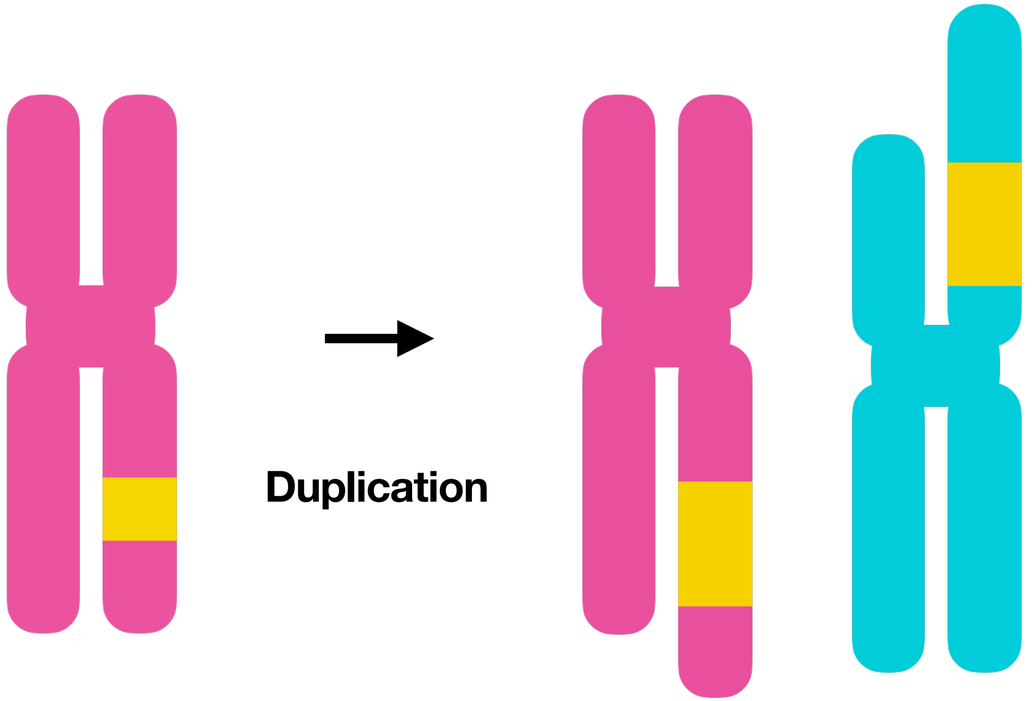
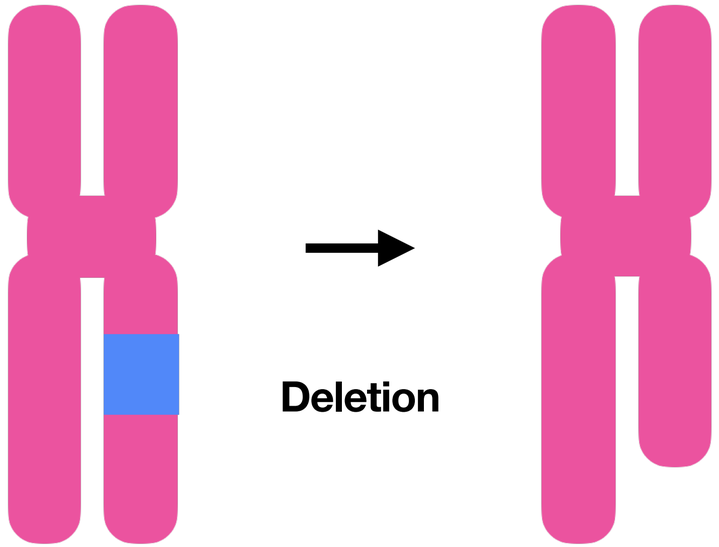
ELIXIR hCNV Community

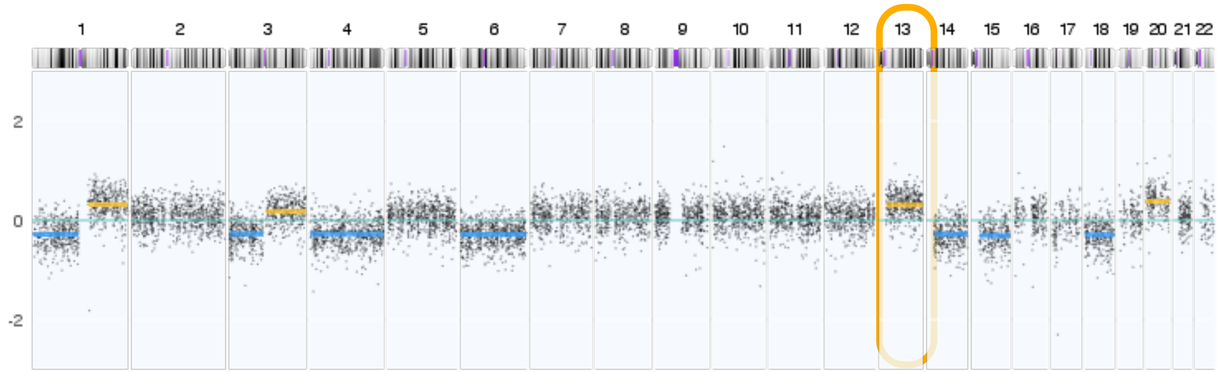
Michael Baudis | ELIXIR hCNV Community Webinar 2024

www.elixir-europe.org

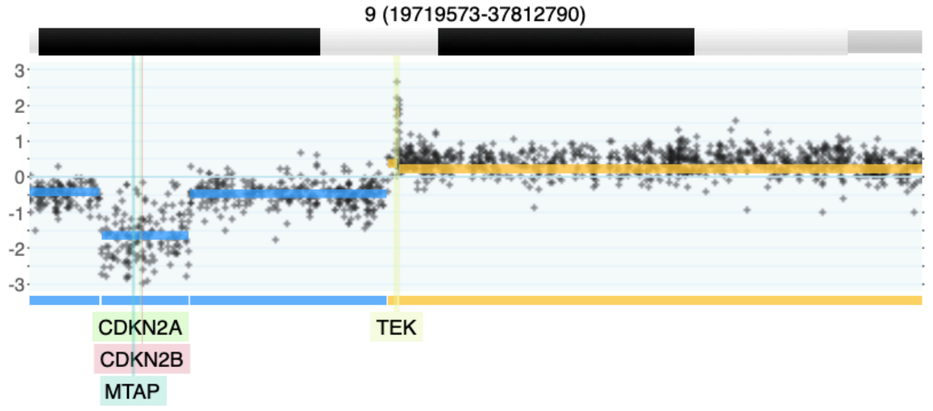
Somatic Copy Number Variation







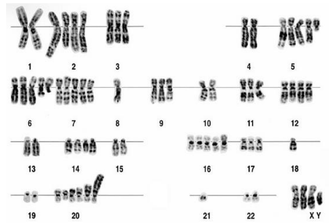
Gain of chromosome arm 13q in colorectal carcinoma

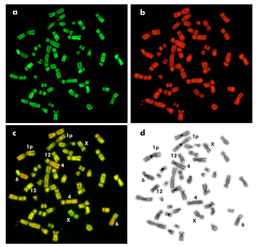


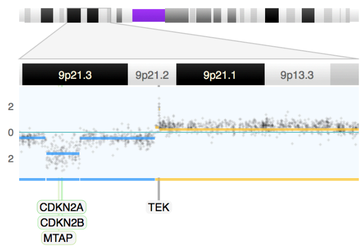
2-event, homozygous deletion in a Glioblastoma

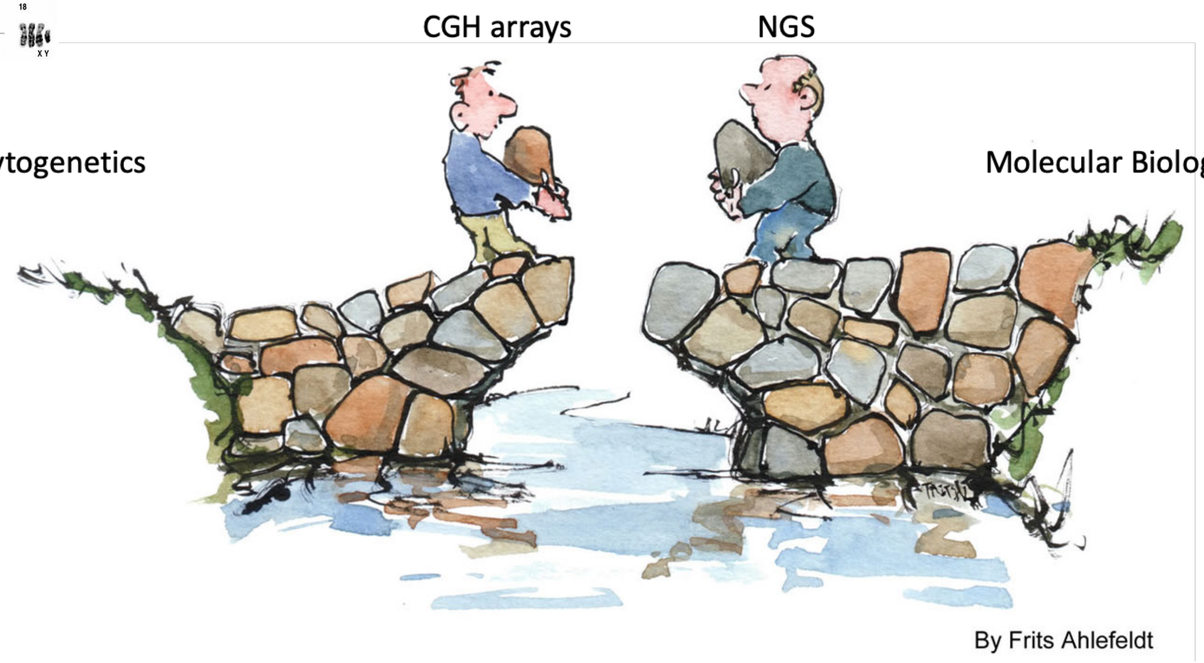
h-CNV scientific context

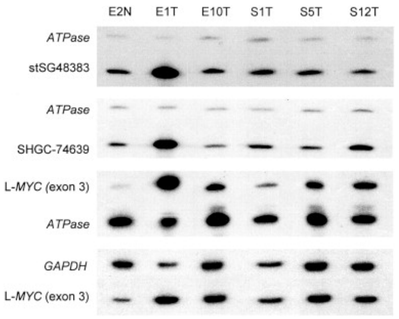
- Structural variants have been the first ones to be detected in humans (late 1950s)
- Genes' mutations shortly followed (Ingram et al. 1957)



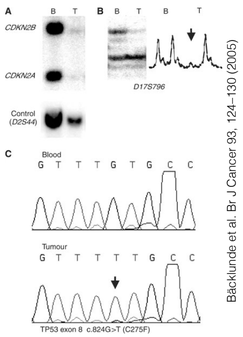








Wu et al. Am J Path 162 (5), P1603-1610, MAY 2003



Bäcklunde et al. Br J Cancer 93, 124–130 (2005)

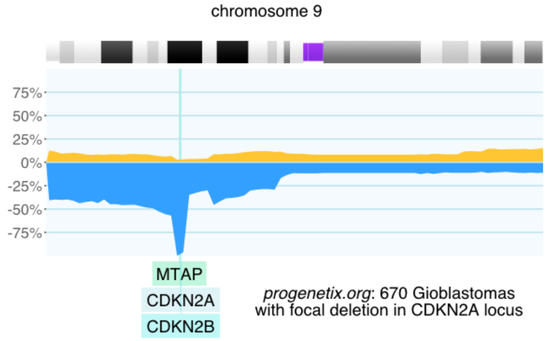
Slide: Christophe Béroud

The hCNV Community

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

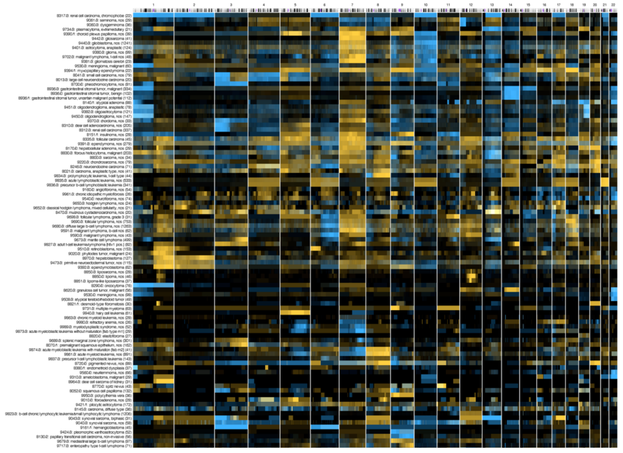
Michael Baudis | ELIXIR hCNV Community Webinar 2024

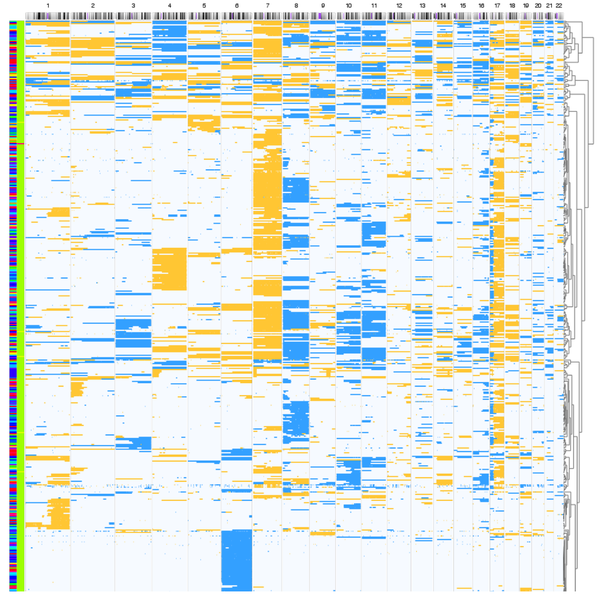
Theoretical Cytogenetics and Oncogenomics Research | Methods | Standards

