

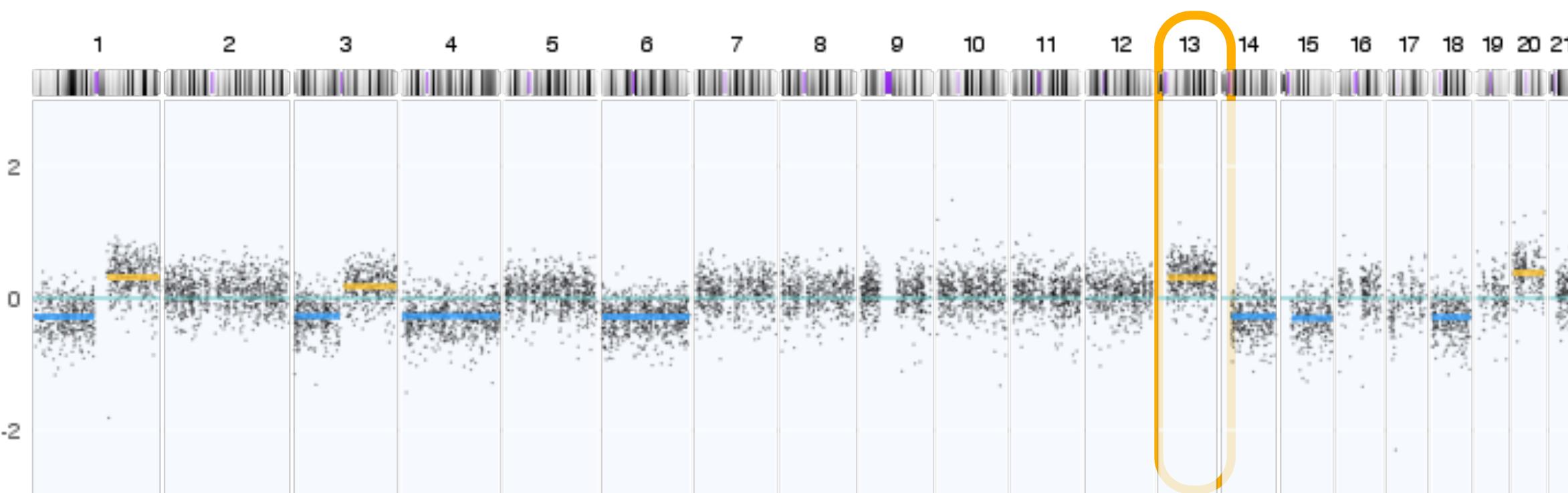
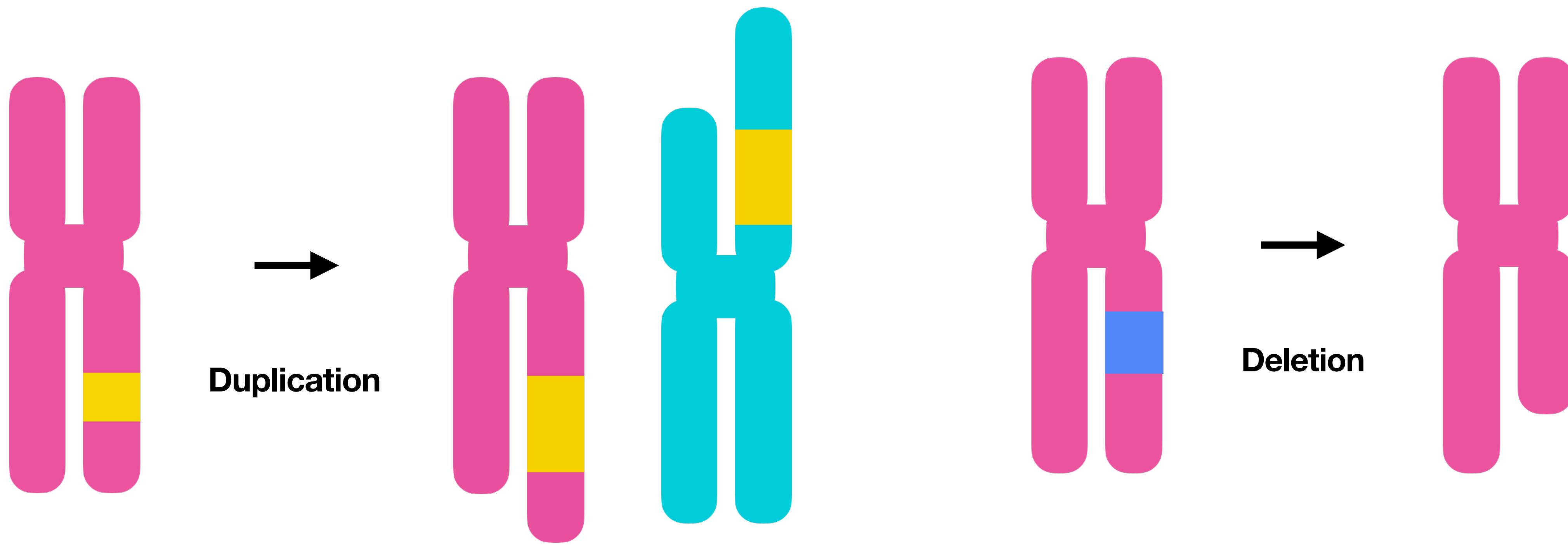


ELIXIR hCNV Community

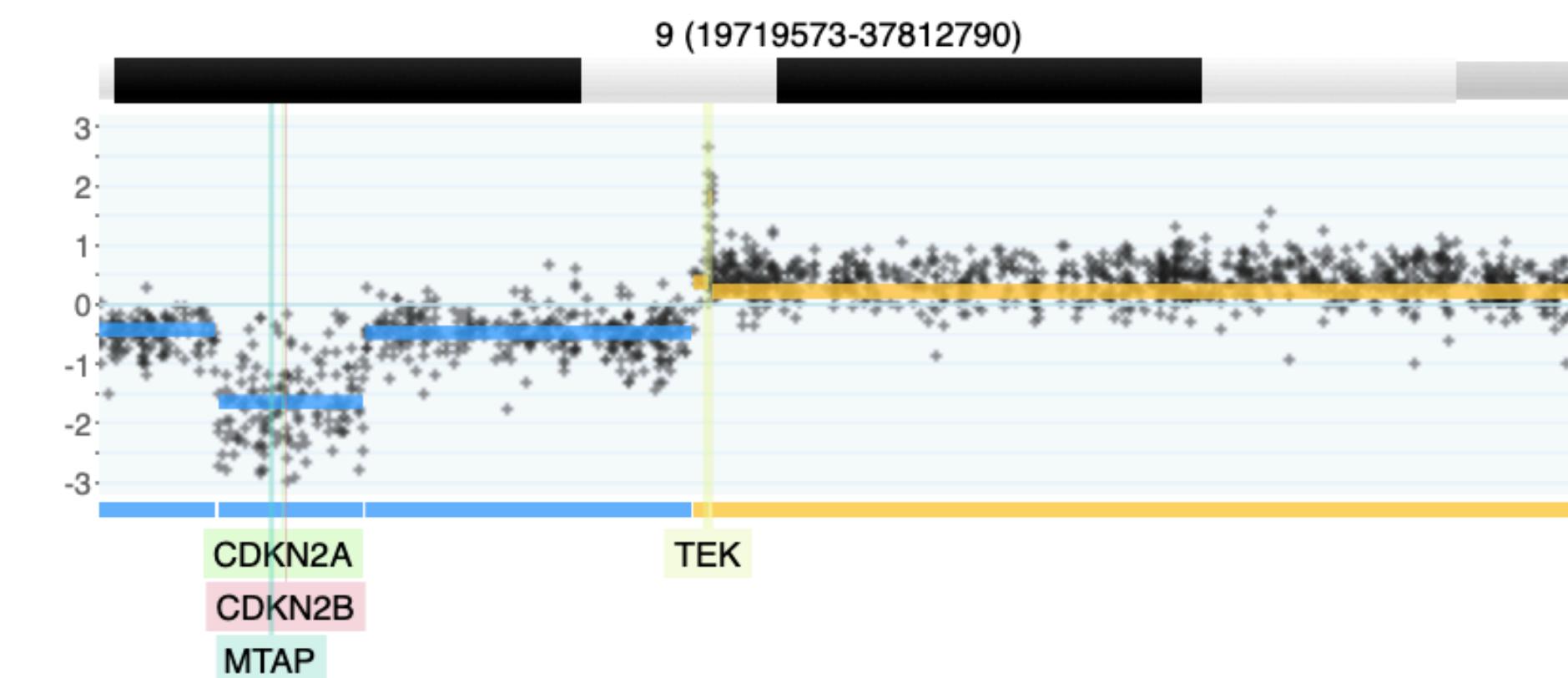
Michael Baudis | ELIXIR hCNV Community Webinar 2024

www.elixir-europe.org

Somatic Copy Number Variation



Gain of chromosome arm 13q in colorectal carcinoma

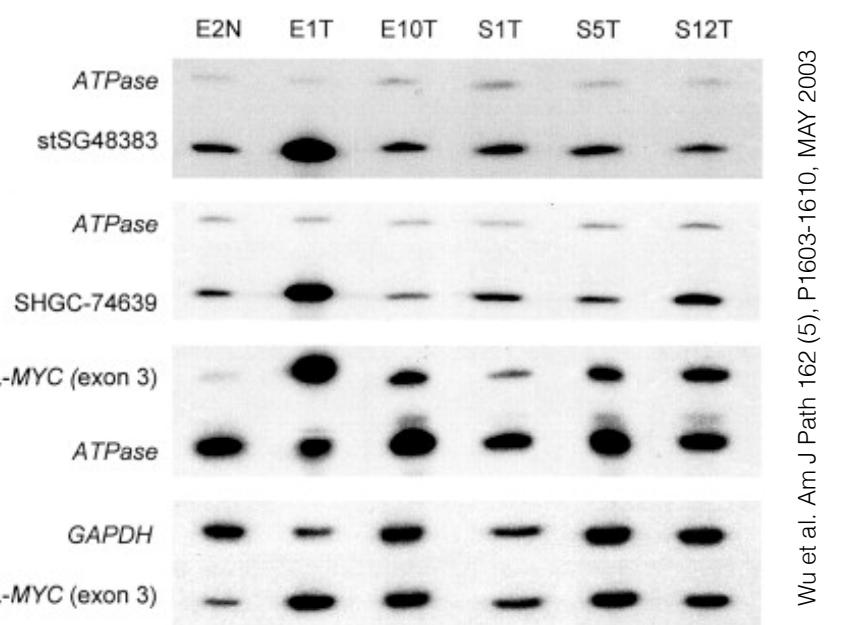
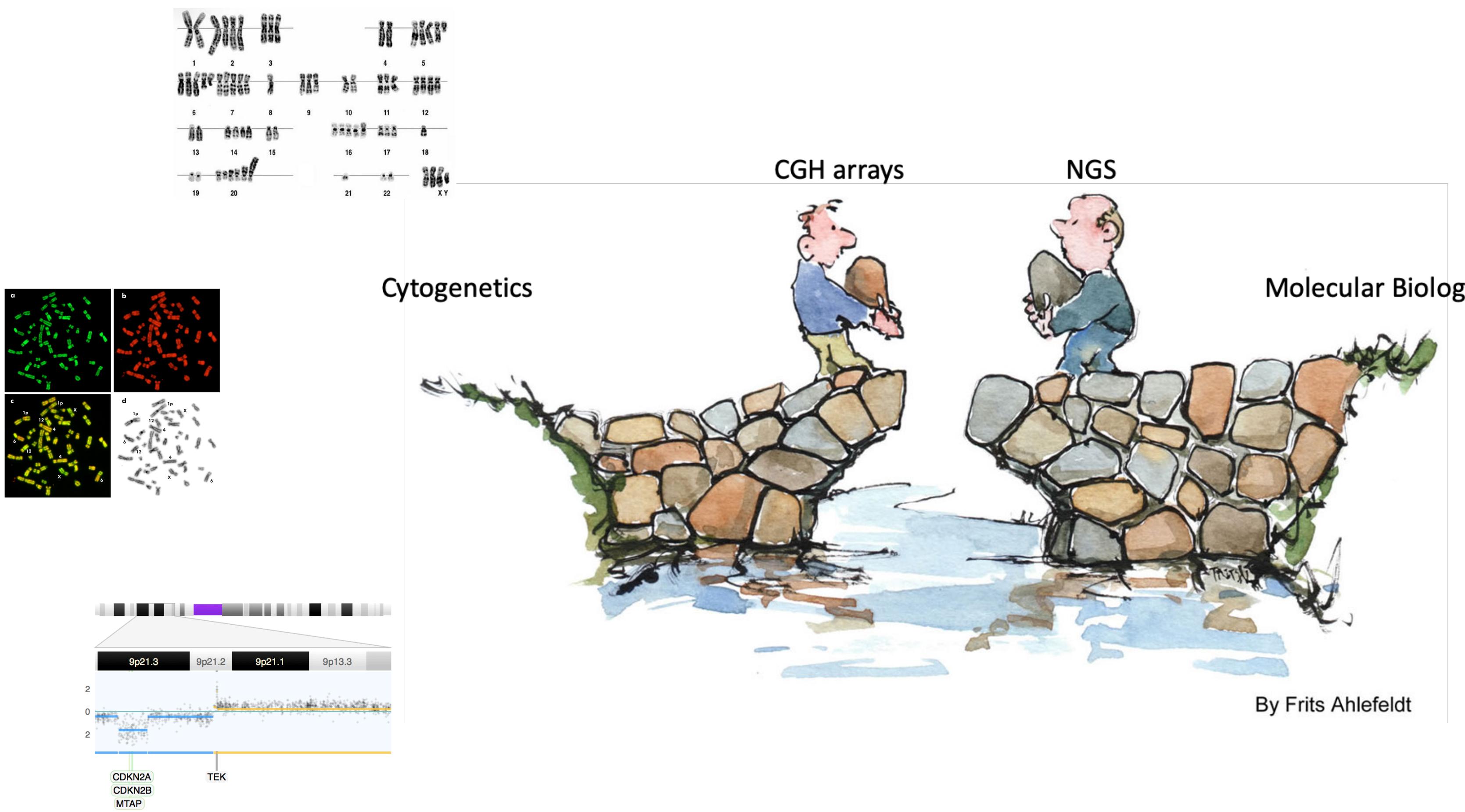


2-event, homozygous deletion in a Glioblastoma

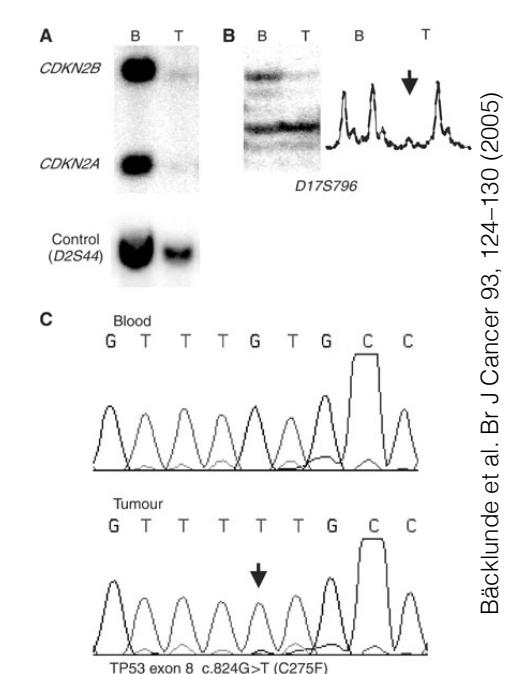
h-CNV scientific context

- 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

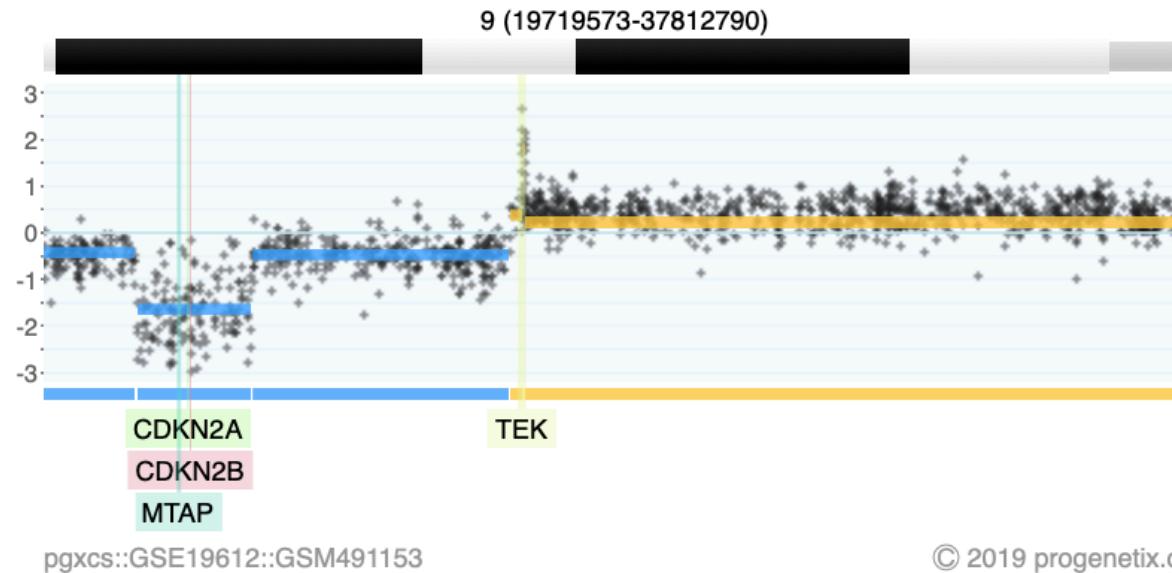


Theoretical Cytogenetics and Oncogenomics

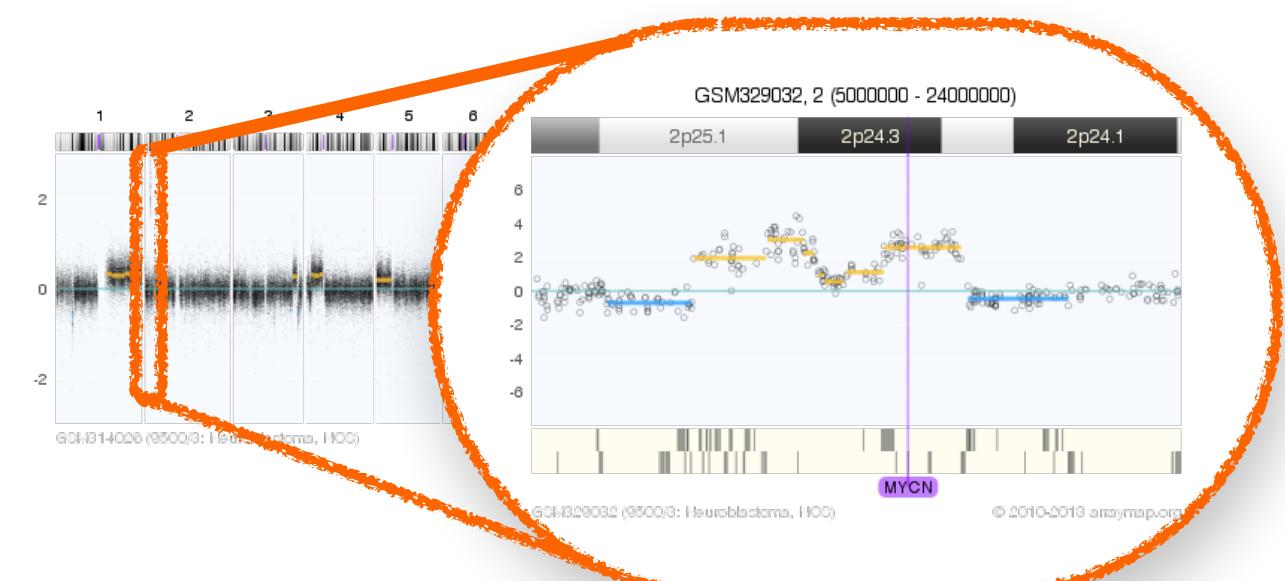
Research | Methods | Standards

Genomic Imbalances in Cancer - Copy Number Variations (CNV)

