

Federated genomic discoveries

Deploying the GA4GH Beacon protocol

Michael Baudis @ GHGA Seminar Series 2024-03-20

Theoretical Cytogenetics and Oncogenomics

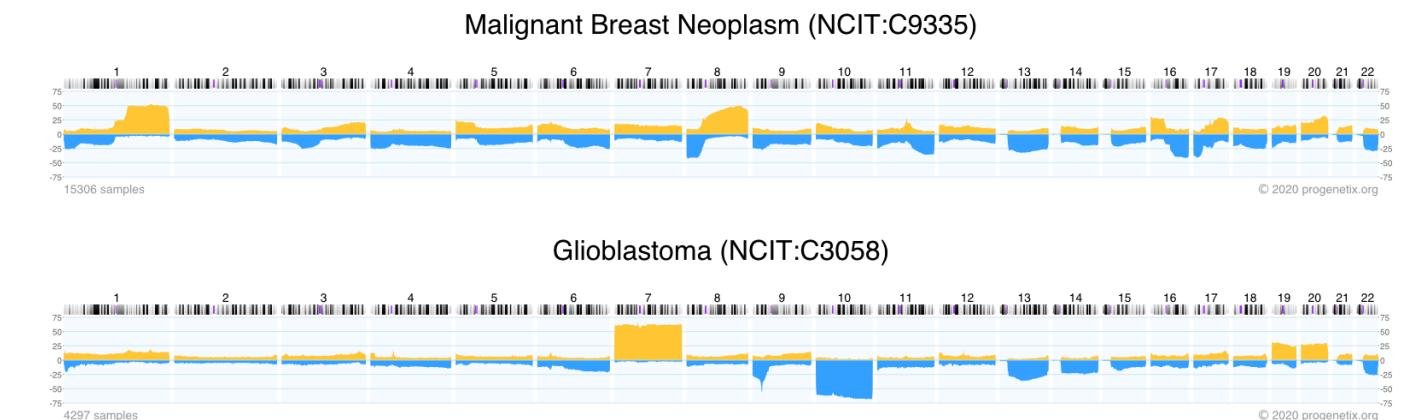
Cancer Genomics | Data Resources | Methods & Standards for Genomics and Personalized Health

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"

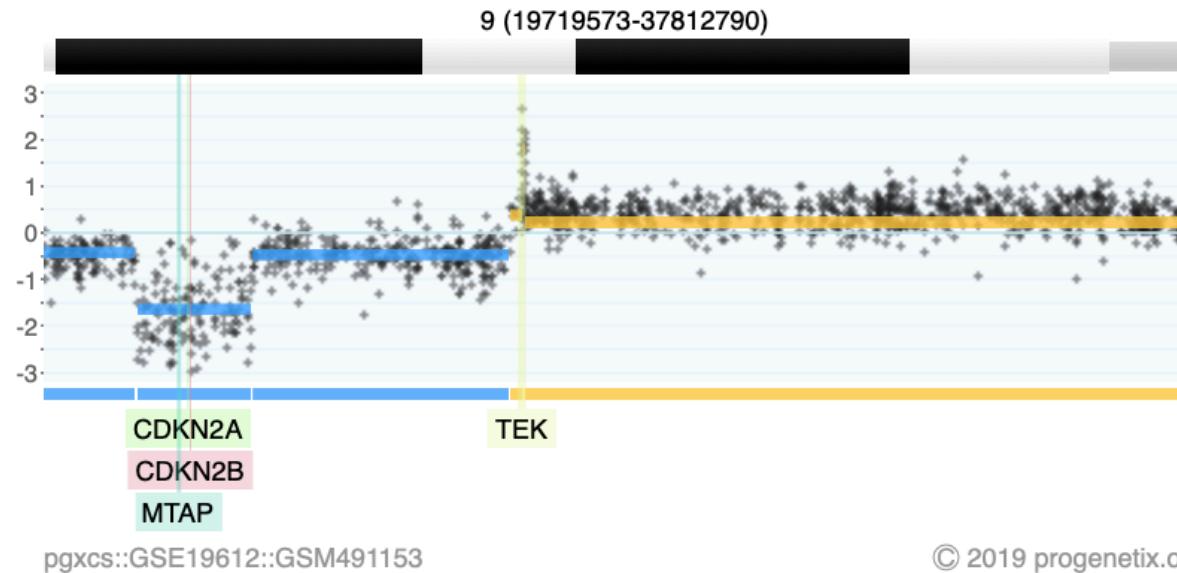


Theoretical Cytogenetics and Oncogenomics

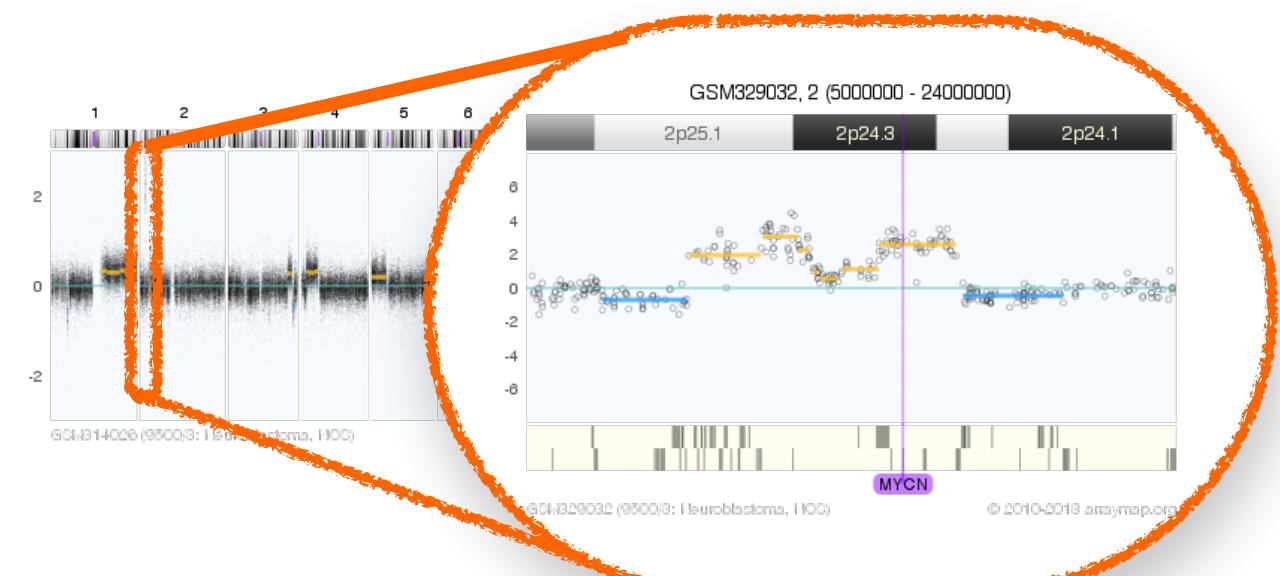
Research | Methods | Standards

Genomic Imbalances in Cancer - Copy Number Variations (CNV)

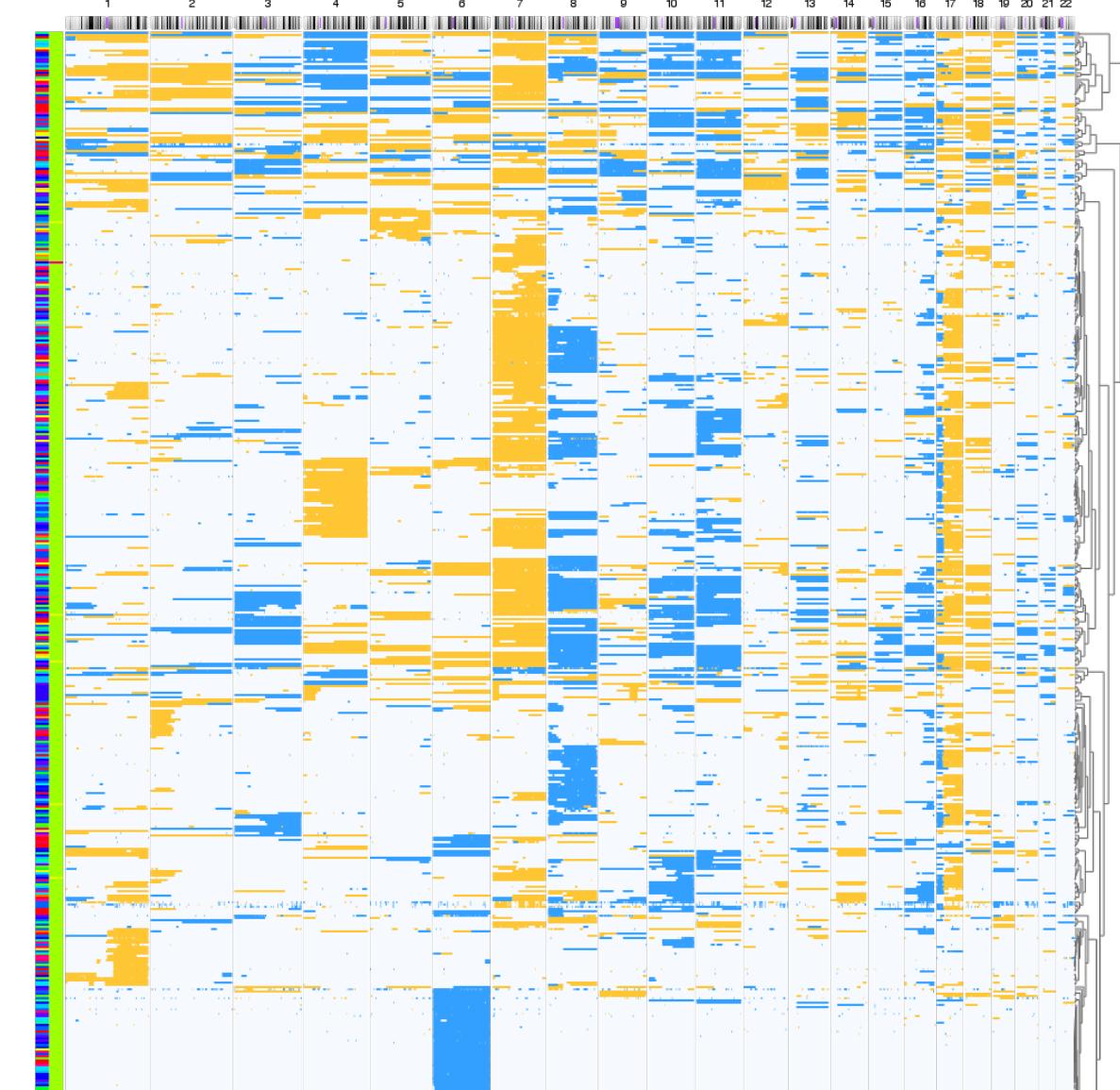
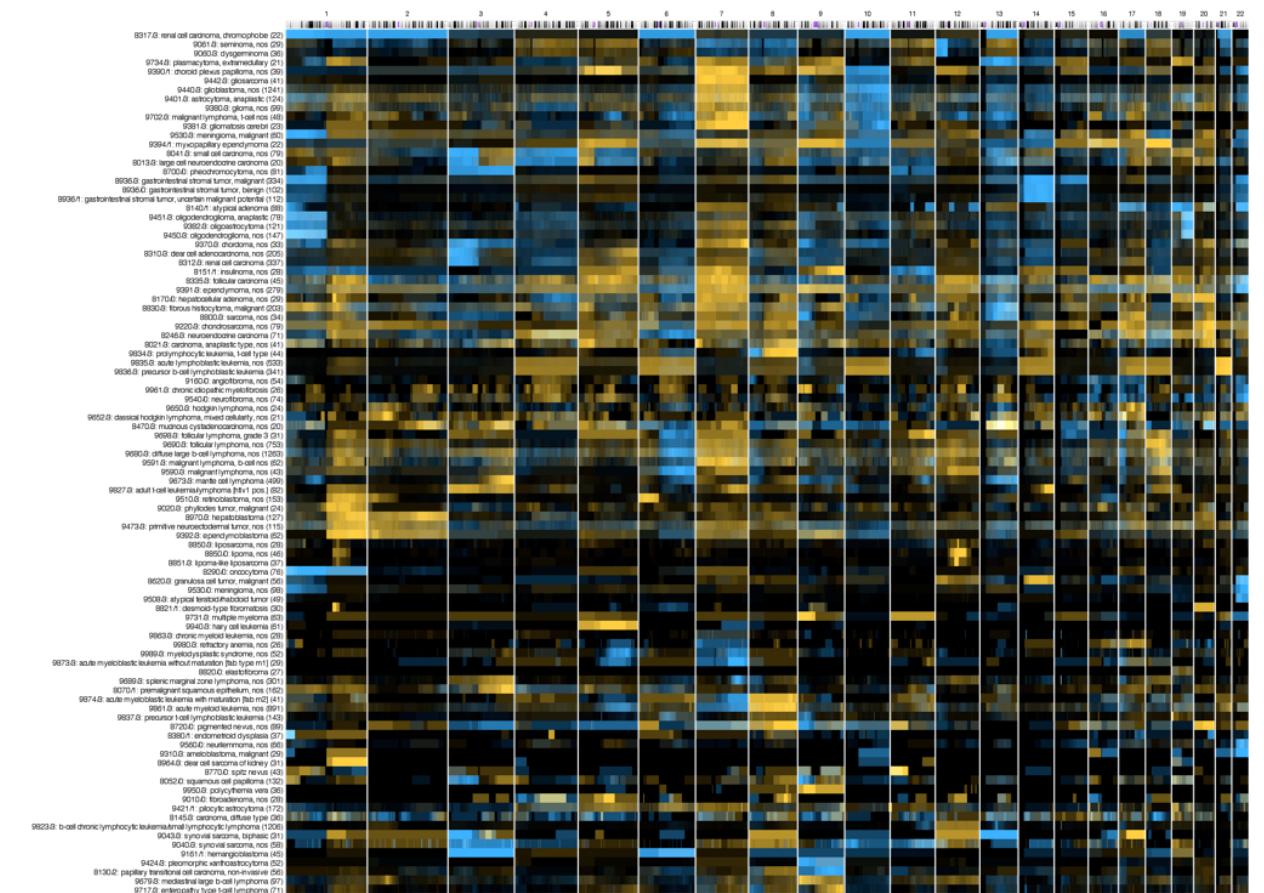
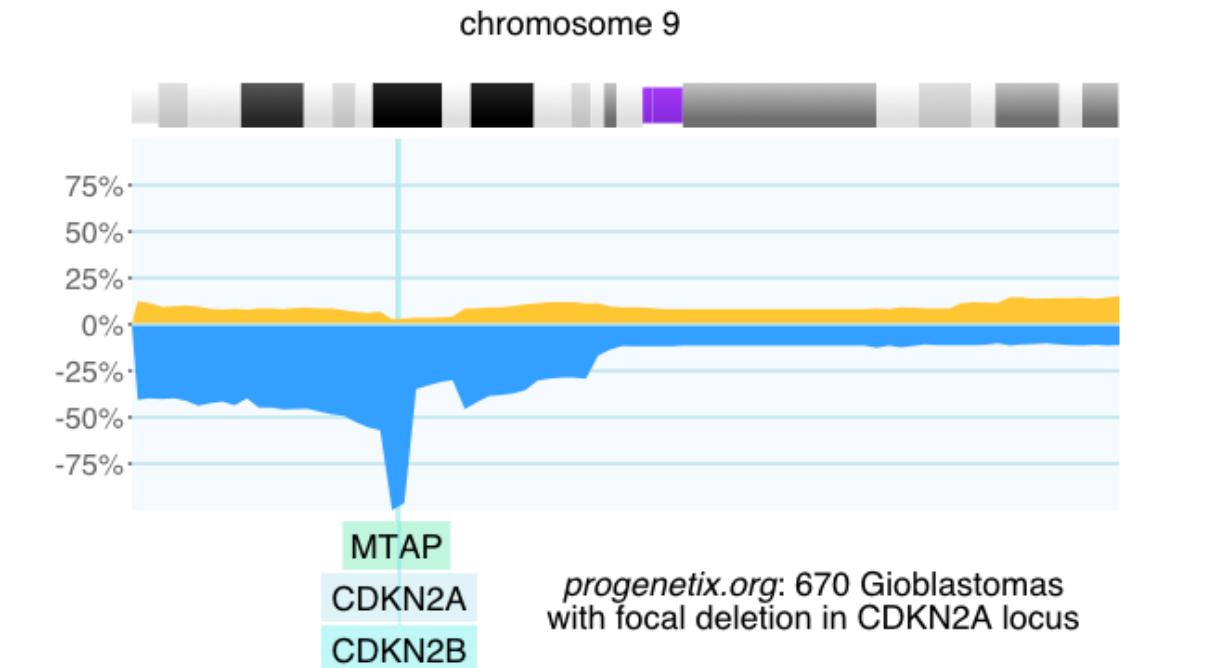
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)



