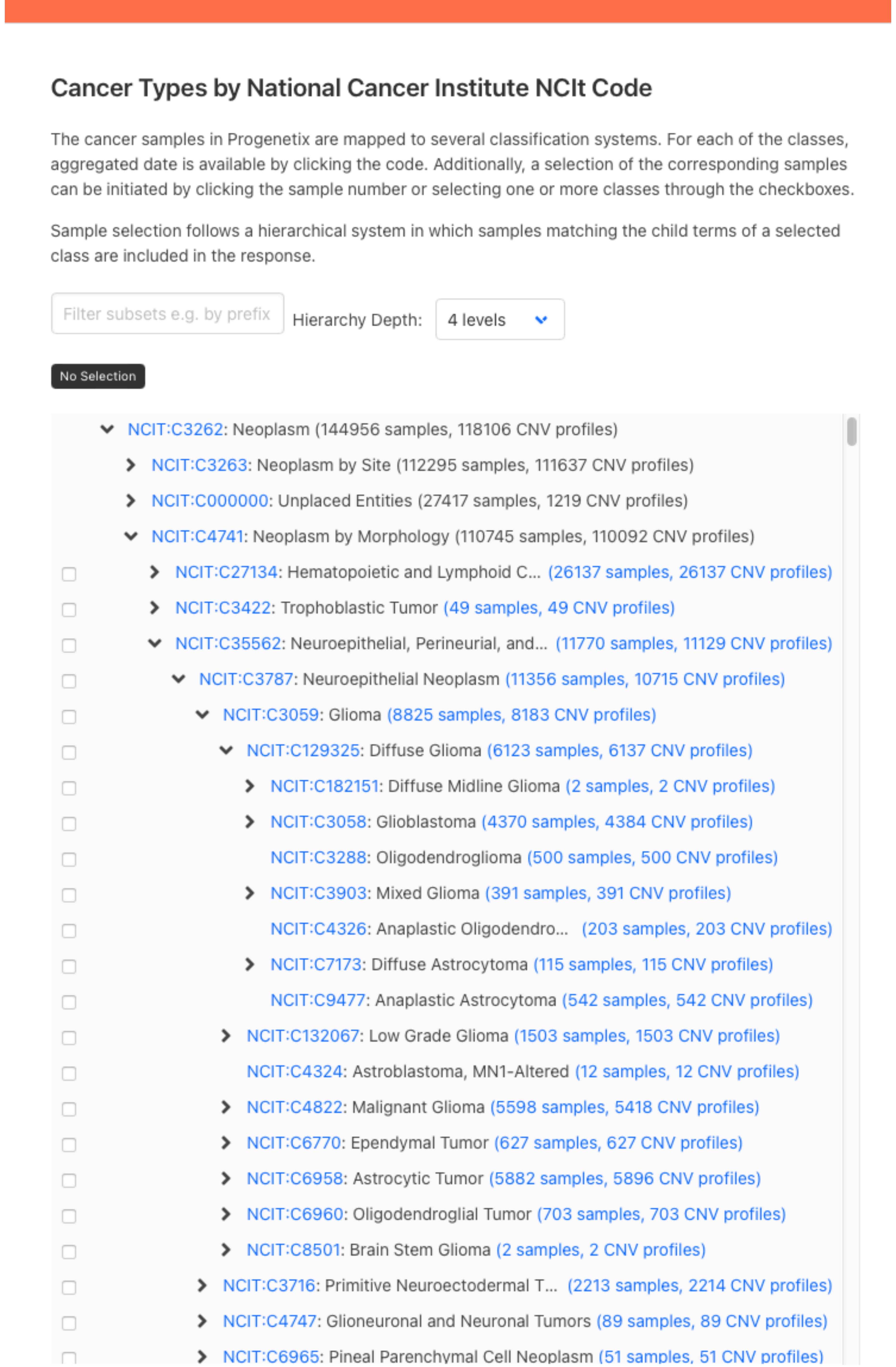
Cancer Genomics Publications

Through the [Publications] page Progenetix provides **4164** annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.

progenetix.org

Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **116'000 cancer CNV profiles**
- more than **800 diagnostic types**
- inclusion of reference datasets (e.g. TCGA)
- 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
- recent addition of SNV data for some series



Cancer Types by National Cancer Institute NCIt Code

The cancer samples in Progenetix are mapped to several classification systems. For each of the classes, aggregated date is available by clicking the code. Additionally, a selection of the corresponding samples can be initiated by clicking the sample number or selecting one or more classes through the checkboxes.

Sample selection follows a hierarchical system in which samples matching the child terms of a selected class are included in the response.

- NCIT:C3262: Neoplasm (144956 samples, 118106 CNV profiles)
  - NCIT:C3263: Neoplasm by Site (112295 samples, 111637 CNV profiles)
  - NCIT:C000000: Unplaced Entities (27417 samples, 1219 CNV profiles)
  - NCIT:C4741: Neoplasm by Morphology (110745 samples, 110092 CNV profiles)
    - ☐ NCIT:C27134: Hematopoietic and Lymphoid C... (26137 samples, 26137 CNV profiles)
    - ☐ NCIT:C3422: Trophoblastic Tumor (49 samples, 49 CNV profiles)
    - ☐ NCIT:C35562: Neuroepithelial, Perineurial, and... (11770 samples, 11129 CNV profiles)
      - ☐ NCIT:C3787: Neuroepithelial Neoplasm (11356 samples, 10715 CNV profiles)
        - ☐ NCIT:C3059: Glioma (8825 samples, 8183 CNV profiles)
          - ☐ NCIT:C129325: Diffuse Glioma (6123 samples, 6137 CNV profiles)
            - ☐ NCIT:C182151: Diffuse Midline Glioma (2 samples, 2 CNV profiles)
            - ☐ NCIT:C3058: Glioblastoma (4370 samples, 4384 CNV profiles)
            - ☐ NCIT:C3288: Oligodendroglioma (500 samples, 500 CNV profiles)
            - ☐ NCIT:C3903: Mixed Glioma (391 samples, 391 CNV profiles)
            - ☐ NCIT:C4326: Anaplastic Oligodendro... (203 samples, 203 CNV profiles)
            - ☐ NCIT:C7173: Diffuse Astrocytoma (115 samples, 115 CNV profiles)
            - ☐ NCIT:C9477: Anaplastic Astrocytoma (542 samples, 542 CNV profiles)
          - ☐ NCIT:C132067: Low Grade Glioma (1503 samples, 1503 CNV profiles)
          - ☐ NCIT:C4324: Astroblastoma, MN1-Altered (12 samples, 12 CNV profiles)
          - ☐ NCIT:C4822: Malignant Glioma (5598 samples, 5418 CNV profiles)
          - ☐ NCIT:C6770: Ependymal Tumor (627 samples, 627 CNV profiles)
          - ☐ NCIT:C6958: Astrocytic Tumor (5882 samples, 5896 CNV profiles)
          - ☐ NCIT:C6960: Oligodendroglial Tumor (703 samples, 703 CNV profiles)
          - ☐ NCIT:C8501: Brain Stem Glioma (2 samples, 2 CNV profiles)
        - ☐ NCIT:C3716: Primitive Neuroectodermal T... (2213 samples, 2214 CNV profiles)
        - ☐ NCIT:C4747: Glioneuronal and Neuronal Tumors (89 samples, 89 CNV profiles)
        - ☐ NCIT:C6965: Pineal Parenchymal Cell Neoplasm (51 samples, 51 CNV profiles)

Universität Zürich UZH

progenetix

SIB Swiss Institute of Bioinformatics

progenetix.org

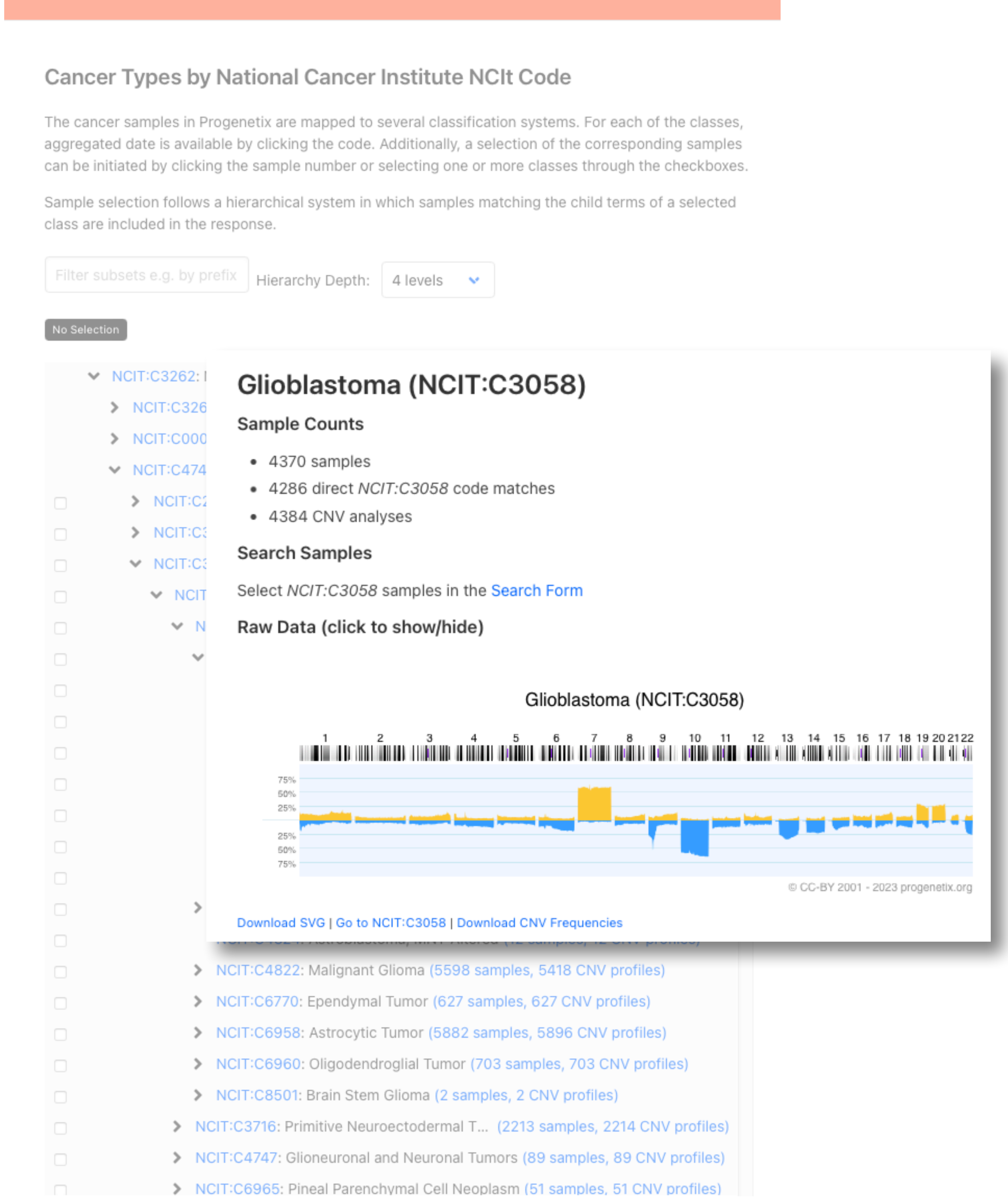
Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **116'000 cancer CNV profiles**
- more than **800 diagnostic types**
- inclusion of reference datasets (e.g. TCGA)
- 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
- recent addition of SNV data for some series

Universität Zürich UZH

progenetix

SIB Swiss Institute of Bioinformatics



Cancer Types by National Cancer Institute NCIt Code

The cancer samples in Progenetix are mapped to several classification systems. For each of the classes, aggregated date is available by clicking the code. Additionally, a selection of the corresponding samples can be initiated by clicking the sample number or selecting one or more classes through the checkboxes.

Sample selection follows a hierarchical system in which samples matching the child terms of a selected class are included in the response.

Filter subsets e.g. by prefix | Hierarchy Depth: 4 levels

No Selection

NCIT:C3262: I
NCIT:C326
NCIT:C000
NCIT:C474
NCIT:C2
NCIT:C3
NCIT:C3
NCIT
N

Glioblastoma (NCIT:C3058)

Sample Counts

- 4370 samples
- 4286 direct *NCIT:C3058* code matches
- 4384 CNV analyses

Search Samples

Select *NCIT:C3058* samples in the Search Form

Raw Data (click to show/hide)

Glioblastoma (NCIT:C3058)

© CC-BY 2001 - 2023 progenetix.org

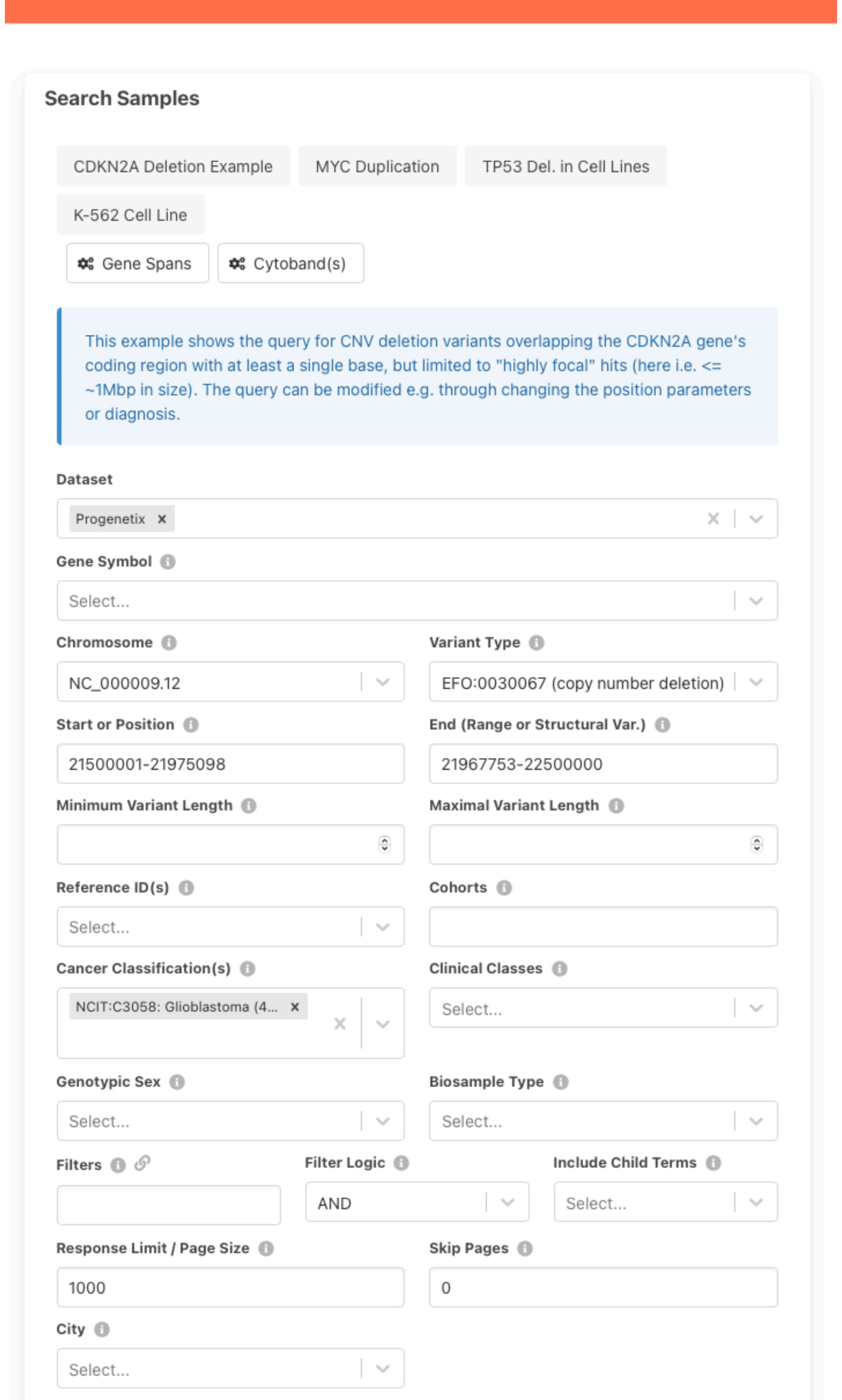
Download SVG | Go to NCIT:C3058 | Download CNV Frequencies

□ NCIT:C4822: Malignant Glioma (5598 samples, 5418 CNV profiles)
□ NCIT:C6770: Ependymal Tumor (627 samples, 627 CNV profiles)
□ NCIT:C6958: Astrocytic Tumor (5882 samples, 5896 CNV profiles)
□ NCIT:C6960: Oligodendroglial Tumor (703 samples, 703 CNV profiles)
□ NCIT:C8501: Brain Stem Glioma (2 samples, 2 CNV profiles)
□ NCIT:C3716: Primitive Neuroectodermal T... (2213 samples, 2214 CNV profiles)
□ NCIT:C4747: Glioneuronal and Neuronal Tumors (89 samples, 89 CNV profiles)
□ NCIT:C6965: Pineal Parenchymal Cell Neoplasm (51 samples, 51 CNV profiles)

progenetix.org

Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **116'000 cancer CNV profiles**
- more than **800 diagnostic types**
- inclusion of reference datasets (e.g. TCGA)
- 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
- recent addition of SNV data for some series



Search Samples

This example shows the query for CNV deletion variants overlapping the CDKN2A gene's coding region with at least a single base, but limited to "highly focal" hits (here i.e. <= ~1Mbp in size). The query can be modified e.g. through changing the position parameters or diagnosis.

Dataset: Progenetix

Gene Symbol: Select...

Chromosome: NC_000009.12

Variant Type: EFO:0030067 (copy number deletion)

Start or Position: 21500001-21975098

End (Range or Structural Var.): 21967753-22500000

Minimum Variant Length:

Maximal Variant Length:

Reference ID(s): Select...

Cohorts:

Cancer Classification(s): NCIT:C3058: Glioblastoma (4...

Clinical Classes: Select...

Genotypic Sex: Select...

Biosample Type: Select...

Filters:

Filter Logic: AND

Include Child Terms: Select...

Response Limit / Page Size: 1000

Skip Pages: 0

City: Select...

Universität Zürich UZH

progenetix

SIB Swiss Institute of Bioinformatics

progenetix.org

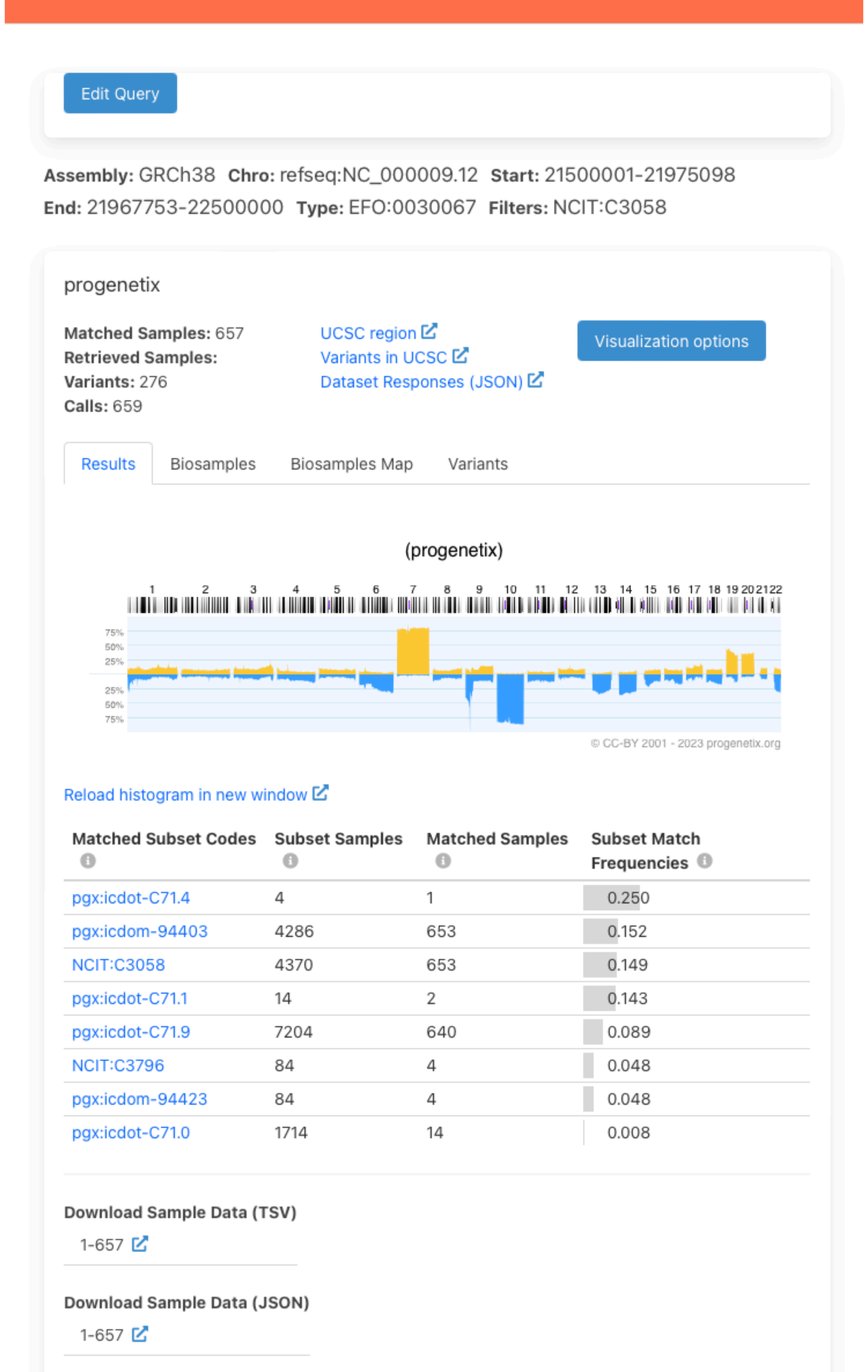
Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **116'000 cancer CNV profiles**
- more than **800 diagnostic types**
- inclusion of reference datasets (e.g. TCGA)
- 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
- recent addition of SNV data for some series

Universität Zürich UZH

progenetix

SIB Swiss Institute of Bioinformatics



Assembly: GRCh38 **Chro:** refseq:NC_000009.12 **Start:** 21500001-21975098 **End:** 21967753-22500000 **Type:** EFO:0030067 **Filters:** NCIT:C3058

progenetix

Matched Samples: 657
Retrieved Samples:
Variants: 276
Calls: 659

UCSC region
Variants in UCSC
Dataset Responses (JSON)

Reload histogram in new window

Matched Subset Codes	Subset Samples	Matched Samples	Subset Match Frequencies
pgx:icdot-C71.4	4	1	0.250
pgx:icdom-94403	4286	653	0.152
NCIT:C3058	4370	653	0.149
pgx:icdot-C71.1	14	2	0.143
pgx:icdot-C71.9	7204	640	0.089
NCIT:C3796	84	4	0.048
pgx:icdom-94423	84	4	0.048
pgx:icdot-C71.0	1714	14	0.008

Download Sample Data (TSV)

1-657

Download Sample Data (JSON)

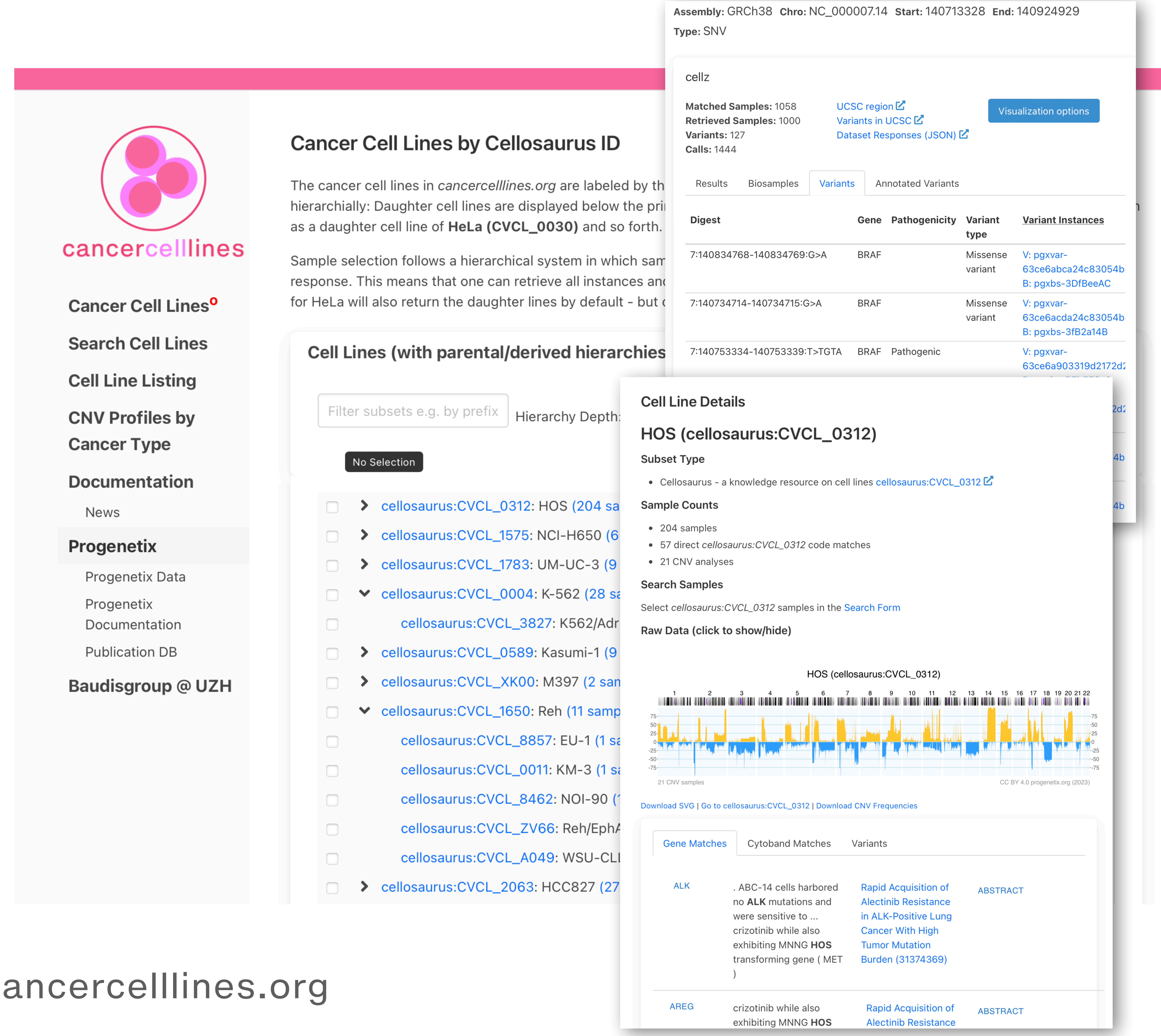
1-657

Cancer Cell Lines

Cancer Genomics Reference Resource

- starting from >5000 cell line CNV profiles
  - 5754 samples | 2163 cell lines
  - 256 different NCIT codes
- genomic mapping of annotated variants and additional data from several resources (ClinVar, CCLE, Cellosaurus...)
  - 16178 cell lines
  - 400 different NCIT codes
- query and data delivery through Beacon v2 API

➡ integration in data federation approaches



Cancer Cell Lines by Cellosaurus ID

The cancer cell lines in *cancercelllines.org* are labeled by th hierarchially: Daughter cell lines are displayed below the pri as a daughter cell line of **HeLa (CVCL_0030)** and so forth.

Sample selection follows a hierarchical system in which sam response. This means that one can retrieve all instances and for HeLa will also return the daughter lines by default - but c

Cell Lines (with parental/derived hierarchies

Filter subsets e.g. by prefix | Hierarchy Depth:

No Selection

- □ cellosaurus:CVCL_0312: HOS (204 sa
- □ cellosaurus:CVCL_1575: NCI-H650 (6
- □ cellosaurus:CVCL_1783: UM-UC-3 (9
- □ cellosaurus:CVCL_0004: K-562 (28 sa
  - □ cellosaurus:CVCL_3827: K562/Adr
- □ cellosaurus:CVCL_0589: Kasumi-1 (9
- □ cellosaurus:CVCL_XK00: M397 (2 sam
- □ cellosaurus:CVCL_1650: Reh (11 samp
  - □ cellosaurus:CVCL_8857: EU-1 (1 sa
  - □ cellosaurus:CVCL_0011: KM-3 (1 sa
  - □ cellosaurus:CVCL_8462: NOI-90 (1
  - □ cellosaurus:CVCL_ZV66: Reh/EphA
  - □ cellosaurus:CVCL_A049: WSU-CLL
- □ cellosaurus:CVCL_2063: HCC827 (27

Assembly: GRCh38 **Chro:** NC_000007.14 **Start:** 140713328 **End:** 140924929

Type: SNV

cellz

Matched Samples: 1058
Retrieved Samples: 1000
Variants: 127
Calls: 1444

UCSC region
Variants in UCSC
Dataset Responses (JSON)

Digest	Gene	Pathogenicity	Variant type	Variant Instances
7:140834768-140834769:G>A	BRAF		Missense variant	V: pgxvar-63ce6abca24c83054b B: pgxbs-3DfBeeAC
7:140734714-140734715:G>A	BRAF		Missense variant	V: pgxvar-63ce6acda24c83054b B: pgxbs-3fB2a14B
7:140753334-140753339:T>TGTA	BRAF	Pathogenic		V: pgxvar-63ce6a903319d2172d2

Cell Line Details

HOS (cellosaurus:CVCL_0312)

Subset Type

- Cellosaurus - a knowledge resource on cell lines cellosaurus:CVCL_0312

Sample Counts

- 204 samples
- 57 direct *cellosaurus:CVCL_0312* code matches
- 21 CNV analyses

Search Samples

Select *cellosaurus:CVCL_0312* samples in the Search Form

Raw Data (click to show/hide)

HOS (cellosaurus:CVCL_0312)

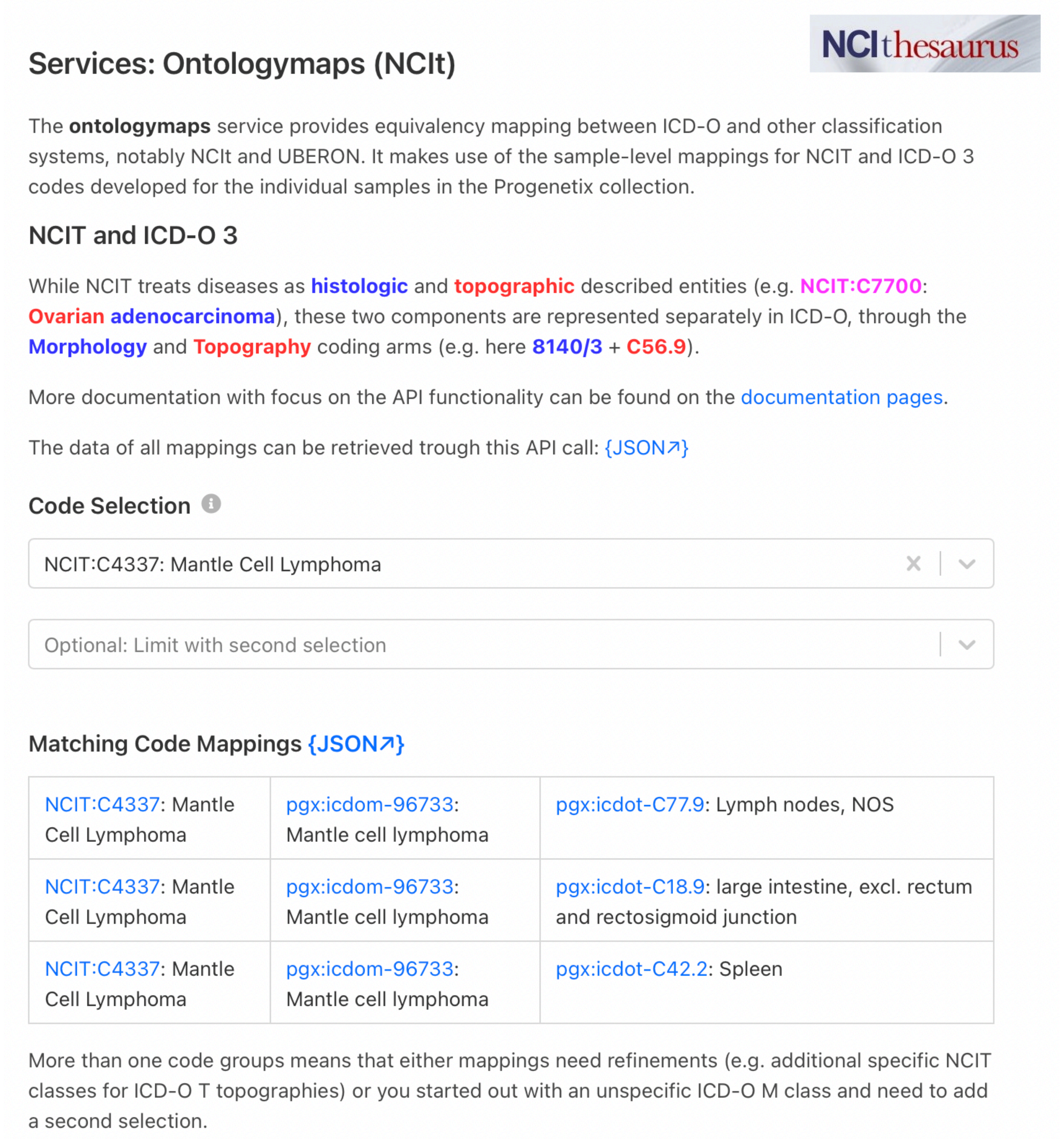
Download SVG | Go to cellosaurus:CVCL_0312 | Download CNV Frequencies

ALK	. ABC-14 cells harbored no **ALK** mutations and were sensitive to ... crizotinib while also exhibiting MNNG **HOS** transforming gene (MET)	Rapid Acquisition of Alectinib Resistance in ALK-Positive Lung Cancer With High Tumor Mutation Burden (31374369)	ABSTRACT
AREG	crizotinib while also exhibiting MNNG **HOS**	Rapid Acquisition of Alectinib Resistance	ABSTRACT

cancercelllines.org

Lead: Rahel Paloots

Ontologies and Classifications



Services: Ontologymaps (NCIt)

NCIthesaurus

The **ontologymaps** service provides equivalency mapping between ICD-O and other classification systems, notably NCIt and UBERON. It makes use of the sample-level mappings for NCIT and ICD-O 3 codes developed for the individual samples in the Progenetix collection.

NCIT and ICD-O 3

While NCIT treats diseases as **histologic** and **topographic** described entities (e.g. **NCIT:C7700**: **Ovarian adenocarcinoma**), these two components are represented separately in ICD-O, through the **Morphology** and **Topography** coding arms (e.g. here **8140/3** + **C56.9**).

More documentation with focus on the API functionality can be found on the documentation pages.

The data of all mappings can be retrieved trough this API call: {JSON↗}

Code Selection

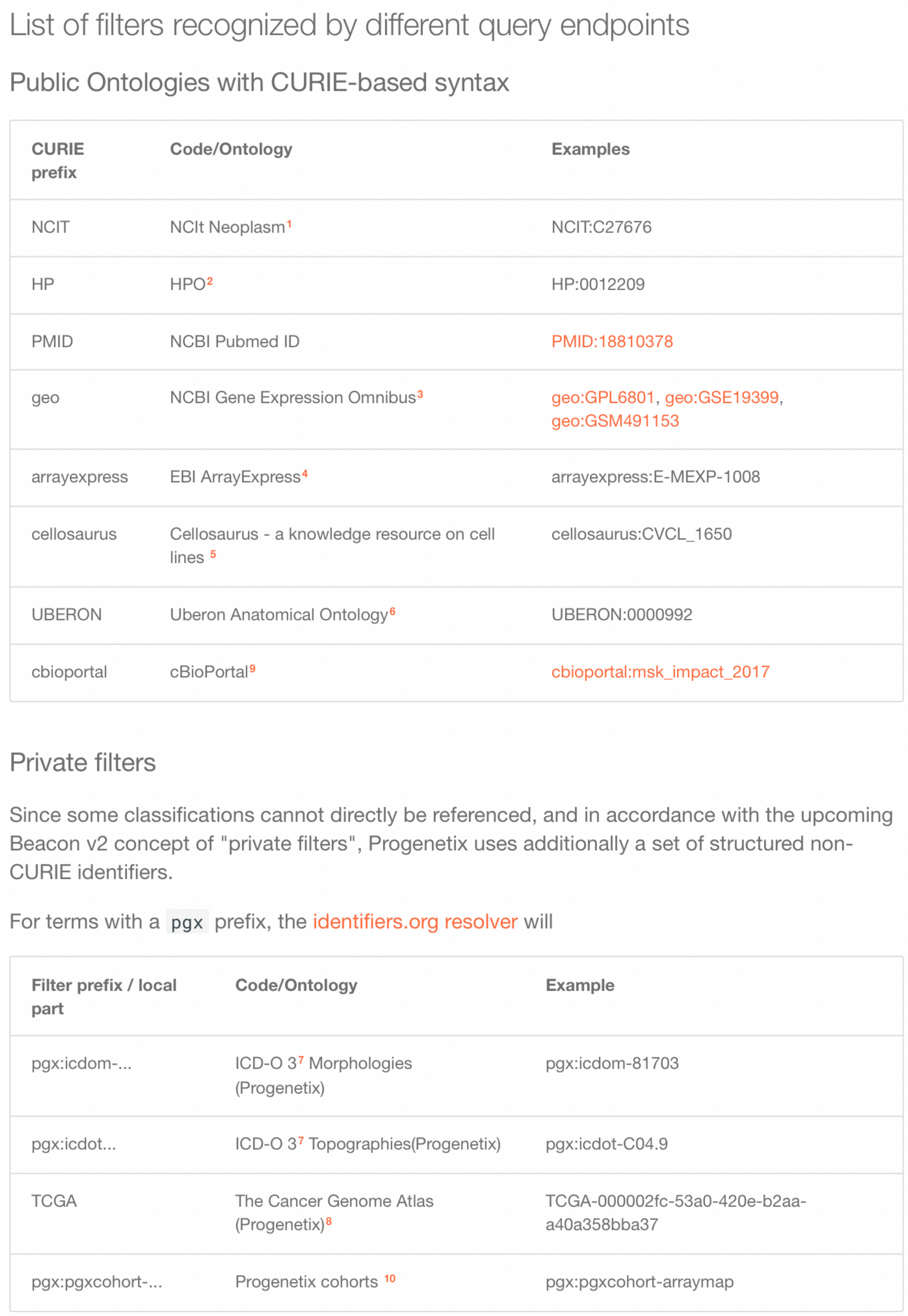
NCIT:C4337: Mantle Cell Lymphoma

Optional: Limit with second selection

Matching Code Mappings {JSON↗}

NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C77.9: Lymph nodes, NOS
NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C18.9: large intestine, excl. rectum and rectosigmoid junction
NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C42.2: Spleen

More than one code groups means that either mappings need refinements (e.g. additional specific NCIT classes for ICD-O T topographies) or you started out with an unspecific ICD-O M class and need to add a second selection.



List of filters recognized by different query endpoints

Public Ontologies with CURIE-based syntax

CURIE prefix	Code/Ontology	Examples
NCIT	NCIt Neoplasm[1]	NCIT:C27676
HP	HPO[2]	HP:0012209
PMID	NCBI Pubmed ID	PMID:18810378
geo	NCBI Gene Expression Omnibus[3]	geo:GPL6801, geo:GSE19399, geo:GSM491153
arrayexpress	EBI ArrayExpress[4]	arrayexpress:E-MEXP-1008
cellosaurus	Cellosaurus - a knowledge resource on cell lines [5]	cellosaurus:CVCL_1650
UBERON	Uberon Anatomical Ontology[6]	UBERON:0000992
cbioportal	cBioPortal[9]	cbioportal:msk_impact_2017

Private filters

Since some classifications cannot directly be referenced, and in accordance with the upcoming Beacon v2 concept of "private filters", Progenetix uses additionally a set of structured non-CURIE identifiers.