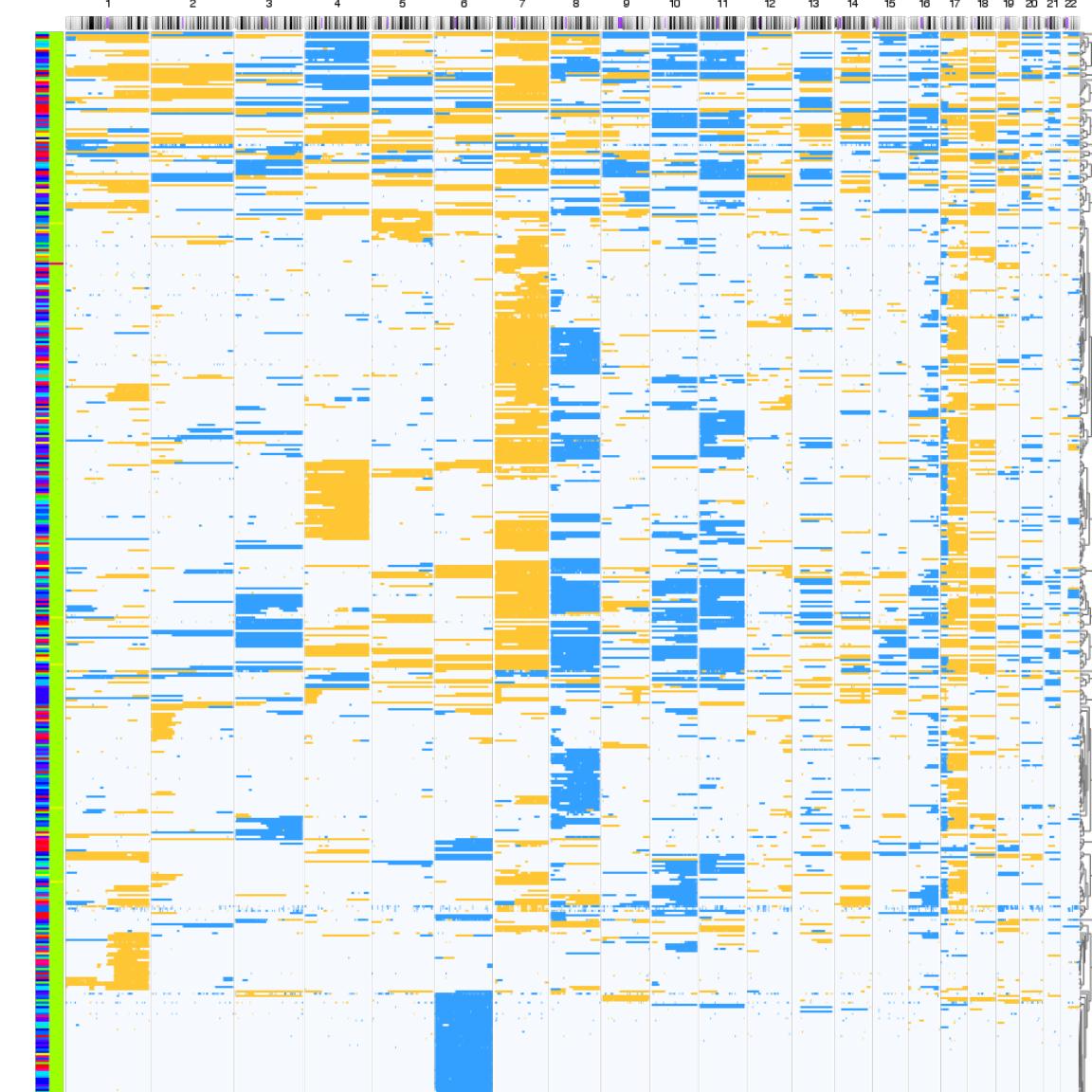
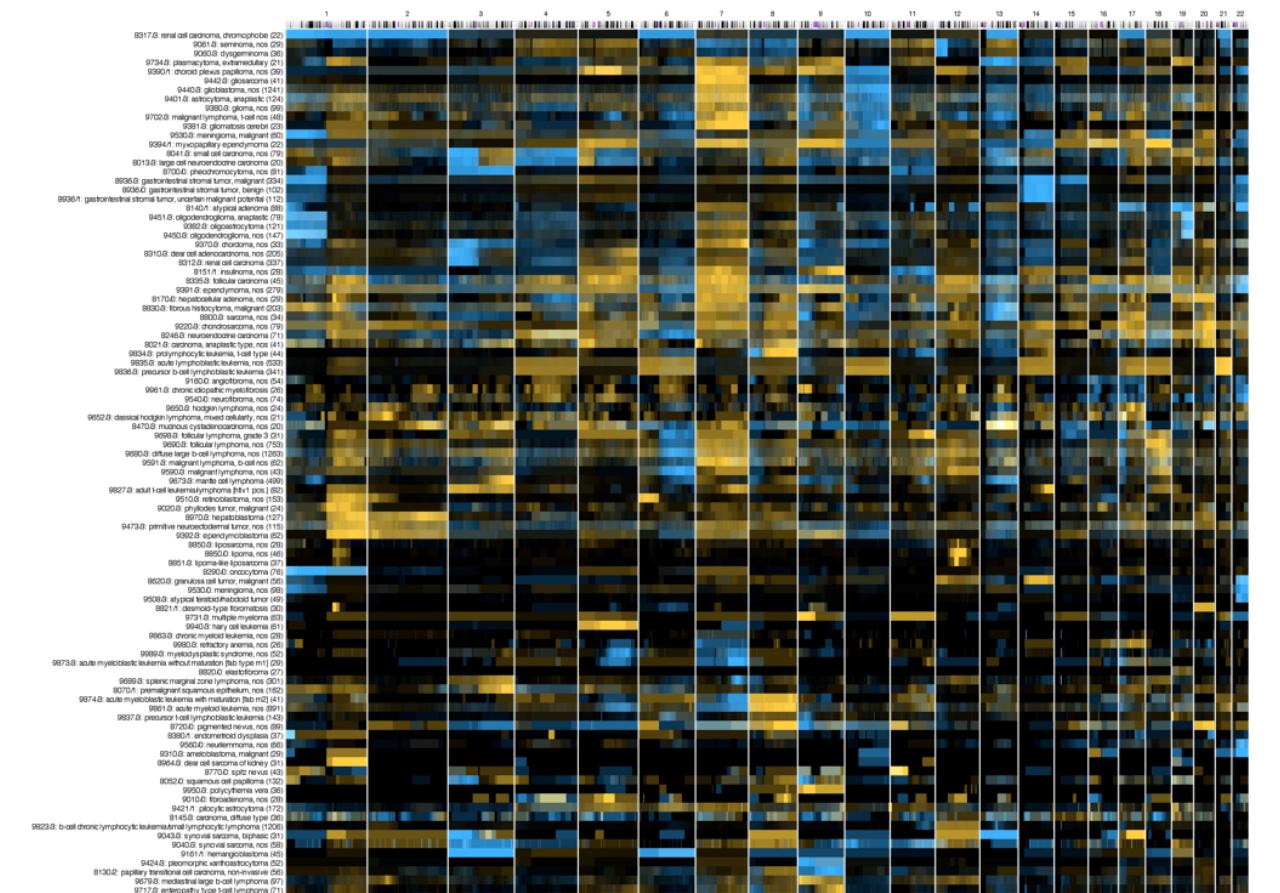
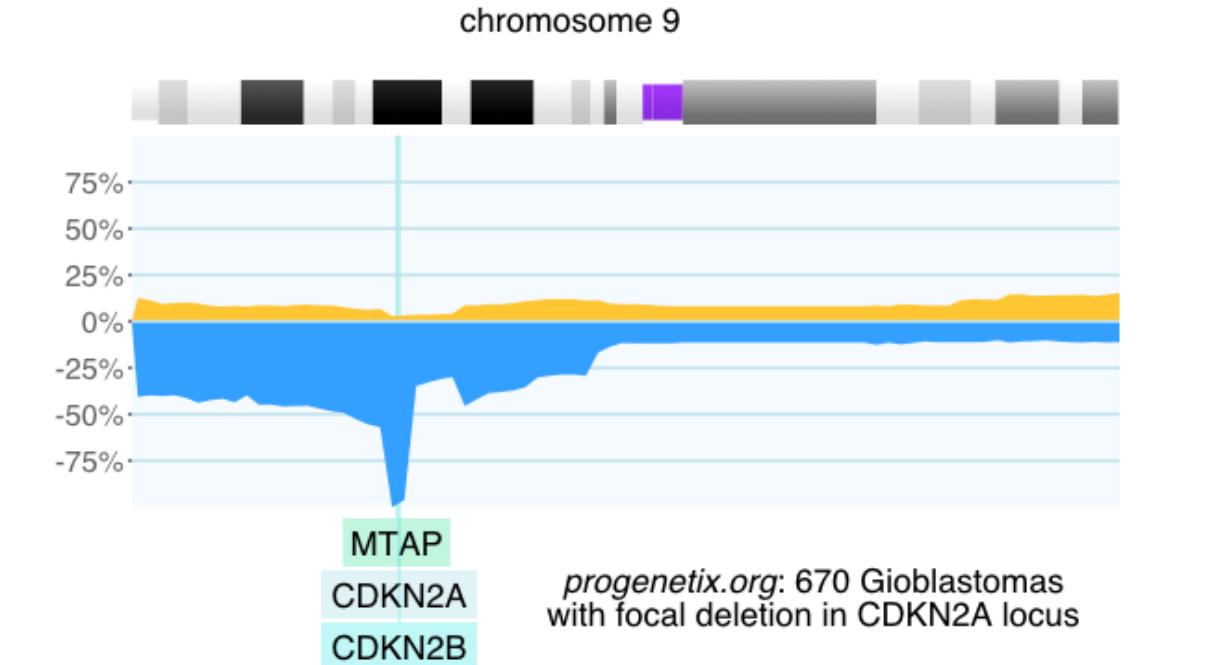
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)



2-event, homozygous deletion in a Glioblastoma

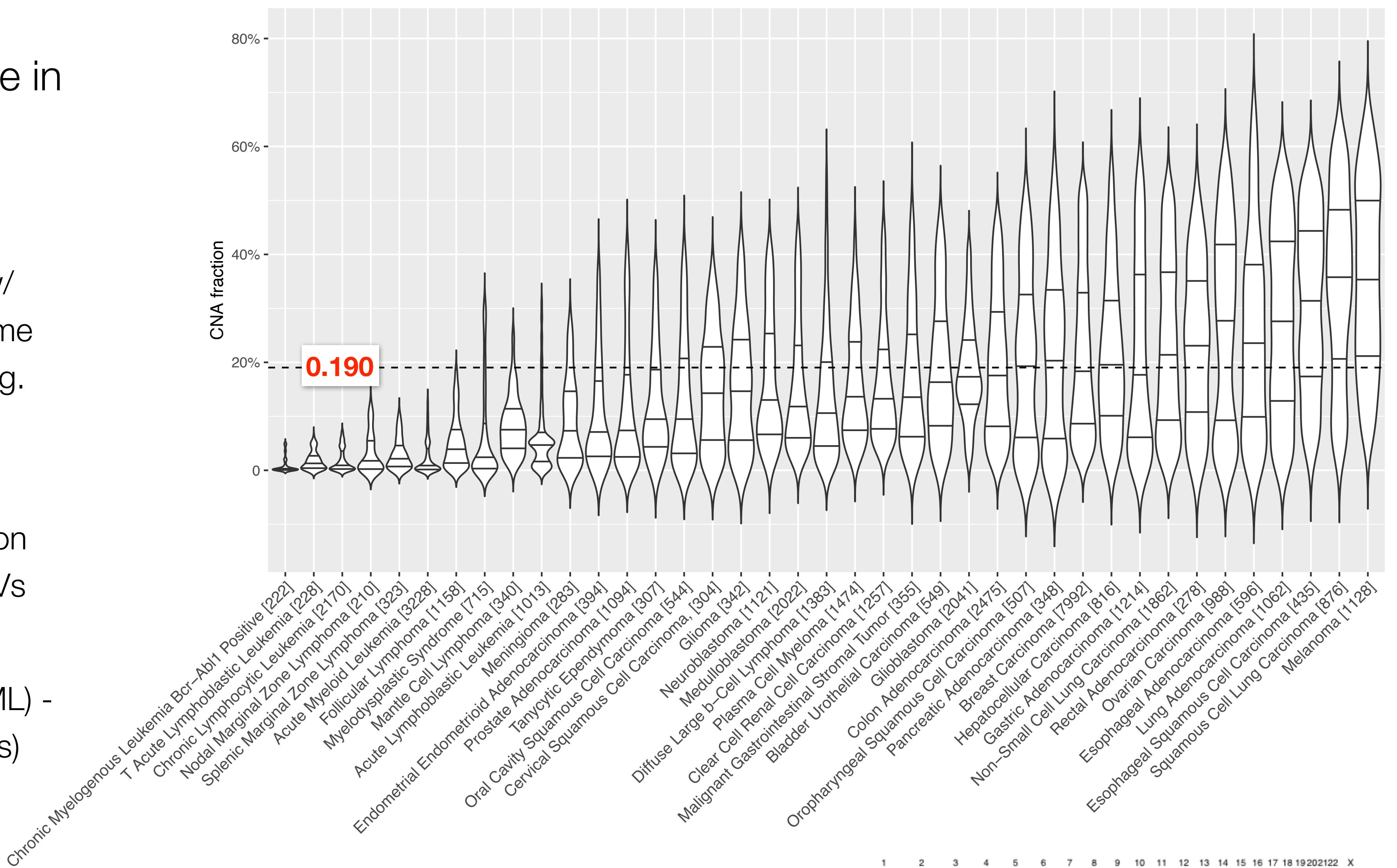


MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

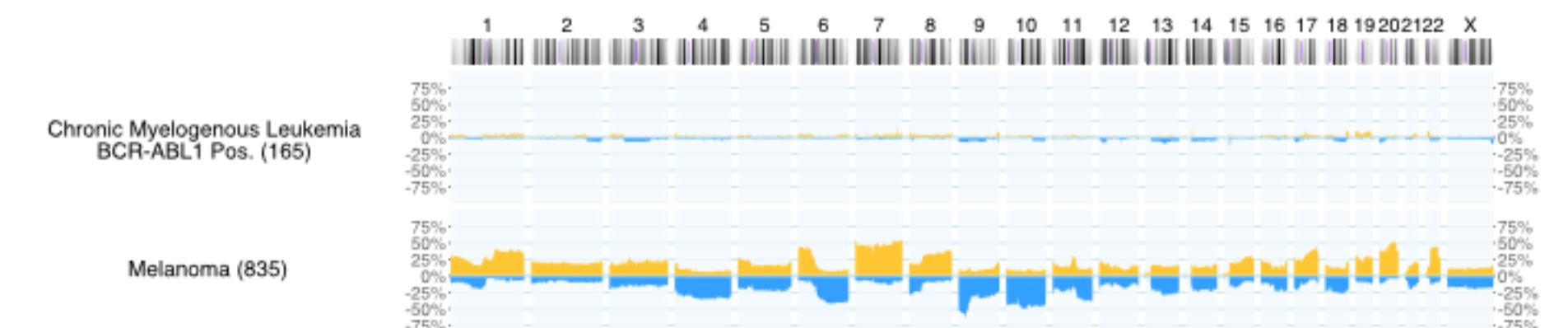


Genome CNV coverage in Cancer Classes

- 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

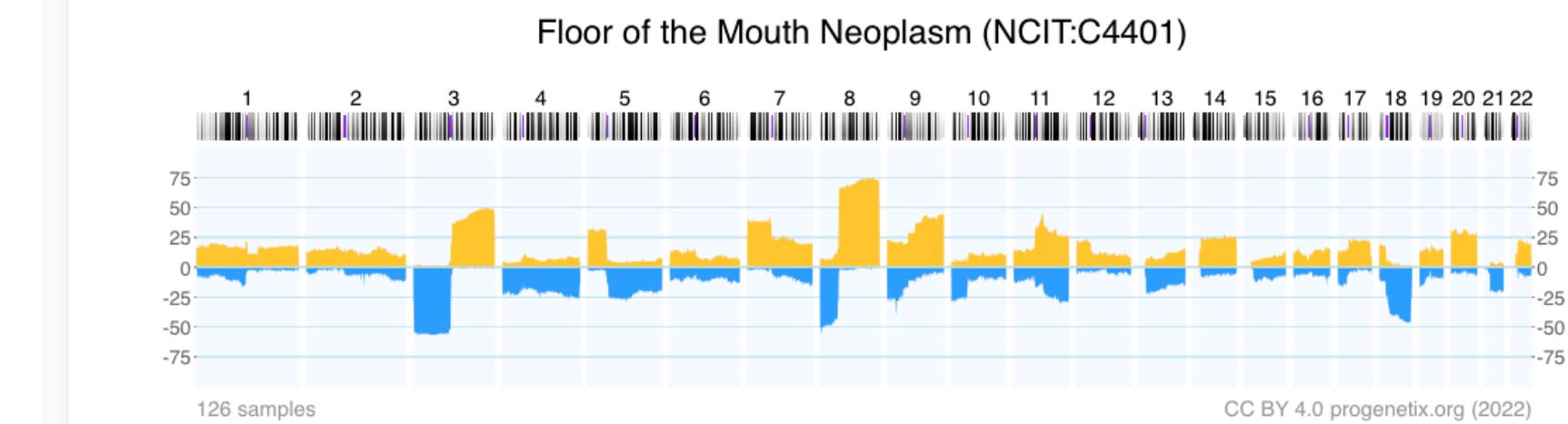
Beacon⁺

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.



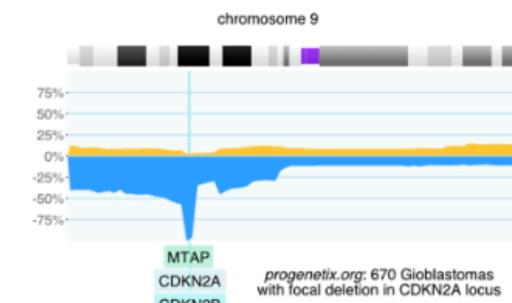
[Download SVG](#) | [Go to NCIT:C4401](#) | [Download CNV Frequencies](#)

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.