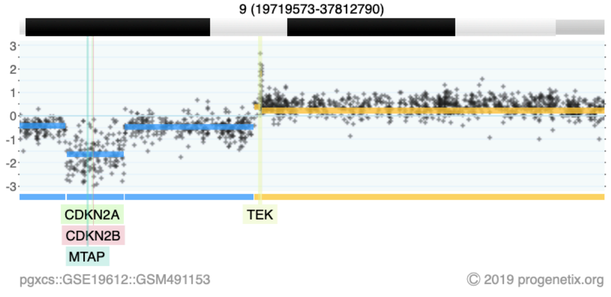


Genomic Imbalances in Cancer - Copy Number Variations (CNV)

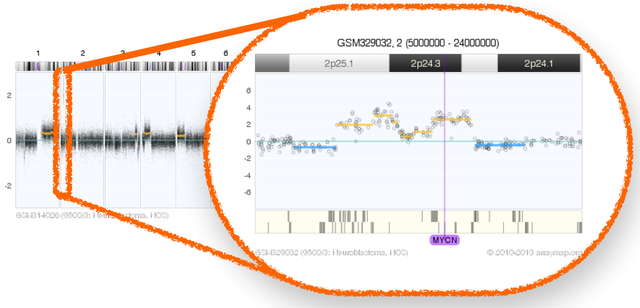
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)







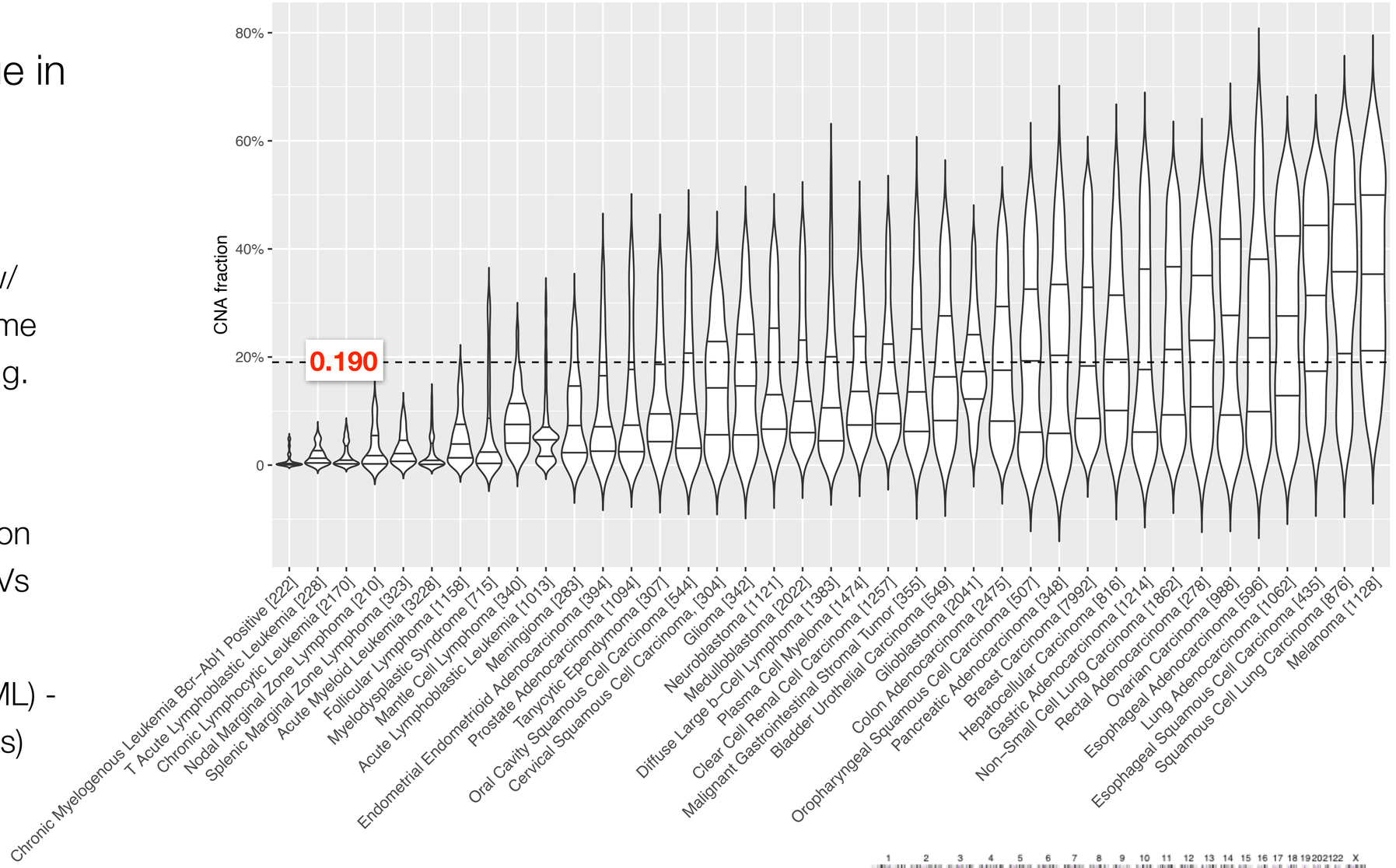
2-event, homozygous deletion in a Glioblastoma



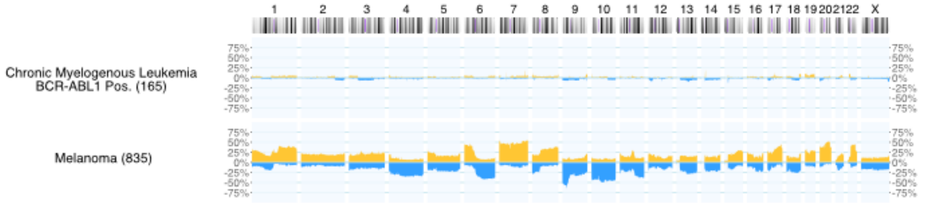
MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

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 =>



progenetix.org

Cancer Genomics Reference Resource

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

Universität Zürich[UZH]

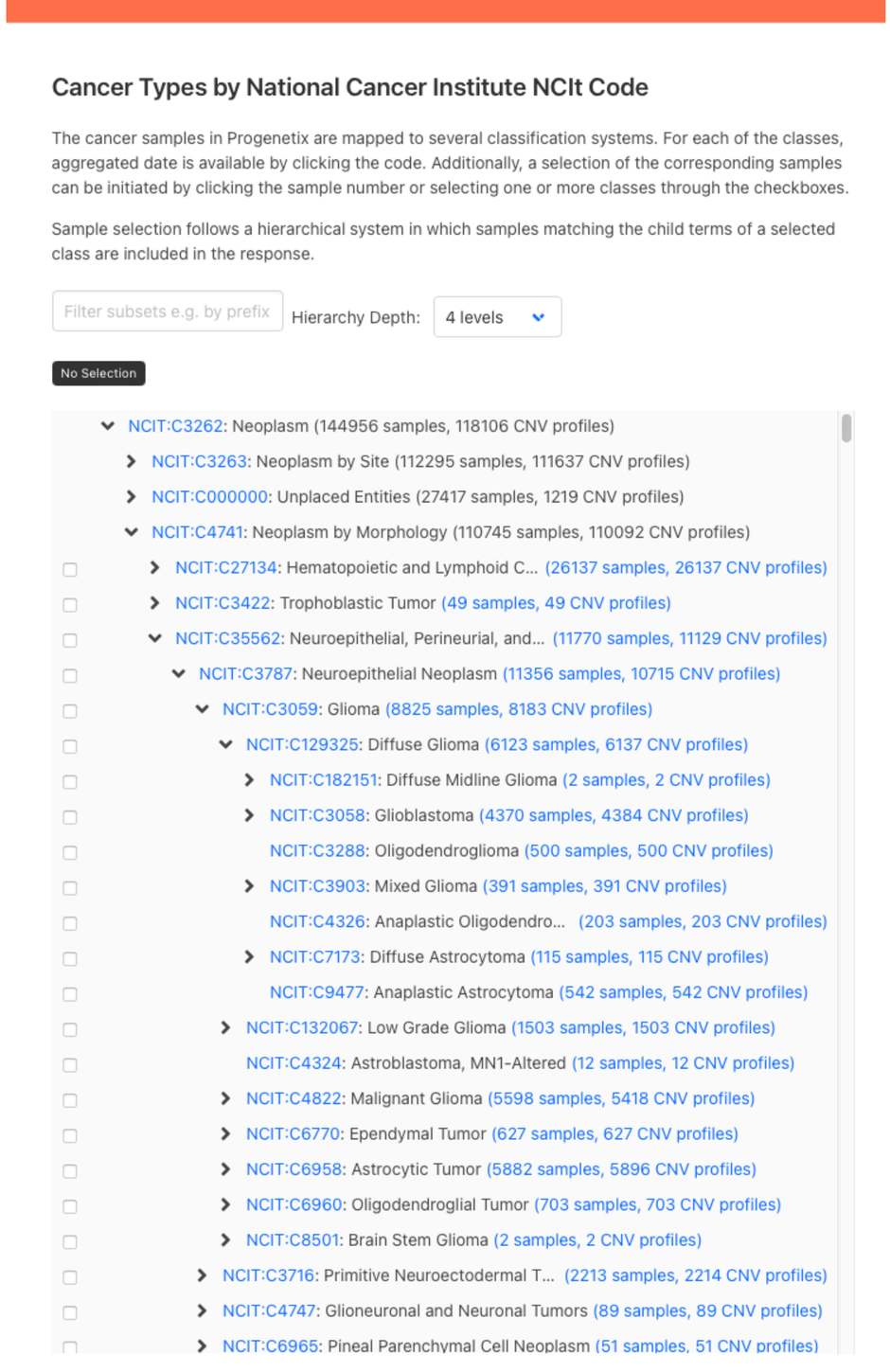
progenetix

SIB Swiss Institute of Bioinformatics

progenetix.org

Cancer Genomics Reference Resource

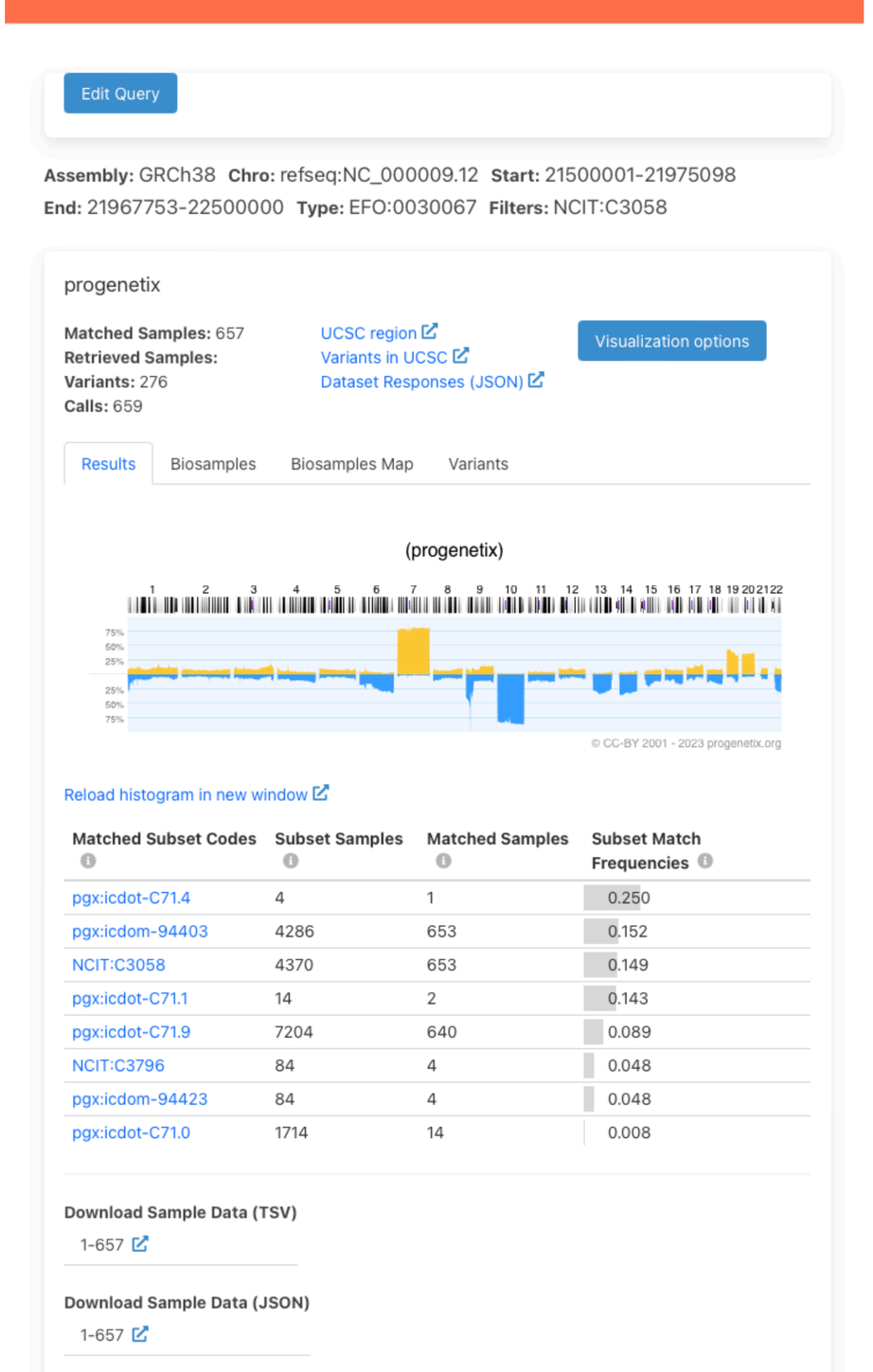
- ***open*** resource for oncogenomic profiles
- over **116'000 cancer CNV profiles**
- more than **800 diagnostic types**
- inclusion of reference datasets (e.g. TCGA)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
-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