For terms with a `pgx` prefix, the identifiers.org resolver will

Filter prefix / local part	Code/Ontology	Example
pgx:icdom-...	ICD-O 3[7] Morphologies (Progenetix)	pgx:icdom-81703
pgx:icdot...	ICD-O 3[7] Topographies(Progenetix)	pgx:icdot-C04.9
TCGA	The Cancer Genome Atlas (Progenetix)[8]	TCGA-000002fc-53a0-420e-b2aa-a40a358bba37
pgx:pgxcohort-...	Progenetix cohorts [10]	pgx:pgxcohort-arraymap

In Progenetix all cancer diagnoses are coded to both NCIt neoplasm codes and ICD-O 3 Morphology + Topography combinations. The matched mappings are provided as lookup-service since neither an official ICD-O ontology nor such a "disease defined by ICD-O M+T" concept is codified anywhere.

progenetix.org

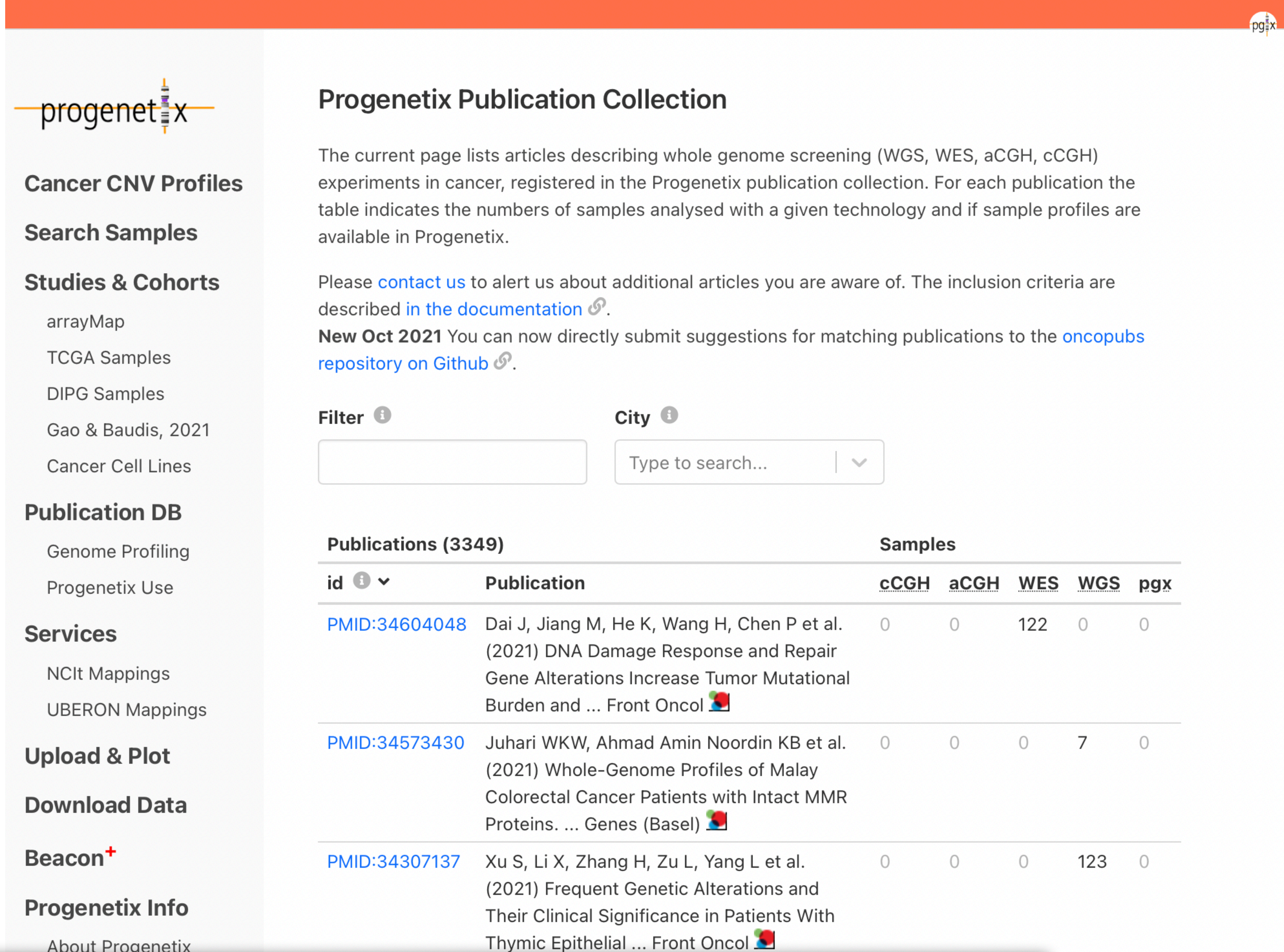
Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **116'000 cancer CNV profiles**
- more than **800 diagnostic types**
- inclusion of reference datasets (e.g. TCGA)
- 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
- recent addition of SNV data for some series

Universität Zürich^UZH

progenetix

SIB Swiss Institute of Bioinformatics



Progenetix Publication Collection

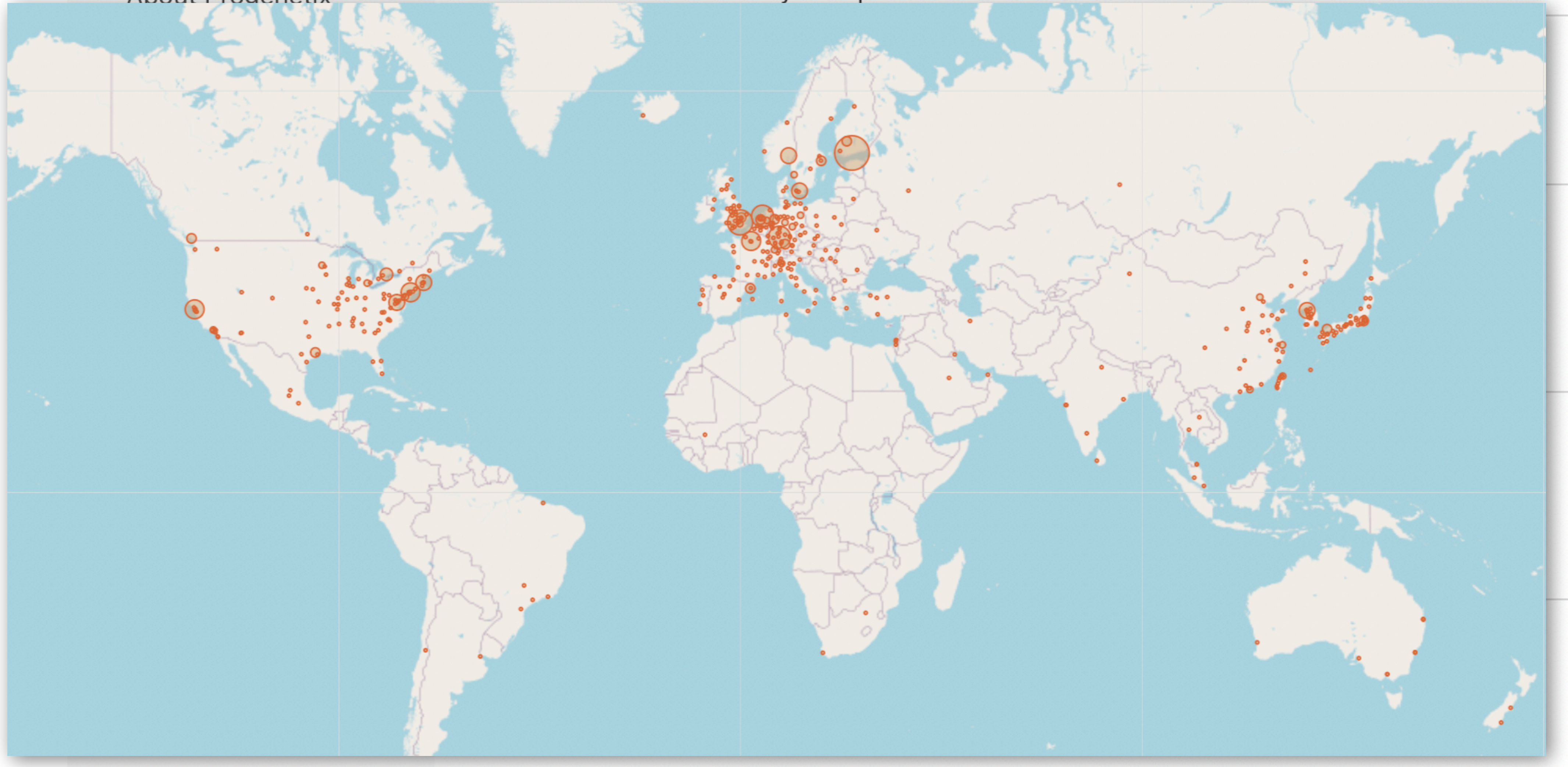
The current page lists articles describing whole genome screening (WGS, WES, aCGH, cCGH) experiments in cancer, registered in the Progenetix publication collection. For each publication the table indicates the numbers of samples analysed with a given technology and if sample profiles are available in Progenetix.

Please contact us to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation.
New Oct 2021 You can now directly submit suggestions for matching publications to the oncopubs repository on Github.

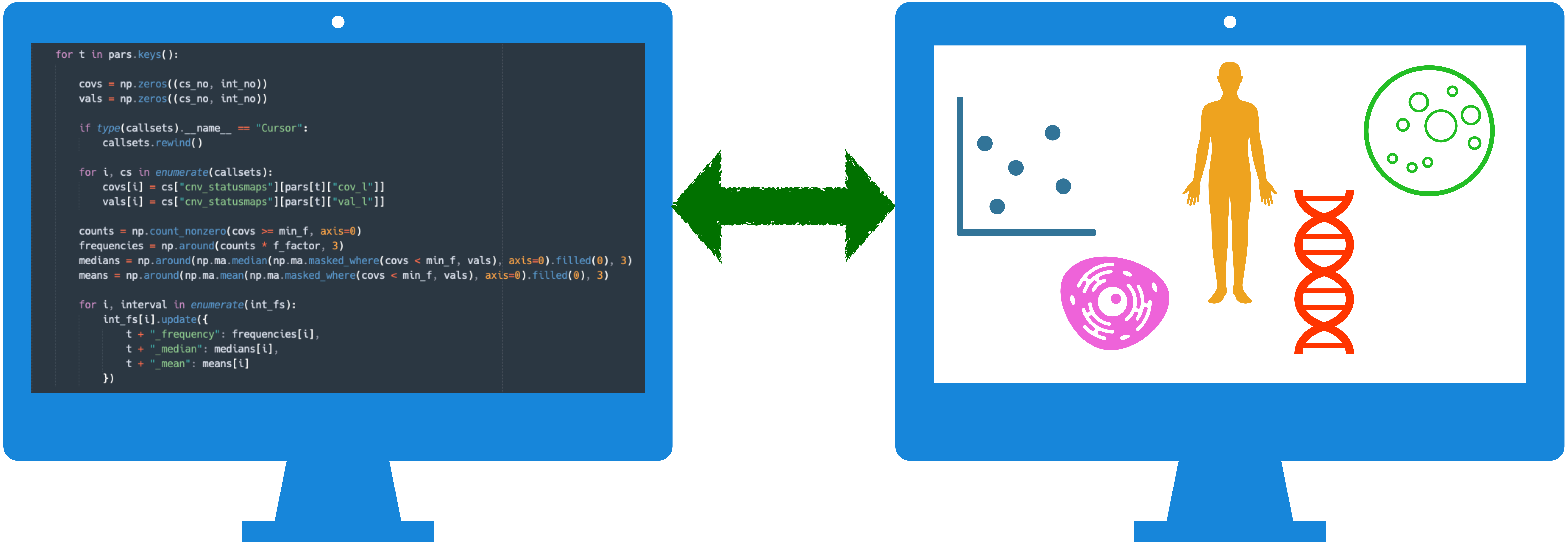
Filter ⓘ

City ⓘ Type to search...

Publications (3349)		**Samples**				
id	**Publication**	**cCGH**	**aCGH**	**WES**	**WGS**	**pgx**
PMID:34604048	Dai J, Jiang M, He K, Wang H, Chen P et al. (2021) DNA Damage Response and Repair Gene Alterations Increase Tumor Mutational Burden and ... Front Oncol	0	0	122	0	0
PMID:34573430	Juhari WKW, Ahmad Amin Noordin KB et al. (2021) Whole-Genome Profiles of Malay Colorectal Cancer Patients with Intact MMR Proteins. ... Genes (Basel)	0	0	0	7	0
PMID:34307137	Xu S, Li X, Zhang H, Zu L, Yang L et al. (2021) Frequent Genetic Alterations and Their Clinical Significance in Patients With Thymic Epithelial ... Front Oncol	0	0	0	123	0
					0	0
					0	0
					135	0
					0	0



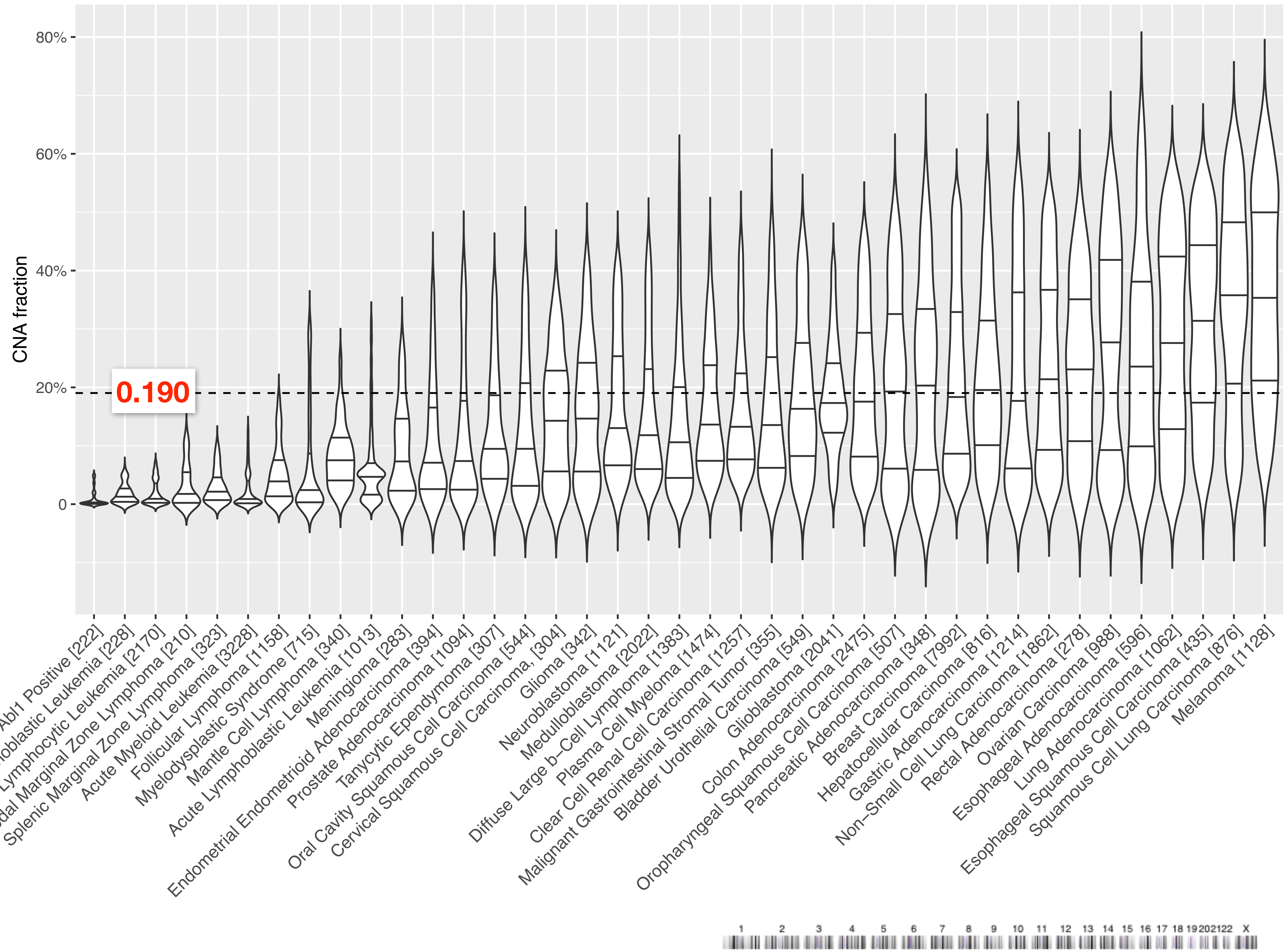
{BioInformaticsScience}



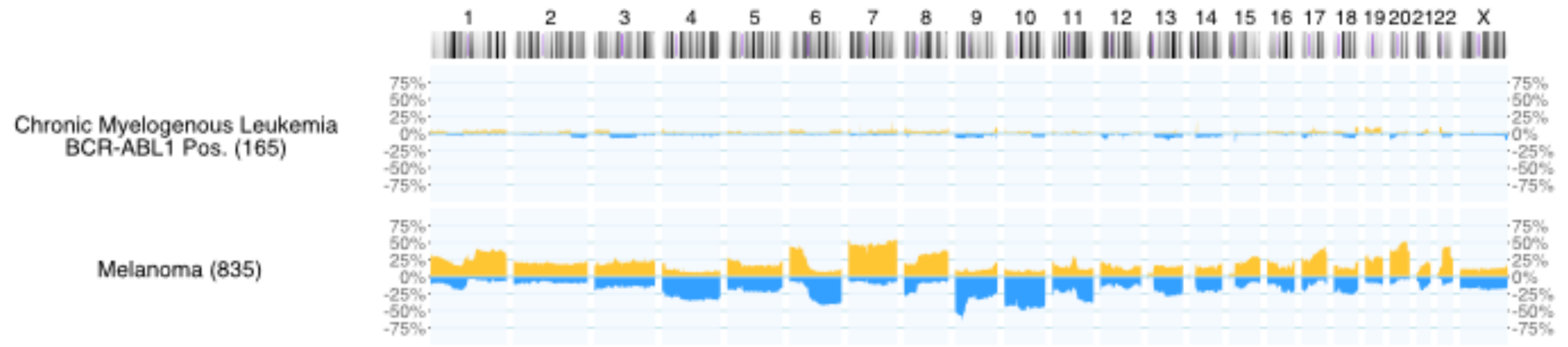
```
for t in pars.keys():

    covs = np.zeros((cs_no, int_no))
    vals = np.zeros((cs_no, int_no))

    if type(callsets).__name__ == "Cursor":
        callsets.rewind()

    for i, cs in enumerate(callsets):
        covs[i] = cs["cnv_statusmaps"][pars[t]["cov_l"]]
        vals[i] = cs["cnv_statusmaps"][pars[t]["val_l"]]

    counts = np.count_nonzero(covs >= min_f, axis=0)
    frequencies = np.around(counts * f_factor, 3)
    medians = np.around(np.ma.median(np.ma.masked_where(covs < min_f, vals), axis=0).filled(0), 3)
    means = np.around(np.ma.mean(np.ma.masked_where(covs < min_f, vals), axis=0).filled(0), 3)

    for i, interval in enumerate(int_fs):
        int_fs[i].update({
            t + "_frequency": frequencies[i],
            t + "_median": medians[i],
            t + "_mean": means[i]
        })
```

Genome CNV coverage in Cancer Classes

- 43654 out of 93640 CNV profiles; filtered for entities w/ >200 samples (removed some entities w/ high CNV rate, e.g. sarcoma subtypes)
- Single-sample CNV profiles were assessed for the fraction of the genome showing CNVs (relative gains, losses)
- range of medians 0.001 (CML) - 0.358 (malignant melanomas)



Lowest / Highest CNV fractions =>



Population stratification in cancer samples based on SNP array data

- Despite extensive somatic mutations of cancer profiling data, consistency between germline and cancer samples reached 97% and 92% for 5 and 26 populations
- Comparison of our benchmarked results with self-reported meta-data estimated a matching rate between 88 % to 92%.
- Ethnicity labels indicated in meta-data are vague compared to the standardized output from our tool

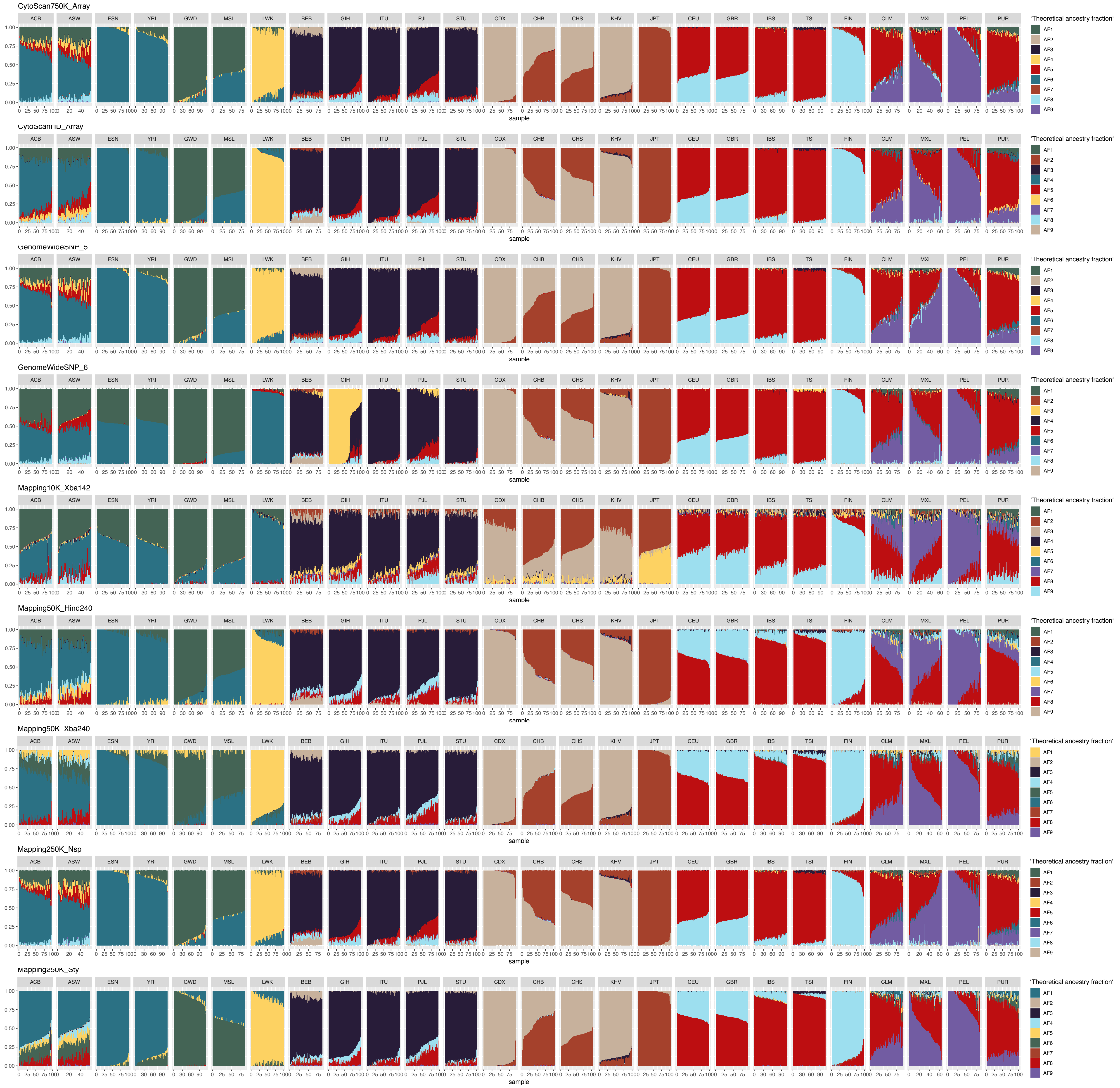


Figure S1 The fraction or contribution of theoretical ancestors (k=9) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are placed neighboring to each other: AFR (1-7), SAS (8-12), EAS (13-17), EUR (18-22), AMR (23-26).

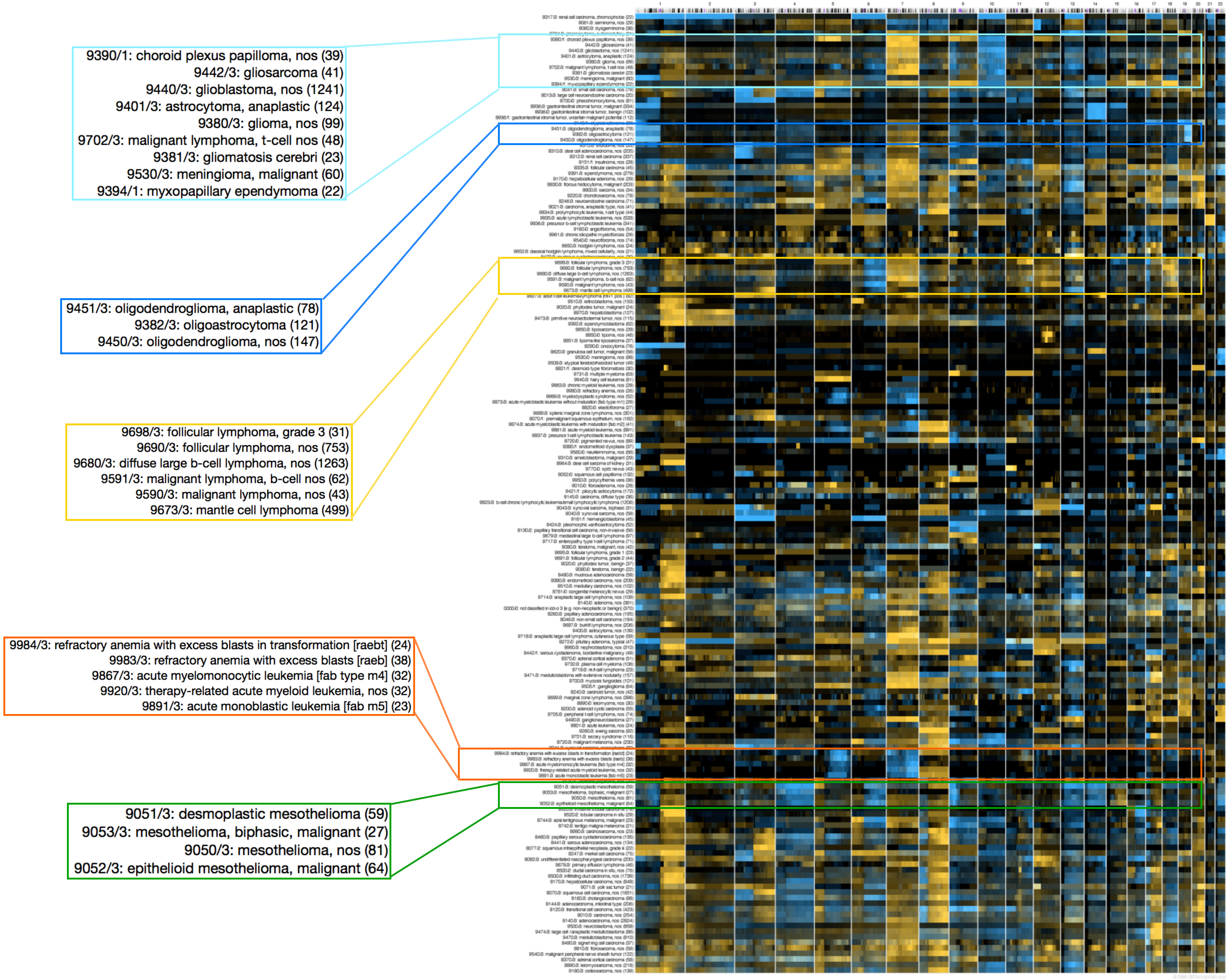
arrayMap

Lead: Qingyao Huang

Somatic Mutations In Cancer: Patterns

Making the case for genomic classifications

Some related cancer entities show similar copy number profiles



9390/1: choroid plexus papilloma, nos (39)
9442/3: gliosarcoma (41)
9440/3: glioblastoma, nos (1241)
9401/3: astrocytoma, anaplastic (124)
9380/3: glioma, nos (99)
9702/3: malignant lymphoma, t-cell nos (48)
9381/3: gliomatosis cerebri (23)
9530/3: meningioma, malignant (60)
9394/1: myxopapillary ependymoma (22)

9451/3: oligodendroglioma, anaplastic (78)
9382/3: oligoastrocytoma (121)
9450/3: oligodendroglioma, nos (147)

9698/3: follicular lymphoma, grade 3 (31)
9690/3: follicular lymphoma, nos (753)
9680/3: diffuse large b-cell lymphoma, nos (1263)
9591/3: malignant lymphoma, b-cell nos (62)
9590/3: malignant lymphoma, nos (43)
9673/3: mantle cell lymphoma (499)

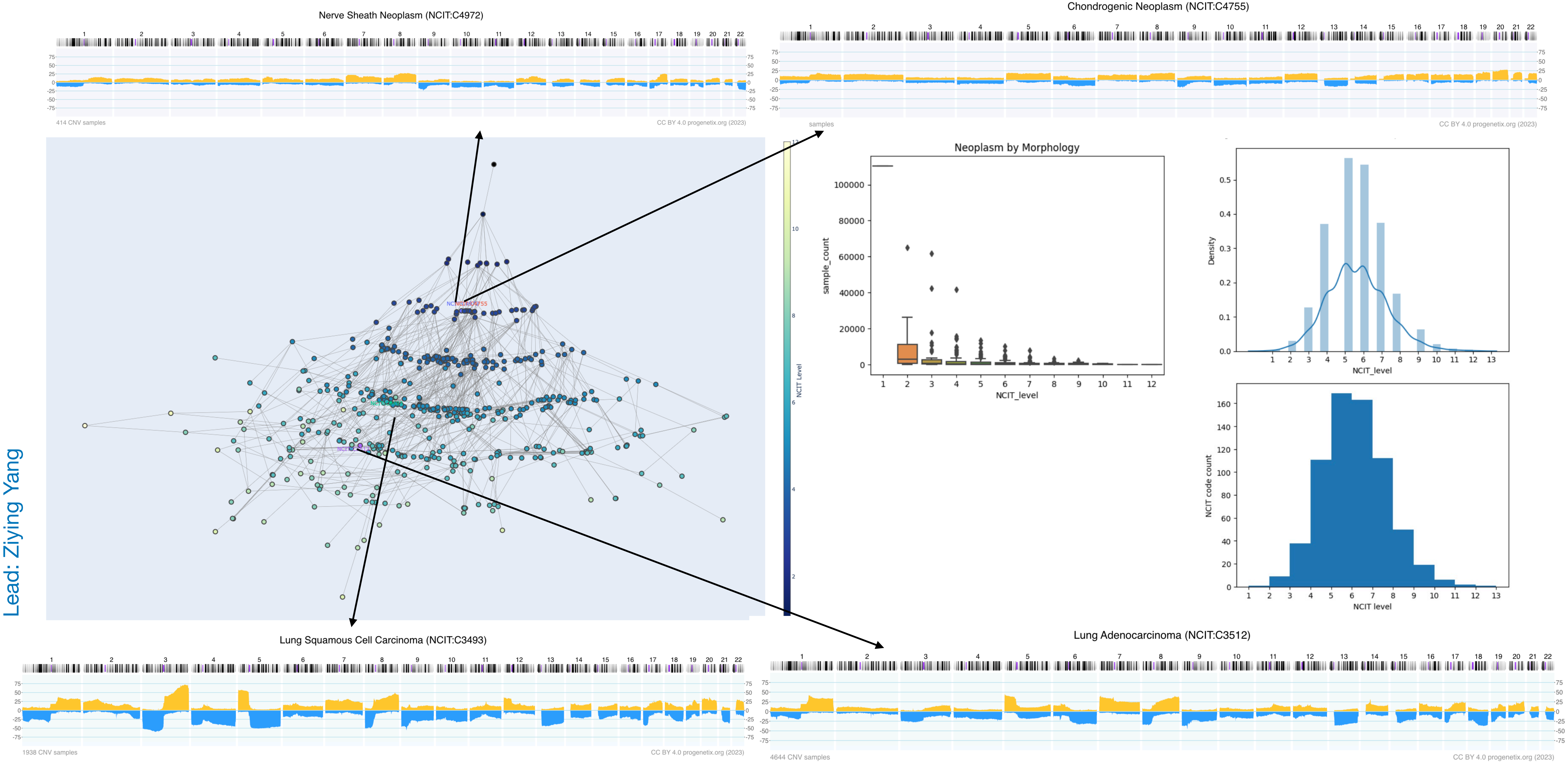
9984/3: refractory anemia with excess blasts in transformation [raebt] (24)
9983/3: refractory anemia with excess blasts [raeb] (38)
9867/3: acute myelomonocytic leukemia [fab type m4] (32)
9920/3: therapy-related acute myeloid leukemia, nos (32)
9891/3: acute monoblastic leukemia [fab m5] (23)

9051/3: desmoplastic mesothelioma (59)
9053/3: mesothelioma, biphasic, malignant (27)
9050/3: mesothelioma, nos (81)
9052/3: epithelioid mesothelioma, malignant (64)

arrayMap

CNV profiles heterogeneity vs cancer classification

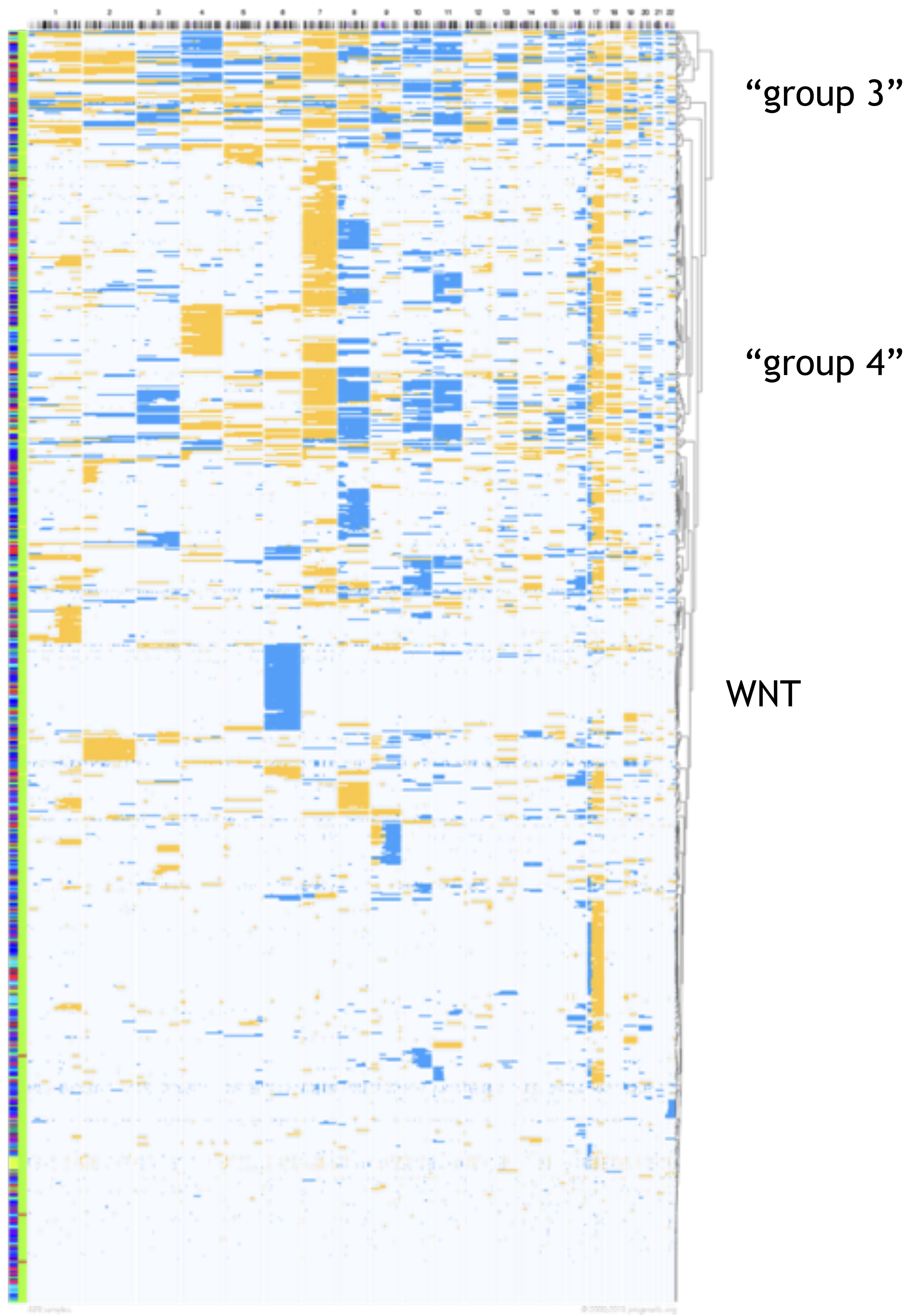
Correspondance of genomic profiles to NCIT cancer hierarchy



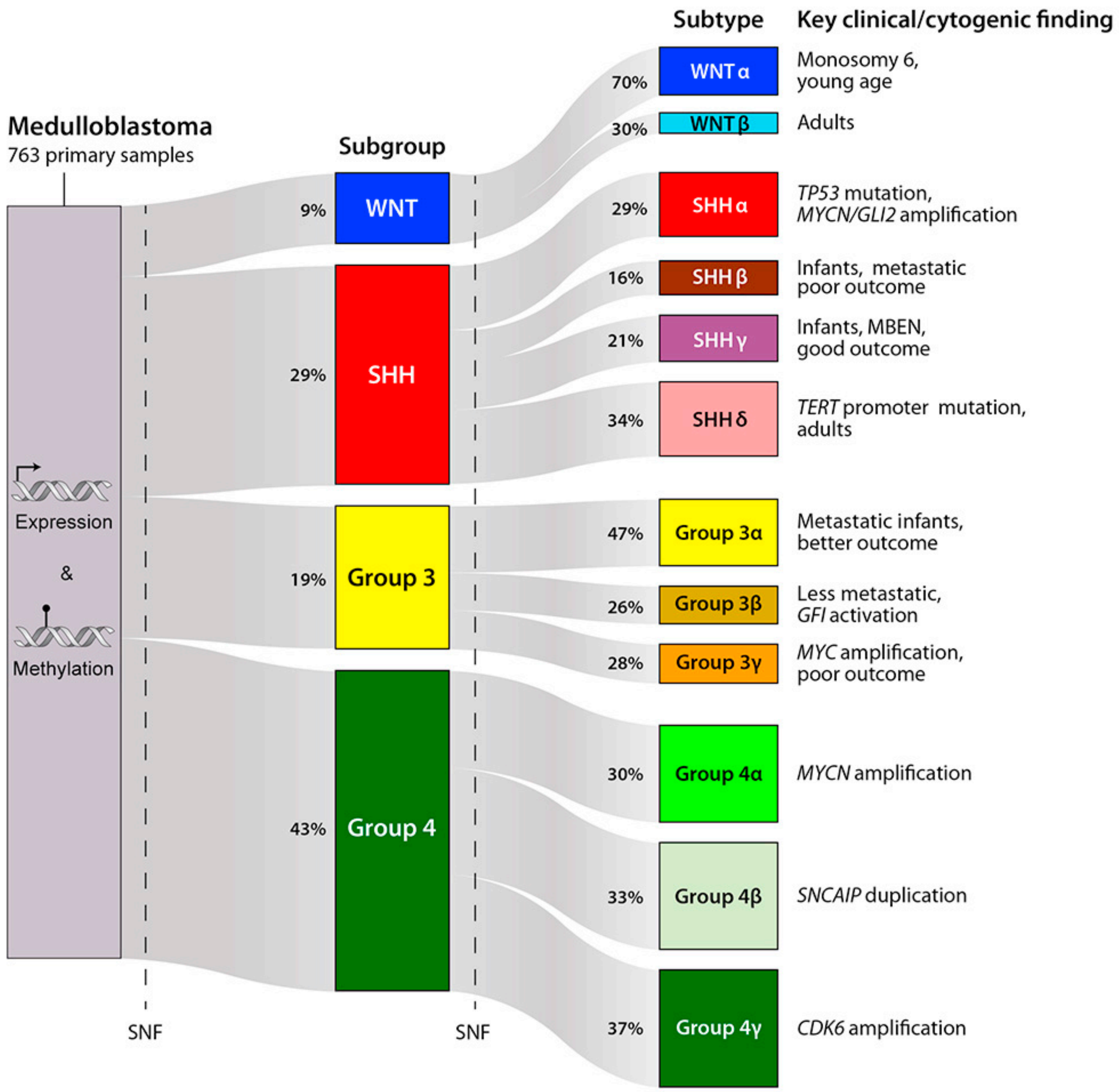
Lead: Ziying Yang

CNA & Cancer heterogeneity

Cancer type definitions can be improved by the addition of molecular parameters as subtype markers or even complete re-evaluation of entity definitions from molecular subtypes with distinct functional mechanisms and clinical trajectories.