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. NCIIt, 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



Swiss Institute of
Bioinformatics

Cancer Types by National Cancer Institute NCIIt Code

The cancer samples in Progenetix are mapped to several classification systems. For each of the classes, aggregated data 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: Oligodendrogloma (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)

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. NCIIt, 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—



Swiss Institute of
Bioinformatics

Edit Query

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) ↗

Visualization options

Results

Biosamples

Biosamples Map

Variants

(progenetix)



© 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 ↗

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

Cancer Cell Lines by Cellosaurus ID

The cancer cell lines in [cancercelllines.org](#) are labeled by their parentage hierarchically: Daughter cell lines are displayed below the primary cell line. For example, HeLa is listed as a daughter cell line of **HeLa (CVCL_0030)** and so forth.

Sample selection follows a hierarchical system in which samples are retrieved based on the selected cell line. For example, selecting "HOS" for HeLa will also return the daughter lines by default - but can be filtered to only return the primary cell line.

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

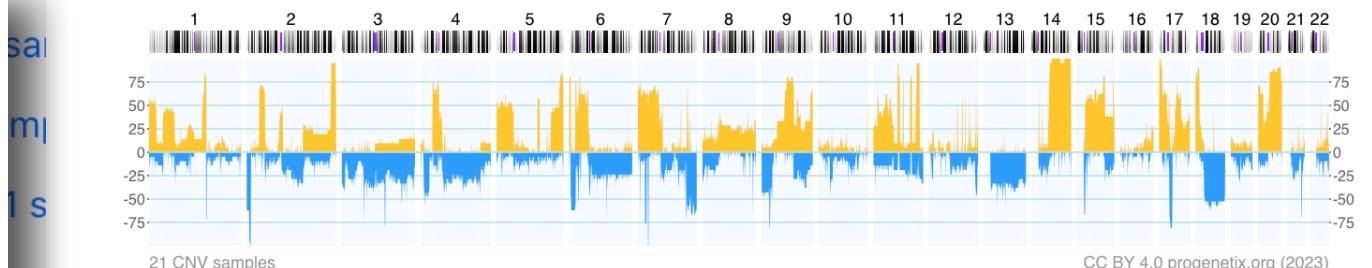
- 204 samples
- 57 direct cellosaurus:CVCL_0312 code matches
- 21 CNV analyses

Search Samples

Select cellosaurus:CVCL_0312 samples in the [Search Form](#)

Raw Data (click to show/hide)

HOS (cellosaurus:CVCL_0312)



[Download SVG](#) | [Go to cellosaurus:CVCL_0312](#) | [Download CNV Frequencies](#)

DATABASE
The Journal of Biological Databases and Curation

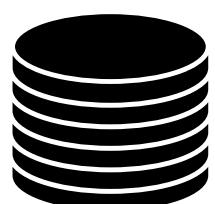
Follow this preprint

	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	

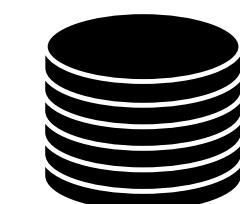
Progenetix Stack



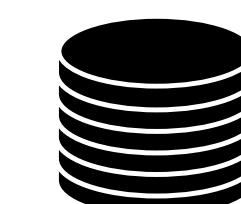
- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
 - ▶ biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads...
- the complete middleware / CGI stack is provided through the **bycon** package
 - ▶ schemas,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



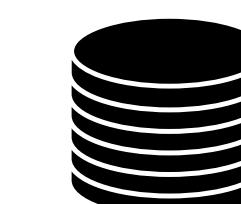
variants



analyses



biosamples



individuals

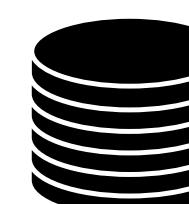


github.com/progenetix/bycon/

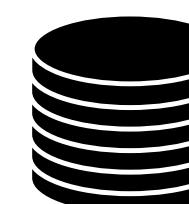


- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
 - ▶ PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703...
- *querybuffer* stores id values of all entities matched by a query and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

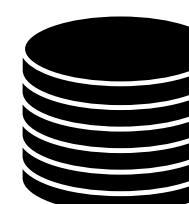
```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
  ObjectId("5bab578d727983b2e00cb505")]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



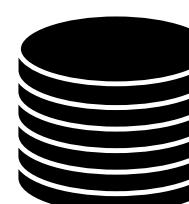
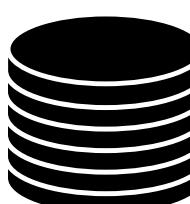
collations



geolocs



genespans publications



qBuffer

Entity collections

Utility collections

{Bio|informatics}Science}

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

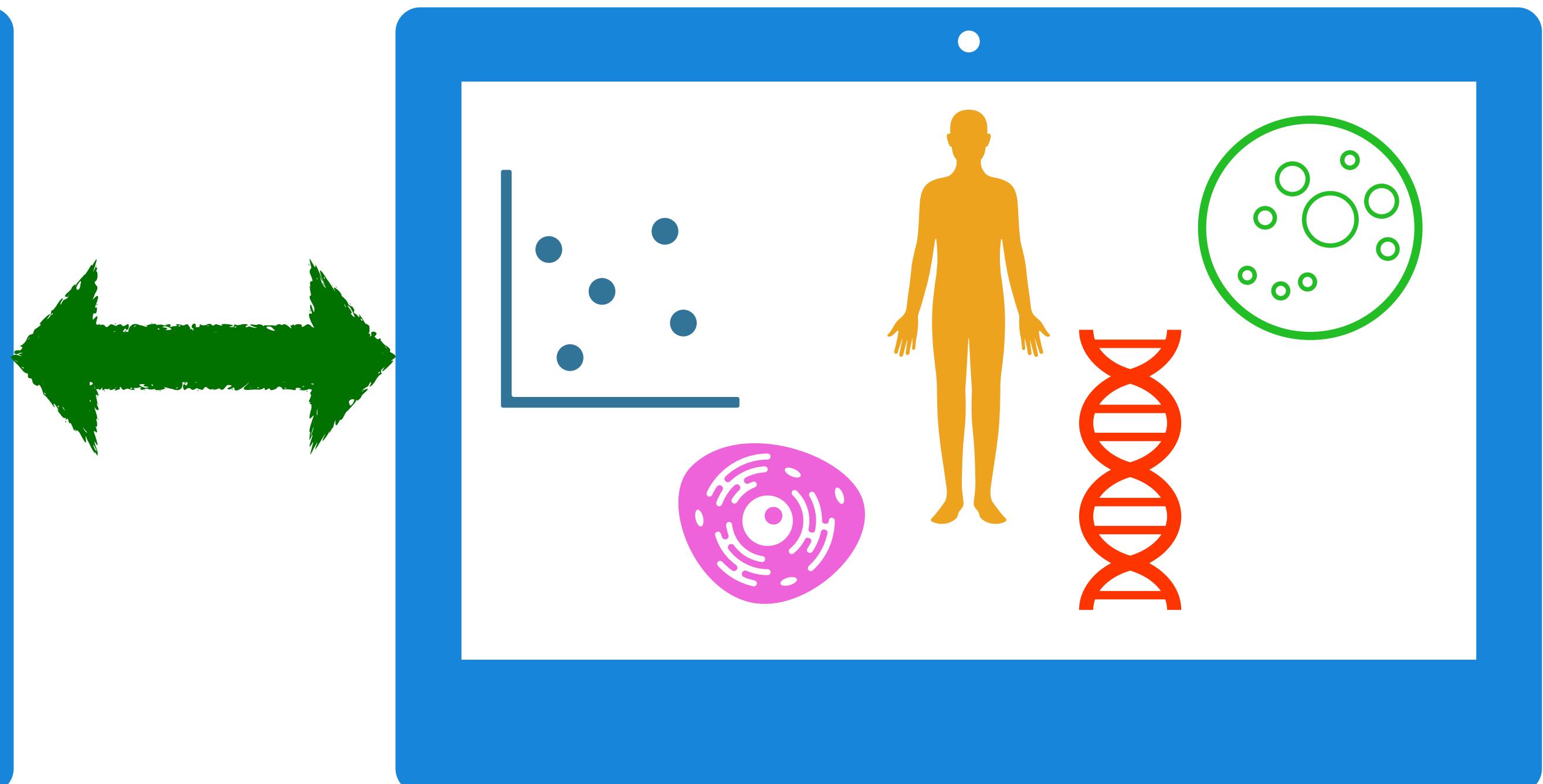
    covs = np.zeros((cs_no, int_no))
    vals = np.zeros((cs_no, int_no))

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

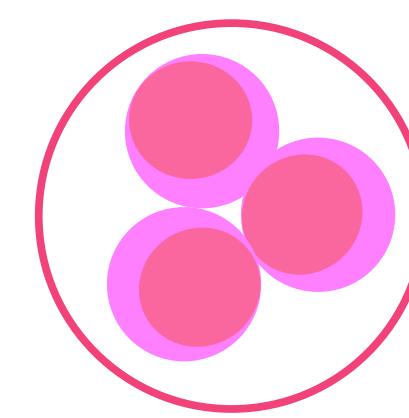
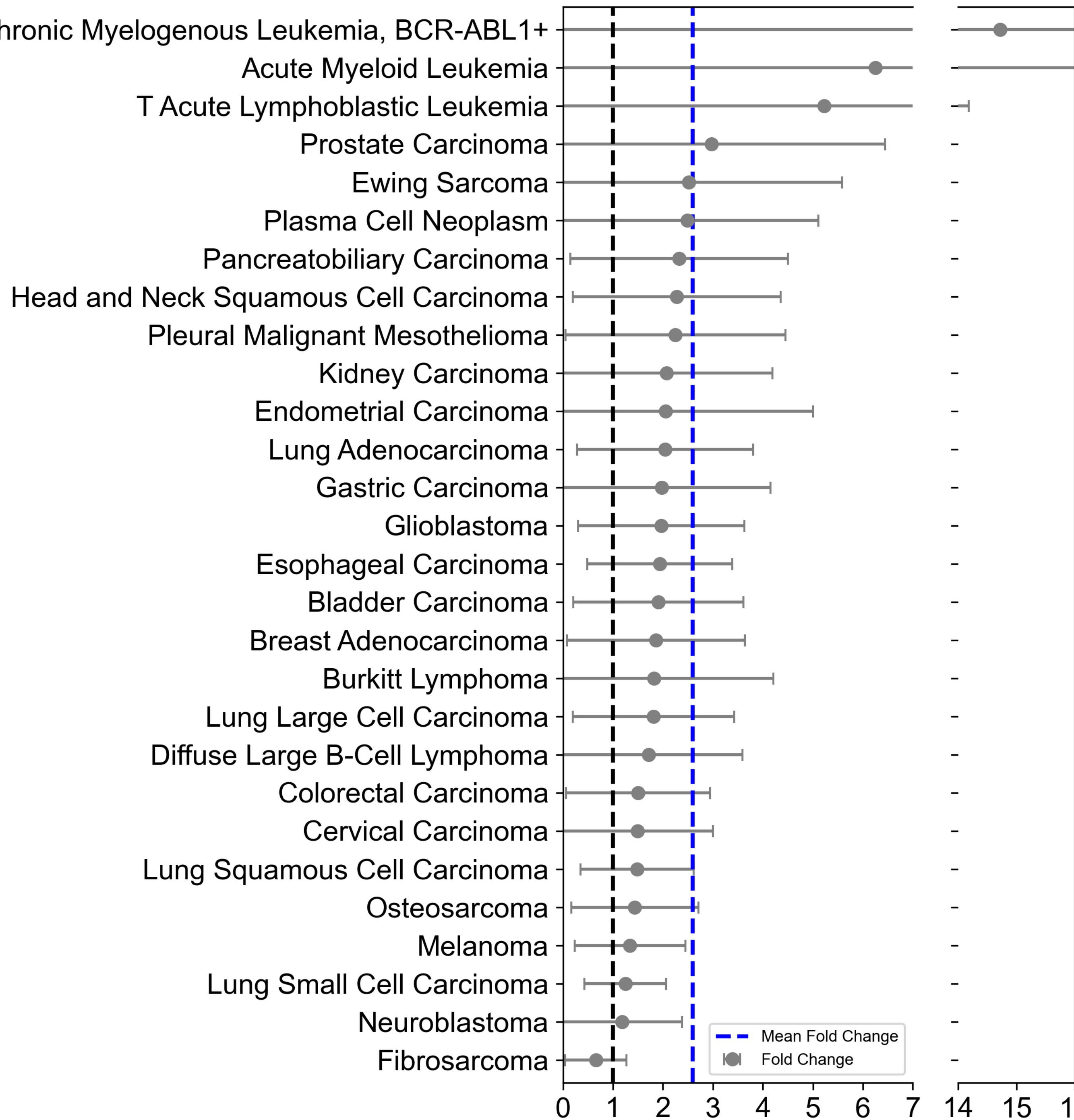
    for i, cs in enumerate(callsets):
        covs[i] = cs["cnv_statusmaps"][pars[t]["cov_l"]]
        vals[i] = cs["cnv_statusmaps"][pars[t]["val_l"]]

    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)

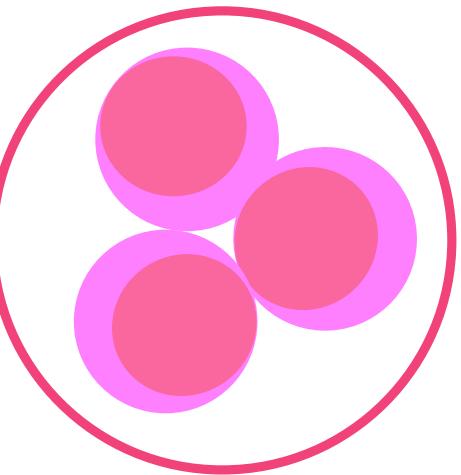
    for i, interval in enumerate(int_fs):
        int_fs[i].update({
            t + "_frequency": frequencies[i],
            t + "_median": medians[i],
            t + "_mean": means[i]
        })
```



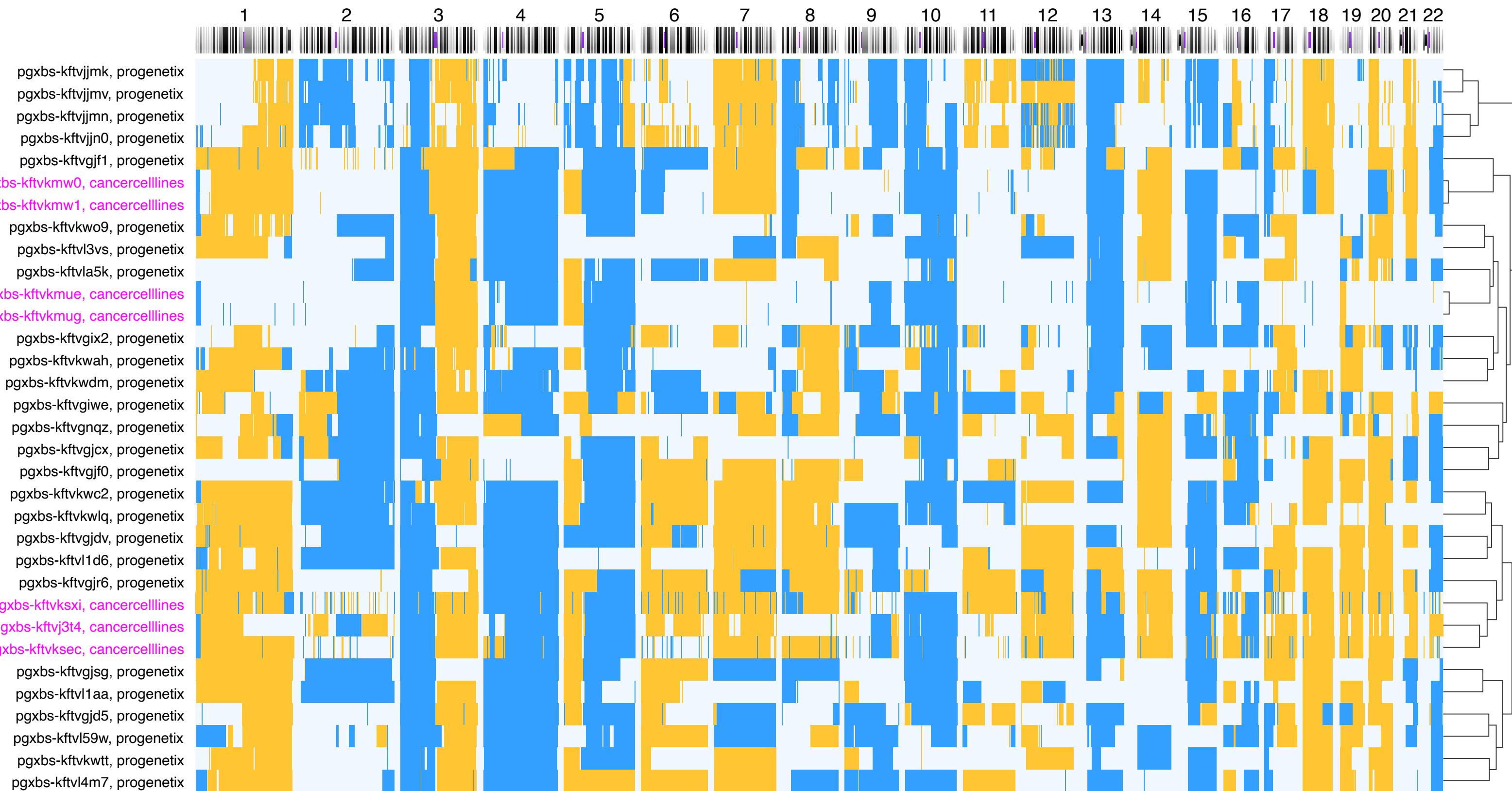
Higher level of CNV coverage of the genomes of cancer cell lines compared to their origins



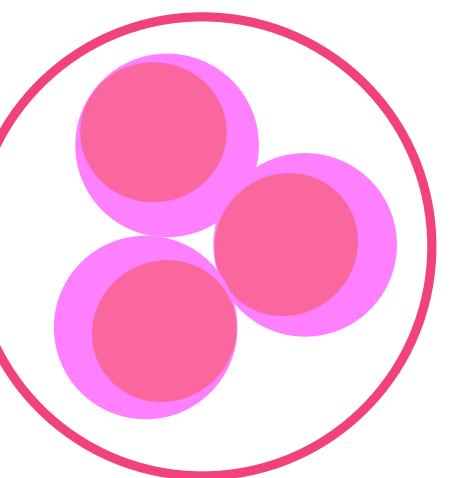
Tumor subpopulations can be matched with highly similar cell lines



- Lung Small Cell Carcinoma Subpopulation
- Cell Lines:
 - CVCL_1140: COR-L279
 - CVCL_1455: NCI-H1105
 - CVCL_1527: NCI-H2107



