



# ELIXIR hCNV Community

Michael Baudis | ELIXIR hCNV Community Webinar 2024

[www.elixir-europe.org](http://www.elixir-europe.org)

## h-CNV Community

Homepage &amp; News

About ...

h-CNV Projects

CNV Annotation Standards

Databases &amp; Resources

CNV References Project

Contacts

Genome Blog

h-CNV @ ELIXIR

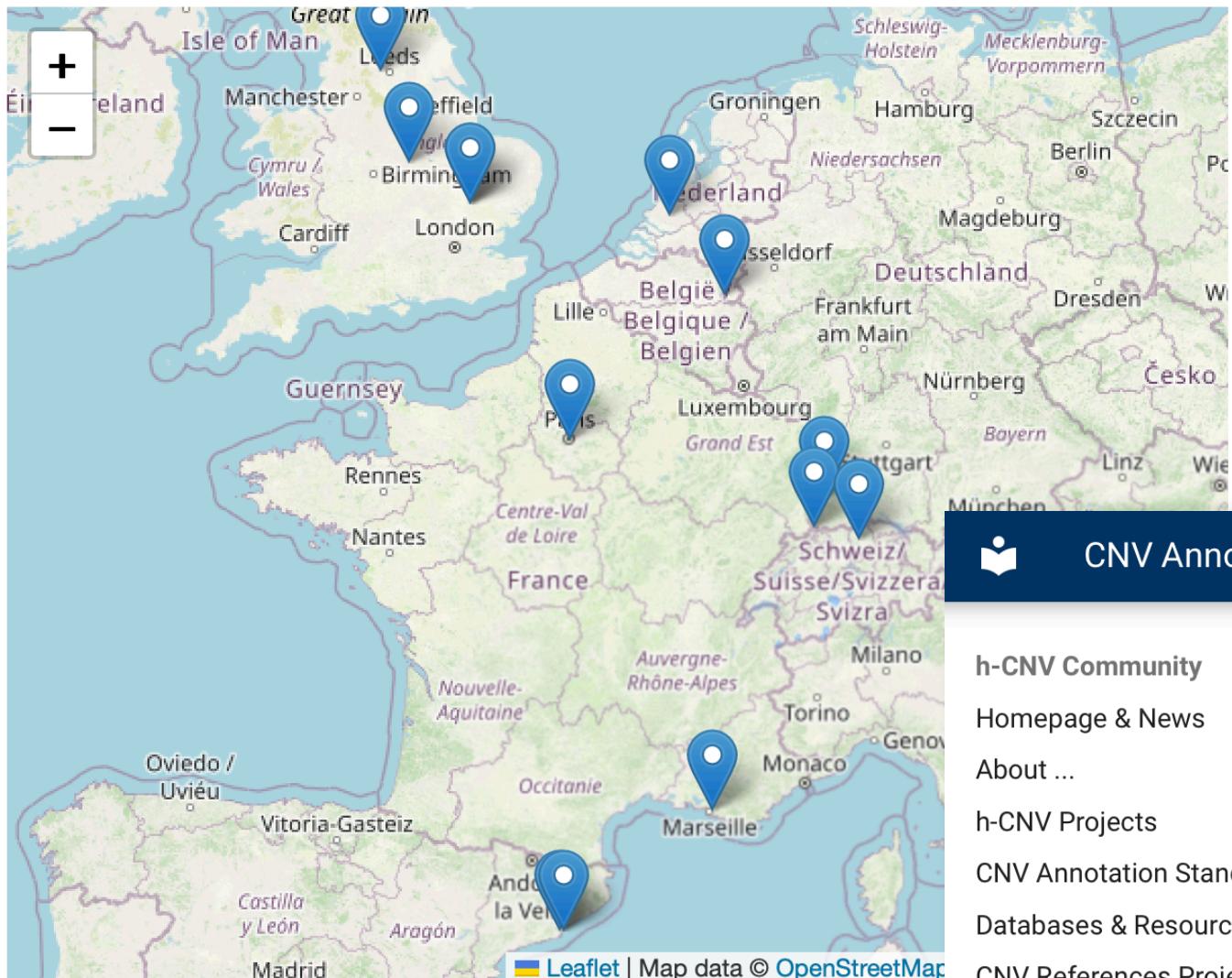
Beacon Project

## ELIXIR Human Copy Number Variation community

Among the different types of inherited and acquired genomic variants, regional genomic copy number variations (CNV) contribute - if measured by affected genomic sequences - contribute by far the largest amount of genomic changes, contributing both to many syndromic diseases as well as the vast majority of human cancers. The [website](#) of the *Human Copy Number Variation*

*Community* (hCNV) is a resource originated in ELIXIR's h-CNV Community Implementation Study (2019-2021) with the aim to provide a resource hub and knowledge exchange space for scientists and practitioners working with - or being interested in - genomic copy number variations in health and diseases.

However, the scope of the community extends beyond CNVs and includes definition of and work with other types of genomic variations with a focus on structural variants.



# ELIXIR hCNV Community

<https://cnvar.org/>

## CNV Annotation Formats

## Search

## h-CNV Community

Homepage &amp; News

About ...

h-CNV Projects

CNV Annotation Standards

Databases &amp; Resources

CNV References Project

Contacts

Genome Blog

h-CNV @ ELIXIR

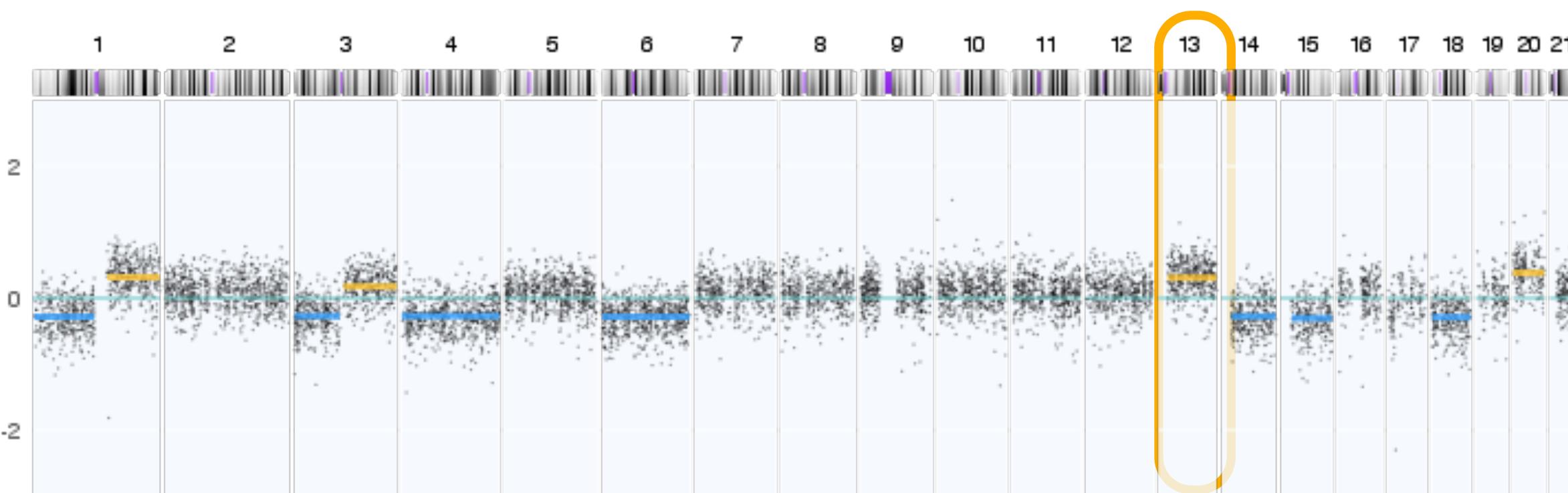
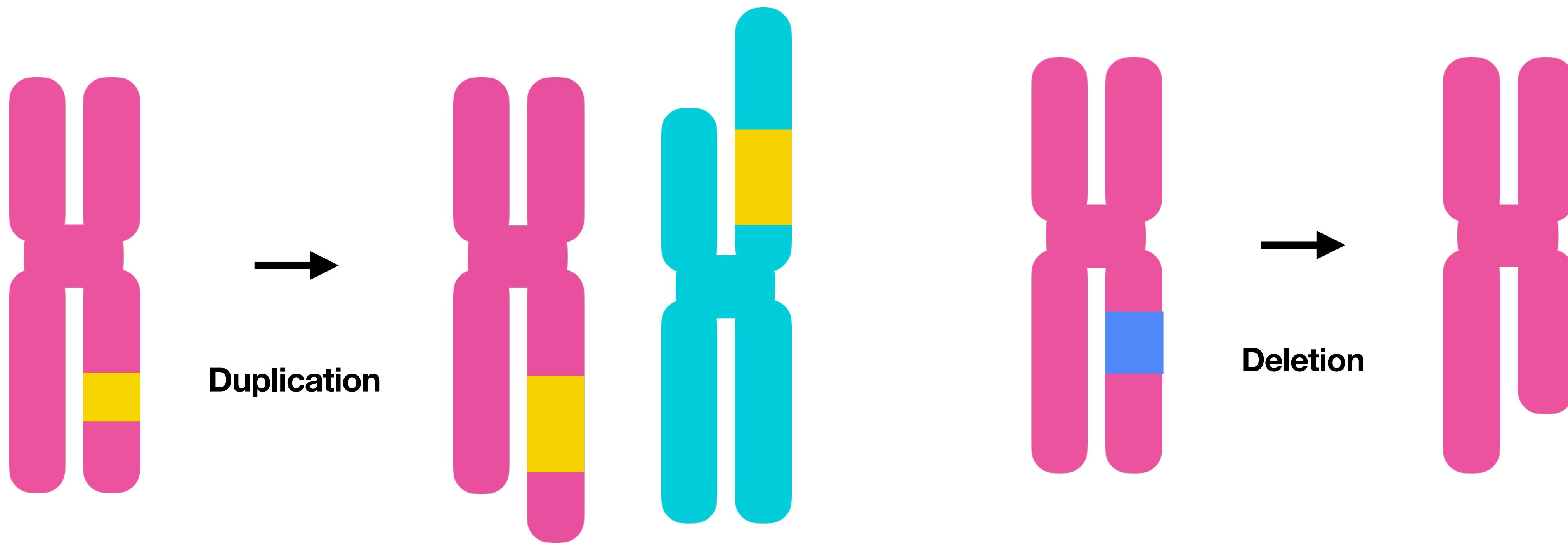
Beacon Project

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

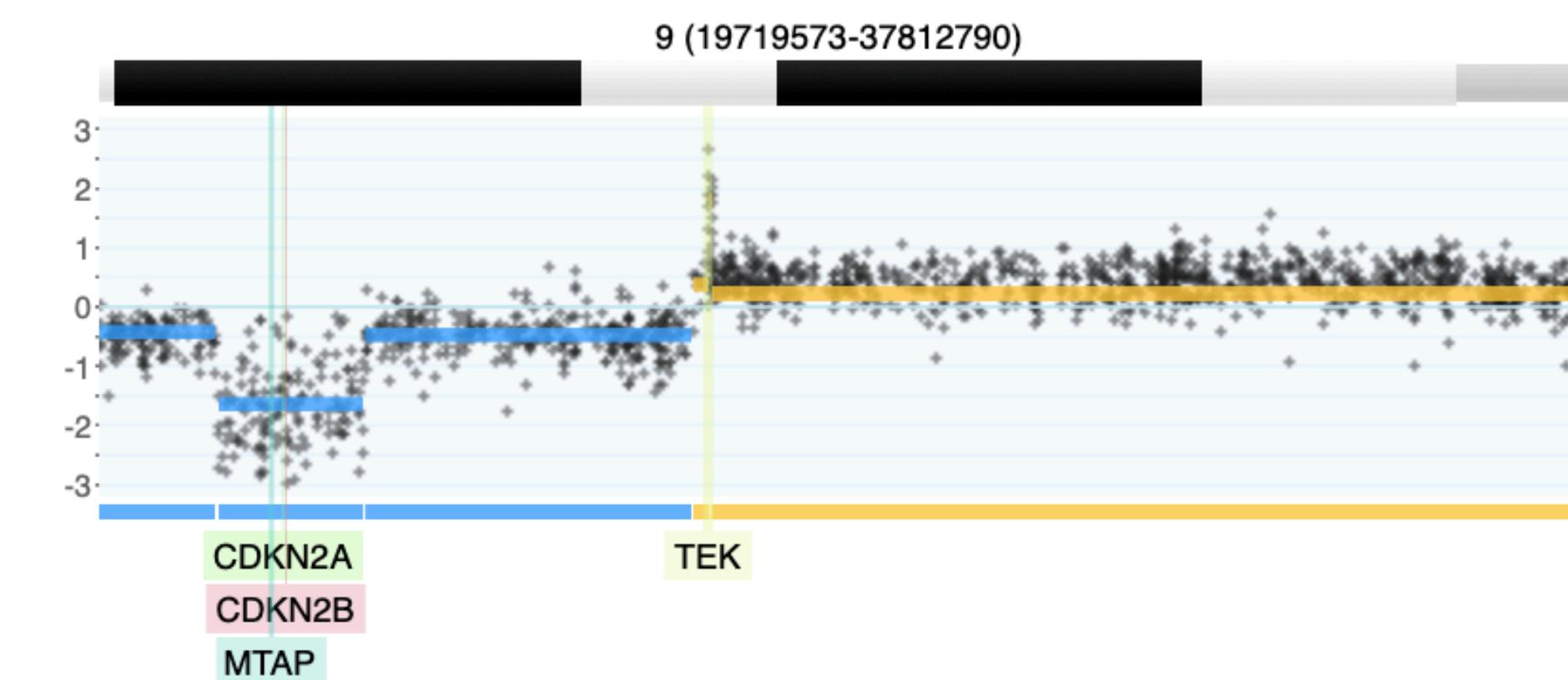
This table is maintained in parallel with the [Beacon v2 documentation](#).

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

# Somatic Copy Number Variation



Gain of chromosome arm 13q in colorectal carcinoma

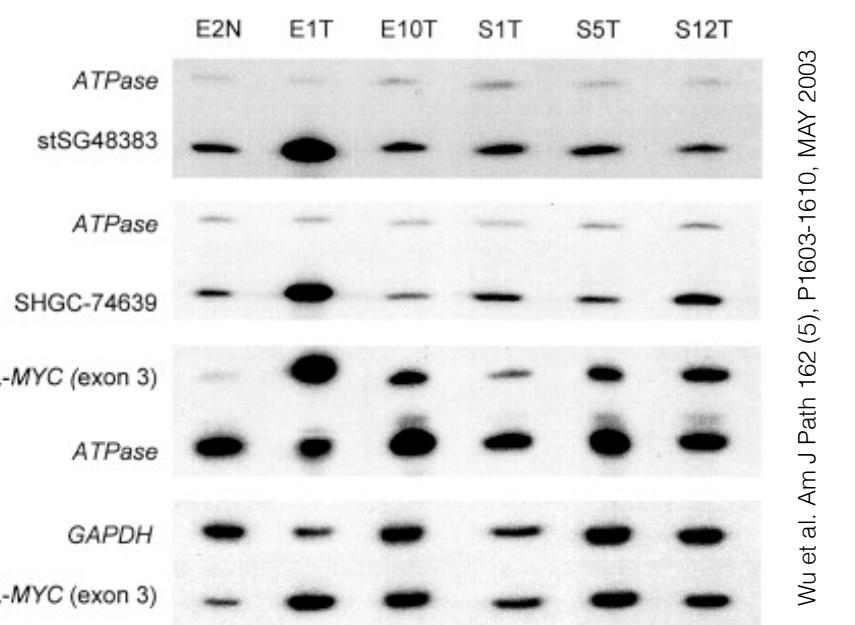
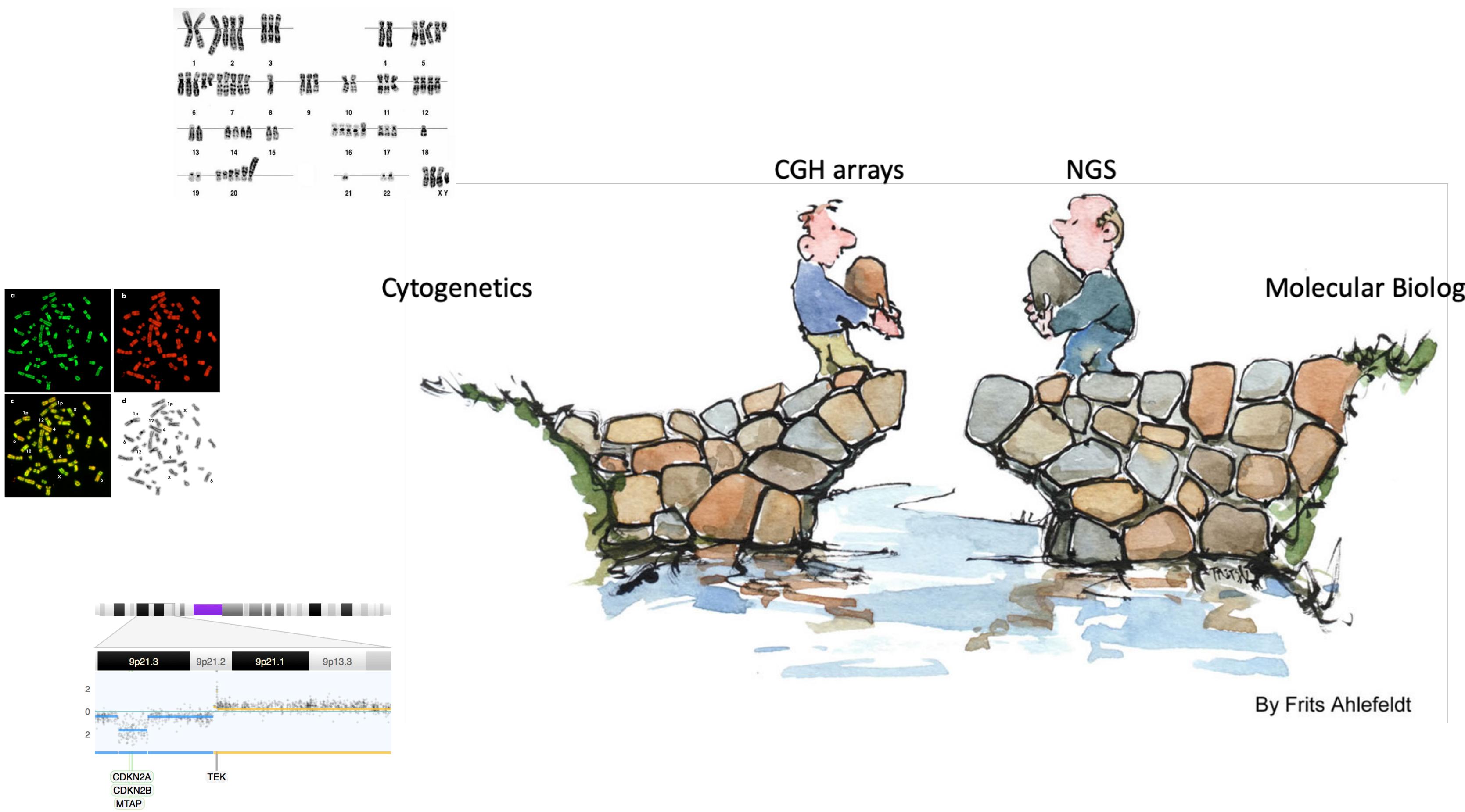