2-event, homozygous deletion in a Glioblastoma



MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)



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

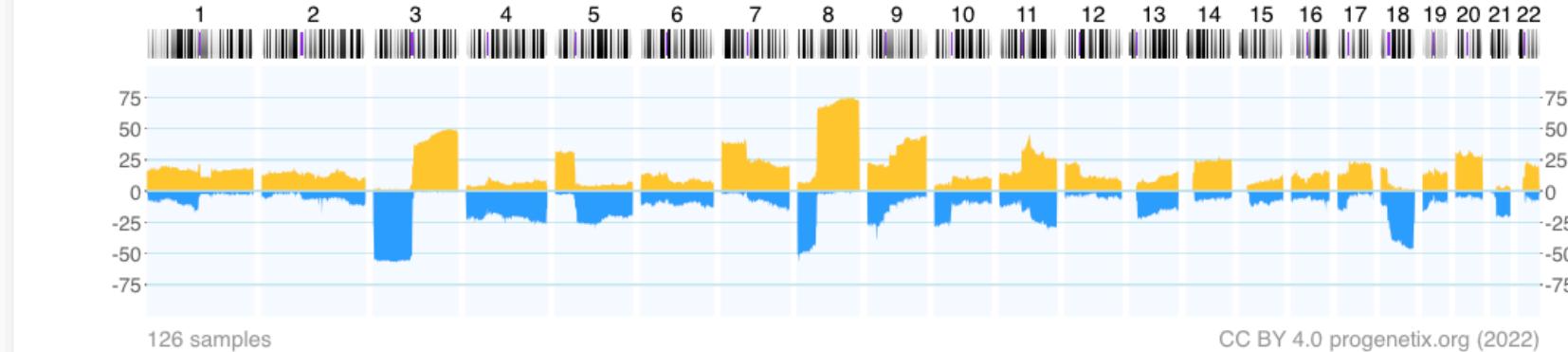
News
Downloads & Use
Cases
Sevices & API

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.

Floor of the Mouth Neoplasm (NCIT:C4401)



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

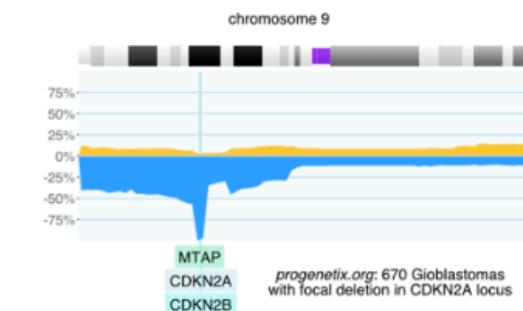
Example for aggregated CNV data in 126 samples in Floor of the Mouth Neoplasm.

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



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

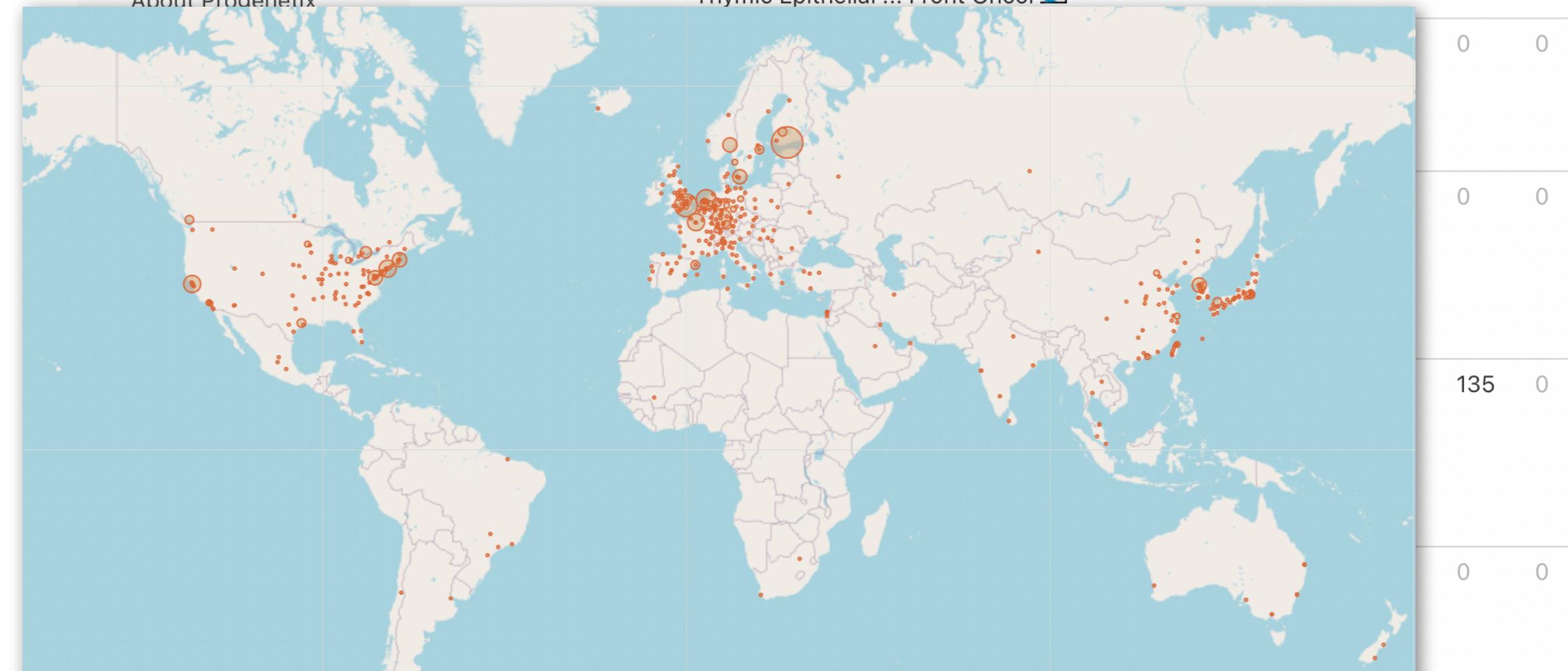
Progenetix Publication Collection

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](#) [↗](#).

Publications (3349)		Samples				
id i ▾	Publication	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



A world map showing the distribution of publications across different countries. The map is color-coded by continent, with red dots representing individual publications. The highest density of publications is visible in North America, Europe, and Asia. A legend on the right side of the map shows the count of publications: 0, 0, 0, 0, 135, 0, 0, 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)

DATABASE
The Journal of Biological Databases and Curation

Assembly: GRCh38 Chro: NC_000007.14 Start: 140713328 End: 140924929

Type: SNV

cellz

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

UCSC region
Variants in UCSC
Dataset Responses (JSON)

Visualization options

Results Biosamples Variants Annotated Variants

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

Cell Line Details

HOS (cellosaurus:CVCL_0312)

Subset Type

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

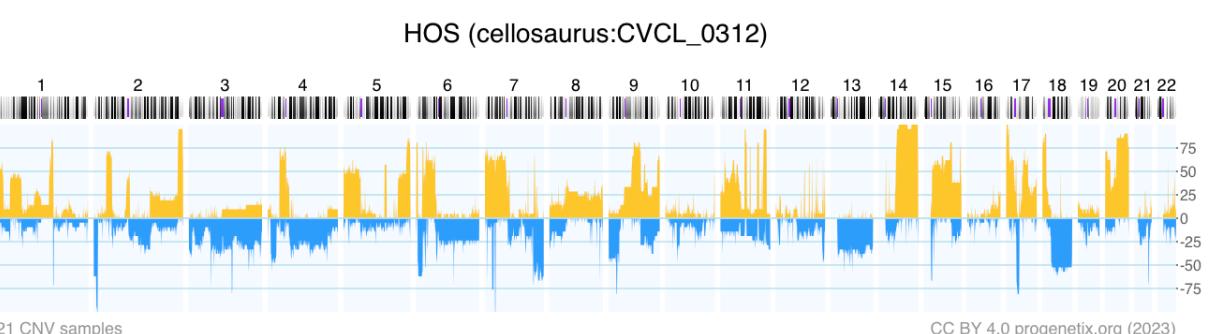
Sample Counts

- 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)

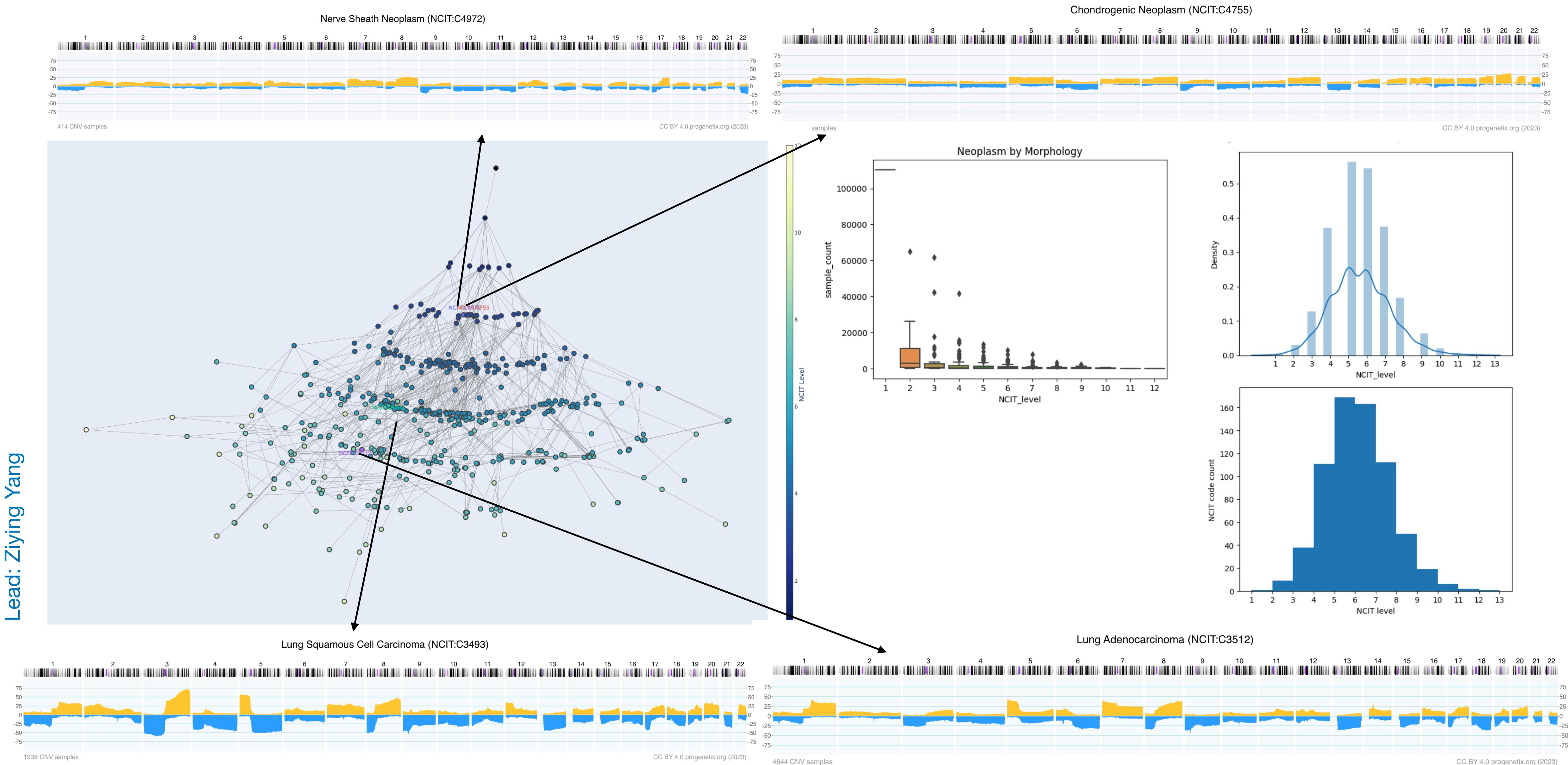


[Download SVG](#) | [Go to cellosaurus:CVCL_0312](#) | [Download CNV Frequencies](#)