progenetix.org

Cancer Genomics Reference Resource

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

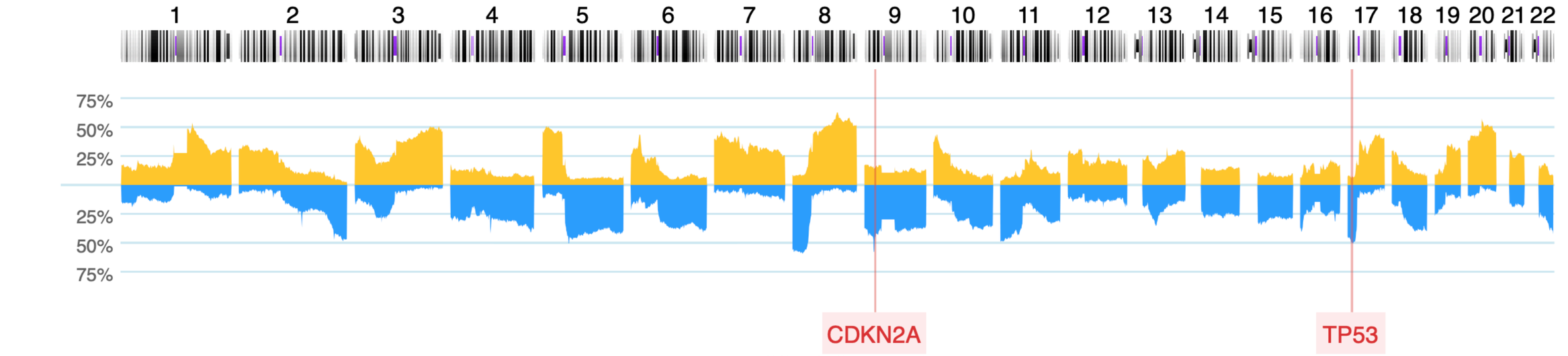


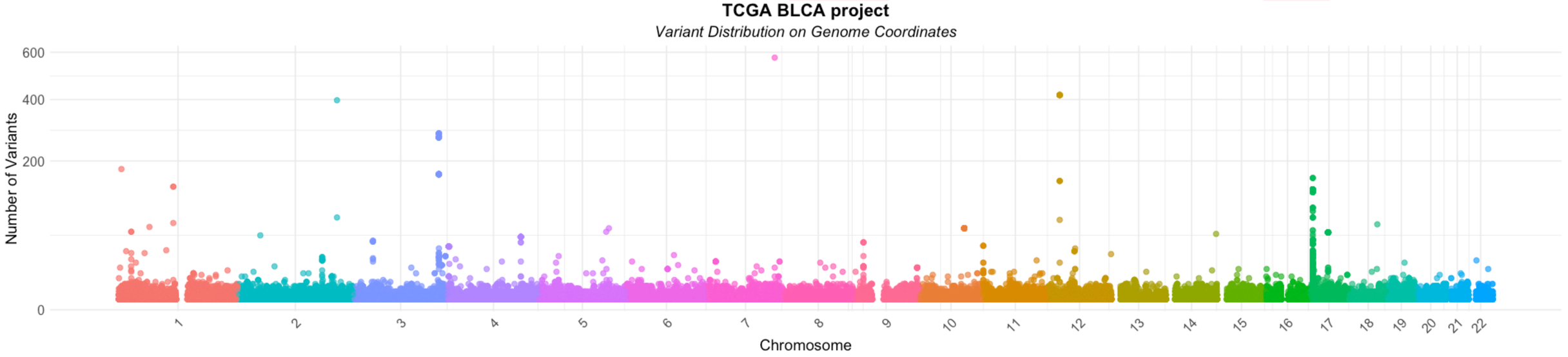
Universität Zürich[UZH]

progenetix

SIB Swiss Institute of Bioinformatics

TCGA BLCA project (pgx:TCGA.BLCA)





