Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

Universität
Zürich UZH

# Genomic Data Mining and The Case for 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

Michael Baudis @ ICLS Colloquium

SIB
Swiss Institute of
Bioinformatics

elixir

# Genome screening at the core of “Personalised Health”

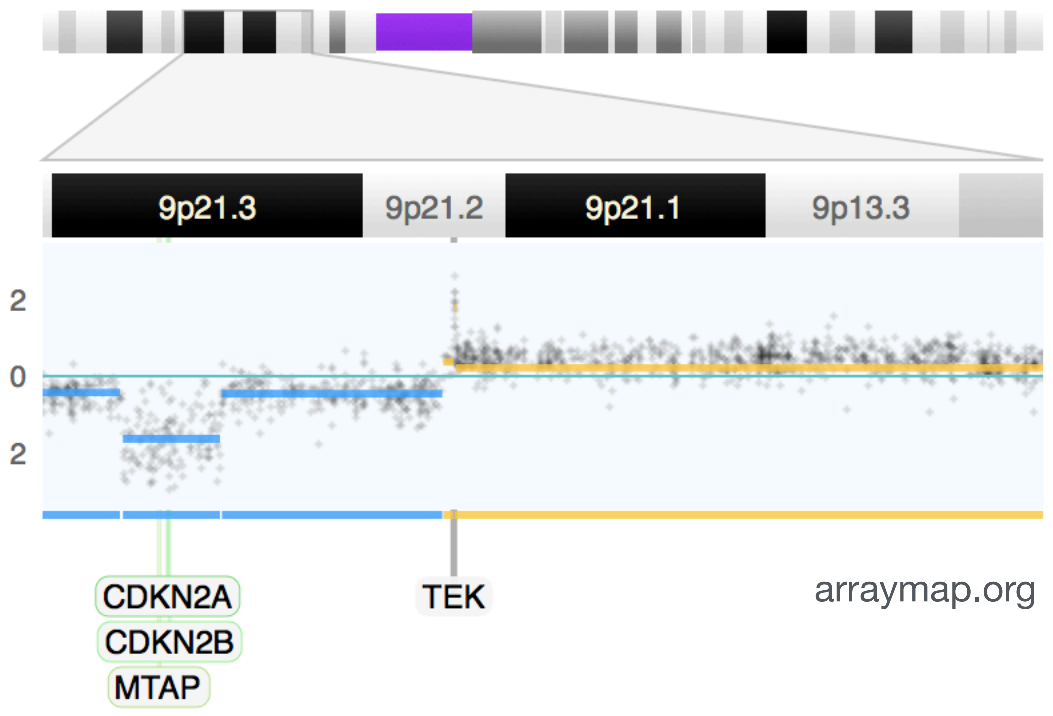
- **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**
- **Standardized data formats** and **exchange protocols** are needed to connect these resources throughout the world, for reciprocal, international **data sharing** and **biocuration** efforts
- Our work @ UZH:
  - ***cancer*** genome repositories
  - biocuration
  - protocols & formats

arrayMap

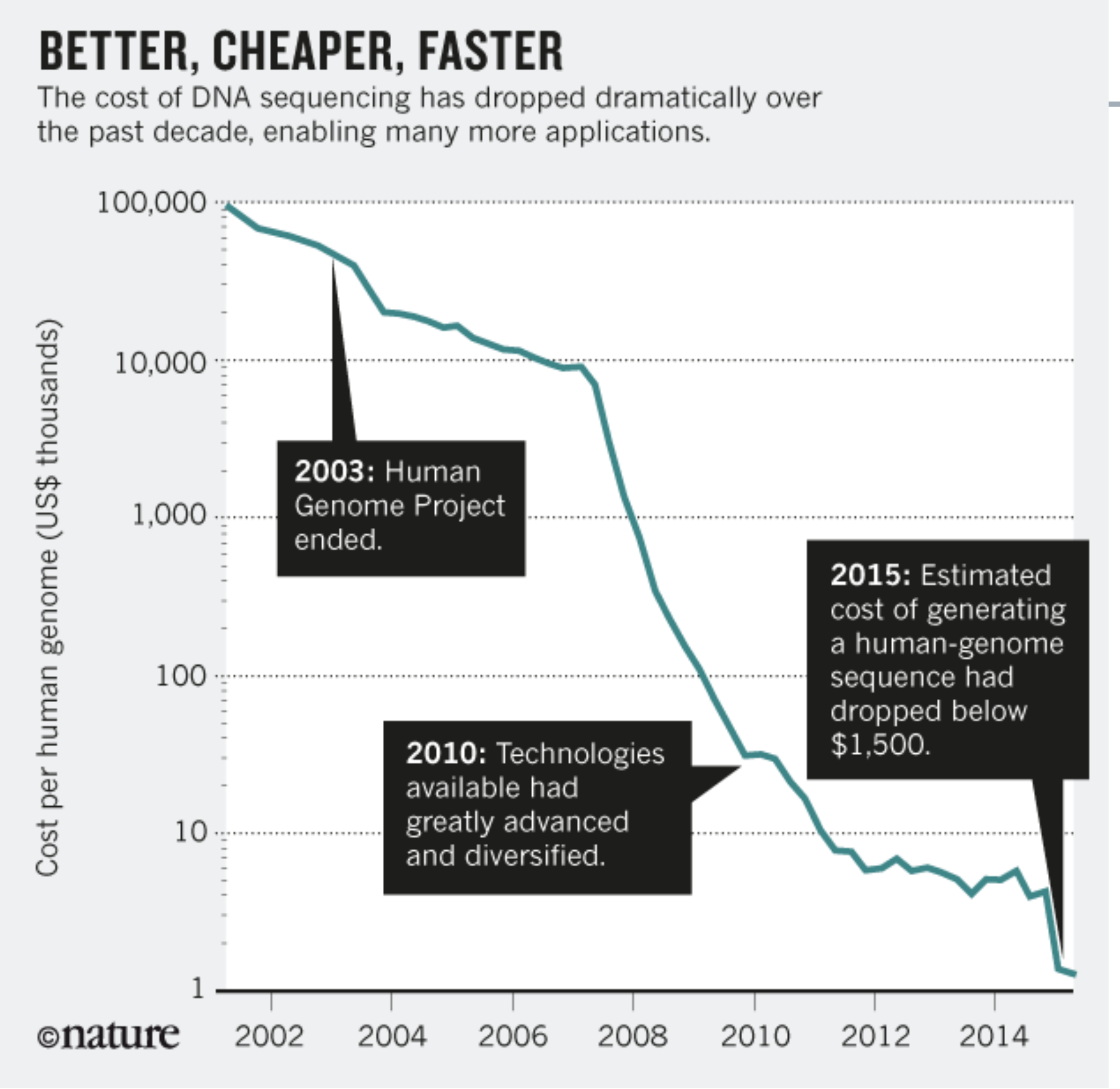
progenetix

elixir

Global Alliance for Genomics & Health



arraymap.org



BETTER, CHEAPER, FASTER

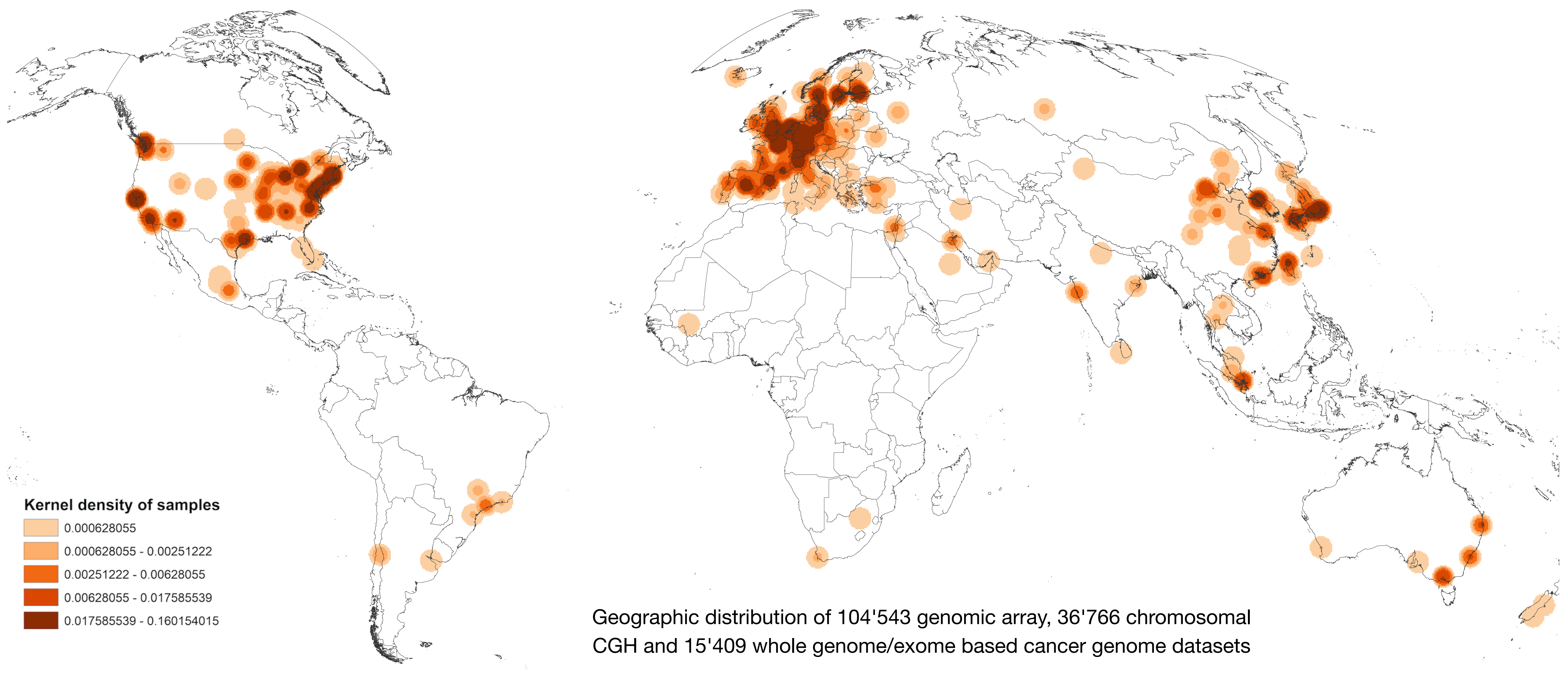
The cost of DNA sequencing has dropped dramatically over the past decade, enabling many more applications.

**2003:** Human Genome Project ended.

**2010:** Technologies available had greatly advanced and diversified.

**2015:** Estimated cost of generating a human-genome sequence had dropped below $1,500.

The future of DNA sequencing. Eric D. Green, Edward M. Rubin & Maynard V. Olson. Nature; 11 October 2017 (News & Views)



**Kernel density of samples**

| | |
|---|---|
| 🟧 | 0.000628055 |
| 🟧 | 0.000628055 - 0.00251222 |
| 🟧 | 0.00251222 - 0.00628055 |
| 🟥 | 0.00628055 - 0.017585539 |
| 🟫 | 0.017585539 - 0.160154015 |

Geographic distribution of 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets

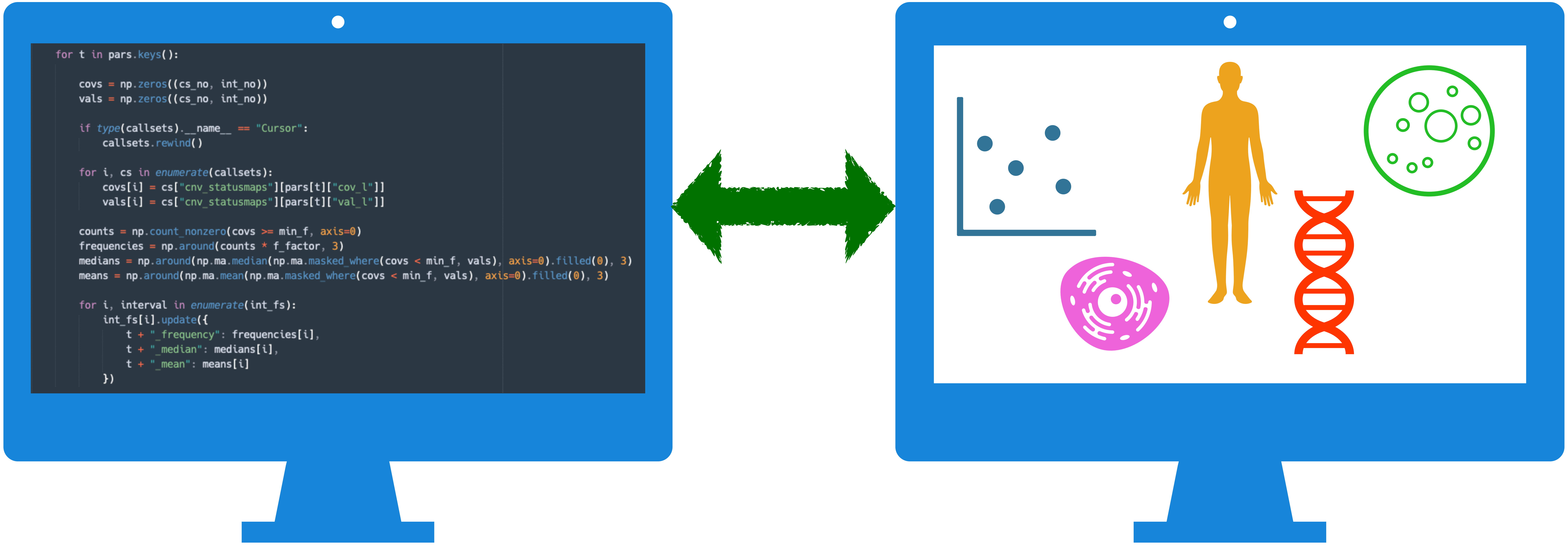
# Theoretical Cytogenetics and Oncogenomics

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

Curators

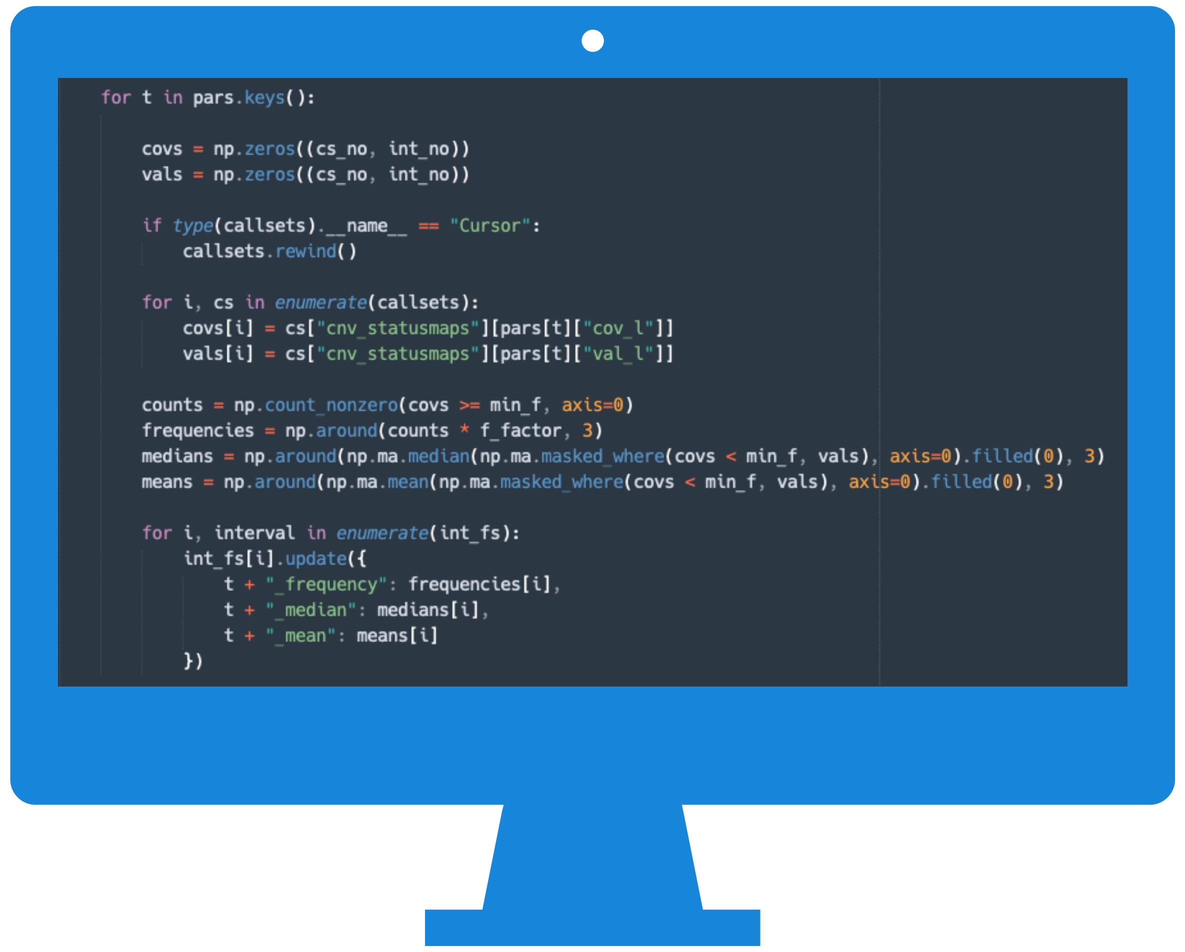
Data ~~Parasites~~

# {BioInformaticsScience}

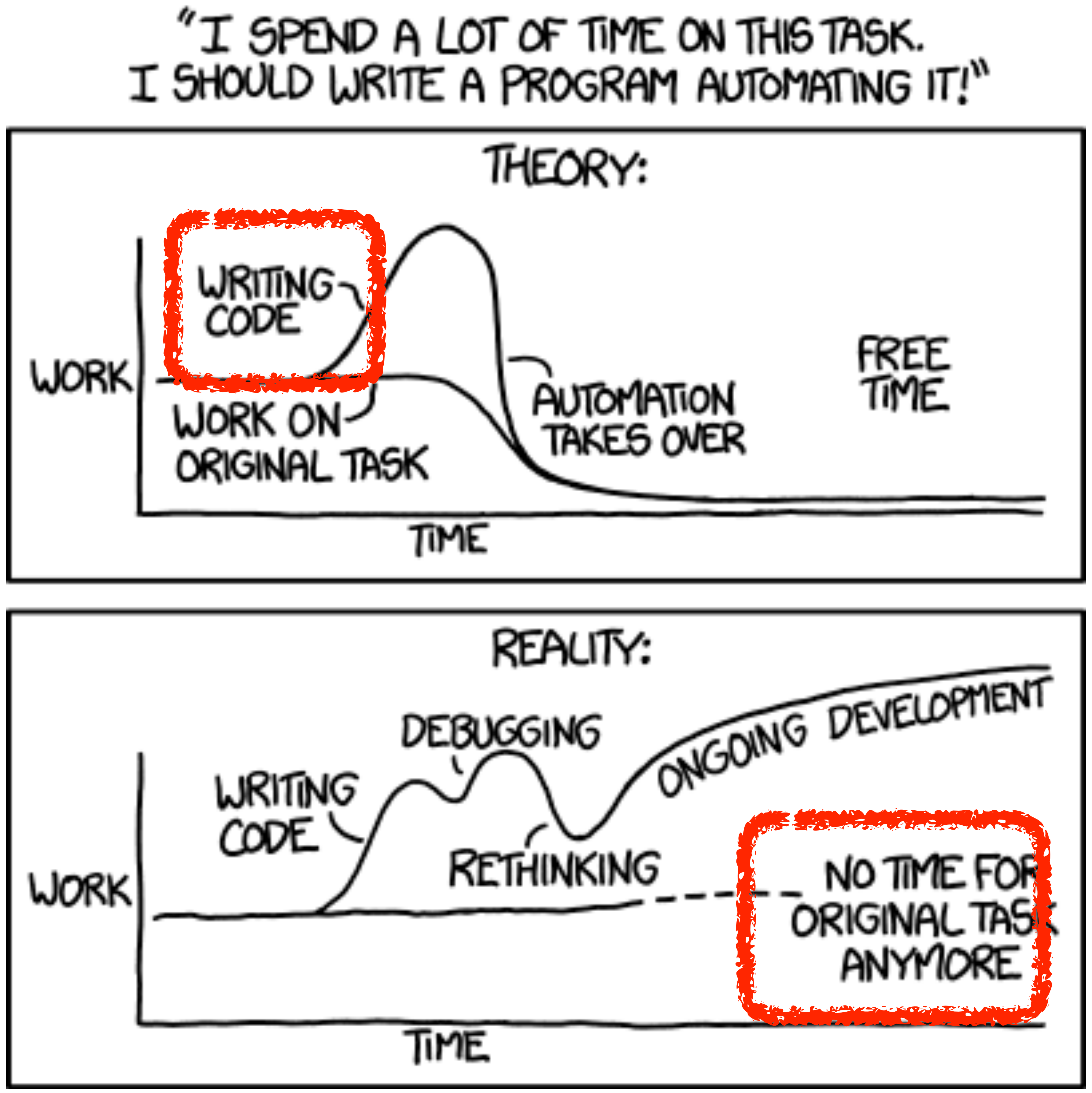


```
for t in pars.keys():

    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]
        })
```