Copy number profiles from 889 primary medulloblastomas

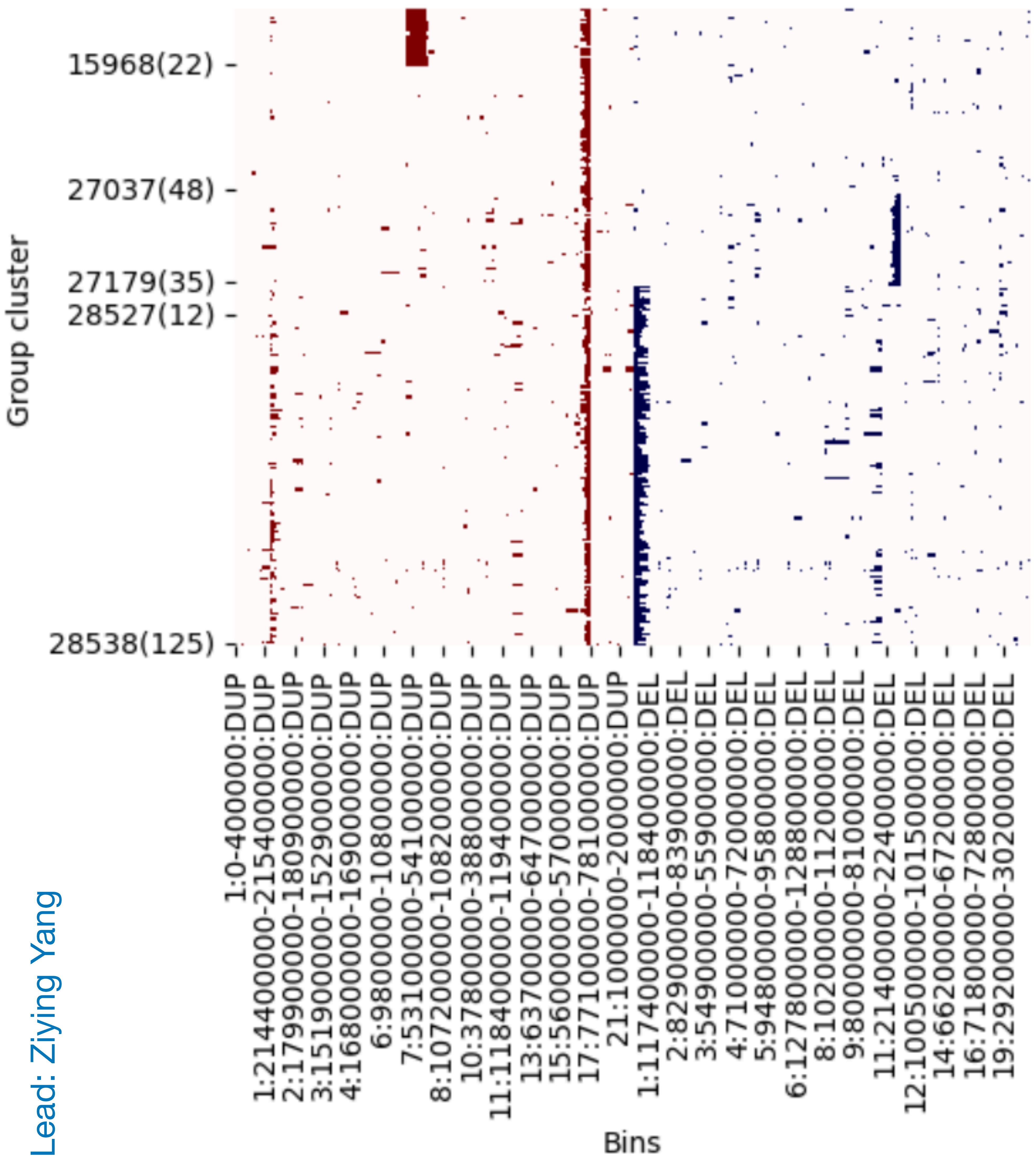


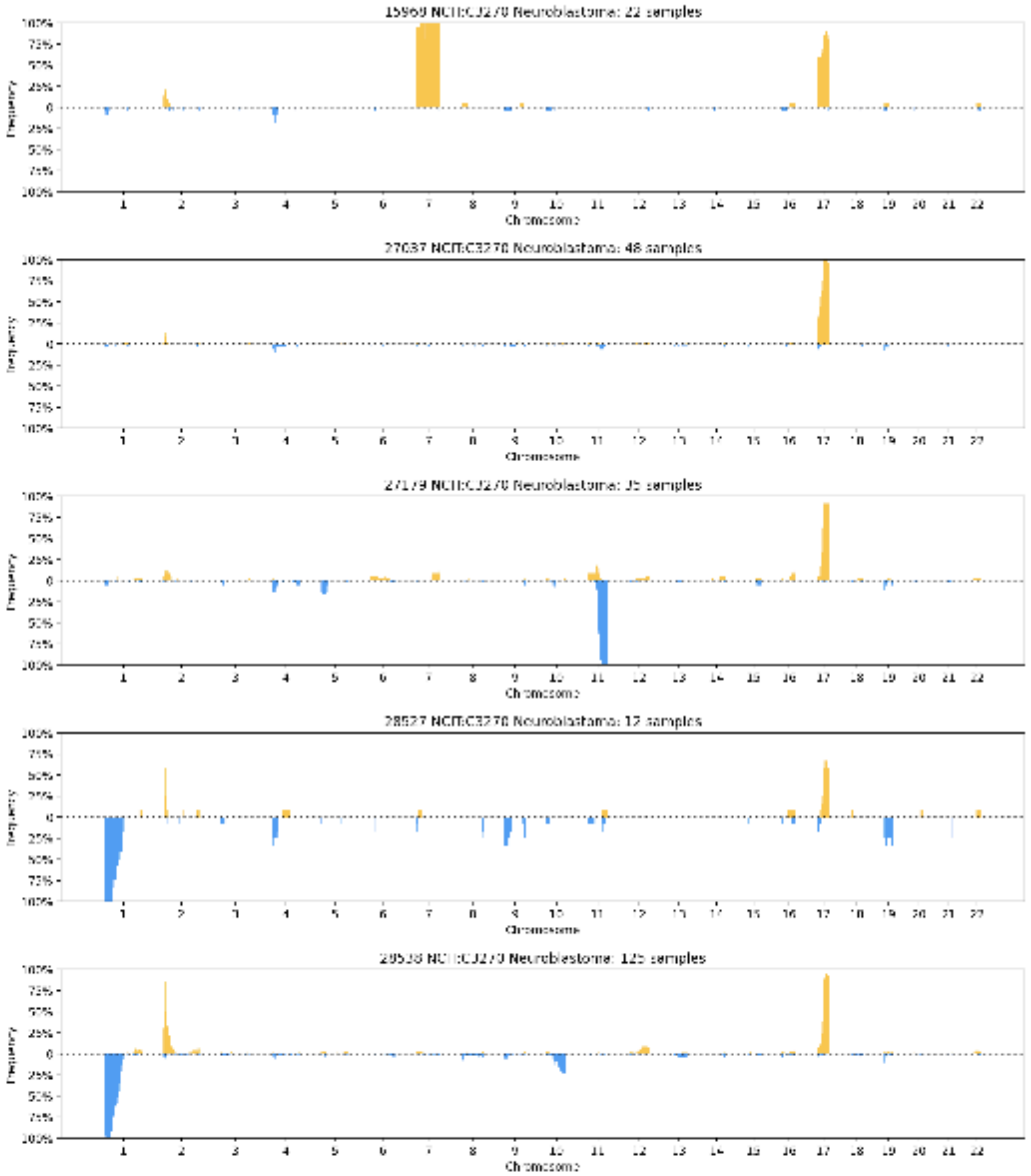
Intertumoral heterogeneity with medulloblastoma subgroups.
Cavalli, Florence MG, et al. "Intertumoral heterogeneity within medulloblastoma subgroups." *Cancer Cell* 31.6 (2017): 737-754.

Lead: Ziying Yang

Results

Entity CNV heterogeneity: Neuroblastoma



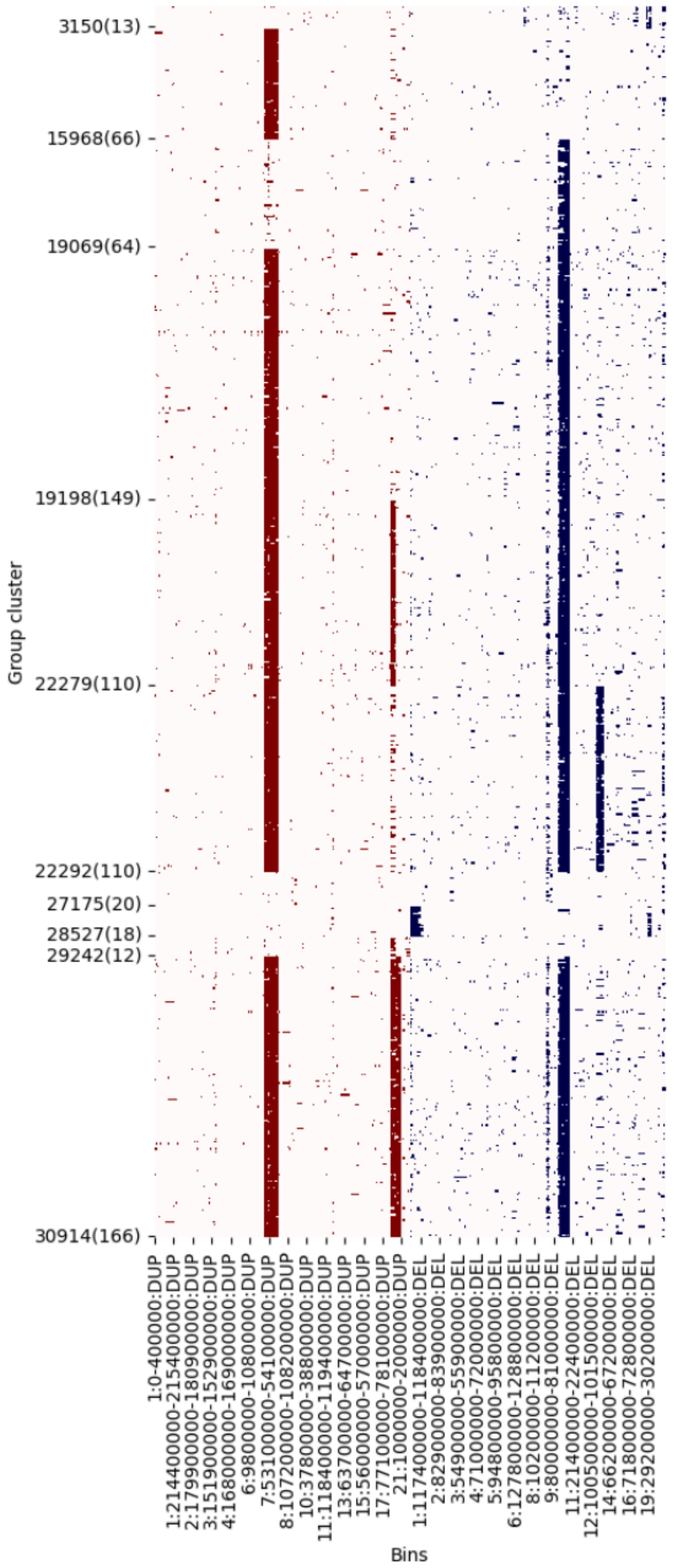


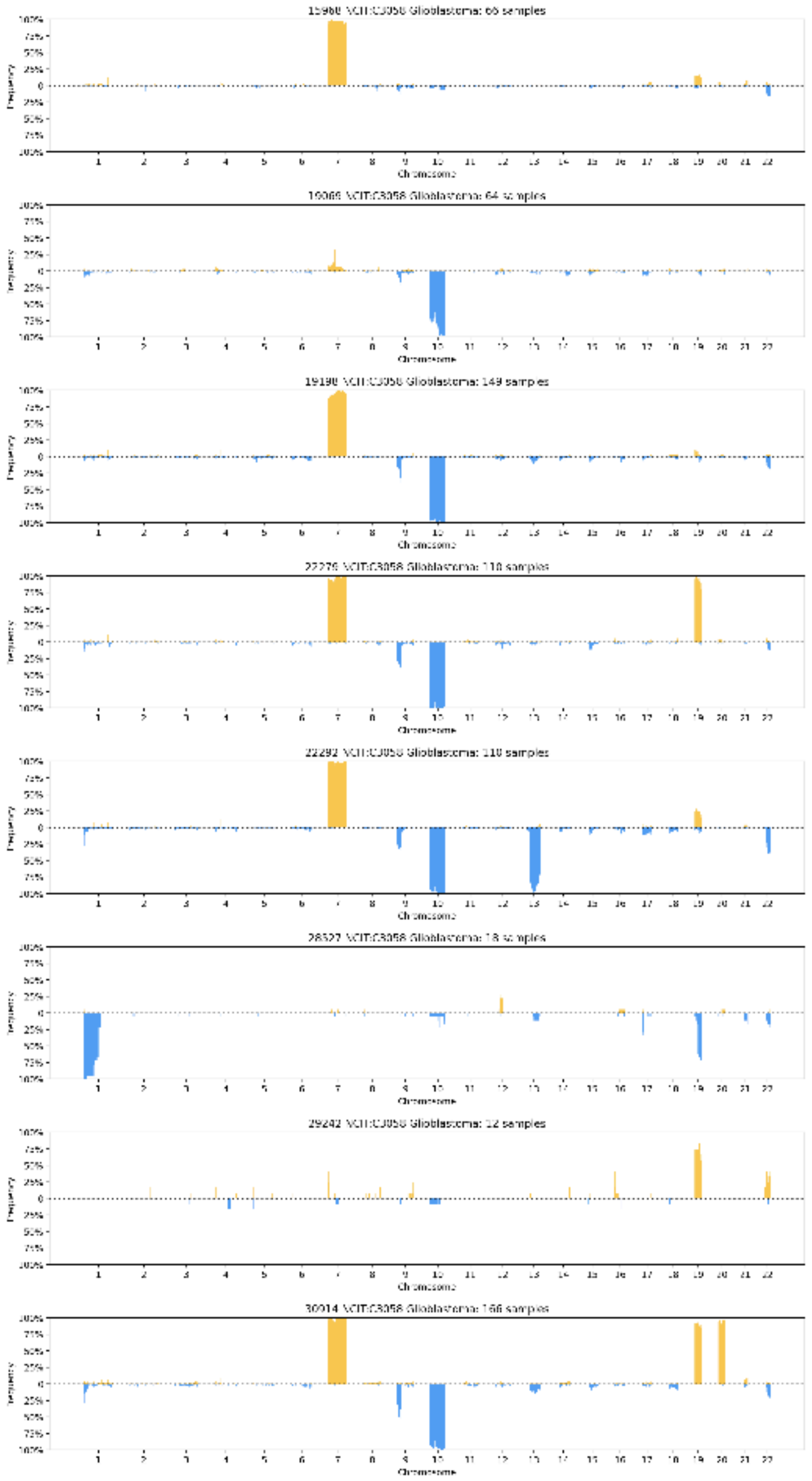
group cluster	CNV features
15968	Dup 7
27037	Dup 17q
27197	Del 11q, Dup 17q
28527	Del 1p
28538	Del 1p, Dup 17q

Lead: Ziying Yang

Results

Entity CNV heterogeneity: Glioblastoma



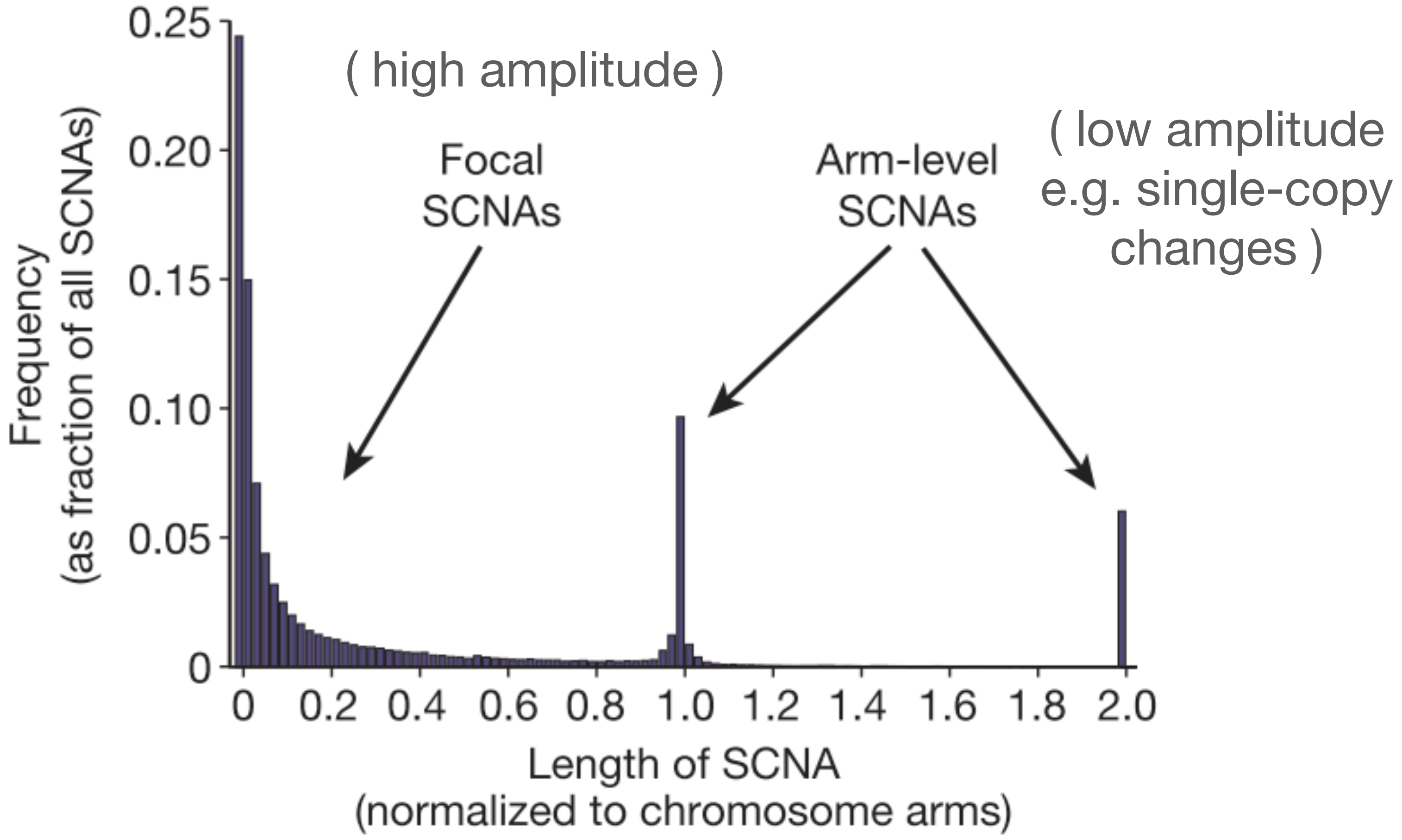


group cluster	CNV features
15968	Dup 7
19069	Del 10
19198	Dup 7, Del 10
22279	Dup 7, Del 10, Dup 19
22292	Dup 7, Del 10, Del 13
28527	Del 1p, Del 19q
29242	Dup 19
30914	Dup 7, Del 10, Dup 19, Dup 20

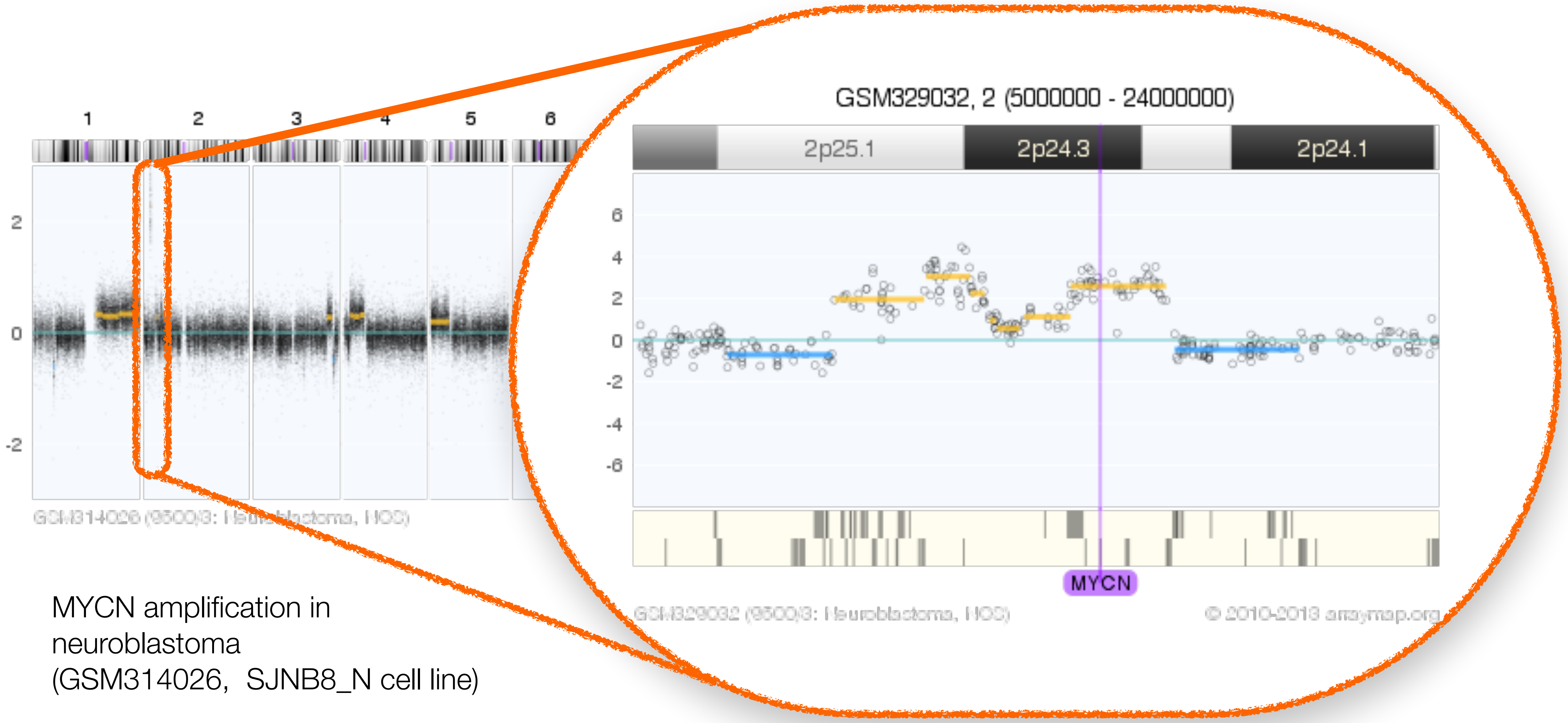
Lead: Ziying Yang

CNV Categorization

different levels of CNV



Rameen et al 2010 Nature



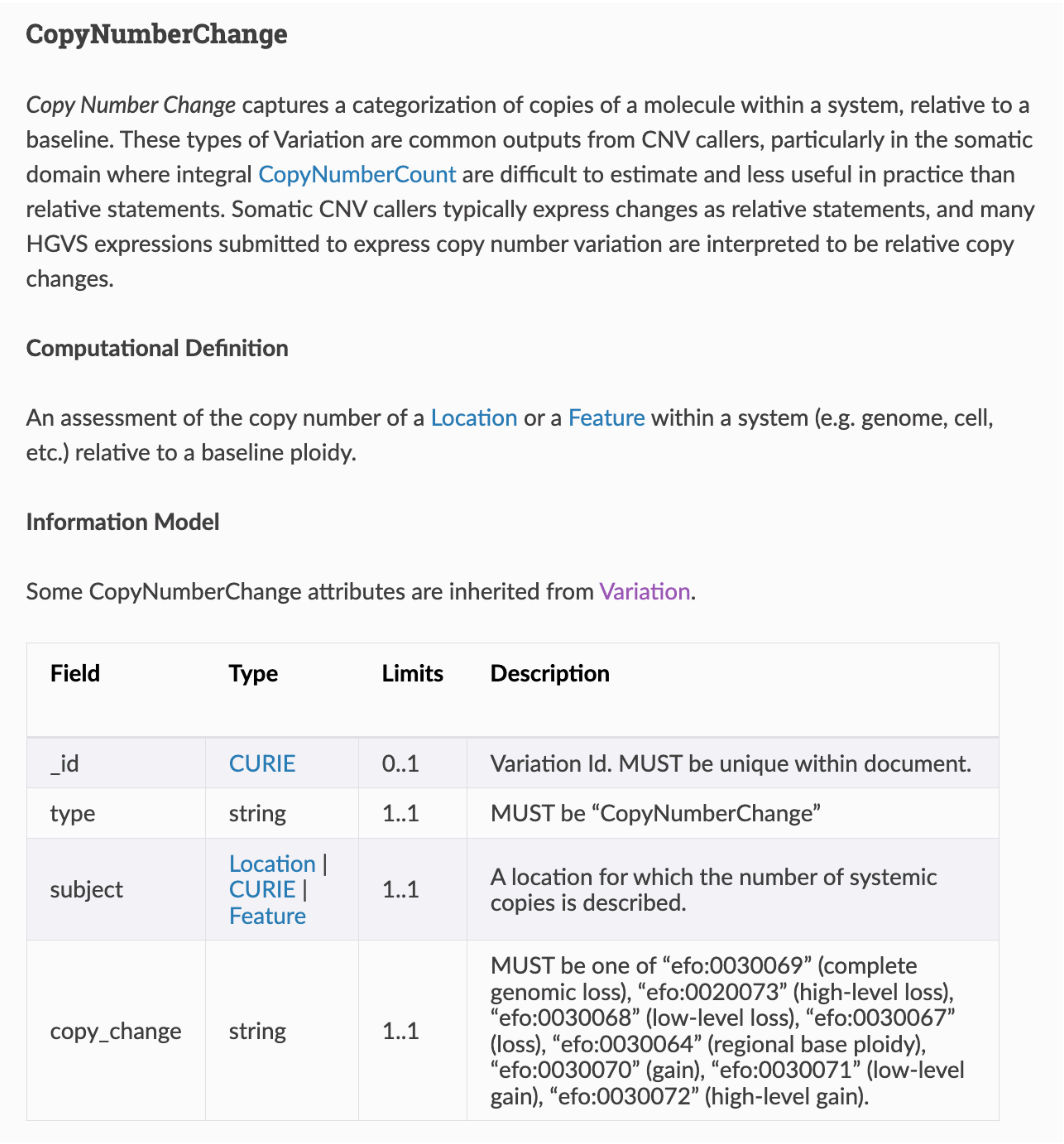
MYCN amplification in neuroblastoma (GSM314026, SJNB8_N cell line)

Lead: Hangjia Zhao

GA4GH Variation Representation Specification

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.



CopyNumberChange

Copy Number Change captures a categorization of copies of a molecule within a system, relative to a baseline. These types of Variation are common outputs from CNV callers, particularly in the somatic domain where integral CopyNumberCount are difficult to estimate and less useful in practice than relative statements. Somatic CNV callers typically express changes as relative statements, and many HGVS expressions submitted to express copy number variation are interpreted to be relative copy changes.

Computational Definition

An assessment of the copy number of a Location or a Feature within a system (e.g. genome, cell, etc.) relative to a baseline ploidy.

Information Model

Some CopyNumberChange attributes are inherited from Variation.

Field	Type	Limits	Description
_id	CURIE	0..1	Variation Id. MUST be unique within document.
type	string	1..1	MUST be “CopyNumberChange”
subject	Location | CURIE | Feature	1..1	A location for which the number of systemic copies is described.
copy_change	string	1..1	MUST be one of “efo:0030069” (complete genomic loss), “efo:0020073” (high-level loss), “efo:0030068” (low-level loss), “efo:0030067” (loss), “efo:0030064” (regional base ploidy), “efo:0030070” (gain), “efo:0030071” (low-level gain), “efo:0030072” (high-level gain).

CNV Term Use Comparison

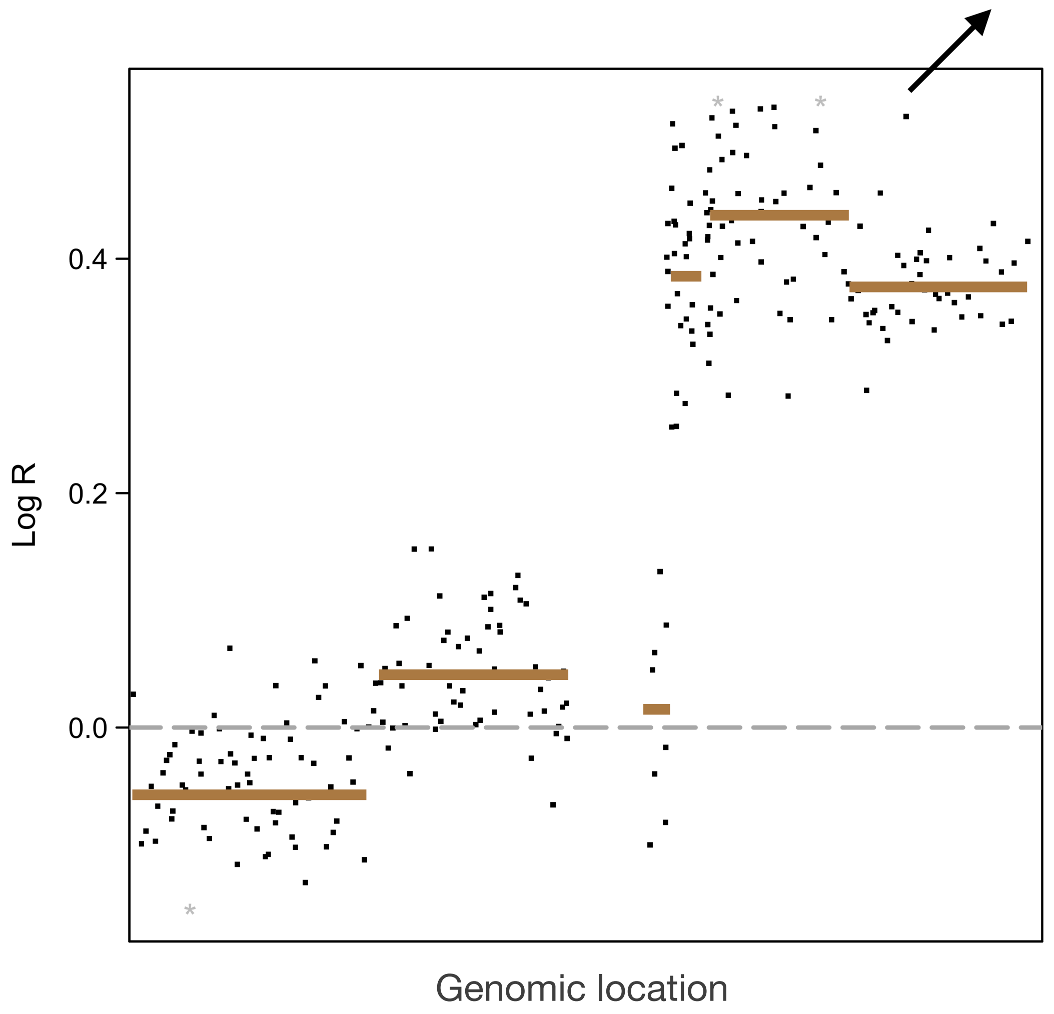
in computational (file/schema) formats

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

labelSeg

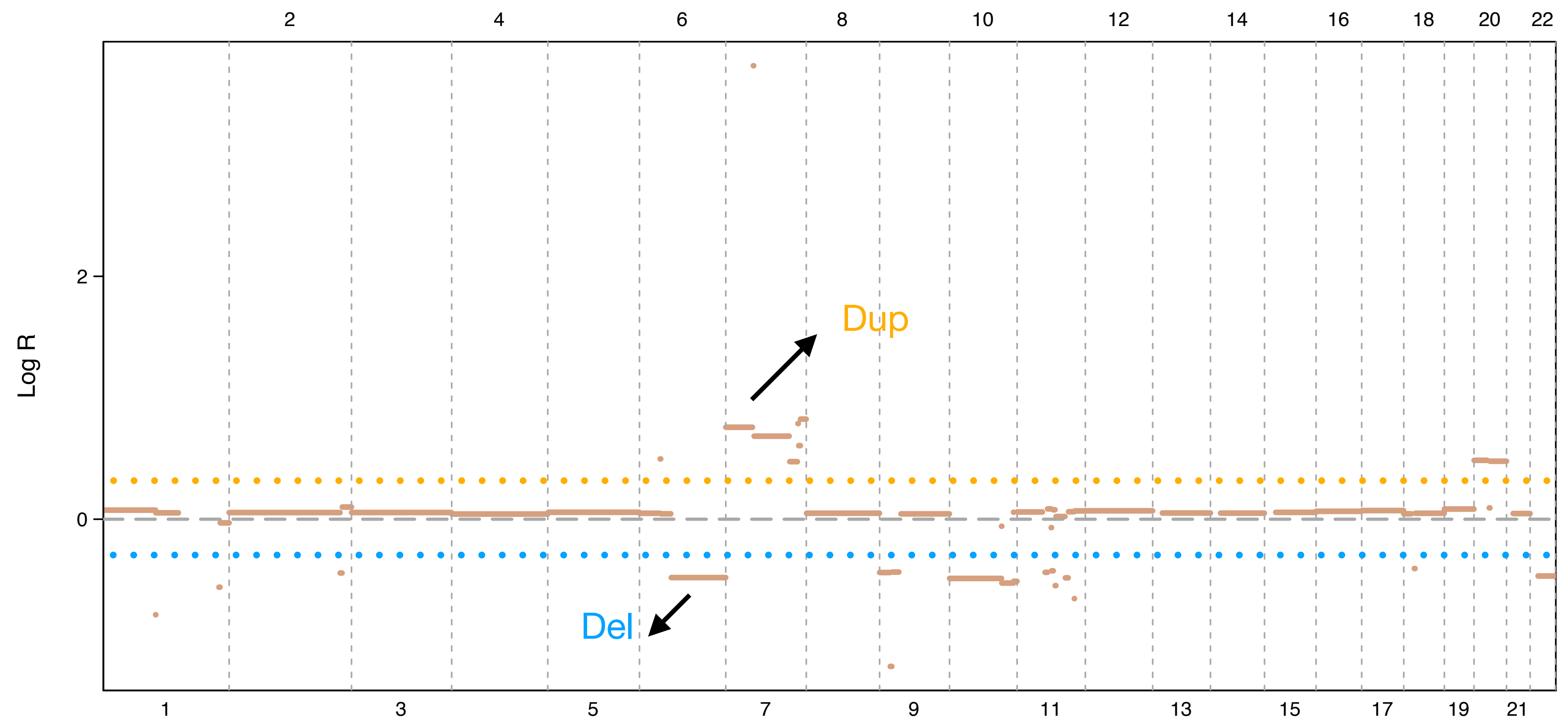
segment annotation for tumor copy number variation profiles

Signal from probes in microarray or from reads in NGS



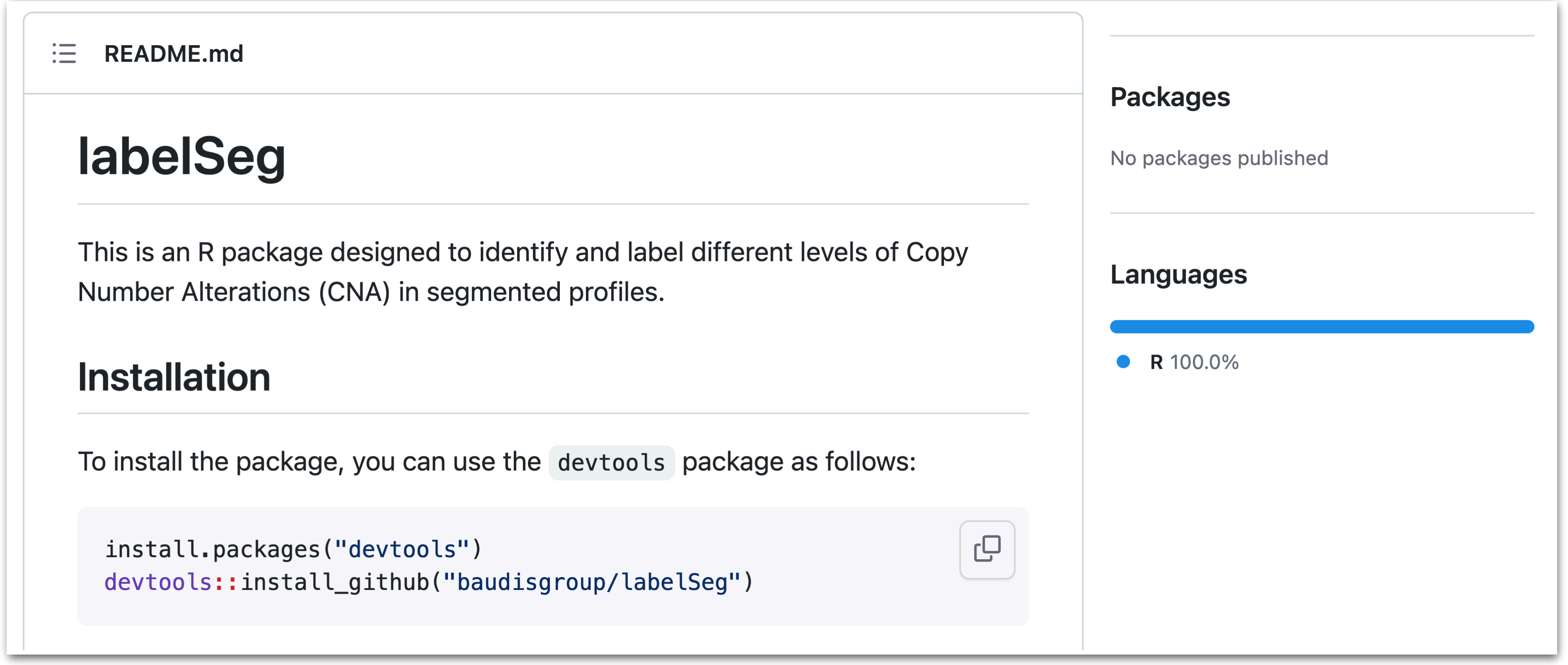
Segmentation

a step to split the chromosomes into regions of equal copy number that accounts for the noise in the data.



threshold for calling

threshold for calling



README.md

labelSeg

This is an R package designed to identify and label different levels of Copy Number Alterations (CNA) in segmented profiles.