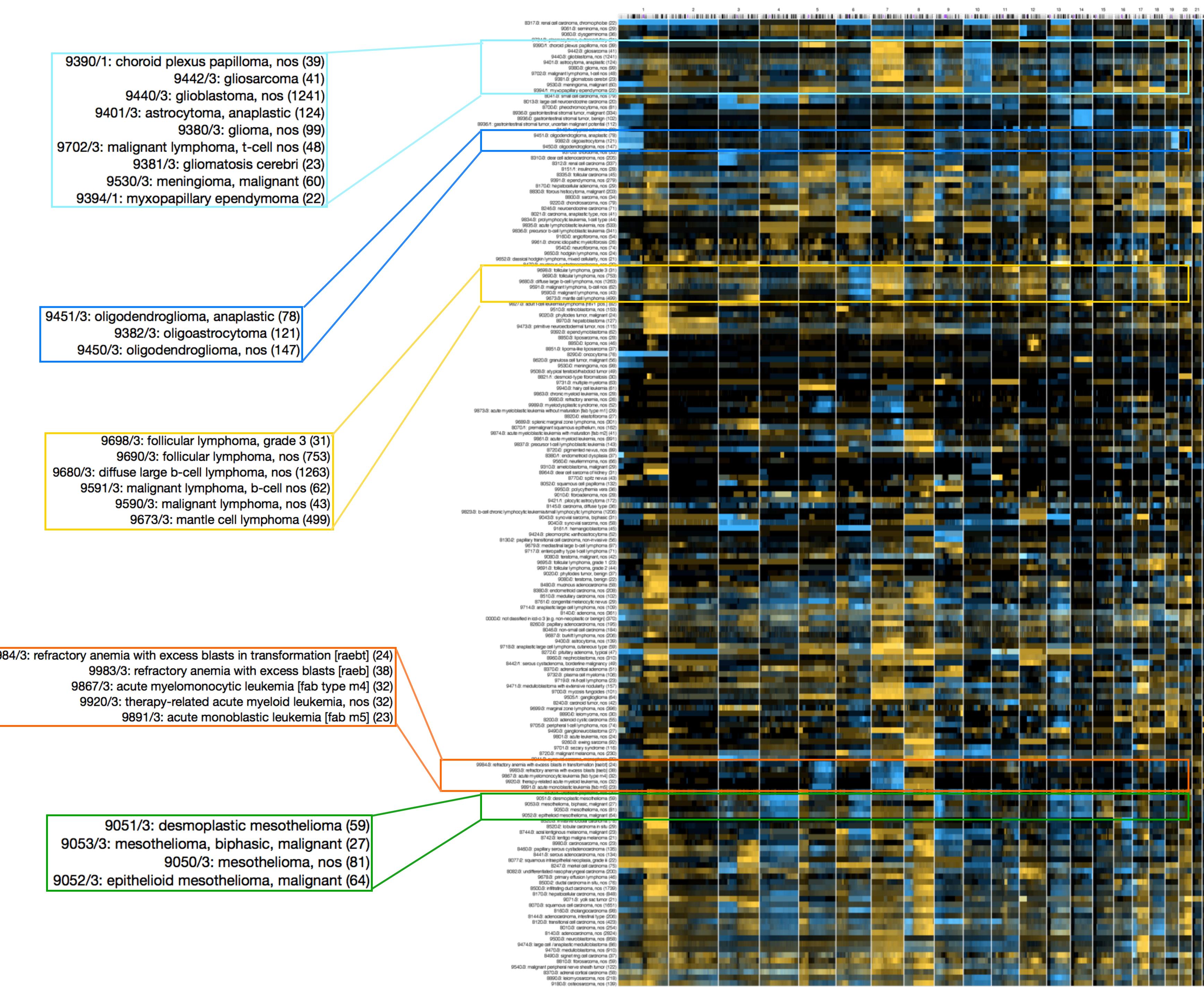
Tumor subpopulations can be matched with highly similar cell lines?!



Somatic Mutations In Cancer: Patterns

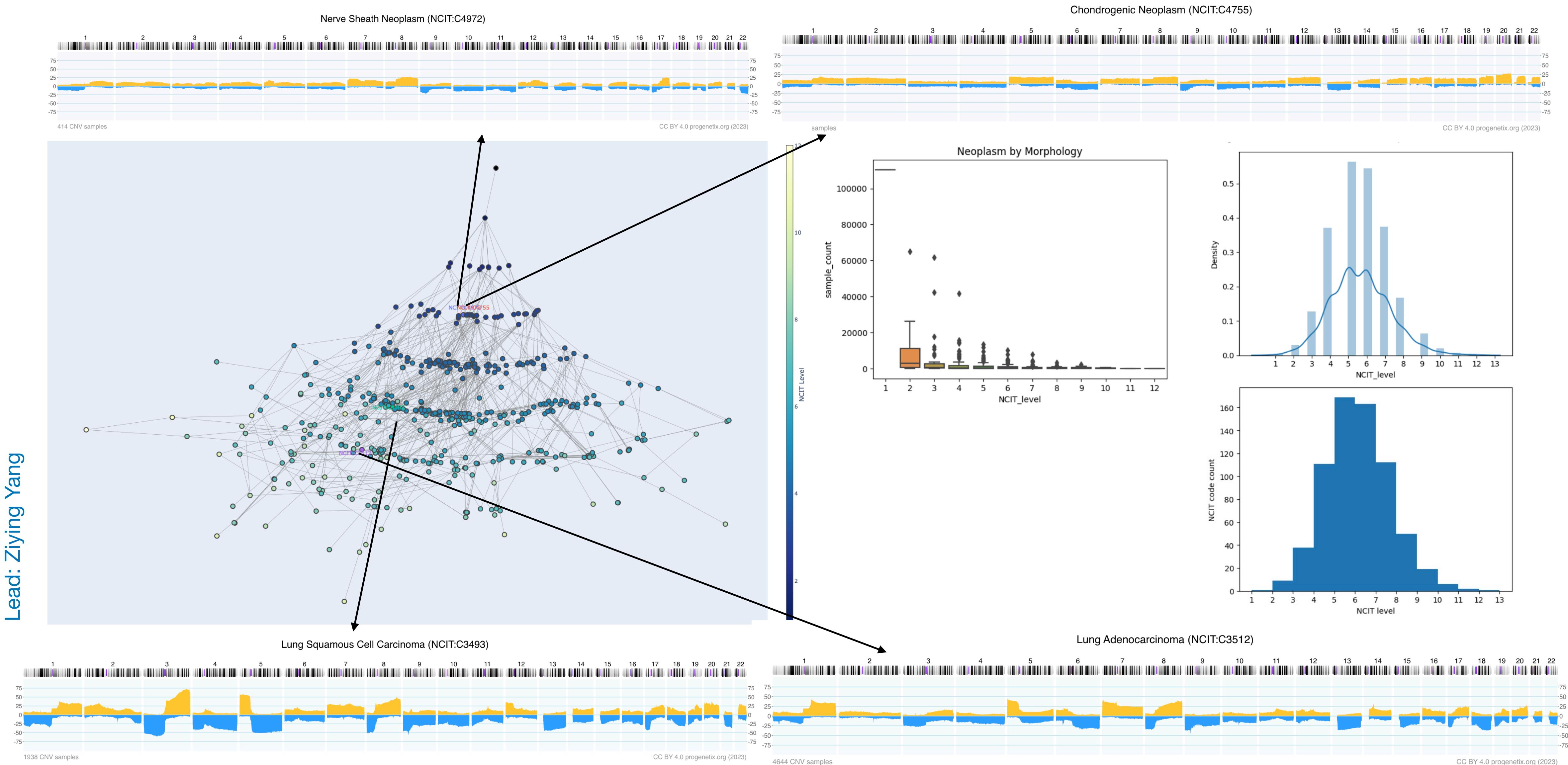
Making the case for genomic classifications

Some related cancer entities show similar copy number profiles



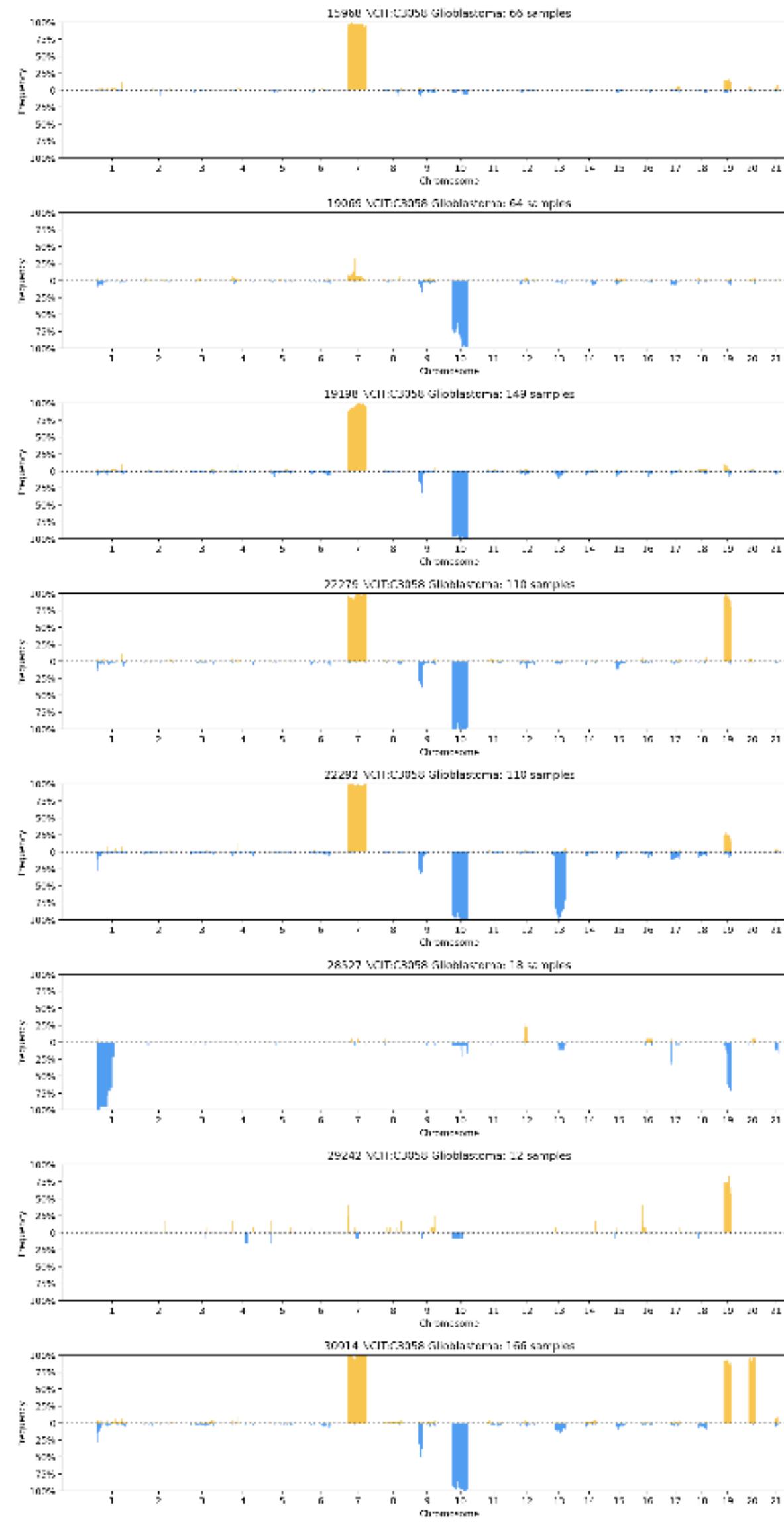
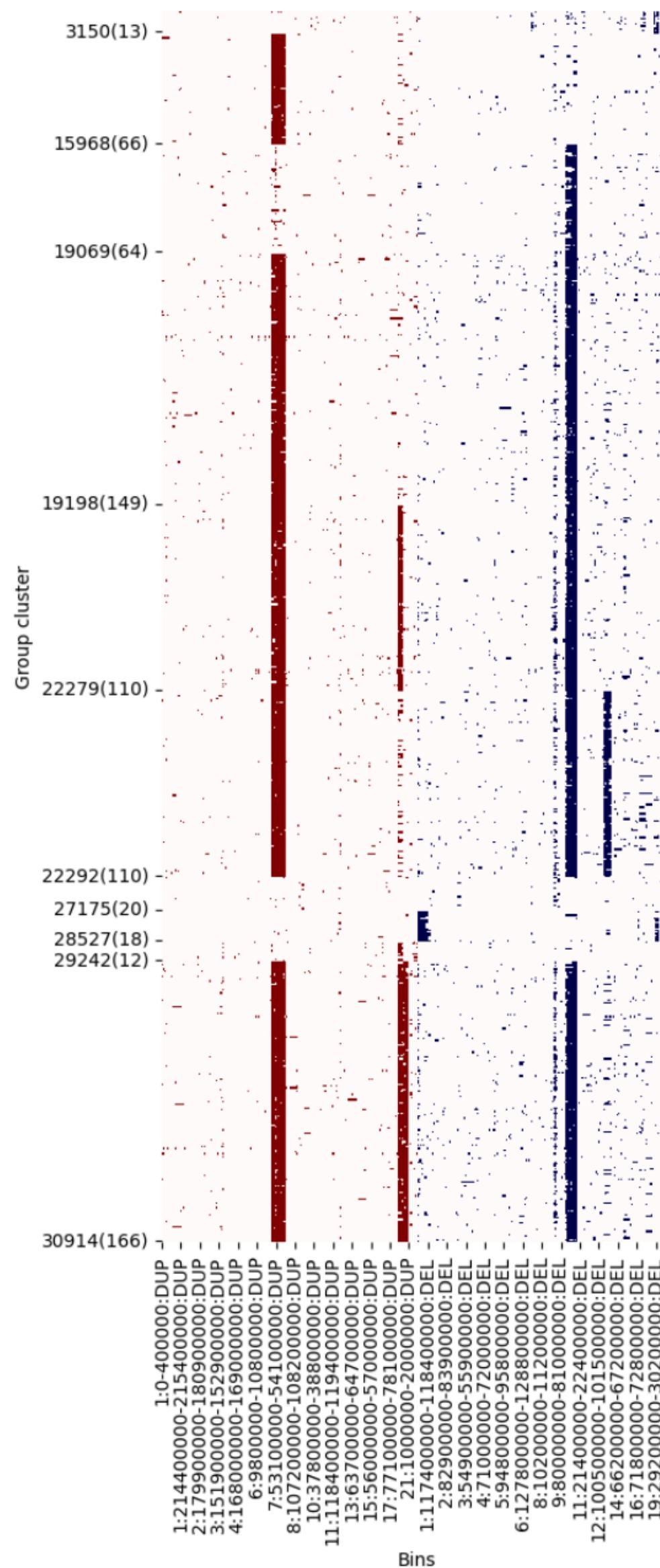
CNV profiles heterogeneity vs cancer classification

Correspondance of genomic profiles to NCIT cancer hierarchy



Results

Entity CNV heterogeneity: Glioblastoma

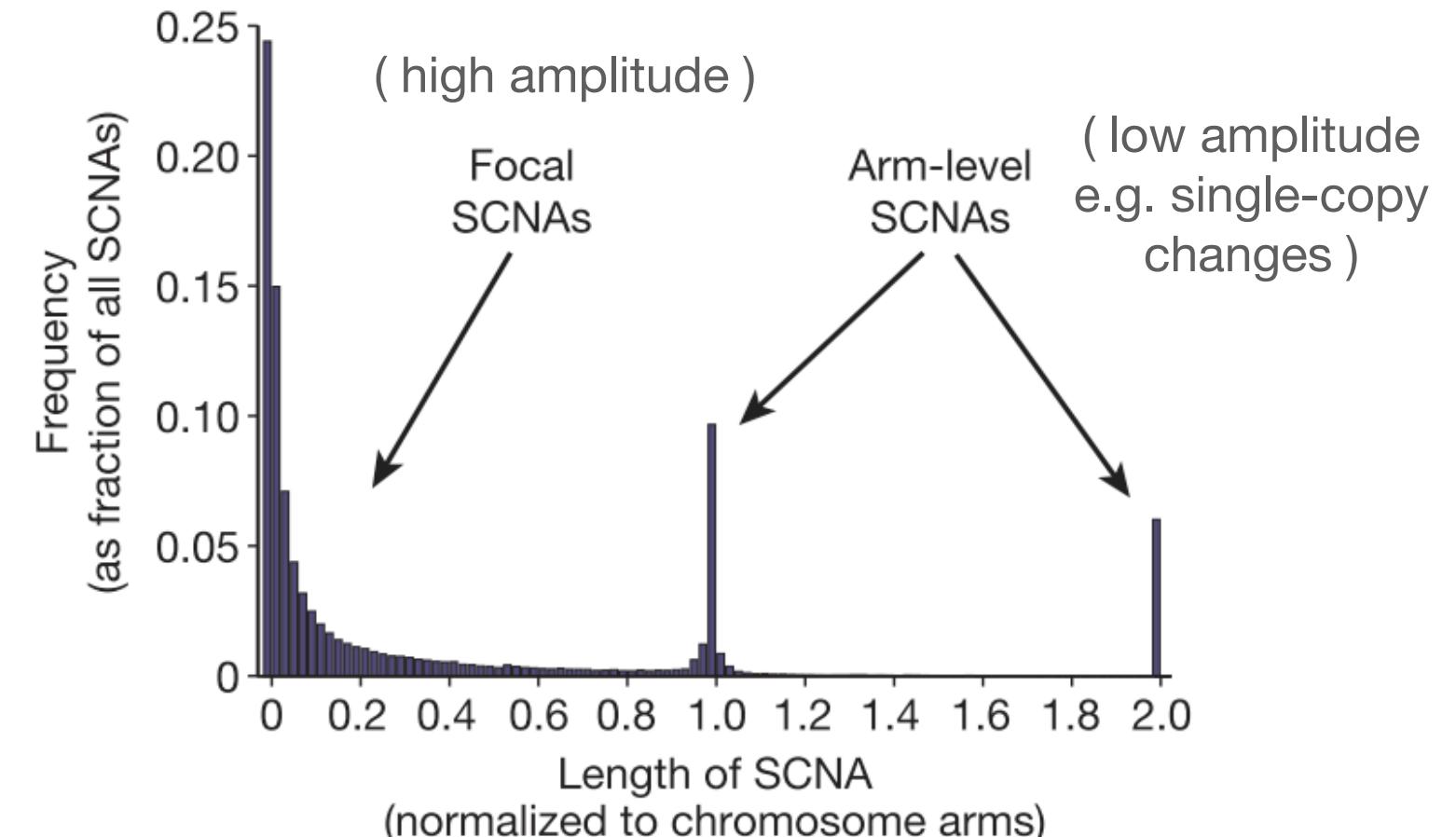


group cluster	CNV features
15968	Dup 7
19069	Del 10
19198	Dup 7, Del 10
22279	Dup 7, Del 10, Dup 19
22292	Dup 7, Del 10, Del 13
27175	Del 1p, Del 19q
28527	Del 1p, Del 19q
29242	Dup 19
30914	Dup 7, Del 10, Dup 19, Dup 20