# {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]
        })
```



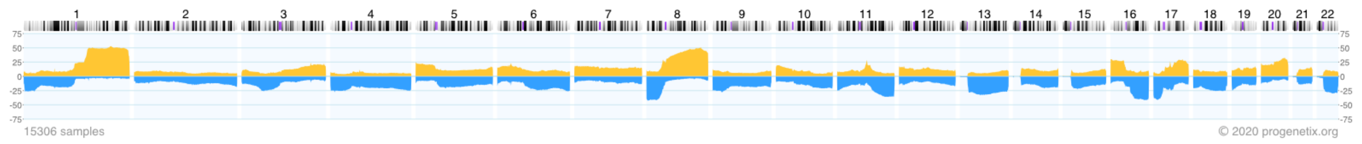
Randall Munroe - XKCD
*https://xkcd.com/1319/*

# Theoretical Cytogenetics and Oncogenomics

## ... but what does this entail @baudisgroup?

- patterns & markers in cancer genomics, especially somatic structural genome variants
- bioinformatics support in collaborative studies

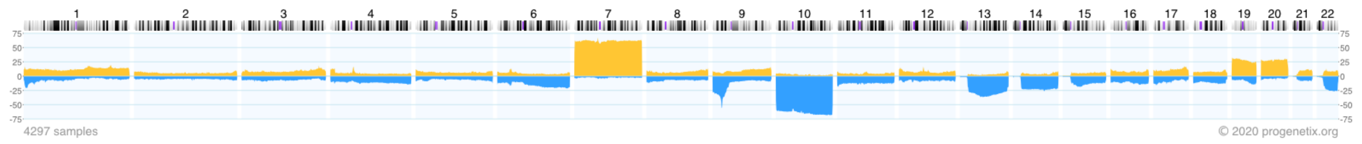
Malignant Breast Neoplasm (NCIT:C9335)



15306 samples

© 2020 progenetix.org

Glioblastoma (NCIT:C3058)

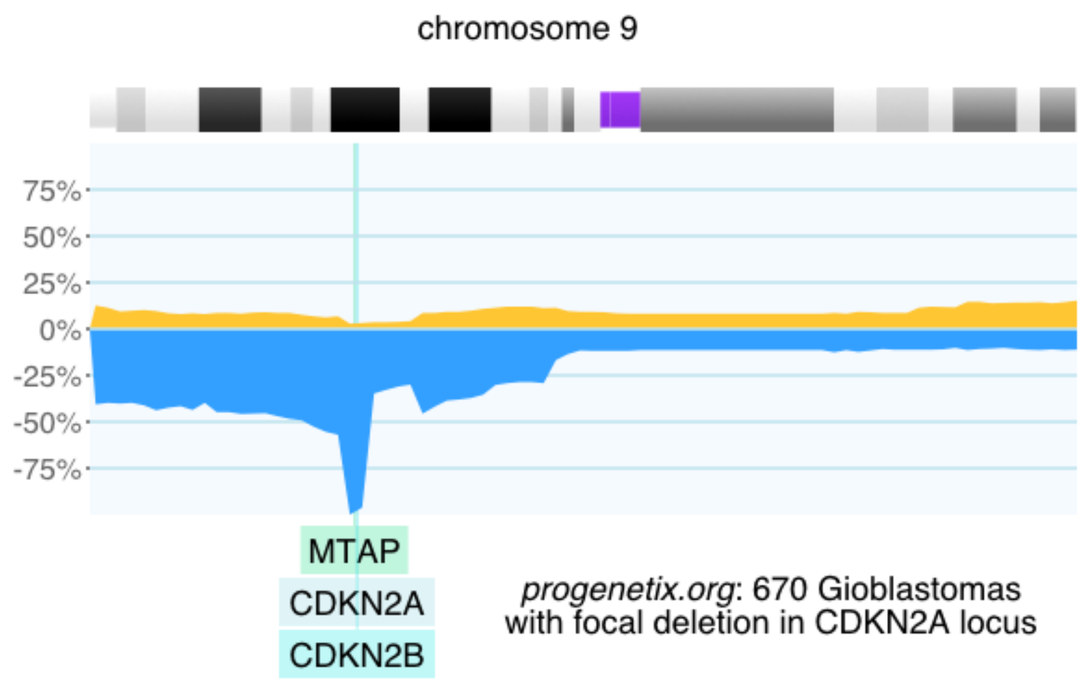


4297 samples

© 2020 progenetix.org

- 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"

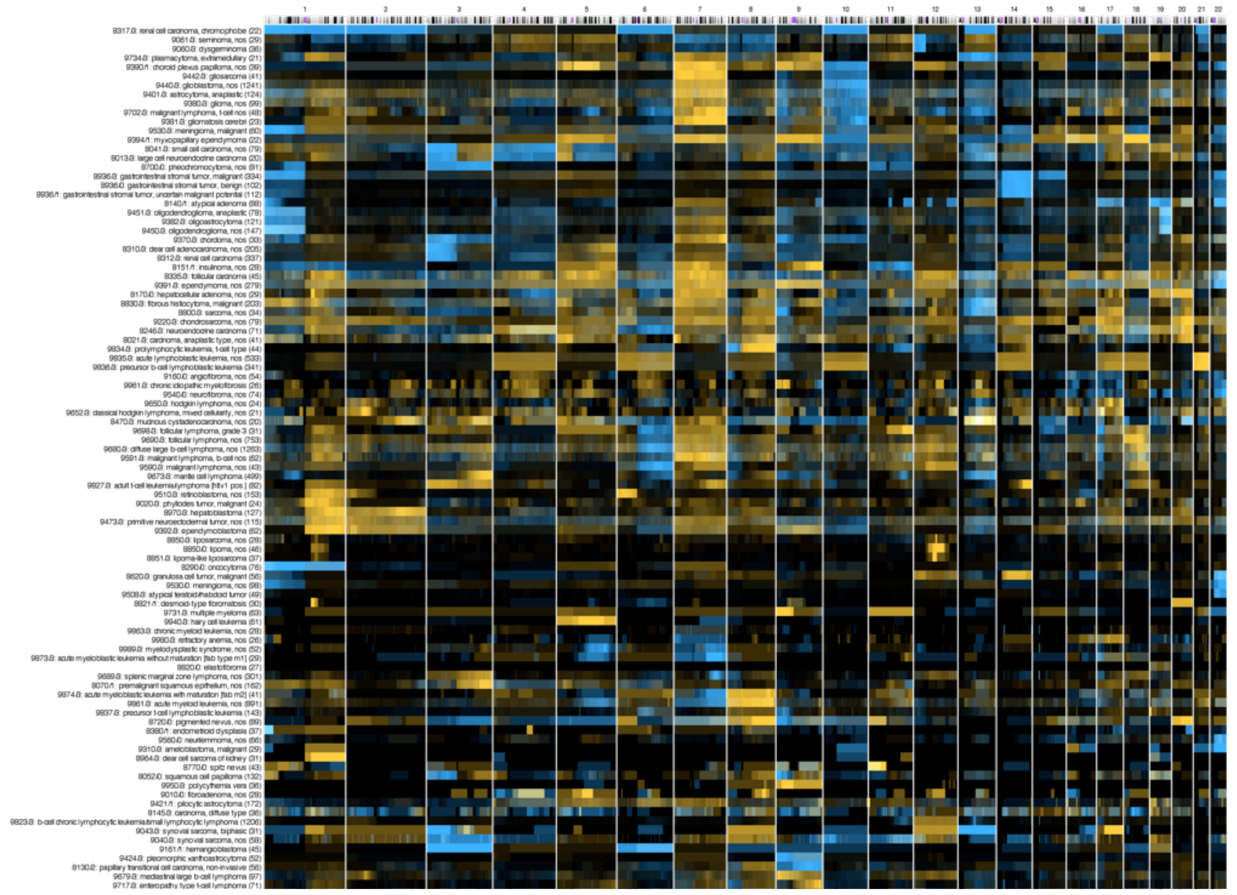
# Theoretical Cytogenetics and Oncogenomics Research | Methods | Standards

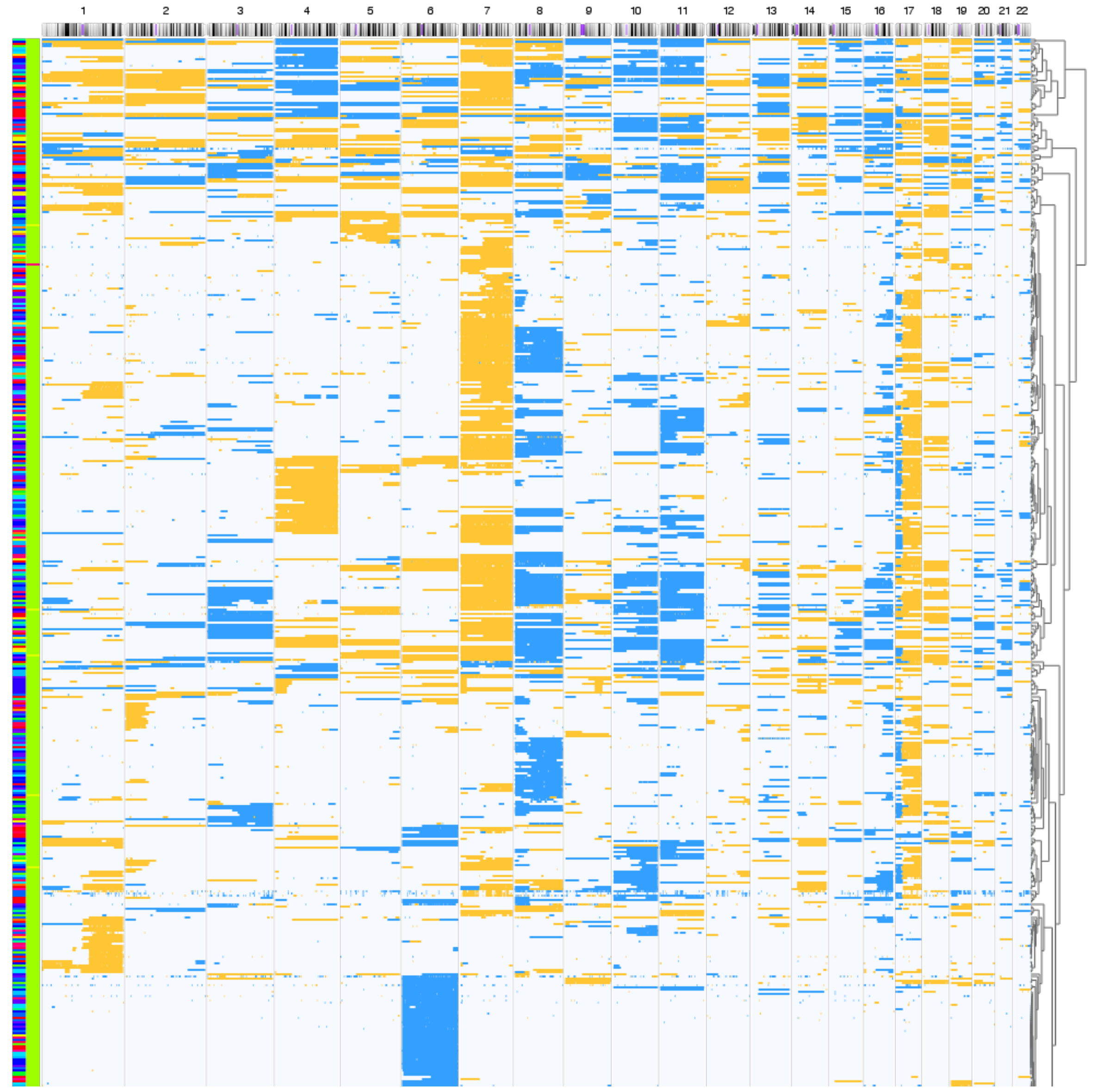


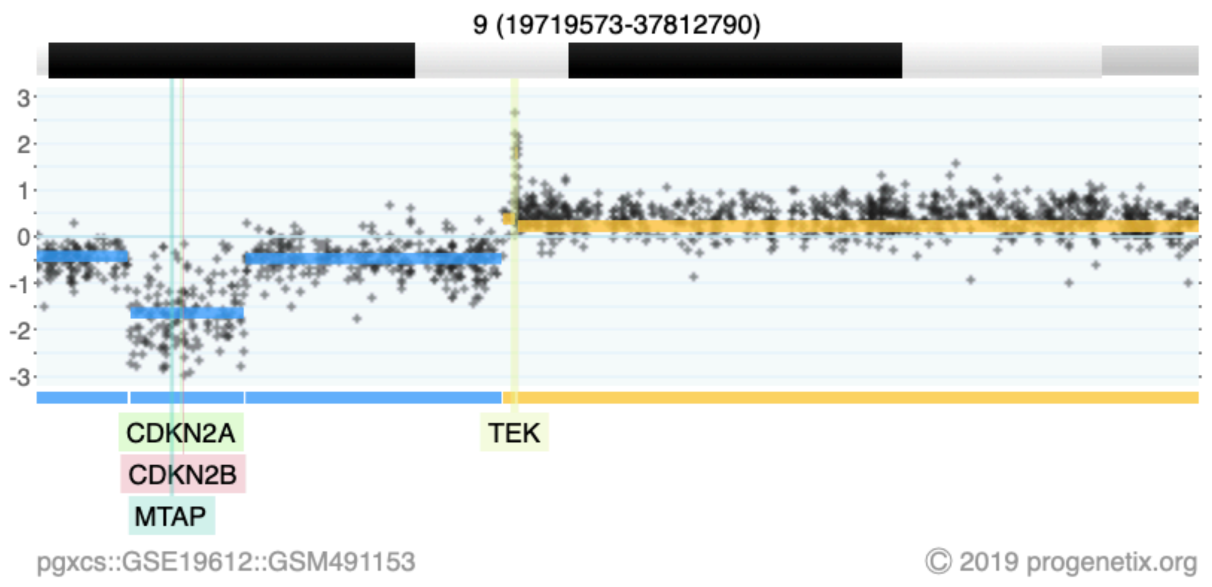
*progenetix.org*: 670 Gioblastomas with focal deletion in CDKN2A locus

## Genomic Imbalances in Cancer - Copy Number Variations (CNV)

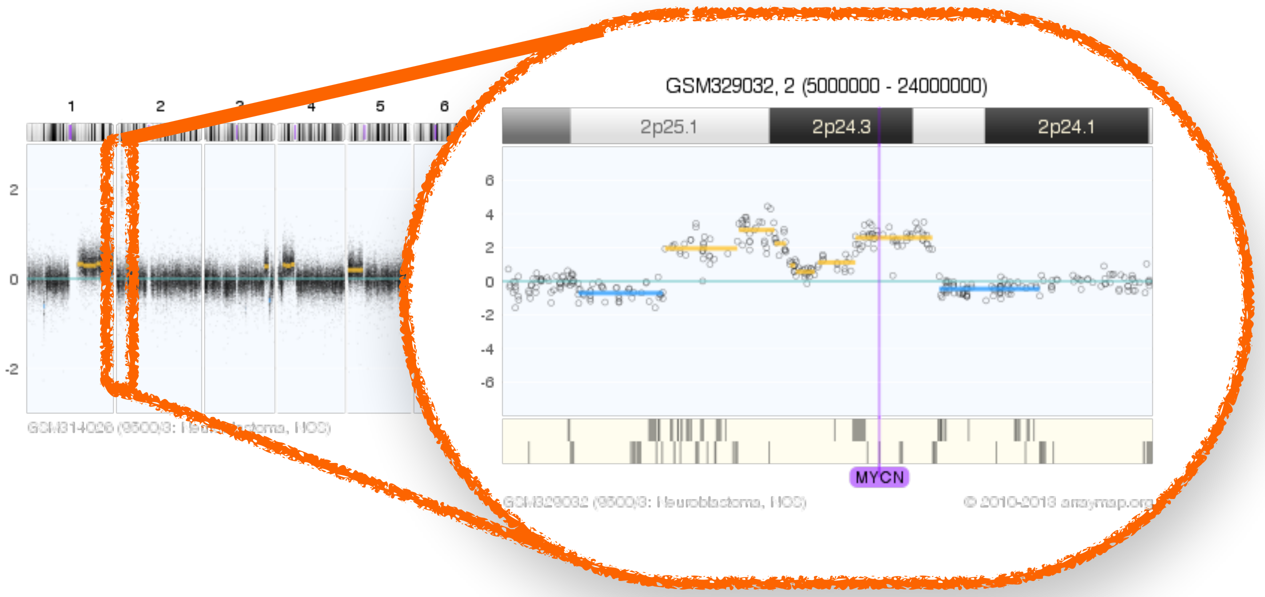
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)







2-event, homozygous deletion in a Glioblastoma



MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

# progenetix.org

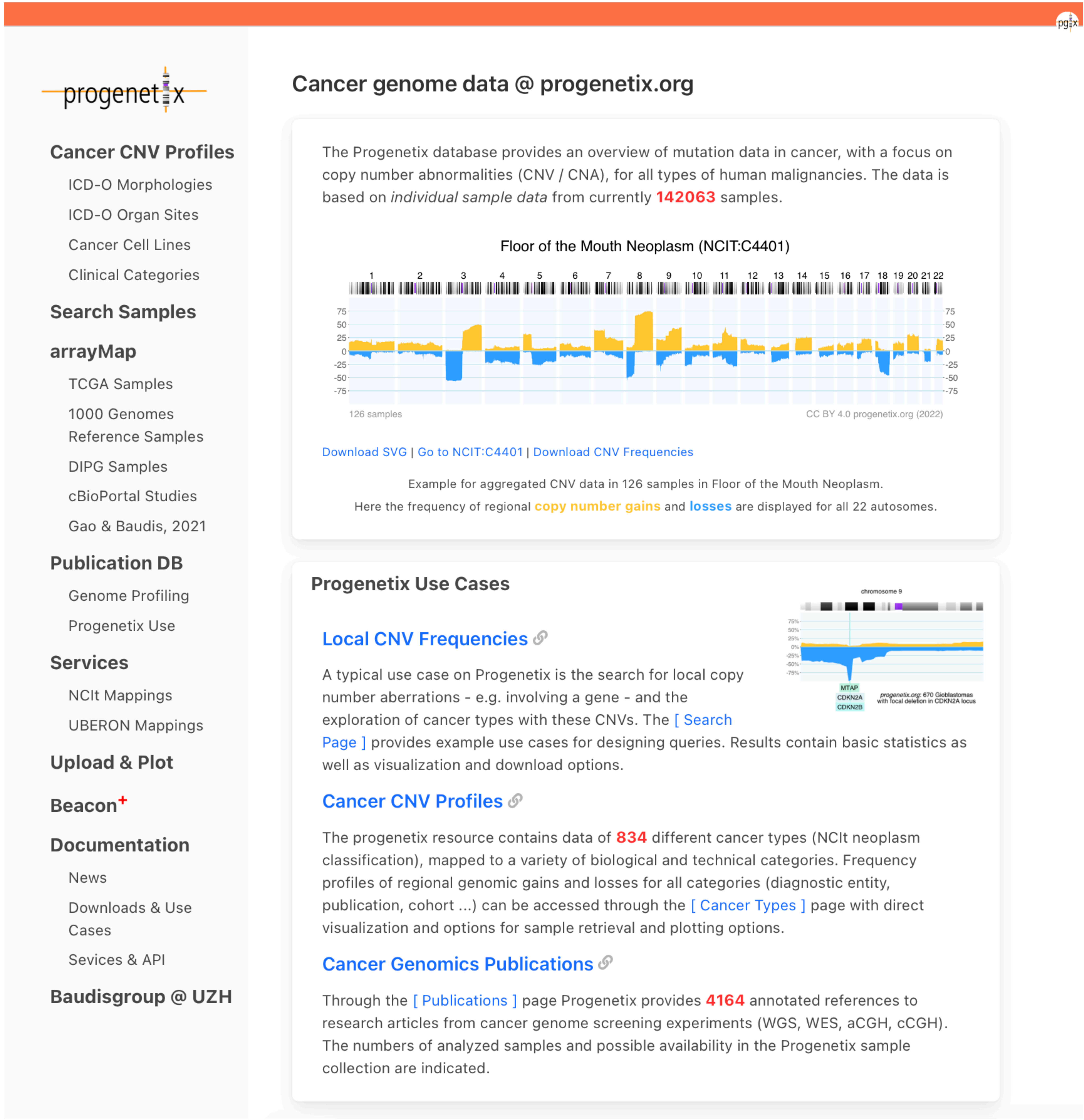
## Cancer Genomics Reference Resource

- ***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
Chromosomes: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
75
50
25
0
-25
-50
-75
126 samples
CC BY 4.0 progenetix.org (2022)
chromosome 9
75%
50%
25%
0%
-25%
-50%
-75%
MTAP
CDKN2A
CDKN2B
progenetix.org: 670 Gioblastomas with focal deletion in CDKN2A locus

Cancer CNV Profiles
- ICD-O Morphologies
- ICD-O Organ Sites
- Cancer Cell Lines
- Clinical Categories

Search Samples

arrayMap
- TCGA Samples
- 1000 Genomes Reference Samples
- DIPG Samples
- cBioPortal Studies
- Gao & Baudis, 2021

Publication DB
- Genome Profiling
- Progenetix Use

Services
- NCIt Mappings
- UBERON Mappings

Upload & Plot

Beacon+

Documentation
- News
- Downloads & Use Cases
- Sevices & API

Baudisgroup @ UZH

### 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.

Floor of the Mouth Neoplasm (NCIT:C4401)

Download SVG | Go to NCIT:C4401 | Download CNV Frequencies

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

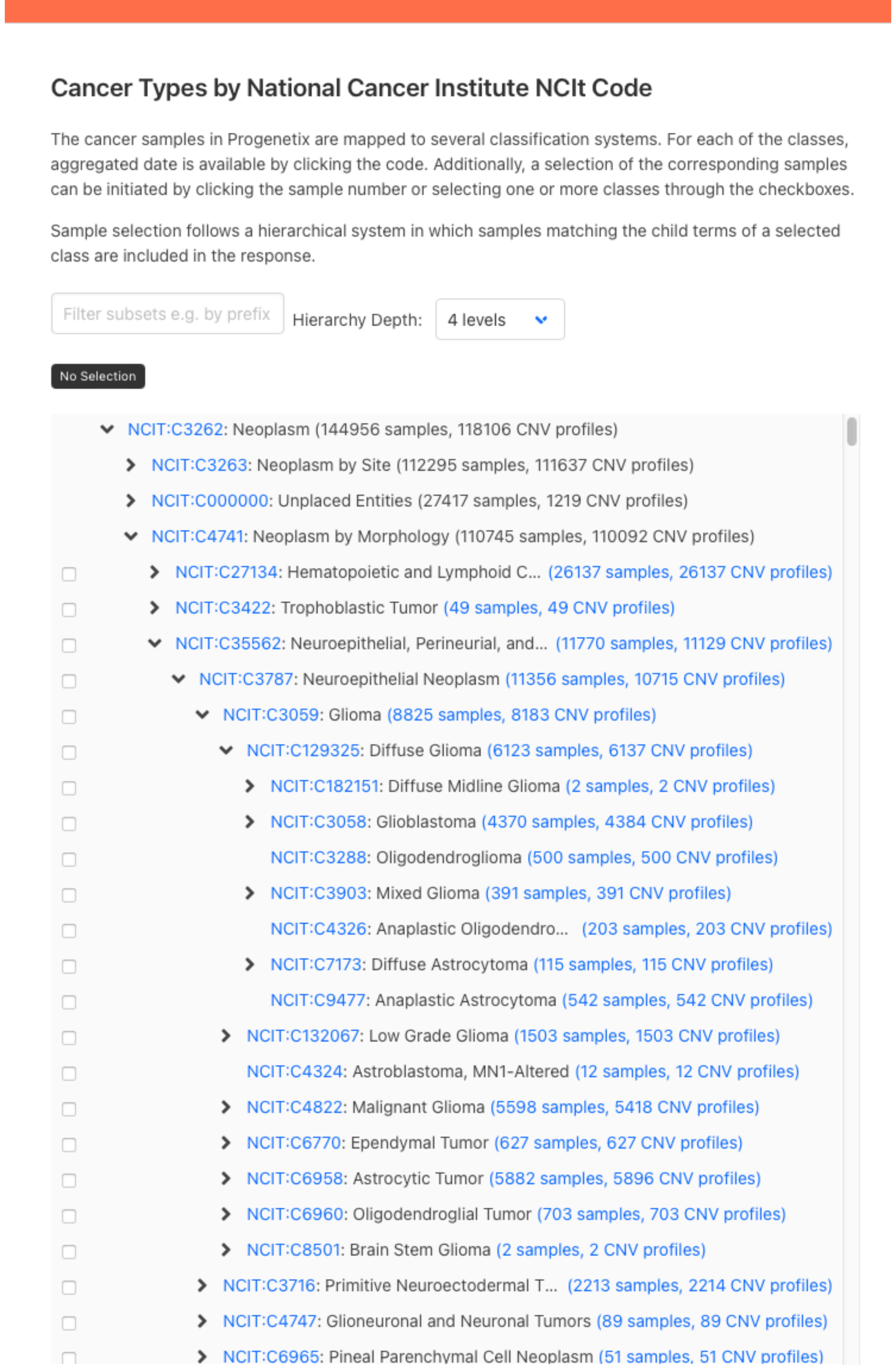
## Cancer Genomics Reference Resource

- ***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

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: 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)

# progenetix.org

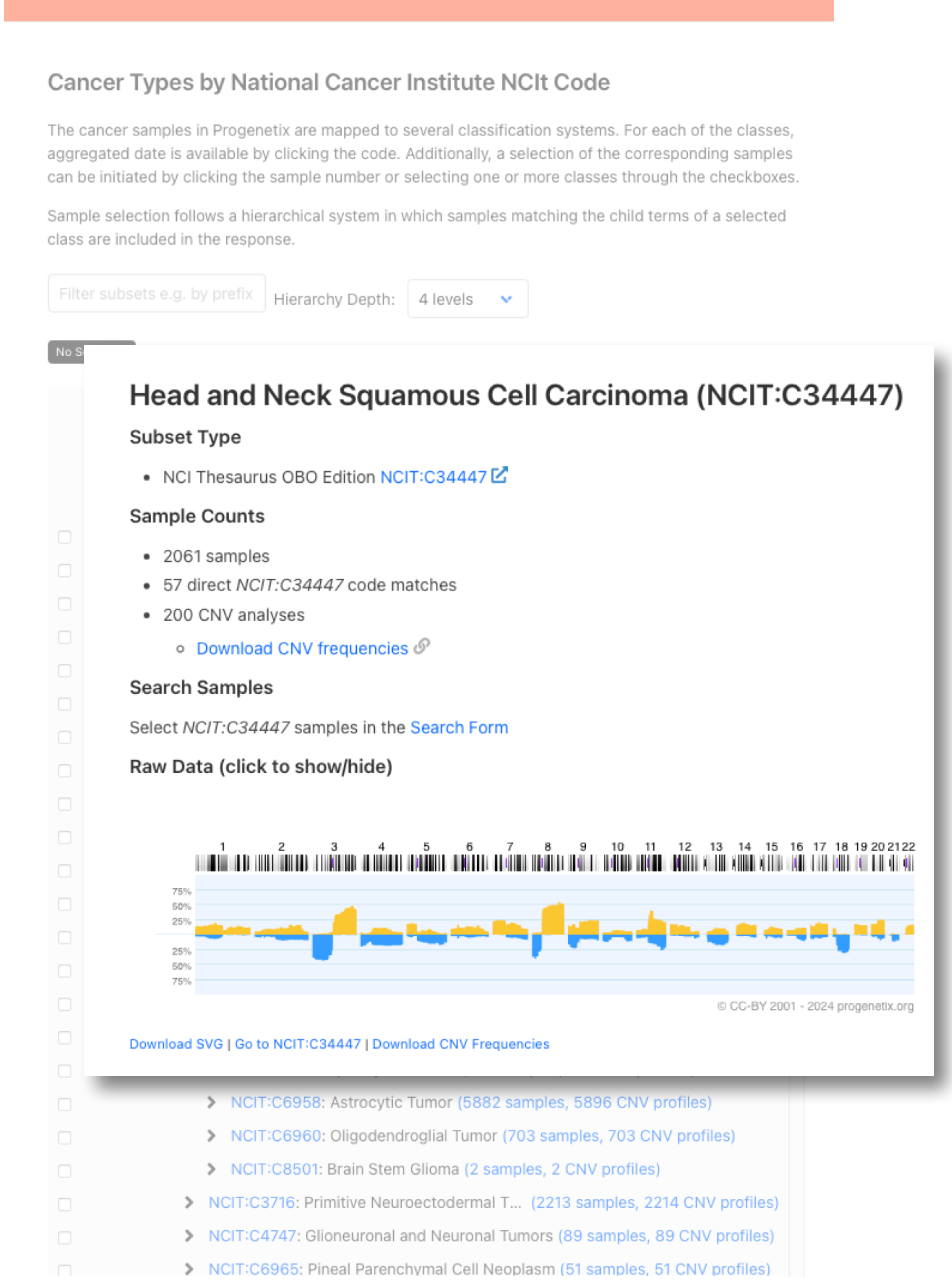
## Cancer Genomics Reference Resource

- ***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 S

Head and Neck Squamous Cell Carcinoma (NCIT:C34447)

Subset Type

- NCI Thesaurus OBO Edition NCIT:C34447

Sample Counts

- 2061 samples
- 57 direct *NCIT:C34447* code matches
- 200 CNV analyses
  - Download CNV frequencies

Search Samples

Select *NCIT:C34447* samples in the Search Form

Raw Data (click to show/hide)

© CC-BY 2001 - 2024 progenetix.org

Download SVG | Go to NCIT:C34447 | Download CNV Frequencies

- 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

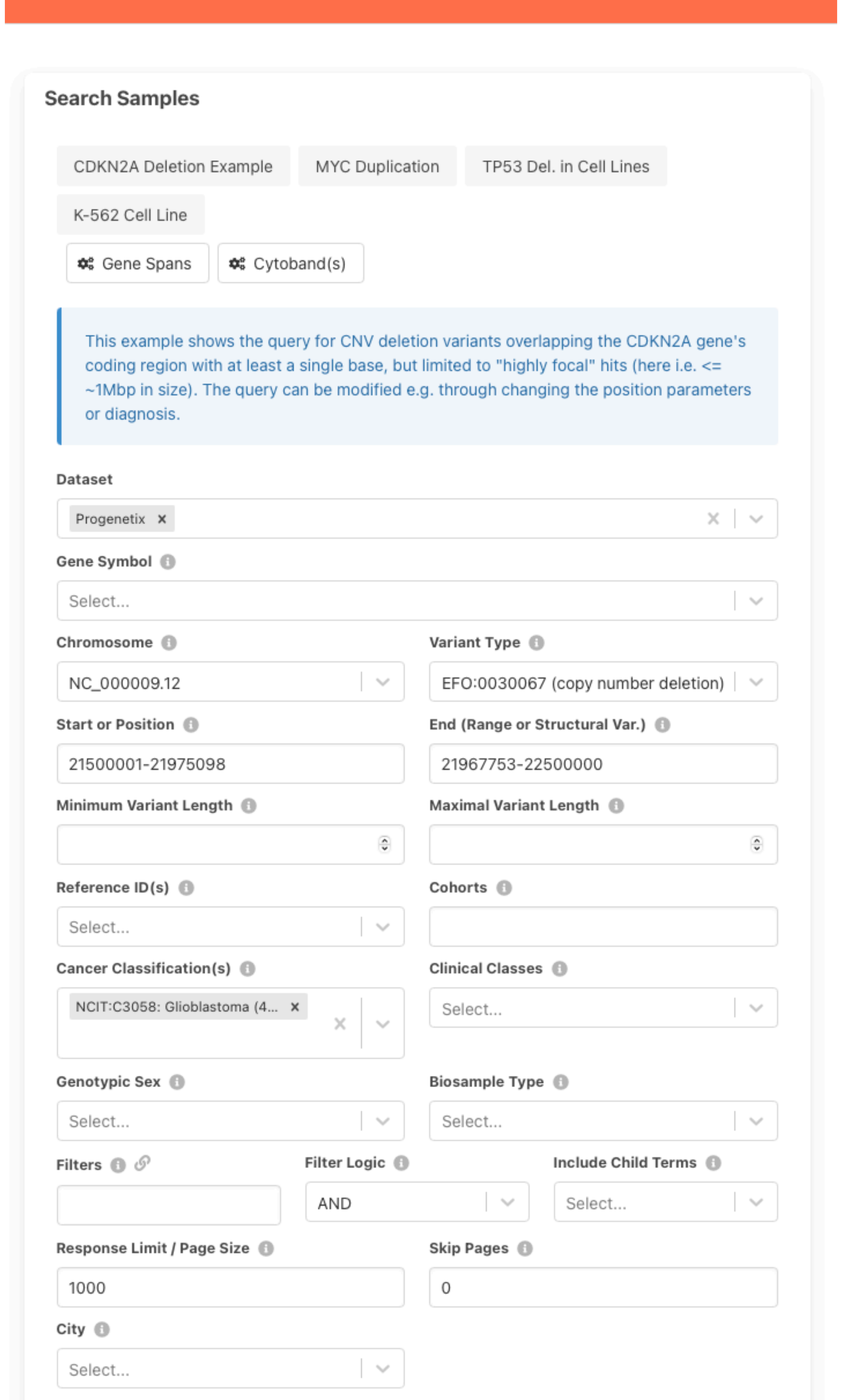
## Cancer Genomics Reference Resource

- ***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

Universität Zürich UZH

progenetix

SIB Swiss Institute of Bioinformatics



