

Structural Genome Variations in Cancer and the Case for Open Data Standards



Michael Baudis

Professor of Bioinformatics
University of Zürich
Swiss Institute of Bioinformatics **SIB**
GA4GH Workstream Co-lead *DISCOVERY*
Co-lead ELIXIR Beacon API Development
Co-lead ELIXIR hCNV Community

Theoretical Cytogenetics and Oncogenomics

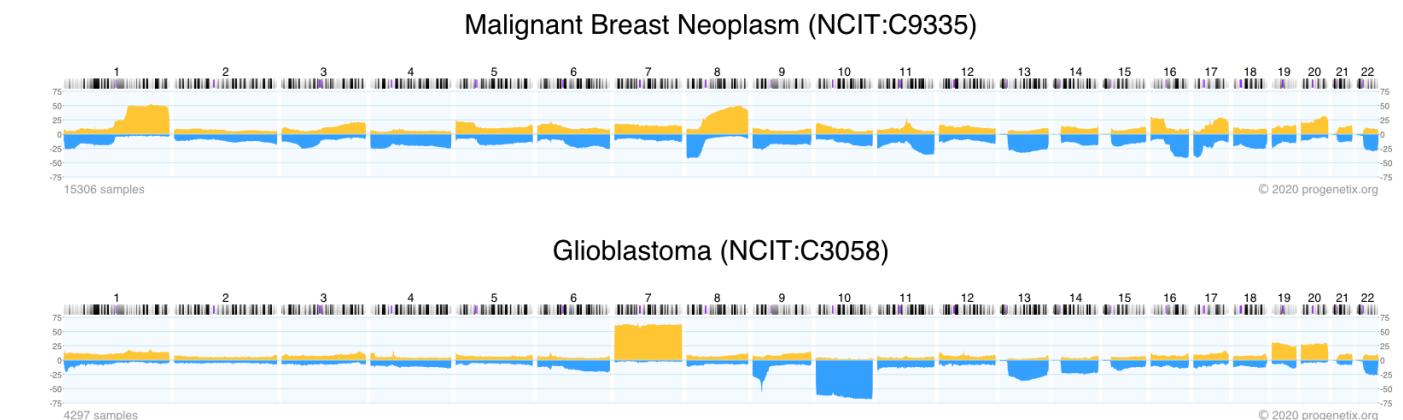
Cancer Genomics | Data Resources | Methods & Standards for Genomics and Personalized Health

Curators
~~Data Parasites~~

Theoretical Cytogenetics and Oncogenomics

... but what does this entail @baudisgroup?

- patterns & markers in cancer genomics, especially somatic structural genome variants
- bioinformatics support in collaborative studies
- reference resources for curated cancer genome variations
- bioinformatics tools & methods
- standards and reference implementations for data sharing in genomics and personalized health
- open research data "ambassadoring"



Bioinformatics & Bioinformaticians are ...



Bioinformatician

strong biological knowledge
provides hypothesis and / or dataset
sufficient statistical and
computational expertise to correctly
use bioinformatics tools & develop
workflows (scripting ...)

expert **user** of informatics tools
may get a Nobel

Bio**informatician**

sufficient biological background
provides statistical, analysis methods
sufficient biological or **medical**
background to understand problems
presented and identify pitfalls and hidden
biases arising from data generation

developer of informatics tools
may get rich

Bioinformatics & Bioinformaticians are ...



Bioinformatician

strong biological knowledge

provides hypothesis and / or dataset

sufficient statistical and
computational expertise to correctly
use bioinformatics tools & develop
workflows (scripting ...)

expert **user** of informatics tools

may get a Nobel

Bio**informatician**

sufficient biological background

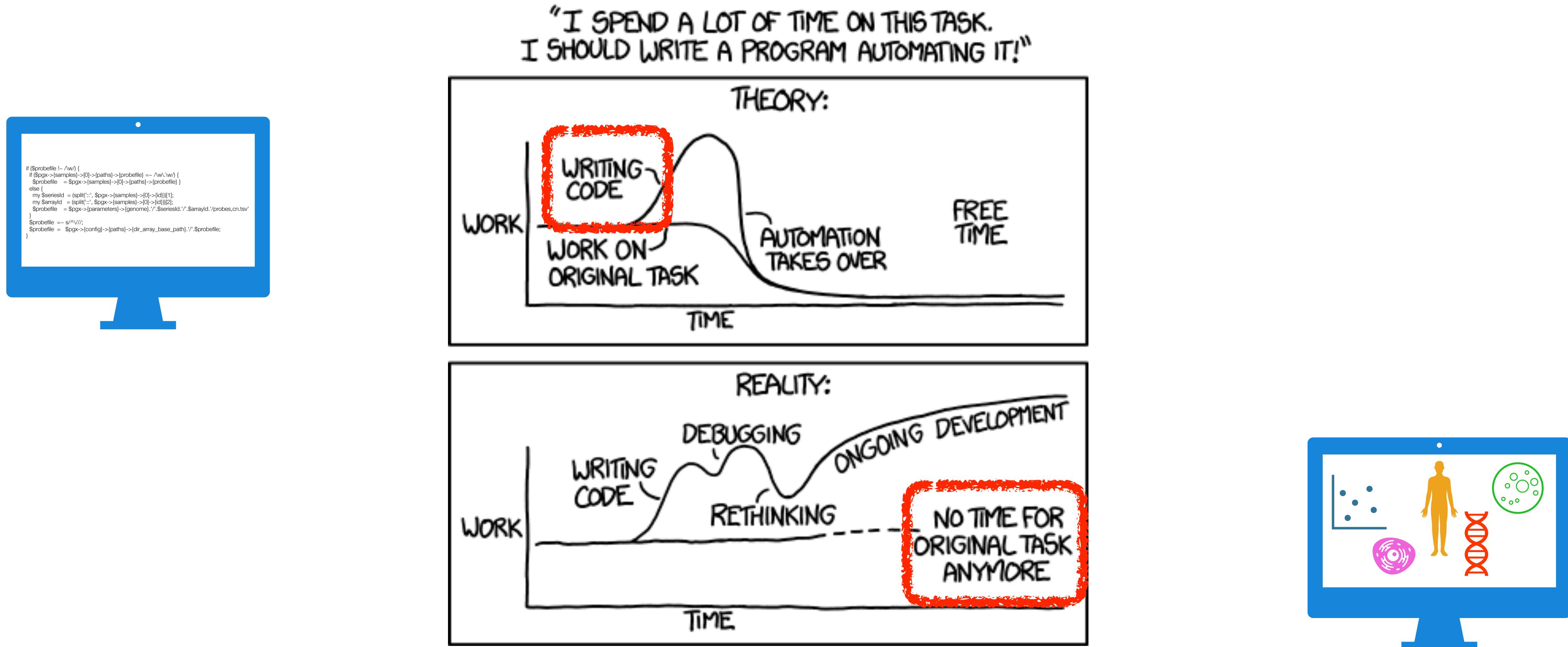
provides statistical, analysis methods

sufficient biological or **medical**
background to understand problems
presented and identify pitfalls and hidden
biases arising from data generation

developer of informatics tools

may get rich

{bio_informatics_science}

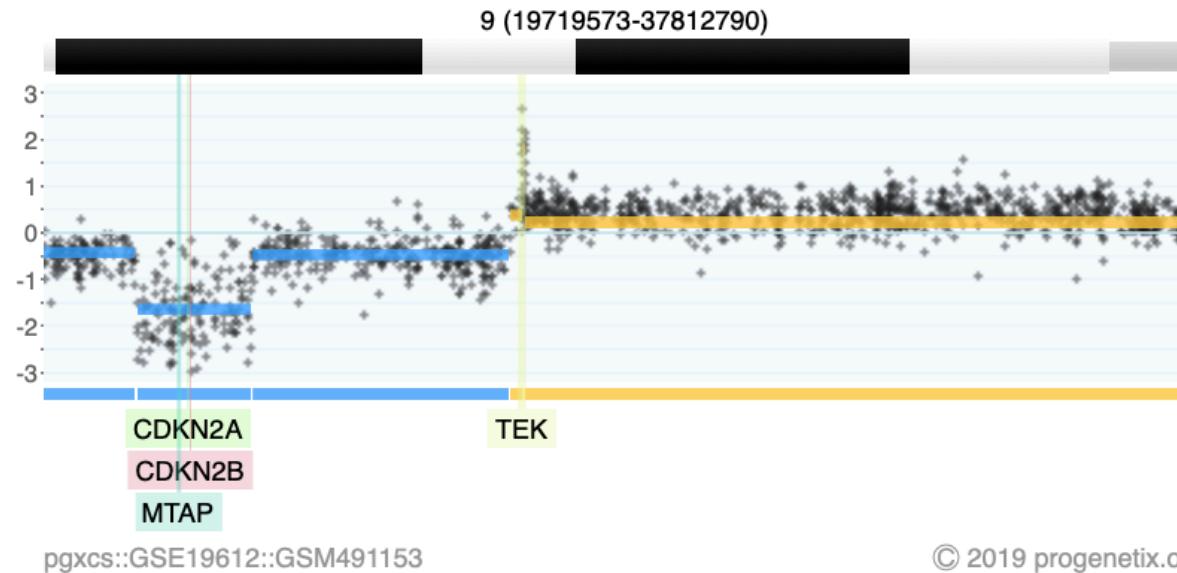


Theoretical Cytogenetics and Oncogenomics

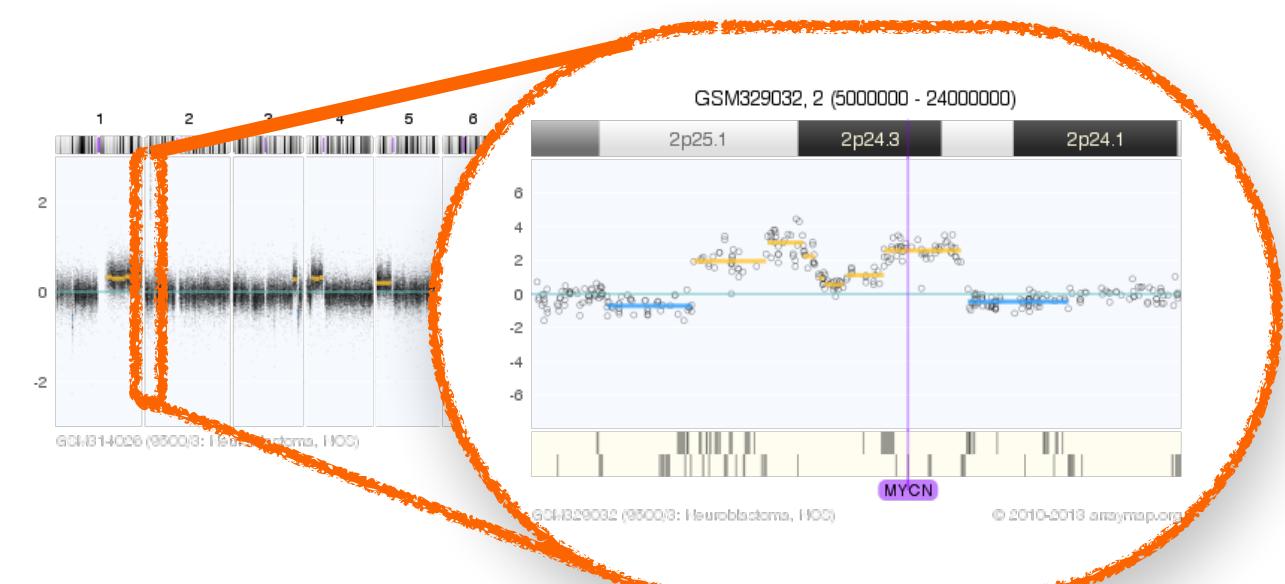
Research | Methods | Standards

Genomic Imbalances in Cancer - Copy Number Variations (CNV)

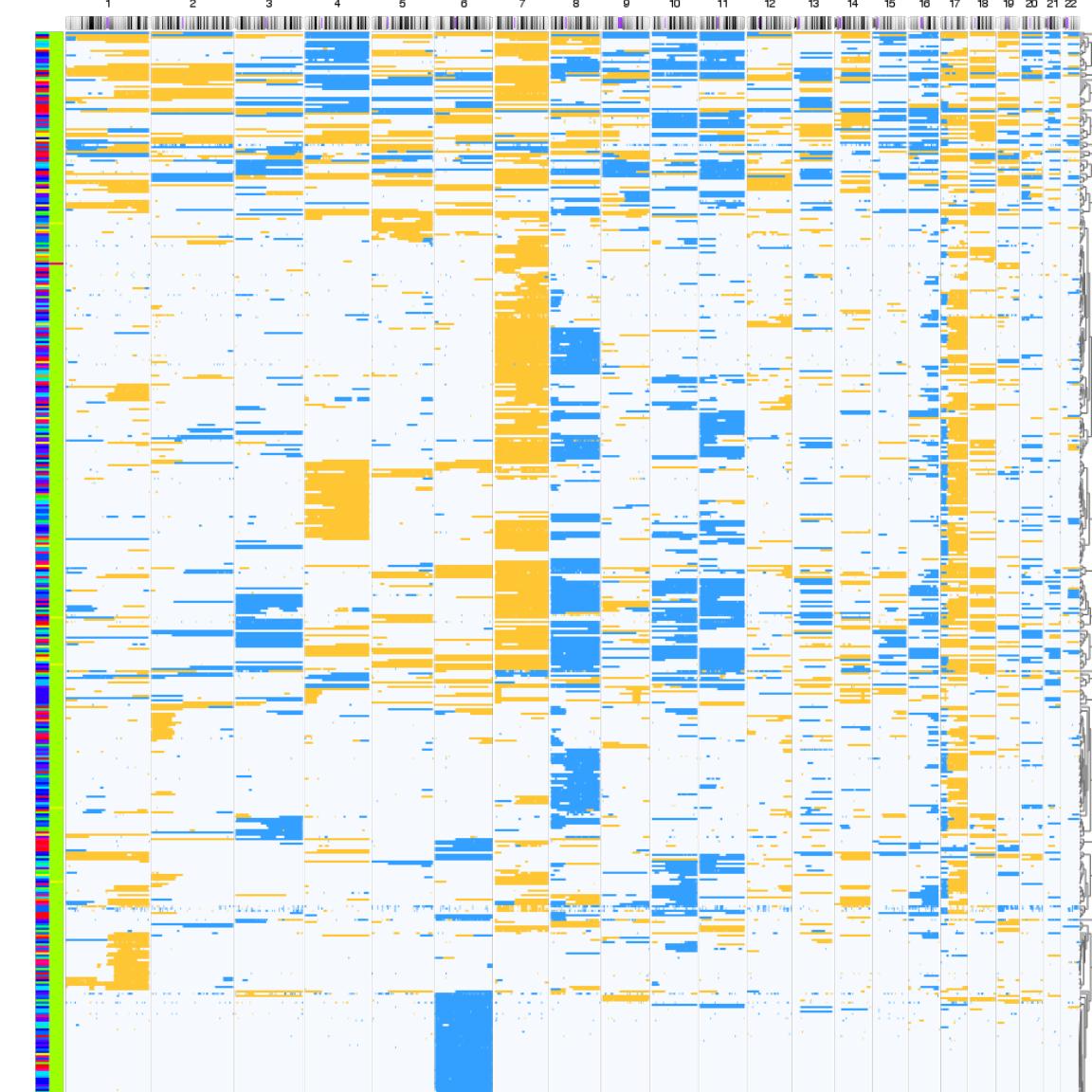
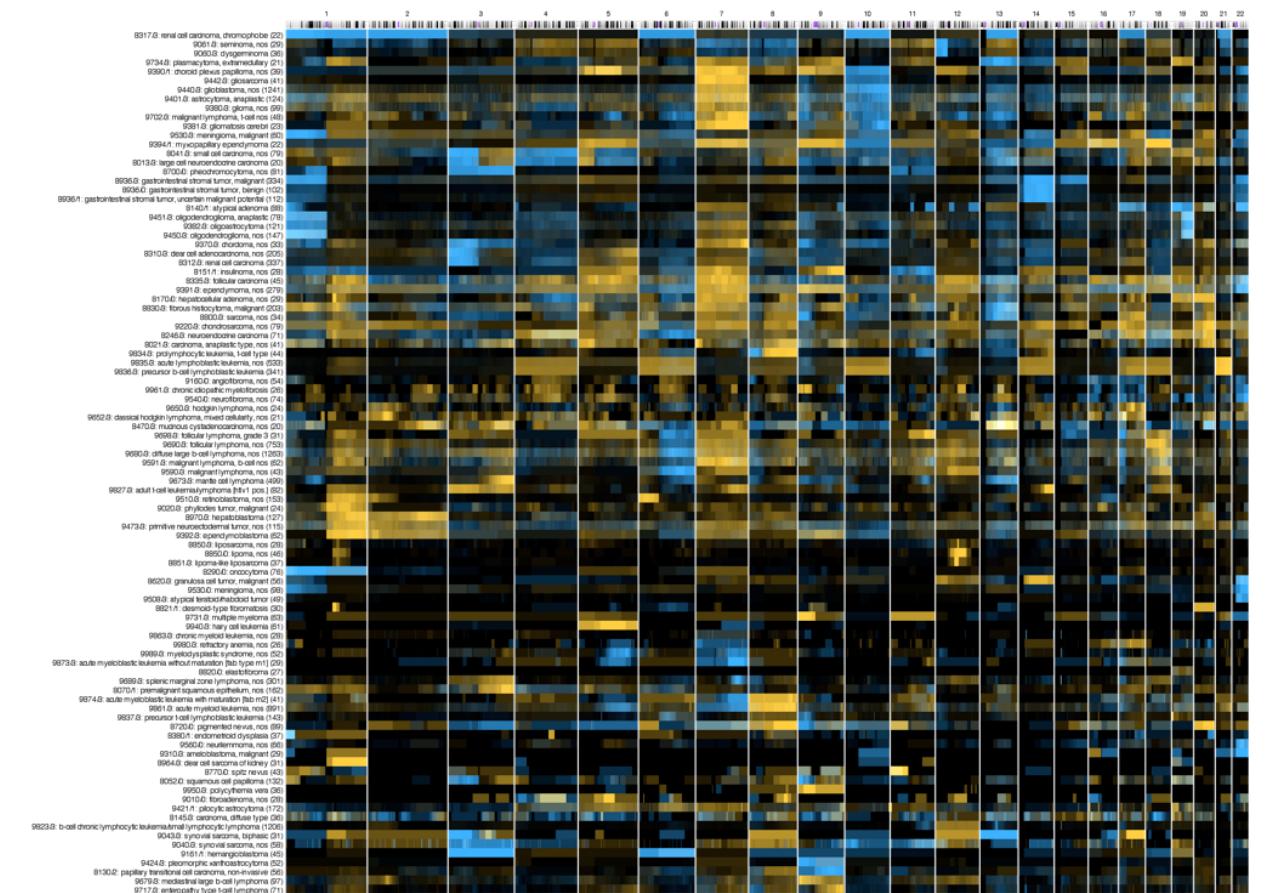
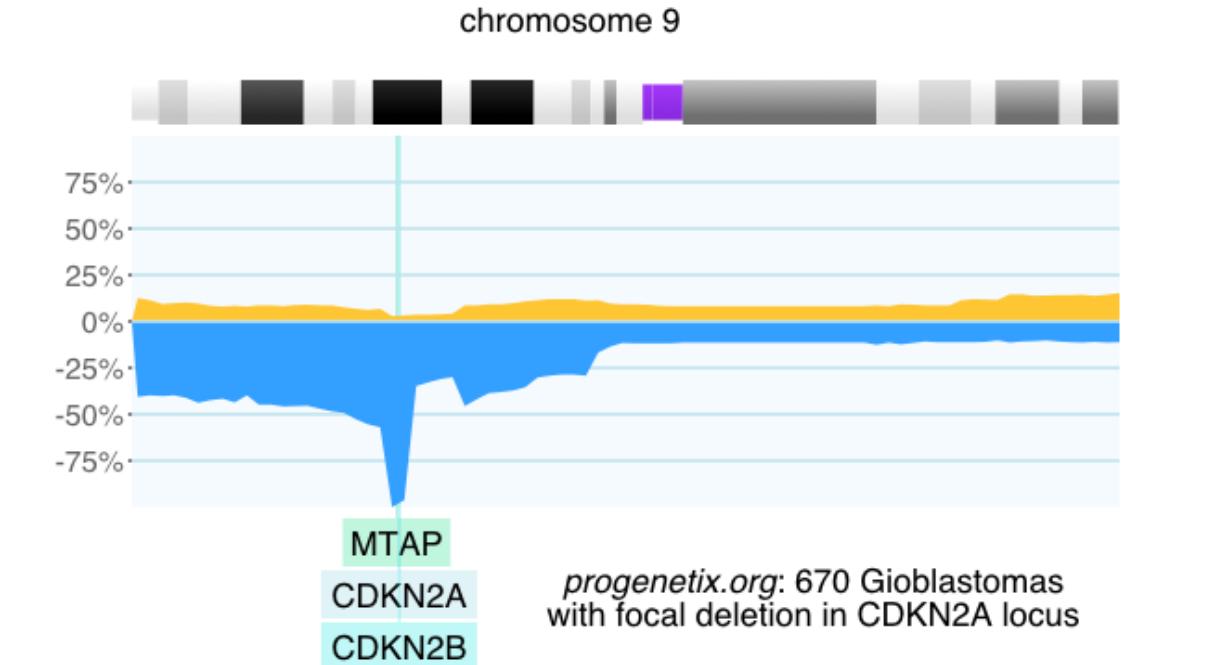
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)



2-event, homozygous deletion in a Glioblastoma

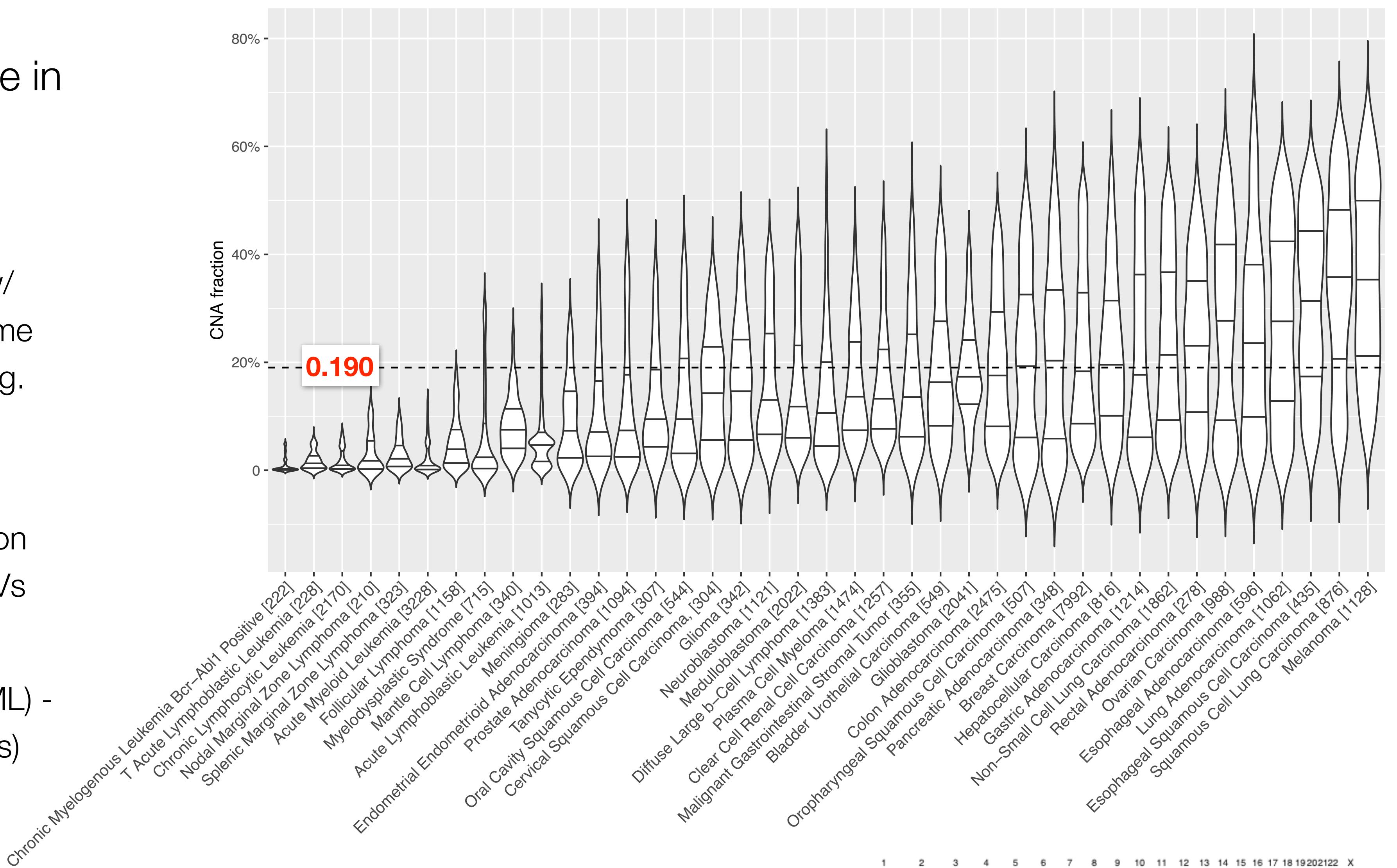


MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

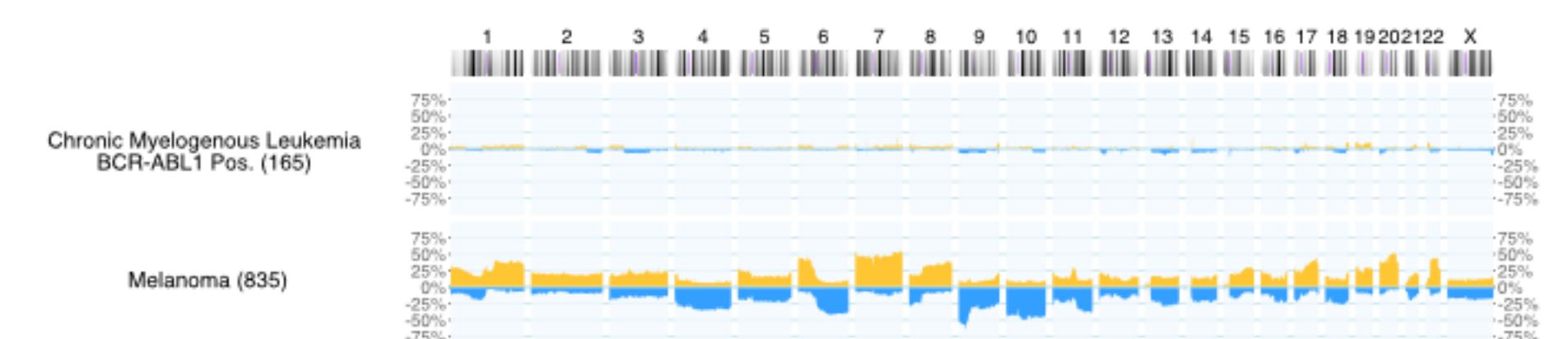


Genome CNV coverage in Cancer Classes

- 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 =>

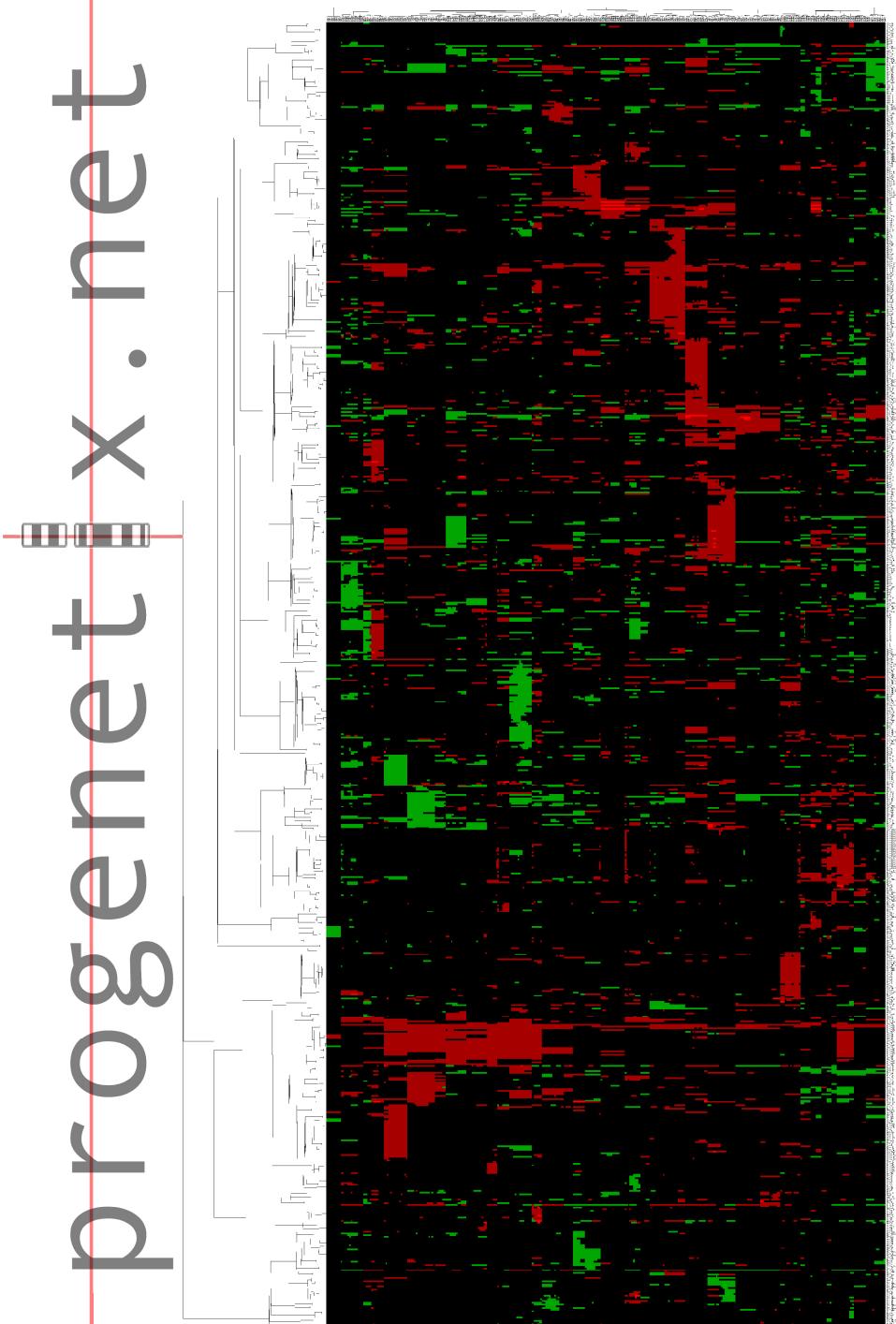


Over the last decade, techniques for the genome wide scanning for genomic imbalances in malignant neoplasia have been developed, e.g. Comparative Genomic Hybridization (CGH).

Currently, no comprehensive online source for CGH data with a standardized format suitable for data mining procedures has been made available for public access. Such a data repository could be valuable in identifying genetic aberration patterns with linkage to specific disease entities, and provide additional information for validating data from large scale expression array experiments.

A case and band specific aberration matrix was selected as most suitable format for the mining of CGH data. The [progenetix.net] data repository was developed to provide the according data to the research community for a growing number of human malignancies.

In the current implementation, two main purposes are being served. First, access to the band specific pattern of chromosomal imbalances allows the instantaneous identification of genomic "hotspots". Second, the band specific aberration matrices can be included in data mining efforts. As an example, the clustering off all informative cases from the current (September 2001) dataset is shown here (online source under www.progenetix.net/bcats/clustered.png).



Data selection

PubMed is searched for publications applying CGH to the analysis of malignant tumors. Articles are selected according to their online availability and the description of genomic imbalances on a per case basis.

Transformation of input data

Chromosomal aberration data is transformed via customized parsing commands to a common format adherent to ISCN 1995 recommendations. In some cases, aberration data was transcribed from graphical representations or provided by the authors.

Data storage

Currently, the primary data is stored in a dedicated "off-line" database. Besides case identifier and ISCN adapted chromosomal imbalance data, tumor classification and source information including the PubMed identifier is recorded. Disease entities are reclassified to ICD-O-3 codes.

Text parsing and generation of aberration matrix

For the generation of the case and band specific aberration matrix, a dedicated text pattern comparison model was developed using Perl. Briefly, for each chromosomal band, the aberration field of each case is searched for a variety of patterns containing aberration information applying to that band. A matrix with currently 324 band resolution is generated, annotating chromosomal gains with "1" and losses with "-1"; localized high-level gains are designated "2".

Website generation

For graphical representation of chromosomal imbalances, HTML pages containing different views of the underlying aberration matrices are generated using Perl. Graphics are implemented using HTML syntax. Besides band specific, whole genomic overviews, chromosome specific pages with links to all involved cases are generated for each ICD-O-3 entity as well as for each registered project. Additionally, those representations are available for several subsets combining related data (e.g. all lymphoid neoplasias, breast carcinoma cases). For each of the groups, the according aberration matrix is linked for download.

Hierarchical clustering of band specific chromosomal imbalances from 999 human neoplasias, contained in the [progenetix.net] collection. Cases without aberrations were excluded.

Progenetix.net: an online repository for molecular cytogenetic aberration data

Michael Baudis^{1,2,*} and Michael L. Cleary²

¹Medizinische Klinik und Poliklinik V der Universität Heidelberg, Germany

²Department of Pathology, Stanford University Medical Center, Stanford, CA 94305, USA

Received on July 5, 2001; revised on July 9, 2001; accepted on July 16, 2001



ABSTRACT

Summary: Through sequencing projects and, more recently, array-based expression analysis experiments, a wealth of genetic data has become accessible via online resources. In contrast, few of the (molecular-) cytogenetic aberration data collected in the last decades are available in a format suitable for data mining procedures. www.progenetix.net is a new online repository for previously published chromosomal aberration data, allowing the addition of band-specific information about chromosomal imbalances to oncologic data analysis efforts.

Availability: <http://www.progenetix.net>

Contact: mbaudis@stanford.edu

Neoplastic transformation and progression is the result of genetic defects arising in normal cells and giving rise to a malignant clone. During the process of oncogenesis, some of the usually multiple steps required for acquisition of the full neoplastic phenotype may represent themselves as numerical or structural abnormalities in the chromosomes of the transformed cells.

Over the last decades, the analysis of chromosomal abnormalities in malignant cells has gained importance in oncologic research as well as in clinical practice. A vast number of genetic abnormalities has been identified in the virtually complete range of human neoplasias. Several attempts have been undertaken for collection and classification of those abnormalities, the most widely recognized being the catalog by Mitelman and co-workers (Mitelman, 1994; online access through <http://cgap.nci.nih.gov/Chromosomes/Mitelman>).

In addition to metaphase analysis of short-term cultivated tumor cells or tumor cell lines, molecular cytogenetic techniques have recently been applied to the analysis of chromosomal abnormalities in primary tumor tissues. One of the more widely used screening techniques is Comparative Genomic Hybridization (CGH; Kallion-

iem et al., 1992; du Manoir et al., 1993). Briefly, this method is based on the competitive *in-situ* hybridization of differentially labeled tumor versus normal genomic DNA to normal human metaphase spreads. The calculation of the intensity ratios of the two fluorochromes gives an overview about relative gains and losses of DNA in the tumor genome with mapping to the respective chromosomal bands. The identification of frequently imbalanced regions in tumor entities may point towards tumor suppressor gene or proto-oncogenes mapping to the respective chromosomal bands. Usually, the result of those experiments is communicated either in text format according to the International System for Cytogenetic Nomenclature (Mitelman, 1995) or graphically, with aberration bars next to chromosomal ideograms for the representation of chromosomal gains and losses.

Because in each experiment CGH analysis covers the whole number of chromosomes, the comparision of data sets from related malignancies could lead to the delineation of common as well as divergent genetic pathways defining the respective malignant phenotypes. Although an extremely large number of malignant tumors has been analyzed using this technique, no comprehensive CGH database with band-specific chromosomal aberration information is publicly available[†].

A minimal requirement for such a database would be the conversion of the text or graphical information used in publications to data tables, representing the information about the aberration status of single chromosomal bands for each case. For the site discussed here, this process includes: (1) the transformation of the published results in a format adapted from the ISCN, and (2) the automatic generation of the band specific aberration table.

Due to format variations of the published data, step 1 consists of the manual conversion of the text data or evaluation and conversion of the graphical representations, respectively. Due to the (in computational terms) odd

[†]Links to a number of online CGH resources with different scopes can be found at www.progenetix.net.

*To whom correspondence should be addressed.

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⁺

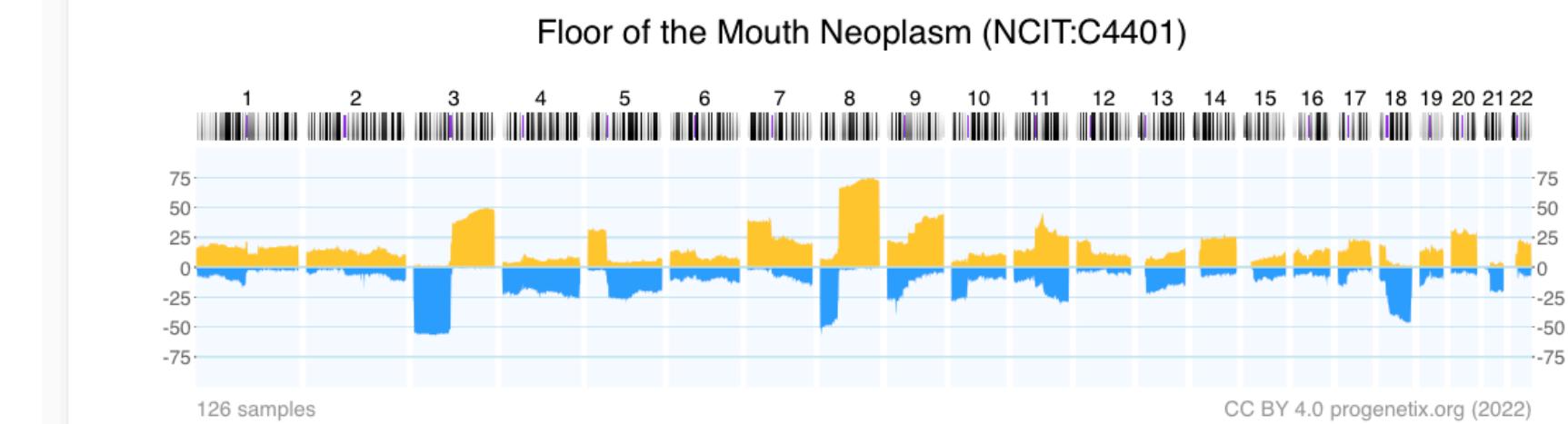
Documentation

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.



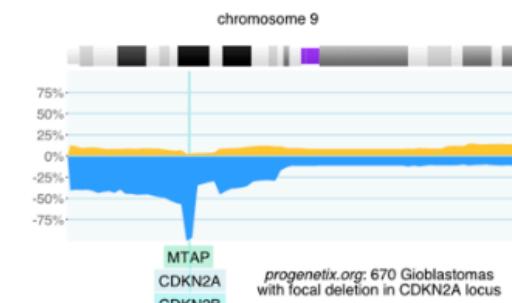
[Download SVG](#) | [Go to NCIT:C4401](#) | [Download CNV Frequencies](#)

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

Cancer Types by National Cancer Institute NCI 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.



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 NCI 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: I
- NCIT:C326
- NCIT:C000
- NCIT:C474
 - NCIT:C2
 - NCIT:C3
 - NCIT:C3
 - NCIT:C3
 - N
- NCIT:C3058: Glioblastoma (NCIT:C3058)
 - Sample Counts
 - 4370 samples
 - 4286 direct NCIT:C3058 code matches
 - 4384 CNV analyses
 - Search Samples
 - Select NCIT:C3058 samples in the [Search Form](#)
 - Raw Data (click to show/hide)
- 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)

Glioblastoma (NCIT:C3058)

Download SVG | Go to NCIT:C3058 | Download CNV Frequencies

© CC-BY 2001 - 2023 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. 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

Search Samples

CDKN2A Deletion Example MYC Duplication TP53 Del. in Cell Lines

K-562 Cell Line

Gene Spans

Cytoband(s)

This example shows the query for CNV deletion variants overlapping the CDKN2A gene's coding region with at least a single base, but limited to "highly focal" hits (here i.e. <= ~1Mbp in size). The query can be modified e.g. through changing the position parameters or diagnosis.

Dataset

Progenetix

Gene Symbol

Select...

Chromosome

NC_000009.12

Variant Type

EFO:0030067 (copy number deletion)

Start or Position

21500001-21975098

End (Range or Structural Var.)

21967753-22500000

Minimum Variant Length

Maximal Variant Length

Reference ID(s)

Select...

Cohorts

Cancer Classification(s)

NCIT:C3058: Glioblastoma (4...

Clinical Classes

Select...

Genotypic Sex

Select...

Biosample Type

Select...

Filters

Filter Logic

AND

Include Child Terms

Select...