2-event, homozygous deletion in a Glioblastoma

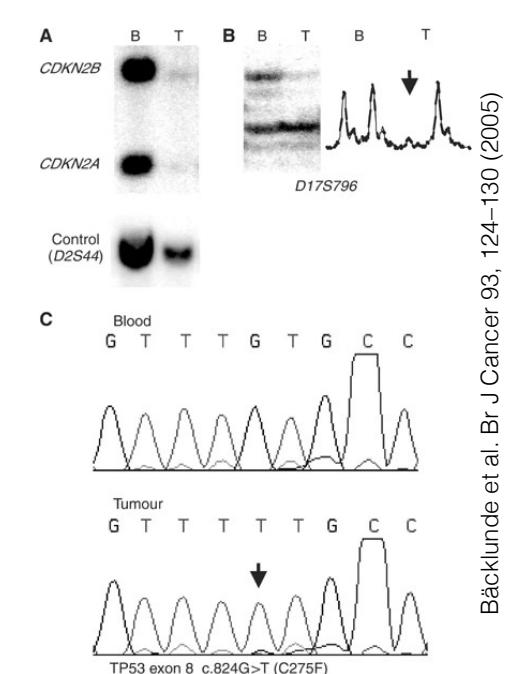
# h-CNV scientific context

- Structural variants have been the first ones to be detected in humans (late 1950s)
- Genes' mutations shortly followed (Ingram et al. 1957)

Slide: Christophe Bérroud



Wu et al. Am J Path 162 (5), P1603-1610, May 2003



Bäcklund et al. Br J Cancer 93, 124-130 (2005)





Universität  
Zürich UZH



progenetix

# The hCNV Community

## CNV profiling resources in cancer genomics & the need for data sharing

Michael Baudis | ELIXIR hCNV Community Webinar 2024

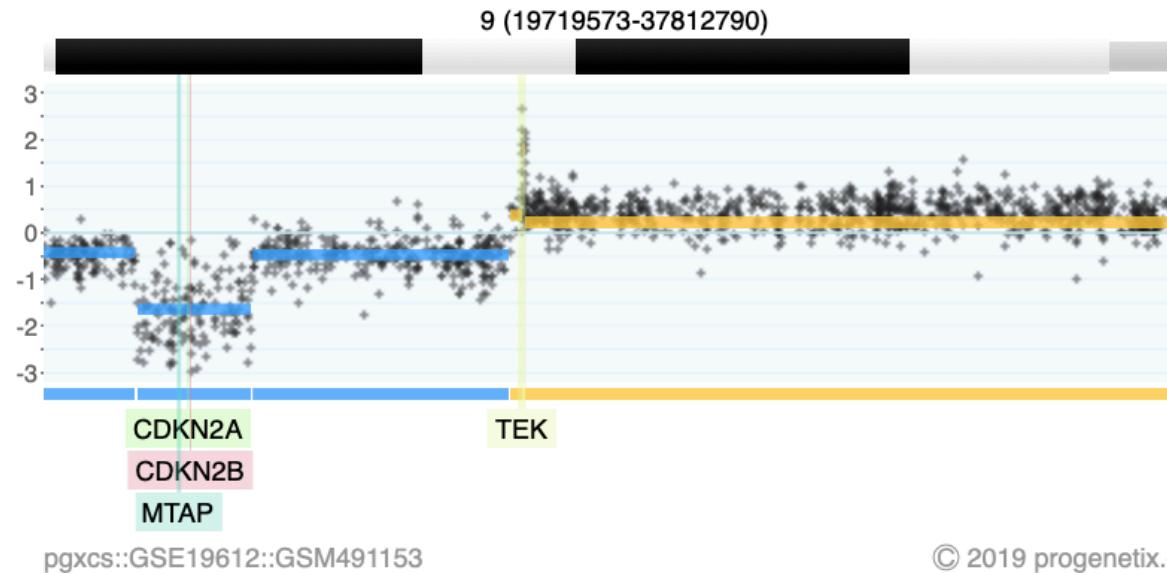


# Theoretical Cytogenetics and Oncogenomics

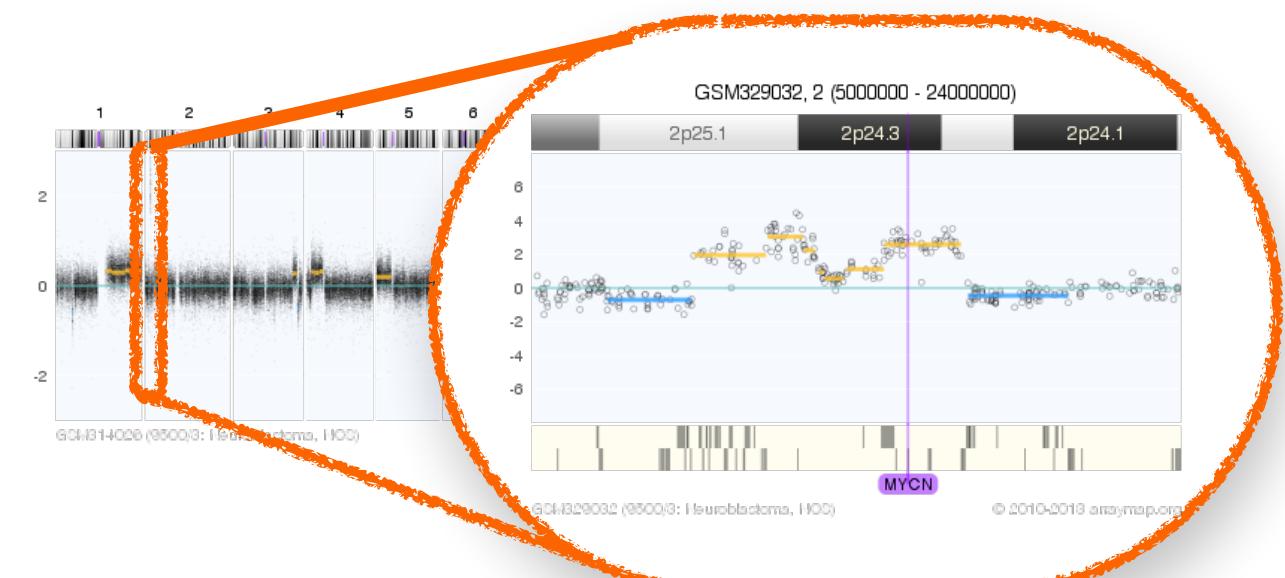
## Research | Methods | Standards

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

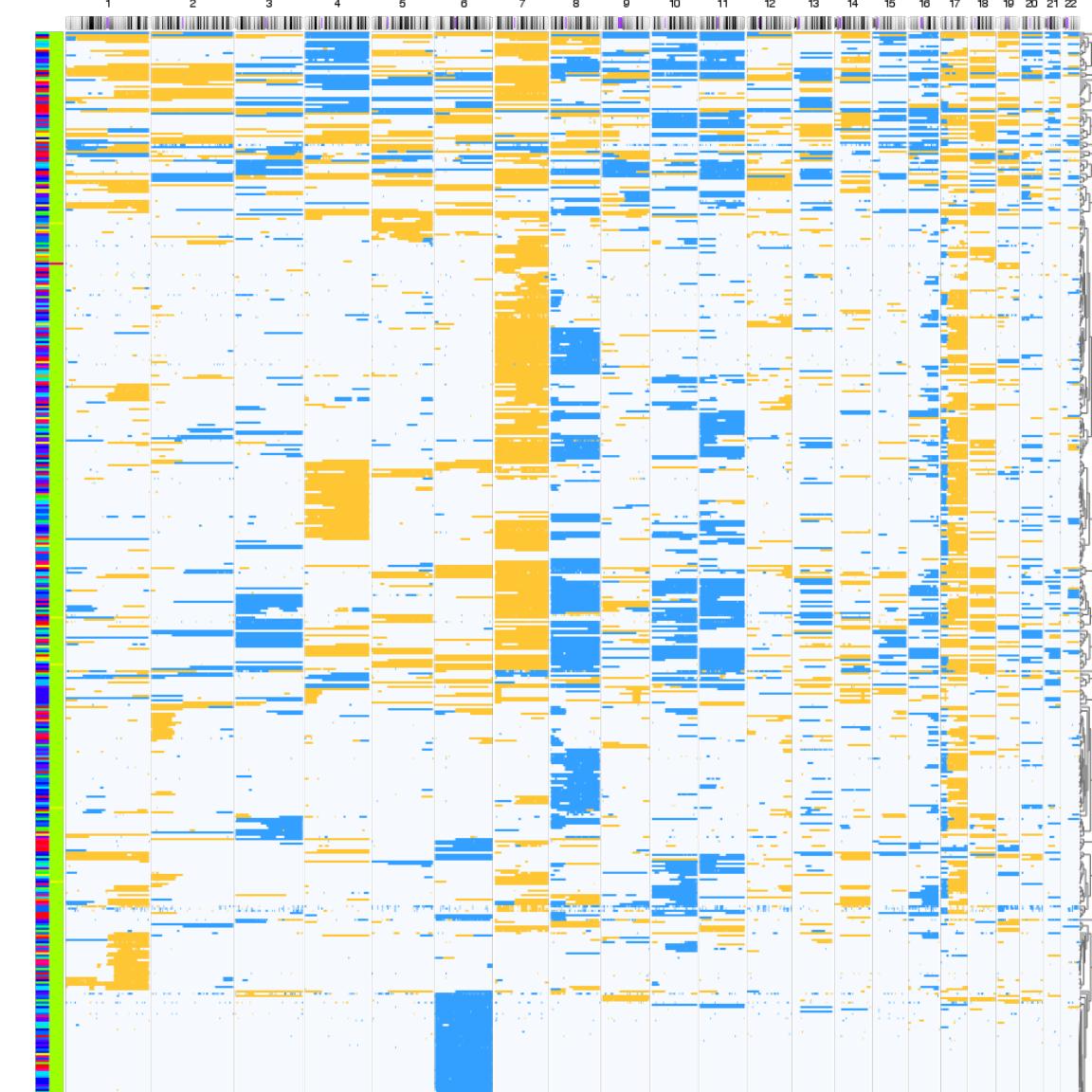
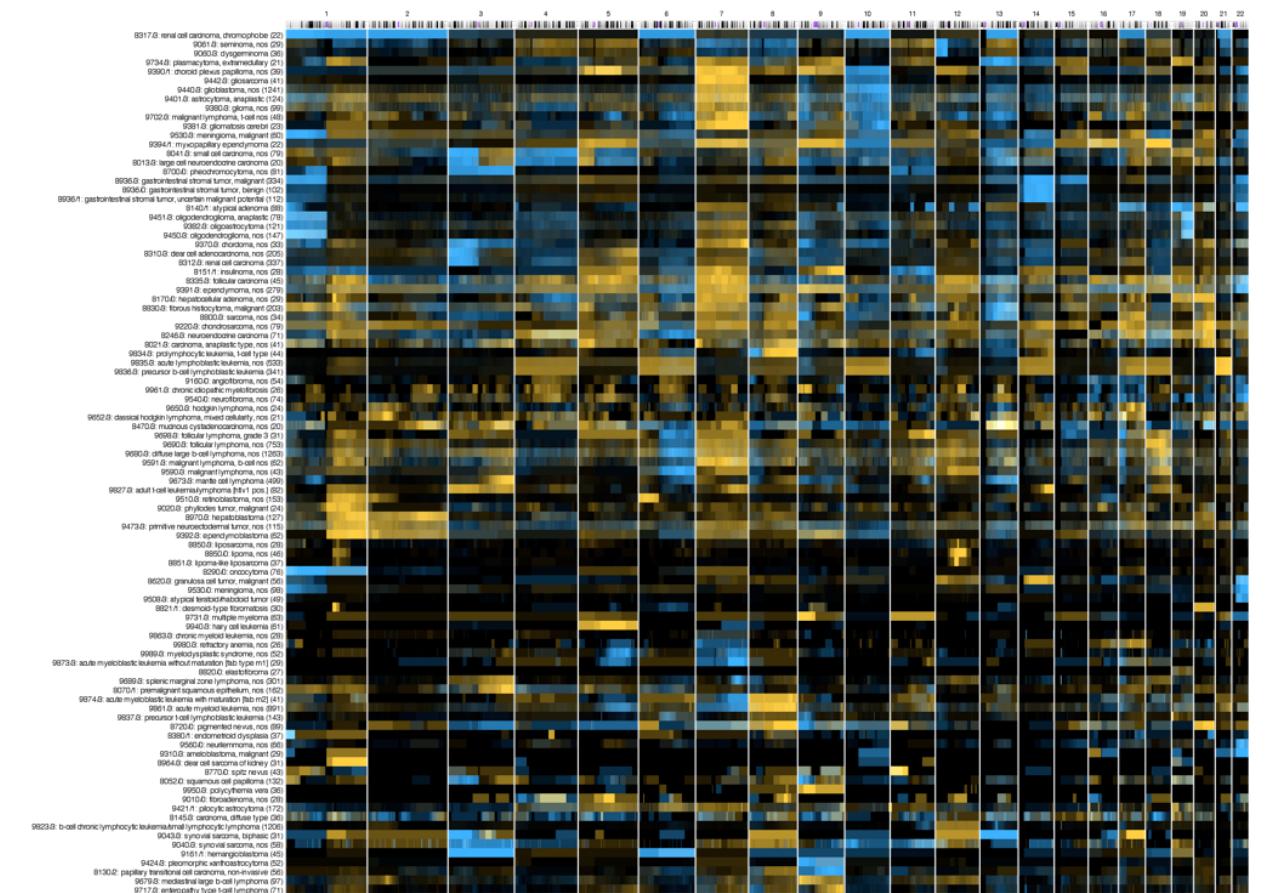
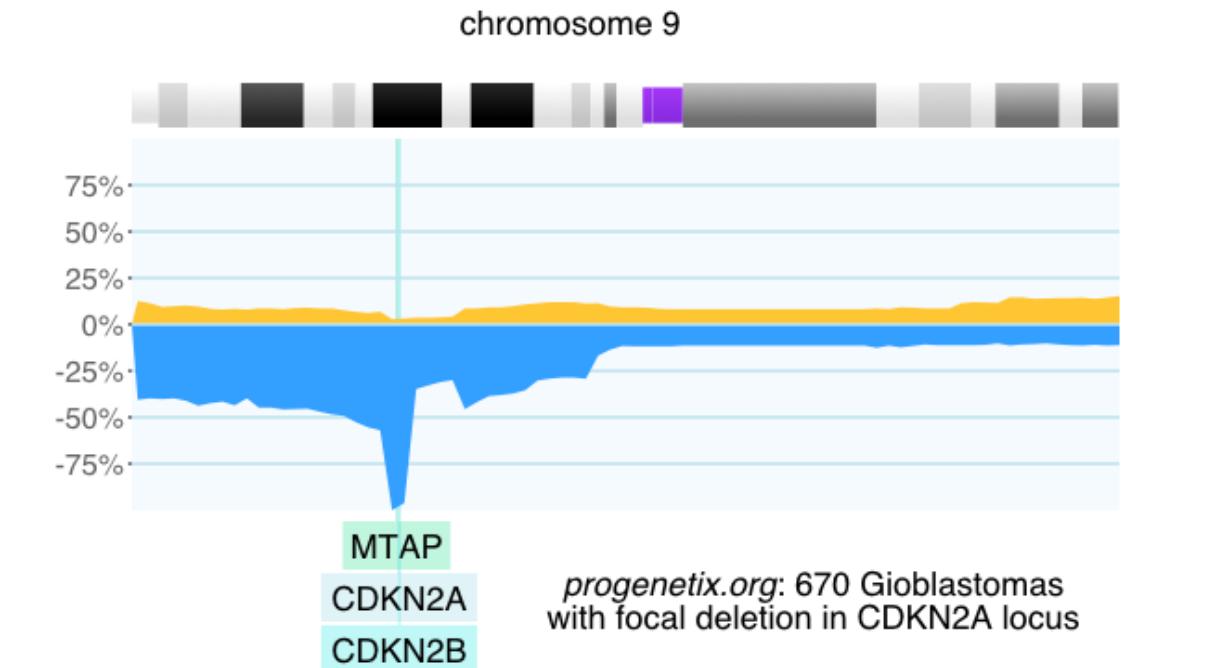
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)



