

# Data *Discovery* and Data *Sharing* in Biomedical Genomics

## Reference Genomics Resources Powered by the GA4GH Beacon Standard



**Michael Baudis**

Professor of Bioinformatics

University of Zürich

Swiss Institute of Bioinformatics **SIB**

Member GA4GH Strategic Leadership Committee

Co-lead ELIXIR Beacon API Development

Co-lead ELIXIR hCNV Community



**Universität  
Zürich**<sup>UZH</sup>



**SIB**  
Swiss Institute of  
Bioinformatics



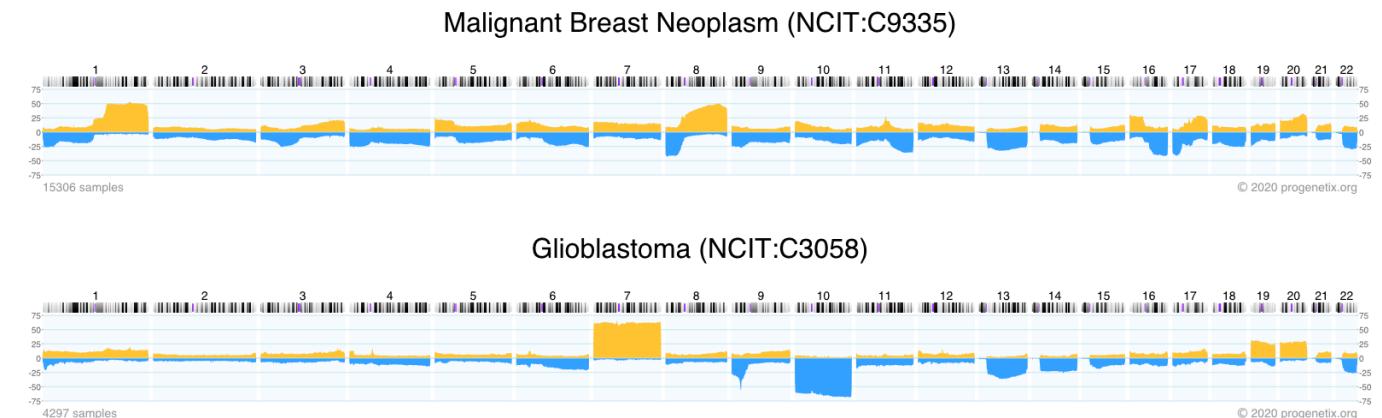
**Global Alliance**  
for Genomics & Health  
*Collaborate. Innovate. Accelerate.*



# Theoretical Cytogenetics and Oncogenomics

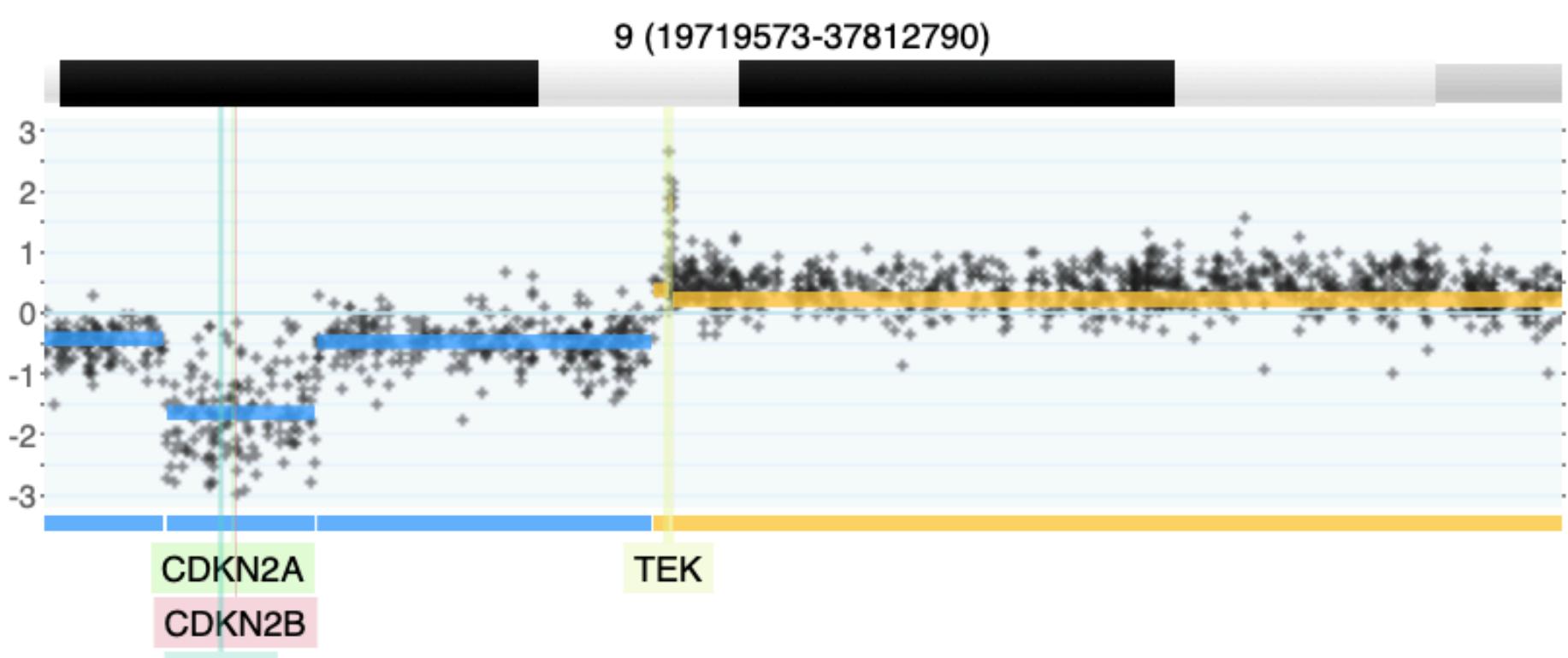
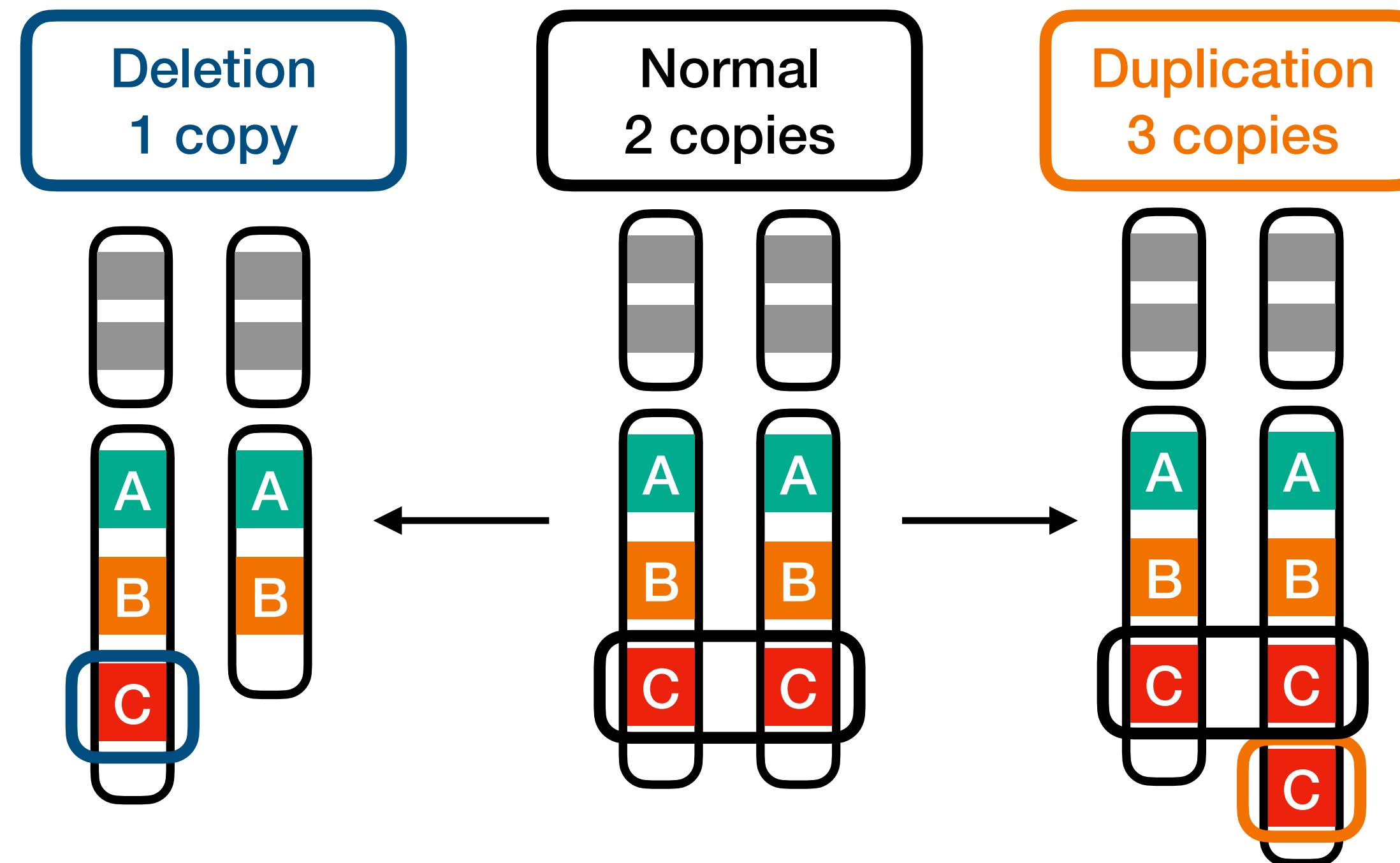
## ... but what does this entail @baudisgroup?

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



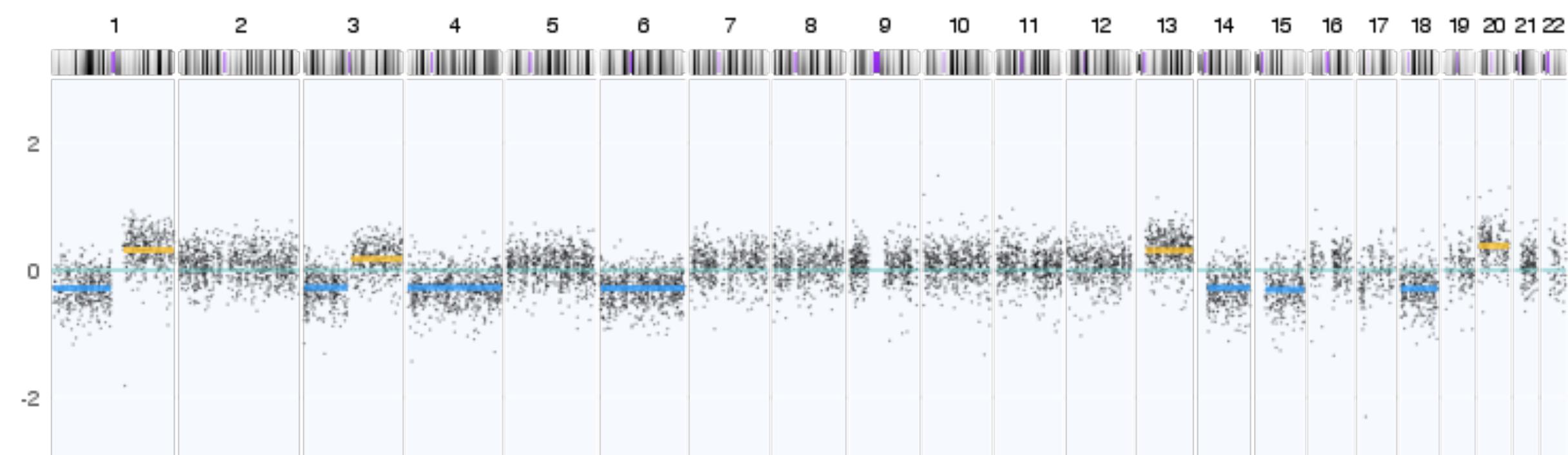
*Curators*  
~~Data Parasites~~

# Copy Number Variant (CNV)



2-event, homozygous deletion in a Glioblastoma

- Intermediate-scale genetic change
- Size: 1kb to multiple megabase
- Additional copies of sequence (**duplications**) and losses of genetic material (**deletions**)



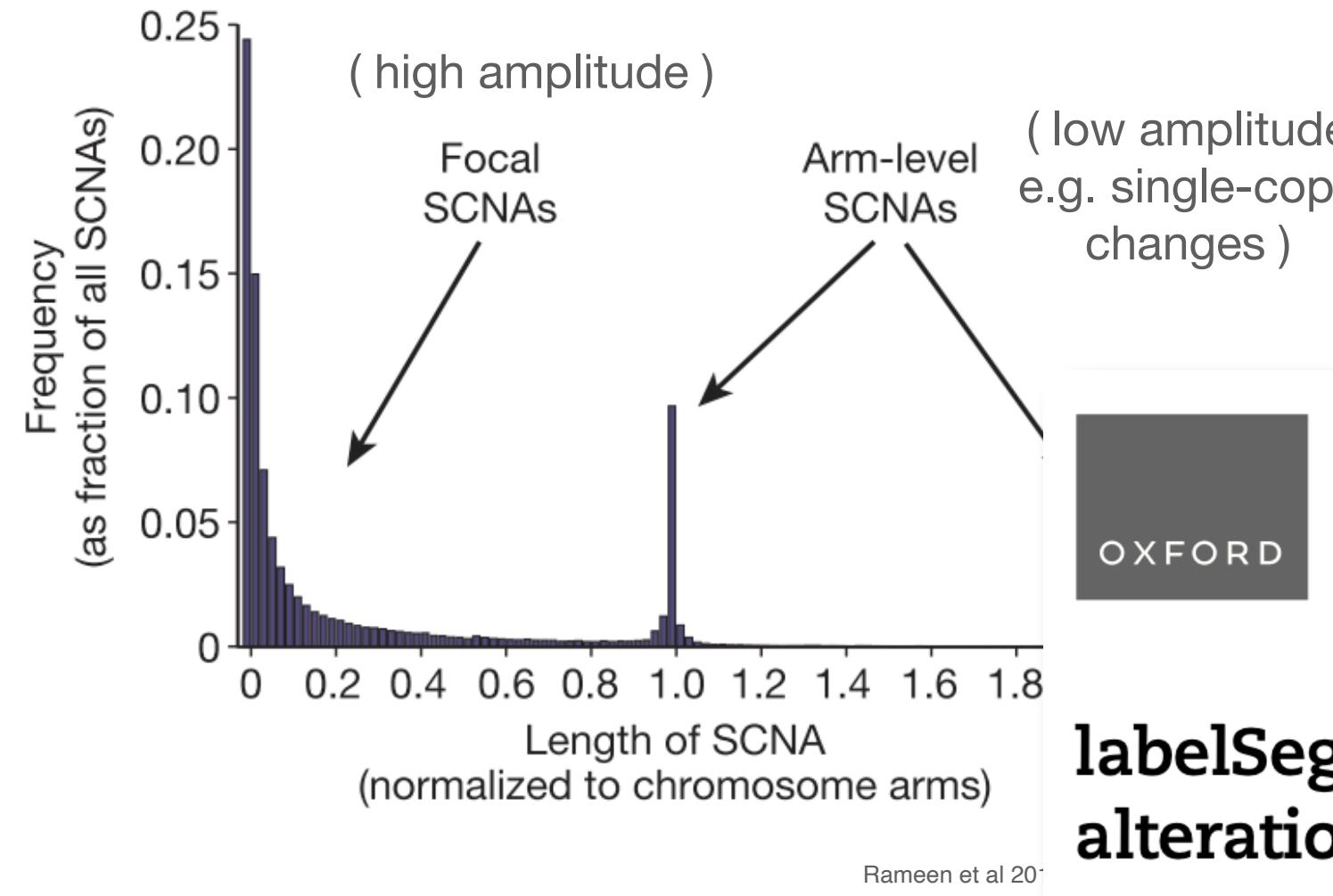
Gain of chromosome arm 13q in colorectal carcinoma

# CNV Categorization Method Supporting GA4GH Standards

GA4GH Variation Representation Specification



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

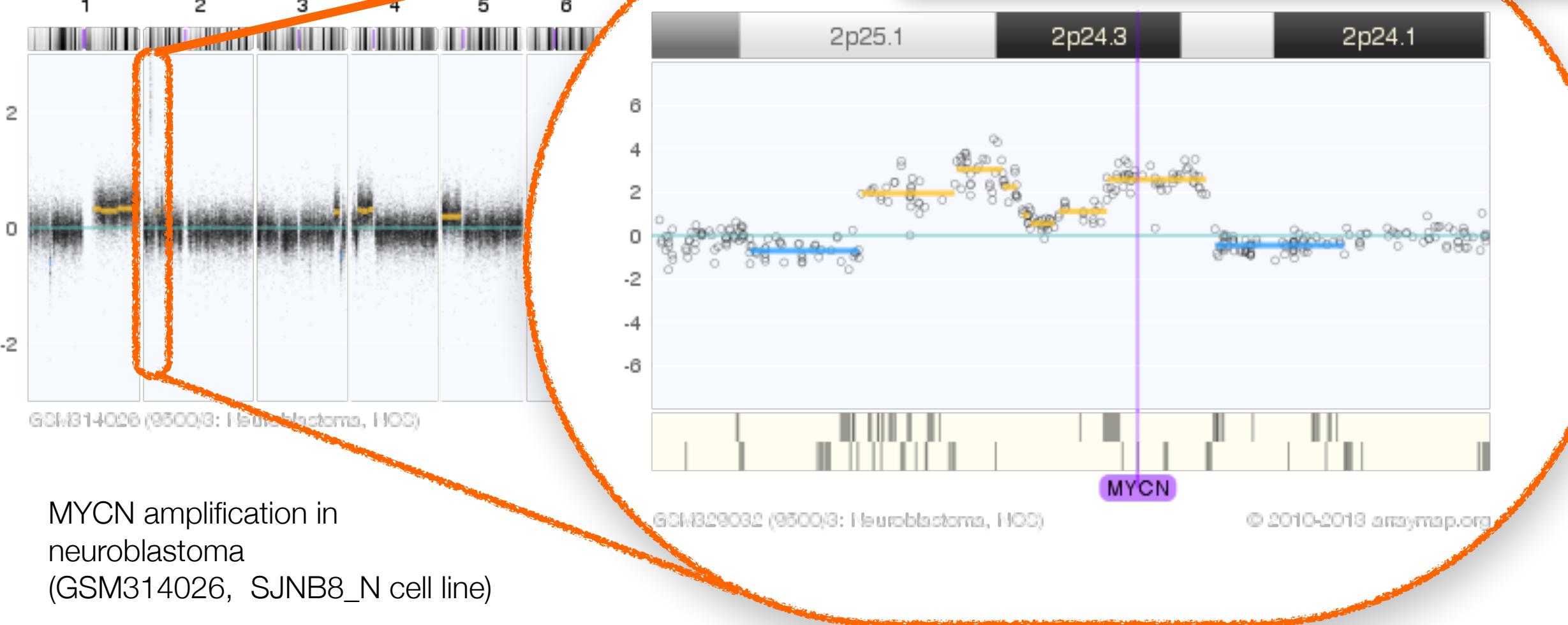


## labelSeg: segment annotation for tumor copy number alteration profiles

Hangjia Zhao and Michael Baudis

Corresponding author: Michael Baudis, Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland.  
Tel.: (+41) 44 635 34 86; E-mail: michael.baudis@mls.uzh.ch