Cancer Cell Lines

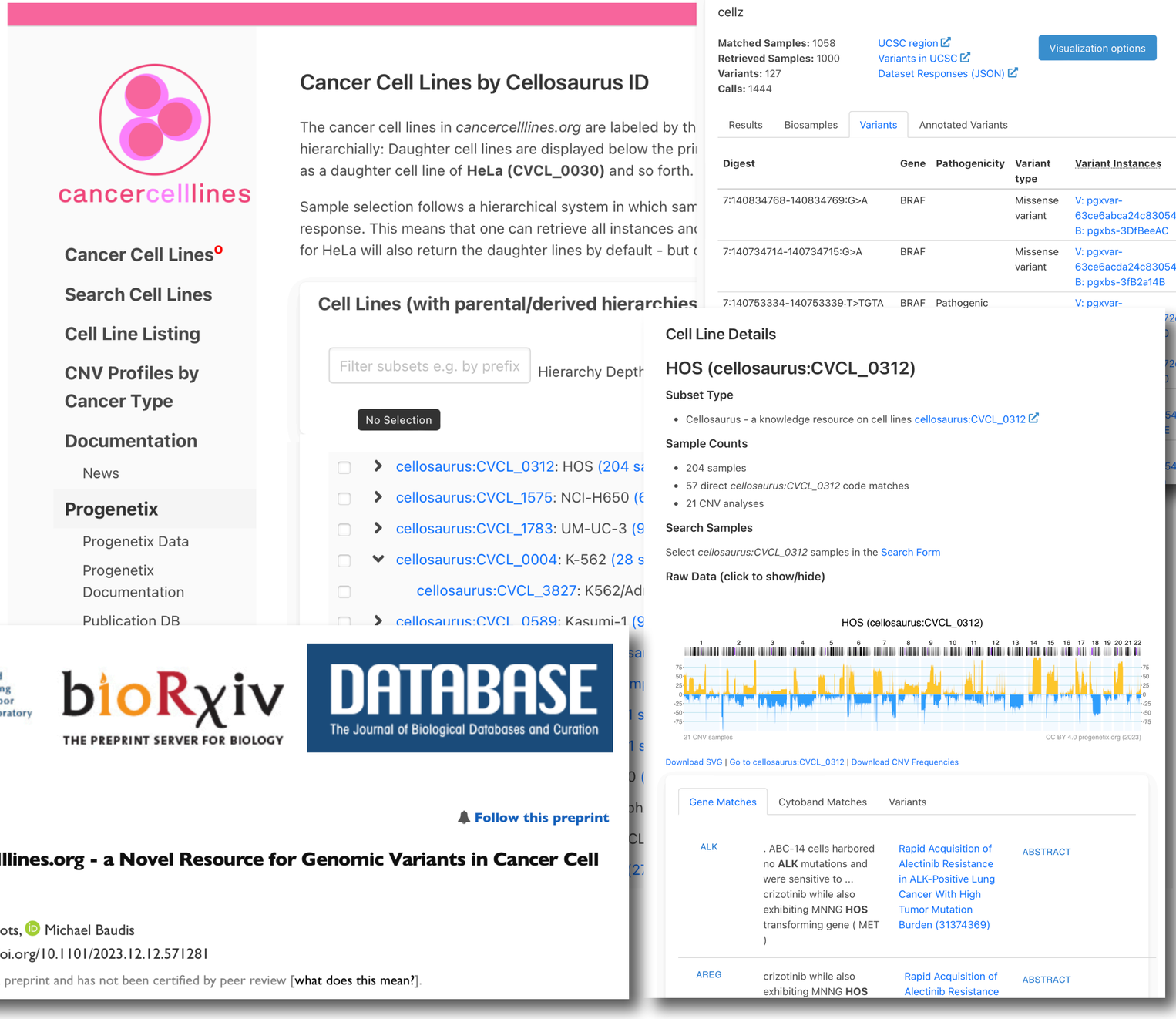
Cancer Genomics Reference Resource

- 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



Progenetix Stack

progenetix

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

React

APACHE HTTP SERVER PROJECT

python™

mongoDB

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

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

variants analyses biosamples individuals

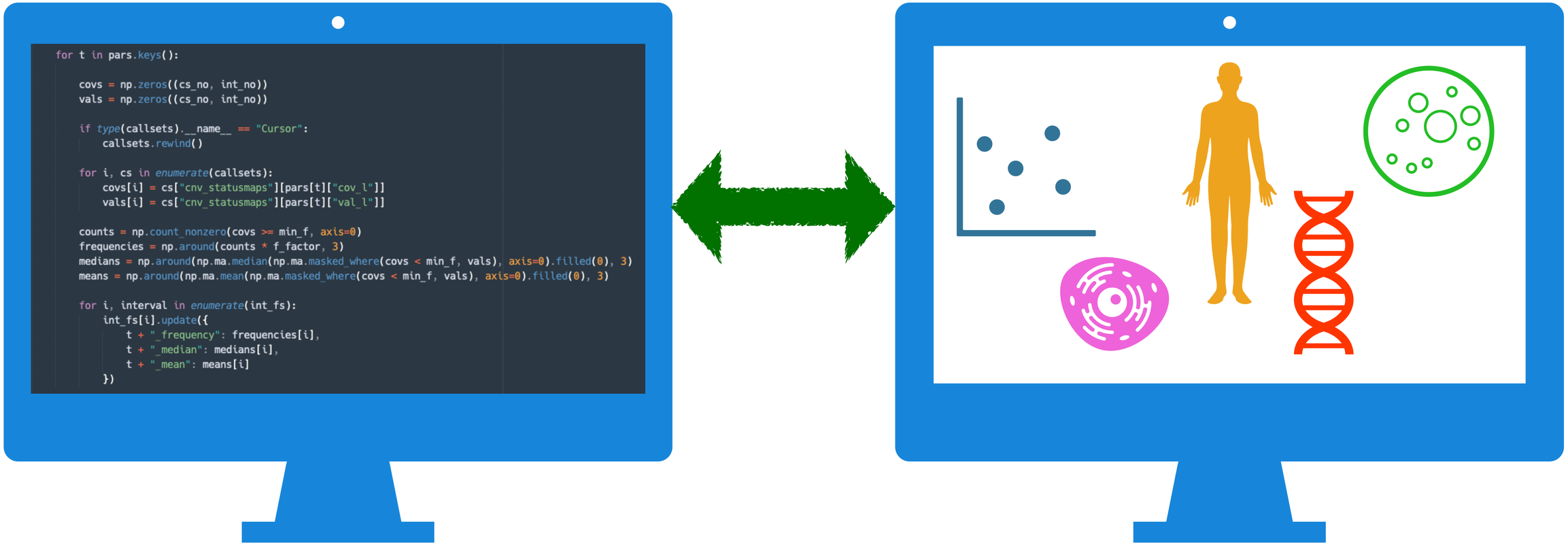
Entity collections

collations geolocs genespans publications qBuffer

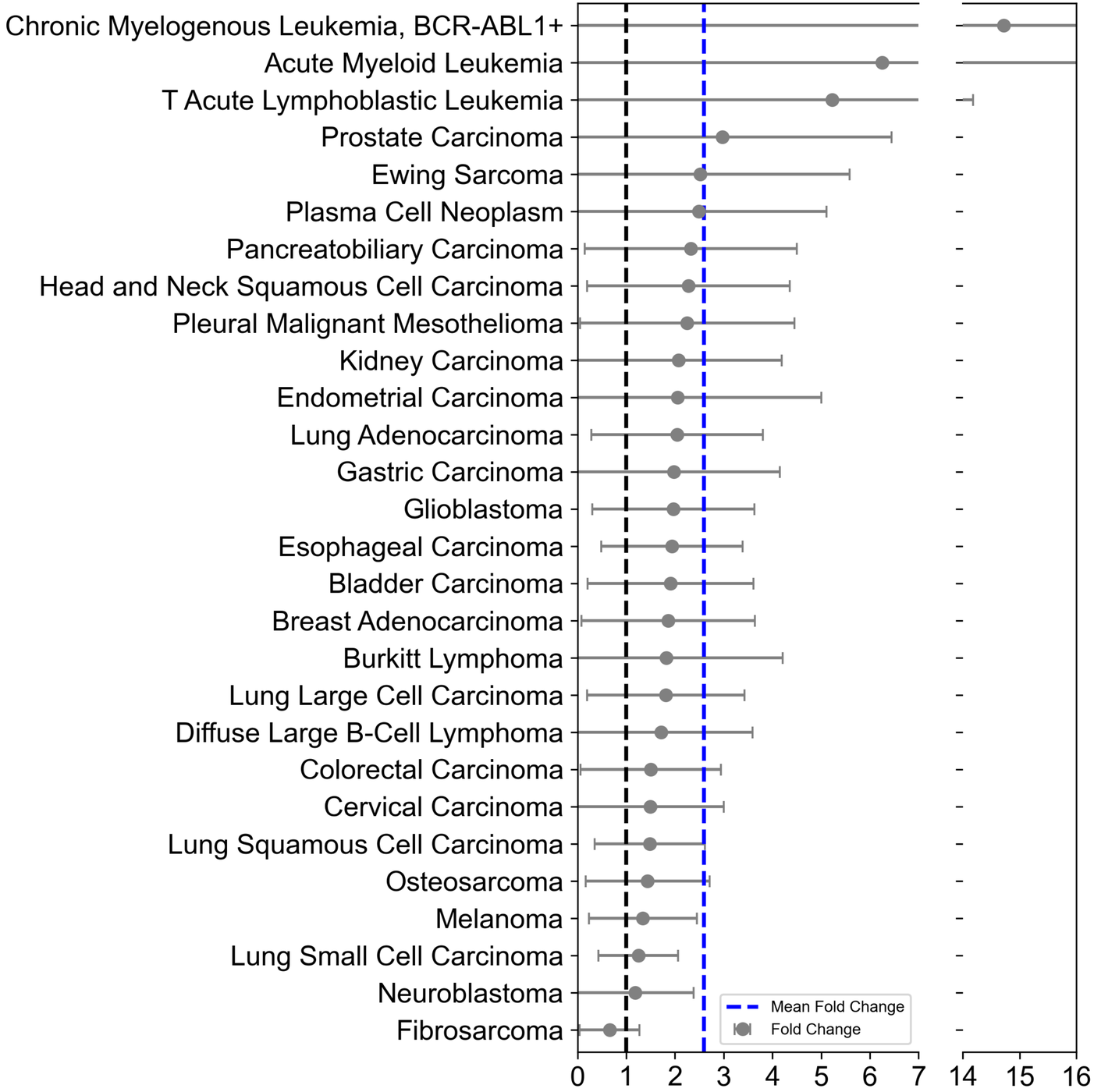
Utility collections

github.com/progenetix/bycon/

{BioInformaticsScience}



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

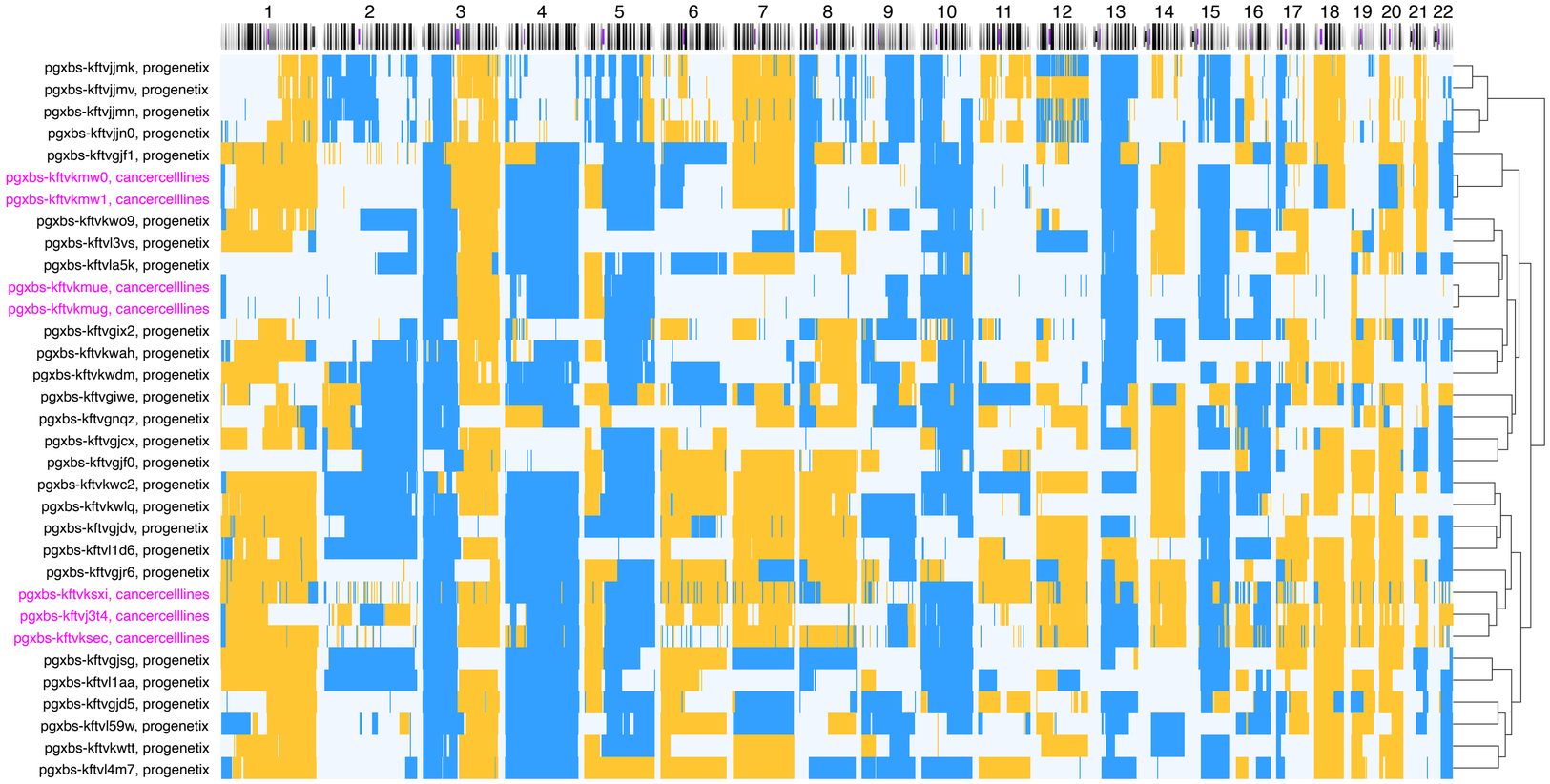


Fold changes between genome CNV coverages of cell lines and tumors

Lead: Rahel Paloots

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

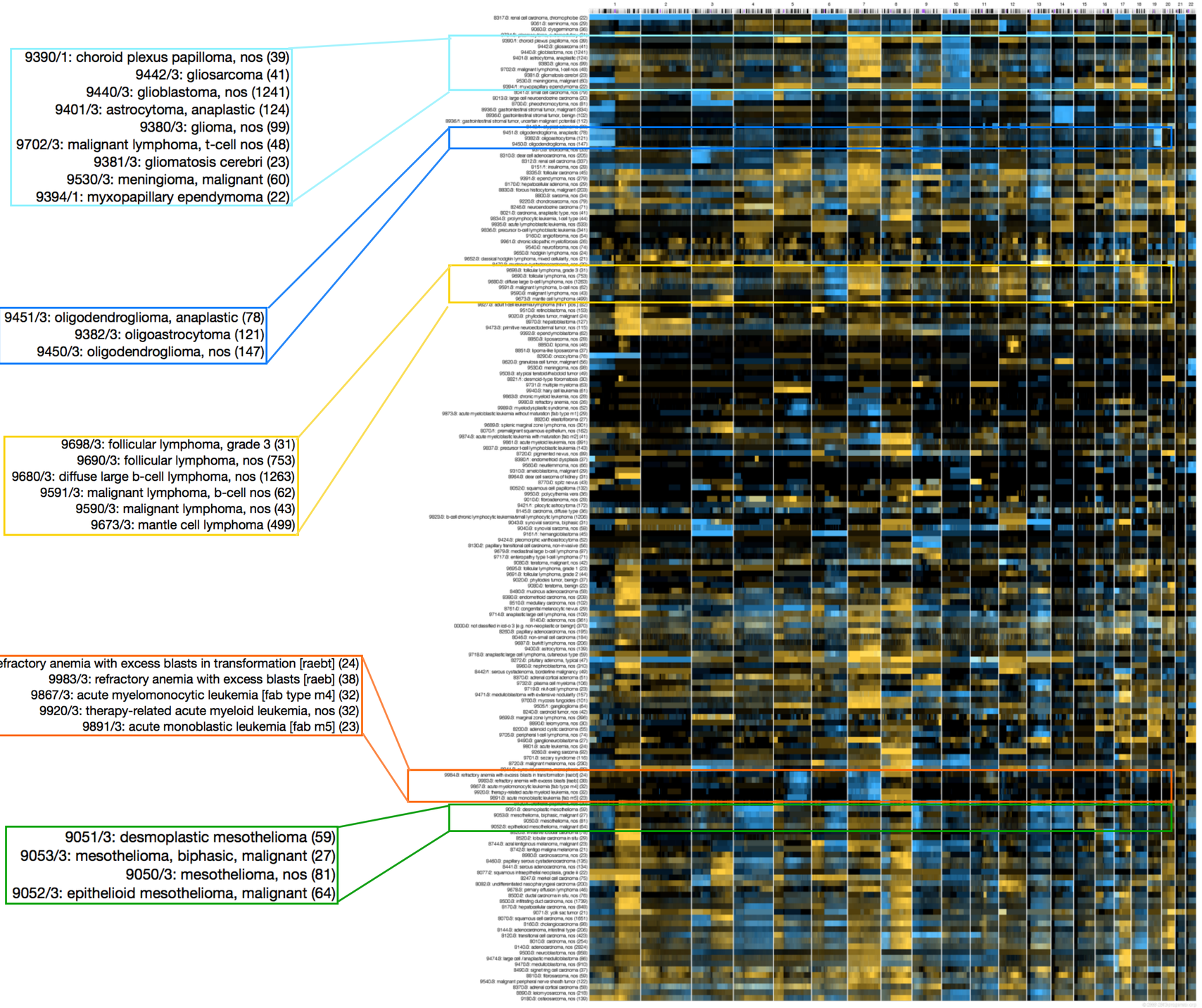


Lead: Rahel Paloots

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

Lead: Rahel Paloots

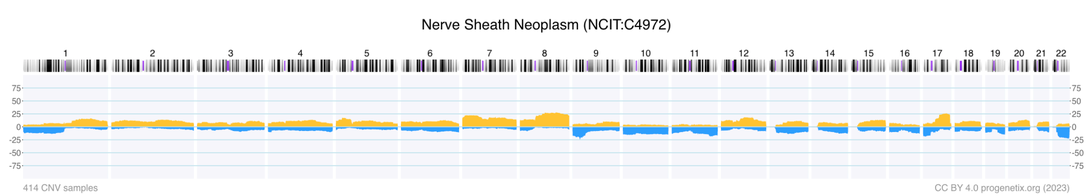
Somatic Mutations In Cancer: Patterns
Making the case for genomic classifications
Some related cancer entities show similar copy number profiles

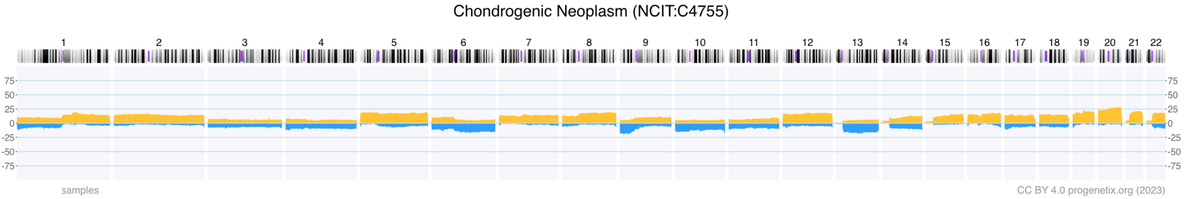


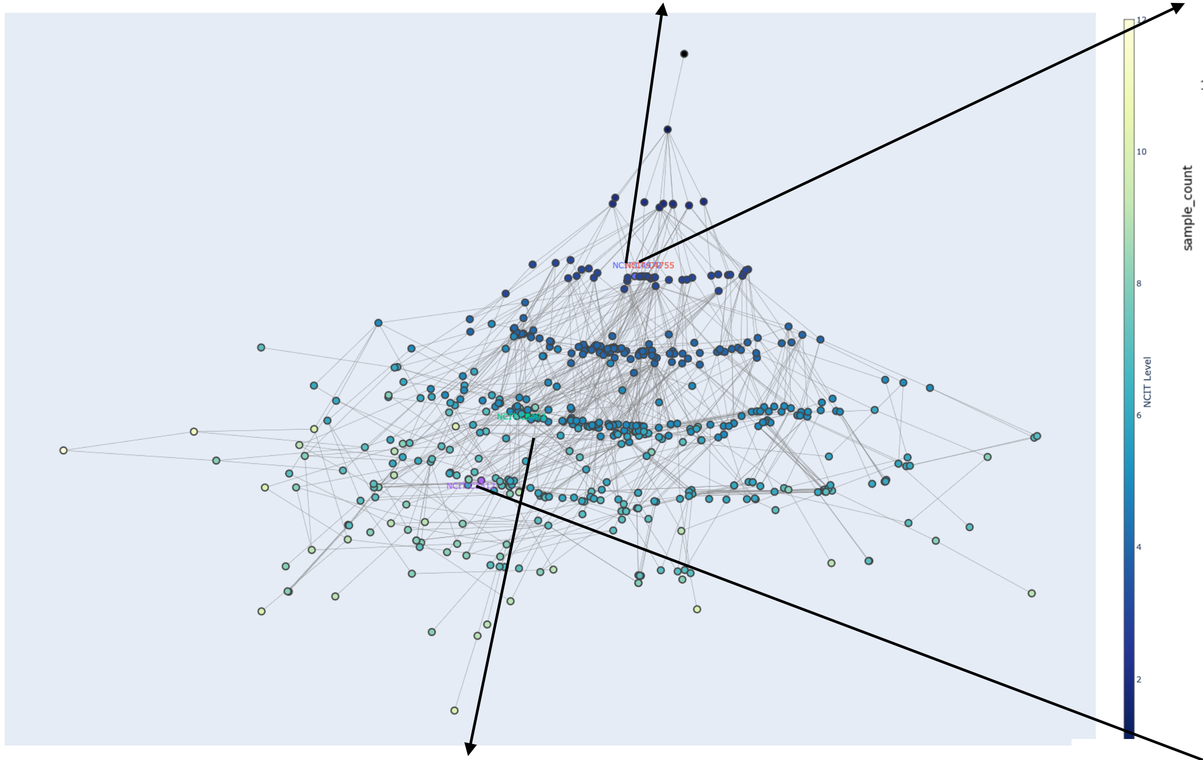
arrayMap

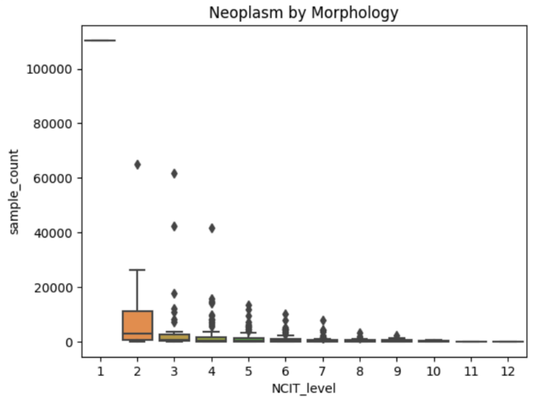
CNV profiles heterogeneity vs cancer classification

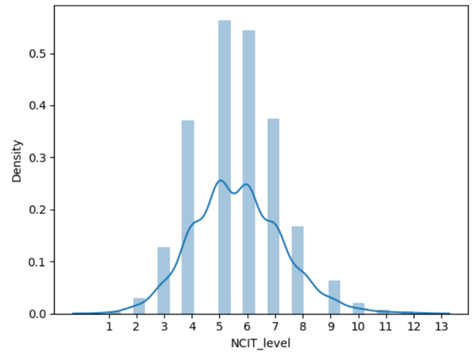
Correspondance of genomic profiles to NCIT cancer hierarchy