Response Limit / Page Size

1000

Skip Pages

0

City

Select...

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) ↗

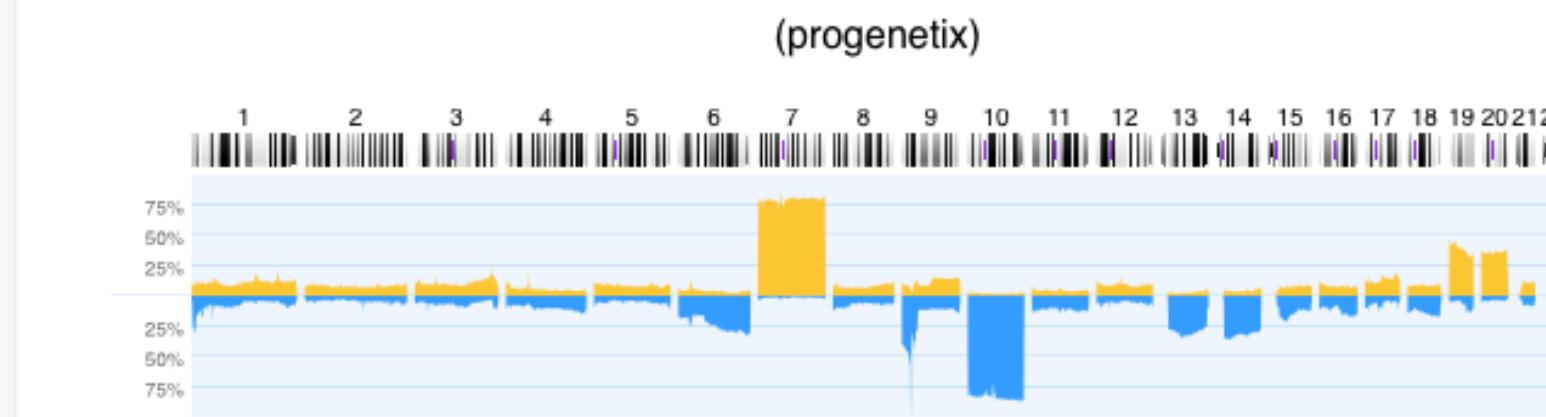
Visualization options

Results

Biosamples

Biosamples Map

Variants



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 ↗

Ontologies and Classifications



Services: Ontologymaps (NCIt)

The **ontologymaps** service provides equivalency mapping between ICD-O and other classification systems, notably NCIt and UBERON. It makes use of the sample-level mappings for NCIT and ICD-O 3 codes developed for the individual samples in the Progenetix collection.

NCIT and ICD-O 3

While NCIT treats diseases as **histologic** and **topographic** described entities (e.g. [NCIT:C7700: Ovarian adenocarcinoma](#)), these two components are represented separately in ICD-O, through the **Morphology** and **Topography** coding arms (e.g. here [8140/3 + C56.9](#)).

More documentation with focus on the API functionality can be found on the [documentation pages](#).

The data of all mappings can be retrieved trough this API call: [{JSON ↗}](#)

Code Selection ⓘ

NCIT:C4337: Mantle Cell Lymphoma

Optional: Limit with second selection

Matching Code Mappings [{JSON ↗}](#)

NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C77.9: Lymph nodes, NOS
NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C18.9: large intestine, excl. rectum and rectosigmoid junction
NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C42.2: Spleen

More than one code groups means that either mappings need refinements (e.g. additional specific NCIT classes for ICD-O T topographies) or you started out with an unspecific ICD-O M class and need to add a second selection.

In Progenetix all cancer diagnoses are coded to both NCIt neoplasm codes and ICD-O 3 Morphology + Topography combinations. The matched mappings are provided as lookup-service since neither an official ICD-O ontology nor such a "disease defined by ICD-O M+T" concept is codified anywhere.

List of filters recognized by different query endpoints

Public Ontologies with CURIE-based syntax

CURIE prefix	Code/Ontology	Examples
NCIT	NCIt Neoplasm ¹	NCIT:C27676
HP	HPO ²	HP:0012209
PMID	NCBI Pubmed ID	PMID:18810378
geo	NCBI Gene Expression Omnibus ³	geo:GPL6801, geo:GSE19399, geo:GSM491153
arrayexpress	EBI ArrayExpress ⁴	arrayexpress:E-MEXP-1008
cellosaurus	Cellosaurus - a knowledge resource on cell lines ⁵	cellosaurus:CVCL_1650
UBERON	Uberon Anatomical Ontology ⁶	UBERON:0000992
cBioPortal	cBioPortal ⁹	cBioPortal:msk_impact_2017

Private filters

Since some classifications cannot directly be referenced, and in accordance with the upcoming Beacon v2 concept of "private filters", Progenetix uses additionally a set of structured non-CURIE identifiers.

For terms with a `pgx` prefix, the [identifiers.org resolver](#) will

Filter prefix / local part	Code/Ontology	Example
pgx:icdom...	ICD-O 3 ⁷ Morphologies (Progenetix)	pgx:icdom-81703
pgx:icdot...	ICD-O 3 ⁷ Topographies(Progenetix)	pgx:icdot-C04.9
TCGA	The Cancer Genome Atlas (Progenetix) ⁸	TCGA-000002fc-53a0-420e-b2aa-a40a358bba37
pgx:pgxcohort...	Progenetix cohorts ¹⁰	pgx:pgxcohort-arraymap

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



Swiss Institute of
Bioinformatics



Cancer CNV Profiles

Search Samples

Studies & Cohorts

arrayMap

TCGA Samples

DIPG Samples

Gao & Baudis, 2021

Cancer Cell Lines

Publication DB

Genome Profiling

Progenetix Use

Services

NCIt Mappings

UBERON Mappings

Upload & Plot

Download Data

Beacon⁺

Progenetix Info

About Progenetix

Progenetix Publication Collection

The current page lists articles describing whole genome screening (WGS, WES, aCGH, cCGH) experiments in cancer, registered in the Progenetix publication collection. For each publication the table indicates the numbers of samples analysed with a given technology and if sample profiles are available in Progenetix.

Please [contact us](#) to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation [↗](#).

New Oct 2021 You can now directly submit suggestions for matching publications to the [oncopubs](#) repository on [Github](#) [↗](#).

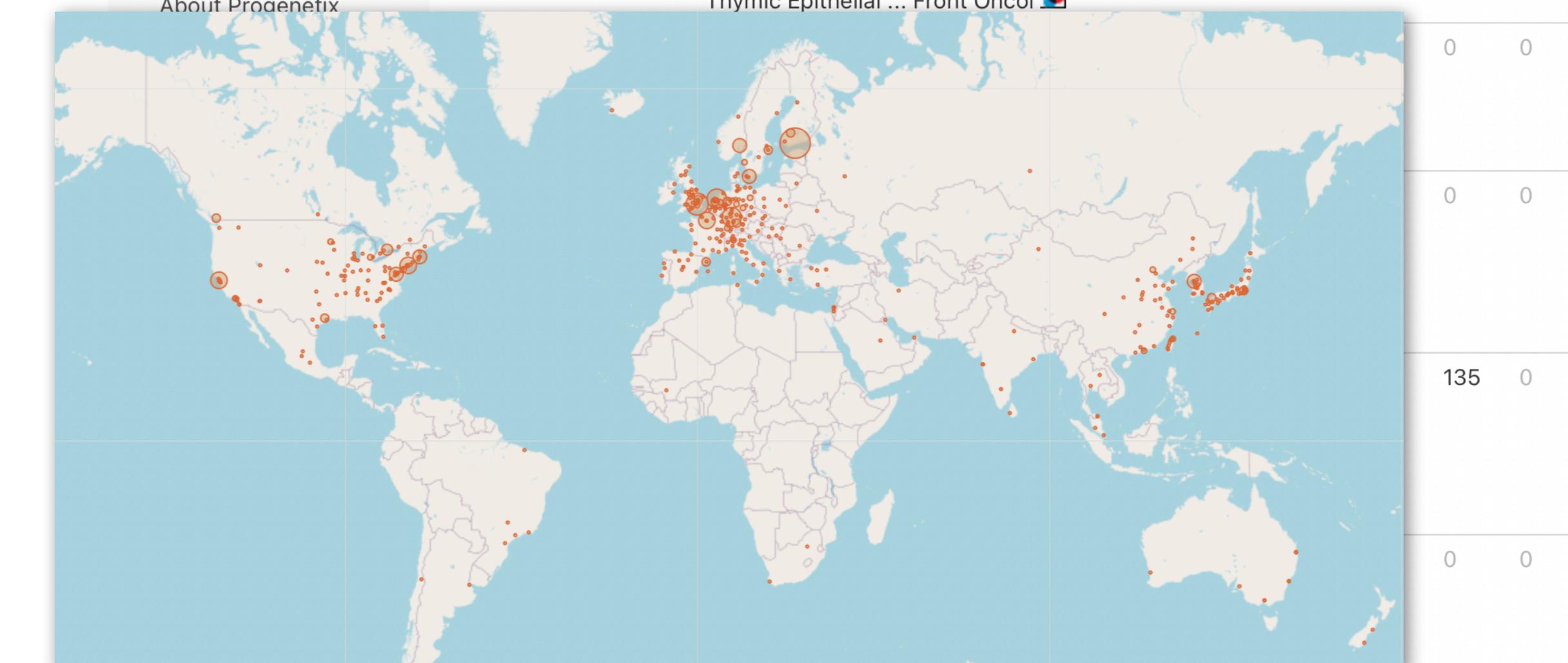
Filter [i](#)

City [i](#)

 Type to search... | [▼](#)

Publications (3349)

id i ▾	Publication	Samples				
		cCGH	aCGH	WES	WGS	pgx
PMID:34604048	Dai J, Jiang M, He K, Wang H, Chen P et al. (2021) DNA Damage Response and Repair Gene Alterations Increase Tumor Mutational Burden and ... <i>Front Oncol</i>	0	0	122	0	0
PMID:34573430	Juhari WKW, Ahmad Amin Noordin KB et al. (2021) Whole-Genome Profiles of Malay Colorectal Cancer Patients with Intact MMR Proteins. ... <i>Genes (Basel)</i>	0	0	0	7	0
PMID:34307137	Xu S, Li X, Zhang H, Zu L, Yang L et al. (2021) Frequent Genetic Alterations and Their Clinical Significance in Patients With Thymic Epithelial ... <i>Front Oncol</i>	0	0	0	123	0



Cancer Cell Lines

Cancer Genomics Reference Resource

- starting from >5000 cell line CNV profiles
 - 5754 samples | 2163 cell lines
 - 256 different NCIT codes
- genomic mapping of annotated variants and additional data from several resources (ClinVar, CCLE, Cellosaurus...)
 - 16178 cell lines
 - 400 different NCIT codes
- query and data delivery through Beacon v2 API

→ integration in data federation approaches

cancercelllines.org

Lead: Rahel Paloots



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?].

The sidebar includes links for Cancer Cell Lines, Search Cell Lines, Cell Line Listing, CNV Profiles by Cancer Type, Documentation, News, and Progenetix (which is currently selected). The main content area features a logo with three pink circles and the text "cancercelllines".

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 as a daughter cell line of HeLa ([CVCL_0030](#)) and so forth.

Sample selection follows a hierarchical system in which sample selection is done at the level of the parent cell line. For example, a search for HeLa will also return the daughter lines by default - but one can also search for a specific daughter line.

Cell Lines (with parental/derived hierarchies)

Filter subsets e.g. by prefix		Hierarchy Depth
No Selection		
<input type="checkbox"/>	> cellosaurus:CVCL_0312: HOS	(204 samples)
<input type="checkbox"/>	> cellosaurus:CVCL_1575: NCI-H650	(6 samples)
<input type="checkbox"/>	> cellosaurus:CVCL_1783: UM-UC-3	(9 samples)
<input type="checkbox"/>	▼ cellosaurus:CVCL_0004: K-562	(28 samples)
<input type="checkbox"/>	cellosaurus:CVCL_3827: K562/Ad	(1 sample)
<input type="checkbox"/>	> cellosaurus:CVCL_0589: Kasumi-1	(9 samples)

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
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 Line Details

HOS (cellosaurus:CVCL_0312)

Subset Type

- Cellosaurus - a knowledge resource on cell lines [cellosaurus:CVCL_0312](#)

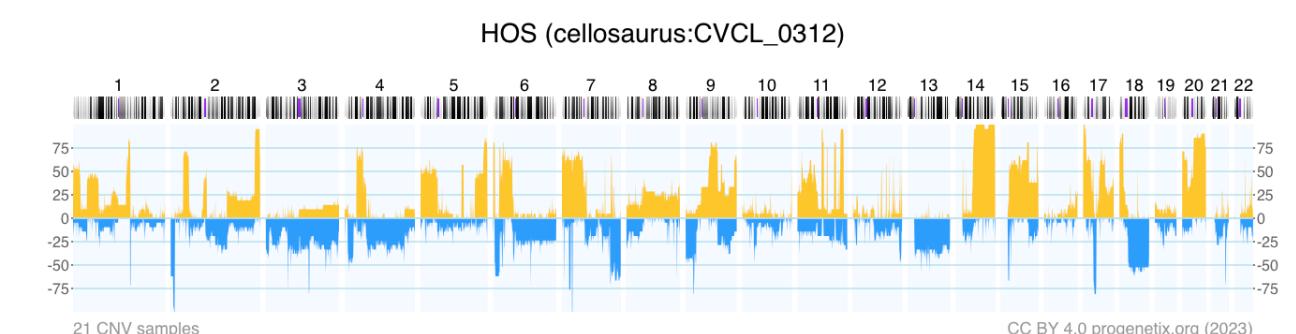
Sample Counts

- 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)



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

Gene Matches	Cytoband Matches	Variants
ALK	. ABC-14 cells harbored no ALK mutations and were sensitive to ... crizotinib while also exhibiting MNNG HOS transforming gene (MET)	Rapid Acquisition of Alectinib Resistance in ALK-Positive Lung Cancer With High Tumor Mutation Burden (31374369)
AREG	crizotinib while also exhibiting MNNG HOS	Rapid Acquisition of Alectinib Resistance

{Bio|informatics}Science}

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

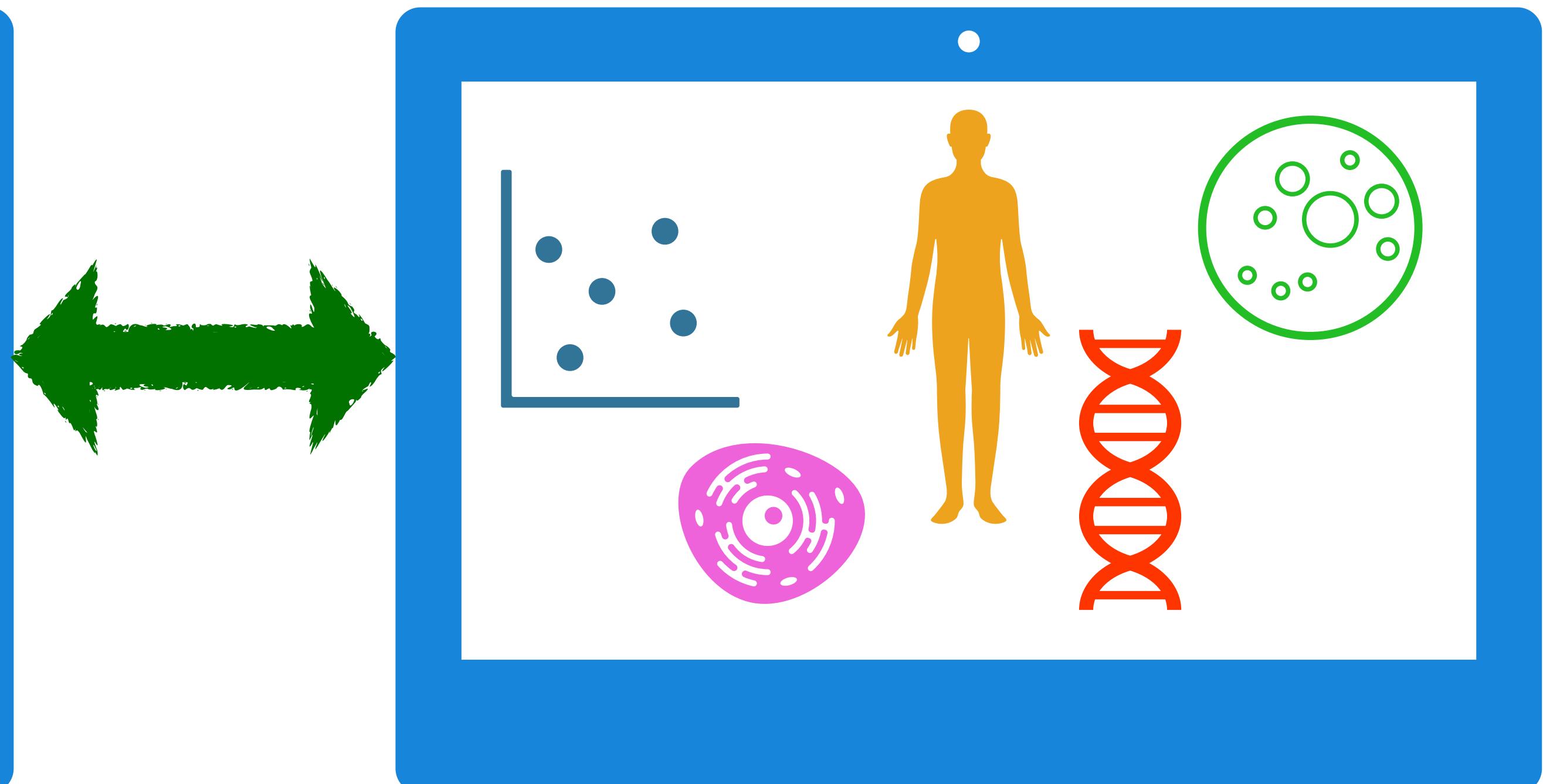
    covs = np.zeros((cs_no, int_no))
    vals = np.zeros((cs_no, int_no))

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

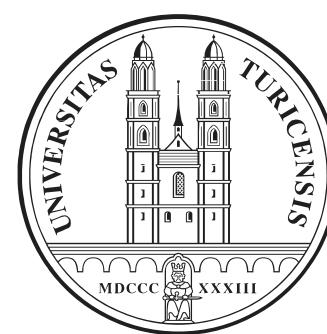
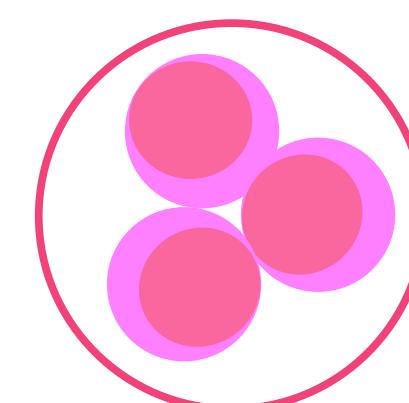
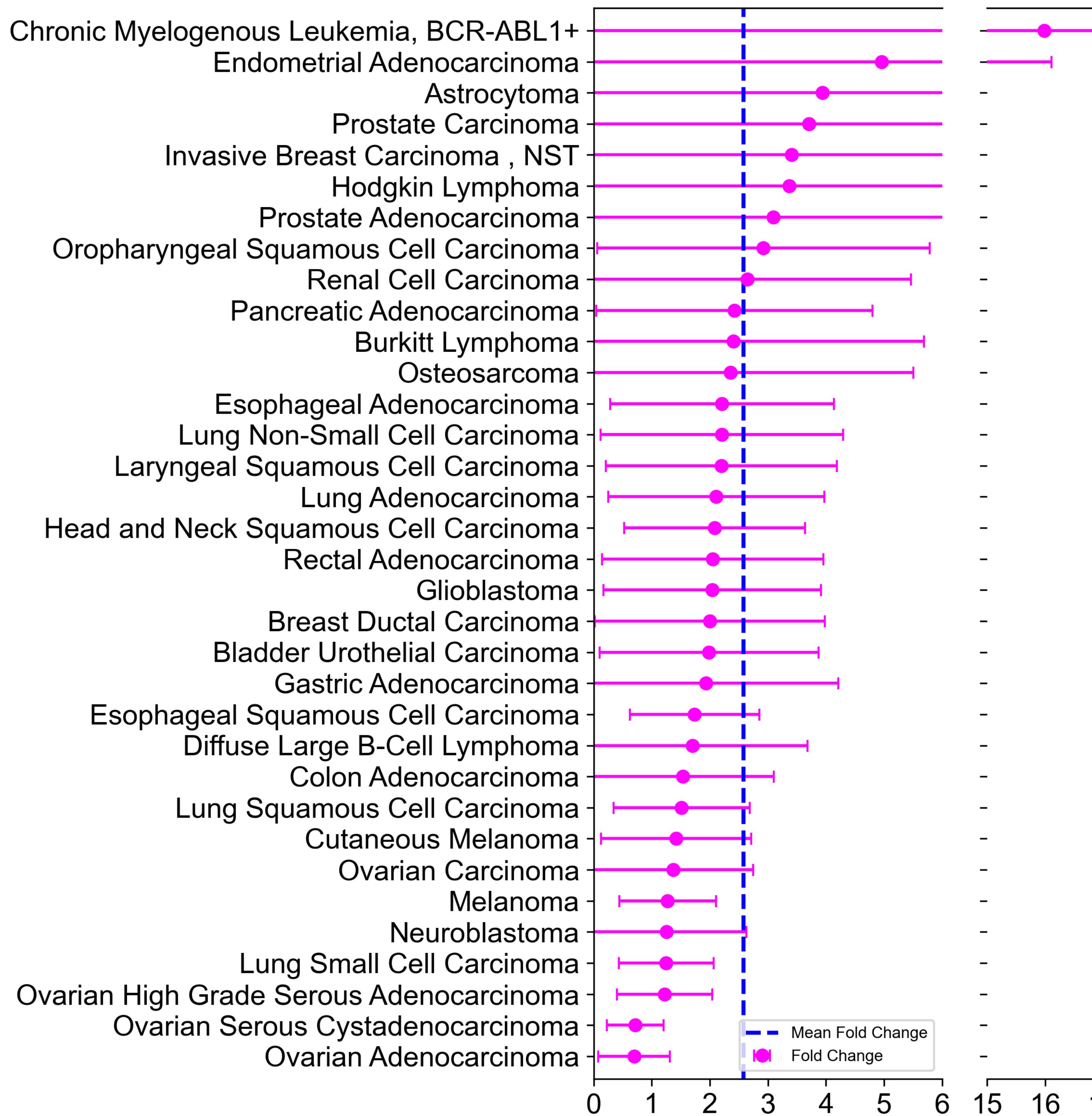
    for i, cs in enumerate(callsets):
        covs[i] = cs["cnv_statusmaps"][pars[t]["cov_l"]]
        vals[i] = cs["cnv_statusmaps"][pars[t]["val_l"]]

    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)