2-event, homozygous deletion in a Glioblastoma

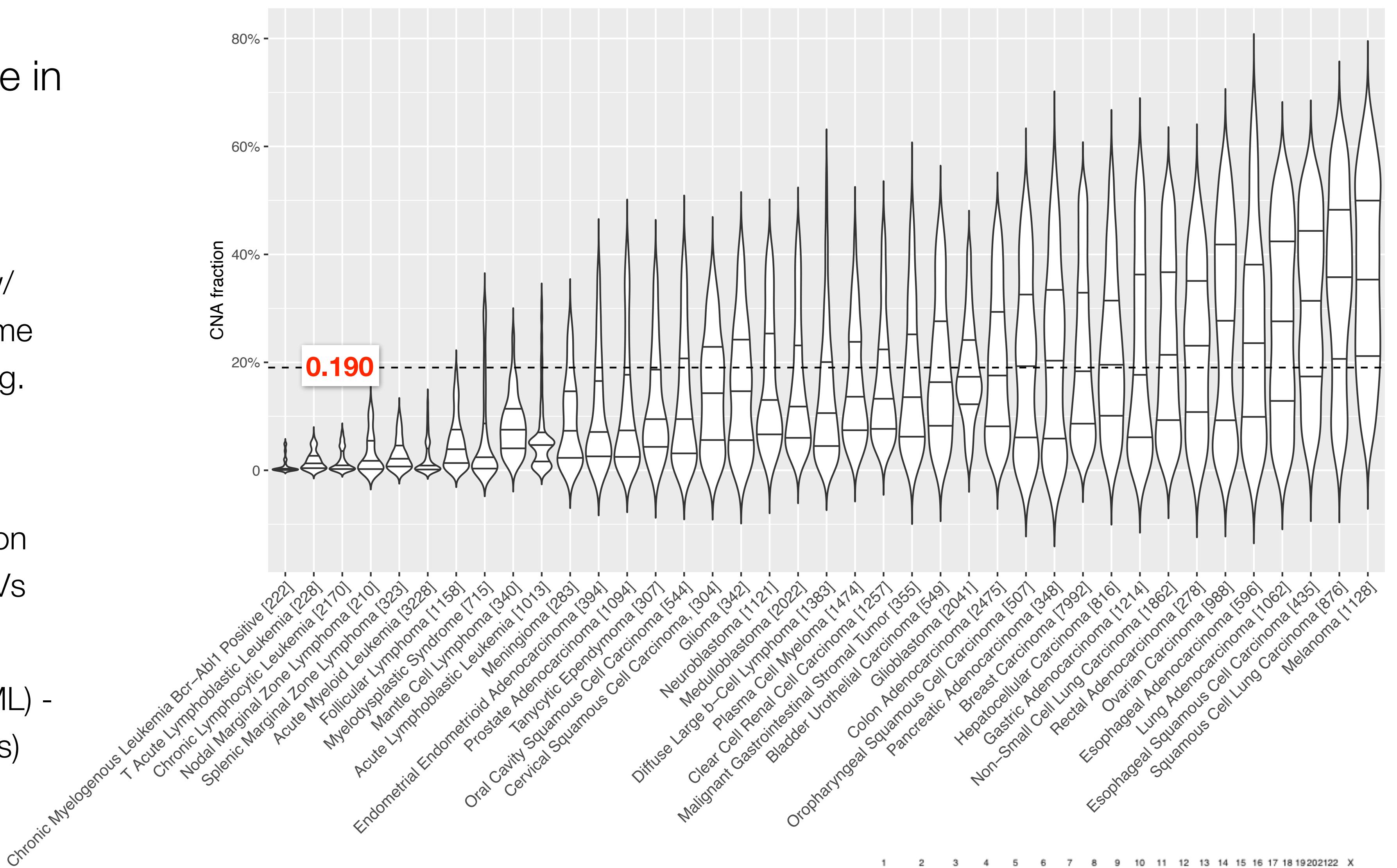


MYCN amplification in neuroblastoma  
(GSM314026, SJNB8\_N cell line)

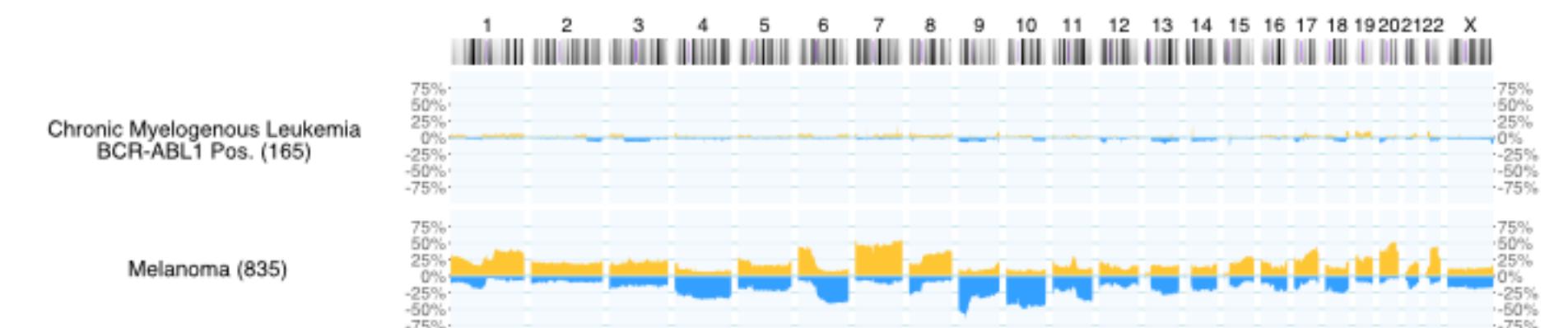


# Genome CNV coverage in Cancer Classes

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



Lowest / Highest CNV fractions =>



## Cancer Genomics Reference Resource

- **open** resource for oncogenomic profiles
- over **116'000 cancer CNV profiles**
- more than **800 diagnostic types**
- inclusion of reference datasets (e.g. TCGA)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services
- recent addition of SNV data for some series



### Cancer 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<sup>+</sup>

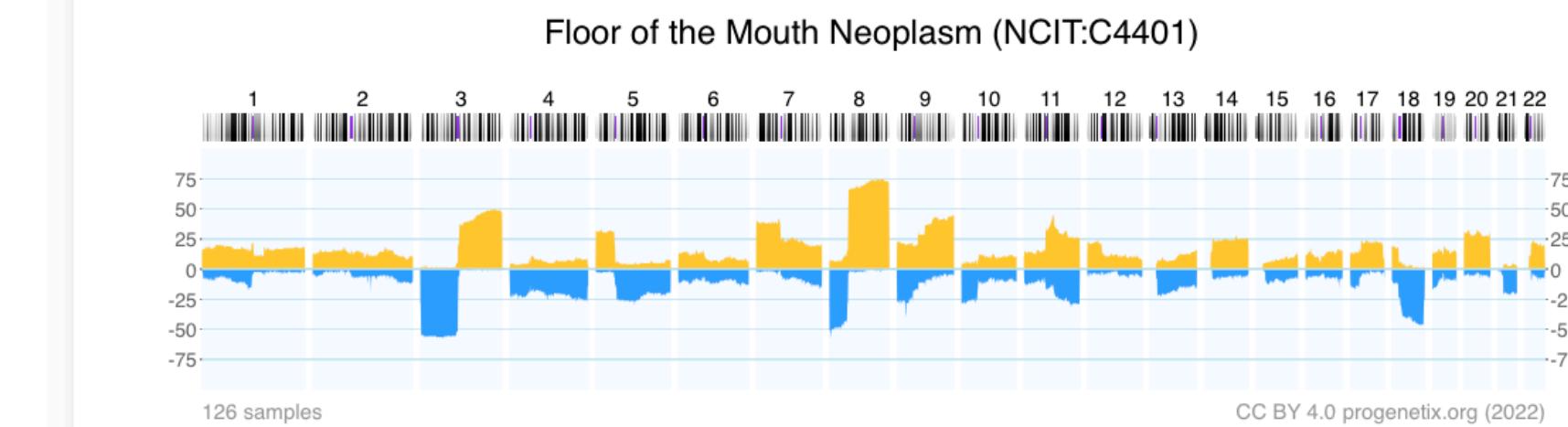
### Documentation

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.



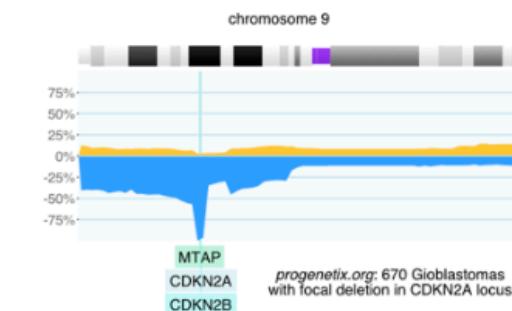
[Download SVG](#) | [Go to NCIT:C4401](#) | [Download CNV Frequencies](#)

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.



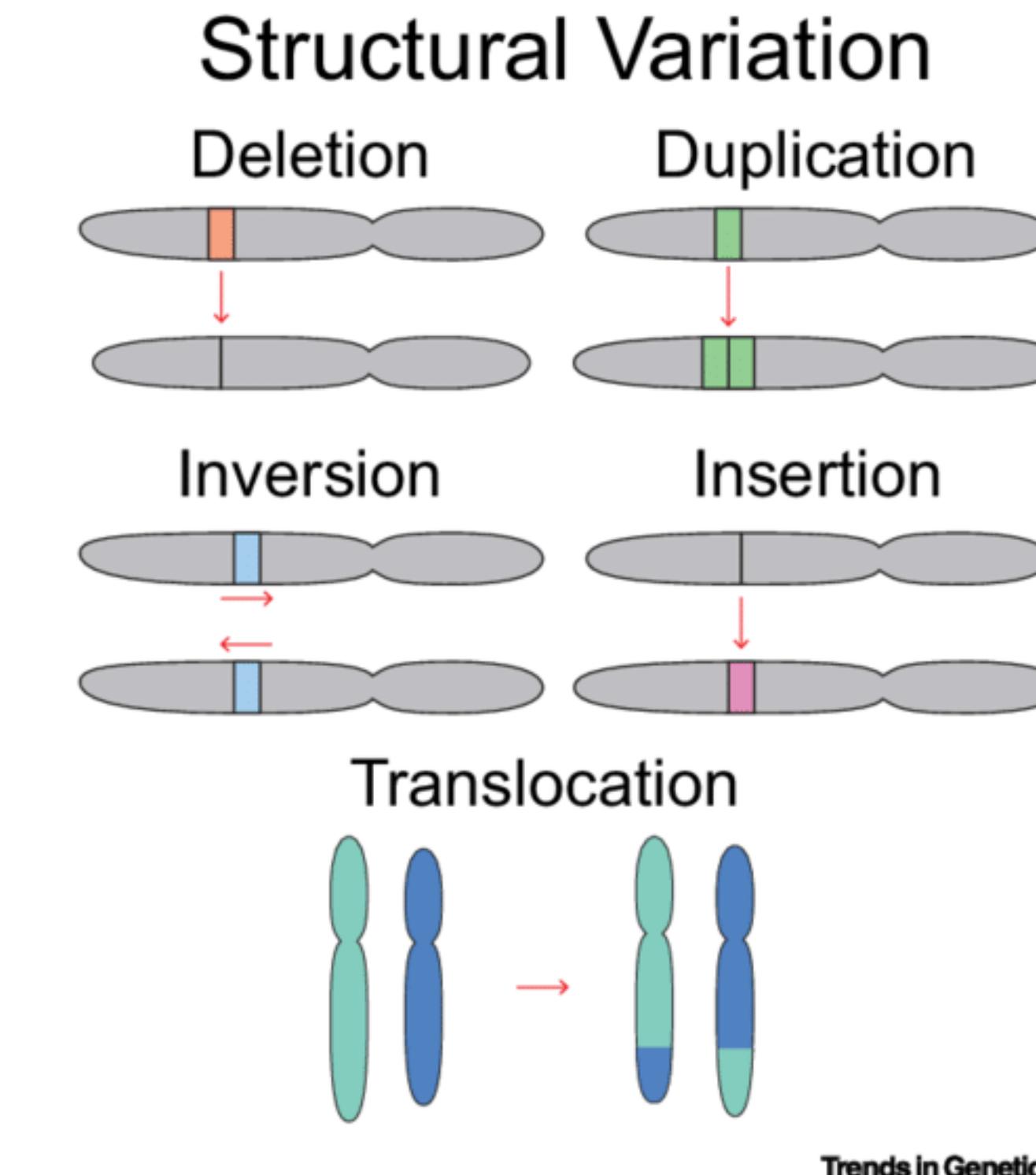
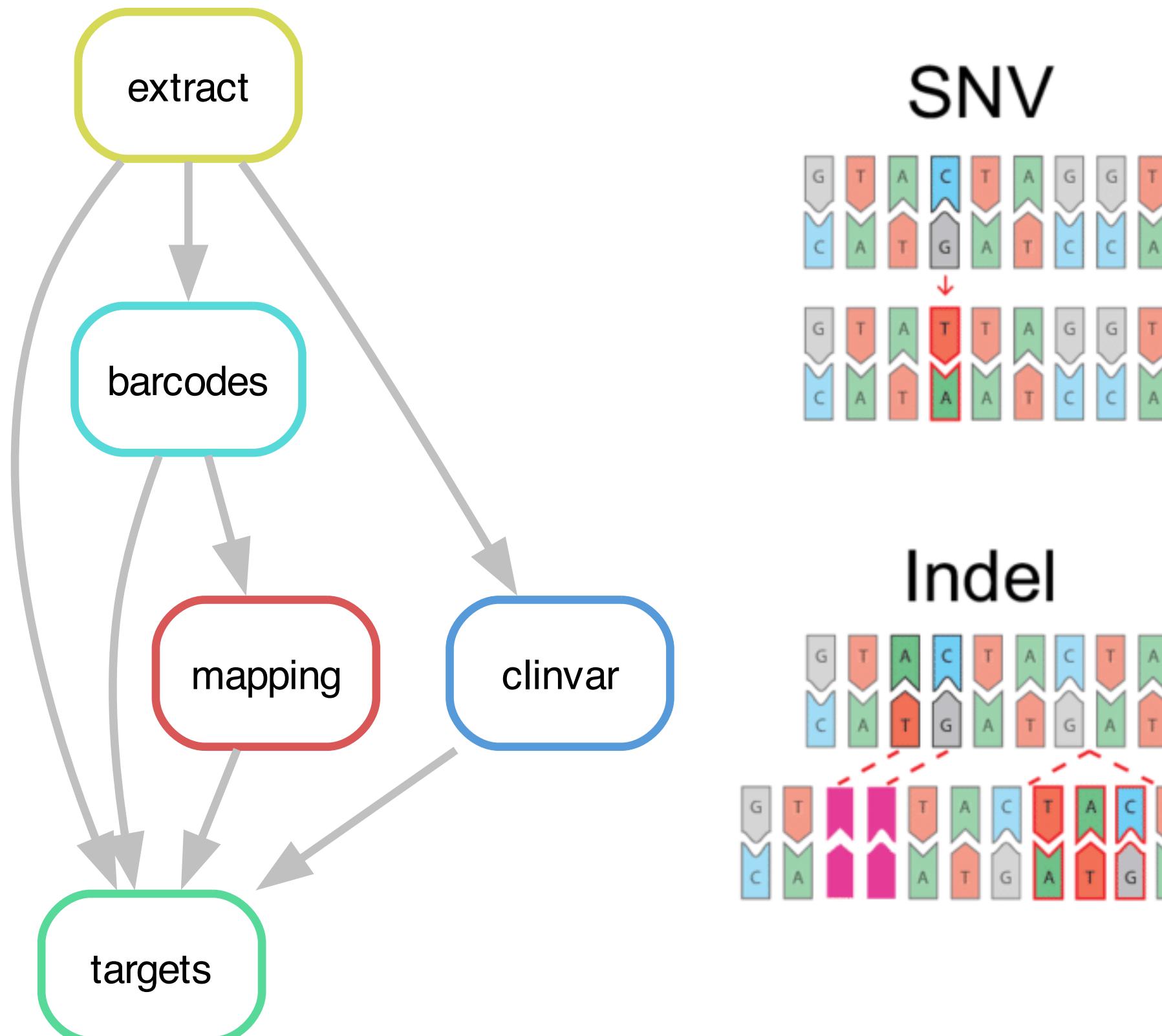
#### Cancer CNV Profiles

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.

- **Integrating SNVs:**
  - Renewed interest due to technological advantages
  - Allow compound variant queries



Nesta, Alex & Tafur, Denisse & Beck, Christine. (2020). Hotspots of structural variation in cancer genomes. Trends in Genetics

Mutation. Trends in Genetics

## TCGA BLCA project (pgx:TCGA.BLCA)



# 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



Cold  
Spring  
Harbor  
Laboratory

**bioRxiv**  
THE PREPRINT SERVER FOR BIOLOGY

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?].