GSM329032, 2 (5000000 - 24000000)



## 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 express copy number variation are interpreted to be relative copy

Briefings in Bioinformatics, 2024, 25(2), 1–12

<https://doi.org/10.1093/bib/bbad541>

Problem Solving Protocol

number of a [Location](#) or a [Feature](#) within a system (e.g. genome, cell,

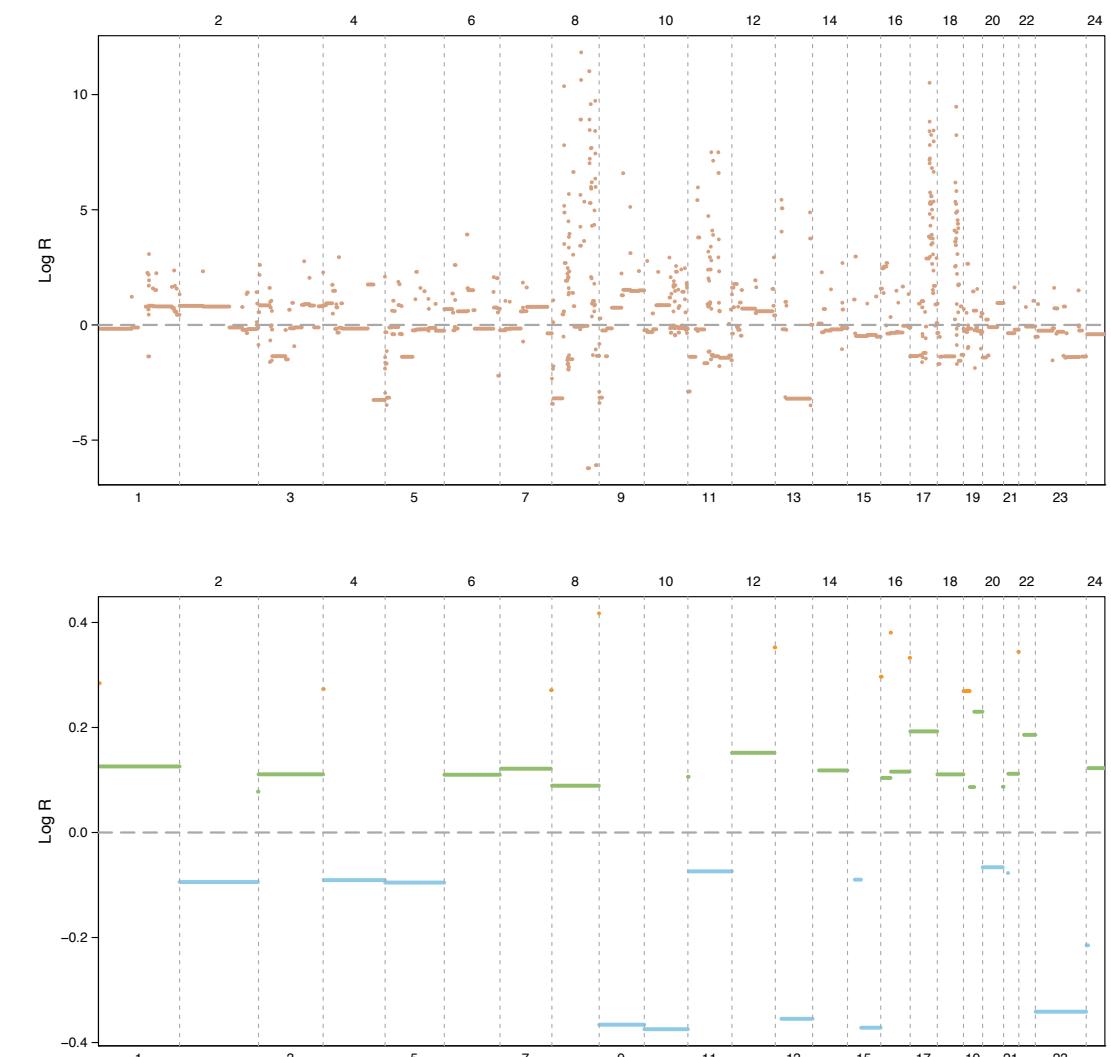
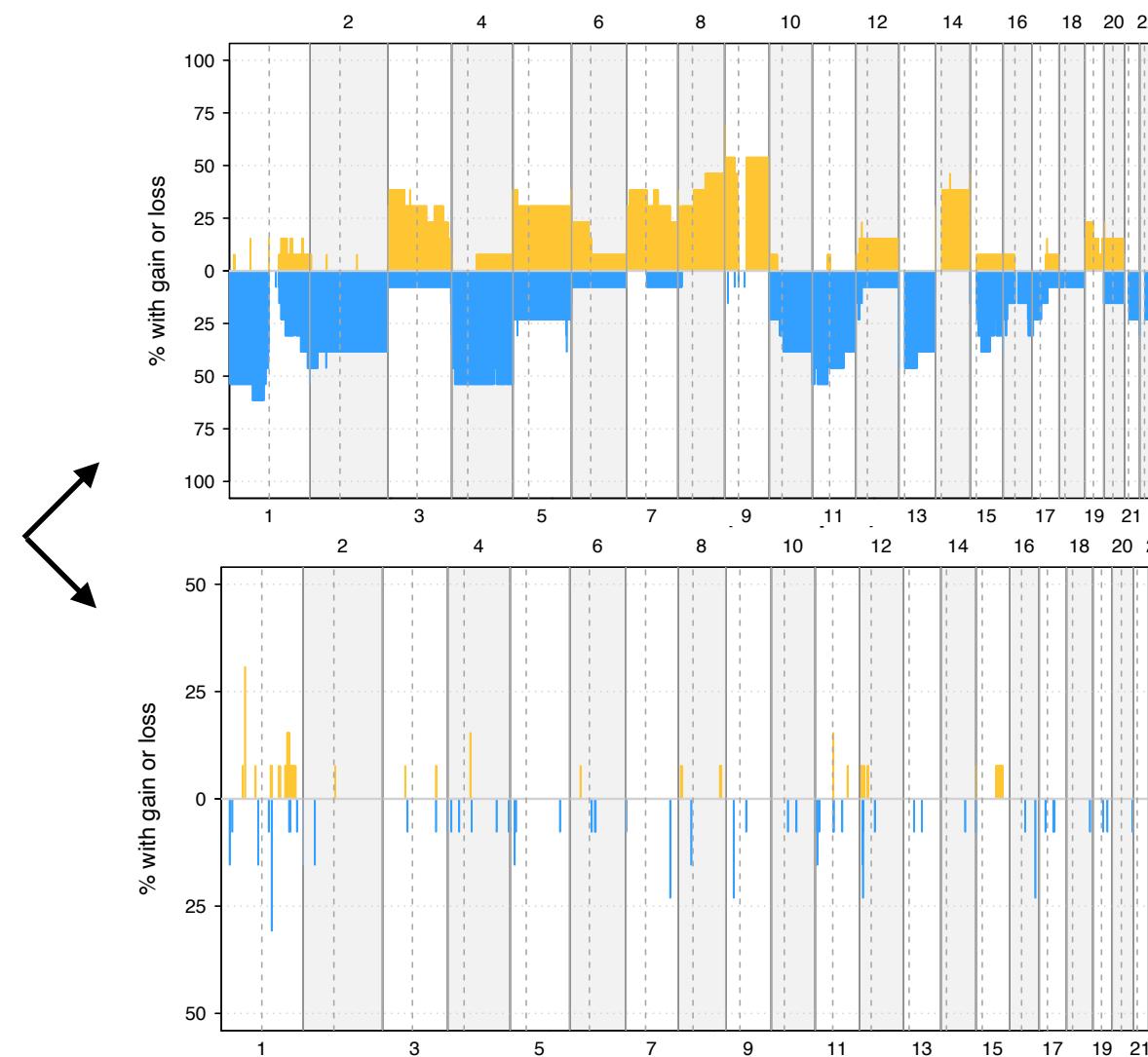
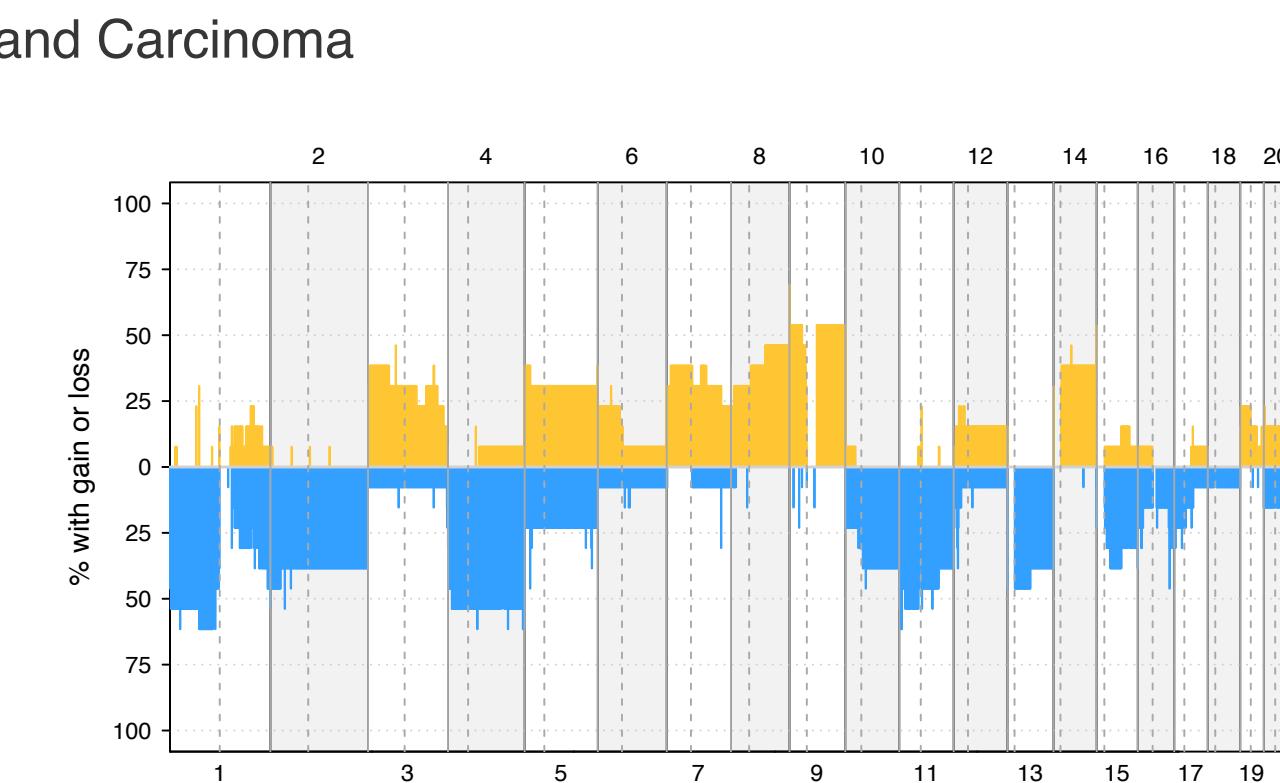
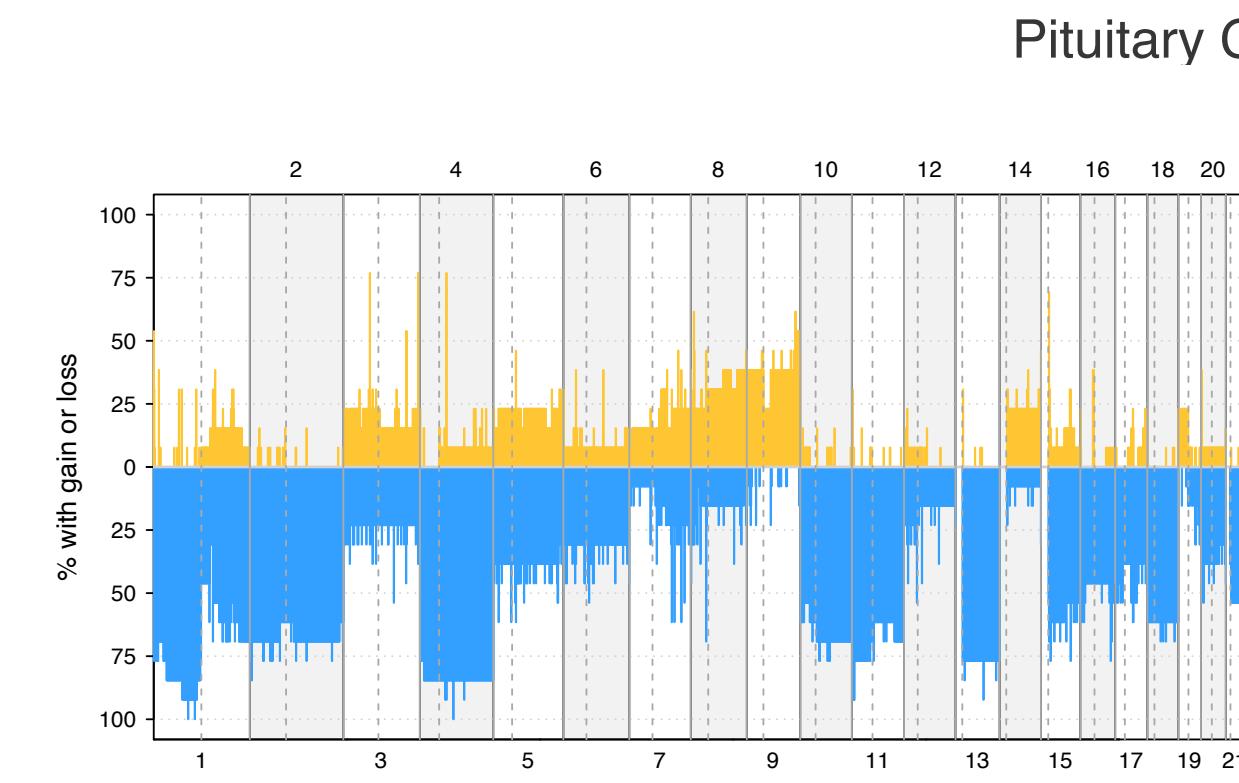
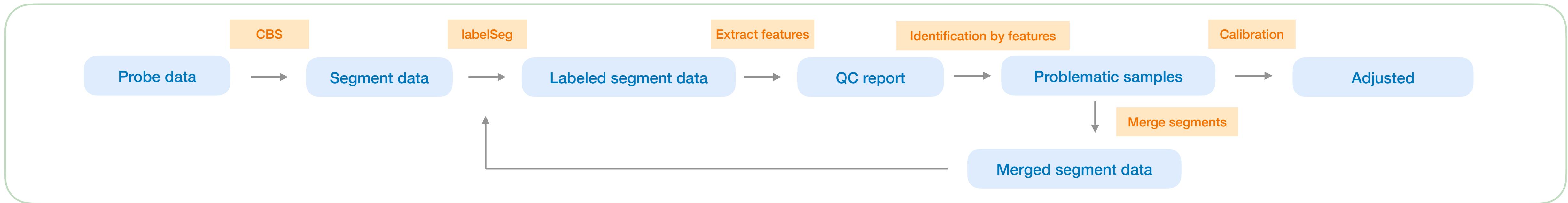
organism, tissue, cell type, individual, population, etc.). Copy number variations are inherited from [Variation](#).

Field	Type	Limits	Description
_id	<a href="#">CURIE</a>	0..1	Variation Id. MUST be unique within document.
type	string	1..1	MUST be "CopyNumberChange"
subject	<a href="#">Location</a>   <a href="#">CURIE</a>   <a href="#">Feature</a>	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).