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

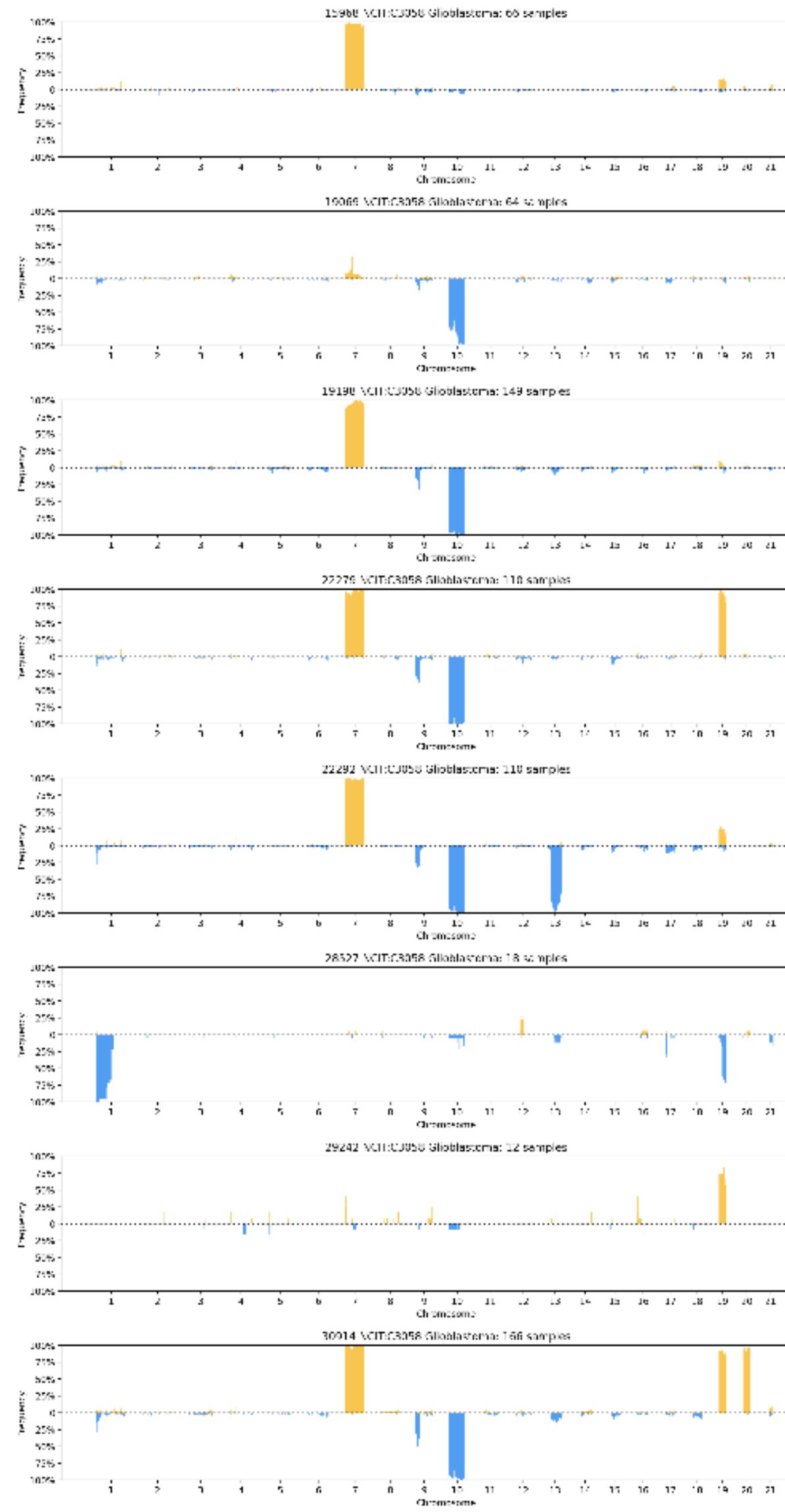
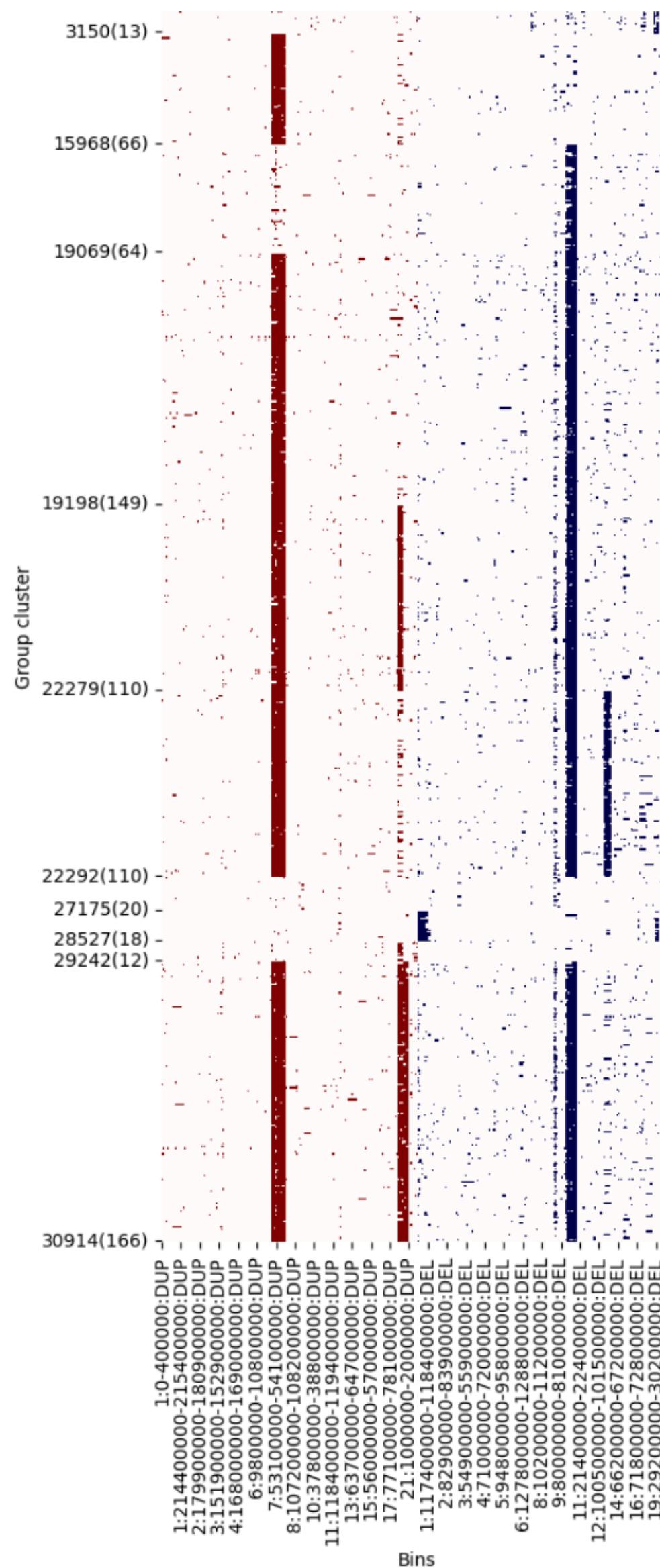
CNV profiles heterogeneity vs cancer classification

Correspondance of genomic profiles to NCIT cancer hierarchy



Results

Entity CNV heterogeneity: Glioblastoma

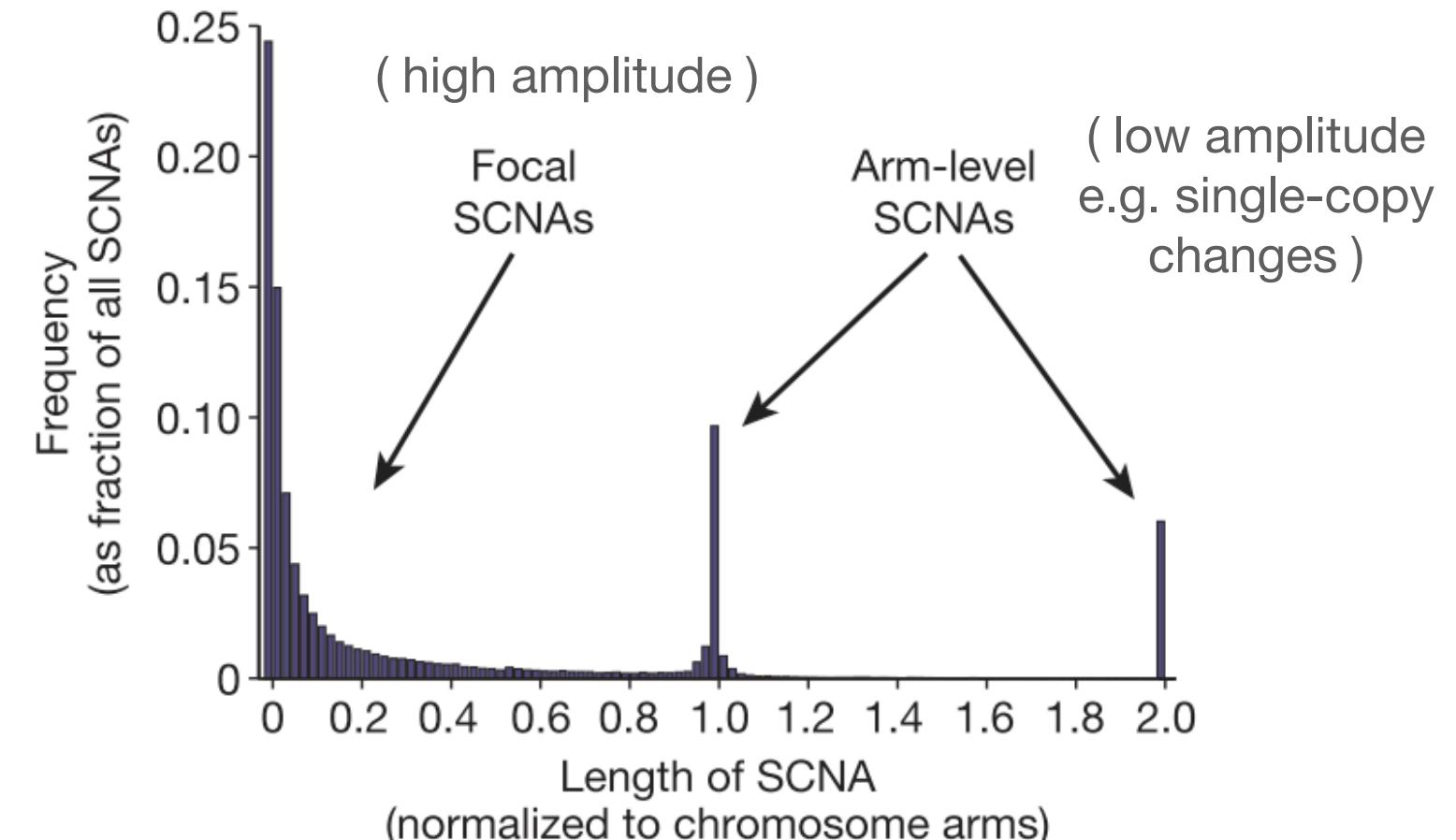


group cluster	CNV features
15968	Dup 7
19069	Del 10
19198	Dup 7, Del 10
22279	Dup 7, Del 10, Dup 19
22292	Dup 7, Del 10, Del 13
27175	Del 1p, Del 19q
28527	Del 1p, Del 19q
29242	Dup 19
30914	Dup 7, Del 10, Dup 19, Dup 20