**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...

# progenetix.org

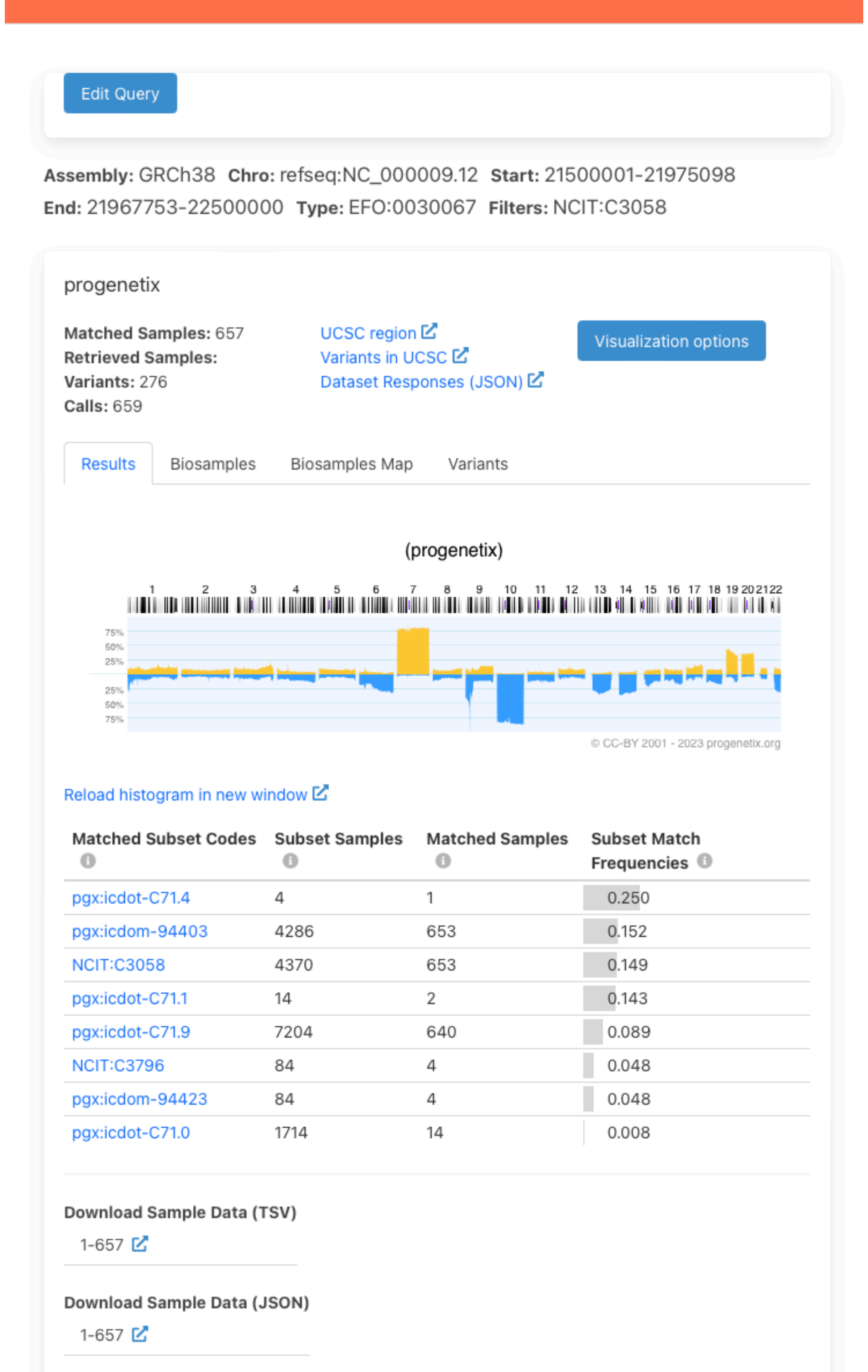
## Cancer Genomics Reference Resource

- ***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)

© CC-BY 2001 - 2023 progenetix.org

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

# progenetix.org

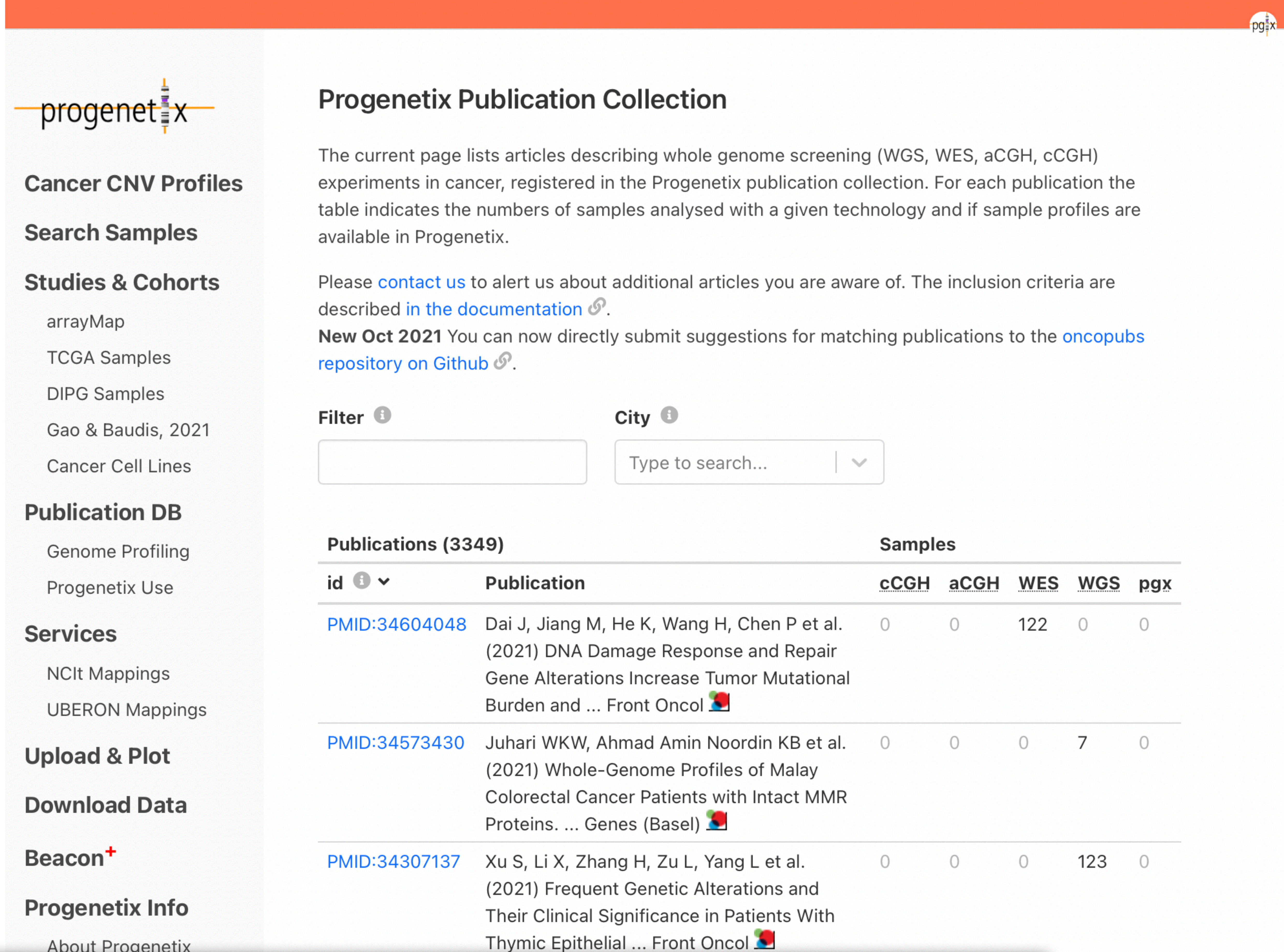
## Cancer Genomics Reference Resource

- ***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.

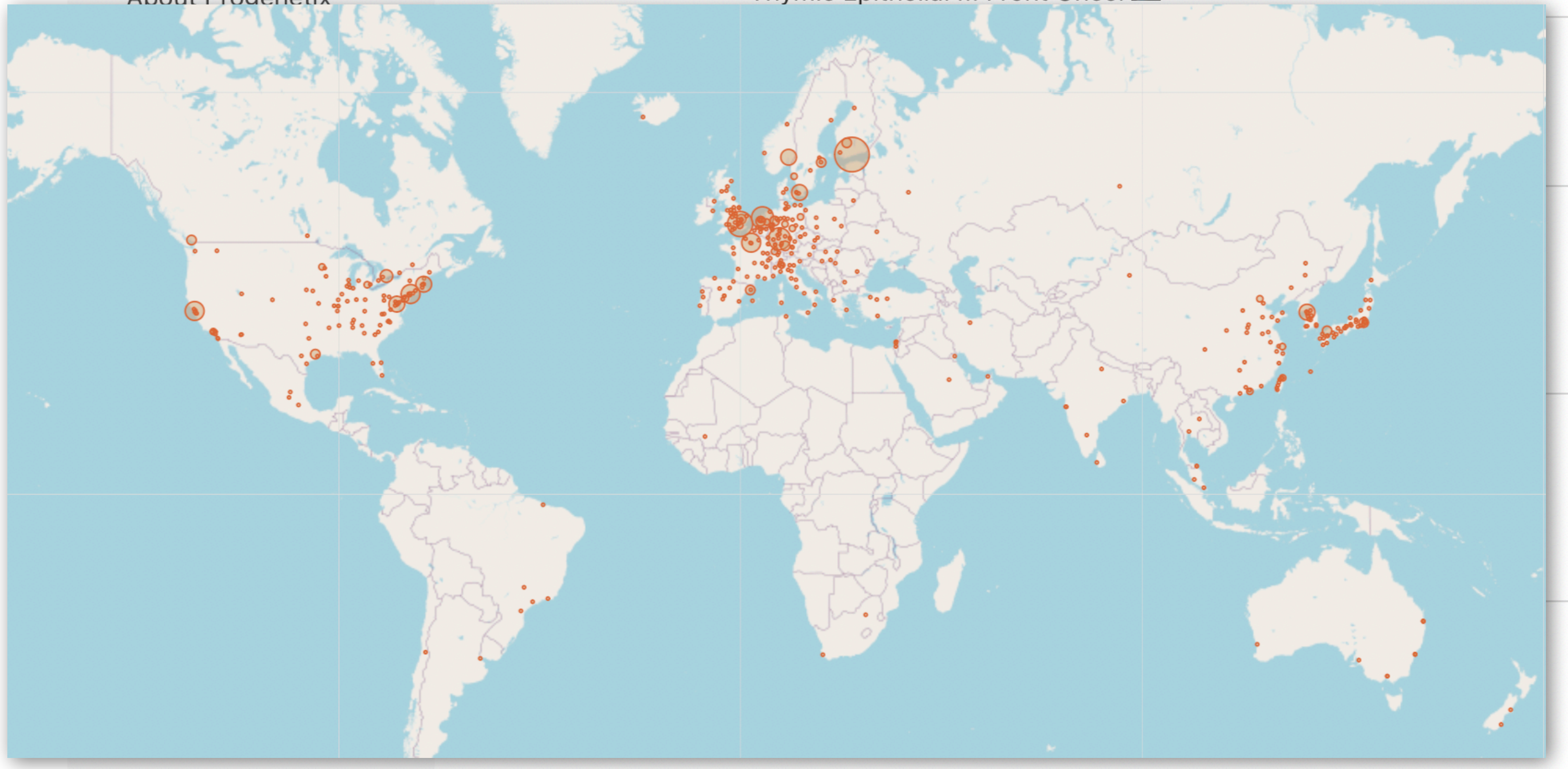
Please contact us to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation.

**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 |



# Cancer Cell Lines

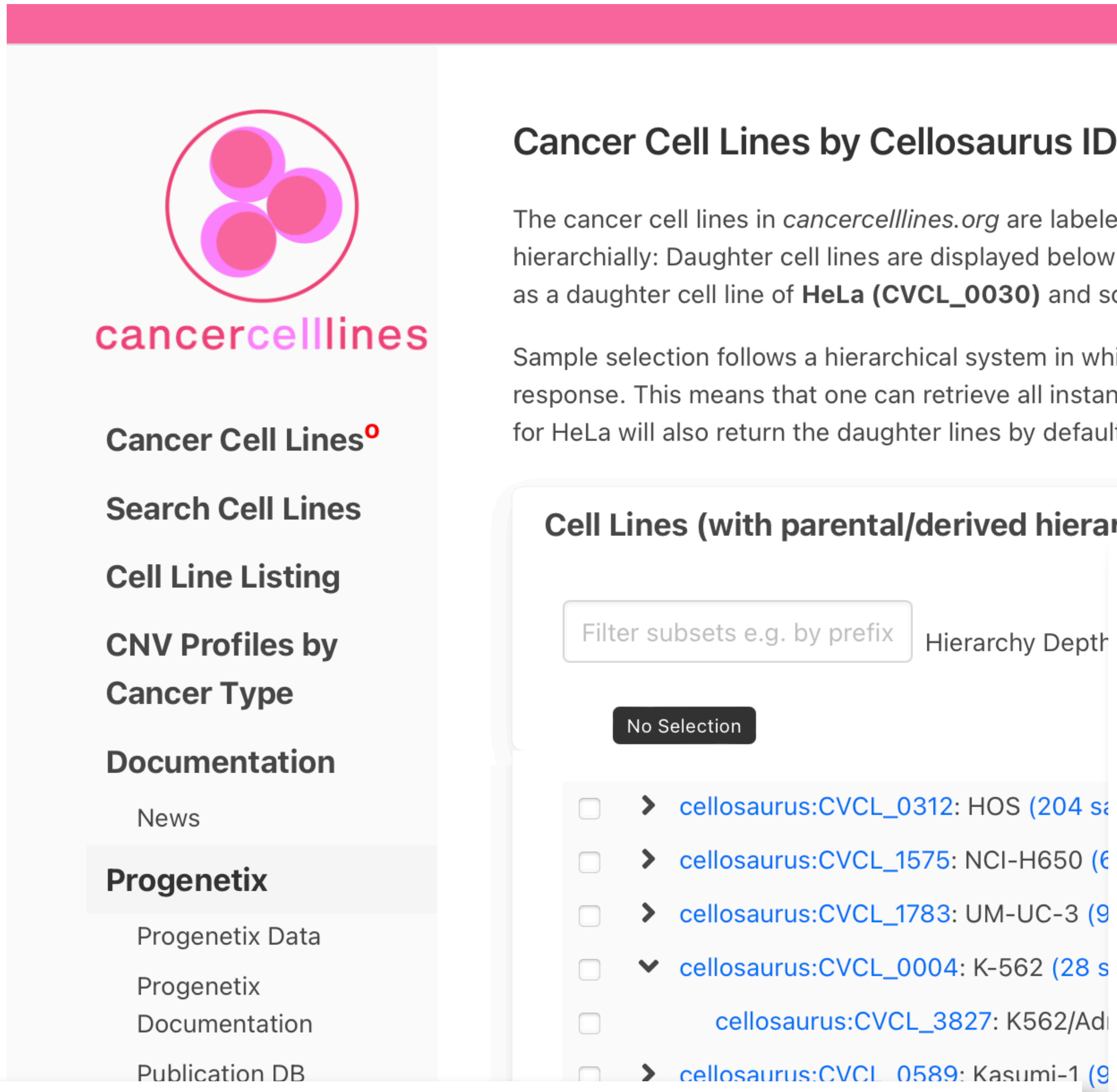
## Cancer Genomics Reference Resource

- 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

cancercelllines.org

Lead: Rahel Paloots



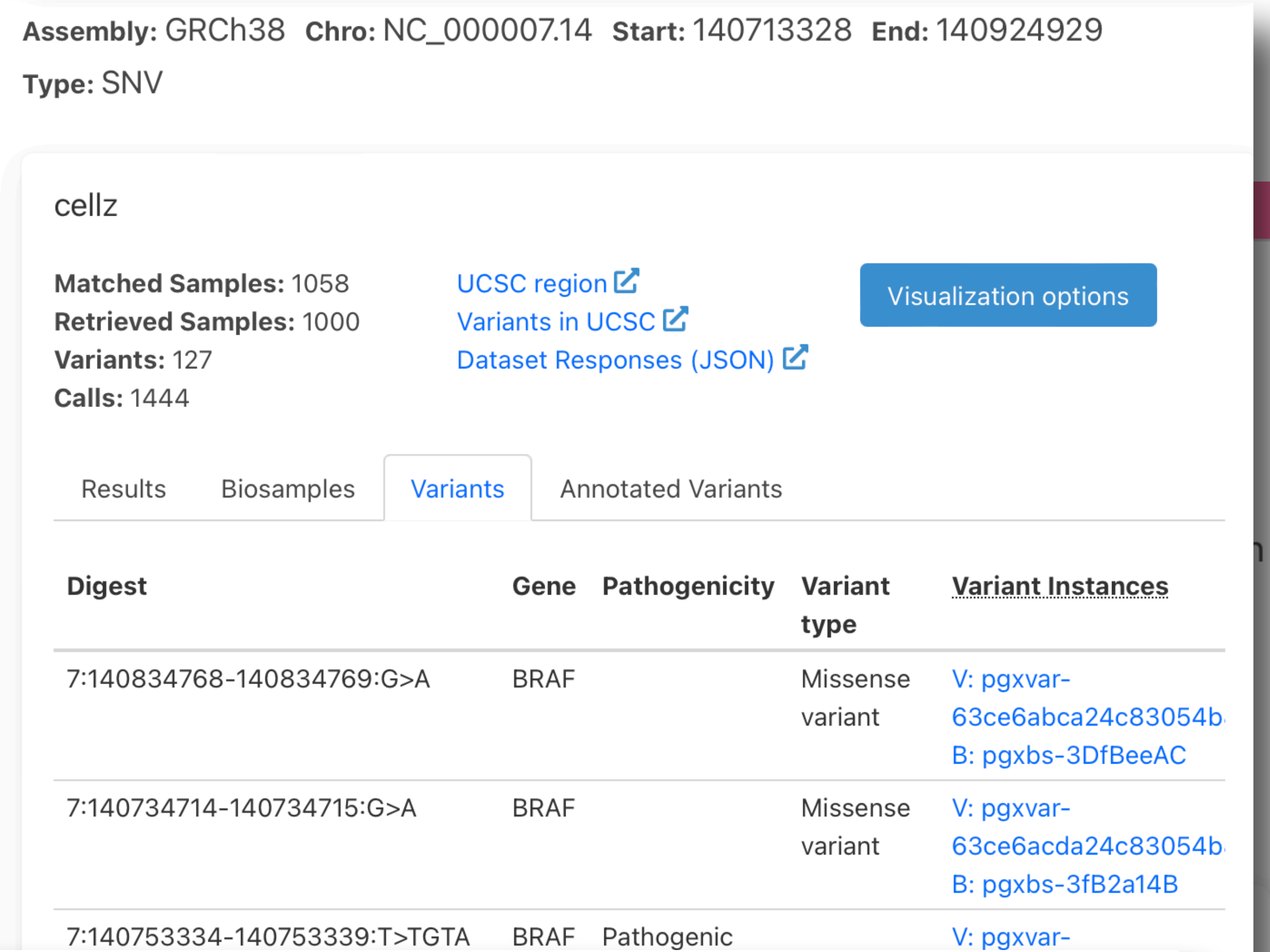
### Cancer Cell Lines by Cellosaurus ID

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

- □ 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 s
  - □ cellosaurus:CVCL_3827: K562/Adr
- □ cellosaurus:CVCL_0589: Kasumi-1 (9

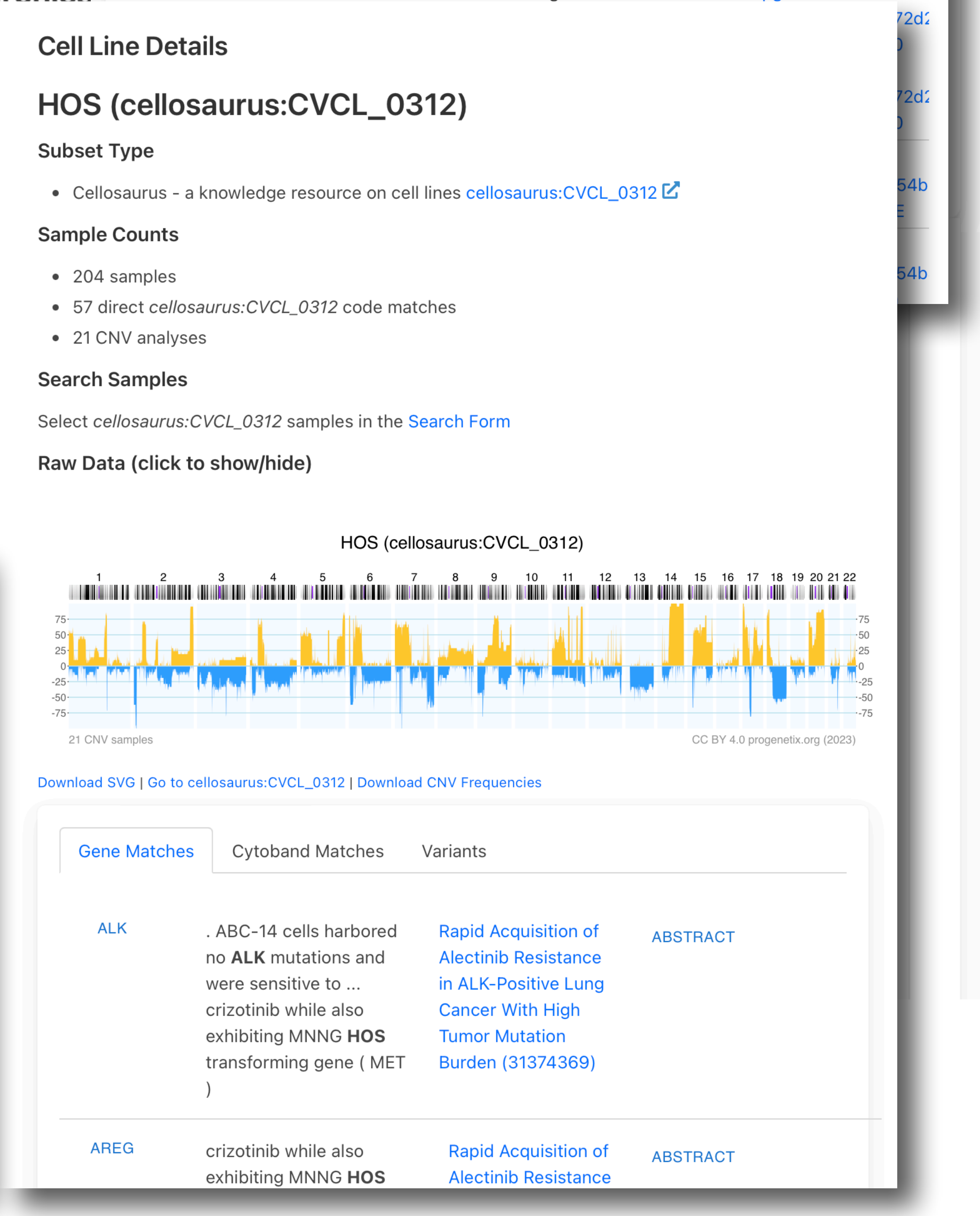


**Assembly:** GRCh38 **Chro:** NC_000007.14 **Start:** 140713328 **End:** 140924929
**Type:** SNV

cellz

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

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



### 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

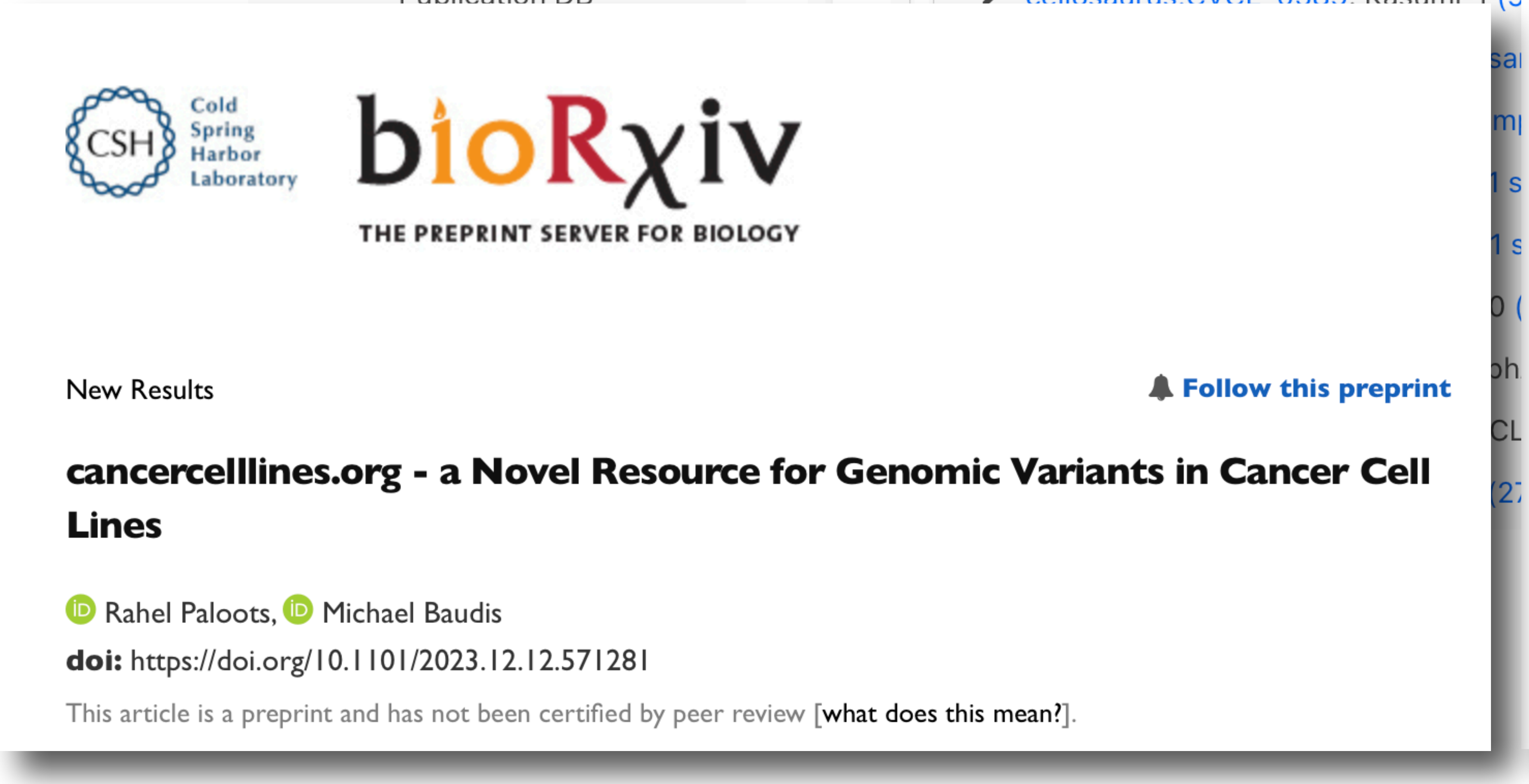
**Search Samples**

Select *cellosaurus:CVCL_0312* samples in the Search Form

**Raw Data (click to show/hide)**

HOS (cellosaurus:CVCL_0312)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
75
50
25
0
-25
-50
-75
21 CNV samples
CC BY 4.0 progenetix.org (2023)
Download SVG | Go to cellosaurus:CVCL_0312 | Download CNV Frequencies
Gene Matches
Cytoband Matches
Variants

| | | | |
|---|---|---|---|
| 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 |


CSH Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY
Follow this preprint

New Results

**cancercelllines.org - a Novel Resource for Genomic Variants in Cancer Cell Lines**

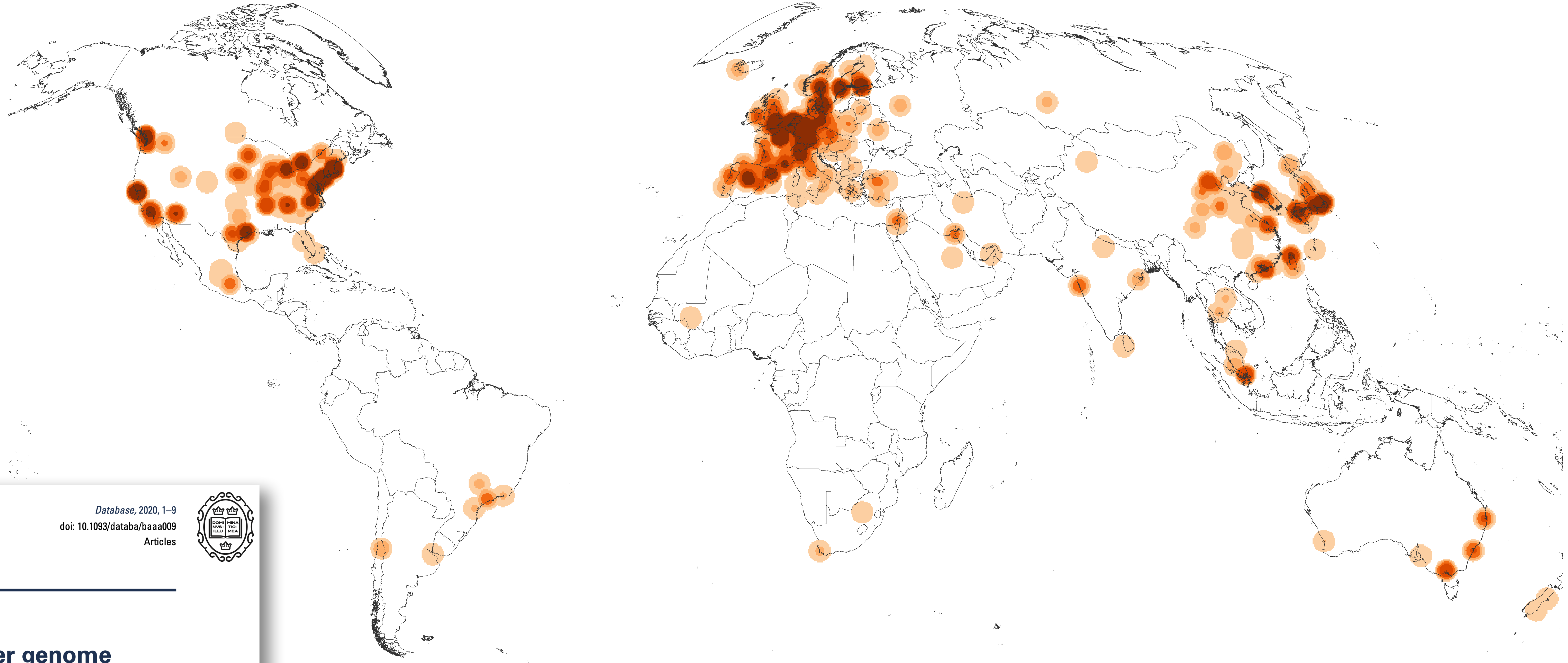
Rahel Paloots, Michael Baudis

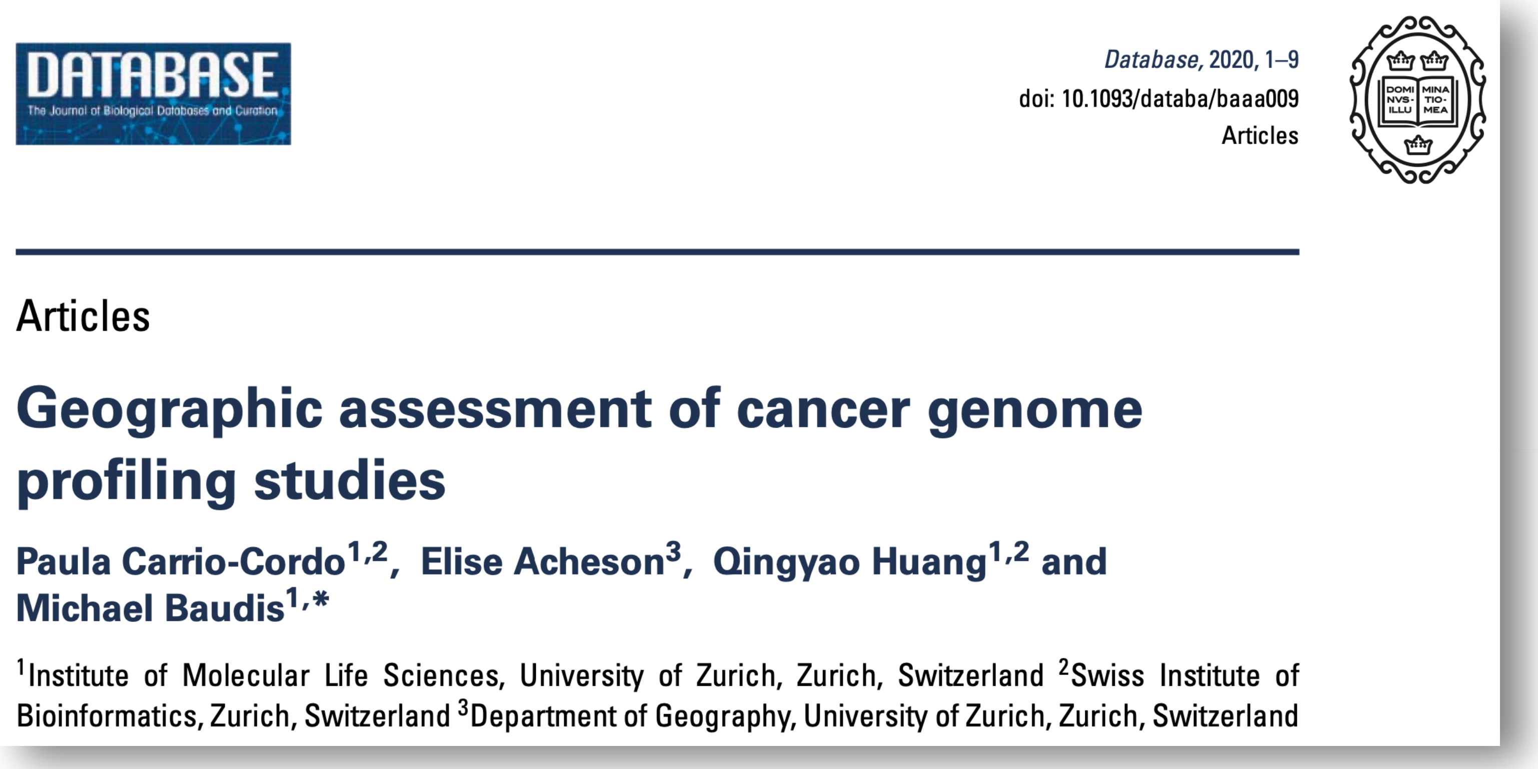
**doi:** https://doi.org/10.1101/2023.12.12.571281

This article is a preprint and has not been certified by peer review [what does this mean?].

# Where does Genomic Data Come From?

## Geographic bias in published cancer genome profiling studies





DATABASE

The Journal of Biological Databases and Curation

*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 Genomic Data Sharing Can...

Demonstrate patterns in health & disease

Increase statistical significance of analyses

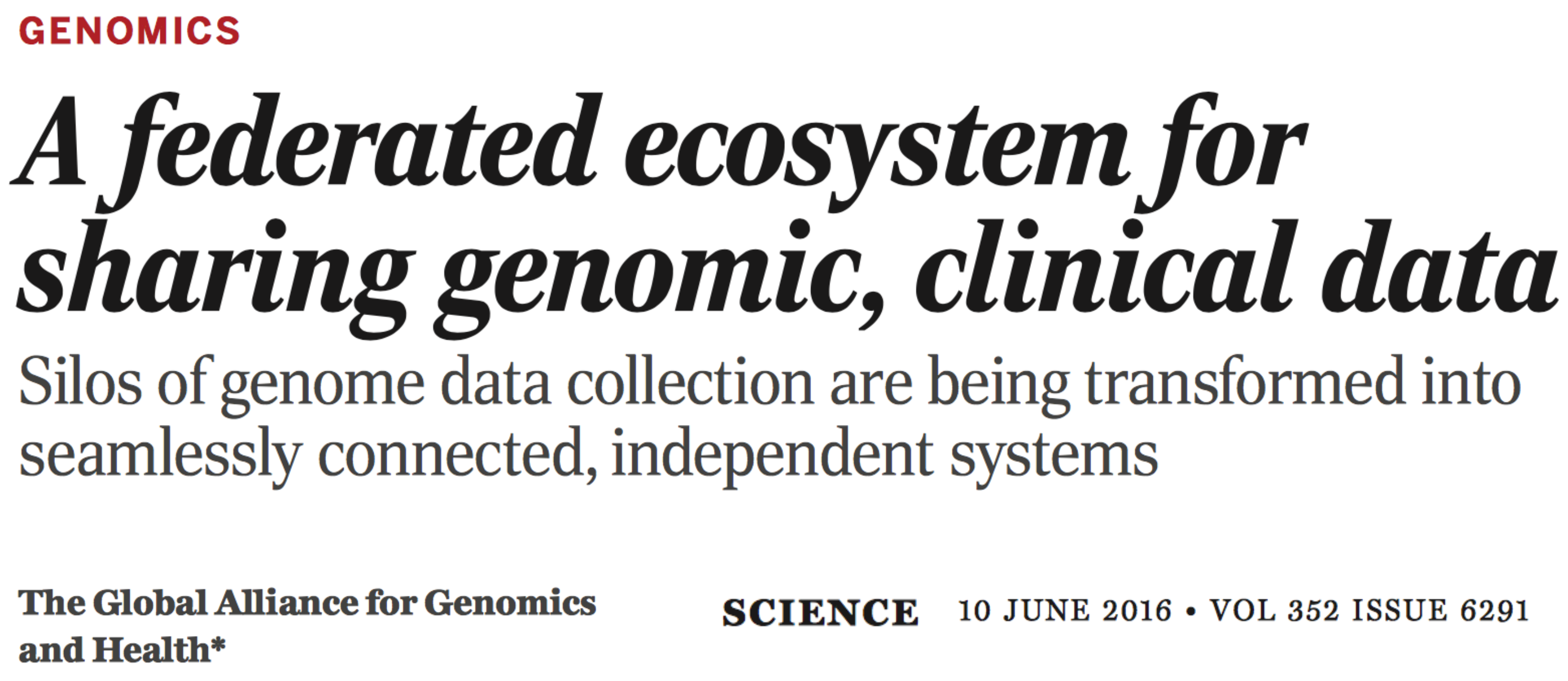
Lead to “stronger” variant interpretations

Increase accurate diagnosis

Advance precision medicine

# 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

# 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

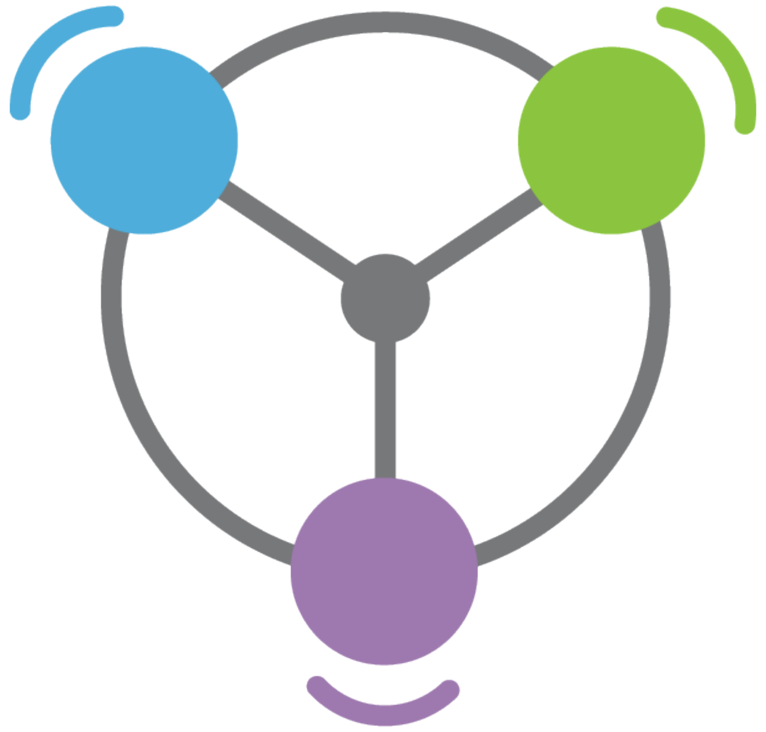
progenetix

**Centralized Genomic Knowledge Bases**

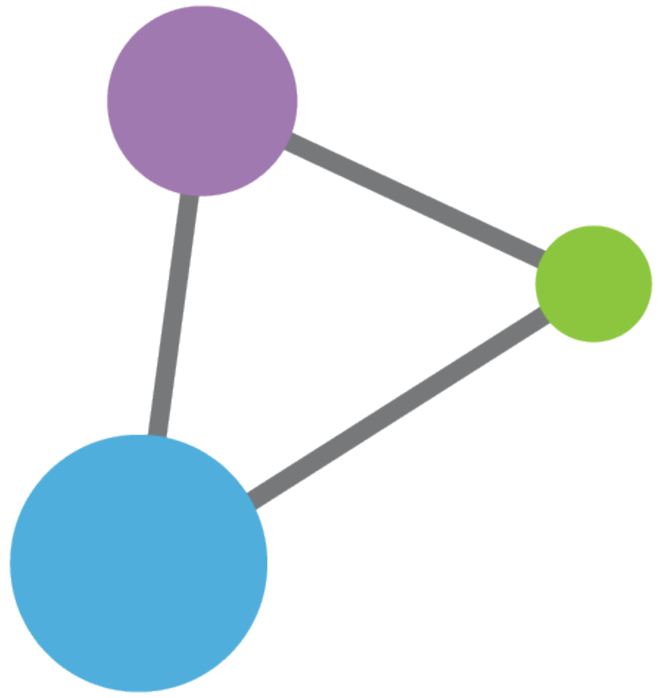
European Genome-Phenome Archive

**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

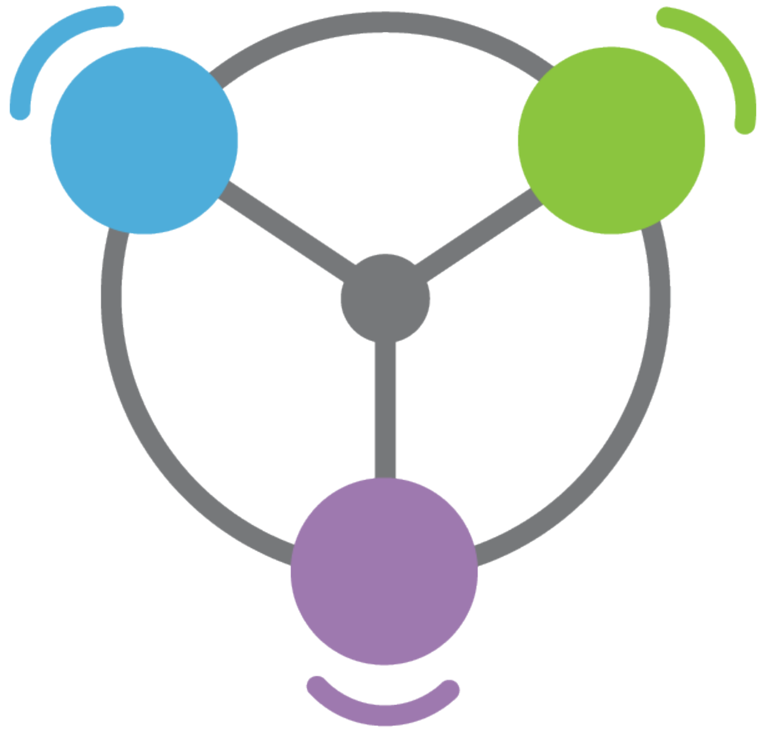
progenetix

**Centralized Genomic Knowledge Bases**

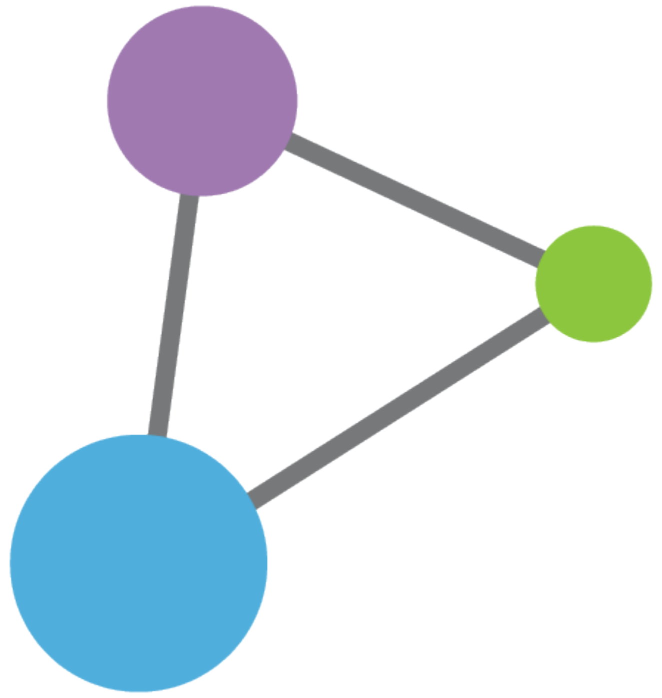
European Genome-Phenome Archive

**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

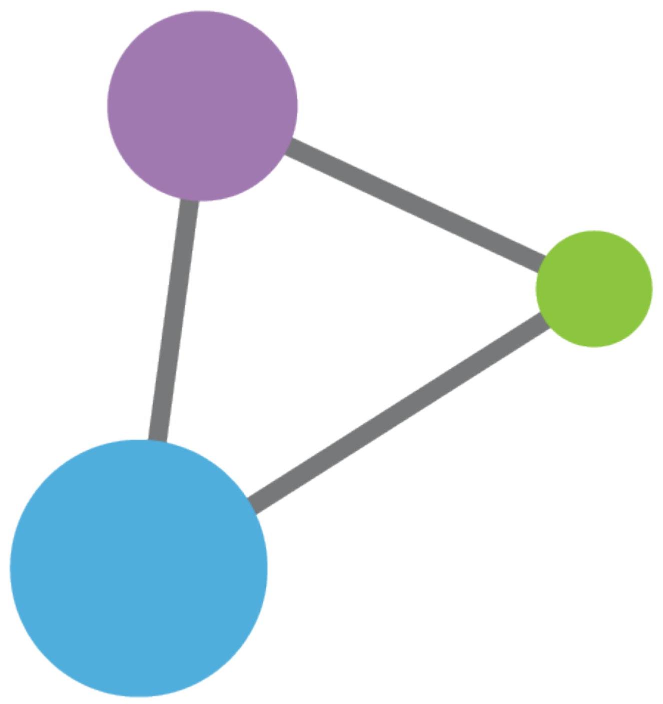