# Pipeline Development

improve CNV calling in large numbers of heterogeneous cancer samples

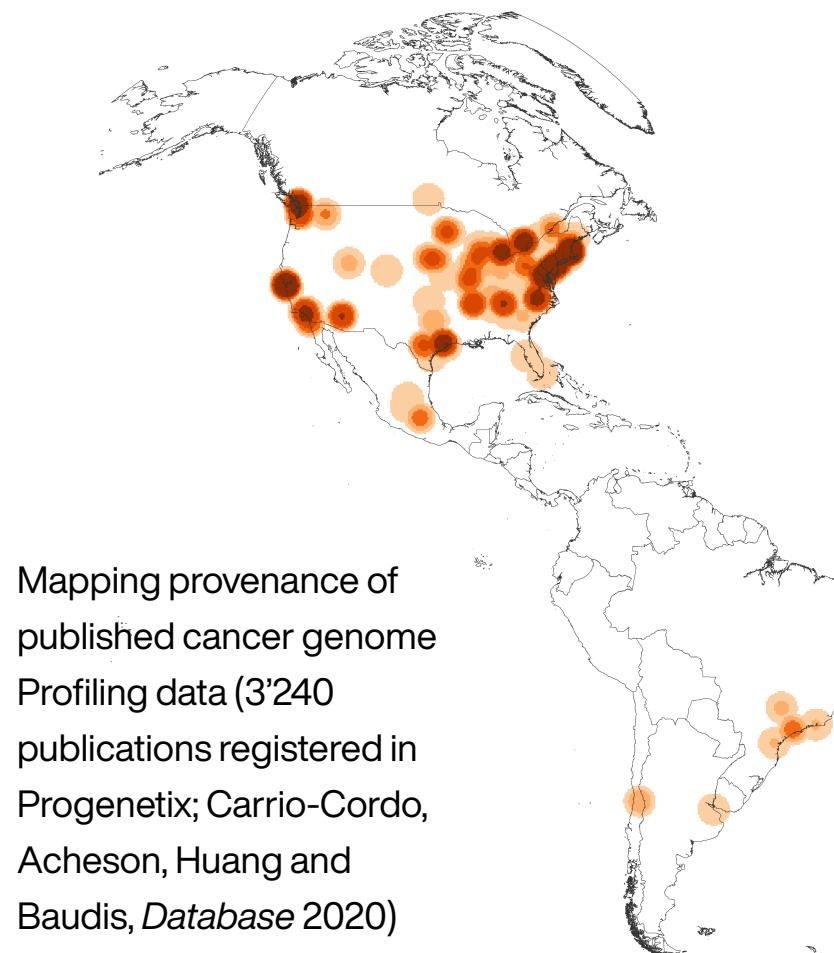
nextflow



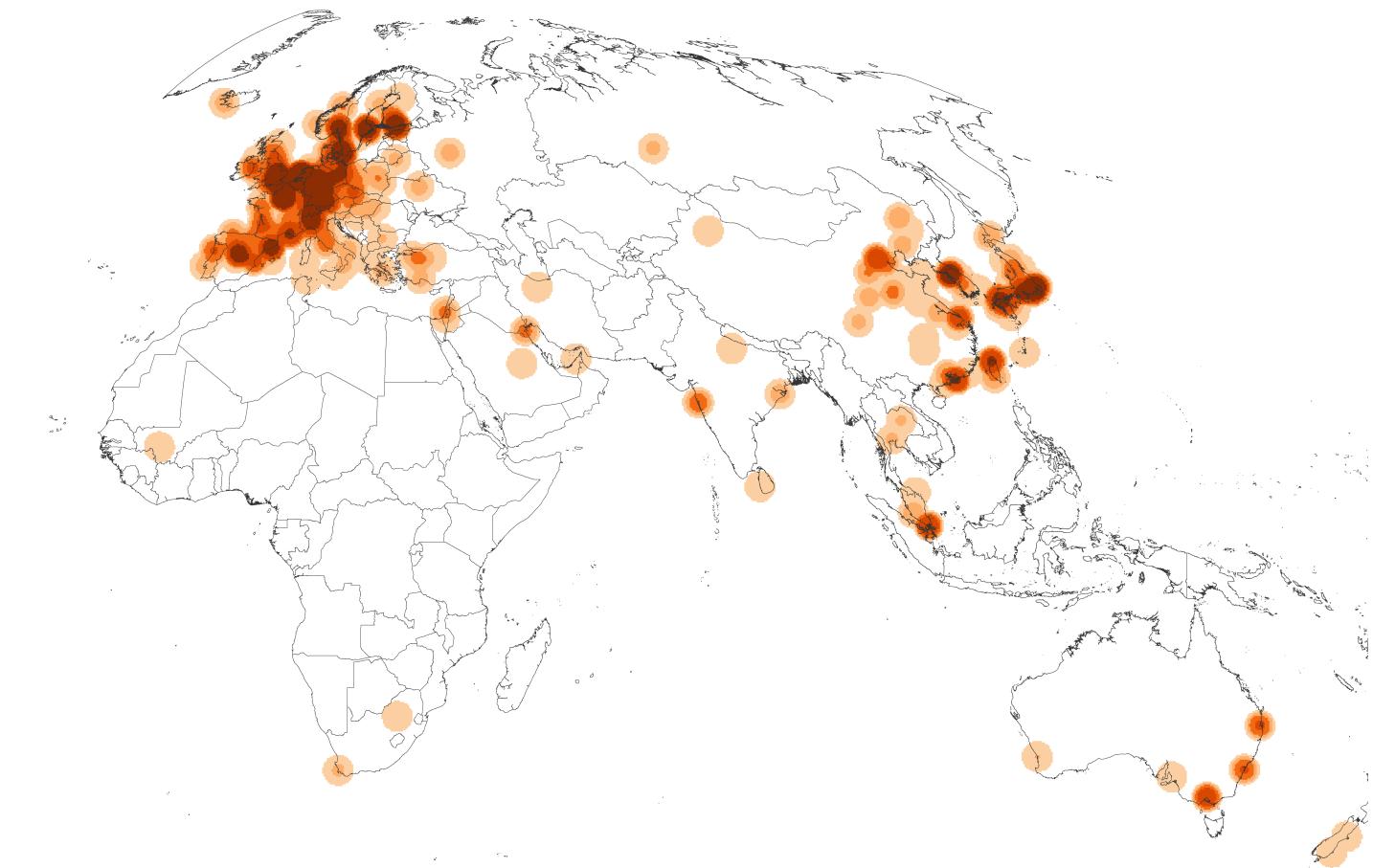
# progenetix.org @ UZH

## Cancer Genomics Reference Resource

- over 240'000 cancer CNV profiles
- more than 1100 diagnostic types
- runs on a **GA4GH Beacon API**
- Open Research Data practices



Mapping provenance of published cancer genome Profiling data (3'240 publications registered in Progenetix; Carrio-Cordo, Acheson, Huang and Baudis, Database 2020)



### Cancer Types by National Cancer Institute NCIt Code

The cancer aggregate can be init

Sample se class are ii

Filter sui

No Selection

#### Glioblastoma (NCIT:C3058)

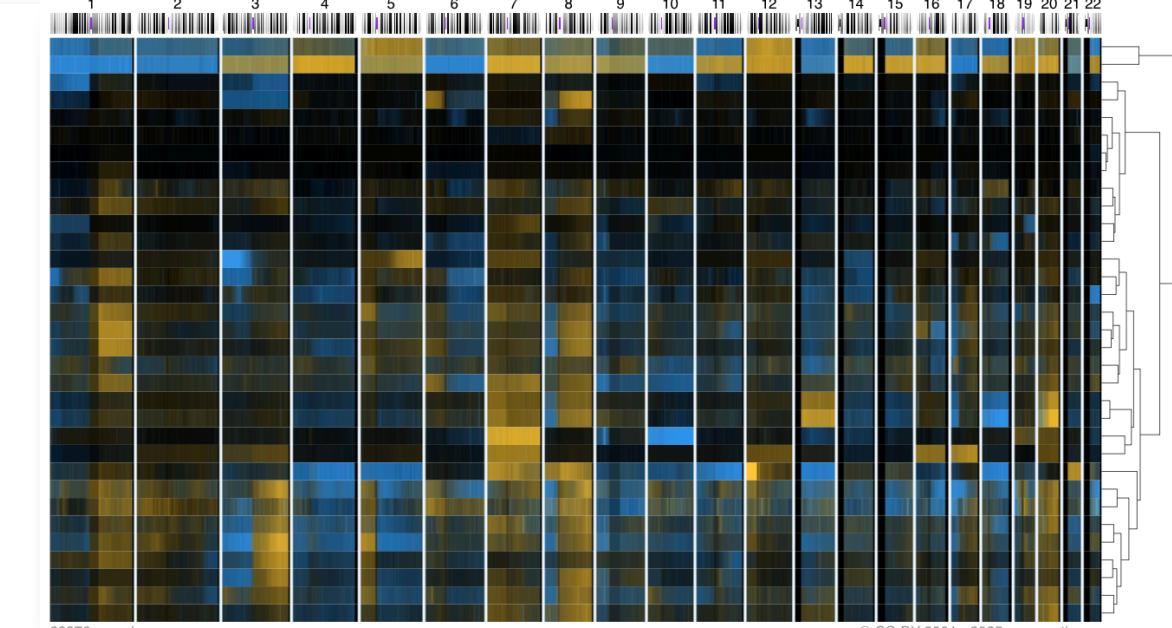
##### Sample Counts

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

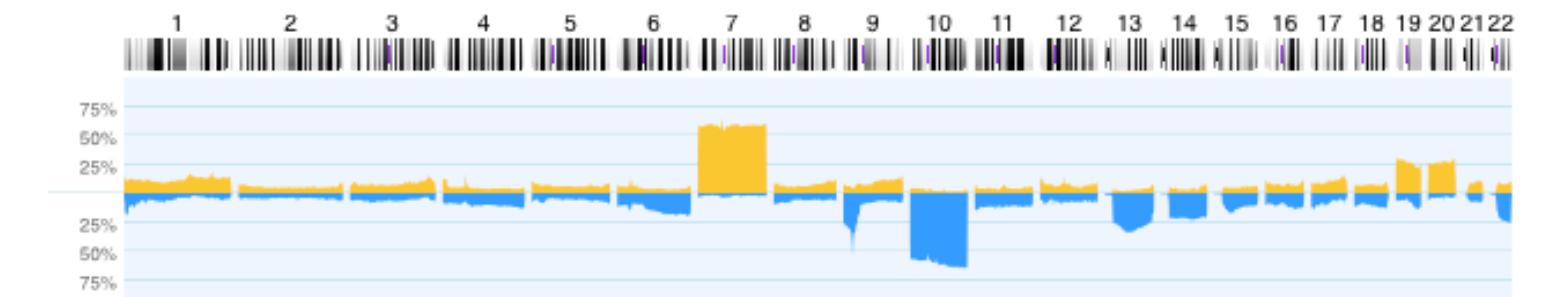
##### Search Samples

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:C3059: Glioma (8825 samples, 8183 CNV profiles)

▼ NCIT:C129325: Diffuse Glioma (6123 samples, 6137 CNV profiles)

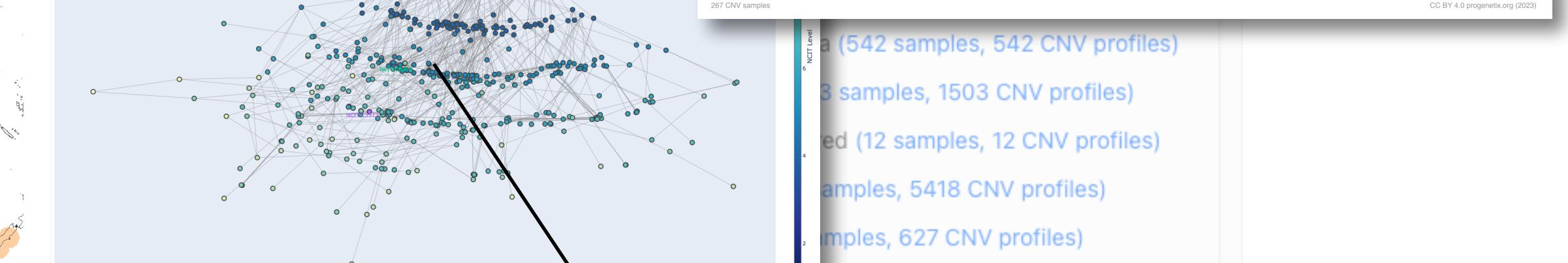
➤ NCIT:C182151: Diffuse Midline Glioma (2 samples, 2 CNV profiles)

▼ NCIT:C4755: Chondrogenic Neoplasm (4384 CNV profiles)

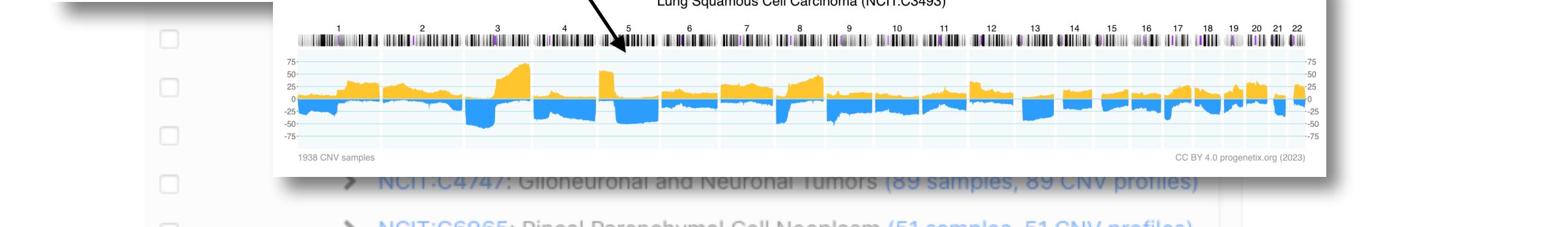
▼ NCIT:C3493: Lung Squamous Cell Carcinoma (1938 samples, 500 CNV profiles)

▼ NCIT:C4747: Glioneuronal and Neuronal Tumors (69 samples, 89 CNV profiles)

▼ NCIT:C6965: Pineal Parenchymal Cell Neoplasm (51 samples, 51 CNV profiles)



Chondrogenic Neoplasm (NCIT:C4755)

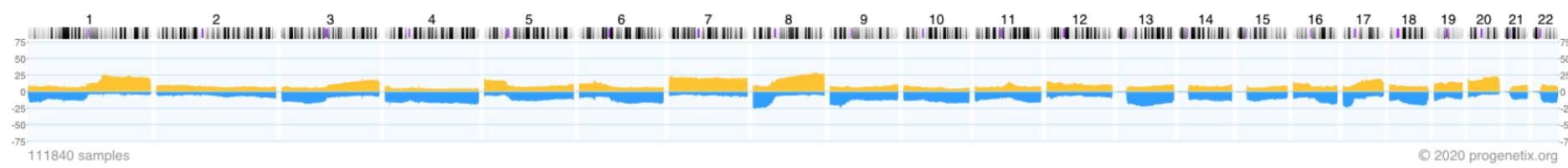


Lung Squamous Cell Carcinoma (NCIT:C3493)

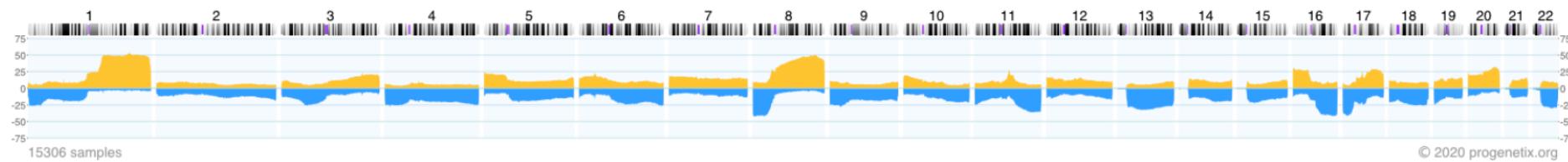
CC BY 4.0 progenetix.org (2023)

## Cancer Genomics Reference Resource

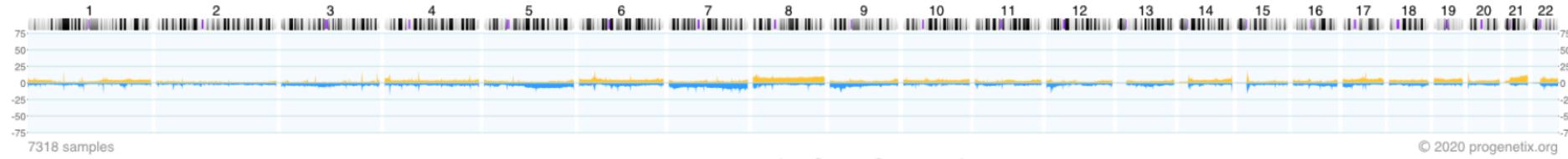
- open resource for oncogenomic profiles
- over 240'000 cancer CNV profiles
- SNV data for some series (e.g. TCGA)
- more than 1100 diagnostic types
- inclusion of reference datasets (e.g. TCGA, GENIE, cBioPortal)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services



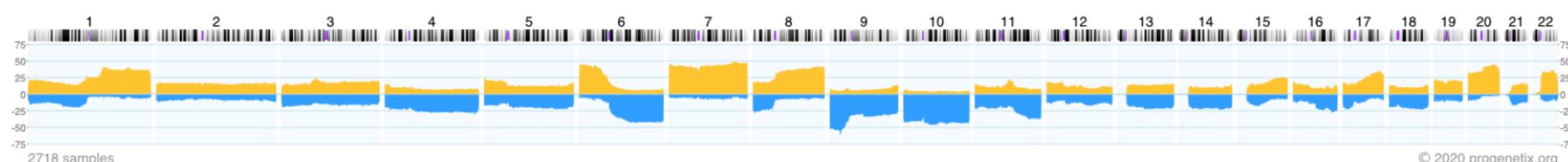
Malignant Breast Neoplasm (NCIT:C9335)



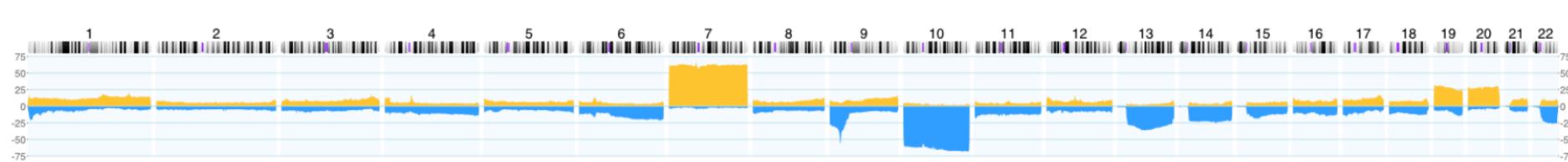
Acute Leukemia (NCIT:C9300)



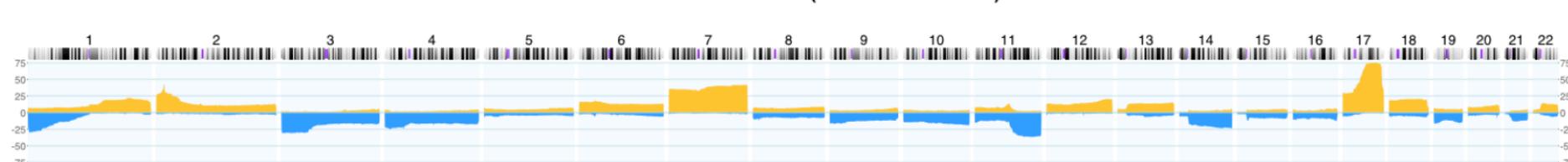
Melanoma (NCIT:C3224)



Glioblastoma (NCIT:C3058)



Neuroblastoma (NCIT:C3270)