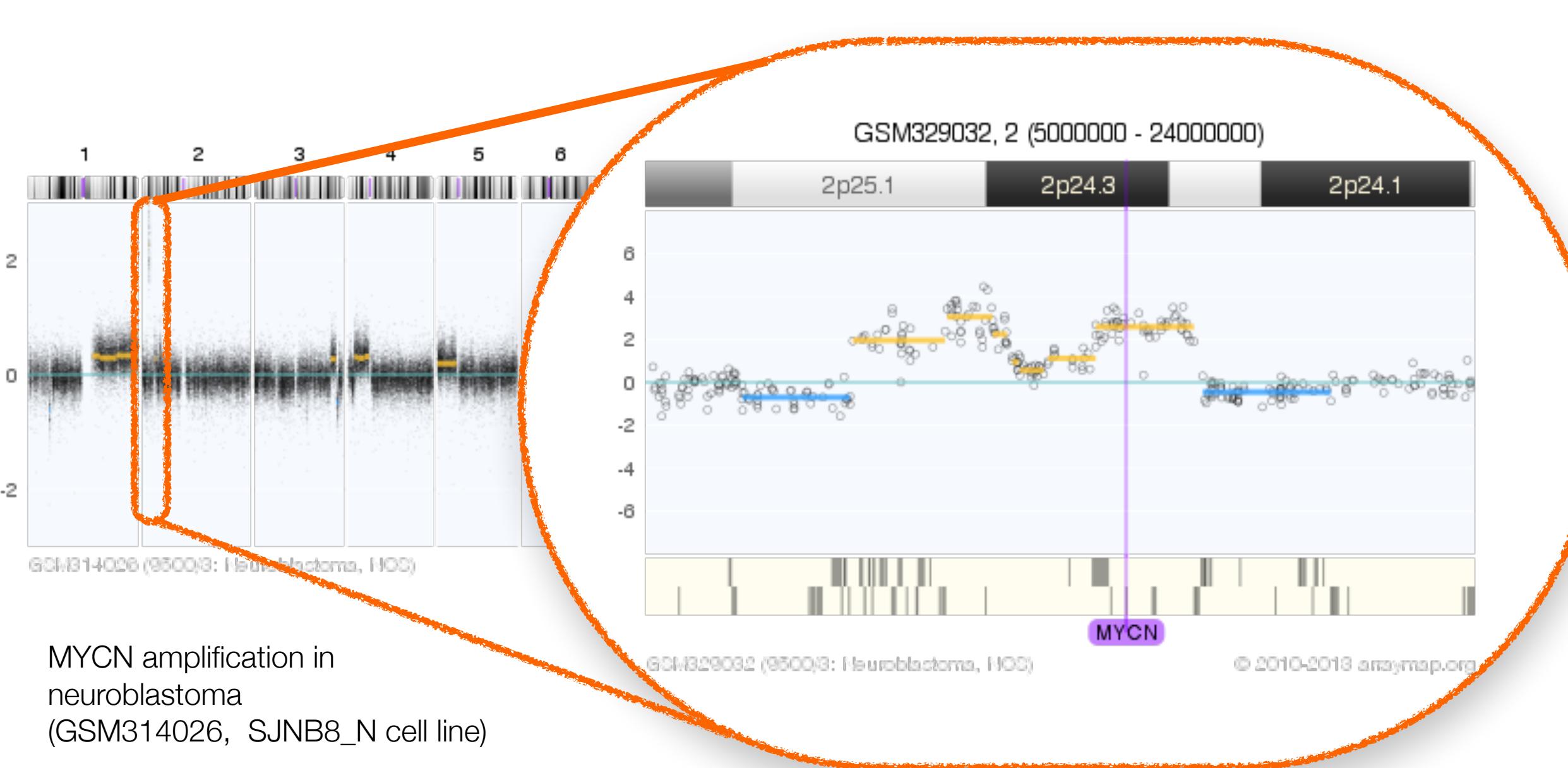


CNV Categorization

different levels of CNV



Rameen et al 2010 Nature



CopyNumberChange

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

Computational Definition

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

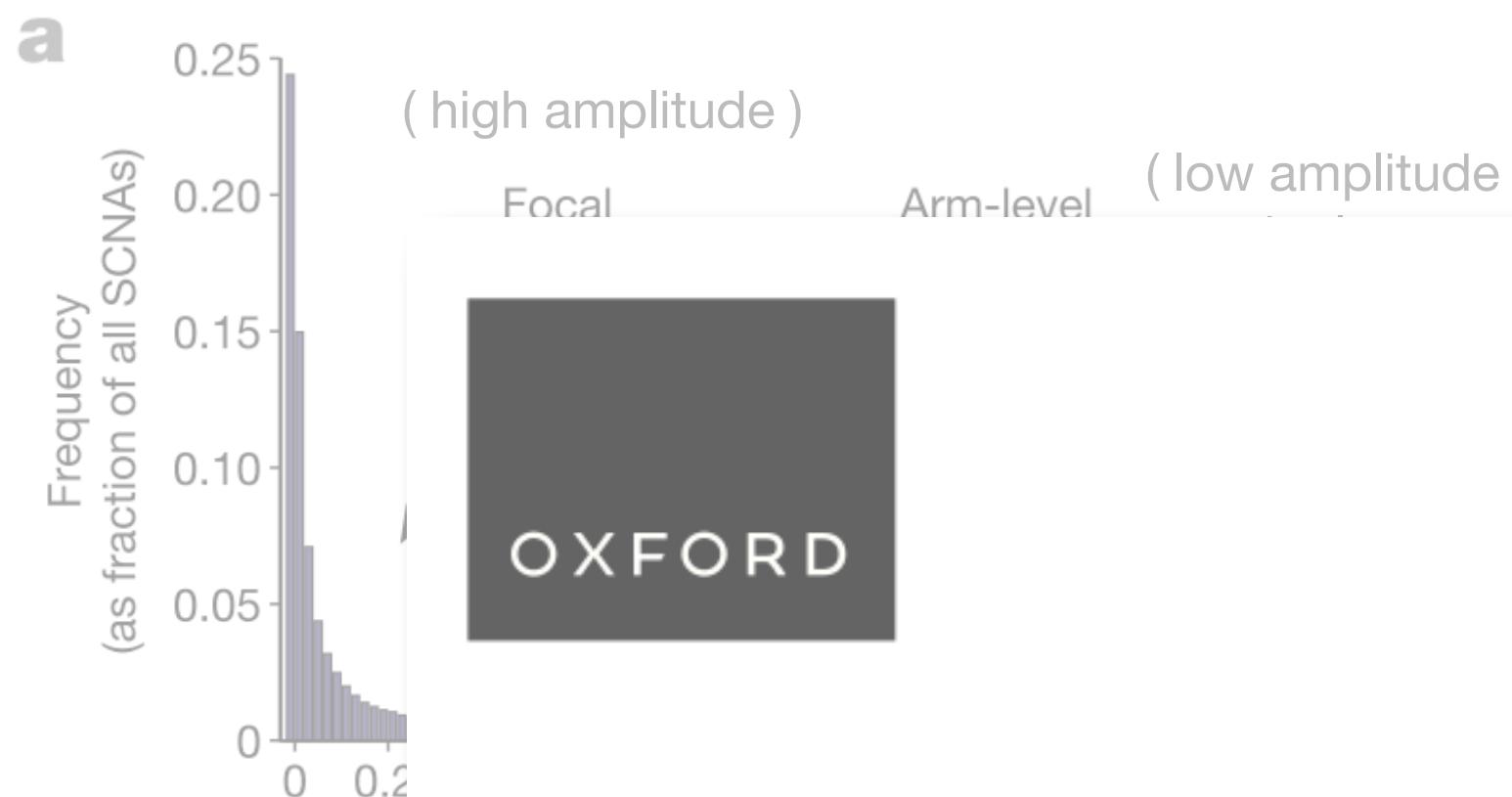
Information Model

Some CopyNumberChange attributes are inherited from **Variation**.

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

CNV 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

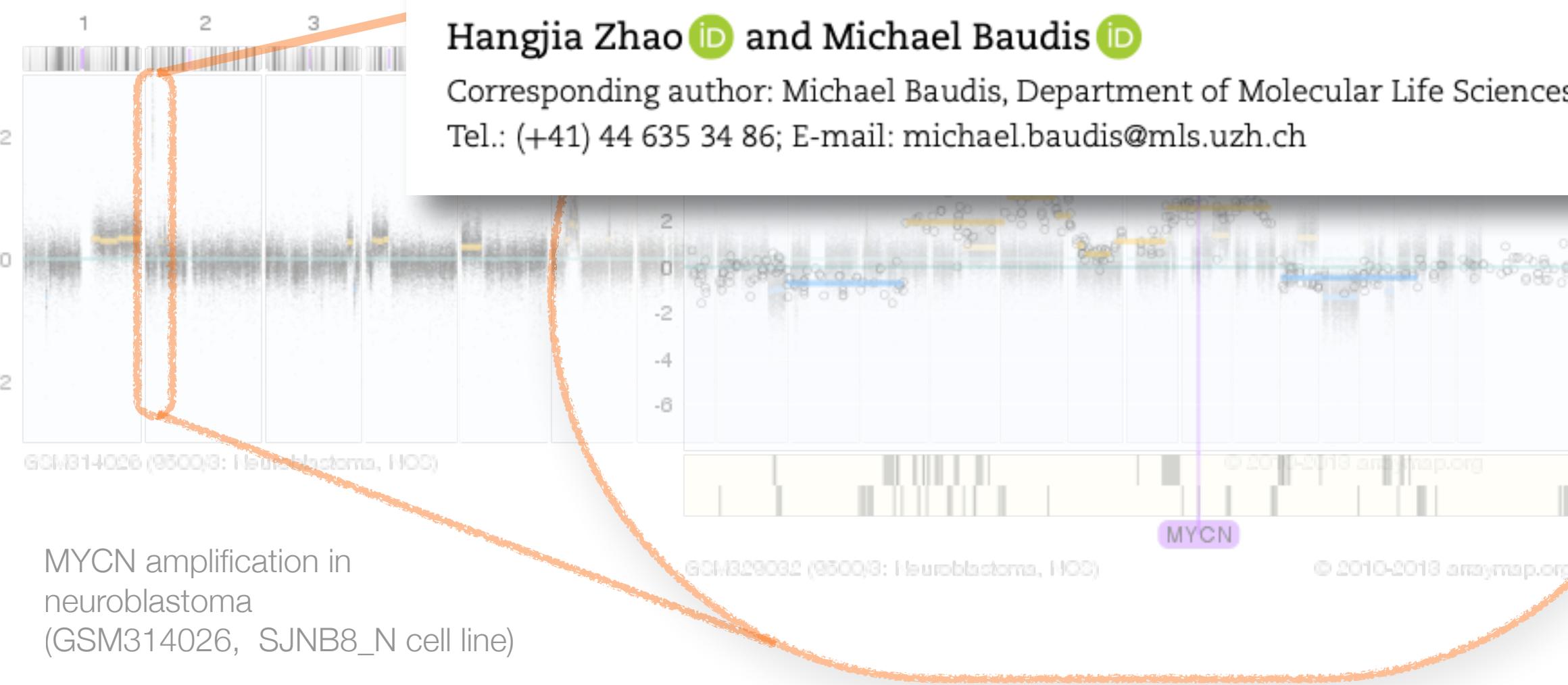
a system (e.g. genome, cell,

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

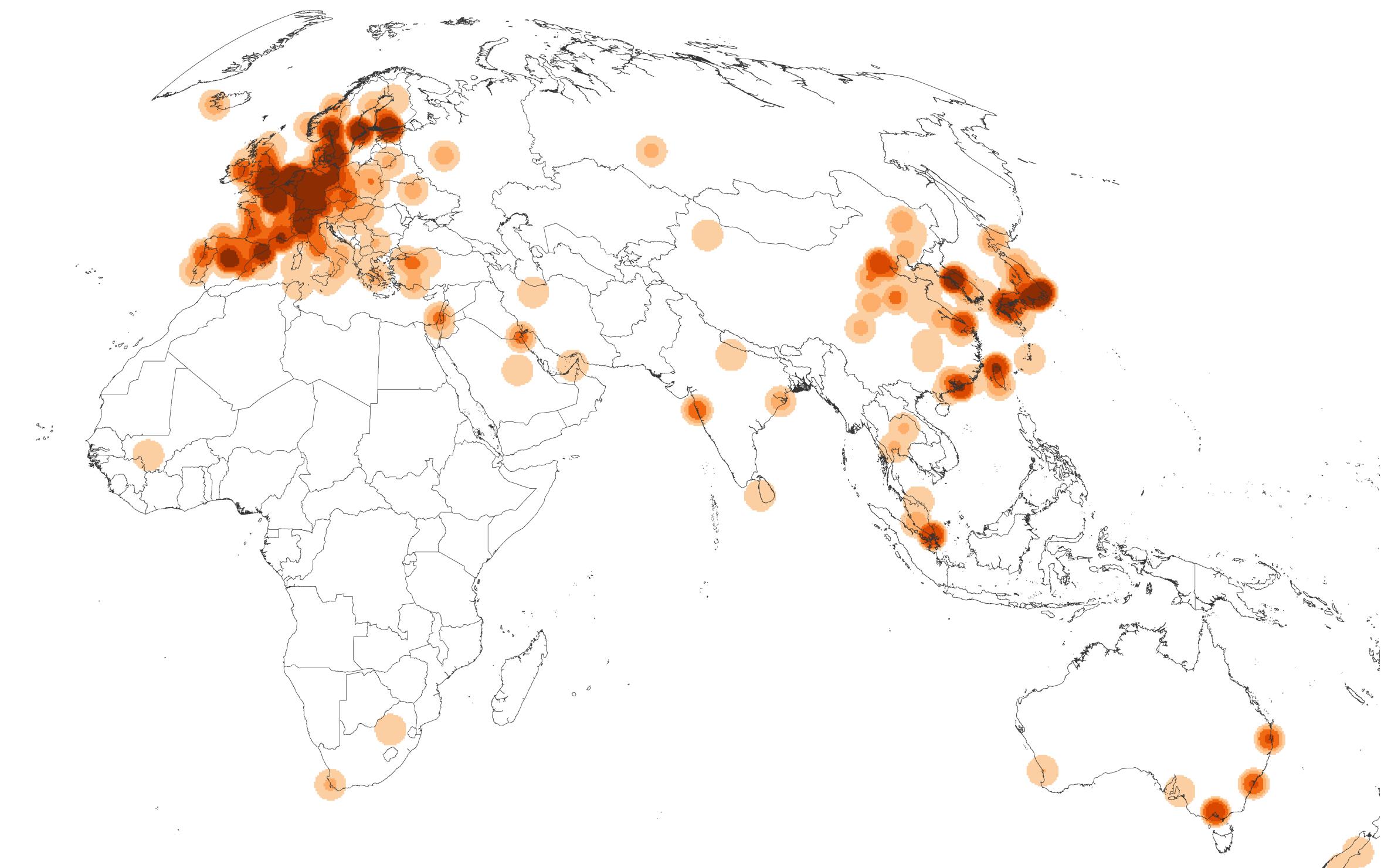
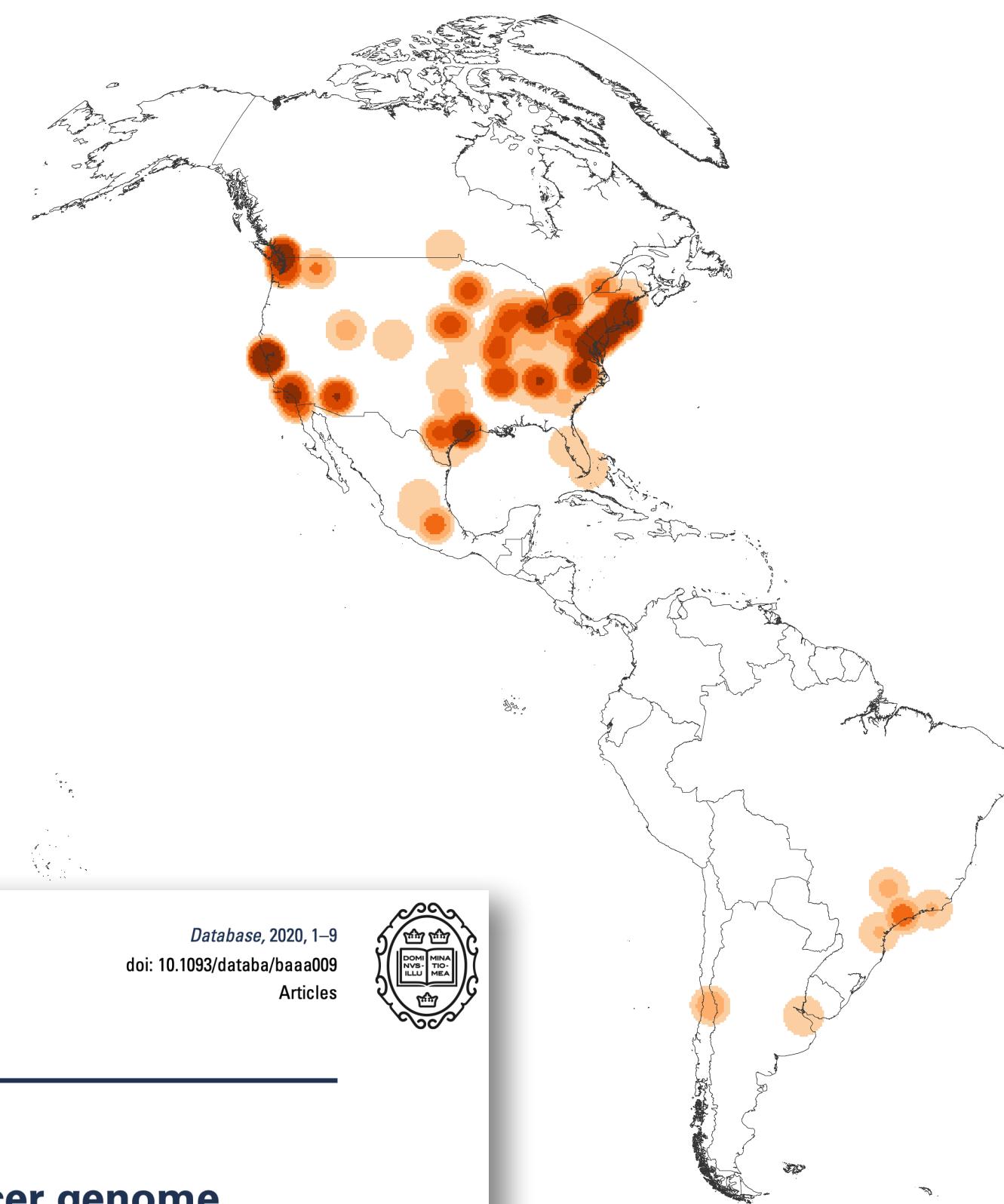
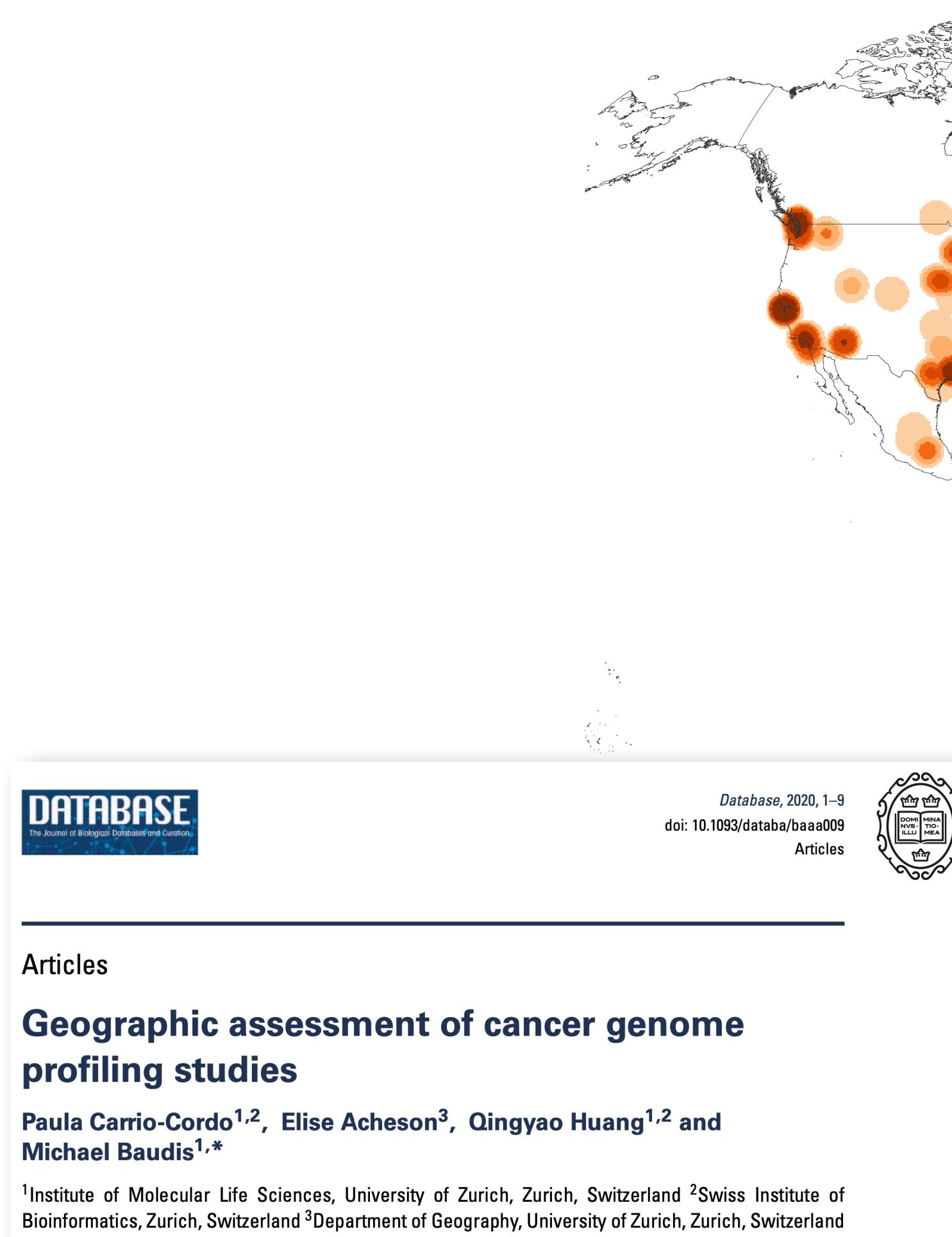
Tel.: (+41) 44 635 34 86; E-mail: michael.baudis@mls.uzh.ch