Installation

To install the package, you can use the `devtools` package as follows:

```
install.packages("devtools")
devtools::install_github("baudisgroup/labelSeg")
```

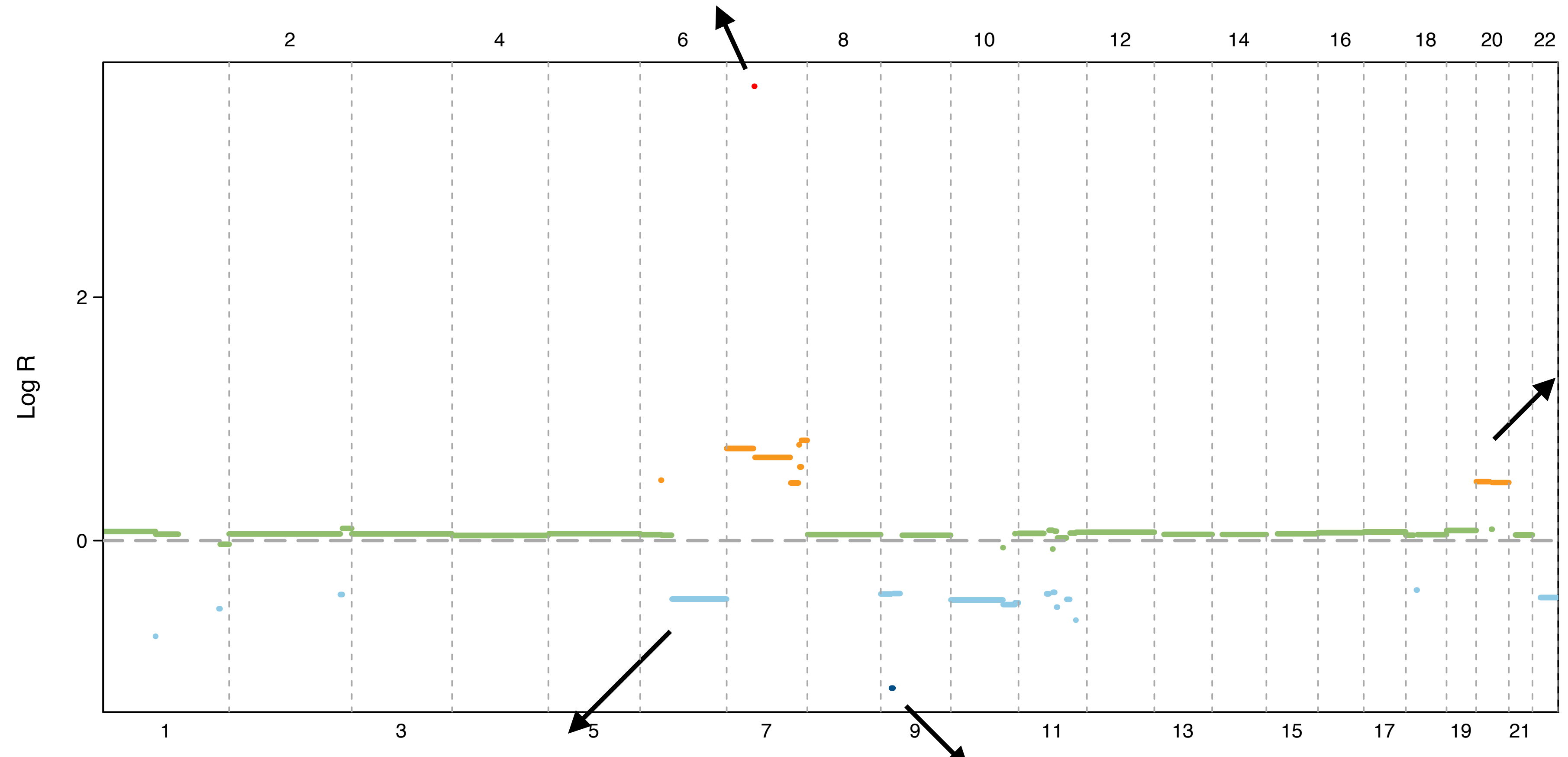
Packages

No packages published

Languages

R 100.0%

High-level dup



Low-level dup

Low-level del

High-level del

Lead: Hangjia Zhao

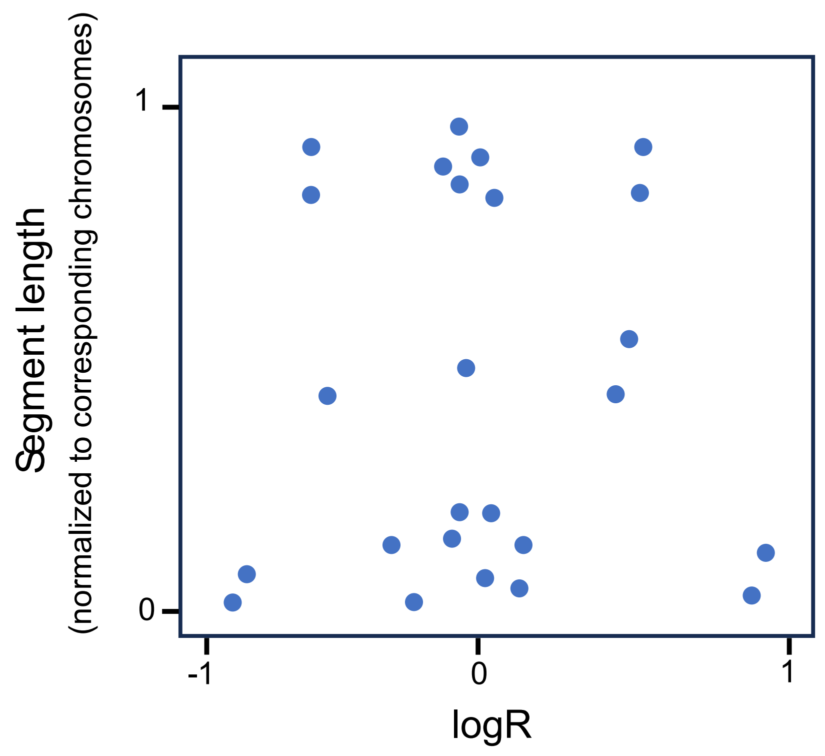
labelSeg

segment annotation for tumor copy number variation profiles

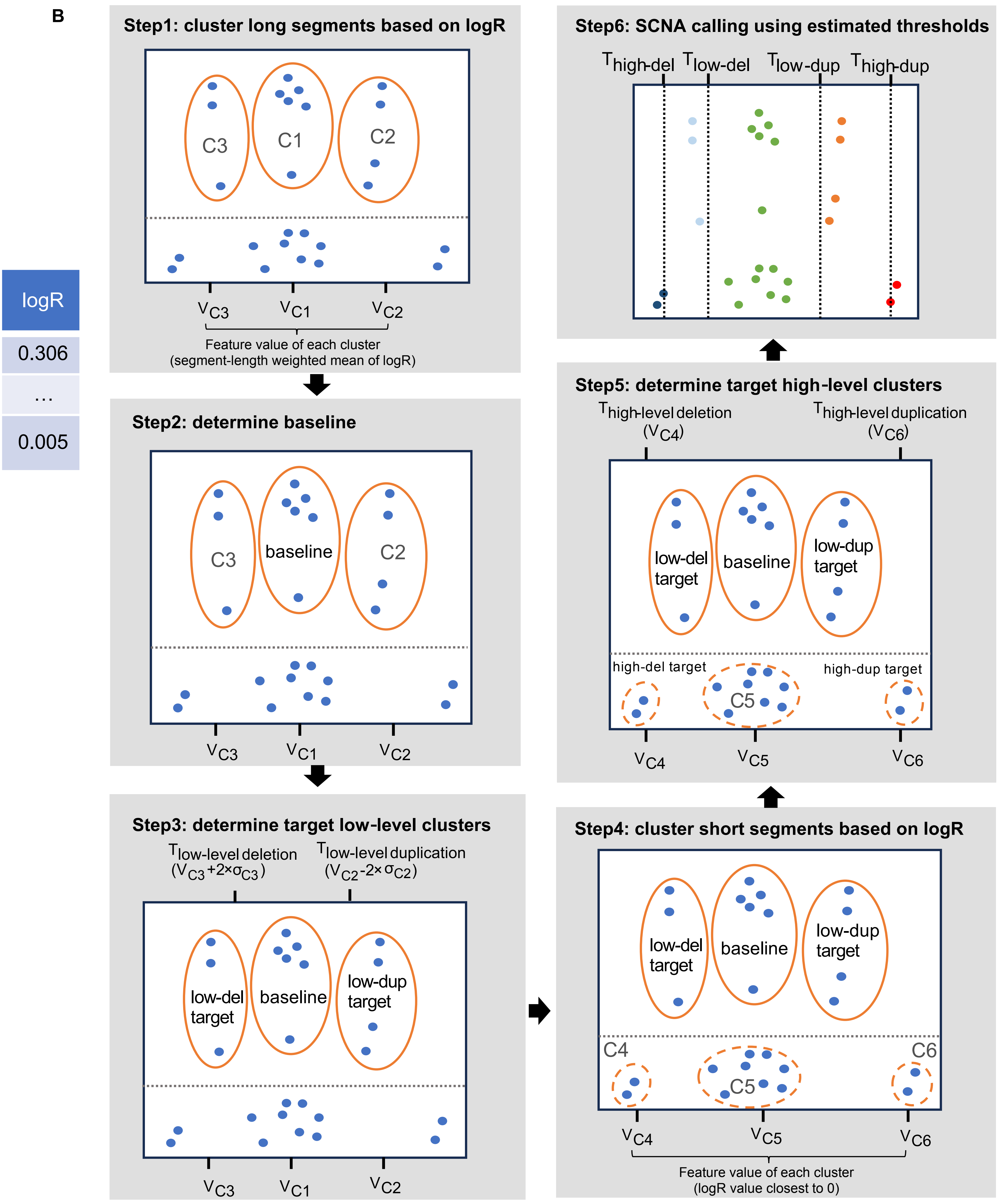
A

INPUT: individual .seg files

Smaple ID	Chr	Start	End	#Mar ker	logR
S1	1	9965785	10911995	450	0.306
…	…	…	…	…	…
S1	X	3236359	155677414	63741	0.005



B


Step1: cluster long segments based on logR

C3
C1
C2
V_{C3}
V_{C1}
V_{C2}
Feature value of each cluster
(segment-length weighted mean of logR)

Step2: determine baseline

Step3: determine target low-level clusters

$T_{\text{low-level deletion}}$
$(V_{C3}+2\times\sigma_{C3})$
$T_{\text{low-level duplication}}$
$(V_{C2}-2\times\sigma_{C2})$
low-del target
baseline
low-dup target
V_{C3}
V_{C1}
V_{C2}

Step4: cluster short segments based on logR

Step5: determine target high-level clusters

Step6: SCNA calling using estimated thresholds

C

OUTPUT: individual .seg files with SCNA labels

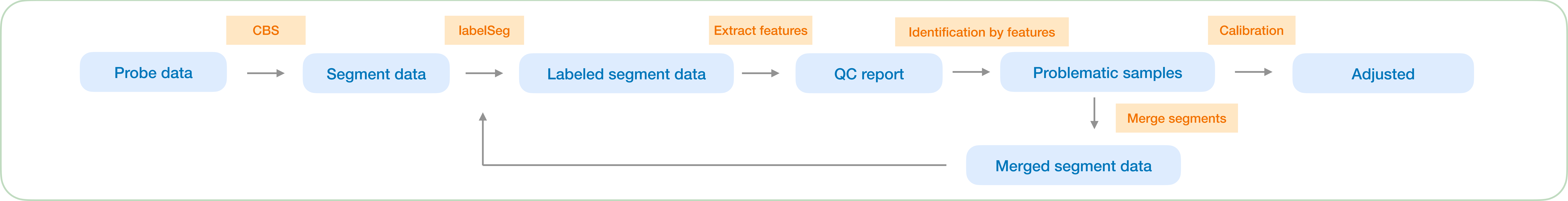
Smaple ID	Chr	Start	End	#Mar ker	logR	Label
S1	1	9965785	10911995	450	0.306	+1
…	…	…	…	…	…	
S1	X	3236359	155677414	63741	0.005	0

Lead: Hangjia Zhao

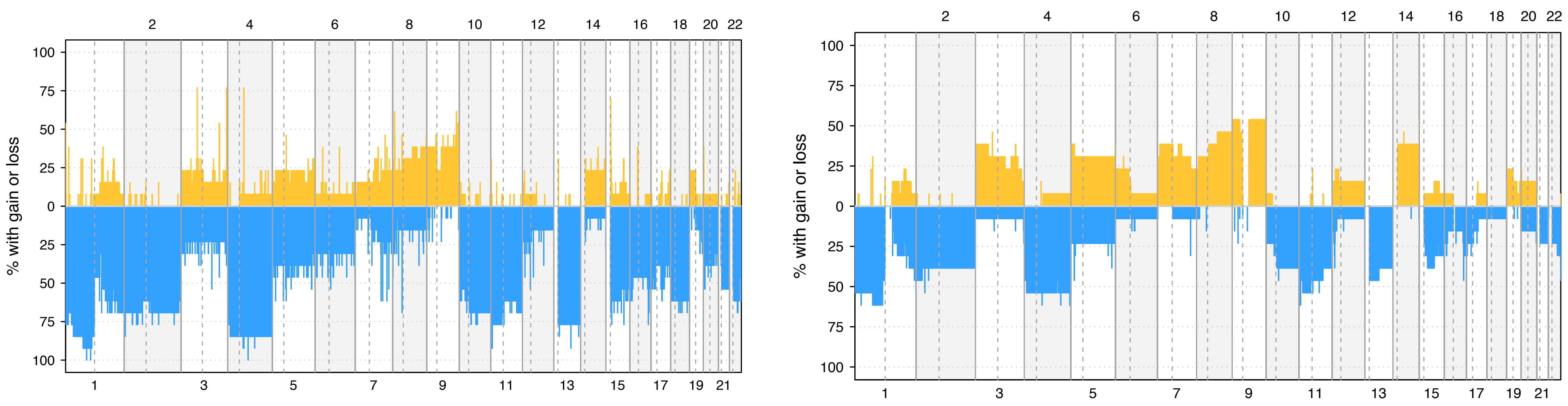
Pipeline Development

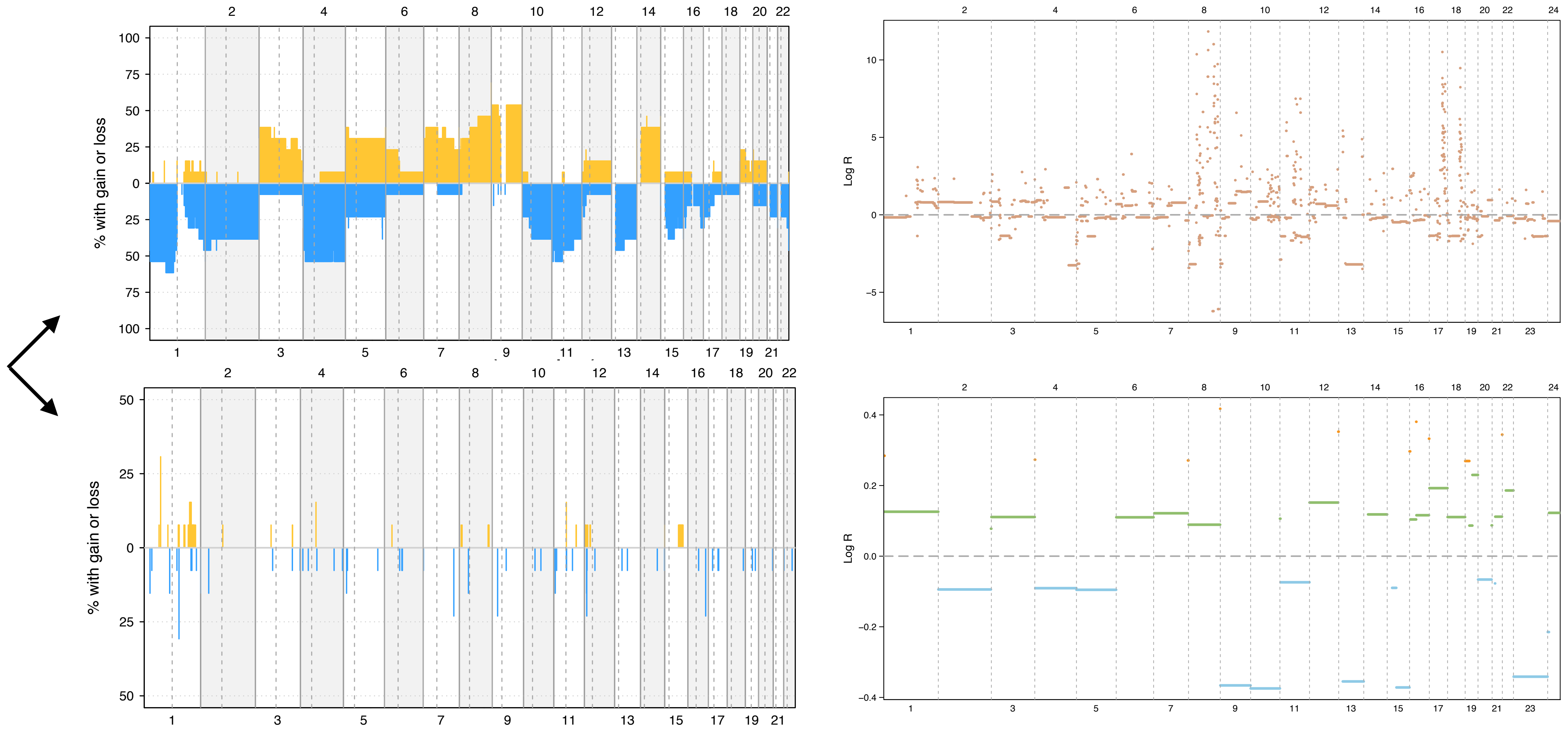
improve CNV calling in large numbers of heterogeneous cancer samples

nextflow



Pituitary Gland Carcinoma

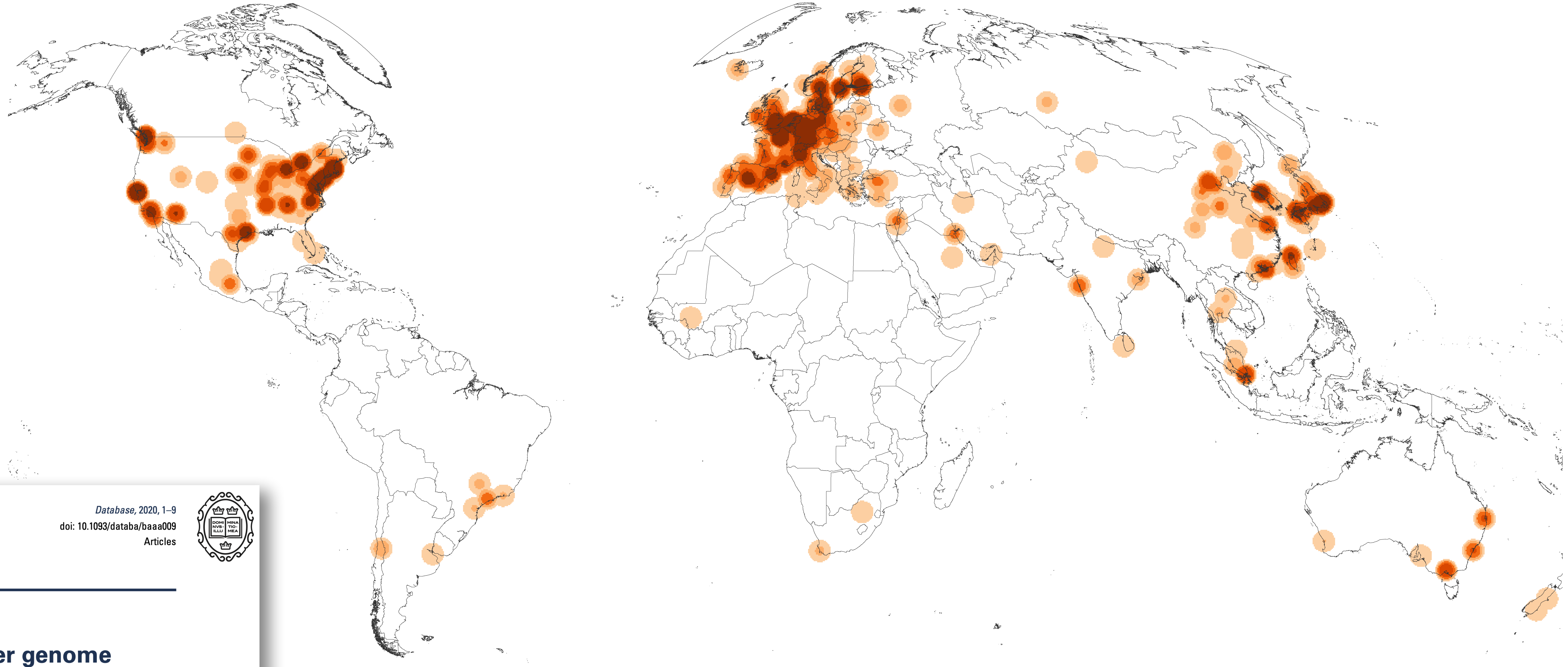


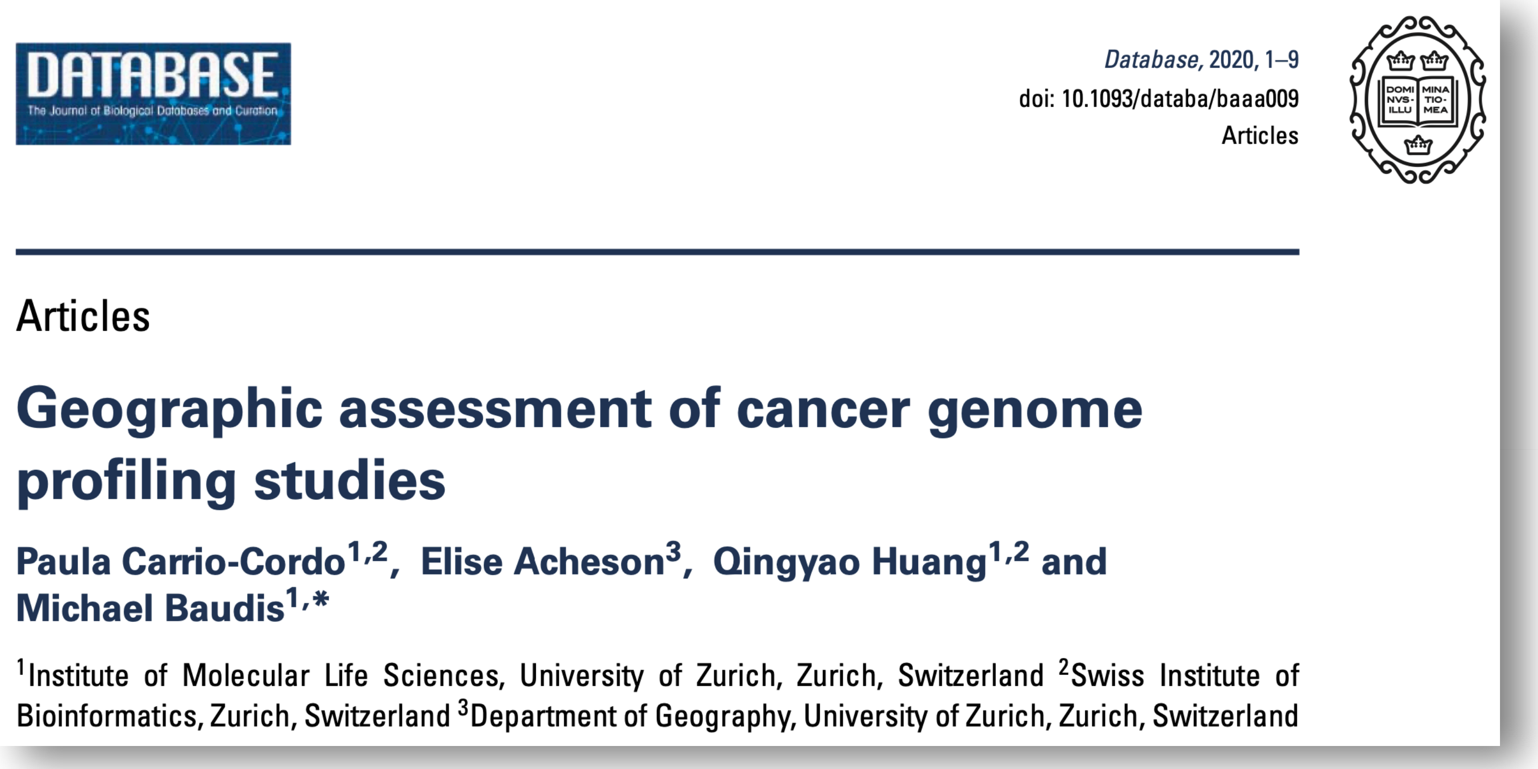


Lead: Hangjia Zhao

Where does Genomic Data Come From?

Geographic bias in published cancer genome profiling studies





DATABASE

Database, 2020, 1–9
doi: 10.1093/databa/baaa009
Articles

Articles

Geographic assessment of cancer genome profiling studies

Paula Carrio-Cordo[1,2], Elise Acheson[3], Qingyao Huang[1,2] and Michael Baudis[1,*]

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland [2]Swiss Institute of Bioinformatics, Zurich, Switzerland [3]Department of Geography, University of Zurich, Zurich, Switzerland

Map of the geographic distribution (by first author affiliation) of the 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets. The numbers are derived from the 3'240 publications registered in the Progenetix database.

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

GENOMICS

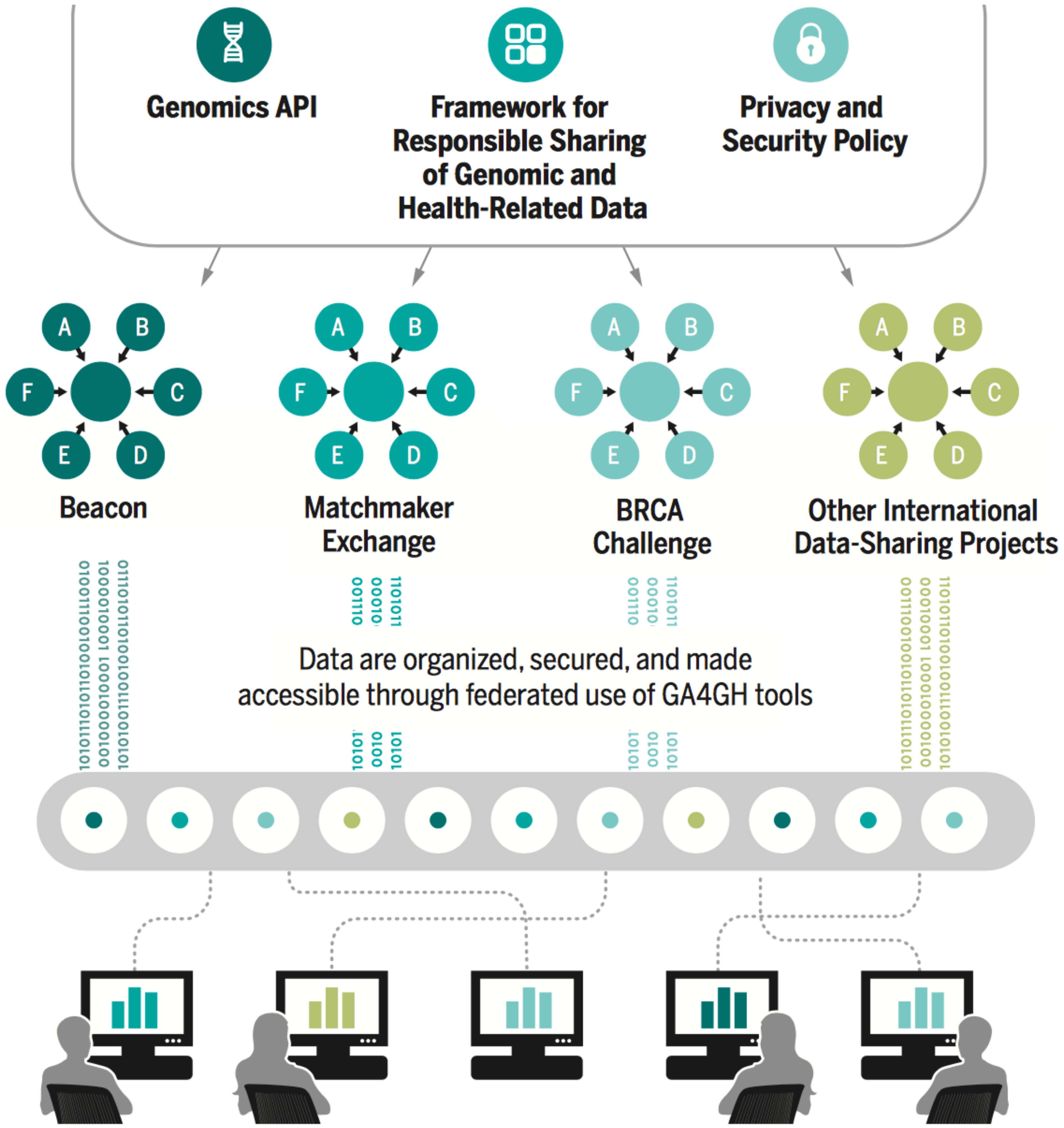
A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems

The Global Alliance for Genomics and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.



Genomics API

Framework for Responsible Sharing of Genomic and Health-Related Data

Privacy and Security Policy

Beacon

Matchmaker Exchange

BRCA Challenge

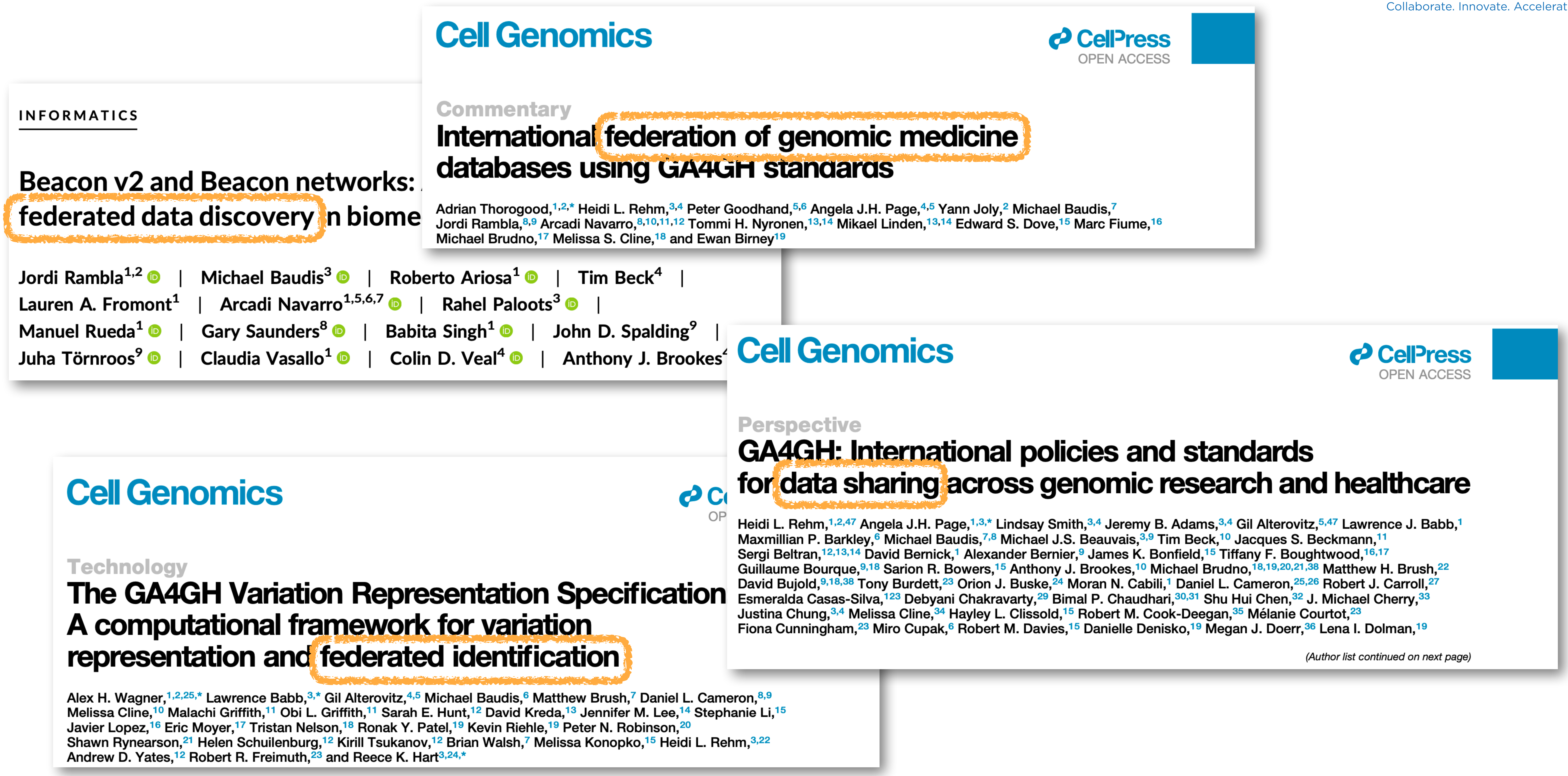
Other International Data-Sharing Projects

Data are organized, secured, and made accessible through federated use of GA4GH tools

Global Alliance for Genomics & Health

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.



INFORMATICS

Beacon v2 and Beacon networks:
federated data discovery n biome

Jordi Rambla[1,2] | Michael Baudis[3] | Roberto Ariosa[1] | Tim Beck[4] |
Lauren A. Fromont[1] | Arcadi Navarro[1,5,6,7] | Rahel Paloots[3] |
Manuel Rueda[1] | Gary Saunders[8] | Babita Singh[1] | John D. Spalding[9] |
Juha Törnroos[9] | Claudia Vasallo[1] | Colin D. Veal[4] | Anthony J. Brookes

Cell Genomics

CellPress OPEN ACCESS

Commentary

International federation of genomic medicine databases using GA4GH standards

Adrian Thorogood,[1,2,*] Heidi L. Rehm,[3,4] Peter Goodhand,[5,6] Angela J.H. Page,[4,5] Yann Joly,[2] Michael Baudis,[7] Jordi Rambla,[8,9] Arcadi Navarro,[8,10,11,12] Tommi H. Nyronen,[13,14] Mikael Linden,[13,14] Edward S. Dove,[15] Marc Fiume,[16] Michael Brudno,[17] Melissa S. Cline,[18] and Ewan Birney[19]

Cell Genomics

CellPress OPEN ACCESS

Perspective

GA4GH: International policies and standards for data sharing across genomic research and healthcare

Heidi L. Rehm,[1,2,47] Angela J.H. Page,[1,3,*] Lindsay Smith,[3,4] Jeremy B. Adams,[3,4] Gil Alterovitz,[5,47] Lawrence J. Babb,[1] Maxmillian P. Barkley,[6] Michael Baudis,[7,8] Michael J.S. Beauvais,[3,9] Tim Beck,[10] Jacques S. Beckmann,[11] Sergi Beltran,[12,13,14] David Bernick,[1] Alexander Bernier,[9] James K. Bonfield,[15] Tiffany F. Boughtwood,[16,17] Guillaume Bourque,[9,18] Sarion R. Bowers,[15] Anthony J. Brookes,[10] Michael Brudno,[18,19,20,21,38] Matthew H. Brush,[22] David Bujold,[9,18,38] Tony Burdett,[23] Orion J. Buske,[24] Moran N. Cabili,[1] Daniel L. Cameron,[25,26] Robert J. Carroll,[27] Esmeralda Casas-Silva,[123] Debyani Chakravarty,[29] Bimal P. Chaudhari,[30,31] Shu Hui Chen,[32] J. Michael Cherry,[33] Justina Chung,[3,4] Melissa Cline,[34] Hayley L. Clissold,[15] Robert M. Cook-Deegan,[35] Mélanie Courtot,[23] Fiona Cunningham,[23] Miro Cupak,[6] Robert M. Davies,[15] Danielle Denisko,[19] Megan J. Doerr,[36] Lena I. Dolman,[19]

(Author list continued on next page)

Cell Genomics

Technology

The GA4GH Variation Representation Specification: A computational framework for variation representation and federated identification

Alex H. Wagner,[1,2,25,*] Lawrence Babb,[3,*] Gil Alterovitz,[4,5] Michael Baudis,[6] Matthew Brush,[7] Daniel L. Cameron,[8,9] Melissa Cline,[10] Malachi Griffith,[11] Obi L. Griffith,[11] Sarah E. Hunt,[12] David Kreda,[13] Jennifer M. Lee,[14] Stephanie Li,[15] Javier Lopez,[16] Eric Moyer,[17] Tristan Nelson,[18] Ronak Y. Patel,[19] Kevin Riehle,[19] Peter N. Robinson,[20] Shawn Rynearson,[21] Helen Schuilenburg,[12] Kirill Tsukanov,[12] Brian Walsh,[7] Melissa Konopko,[15] Heidi L. Rehm,[3,22] Andrew D. Yates,[12] Robert R. Freimuth,[23] and Reece K. Hart[3,24,*]

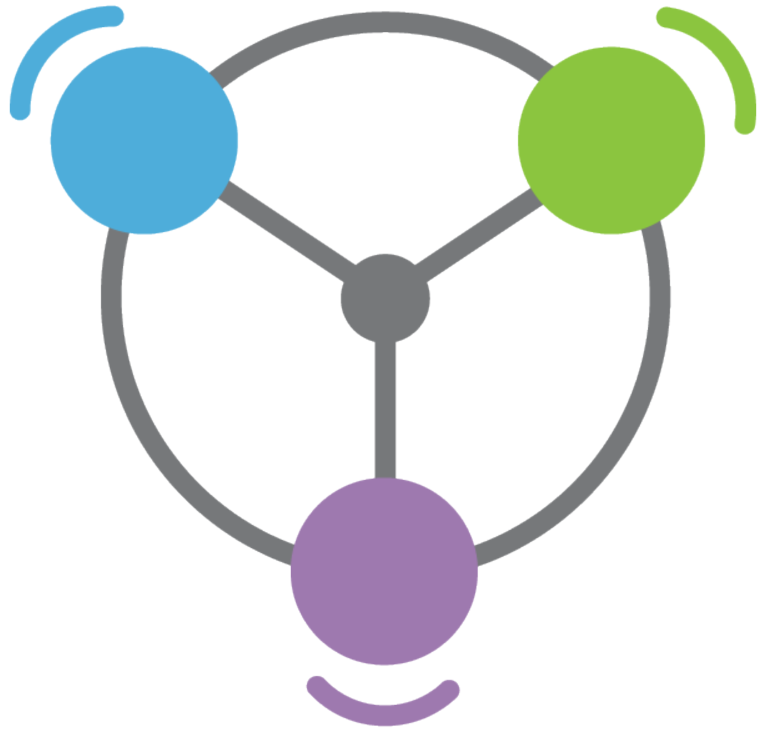
Global Genomic Data Sharing Can...