**Centralized Genomic Knowledge Bases**

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

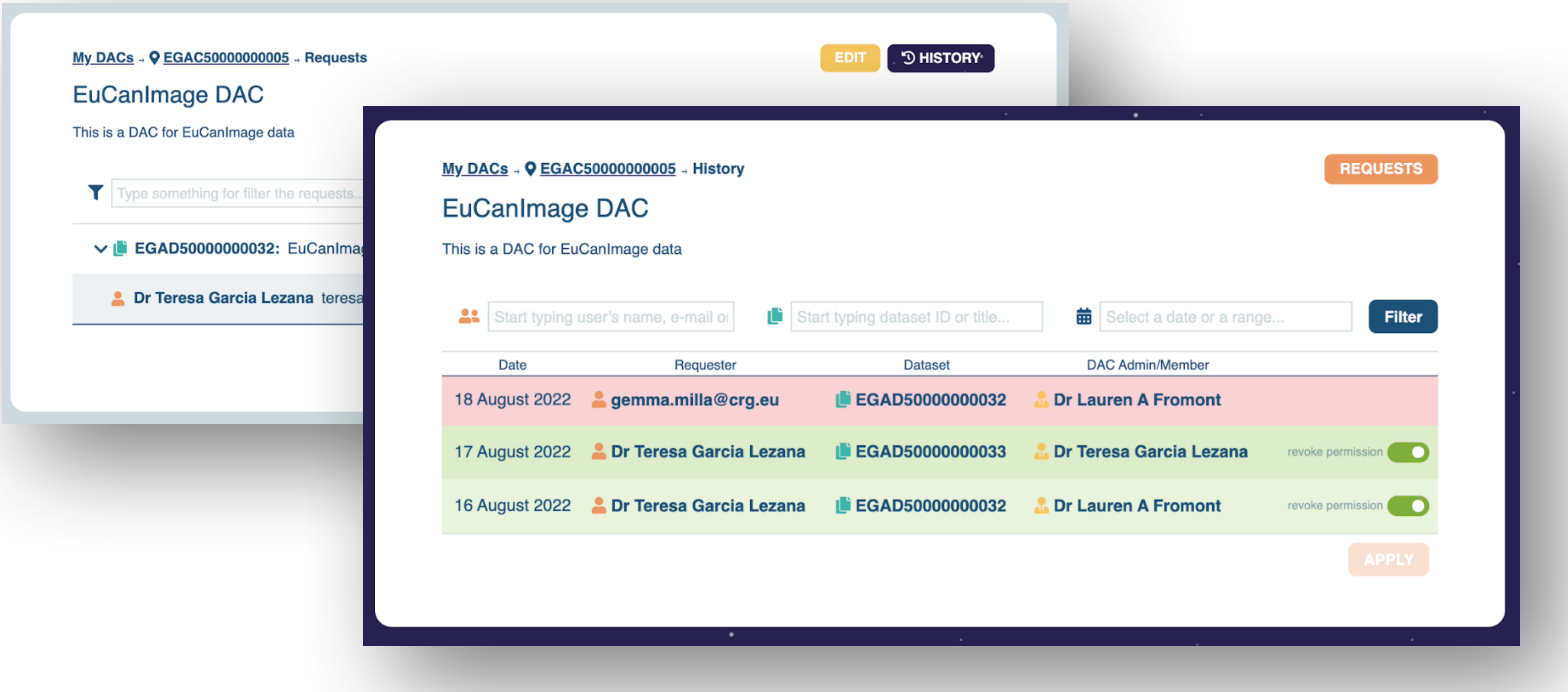
Global Alliance for Genomics & Health DRIVER PROJECT

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



My DACs → EGAC50000000005 → Requests

EuCanImage DAC

This is a DAC for EuCanImage data

Type something for filter the requests...

EGAD50000000032: EuCanImag

Dr Teresa Garcia Lezana teresa

My DACs → EGAC50000000005 → History

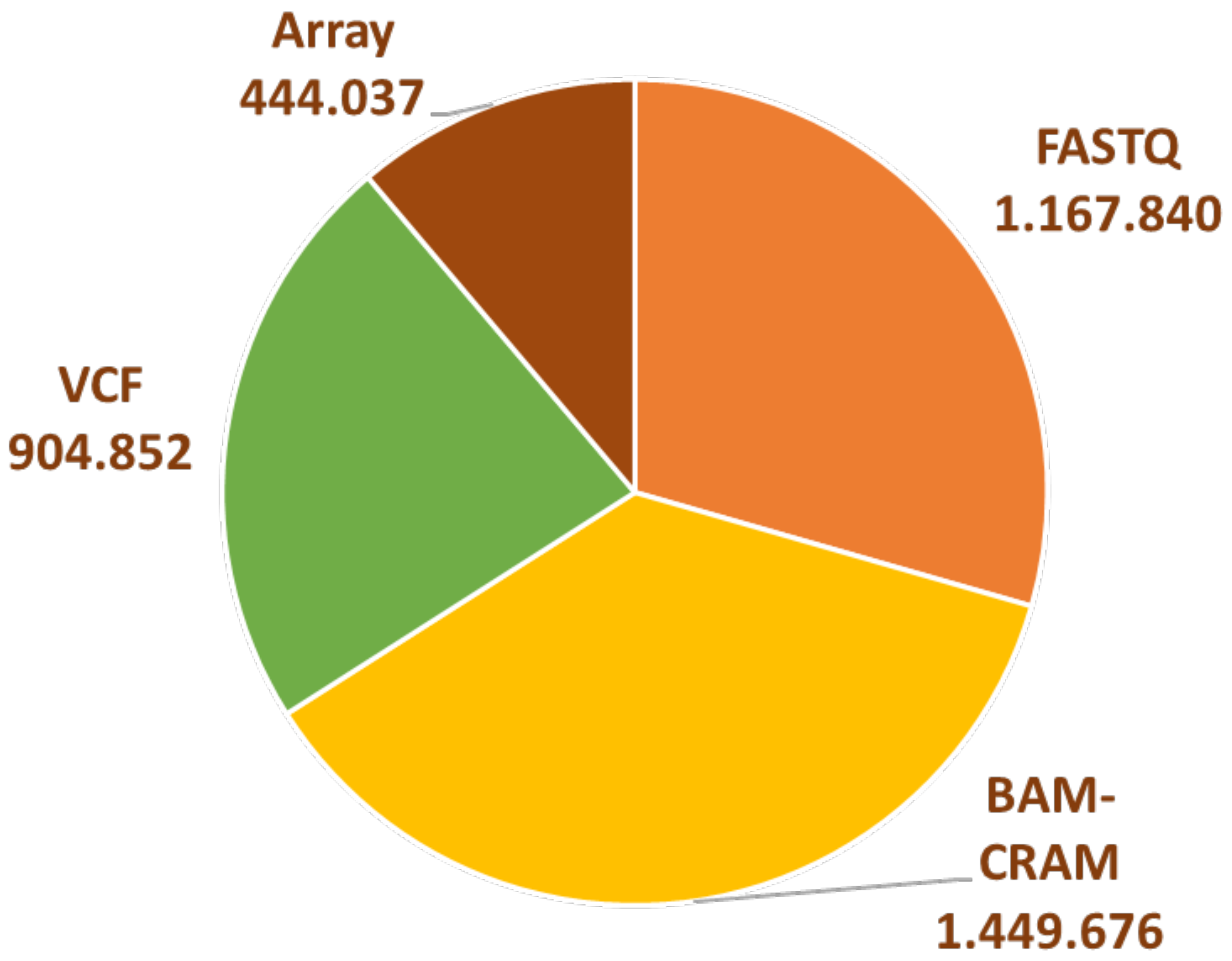
EuCanImage DAC

This is a DAC for EuCanImage data

Start typing user's name, e-mail o | Start typing dataset ID or title... | Select a date or a range...

| 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

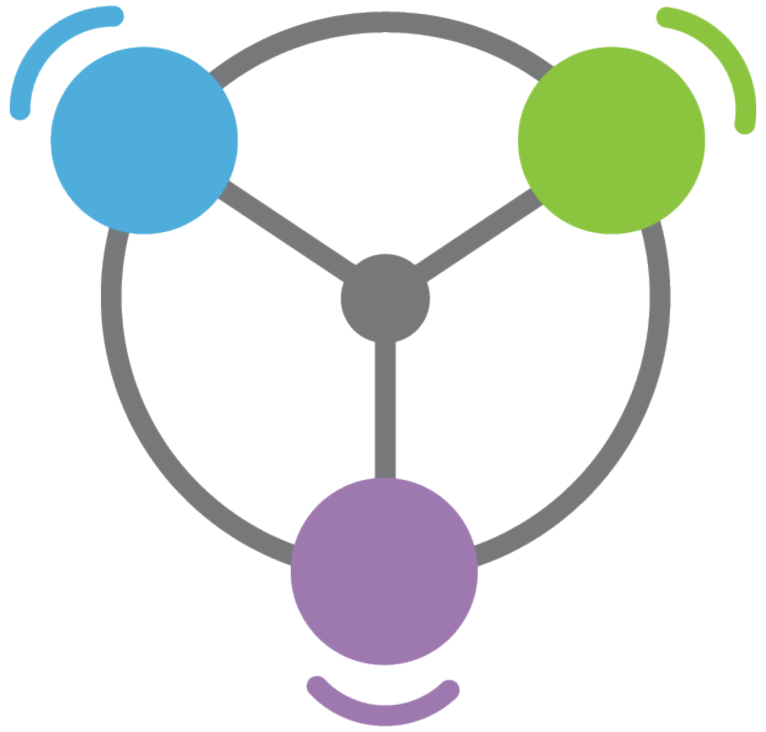
progenetix

**Centralized Genomic Knowledge Bases**

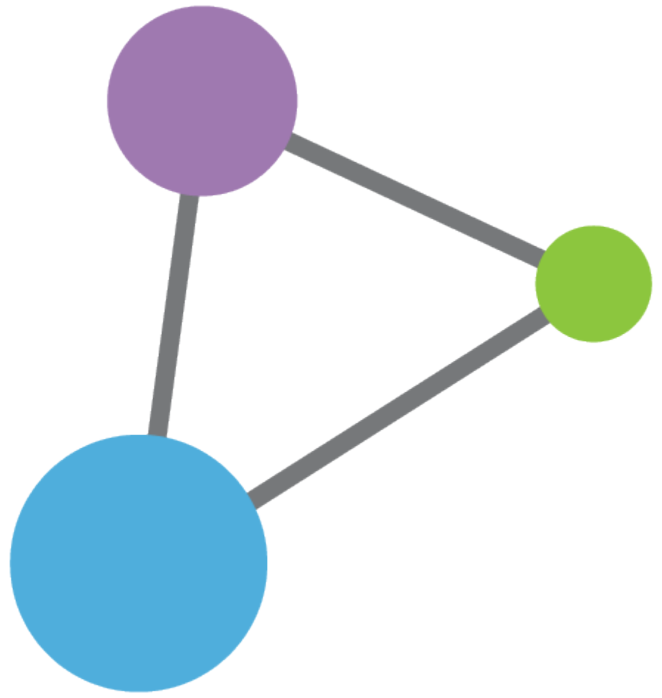
European Genome-Phenome Archive

**Data Commons**
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network



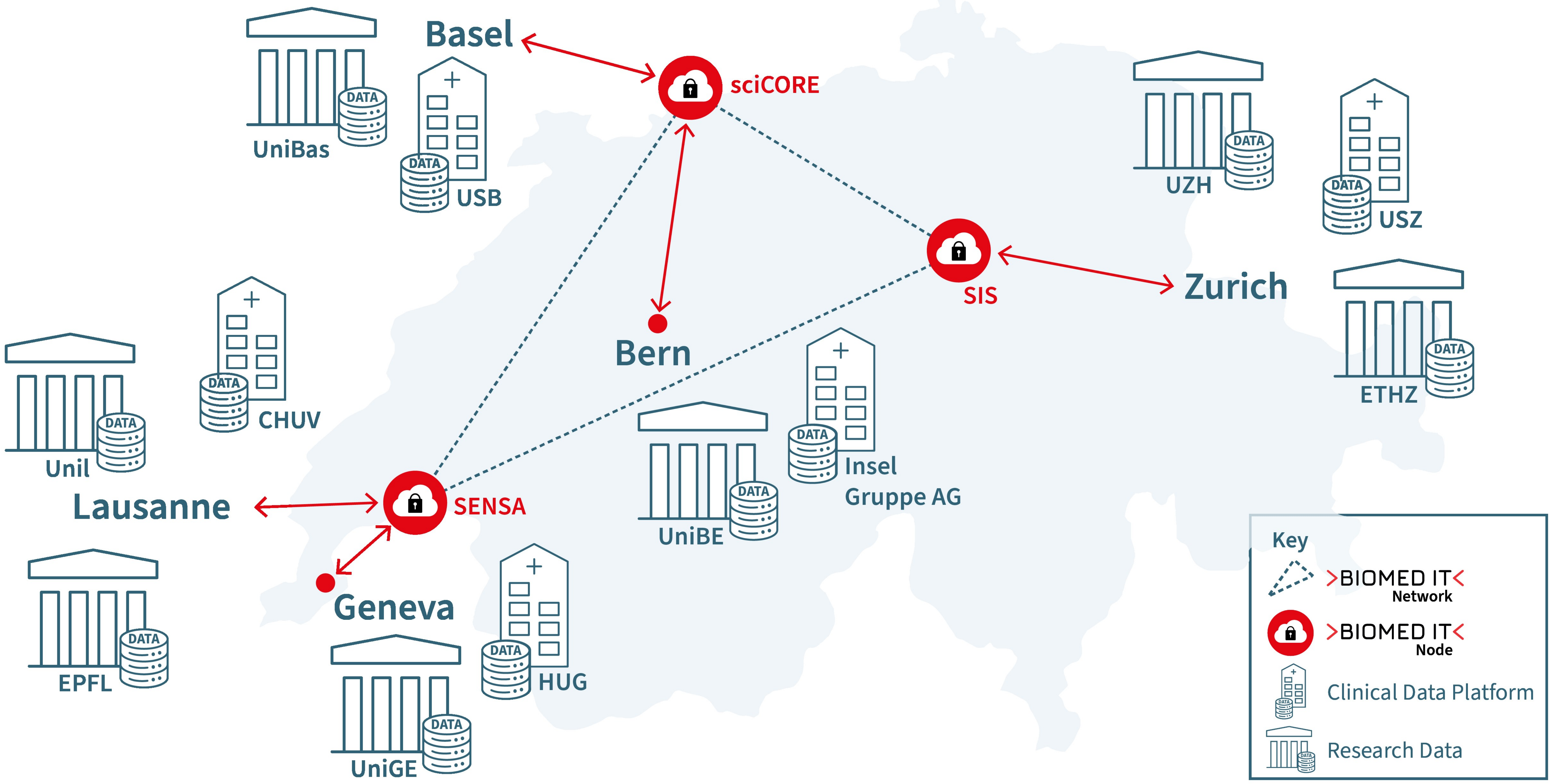
**Hub and Spoke**
Common data elements, access, and usage rules



**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

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

swissuniversities

Universitäre Medizin Schweiz Médecine Universitaire Suisse

**SIB Personalized Health Informatics Group**

SPHN Data Coordination Center (DCC)
BioMedIT Network

# Different Approaches to Data Sharing

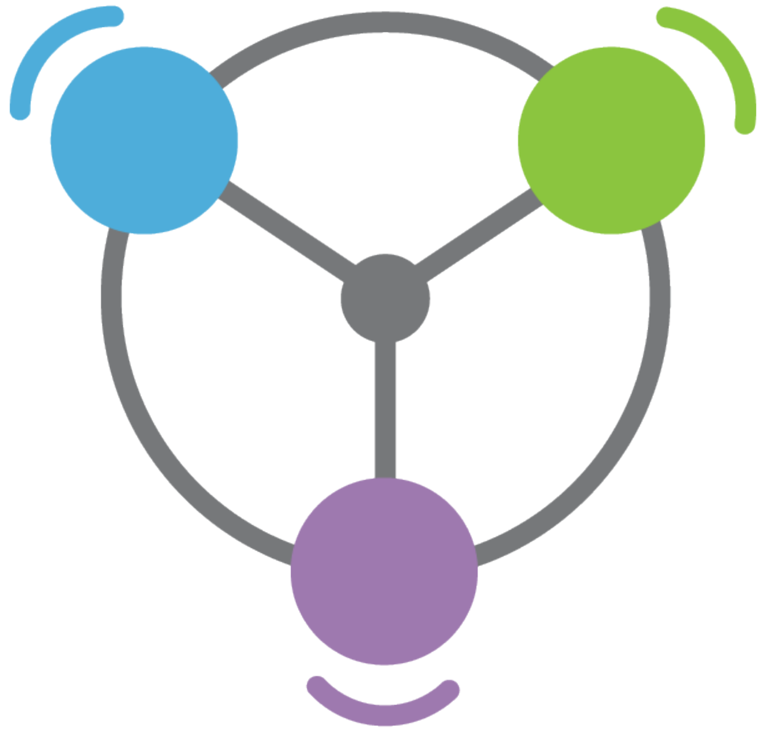
progenetix

**Centralized Genomic Knowledge Bases**

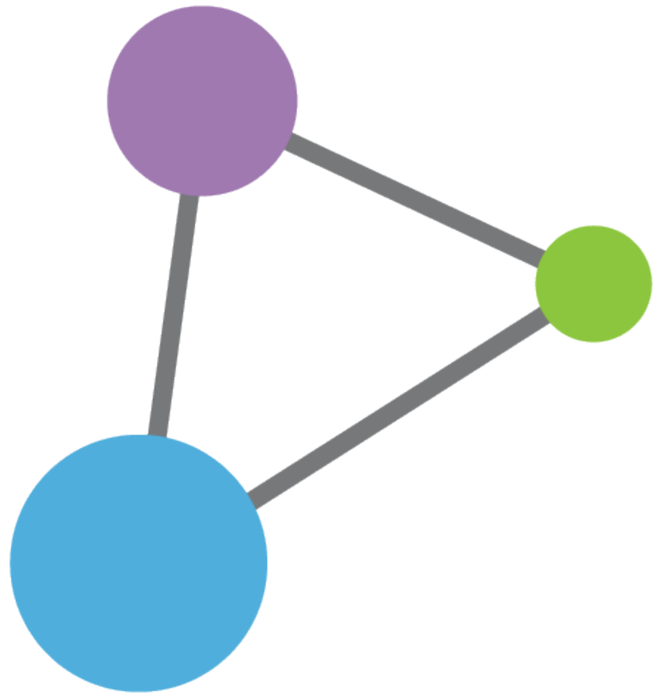
European Genome-Phenome Archive

**Data Commons**
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network



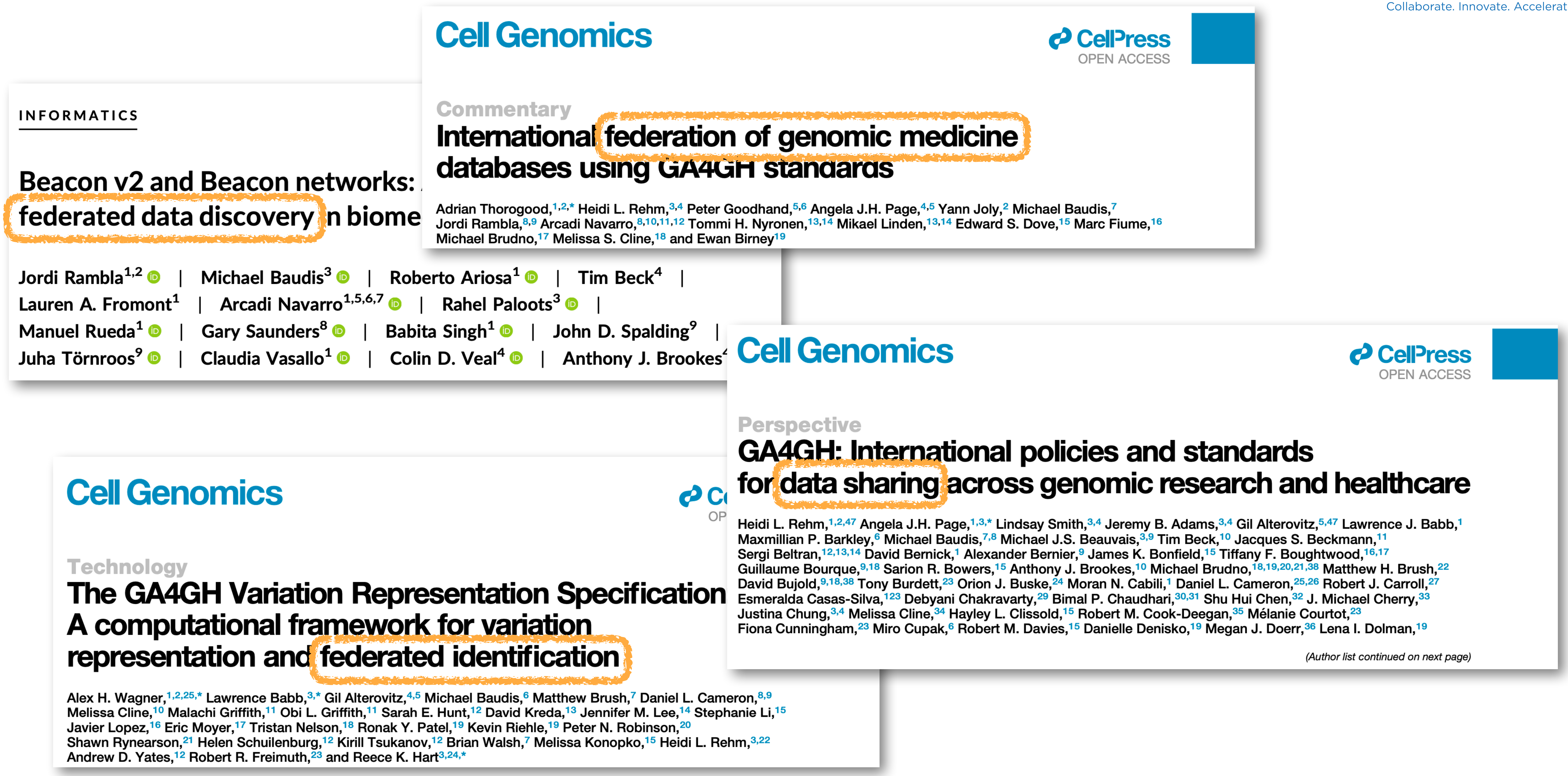
**Hub and Spoke**
Common data elements, access, and usage rules



**Linkage of distributed and disparate datasets**

**Federation**

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

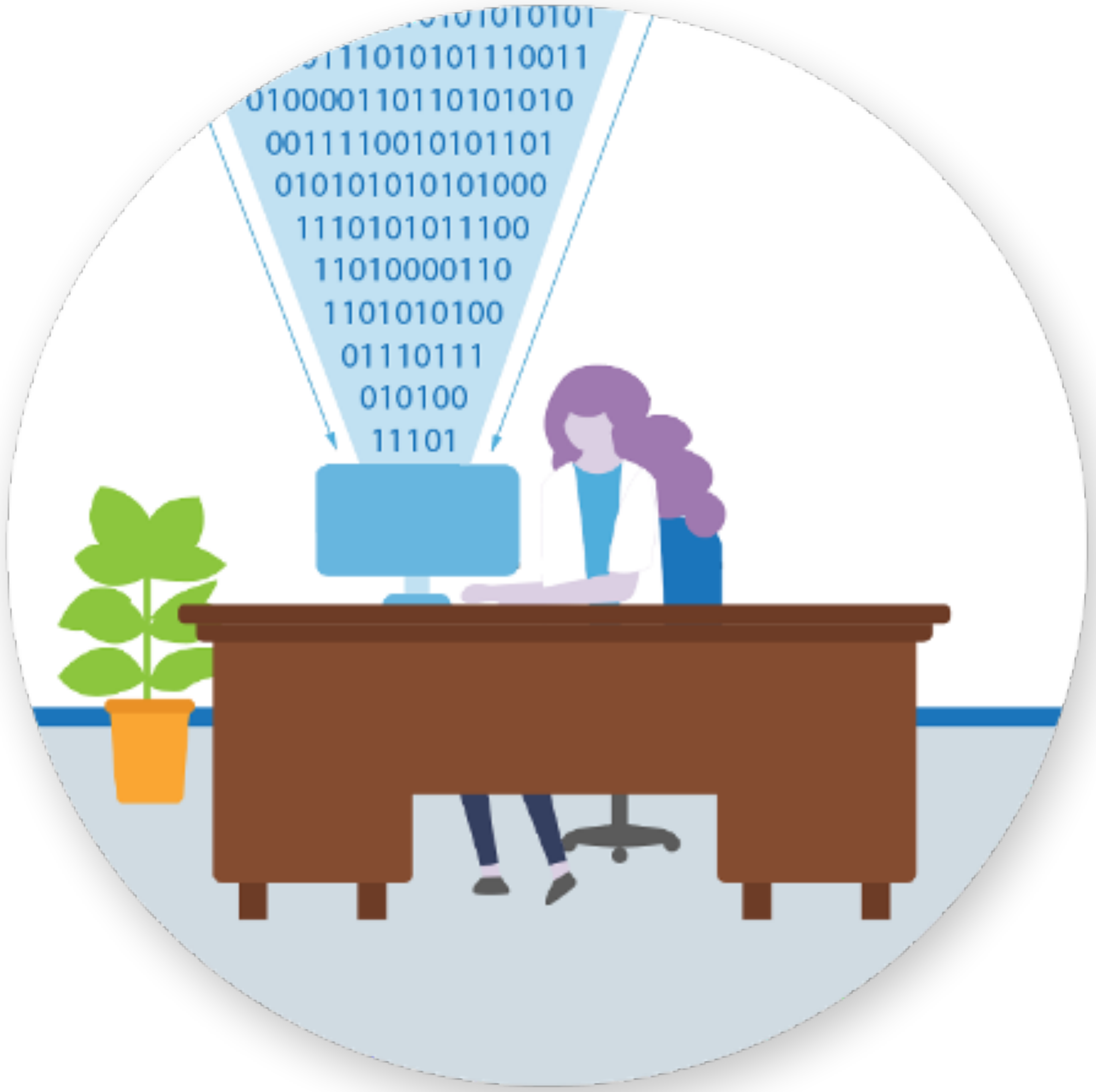
Technology

## 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,*]

# A New Paradigm for Data Sharing

FROM



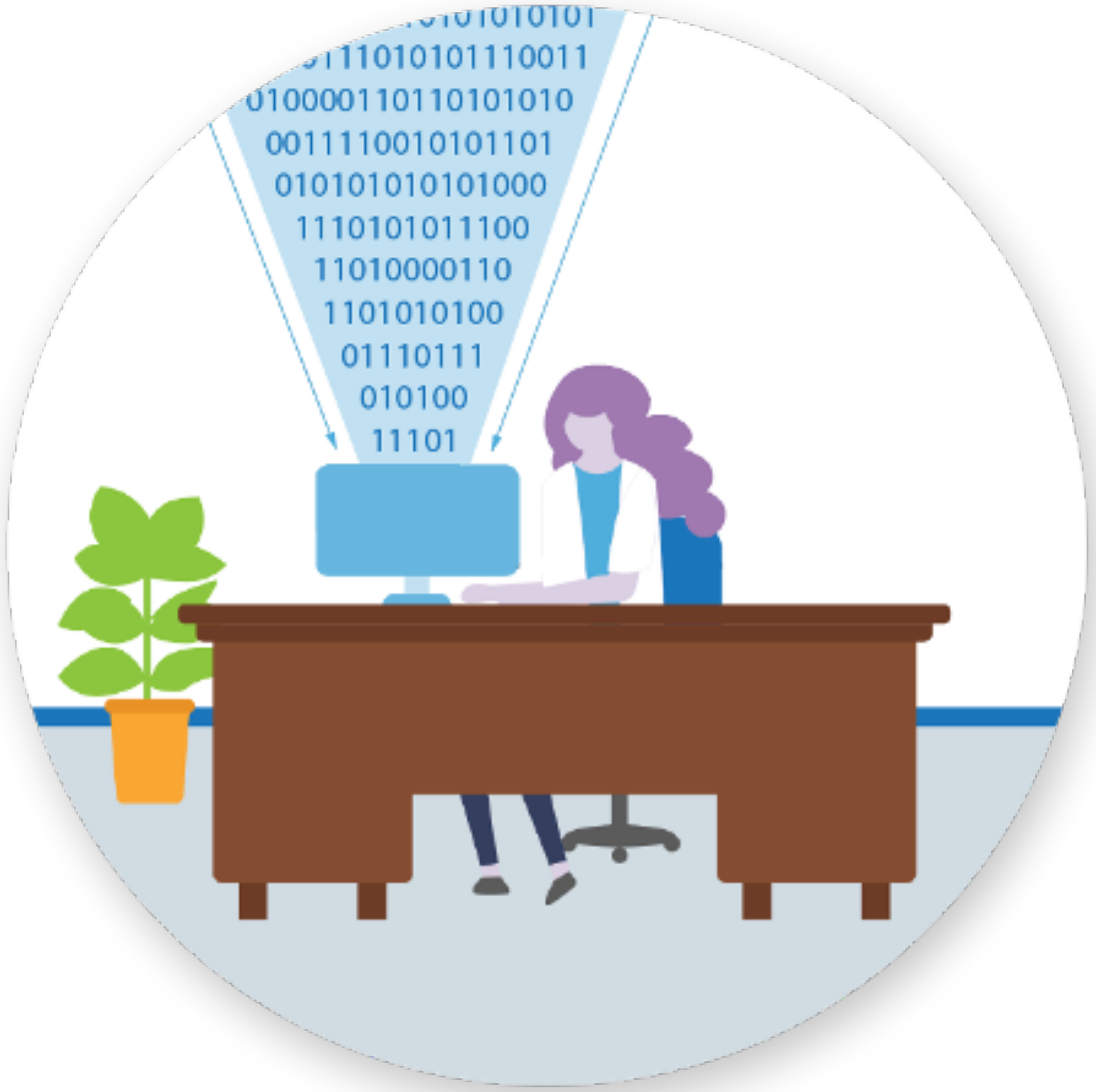
Data Copying

TO

Data Visiting

# A New Paradigm for Data Sharing

FROM

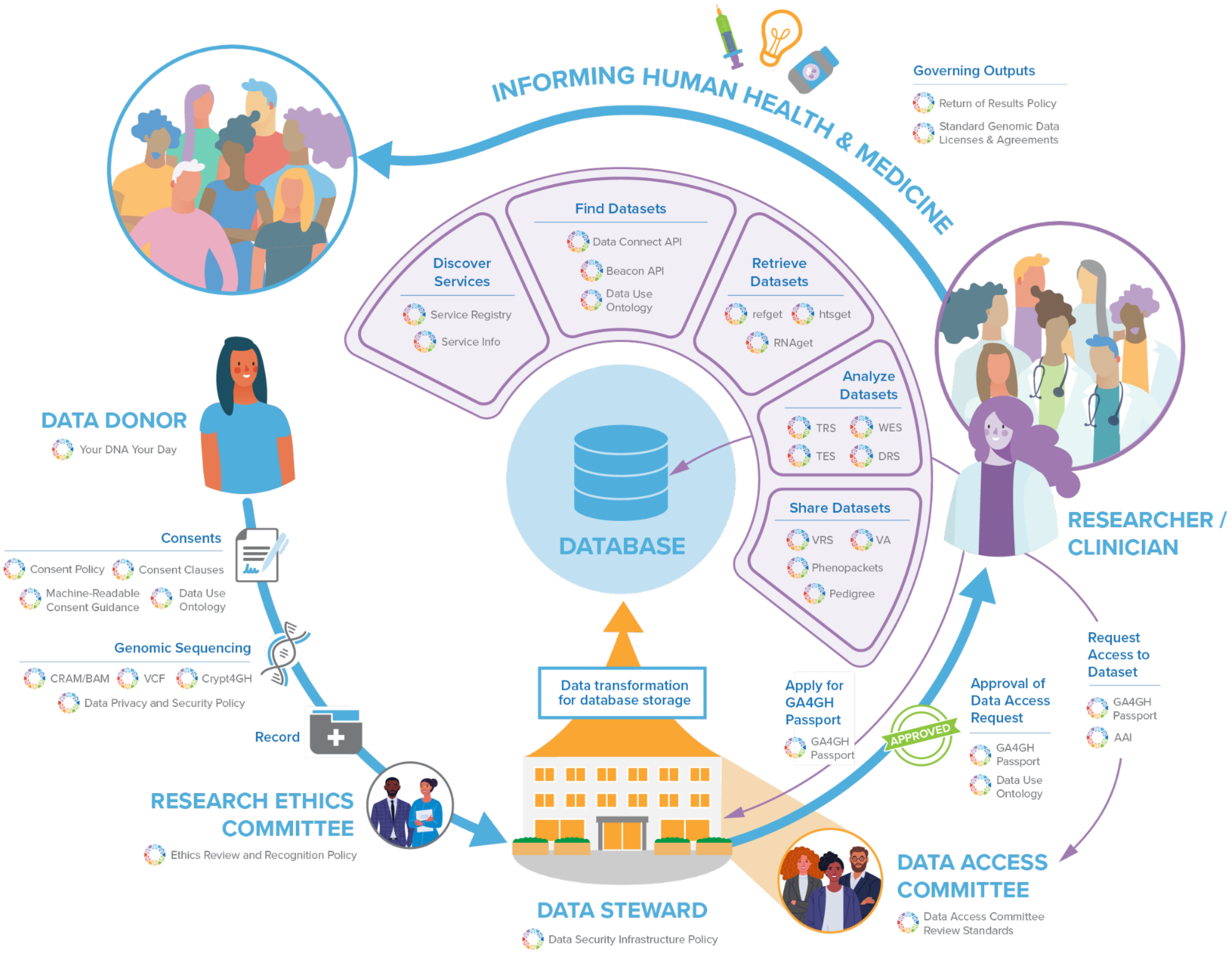


**Data Copying**

STANDARDS →

TO

**Data Visiting**



INFORMING HUMAN HEALTH & MEDICINE

Governing Outputs

- Return of Results Policy
- Standard Genomic Data Licenses & Agreements

DATA DONOR

- Your DNA Your Day

Consents

- Consent Policy
- Consent Clauses
- Machine-Readable Consent Guidance
- Data Use Ontology

Genomic Sequencing

- CRAM/BAM
- VCF
- Crypt4GH
- Data Privacy and Security Policy

Record

RESEARCH ETHICS COMMITTEE

- Ethics Review and Recognition Policy

Discover Services

- Service Registry
- Service Info

Find Datasets

- Data Connect API
- Beacon API
- Data Use Ontology

Retrieve Datasets

- refget
- htsget
- RNAget

Analyze Datasets

- TRS
- WES
- TES
- DRS

Share Datasets

- VRS
- VA
- Phenopackets
- Pedigree

DATABASE

Data transformation for database storage

DATA STEWARD

- Data Security Infrastructure Policy

RESEARCHER / CLINICIAN

Request Access to Dataset

- GA4GH Passport
- AAI

Approval of Data Access Request

- GA4GH Passport
- Data Use Ontology

APPROVED

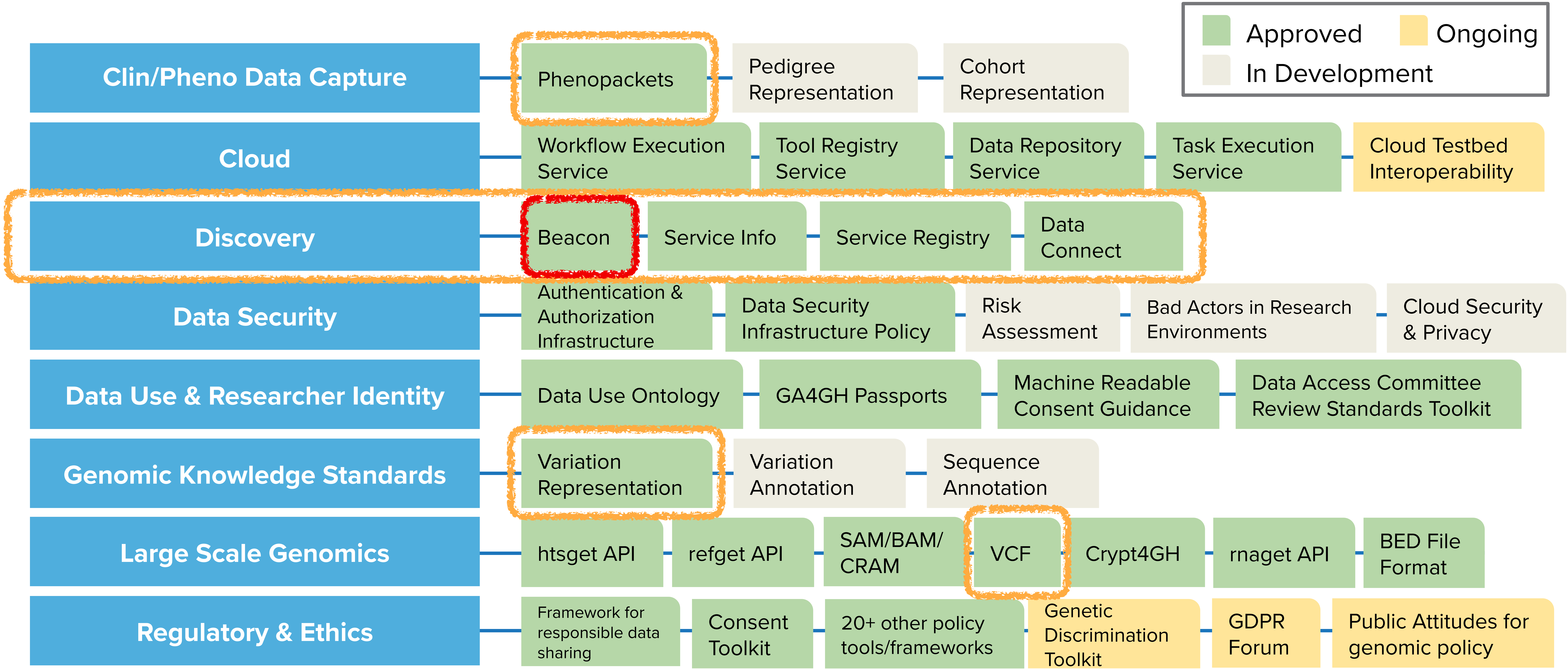
Apply for GA4GH Passport

- GA4GH Passport

DATA ACCESS COMMITTEE

- Data Access Committee Review Standards

# Overview of GA4GH standards and frameworks



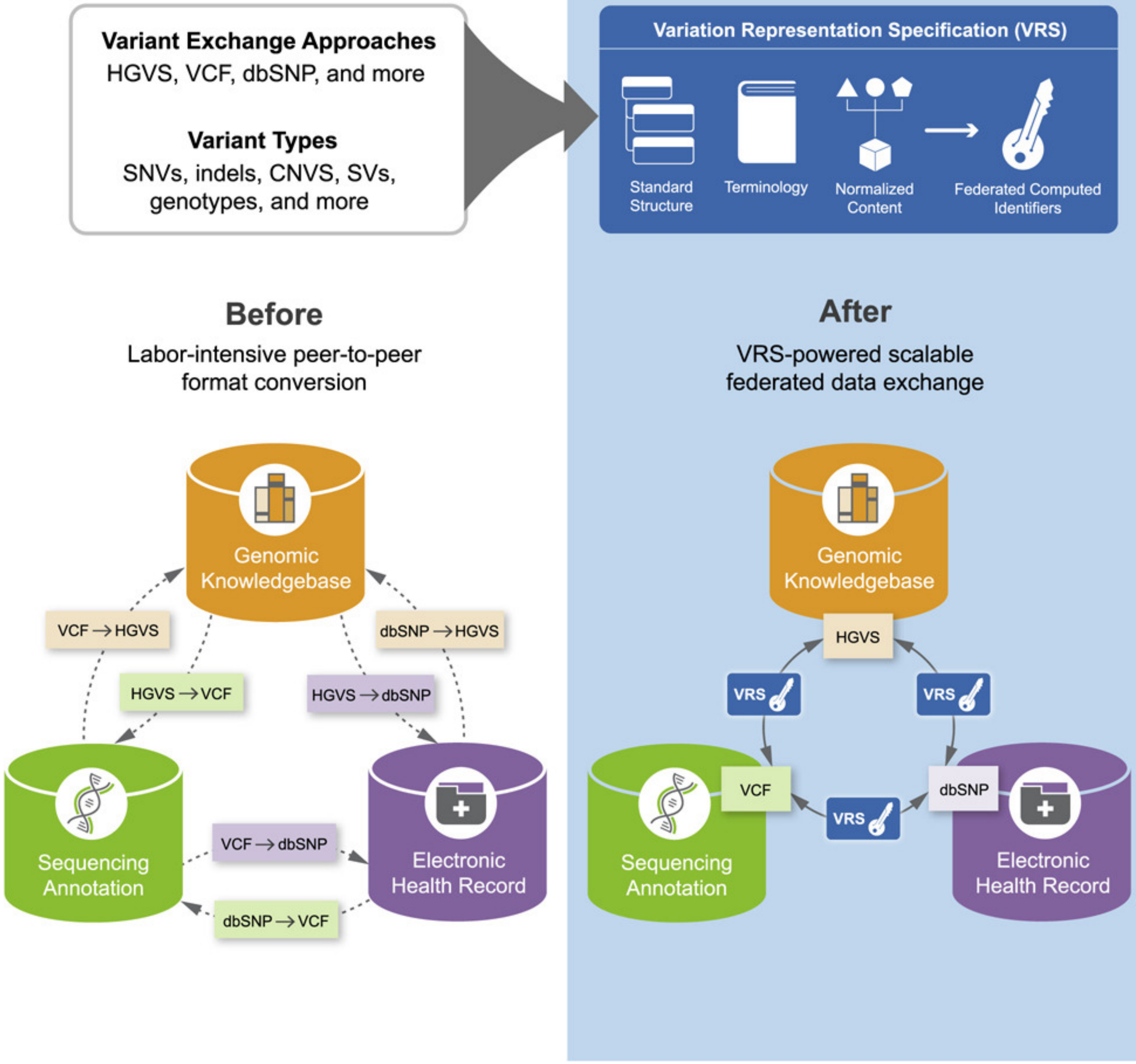