Cancer Cell Lines

Search Cell Lines

Cell Line Listing

CNV Profiles by  
Cancer Type

Documentation

News

Progenetix

Progenetix Data

Progenetix

Documentation

Publication DB

Assembly: GRCh38 Chro: NC\_000007.14 Start: 140713328 End: 140924929

Type: SNV

cellz

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

UCSC region  
Variants in UCSC  
Dataset Responses (JSON)

Visualization options

Results Biosamples Variants Annotated Variants

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

Cell Lines (with parental/derived hierarchies)

Filter subsets e.g. by prefix

Hierarchy Depth

No Selection

- cellosaurus:CVCL\_0312: HOS (204 samples)
- cellosaurus:CVCL\_1575: NCI-H650 (6 samples)
- cellosaurus:CVCL\_1783: UM-UC-3 (9 samples)
- cellosaurus:CVCL\_0004: K-562 (28 samples)
- cellosaurus:CVCL\_3827: K562/Ad (1 sample)
- cellosaurus:CVCL\_0589: Kasumi-1 (9 samples)

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)

CC BY 4.0 progenetix.org (2023)

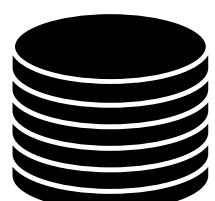
[Download SVG](#) | Go to cellosaurus:CVCL\_0312 | [Download CNV Frequencies](#)

Gene Matches	Cytoband Matches	Variants	Abstract
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

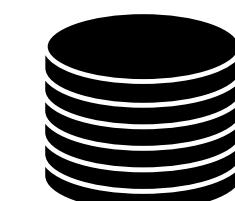
# Progenetix Stack



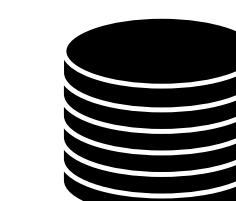
- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
  - ▶ biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads...
- the complete middleware / CGI stack is provided through the **bycon** package
  - ▶ schemas, 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



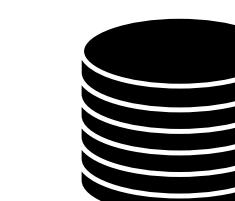
variants



analyses



biosamples

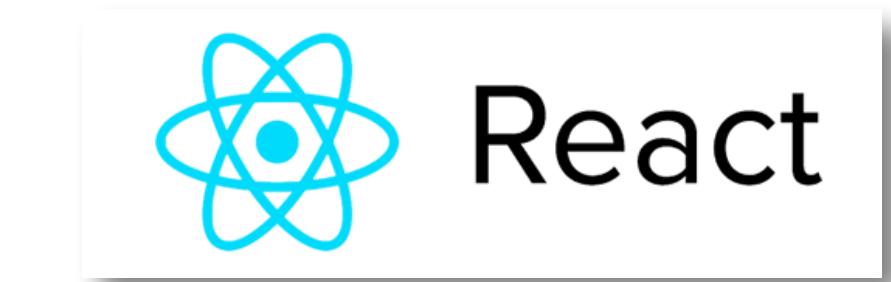


individuals



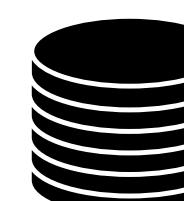
[github.com/progenetix/bycon/](https://github.com/progenetix/bycon/)

Entity collections

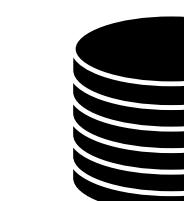


- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
  - ▶ PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703...
- *querybuffer* stores id values of all entities matched by a query and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

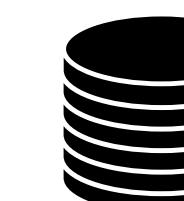
```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
  ObjectId("5bab578d727983b2e00cb505")]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



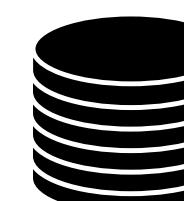
collations



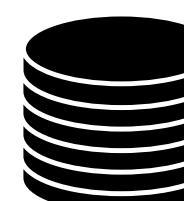
geolocs



genespans publications



publications



qBuffer

Utility collections

# {Bio|informatics}Science}

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

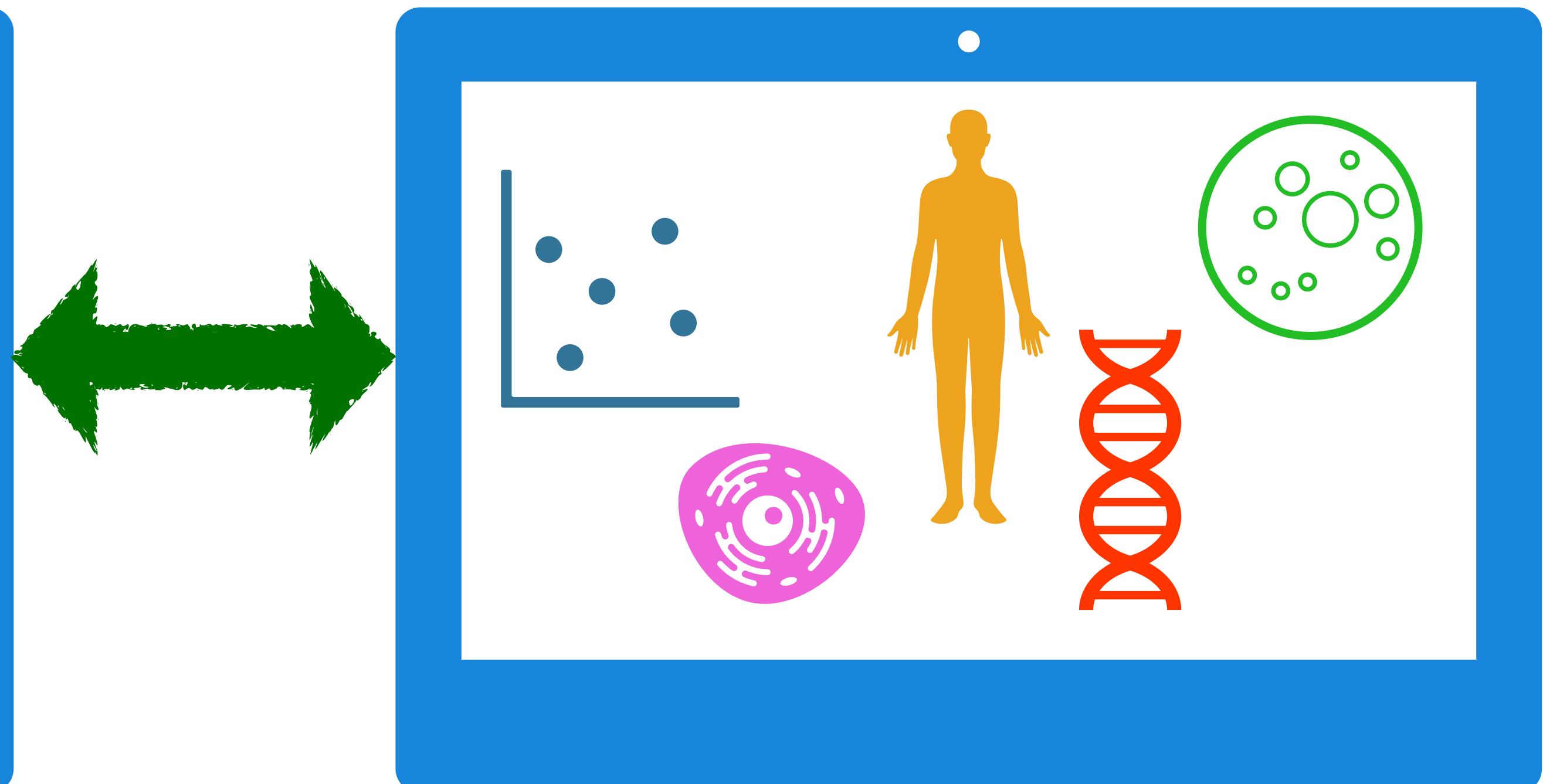
    covs = np.zeros((cs_no, int_no))
    vals = np.zeros((cs_no, int_no))

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

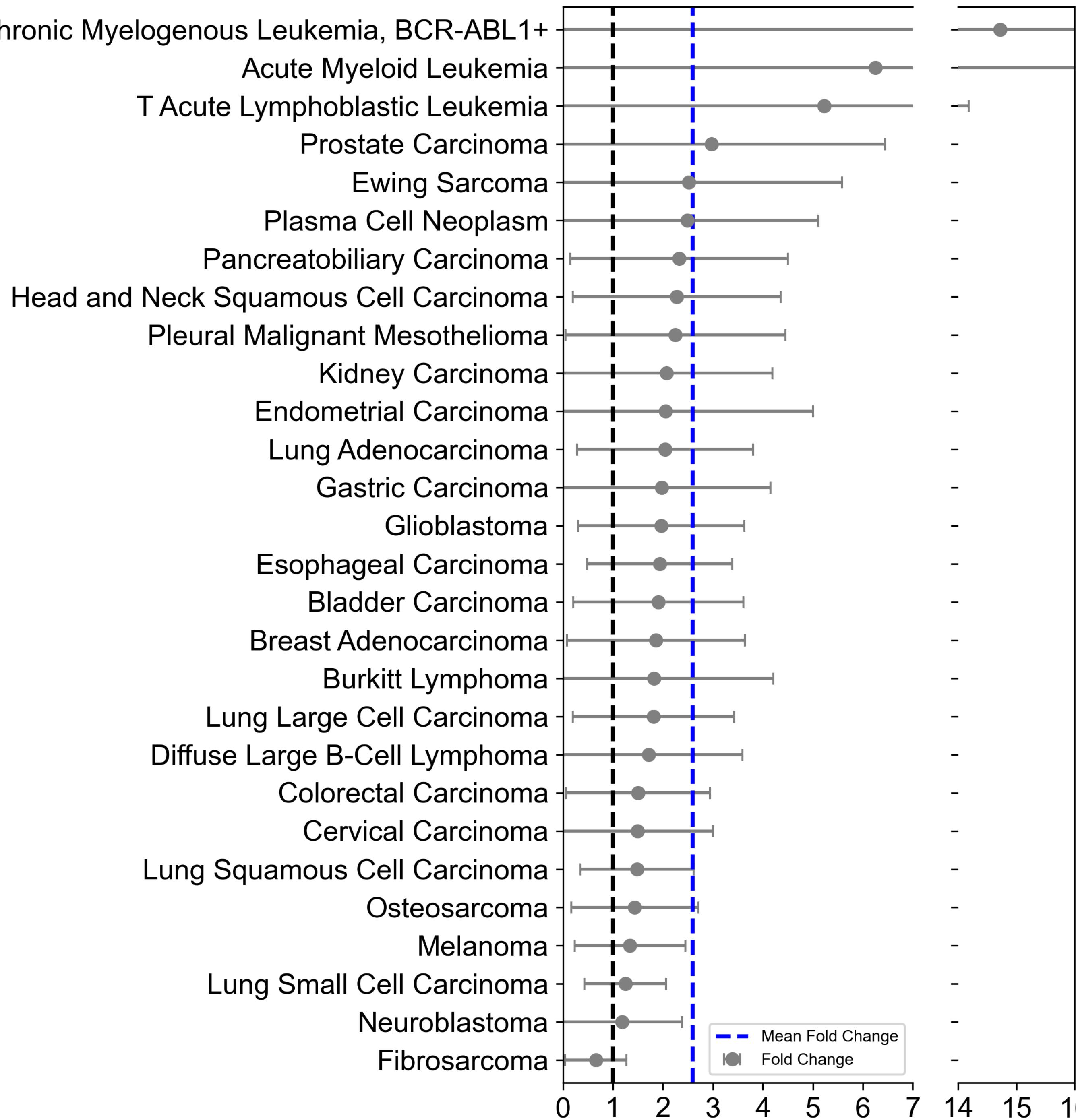
    for i, cs in enumerate(callsets):
        covs[i] = cs["cnv_statusmaps"][pars[t]["cov_l"]]
        vals[i] = cs["cnv_statusmaps"][pars[t]["val_l"]]

    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)

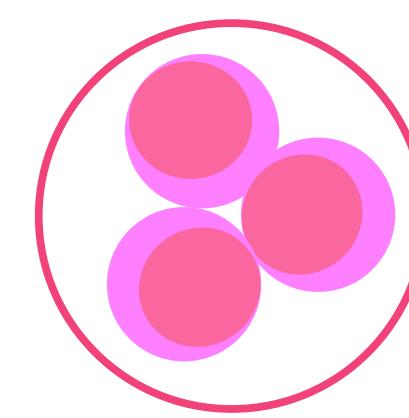
    for i, interval in enumerate(int_fs):
        int_fs[i].update({
            t + "_frequency": frequencies[i],
            t + "_median": medians[i],
            t + "_mean": means[i]
        })
```



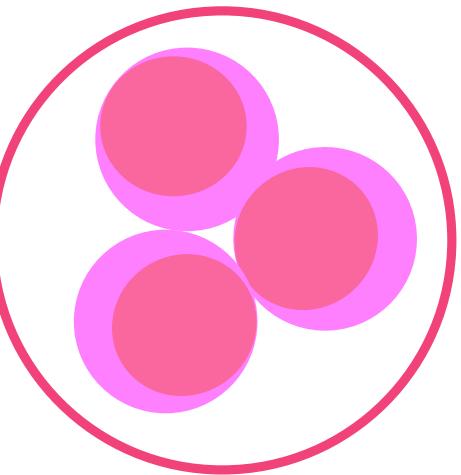
# Higher level of CNV coverage of the genomes of cancer cell lines compared to their origins



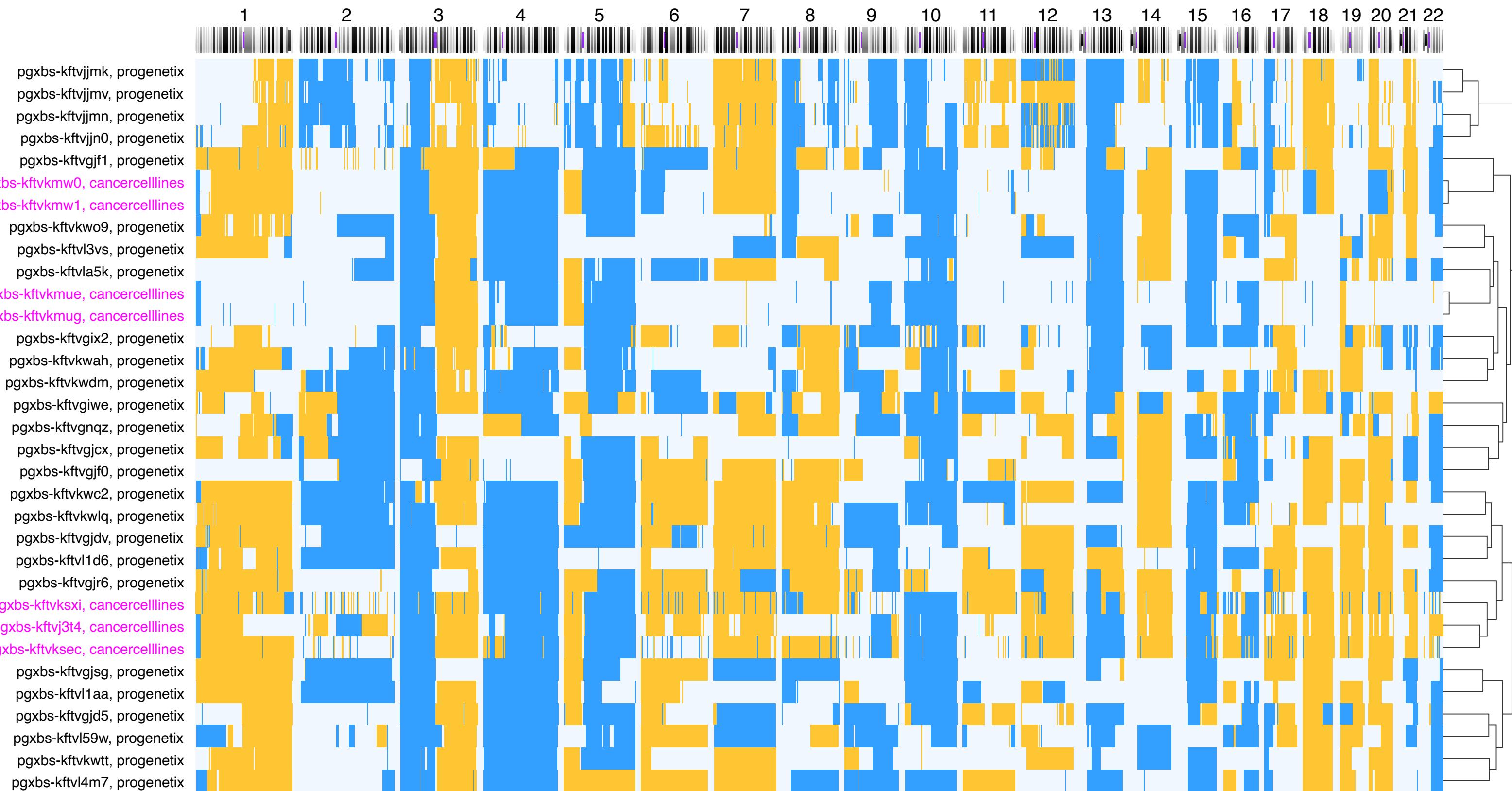
Fold changes between genome CNV coverages of cell lines and tumors



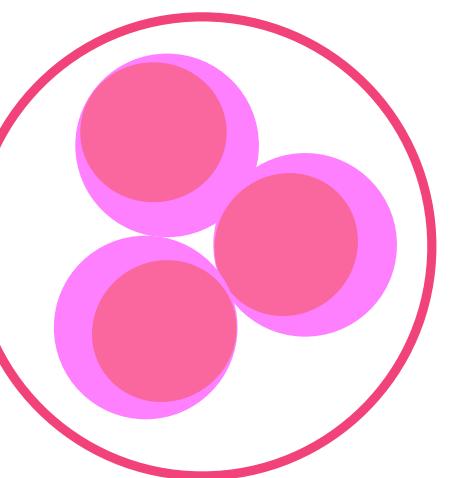
# Tumor subpopulations can be matched with highly similar cell lines



- Lung Small Cell Carcinoma Subpopulation
- Cell Lines:
  - CVCL\_1140: COR-L279
  - CVCL\_1455: NCI-H1105
  - CVCL\_1527: NCI-H2107



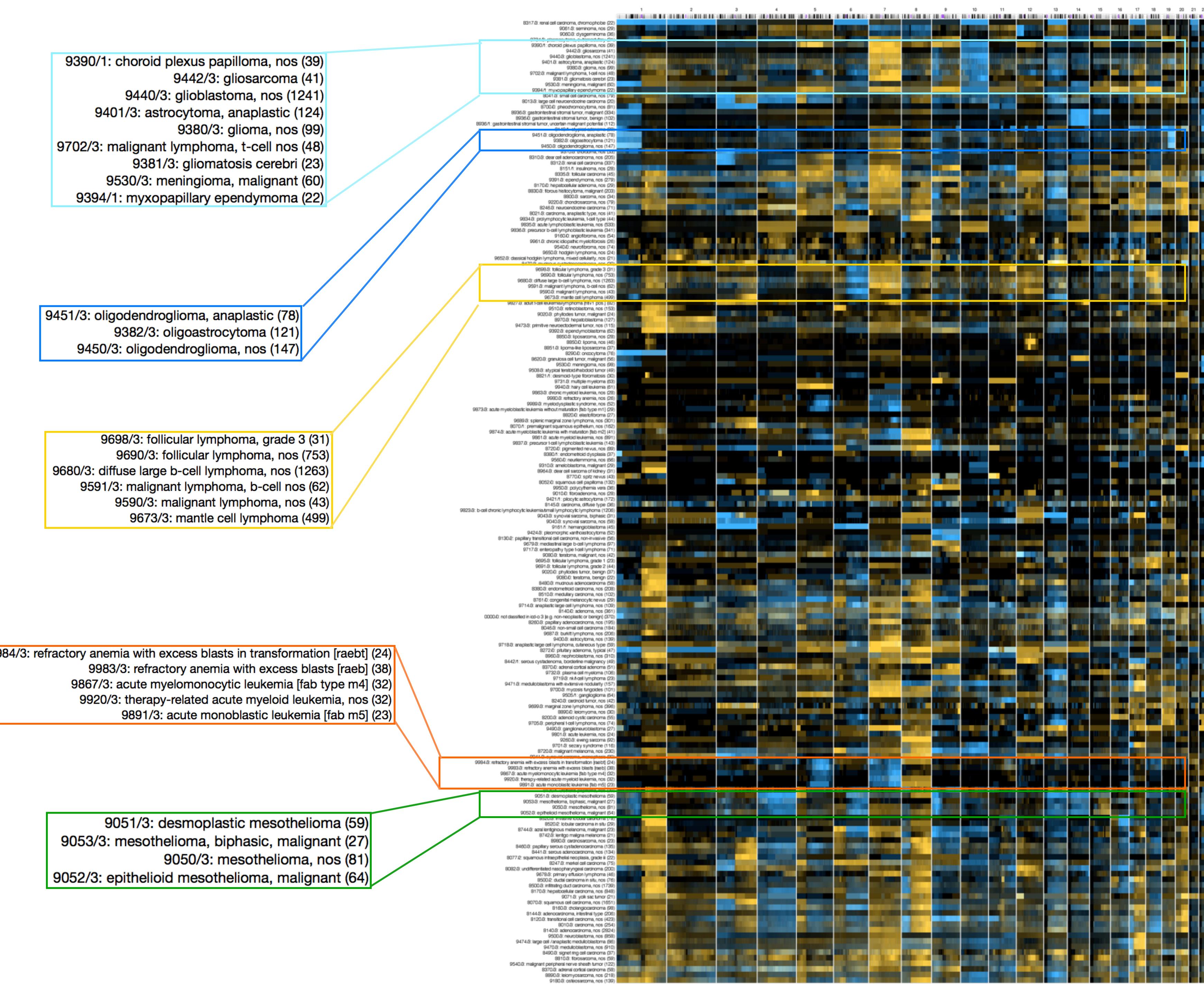
# Tumor subpopulations can be matched with highly similar cell lines?!