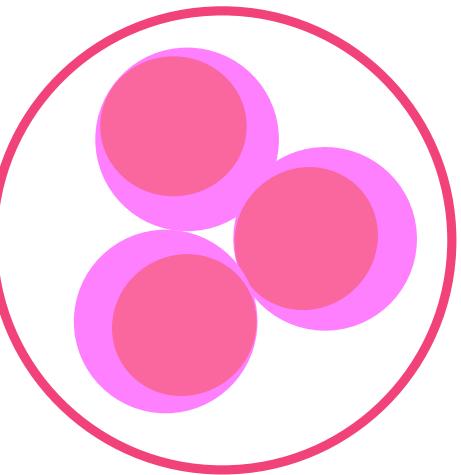
    for i, interval in enumerate(int_fs):
        int_fs[i].update({
            t + "_frequency": frequencies[i],
            t + "_median": medians[i],
            t + "_mean": means[i]
        })
```



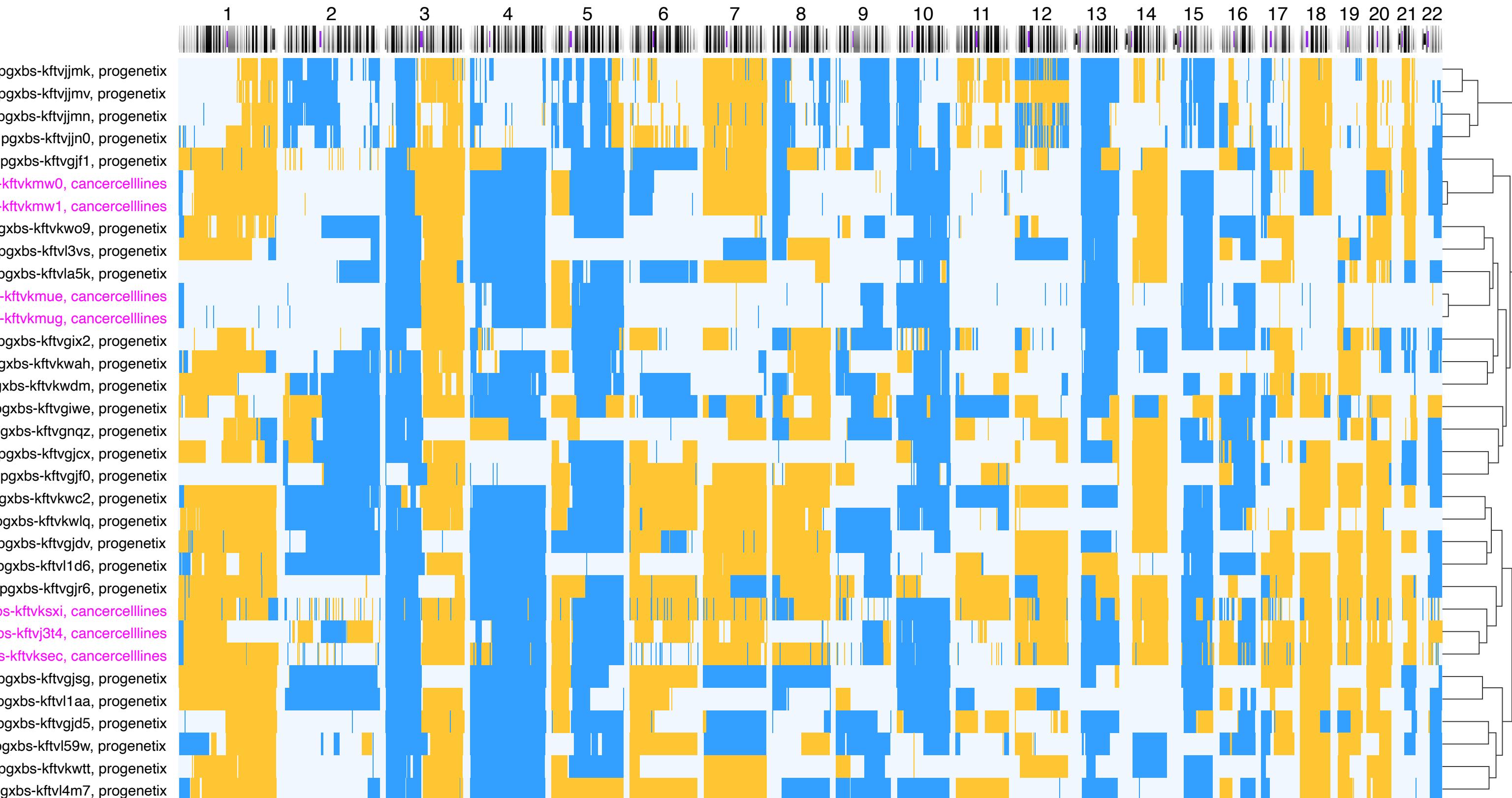
Higher level of CNV coverage of the genomes of cancer cell lines compared to their origins



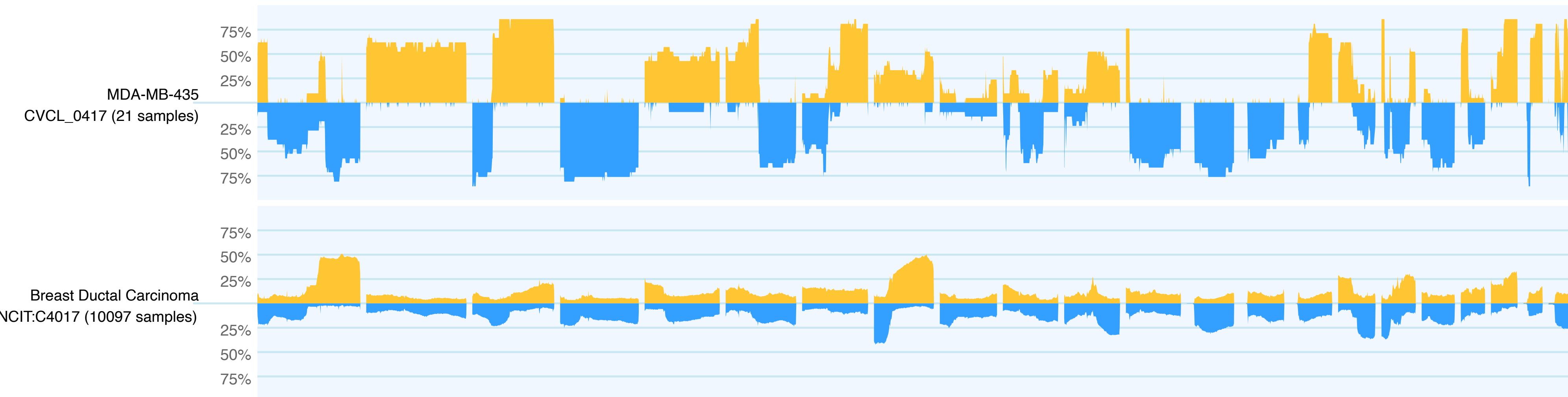
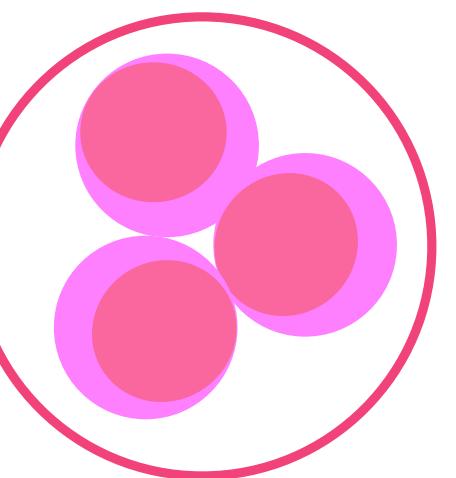
Tumor subpopulations can be matched with highly similar cell lines



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



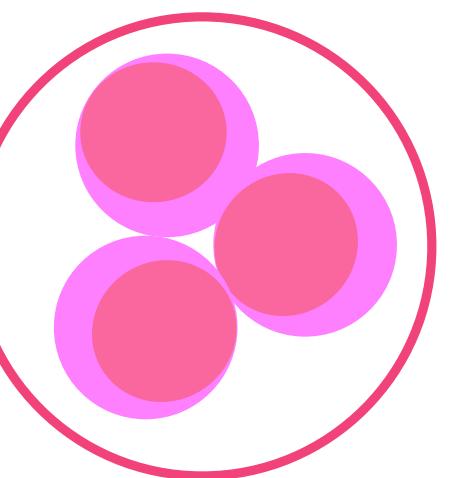
Tumor subpopulations can be matched with highly similar cell lines?!



Lead: Rahel Paloots



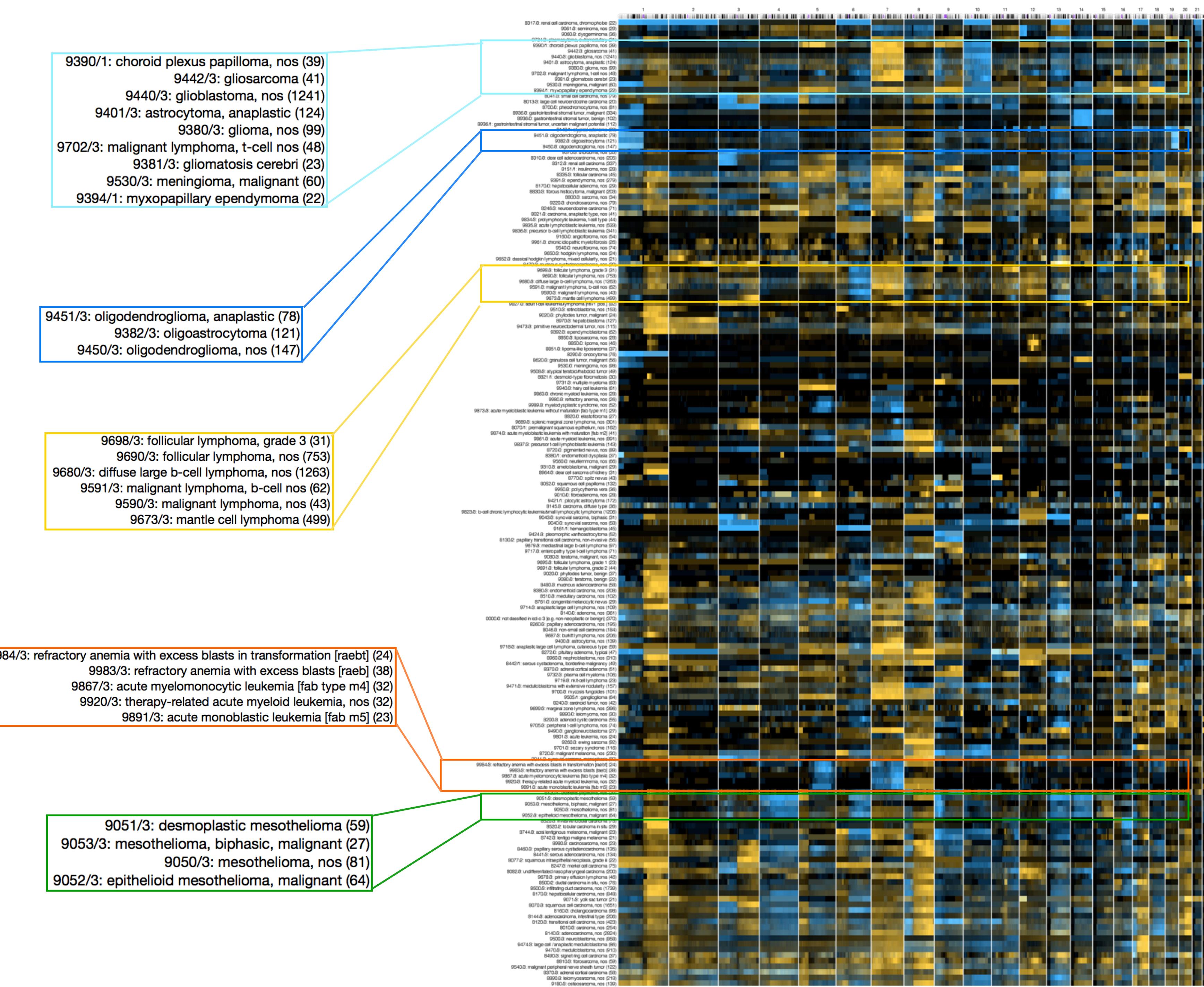
Tumor subpopulations can be matched with highly similar cell lines?!



Somatic Mutations In Cancer: Patterns

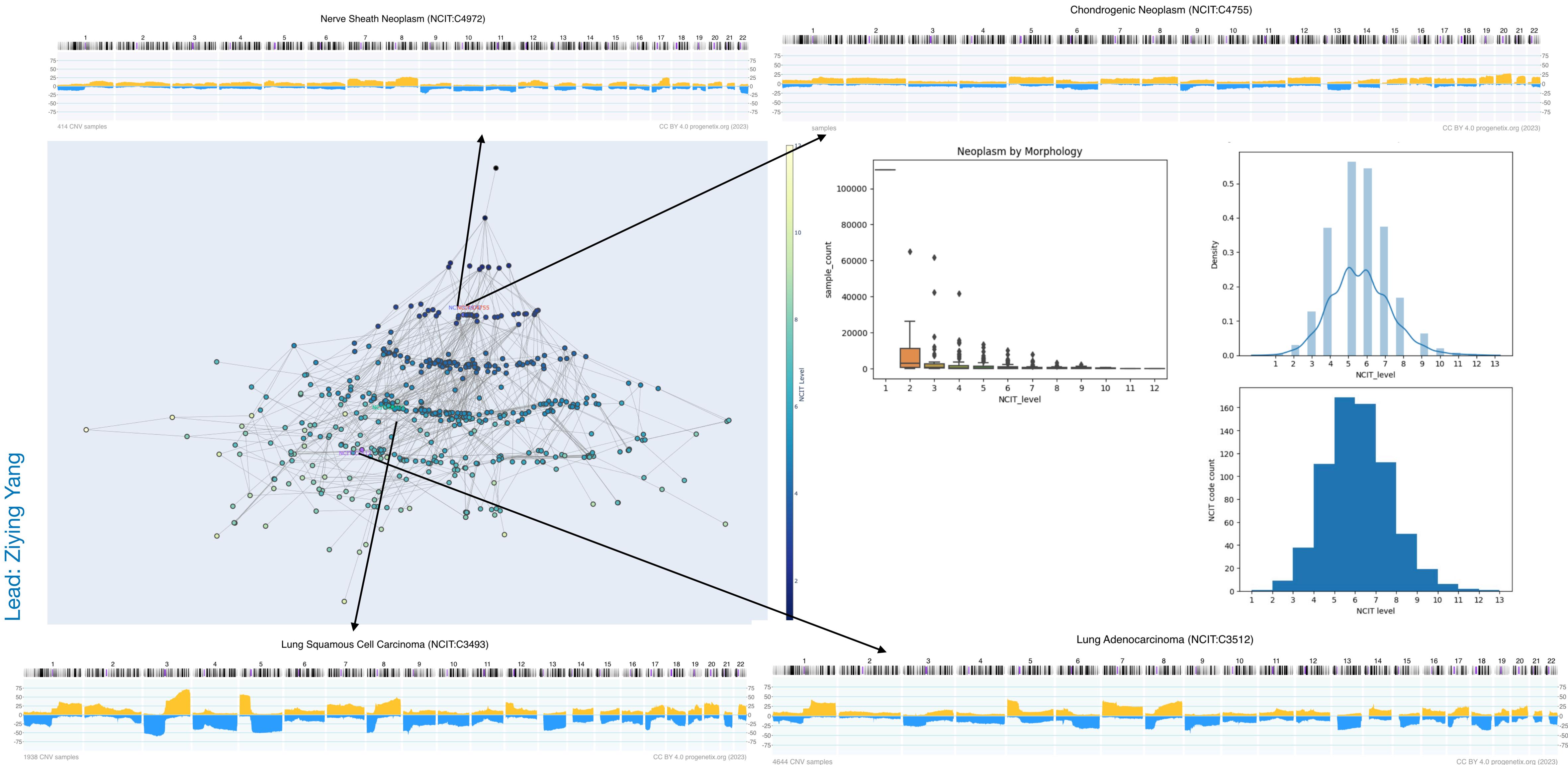
Making the case for genomic classifications

Some related cancer entities show similar copy number profiles



CNV profiles heterogeneity vs cancer classification

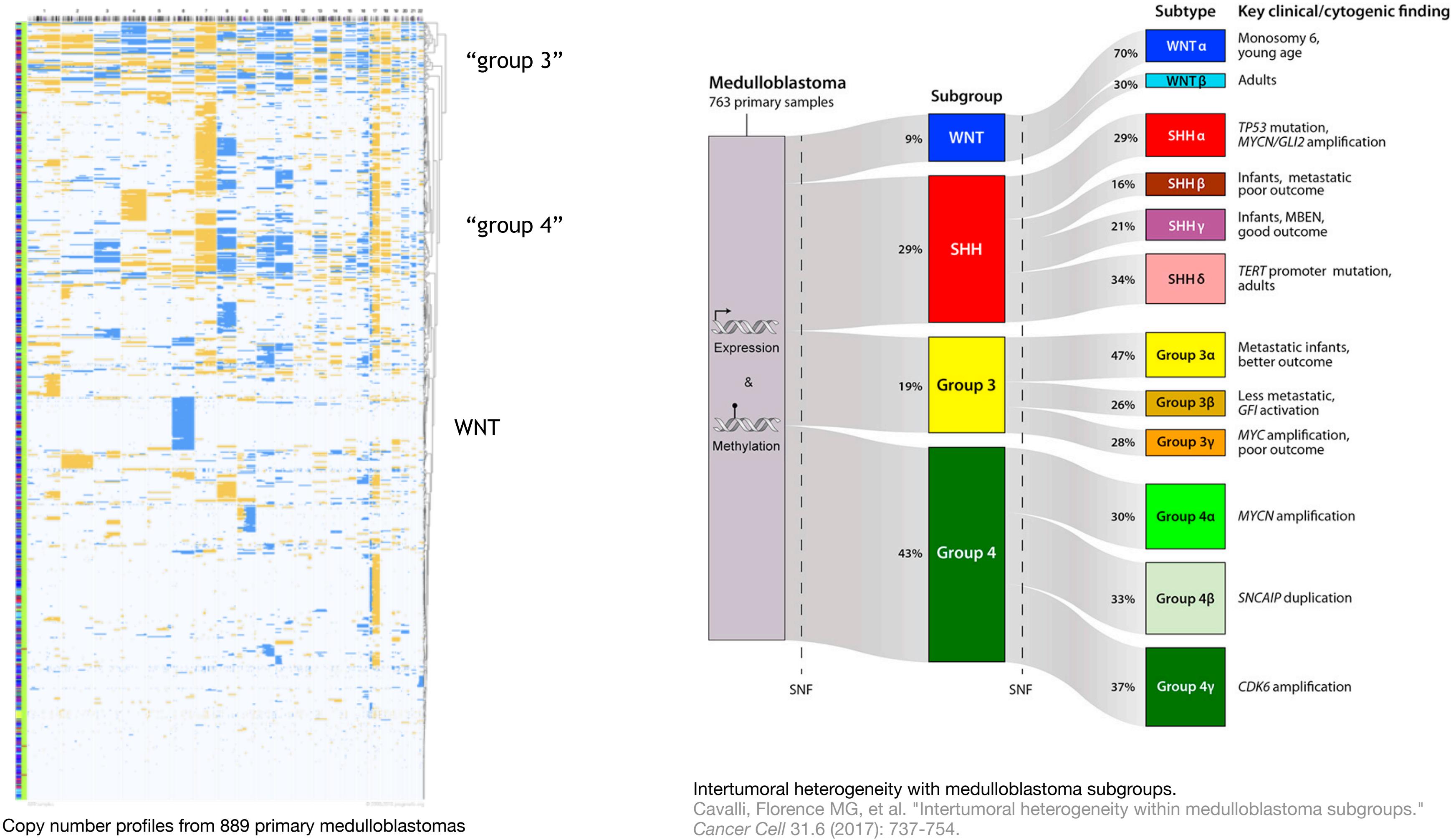
Correspondance of genomic profiles to NCIT cancer hierarchy



CNA & Cancer heterogeneity

Cancer type definitions can be improved by the addition of molecular parameters as subtype markers or even complete re-evaluation of entity definitions from molecular subtypes with distinct functional mechanisms and clinical trajectories.

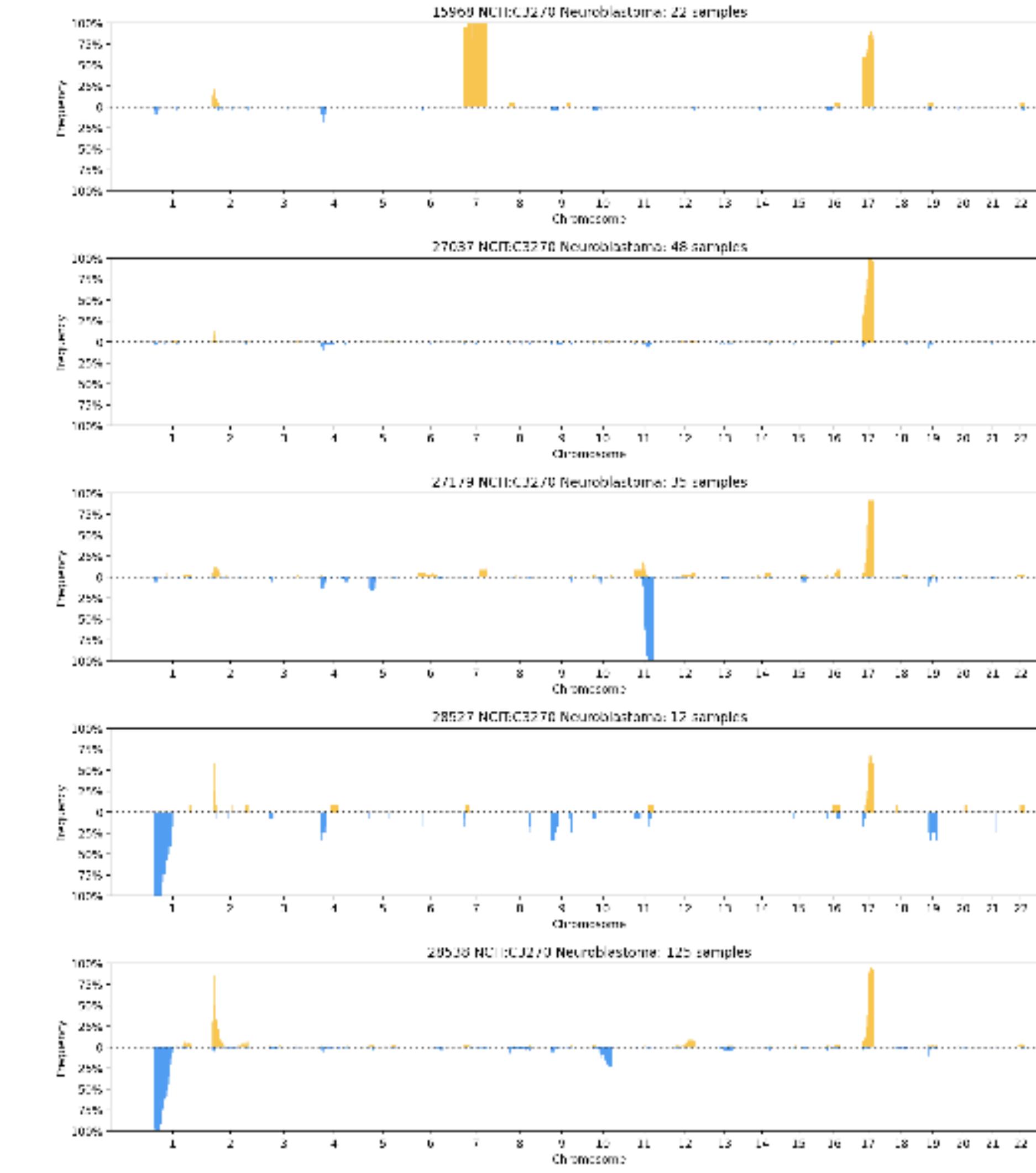
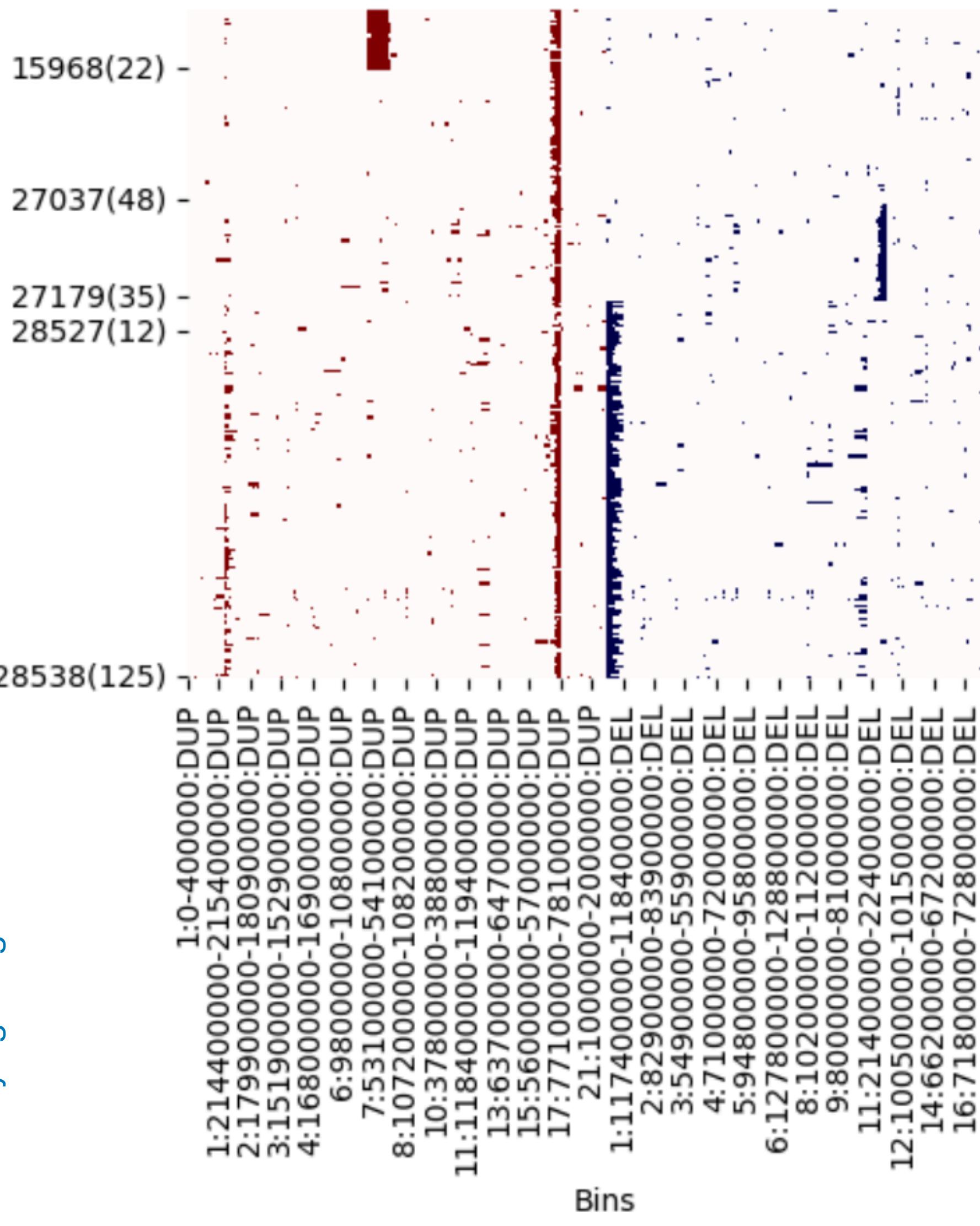
Lead: Ziyang Yang



Results

Entity CNV heterogeneity: Neuroblastoma

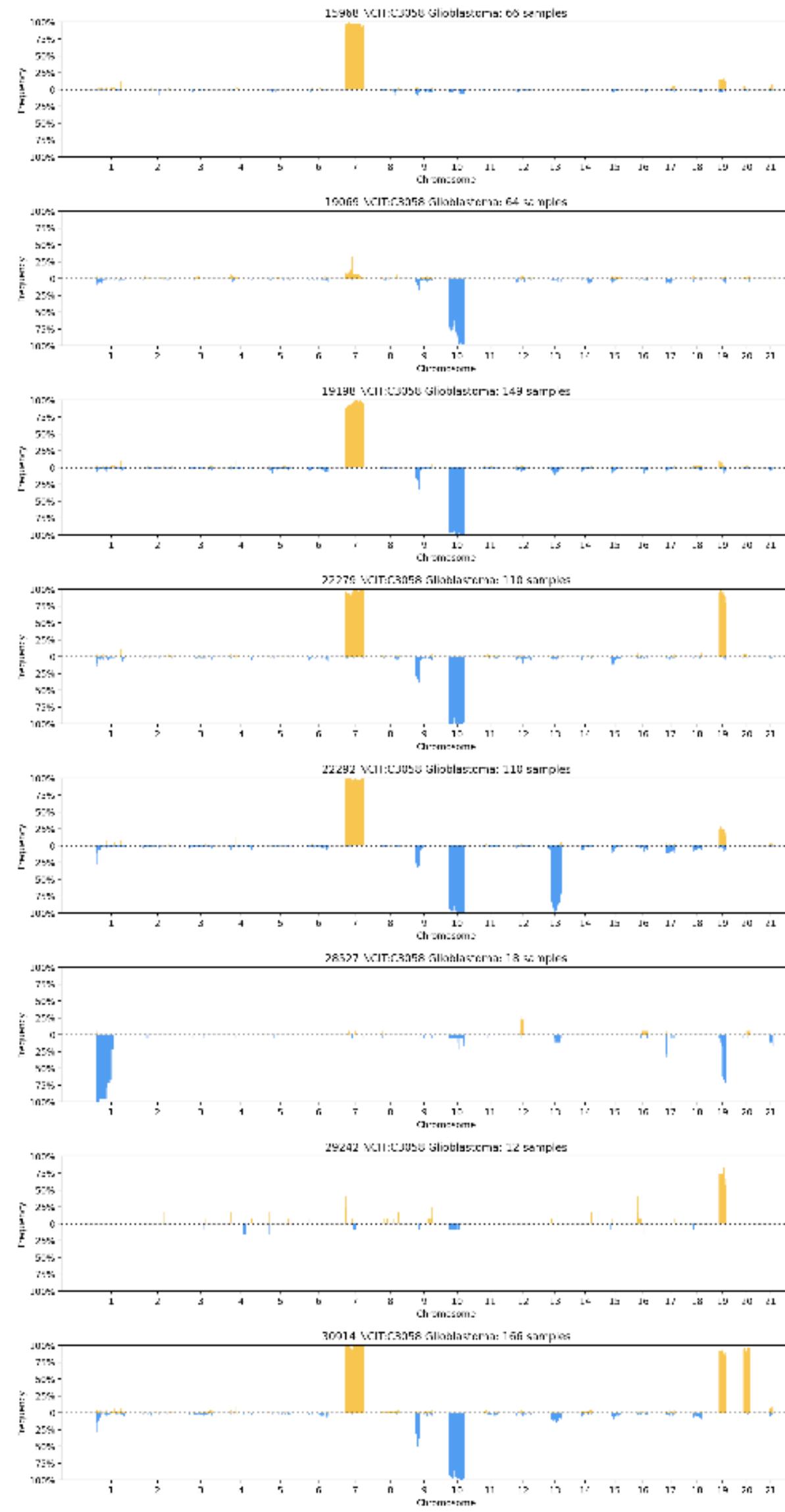
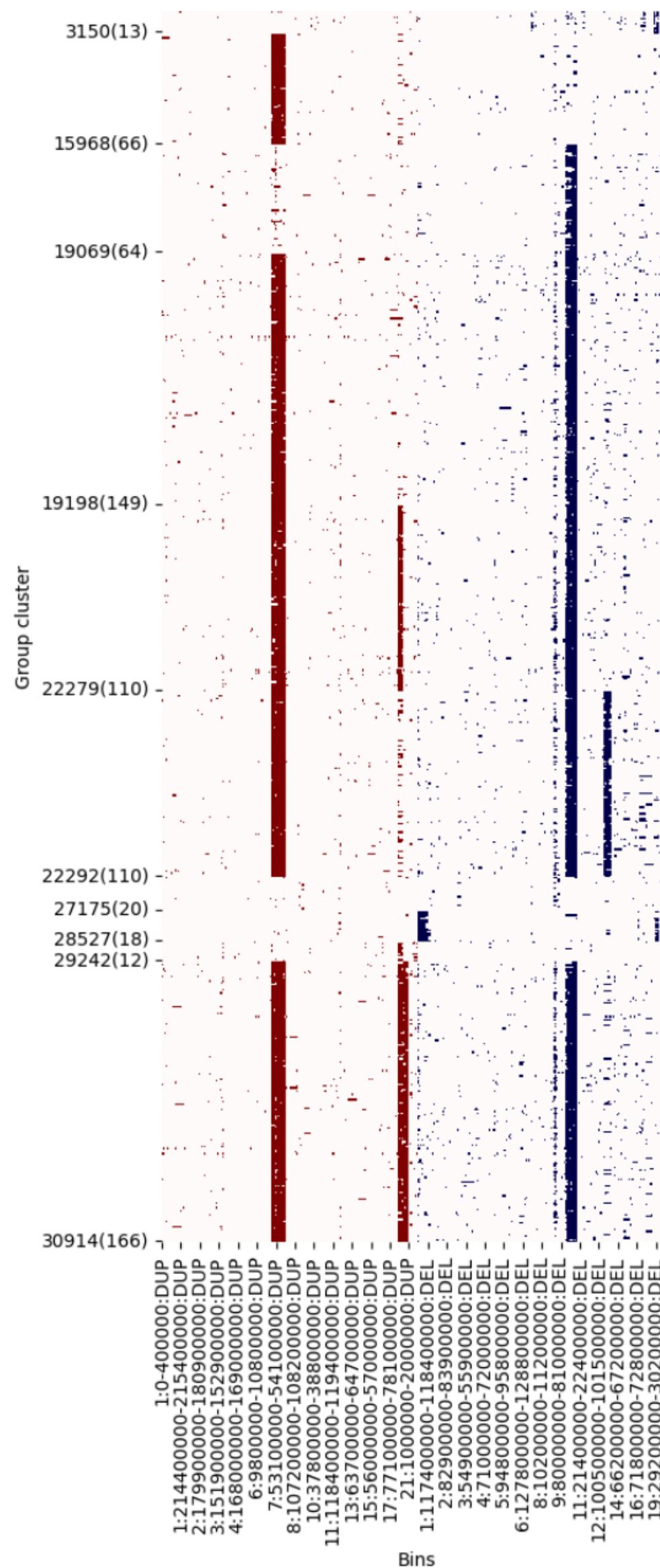
Lead: Ziyi Yang



group cluster	CNV features
15968	Dup 7
27037	Dup 17q
27179	Del 11q, Dup 17q
28527	Del 1p
28538	Del 1p, Dup 17q

Results

Entity CNV heterogeneity: Glioblastoma

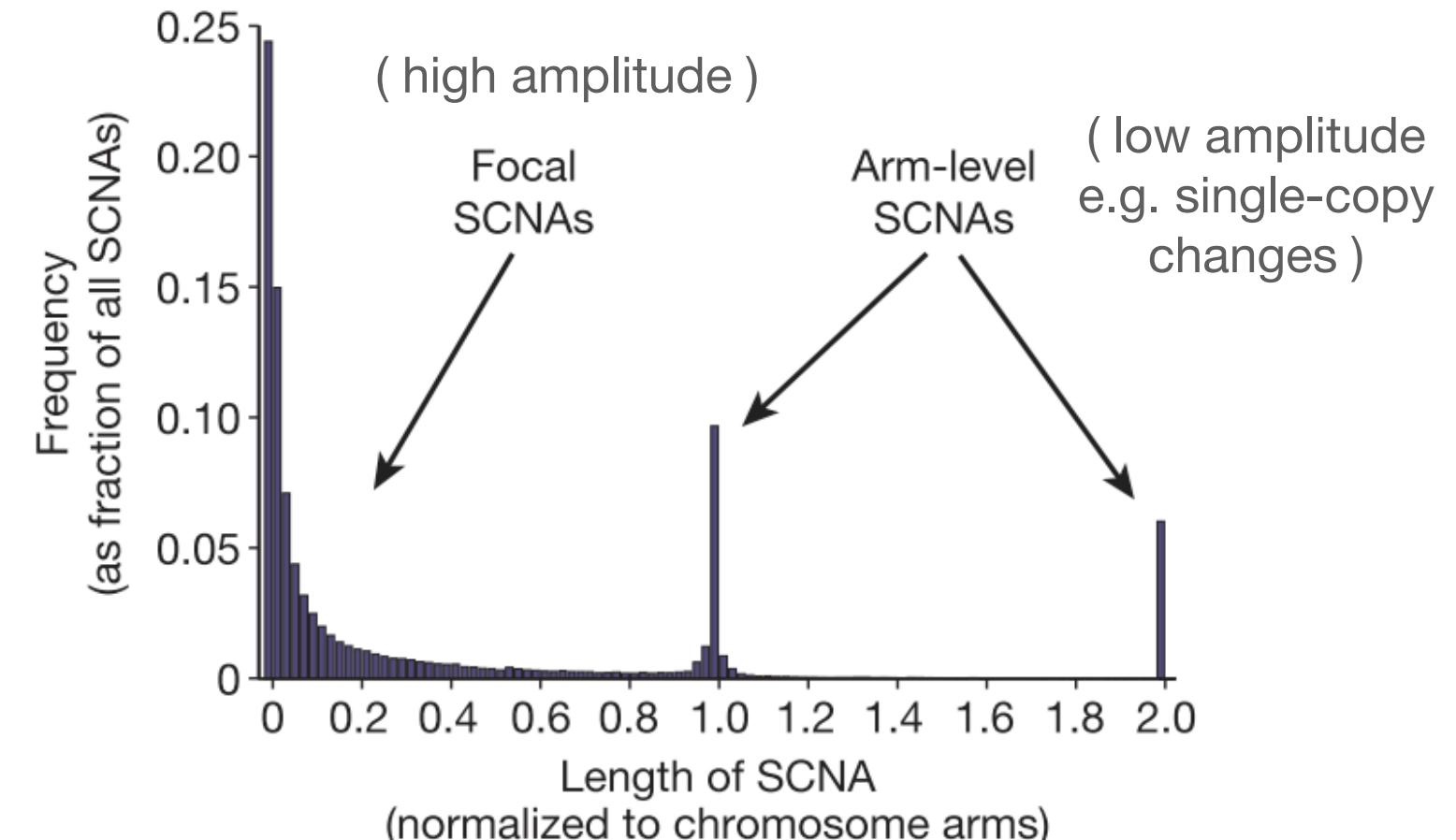


group cluster	CNV features
15968	Dup 7
19069	Del 10
19198	Dup 7, Del 10
22279	Dup 7, Del 10, Dup 19
22292	Dup 7, Del 10, Del 13
27175	Del 1p, Del 19q
28527	Del 1p, Del 19q
29242	Dup 19
30914	Dup 7, Del 10, Dup 19, Dup 20

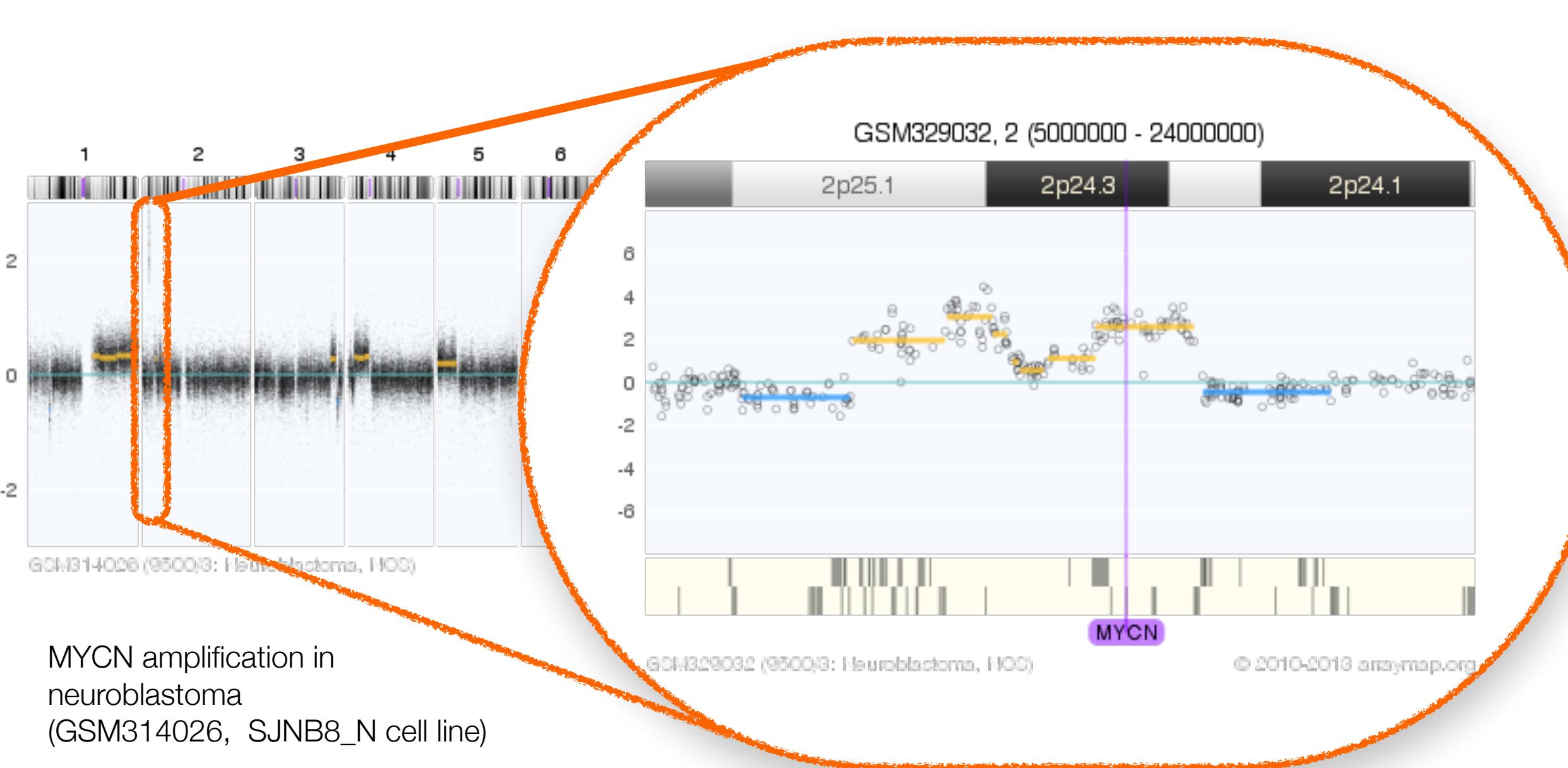


CNV Categorization

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 Term Use Comparison

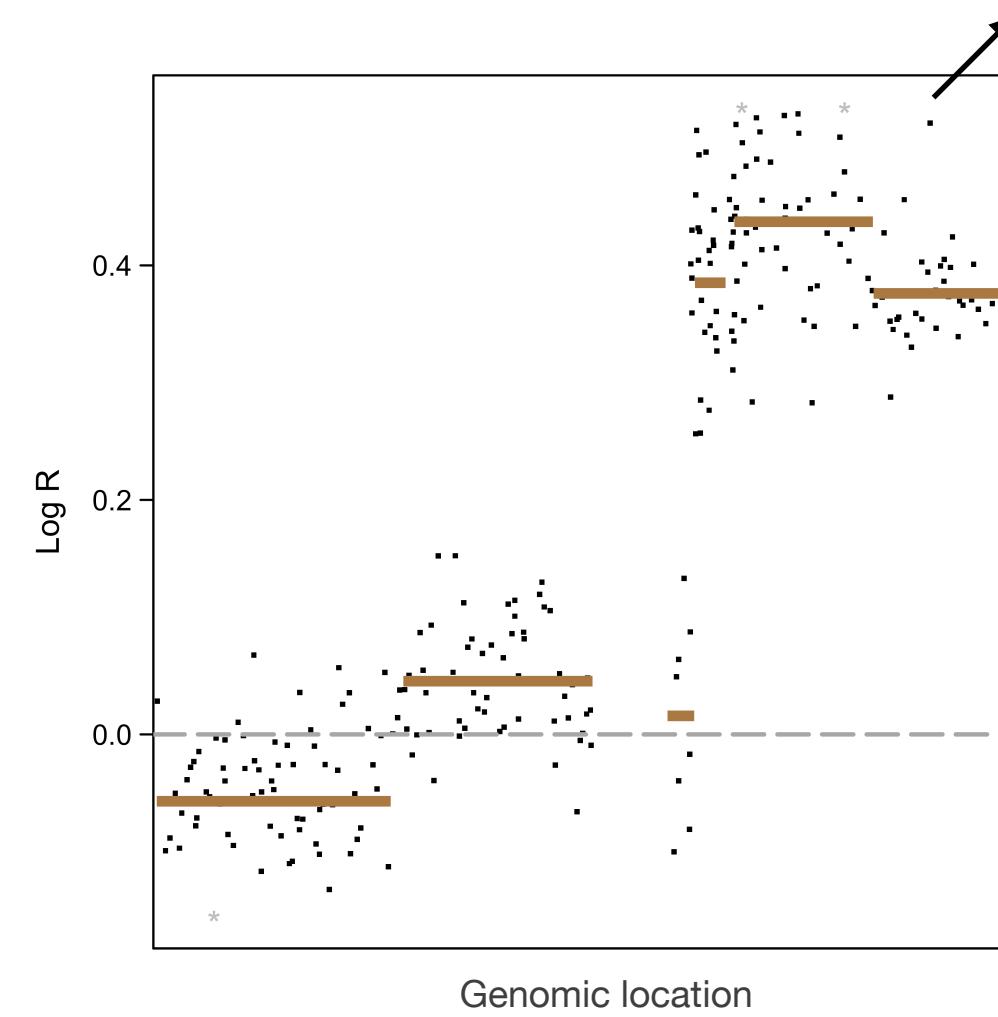
in computational (file/schema) formats

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

labelSeg

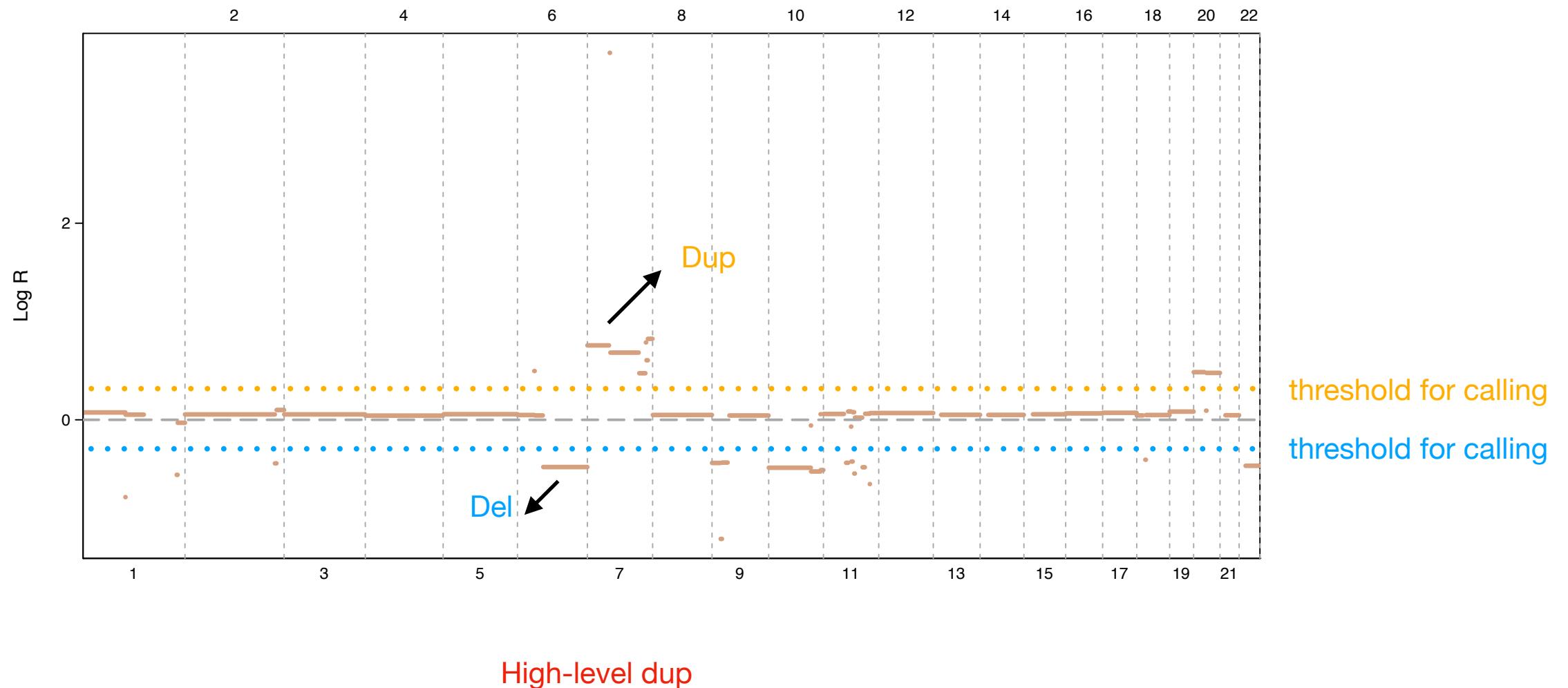
segment annotation for tumor copy number variation profiles

Signal from probes in microarray or from reads in NGS



Segmentation

a step to split the chromosomes into regions of equal copy number that accounts for the noise in the data.



README.md

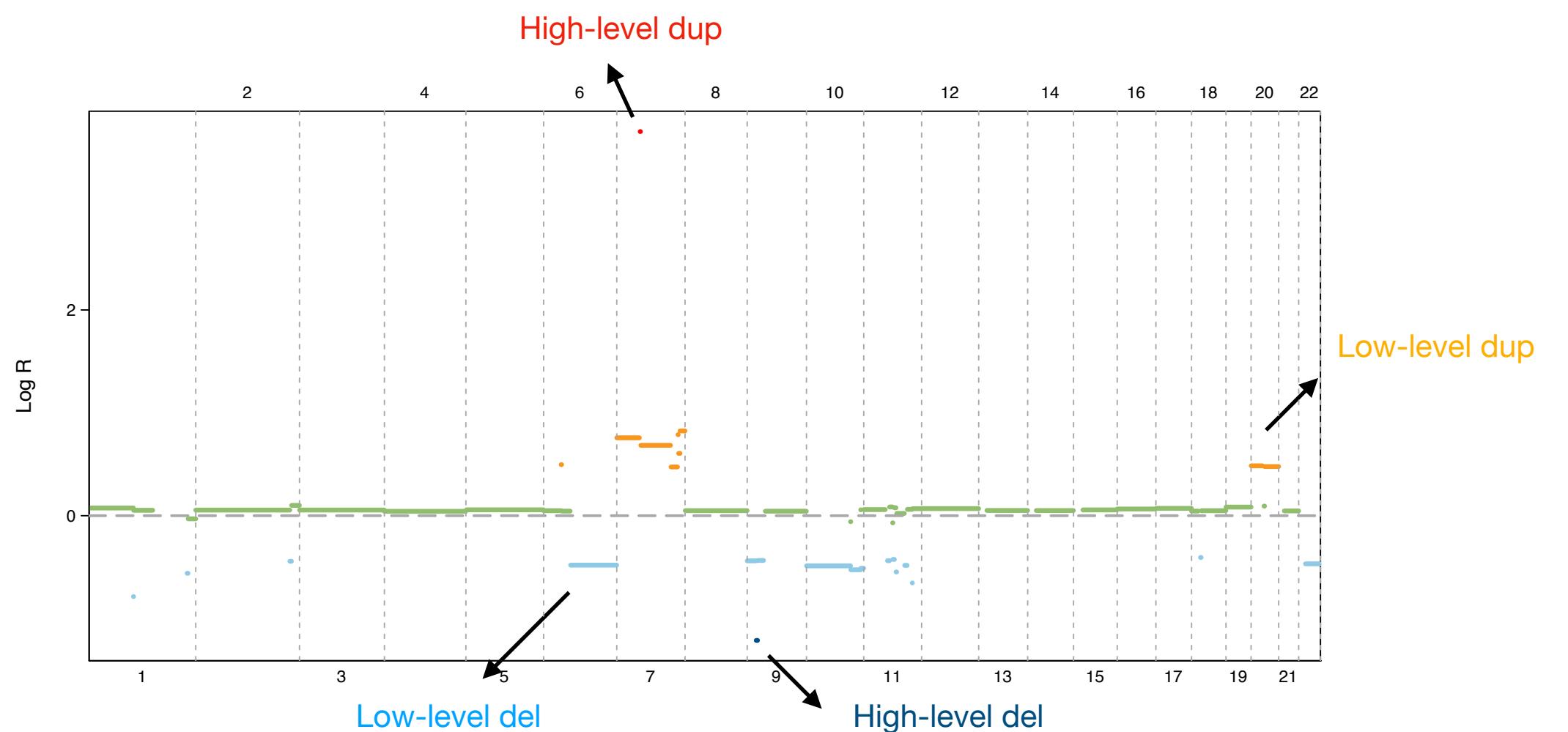
labelSeg

This is an R package designed to identify and label different levels of Copy Number Alterations (CNA) in segmented profiles.

Installation

To install the package, you can use the `devtools` package as follows:

```
install.packages("devtools")
devtools::install_github("baudisgroup/labelSeg")
```



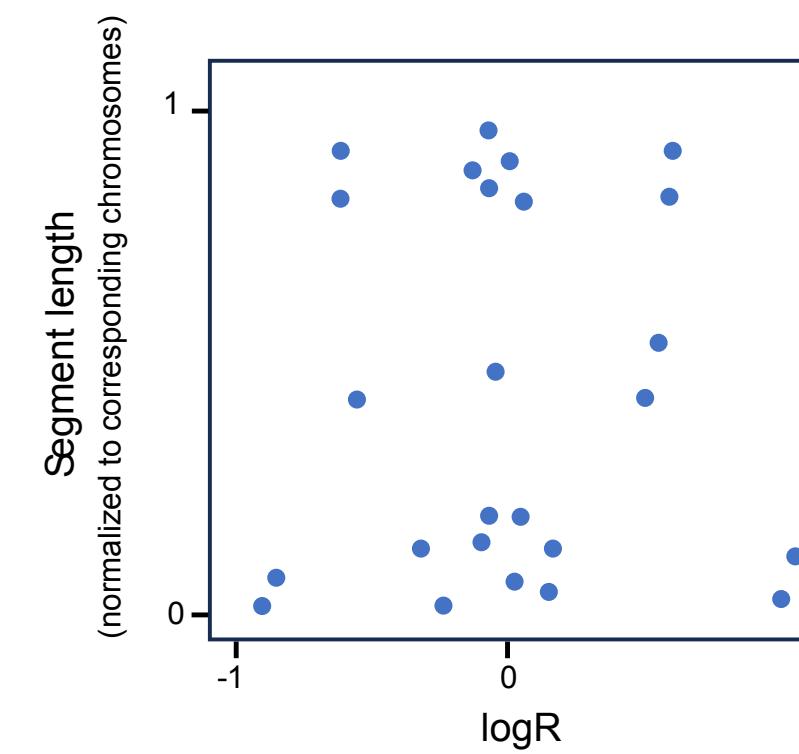
labelSeg

segment annotation for tumor copy number variation profiles

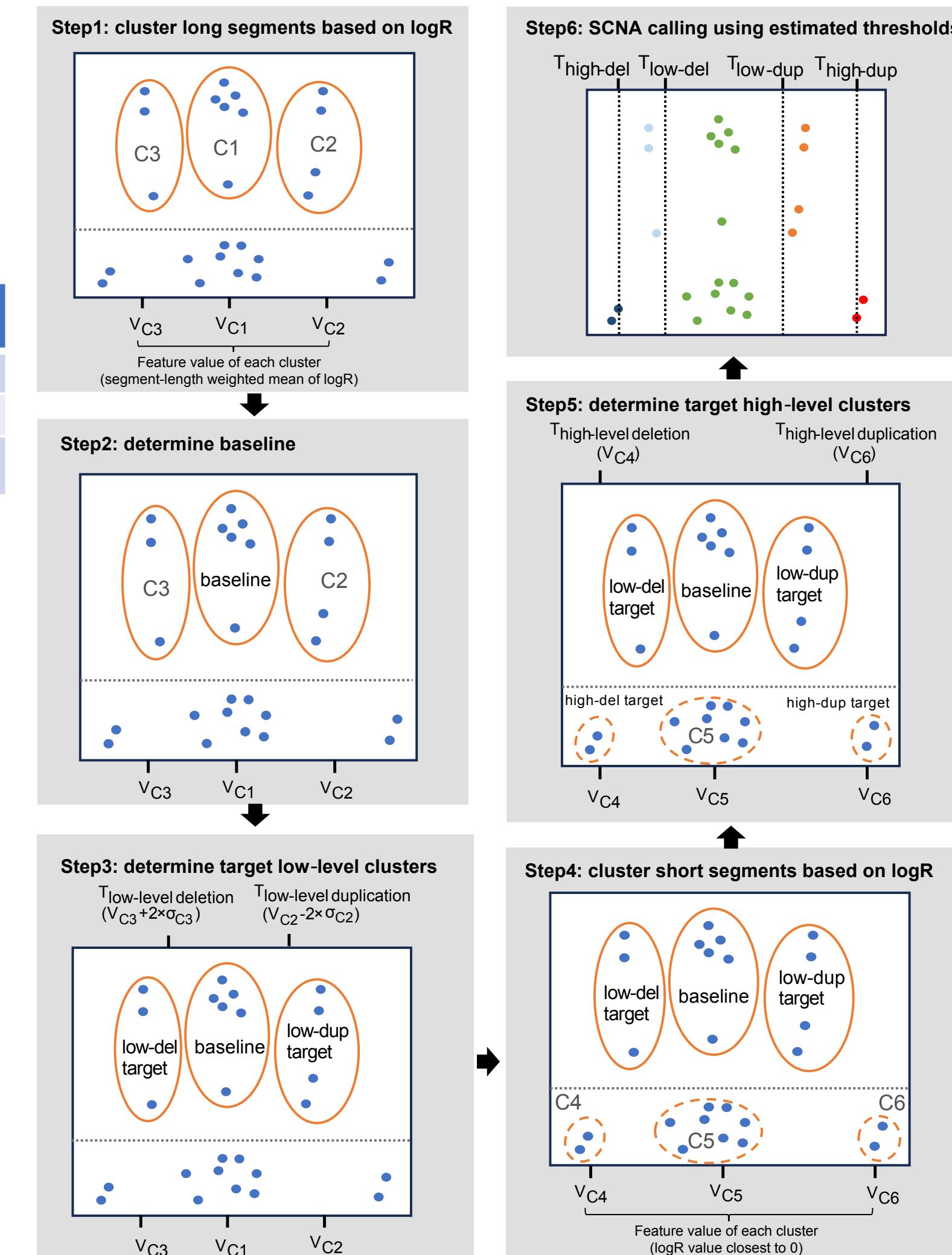
A

INPUT: individual .seg files

Sample ID	Chr	Start	End	#Marker	logR
S1	1	9965785	10911995	450	0.306
...
S1	X	3236359	155677414	63741	0.005



B



C

OUTPUT: individual .seg files with SCNA labels

Sample ID	Chr	Start	End	#Marker	logR	Label
S1	1	9965785	10911995	450	0.306	+1
...
S1	X	3236359	155677414	63741	0.005	0



bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

Follow this preprint

labelSeg: segment annotation for tumor copy number alteration profiles

id Hangjia Zhao, id Michael Baudis

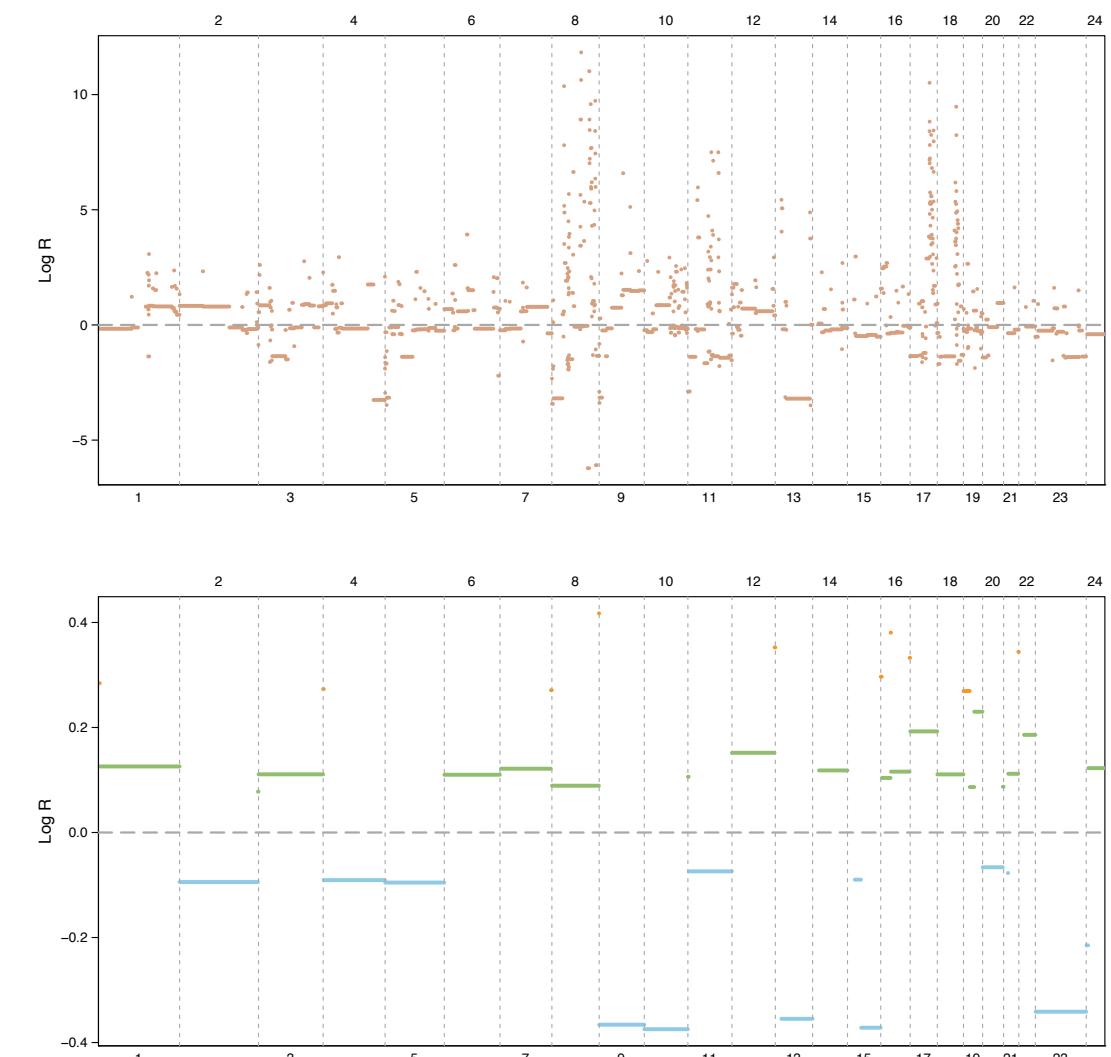
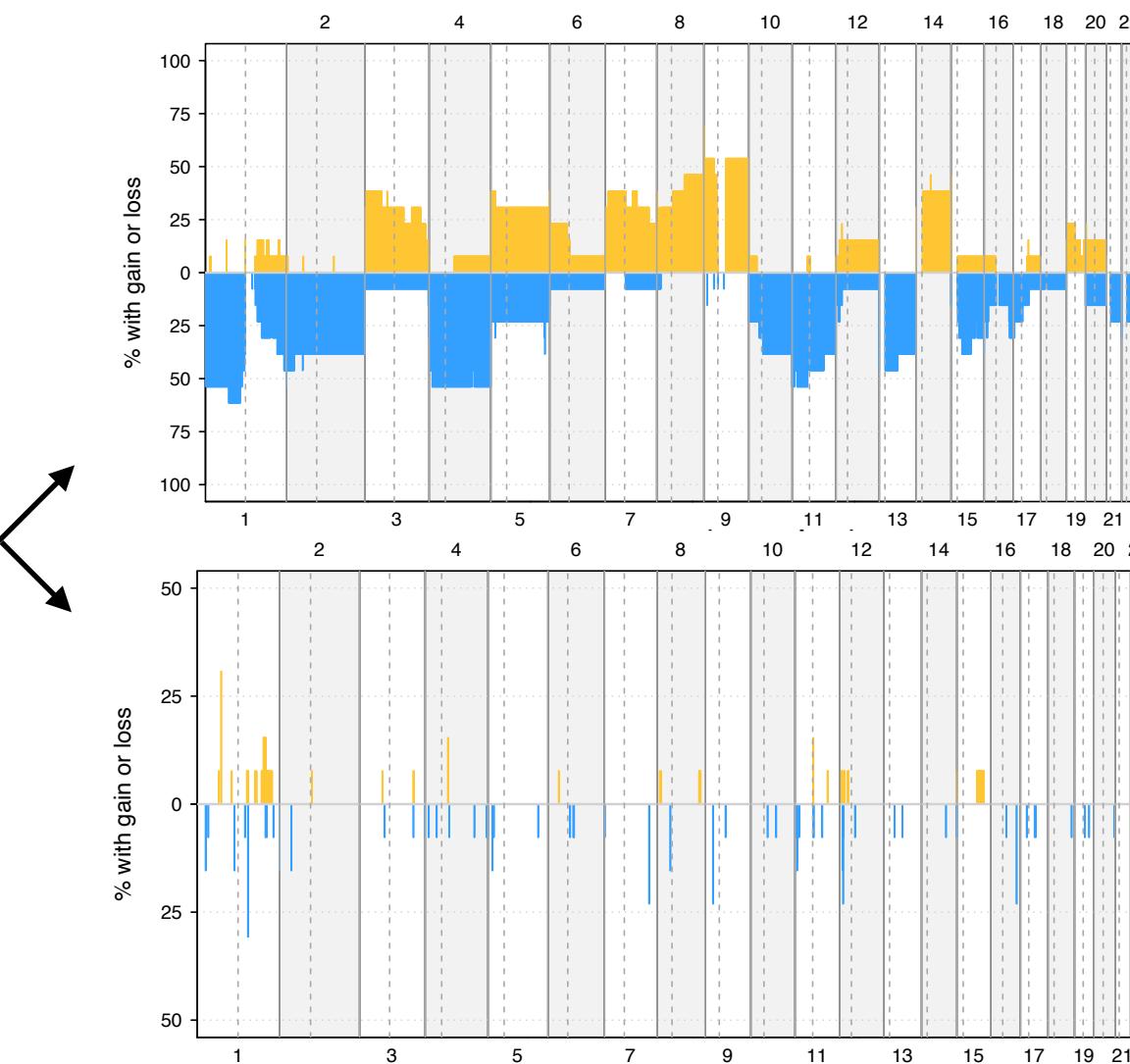
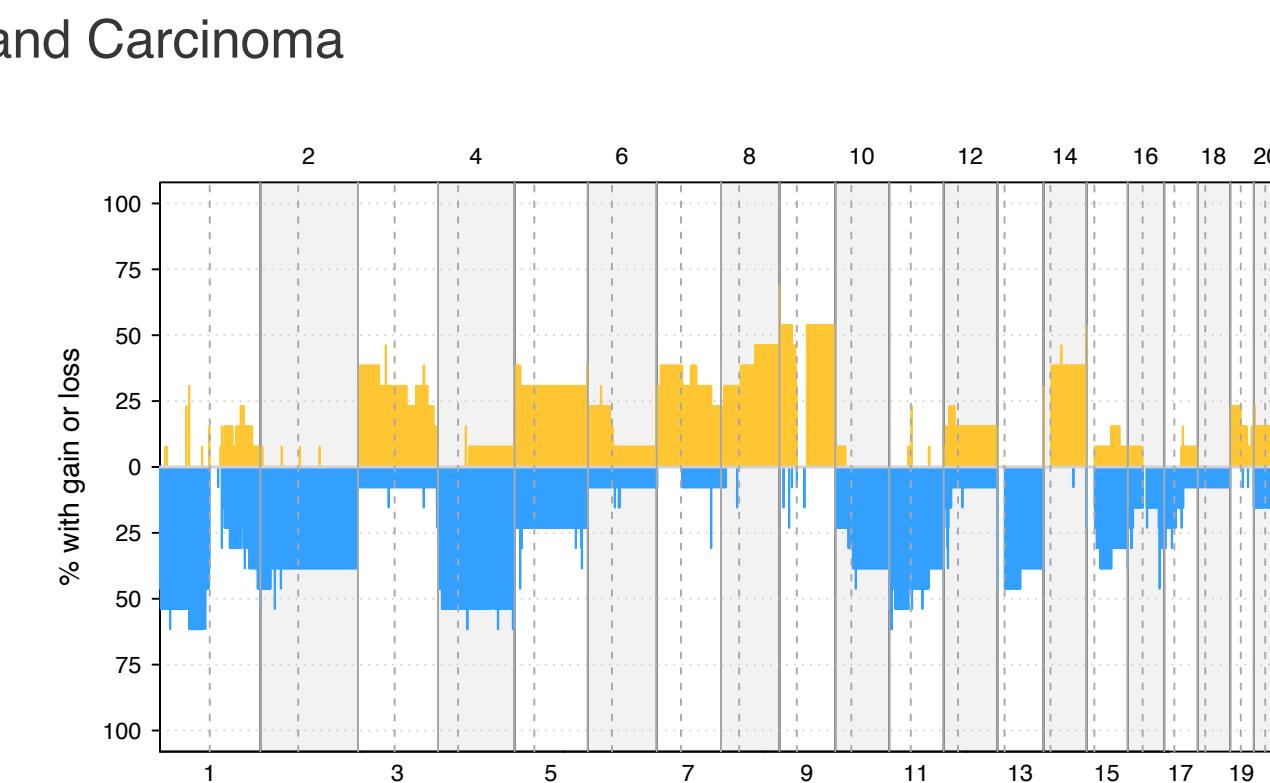
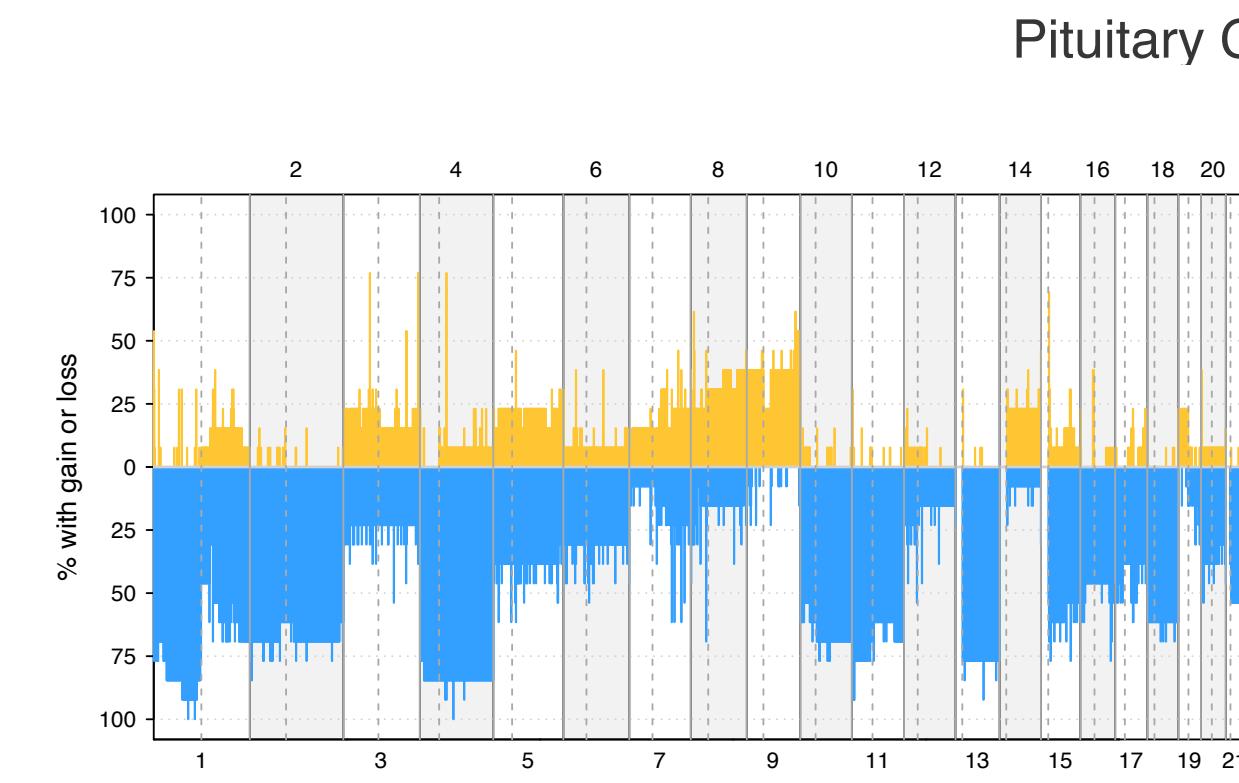
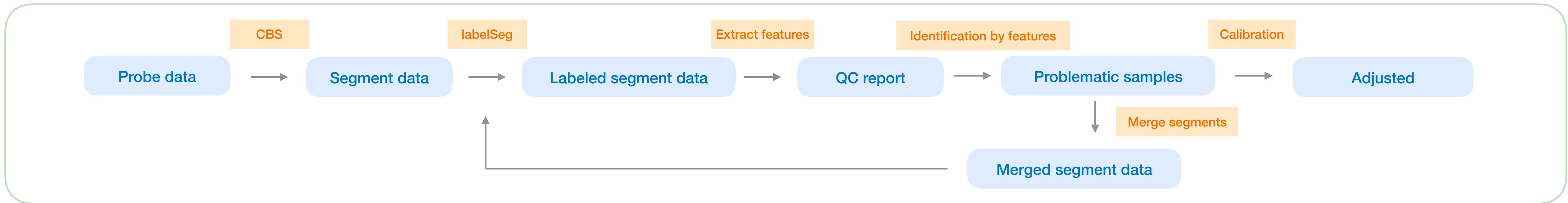
doi: <https://doi.org/10.1101/2023.05.17.541097>

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

Pipeline Development

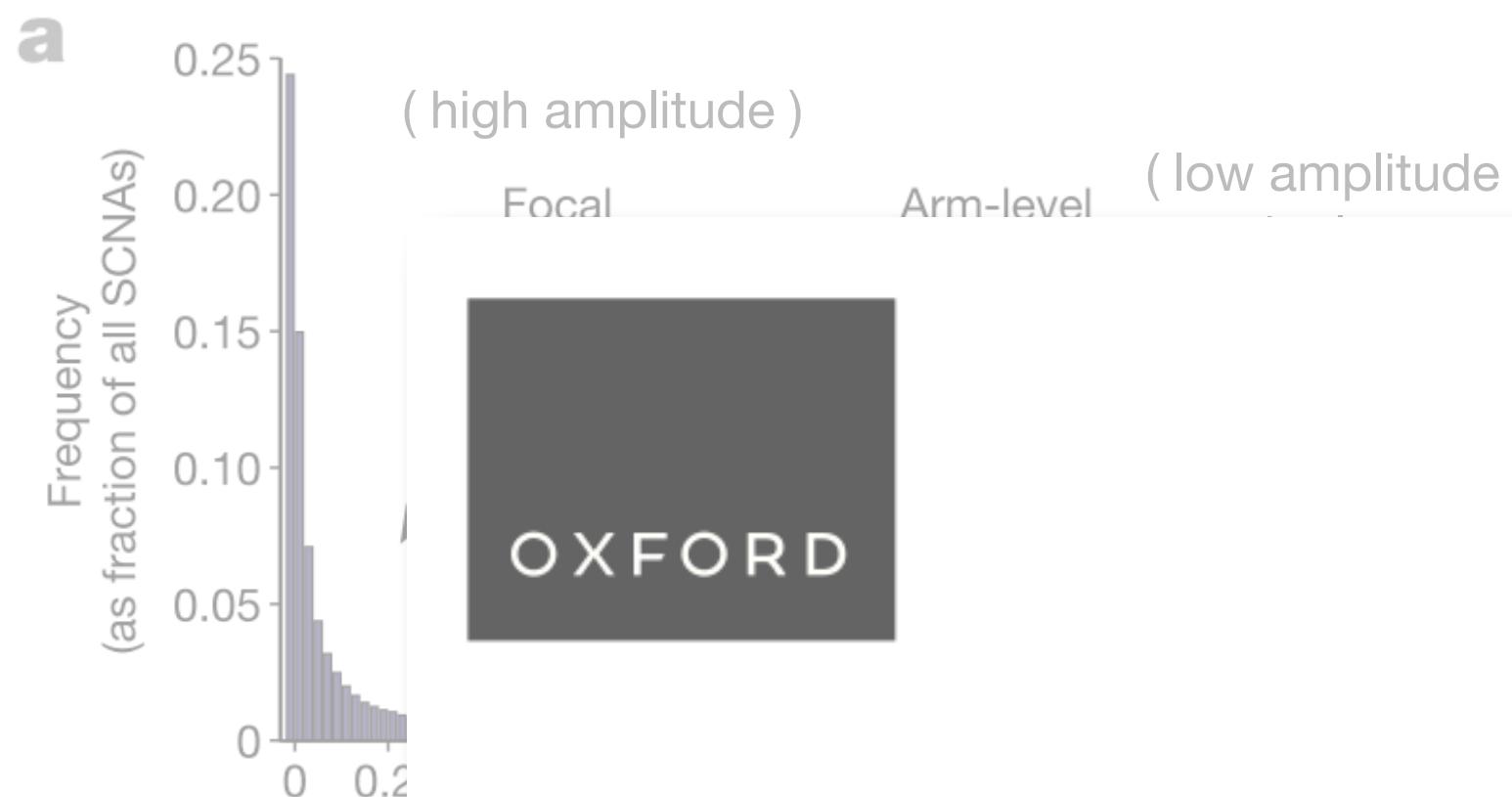
improve CNV calling in large numbers of heterogeneous cancer samples

nextflow



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

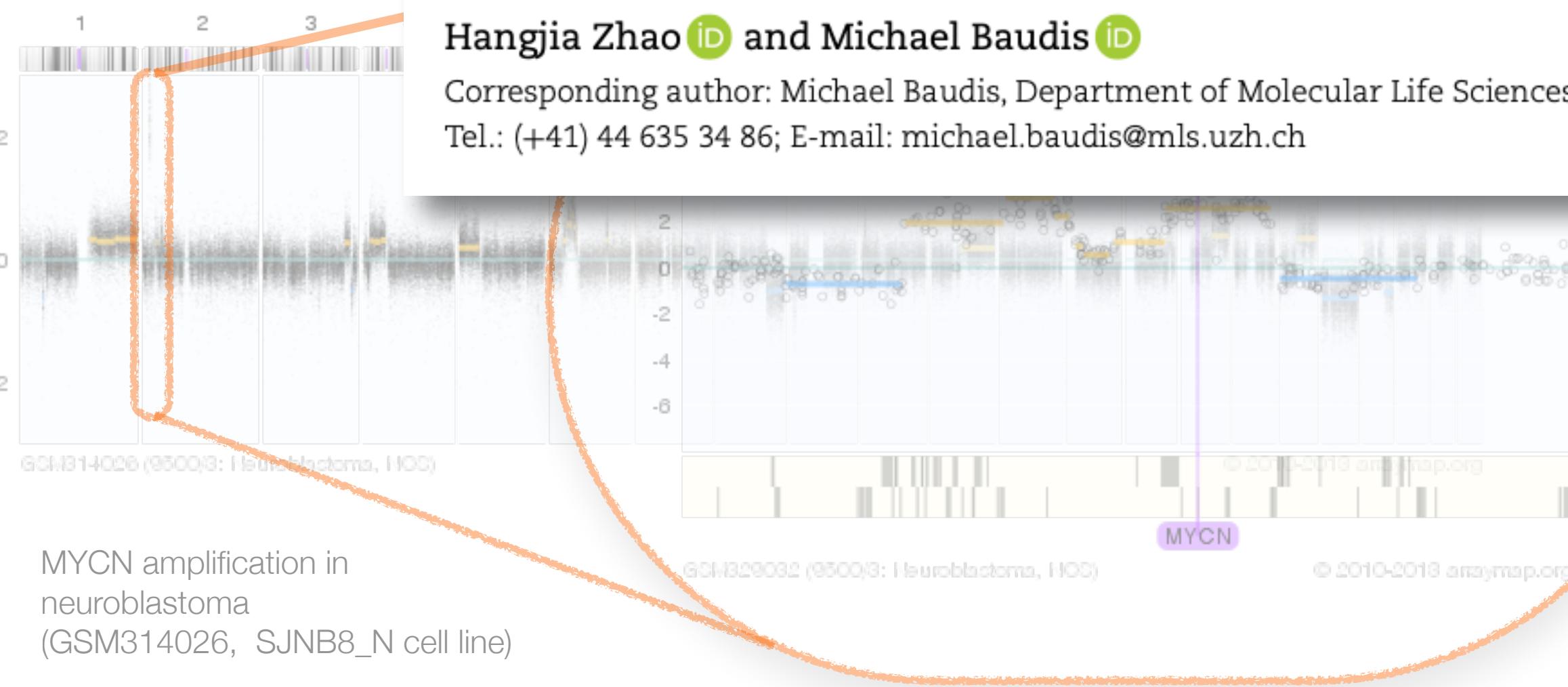
a system (e.g. genome, cell,

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

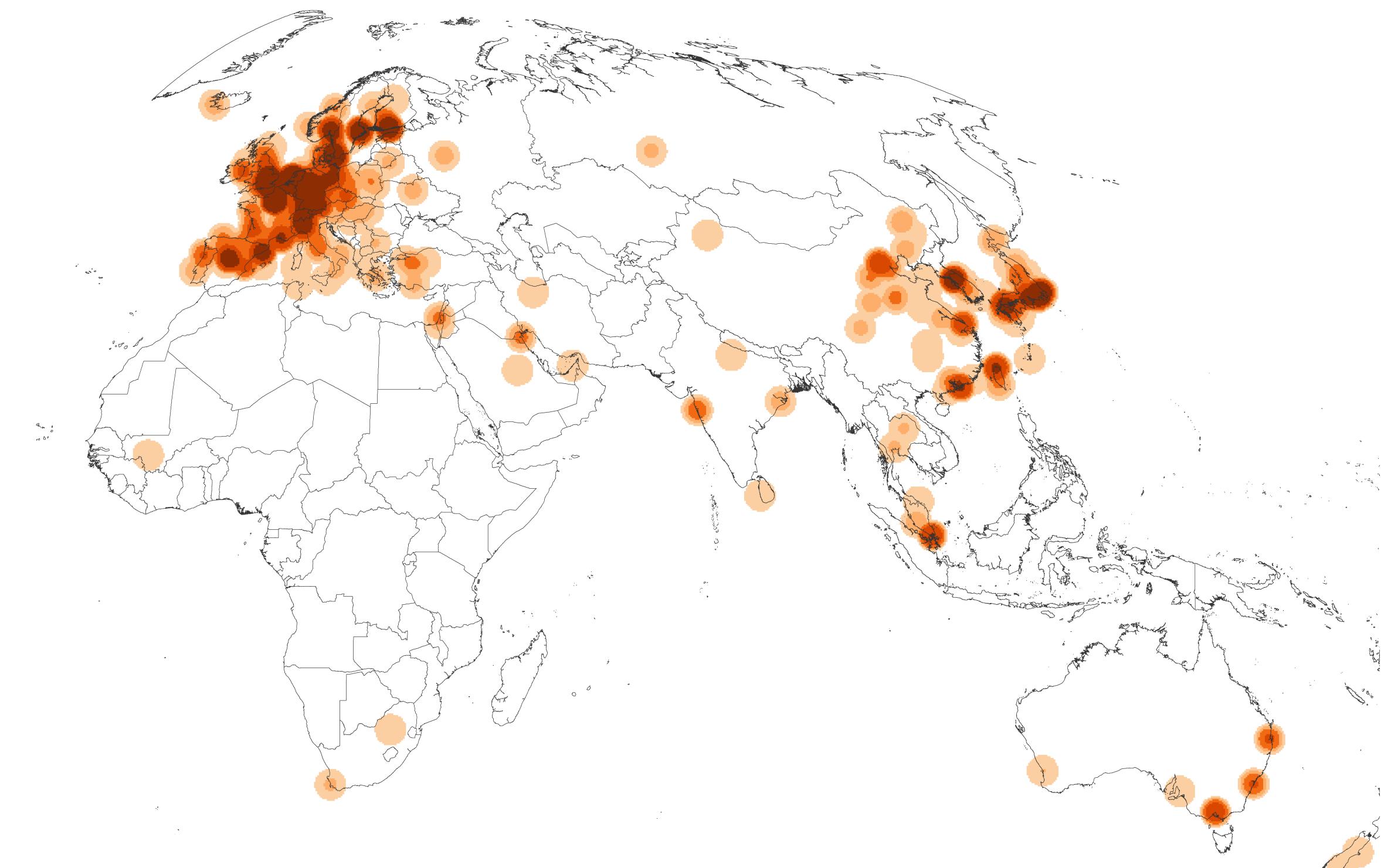
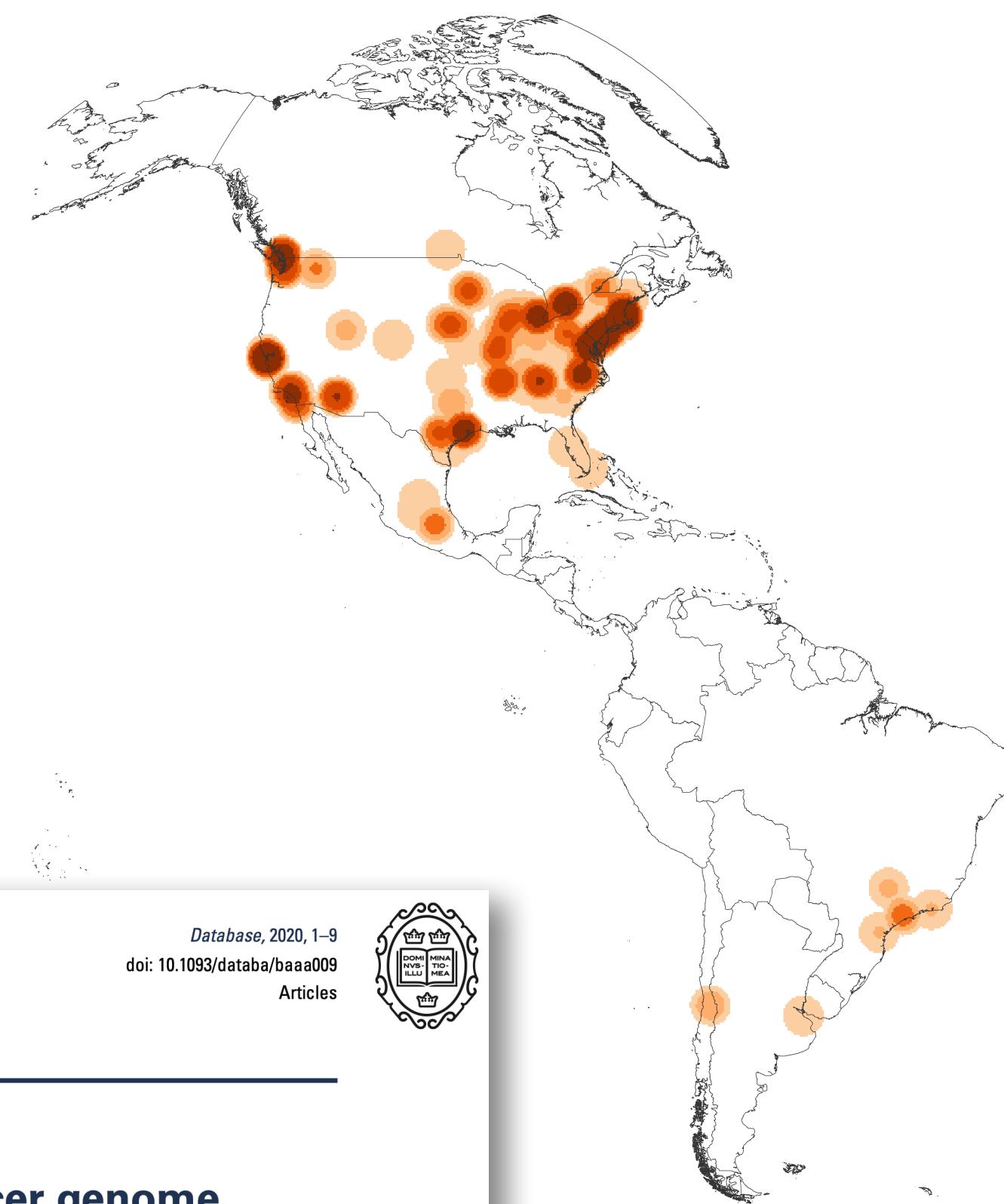
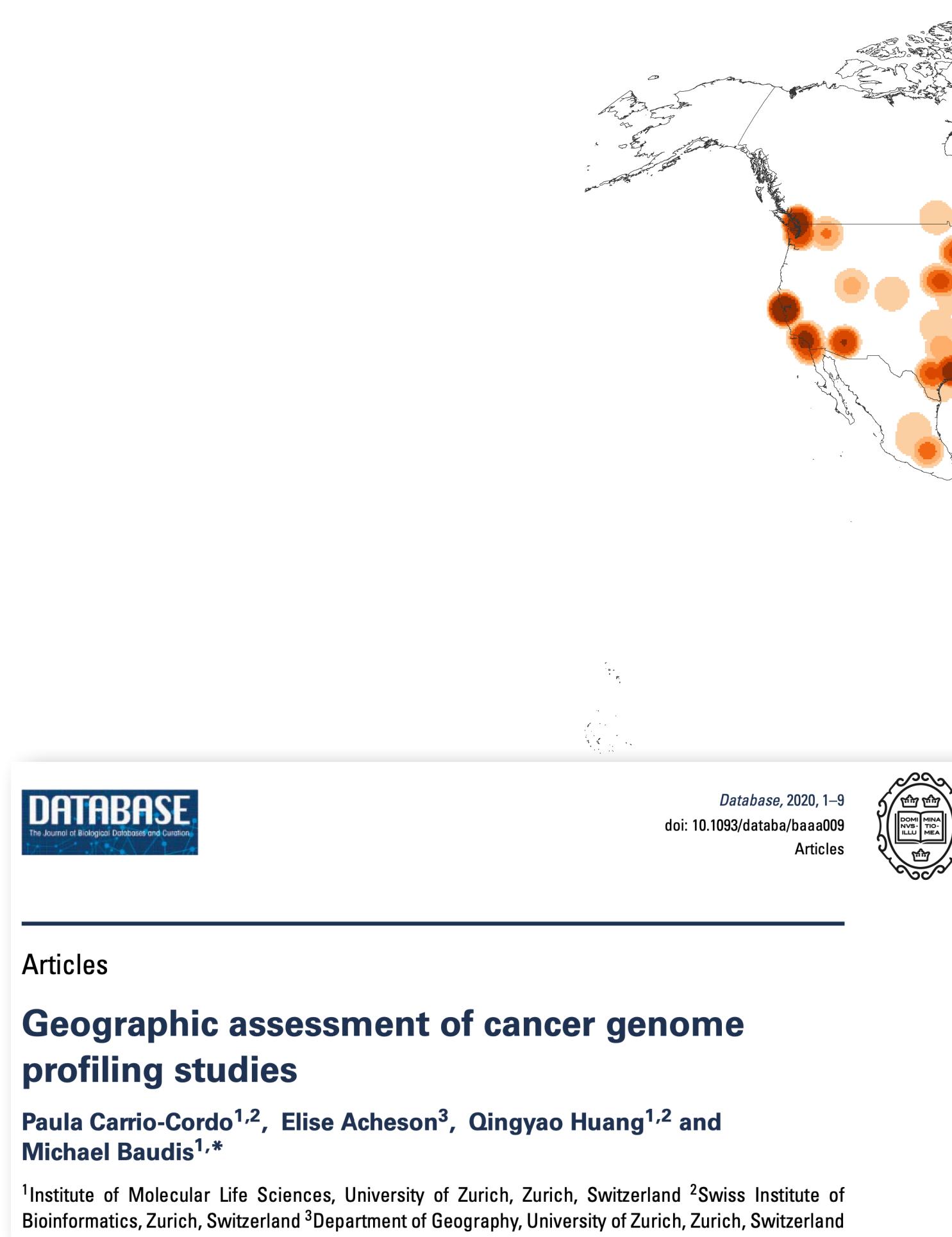
Tel.: (+41) 44 635 34 86; E-mail: michael.baudis@mls.uzh.ch



_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.



Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

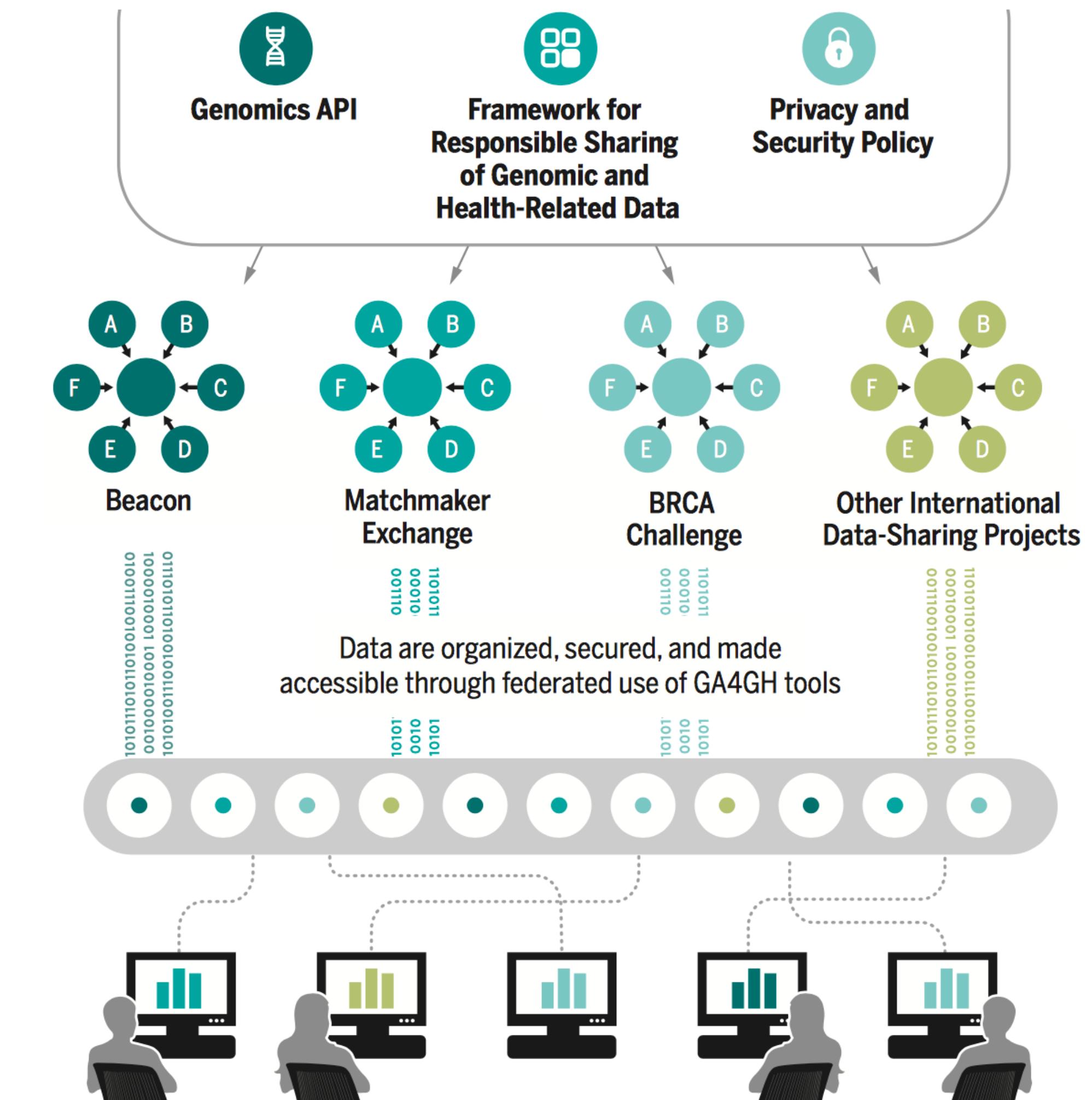


GENOMICS

A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems

A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.



INFORMATICS

Beacon v2 and Beacon networks: federated data discovery in biome

Commentary

International federation of genomic medicine databases using GA4GH standards

Adrian Thorogood,^{1,2,*} Heidi L. Rehm,^{3,4} Peter Goodhand,^{5,6} Angela J.H. Page,^{4,5} Yann Joly,² Michael Baudis,⁷ Jordi Rambla,^{8,9} Arcadi Navarro,^{8,10,11,12} Tommi H. Nyronen,^{13,14} Mikael Linden,^{13,14} Edward S. Dove,¹⁵ Marc Fiume,¹⁶ Michael Brudno,¹⁷ Melissa S. Cline,¹⁸ and Ewan Birney¹⁹

Jordi Rambla^{1,2} | Michael Baudis³ | Roberto Ariosa¹ | Tim Beck⁴ |
 Lauren A. Fromont¹ | Arcadi Navarro^{1,5,6,7} | Rahel Paloots³ |
 Manuel Rueda¹ | Gary Saunders⁸ | Babita Singh¹ | John D. Spalding⁹ |
 Juha Törnroos⁹ | Claudia Vasallo¹ | Colin D. Veal⁴ | Anthony J. Brookes⁴

Cell Genomics

Technology

The GA4GH Variation Representation Specification A computational framework for variation representation and federated identification

Alex H. Wagner,^{1,2,25,*} Lawrence Babb,^{3,*} Gil Alterovitz,^{4,5} Michael Baudis,⁶ Matthew Brush,⁷ Daniel L. Cameron,^{8,9} Melissa Cline,¹⁰ Malachi Griffith,¹¹ Obi L. Griffith,¹¹ Sarah E. Hunt,¹² David Kreda,¹³ Jennifer M. Lee,¹⁴ Stephanie Li,¹⁵ Javier Lopez,¹⁶ Eric Moyer,¹⁷ Tristan Nelson,¹⁸ Ronak Y. Patel,¹⁹ Kevin Riehle,¹⁹ Peter N. Robinson,²⁰ Shawn Rynearson,²¹ Helen Schuilenburg,¹² Kirill Tsukanov,¹² Brian Walsh,⁷ Melissa Konopko,¹⁵ Heidi L. Rehm,^{3,22} Andrew D. Yates,¹² Robert R. Freimuth,²³ and Reece K. Hart^{3,24,*}

Cell Genomics

Perspective

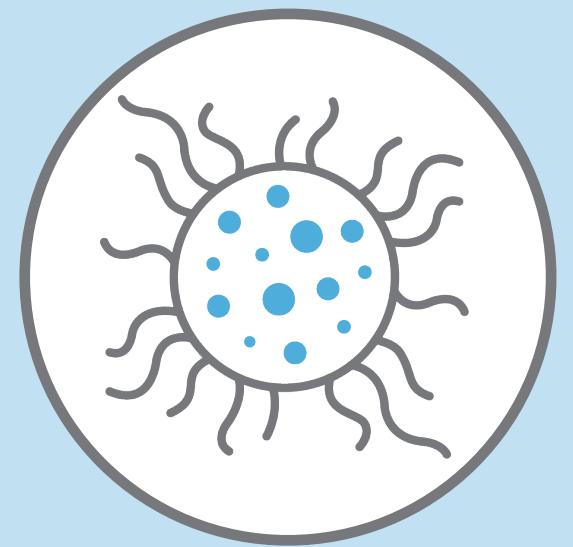
GA4GH: International policies and standards for data sharing across genomic research and healthcare

Heidi L. Rehm,^{1,2,47} Angela J.H. Page,^{1,3,*} Lindsay Smith,^{3,4} Jeremy B. Adams,^{3,4} Gil Alterovitz,^{5,47} Lawrence J. Babb,¹ Maxmillian P. Barkley,⁶ Michael Baudis,^{7,8} Michael J.S. Beauvais,^{3,9} Tim Beck,¹⁰ Jacques S. Beckmann,¹¹ Sergi Beltran,^{12,13,14} David Bernick,¹ Alexander Bernier,⁹ James K. Bonfield,¹⁵ Tiffany F. Boughtwood,^{16,17} Guillaume Bourque,^{9,18} Sarion R. Bowers,¹⁵ Anthony J. Brookes,¹⁰ Michael Brudno,^{18,19,20,21,38} Matthew H. Brush,²² David Bujold,^{9,18,38} Tony Burdett,²³ Orion J. Buske,²⁴ Moran N. Cabili,¹ Daniel L. Cameron,^{25,26} Robert J. Carroll,²⁷ Esmeralda Casas-Silva,¹²³ Debyani Chakravarty,²⁹ Bimal P. Chaudhari,^{30,31} Shu Hui Chen,³² J. Michael Cherry,³³ Justina Chung,^{3,4} Melissa Cline,³⁴ Hayley L. Clissold,¹⁵ Robert M. Cook-Deegan,³⁵ Mélanie Courtot,²³ Fiona Cunningham,²³ Miro Cupak,⁶ Robert M. Davies,¹⁵ Danielle Denisko,¹⁹ Megan J. Doerr,³⁶ Lena I. Dolman,¹⁹

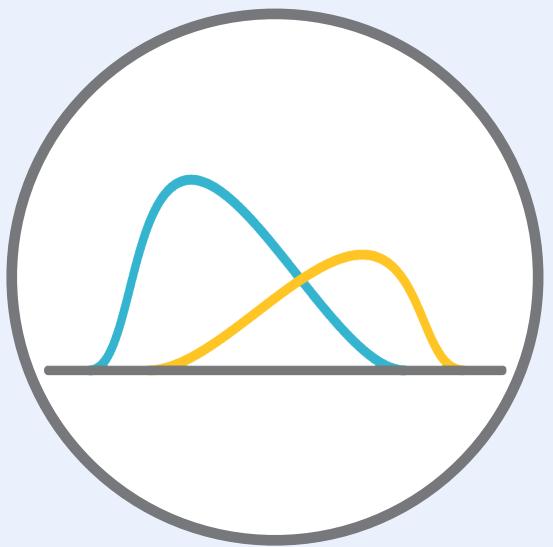
(Author list continued on next page)



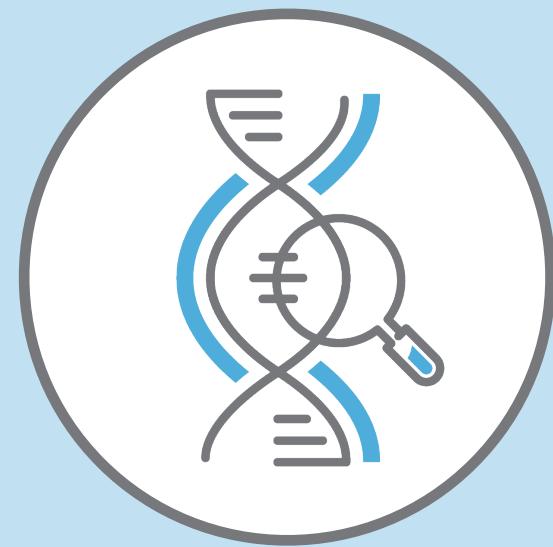
Global Genomic Data Sharing Can...



Demonstrate patterns in health & disease



Increase statistical significance of analyses



Lead to “stronger” variant interpretations



Increase accurate diagnosis



Advance precision medicine

Different Approaches to Data Sharing



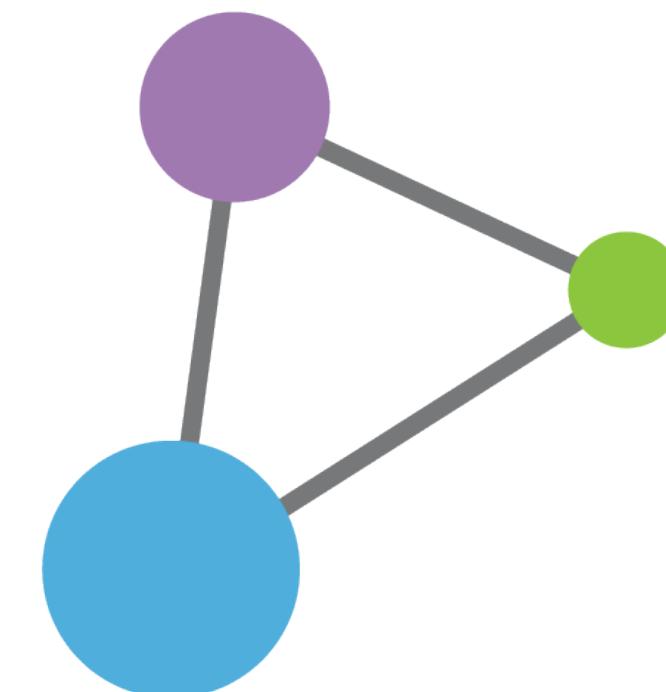
Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets



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



Linkage of distributed and disparate datasets

Different Approaches to Data Sharing



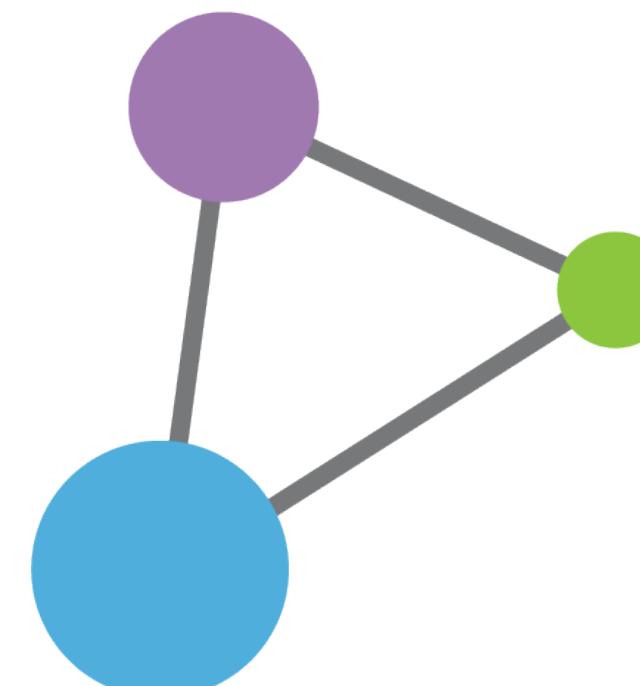
Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets

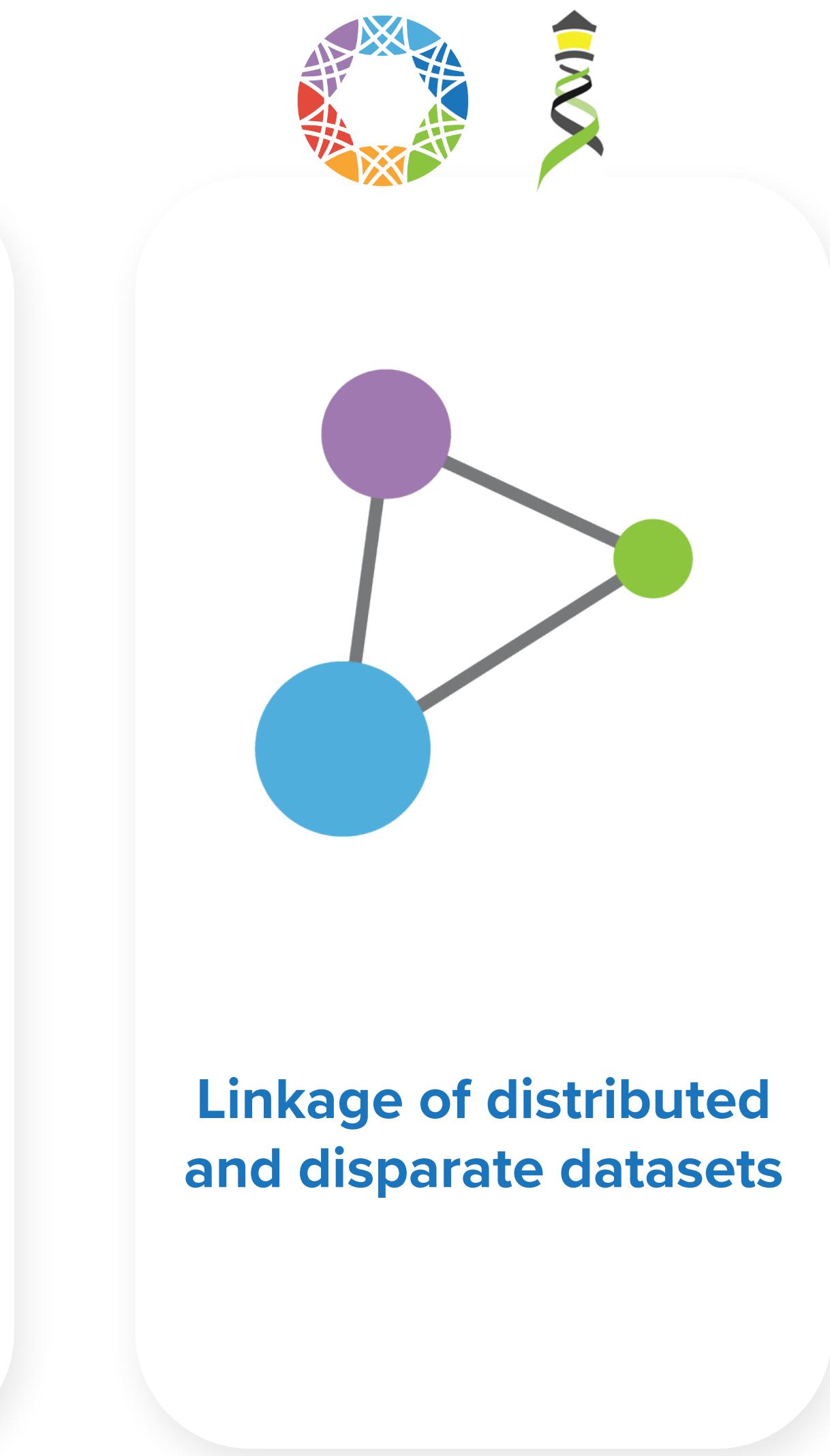
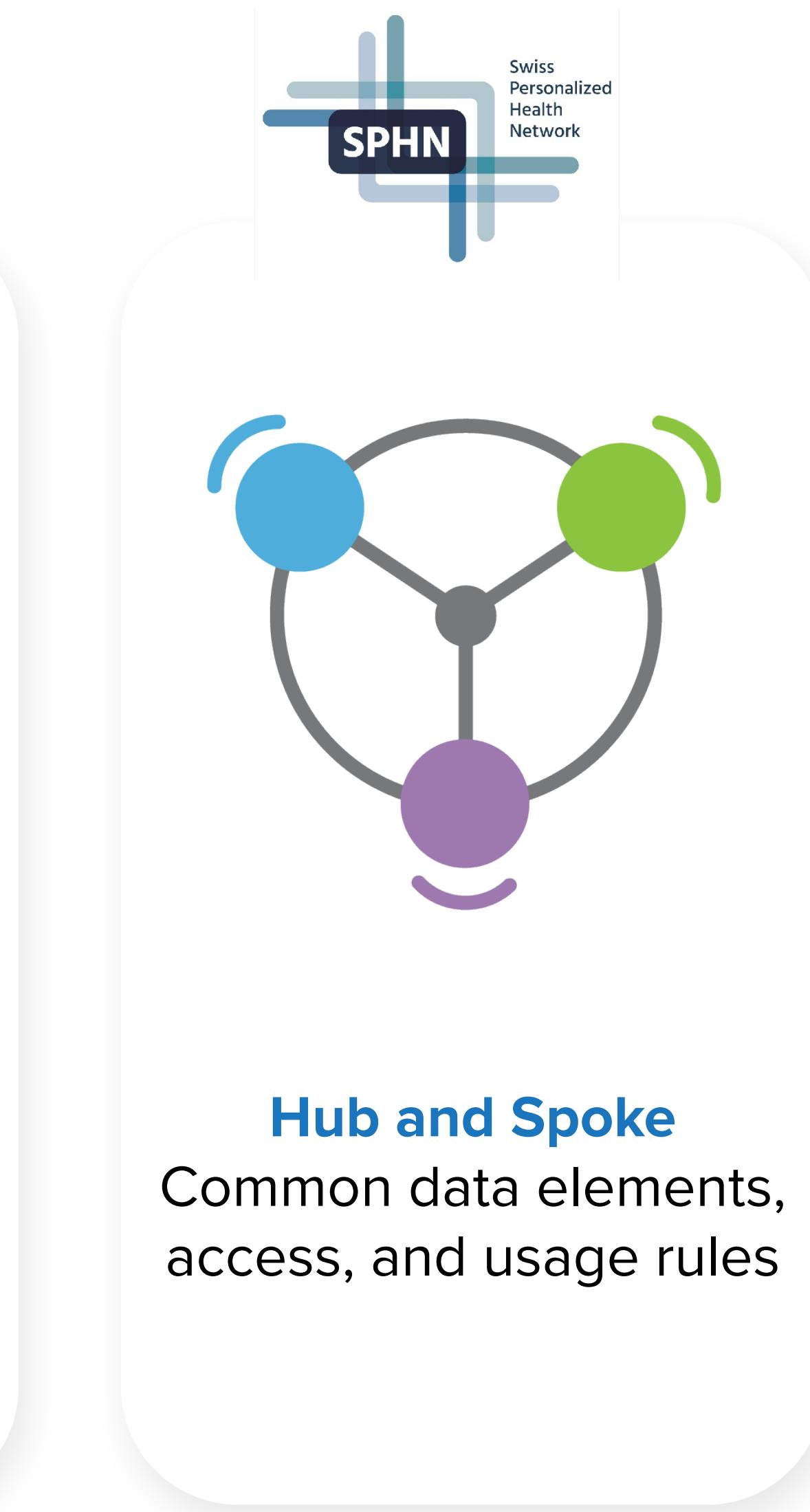
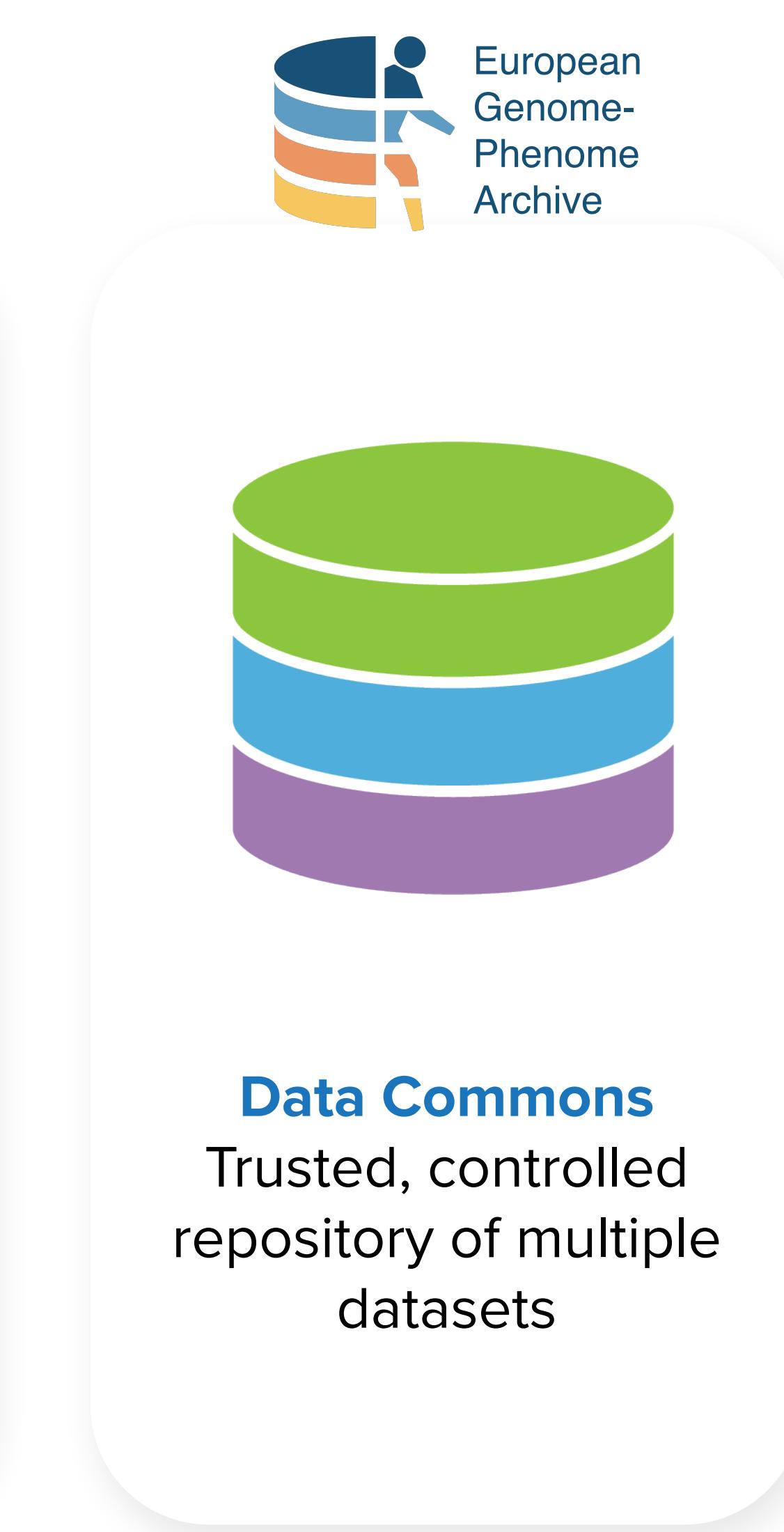
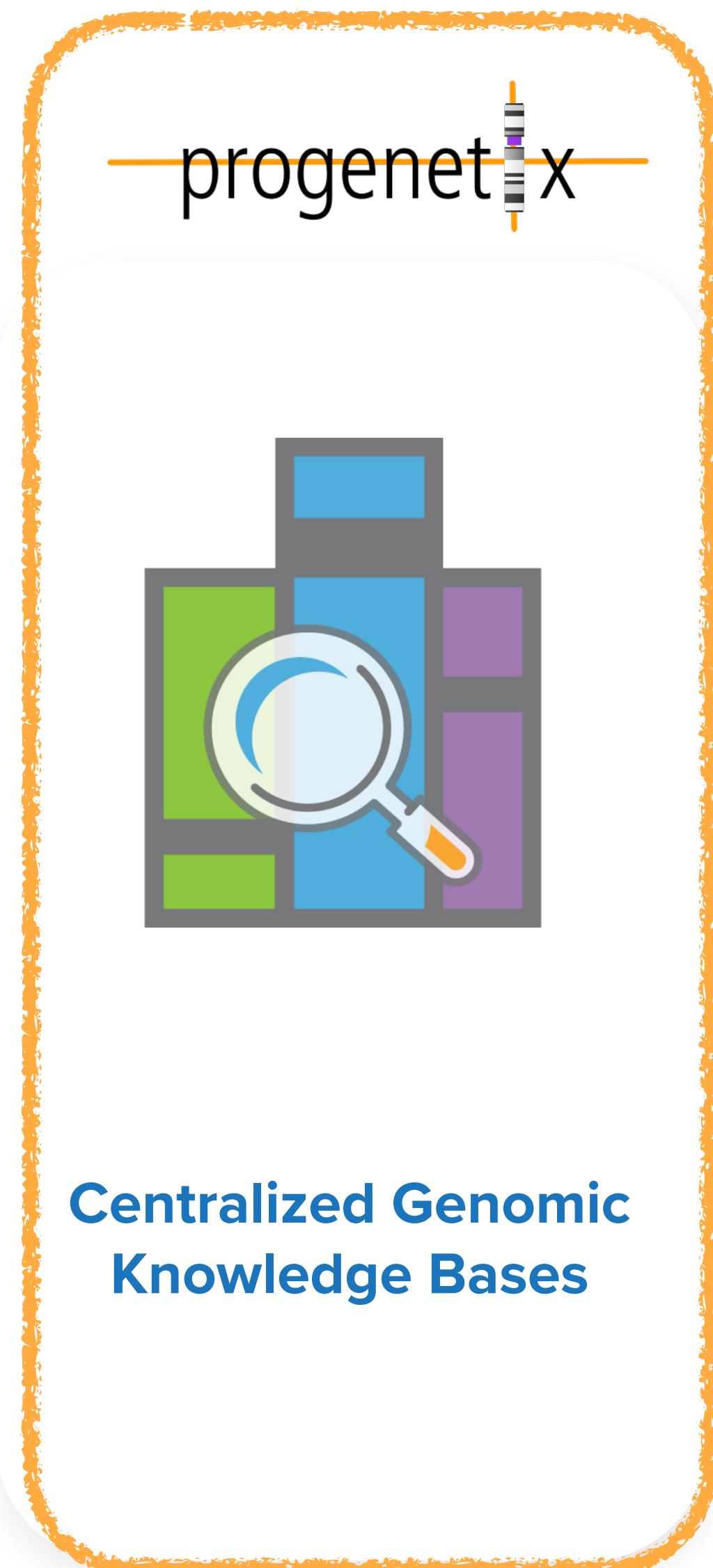


Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

Different Approaches to Data Sharing



Different Approaches to Data Sharing



Centralized Genomic Knowledge Bases



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



Linkage of distributed and disparate datasets

The EGA



Long term secure archive for human biomedical research sensitive data, with focus on reuse of the data for further research (or “*broad and responsible use of genomic data*”)

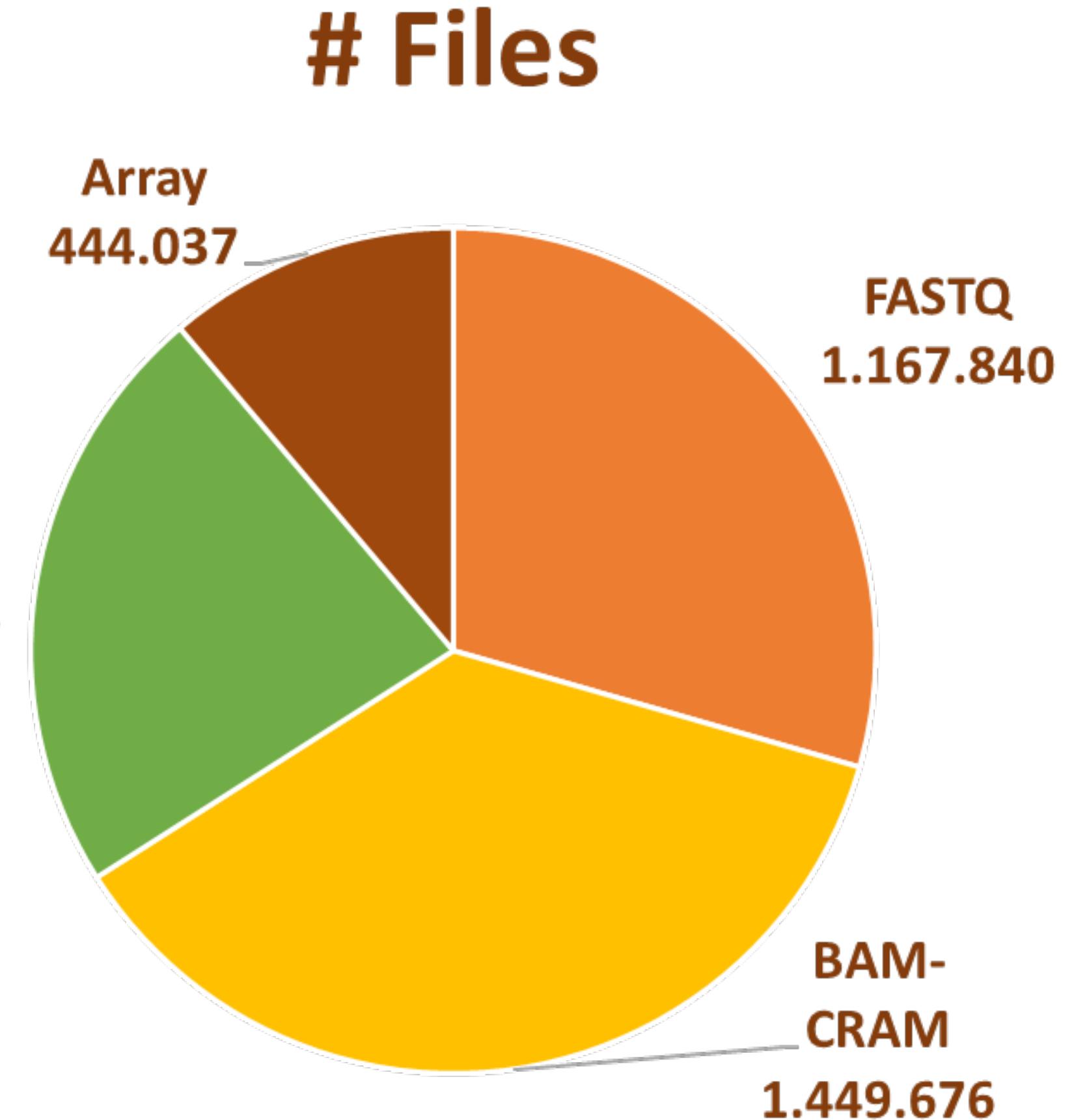


The EGA



- EGA “owns” nothing; data controllers tell who is authorized to access ***their*** datasets
- EGA admins provide smooth “all or nothing” data sharing process