🟩 Approved 🟨 Ongoing ⬜ In Development

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 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 | |

# GA4GH VRS

## Bringing consistency to genomic variation representation

- The GA4GH Variation Representation Specification ("VRS"):
  - ➡ … is a computational framework for representing biomolecular variation
  - ➡ … enables computable identification of variation supporting federated data exchange
  - ➡ … continues to evolve as an open-source, community-driven standard of the GA4GH



https://www.cell.com/cell-genomics/fulltext/S2666-979X(21)00034-3

# Phenopackets v2

Phenopackets is a standard schema for sharing phenotypic information.

**Approved:** June 24, 2021

## Example Users

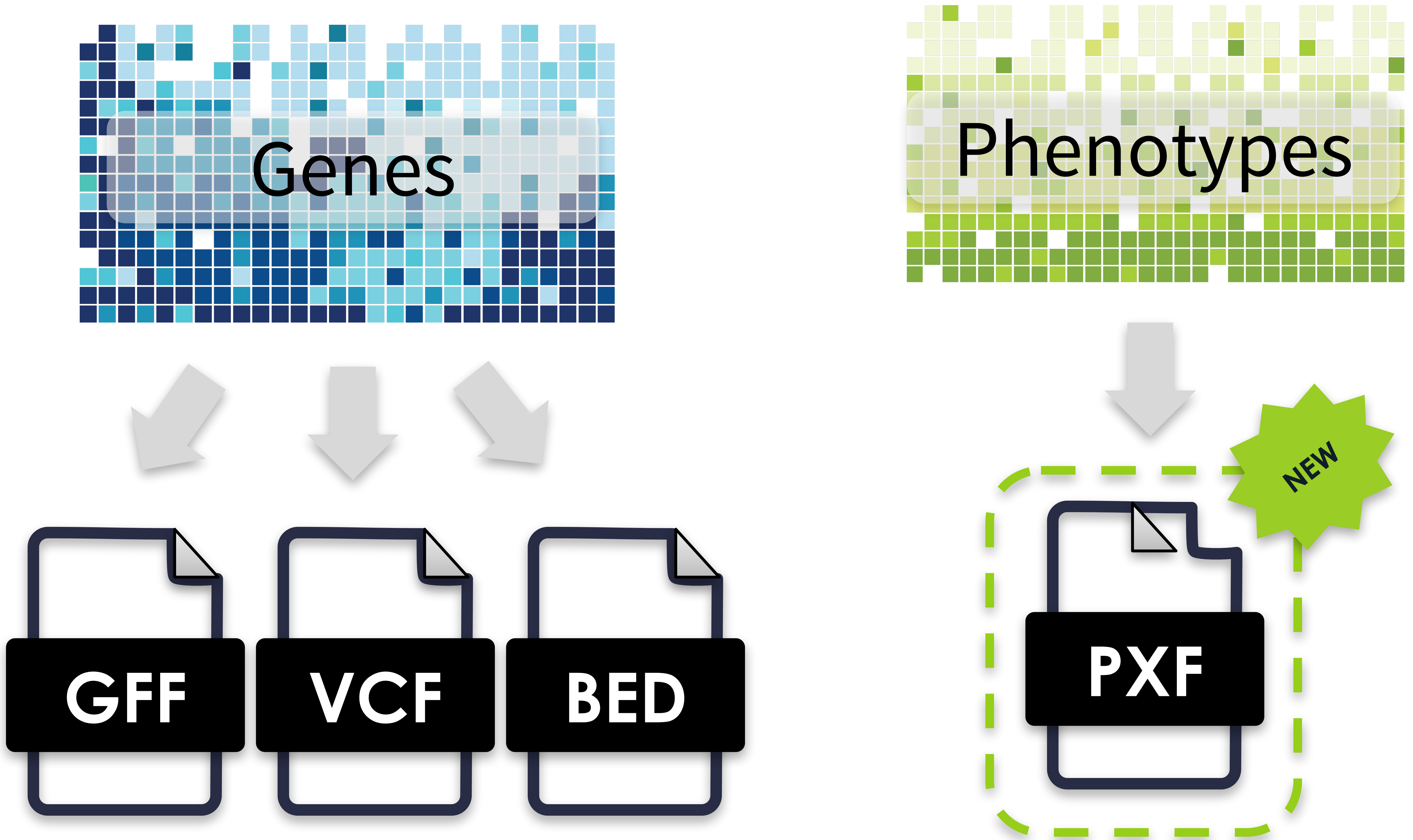
Cafe Variome

NIH

AMED

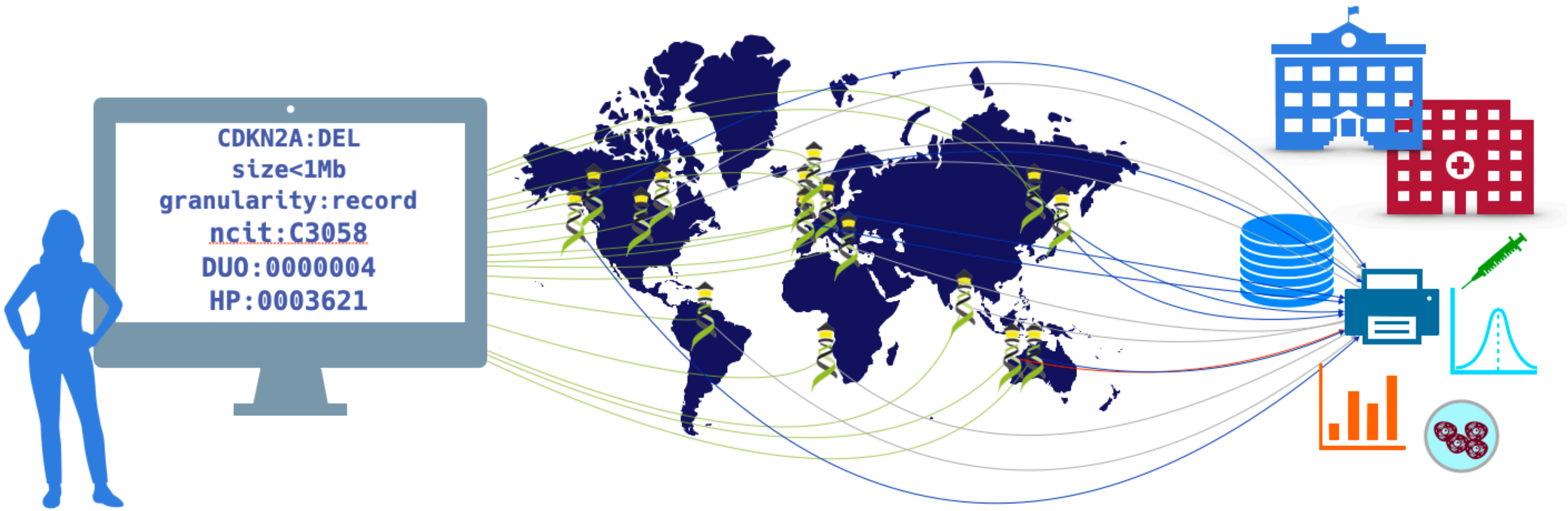
PHENOTIPS™

RD Connect



Global Alliance
for Genomics & Health

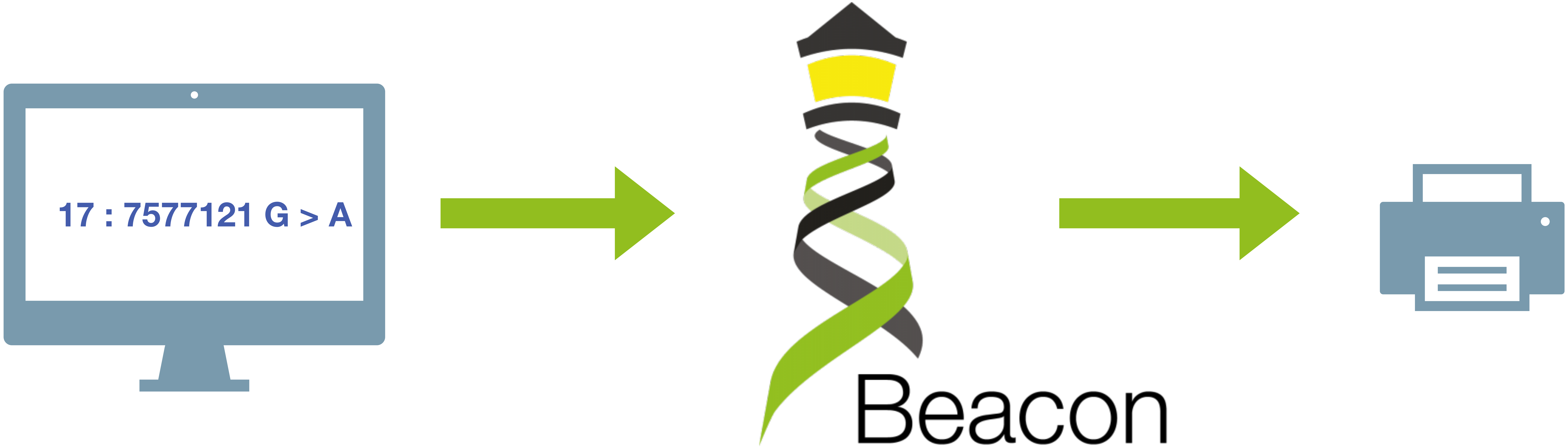
Collaborate. Innovate. Accelerate.



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

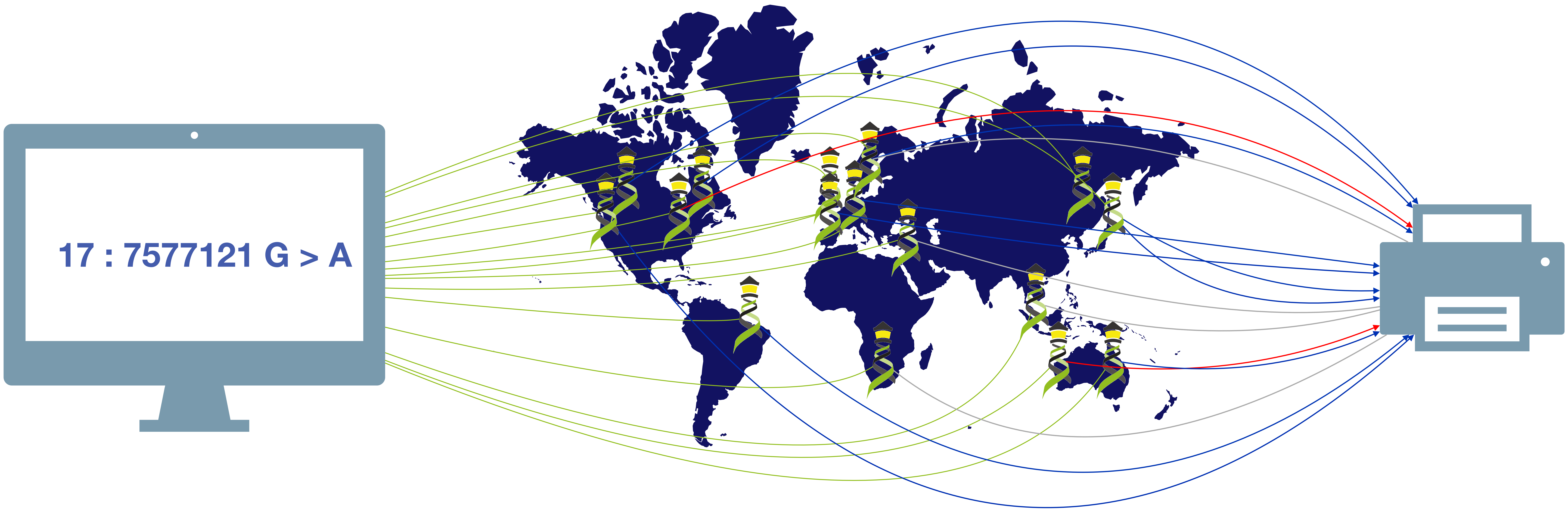
# The GA4GH Beacon Protocol

## Federating Genomic Discoveries



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

**YES** | **NO** | **\0**



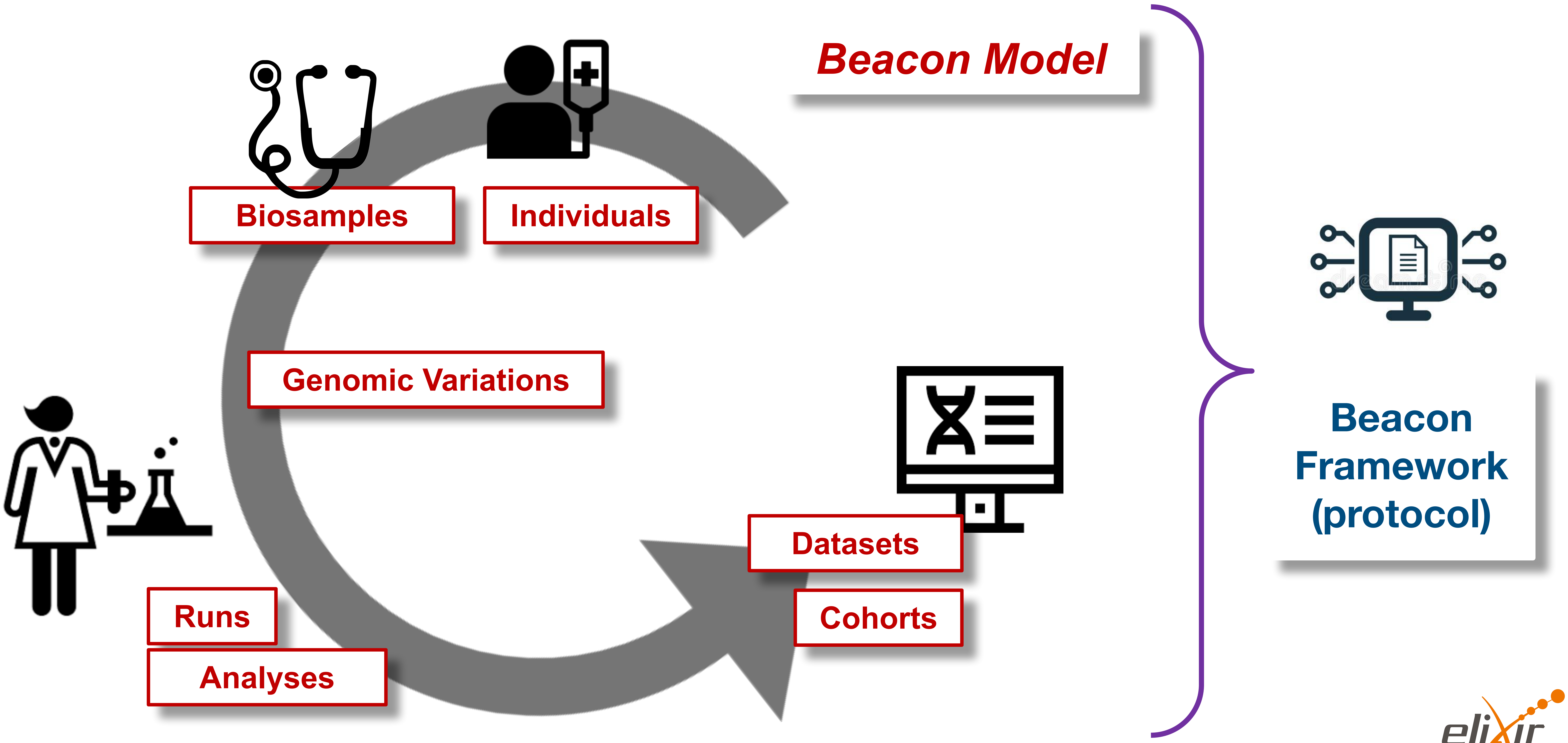
Have you seen this variant? It came up in my patient and we don't know if this is a common SNP or worth following up.

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



*Beacon Model*

Biosamples

Individuals

Genomic Variations

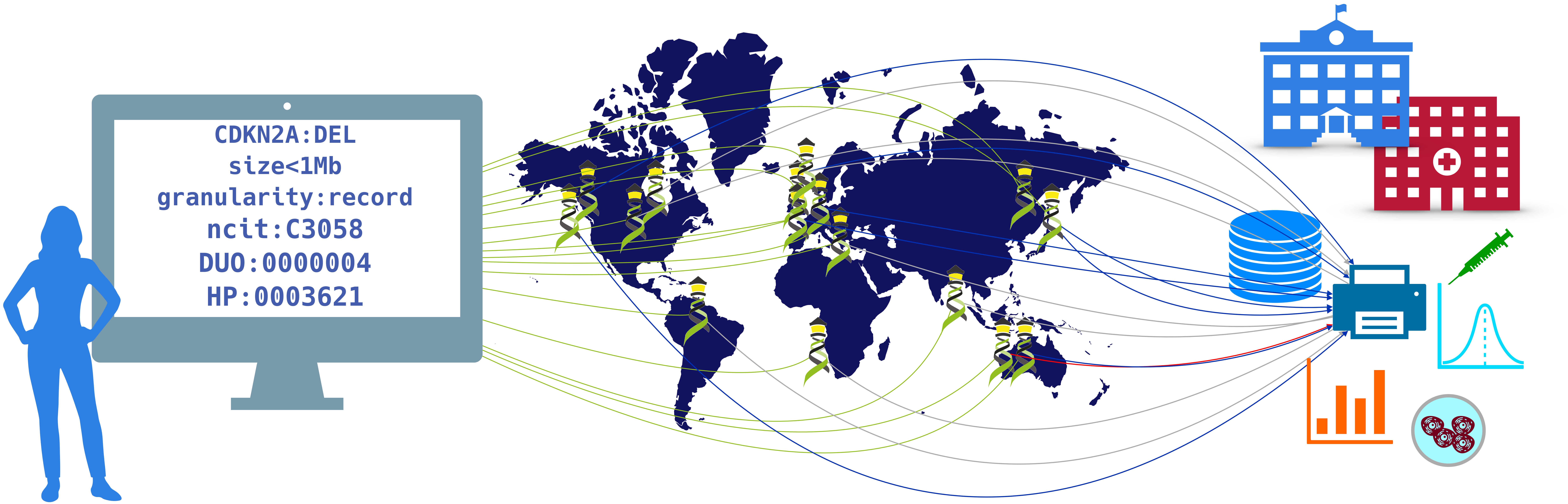
Runs

Analyses

Datasets

Cohorts

Beacon Framework (protocol)



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

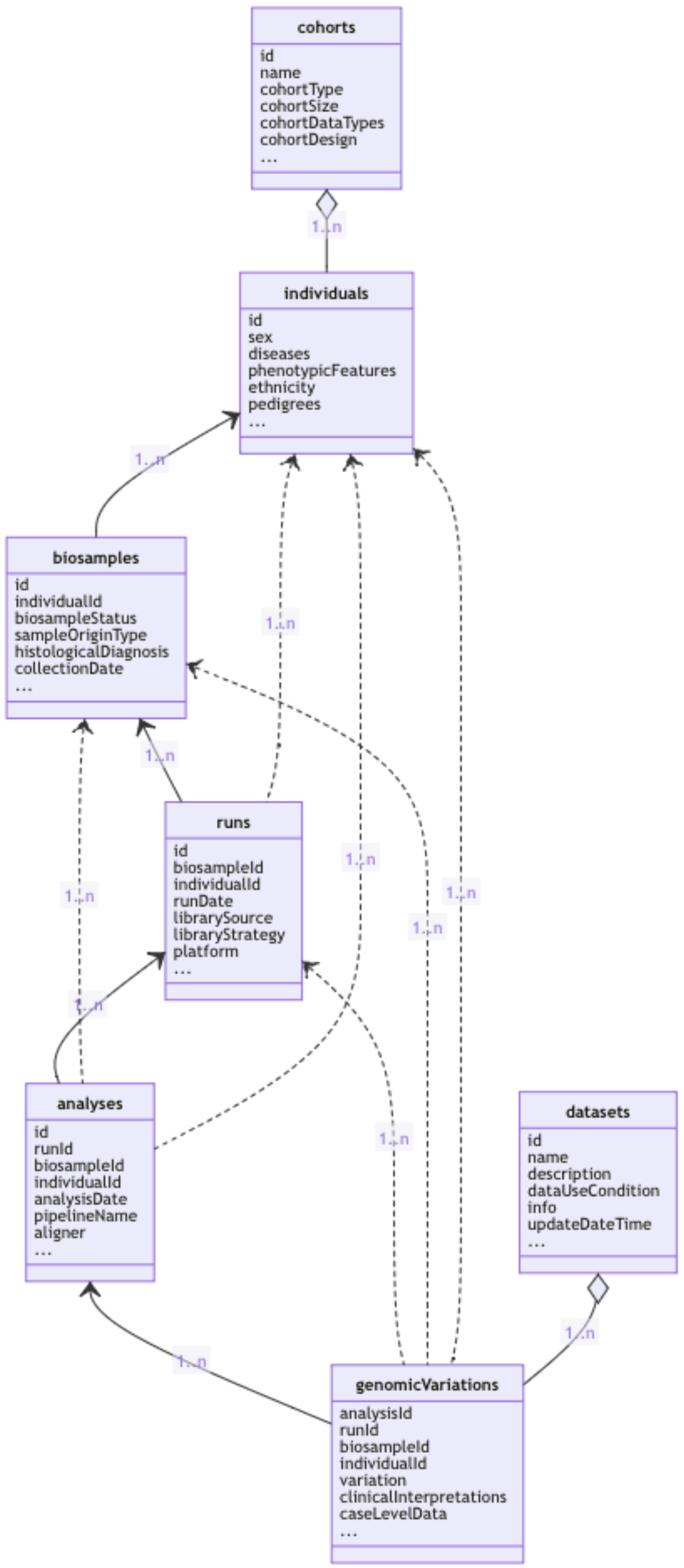
Can you provide data about focal deletions in CDKN2A in Glioblastomas from juvenile patients with unrestricted access?

# Beacon *v2* API

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

# Beacon Default v2 Model

- The Beacon ***framework*** describes the overall structure of the API requests, responses, parameters, the common components, etc.
- Beacon ***models*** describe the set of concepts included in a Beacon, like individual or biosample, and also the relationships between them.
- Besides logical concepts, the Beacon ***models*** represent the schemas for data delivery in “record” granularity
- Beacon explicitly allows the use of *other models* besides its *version specific default*.
- Adherence to a shared ***model*** empowers federation
- Use of the ***framework*** w/ different models extends adoption



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

# Request Components

## Deparsing the Beacon v2 Example

- query against genomic variations, no matter how they are stored
- copy number deletion, as indicated through the VCF symbolic allele `DEL` expression
- a combination of `genId` (server side gene data) OR
- a range query and `variantMaxLength`, or positional (`start`, `end`)
- a filter for the Glioblastoma diagnosis, as NCIT term NCIT:C3058
- as an HPO term for "juvenile" HP:0003621
- full data access as per DUO:0000004

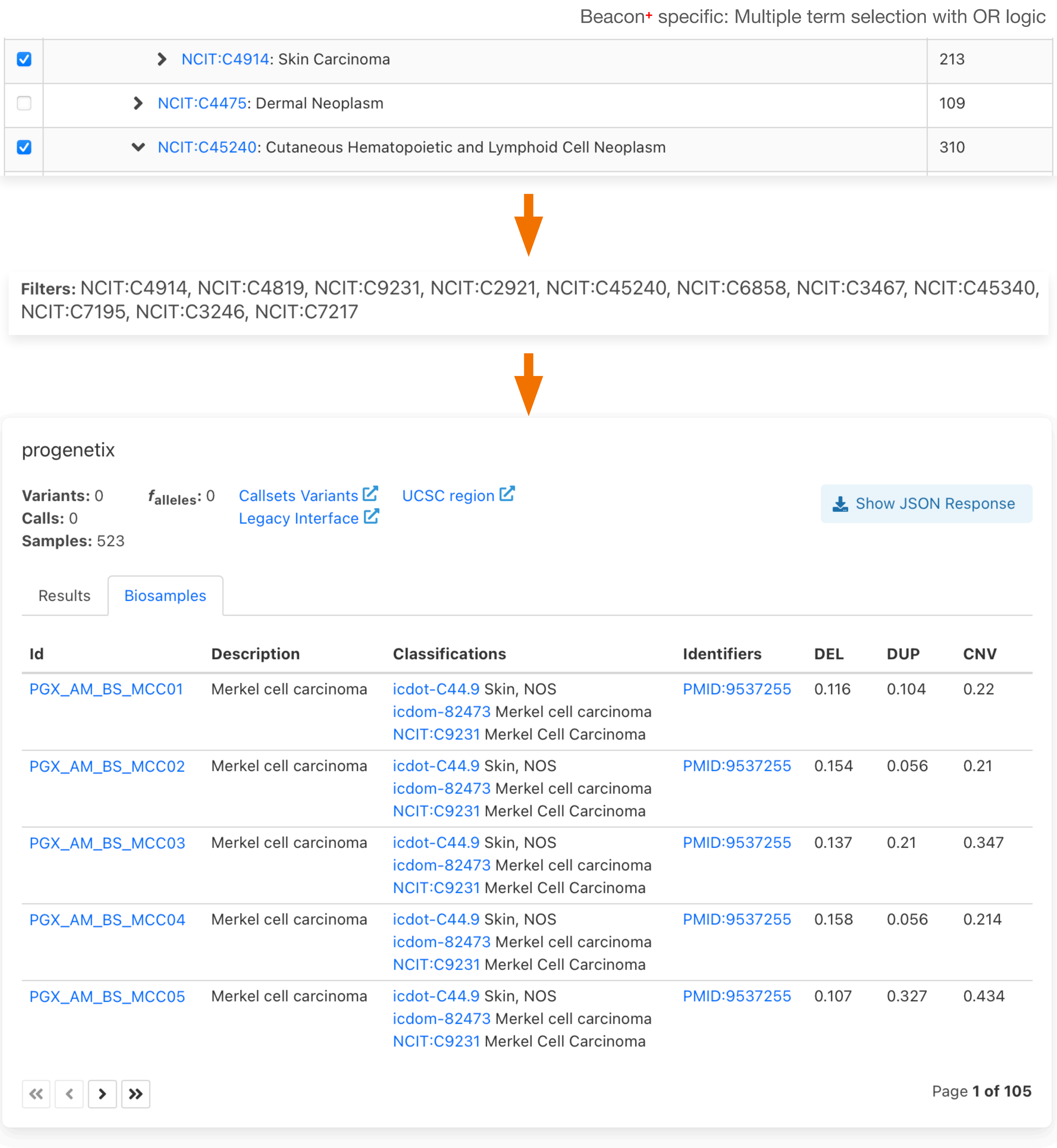
# Beacon v2 Filters

## 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



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
**Calls:** 0
**Samples:** 523

$f_{alleles}$: 0

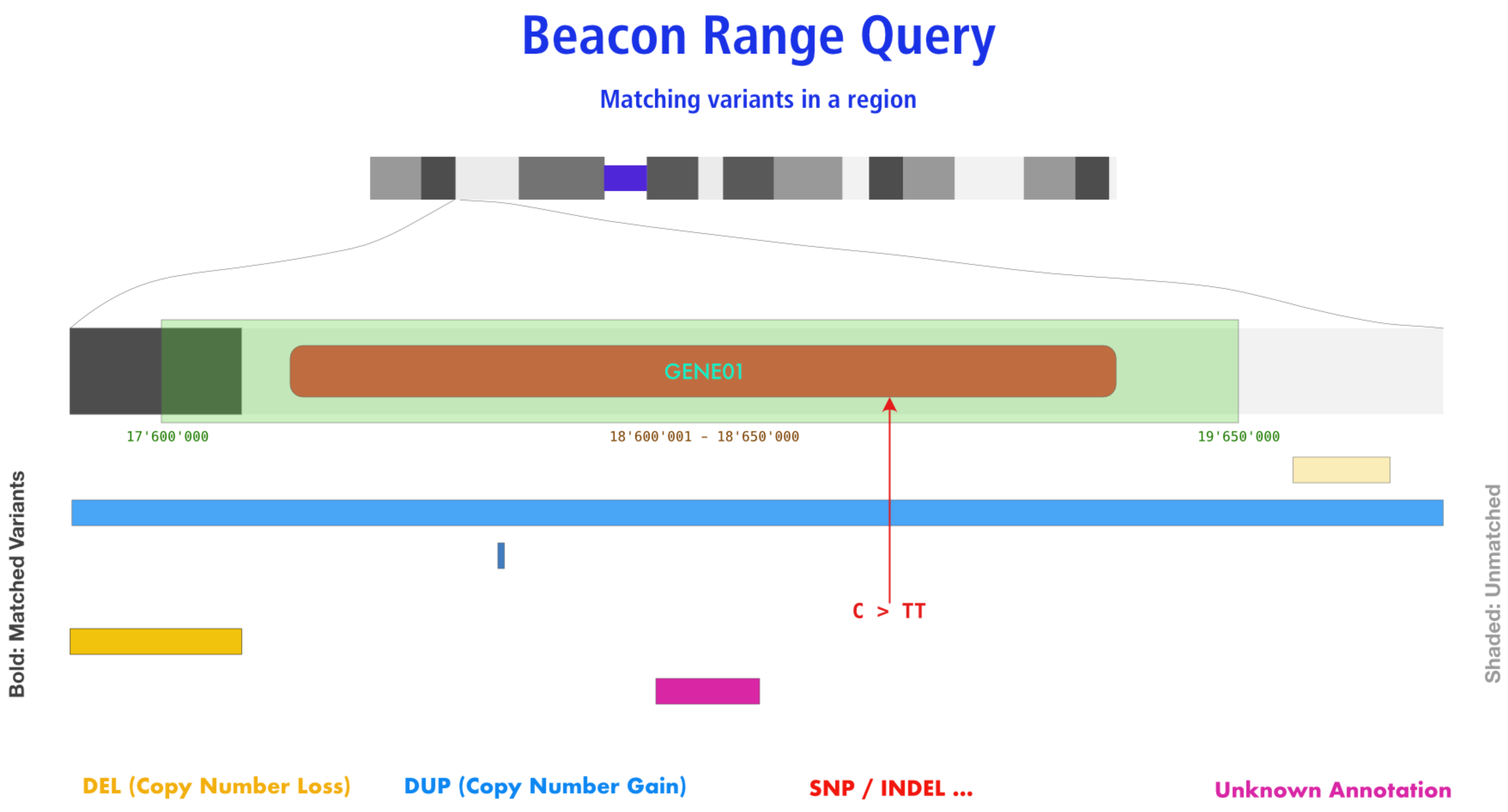
| Id | Description | Classifications | Identifiers | DEL | DUP | CNV |