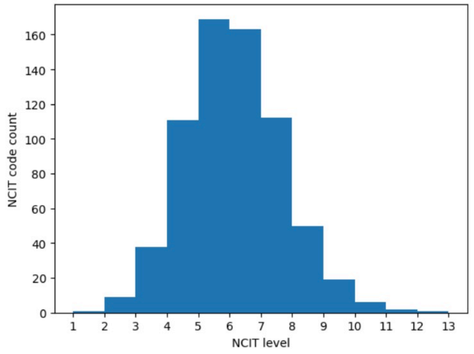


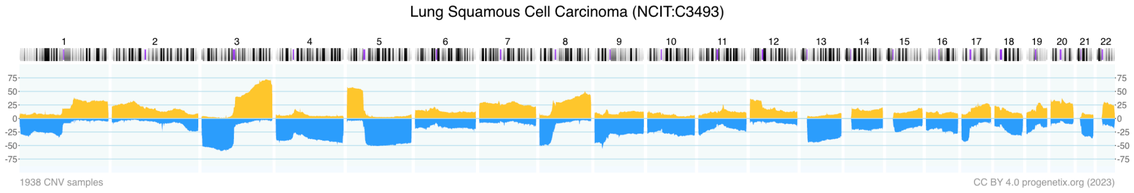


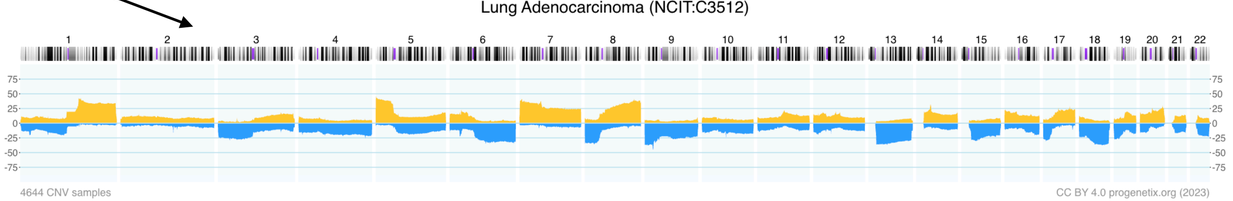








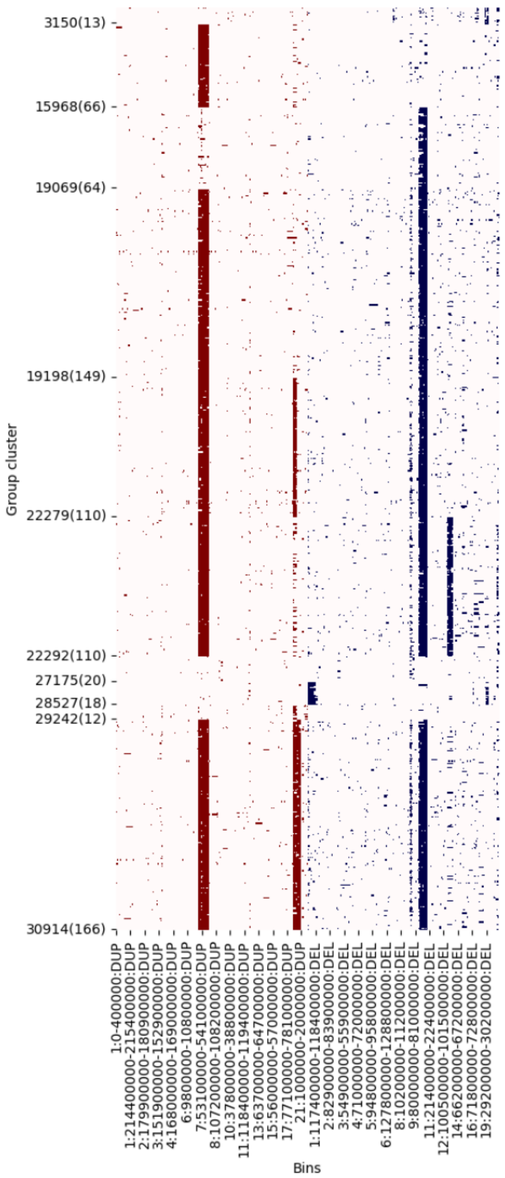


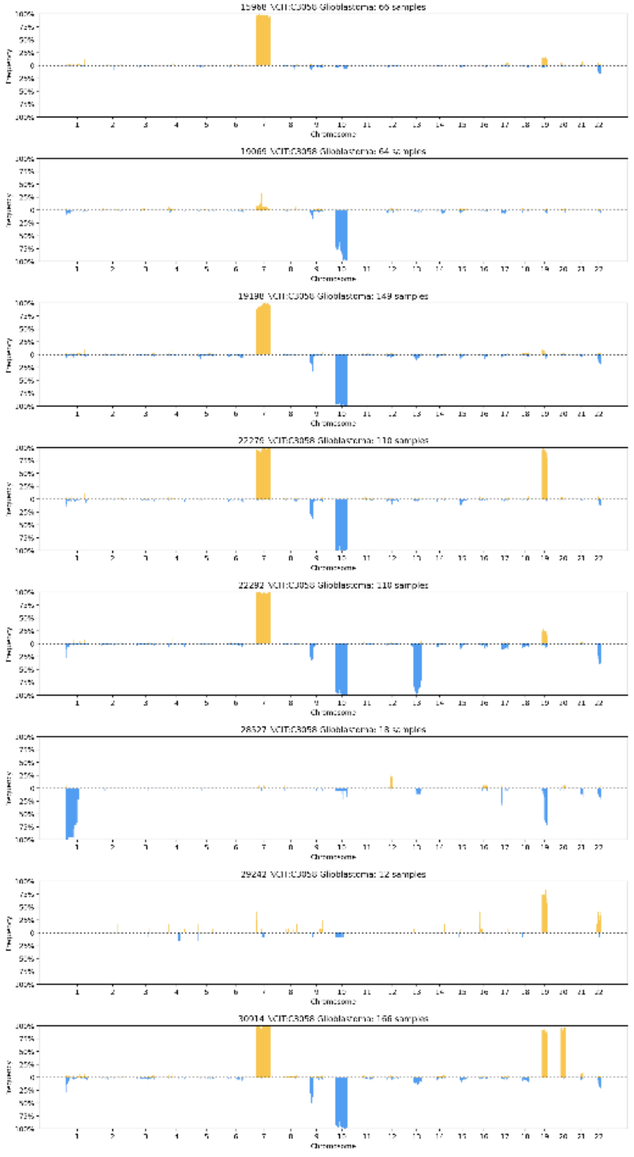


Lead: Ziying Yang

Results

Entity CNV heterogeneity: Glioblastoma



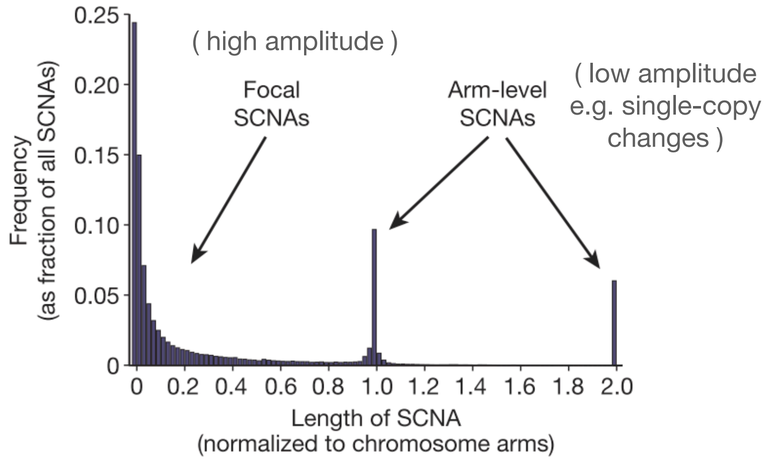


group cluster	CNV features
15968	Dup 7
19069	Del 10
19198	Dup 7, Del 10
22279	Dup 7, Del 10, Dup 19
22292	Dup 7, Del 10, Del 13
28527	Del 1p, Del 19q
29242	Dup 19
30914	Dup 7, Del 10, Dup 19, Dup 20

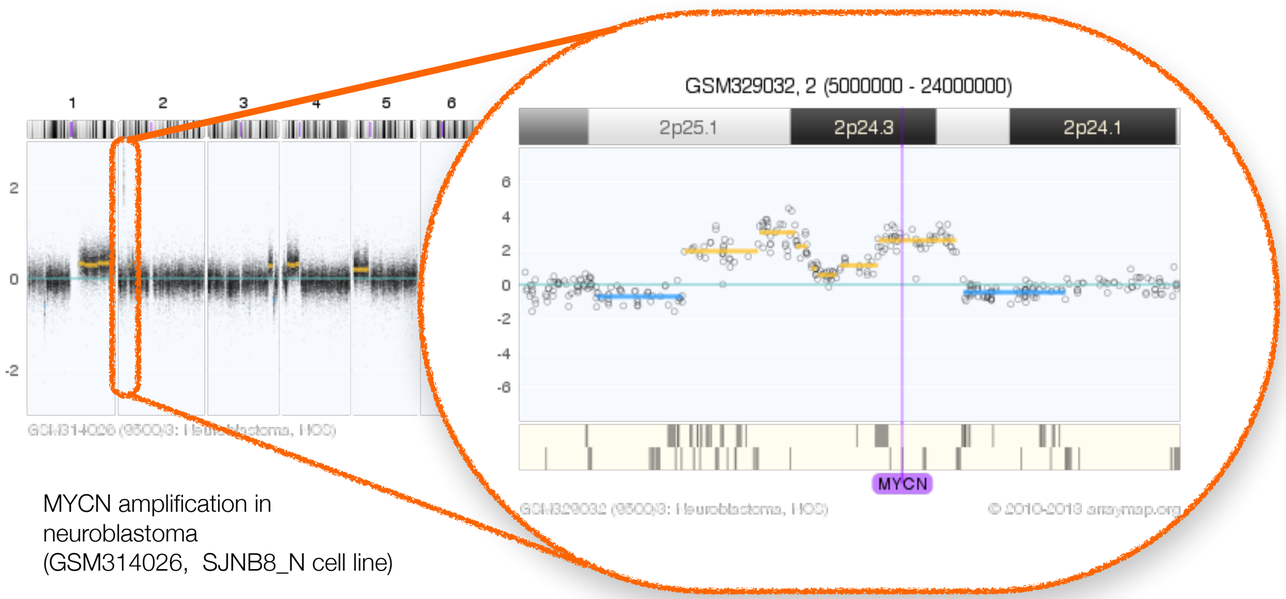
Lead: Ziying Yang

CNV Categorization

different levels of CNV



Rameen et al 2010 Nature



MYCN amplification in neuroblastoma (GSM314026, SJNB8_N cell line)

Lead: Hangjia Zhao

GA4GH Variation Representation Specification

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

CopyNumberChange

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

Computational Definition

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

Information Model

Some CopyNumberChange attributes are inherited from Variation.

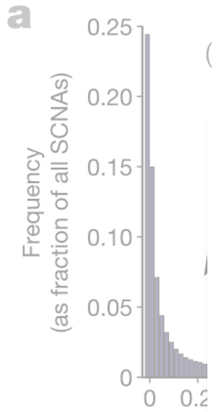
Field	Type	Limits	Description
_id	CURIE	0..1	Variation Id. MUST be unique within document.
type	string	1..1	MUST be “CopyNumberChange”
subject	Location | CURIE | Feature	1..1	A location for which the number of systemic copies is described.
copy_change	string	1..1	MUST be one of “efo:0030069” (complete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

CNV Categorization

different levels of CNV

GA4GH Variation Representation Specification

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



(high amplitude)

Focal

Arm-level

(low amplitude

OXFORD

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

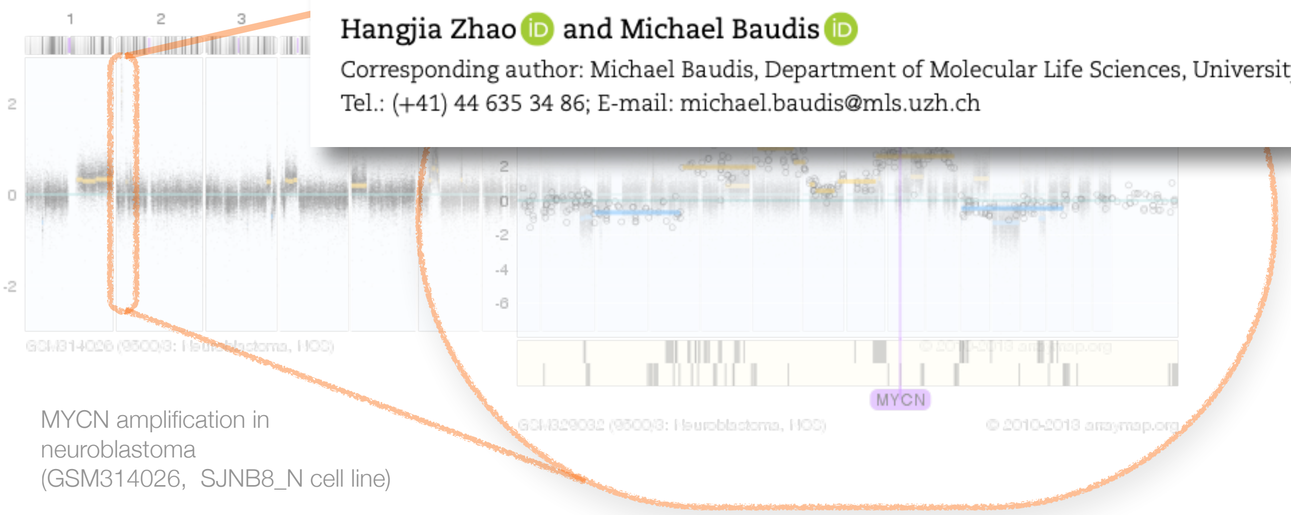
https://doi.org/10.1093/bib/bbad541

Problem Solving Protocol

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



MYCN amplification in neuroblastoma (GSM314026, SJNB8_N cell line)

CopyNumberChange

Copy Number Change captures a categorization of copies of a molecule within a system, relative to a
allers, particularly in the somatic
nd less useful in practice than
s relative statements, and many
nterpreted to be relative copy