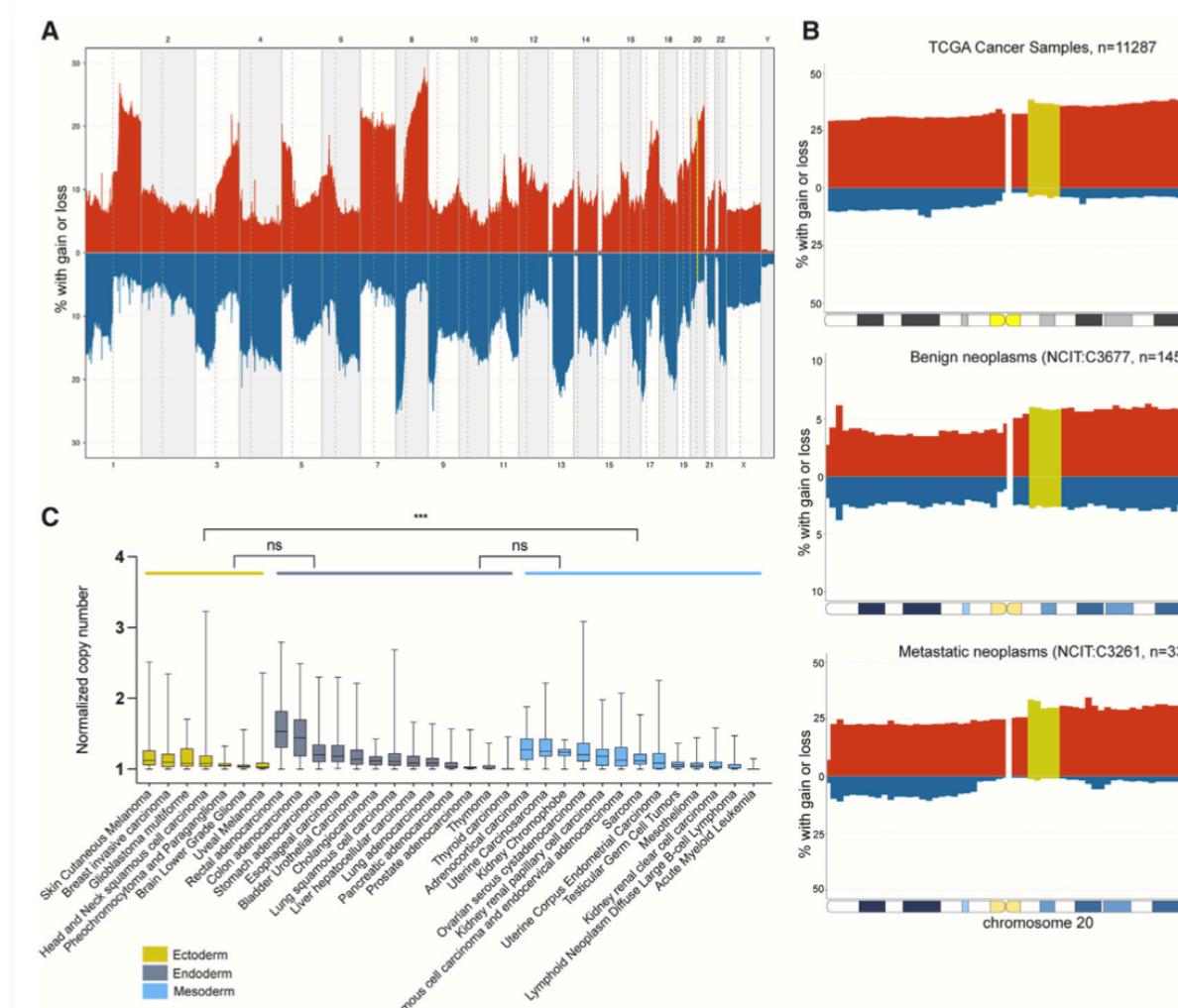


- Benign Urinary Tract Neoplasm: NCIT:C192667 (3 samples, 3 CNV profiles)
- Benign Kidney Neoplasm: NCIT:C4778 (95 samples, 95 CNV profiles)
- Benign Reproductive System Neoplasm: NCIT:C192667 (145 samples, 145 CNV profiles)
  - Benign Female Reproductive System Neoplasm: NCIT:C192667 (145 samples, 145 CNV profiles)
- Malignant Genitourinary System Neoplasms: NCIT:C192667 (20567 samples, 22154 CNV profiles)
  - Metastatic Malignant Genitourinary System Neoplasms: NCIT:C192667 (2 samples, 2 CNV profiles)
    - Metastatic Genitourinary System Carcinoma: NCIT:C192667 (2 samples, 2 CNV profiles)
  - Genitourinary System Carcinoma: NCIT:C192667 (19462 samples, 20921 CNV profiles)
    - Metastatic Genitourinary System Carcinoma: NCIT:C192667 (2 samples, 2 CNV profiles)
    - Female Reproductive System Carcinoma: NCIT:C192667 (5746 samples, 5974 CNV profiles)
    - Male Reproductive System Carcinoma: NCIT:C192667 (7022 samples, 7808 CNV profiles)
    - Urinary System Carcinoma: NCIT:C188667 (6694 samples, 7139 CNV profiles)
  - Recurrent Malignant Genitourinary System Neoplasms: NCIT:C192667 (3 samples, 3 CNV profiles)

# Progenetix Use

- CNV data is used e.g. as reference data in cancer genomics studies
- diagnosis specific CNV profiles serve as "fast look-up" in clinical genomics laboratories
- we loosely track publications in our literature database but there is no systematic check-back mechanism...

Example: 2025 article using Progenetix' *pgxRpi* Beacon/R interface to retrieve & visualize 117'587 cancer CNV profiles for a study into pluripotent stem cells' genomics



Progenetix References

arrayMap progenetix cancercellines

Articles Citing - or Using - Progenetix

This page lists articles which we found to have made use of, or referred to, the Progenetix resource ecosystem. These articles may not necessarily contain original case profiles themselves. Please contact us to alert us about additional articles you are aware of. Also, you can now directly submit suggestions for matching publications to the oncopubs repository on Github.

Filter

Publications (121)	Samples		
id	Publication	Genomes	pgx
PMID:38157850	Krivec N, Ghosh MS et al. (2024) Gains of 20q11.21 in human pluripotent stem cells: Insights from cancer research. ... Stem Cell Reports	0	0
PMID:37627037	Austin BK, Firooz A, Valafar H et al. (2023) An Updated Overview of Existing Cancer Databases and Identified Needs. Biology (Basel)	0	0
PMID:37393410	Liu SC, Wang CI, Liu TT, Tsang NM et al. (2023) A 3-gene signature comprising CDH4, STAT4 and EBV-encoded LMP1 for early diagnosis ... Discov Oncol	0	0

## Stem Cell Reports Review



OPEN ACCESS

### Gains of 20q11.21 in human pluripotent stem cells: Insights from cancer research

Nuša Krivec,<sup>1,2</sup> Manjusha S. Ghosh,<sup>1,2</sup> and Claudia Spits<sup>1,2,\*</sup>

<sup>1</sup>Research Group Reproduction and Genetics, Faculty of Medicine and Pharmacy, Vrije Universiteit Brussel, Brussels, Laarbeeklaan 103, 1090 Brussels, Belgium

<sup>2</sup>These authors contributed equally.

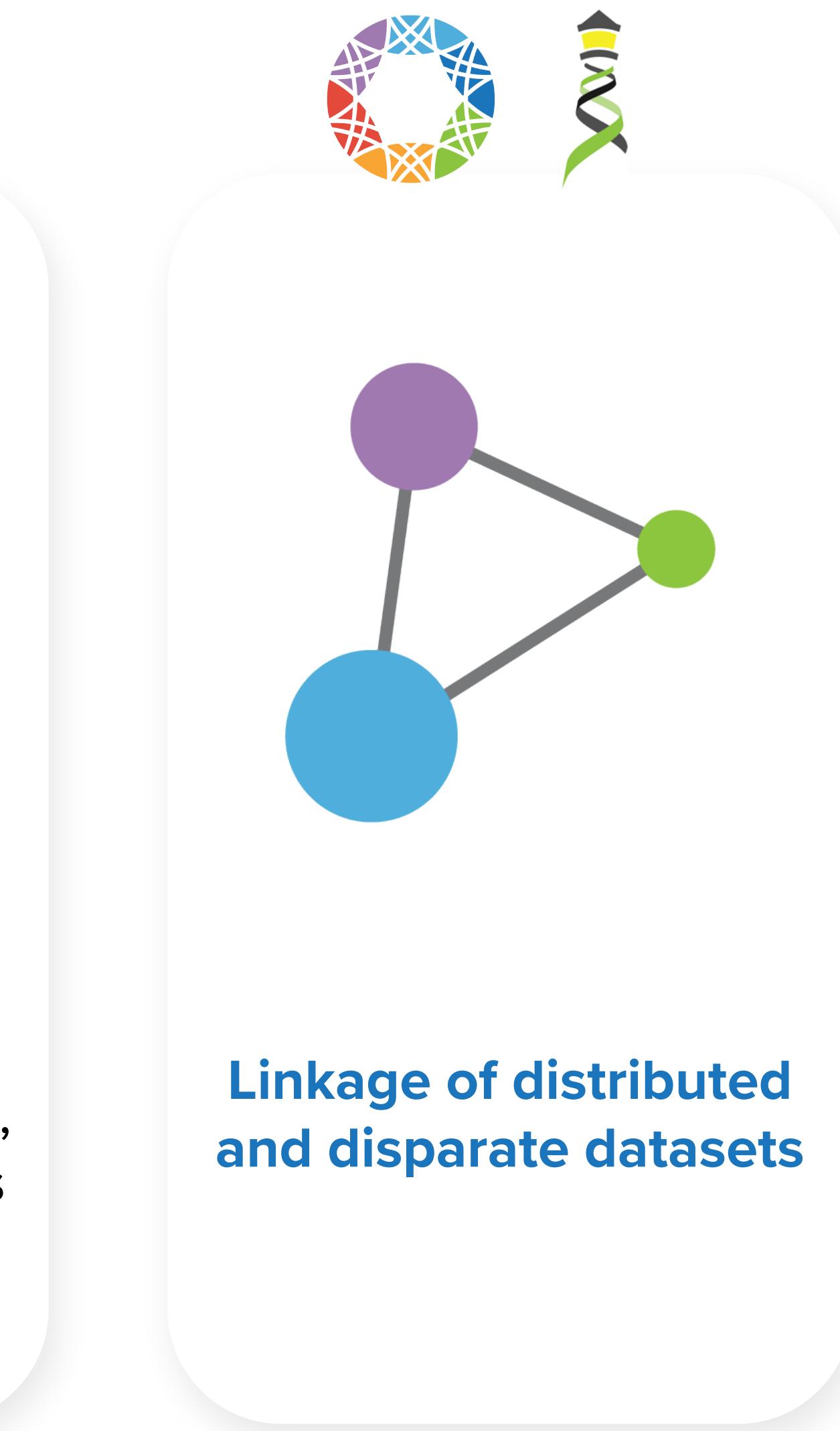
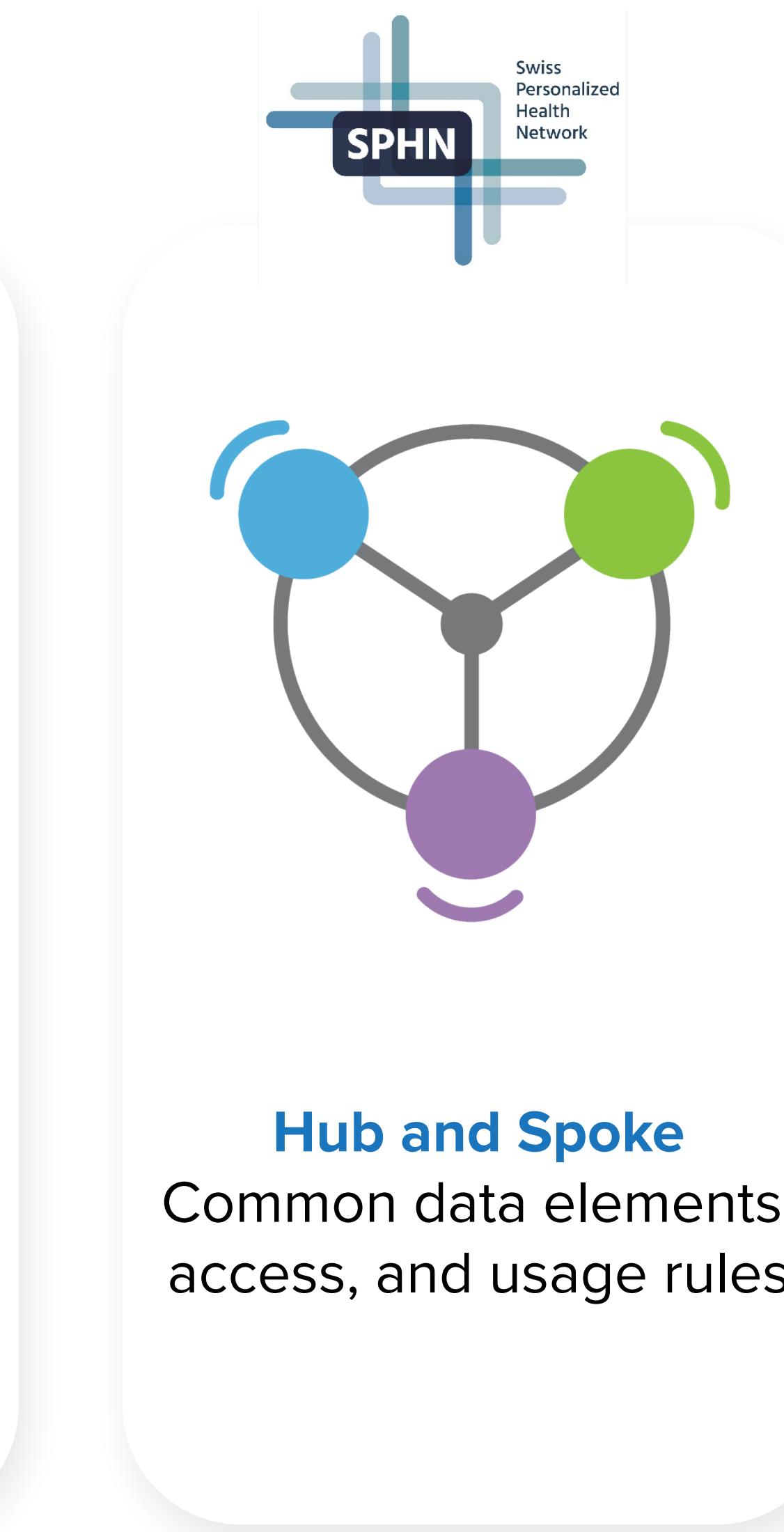
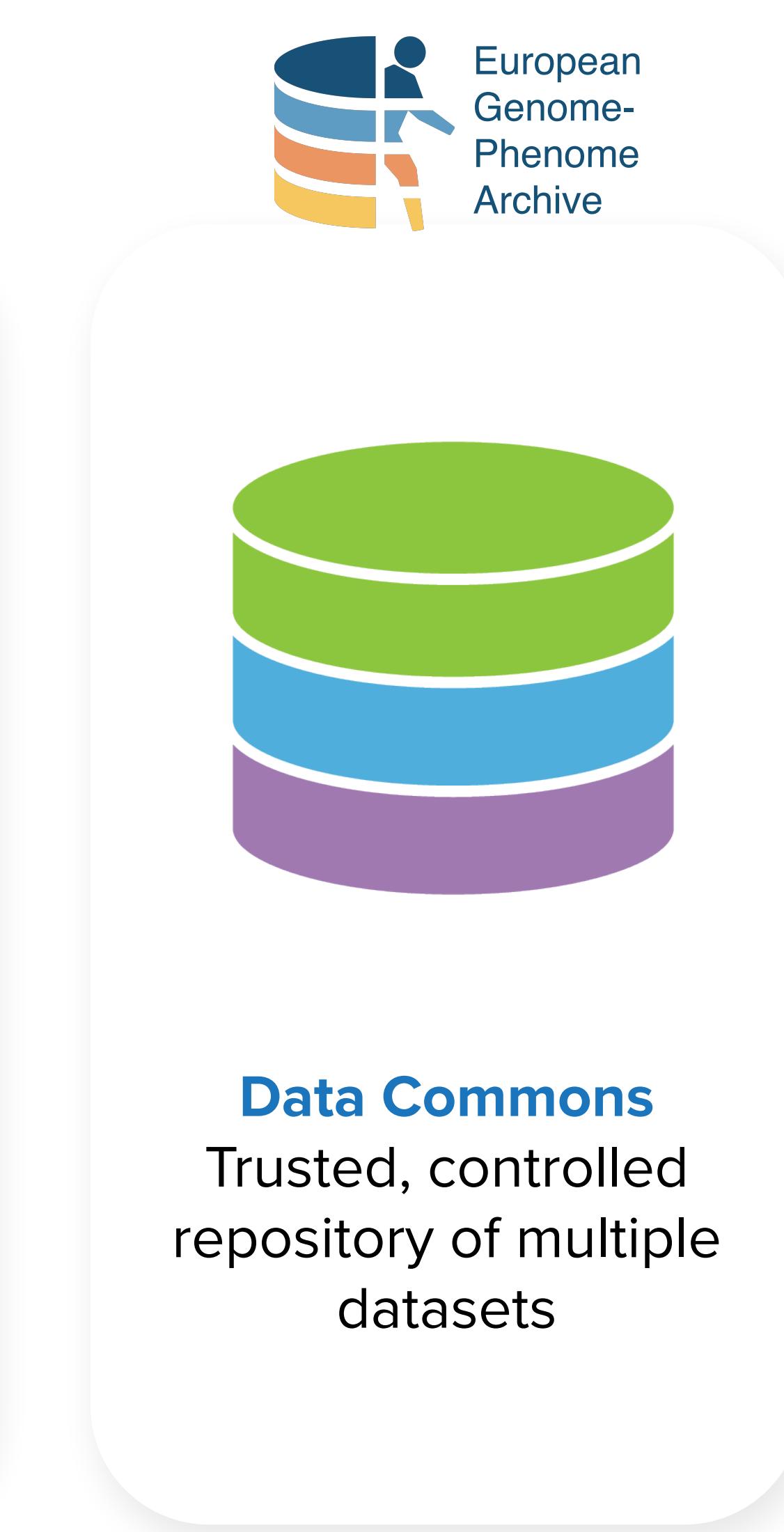
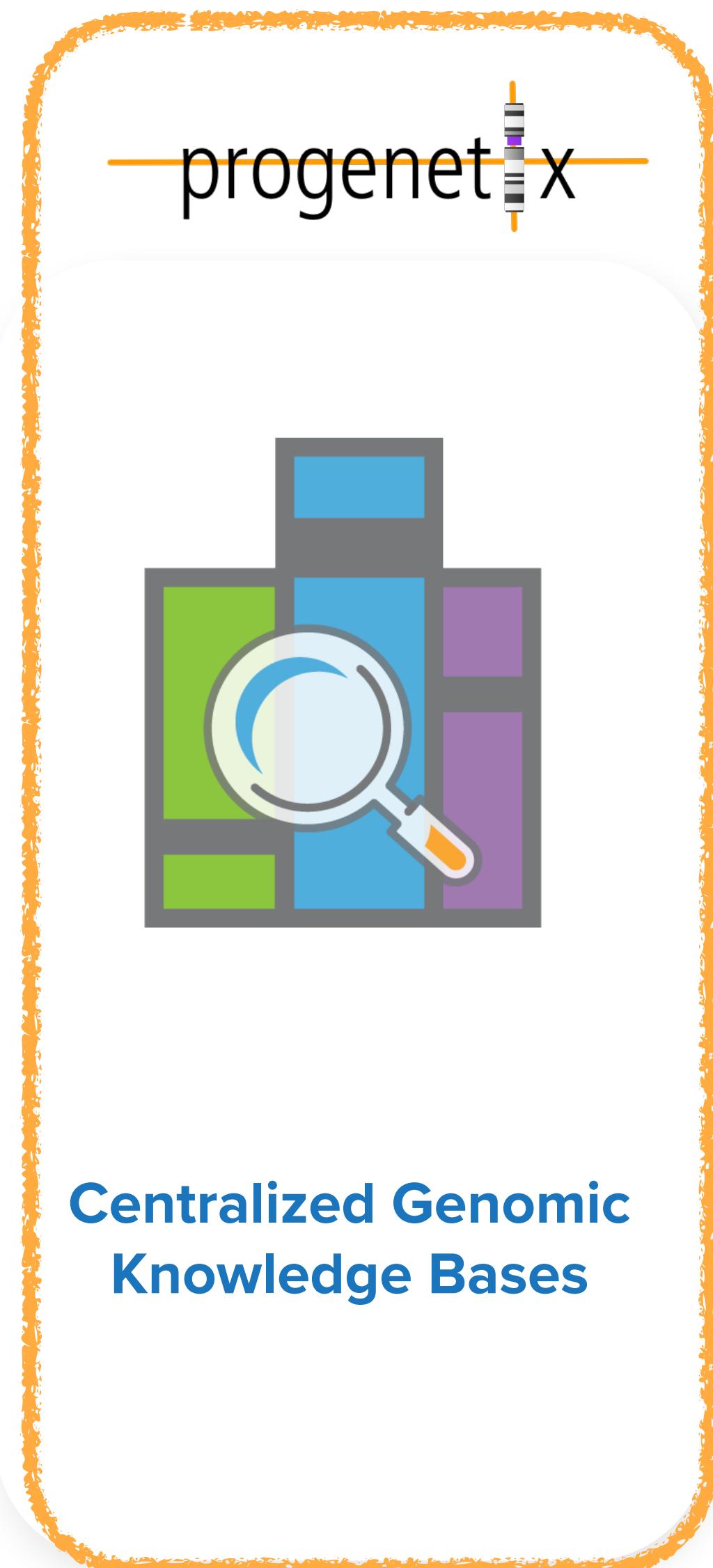
\*Correspondence: claudia.spits@vub.be  
<https://doi.org/10.1016/j.stemcr.2023.11.013>

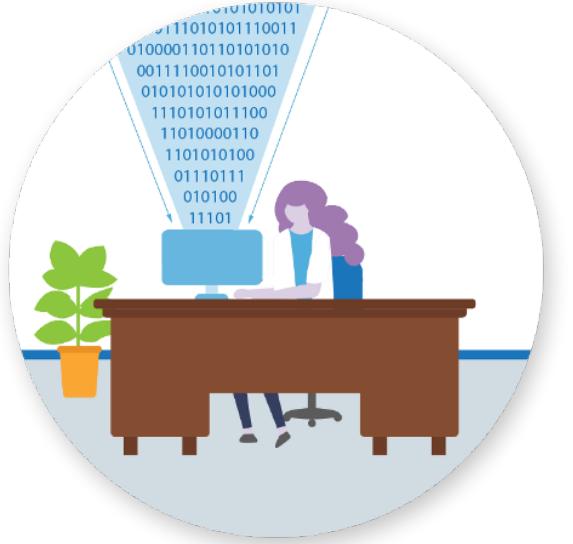
#### Figure 2. Copy-number alterations of human chromosome 20q11.21 in cancers

(A) Aggregated copy-number variation (CNV) data of 117,587 neoplasms (NCIT: C3262) from the Progenetix database (Huang et al., 2021) were plotted using R library pgxRpi. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079–35871578 is marked in moss green. NCIT, National Cancer Institute Thesaurus.