|---|---|---|---|---|---|---|
| PGX_AM_BS_MCC01 | Merkel cell carcinoma | icdot-C44.9 Skin, NOS<br>icdom-82473 Merkel cell carcinoma<br>NCIT:C9231 Merkel Cell Carcinoma | PMID:9537255 | 0.116 | 0.104 | 0.22 |
| PGX_AM_BS_MCC02 | Merkel cell carcinoma | icdot-C44.9 Skin, NOS<br>icdom-82473 Merkel cell carcinoma<br>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<br>icdom-82473 Merkel cell carcinoma<br>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<br>icdom-82473 Merkel cell carcinoma<br>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<br>icdom-82473 Merkel cell carcinoma<br>NCIT:C9231 Merkel Cell Carcinoma | PMID:9537255 | 0.107 | 0.327 | 0.434 |

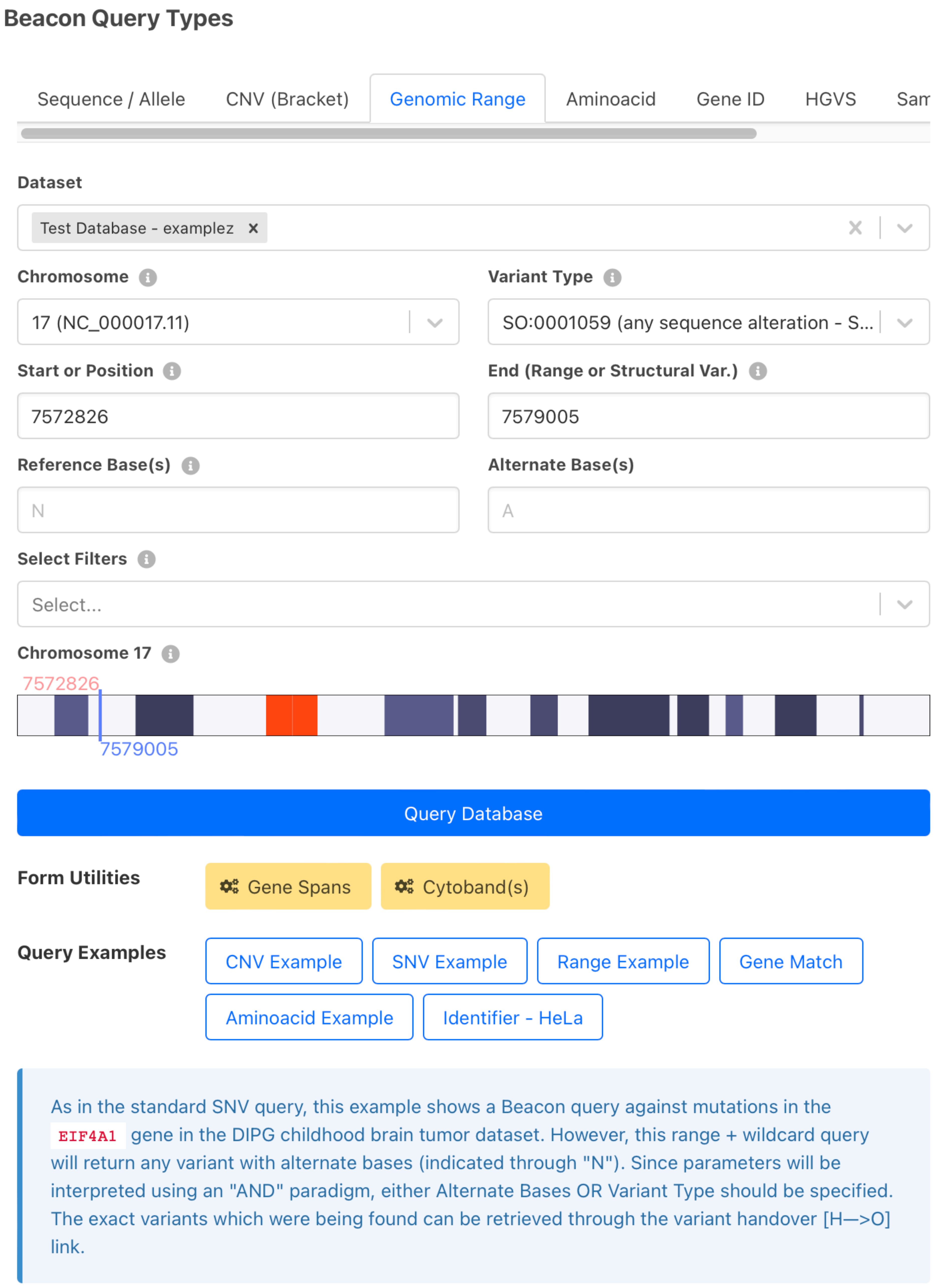
Page **1 of 105**

# Beacon Queries

## Range ("anything goes") Request

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





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

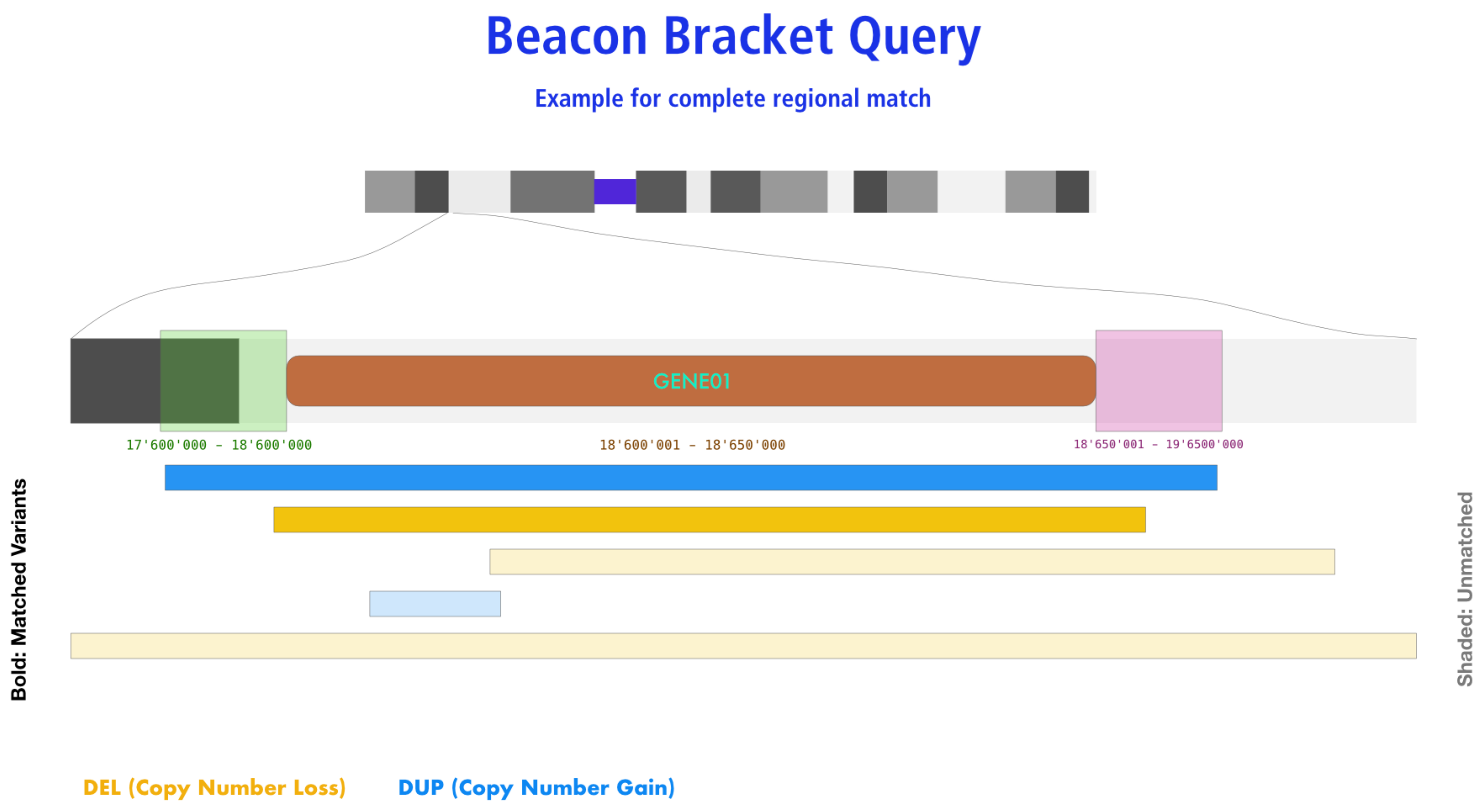
Query Examples

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

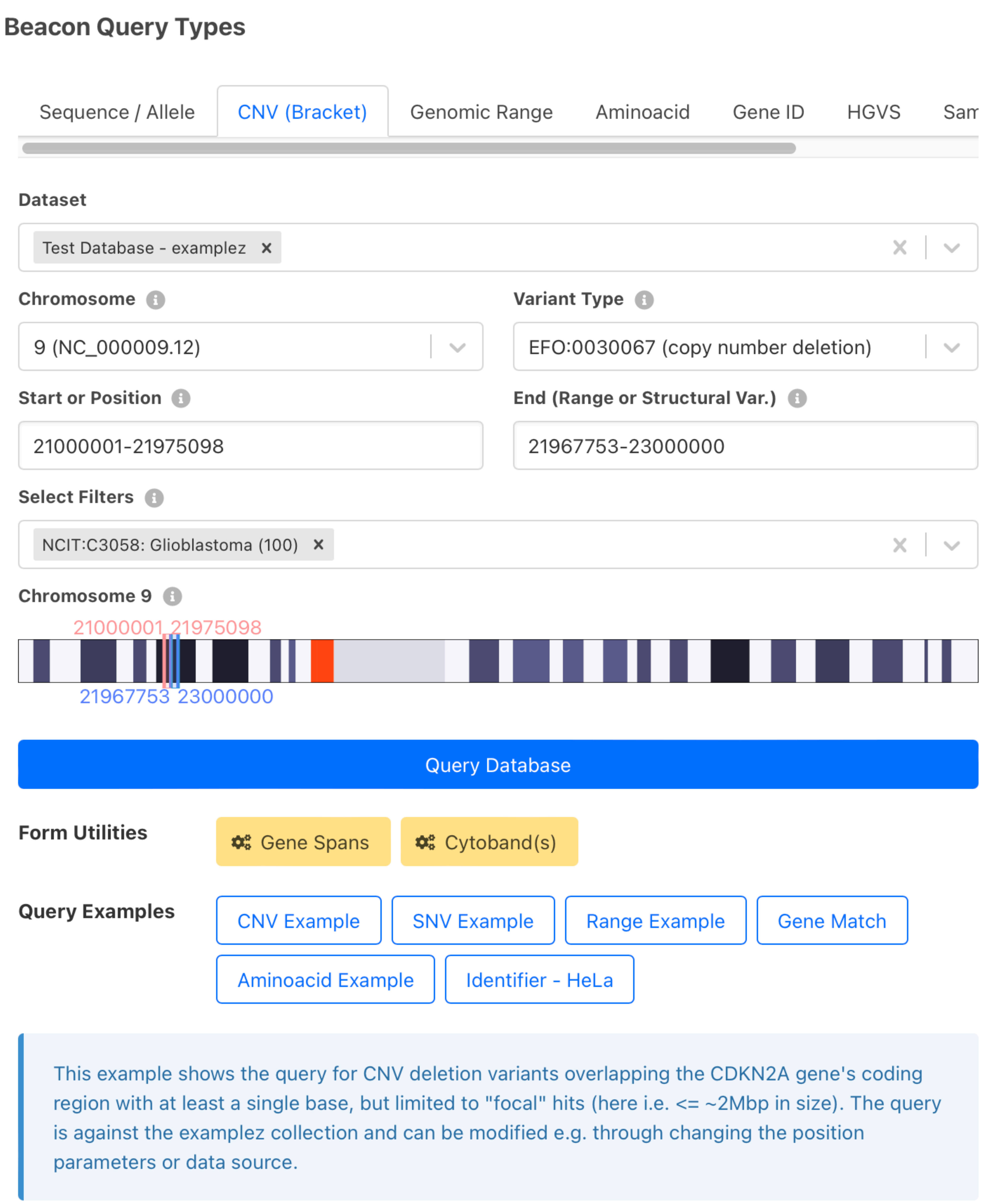
## Bracket ("CNV") Query

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



Beacon Bracket Query

Example for complete regional match



Beacon Query Types

Sequence / Allele | CNV (Bracket) | Genomic Range | Aminoacid | Gene ID | HGVS | Sar

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.

# CNV Term Use Comparison

## in computational (file/schema) formats

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

Universität Zürich^UZH

Global Alliance for Genomics & Health

# Progenetix and GA4GH Beacon

## Implementation driven development of a GA4GH standard

progenetix

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

# Progenetix & Beacon

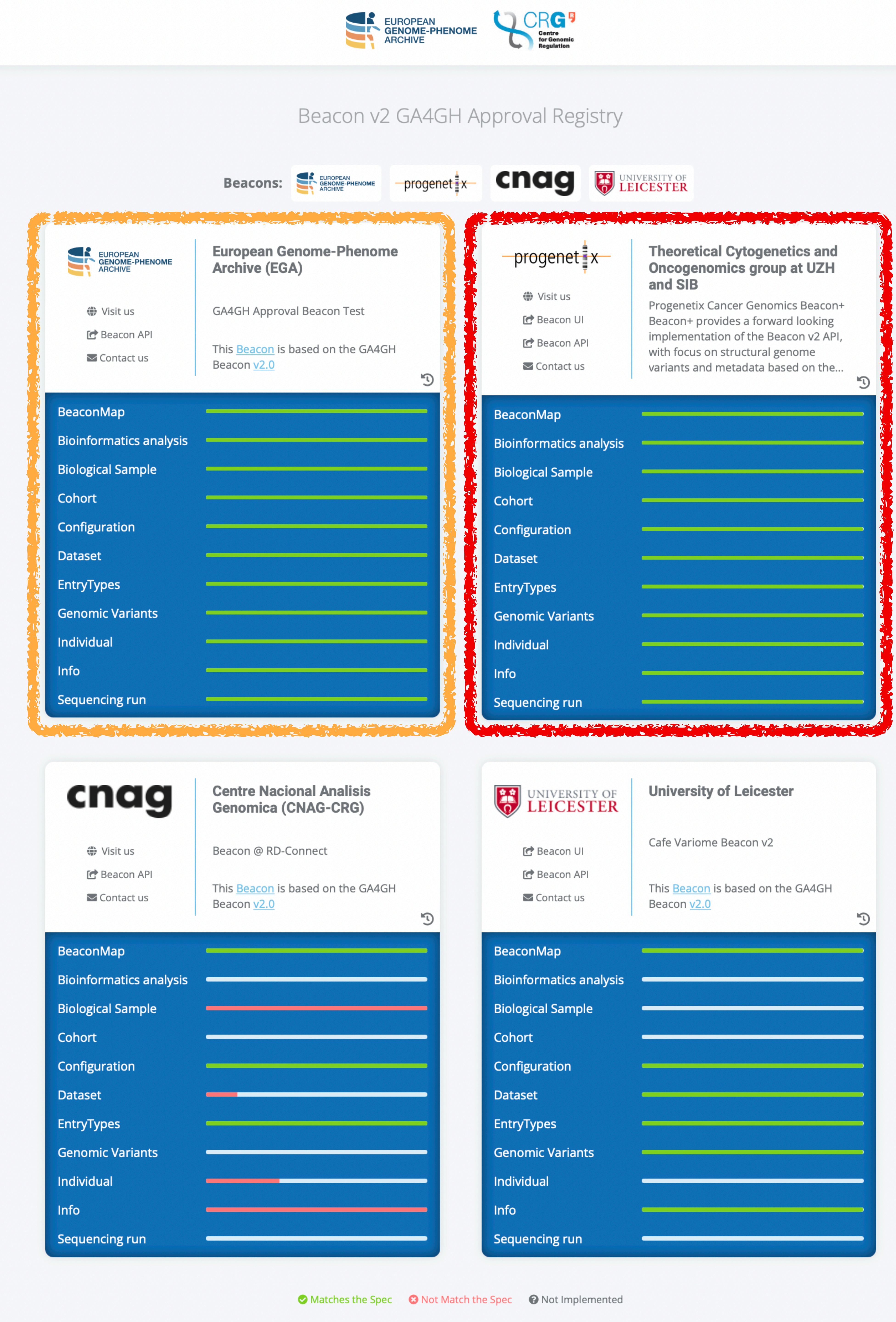
## Implementation driven standards development

- 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 response verifier at time of GA4GH approval Spring 2022

# 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

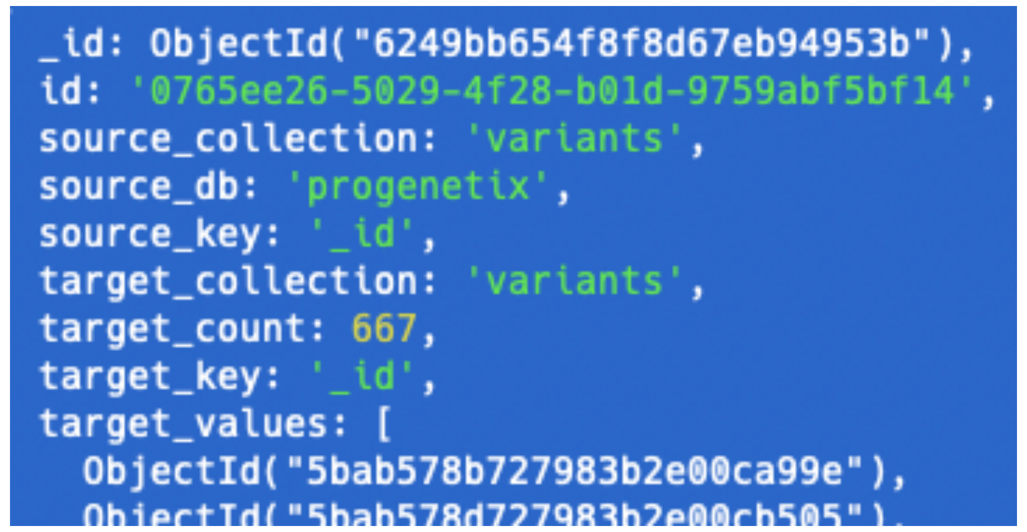
React

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 *bycon*

## Putting the + into Beacon ...

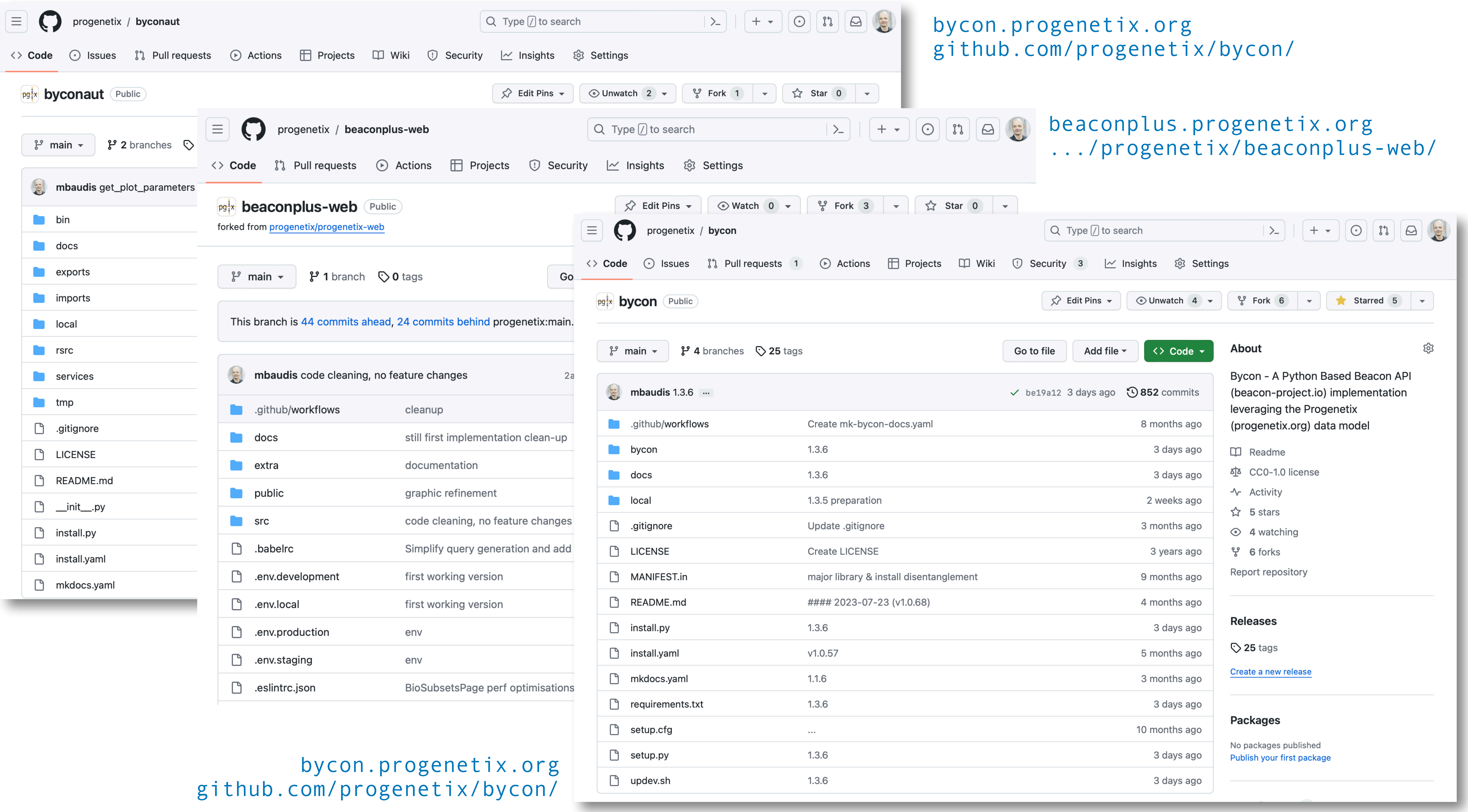
- support & use of standard Beacon v2 PUT & GET variant queries, filters and meta parameters
  - ➡ variant parameters, geneId, lengths, EFO, SO & VCF CNV types, pagination
  - ➡ widespread, self-scoping filter use for bio-, technical- and and id parameters with switch for descending terms use (globally 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
  - ➡ testing of **cytogenetic events**
  - ➡ **multi-variant queries**, i.e. option to supply multiple variant queries of same or different types which are **intersected at the biosample level**
- ╮(╯ ▽ ╰)╭ only rudimentary/test implementation of authentication on this open dataset

***bycon* provides additional services and output formats through *byconaut* & `/services` path 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

byconaut Public

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

beaconplus-web Public

forked from progenetix/progenetix-web

main · 1 branch · 0 tags

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

bycon Public

main · 4 branches · 25 tags

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

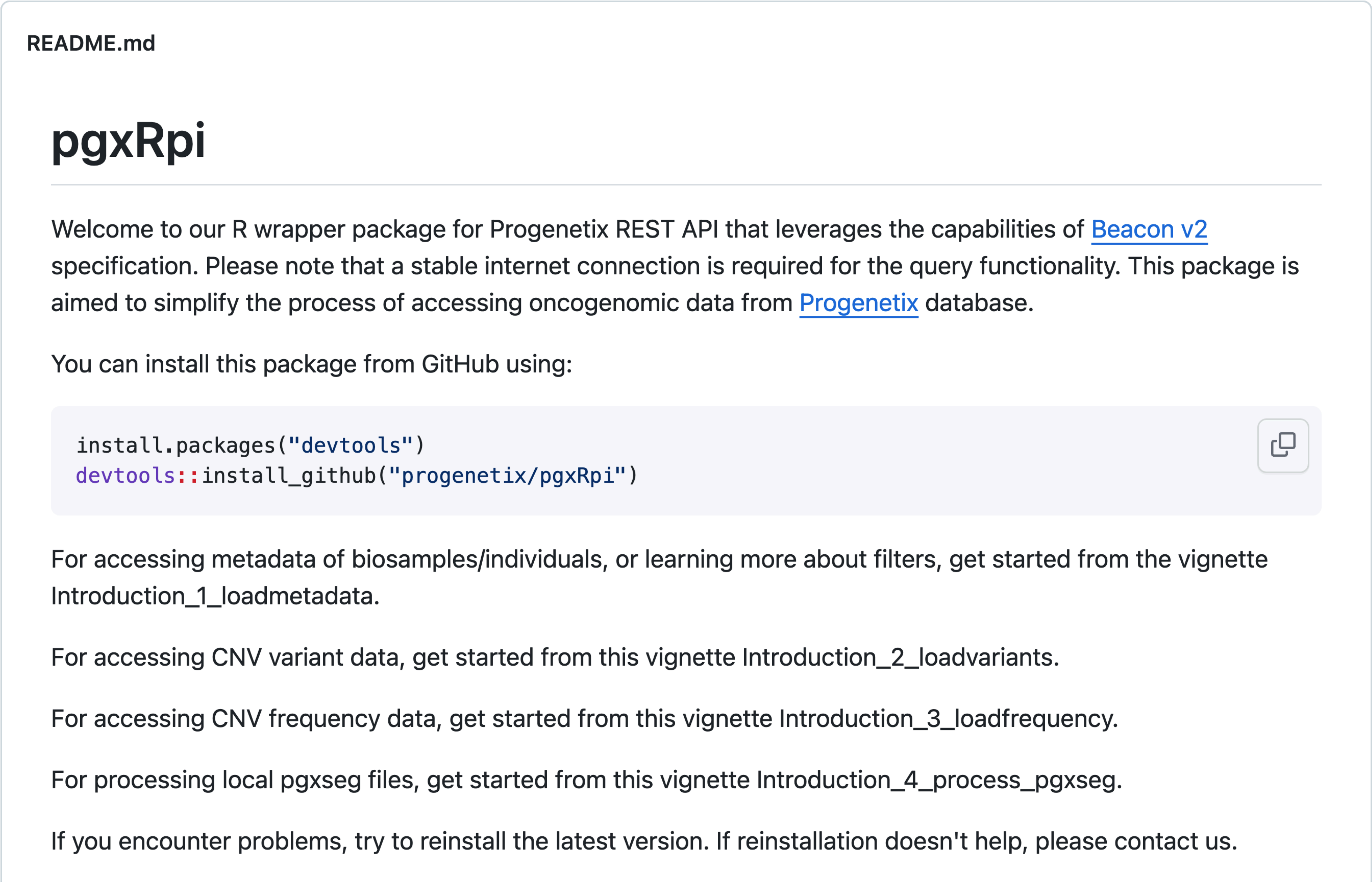
bycon.progenetix.org
github.com/progenetix/bycon/

# 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")
```