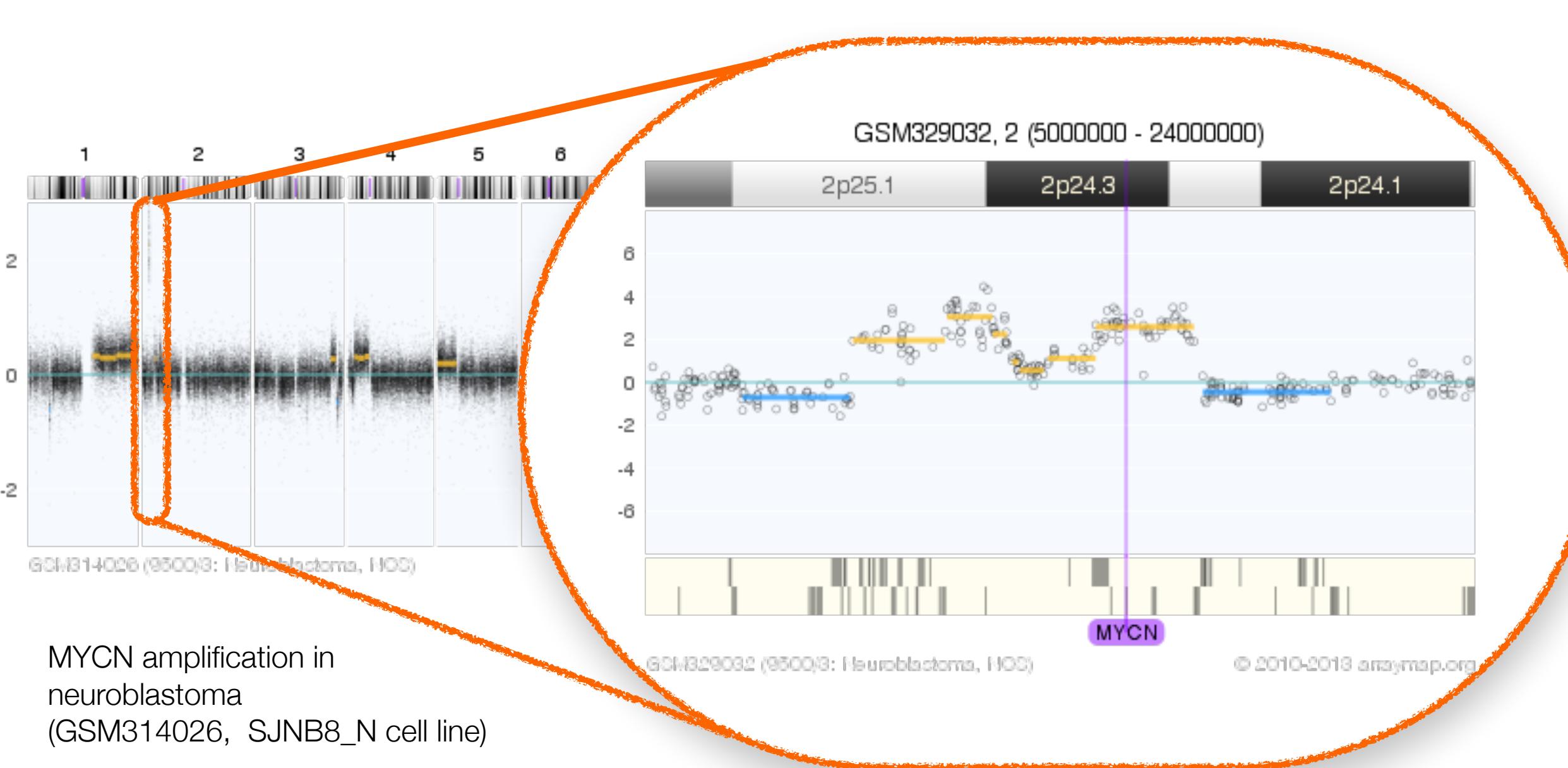


CNV Categorization

different levels of CNV



Rameen et al 2010 Nature



Lead: Hangjia Zhao

CopyNumberChange

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

Computational Definition

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

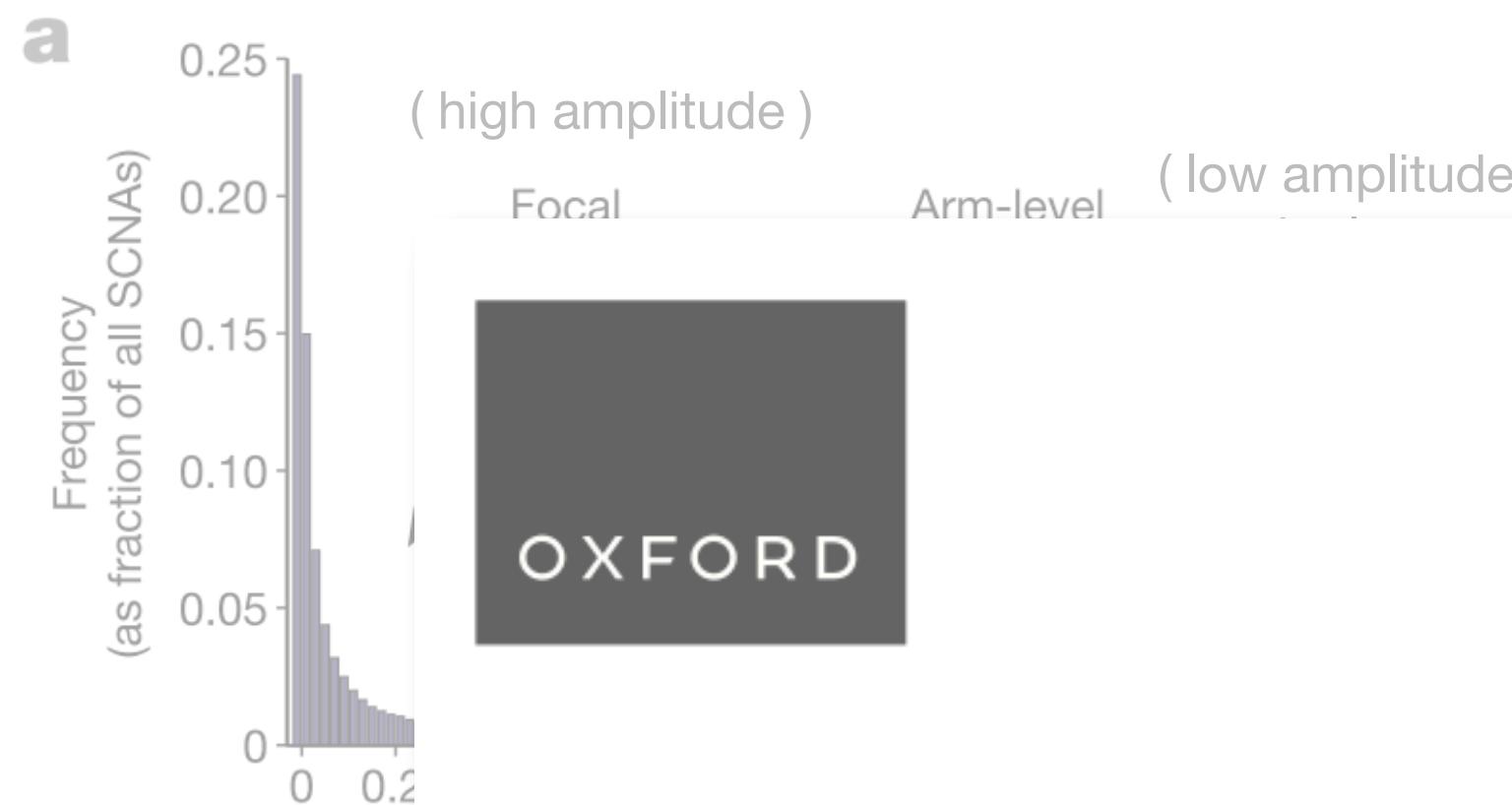
Information Model

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

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

CNV Categorization

different levels of CNV



CopyNumberCham

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

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

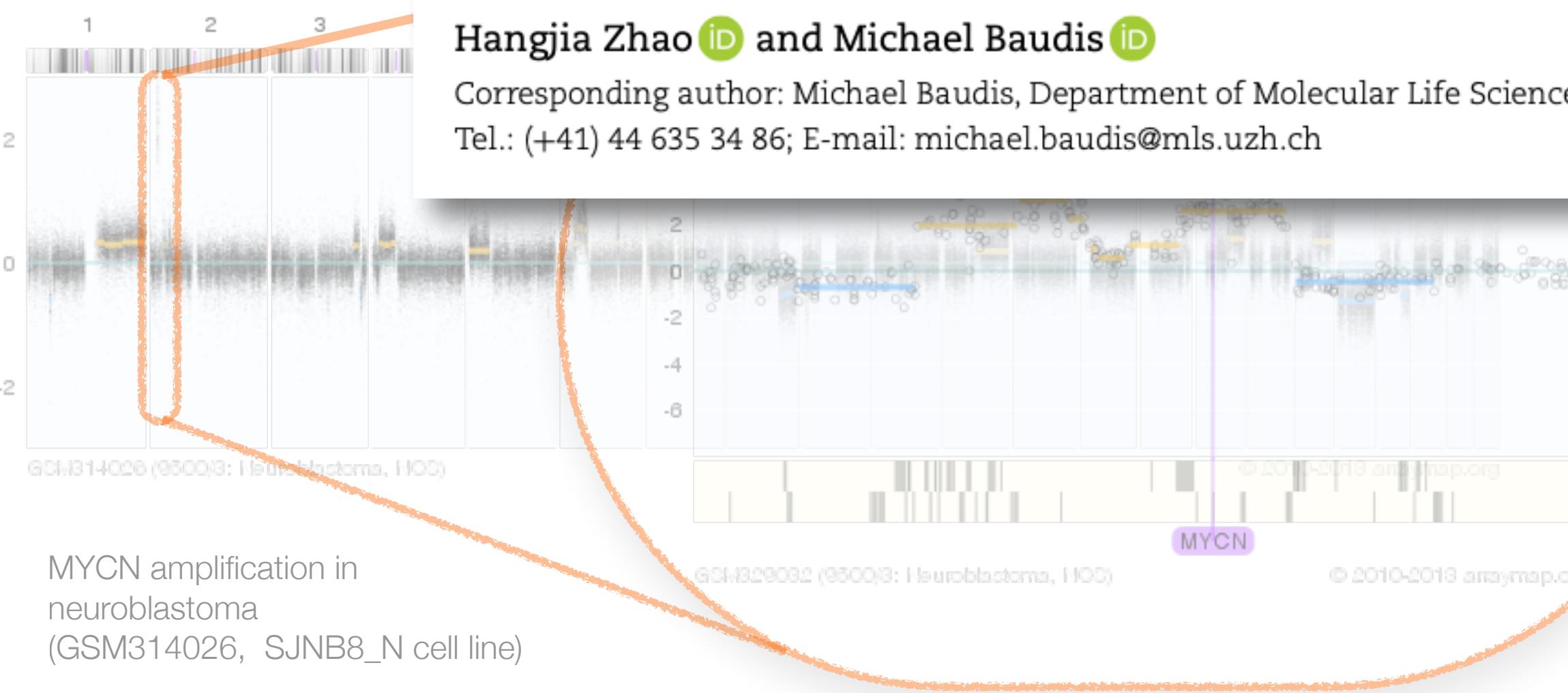
a system (e.g. genome, cell,

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

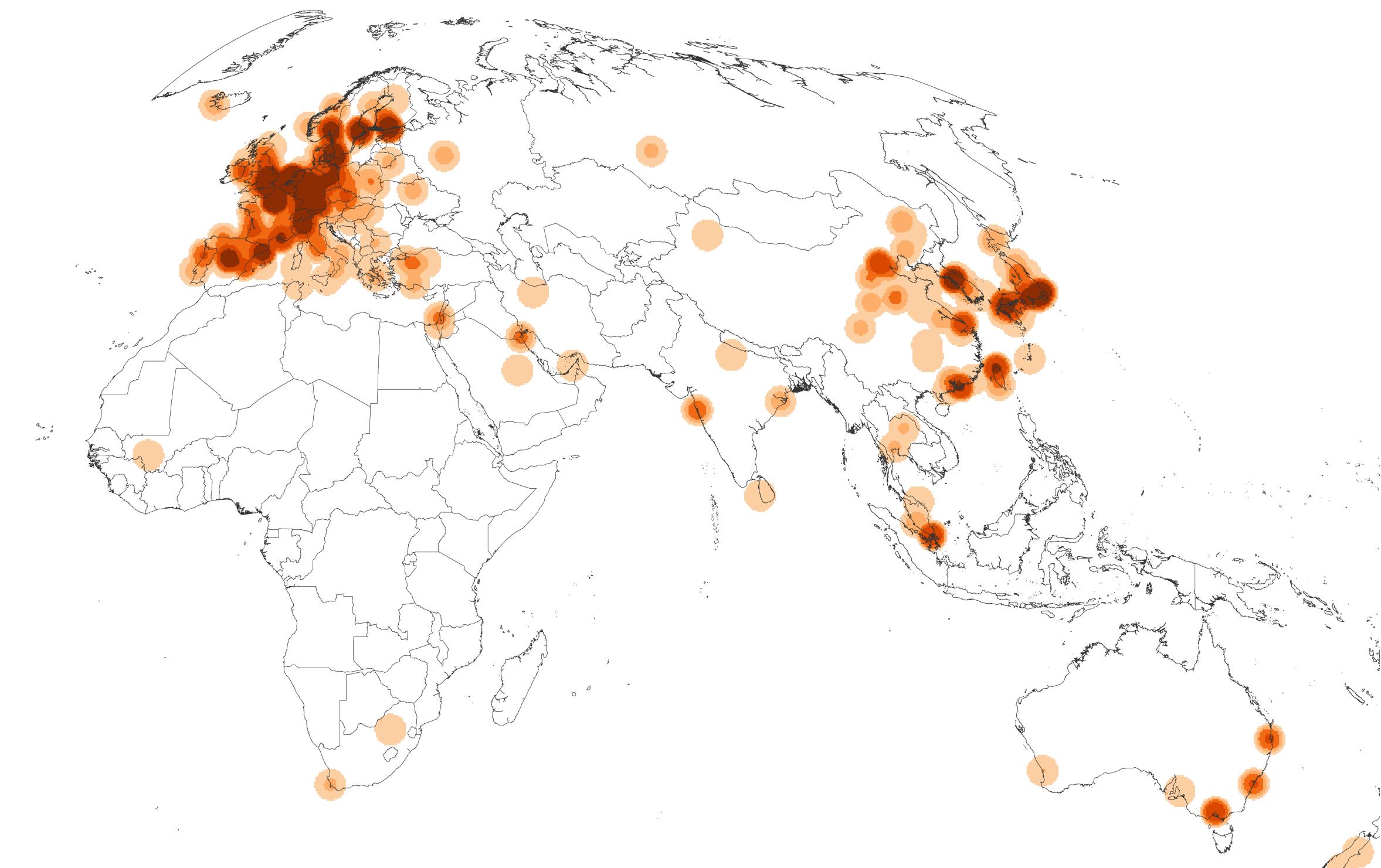
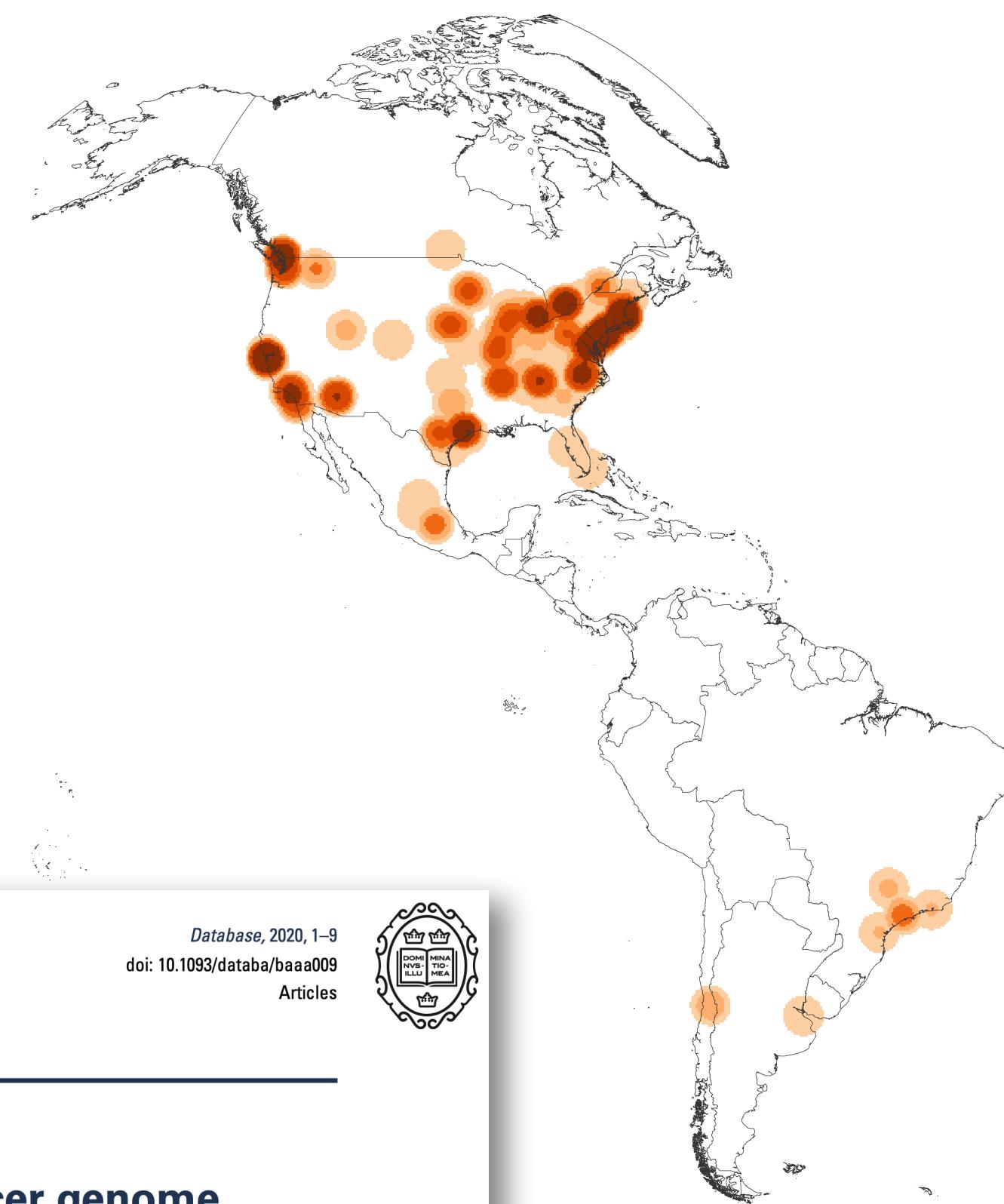
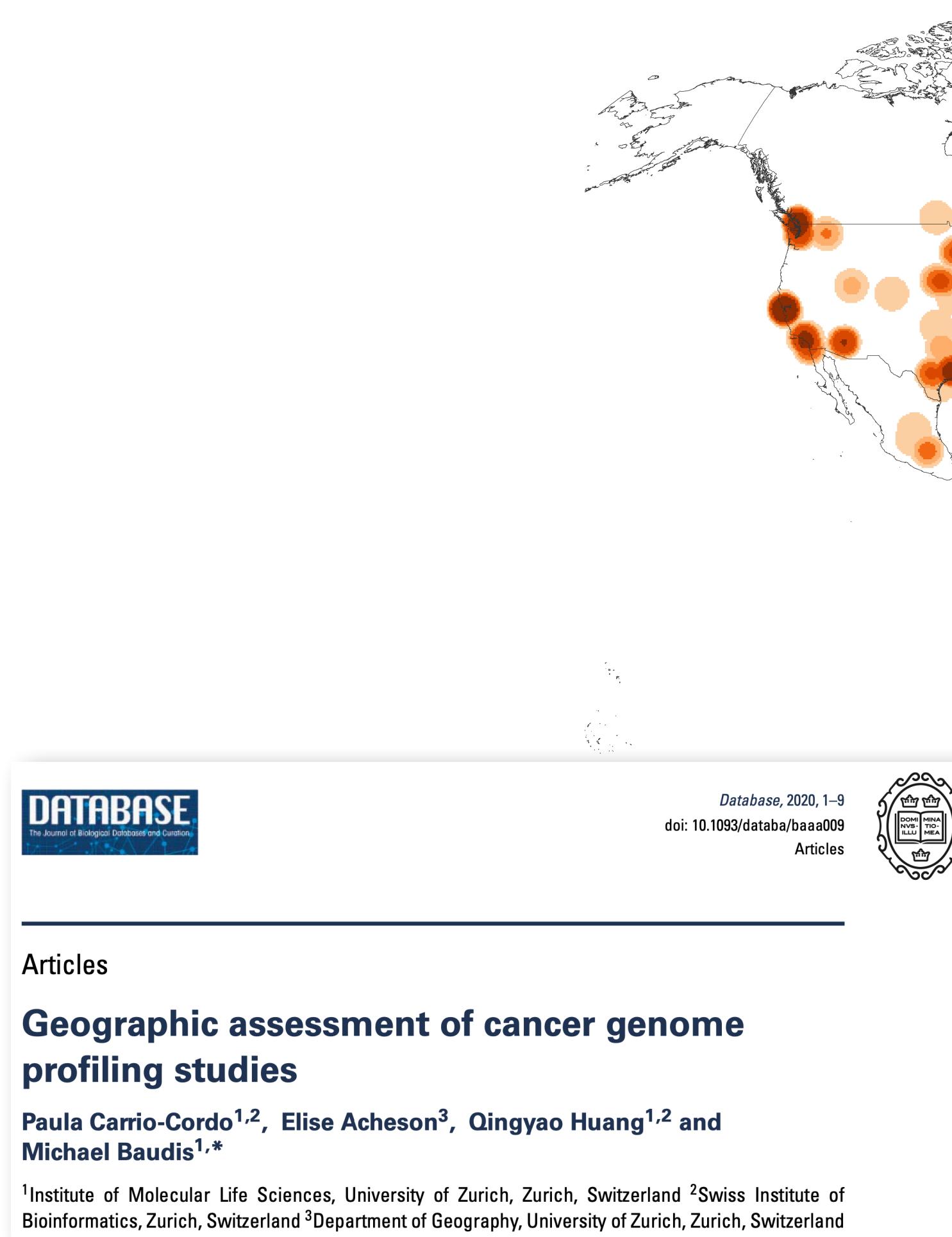
Tel.: (+41) 44 635 34 86; E-mail: michael.baudis@mls.uzh.ch