- Demonstrate patterns in health & disease
- Increase statistical significance of analyses
- Lead to “stronger” variant interpretations
- Increase accurate diagnosis
- Advance precision medicine

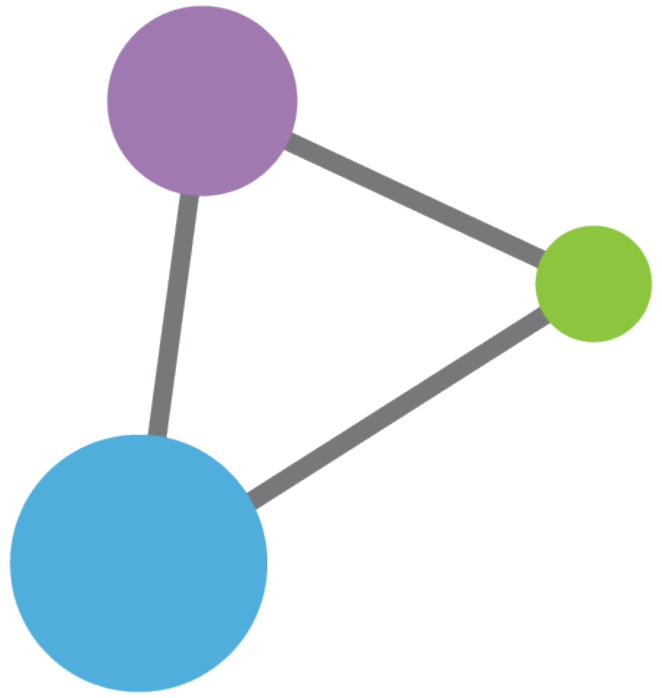
Different Approaches to Data Sharing

Centralized Genomic Knowledge Bases

Data Commons
Trusted, controlled repository of multiple datasets



Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

Different Approaches to Data Sharing

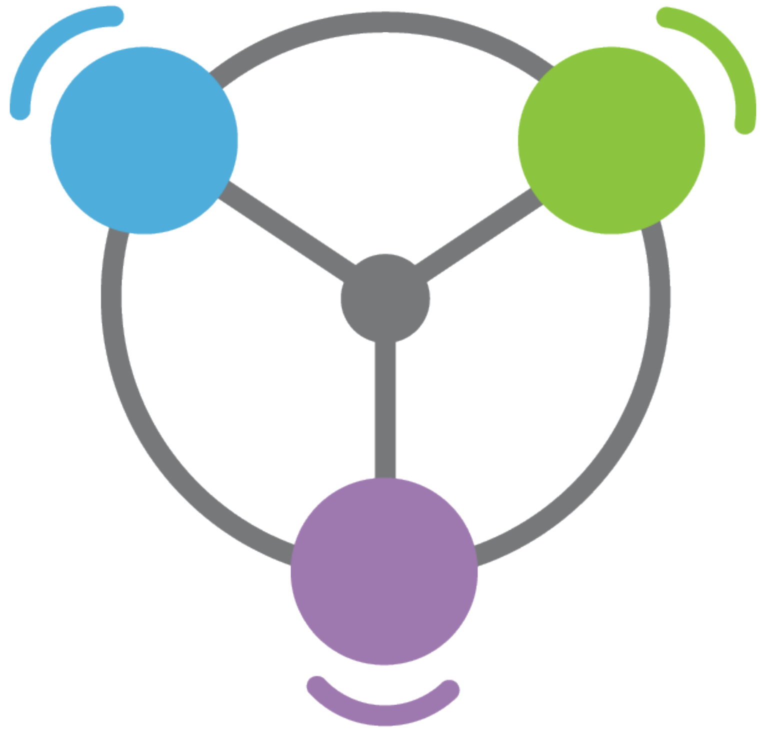
progenetix

Centralized Genomic Knowledge Bases

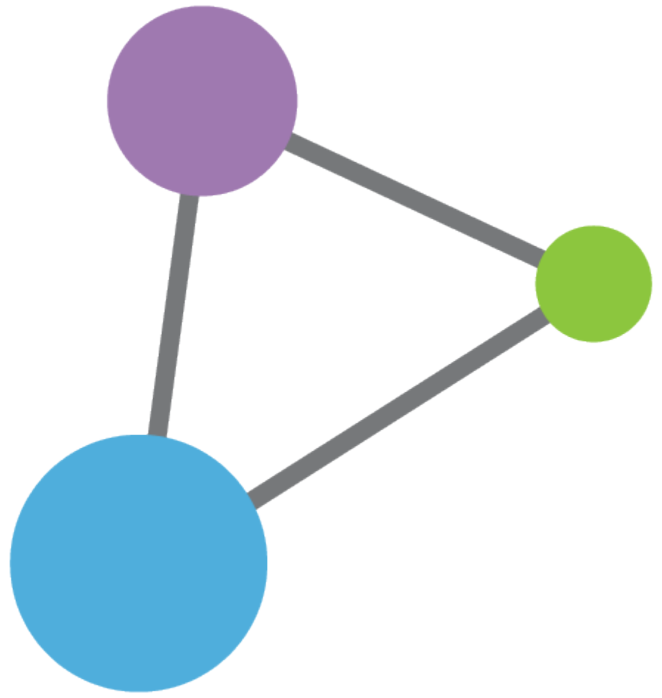
European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network



Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

Different Approaches to Data Sharing

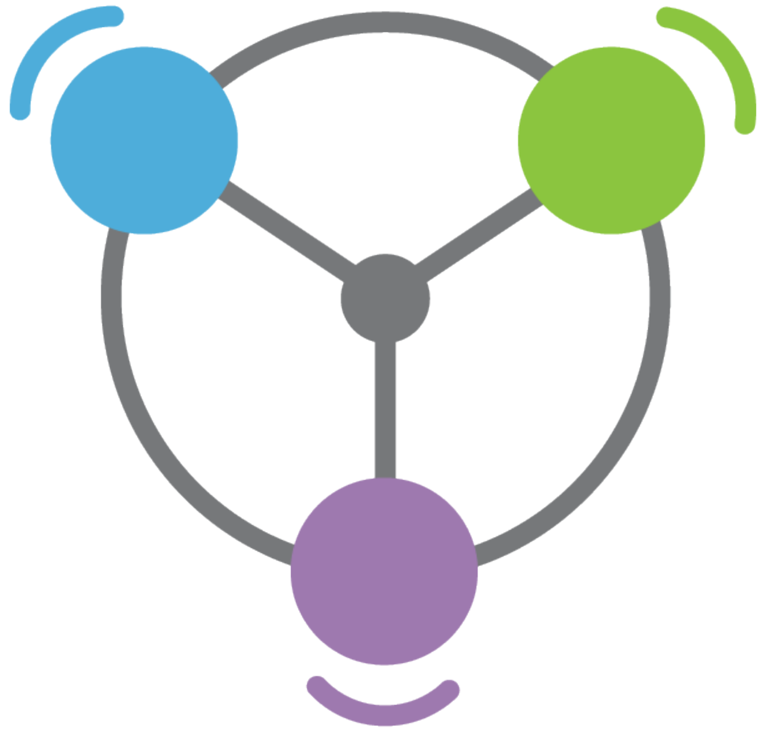
progenetix

Centralized Genomic Knowledge Bases

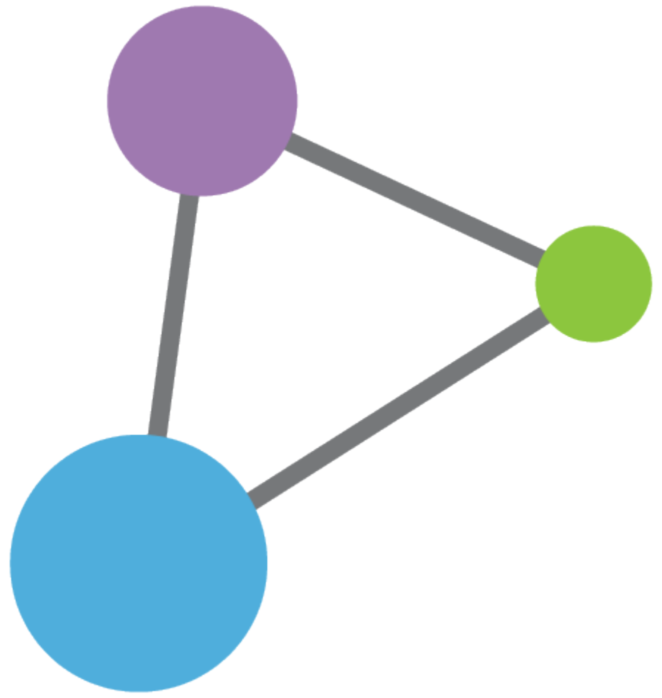
European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network



Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

Different Approaches to Data Sharing

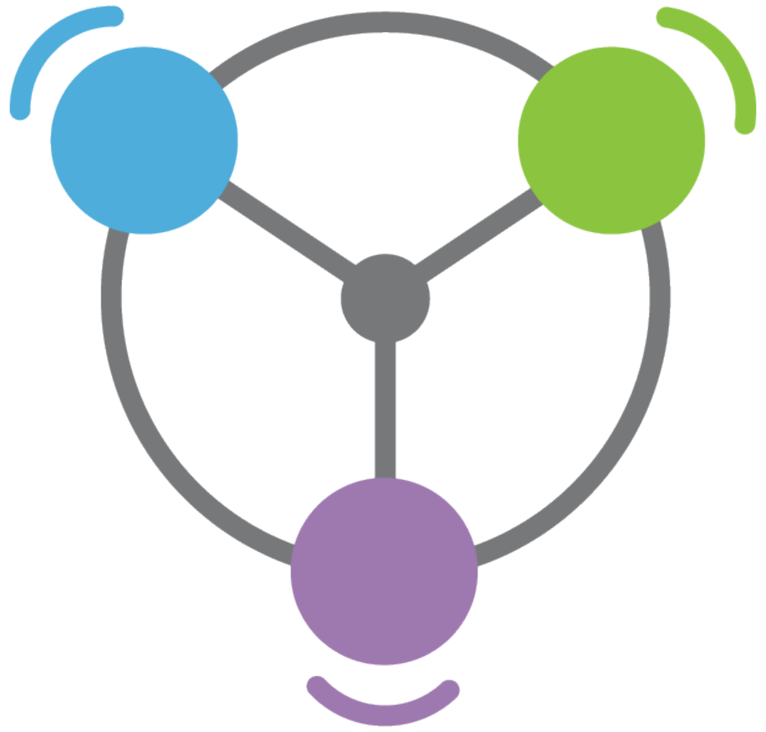
progenetix

Centralized Genomic Knowledge Bases

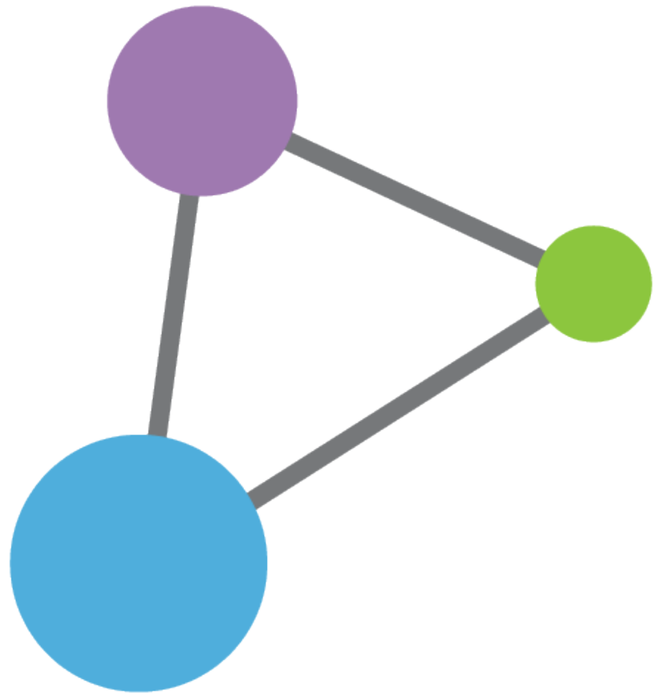
European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network



Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

The EGA

European Genome-Phenome Archive

Long term secure archive for human biomedical research sensitive data, with focus on reuse of the data for further research (or “*broad and responsible use of genomic data*”)

EUROPEAN GENOME-PHENOME ARCHIVE

CRG Centre for Genomic Regulation

EMBL-EBI

Global Alliance for Genomics & Health

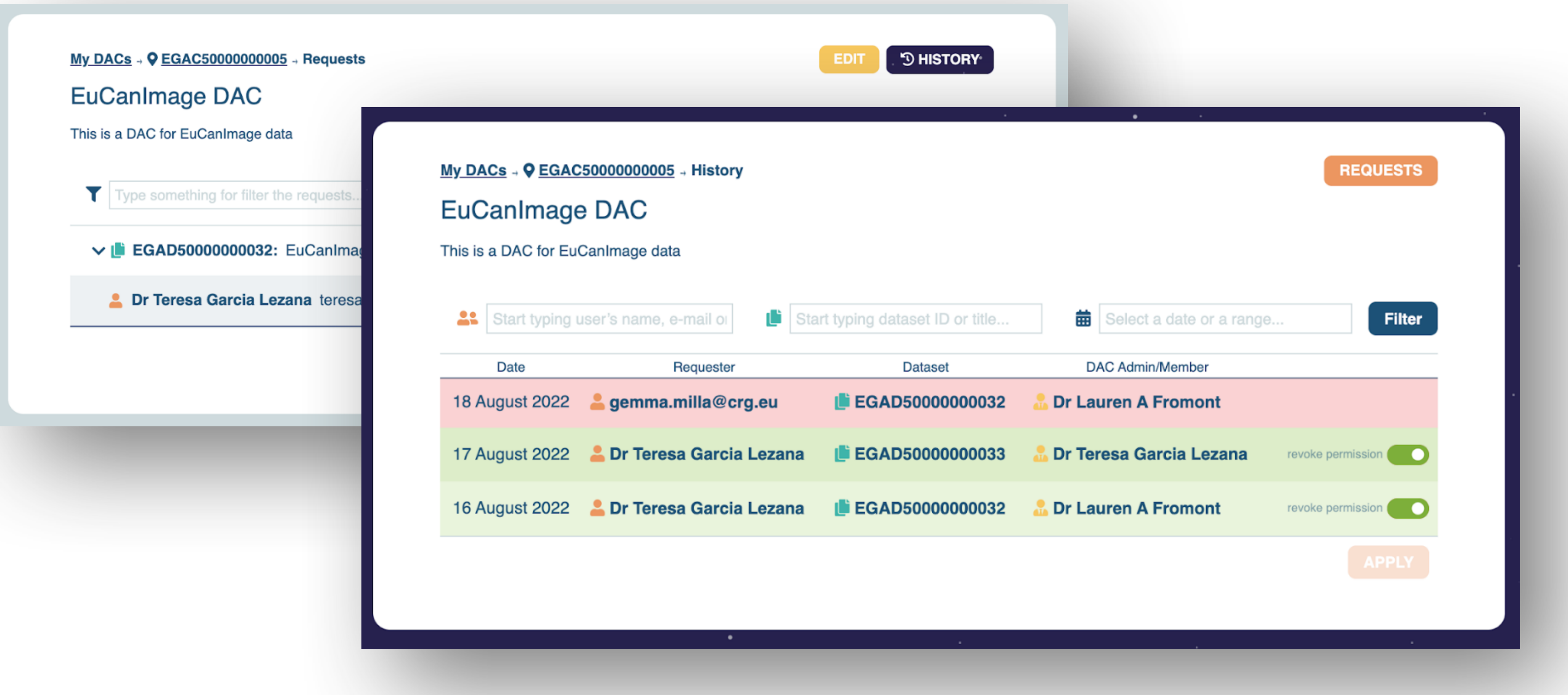
DRIVER PROJECT

Slide: adapted from Jordi Rambla@ GA4GH 2023

The EGA

European Genome-Phenome Archive

- EGA “owns” nothing; data controllers tell who is authorized to access ***their*** datasets
- EGA admins provide smooth “all or nothing" data sharing process


EDIT
HISTORY
Type something for filter the requests...
REQUESTS
Start typing user's name, e-mail o
Start typing dataset ID or title...
Select a date or a range...
Filter
APPLY

My DACs → EGAC50000000005 → Requests

EuCanImage DAC

This is a DAC for EuCanImage data

EGAD50000000032: EuCanImag

Dr Teresa Garcia Lezana teresa

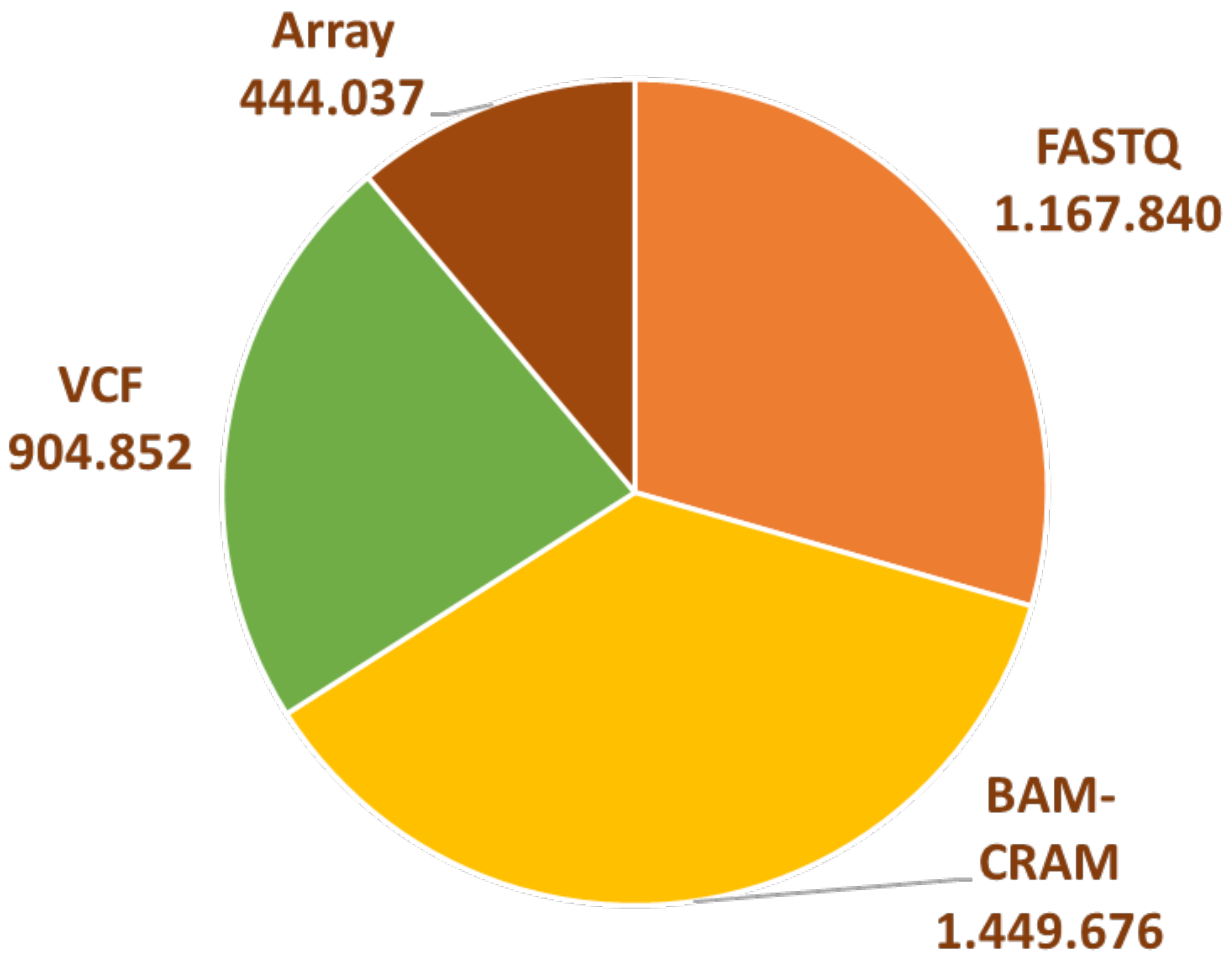
My DACs → EGAC50000000005 → History

EuCanImage DAC

This is a DAC for EuCanImage data

Date	Requester	Dataset	DAC Admin/Member	
18 August 2022	gemma.milla@crg.eu	EGAD50000000032	Dr Lauren A Fromont	
17 August 2022	Dr Teresa Garcia Lezana	EGAD50000000033	Dr Teresa Garcia Lezana	revoke permission
16 August 2022	Dr Teresa Garcia Lezana	EGAD50000000032	Dr Lauren A Fromont	revoke permission

Files



4,328 Studies released
10,470 Datasets
2,309 Data Access Committees

Slide: adapted from Jordi Rambla@ GA4GH 2023

Different Approaches to Data Sharing

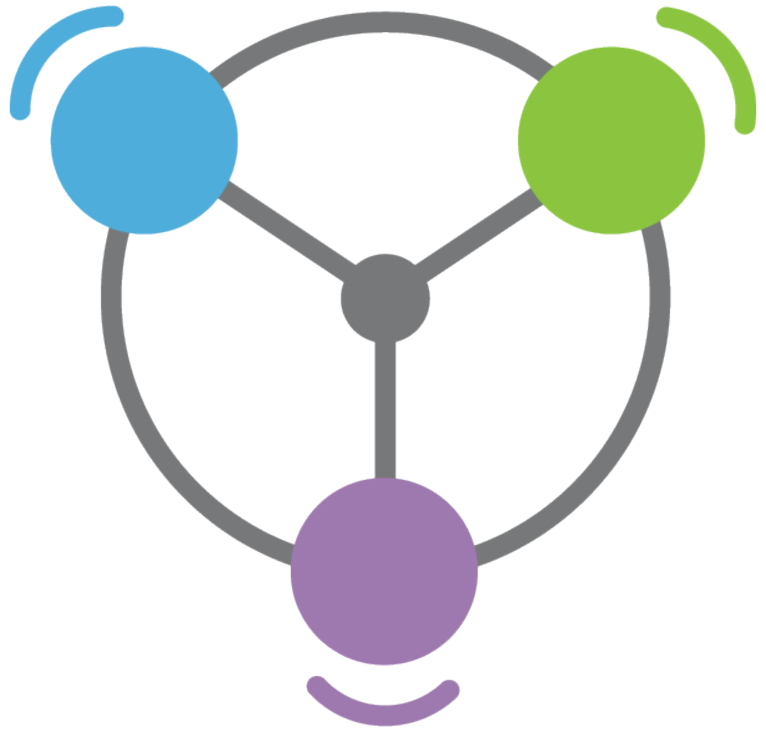
progenetix

Centralized Genomic Knowledge Bases

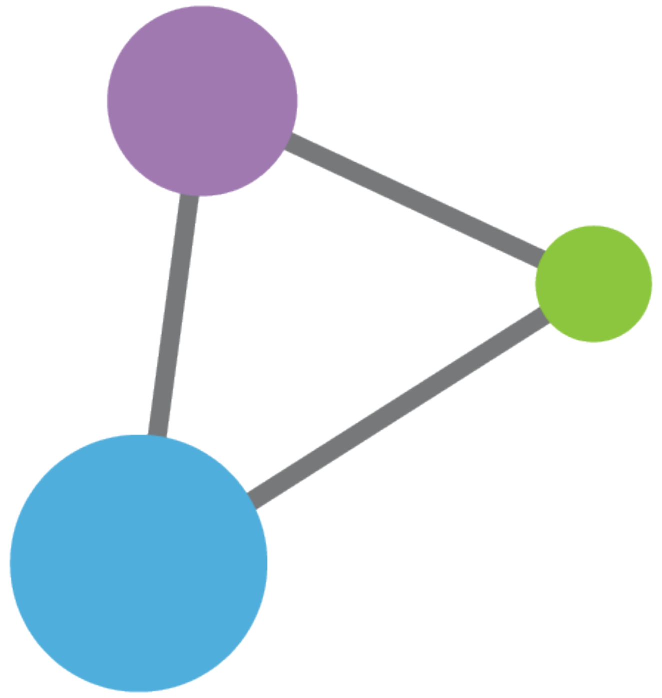
European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network



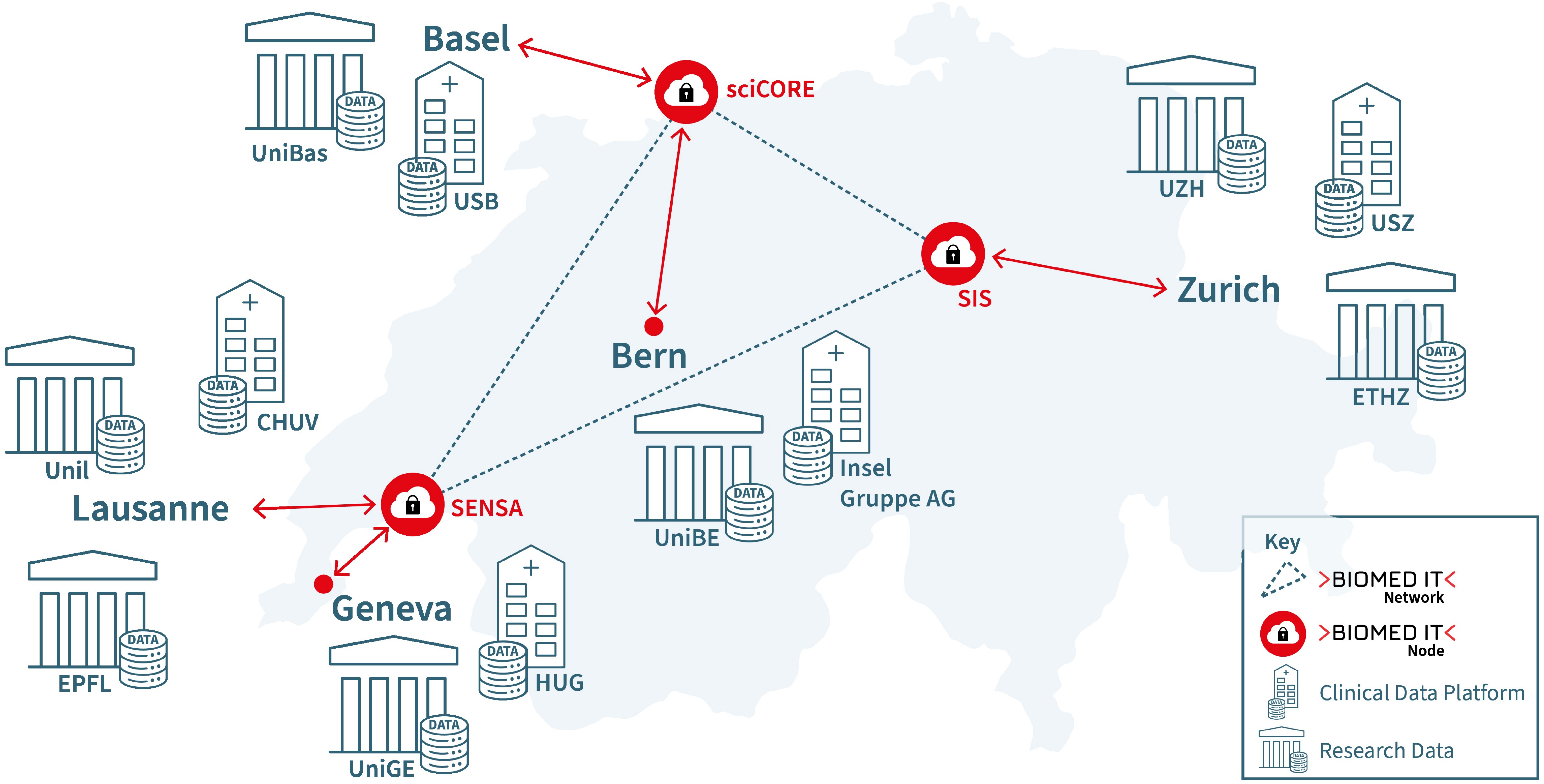
Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

The Swiss Personalized Health Network

Swiss Personalized Health Network SPHN



Strategic Focus Area
Personalized Health and Related Technologies

ehealthsuisse

THE LOOP ZURICH MEDICAL RESEARCH CENTER

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

Personalized Health Alliance
Basel-Zurich

SWISS BIOBANKING PLATFORM

SCTO

SAKK

SSPH+ SWISS SCHOOL OF PUBLIC HEALTH

life sciences cluster basel

swissuniversities

Universitäre Medizin Schweiz
Médecine Universitaire Suisse

SIB

Personalized Health Informatics Group

SPHN Data Coordination Center (DCC)
BioMedIT Network

University Hospital Basel

USZ Universitäts Spital Zürich

HUG Hôpitaux Universitaires Genève

CHUV Centre hospitalier universitaire vaudois

INSELSPITAL
UNIVERSITÄTSSPITAL BERN
HOPITAL UNIVERSITAIRE DE BERNE
BERN UNIVERSITY HOSPITAL

Different Approaches to Data Sharing

progenetix

Centralized Genomic Knowledge Bases

European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network

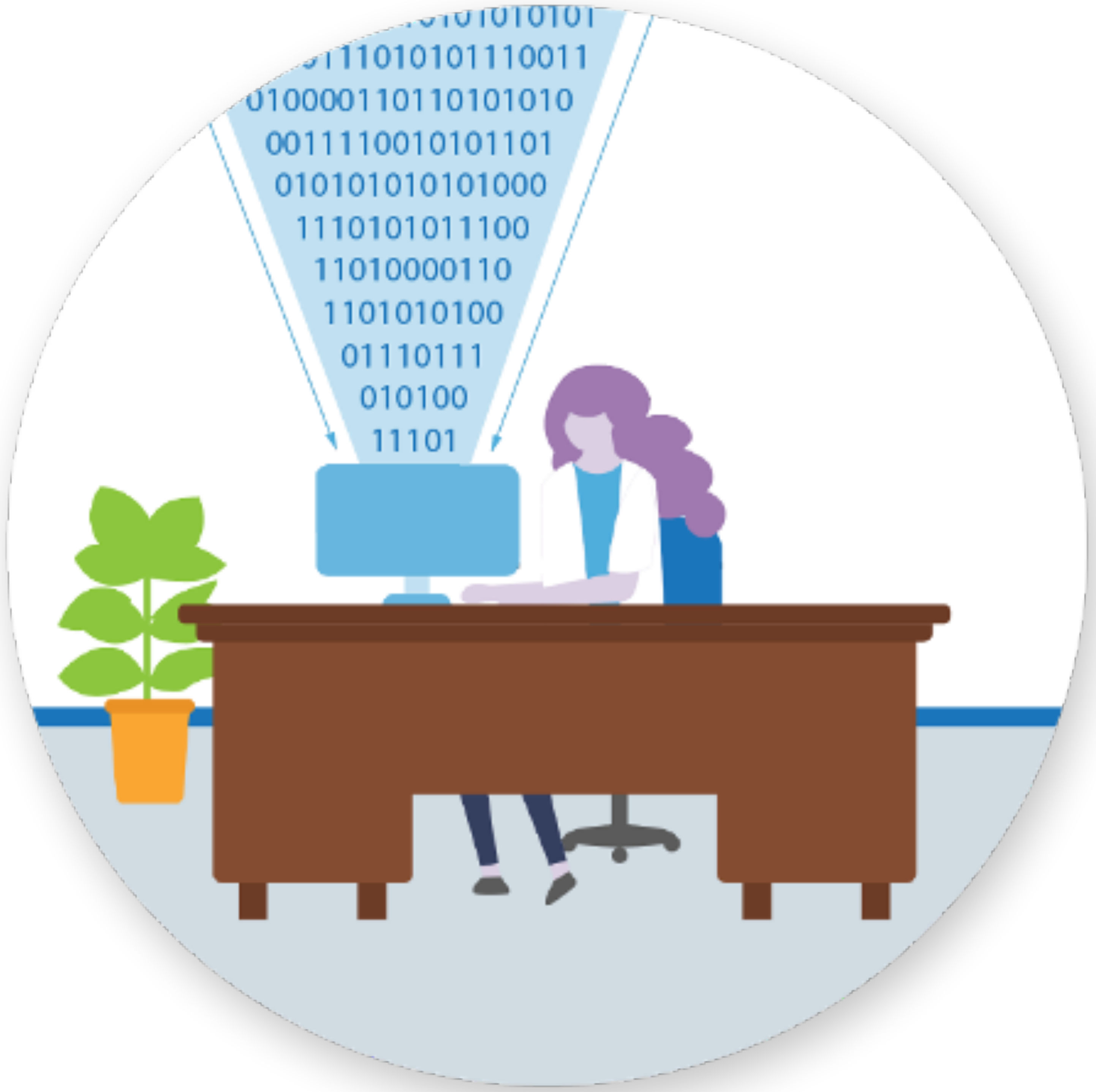
Hub and Spoke
Common data elements, access, and usage rules

Linkage of distributed and disparate datasets

Federation

A New Paradigm for Data Sharing

FROM



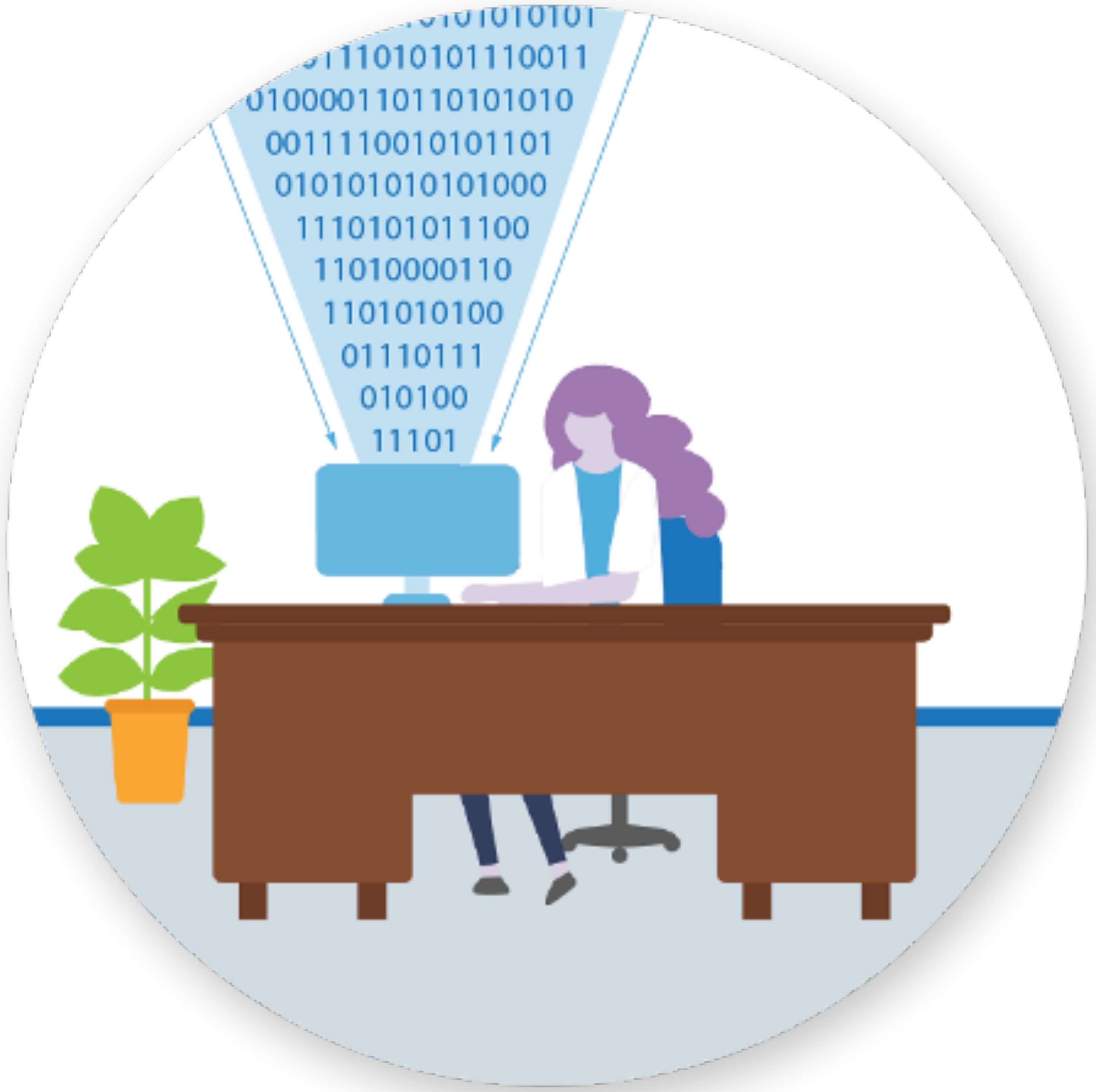
Data Copying

TO

Data Visiting

A New Paradigm for Data Sharing

FROM



Data Copying

STANDARDS

TO

Data Visiting

Overview of GA4GH standards and frameworks

🟩 Approved 🟨 Ongoing
⬜ In Development

- **Clin/Pheno Data Capture**: 🟩 Phenopackets; ⬜ Pedigree Representation; ⬜ Cohort Representation
- **Cloud**: 🟩 Workflow Execution Service; 🟩 Tool Registry Service; 🟩 Data Repository Service; 🟩 Task Execution Service; 🟨 Cloud Testbed Interoperability
- **Discovery**: 🟩 Beacon; 🟩 Service Info; 🟩 Service Registry; 🟩 Data Connect
- **Data Security**: 🟩 Authentication & Authorization Infrastructure; 🟩 Data Security Infrastructure Policy; ⬜ Risk Assessment; ⬜ Bad Actors in Research Environments; ⬜ Cloud Security & Privacy
- **Data Use & Researcher Identity**: 🟩 Data Use Ontology; 🟩 GA4GH Passports; 🟩 Machine Readable Consent Guidance; 🟩 Data Access Committee Review Standards Toolkit
- **Genomic Knowledge Standards**: 🟩 Variation Representation; ⬜ Variation Annotation; ⬜ Sequence Annotation
- **Large Scale Genomics**: 🟩 htsget API; 🟩 refget API; 🟩 SAM/BAM/ CRAM; 🟩 VCF; 🟩 Crypt4GH; 🟩 rnaget API; 🟩 BED File Format
- **Regulatory & Ethics**: 🟩 Framework for responsible data sharing; 🟩 Consent Toolkit; 🟩 20+ other policy tools/frameworks; 🟨 Genetic Discrimination Toolkit; 🟨 GDPR Forum; 🟨 Public Attitudes for genomic policy

Phenopackets v2

Phenopackets is a standard schema for sharing phenotypic information.

Approved: June 24, 2021

Example Users

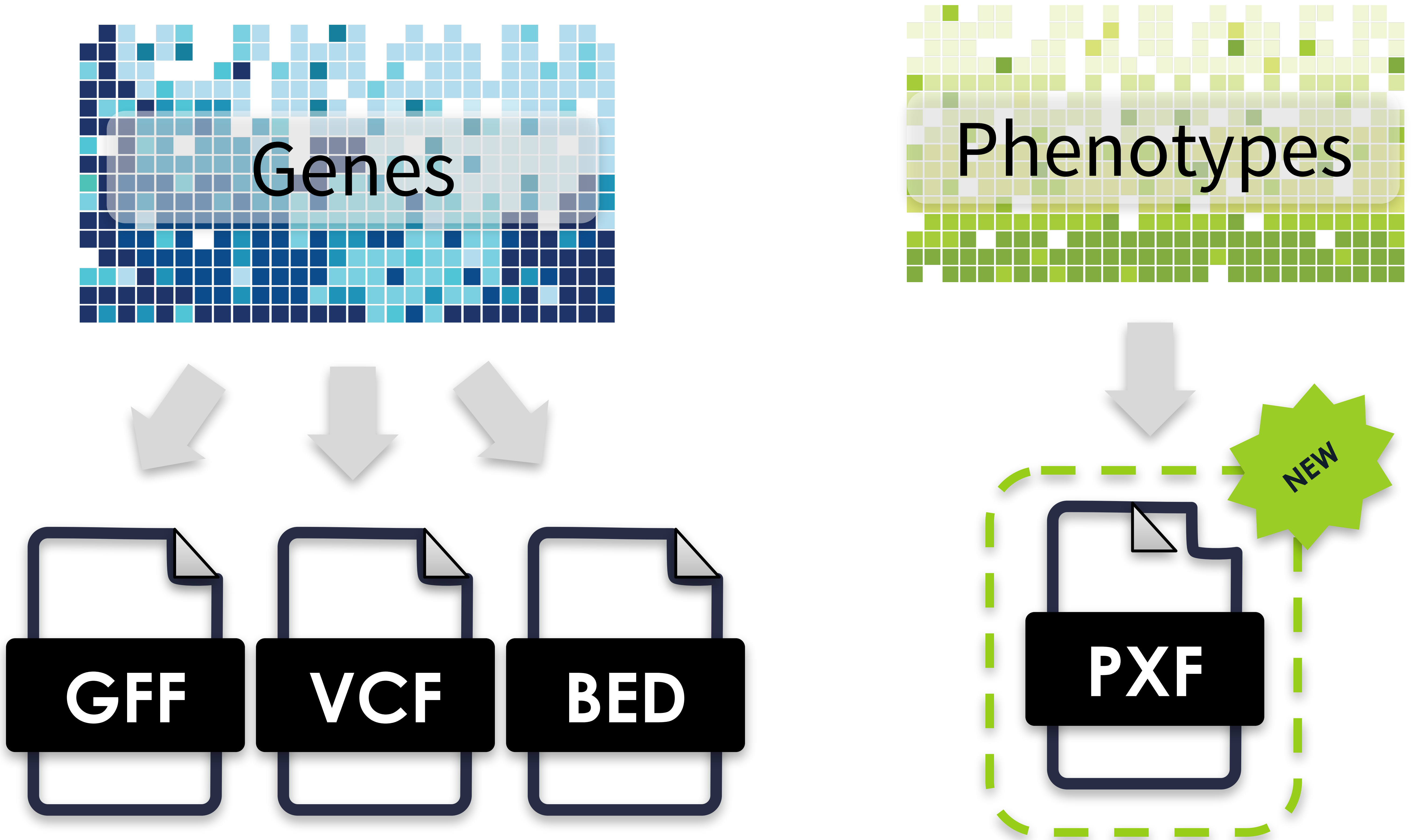
Cafe Variome

NIH

AMED

PHENOTIPS™

RD Connect



VCF/BCF

The Variant Call Format (VCF) specifies the format of a text file used in bioinformatics for storing gene sequence variations. The Binary Call Format (BCF) is the Binary equivalent, smaller and more efficient to process.