(B) Top to bottom: Aggregated CNV data of 11,287 TCGA cancer samples, 336 metastatic neoplasms (NCIT: C3261), and 1,455 benign neoplasms (NCIT: C3677) from the Progenetix database (Huang et al., 2021), respectively, were plotted using R library pgxRpi. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079–35871578 is marked in moss green.

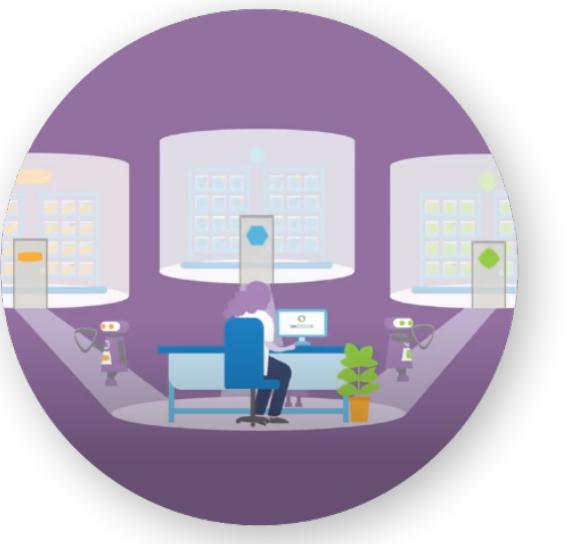
# Different Approaches to Data Sharing





## GA4GH STANDARDS

From Data Copying to  
Data Discovery & Visiting



- **Growing genomics data:** research, clinics and national projects
- **Data federation** allows utilization of rare and population related data **for research and precision medicine**
- The Global Alliance for Genomics and Health **GA4GH** started operation in 2014 to promote the “responsible sharing of genomic and health-related data” through **standards and policies**

→ UZH and SIB are founding members of GA4GH

**Cell Genomics**

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

Alex H. Wagner,<sup>1,2,\*</sup> Lawrence Babb,<sup>3,\*</sup> Gil Alterovitz,<sup>4,5</sup> Michael Baudis,<sup>6</sup> Matthew Brush,<sup>7</sup> Daniel Melissa Cline,<sup>10</sup> Malachi Griffith,<sup>11</sup> Obi L. Griffith,<sup>11</sup> Sarah E. Hunt,<sup>12</sup> David Kreda,<sup>13</sup> Jennifer M. Lee, Javier Lopez,<sup>16</sup> Eri Shaw Nyeonar,<sup>14</sup> Andrew D. Yates,<sup>11</sup> **Cell Genomics**

**INFORMATICS**

Human Mutation HGV WILEY

Beacon v2 and Beacon networks: A “lingua franca” for federated data discovery in biomedical genomics, and beyond

Jordi Rambla<sup>1,2</sup> | Michael Baudis<sup>3</sup> | Roberto Ariosa<sup>1</sup> | Tim Beck<sup>4</sup> | Lauren A. Froehlich<sup>5</sup> | Manuel Rueda<sup>6</sup> | Juha Törnroos<sup>7</sup>

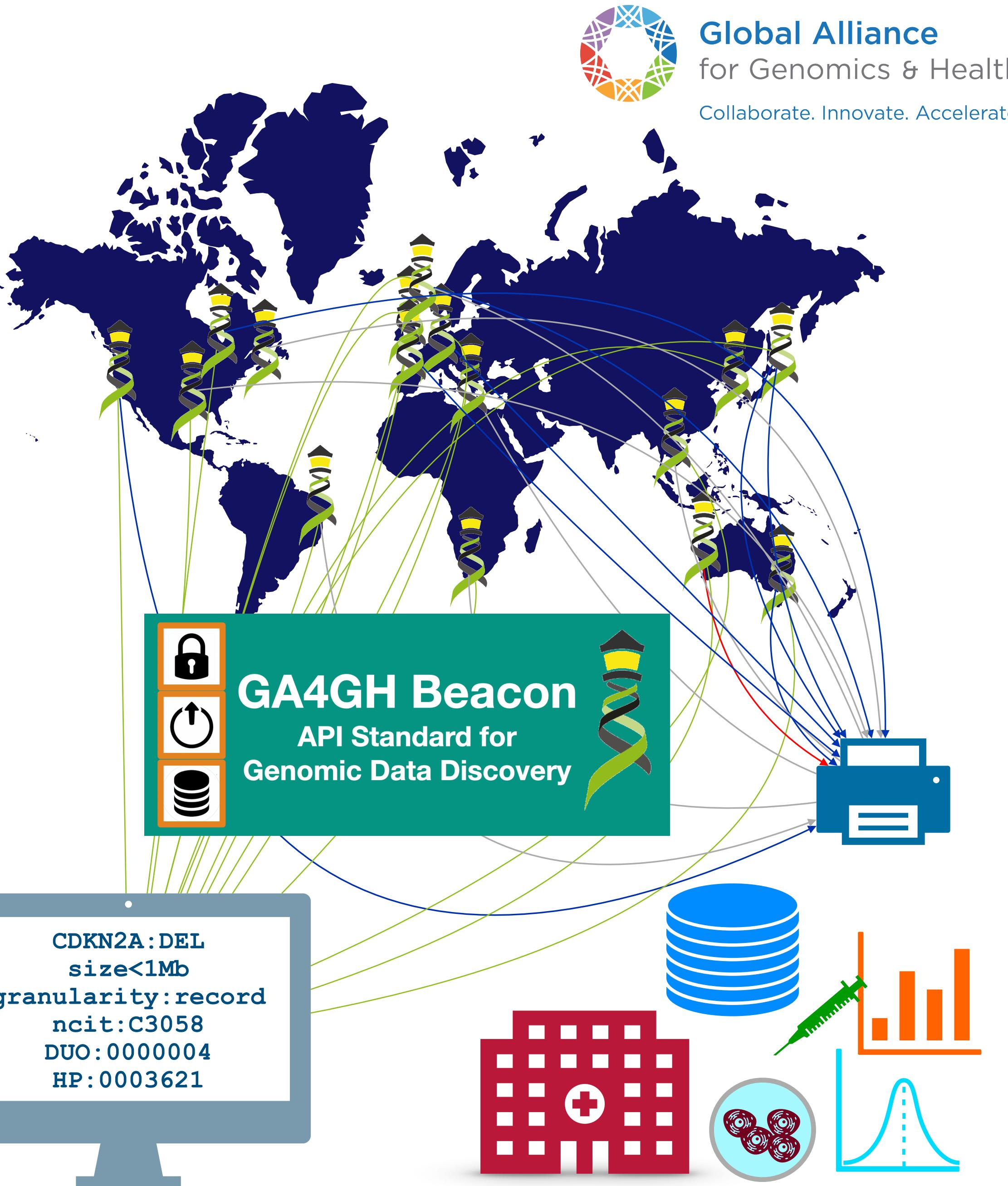
**Cell Genomics**

**Commentary**  
International federation of genomic medicine databases using GA4GH standards

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

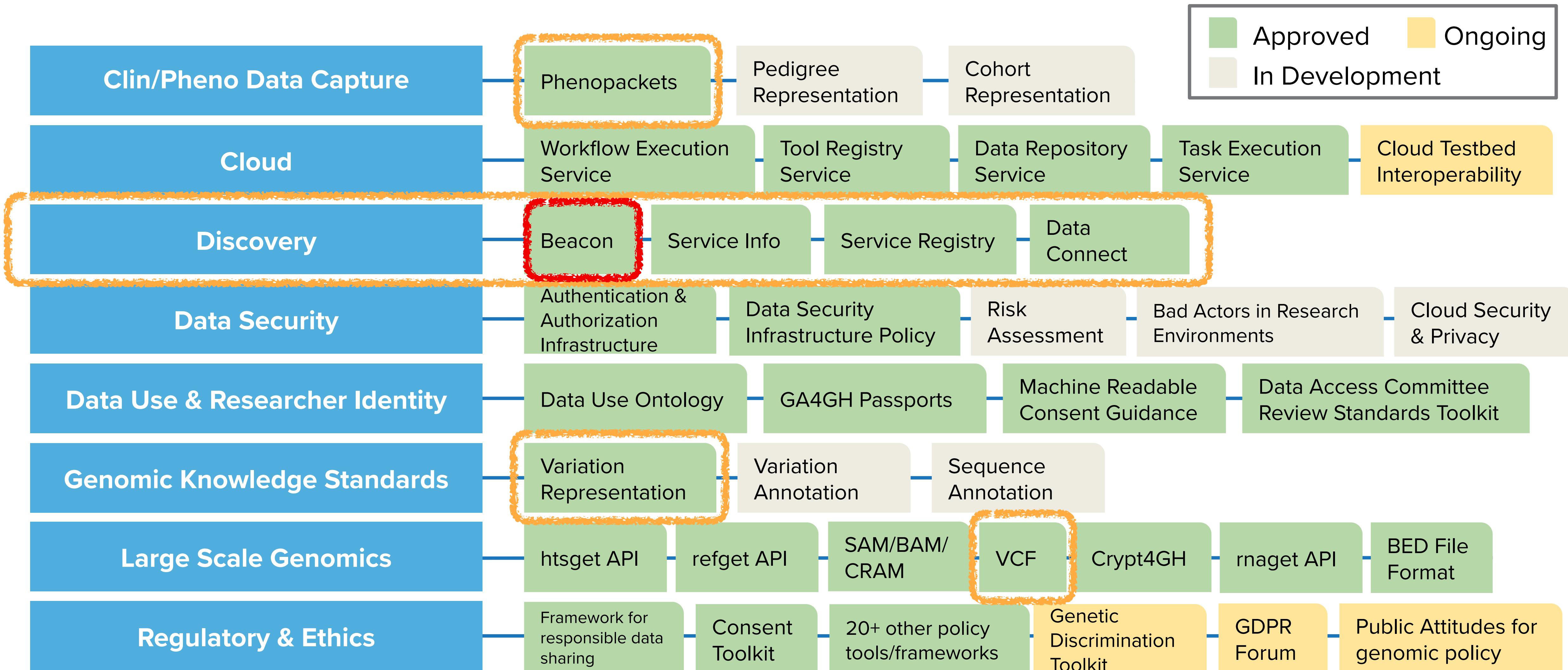
**Perspective**  
**GA4GH: International policies and standards for data sharing across genomic research and healthcare**

Heidi L. Rehm,<sup>1,2,47</sup> Angela J.H. Page,<sup>1,3,\*</sup> Lindsay Smith,<sup>3,4</sup> Jeremy B. Adams,<sup>3,4</sup> Gil Alterovitz,<sup>5,47</sup> Lawrence J. Babb,<sup>1</sup> Maximilian P. Barkley,<sup>6</sup> Michael Baudis,<sup>7,8</sup> Michael J.S. Beauvais,<sup>3,9</sup> Tim Beck,<sup>10</sup> Jacques S. Beckmann,<sup>11</sup> Sergi Beltran,<sup>12,13,14</sup> David Bernick,<sup>1</sup> Alexander Bernier,<sup>9</sup> James K. Bonfield,<sup>15</sup> Tiffany F. Boughtwood,<sup>16,17</sup> Guillaume Bourque,<sup>9,18</sup> Sarien R. Bowers,<sup>15</sup> Anthony J. Brookes,<sup>10</sup> Michael Brudno,<sup>18,19,20,21,38</sup> Matthew H. Brush,<sup>22</sup>



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

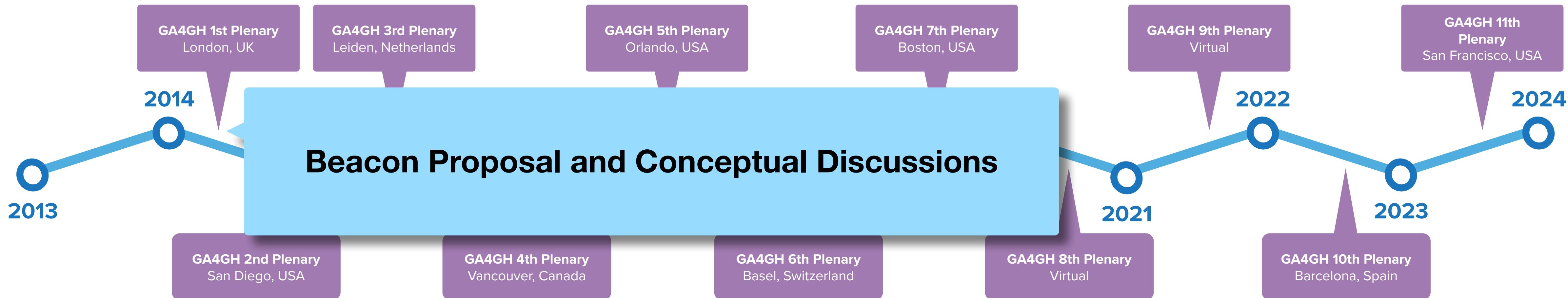
# Overview of GA4GH standards and frameworks



# GA4GH timeline



Global Alliance  
for Genomics & Health



Pre-launch	Building momentum	GA4GH Connect	Gap analysis	Strategic Refresh
 <p>73 partners sign a letter of intent to form an alliance</p>	 <p><b>Global Alliance</b> for Genomics &amp; Health Collaborate. Innovate. Accelerate.</p> <p><b>Formal launch of GA4GH</b></p> <p>Published <i>Framework for Responsible Sharing of Genomic and Health-Related Data</i></p> <p>Formed four working groups</p> <p>Developed three demonstration projects</p>	 <p>Launch of <b>GA4GH Connect</b> and Strategic Roadmap</p> <p>Formation of new organizational structure consisting of eight Work Streams and over twenty Driver Projects</p>	<p><b>Gap analysis</b> identifies three community imperatives</p> <ul style="list-style-type: none"><li> Interoperability and alignment</li><li> Implementation support</li><li> Engaging with healthcare and clinical standards</li></ul>	 <p><b>Strategic refresh</b> introduces updates to GA4GH to better meet the three community imperatives</p>

## Beacon v1 Development

2014

GA4GH founding event; Jim Ostell proposes Beacon concept including "more features ... version 2"

2015

- beacon-network.org aggregator created by DNAstack

2016

- Beacon v0.3 release
- work on queries for structural variants (brackets for fuzzy start and end parameters...)