_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

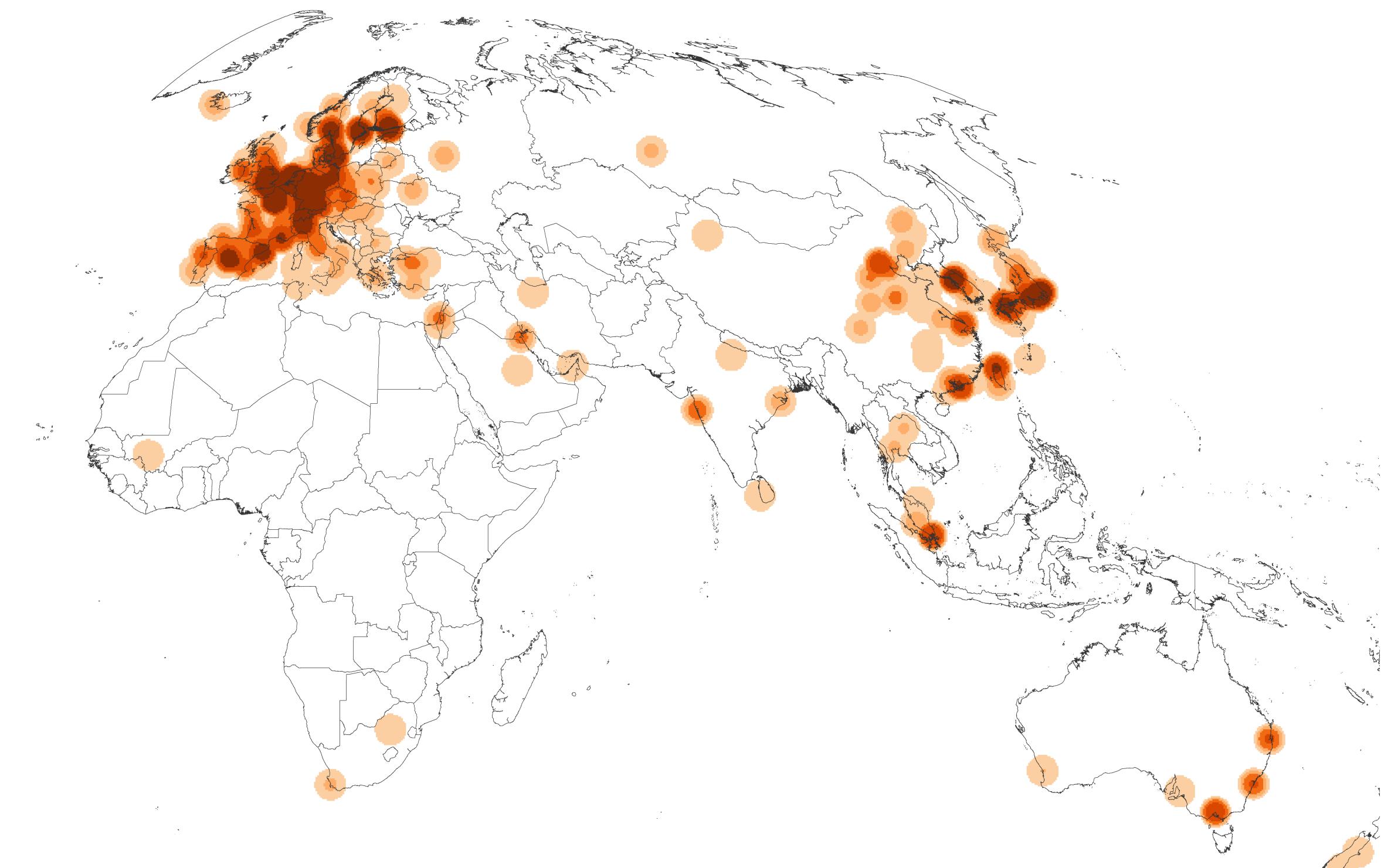
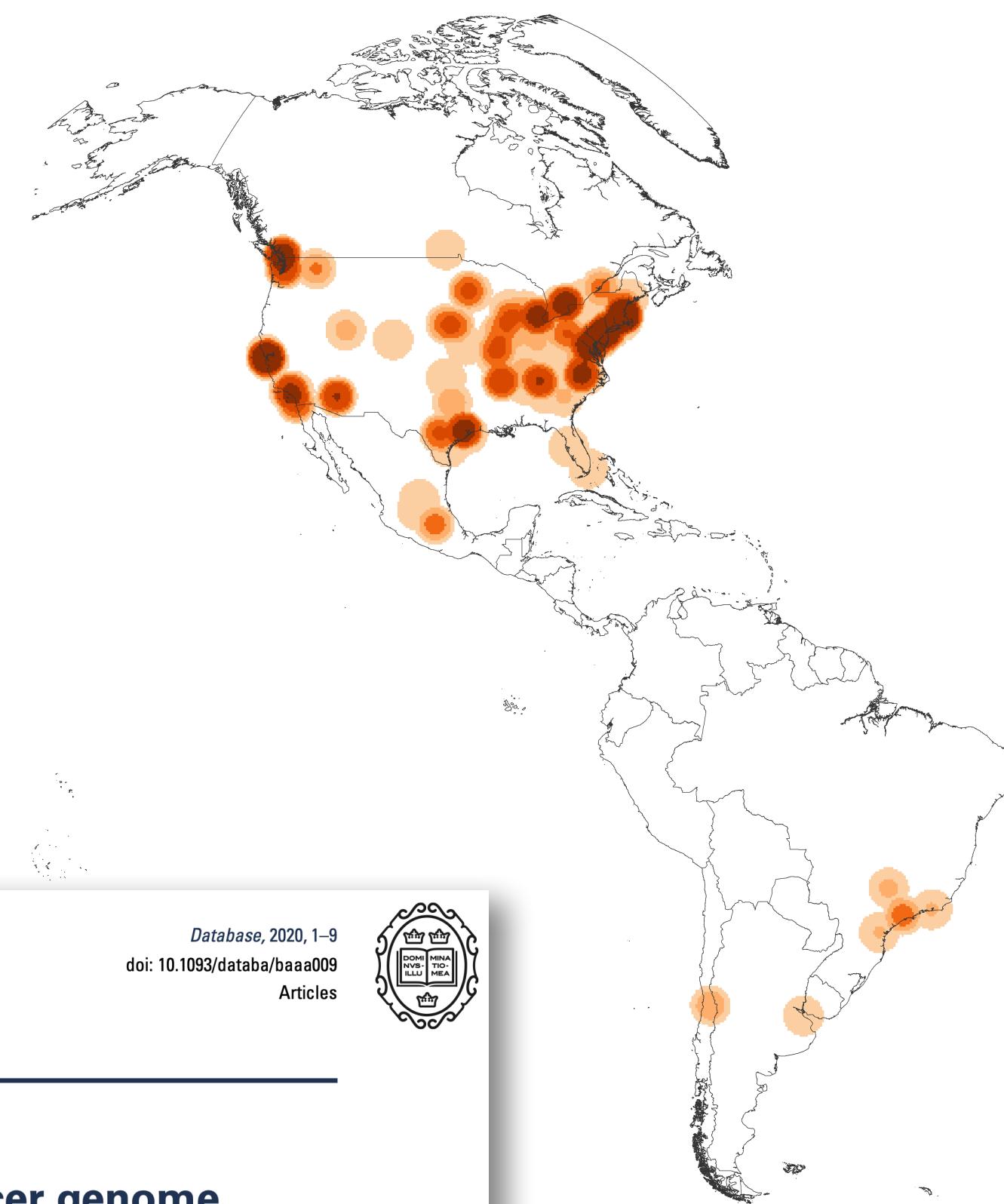
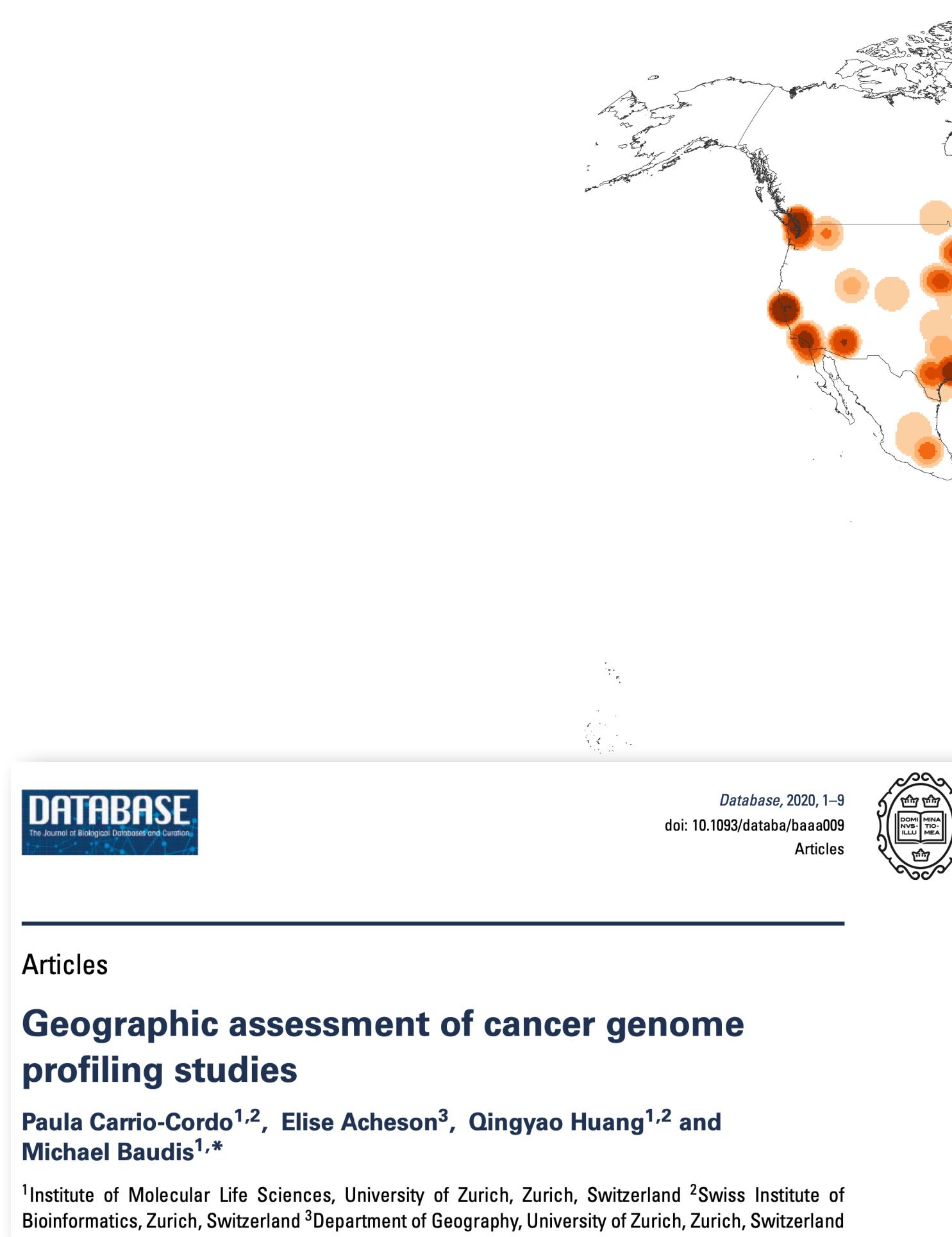
Standards for CNV annotation and their use in data discovery protocols

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.