Software Libraries: htslib | htsjdk

Tools: Samtools | BCFtools

Databases: European Variation Archive (EVA) | dbGAP | dbSNP | 1000 Genomes Projects / IGSR

Genome Browsers: ENSEMBL | JBrowse | UCSC Genome Browser

Example Users

All of Us RESEARCH PROGRAM

BROAD INSTITUTE

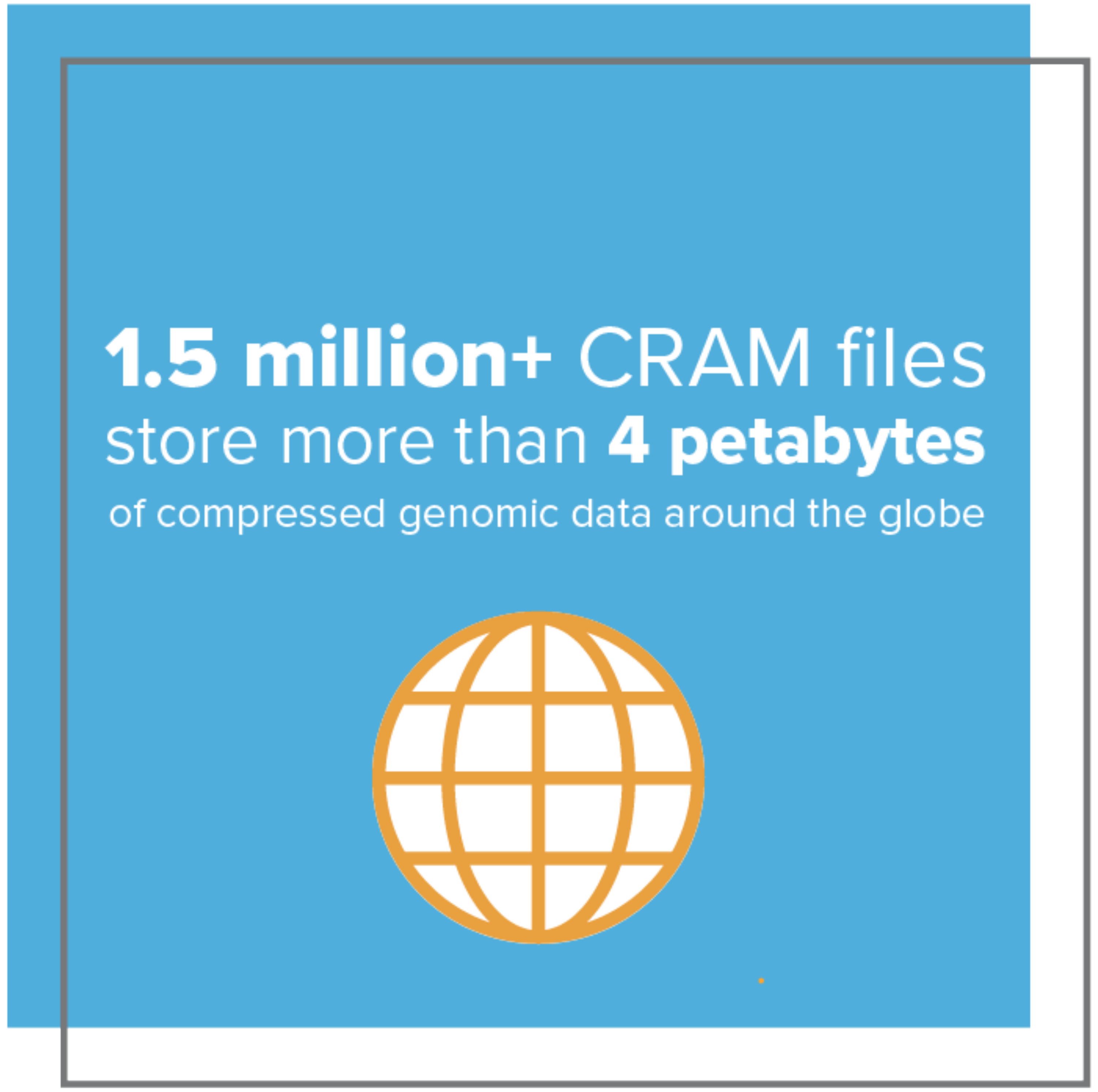
elixir

EMBL-EBI

Genomics england

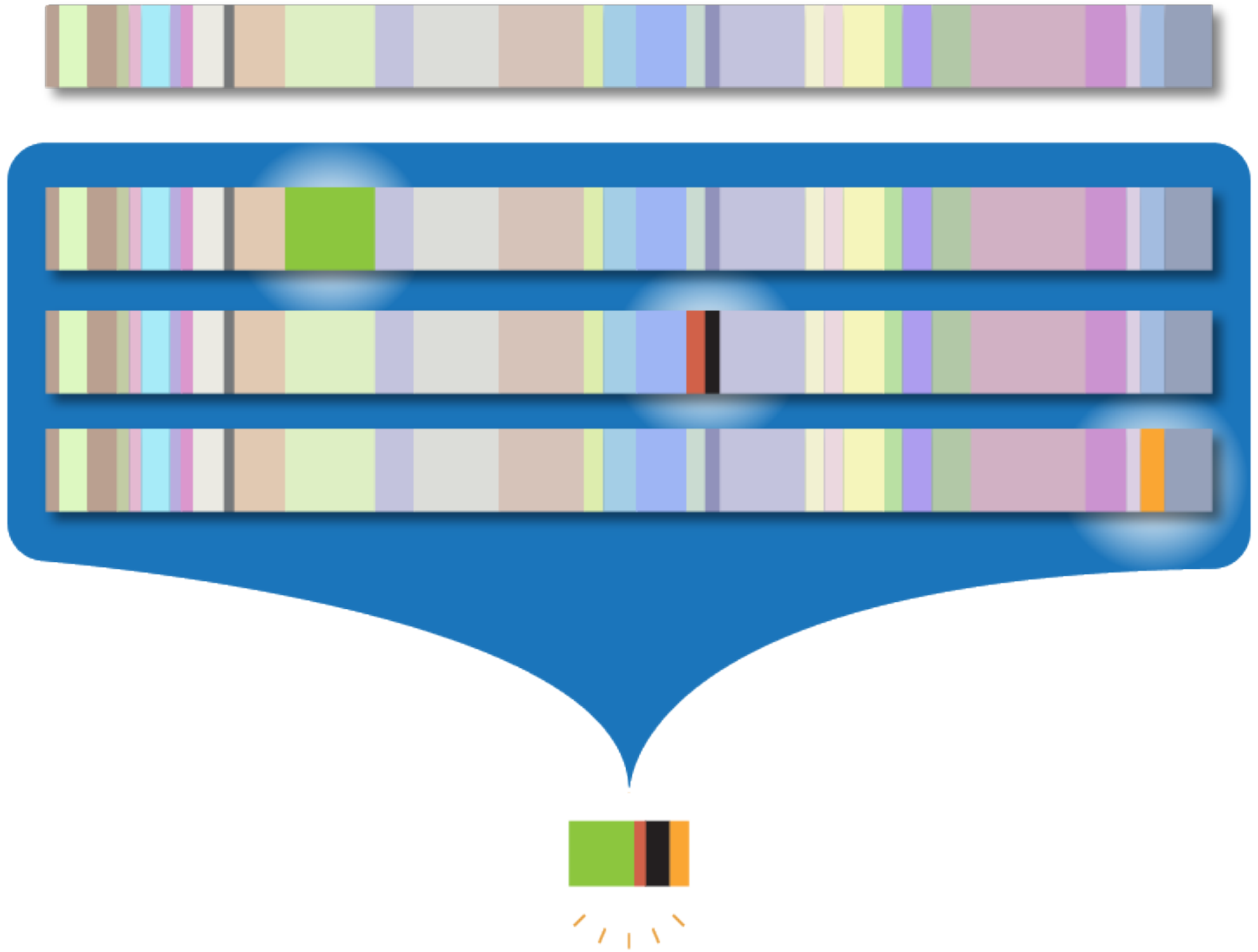
CRAM

CRAM is a file format for storing compressed genomic data. To make files small and efficient, the algorithm compresses information by only storing the parts that are different from the reference human genome.



1.5 million+ CRAM files store more than **4 petabytes** of compressed genomic data around the globe

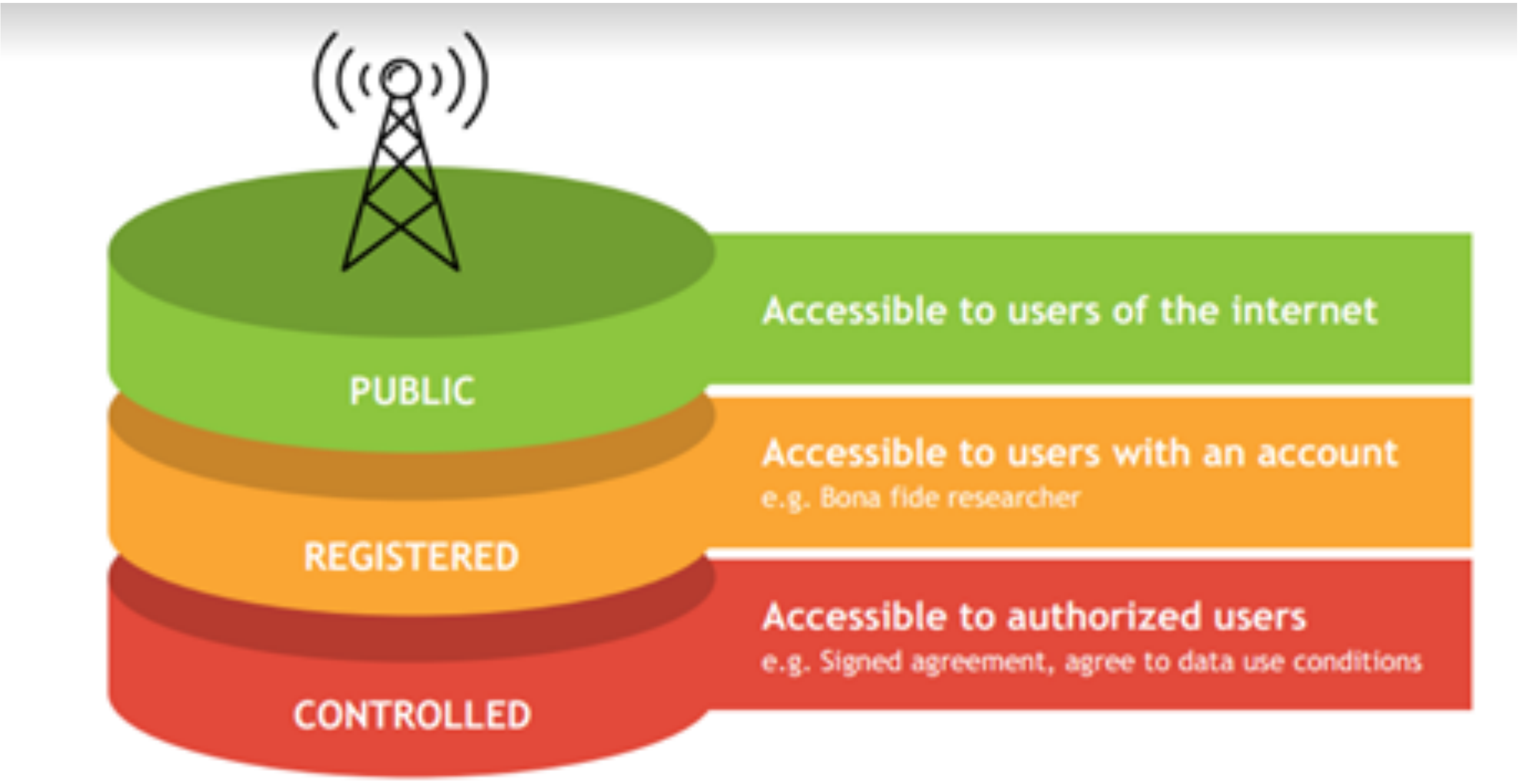
CRAM compresses data by only storing the difference.



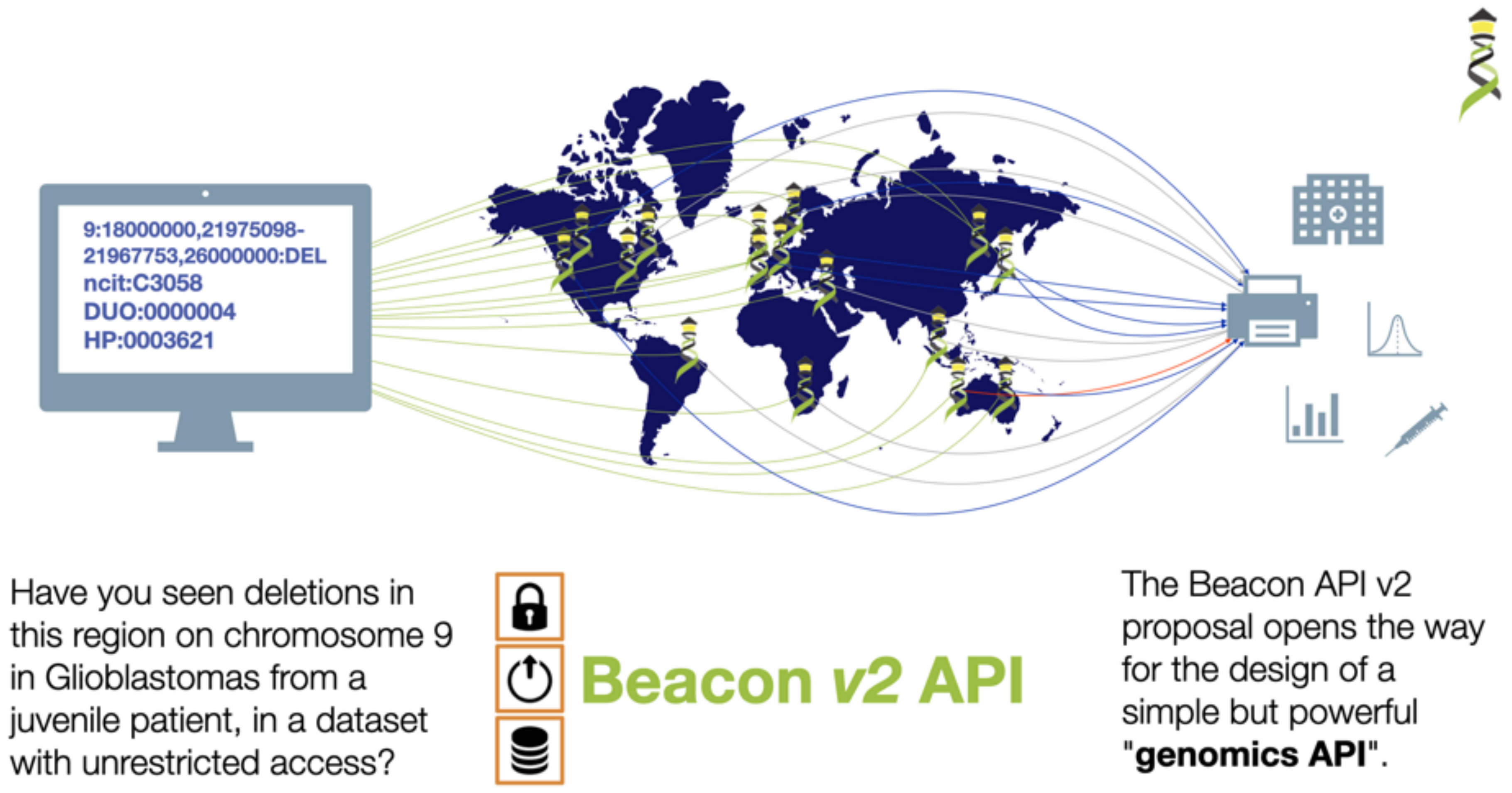
Beacon API v2

The Beacon API can be implemented as a web-accessible service that users may query for information about a specific allele.

Approved: April 21, 2022



PUBLIC
Accessible to users of the internet
REGISTERED
Accessible to users with an account
e.g. Bona fide researcher
CONTROLLED
Accessible to authorized users
e.g. Signed agreement, agree to data use conditions



Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

The Beacon API v2 proposal opens the way for the design of a simple but powerful "**genomics API**".

Example Users

EMBL-EBI

Australian Genomics

SciLifeLab

elixir

University of California Santa Cruz

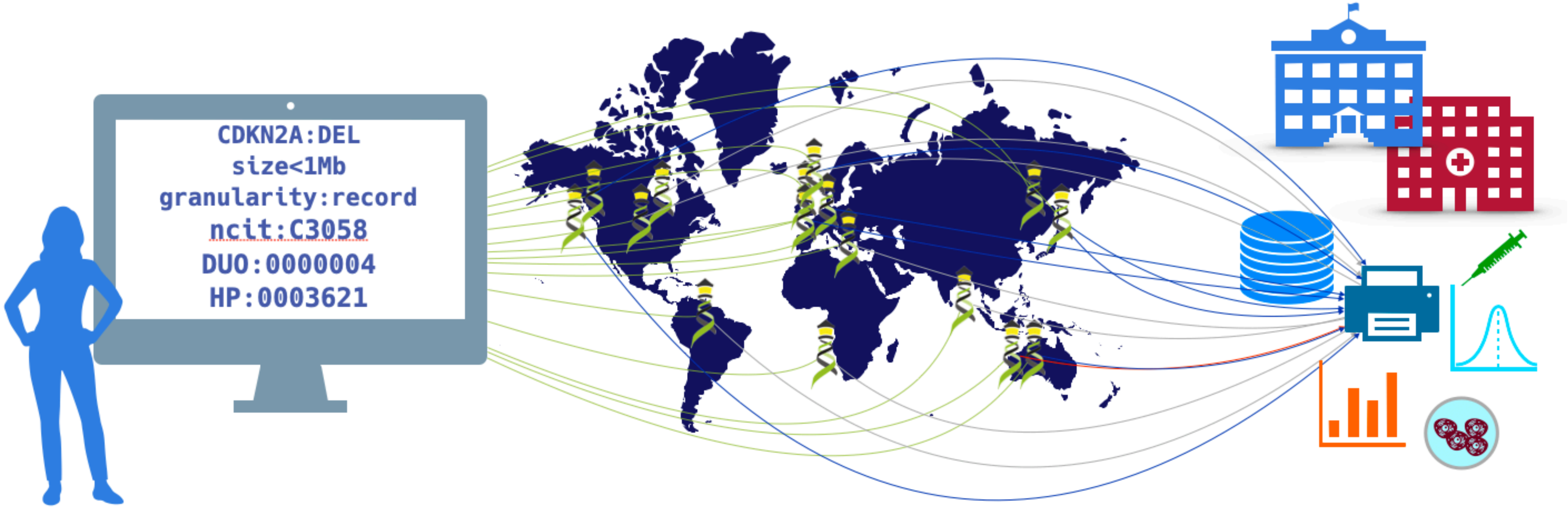
Broad Institute

European Genome-Phenome Archive

International Cancer Genome Consortium

Global Alliance
for Genomics & Health

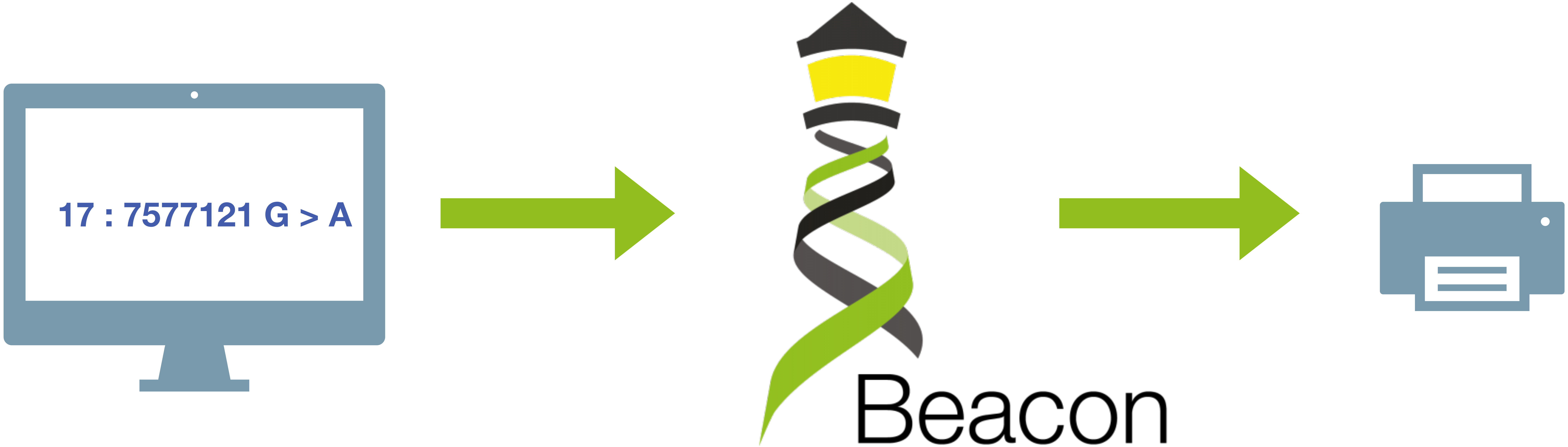
Collaborate. Innovate. Accelerate.



CDKN2A:DEL
size<1Mb
granularity:record
ncit:C3058
DUO:0000004
HP:0003621

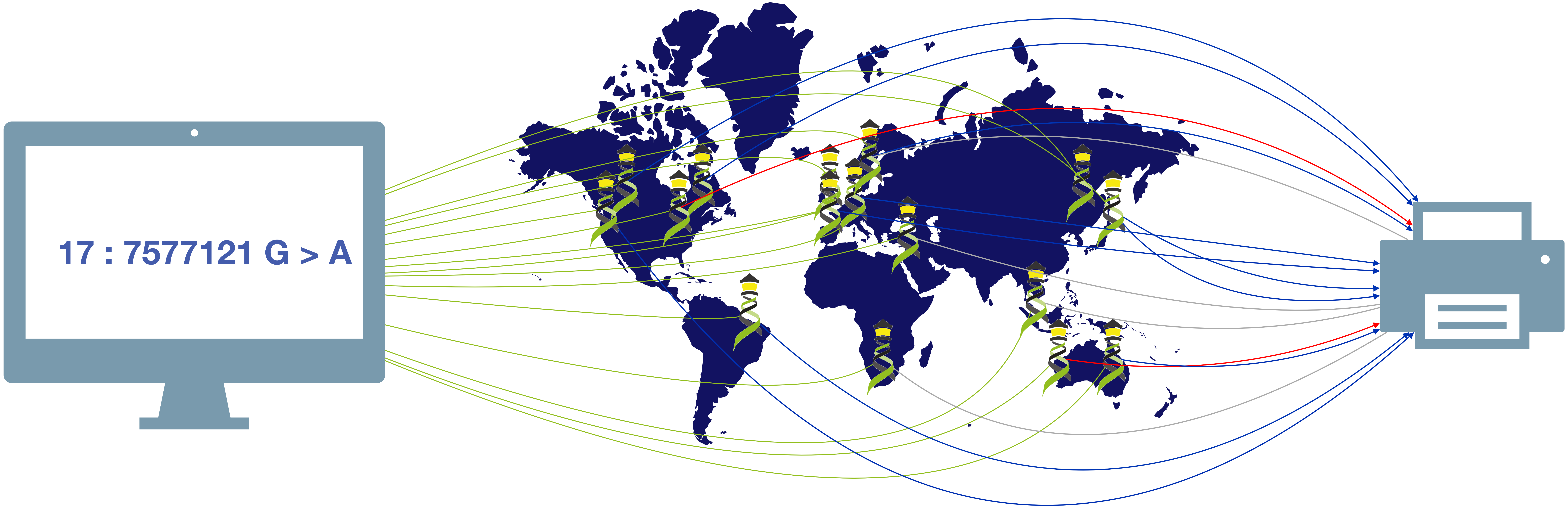
The GA4GH Beacon Protocol

Federating Genomic Discoveries



A ***Beacon*** answers a query for a specific genome variant against individual or aggregate genome collections

YES | NO | \0



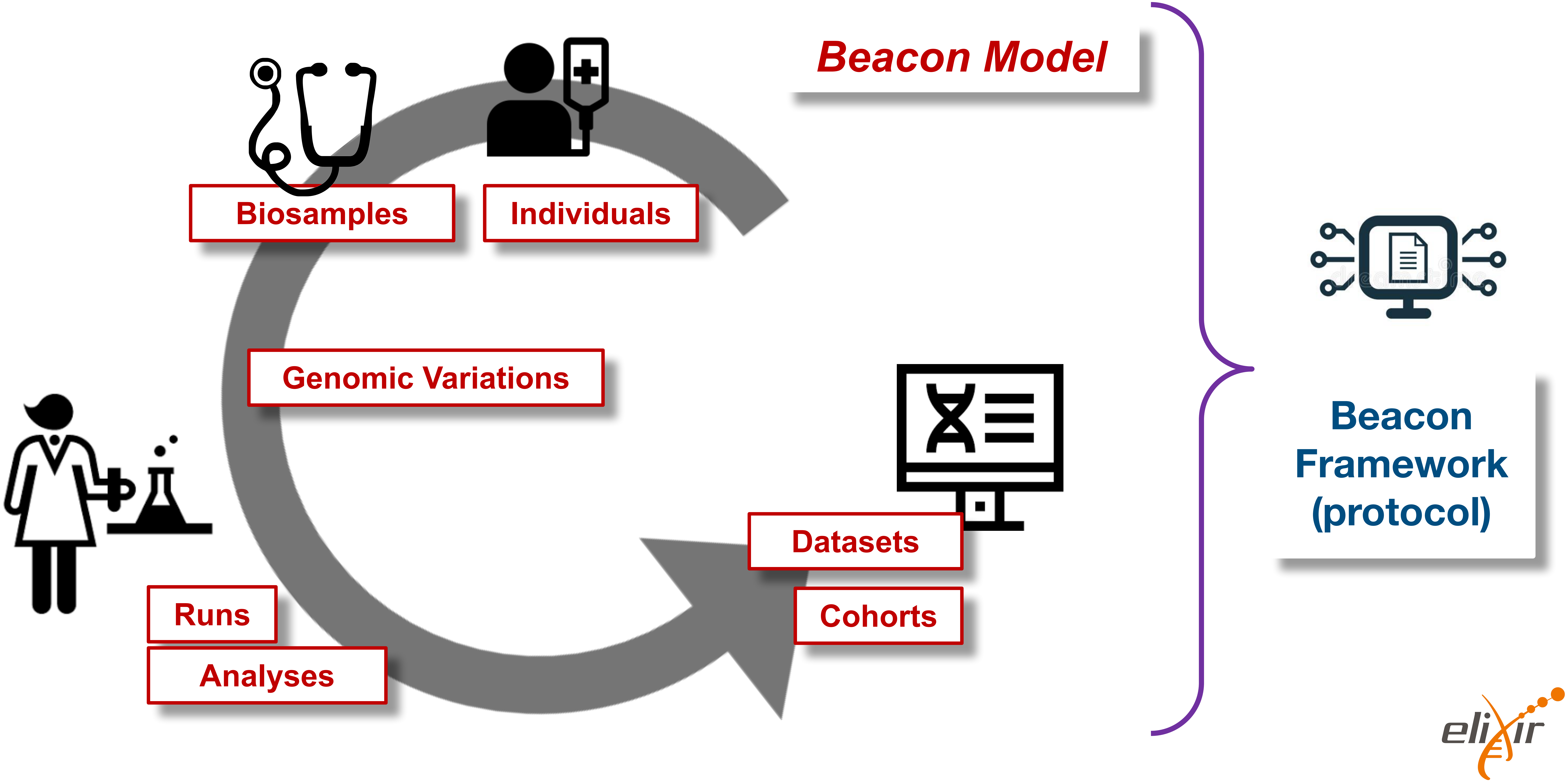
Have you seen this variant? It came up in my patient and we don't know if this is a common SNP or worth following up.

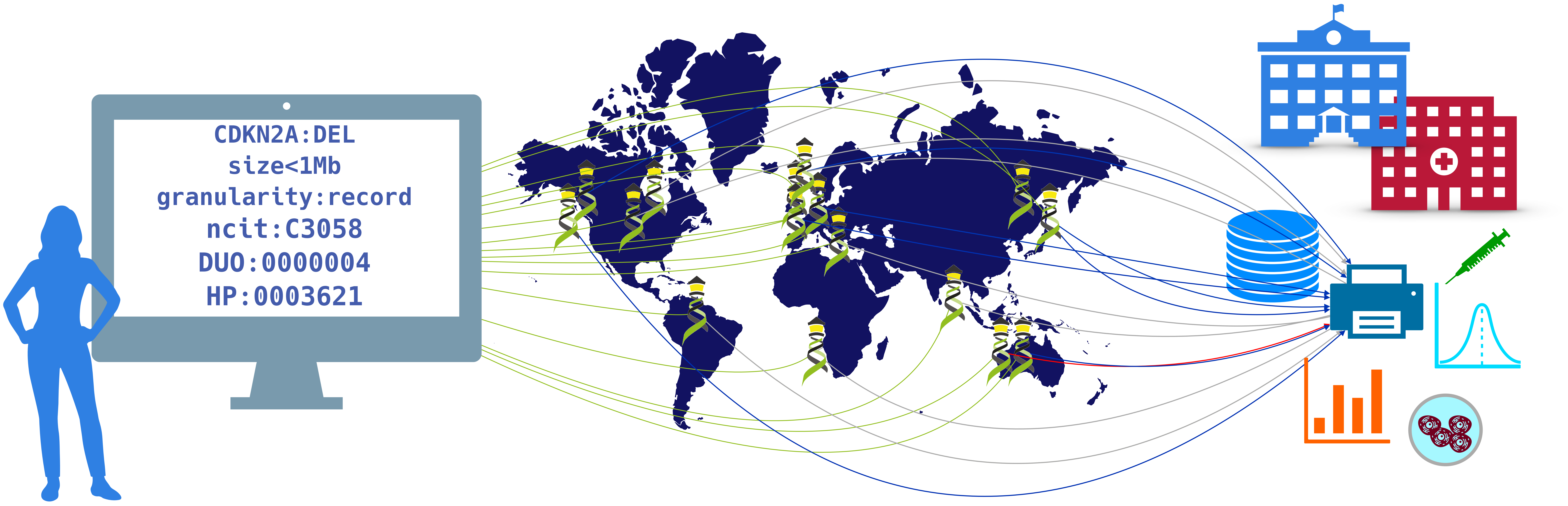
A Beacon network federates *genome variant queries* across databases that support the ***Beacon API***

Here: The variant has been found in **few** resources, and those are from **disease** specific **collections**.

Beacon v2

docs.genomebeacons.org





```
CDKN2A:DEL
size<1Mb
granularity:record
ncit:C3058
DUO:0000004
HP:0003621
```

Can you provide data about focal deletions in CDKN2A in Glioblastomas from juvenile patients with unrestricted access?

Beacon *v2* API

The Beacon API v2 represents a simple but powerful **genomics API** for ***federated*** data discovery and retrieval

Universität Zürich^UZH

Global Alliance for Genomics & Health

Progenetix and GA4GH Beacon

Implementation driven development of a GA4GH standard

progenetix

Progenetix & Beacon

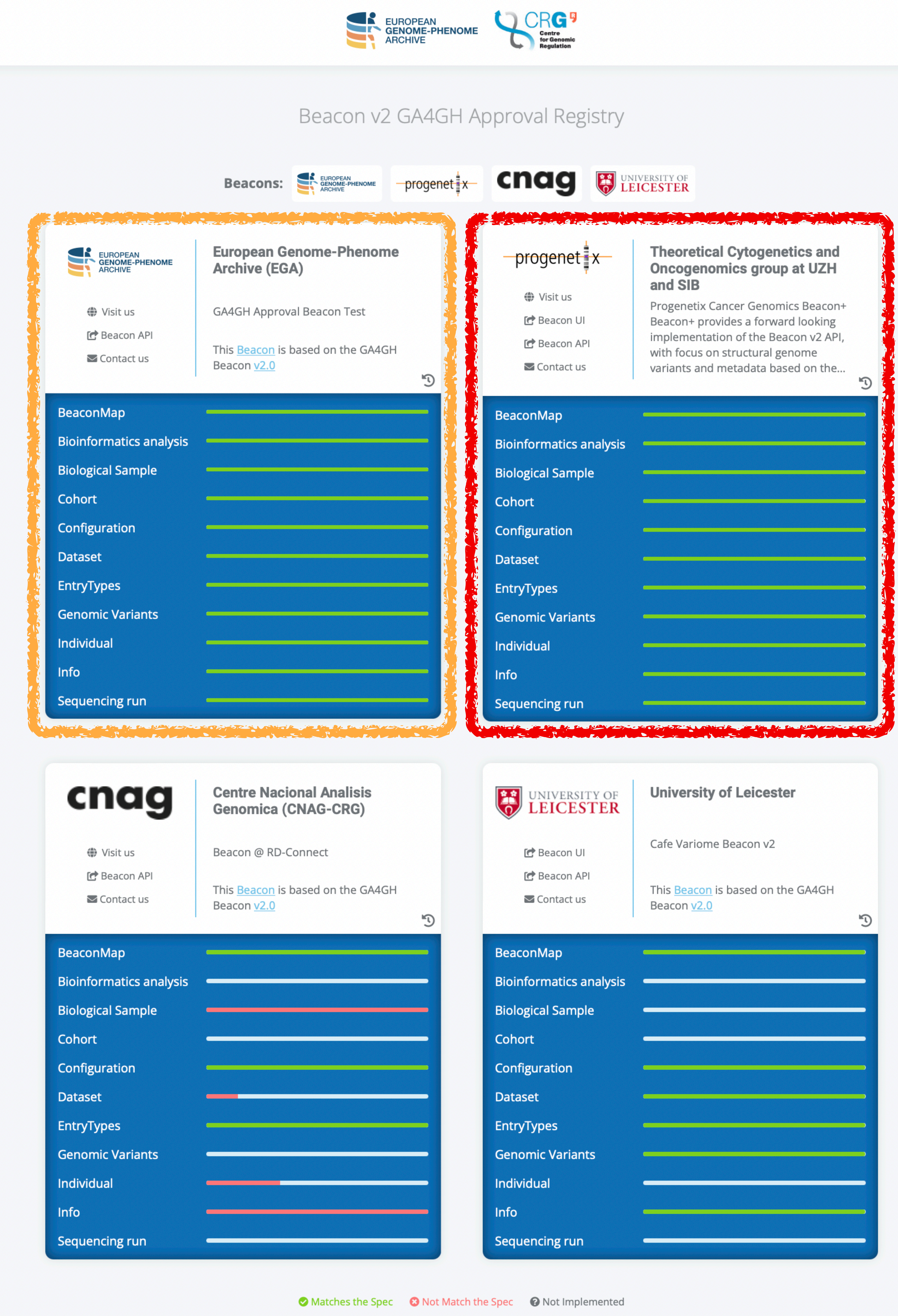
Implementation driven standards development

- Progenetix Beacon+ has served as implementation driver since 2016
- prototyping of advanced Beacon features such as
  - ➡ structural variant queries
  - ➡ data handovers
  - ➡ Phenopackets integration

Global Alliance for Genomics & Health

progenetix

elixir



EUROPEAN GENOME-PHENOME ARCHIVE · CRG Centre for Genomic Regulation

Beacon v2 GA4GH Approval Registry

Beacons: EUROPEAN GENOME-PHENOME ARCHIVE · progenetix · cnag · UNIVERSITY OF LEICESTER

European Genome-Phenome Archive (EGA)

Visit us · Beacon API · Contact us

GA4GH Approval Beacon Test

This Beacon is based on the GA4GH Beacon v2.0

Theoretical Cytogenetics and Oncogenomics group at UZH and SIB

Visit us · Beacon UI · Beacon API · Contact us

Progenetix Cancer Genomics Beacon+ Beacon+ provides a forward looking implementation of the Beacon v2 API, with focus on structural genome variants and metadata based on the...

Centre Nacional Analisis Genomica (CNAG-CRG)

Visit us · Beacon API · Contact us

Beacon @ RD-Connect

This Beacon is based on the GA4GH Beacon v2.0

University of Leicester

Beacon UI · Beacon API · Contact us

Cafe Variome Beacon v2

This Beacon is based on the GA4GH Beacon v2.0

	European Genome-Phenome Archive (EGA)	Theoretical Cytogenetics and Oncogenomics group at UZH and SIB	Centre Nacional Analisis Genomica (CNAG-CRG)	University of Leicester
BeaconMap	🟩	🟩	🟩	🟩
Bioinformatics analysis	🟩	🟩	⬜	⬜
Biological Sample	🟩	🟩	🟥	⬜
Cohort	🟩	🟩	⬜	⬜
Configuration	🟩	🟩	🟩	🟩
Dataset	🟩	🟩	🟥⬜	🟩
EntryTypes	🟩	🟩	🟩	🟩
Genomic Variants	🟩	🟩	⬜	🟩
Individual	🟩	🟩	🟥⬜	⬜
Info	🟩	🟩	🟥	🟩
Sequencing run	🟩	🟩	⬜	⬜

🟩 Matches the Spec 🟥 Not Match the Spec ⬜ Not Implemented

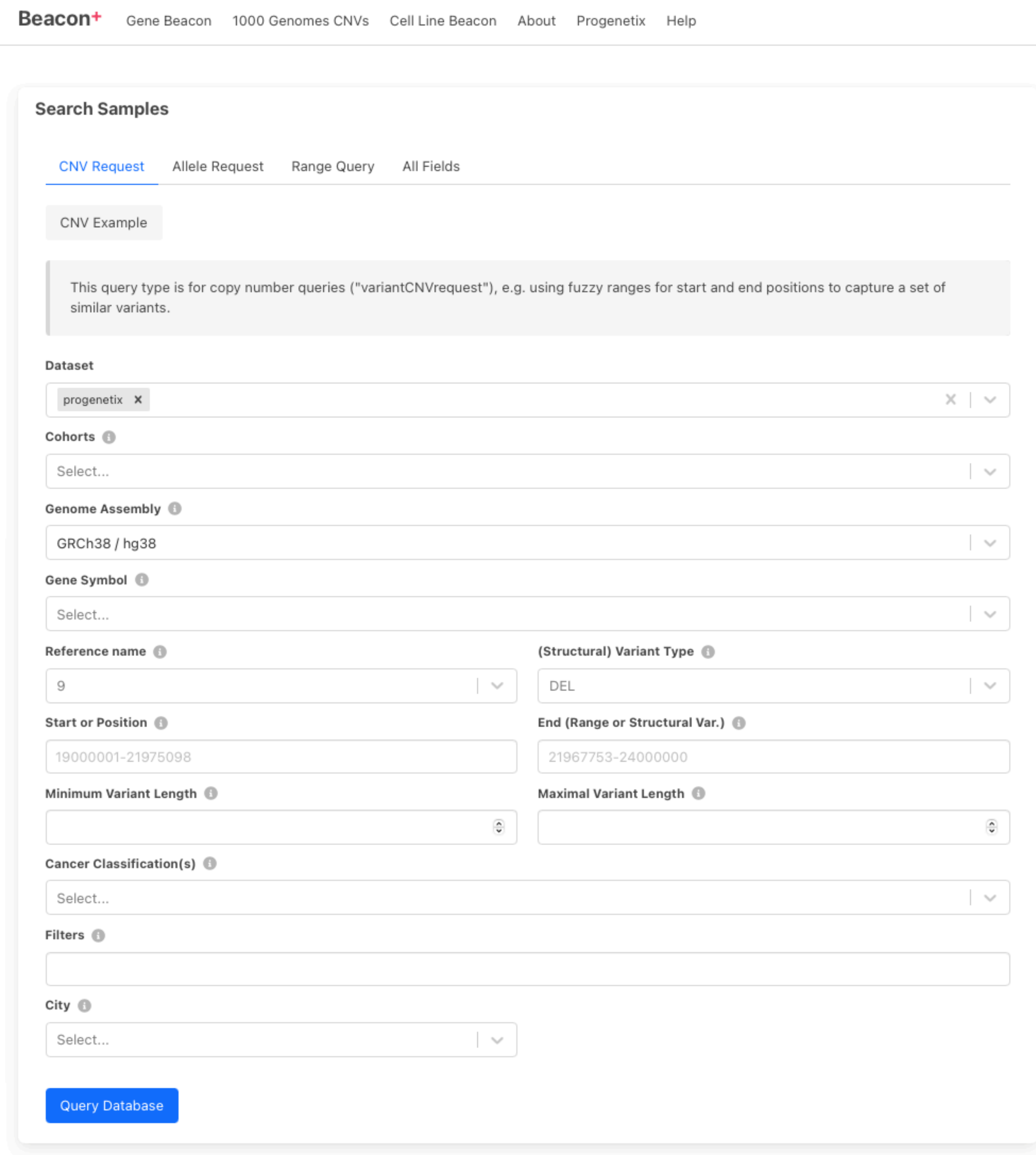
Beacon protocol verifier at time of GA4GH approval Spring 2022

	Beacon v1 Development	Beacon v2 Development	Related ...
2014	GA4GH founding event; Jim Ostell proposes Beacon concept including "more features ... version 2"		
2015	• beacon-network.org aggregator created by DNAstack		• ELIXIR starts Beacon project support
2016	• Beacon v0.3 release • work on queries for structural variants (brackets for fuzzy start and end parameters...)	• Beacon+ concept implemented on progenetix.org • concepts from GA4GH Metadata (ontologies…) • entity-scoped query parameters ("individual.age")	
2017	• OpenAPI implementation • integrating CNV parameters (e.g. "startMin, statMax")	• Beacon+ demos "handover" concept	• GA4GH re-structuring (workstreams...) • Beacon part of Discovery WS
2018	• Beacon v0.4 release in January; feature release for GA4GH approval process • GA4GH Beacon v1 approved at Oct plenary		• new Beacon website (March)
2019	• ELIXIR Beacon Network	• Beacon hackathon Stockholm; settling on "filters" • Barcelona goes Zurich developers meeting • Beacon API v2 Kick off • adopting "handover" concept	• Beacon publication at Nature Biotechnology
2020		• "Scouts" teams working on different aspects - filters, genomic variants, compliance ... • discussions w/ clinical stakeholders	
2021		• framework + models concept implemented • range and bracket queries, variant length parameters • starting of GA4GH review process	• Phenopackets v2 approved
2022		• further changes esp. in default model, aligning with Phenopackets and VRS • unified beacon-v2 code & docs repository • Beacon v2 approved at Apr GA4GH Connect	• *docs.genomebeacons.org*

Beacon+ by Progenetix

From Beacon Query to Explorative Analyses of CNV Patterns

- Since 2016 the Progenetix resource has been used to model options for Beacon development
  - 138334 individual samples from 698 cancer types
- The consistent use of hierarchical diagnostic codes allows the use of Beacon "filters" for histopathological/clinically scoped queries
- Beacon's handover protocols can be utilized for data retrieval and, well, handing over to additional services, e.g.
  - downloads
  - visualization
  - use of external services (UCSC browser display...)