A screenshot of the EGA DAC interface. At the top, it says "My DACs - EGAC5000000005 - Requests" and "EuCanImage DAC". Below that, it says "This is a DAC for EuCanImage data". A search bar says "Type something for filter the requests...". A "REQUESTS" button is visible. The main area shows a table of requests with columns: Date, Requester, Dataset, and DAC Admin/Member. The table contains three rows of data: 18 August 2022 (gemma.milla@crg.eu), 17 August 2022 (Dr Teresa Garcia Lezana), and 16 August 2022 (Dr Teresa Garcia Lezana). Each row includes a "revoke permission" toggle switch. At the bottom right of the table is an "APPLY" button.



4,328 Studies released
10,470 Datasets
2,309 Data Access Committees

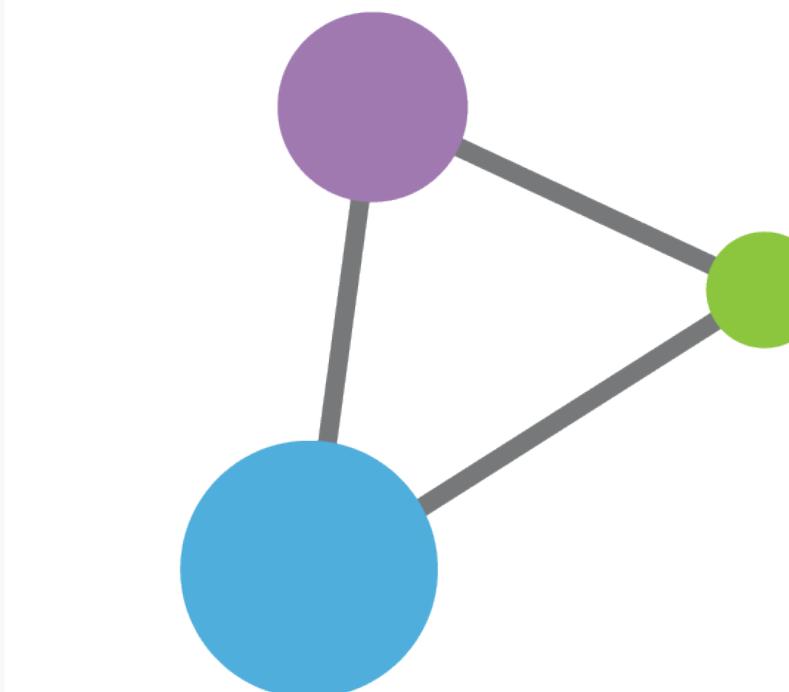
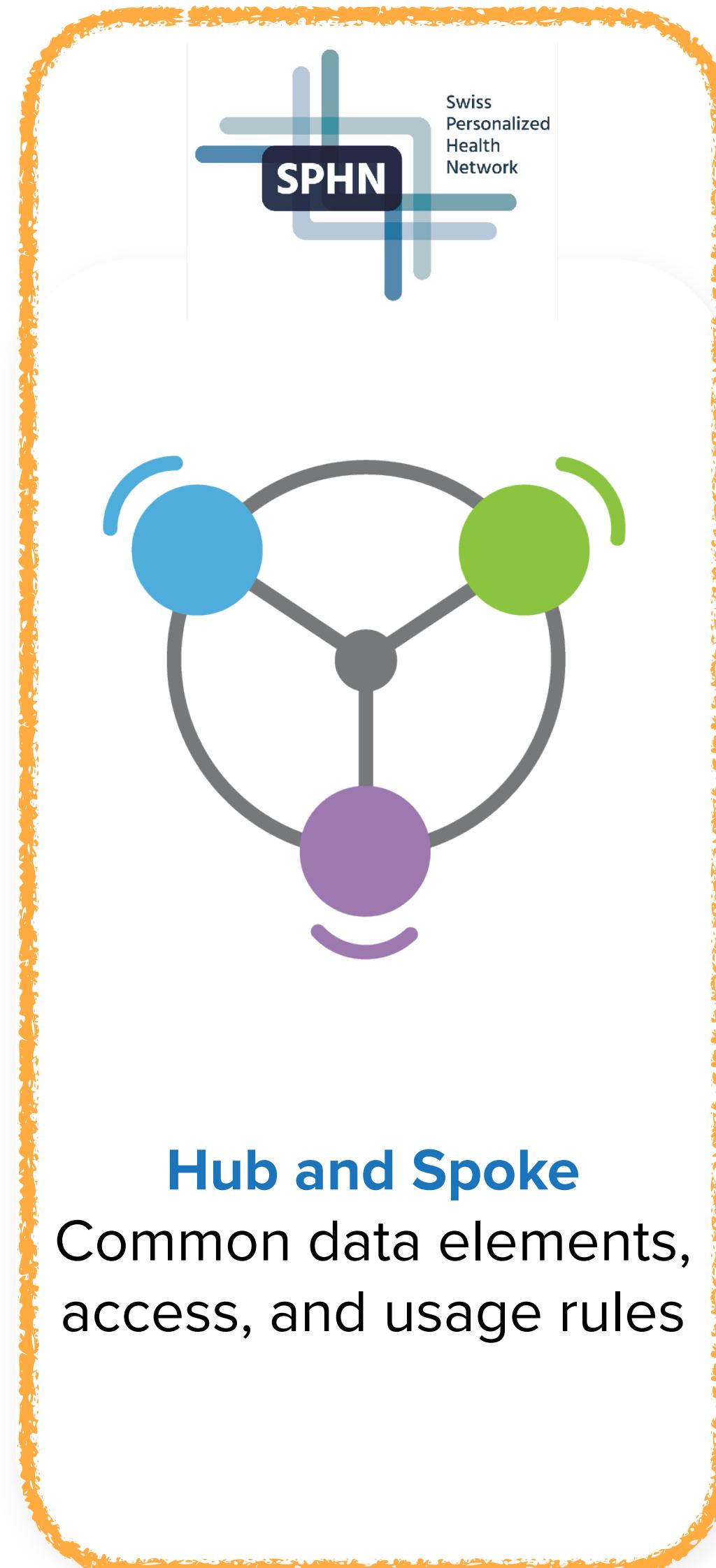
Different Approaches to Data Sharing



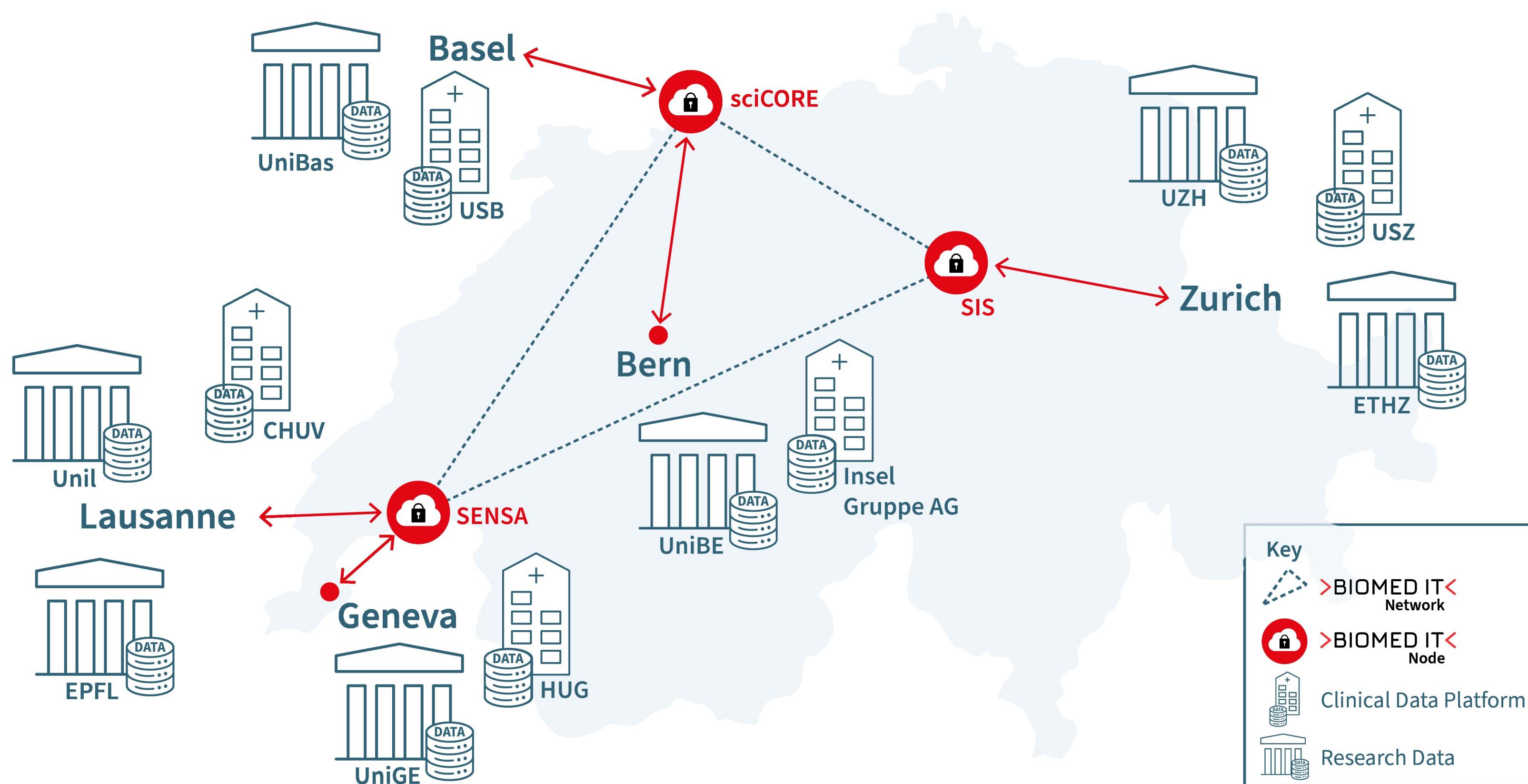
Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets



The Swiss Personalized Health Network



 **Personalized Health Informatics Group**
SPHN Data Coordination Center (DCC)
BioMedIT Network

 University Hospital
Basel
 Centre hospitalier
universitaire vaudois

 **USZ** Universitäts
Spital Zürich
 **HUG** Hôpitaux
Universitaires
Genève
 **INSELSPITAL**
UNIVERSITÄTSSPITAL BERN
HOPITAL UNIVERSITAIRE DE BERNE
BERN UNIVERSITY HOSPITAL

Strategic Focus Area
Personalized Health and Related Technologies

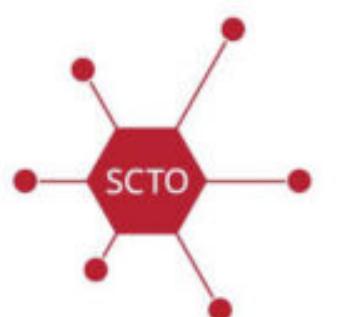
ehealthsuisse

THE LOOP
ZURICH
MEDICAL
RESEARCH
CENTER

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

Personalized Health Alliance
Basel-Zurich

 **SWISS BIOBANKING PLATFORM**



 **SAKK**
WE BRING PROGRESS TO CANCER CARE

 **SSPH+**
SWISS SCHOOL OF
PUBLIC HEALTH

 **swissuniversities**

 **Université Medizin Schweiz
Médecine Universitaire Suisse**



Different Approaches to Data Sharing



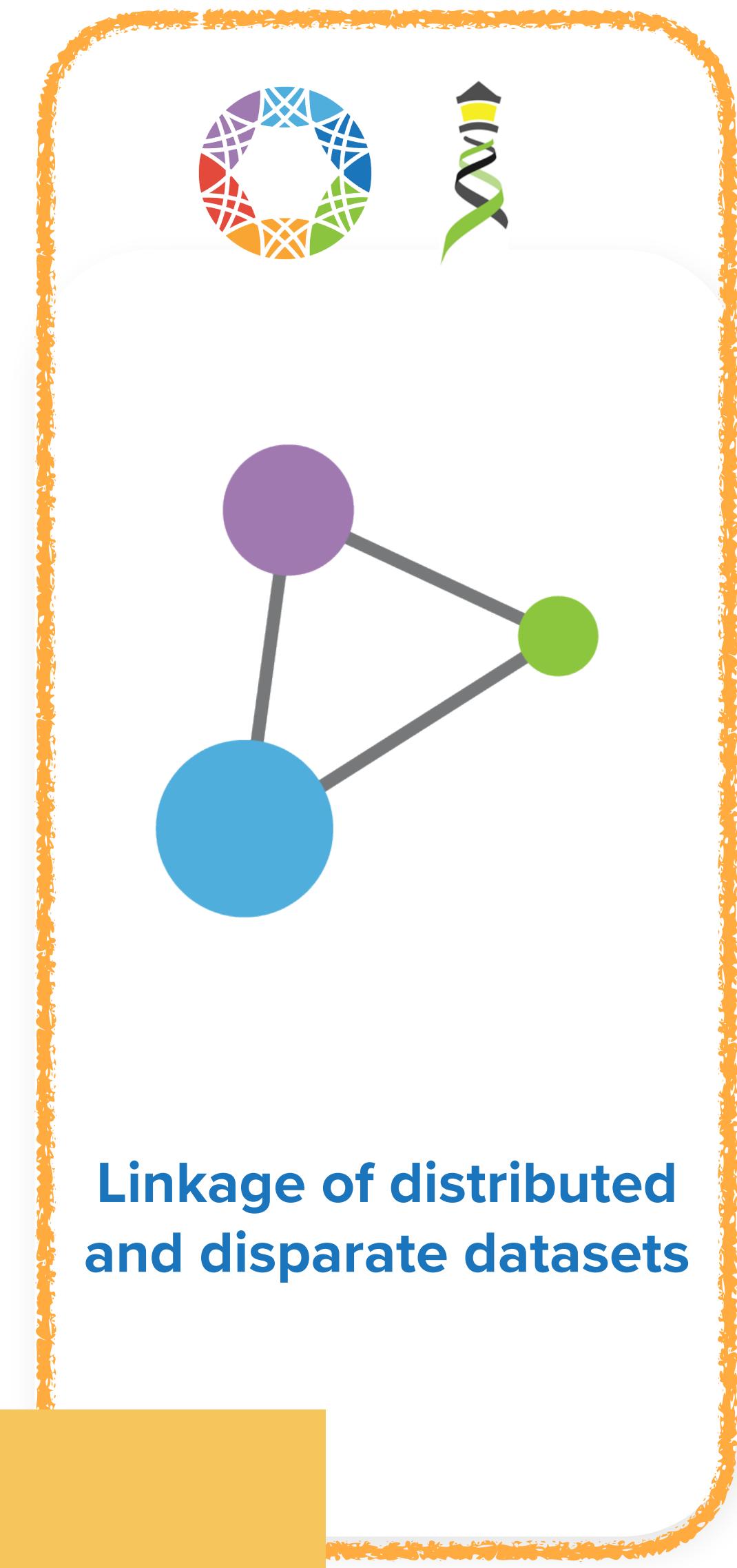
Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets



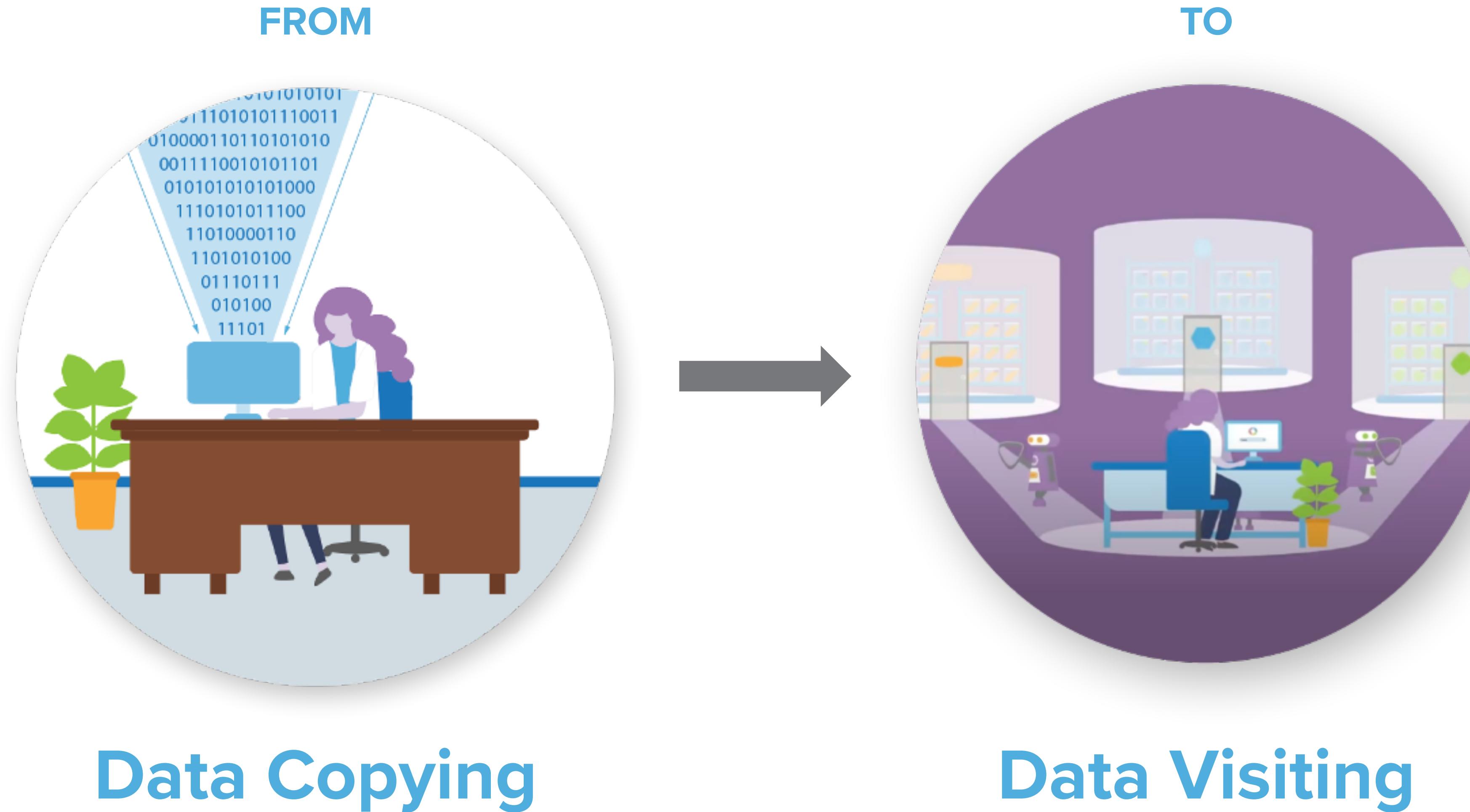
Hub and Spoke
Common data elements, access, and usage rules



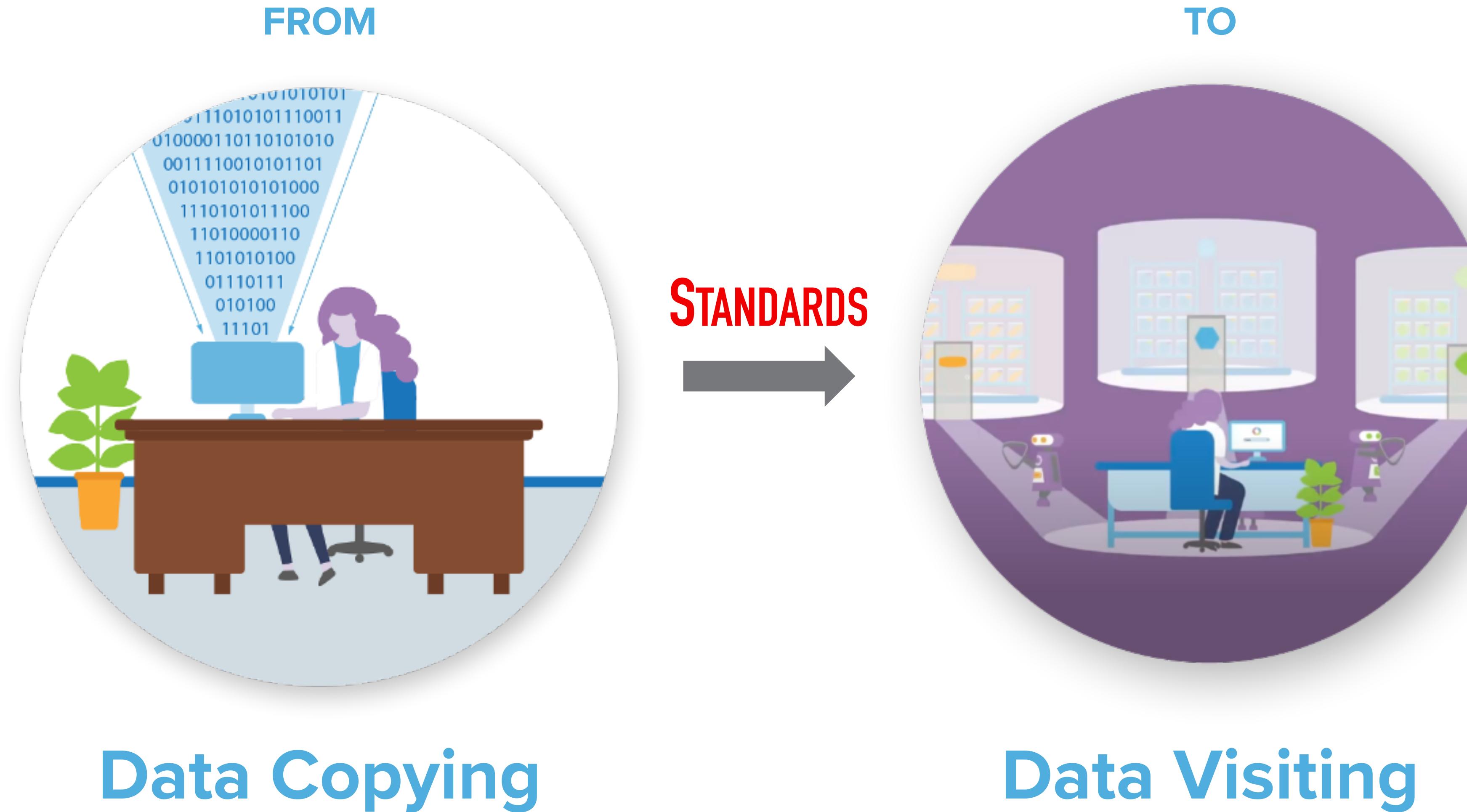
Linkage of distributed and disparate datasets

Federation

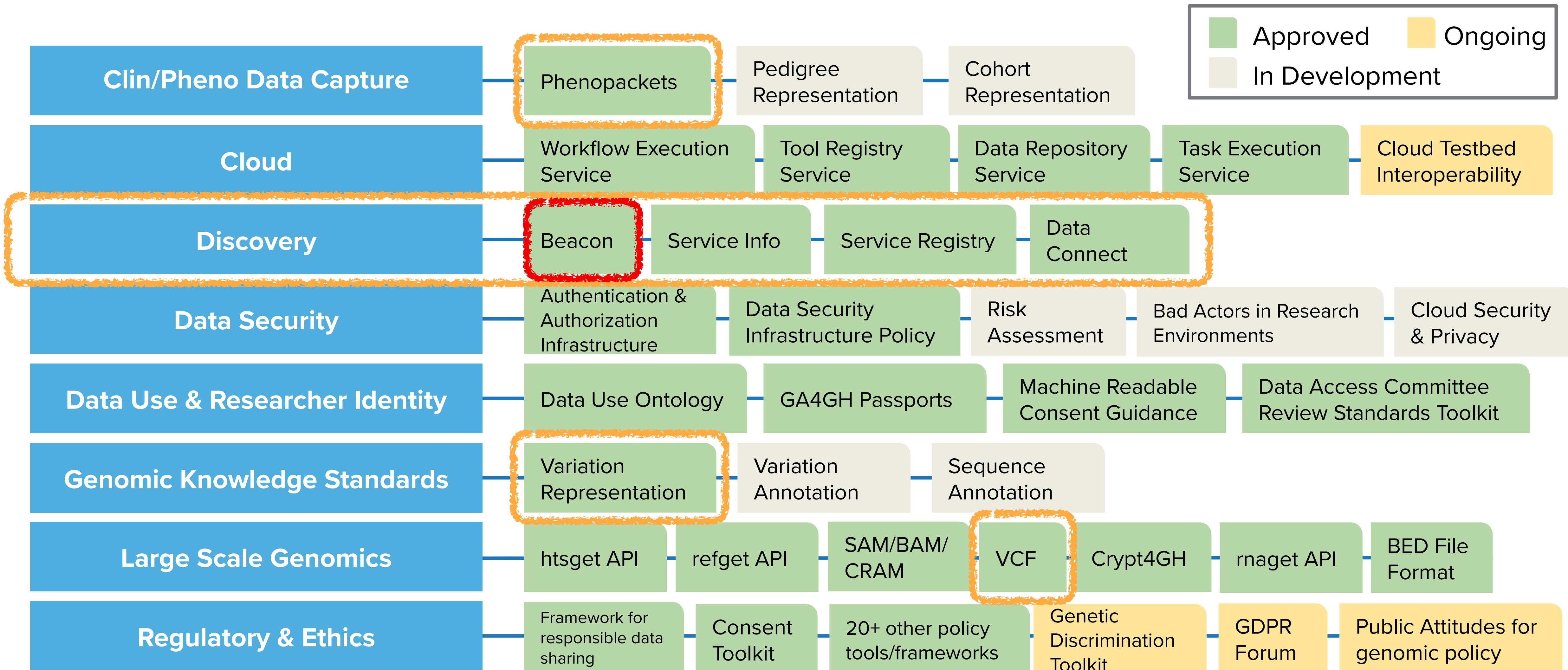
A New Paradigm for Data Sharing



A New Paradigm for Data Sharing



Overview of GA4GH standards and frameworks



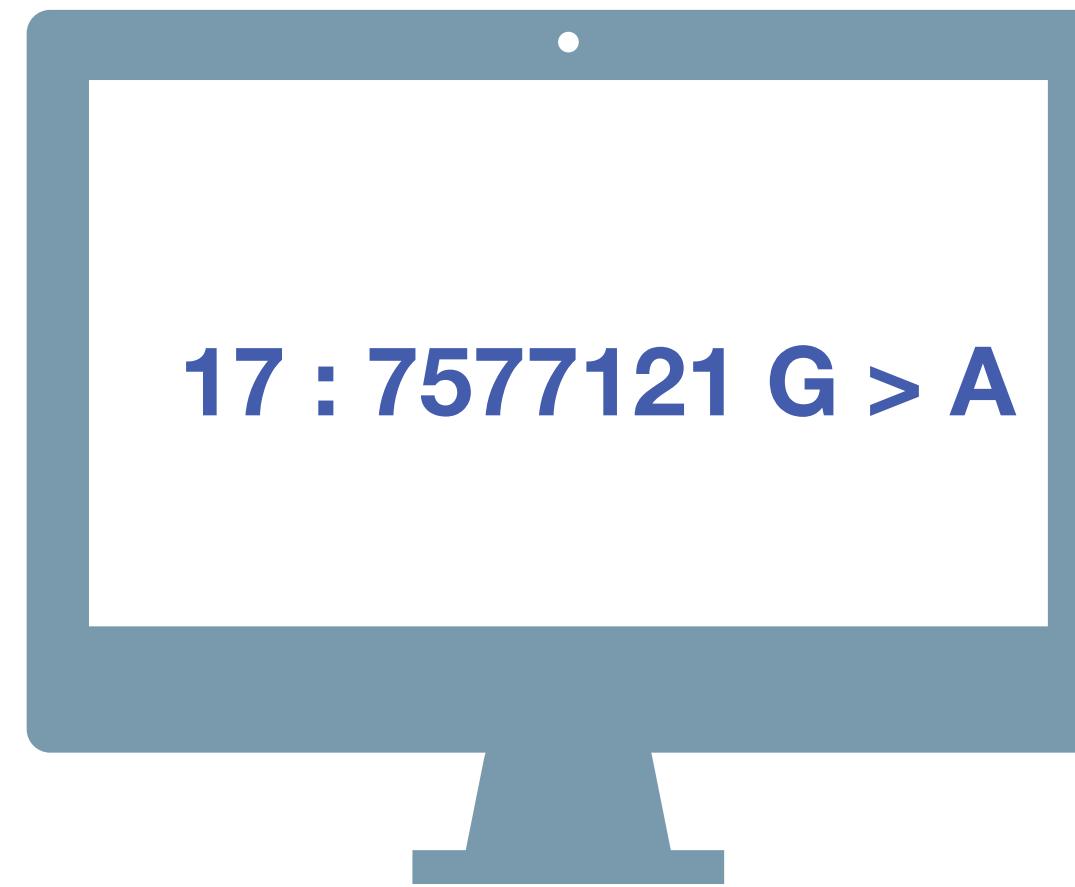


Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.



The GA4GH Beacon Protocol

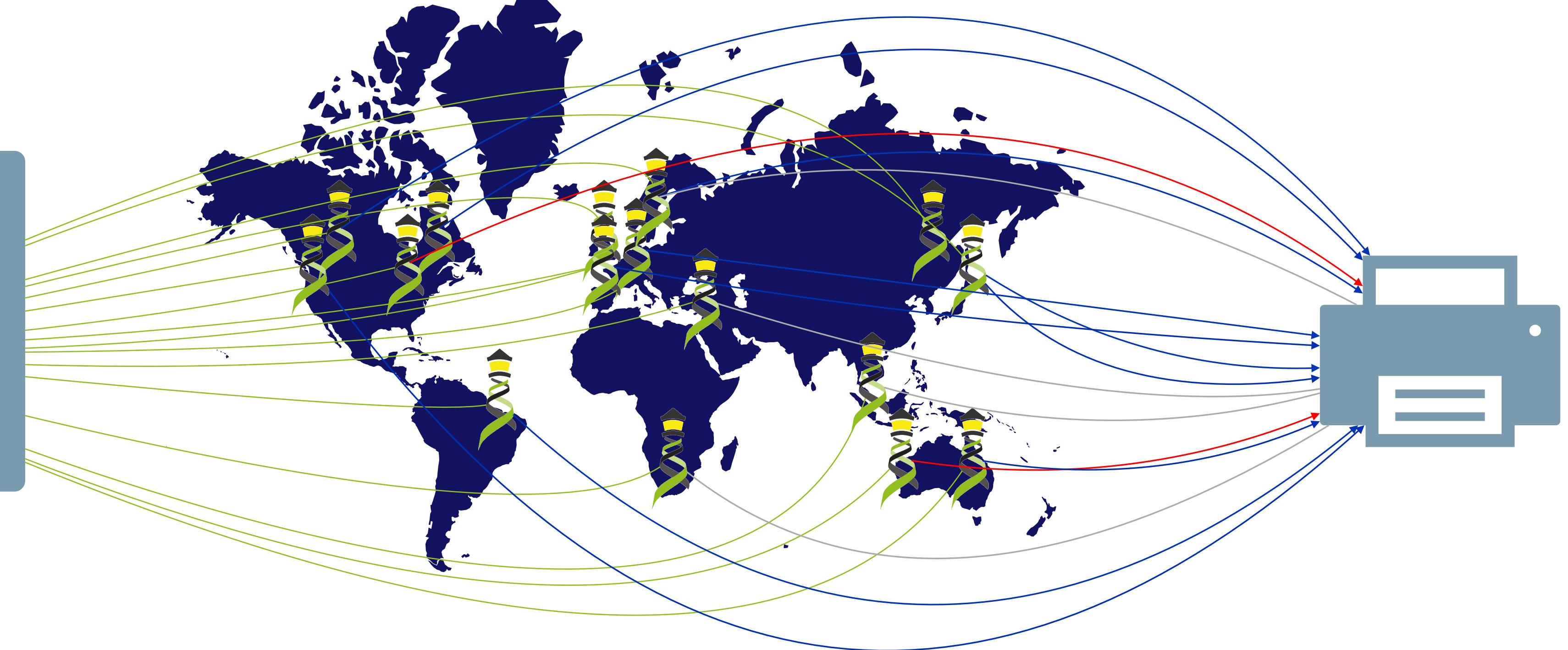
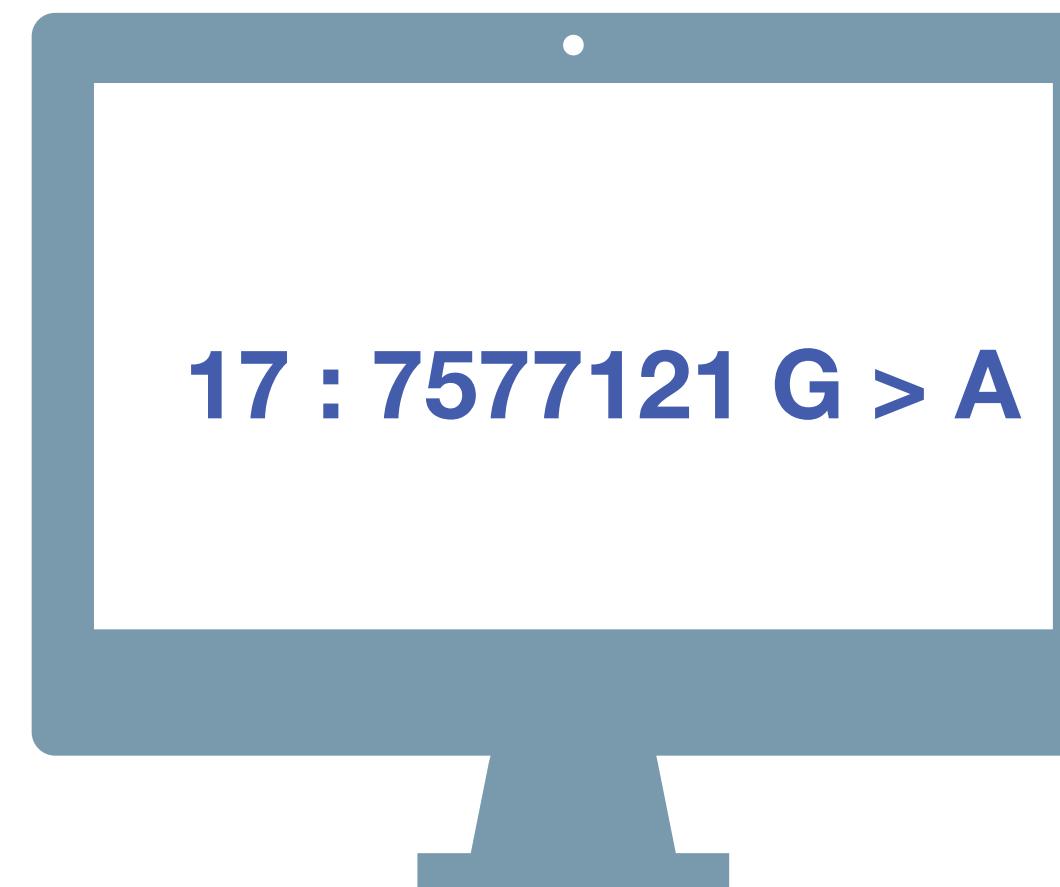
Federating Genomic Discoveries



Beacon

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

YES | NO | \0



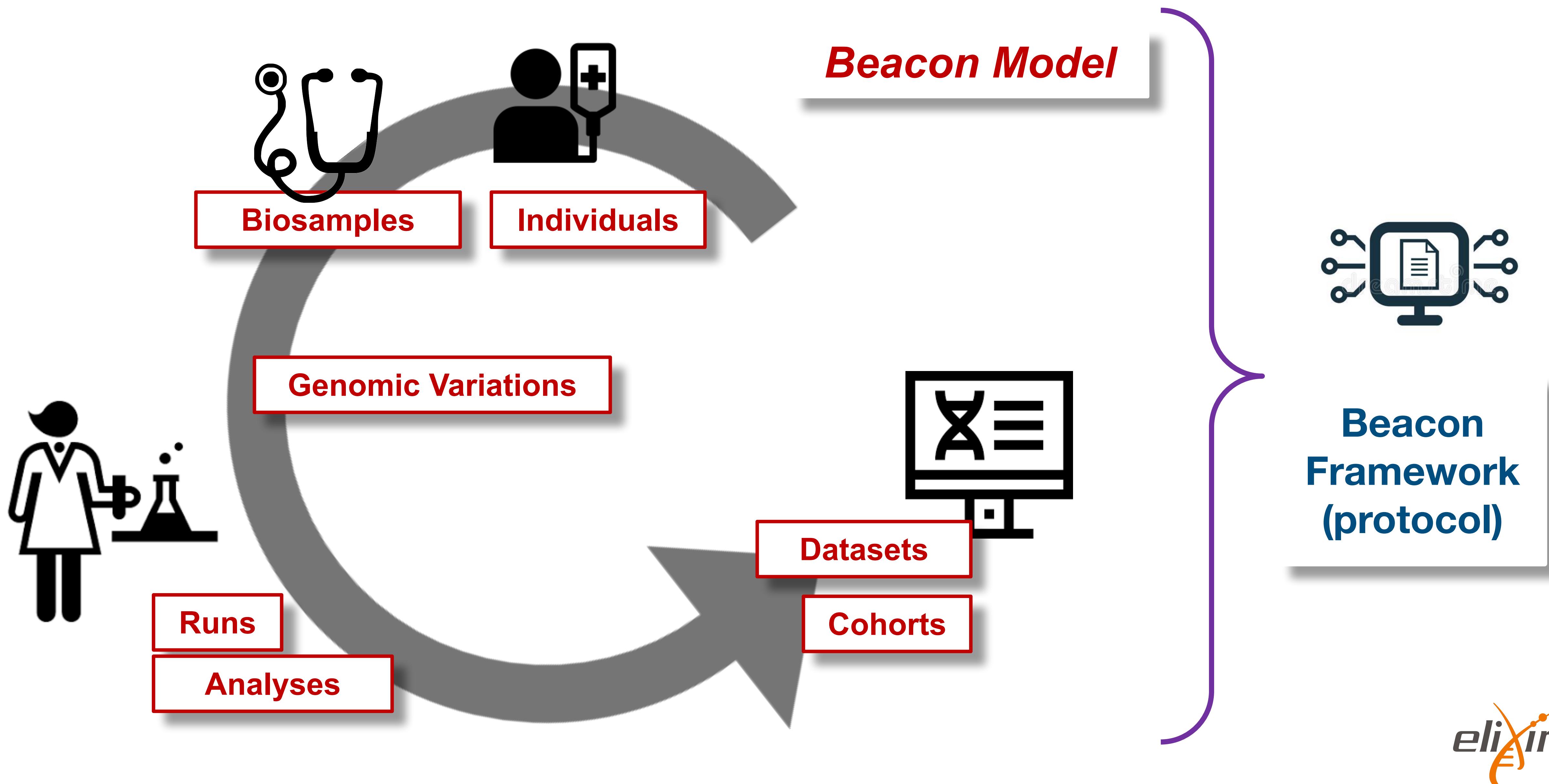
Have you seen this variant?
It came up in my patient
and we don't know if this is
a common SNP or worth
following up.

A Beacon network federates
genome variant queries
across databases that
support the **Beacon API**

Here: The variant has
been found in **few**
resources, and those
are from **disease**
specific **collections**.

Beacon v2

docs.genomebeacons.org



Beacon v1 Development

2014

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

2015

- beacon-network.org aggregator created by DNAstack

2016

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

2017

- OpenAPI implementation
- integrating CNV parameters (e.g. "startMin, statMax")

2018

- Beacon v0.4 release in January; feature release for GA4GH approval process
- GA4GH Beacon v1 approved at Oct plenary

2019

- ELIXIR Beacon Network

2020

2021

2022

Beacon v2 Development

Related ...

- ELIXIR starts Beacon project support

- GA4GH re-structuring (workstreams...)
- Beacon part of Discovery WS

- new Beacon website (March)

- Beacon publication at Nature Biotechnology

- Phenopackets v2 approved

- docs.genomebeacons.org

- Beacon+ concept implemented on progenetix.org
- concepts from GA4GH Metadata (ontologies...)
- entity-scoped query parameters ("individual.age")

- Beacon+ demos "handover" concept

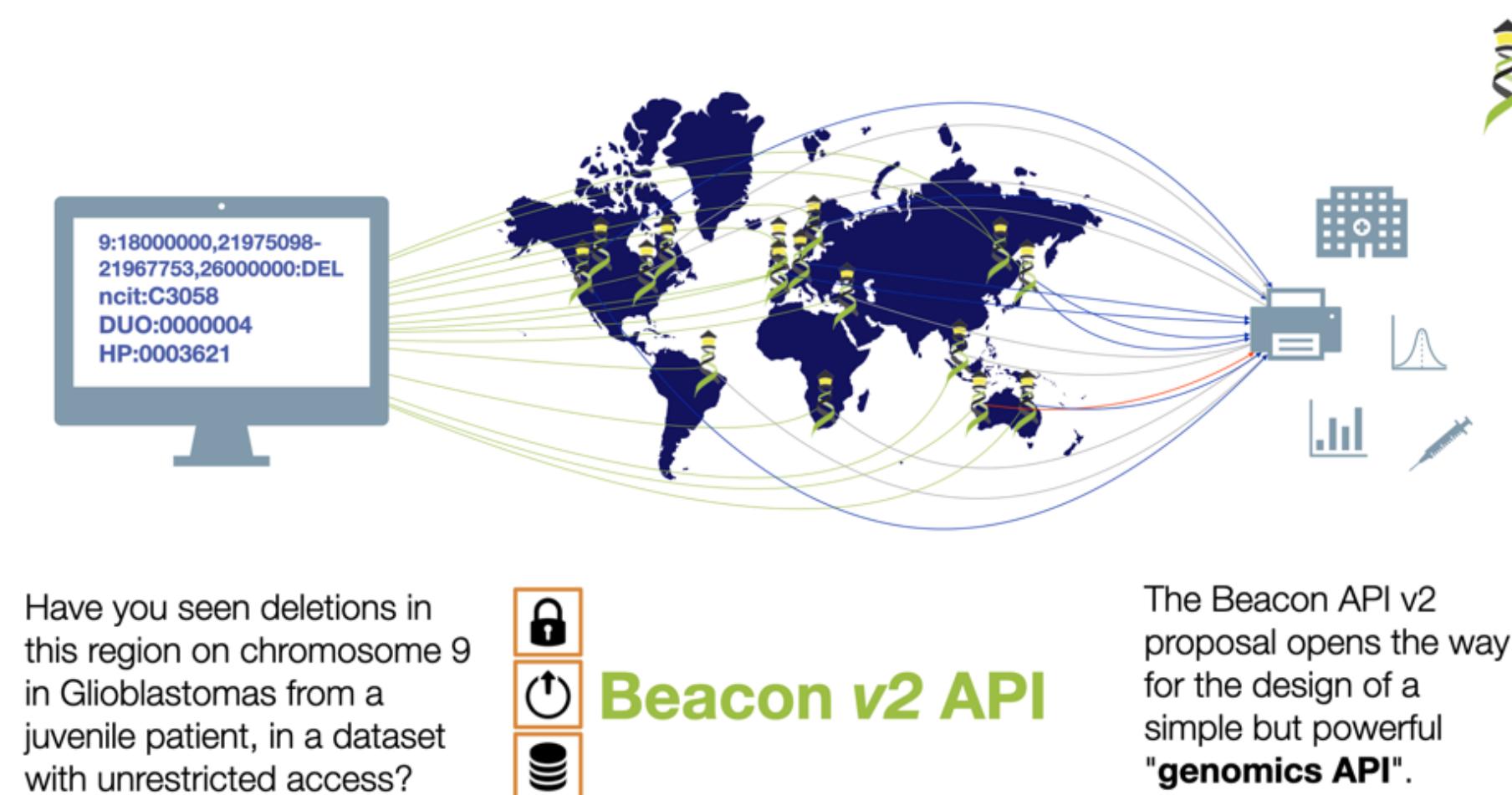
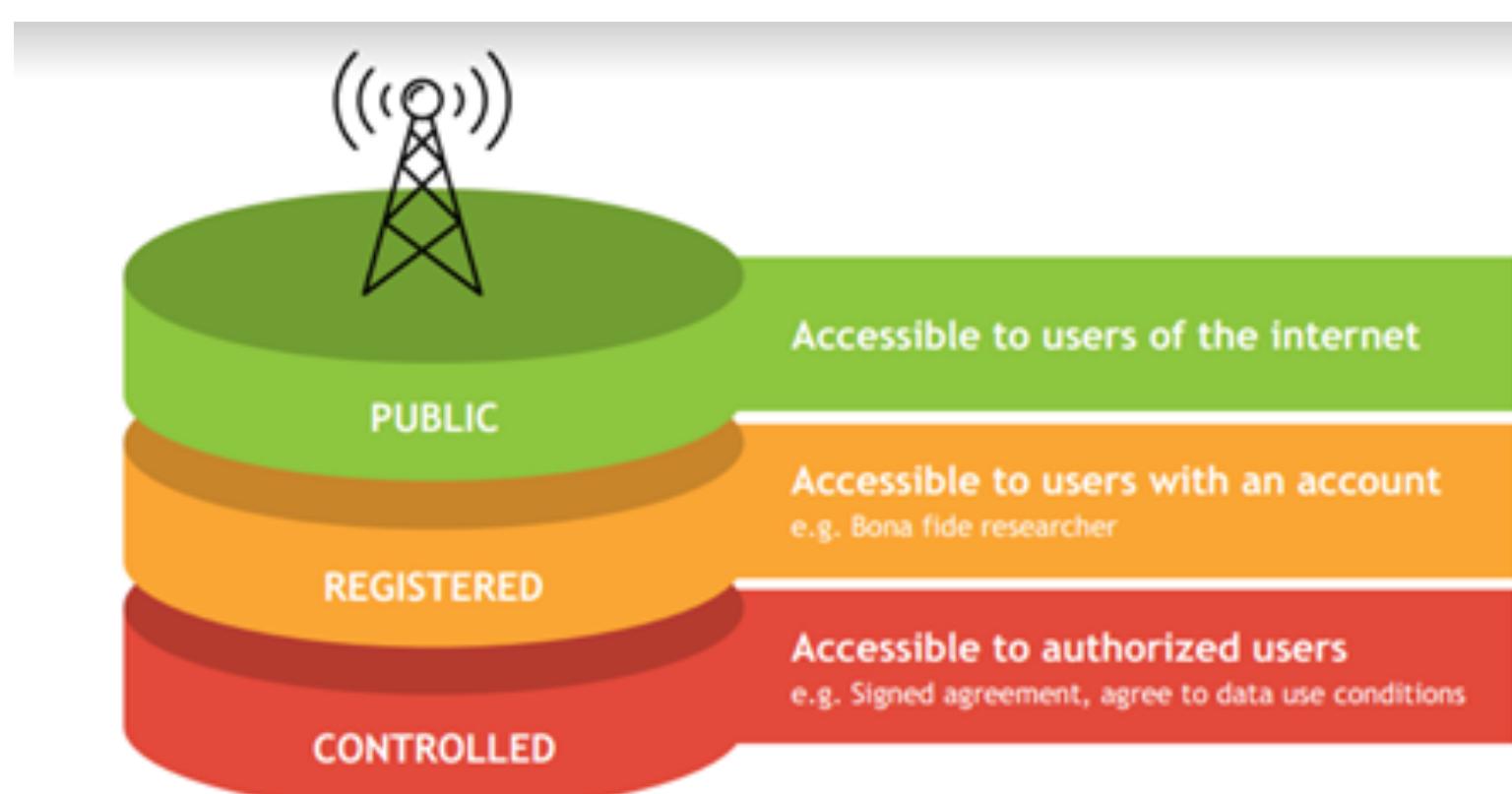
- Beacon hackathon Stockholm; settling on "filters"
- Barcelona goes Zurich developers meeting
- Beacon API v2 Kick off
- adopting "handover" concept
- "Scouts" teams working on different aspects - filters, genomic variants, compliance ...
- discussions w/ clinical stakeholders

- framework + models concept implemented
- range and bracket queries, variant length parameters
- starting of GA4GH review process
- further changes esp. in default model, aligning with Phenopackets and VRS
- unified beacon-v2 code & docs repository
- Beacon v2 approved at Apr GA4GH Connect

Beacon API v2

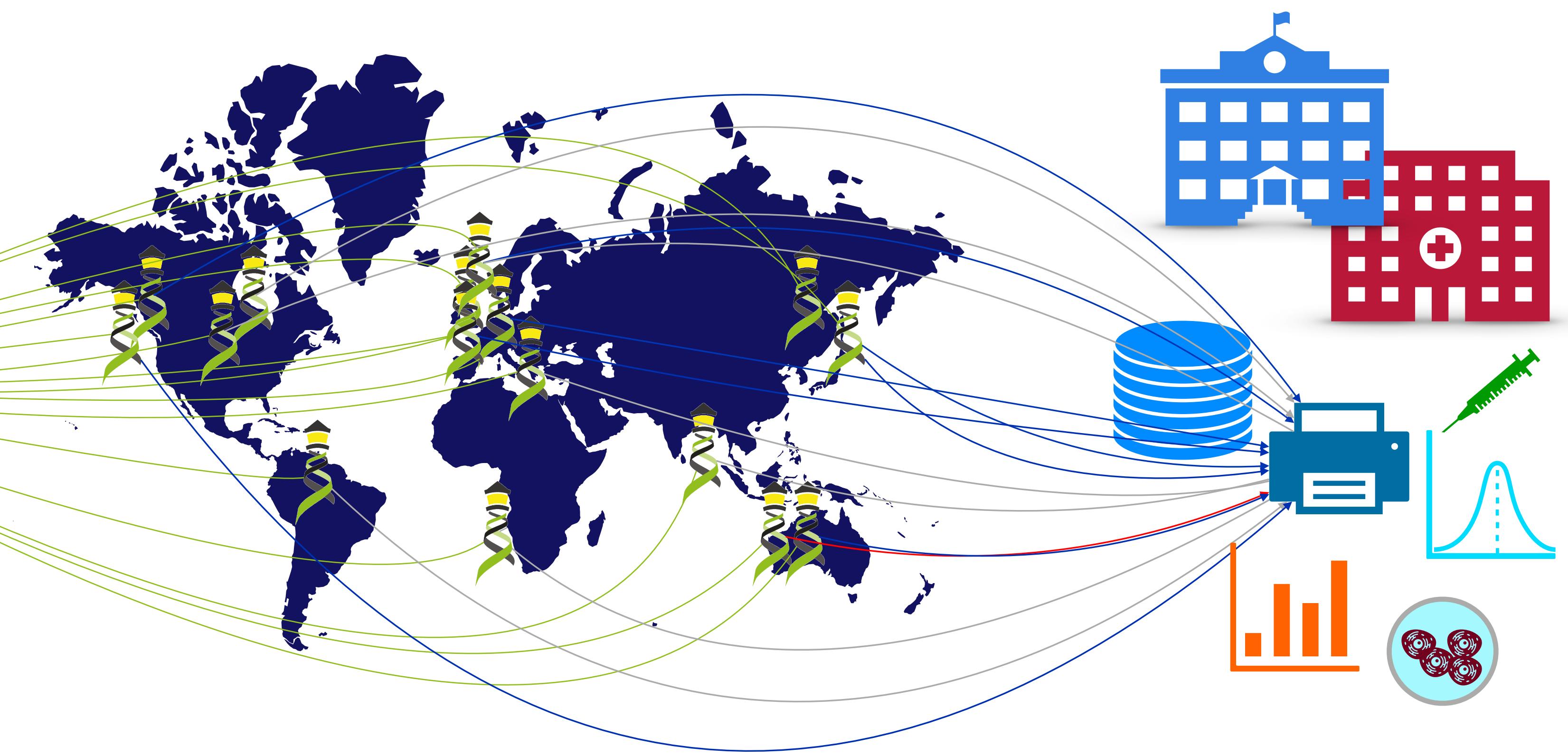
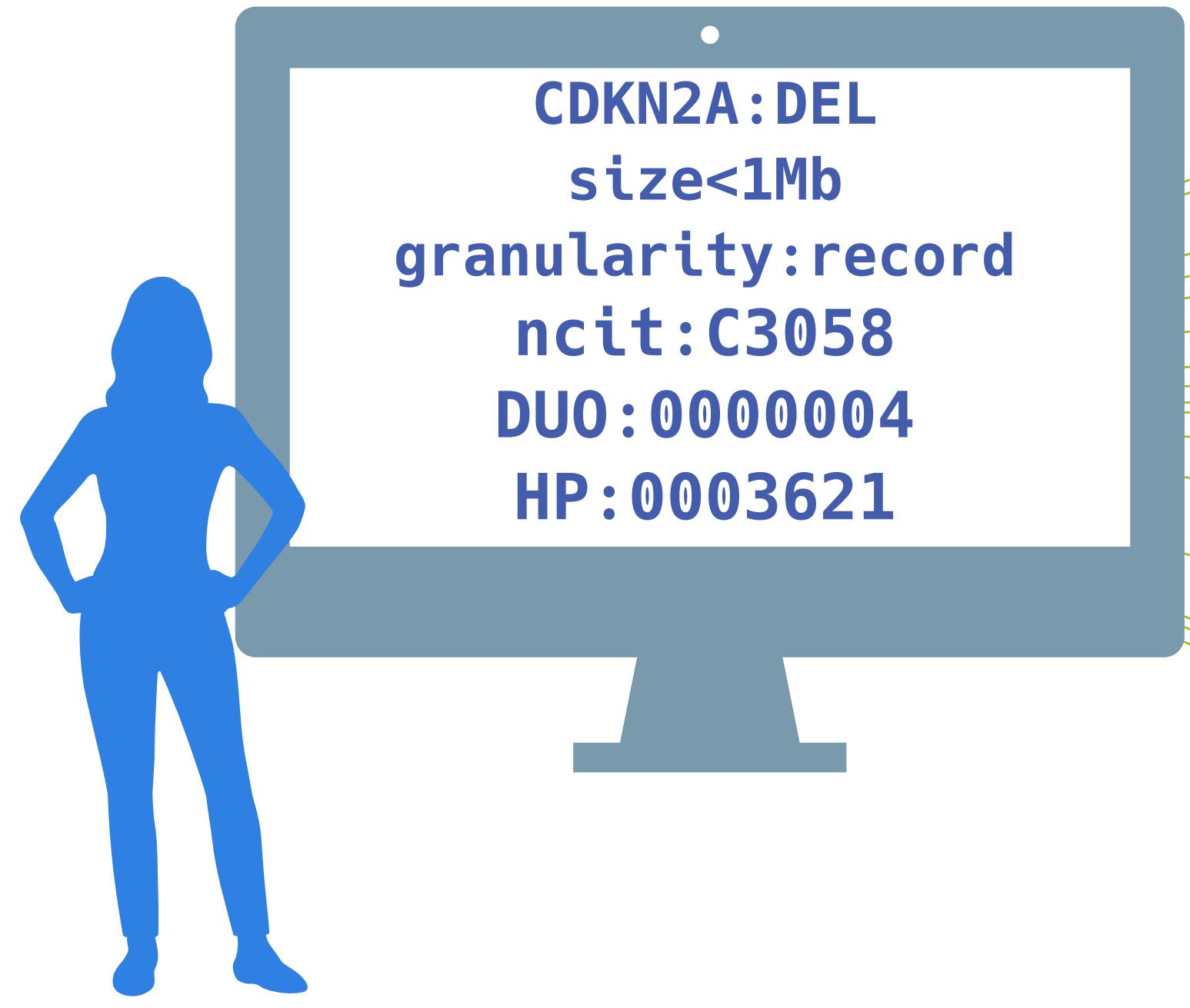
The Beacon API can be implemented as a web-accessible service that users may query for information about a specific allele.

Approved: April 21, 2022

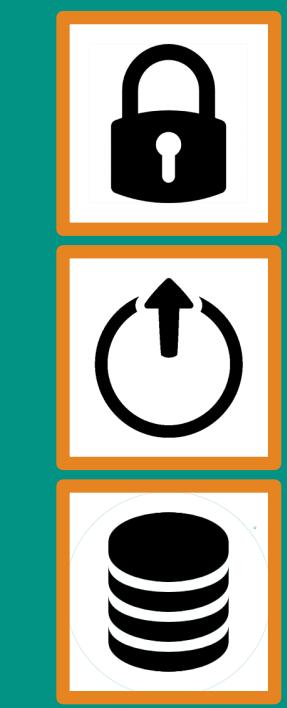


Example Users





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



Beacon v2 API

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

Progenetix and GA4GH Beacon

Implementation driven development of a GA4GH standard



Progenetix & Beacon

Implementation driven standards development

- Progenetix Beacon+ has served as implementation driver since 2016
 - prototyping of advanced Beacon features such as
 - structural variant queries
 - data handovers
 - Phenopackets integration



progenetix



Beacon v2 GA4GH Approval Registry

European Genome-Phenome Archive (EGA)

GA4GH Approval Beacon Test

This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

BeaconMap	
Bioinformatics analysis	
Biological Sample	
Cohort	
Configuration	
Dataset	
EntryTypes	
Genomic Variants	
Individual	
Info	
Sequencing run	

Theoretical Cytogenetics and Oncogenomics group at UZH and SIB

Progenetix Cancer Genomics Beacon+ Beacon+ provides a forward looking implementation of the Beacon v2 API, with focus on structural genome variants and metadata based on the...