_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

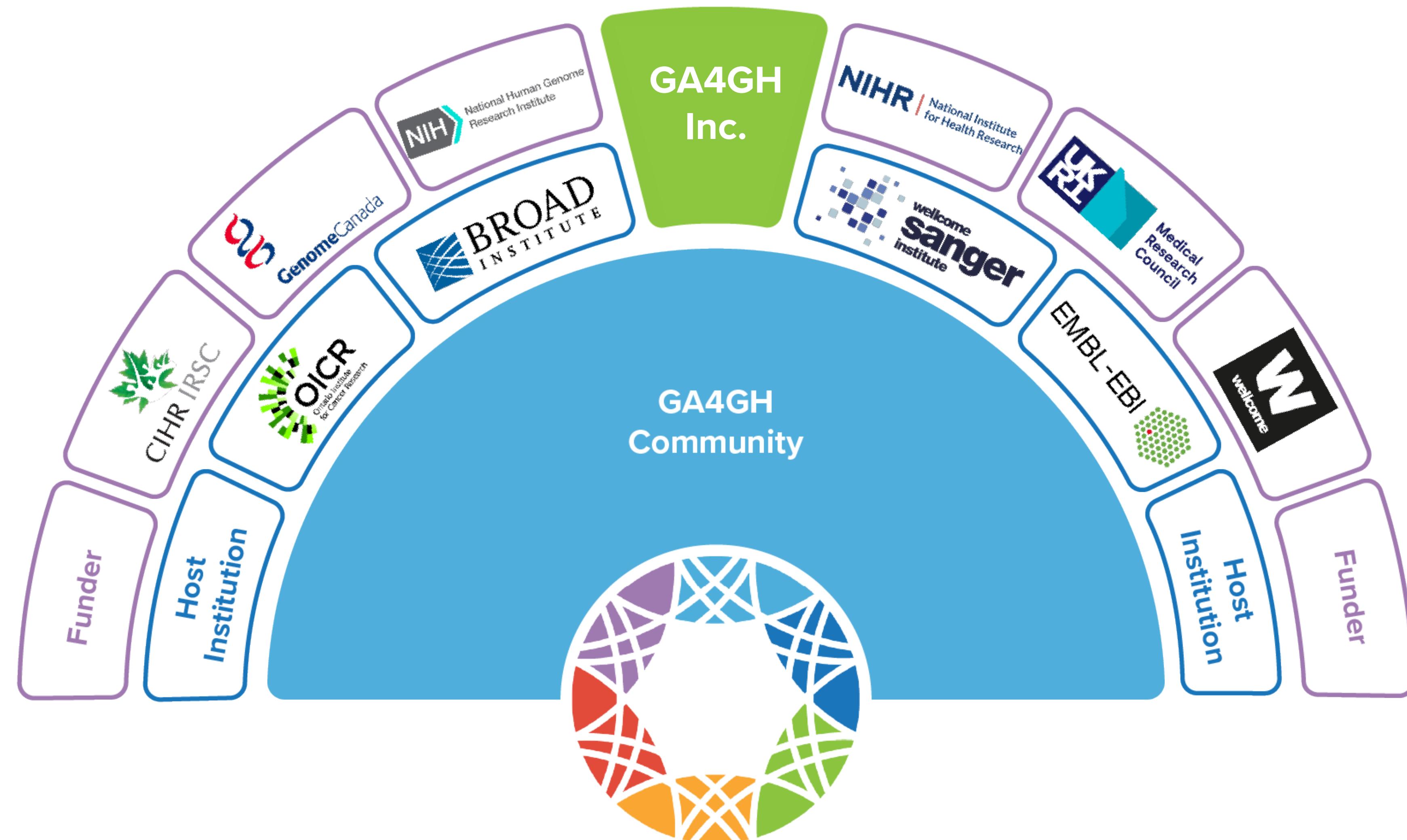
The Global Alliance for Genomics
and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

Organization

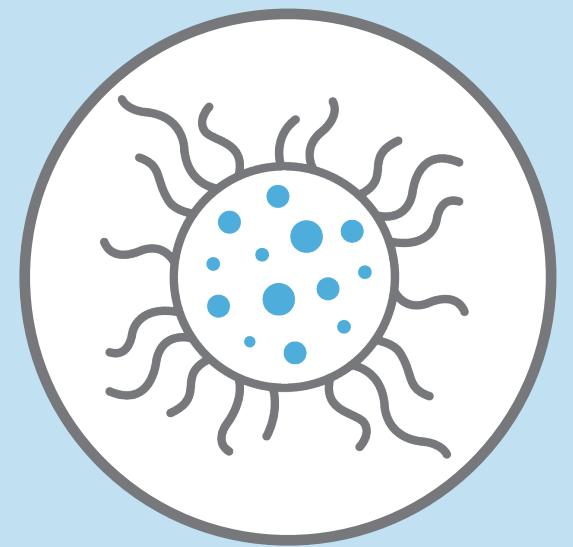


Global Alliance
for Genomics & Health

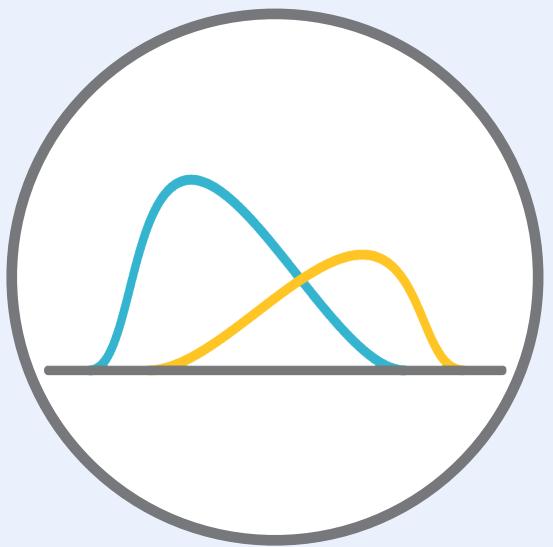




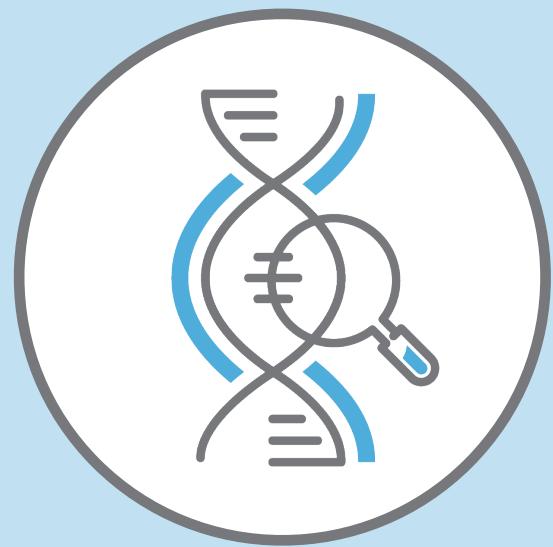
Global Genomic Data Sharing Can...



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



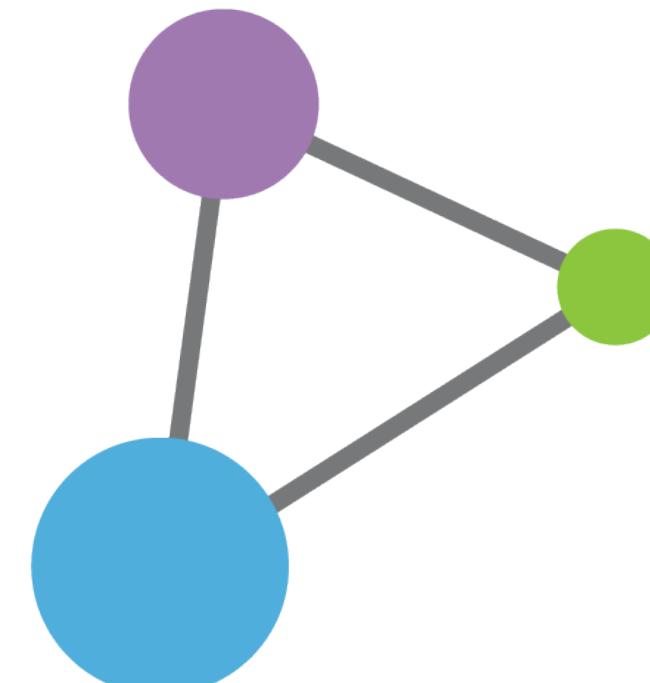
Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets

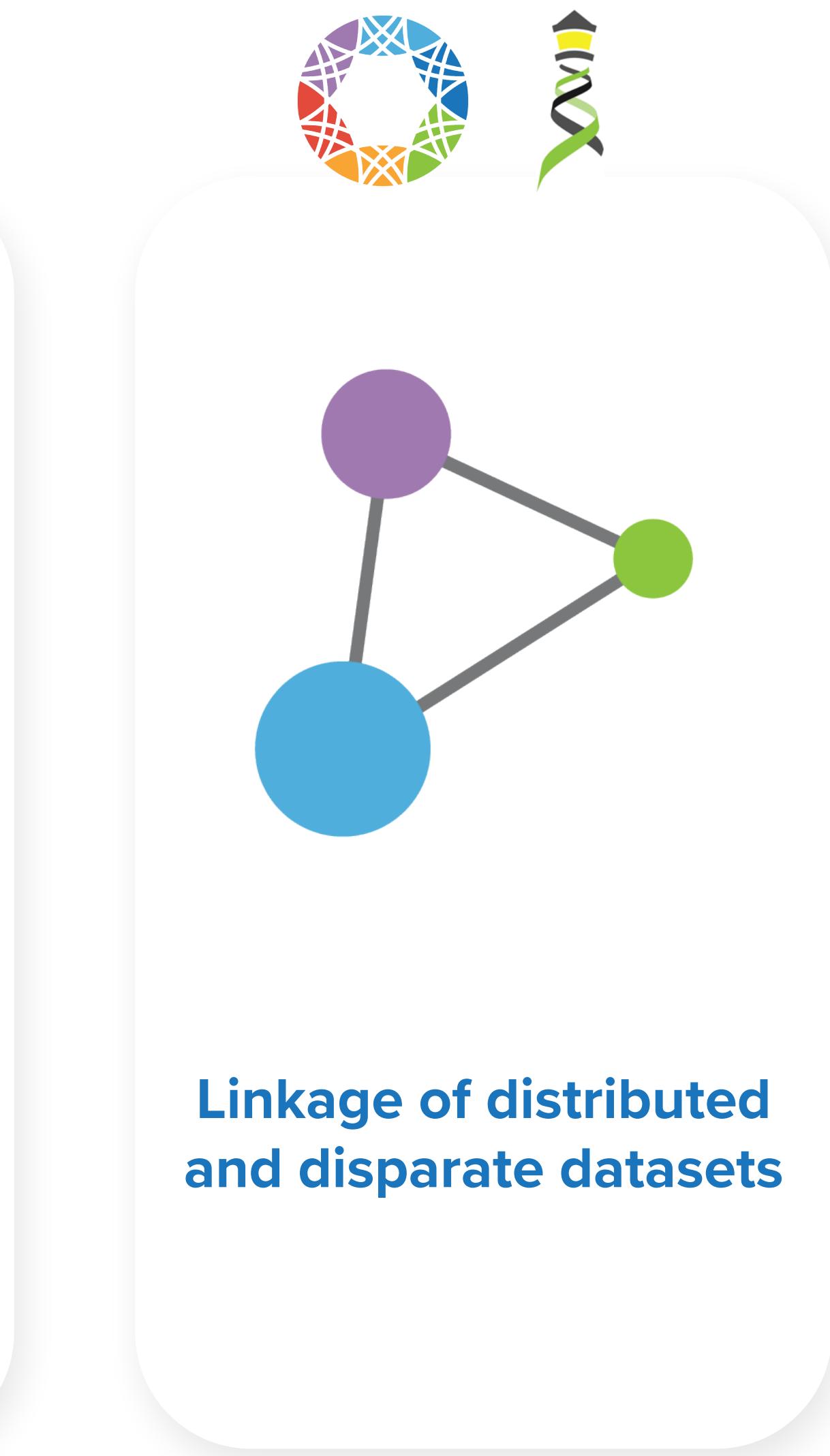
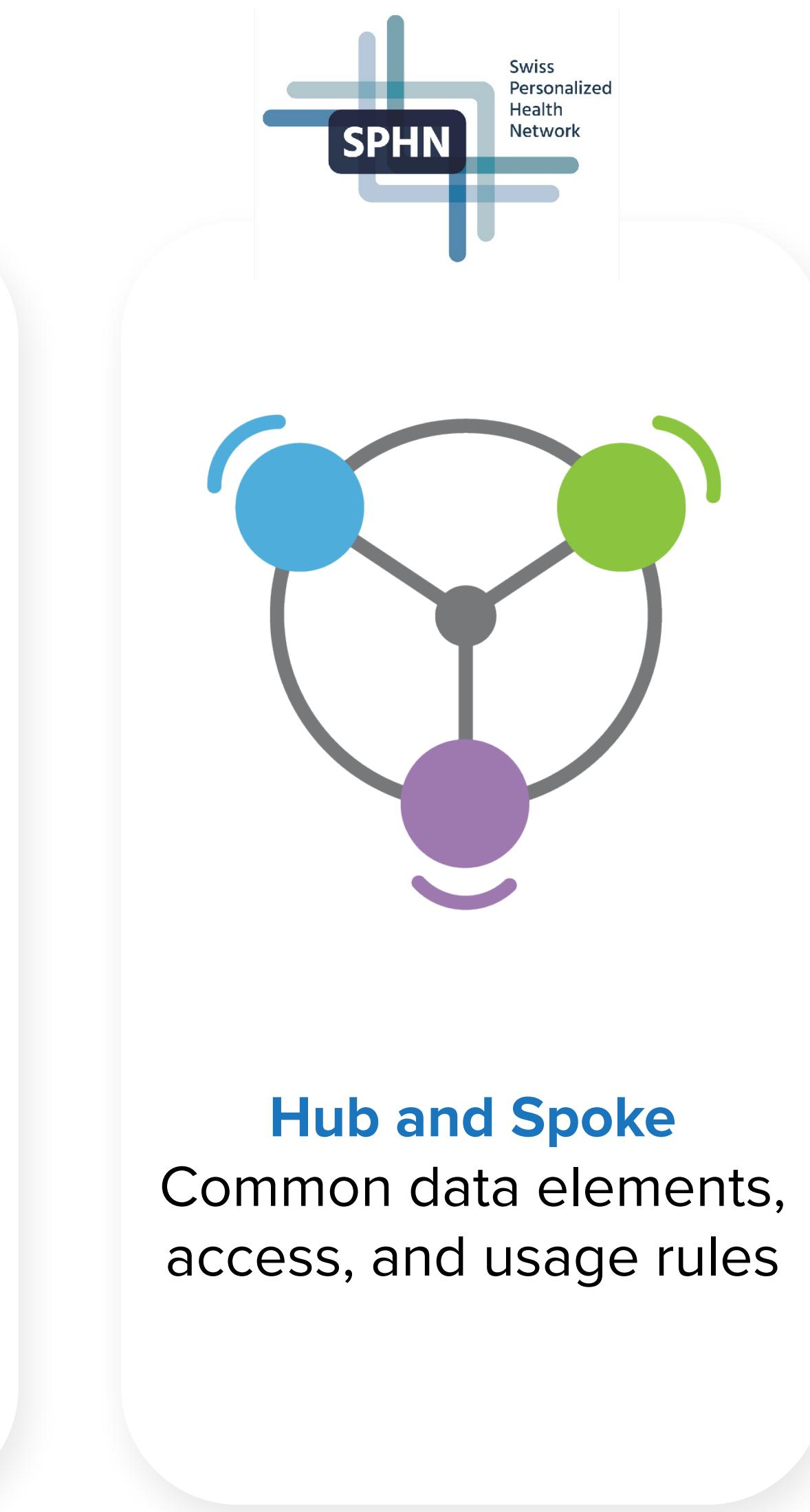
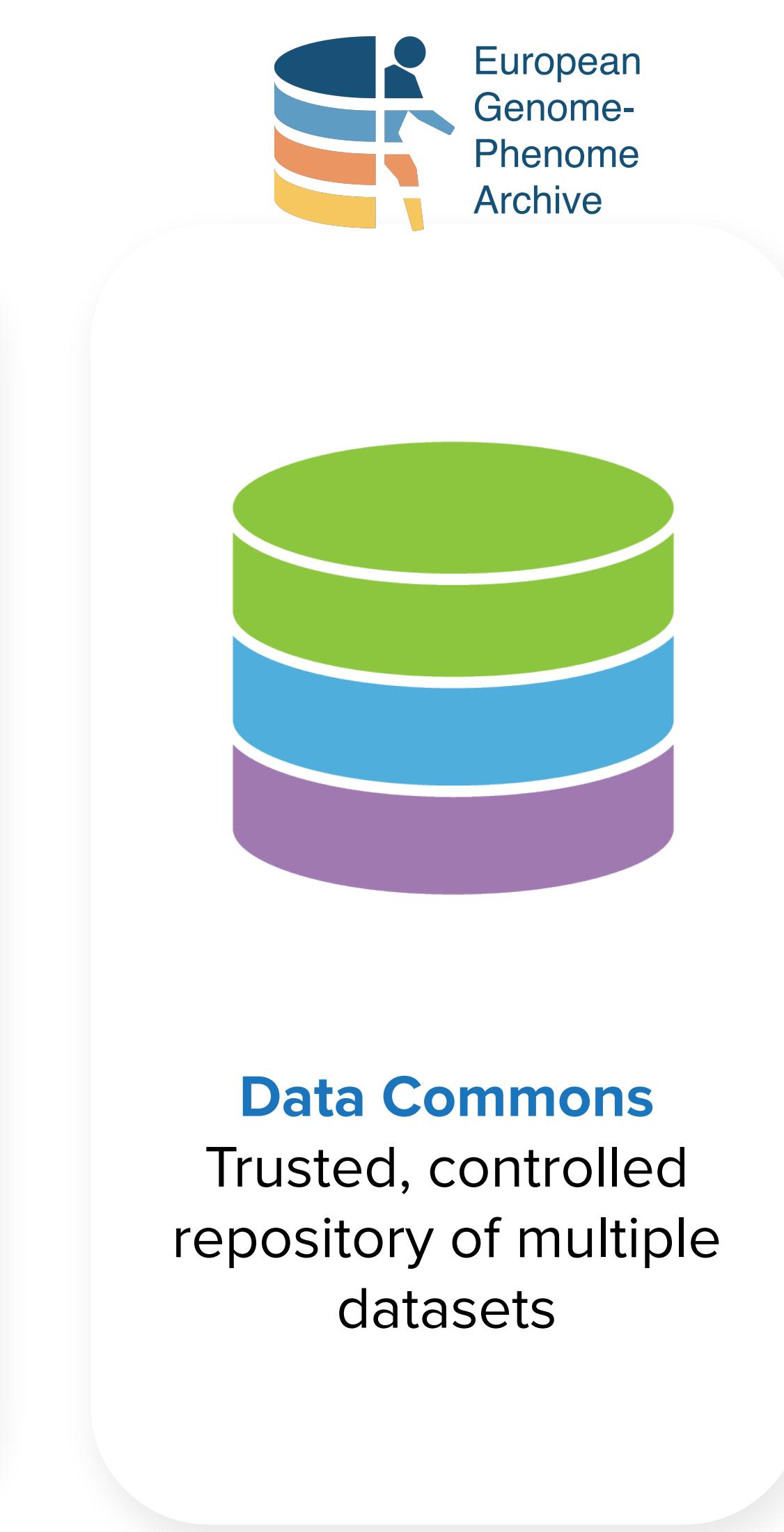


Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

Different Approaches to Data Sharing



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*”)