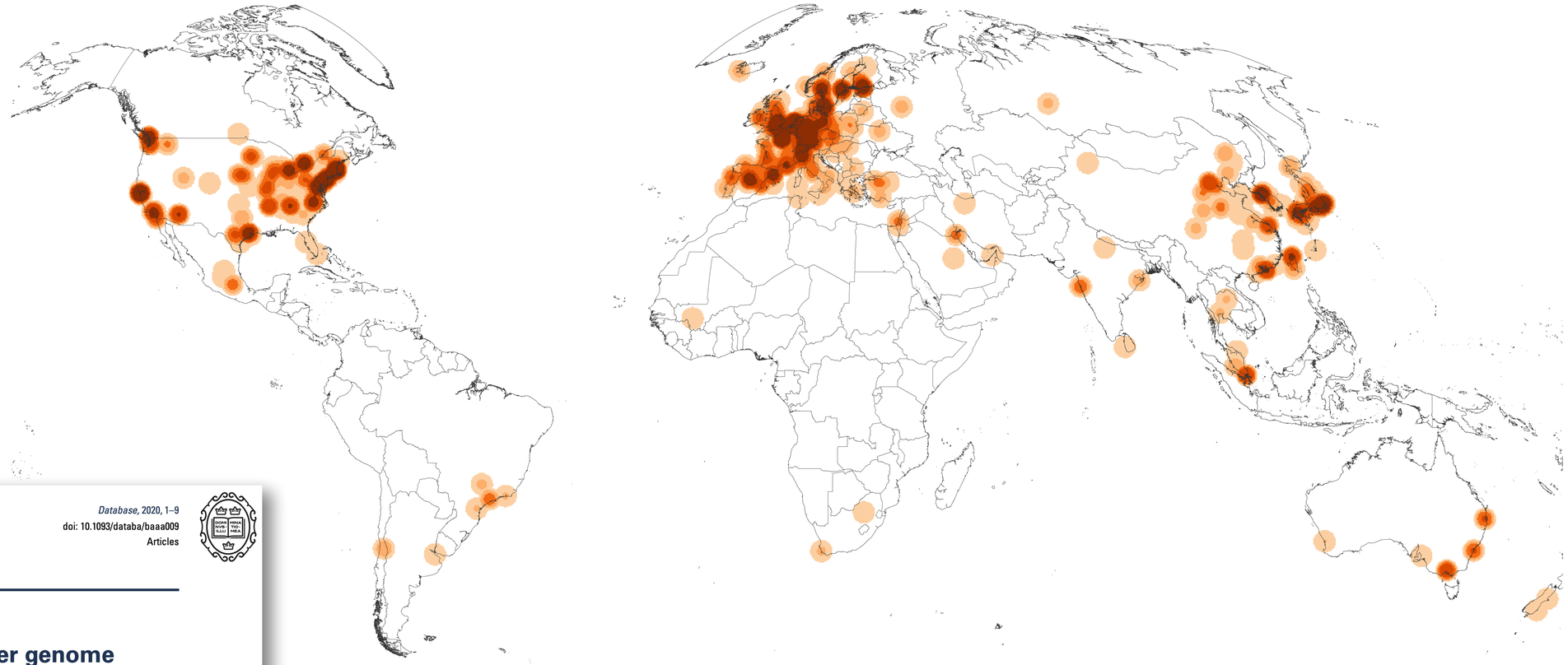
a system (e.g. genome, cell,

_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).

Lead: Hangjia Zhao

Where does Genomic Data Come From?

Geographic bias in published cancer genome profiling studies



DATABASE
The Journal of Biological Databases and Curation

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

Articles

Geographic assessment of cancer genome profiling studies

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

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

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

Universität Zürich[UZH]

progenetix

The hCNV Community

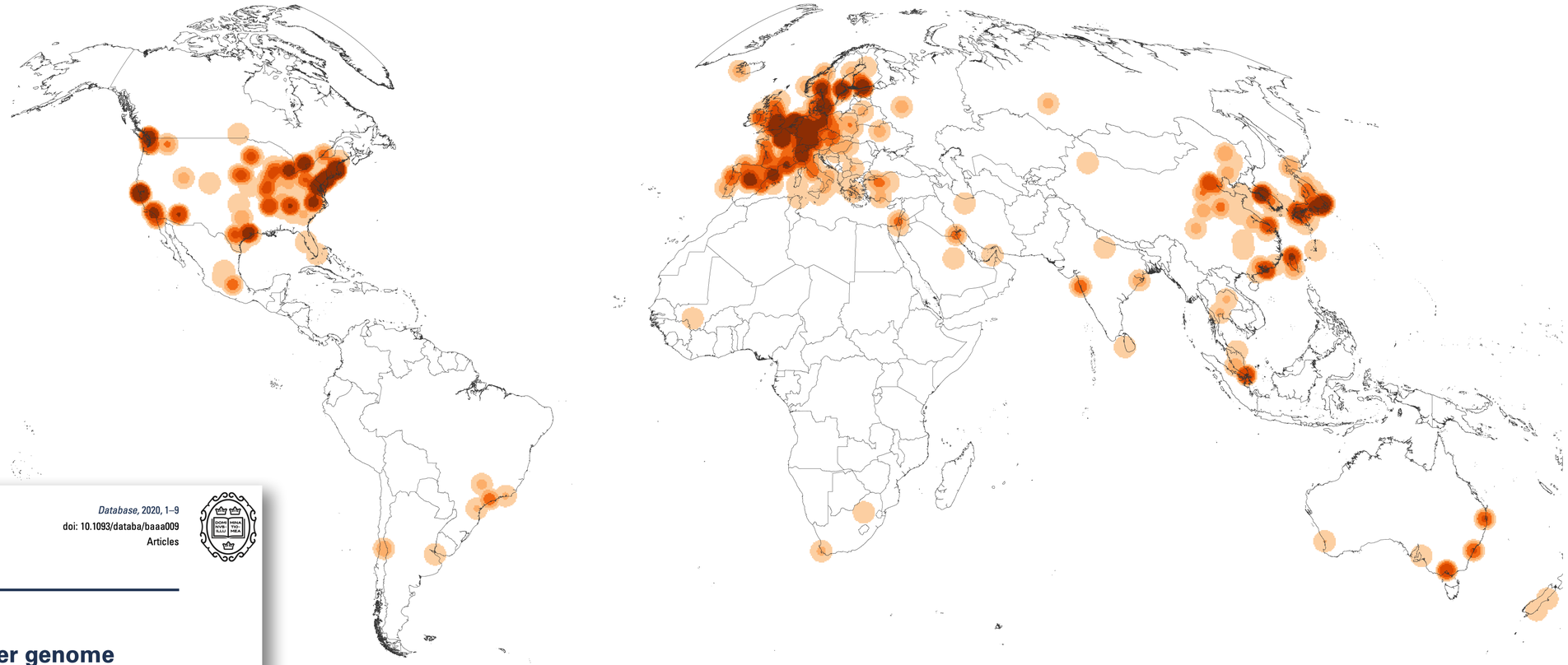
Standards for CNV annotation and their use in data discovery protocols

Michael Baudis | ELIXIR hCNV Community Webinar 2024

elixir

Where does Genomic Data Come From?

Geographic bias in published cancer genome profiling studies



DATABASE
The Journal of Biological Databases and Curation

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

Articles

Geographic assessment of cancer genome profiling studies

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

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

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

Different Approaches to Data Sharing

progenetix

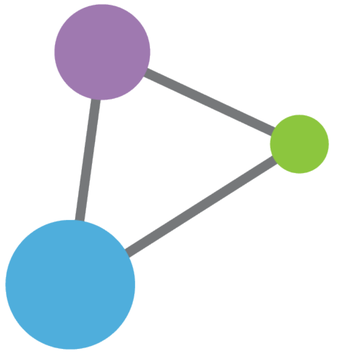
European Genome-Phenome Archive

SPHN Swiss Personalized Health Network

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

FROM

Data Copying

TO

Data Visiting

A New Paradigm for Data Sharing

FROM

STANDARDS

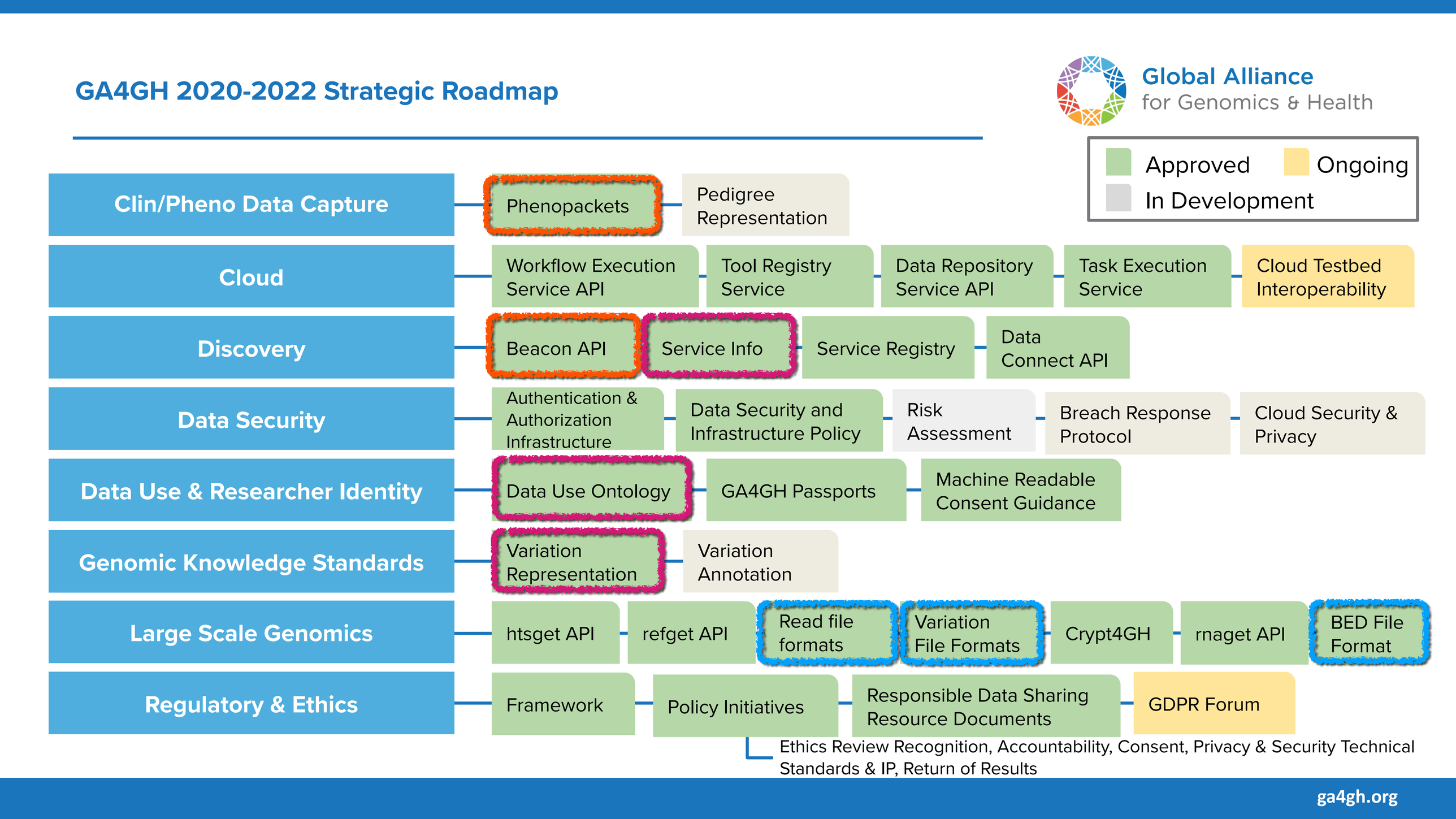
TO

Data Copying

Data Visiting

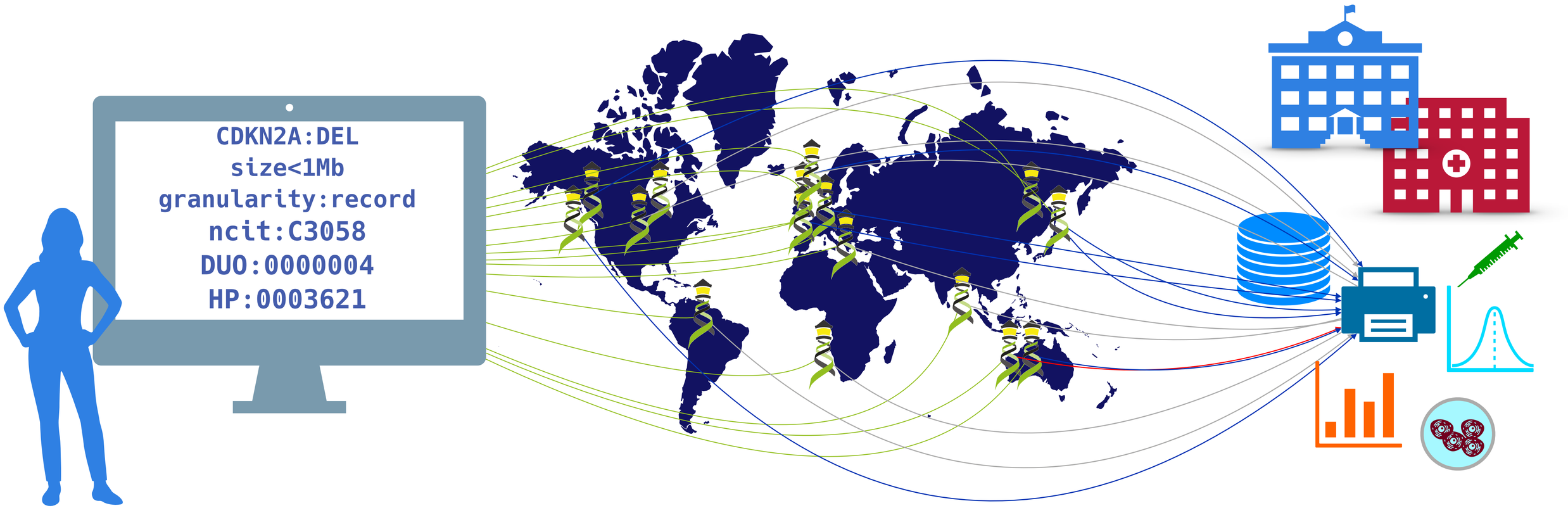
GA4GH 2020-2022 Strategic Roadmap

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`	here a chromosome name, but could be any sequence identifier
`start`	a genomic position defined as using a 0-base, interbase format
`referenceBases`	a sequence in the reference genome
`alternateBases`	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