2017

- OpenAPI implementation
- integrating **CNV parameters** (e.g. "startMin, statMax")
- Beacon v0.4 release in January; feature release for GA4GH approval process
- **GA4GH Beacon v1 approved** at Oct plenary

2018

- ELIXIR Beacon Network

2019



2020

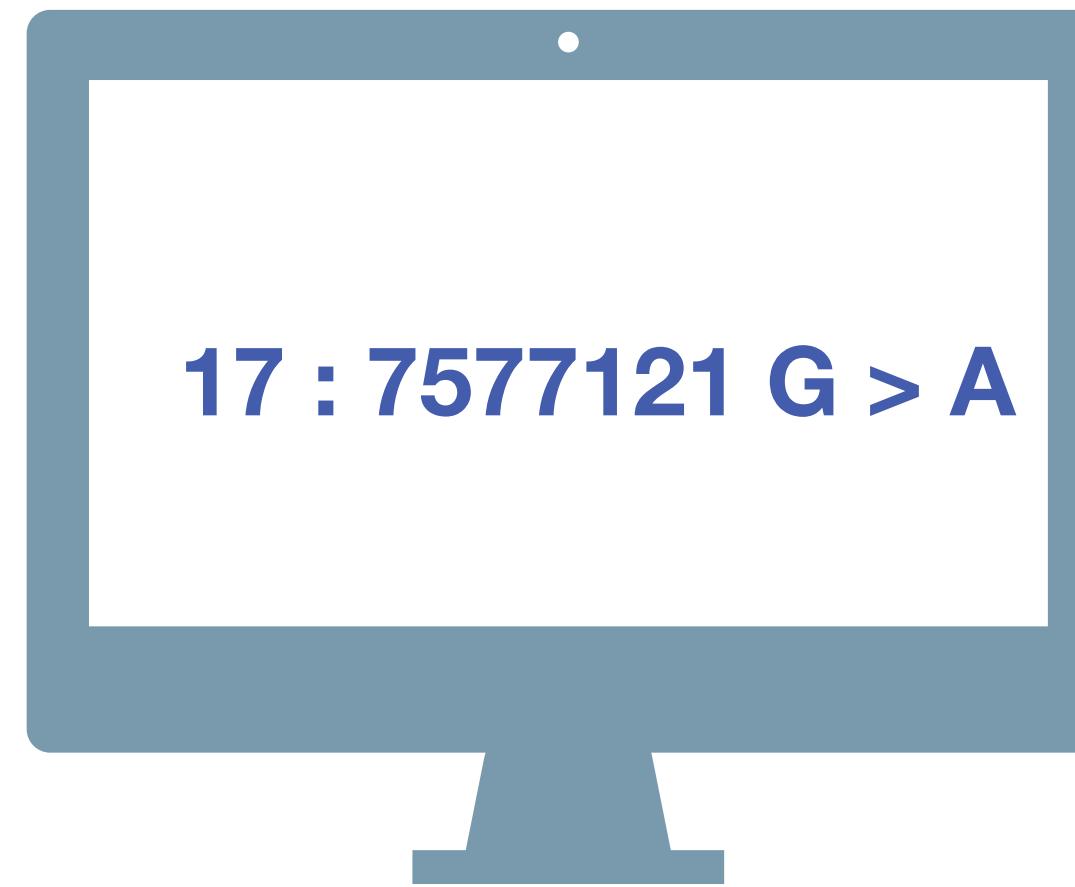
## Beacon v2 Development

2022

- Beacon+ concept implemented @ [progenetix.org](#)
- concepts from GA4GH Metadata (ontologies...)
- entity-scoped query parameters ("individual.age")
- Beacon+ demos "handover" concept
- Beacon hackathon Stockholm; settling on **filters**
- Barcelona goes Zurich developers meeting
- Beacon API v2 Kick off
- adopting "handover" concept
- "Scouts" teams working on different aspects - filters, genomic variants, compliance ...
- **framework + models** concept implemented
- range and bracket queries, variant length
- starting of GA4GH review process
- changes in default model, aligning with Phenopackets and VRS
- unified beacon-v2 code & docs repository
- **Beacon v2 approved** at April GA4GH Connect

## Related ...

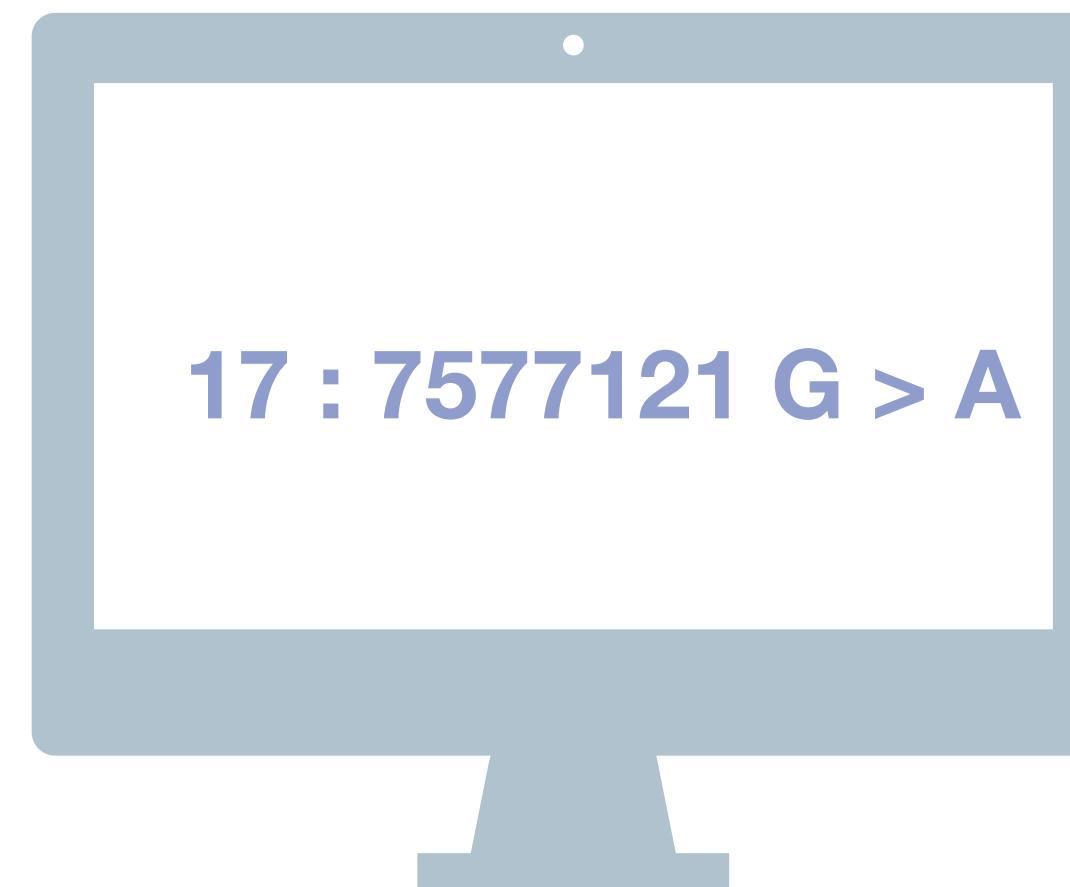
- ELIXIR starts Beacon project support
- GA4GH re-structuring (workstreams...)
- Beacon part of Discovery WS
- new Beacon website (March)
- Beacon publication at Nature Biotechnology
- Phenopackets v2 approved
- [docs.genomebeacons.org](#)



# Beacon

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

**YES | NO | \0**

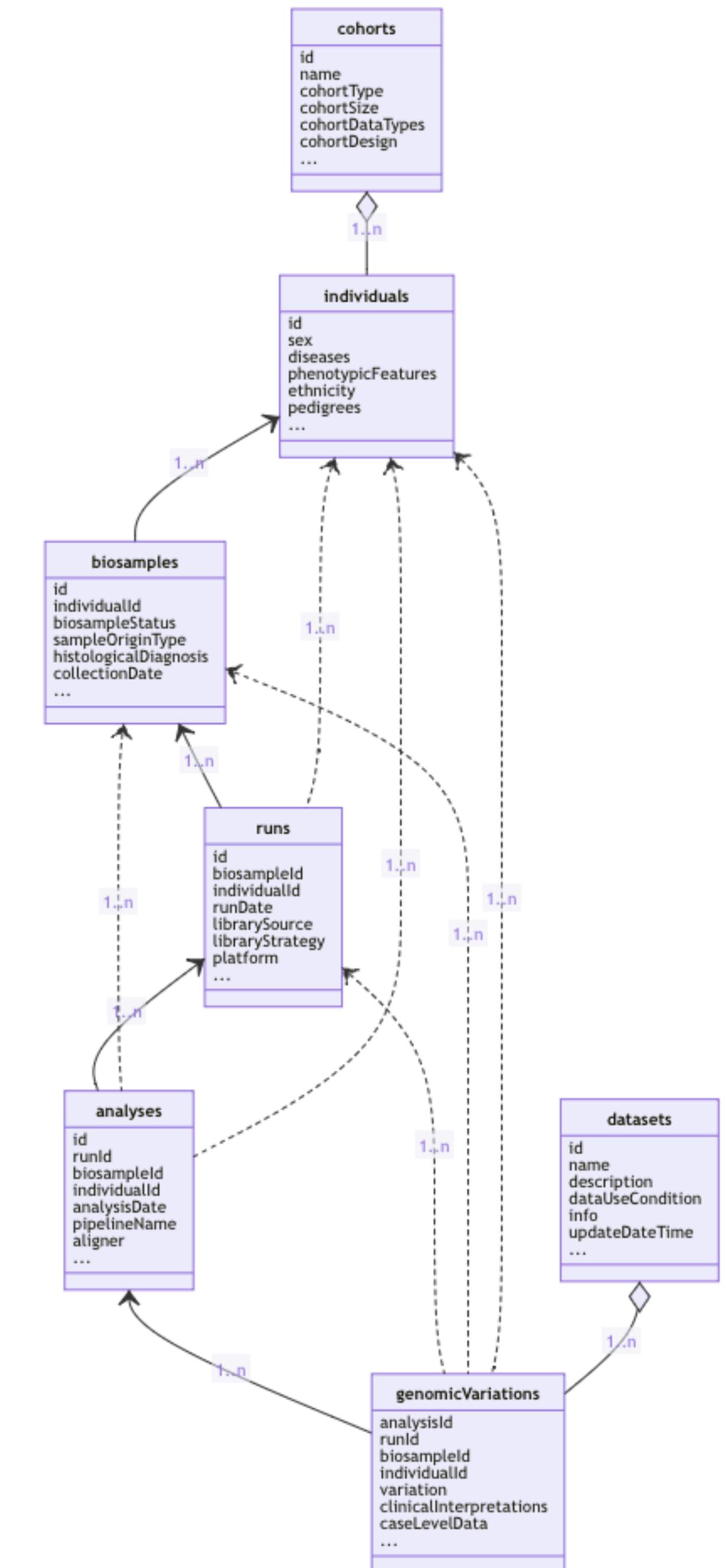


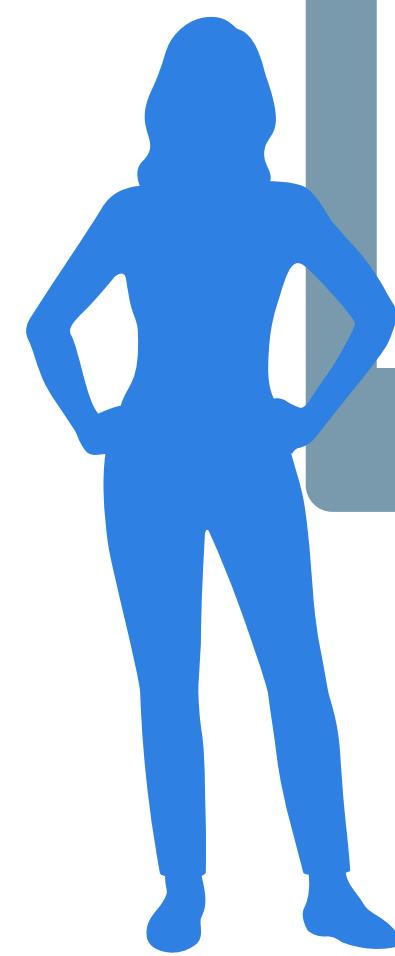
A **Beacon** answers a query for a specific genome variant against individual or aggregate genome collections

**YES | NO | \0**

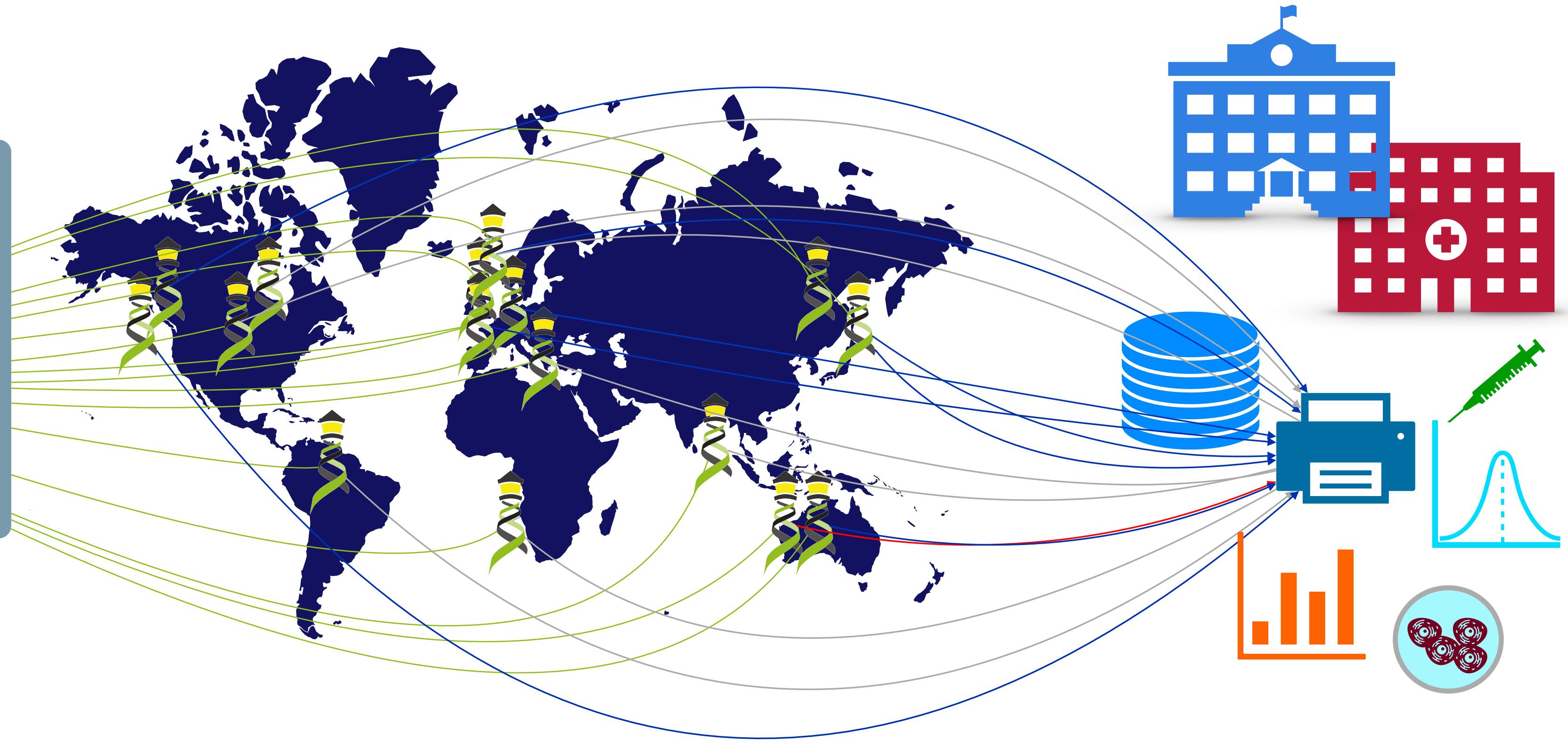
# Beacon Default v2 Model

- 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



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

**But... Can you really do that with the Beacon default model? Let's explore...**

# Request Components

## Deparsing the Beacon v2 Example

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

- query against genomic variations, no matter how they are stored
- copy number deletion, as indicated through the VCF symbolic allele **DEL** expression
  - preferably replaced by VRS v2 term
- 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**

# Request Components

## Deparsing the Beacon v2 Example

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

- Where can these parameters be applied in the current default model?
  - ▶ variant parameters [/g\\_variants](#)
  - ▶ histologicalDiagnosis [/biosamples](#)
  - ▶ dataUseConditions [/dataset](#)
  - ▶ ageOfOnset term [/individual](#)
- Such a request requires the application of filtering terms to other entities than the ones indicated by their entry type
  - Variant parameters always need **aggregation**
  - ... this does not prohibit simple implementations

# Standards Development & Implementation: CNV Terms

## in computational (file/schema) formats

```

- EFO:0030064
- EFO:0030067
  |- EFO:0030068
  \- EFO:0020073
    \- EFO:0030069
- EFO:0030070
  |- EFO:0030071
  \- EFO:0030072

```

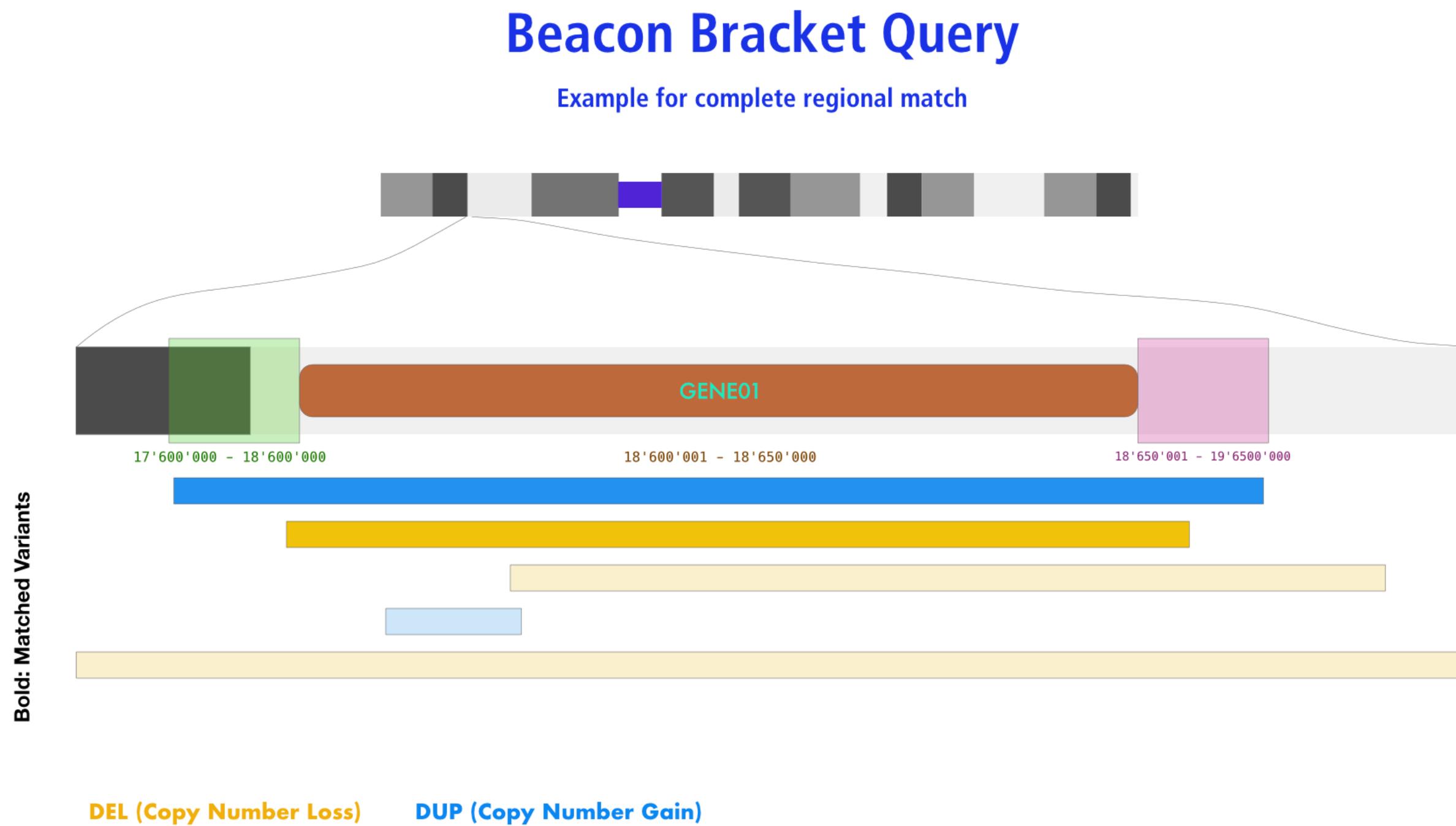
- Several competing concepts when annotating "systemic" variants such as CNV
- A vocabulary reflecting current consensus about CNV "level" status types served as basis for copyNumberVariation class in GA4GH VRS
- Progenetix provides a use case for CNV profiles of ~240'000 cancer CNV profiles encoded in VRS v2 and Beacon API support