Search Samples

This query type is for copy number queries ("variantCNVrequest"), e.g. using fuzzy ranges for start and end positions to capture a set of similar variants.

Dataset: progenetix

Cohorts: Select...

Genome Assembly: GRCh38 / hg38

Gene Symbol: Select...

Reference name: 9

(Structural) Variant Type: DEL

Start or Position: 19000001-21975098

End (Range or Structural Var.): 21967753-24000000

Minimum Variant Length:

Maximal Variant Length:

Cancer Classification(s): Select...

Filters:

City: Select...

Beacon v2 Filters

Example: Use of hierarchical classification systems (here NCIt neoplasm core)

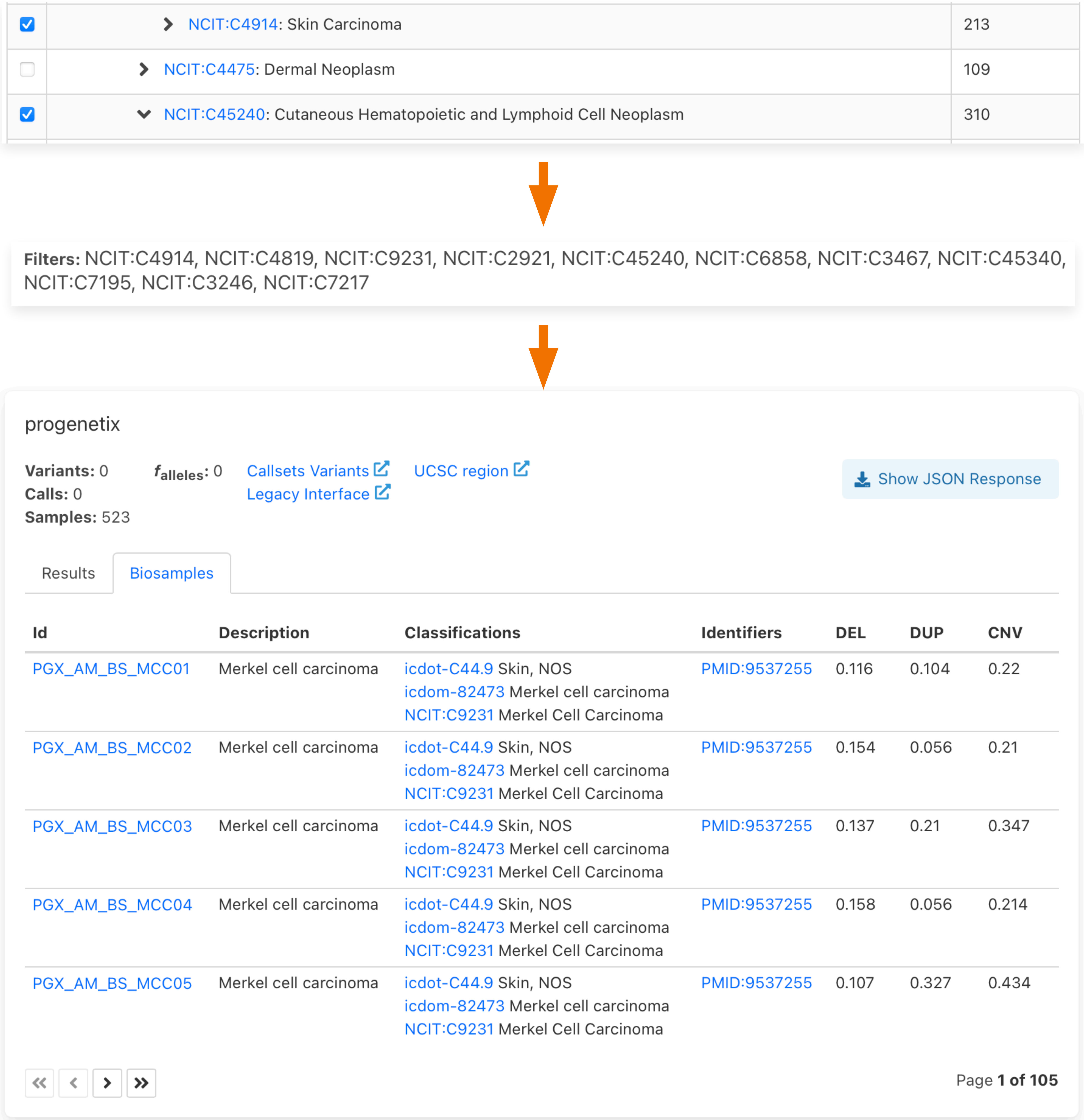
- Beacon v2 "filters" assumes inclusion of child terms when using hierarchical classifications

  ➡ implicit *OR* with otherwise assumed *AND*

- implementation of hierarchical annotations overcomes some limitatiions of "fuzzy" disease annotations

progenetix

Beacon+ specific: Multiple term selection with OR logic



☑	› NCIT:C4914: Skin Carcinoma	213
□	› NCIT:C4475: Dermal Neoplasm	109
☑	⌄ NCIT:C45240: Cutaneous Hematopoietic and Lymphoid Cell Neoplasm	310

Filters: NCIT:C4914, NCIT:C4819, NCIT:C9231, NCIT:C2921, NCIT:C45240, NCIT:C6858, NCIT:C3467, NCIT:C45340, NCIT:C7195, NCIT:C3246, NCIT:C7217

progenetix

Variants: 0 $f_{alleles}$: 0
Calls: 0
Samples: 523

Id	Description	Classifications	Identifiers	DEL	DUP	CNV
PGX_AM_BS_MCC01	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.116	0.104	0.22
PGX_AM_BS_MCC02	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.154	0.056	0.21
PGX_AM_BS_MCC03	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.137	0.21	0.347
PGX_AM_BS_MCC04	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.158	0.056	0.214
PGX_AM_BS_MCC05	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.107	0.327	0.434

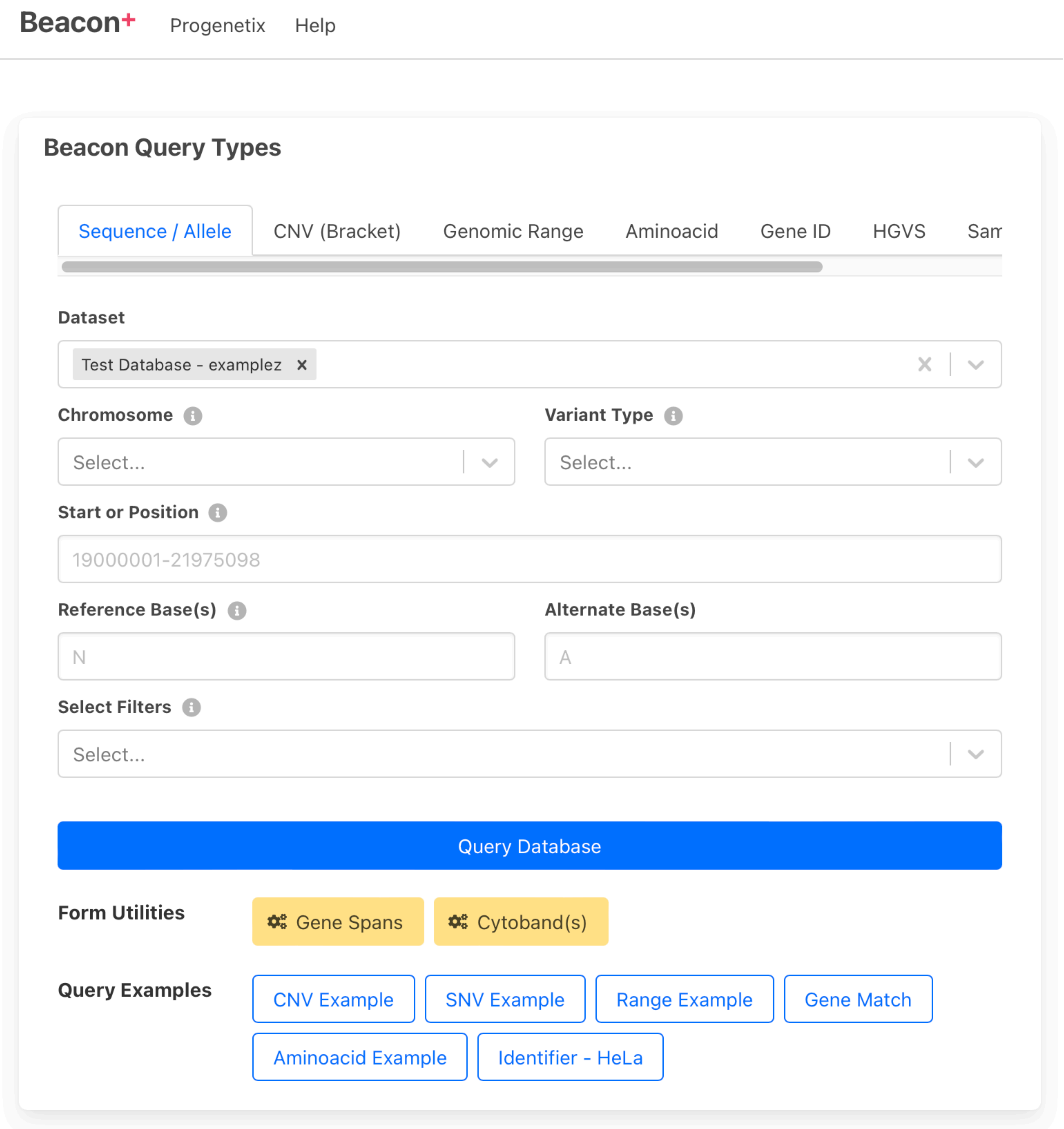
Page **1 of 105**

Beacon Queries

Implementation of Current Options

- (so far) the Beacon model does not define explicit query types
- disambiguation of parameters is left to implementers
- implicit query types:
    - ➡ allele/sequence query
    - ➡ range query, w/ or w/o additional parameters
    - ➡ bracket query (e.g. sized CNVs)
    - ➡ aminoacid, HGVS, gene

beaconplus.progenetix.org



Beacon Query Types

Dataset: Test Database - examplez ×

Chromosome: Select...

Variant Type: Select...

Start or Position: 19000001-21975098

Reference Base(s): N

Alternate Base(s): A

Select Filters: Select...

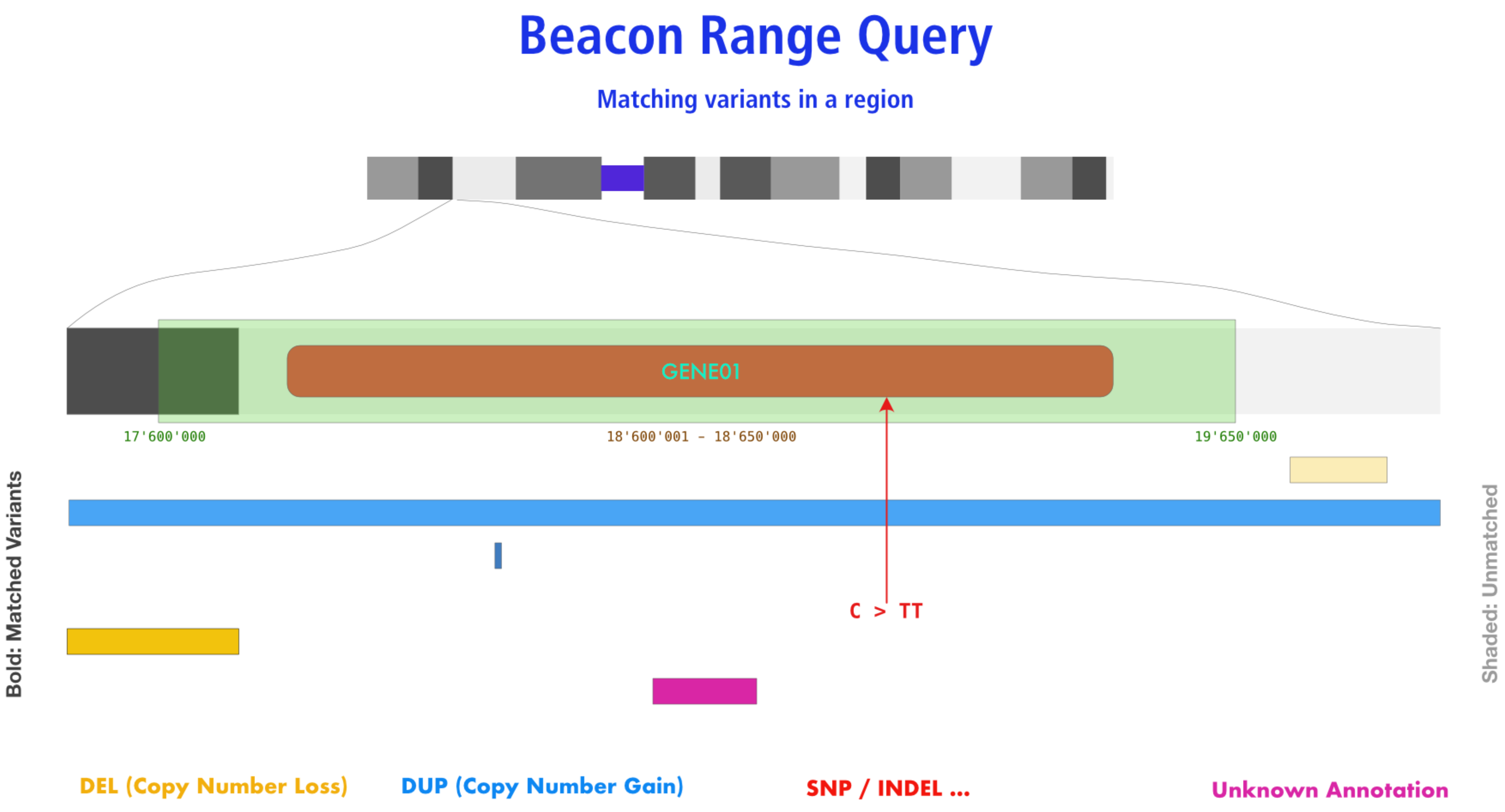
Form Utilities

Query Examples

Beacon Queries

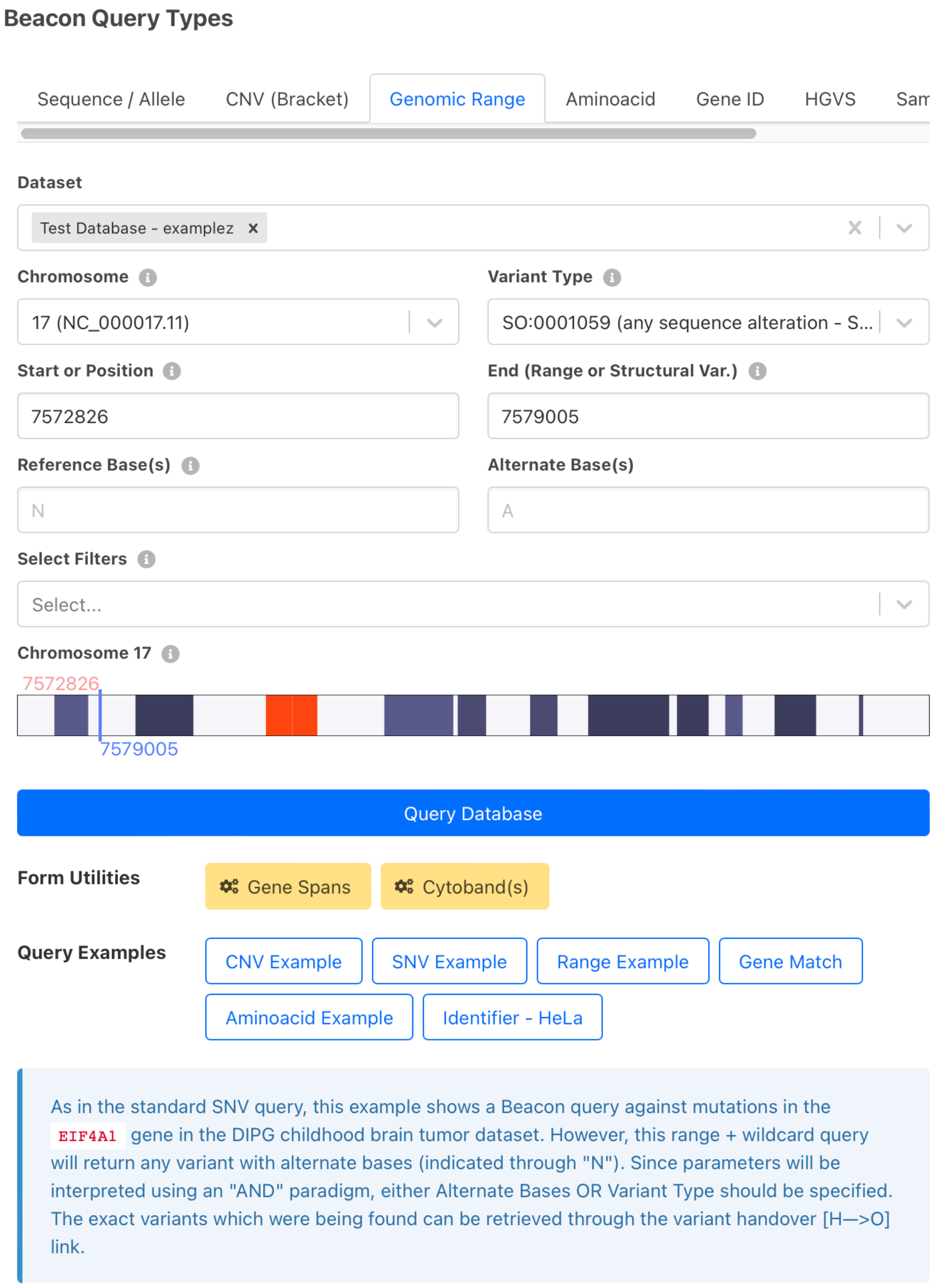
Range ("anything goes") Request

- defined through the use of 1 start, 1 end
- any variant... but can be limited by type etc.



Beacon Range Query

Matching variants in a region



Beacon Query Types

Dataset: Test Database - examplez

Chromosome: 17 (NC_000017.11)

Variant Type: SO:0001059 (any sequence alteration - S...

Start or Position: 7572826

End (Range or Structural Var.): 7579005

Reference Base(s): N

Alternate Base(s): A

Select Filters: Select...

Chromosome 17

Form Utilities: Gene Spans, Cytoband(s)

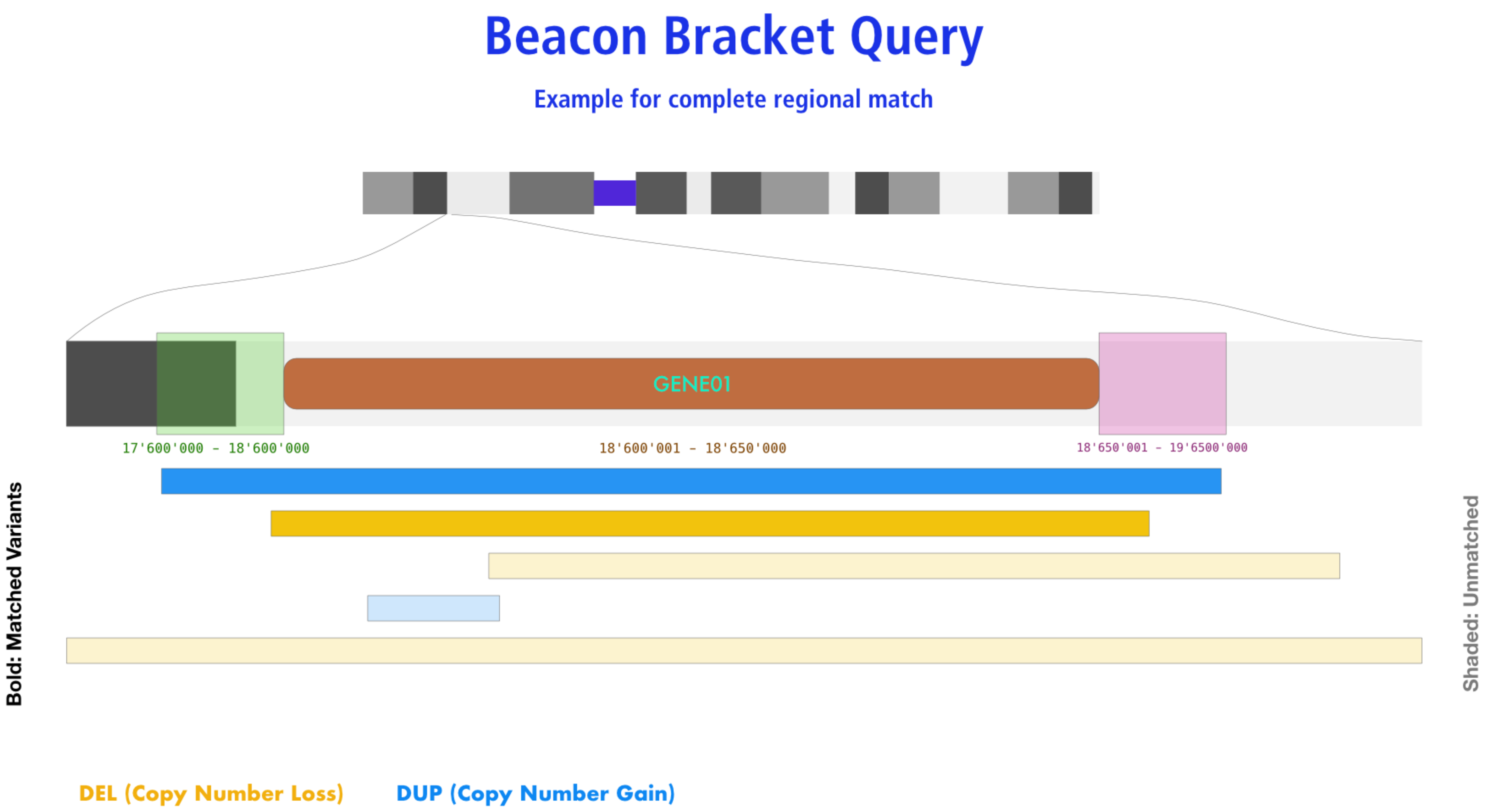
Query Examples: CNV Example, SNV Example, Range Example, Gene Match, Aminoacid Example, Identifier - HeLa

As in the standard SNV query, this example shows a Beacon query against mutations in the `EIF4A1` gene in the DIPG childhood brain tumor dataset. However, this range + wildcard query will return any variant with alternate bases (indicated through "N"). Since parameters will be interpreted using an "AND" paradigm, either Alternate Bases OR Variant Type should be specified. The exact variants which were being found can be retrieved through the variant handover [H—>O] link.

Beacon Queries

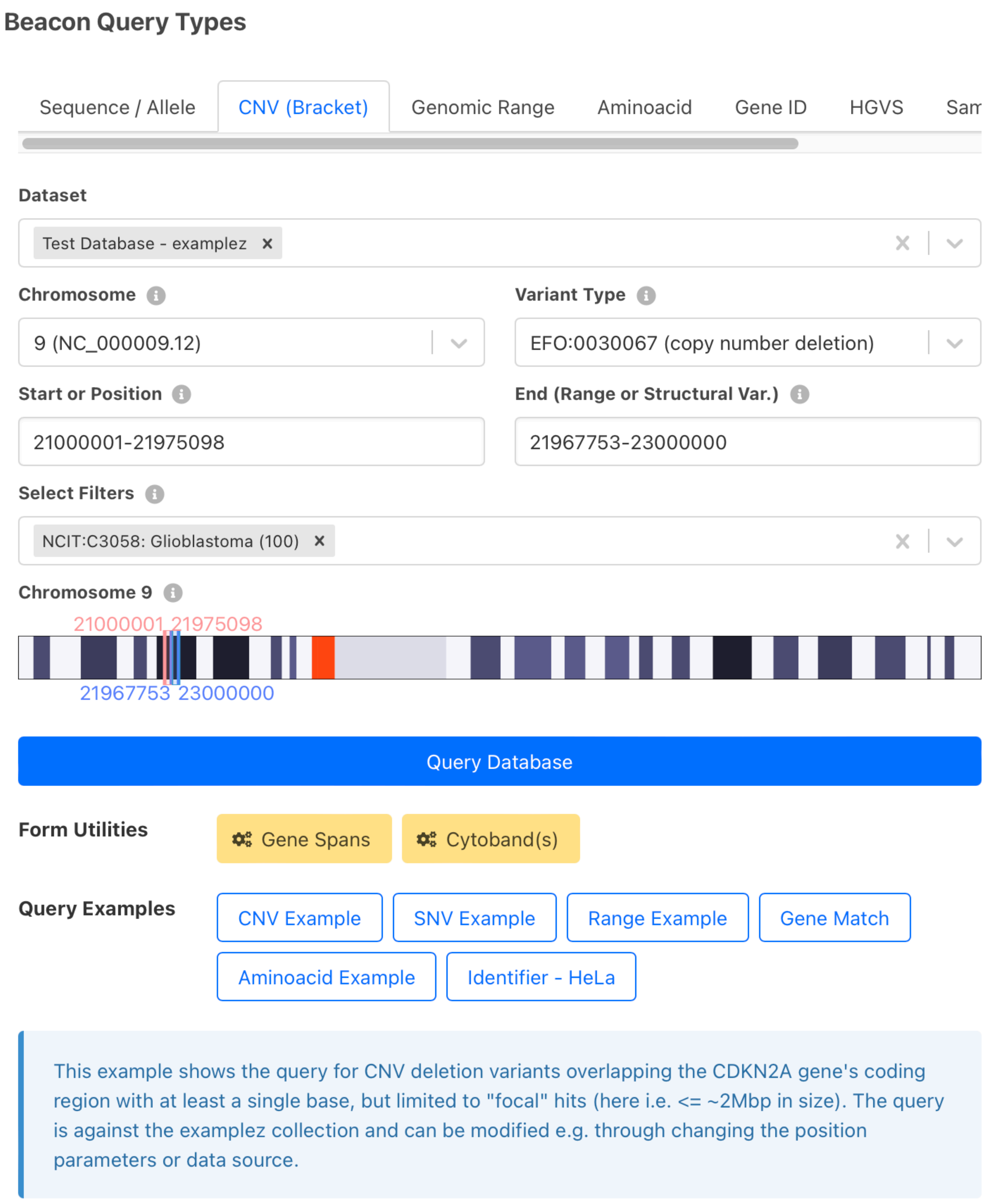
Bracket ("CNV") Query

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



Beacon Bracket Query

Example for complete regional match



Beacon Query Types

Dataset: Test Database - examplez

Chromosome: 9 (NC_000009.12)

Variant Type: EFO:0030067 (copy number deletion)

Start or Position: 21000001-21975098

End (Range or Structural Var.): 21967753-23000000

Select Filters: NCIT:C3058: Glioblastoma (100)

Chromosome 9

This example shows the query for CNV deletion variants overlapping the CDKN2A gene's coding region with at least a single base, but limited to "focal" hits (here i.e. <= ~2Mbp in size). The query is against the examplez collection and can be modified e.g. through changing the position parameters or data source.

Progenetix Stack

progenetix

- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
  - biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads…
- the complete middleware / CGI stack is provided through the *bycon* package
  - schemas, query stack, data transformation (e.g. Phenopackets generation)…
- data collections mostly correspond to the main Beacon default model entities
  - no separate *runs* collection; integrated w/ analyses
  - *variants* are stored per observation instance

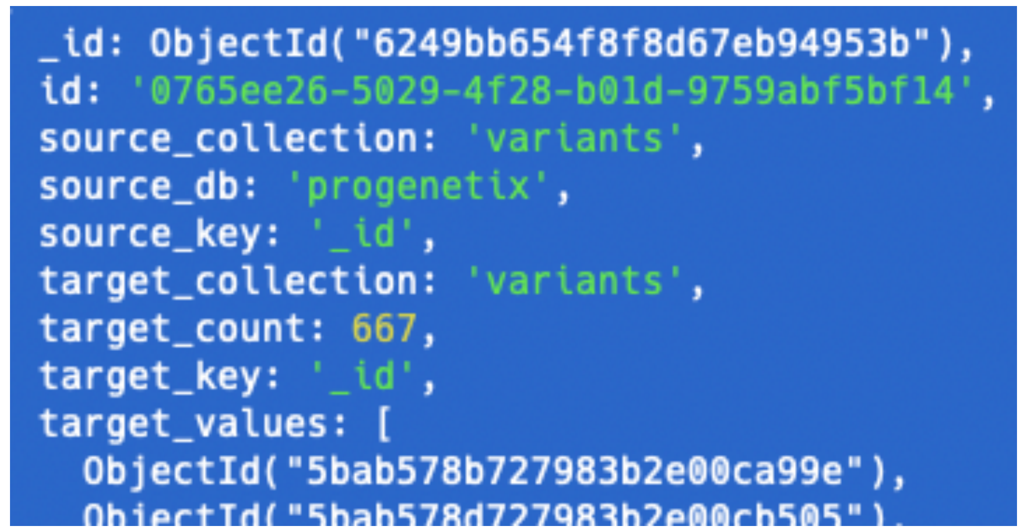
React

APACHE HTTP SERVER PROJECT

python™

mongoDB

- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
  - PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703…
- *querybuffer* stores id values of all entities matched by a query and provides the corresponding access handle for **handover** generation



```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
  ObjectId("5bab578d727983b2e00cb505"),
```

variants

analyses

biosamples

individuals

Entity collections

collations

geolocs

genespans

publications

qBuffer

Utility collections

Beacon v2 Conformity and Extensions in Progenetix

Putting the + into Beacon ...

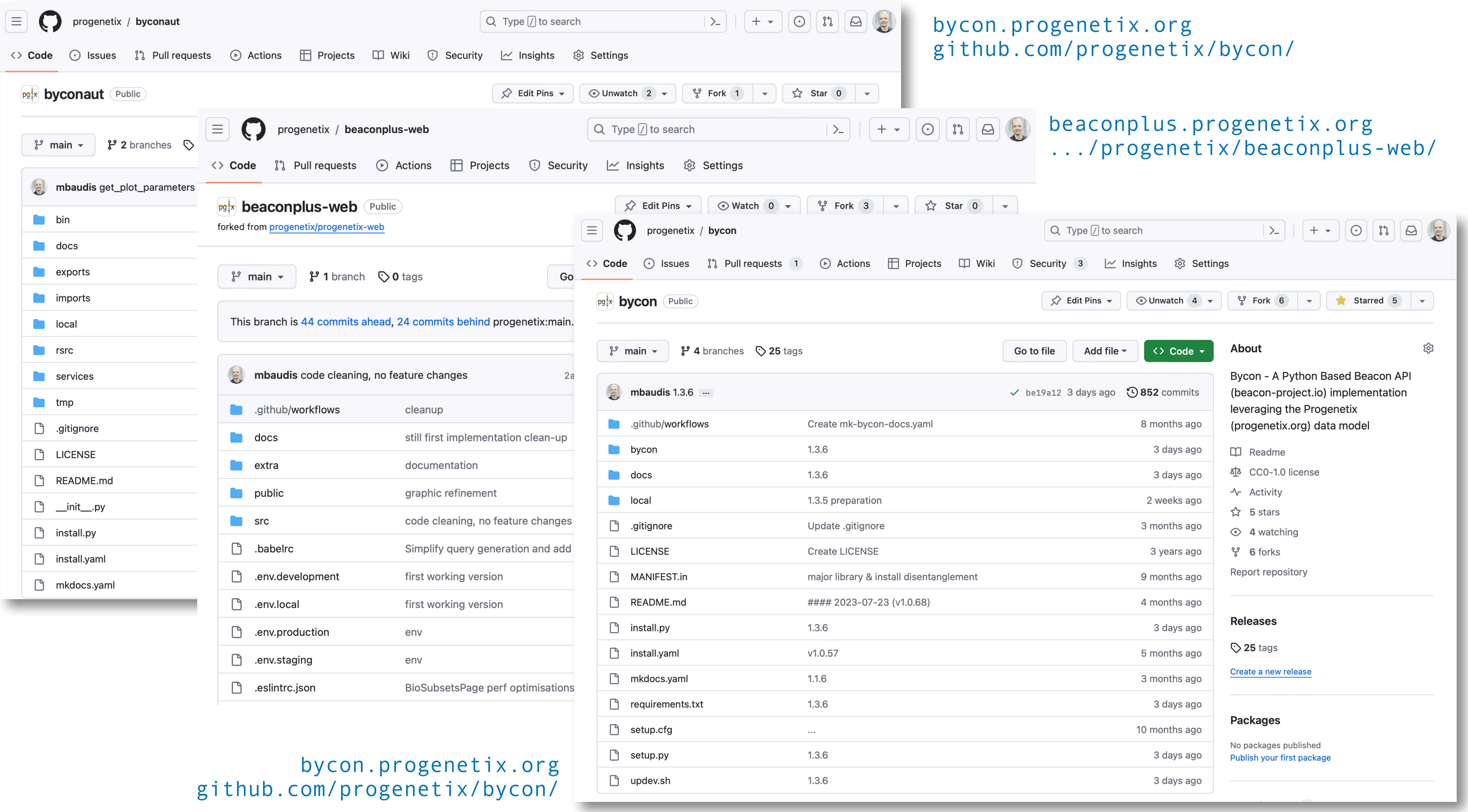
- support & use of standard Beacon v2 PUT & GET variant queries, filters and meta parameters
  - ➡ variant parameters, geneId, lengths, EFO & VCF CNV types, pagination
  - ➡ widespread, self-scoping filter use for bio-, technical- and and id parameters with switch for descending terms use (globally or per term if using POST)
- extensive use of handovers
  - ➡ asynchronous delivery of e.g. variant and sample data, data plots
- + optional use of OR logic for filter combinations (global)
- + extension of query parameters
  - ➡ geographic queries incl. $geonear and use of GeoJSON in schemas
- ╮(╯▽╰)╭ no implementation of authentication on this open dataset

Progenetix provides a number of additional services and output formats which are initiated over the / services path or provided as request parameters and are not considered Beacon extensions (though they follow the syntax where possible).

progenetix

bycon.progenetix.org
github.com/progenetix/bycon/

beaconplus.progenetix.org
.../progenetix/beaconplus-web/



progenetix / byconaut

Code · Issues · Pull requests · Actions · Projects · Wiki · Security · Insights · Settings

byconaut Public

Edit Pins · Unwatch 2 · Fork 1 · Star 0

main · 2 branches

mbaudis get_plot_parameters

bin
docs
exports
imports
local
rsrc
services
tmp
.gitignore
LICENSE
README.md
__init__.py
install.py
install.yaml
mkdocs.yaml

progenetix / beaconplus-web

Code · Pull requests · Actions · Projects · Security · Insights · Settings

beaconplus-web Public

forked from progenetix/progenetix-web

Edit Pins · Watch 0 · Fork 3 · Star 0

main · 1 branch · 0 tags · Go

This branch is 44 commits ahead, 24 commits behind progenetix:main.

mbaudis code cleaning, no feature changes 2a

.github/workflows	cleanup
docs	still first implementation clean-up
extra	documentation
public	graphic refinement
src	code cleaning, no feature changes
.babelrc	Simplify query generation and add
.env.development	first working version
.env.local	first working version
.env.production	env
.env.staging	env
.eslintrc.json	BioSubsetsPage perf optimisations

progenetix / bycon

Code · Issues · Pull requests 1 · Actions · Projects · Wiki · Security 3 · Insights · Settings

bycon Public

Edit Pins · Unwatch 4 · Fork 6 · Starred 5

main · 4 branches · 25 tags · Go to file · Add file · Code

mbaudis 1.3.6 ... ✓ be19a12 3 days ago 852 commits

.github/workflows	Create mk-bycon-docs.yaml	8 months ago
bycon	1.3.6	3 days ago
docs	1.3.6	3 days ago
local	1.3.5 preparation	2 weeks ago
.gitignore	Update .gitignore	3 months ago
LICENSE	Create LICENSE	3 years ago
MANIFEST.in	major library & install disentanglement	9 months ago
README.md	#### 2023-07-23 (v1.0.68)	4 months ago
install.py	1.3.6	3 days ago
install.yaml	v1.0.57	5 months ago
mkdocs.yaml	1.1.6	3 months ago
requirements.txt	1.3.6	3 days ago
setup.cfg	...	10 months ago
setup.py	1.3.6	3 days ago
updev.sh	1.3.6	3 days ago

About

Bycon - A Python Based Beacon API (beacon-project.io) implementation leveraging the Progenetix (progenetix.org) data model

Readme
CC0-1.0 license
Activity
5 stars
4 watching
6 forks
Report repository

Releases

25 tags
Create a new release

Packages

No packages published
Publish your first package

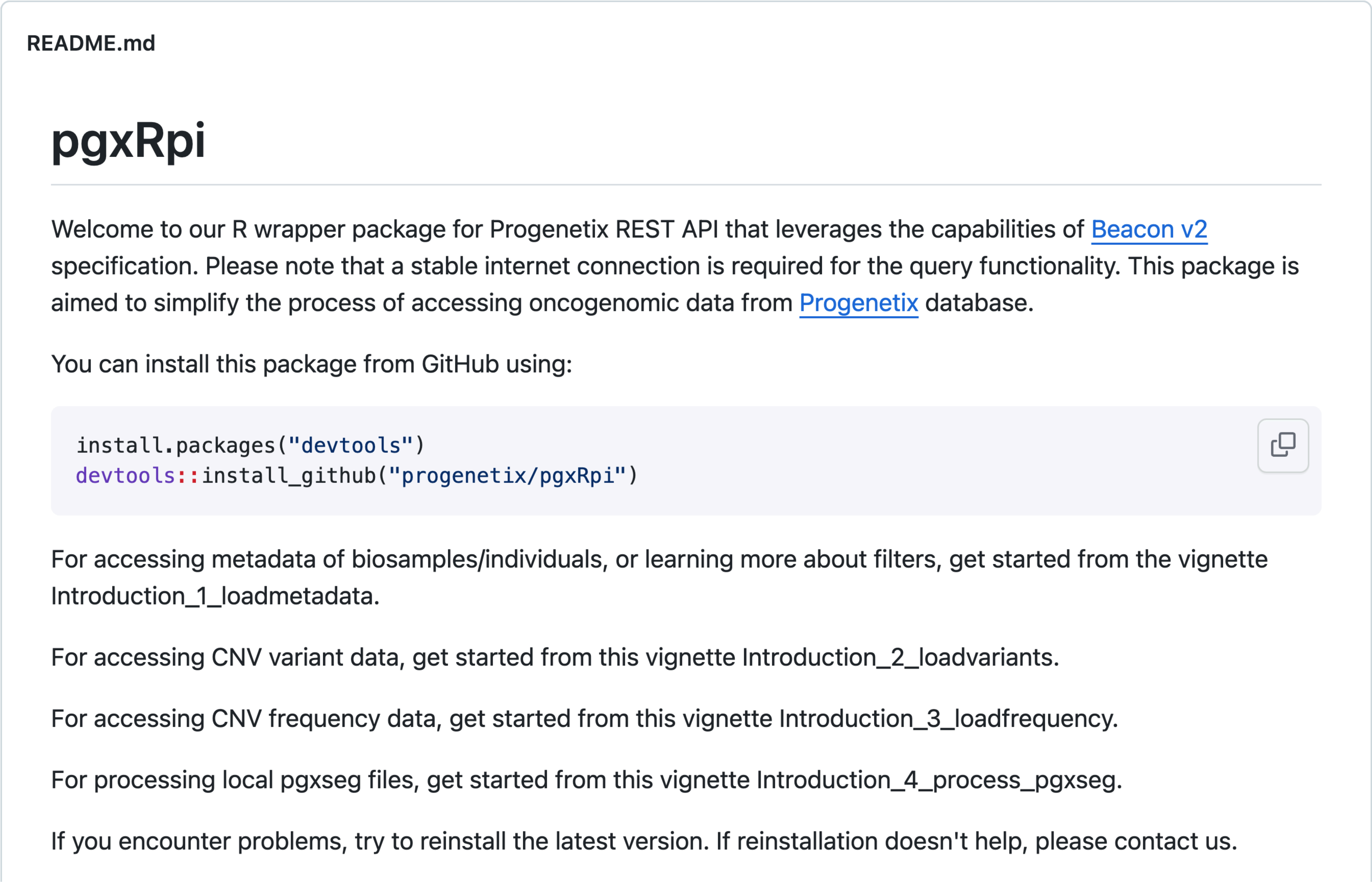
bycon.progenetix.org
github.com/progenetix/bycon/

pgxRpi

An interface API for analyzing Progenetix CNV data in R using the Beacon+ API

GitHub: https://github.com/progenetix/pgxRpi

Bioconductor



README.md

pgxRpi

Welcome to our R wrapper package for Progenetix REST API that leverages the capabilities of Beacon v2 specification. Please note that a stable internet connection is required for the query functionality. This package is aimed to simplify the process of accessing oncogenomic data from Progenetix database.