BeaconMap	
Bioinformatics analysis	
Biological Sample	
Cohort	
Configuration	
Dataset	
EntryTypes	
Genomic Variants	
Individual	
Info	
Sequencing run	

Centre Nacional Analisis Genomica (CNAG-CRG)

Beacon @ RD-Connect

This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

BeaconMap	
Bioinformatics analysis	
Biological Sample	
Cohort	
Configuration	
Dataset	
EntryTypes	
Genomic Variants	
Individual	
Info	
Sequencing run	

University of Leicester

Cafe Variome Beacon v2

This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

BeaconMap	
Bioinformatics analysis	
Biological Sample	
Cohort	
Configuration	
Dataset	
EntryTypes	
Genomic Variants	
Individual	
Info	
Sequencing run	

Beacon protocol response verifier at time of GA4GH approval Spring 2022

Beacon protocol response verifier at time of GA4GH approval Spring 2022

Beacon+ by Progenetix

From Beacon Query to Explorative Analyses of CNV Patterns

- Since 2016 the Progenetix resource has been used to model options for Beacon development
 - 138334 individual samples from 698 cancer types
- The consistent use of hierarchical diagnostic codes allows the use of Beacon "filters" for histopathological/clinically scoped queries
- Beacon's handover protocols can be utilized for data retrieval and, well, handing over to additional services, e.g.
 - downloads
 - visualization
 - use of external services (UCSC browser display...)



Search Samples

CNV Request Allele Request Range Query All Fields

CNV Example

This query type is for copy number queries ("variantCNVrequest"), e.g. using fuzzy ranges for start and end positions to capture a set of similar variants.

Dataset: progenetix

Cohorts: Select...

Genome Assembly: GRCh38 / hg38

Gene Symbol: Select...

Reference name: 9 **(Structural) Variant Type**: DEL

Start or Position: 19000001-21975098 **End (Range or Structural Var.)**: 21967753-24000000

Minimum Variant Length: **Maximal Variant Length**:

Cancer Classification(s): Select...

Filters:

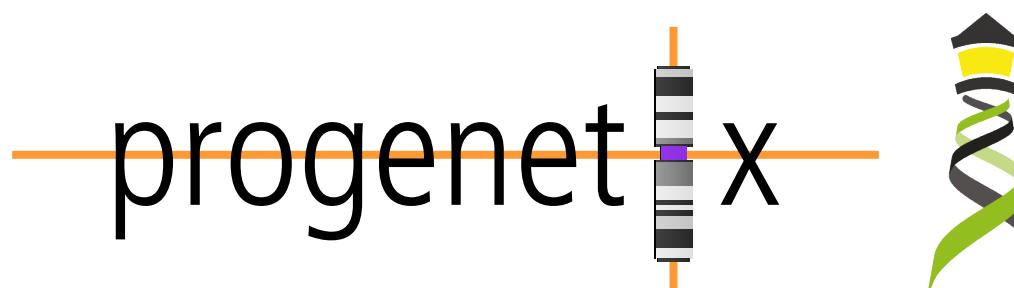
City: Select...

Query Database

Beacon v2 Filters

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

- Beacon v2 "filters" assumes inclusion of child terms when using hierarchical classifications
 - implicit *OR* with otherwise assumed *AND*
 - implementation of hierarchical annotations overcomes some limitations of "fuzzy" disease annotations



Beacon+ specific: Multiple term selection with OR logic

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

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



progenetix

Variants: 0 *f*alleles: 0 Callsets Variants ↗ UCSC region ↗ Calls: 0 Legacy Interface ↗ Samples: 523 [Show JSON Response](#)

Results **Biosamples**

Id	Description	Classifications	Identifiers	DEL	DUP	CNV
PGX_AM_BS_MCC01	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.116	0.104	0.22
PGX_AM_BS_MCC02	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.154	0.056	0.21
PGX_AM_BS_MCC03	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.137	0.21	0.347
PGX_AM_BS_MCC04	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.158	0.056	0.214
PGX_AM_BS_MCC05	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.107	0.327	0.434

Beacon Queries

Implementation of Current Options

- (so far) the Beacon model does not define explicit query types
- disambiguation of parameters is left to implementers
- implicit query types:
 - allele/sequence query
 - range query, w/ or w/o additional parameters
 - bracket query (e.g. sized CNVs)
 - aminoacid, HGVS, gene

Beacon+ Progenetix Help

Beacon Query Types

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

Dataset
Test Database - examplez x | ▾

Chromosome i Variant Type i
Select... | Select...

Start or Position i
19000001-21975098

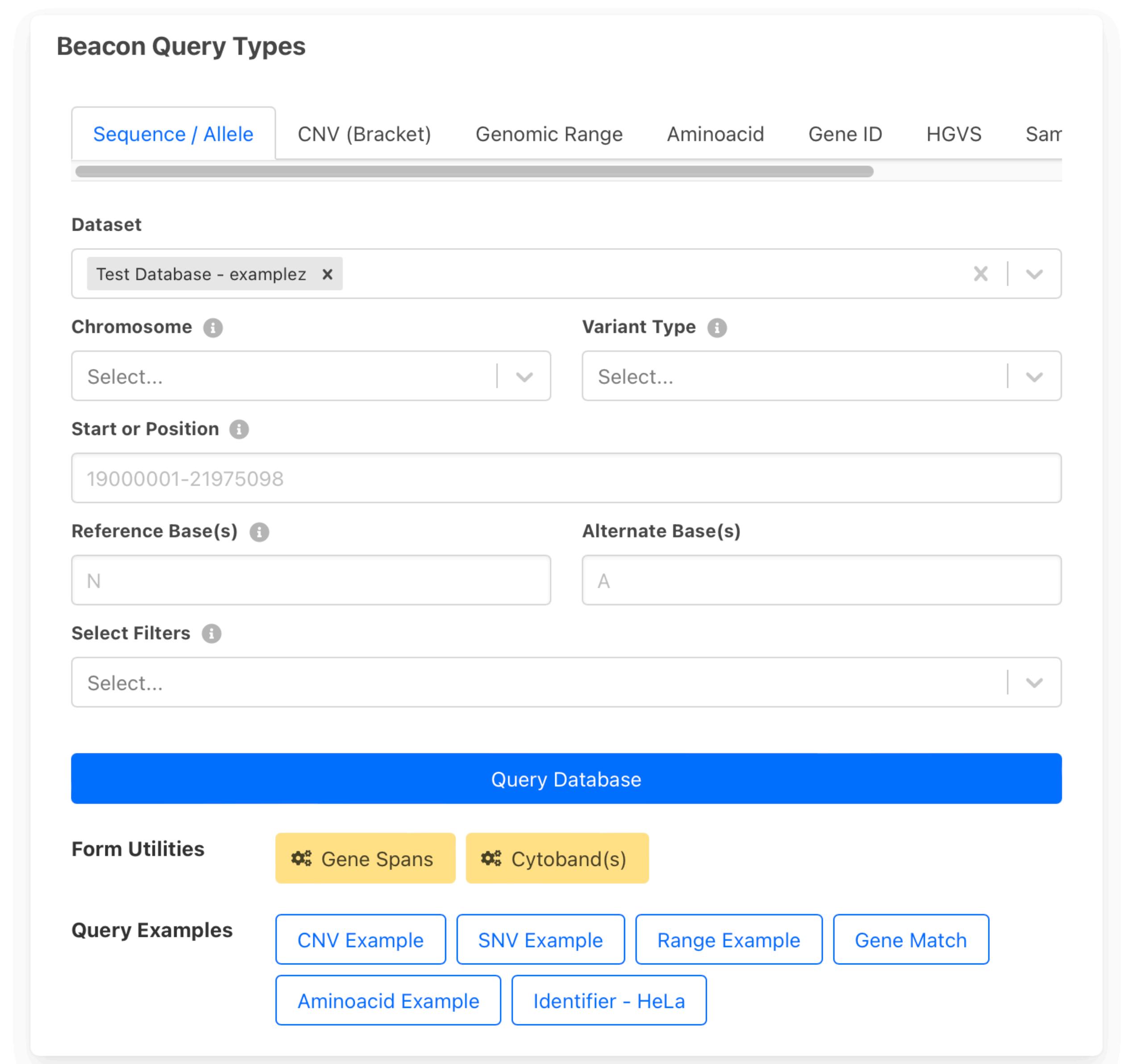
Reference Base(s) i Alternate Base(s)
N A

Select Filters i
Select...

Query Database

Form Utilities Gene Spans Cytoband(s)

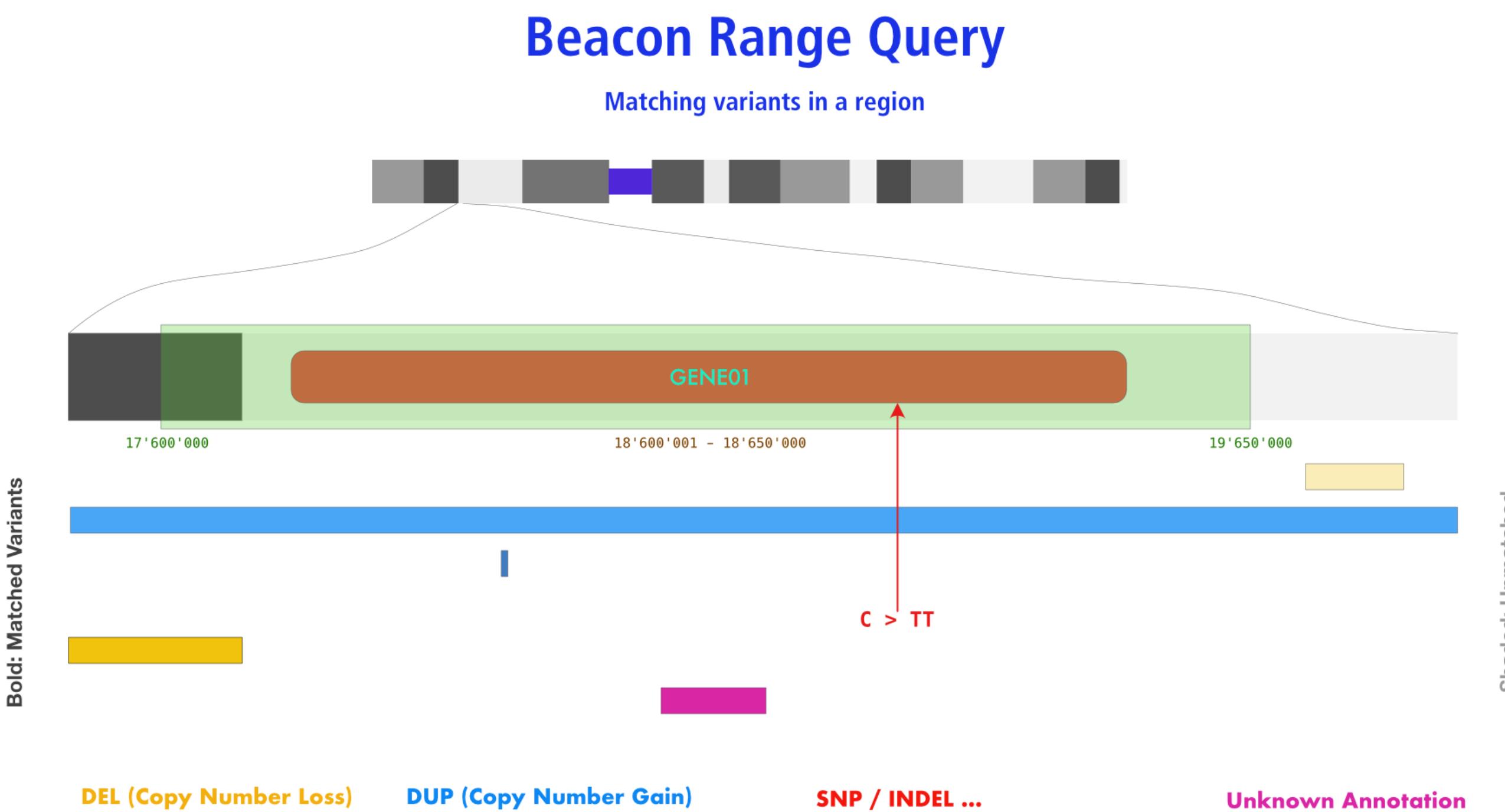
Query Examples CNV Example SNV Example Range Example Gene Match
Aminoacid Example Identifier - HeLa



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 Query Types

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

Dataset

Test Database - examplez X | ▼

Chromosome i **Variant Type** i

17 (NC_000017.11) | ▼ SO:0001059 (any sequence alteration - S...) | ▼

Start or Position i **End (Range or Structural Var.)** i

7572826 7579005

Reference Base(s) i **Alternate Base(s)**

N A

Select Filters i

Select... | ▼

Chromosome 17 i

7572826  7579005

Query Database

Form Utilities ⚙️ Gene Spans ⚙️ Cytoband(s)

Query Examples

[CNV Example](#) [SNV Example](#) [Range Example](#) [Gene Match](#)

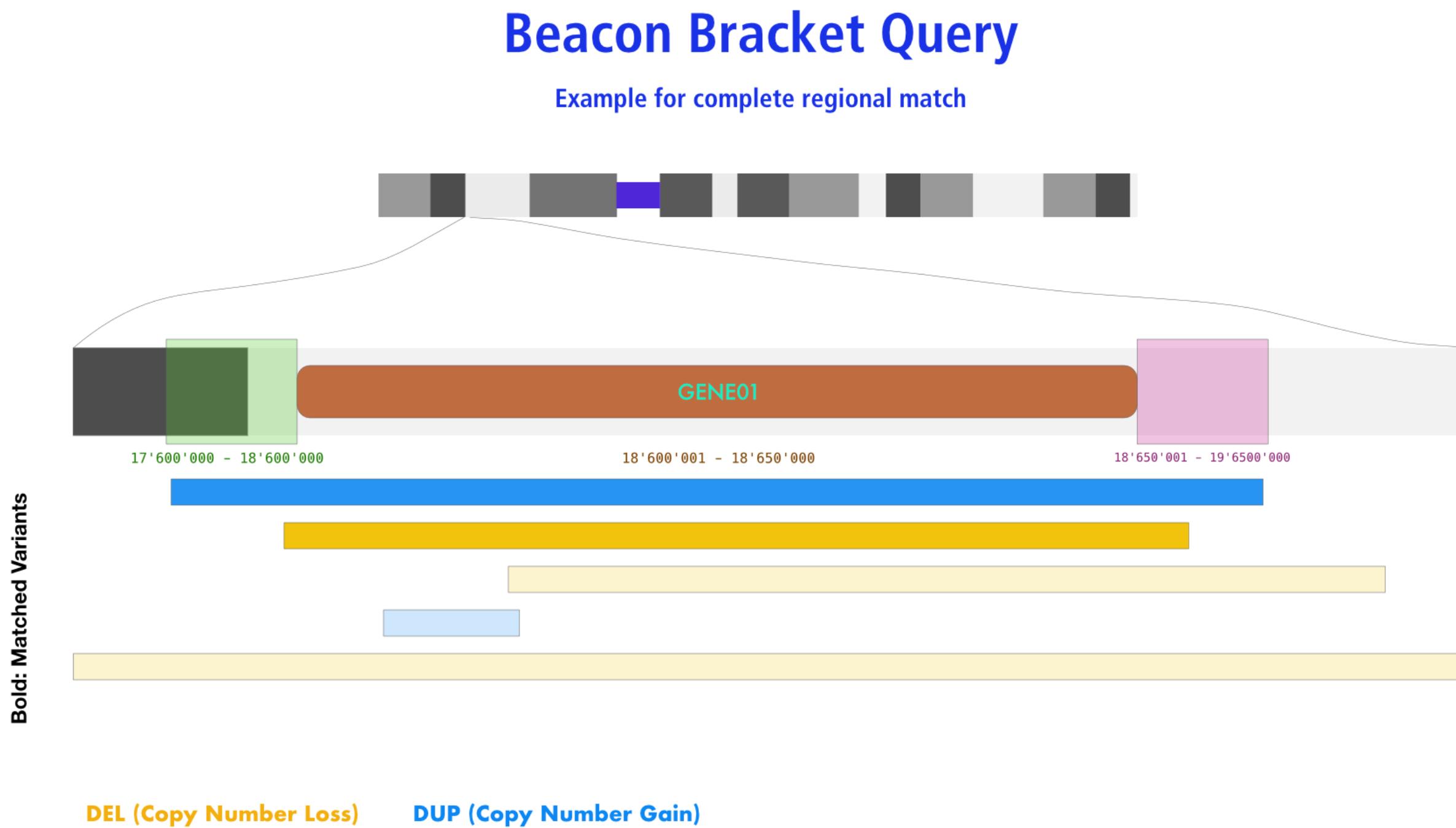
[Aminoacid Example](#) [Identifier - HeLa](#)

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](#)

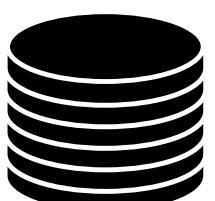
[Identifier - HeLa](#)

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.

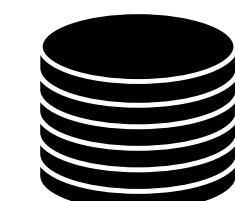
Progenetix Stack



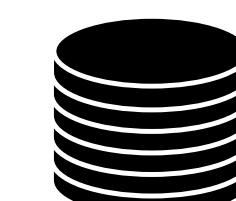
- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
 - ▶ biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads...
- the complete middleware / CGI stack is provided through the *bycon* package
 - ▶ schemas, query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



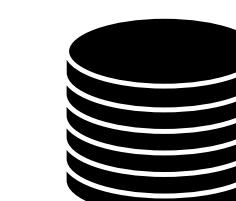
variants



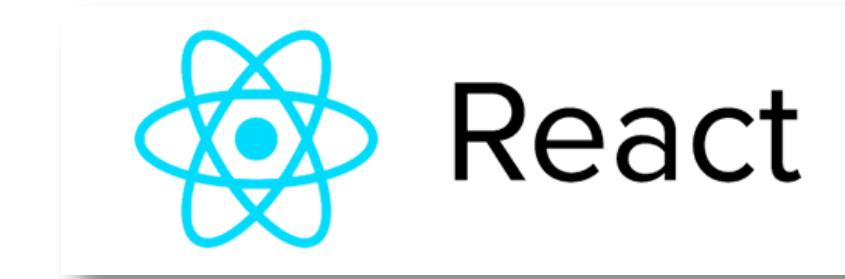
analyses



biosamples



individuals



- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**

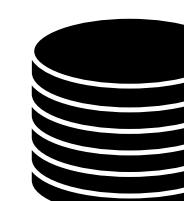


- ▶ PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703...

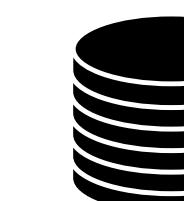


- *querybuffer* stores id values of all entities matched by a query and provides the corresponding access handle for **handover** generation

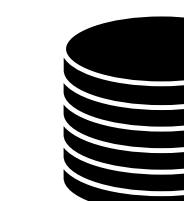
```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
  ObjectId("5bab578d727983b2e00cb505")]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