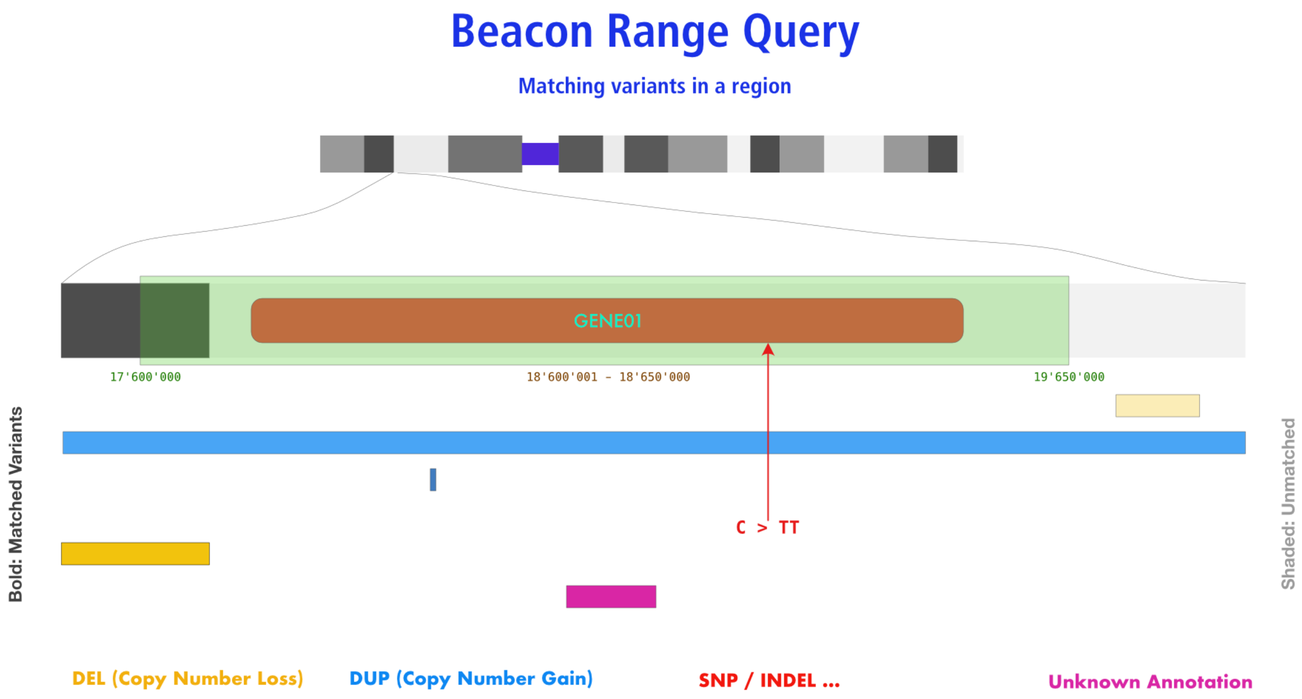
This table is maintained in parallel with the Beacon v2 documentation.

EFO	Beacon	VCF	SO	GA4GH VRS[1]	Notes
EFO:0030070 copy number gain	DUP [2] or EFO:0030070	DUP 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
EFO:0030069 complete genomic deletion	DEL [2] or EFO:0030069	DEL SVCLAIM=D [3]	SO:0001743 copy_number_loss	EFO:0030069 complete genomic loss	complete genomic deletion (e.g. homozygous deletion on a bi-allelic genome region)

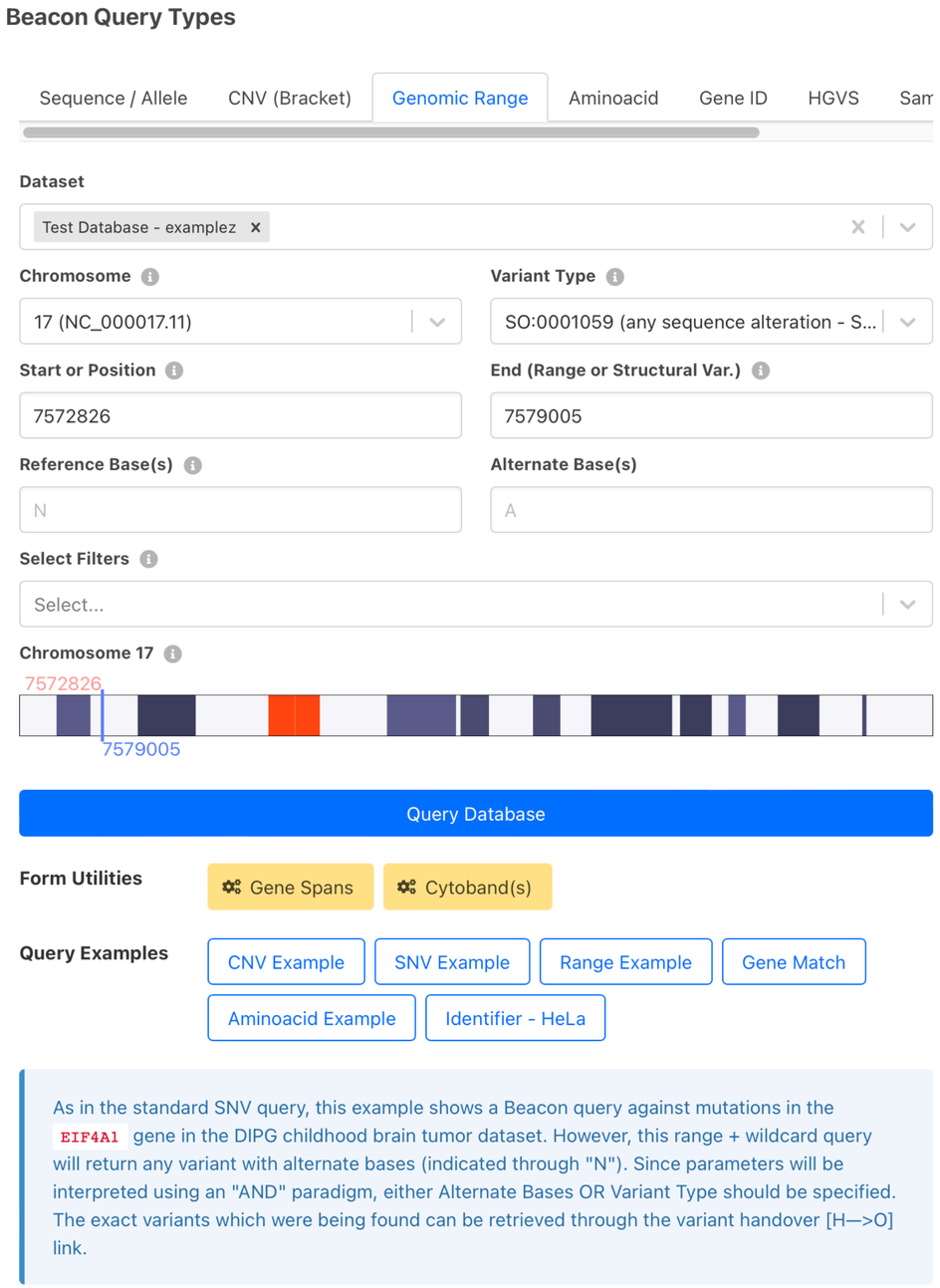
Beacon Queries

Range ("anything goes") Request

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



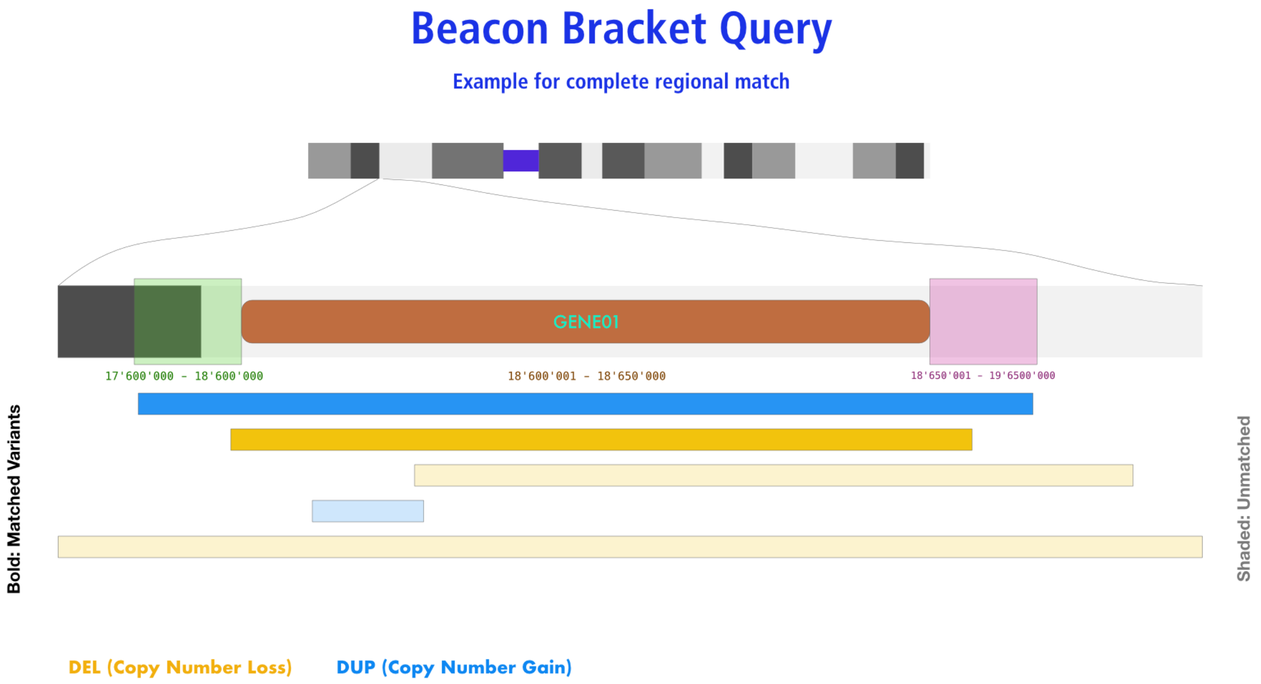
beaconplus.progenetix.org

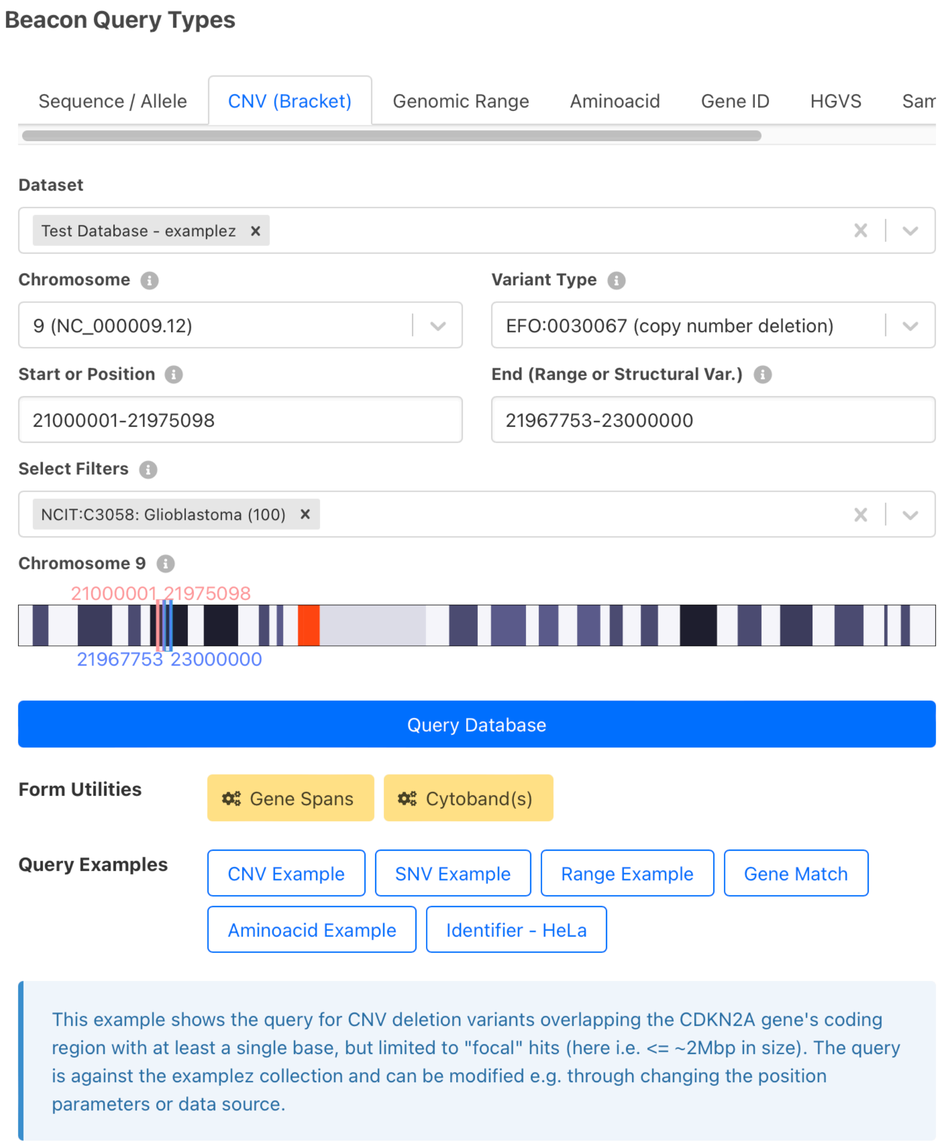


Beacon Queries

Bracket ("CNV") Query

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



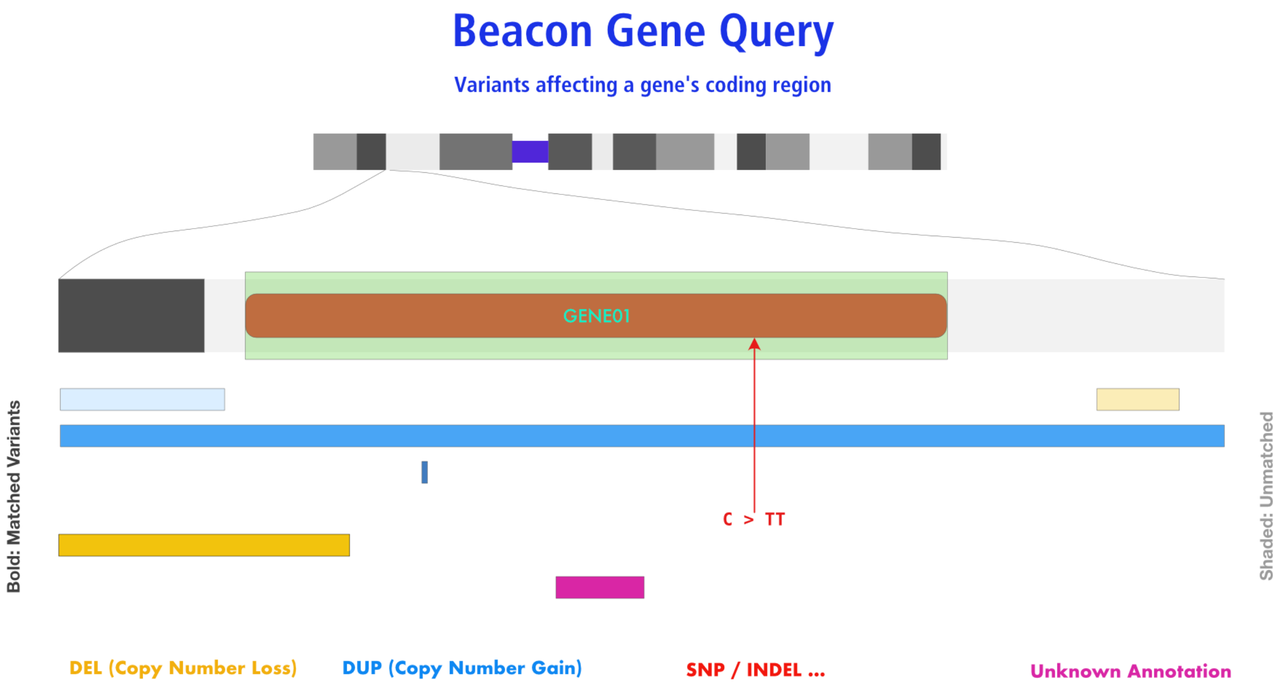


beaconplus.progenetix.org

Beacon Queries

Gene Request

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



beaconplus.progenetix.org

Beacon Query Types

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

Dataset

Cancer Cell Lines Collection ×

Gene Symbol

CDK2 (12:55966830-55972789)

Variant Type

Select...