# Somatic Mutations In Cancer: Patterns

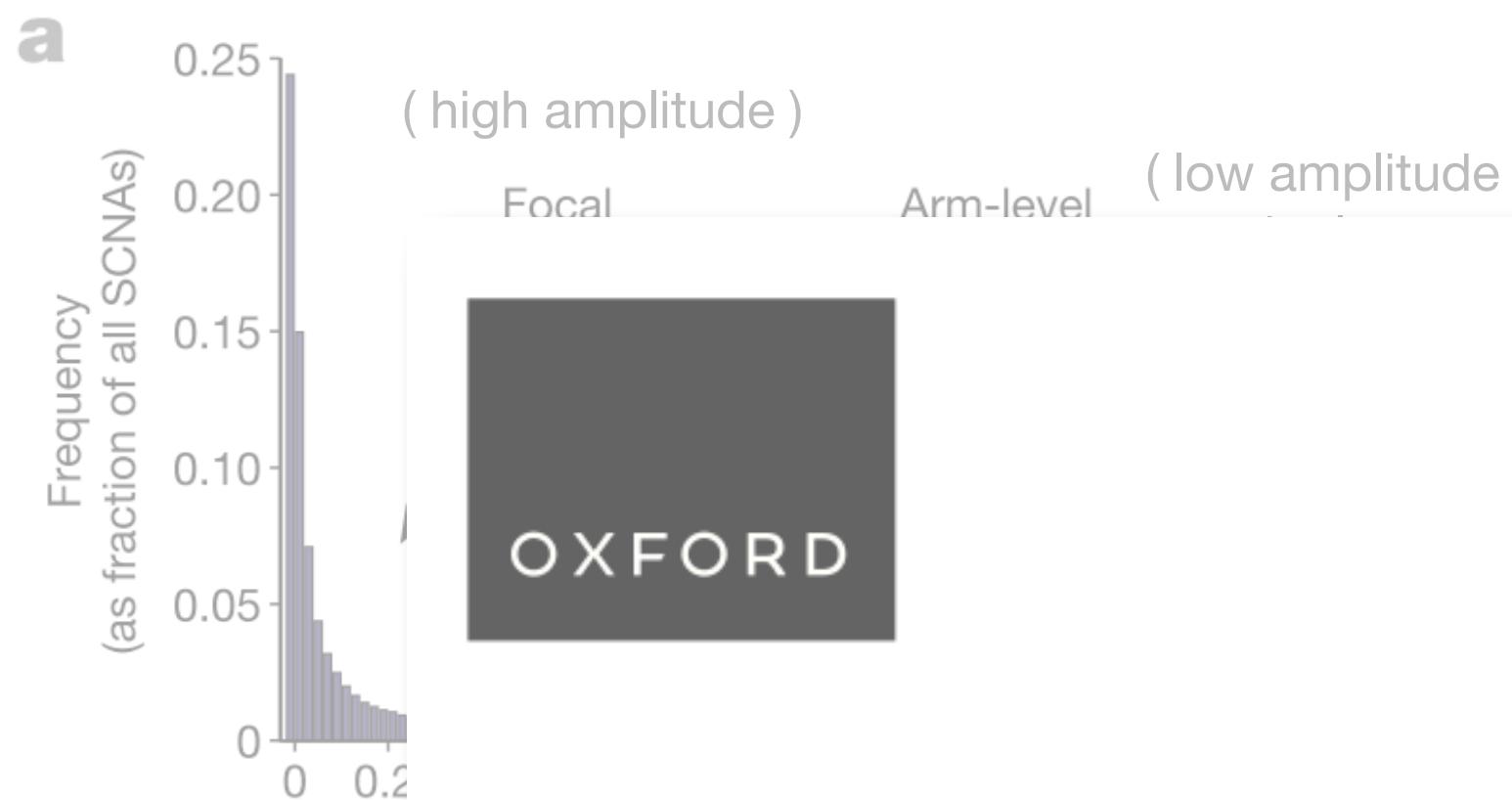
## Making the case for genomic classifications

Some related cancer entities show similar copy number profiles



# CNV Categorization

## different levels of CNV



## CopyNumberChange

*Copy Number Change* captures a categorization of copies of a molecule within a system relative to a

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

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

## Problem Solving Protocol

rule within a system, relative to a  
allers, particularly in the somatic  
and less useful in practice than  
is relative statements, and many  
interpreted to be relative copy

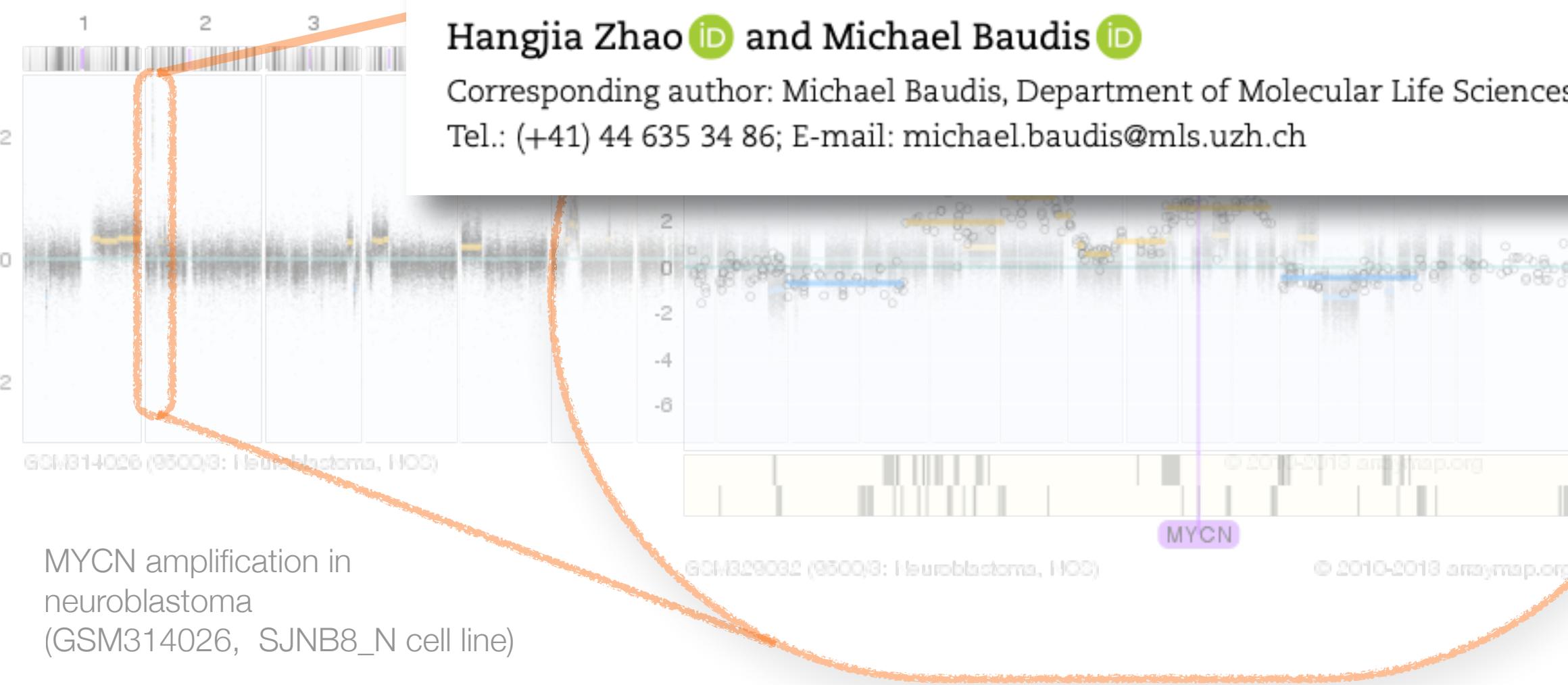
a system (e.g. genome, cell,

**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

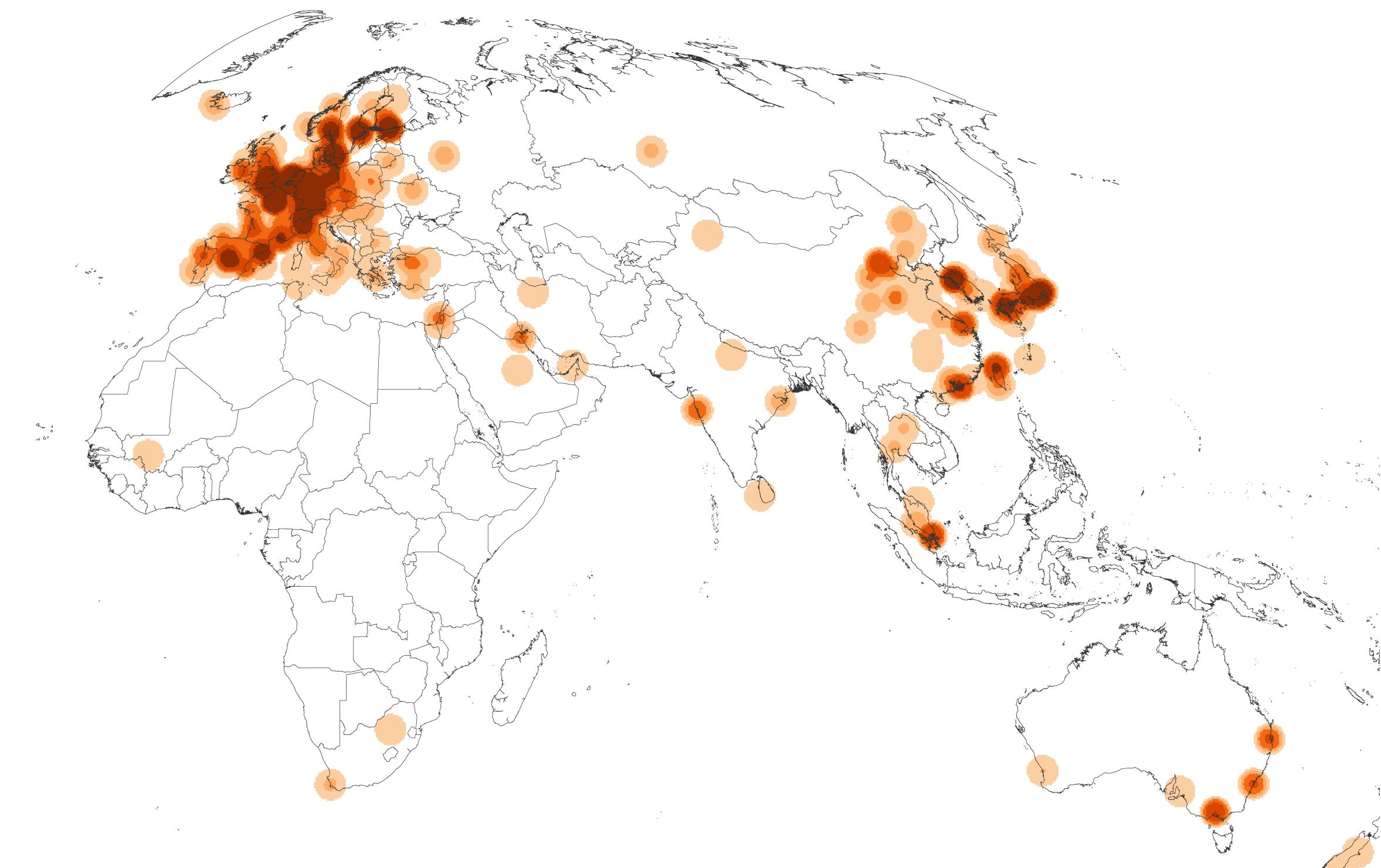
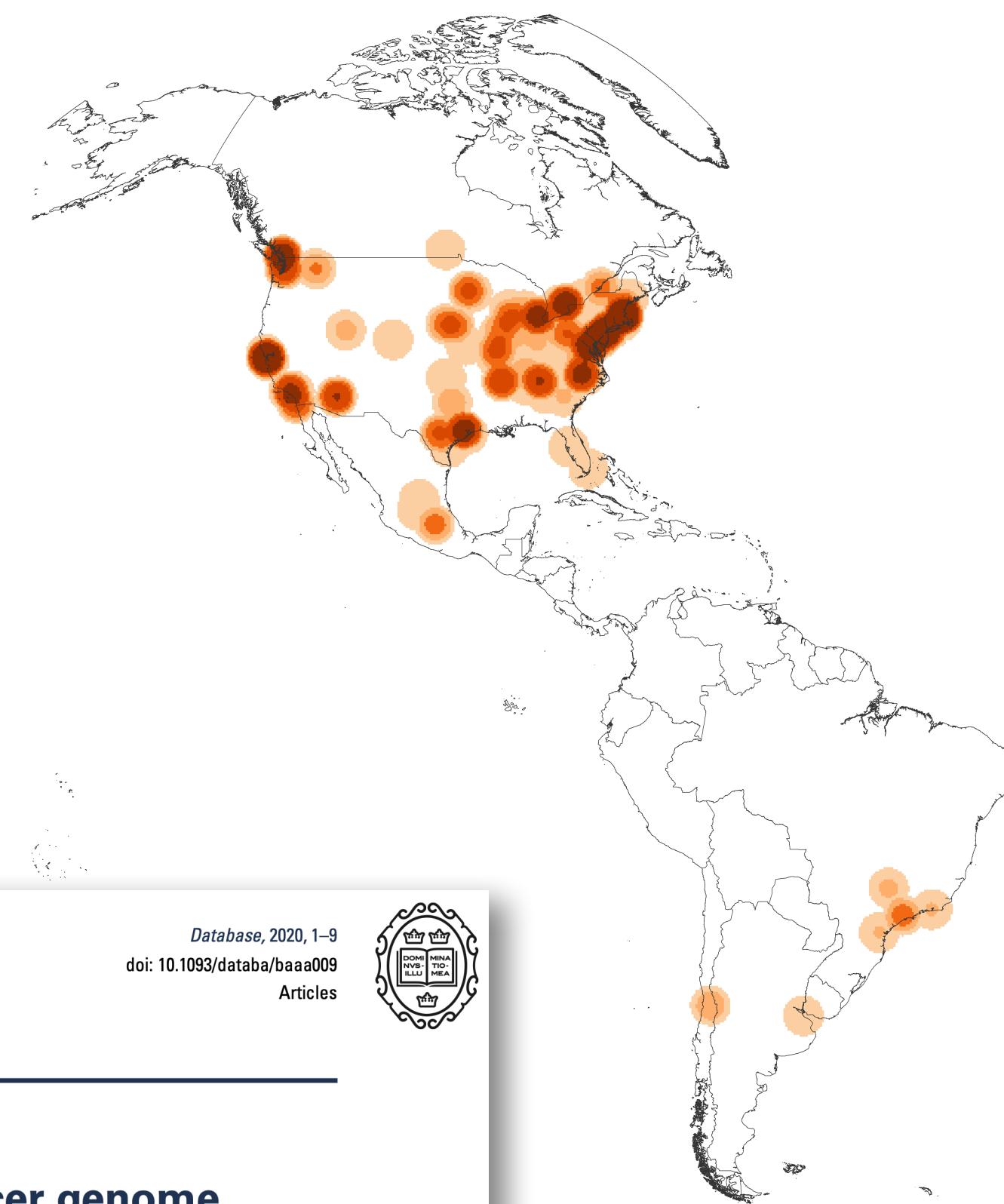
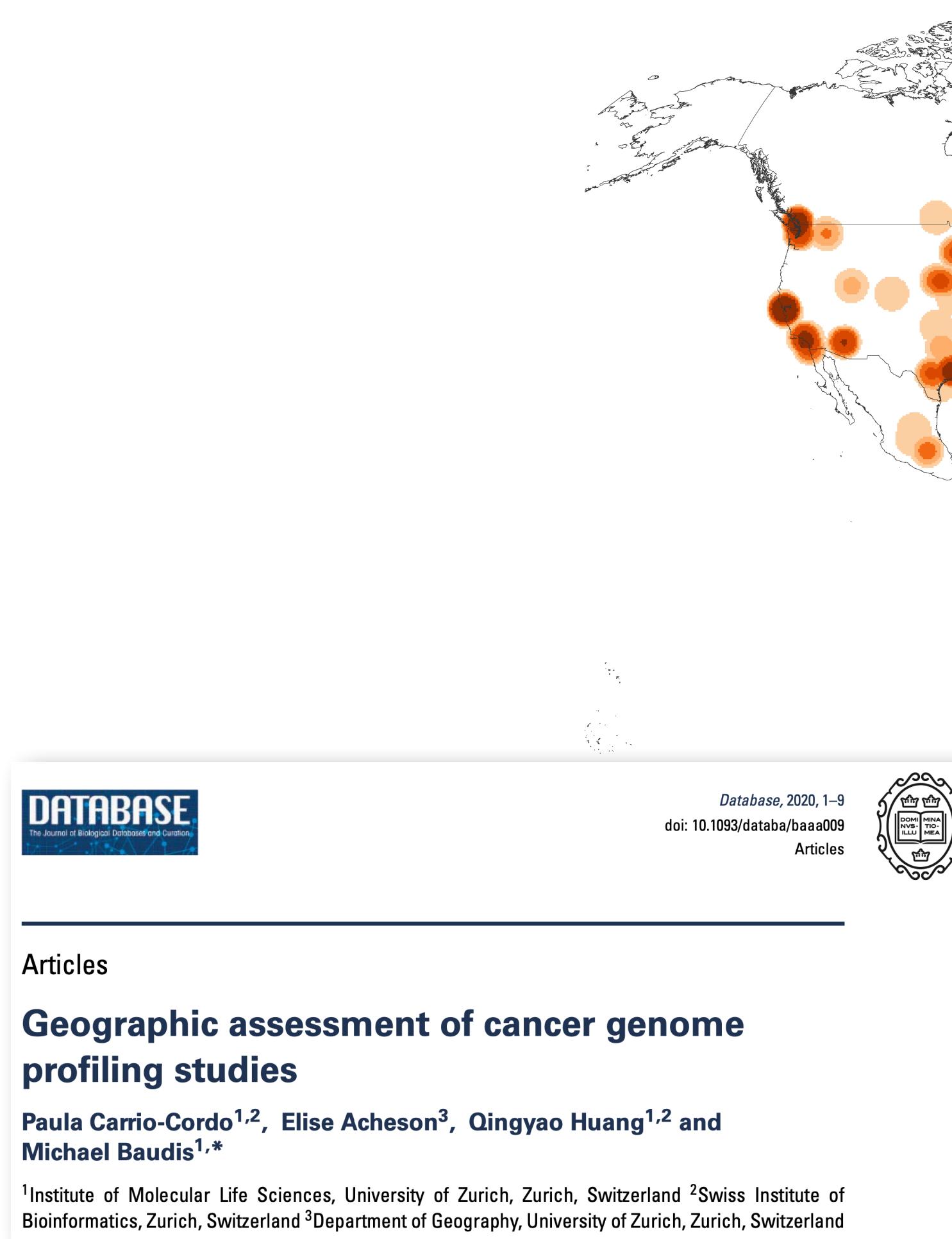
Tel.: (+41) 44 635 34 86; E-mail: michael.baudis@mls.uzh.ch



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

# Where does Genomic Data Come From?

## Geographic bias in published cancer genome profiling studies



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.