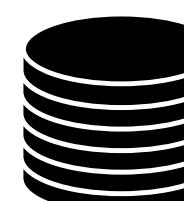
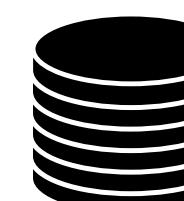
collations



geolocs



genespans publications



qBuffer



Entity collections

Utility collections

progenetix / byconaut

Type ⌘ to search

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

byconaut Public

Edit Pins Unwatch 2 Fork 1 Star 0

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

progenetix / beaconplus-web

Type ⌘ to search

Code Pull requests Actions Projects Security Insights Settings

mbaudis get_plot_parameters

bin docs exports imports local rsrc services tmp .gitignore LICENSE README.md __init__.py install.py install.yaml mkdocs.yaml

2 branches

main

beaconplus-web Public forked from progenetix/progenetix-web

main 1 branch 0 tags

This branch is 44 commits ahead, 24 commits behind progenetix:main.

beaconplus.progenetix.org
.../progenetix/beaconplus-web/

progenetix / bycon

Type ⌘ to search

Code Issues Pull requests 1 Actions Projects Wiki Security 3 Insights Settings

bycon Public

Edit Pins Unwatch 4 Fork 6 Starred 5

main 4 branches 25 tags

Go to file Add file Code

mbaudis 1.3.6 ... be19a12 3 days ago 852 commits

File	Commit	Date
.github/workflows	Create mk-bycon-docs.yaml	8 months ago
bycon	1.3.6	3 days ago
docs	1.3.6	3 days ago
local	1.3.5 preparation	2 weeks ago
.gitignore	Update .gitignore	3 months ago
LICENSE	Create LICENSE	3 years ago
MANIFEST.in	major library & install disentanglement	9 months ago
README.md	#### 2023-07-23 (v1.0.68)	4 months ago
install.py	1.3.6	3 days ago
install.yaml	v1.0.57	5 months ago
mkdocs.yaml	1.1.6	3 months ago
requirements.txt	1.3.6	3 days ago
setup.cfg	...	10 months ago
setup.py	1.3.6	3 days ago
updev.sh	1.3.6	3 days ago

About

Bycon - A Python Based Beacon API (beacon-project.io) implementation leveraging the Progenetix (progenetix.org) data model

Readme CC0-1.0 license Activity 5 stars 4 watching 6 forks Report repository

Releases

25 tags Create a new release

Packages

No packages published Publish your first package

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

pgxRpi

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

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

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>.

pgxRpi

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

The screenshot shows a GitHub repository page for 'pgxRpi'. The repository is public, has 1 branch, and 0 tags. The main branch has 87 commits from user 'hangjiaz' with the message 'version bump'. The commits are listed in descending order of age. A red arrow points to the commit log.

Commit	Message	Date
R	adapt to beacon variant export	4 days ago
data-raw	add data documentation; optimise get method which causes error in ...	3 months ago
data	change based on opinion from bioc reviewer	last month
inst	add data documentation; optimise get method which causes error in ...	3 months ago
man	documentation for parameters using match.arg	last month
tests	adapt to API change; optimise code logic	last month
vignettes	documentation for parameters using match.arg	last month
.Rbuildignore	change vignette builder; build vignette	2 months ago
.gitignore	modify gitignore	2 months ago
DESCRIPTION	version bump	4 days ago
NAMESPACE	add pgxFiler; change variant query logic and url; other code change...	last month
NEWS.md	add pgxFiler; change variant query logic and url; other code change...	last month
README.md	add pgxFiler; change variant query logic and url; other code change...	last month

2 Retrieve metadata of samples

2.1 Relevant parameters

type, filters, filterLogic, individual_id, biosample_id, codematches, limit, skip

2.2 Search by filters

Filters are a significant enhancement to the [Beacon](#) query API, providing a mechanism for specifying rules to select records based on their field values. To learn more about how to utilize filters in Progenetix, please refer to the [documentation](#).

The `pgxFiler` function helps access available filters used in Progenetix. Here is the example use:

```
# access all filters
all_filters <- pgxFiler()
# get all prefix
all_prefix <- pgxFiler(return_all_prefix = TRUE)
# access specific filters based on prefix
ncit_filters <- pgxFiler(prefix="NCIT")
head(ncit_filters)
#> [1] "NCIT:C28076" "NCIT:C18000" "NCIT:C14158" "NCIT:C14161" "NCIT:C28077"
#> [6] "NCIT:C28078"
```

The following query is designed to retrieve metadata in Progenetix related to all samples of lung adenocarcinoma, utilizing a specific type of filter based on an NCIt code as an ontology identifier.

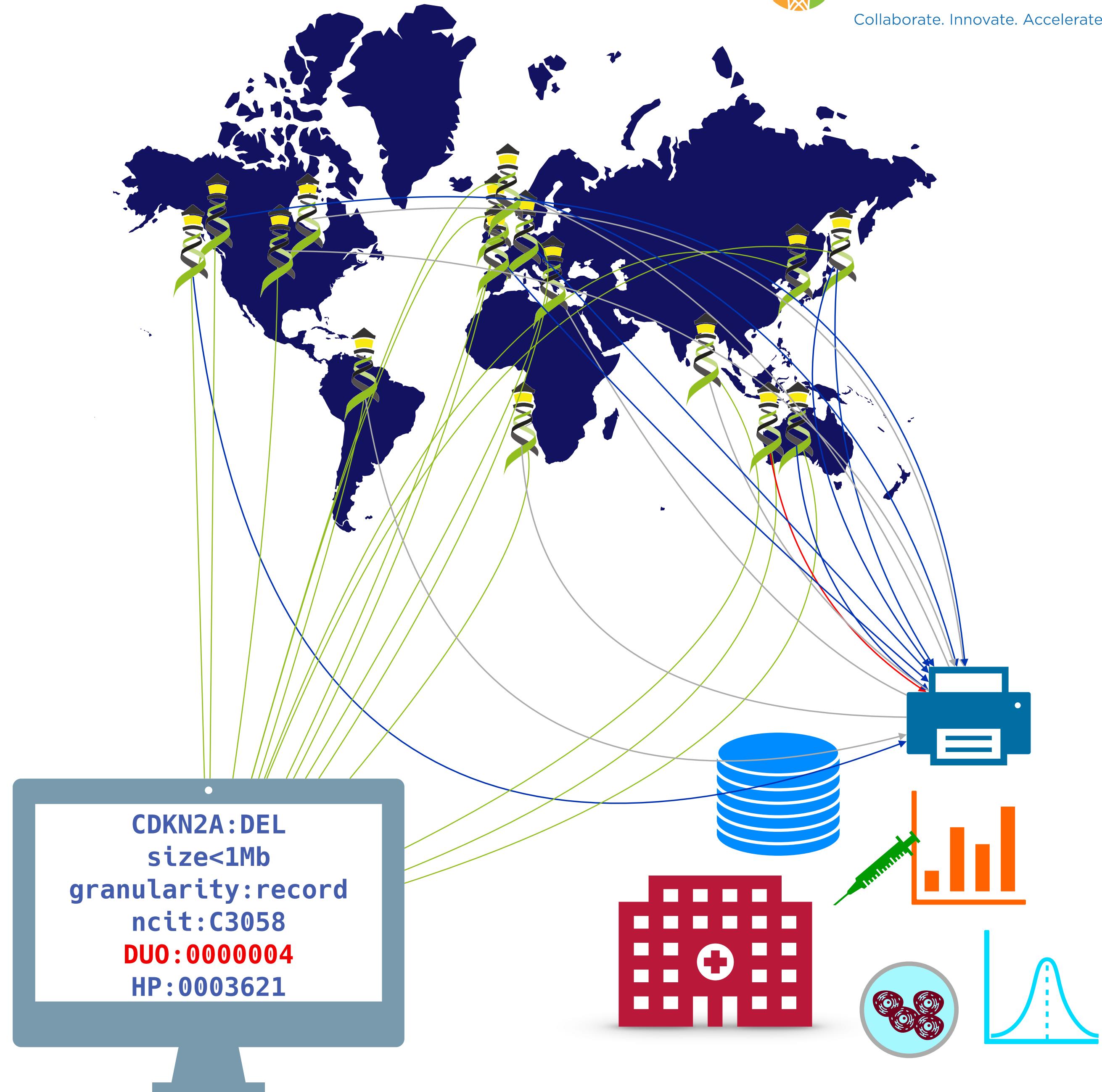
```
biosamples <- pgxLoader(type="biosample", filters = "NCIT:C3512")
# data looks like this
biosamples[c(1700:1705),]
#>   biosample_id group_id group_label individual_id callset_ids
#> 1700 pgxbs-kftvjjhx    NA      NA pgxind-kftx5fyd pgxcs-kftwjevi
#> 1701 pgxbs-kftvjjhz    NA      NA pgxind-kftx5fyf pgxcs-kftwjew0
#> 1702 pgxbs-kftvjjj1    NA      NA pgxind-kftx5fyh pgxcs-kftwjewi
#> 1703 pgxbs-kftvjjn2    NA      NA pgxind-kftx5g4r pgxcs-kftwjg5r
#> 1704 pgxbs-kftvjjn4    NA      NA pgxind-kftx5g4t pgxcs-kftwjg6q
#> 1705 pgxbs-kftvjjn5    NA      NA pgxind-kftx5g4v pgxcs-kftwjg78
```

What Can You Do?

- implement procedures and standards supporting **data discovery** (FAIR principles) and federation approaches
- forward looking consent and data protection models adhering to **ORD** principles ("as secure as necessary, as open as possible")
- **support** and/or get involved with international **data standards** efforts and projects



Collaborate!

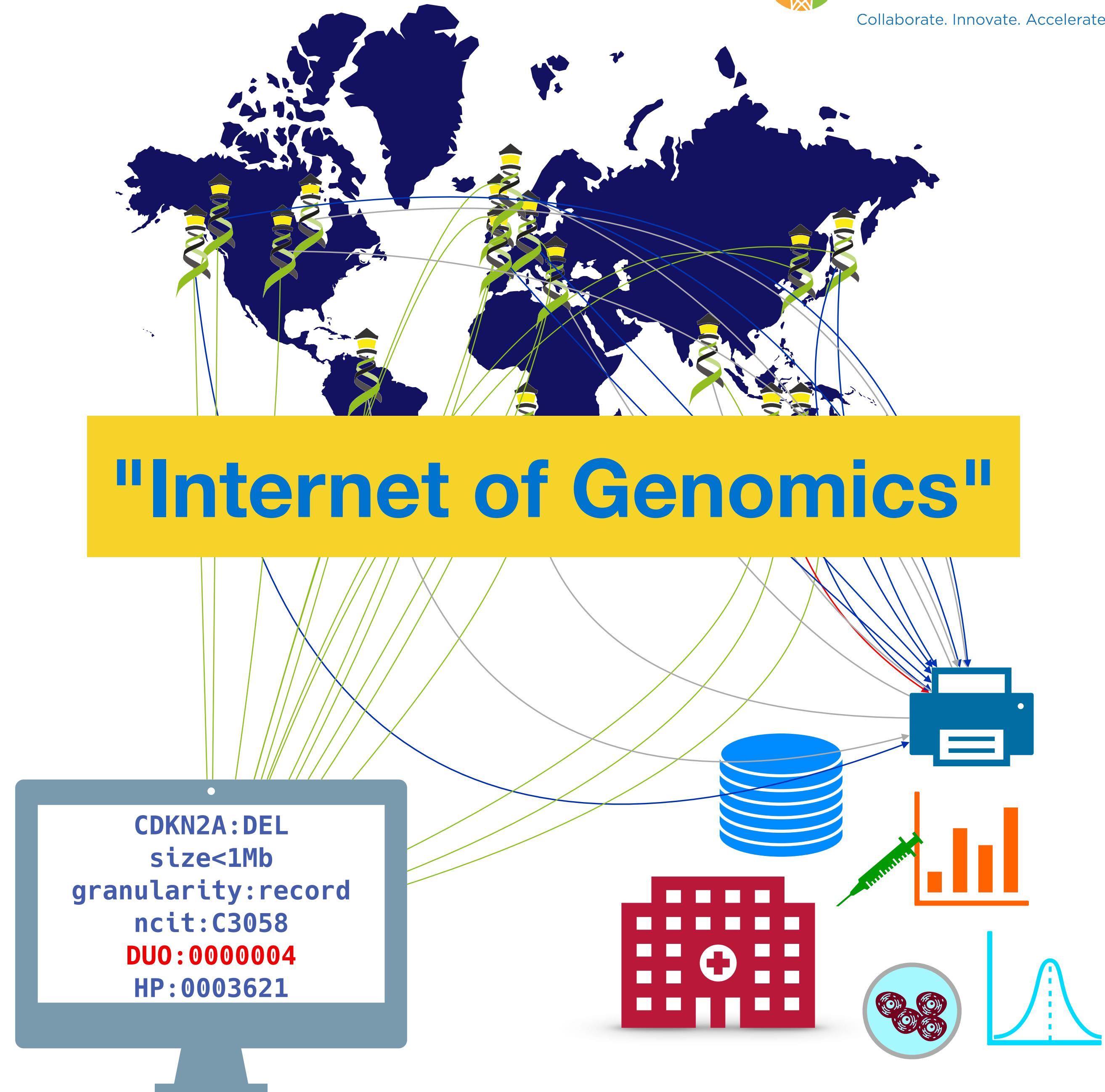


What Can You Do?

- implement procedures and standards supporting **data discovery** (FAIR principles) and federation approaches
- forward looking consent and data protection models adhering to **ORD** principles ("as secure as necessary, as open as possible")
- **support** and/or get involved with international **data standards** efforts and projects



Collaborate!



h-CNV Community

Homepage & News

About ...

h-CNV Projects

CNV Annotation Standards

Databases & Resources

CNV References Project

Contacts

Genome Blog

h-CNV @ ELIXIR

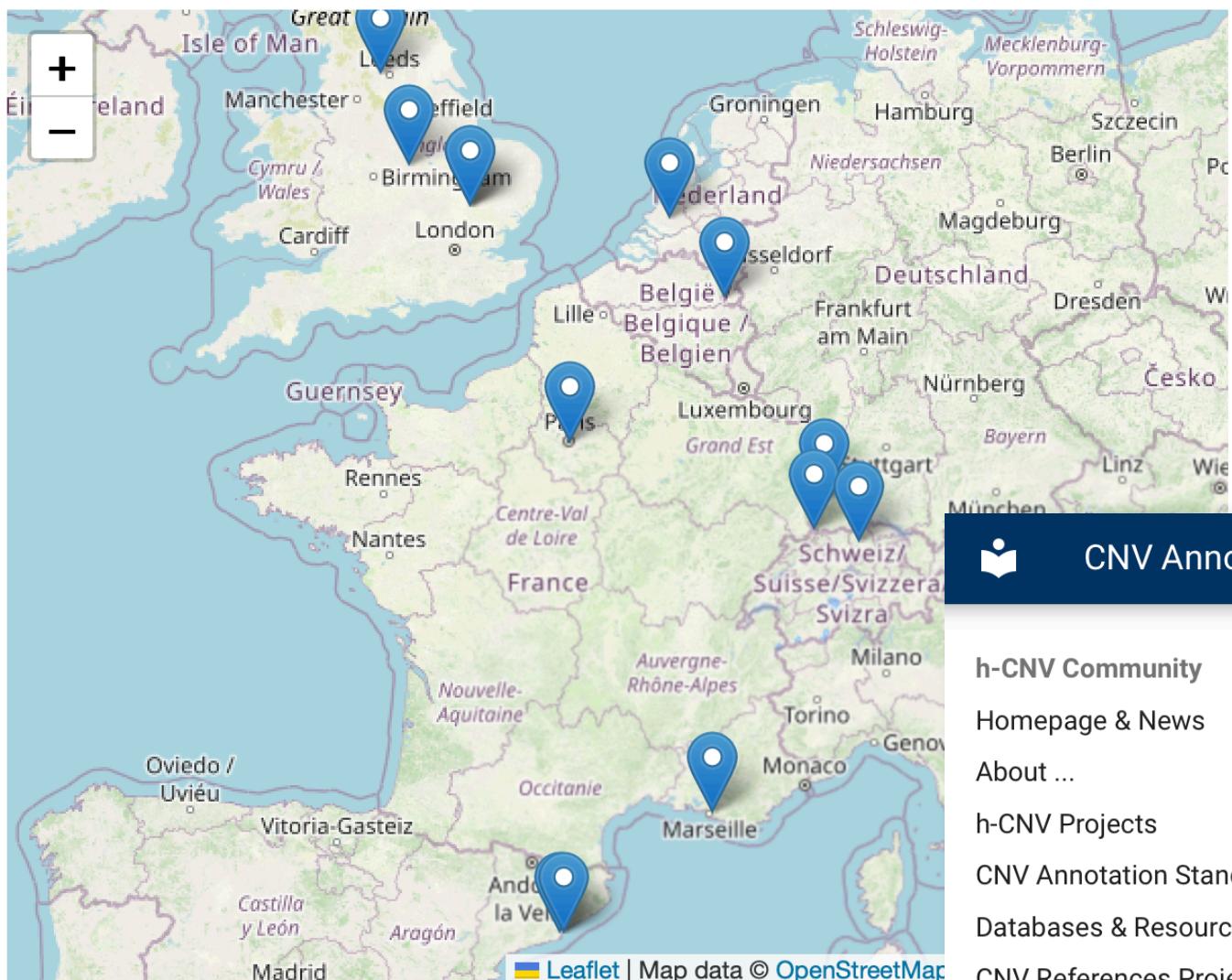
Beacon Project

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

elxir

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



Jordi Rambla
Arcadi Navarro
Roberto Ariosa
Manuel Rueda
Lauren Fromont
Mauricio Moldes
Claudia Vasallo
Babita Singh
Sabela de la Torre
Marta Ferri
Fred Haziza



Juha Törnroos
Teemu Kataja
Ikkka Lappalainen
Dylan Spalding



Tony Brookes

Tim Beck

Colin Veal

Tom Shorter



Michael Baudis

Rahel Paloots

Hangjia Zhao

Ziying Yang

Bo Gao

Qingyao Huang



Augusto Rendon

Ignacio Medina

Javier López

Jacobo Coll

Antonio Rueda



centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

Sergi Beltran

Carles Hernandez



David Salgado



Salvador Capella

Dmitry Repchevski

JM Fernández



Laura Furlong

Janet Piñero



Serena Scollen

Gary Saunders

Giselle Kerry

David Lloyd



Nicola Mulder

Mamana

Mbiyavanga

Ziyaad Parker



David

Torrents



Dean Hartley



Fundación Progreso y Salud
CONSEJERÍA DE SALUD

Joaquin Dopazo

Javier Pérez

J.L. Fernández

Gema Roldan



Thomas Keane

Melanie Courtot

Jonathan Dursi



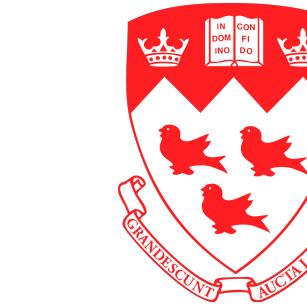
Heidi Rehm

Ben Hutton



Toshiaki

Katayama



Stephane Dyke

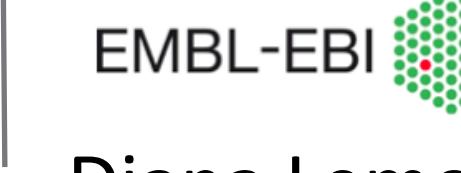


Marc Fiume

Miro Cupak



Melissa Cline



Diana Lemos



GA4GH Phenopackets

Peter Robinson
Jules Jacobsen



GA4GH VRS
Alex Wagner
Reece Hart

Beacon PRC

Alex Wagner
Jonathan Dursi
Mamana Mbiyavanga
Alice Mann
Neerjah Skantharajah



The Beacon team through the ages

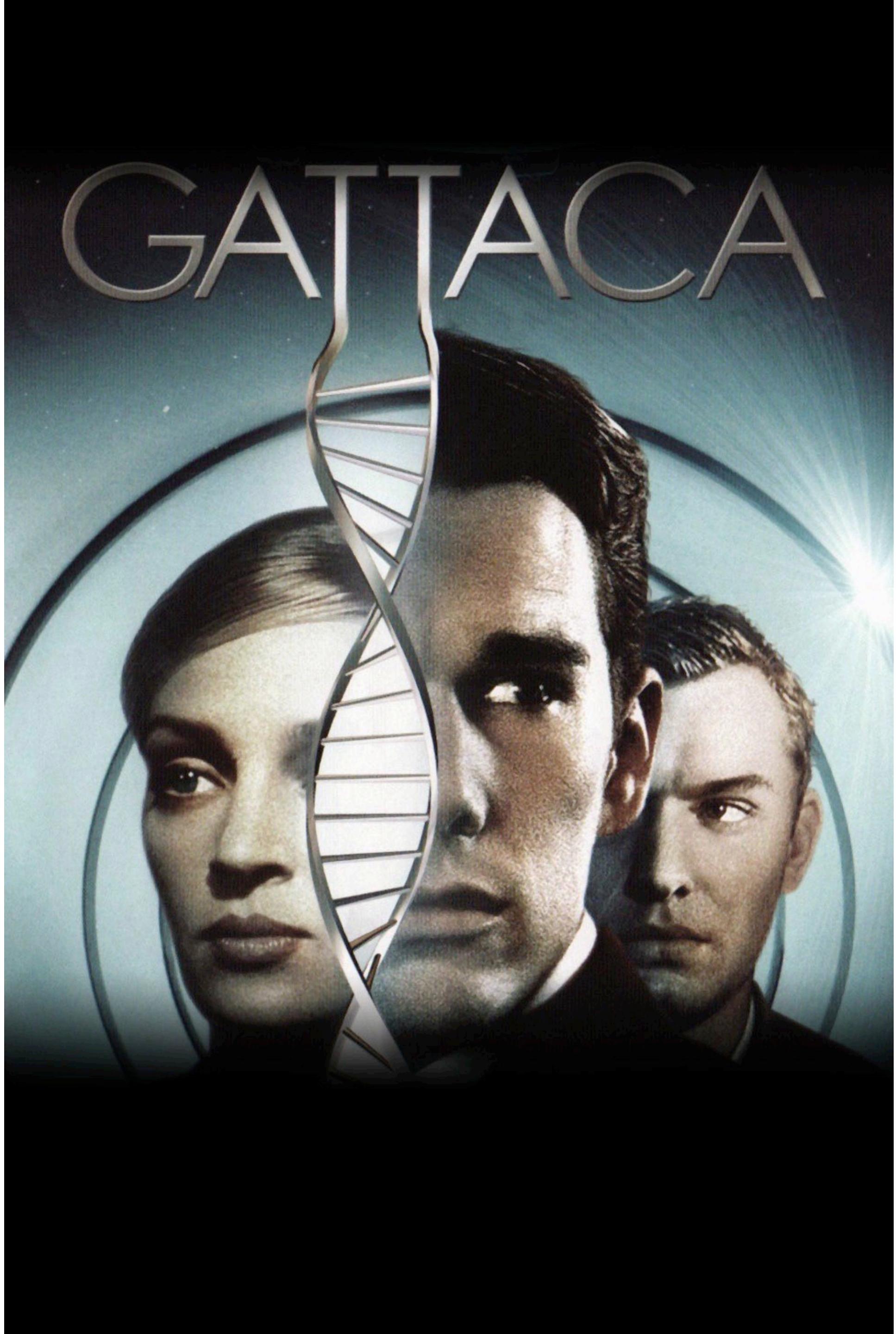
Genomic Data & Privacy

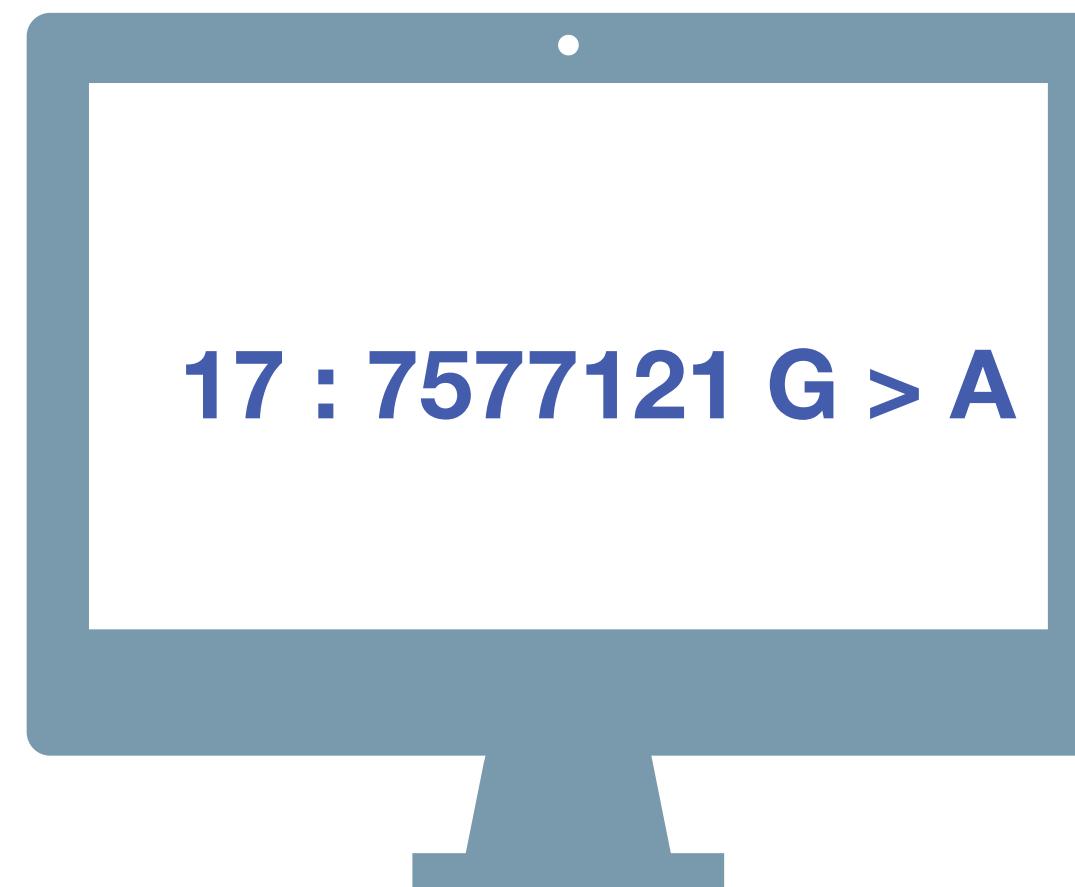
Risks & opportunities

Gattaca (1997)

A genetically inferior man assumes the identity of a superior one in order to pursue his lifelong dream of space travel.