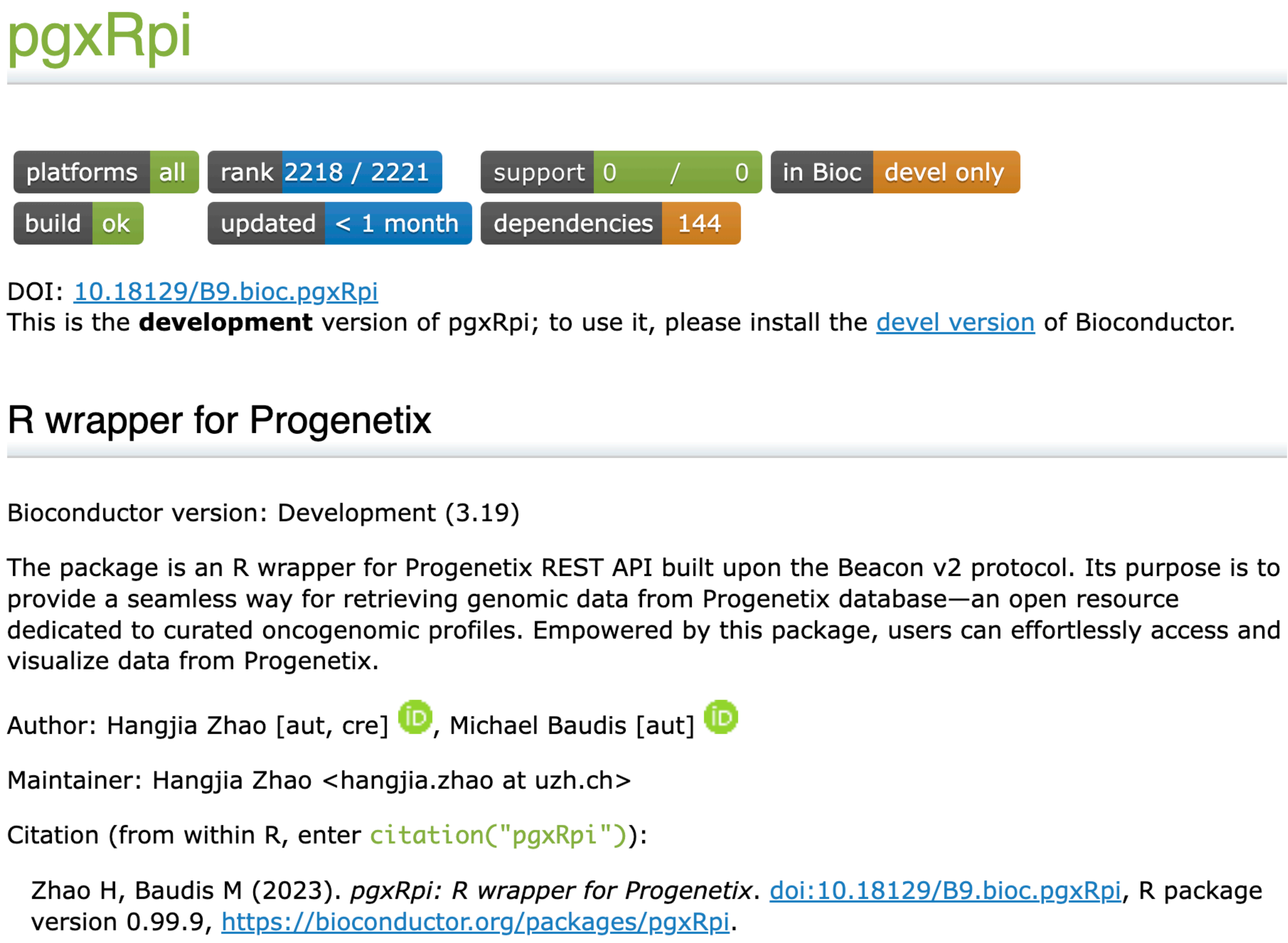
For accessing metadata of biosamples/individuals, or learning more about filters, get started from the vignette Introduction_1_loadmetadata.