Universität  
Zürich UZH



progenet X

# The hCNV Community

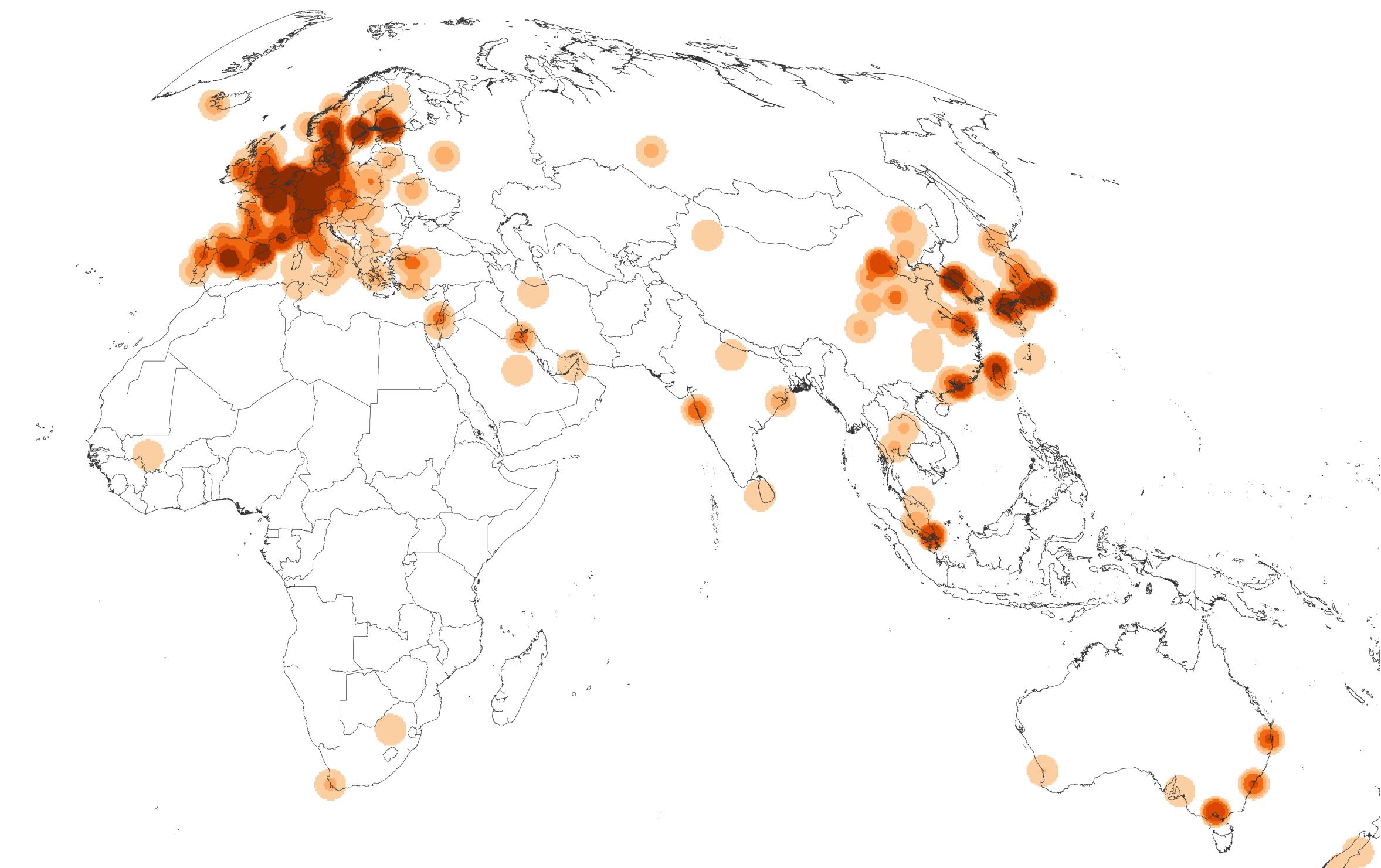
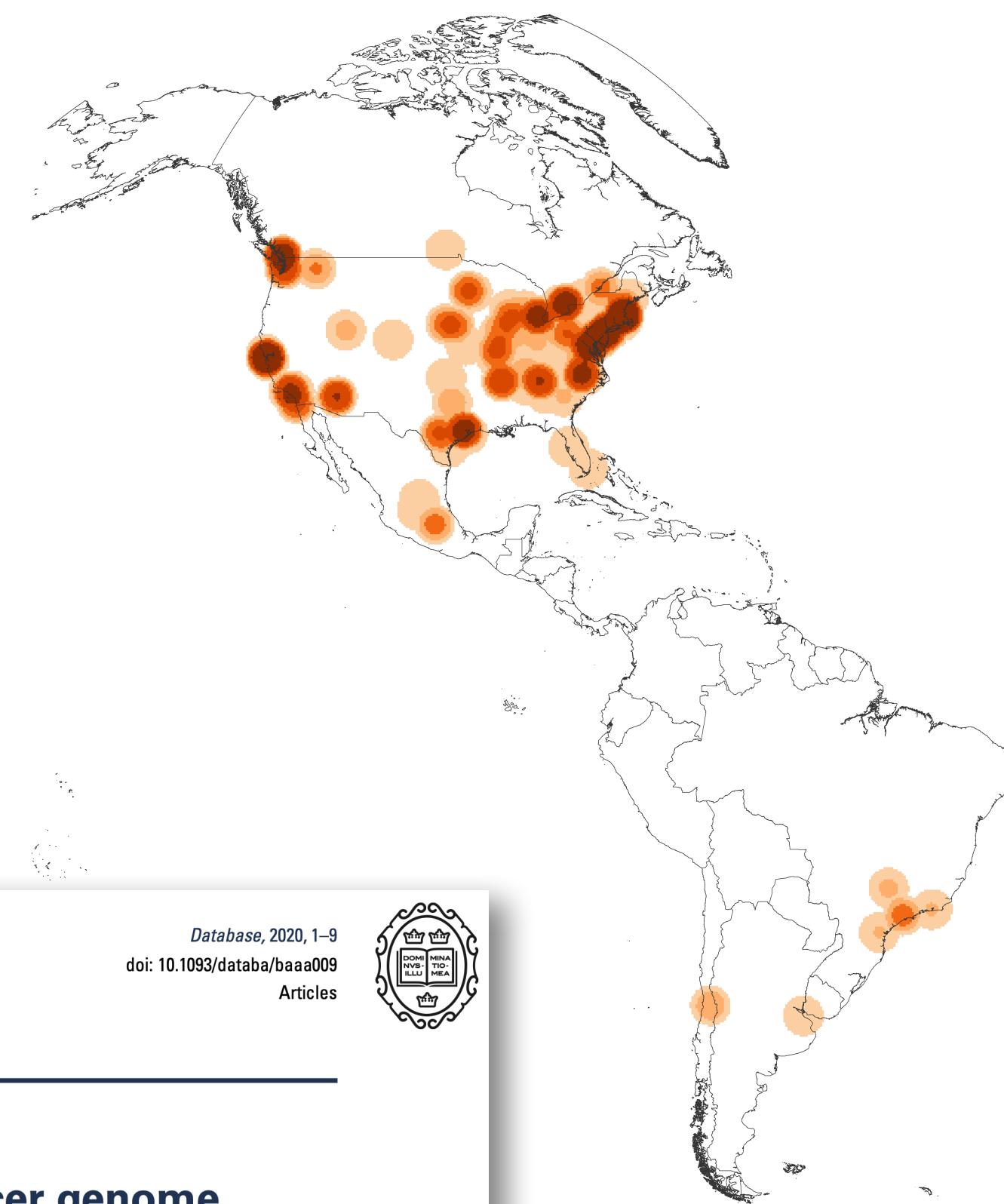
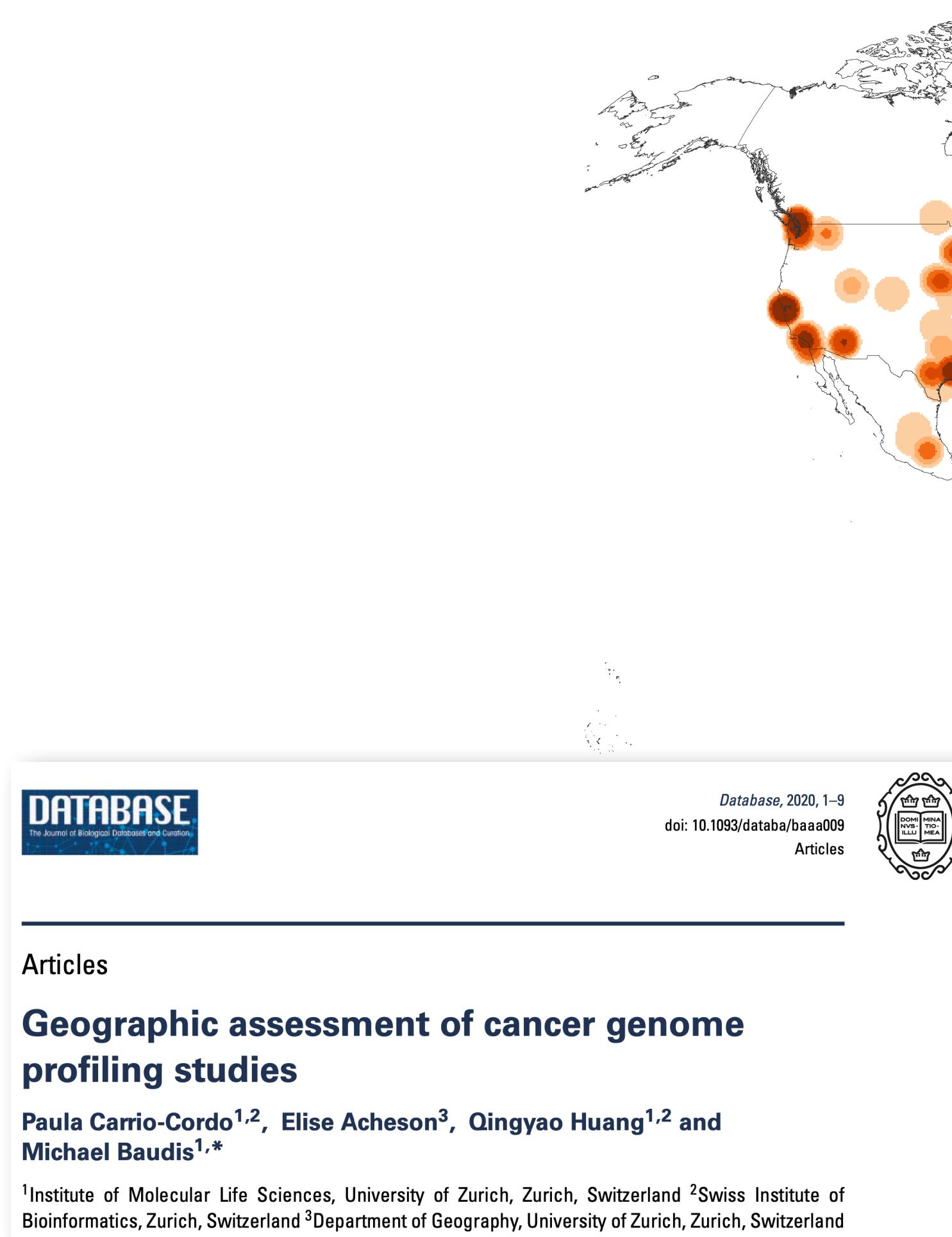
## Standards for CNV annotation and their use in data discovery protocols

Michael Baudis | ELIXIR hCNV Community Webinar 2024



# Where does Genomic Data Come From?

## Geographic bias in published cancer genome profiling studies



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.