- genetic determinism
 - ▶ main character has been determined to be unsuitable for complex jobs based on genetic analysis
- genetic identification
 - ▶ the use of genetic sampling for personal identification is daily routine





Beacon

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

YES | NO | \0



Genome Beacons Compromise Security?

Querying for thousands of specific SNV occurrences in a genomic data pool can identify individuals in an anonymized genomic data collection

Stanford researchers identify potential security hole in genomic data-sharing network

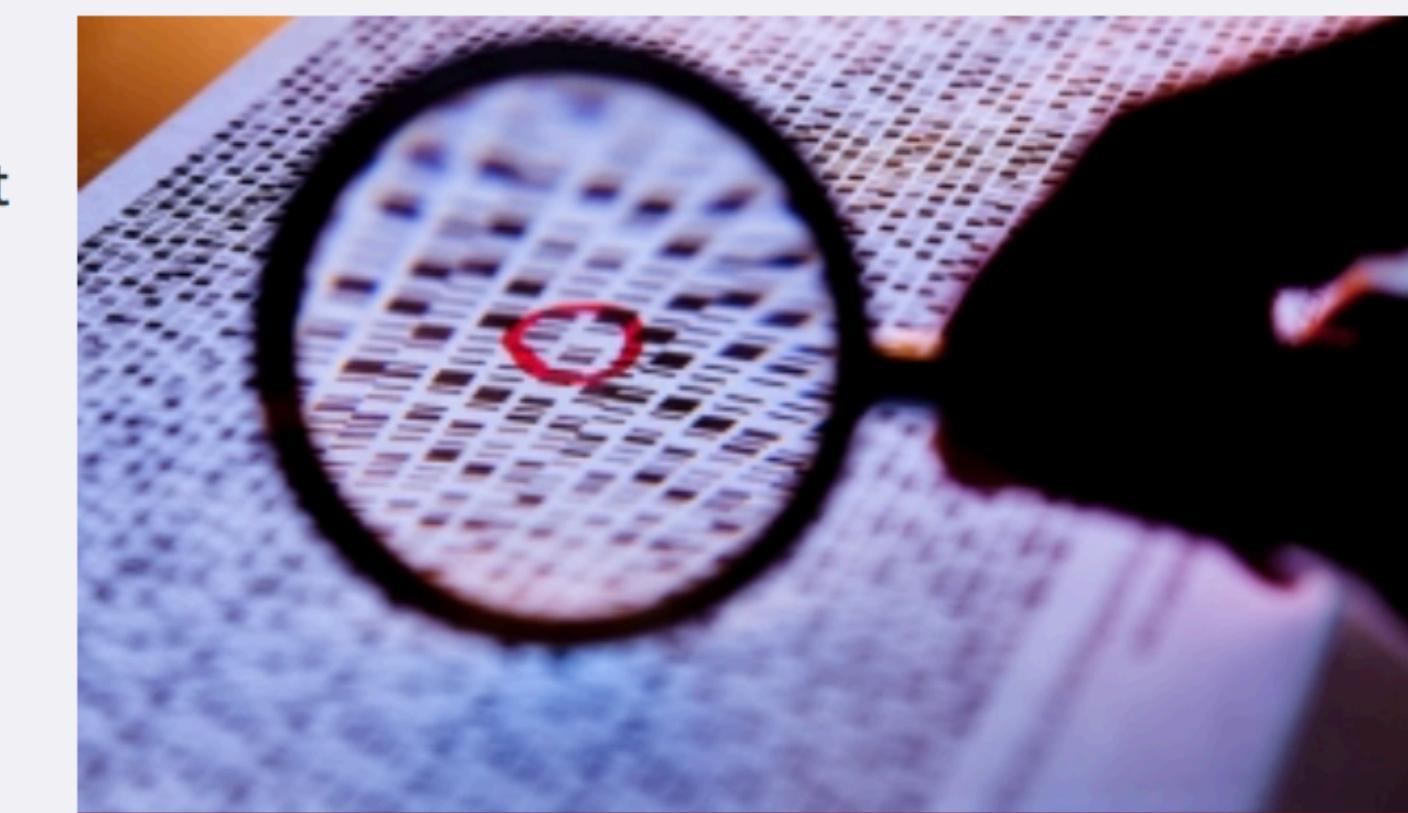
Hackers with access to a person's genome might find out if that genome is in an international network of disease databases.

OCT 29
2015

Sharing genomic information among researchers is critical to the advance of biomedical research. Yet genomic data contains identifiable information and, in the wrong hands, poses a risk to individual privacy. If someone had access to your genome sequence — either directly from your saliva or other tissues, or from a popular genomic information service — they could check to see if you appear in a database of people with certain medical conditions, such as heart disease, lung cancer or autism.

Work by a pair of researchers at the [Stanford University School of Medicine](#) makes that genomic data more secure. [Suyash Shringarpure](#), PhD, a postdoctoral scholar in genetics, and [Carlos Bustamante](#), PhD, a professor of genetics, have demonstrated a technique for hacking a network of global genomic databases and how to prevent it. They are working with investigators from the Global Alliance for Genomics and Health on implementing preventive measures.

The work, published Oct. 29 in *The American Journal of Human Genetics*, also bears importantly on the larger question of how to analyze mixtures of genomes, such as those from different people at a crime scene.



Stanford researchers are working with the Global Alliance for Genomics and Health to make genomic information in the Beacon Project more secure.
Science photo/Shutterstock

IDENTIFICATION OF INDIVIDUALS FROM MIXED COLLECTIONS USING RARE ALLELES

Privacy Risks from Genomic Data-Sharing Beacons

Suyash S. Shringarpure^{1,*} and Carlos D. Bustamante^{1,*}

The human genetics community needs robust protocols that enable secure sharing of genomic data from participants in genetic research. Beacons are web servers that answer allele-presence queries—such as “Do you have a genome that has a specific nucleotide (e.g., A) at a specific genomic position (e.g., position 11,272 on chromosome 1)?”—with either “yes” or “no.” Here, we show that individuals in a beacon are susceptible to re-identification even if the only data shared include presence or absence information about alleles in a beacon. Specifically, we propose a likelihood-ratio test of whether a given individual is present in a given genetic beacon. Our test is not dependent on allele frequencies and is the most powerful test for a specified false-positive rate. Through simulations, we showed that in a beacon with 1,000 individuals, re-identification is possible with just 5,000 queries. Relatives can also be identified in the beacon. Re-identification is possible even in the presence of sequencing errors and variant-calling differences. In a beacon constructed with 65 European individuals from the 1000 Genomes Project, we demonstrated that it is possible to detect membership in the beacon with just 250 SNPs. With just 1,000 SNP queries, we were able to detect the presence of an individual genome from the Personal Genome Project in an existing beacon. Our results show that beacons can disclose membership and implied phenotypic information about participants and do not protect privacy *a priori*. We discuss risk mitigation through policies and standards such as not allowing anonymous pings of genetic beacons and requiring minimum beacon sizes.

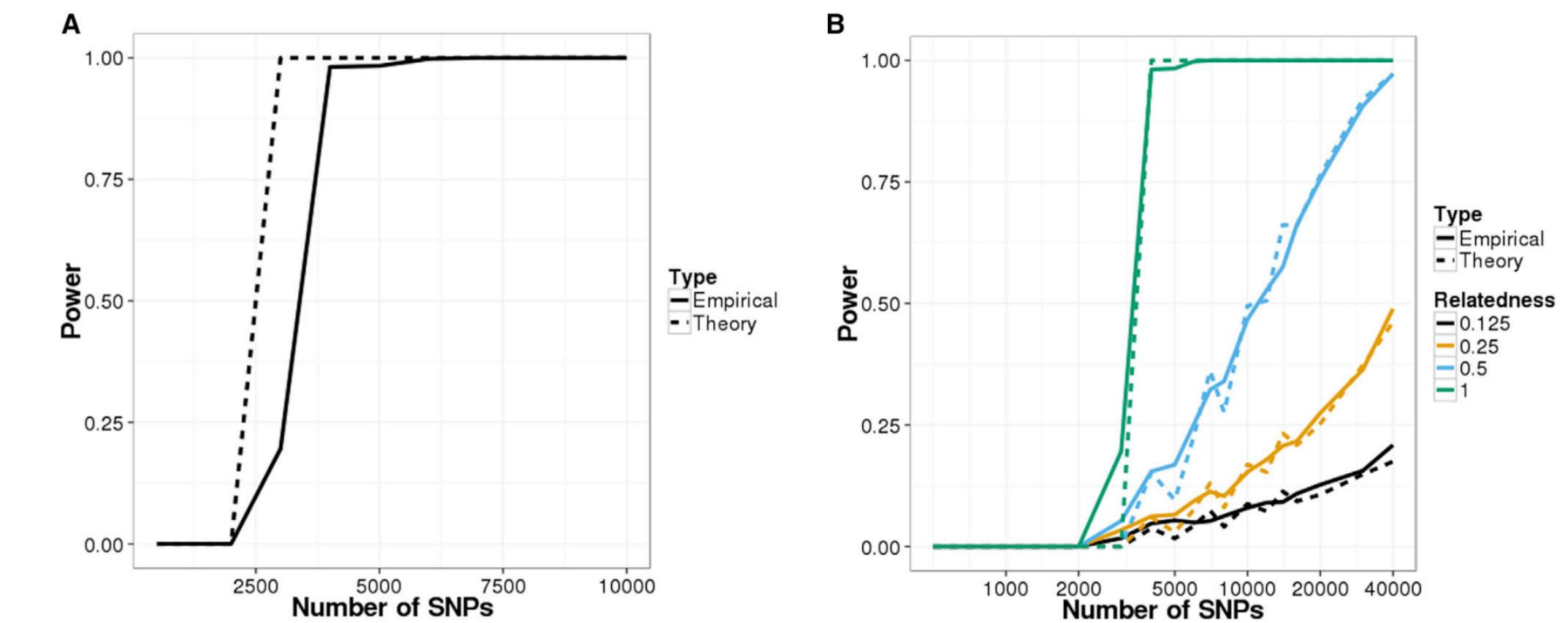


Figure 1. Power of Re-identification Attacks on Beacons Constructed with Simulated Data
Power curves for the likelihood-ratio test (LRT) on (A) a simulated beacon with 1,000 individuals and (B) detecting relatives in the simulated beacon. The false-positive rate was set to 0.05 for all scenarios.

- ▶ rare allelic variants can be used to identify an individual (or her relatives) in a genome collection without having access to individual datasets
- ▶ however, such an approach requires previous knowledge about the individual's SNPs

Direct to Consumer DNA Analyses

Population Background, Family Trees, Traits & Disease Risks...

Enorme Ersparnis!

Letzte Chance DNA-Weihnachtsaktion

Nur 39 CHF ~~89 CHF~~



Our Service Learn

SIGN IN REGISTER K

Your DNA

Jacqueline 100%

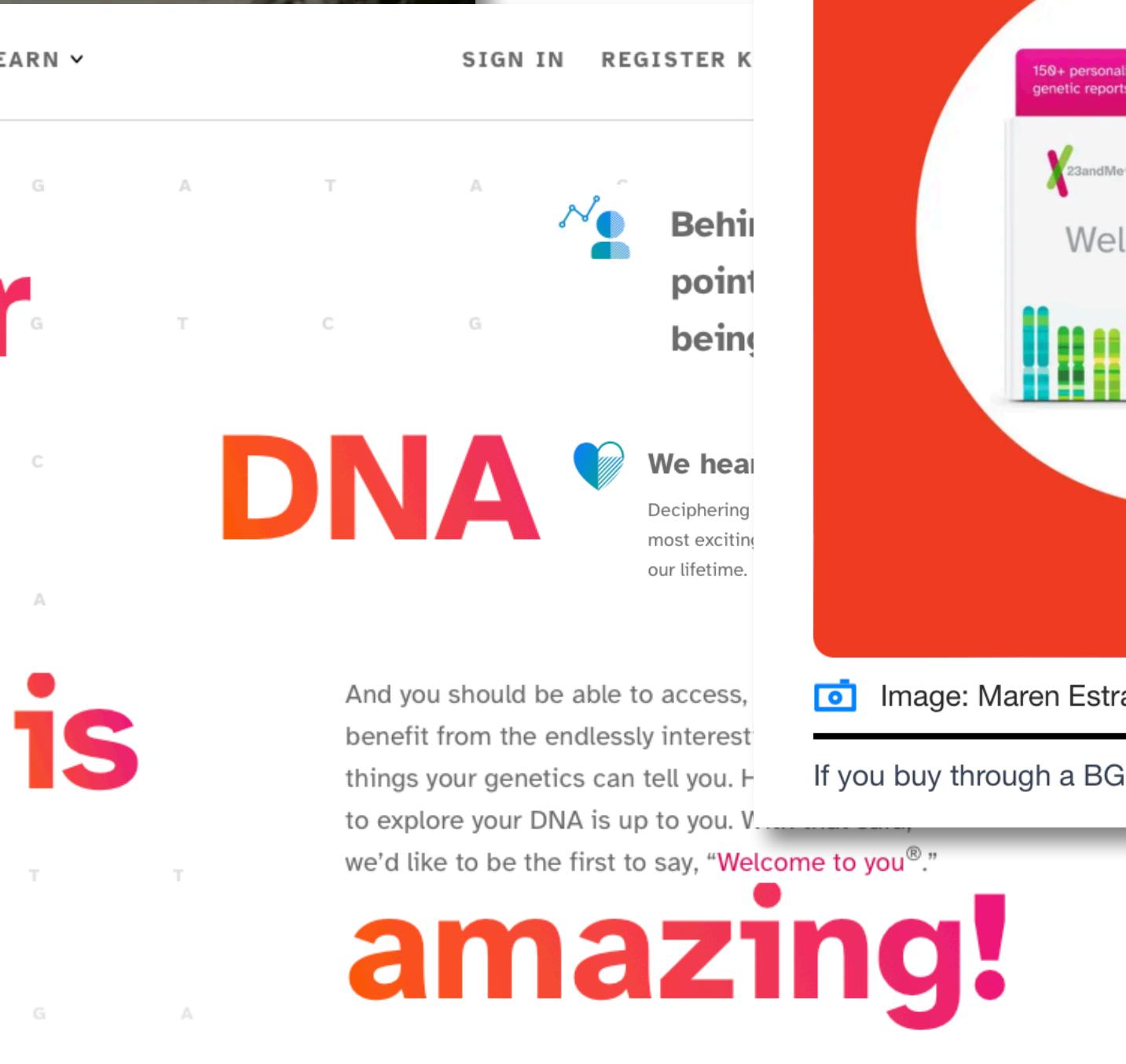
European 50.1%

- British & Irish 39.7%
- French & German 7.0%
- Broadly Northwestern European 3.2%
- Scandinavian 0.2%

East Asian & Indigenous American 49.9%

- Vietnamese 46.3%
- Indonesian, Thai & Myanma 1.5%
- Chinese 0.5%

is amazing!



By the numbers

2006

The year we set out to make DNA more accessible and meaningful for all.

12M+

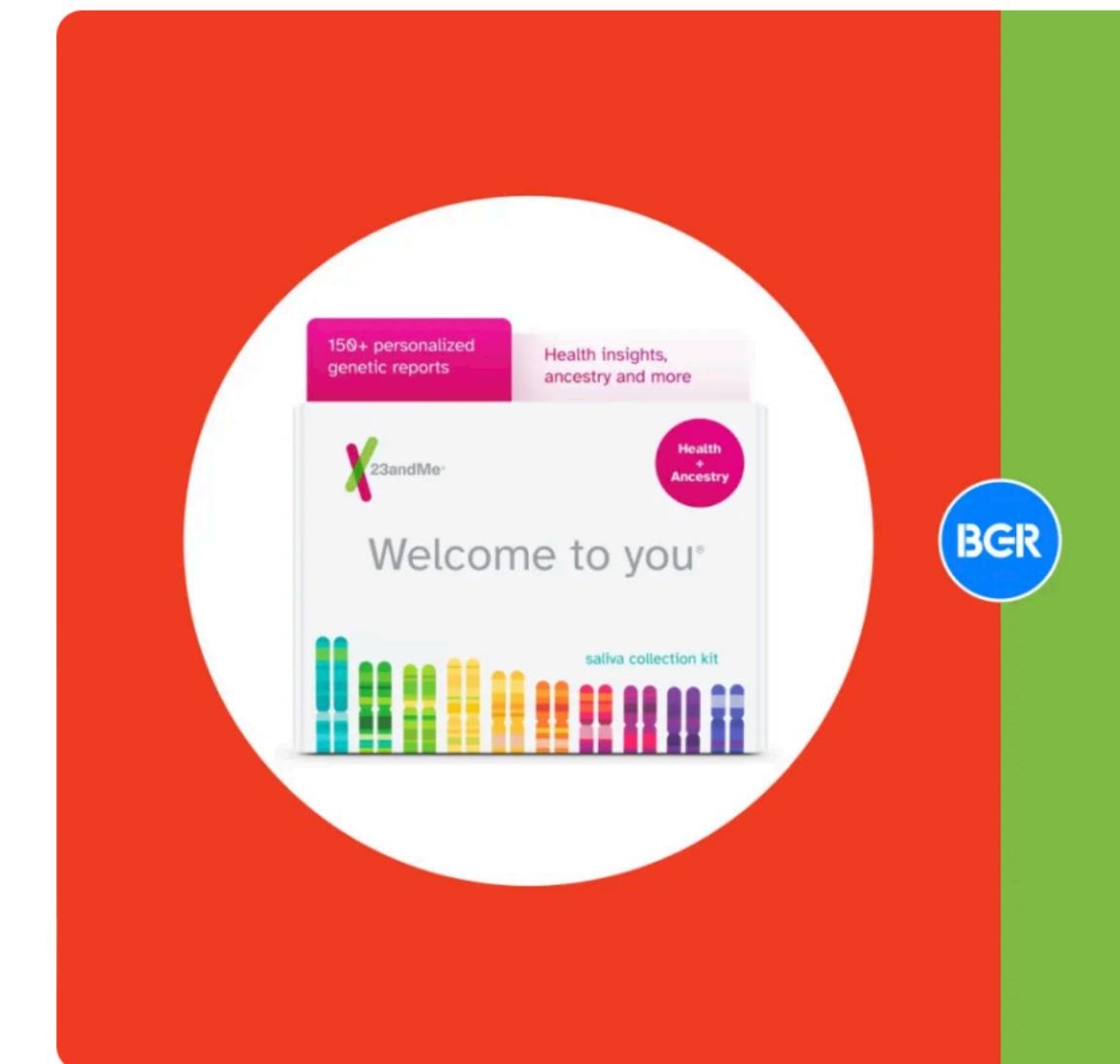
The number of DNA kits we've sold in that time.

Best DNA test kits on sale for Cyber Monday 2023



By Maren Estrada 

Published Nov 26th, 2023 11:11AM EST



ancestry GENEALOGY DNA

FREE TRIAL SIGN IN EN

What would you like to learn about your family history? Select all that apply

Details about my ancestors My origins I'm not really sure what I can discover on Ancestry

Skip Next

Dismiss

HOLIDAY SALE

Inherited from Parent 1 Inherited from Parent 2

Build a family tree and uncover your story. Start a free trial

Save up to \$50* on DNA offers for a limited time. Start saving

*Ends 31 Dec 2022. Terms apply. Pricing for U.S. customers only.

Bring the generations together with a gift from Ancestry®.



“We’re an information economy. They teach you that in school. What they don’t tell you is that it’s impossible to move, to live, to operate at any level without leaving traces, bits, seemingly meaningless fragments of personal information. Fragments that can be retrieved, amplified . . .”

–William Gibson in "Johnny Mnemonic" (1986)

Phenotyping from DNA

From DNA to "Wanted" Posters?

- association of genomic variants with phenotypic data collection
- while hair, eye color are easy targets not useful for relevant phenotypic features especially if large environmental component
- huge biases based on input/collection data
- Belgium and Germany do not allow forensic DNA phenotyping
- Switzerland: Bundesrat decision on 2020-12-04 to allow phenotyping for law enforcement purposes



"When the New York Times ran an informal test of the Paragon system with one of its reporters, it failed badly." (ACLU.org)



Hi Michael,

Good news! We've discovered new DNA Matches for you.

- Commercial, "Direct to Customer" DNA analyses are provided through independent sites and such affiliated to genealogy services (MyHeritage, Ancestry.com, 23andMe...)
- Genealogy sites identify individuals with matching haplotype blocks & provide a prediction about degree of genetic relation
- Law enforcement agencies (and who else?!) can send individual SNP profiles (e.g. recovered from evidence many years after a crime) using a *Jane Doe* identity, to identify relatives of the suspect - **long range familial search**

Long-Range Familial Searches

Daily Journal

Helping Northeast Mississippi Grow!

We're donating a portion of every 1-year or 6-month subscription to Tupelo High Band Boosters!
842-2613 or djournal.com/subscribe
New home delivery subscriptions only | Offer ends June 30



SUBSCRIBE

ALL SEC Devaughn had never been a suspect until genetic genealogy put police on his trail several months ago. Earlier this year, police sent the DNA profile to Parabon, a private genetics company, to compare the suspect's DNA sample to a public genealogy DNA database looking for people with similar DNA profiles who might be kin to the suspect. That eventually led authorities to look at Devaughn.

Rienzi man charged with 1990 Starkville murder

By William Moore Daily Journal 15 hrs ago Comments

© Copyright 2018 Daily Journal, 1242 S Green St Tupelo, MS



The New York Times

How a Genealogy Site Led to the Front Door of the Golden State Killer Suspect

Investigators used DNA from crime scenes that had been stored all these years and plugged the genetic profile of the suspected assailant into an online genealogy database. One such service, GEDmatch, said in a statement on Friday that law enforcement officials had used its database to crack the case. Officers found distant relatives of Mr. DeAngelo's and, despite his years of eluding the authorities, traced their DNA to his front door.

The New York Times, April 26, 2018

Attacks Associated With the Golden State Killer



DNA & Law Enforcement

Legal minefields, hard to avoid?

- "...when police in Edmonton, Canada, released a suspect's image, the **crude graphic** ... came **from the suspect's DNA**."
- "...every time a **family member** sends in their swab, they're sending in your data too..."
- "...**many players** in this growing movement offer to translate our genetic code into phenotypes (that is, observable features like eye color), often with **scant commitment to scientific accuracy**..."
- "...Veering into **pseudoscience**, they are a modern **sales pitch for** the long-discredited **phrenology** of the past. They wrongly treat race as a biological fact, rather than the social construct that it is. And in the process, they open all the **flaws of facial recognition** to new realms..."
- "...we first have to change our focus from preventing DNA collection to **preventing misuse** and managing access..."
- "The answer is simple: **Ban DNA searches** ... beyond the types of one-to-one DNA tests that are subject to judicial oversight..."



Cops Might Already Have Your DNA, Without Your Consent

| FAREWELL PRIVACY |

We've entered the era of genetic surveillance and nothing—not even our own cells—is off-limits.

Albert Fox Cahn | Ayesha Rasheed

Published Nov. 14, 2022 4:51AM ET

"The unchallenged expansion of DNA collection and law enforcement misuse of the data has also spurred a surge in DNA surveillance startups."

But genotyping itself is for professional labs, right?

Rapid re-identification of human samples

...

We developed a rapid, inexpensive, and portable strategy to re-identify human DNA using the MinION. Our strategy requires only ~60 min preparation and 5-30 minutes of MinION sequencing, works with low input DNA, and enables familial searches using Direct-to-Consumer genomic reference datasets. This method can be implemented in a variety of fields:



Forensics

Identification of abandoned material using DNA fingerprinting is a common practice. The main challenge currently being: time. Our method allows rapid sample preparation at the crime scene (see movie). We envision that the method can be adopted in the field for rapid checks, after a mass disaster, and can be adopted in border control to fight human trafficking.



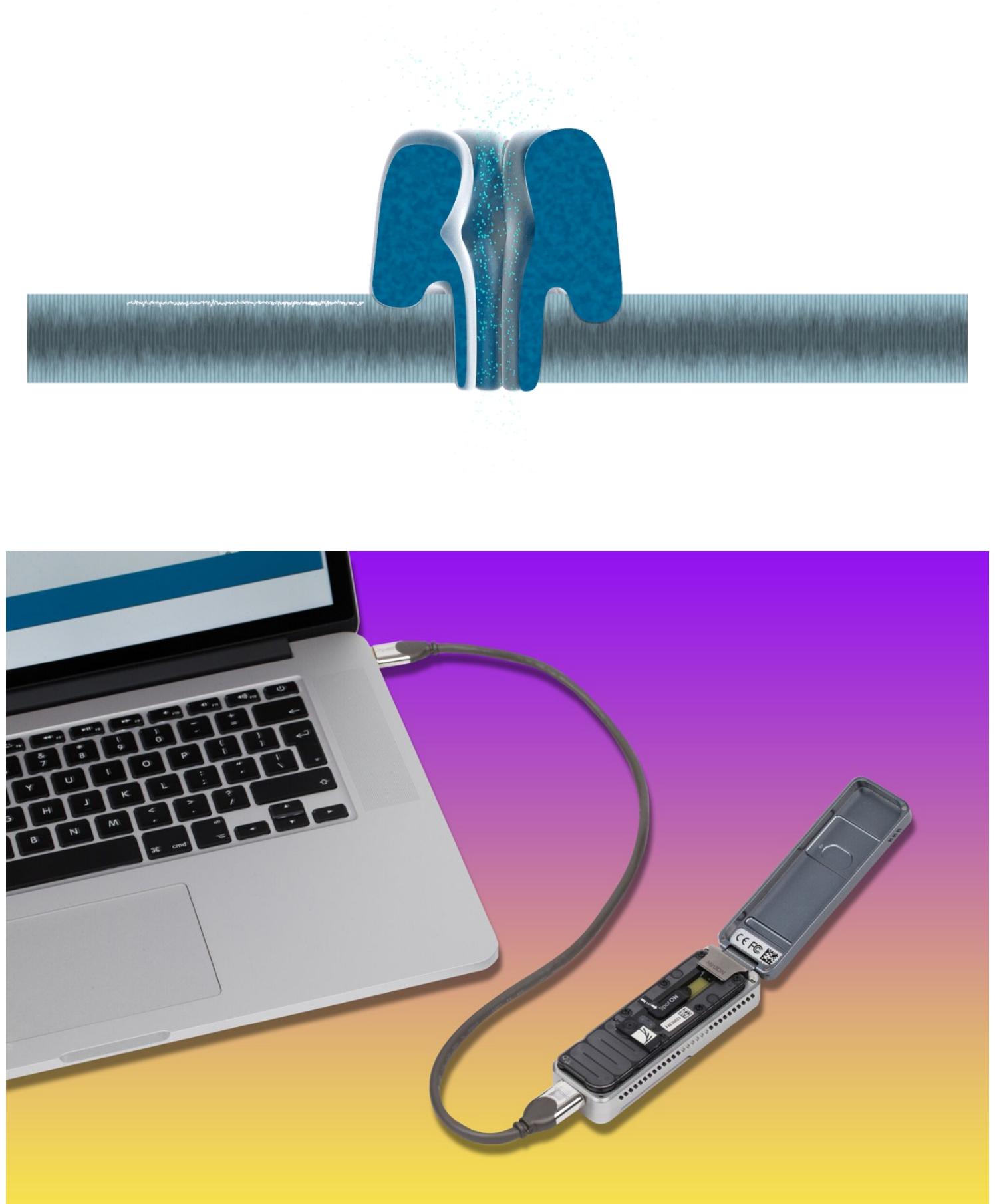
Clinic

Clinics process many samples, either for analysis or, for example, organ donations. These samples are DNA fingerprinted to prevent sample mix-up mistakes. Our method can be implemented in the clinic for rapid sanity-check of all incoming samples.



Cell line identification

Cross contamination of cell lines in science is a major problem. It results in unrepeatable data, and clinical trials based on inaccurate findings. This problem costs billions of dollars per year. We envision labs can adopt our identification method to ensure the purity of the cell line, and detect contamination.



The MinION (Oxford Nanopore)
Source: Sophie Zaaijer

DEMOCRATIZING DNA FINGERPRINTING

Sophie Zaaijer, Assaf Gordon, Robert Piccone, Daniel Speyer, Yaniv Erlich, 2016
ddf.teamerlich.org



DNA sequencing for identification/fingerprinting soon “commodity” technology (in contrast with technological/data challenges in “precision medicine”)

MinION by Oxford Nanopore Technologies



The MinION is the smallest DNA sequencer currently around. It's the size of a Mars bar, and can be simply plugged into a laptop with a USB3.0 port.

For more information about the MinION please click:
[Oxford Nanopore Technologies](#)

Bento Lab



The Bento lab is a miniature lab with a centrifuge, thermocycler and a electrophoresis compartment.

For more information about the Bento-lab please click:
[Bento Lab](#)



Generalkonsent

BENEFIT

BLOCKCHAIN

HEALTH

PRIVACY

CONSENT

SECURITY

HACKERS

ACCESS

Right to Research

LAWS

Genetic
Information
Nondiscrimination
Act

Health
Insurance
Portability and
Accountability
Act

SAFETY

CRYPTOGRAPHY

Share YOUR Genome data?

- The Beacon concept - balanced approach for accessing genome variant data from internationally distributed resources
- However: Genome data has the inherent “risk” of being identified and linked to a person

Solutions from Technology or Society? Discourse!

Welcome to openSNP

The screenshot shows the openSNP website homepage. At the top, there is a banner for MyHeritage DNA's Valentine's Day DNA SALE, offering 59€ per kit for 2+ kits. Below the banner, there is a section for "For Genotyping Users" with a "Upload Your Genotyping File" button. To the right, there is a large image of a DNA test kit and some text about phenotypes. On the far right, there is a sidebar with news and a "Order now" button.

The screenshot shows two side-by-side website snippets. On the left is the 23andMe website, featuring a "Welcome to you" message and a "saliva collection kit". On the right is the AncestryDNA website, which includes a headline stating "THE AVERAGE BRITISH PERSON'S DNA IS ONLY 36% BRITISH". It also features buttons for "GROW YOUR TREE" and "Discover DISCOVER".

How can a DNA firm lose half its users' data to 'Jew-hating' hackers?

Dark-web criminals cited the head of 23andMe's faith after a raid on the details of 6.9 million people — including her Google-founding ex. Now the lawsuits are coming

FAMILY MATTERS

Hackers stole ancestry data of 6.9 million users, 23andMe finally confirmed

Majority of impacted users are now being notified

ASHLEY BELANGER - 12/4/2023, 11:48 PM

Welcome to you

saliva collection kit

order now **USD\$99**

Bloomberg / Contributor | Bloomberg

It has now been confirmed that an additional **6.9 million 23andMe users had ancestry data stolen** after hackers accessed thousands of accounts by likely reusing previously leaked passwords.

... Wired estimated that "at least a million data points from 23andMe accounts" that were "exclusively about Ashkenazi Jews" and data points from "hundreds of thousands of users of Chinese descent" seemed to be exposed.

a spokesperson to confirm that two groups of opted into the **DNA Relatives feature** had their a stolen.

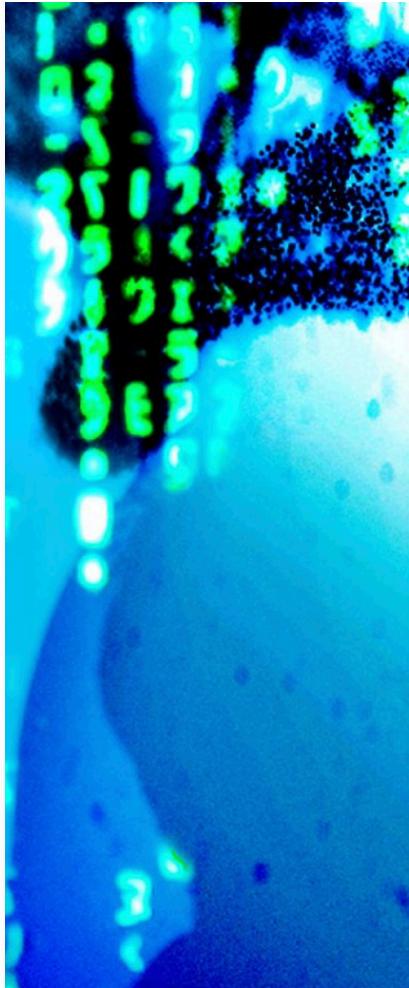
scribes the DNA Relatives feature as ... u to find and connect with genetic relatives and about your family." By **opting in**, users hope to ily members by willingly giving others access to like their birth year, current location, and names and birth locations. Users can opt out at

... about 5.5 million users, was hacked after opting in to automatically sharing information with DNA Relatives, including their "**name, birth year, relationship labels, the percentage of DNA shared with relatives, ancestry reports, and self-reported location**," TechCrunch reported.

... about 1.4 million users, shared "Family Tree profile information" ... including display names, relationship labels, birth year, and self-reported location, TechCrunch reported.

How can a DNA firm lose half its users' data to 'Jew-hating' hackers?

Dark-web criminals cited the head of 23andMe's faith after a raid on the details of 6.9 million people — including her Google-founding ex. Now the lawsuits are coming



FAMILY MATTERS —

Hackers stole ancestry data of 6.9 million users, 23andMe finally confirmed

Majority of impacted users are now being notified.

ASHLEY BELANGER - 12/4/2023, 11:48 PM



It has now been confirmed that an additional **6.9 million 23andMe users had ancestry data stolen** after hackers accessed thousands of accounts by likely reusing previously leaked passwords.

... Wired estimated that "at least a million data points from 23andMe accounts" that were "exclusively about Ashkenazi Jews" and data points from "hundreds of thousands of users of Chinese descent" seemed to be exposed.

... prompting a spokesperson to confirm that two groups of users who opted into the **DNA Relatives feature** had their personal data stolen.

23andMe describes the DNA Relatives feature as ... "allowing you to find and connect with genetic relatives and learn more about your family." By **opting in**, users hope to find lost family members by **willingly** giving others access to information like their birth year, current location, and ancestors' names and birth locations. Users can opt out at any time ...

... about 5.5 million users, was hacked after opting in to automatically sharing information with DNA Relatives, including their "**name, birth year, relationship labels, the percentage of DNA shared with relatives, ancestry reports, and self-reported location**," TechCrunch reported.

... about 1.4 million users, shared "Family Tree profile information" ... including display names, relationship labels, birth year, and self-reported location, TechCrunch reported.

How can a DNA firm users' data to 'Jew-haters'

Dark-web criminals cited the head of 23andMe as the source of the details of 6.9 million people — including the names and addresses of 2.5 million U.S. citizens. Now the lawsuits are coming.

WSJ Barron's MarketWatch IBD

THE WALL STREET JOURNAL.

SIGN IN

SUBSCRIBE

FAMILY MATTERS
Hackers stole data from 23andMe users, 23andMe says

Majority of impacted users opted in

ASHLEY BELANGER - 12/4/2023, 11:48 AM



23andMe's Fall From \$6 Billion to Nearly \$0

From celebrity 'spit parties' to a drop in the bucket: The once-hot DNA-testing company is struggling to profit

Anne Wojcicki of 23andMe, center, remotely rang the Nasdaq opening bell the day the company went public in 2021. PETER DASILVA/REUTERS

By [Rolle Winkler](#) [Follow](#)

Jan. 31, 2024 at 5:30 am ET

It has now been confirmed that an additional 6.9 million users' data stolen after hackers obtained it by likely reusing previously

stolen a million data points from the "exclusively about Ashkenazi" hundreds of thousands of users to be exposed.

can to confirm that two groups of A Relatives feature had their

Relatives feature as ... connect with genetic relatives and " By opting in, users hope to willingly giving others access to year, current location, and locations. Users can opt out at

as hacked after opting in to connection with DNA Relatives, year, relationship labels, the with relatives, ancestry location," TechCrunch reported.

shared "Family Tree profile information, including display names, relationship labels, birth year, and self-reported location, TechCrunch reported.



Universal Declaration of Human Rights (1948)

27(1)

“The Right to Science”

“Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and **to share in scientific advancement and its benefits.**”

27(2)

“The Right to Recognition”

“Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author.”



Universität
Zürich^{UZH}



Swiss Institute of
Bioinformatics

elixir