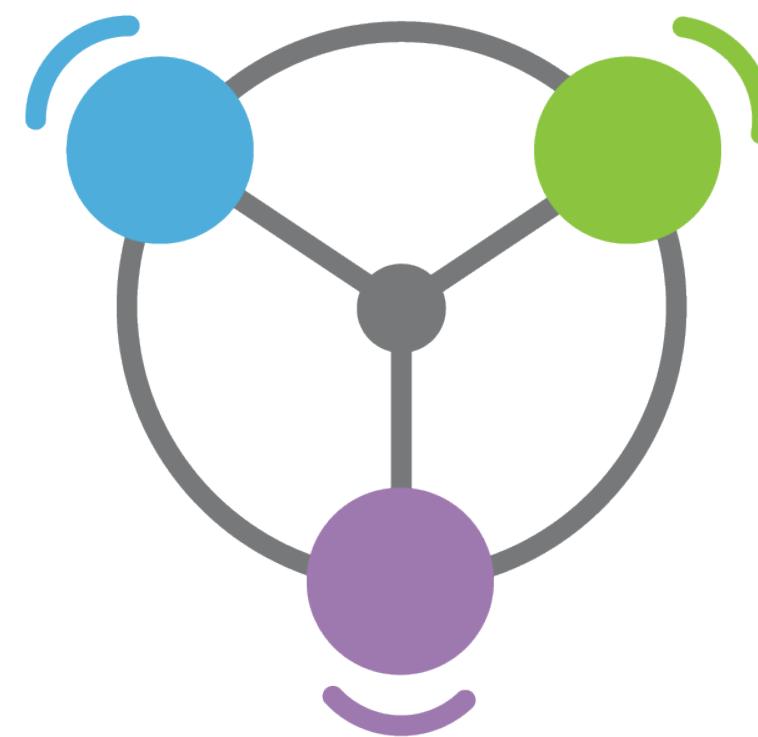
Different Approaches to Data Sharing



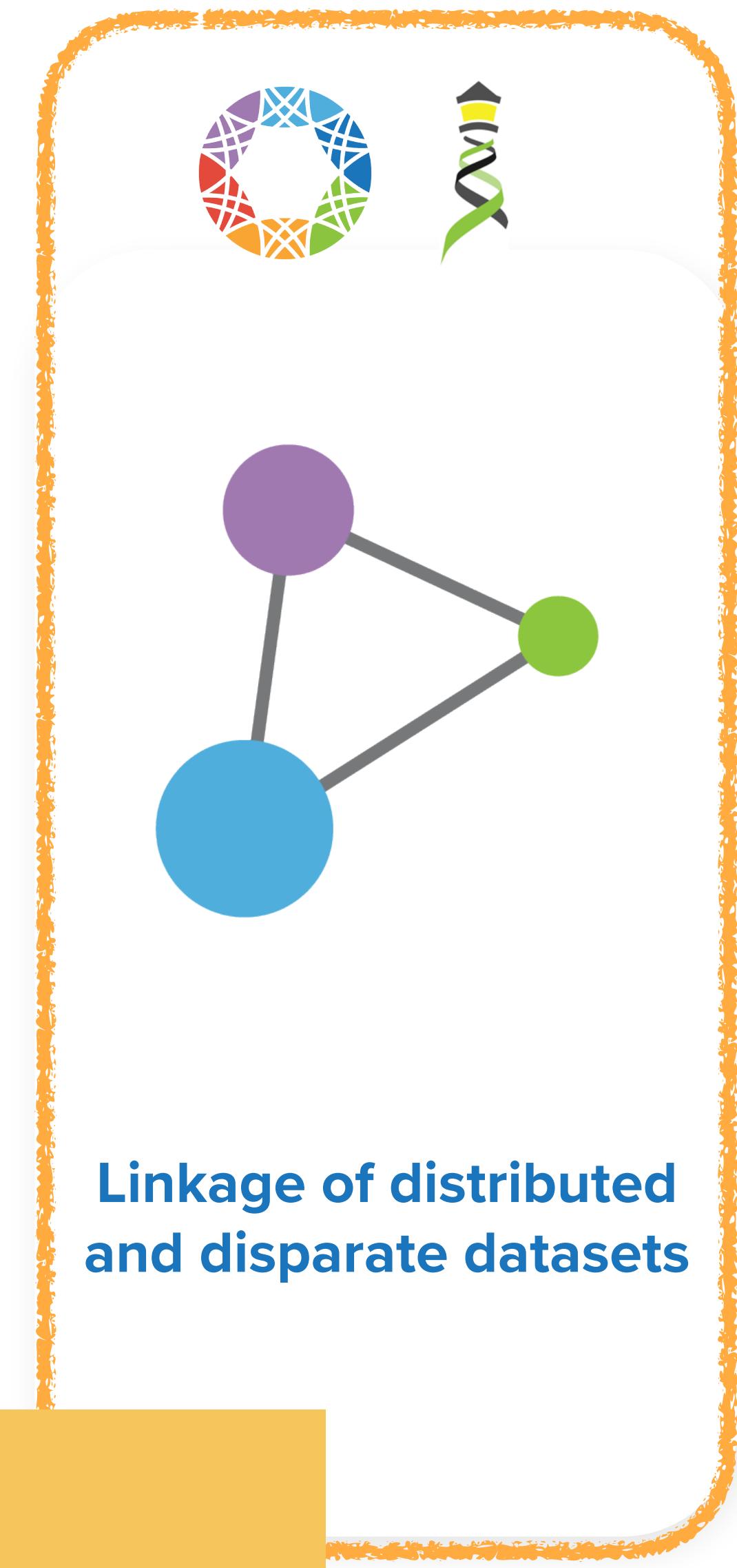
Centralized Genomic Knowledge Bases



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

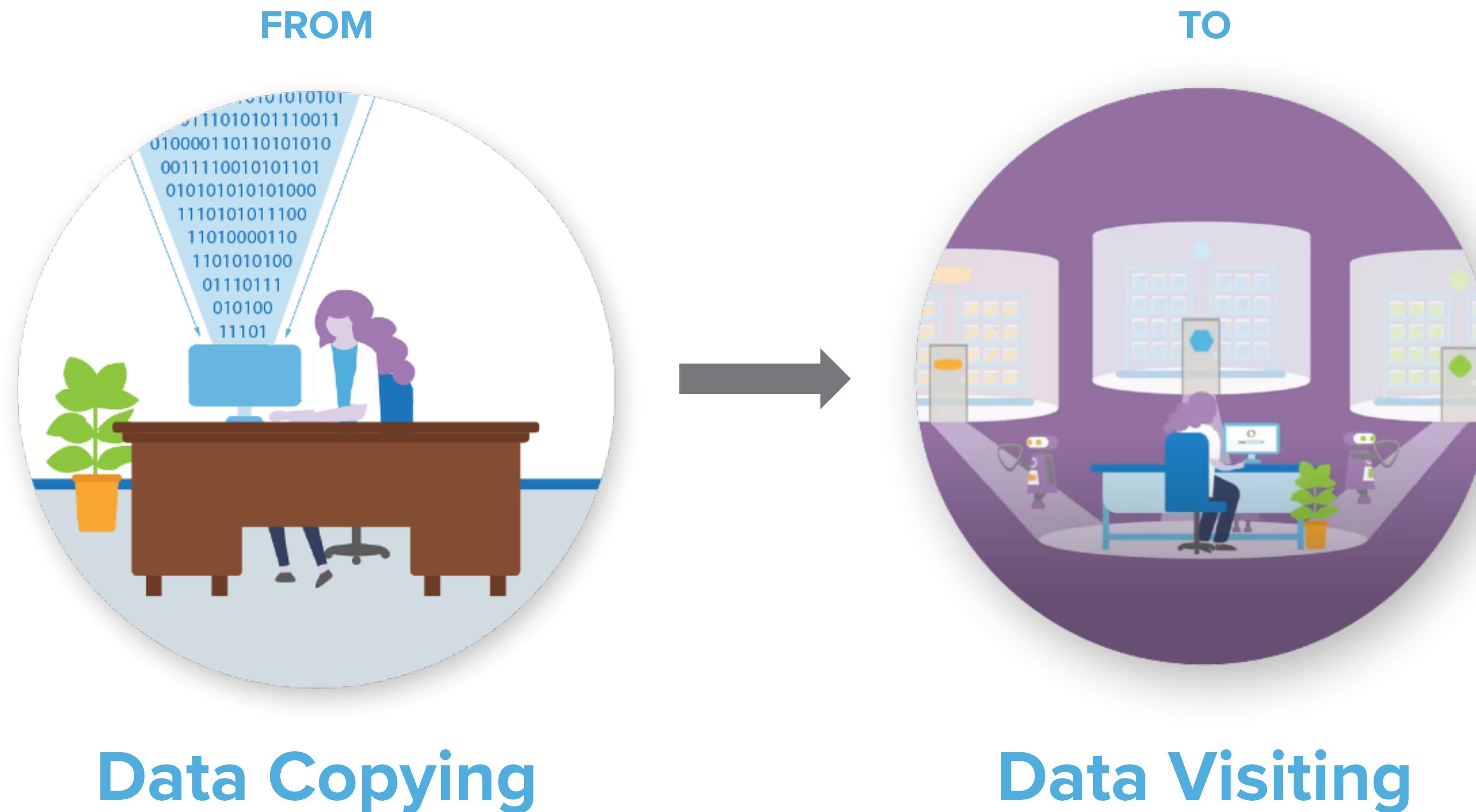
Beacon v2

Federated Genomics



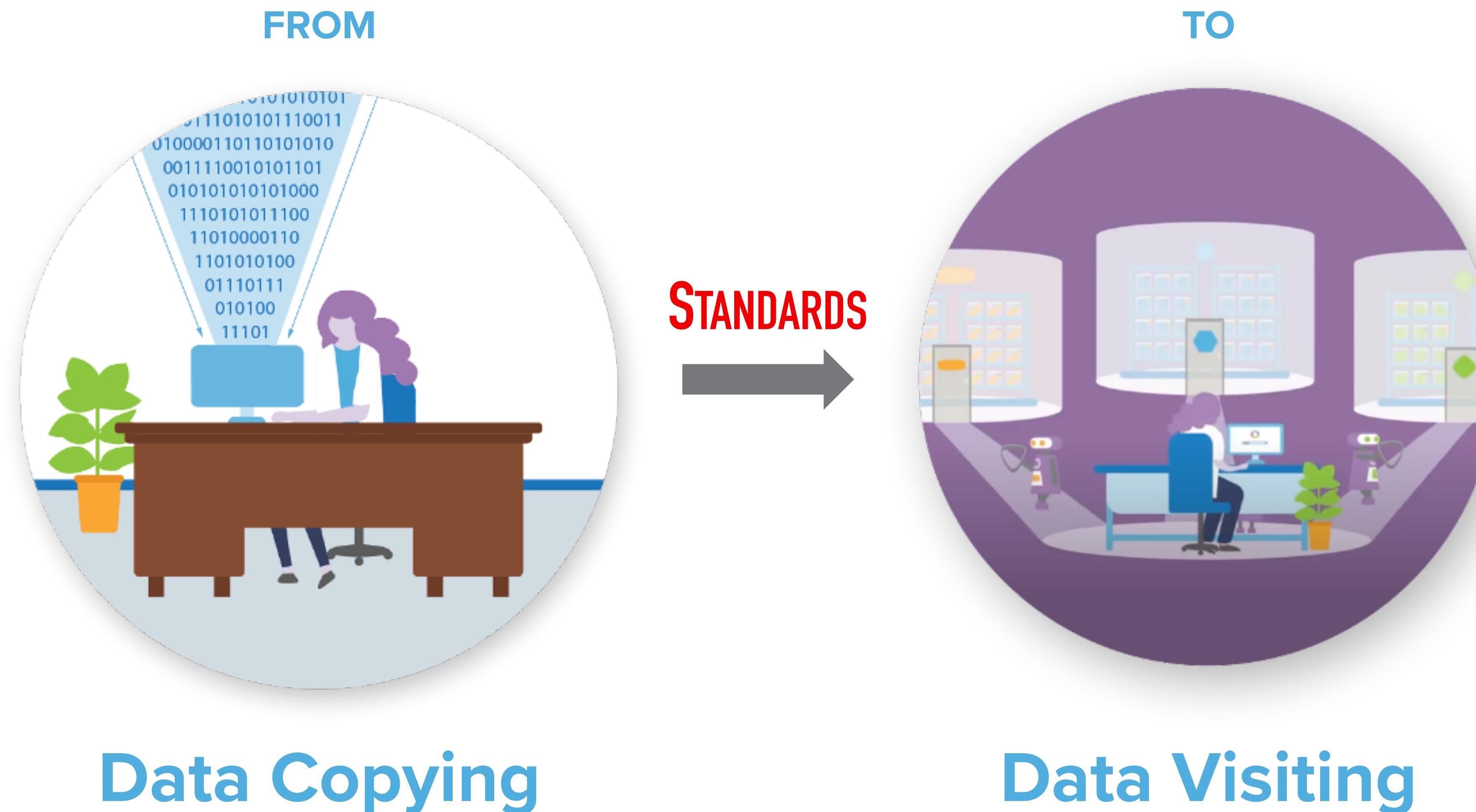


A New Paradigm for Data Sharing





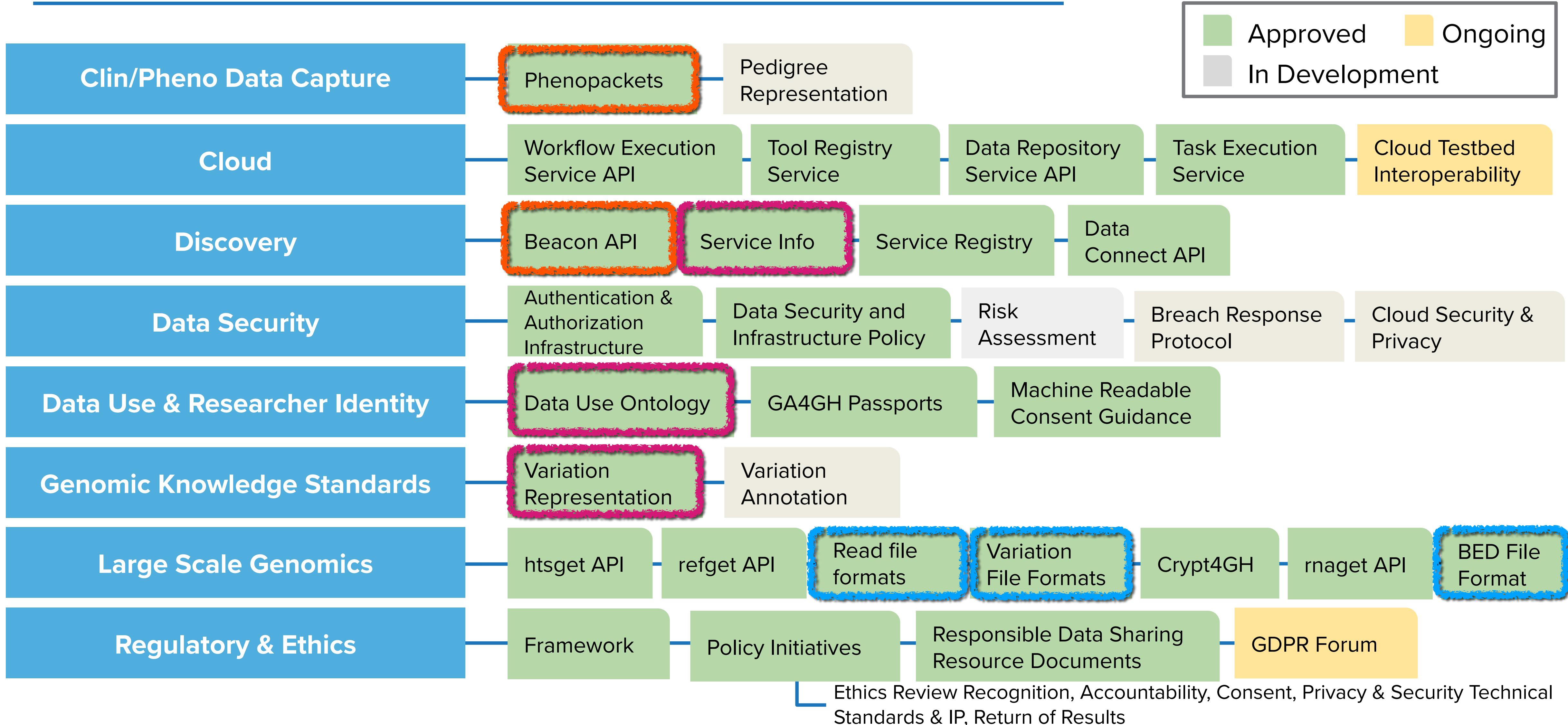
A New Paradigm for Data Sharing



GA4GH 2020-2022 Strategic Roadmap



Global Alliance
for Genomics & Health

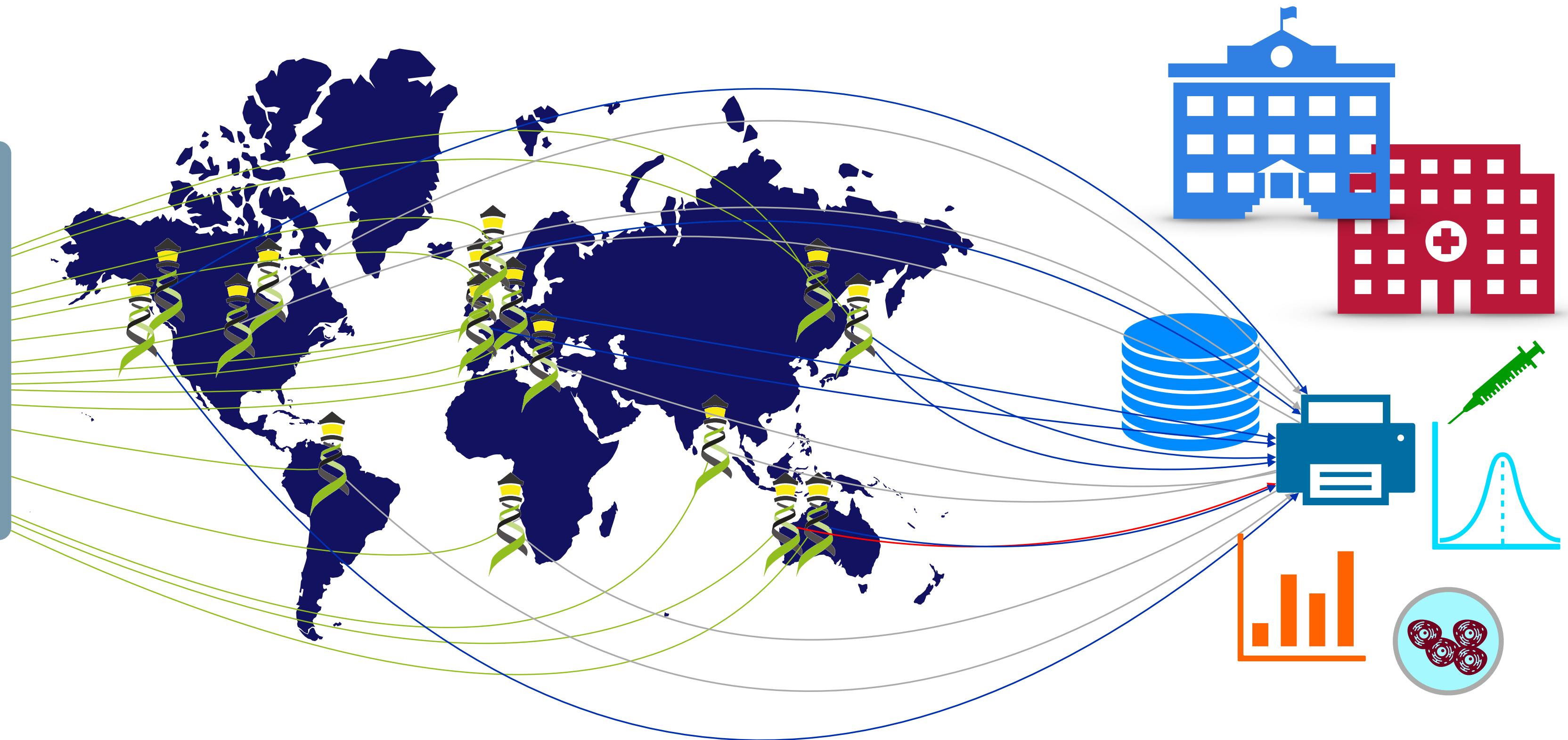




The basic Beacon v2 query asks for the replacement of a sequence by a different one of equal or different length.

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

Beacon Queries

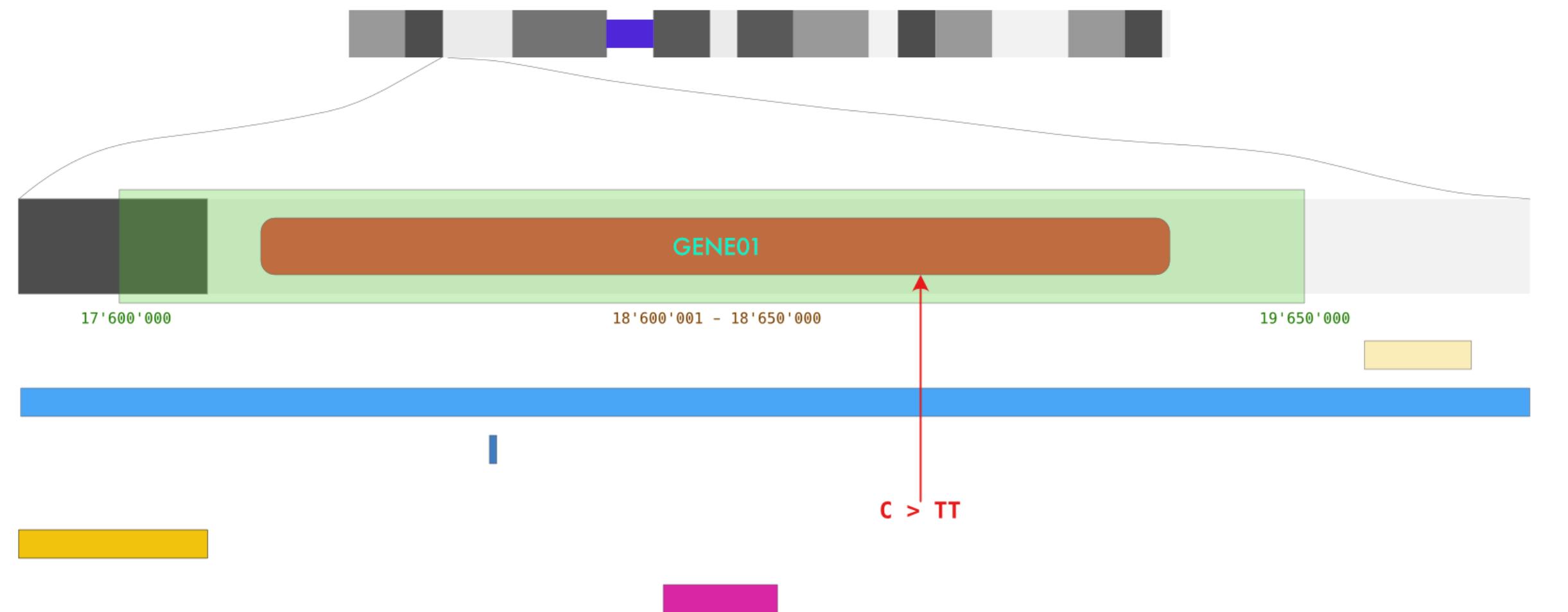
Range ("anything goes") Request

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

Beacon Range Query

Matching variants in a region

Bold: Matched Variants



DEL (Copy Number Loss)

DUP (Copy Number Gain)

SNP / INDEL ...