# Different Approaches to Data Sharing



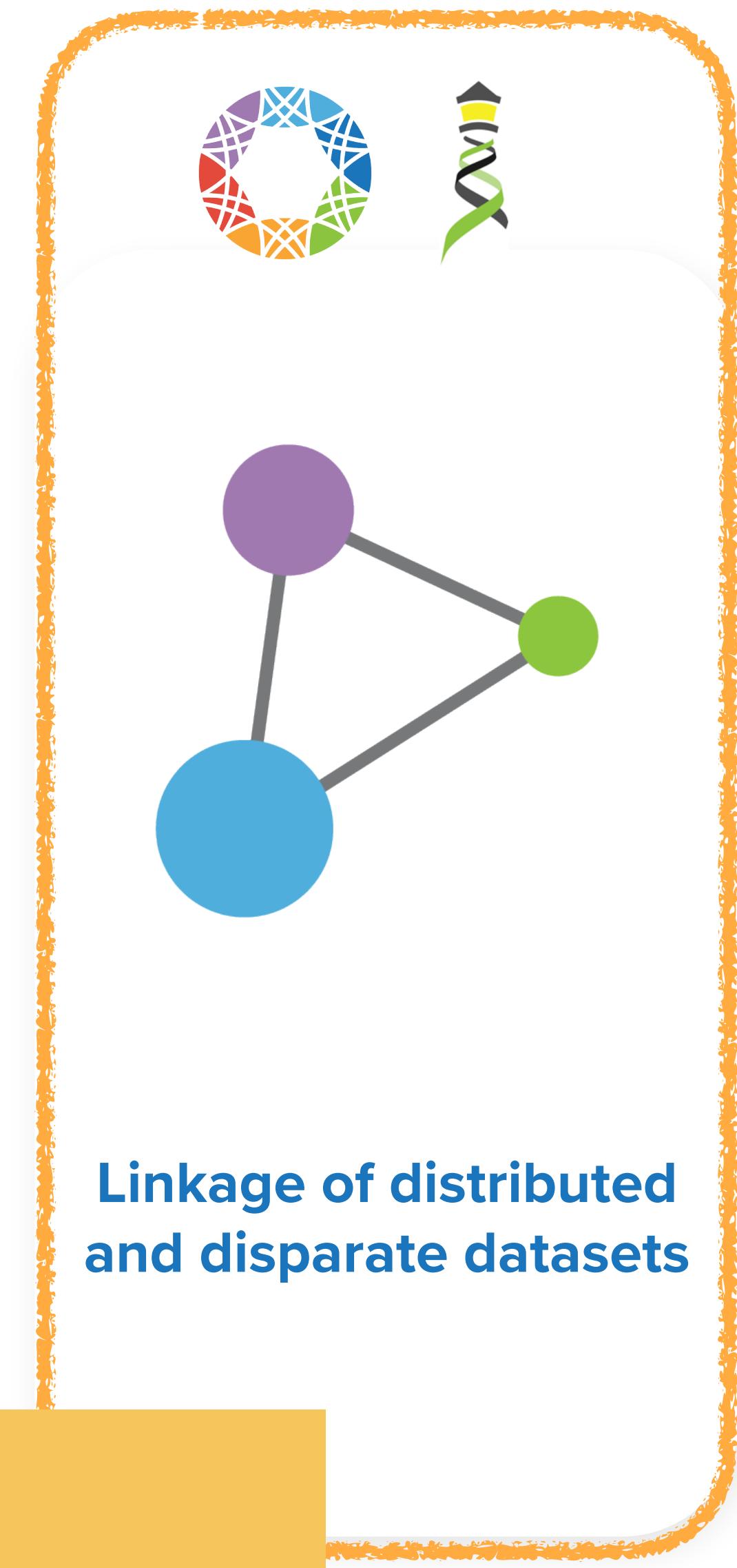
**Centralized Genomic Knowledge Bases**



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



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



**Linkage of distributed and disparate datasets**

**Federation**

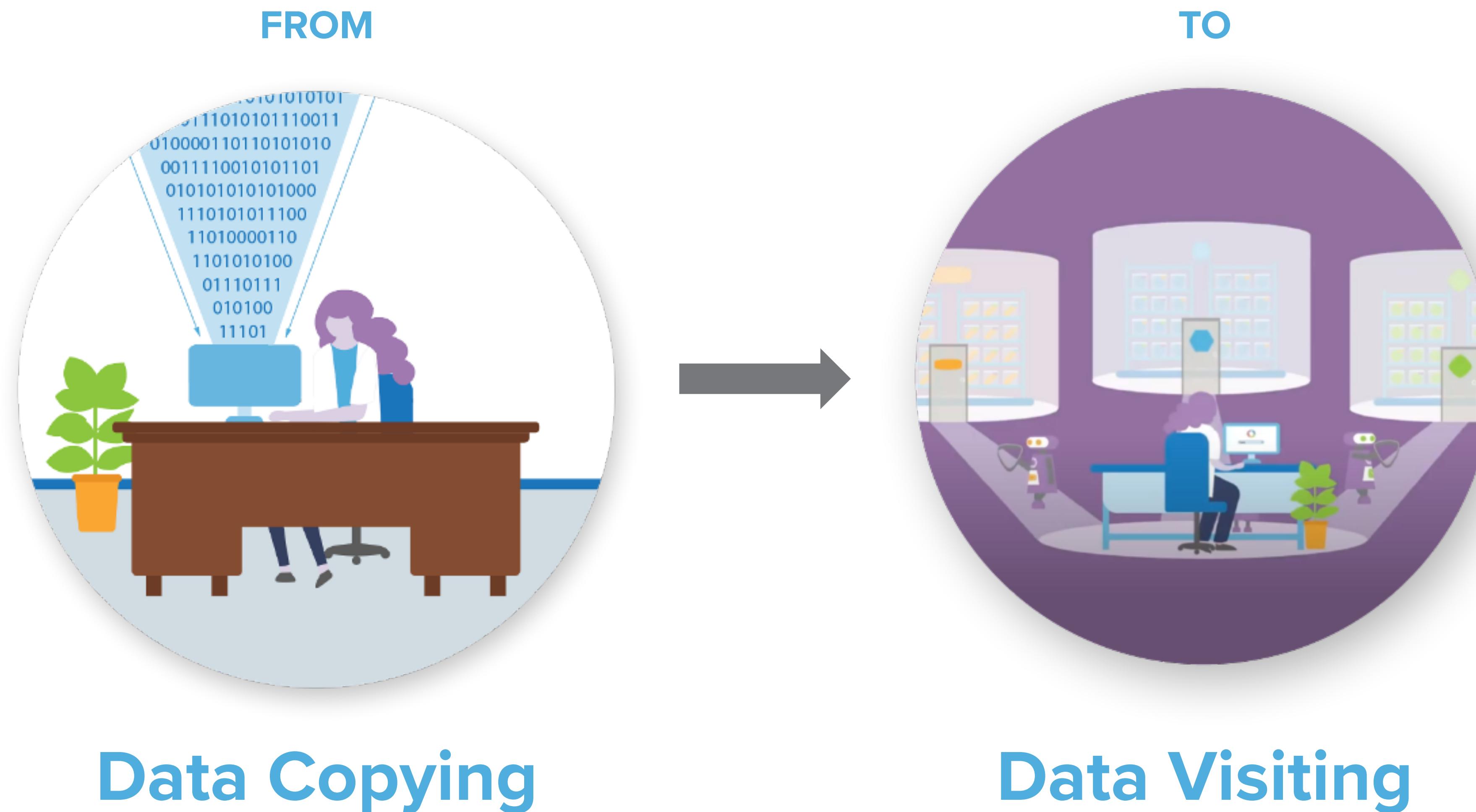
# **Beacon v2**

## **Federated Genomics**



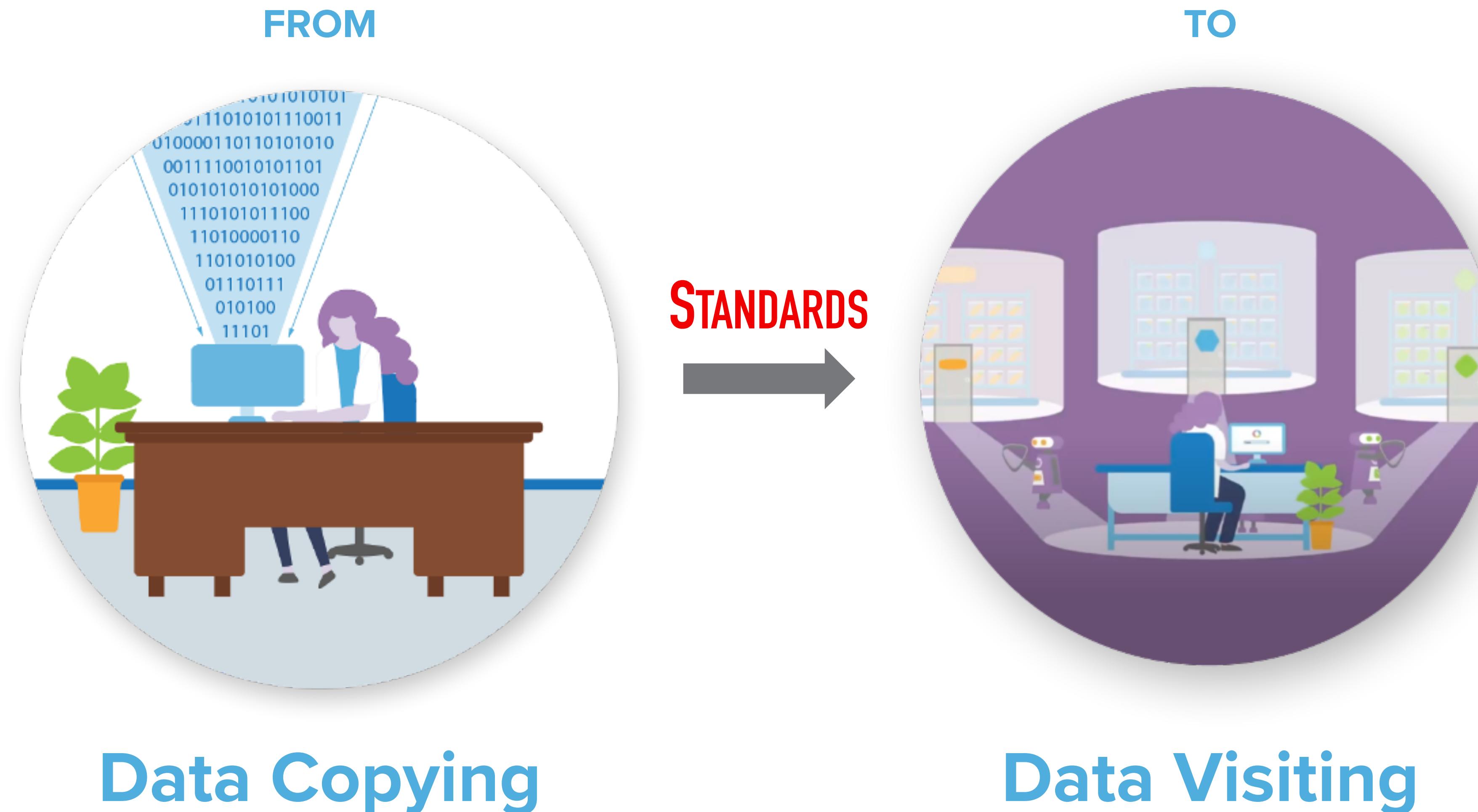


# A New Paradigm for Data Sharing





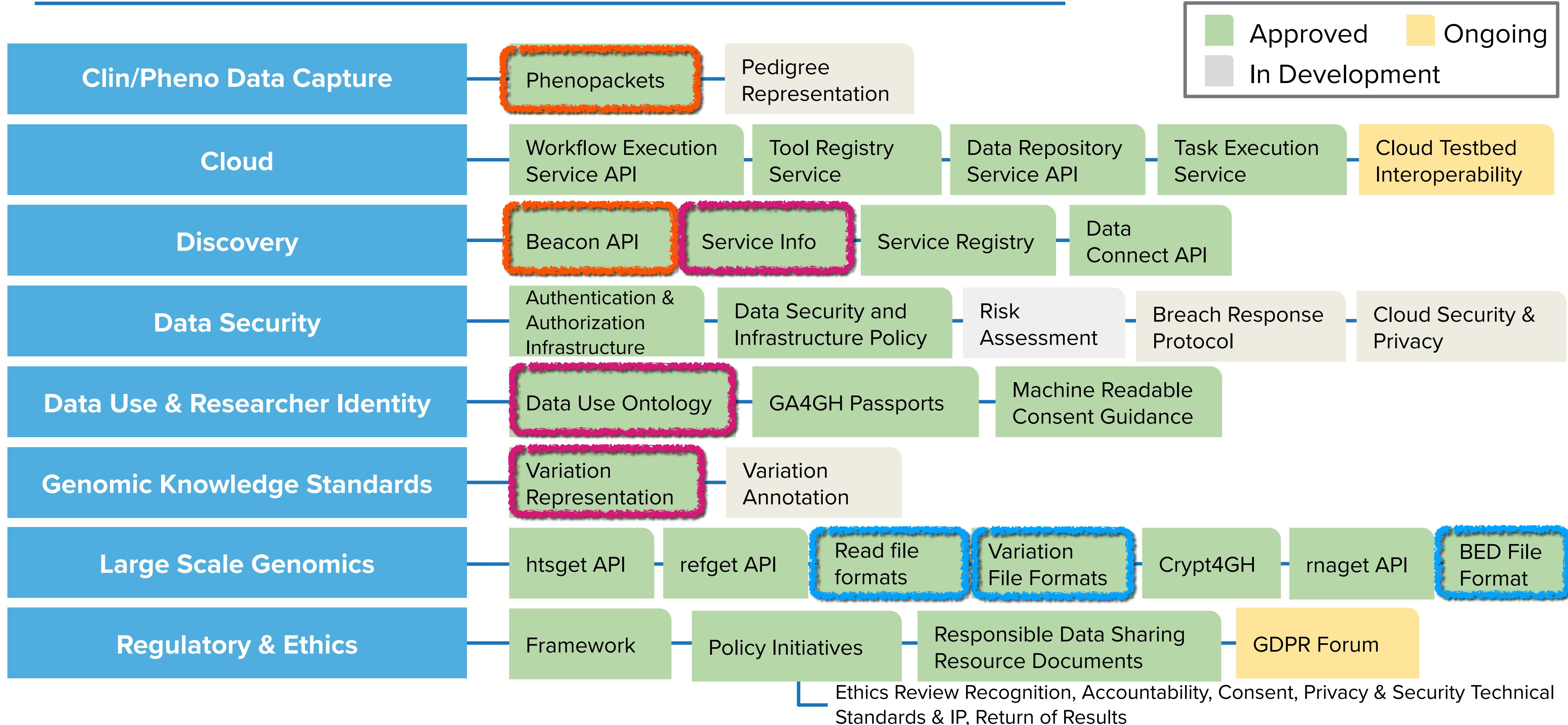
# A New Paradigm for Data Sharing



# GA4GH 2020-2022 Strategic Roadmap



**Global Alliance**  
for Genomics & Health

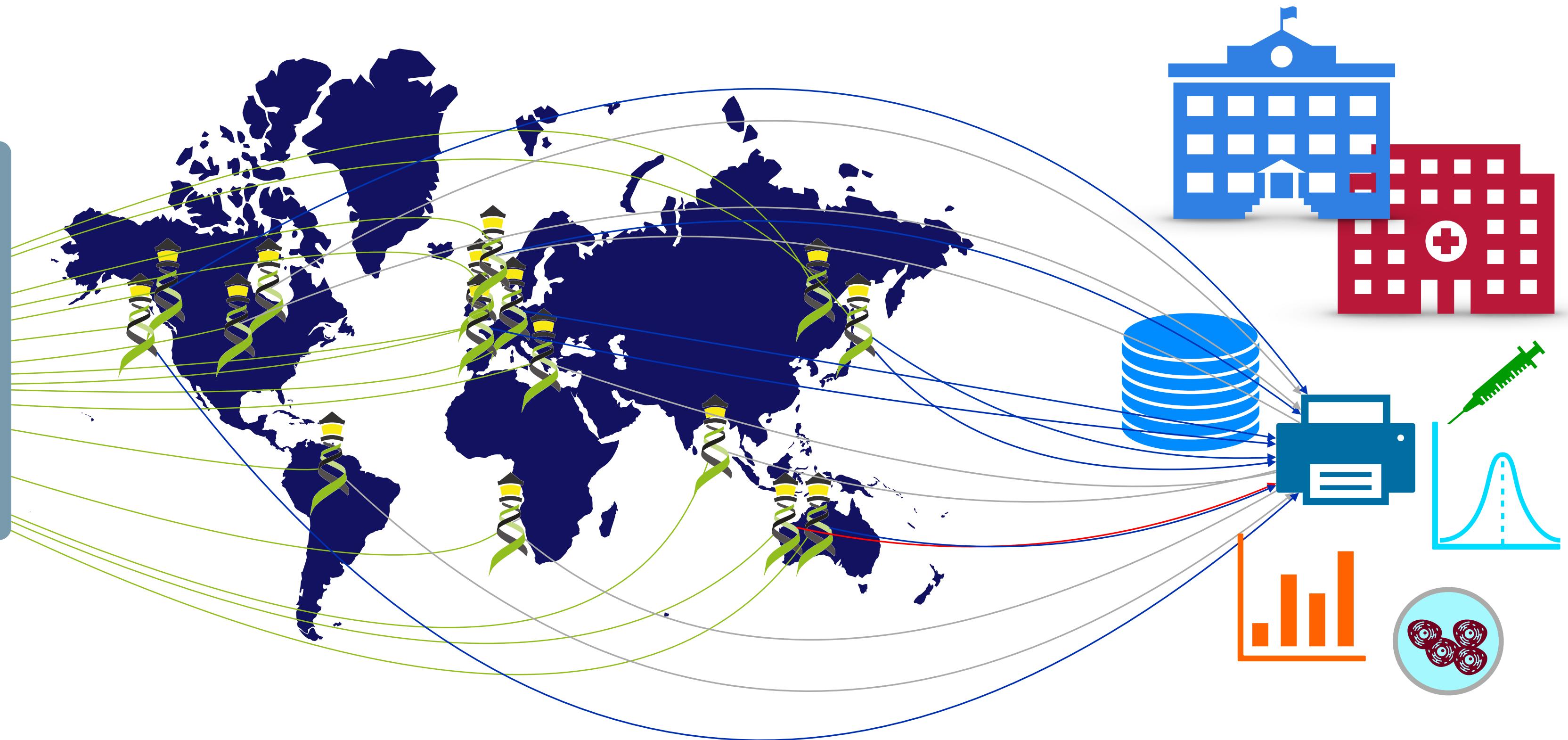




The basic Beacon v2 query asks for the replacement of a sequence by a different one of equal or different length.

`referenceName`  
`start`  
`referenceBases`  
`alternateBases`

here a chromosome name, but could be any sequence identifier  
a genomic position defined as using a 0-base, interbase format  
a sequence in the reference genome  
a sequence replacing the reference\_sequence



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

# The Beacon v2 Standard Supports Data Discovery to Support Federated Biomedical Genomics

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

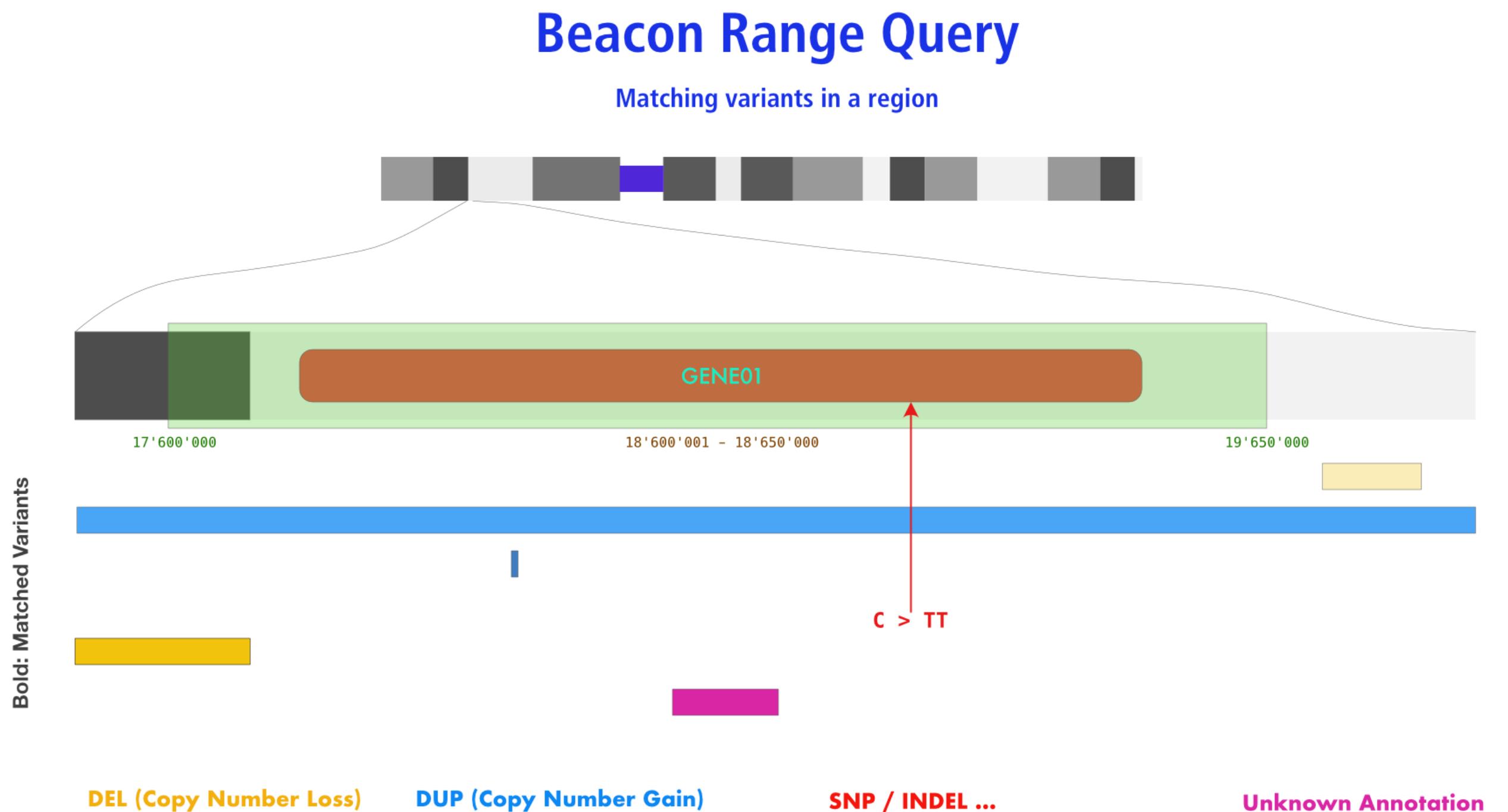
This table is maintained in parallel with the [Beacon v2 documentation](#).