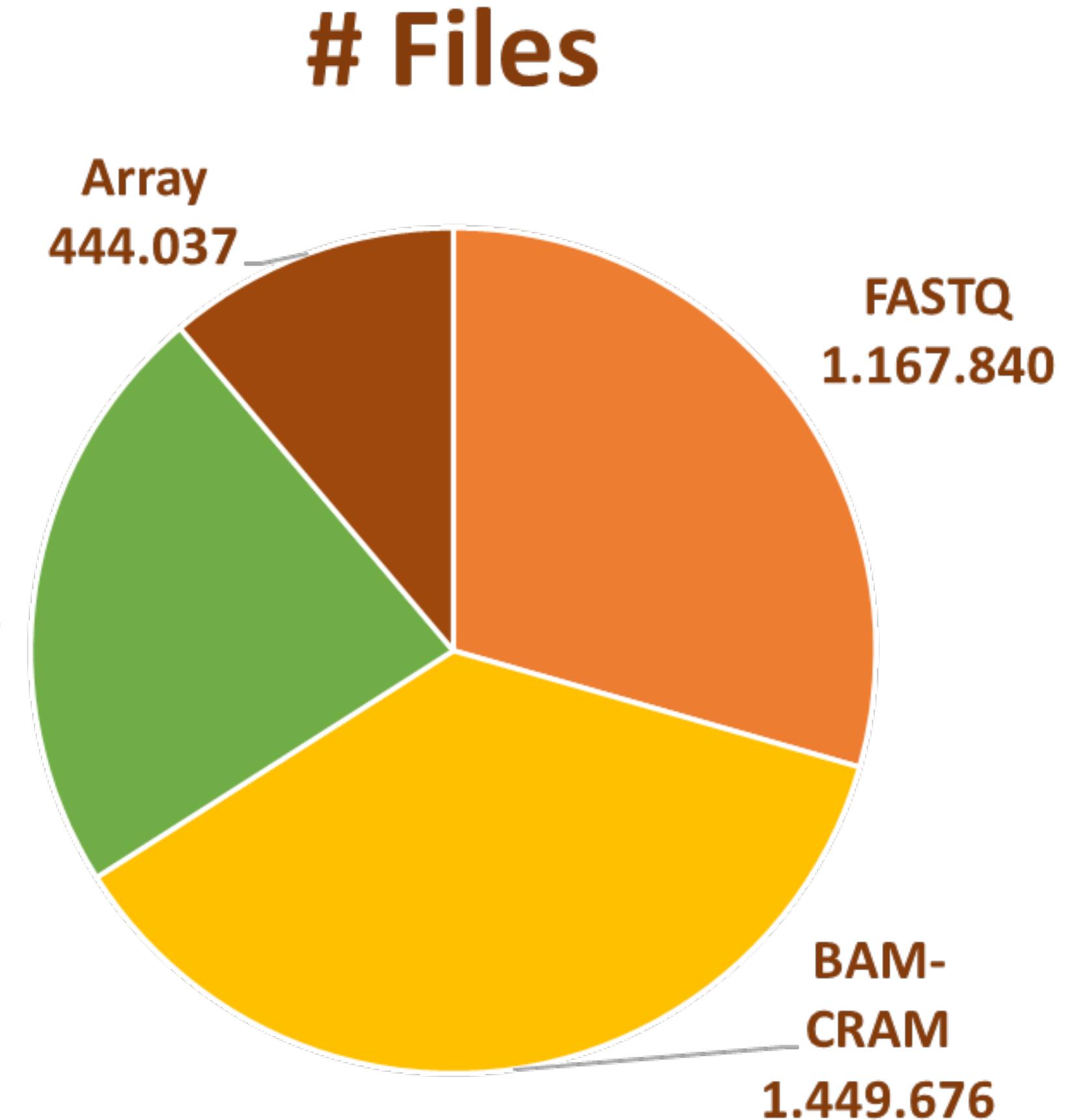


The EGA



- 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

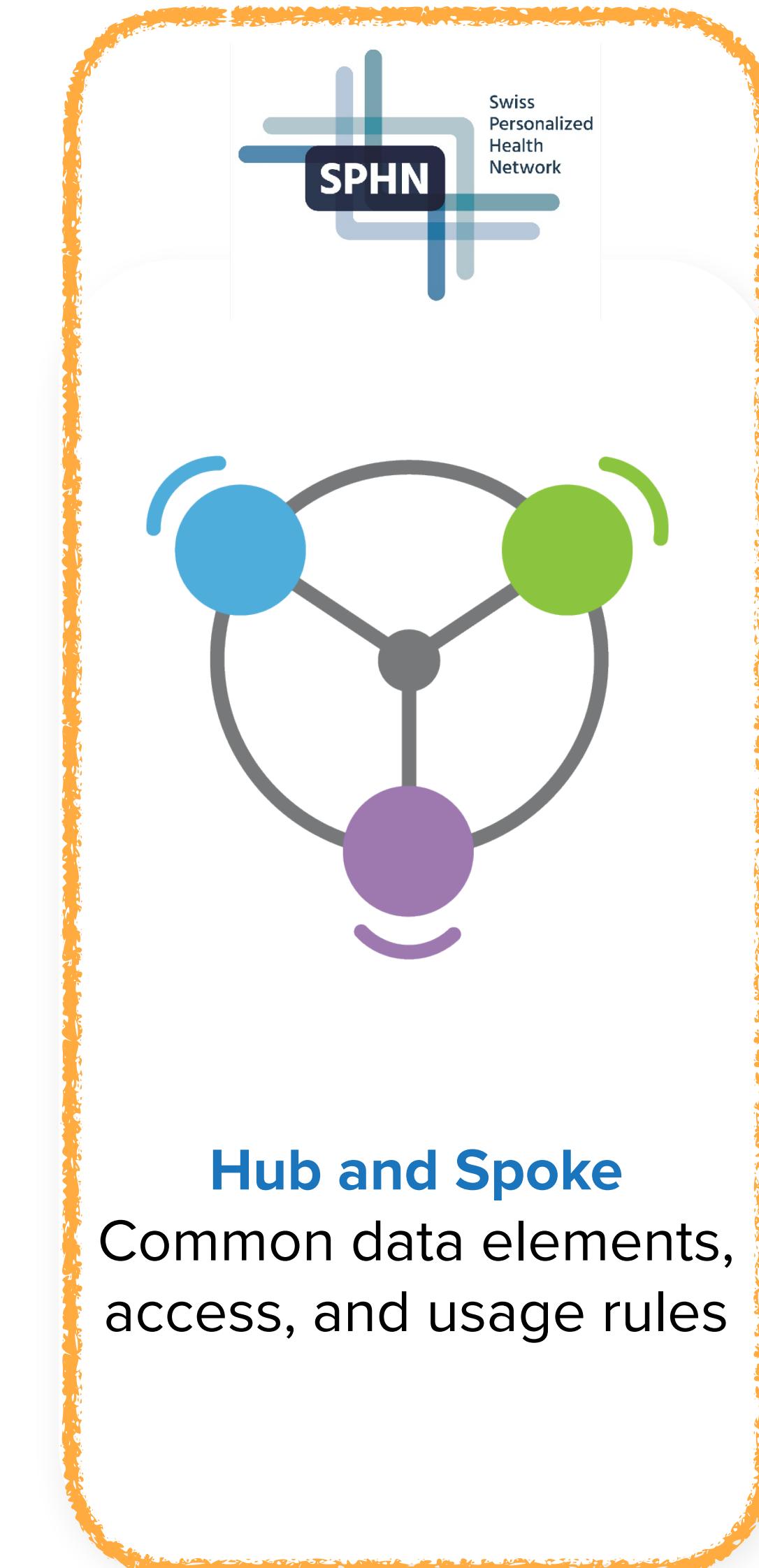
Different Approaches to Data Sharing



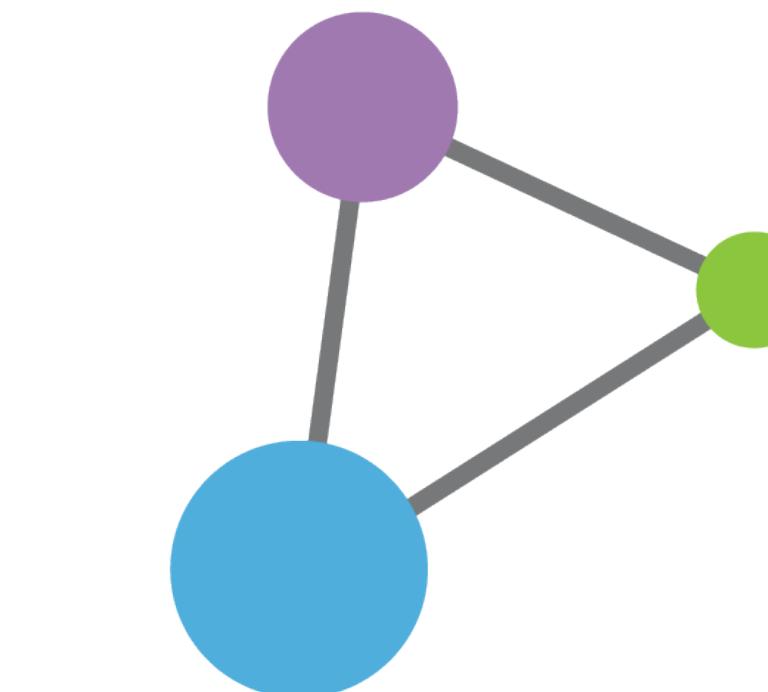
Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets

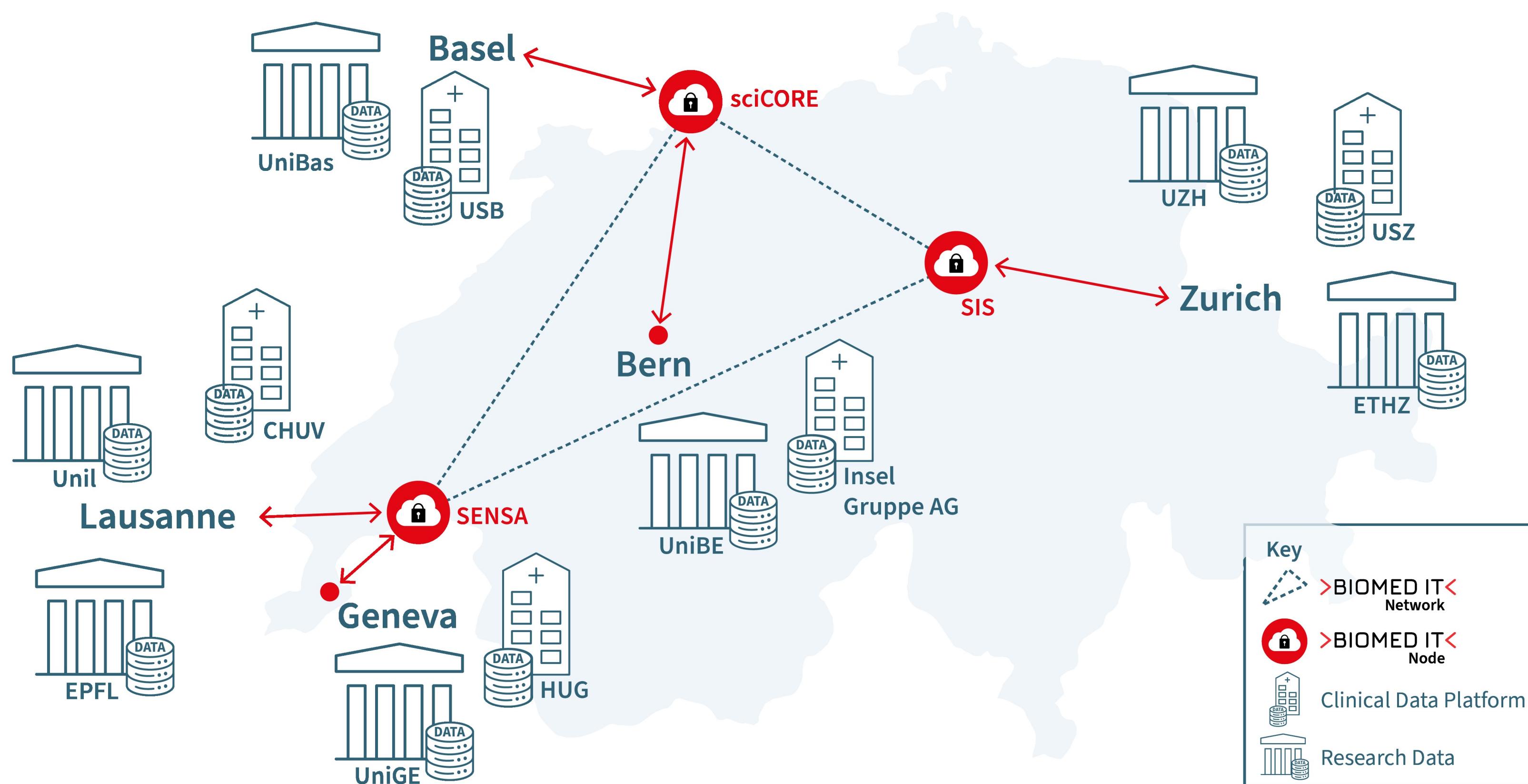


Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate 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