Min Variant Length | **Max Variant Length** | **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

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

Bioconductor

README.md

pgxRpi

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

You can install this package from GitHub using:

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

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

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

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

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

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

pgxRpi

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

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

R wrapper for Progenetix

Bioconductor version: Development (3.19)

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

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

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

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

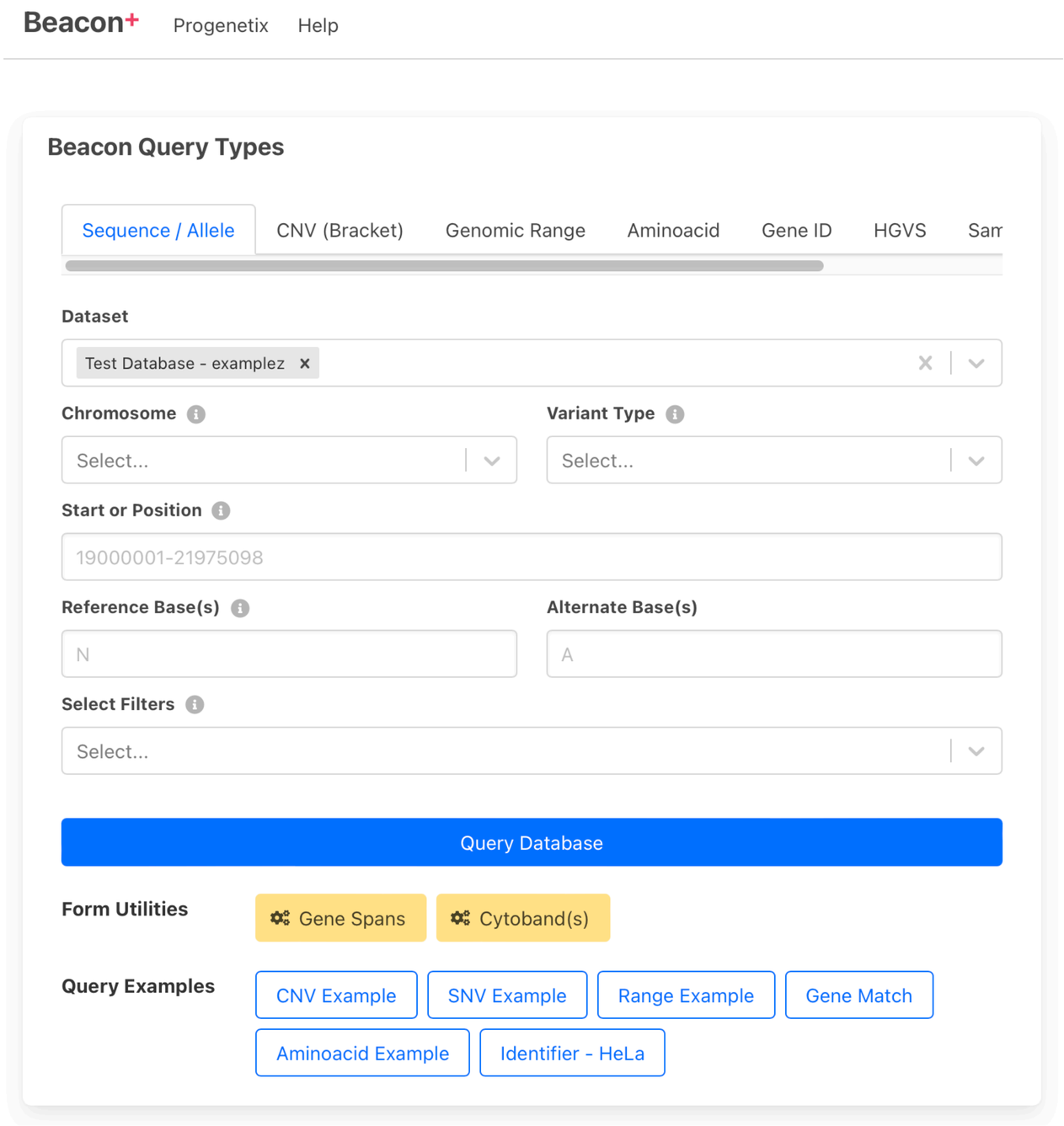
Zhao H, Baudis M (2023). *pgxRpi: R wrapper for Progenetix*. doi:10.18129/B9.bioc.pgxRpi, R package version 0.99.9, https://bioconductor.org/packages/pgxRpi.

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**

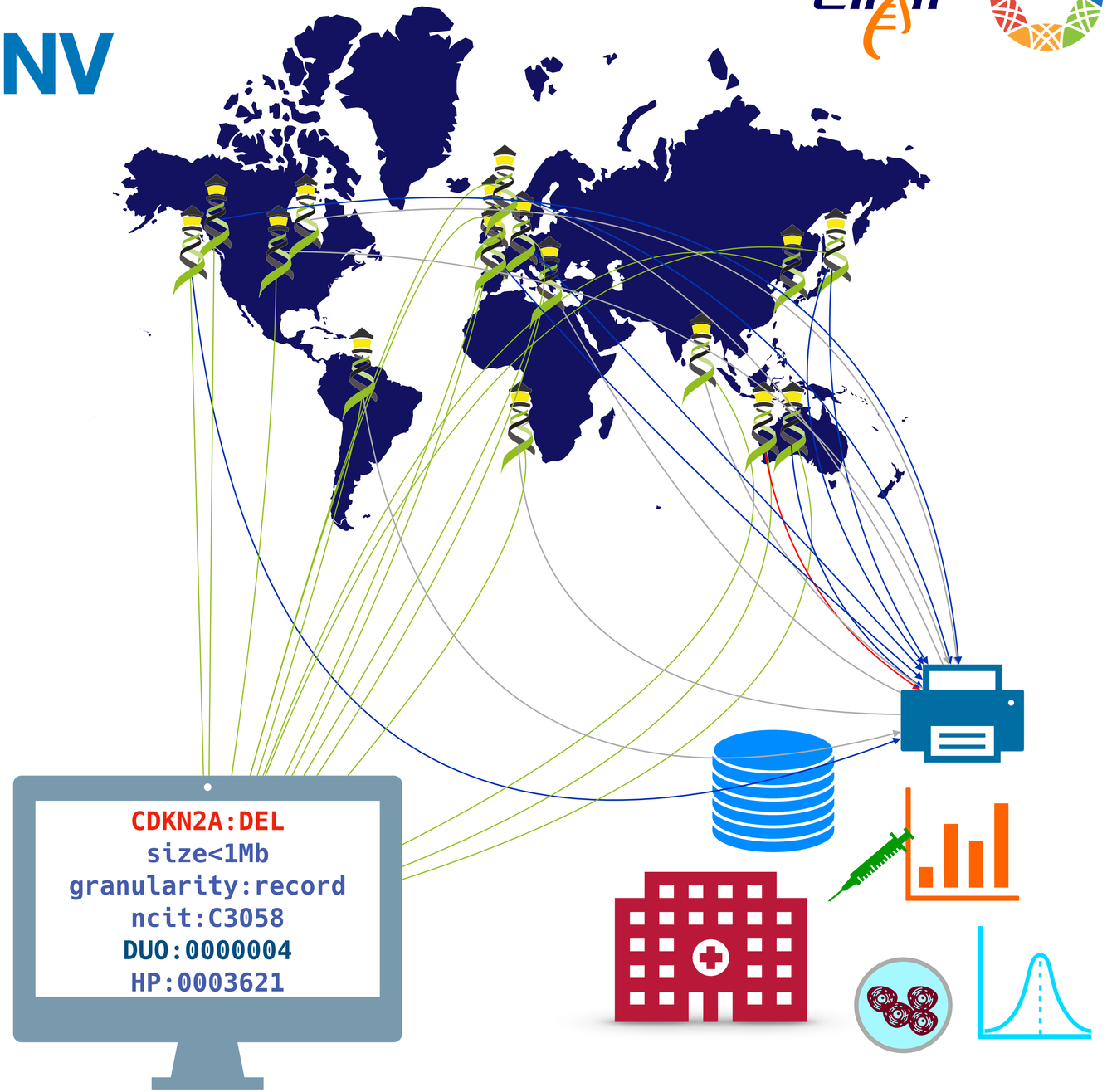
➡ **Beacon & hCNV Scout Team**



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

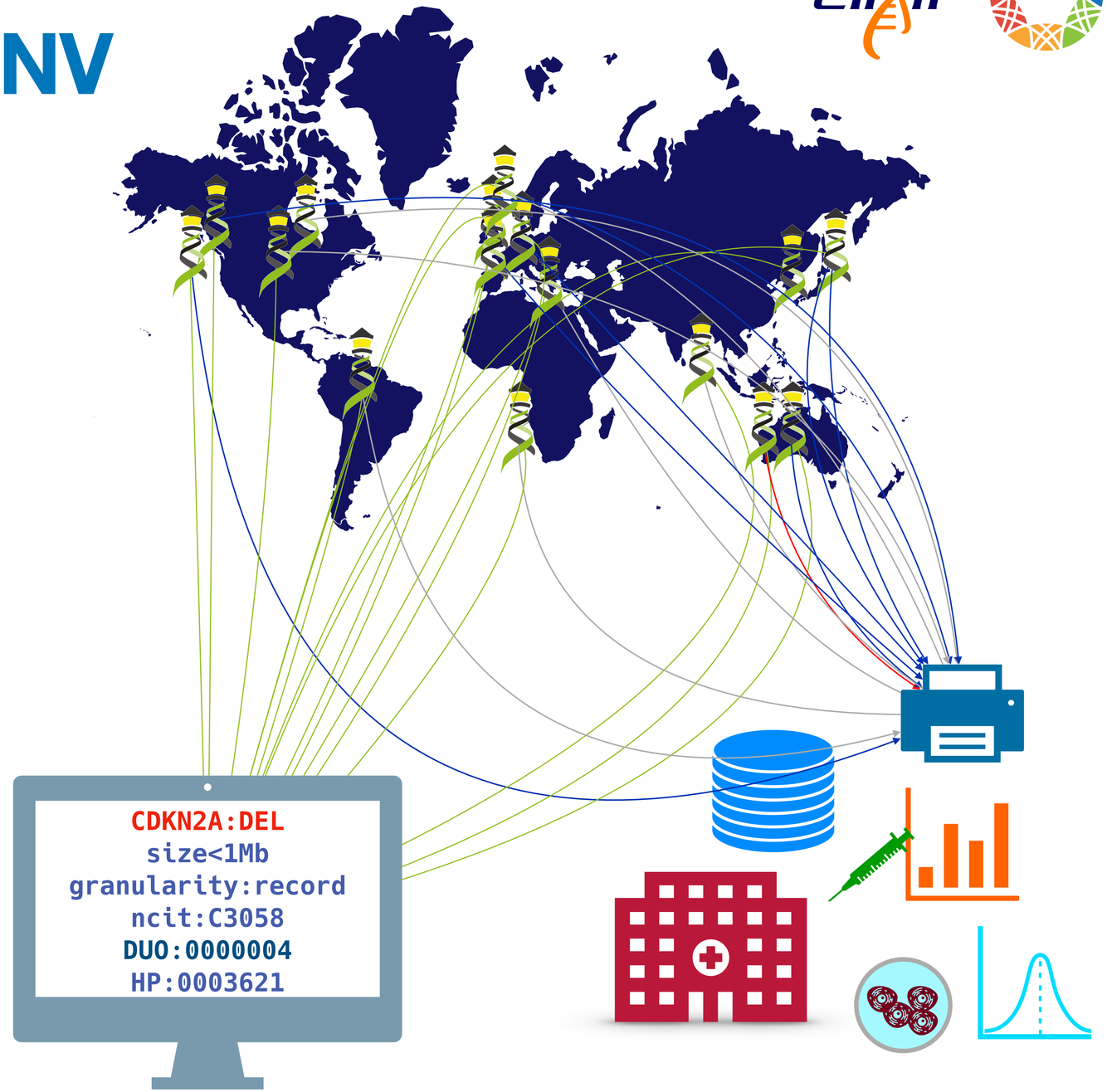
➡ **Collaborate w/ hCNV!**



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!**



ELIXIR hCNV Community

https://cnvar.org/