GA4GH VRS v2	Beacon v2	VCF v4.4	SO
<b>gain</b> (EFO:0030070)	DUP or <a href="#">EFO:0030070</a>	DUP SVCLAIM=D	<a href="#">SO:0001742</a> copy_number_gain
<b>low-level gain</b> (EFO:0030071)	DUP or <a href="#">EFO:0030071</a>	DUP SVCLAIM=D	<a href="#">SO:0001742</a> copy_number_gain
<b>high-level gain</b> (EFO:0030072)	DUP or <a href="#">EFO:0030072</a>	DUP SVCLAIM=D	<a href="#">SO:0001742</a> copy_number_gain
<b>high-level gain</b> (EFO:0030072)	DUP or EFO:0030073 focal genome amplification	DUP SVCLAIM=D	<a href="#">SO:0001742</a> copy_number_gain
<b>loss</b> (EFO:0030067)	DEL or <a href="#">EFO:0030067</a>	DEL SVCLAIM=D	<a href="#">SO:0001743</a> copy_number_loss
<b>low-level loss</b> (EFO:0030068)	DEL or <a href="#">EFO:0030068</a>	DEL SVCLAIM=D	<a href="#">SO:0001743</a> copy_number_loss
<b>high-level loss</b> (EFO:0020073)	DEL or <a href="#">EFO:0020073</a>	DEL SVCLAIM=D	<a href="#">SO:0001743</a> copy_number_loss
<b>complete genomic loss</b> (EFO:0030069)	DEL or <a href="#">EFO:0030069</a>	DEL SVCLAIM=D	<a href="#">SO:0001743</a> copy_number_loss

# Variation Queries

## Bracket ("CNV") Query

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



### Beacon Query Types

Sequence / Allele   **CNV (Bracket)**   Genomic Range   Aminoacid   Gene ID   HGVS   Sam

#### Dataset

Test Database - examplez X | ▼

#### Chromosome i

9 (NC\_000009.12) | ▼

#### Variant Type i

EFO:0030067 (copy number deletion) | ▼

#### Start or Position i

21000001-21975098

#### End (Range or Structural Var.) i

21967753-23000000

#### Select Filters i

NCIT:C3058: Glioblastoma (100) X | ▼

#### Chromosome 9 i

21000001-21975098



### Query Database

#### Form Utilities

Gene Spans

Cytoband(s)

#### Query Examples

[CNV Example](#)

[SNV Example](#)

[Range Example](#)

[Gene Match](#)

[Aminoacid Example](#)

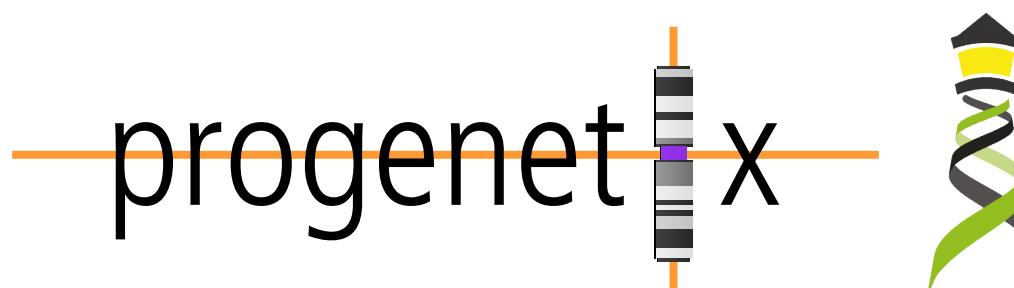
[Identifier - HeLa](#)

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.

# Beacon v2 Filters

# **Example: Use of hierarchical classification systems (here NCI neoplasm core)**

- Beacon v2 relies heavily on "filters"
    - ontology term / CURIE
    - alphanumeric
    - custom
  - Beacon v2 "filters" assumes inclusion of child terms when using hierarchical classifications
    - implicit *OR* with otherwise assumed *AND*
  - implementation of hierarchical annotations overcomes some limitations of "fuzzy" disease annotations



Beacon+ specific: Multiple term selection with OR logic

<input checked="" type="checkbox"/>	> <a href="#">NCIT:C4914: Skin Carcinoma</a>	213
<input type="checkbox"/>	> <a href="#">NCIT:C4475: Dermal Neoplasm</a>	109
<input checked="" type="checkbox"/>	> <a href="#">NCIT:C45240: Cutaneous Hematopoietic and Lymphoid Cell Neoplasm</a>	310

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



Female patients with focal high-level deletion in the CDKN2A locus in Glioblastoma tumors

```
GET /beacon/individuals/  
?variantType=EFO%3A0020073  
&referenceName=NC_000009.12  
&start=21000000,21975098  
&end=21967753,23000000  
&filters=NCIT%3AC3058,NCIT%3AC16576
```

Server-side intersect of

- variants from genomic bracket query
- histology from biosample (NCIT:C3058)
- sex from individual (NCIT:C16576)

# **ELIXIR Beacon Network**

# Begriffsbestimmung

## The right expressions help to conceptualize...

- **Beacon:** The protocol/API, with framework and default model
- **beacon:** Implementation of Beacon
  - using the Beacon v2 framework & supporting at minimum boolean responses
  - suggested support of Beacon v2 default model but can choose other
- Beacon **Aggregator:** service distributes queries to beacons and aggregates responses into a single Beacon response
  - potential to liftover genomes, remap filtering terms, translate between protocol versions...
  - entry point to or potentially itself node in a ...
- Beacon **Network:** Set of beacons with shared entry point for distributed queries and aggregated response delivery
  - "true" beacon networks should have managed aspects - scope, term use...
  - networks may combine mixes of internal (protected, rich data, additional extensions...) and external interfaces

# WP4 Beacon Network

## Leads:

Jordi Rambla (ES)

Michael Baudis (CH)

Jaakko Leinonen (FI)



## Objectives

This Work Package will

- Maintain an **operational Beacon Network service**, with ELIXIR lead participation, including a service level target for availability and user service and incident response.
- Deliver transparent and responsive governance structures that appropriately represent stakeholders, delineate strategy, and **react to user feedback**.
- Provide an approach to change management which ensures **ongoing development** of the service through other activities is integrated into the live service with minimum impact on existing users and dependencies.



- ELIXIR Beacon Network has been *prototyped* in a previous IS
- This CoS will:
  - Contribute to the costs of running a service in production
  - Solidify the governing bodies
    - Strategic
    - Operational
  - While keeping the running costs to a minimum
  - Allowing for take over by other partners if ever necessary
- Challenges
  - Relies on partner willingness to run the service
    - Only a part of the operational costs are covered
    - Additional funds should be guaranteed by the partners
  - Funding for development or **software maintenance** is not included, but these activities are **clearly necessary**



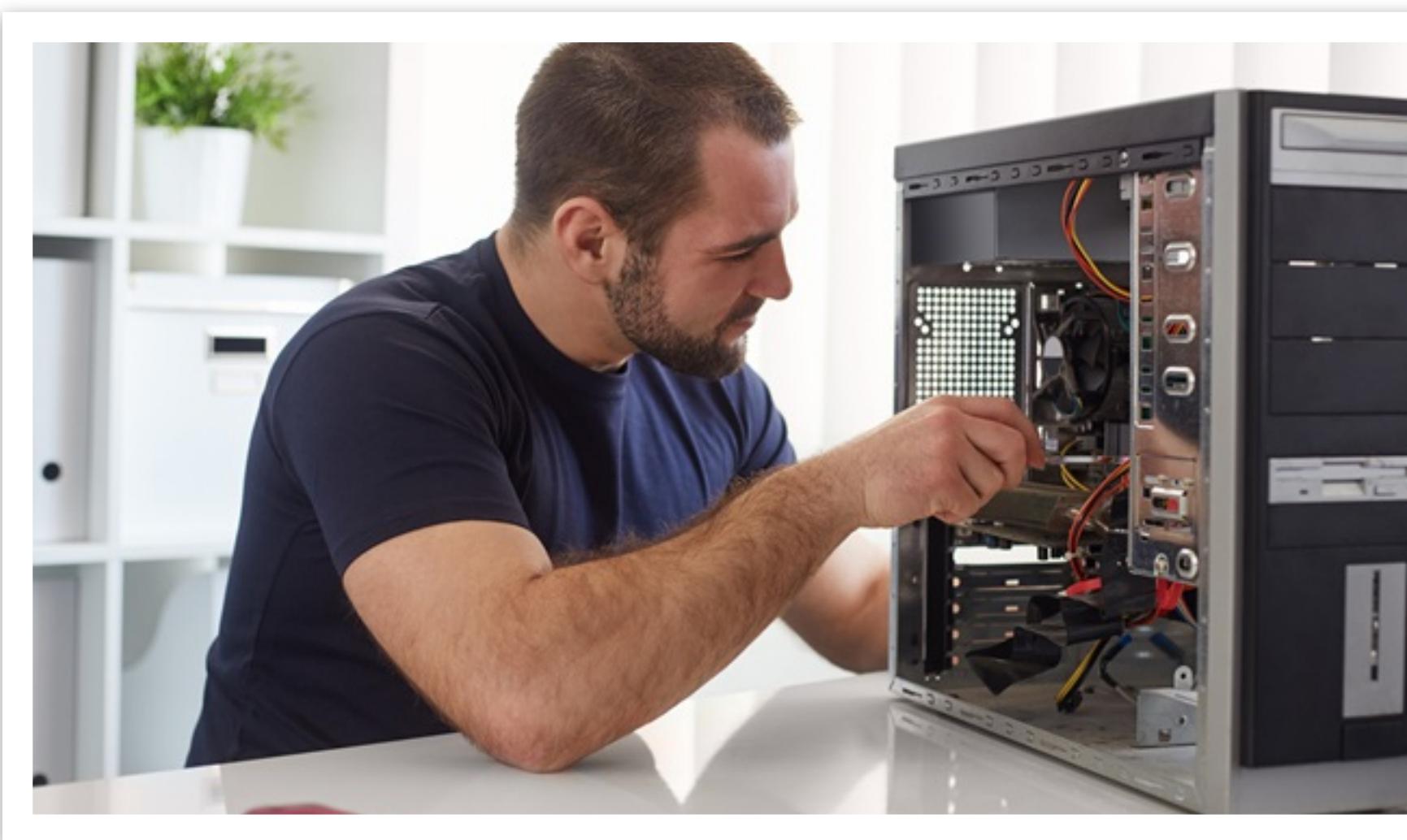
- ELIXIR Beacon Network in alignment with core stakeholders and external contributors to provide a working service with maximum capability profile
  - as GA4GH standard external input but also requirements to be accommodated
- importance of partner engagement since stakeholders (e.g. GDI ...) currently lack **live** multi-node data scenarios but have a projected need
- high relevance of further **protocol development** to accommodate use cases **within** protocol specifications
  - RNAseq, pathogens, imaging ...
  - massive opportunity to network people and data across domains through shared framework
- limited resources for "engage and ingest" work which is relevant for the global reach and - conversely - recognition and use of this ELIXIR standard and service

**Beacon is a core ELIXIR asset. The Beacon Network provides a unique infrastructure service and is an attractor for cross-domain engagement**



# Beacon v2 deployment

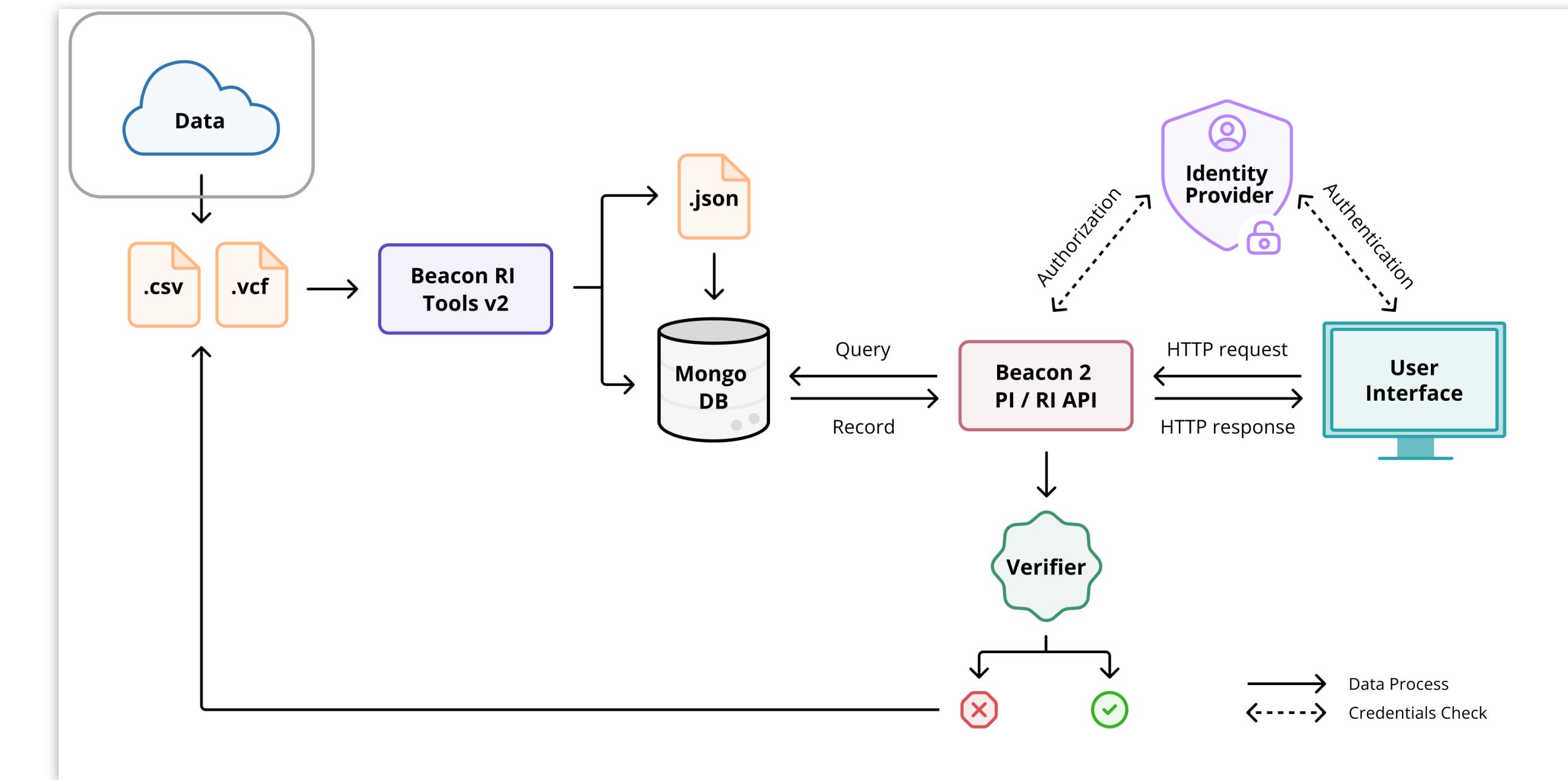
## Build it yourself



### Beacon v2 API

<https://github.com/ga4gh-beacon/beacon-v2>

## Toolkit for production environments



### Beacon v2 Production Implementation (released Oct 2024)

<https://github.com/ga4gh-beacon/beacon-v2>

# bycon Beacon

## Implementation driven standards development

- Progenetix' Beacon+ has served as implementation driver since 2016
- the *bycon* package is used to prototype advanced Beacon features such as
  - structural variant queries
  - data handovers
  - Phenopackets integration
  - variant co-occurrences
  - ...

Beacon protocol response verifier at time of GA4GH approval Spring 2022

Beacon v2 GA4GH Approval Registry

Beacons:    

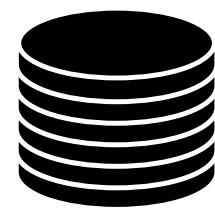
Category	EGA	progenetix	cnag	University of Leicester
BeaconMap	Green	Green	Green	Green
Bioinformatics analysis	Green	Green	Green	Green
Biological Sample	Green	Red	Red	Green
Cohort	Green	Green	Green	Green
Configuration	Green	Green	Green	Green
Dataset	Green	Red	Red	Green
EntryTypes	Green	Green	Green	Green
Genomic Variants	Green	Green	Green	Green
Individual	Green	Red	Red	Green
Info	Green	Red	Red	Green
Sequencing run	Green	Green	Green	Green

Legend:  Matches the Spec  Not Match the Spec  Not Implemented

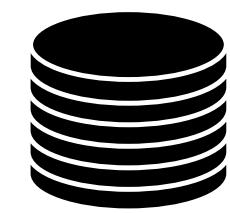
# *bycon* based Beacon+ Stack

progenetix

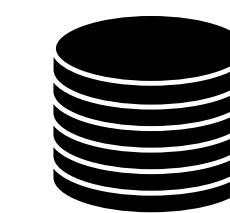
- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
  - ▶ [pubmed:10027410](#), [NCIT:C3222](#), [pgx:cohort-TCGA](#), [pgx:icdom-94703...](#)
  - ▶ precomputed frequencies per collection informative e.g. in form autfills
- *querybuffer* stores id values of all entities matched by a query and provides the corresponding **accessid** for **handover** generation
- complete query aggregation; i.e. individual queries are run against the corresponding entities and ids are intersected
  - retrieval of any entity, e.g. all individuals which have queried variants analyzed on a given platform
  - allows multi-variant queries, i.e. all bio samples or individuals which had matches of all of the individual variant queries



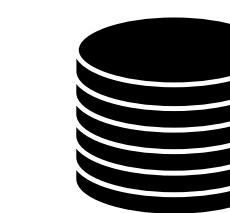
variants



analyses



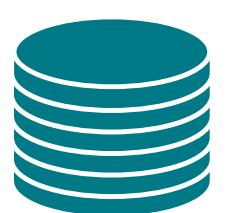
biosamples



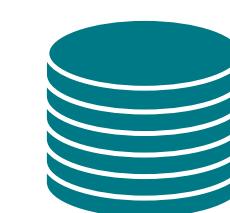
individuals



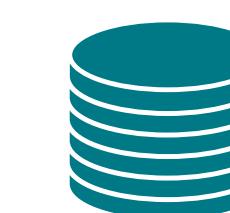
collations



geolocs



genespans

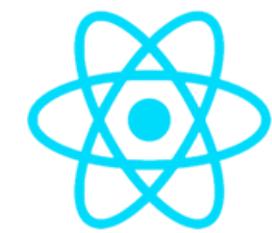


qBuffer

Entity collections

Utility collections

[github.com/progenetix/bycon](https://github.com/progenetix/bycon)



React

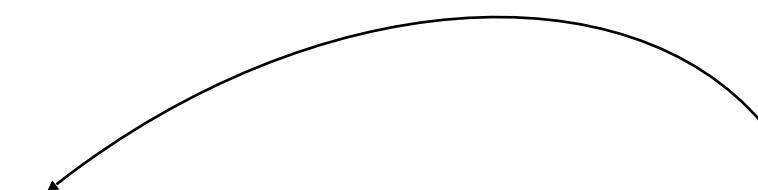




## Beacon as a global standard



### Beacon Scouts



### Real-world needs

Cancer

Common diseases

Rare Diseases

...

- **Beacon Filters** – improve current filter solutions
- **Beacon Cohorts** – develop aggregated request and response (e.g. counts by sex and age)
- **Beacon Variants** – expand specification to cover new use cases and typed queries
- **Beacon Dev** – improve API (cleaning code, GitHub issues)
- **Beacon Matchmaking** – implementation in matchmaking use cases

# Beacon Scouts

## Finding the Paths to Beacon's Future

### ● Genomic Variation Scouts

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

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

GA4GH Beacon Genomic Variation Query Standards

Search GitHub elixir

## Beacon VQS Requests

The `VQSRequest` type represents the generic collection of variant parameters supported in Beacon v2+ requests. These include parameters with close alignment to VRS v2 concepts and replacing some Beacon v1/v2 generics with tighter definitions (e.g. `referenceAccession` instead of `referenceName` and `accession` or `copyChange` for a specific subset of former `variantType` values) but also keep some concepts beyond VRS scope or specifically geared towards query applications (`geneId`, `sequenceLength`)

For the parameter definitions please see the [requestParameterComponents page](#).