Unknown Annotation

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

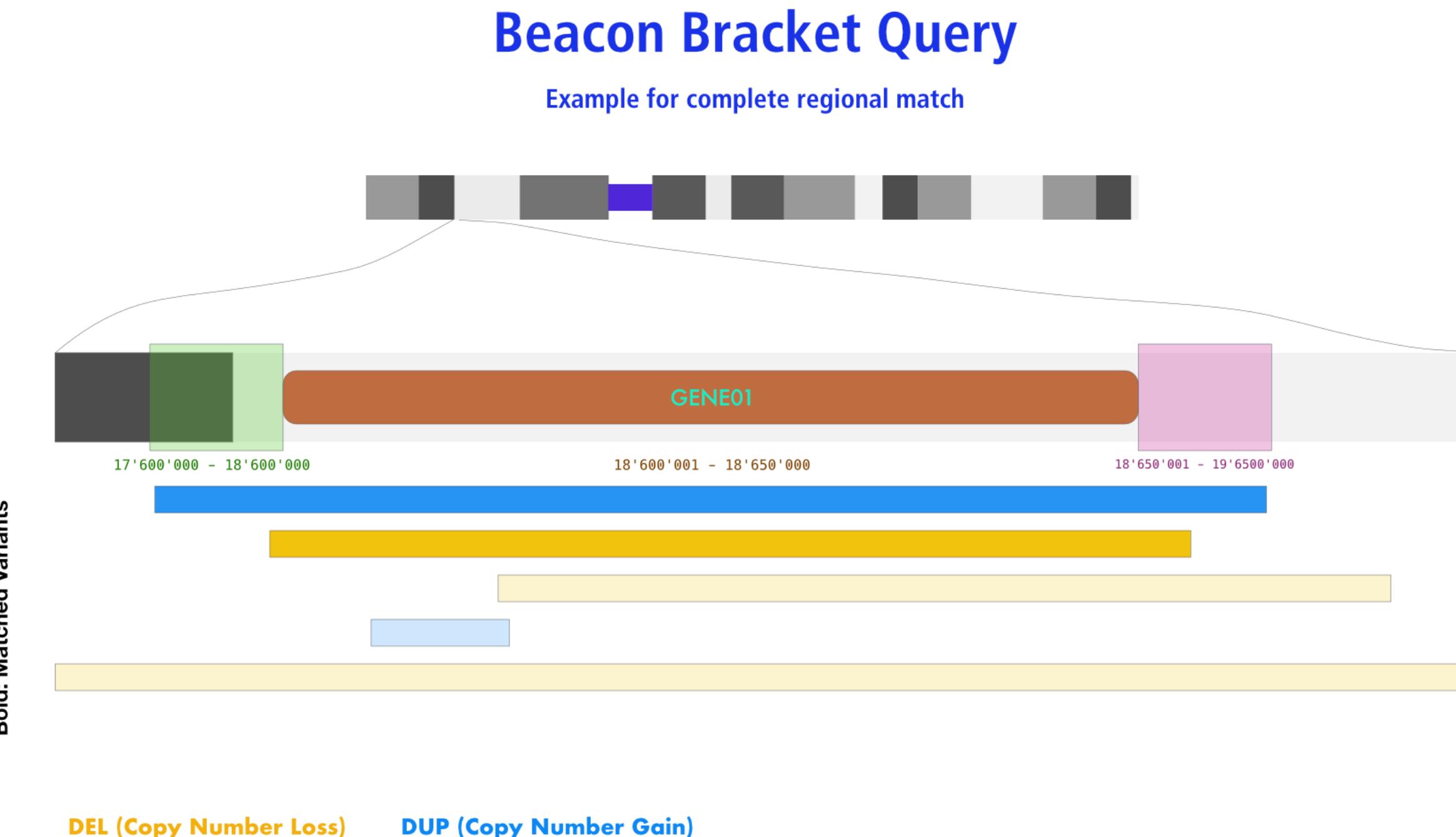
CNV Example SNV Example Range Example Gene Match
Aminoacid Example Identifier - HeLa

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](#)

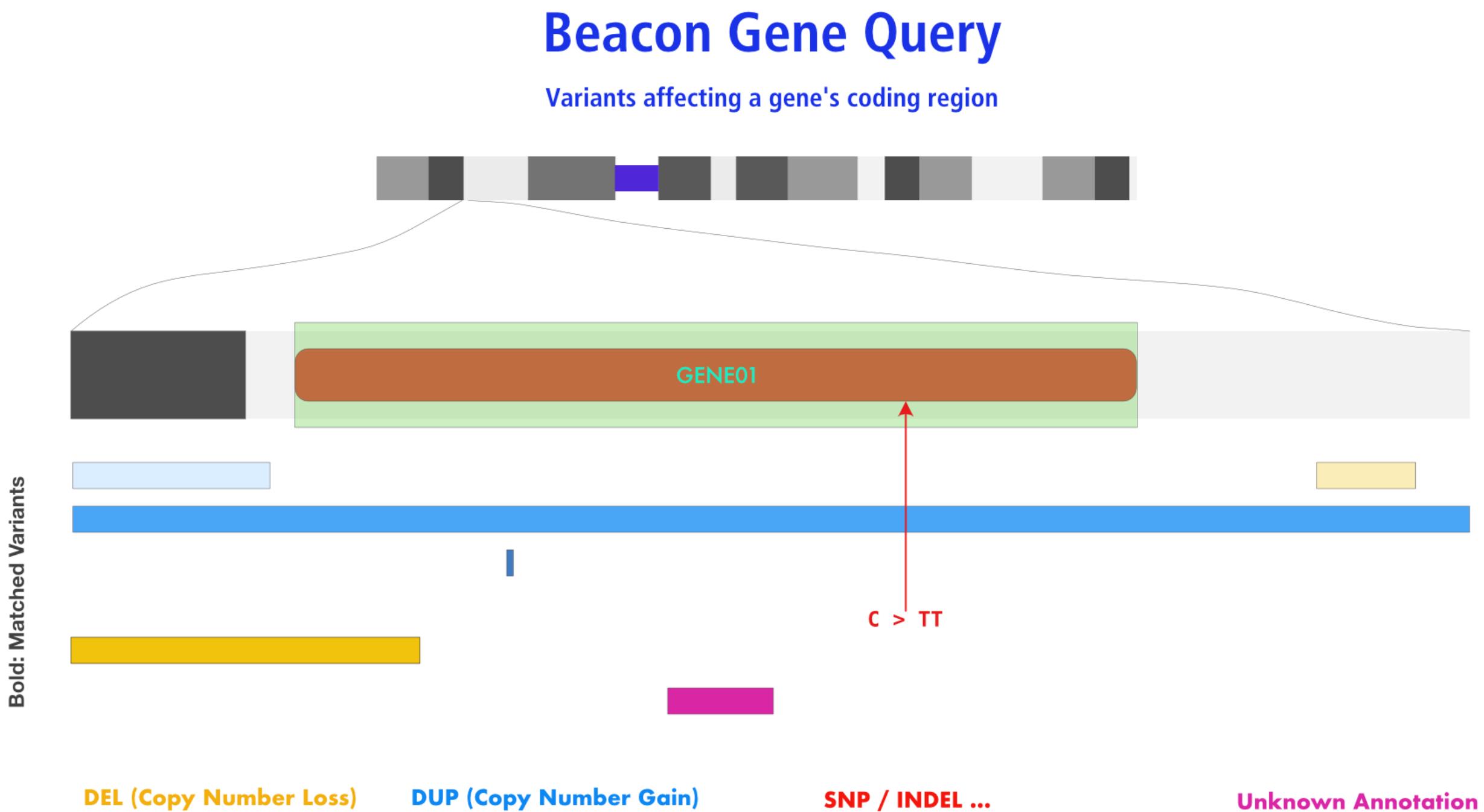
[Identifier - HeLa](#)

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. <= ~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 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>

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

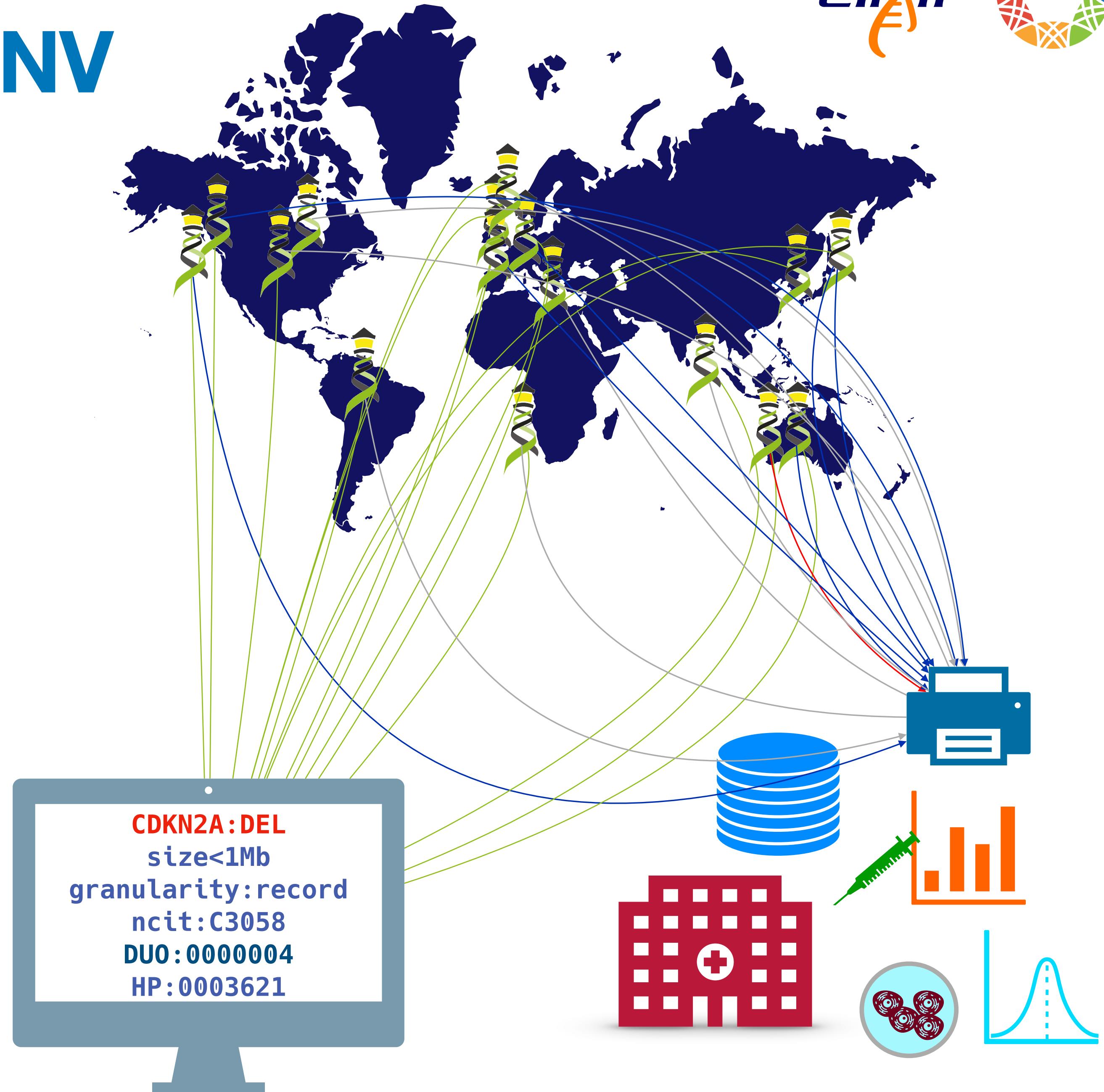
Form Utilities: Gene Spans Cytoband(s)

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

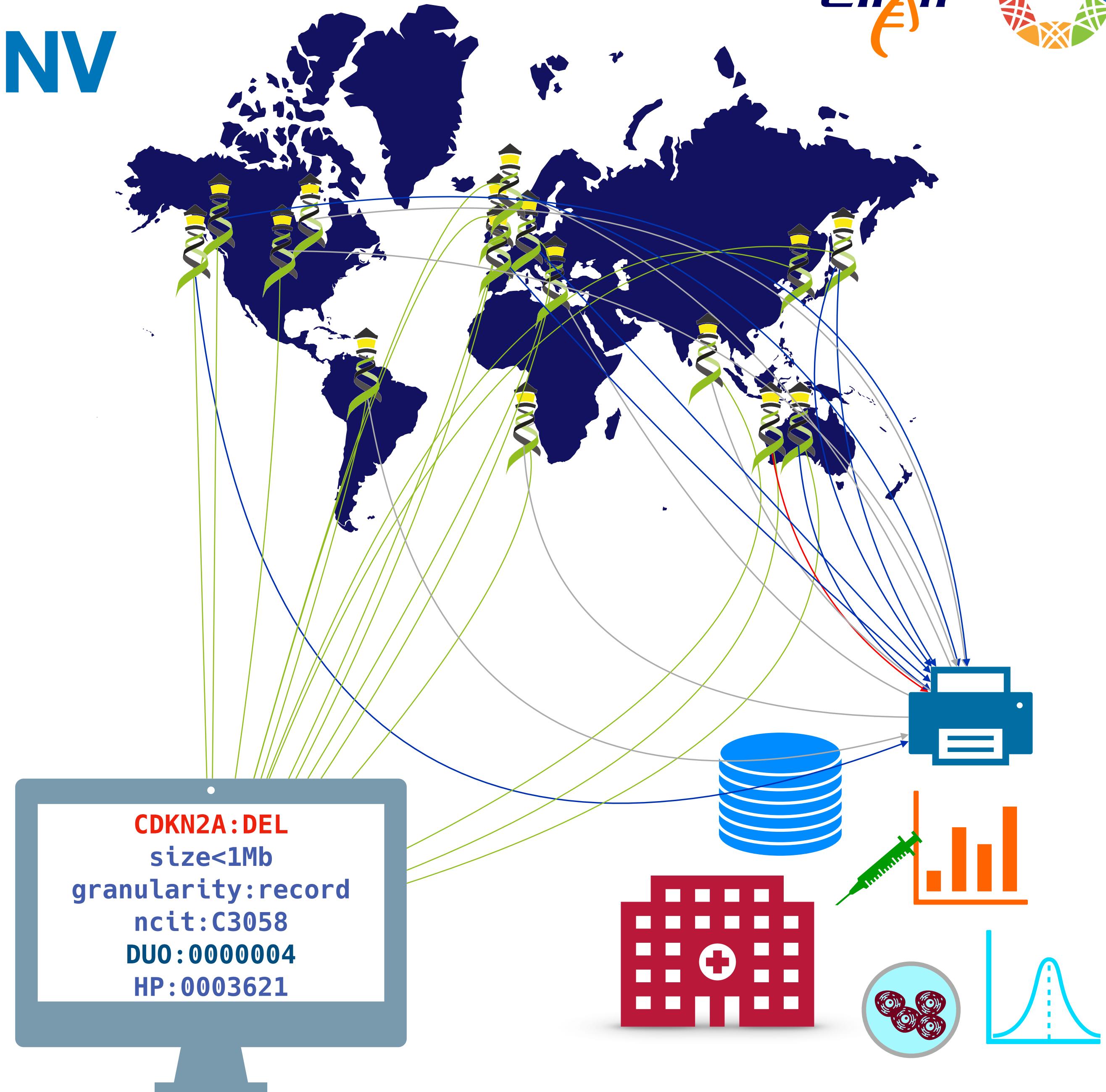
→ Collaborate w/ hCNV!



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!



h-CNV Community

Homepage & News

About ...

h-CNV Projects

CNV Annotation Standards

Databases & Resources

CNV References Project

Contacts

Genome Blog

h-CNV @ ELIXIR

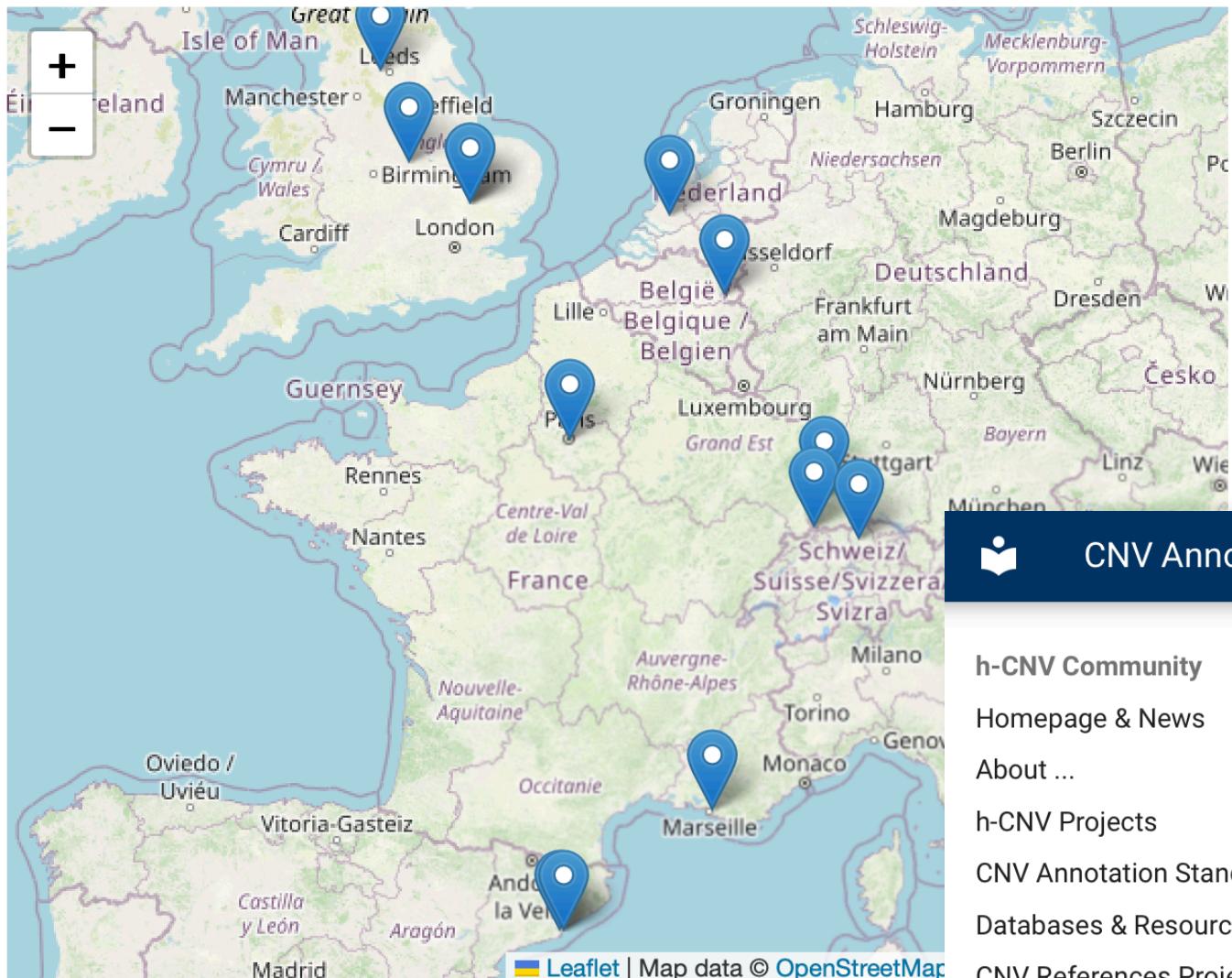
Beacon Project

ELIXIR Human Copy Number Variation community

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

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

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



ELIXIR hCNV Community

<https://cnvar.org/>

CNV Annotation Formats

Search

hcnv.github.io
star 0 fork 6

elixir

h-CNV Community

Homepage & News

About ...

h-CNV Projects

CNV Annotation Standards

Databases & Resources

CNV References Project

Contacts

Genome Blog

h-CNV @ ELIXIR

Beacon Project

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

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

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