You can install this package from GitHub using:

```
install.packages("devtools")
devtools::install_github("progenetix/pgxRpi")
```

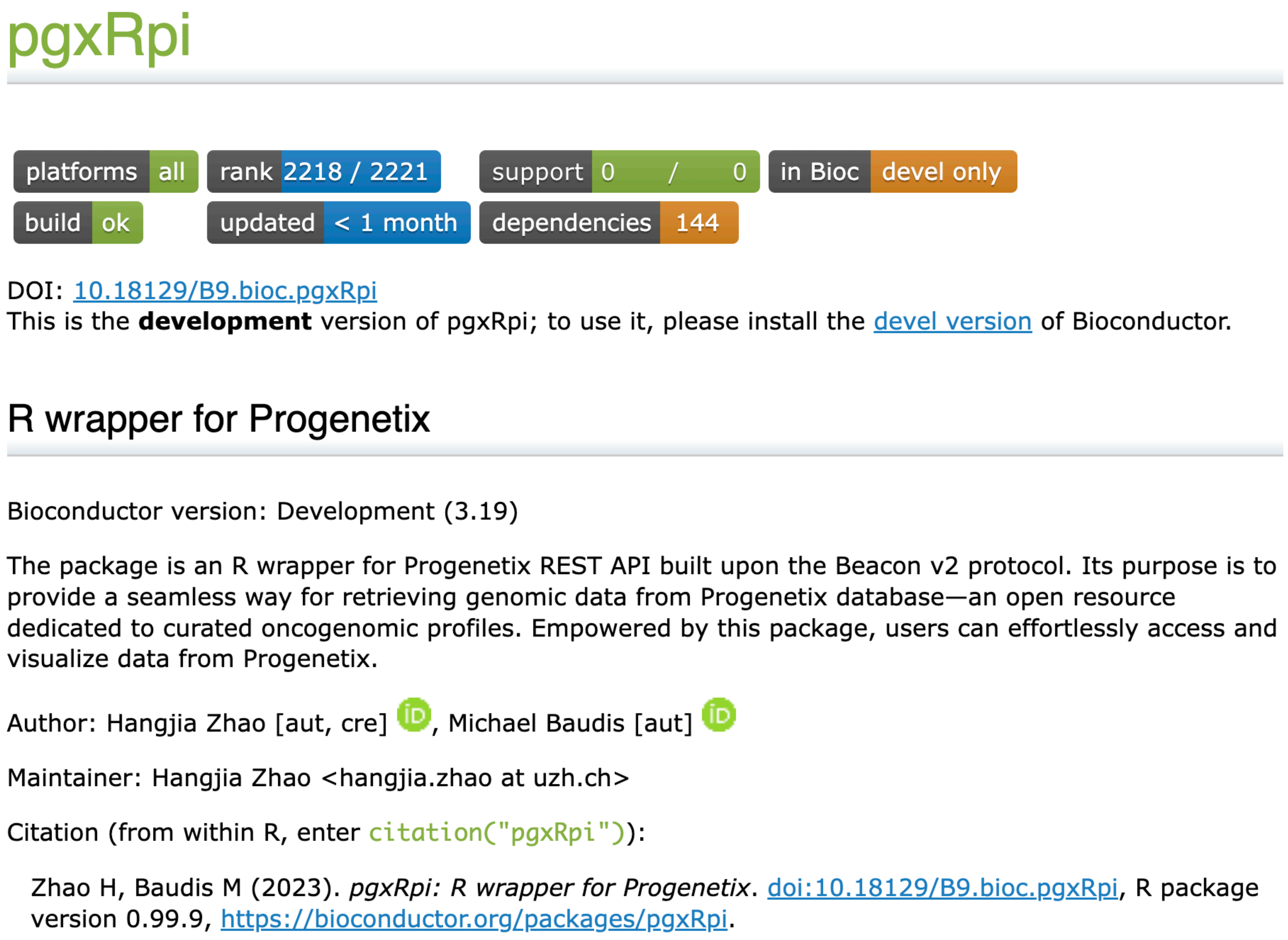
For accessing metadata of biosamples/individuals, or learning more about filters, get started from the vignette Introduction_1_loadmetadata.

For accessing CNV variant data, get started from this vignette Introduction_2_loadvariants.

For accessing CNV frequency data, get started from this vignette Introduction_3_loadfrequency.

For processing local pgxseg files, get started from this vignette Introduction_4_process_pgxseg.

If you encounter problems, try to reinstall the latest version. If reinstallation doesn't help, please contact us.



pgxRpi

platforms all | rank 2218 / 2221 | support 0 / 0 | in Bioc devel only
build ok | updated < 1 month | dependencies 144

DOI: 10.18129/B9.bioc.pgxRpi
This is the **development** version of pgxRpi; to use it, please install the devel version of Bioconductor.

R wrapper for Progenetix

Bioconductor version: Development (3.19)

The package is an R wrapper for Progenetix REST API built upon the Beacon v2 protocol. Its purpose is to provide a seamless way for retrieving genomic data from Progenetix database—an open resource dedicated to curated oncogenomic profiles. Empowered by this package, users can effortlessly access and visualize data from Progenetix.

Author: Hangjia Zhao [aut, cre], Michael Baudis [aut]

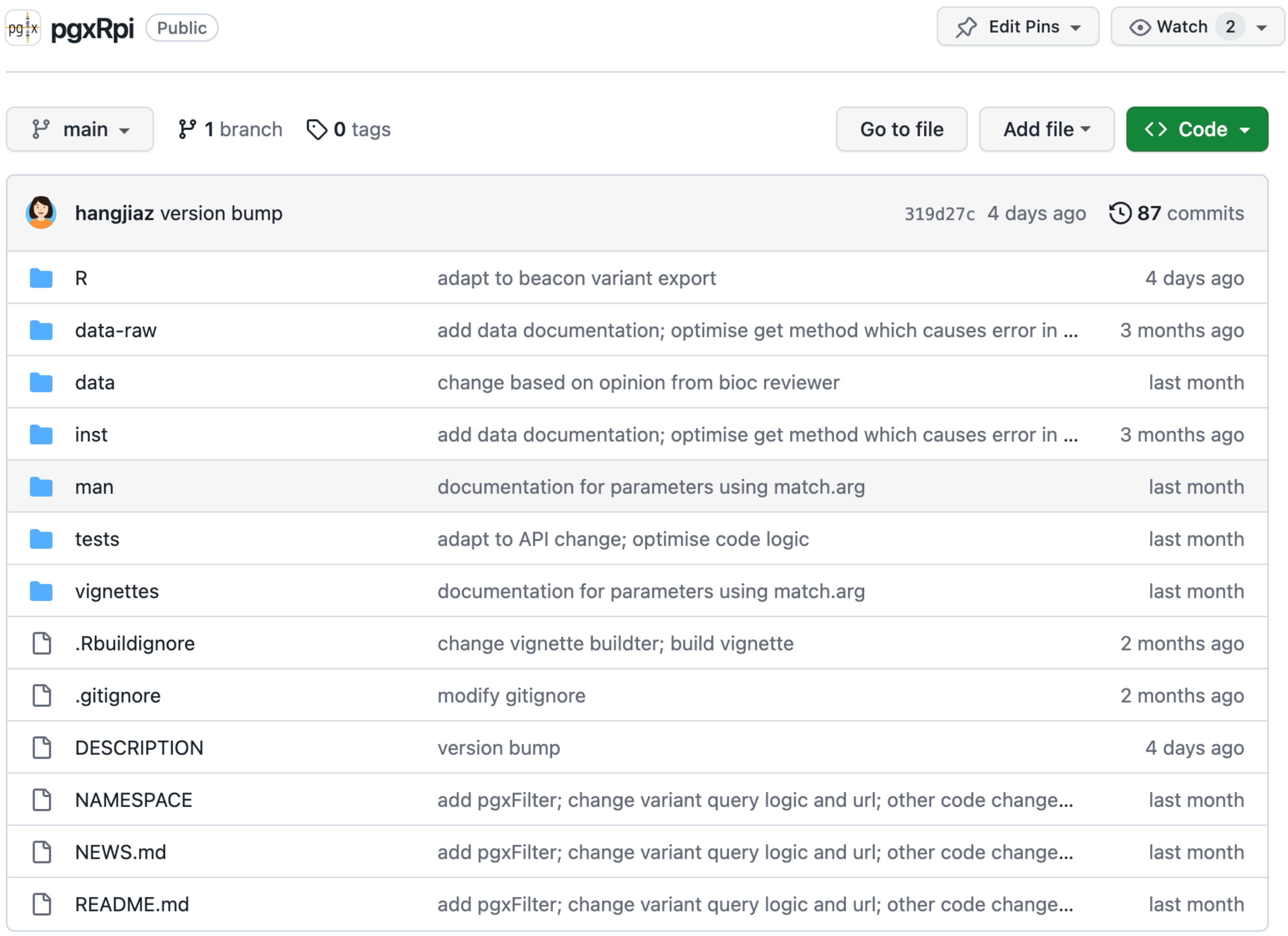
Maintainer: Hangjia Zhao <hangjia.zhao at uzh.ch>

Citation (from within R, enter `citation("pgxRpi")`):

Zhao H, Baudis M (2023). *pgxRpi: R wrapper for Progenetix*. doi:10.18129/B9.bioc.pgxRpi, R package version 0.99.9, https://bioconductor.org/packages/pgxRpi.

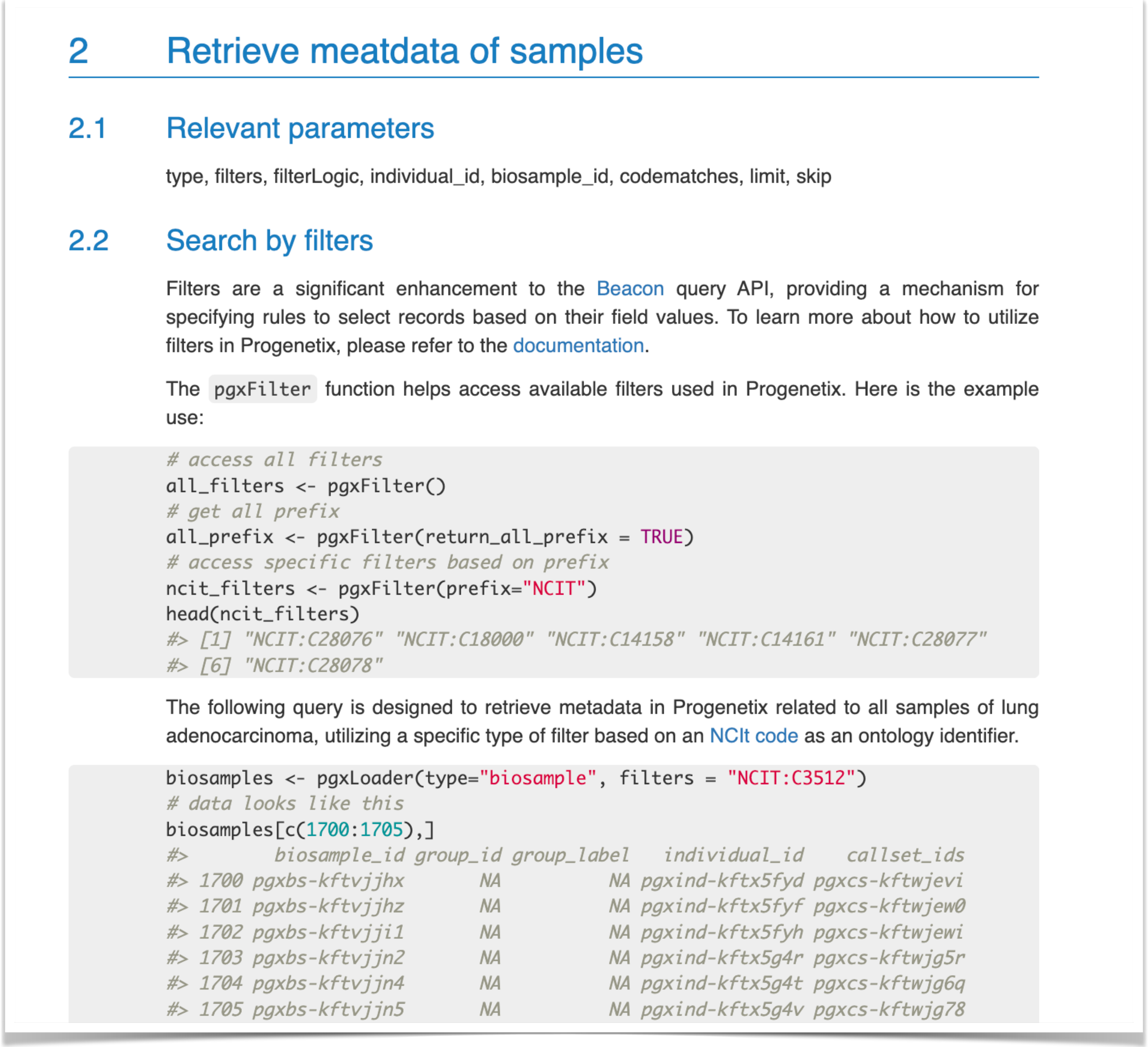
pgxRpi

An interface API for analyzing Progenetix CNV data in R using the Beacon+ API



hangjiaz version bump 319d27c 4 days ago 87 commits

R	adapt to beacon variant export	4 days ago
data-raw	add data documentation; optimise get method which causes error in ...	3 months ago
data	change based on opinion from bioc reviewer	last month
inst	add data documentation; optimise get method which causes error in ...	3 months ago
man	documentation for parameters using match.arg	last month
tests	adapt to API change; optimise code logic	last month
vignettes	documentation for parameters using match.arg	last month
.Rbuildignore	change vignette buildter; build vignette	2 months ago
.gitignore	modify gitignore	2 months ago
DESCRIPTION	version bump	4 days ago
NAMESPACE	add pgxFilter; change variant query logic and url; other code change...	last month
NEWS.md	add pgxFilter; change variant query logic and url; other code change...	last month
README.md	add pgxFilter; change variant query logic and url; other code change...	last month



2 Retrieve meatdata of samples

2.1 Relevant parameters

type, filters, filterLogic, individual_id, biosample_id, codematches, limit, skip

2.2 Search by filters

Filters are a significant enhancement to the Beacon query API, providing a mechanism for specifying rules to select records based on their field values. To learn more about how to utilize filters in Progenetix, please refer to the documentation.

The `pgxFilter` function helps access available filters used in Progenetix. Here is the example use:

```
# access all filters
all_filters <- pgxFilter()
# get all prefix
all_prefix <- pgxFilter(return_all_prefix = TRUE)
# access specific filters based on prefix
ncit_filters <- pgxFilter(prefix="NCIT")
head(ncit_filters)
#> [1] "NCIT:C28076" "NCIT:C18000" "NCIT:C14158" "NCIT:C14161" "NCIT:C28077"
#> [6] "NCIT:C28078"
```

The following query is designed to retrieve metadata in Progenetix related to all samples of lung adenocarcinoma, utilizing a specific type of filter based on an NCIt code as an ontology identifier.

```
biosamples <- pgxLoader(type="biosample", filters = "NCIT:C3512")
# data looks like this
biosamples[c(1700:1705),]
#>         biosample_id group_id group_label   individual_id     callset_ids
#> 1700 pgxbs-kftvjjhx       NA          NA pgxind-kftx5fyd pgxcs-kftwjevi
#> 1701 pgxbs-kftvjjhz       NA          NA pgxind-kftx5fyf pgxcs-kftwjew0
#> 1702 pgxbs-kftvjji1       NA          NA pgxind-kftx5fyh pgxcs-kftwjewi
#> 1703 pgxbs-kftvjjn2       NA          NA pgxind-kftx5g4r pgxcs-kftwjg5r
#> 1704 pgxbs-kftvjjn4       NA          NA pgxind-kftx5g4t pgxcs-kftwjg6q
#> 1705 pgxbs-kftvjjn5       NA          NA pgxind-kftx5g4v pgxcs-kftwjg78
```

What Can You Do?

- implement procedures and standards supporting **data discovery** (FAIR principles) and federation approaches
- forward looking consent and data protection models adhering to **ORD** principles ("*as secure as necessary, as open as possible*")
- **support** and/or get involved with international **data standards** efforts and projects

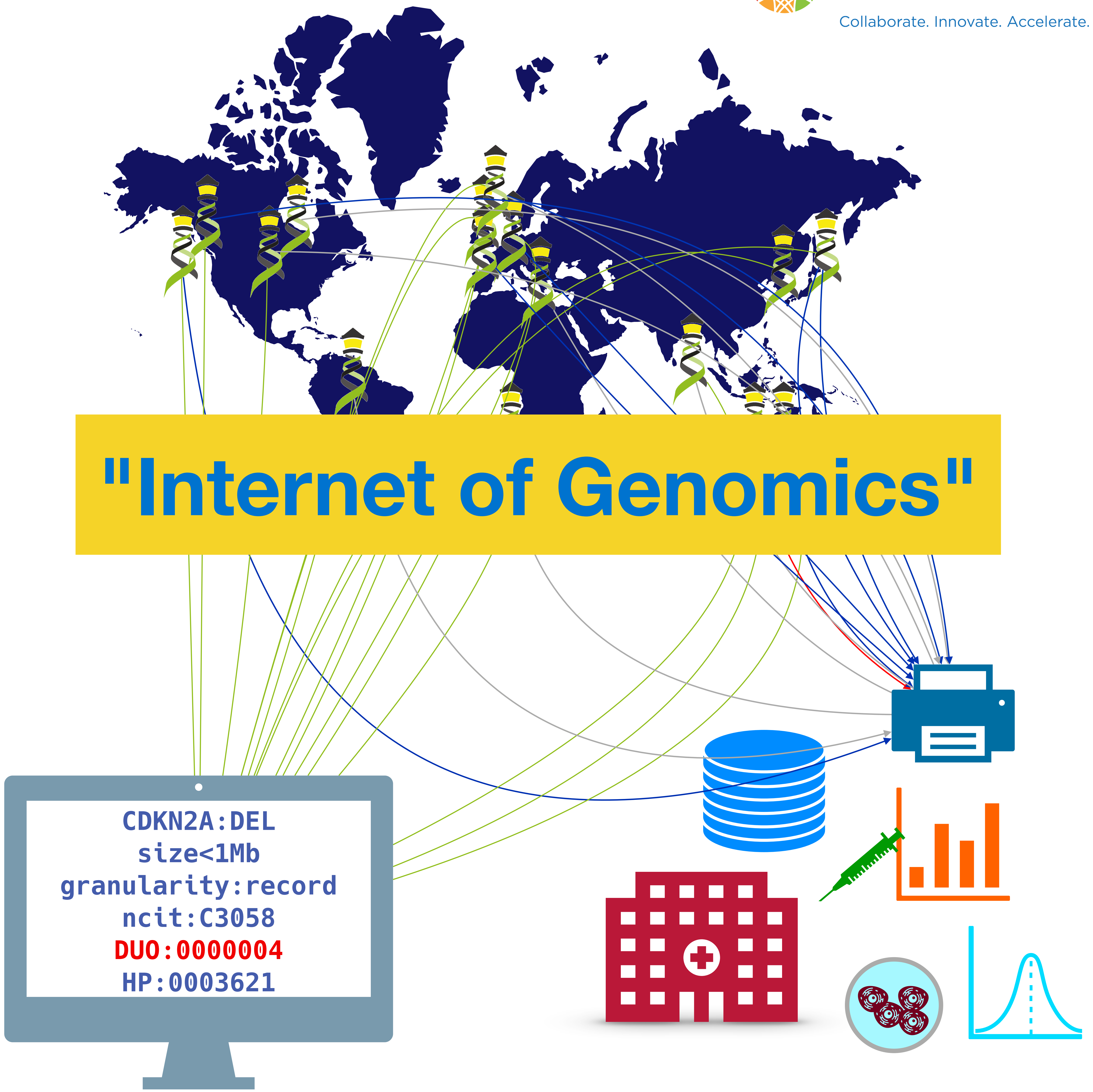
➡ **Collaborate!**

CDKN2A:DEL
size<1Mb
granularity:record
ncit:C3058
DUO:0000004
HP:0003621

What Can You Do?

- implement procedures and standards supporting **data discovery** (FAIR principles) and federation approaches
- forward looking consent and data protection models adhering to **ORD** principles ("*as secure as necessary, as open as possible*")
- **support** and/or get involved with international **data standards** efforts and projects

➡ **Collaborate!**



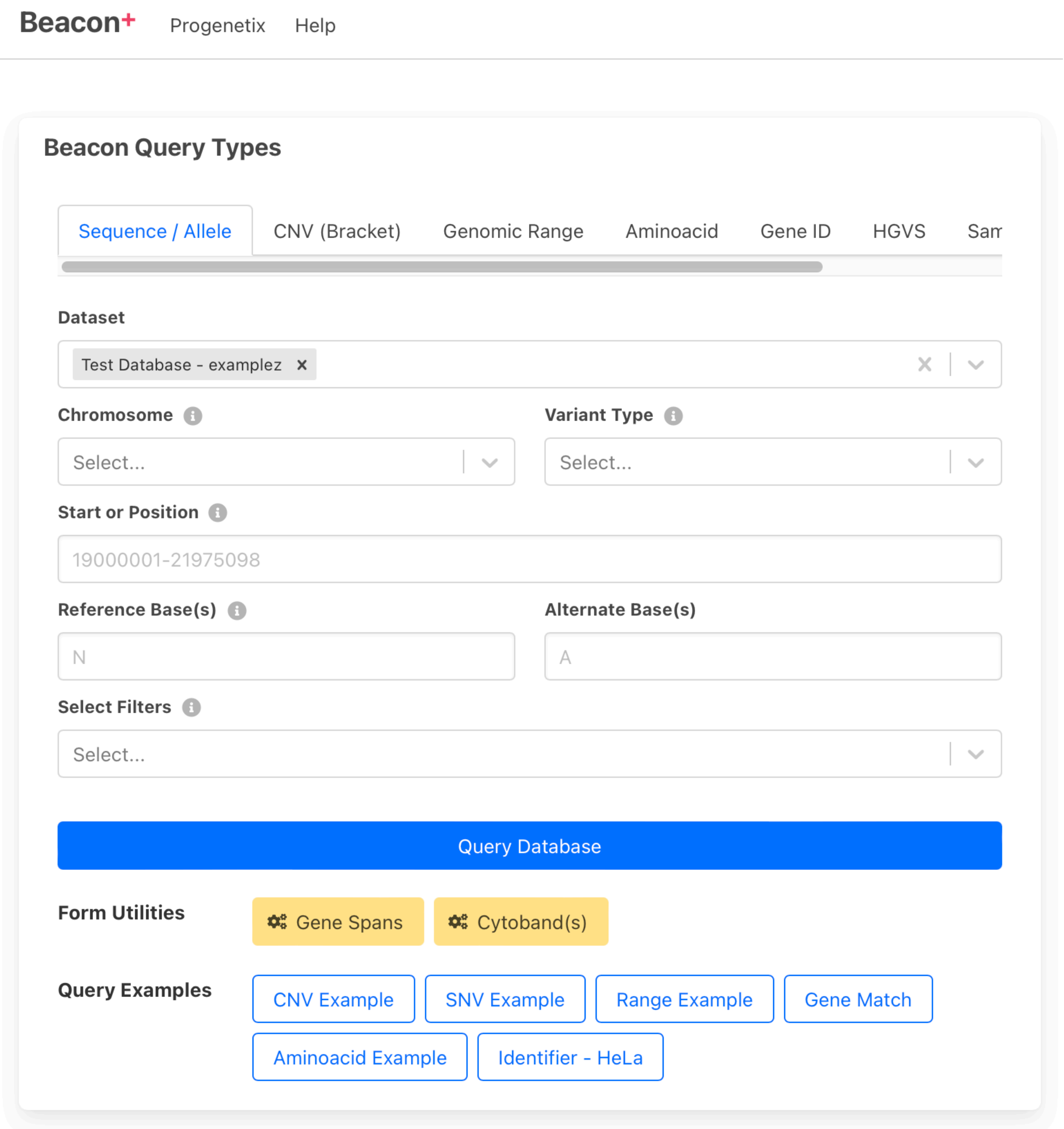
"Internet of Genomics"

Beacon Queries

Missing or ill defined options

- translocations are in principle possible (start bracket with "referenceName" and end bracket with "mateName") but not yet documented / battle tested
- functional elements?
- exon hits beyond specifying individual ones by sequence
- tandem dups ...

➡ **Beacon & hCNV Scout Team**



Beacon+ Progenetix Help

Beacon Query Types

Dataset: Test Database - examplez

Chromosome: Select...

Variant Type: Select...

Start or Position: 19000001-21975098

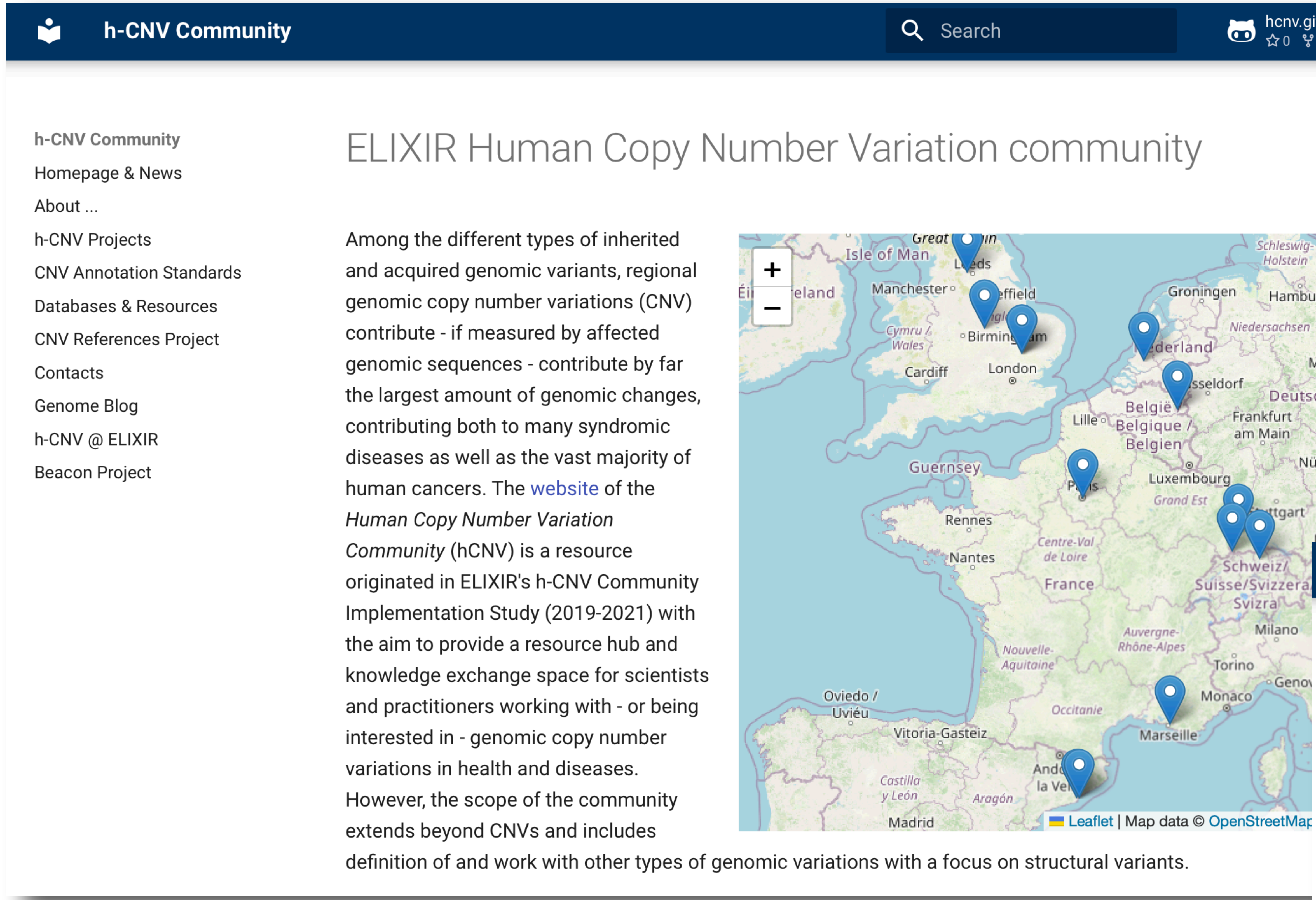
Reference Base(s): N

Alternate Base(s): A

Select Filters: Select...

ELIXIR hCNV Community

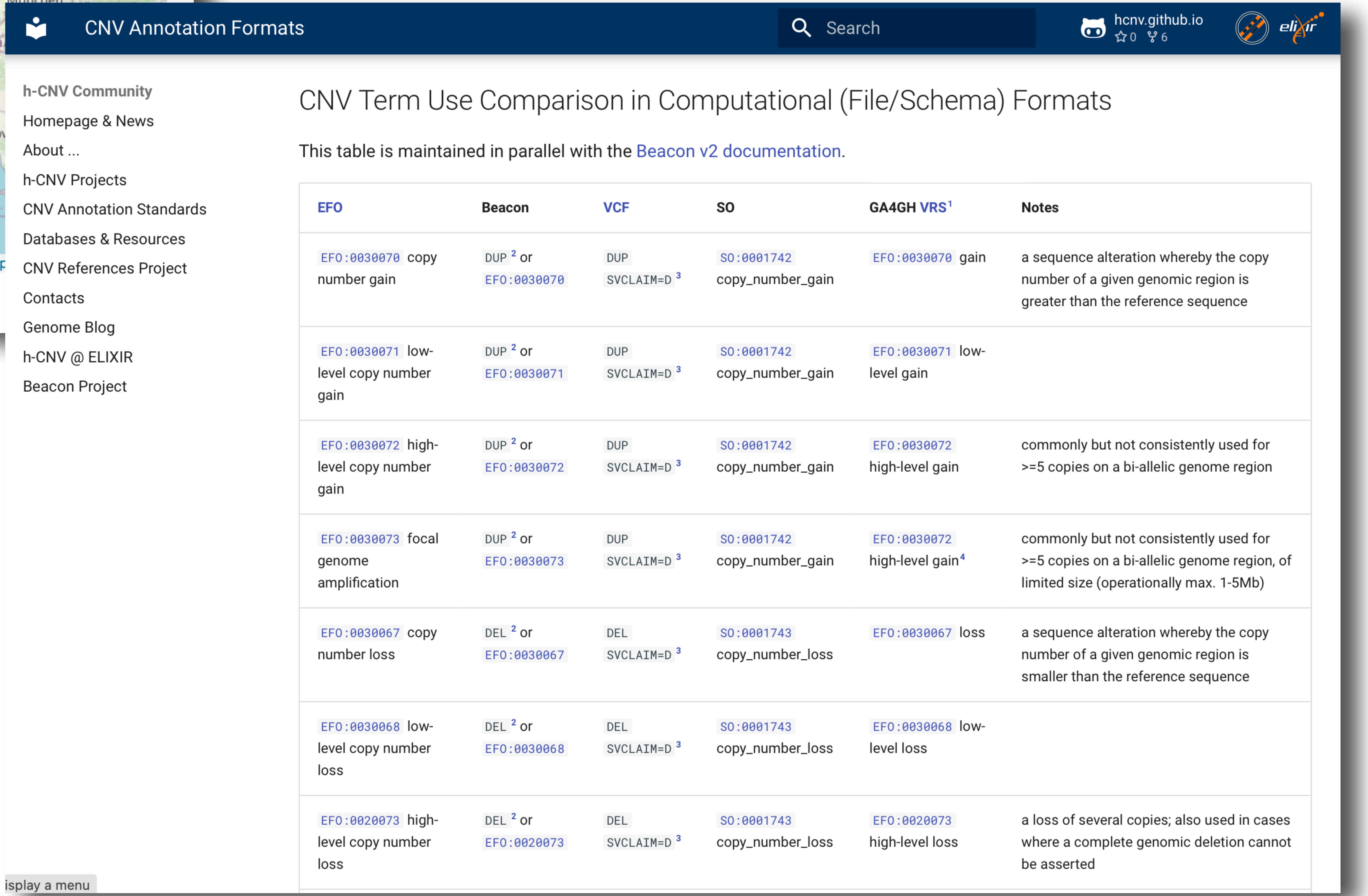
https://cnvar.org/



h-CNV Community
Homepage & News
About ...
h-CNV Projects
CNV Annotation Standards
Databases & Resources
CNV References Project
Contacts
Genome Blog
h-CNV @ ELIXIR
Beacon Project

ELIXIR Human Copy Number Variation community

Among the different types of inherited and acquired genomic variants, regional genomic copy number variations (CNV) contribute - if measured by affected genomic sequences - contribute by far the largest amount of genomic changes, contributing both to many syndromic diseases as well as the vast majority of human cancers. The website of the *Human Copy Number Variation Community* (hCNV) is a resource originated in ELIXIR's h-CNV Community Implementation Study (2019-2021) with the aim to provide a resource hub and knowledge exchange space for scientists and practitioners working with - or being interested in - genomic copy number variations in health and diseases. However, the scope of the community extends beyond CNVs and includes definition of and work with other types of genomic variations with a focus on structural variants.



h-CNV Community
Homepage & News
About ...
h-CNV Projects
CNV Annotation Standards
Databases & Resources
CNV References Project
Contacts
Genome Blog
h-CNV @ ELIXIR
Beacon Project

CNV Term Use Comparison in Computational (File/Schema) Formats

This table is maintained in parallel with the Beacon v2 documentation.

EFO	Beacon	VCF	SO	GA4GH VRS[1]	Notes
EFO:0030070 copy number gain	DUP [2] or EFO:0030070	DUP SVCLAIM=D [3]	SO:0001742 copy_number_gain	EFO:0030070 gain	a sequence alteration whereby the copy number of a given genomic region is greater than the reference sequence
EFO:0030071 low-level copy number gain	DUP [2] or EFO:0030071	DUP SVCLAIM=D [3]	SO:0001742 copy_number_gain	EFO:0030071 low-level gain	
EFO:0030072 high-level copy number gain	DUP [2] or EFO:0030072	DUP SVCLAIM=D [3]	SO:0001742 copy_number_gain	EFO:0030072 high-level gain	commonly but not consistently used for >=5 copies on a bi-allelic genome region
EFO:0030073 focal genome amplification	DUP [2] or EFO:0030073	DUP SVCLAIM=D [3]	SO:0001742 copy_number_gain	EFO:0030072 high-level gain[4]	commonly but not consistently used for >=5 copies on a bi-allelic genome region, of limited size (operationally max. 1-5Mb)
EFO:0030067 copy number loss	DEL [2] or EFO:0030067	DEL SVCLAIM=D [3]	SO:0001743 copy_number_loss	EFO:0030067 loss	a sequence alteration whereby the copy number of a given genomic region is smaller than the reference sequence
EFO:0030068 low-level copy number loss	DEL [2] or EFO:0030068	DEL SVCLAIM=D [3]	SO:0001743 copy_number_loss	EFO:0030068 low-level loss	
EFO:0020073 high-level copy number loss	DEL [2] or EFO:0020073	DEL SVCLAIM=D [3]	SO:0001743 copy_number_loss	EFO:0020073 high-level loss	a loss of several copies; also used in cases where a complete genomic deletion cannot be asserted

The Beacon team through the ages

EUROPEAN GENOME-PHENOME ARCHIVE

CRG Centre for Genomic Regulation

Jordi Rambla
Arcadi Navarro
Roberto Ariosa
Manuel Rueda
Lauren Fromont
Mauricio Moldes
Claudia Vasallo
Babita Singh
Sabela de la Torre
Marta Ferri
Fred Haziza

CSC

Juha Törnroos
Teemu Kataja
Ilkka Lappalainen
Dylan Spalding

UNIVERSITY OF LEICESTER

Cafe Variome Central

Tony Brookes
Tim Beck
Colin Veal
Tom Shorter

SPHN Swiss Personalized Health Network

University of Zurich UZH

Michael Baudis
Rahel Paloots
Hangjia Zhao
Ziying Yang
Bo Gao
Qingyao Huang

Genomics england

Augusto Rendon
Ignacio Medina
Javier López
Jacobo Coll
Antonio Rueda

cnag centre nacional d'anàlisi genòmica centro nacional de análisis genómico

Sergi Beltran
Carles Hernandez

Inserm Institut national de la santé et de la recherche médicale

David Salgado

Barcelona Supercomputing Center Centro Nacional de Supercomputación

Salvador Capella
Dmitry Repchevski
JM Fernández

DisGeNET

Laura Furlong
Janet Piñero

elixir B1MG

Serena Scollen
Gary Saunders
Giselle Kerry
David Lloyd

H3Africa Human Heredity & Health in Africa

Nicola Mulder
Mamana Mbiyavanga
Ziyaad Parker

EU Can CAN.

David Torrents

AUTISM SPEAKS

Dean Hartley

JUNTA DE ANDALUCIA

Fundación Progreso y Salud CONSEJERÍA DE SALUD

Joaquin Dopazo
Javier Pérez
J.L. Fernández
Gema Roldan

CINECA

Thomas Keane
Melanie Courtot
Jonathan Dursi

Heidi Rehm
Ben Hutton

GEM Japan

Toshiaki Katayama

Stephane Dyke

DNASTACK

Marc Fiume
Miro Cupak

BRCA EXCHANGE

Melissa Cline

ENA

EMBL-EBI

Diana Lemos

EUROPEAN JOINT PROGRAMME RARE DISEASES

VICC Variant Interpretation for Cancer Consortium

GA4GH Phenopackets
Peter Robinson
Jules Jacobsen

GA4GH VRS
Alex Wagner
Reece Hart

Beacon PRC
Alex Wagner
Jonathan Dursi
Mamana Mbiyavanga

Alice Mann
Neerjah Skantharajah

elixir

Universität Zürich UZH

SIB Swiss Institute of Bioinformatics

elixir

Universität Zürich^UZH

SIB

Swiss Institute of
Bioinformatics

elixir