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>

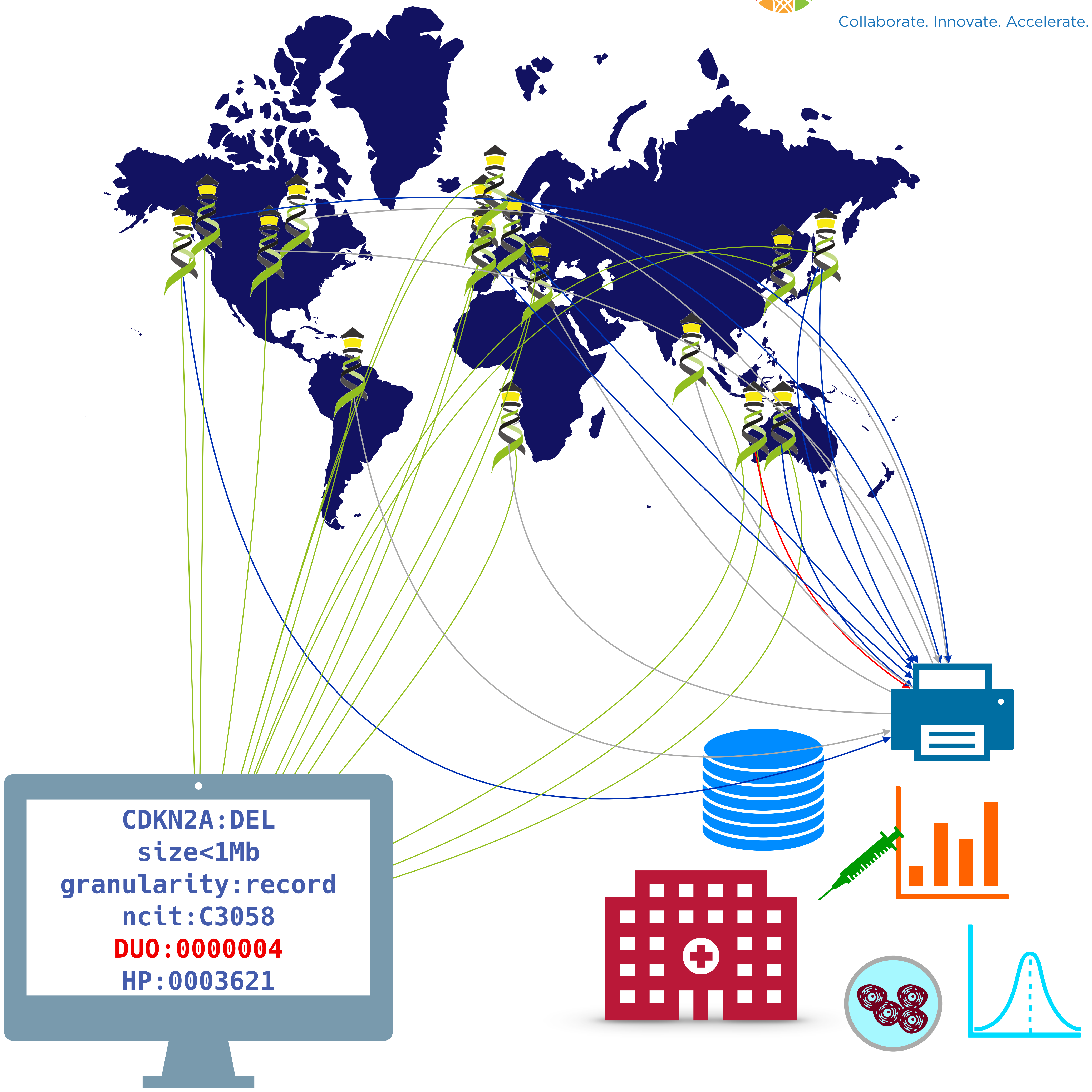
Citation (from within R, enter `citation("pgxRpi")`):

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.

# 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!**

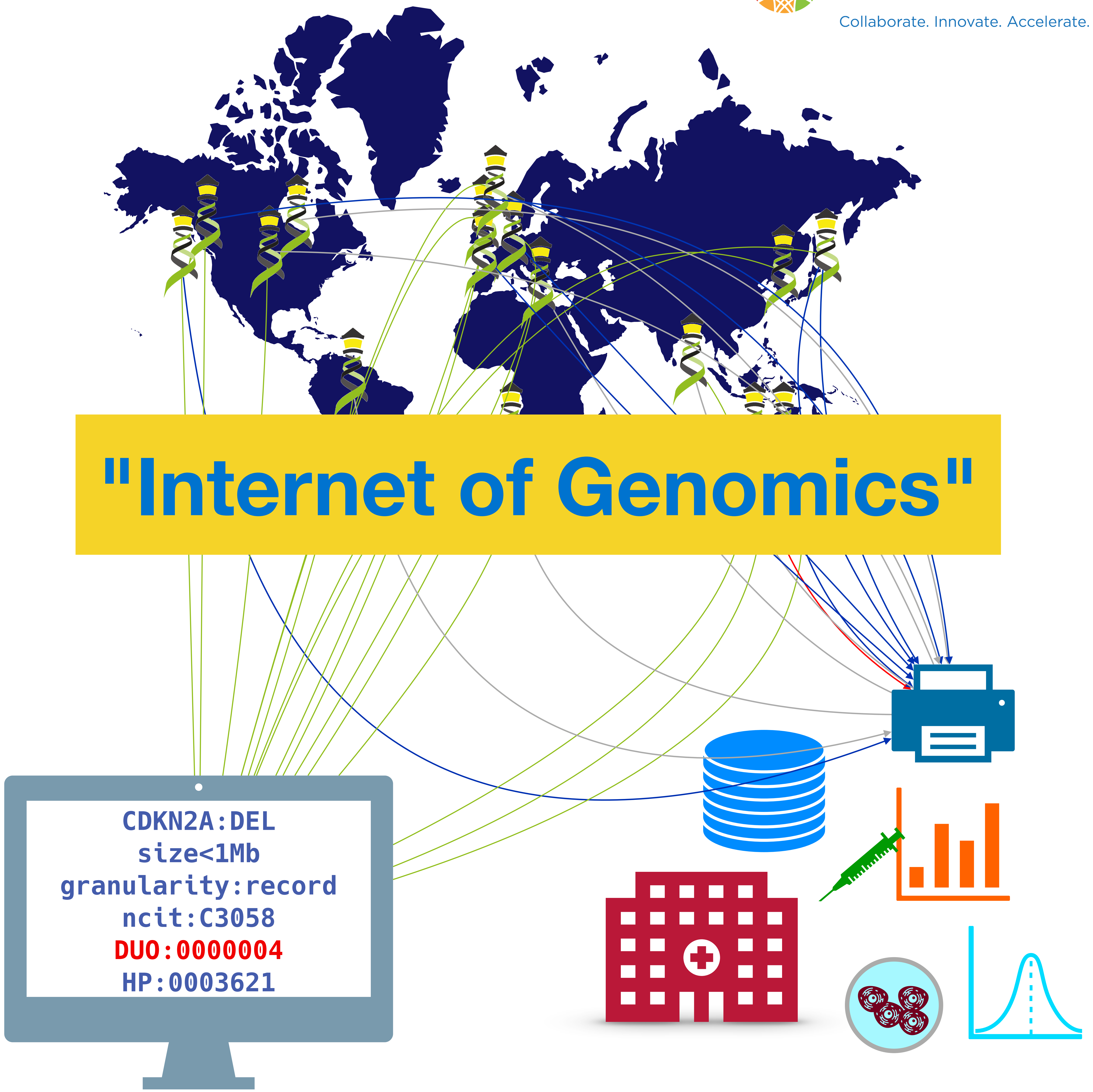


CDKN2A:DEL
size<1Mb
granularity:record
ncit:C3058
DUO:0000004
HP:0003621

# 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"

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

# 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

BSC 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

Toshiaki Katayama

GEM Japan

**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

GA4GH Connect Ascona 2024

Universität Zürich UZH

SIB Swiss Institute of Bioinformatics

elixir