### VQSRequest Parameters

```
requestProfileId: ./requestParameterComponents.yaml#/defs/RequestProfileId
referenceAccession: ./requestParameterComponents.yaml#/defs/RefgetAccession
start: ./requestParameterComponents.yaml#/defs/SequenceStart
end: ./requestParameterComponents.yaml#/defs/SequenceEnd
sequence: ./requestParameterComponents.yaml#/defs/Sequence
copyChange: ./requestParameterComponents.yaml#/defs/CopyChange
adjacencyAccession: ./requestParameterComponents.yaml#/defs/AdjacencyAccession
adjacencyStart: ./requestParameterComponents.yaml#/defs/AdjacencyStart
adjacencyEnd: ./requestParameterComponents.yaml#/defs/AdjacencyEnd
repeatSubunitCount: ./requestParameterComponents.yaml#/defs/RepeatSubunitCount
repeatSubunitLength: ./requestParameterComponents.yaml#/defs/RepeatSubunitLength
geneId: ./requestParameterComponents.yaml#/defs/GeneId
aminoacidChange: ./requestParameterComponents.yaml#/defs/AminoacidChange
genomicAlleleShortForm:
./requestParameterComponents.yaml#/defs/GenomicAlleleShortForm
sequenceLength: ./requestParameterComponents.yaml#/defs/SequenceLength
vrsType: ./requestParameterComponents.yaml#/defs/VRStype
```

Table of contents

- VQSRequest Parameters
- Beacon v2+/VQS "VRSified"
- Request Examples
  - Copy number gains involving the whole locus chr2:54,700,000-63,900,000
  - Focal high-level deletion involving the CDKN2A locus
  - Find t(8;14)(q24;q32) translocations
  - CAG repeat in the first exon of the huntingtin gene (HTT)
  - CAG repeat in the first exon of the huntingtin gene (HTT)
  - CGG trinucleotide repeat expansion in the FMR1 gene
  - Query for a focal deletion involving TP53

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

# VQS - Variant Query Standard

## VRS aligned typed queries

- Typed queries
  - query schemas with defined set of (required and optional) parameters
    - ▶ can be verified
    - ▶ profile ids can be advertised by beacons
- VRS aligned
  - explicit reference to VRS types
  - ... but differ in (some) parameter use since query NE representation
- Expanding library
  - adjacency, repeats...

```
vQScopyChangeRequest:  
  description: |-  
    A typical Beacon v2.n request for copy number variation.  
    Approximate positions for CNV start and end regions are of  
    `Range` type. The `copyChange` parameter indicates the  
    genomic copy number (pls. refer to the class definition).  
  type: object  
  properties:  
    requestProfile:  
      const: VQScopyChangeRequest  
    referenceAccession:  
      $ref: "./requestParameterComponents.yaml#/defs/referenceAccession"  
    startRange:  
      $ref: "./requestParameterComponents.yaml#/defs/startRange"  
    endRange:  
      $ref: "./requestParameterComponents.yaml#/defs/endRange"  
    copyChange:  
      $ref: "./requestParameterComponents.yaml#/defs/copyChange"  
    sequenceLength:  
      $ref: "./requestParameterComponents.yaml#/defs/sequenceLength"  
  vrsType:  
    const: CopyNumberChange  
  required:  
    - requestProfile  
    - referenceAccession  
    - startRange  
    - endRange  
    - copyChange
```

```
requestProfile: VQScopyChangeRequest  
referenceAccession: refseq:NC_00002.12  
start:  
  o 21000001  
  o 21975098  
end:  
  o 21967753  
  o 23000000  
copyChange: EFO:0020073  
vrsType: CopyNumberChange
```

```
requestProfile: VQSadjacencyRequest  
referenceAccession: refseq:NC_00008.11  
start: 116700000  
end: 145138636  
adjacencyAccession: refseq:NC_00014.9  
adjacencyStart: 89300000  
adjacencyEnd: 107043718  
vrsType: Adjacency
```

```
requestProfile: VQSSequenceRepeatRequest  
geneId: HTT  
repeatSubunitLength: 3  
sequenceLength:  
  o 105  
  o 750  
vrsType: ReferenceLengthExpression
```

# Aggregation / Summaries

# Data Summaries

- Aggregated results offer a way for beacons to
  - ▶ de-couple information about the resource's data content from samples and individuals
  - ▶ highlight relevant features and data
  - ▶ support performant front-ends/dashboards
- *Distributions* for one and two-dimensional counts or (potentially) value representations
  - ▶ value selection, binning/bucketing through parameters
- While custom dashboards are standard feature of genomic data resources, Beacon protocol supports empowerment

**Federation => Aggregation<sup>^2</sup>**

```
- concepts:  
- property: histological_diagnosis.id  
  scope: biosample  
  description: Histological diagnoses in matched biosamples  
  distribution:  
- conceptValues:  
  - id: NCIT:C132256  
    label: Unspecified Tissue  
    count: 18313  
- conceptValues:  
  - id: NCIT:C3512  
    label: Lung Adenocarcinoma  
    count: 13657  
- conceptValues:  
  - id: NCIT:C4017  
    label: Breast Ductal Carcinoma  
    count: 9634  
- conceptValues:  
  - id: NCIT:C2919  
    label: Prostate Adenocarcinoma  
    count: 9115  
...  
- id: ageAtDiagnosis  
  label: Age at Diagnosis  
  sorted: true  
  concepts:  
- property: index_disease.age  
  scope: individual  
  splits:  
  - P0D  
  - P1Y  
  - P18M  
  - P18Y  
  - P120Y  
  description: Follow-up time after diagnosis  
  distribution:  
- conceptValues:  
  - id: '>= P6M'  
    label: '>= P6M'  
    count: 3872  
- conceptValues:  
  - id: '>= P1Y'  
    label: '>= P1Y'  
    count: 6041  
...
```

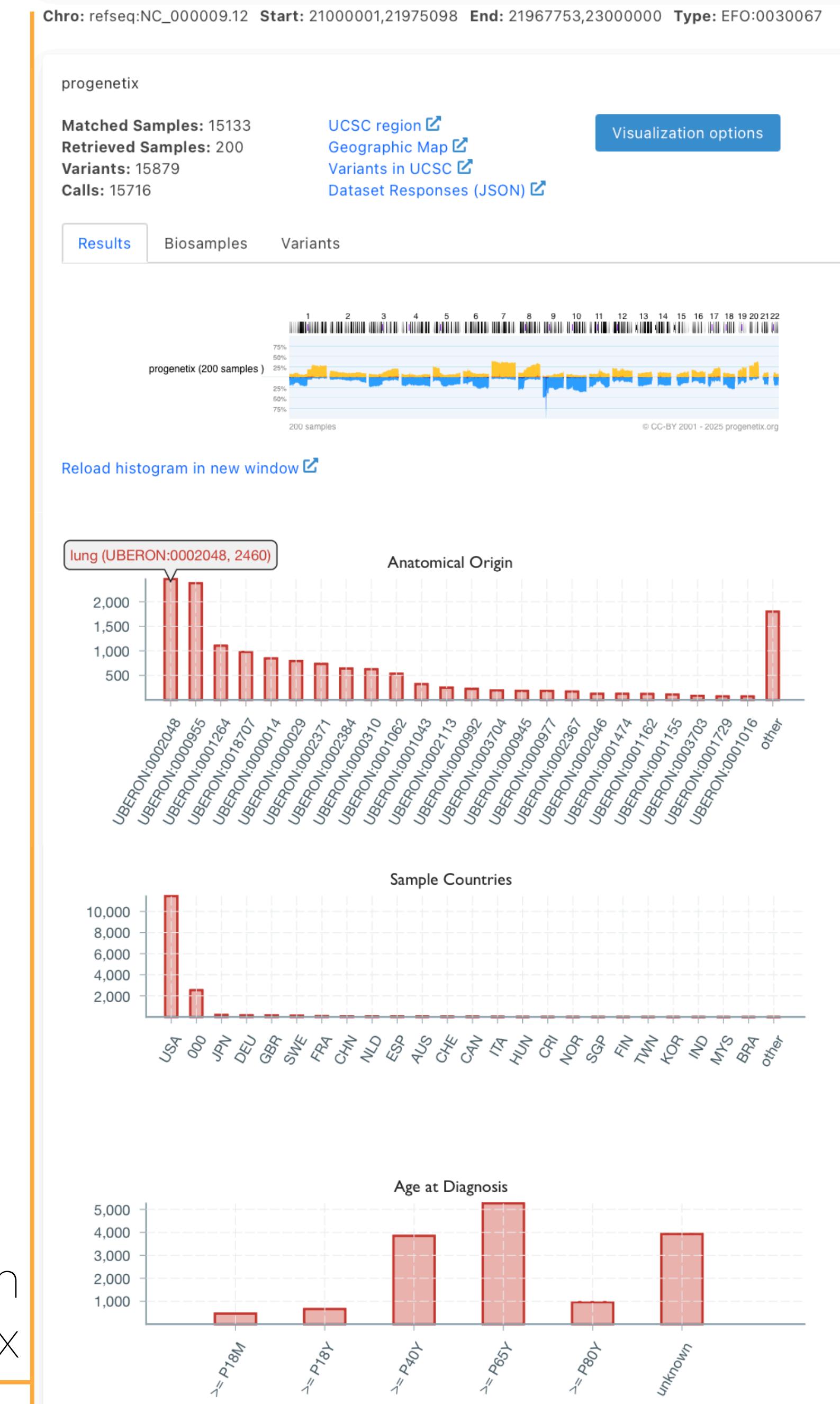
# Data Summaries

## Representation of aggregated results in Beacon responses

- Aggregated results offer a way for beacons to
  - ▶ de-couple information about the resource's data content from samples and individuals
  - ▶ highlight relevant features and data
  - ▶ support performant front-ends/dashboards
- *Distributions* for one and two-dimensional counts or (potentially) value representations
  - ▶ value selection, binning/bucketing through parameters
- While custom dashboards are standard feature of genomic data resources, Beacon protocol supports empowerment

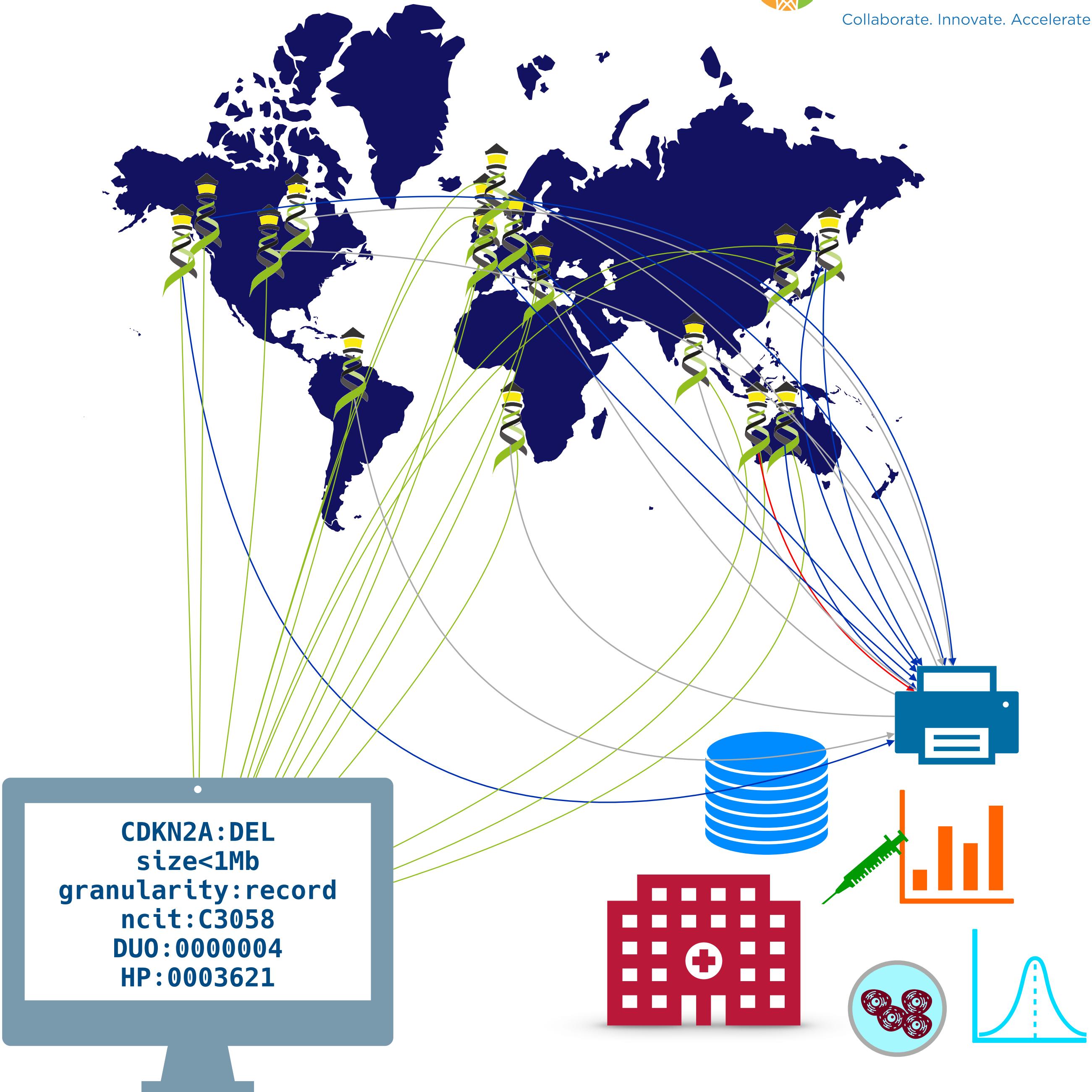
Federation => Aggregation<sup>^2</sup>

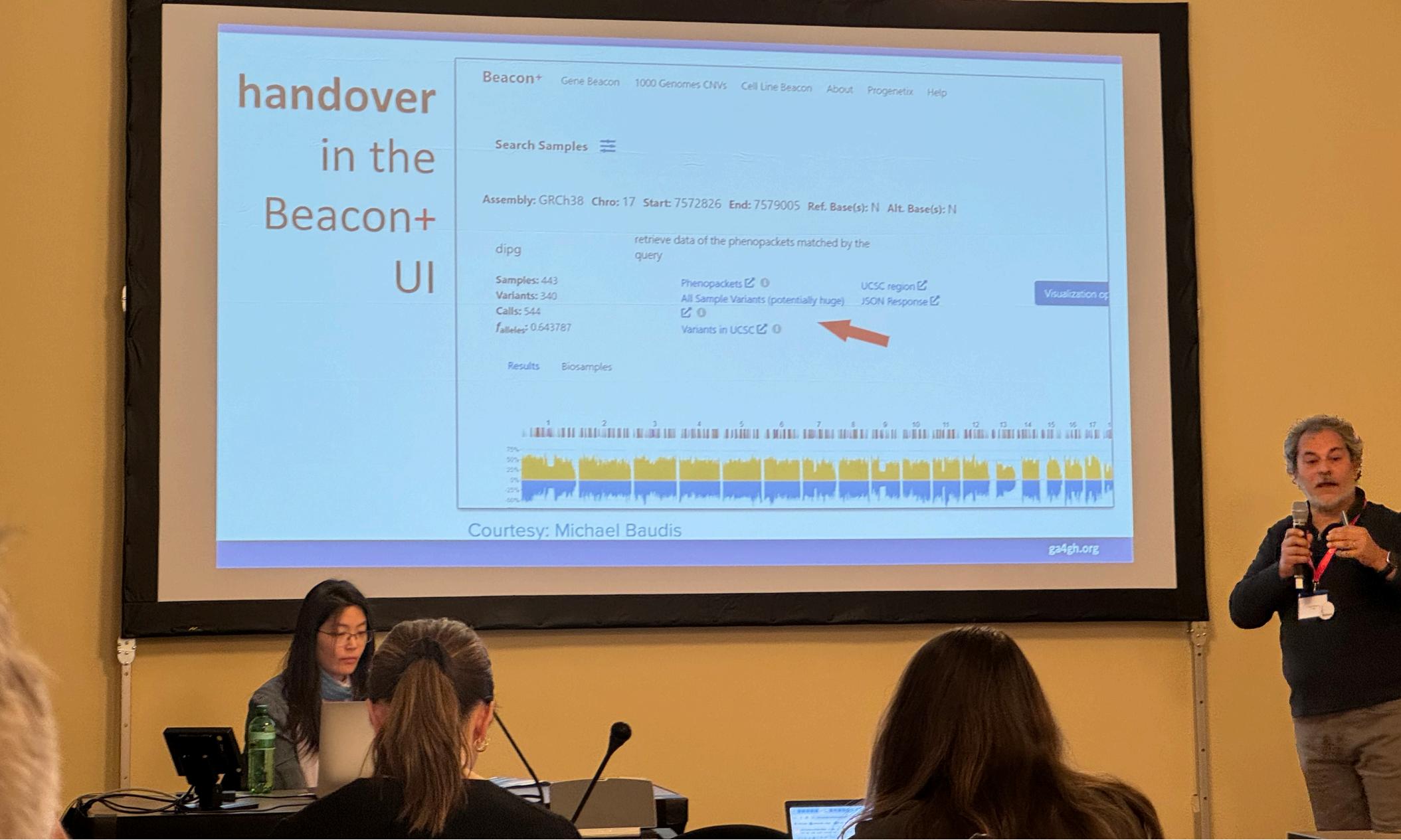
Summary aggregations implemented using a Beacon aggregated response prototype in Progenetix



# BioTAPESTRY?

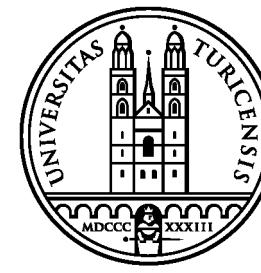
- host exchange visitors or project partners for
  - implementing Beacons via **bycon**
  - porting workflows and tools to Galaxy (focus on CNV analyses)
  - text mining and metadata mappings for Progenetix
  - Beacon implementations for novel use cases
- visit for presentation, workshops, projects
  - teaching about Beacon, CNVs ..
  - group members for galaxyfying work





The Global Alliance for Genomics and Health (GA4GH) gathered for the 2024 [April Connect meeting](#) in Ascona, Switzerland and online from 21 to 24 April. The GA4GH Connect meetings provide an opportunity for contributors to advance the GA4GH Road Map, showcase GA4GH standards and policies in action, and gather feedback on product development and community needs. The meeting brought together 103 in-person attendees and 312 virtual attendees for updates from Work Streams and Driver Projects, breakout sessions, and themed events.





University of  
Zurich<sup>UZH</sup>  
Department of Molecular Life Sciences