EFO	Beacon	VCF	SO	GA4GH VRS <sup>1</sup>	Notes
<a href="#">EFO:0030070</a> copy number gain	DUP <sup>2</sup> or <a href="#">EFO:0030070</a>	DUP	<a href="#">SO:0001742</a> copy_number_gain	<a href="#">EFO:0030070</a> gain	a sequence alteration whereby the copy number of a given genomic region is greater than the reference sequence
<a href="#">EFO:0030071</a> low-level copy number gain	DUP <sup>2</sup> or <a href="#">EFO:0030071</a>	DUP	<a href="#">SO:0001742</a> copy_number_gain	<a href="#">EFO:0030071</a> low-level gain	
<a href="#">EFO:0030072</a> high-level copy number gain	DUP <sup>2</sup> or <a href="#">EFO:0030072</a>	DUP	<a href="#">SO:0001742</a> copy_number_gain	<a href="#">EFO:0030072</a> high-level gain	commonly but not consistently used for >=5 copies on a bi-allelic genome region
<a href="#">EFO:0030073</a> focal genome amplification	DUP <sup>2</sup> or <a href="#">EFO:0030073</a>	DUP	<a href="#">SO:0001742</a> copy_number_gain	<a href="#">EFO:0030072</a> high-level gain <sup>4</sup>	commonly but not consistently used for >=5 copies on a bi-allelic genome region, of limited size (operationally max. 1-5Mb)
<a href="#">EFO:0030067</a> copy number loss	DEL <sup>2</sup> or <a href="#">EFO:0030067</a>	DEL	<a href="#">SO:0001743</a> copy_number_loss	<a href="#">EFO:0030067</a> loss	a sequence alteration whereby the copy number of a given genomic region is smaller than the reference sequence
<a href="#">EFO:0030068</a> low-level copy number loss	DEL <sup>2</sup> or <a href="#">EFO:0030068</a>	DEL	<a href="#">SO:0001743</a> copy_number_loss	<a href="#">EFO:0030068</a> low-level loss	
<a href="#">EFO:0020073</a> high-level copy number loss	DEL <sup>2</sup> or <a href="#">EFO:0020073</a>	DEL	<a href="#">SO:0001743</a> copy_number_loss	<a href="#">EFO:0020073</a> high-level loss	a loss of several copies; also used in cases where a complete genomic deletion cannot be asserted
<a href="#">EFO:0030069</a> complete genomic deletion	DEL <sup>2</sup> or <a href="#">EFO:0030069</a>	DEL	<a href="#">SO:0001743</a> copy_number_loss	<a href="#">EFO:0030069</a> complete genomic loss	complete genomic deletion (e.g. homozygous deletion on a bi-allelic genome region)

# Beacon Queries

## Range ("anything goes") Request

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



beaconplus.progenetix.org

### Beacon Query Types

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

#### Dataset

Test Database - examplez X

#### 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

7572826  
7579005

### Query Database

#### Form Utilities

Gene Spans    Cytoband(s)

#### Query Examples

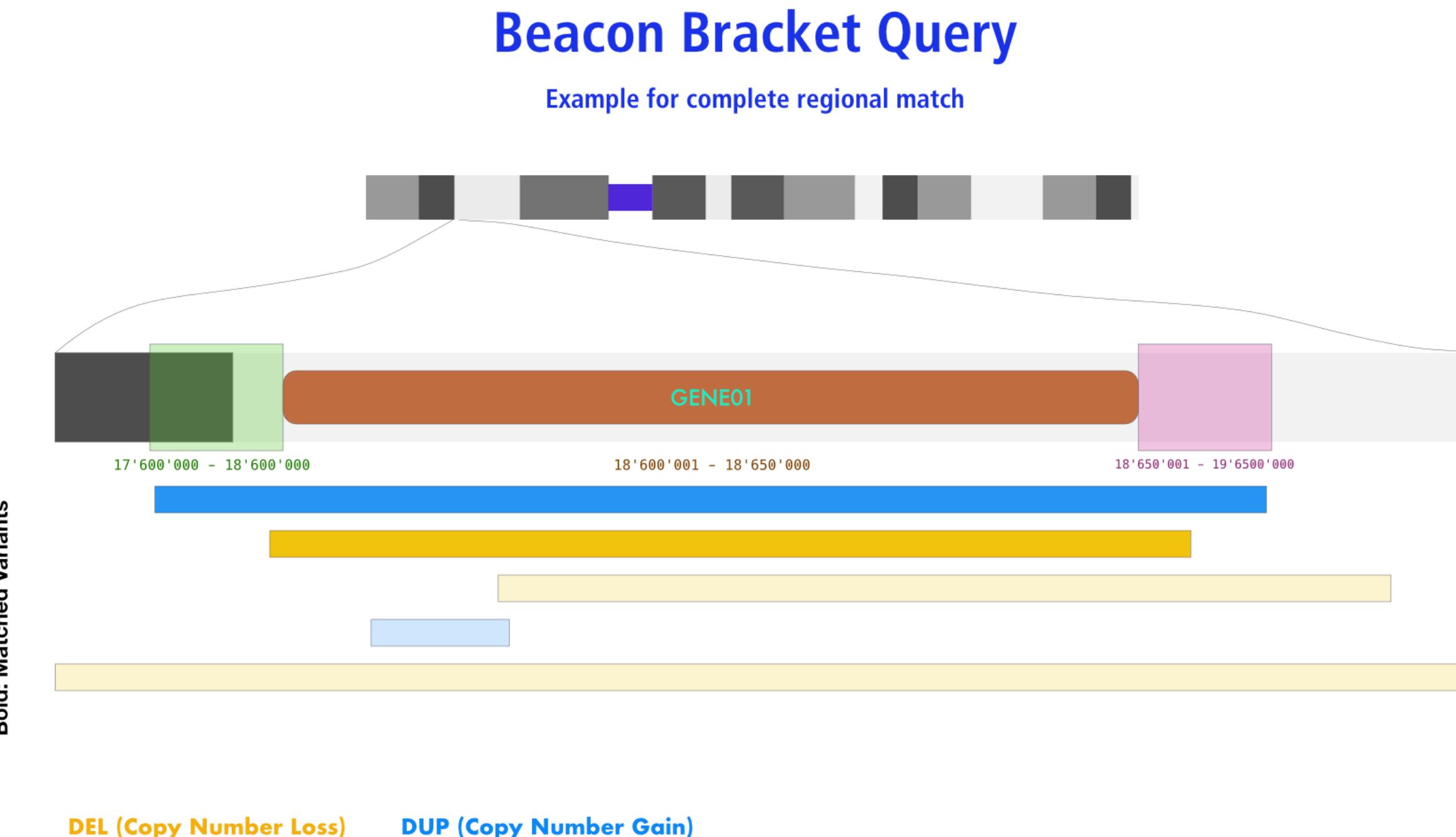
CNV Example   SNV Example   Range Example   Gene Match  
Aminoacid Example   Identifier - HeLa

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

# Beacon Queries

## Bracket ("CNV") Query

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



### Beacon Query Types

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

#### Dataset

Test Database - examplez X | ▼

#### 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) X | ▼

#### Chromosome 9

21000001-21975098



### Query Database

#### Form Utilities

⚙️ Gene Spans

⚙️ Cytoband(s)

#### Query Examples

[CNV Example](#)

[SNV Example](#)

[Range Example](#)

[Gene Match](#)

[Aminoacid Example](#)

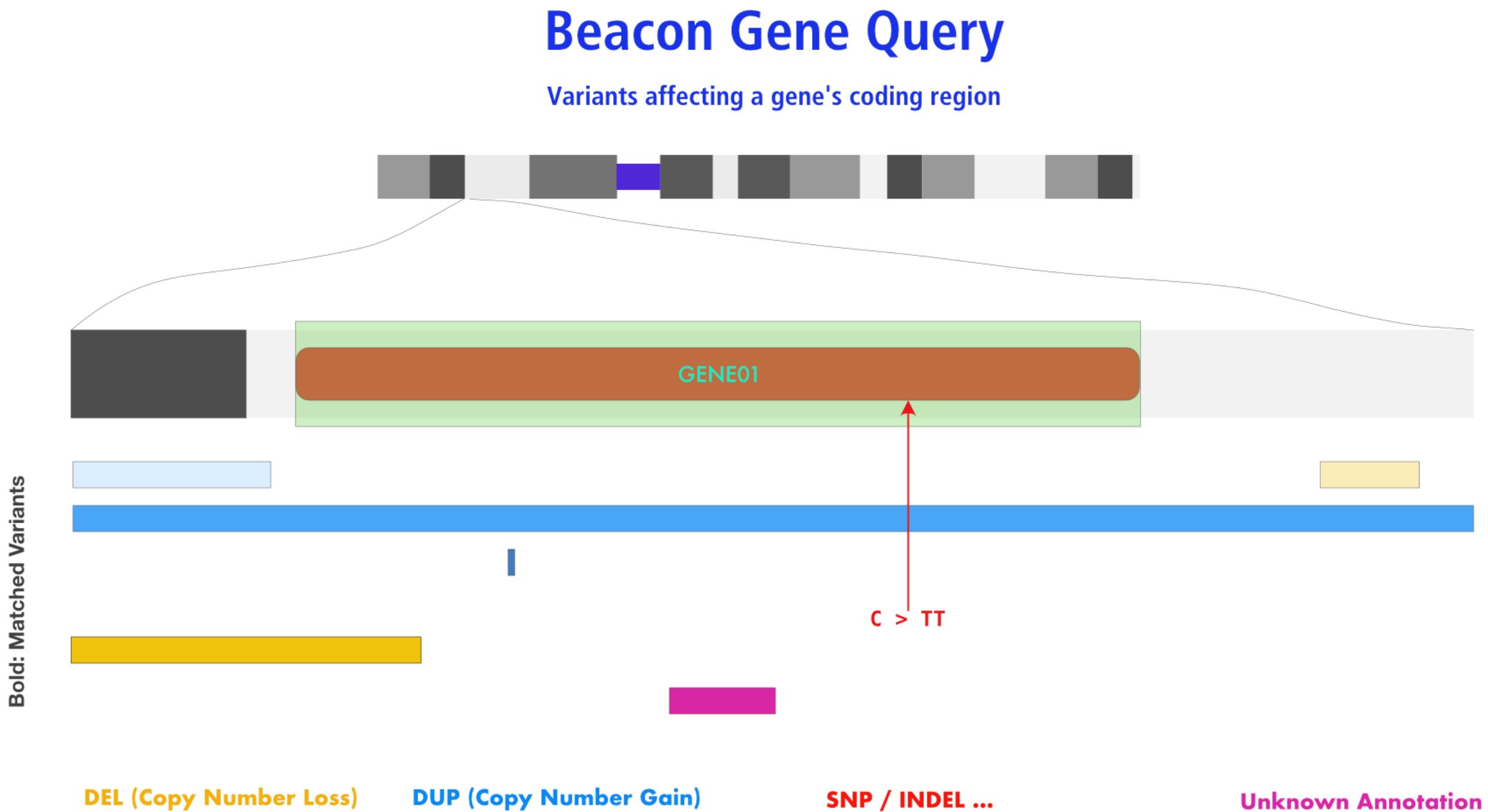
[Identifier - HeLa](#)

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.  $\leq \sim 2\text{Mbp}$  in size). The query is against the examplez collection and can be modified e.g. through changing the position parameters or data source.

# Beacon Queries

## Gene Request

- defined through a (HUGO) gene symbol
- assuming hit on the gene's CDR but YMMV



## Beacon Query Types

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

### Dataset

Cancer Cell Lines Collection x | ▾

### Gene Symbol i

CDK2 (12:55966830-55972789) x | ▾

### Variant Type i

Select...

### Min Variant Length i

### Max Variant Length i

### Alternate Base(s)

A

### Select Filters i

Select...

### Query Database

### Form Utilities

Gene Spans

Cytoband(s)

### Query Examples

[CNV Example](#)

[SNV Example](#)

[Range Example](#)

[Gene Match](#)

[Aminoacid Example](#)

[Identifier - HeLa](#)

Beacons in v2 can support the discovery of variants with overlap with the genomic location of a gene, indicated by its symbol (e.g. `CDK2` ). Additional parameters can *optionally* be used to make matches more specific:

- `variantMinLength` and `variantMaxLength` to limit matched CNV sizes
- `genomicAlleleShortForm` (e.g. `V600E` with `BRAF` )
- `variantType` and `alternateBases` to specify variants

# pgxRpi

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

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

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.

Bioconductor

### 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](https://doi.org/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@uzh.ch](mailto:hangjia.zhao@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](https://doi.org/10.18129/B9.bioc.pgxRpi), R package version 0.99.9, <https://bioconductor.org/packages/pgxRpi>.

# Beacon Queries

## Missing or ill defined options

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

→ **Beacon & hCNV Scout Team**

**Beacon Query Types**

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

Dataset: Test Database - examplez | X | ▾

Chromosome: Select... Variant Type: Select...

Start or Position: 19000001-21975098

Reference Base(s): N Alternate Base(s): A

Select Filters: Select...

**Query Database**

Form Utilities: Gene Spans Cytoband(s)

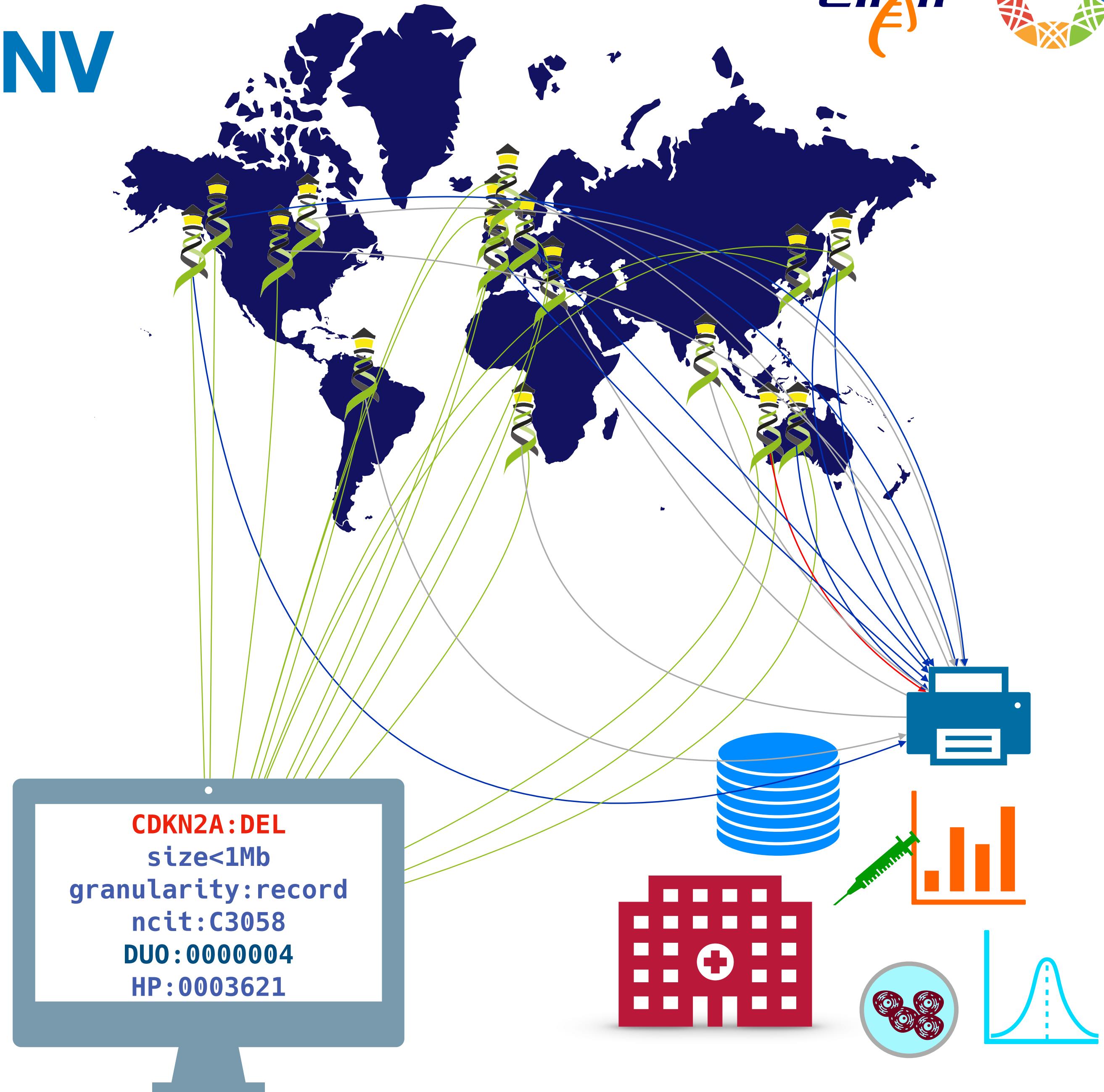
Query Examples: CNV Example SNV Example Range Example Gene Match

Aminoacid Example Identifier - HeLa

# Why to engage with hCNV

- most genomic projects (rare diseases, cancer ...) should require CNV analysis components
- recent standards by hCNV & GA4GH solve problems and help with misconceptions
- the Beacon project is a target for CNV standards testing and use
- we need challenges by communities to demonstrate the current state of the art - and improve it

→ Collaborate w/ hCNV!



# Why to engage with hCNV

- work on benchmarking of human CNV Datasets for clinical applications
- improve annotation of CNV features
- define standard protocols for federated-learning within Elixir
- develop or use CNV workflows on Galaxy
- implement data discovery of genomic, clinical and cohort data over the Beacon protocol

→ Collaborate w/ hCNV!