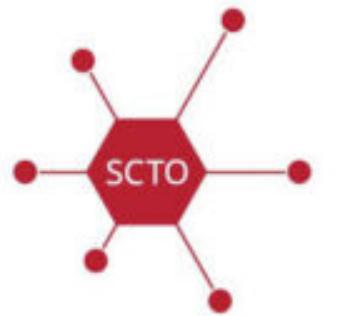
ehealthsuisse

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

life sciences cluster basel

 **Université Medizin Schweiz
Médecine Universitaire Suisse**

swissuniversities



Different Approaches to Data Sharing



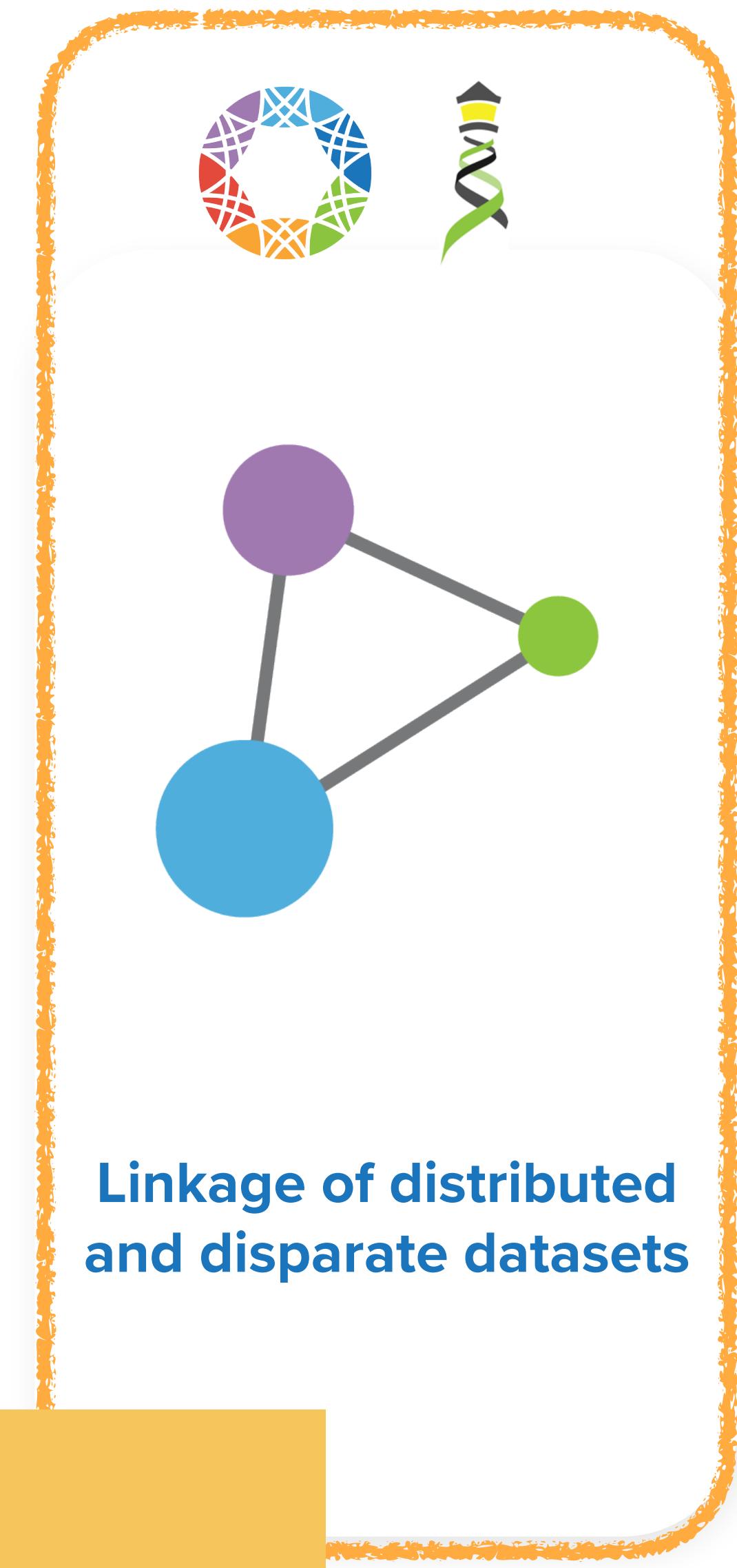
Centralized Genomic Knowledge Bases



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

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

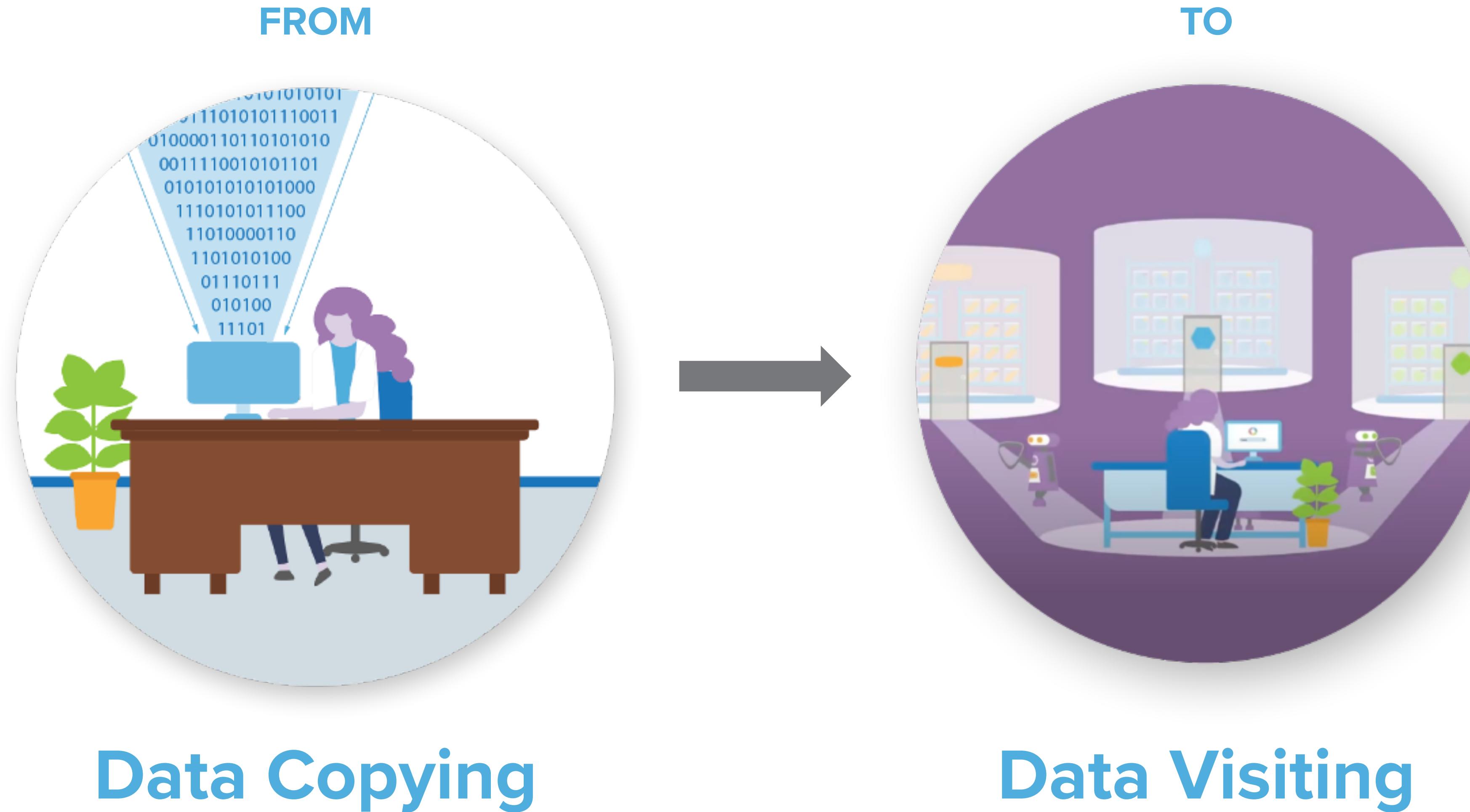
Perspective

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

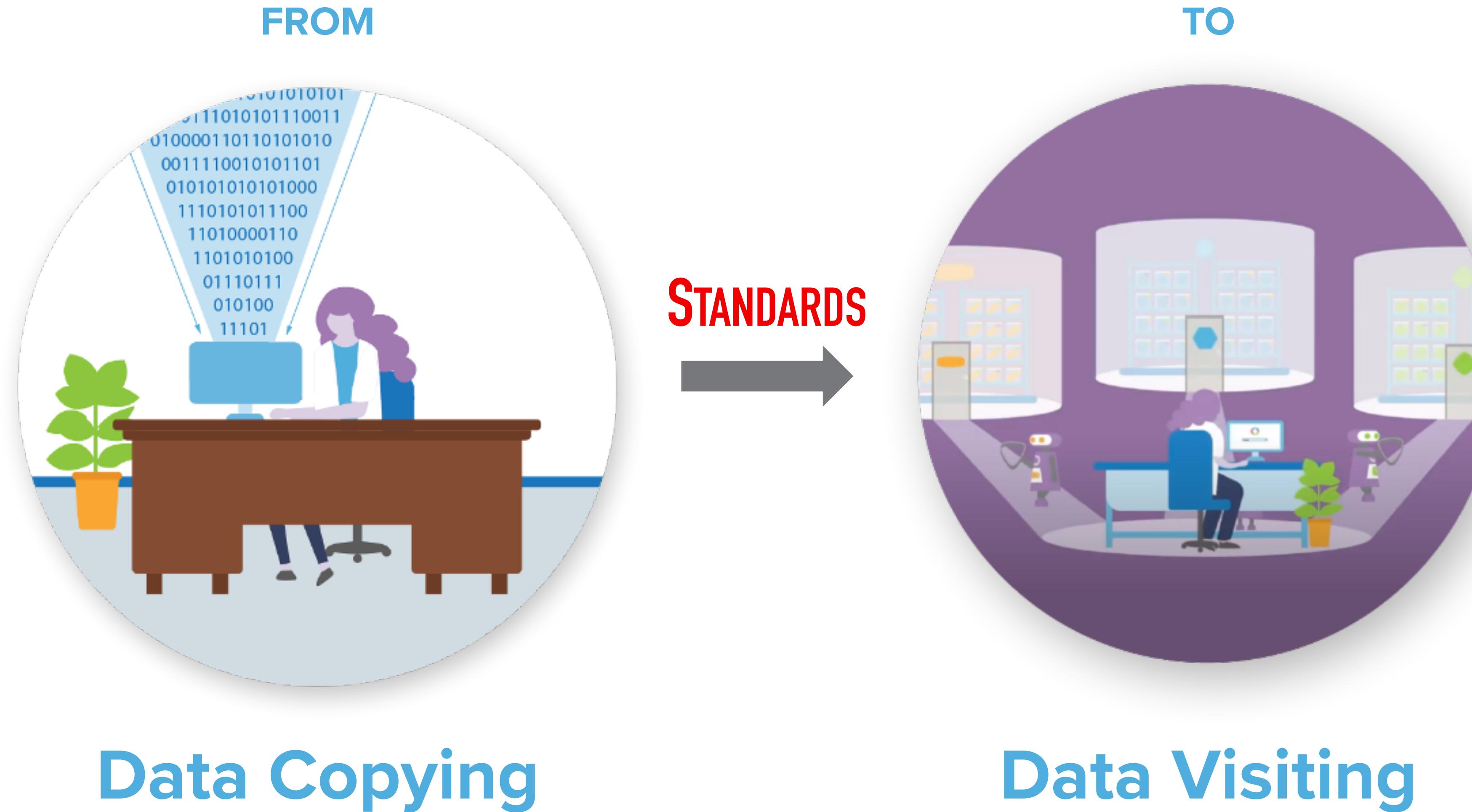
Heidi L. Rehm,^{1,2,47} Angela J.H. Page,^{1,3,*} Lindsay Smith,^{3,4} Jeremy B. Adams,^{3,4} Gil Alterovitz,^{5,47} Lawrence J. Babb,¹ Maxmillian P. Barkley,⁶ Michael Baudis,^{7,8} Michael J.S. Beauvais,^{3,9} Tim Beck,¹⁰ Jacques S. Beckmann,¹¹ Sergi Beltran,^{12,13,14} David Bernick,¹ Alexander Bernier,⁹ 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

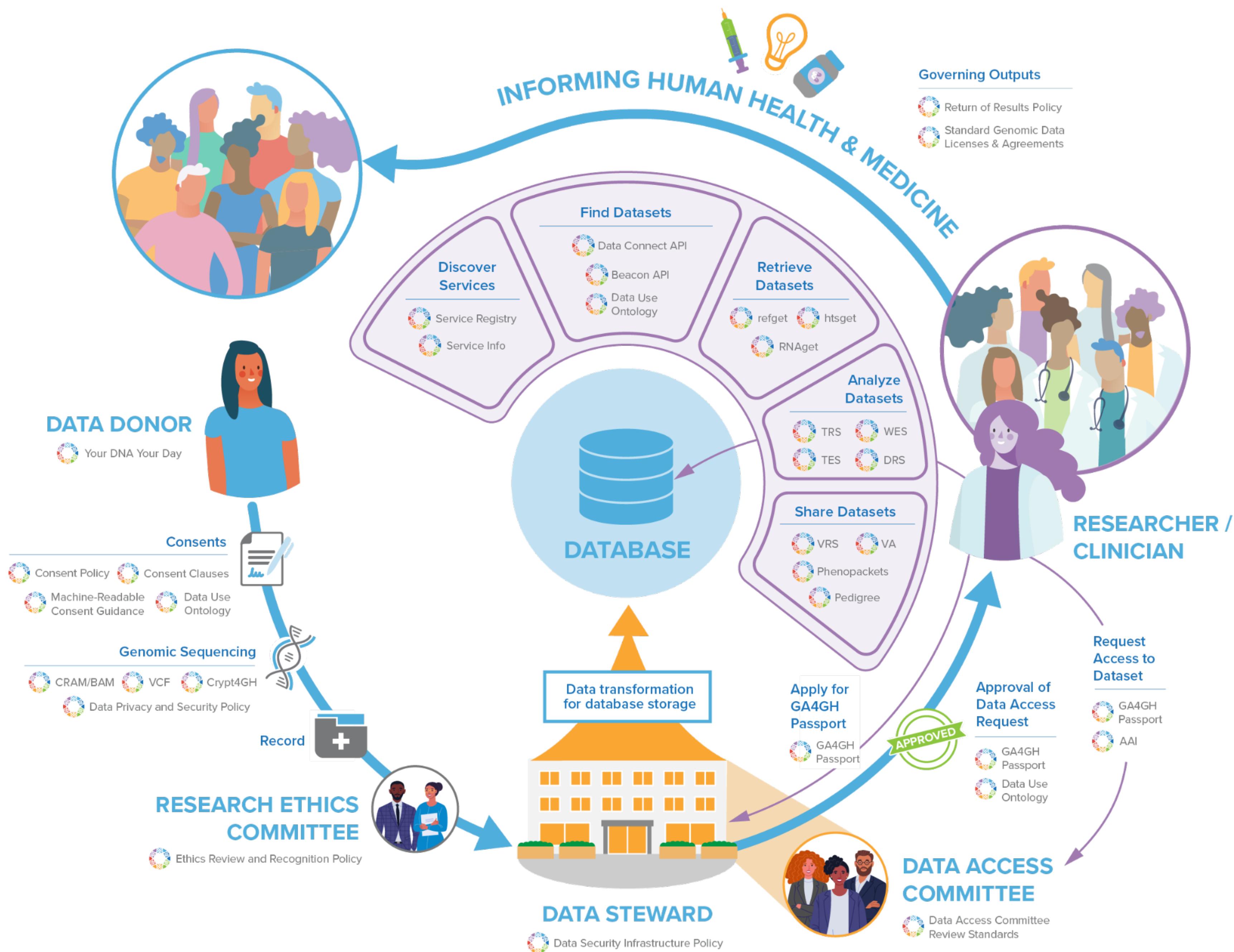
(Author list continued on next page)

A New Paradigm for Data Sharing

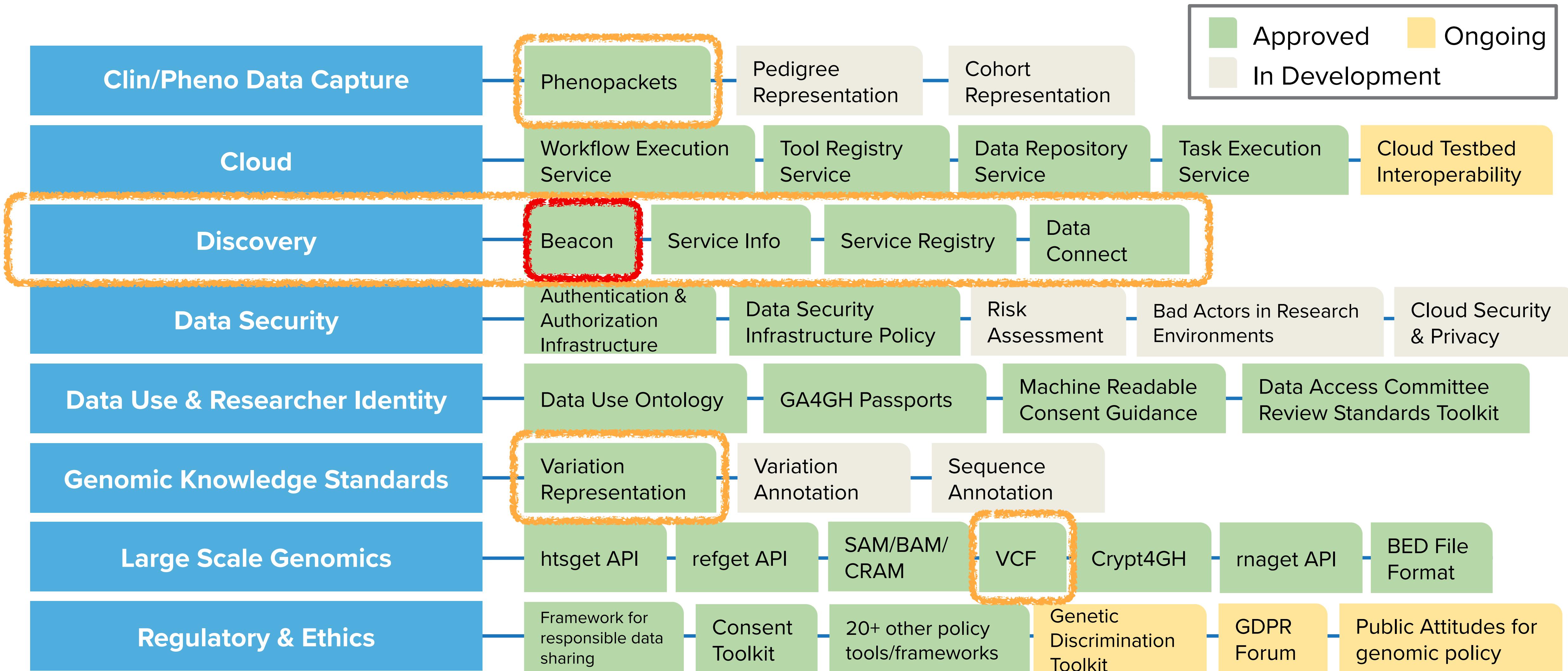


A New Paradigm for Data Sharing





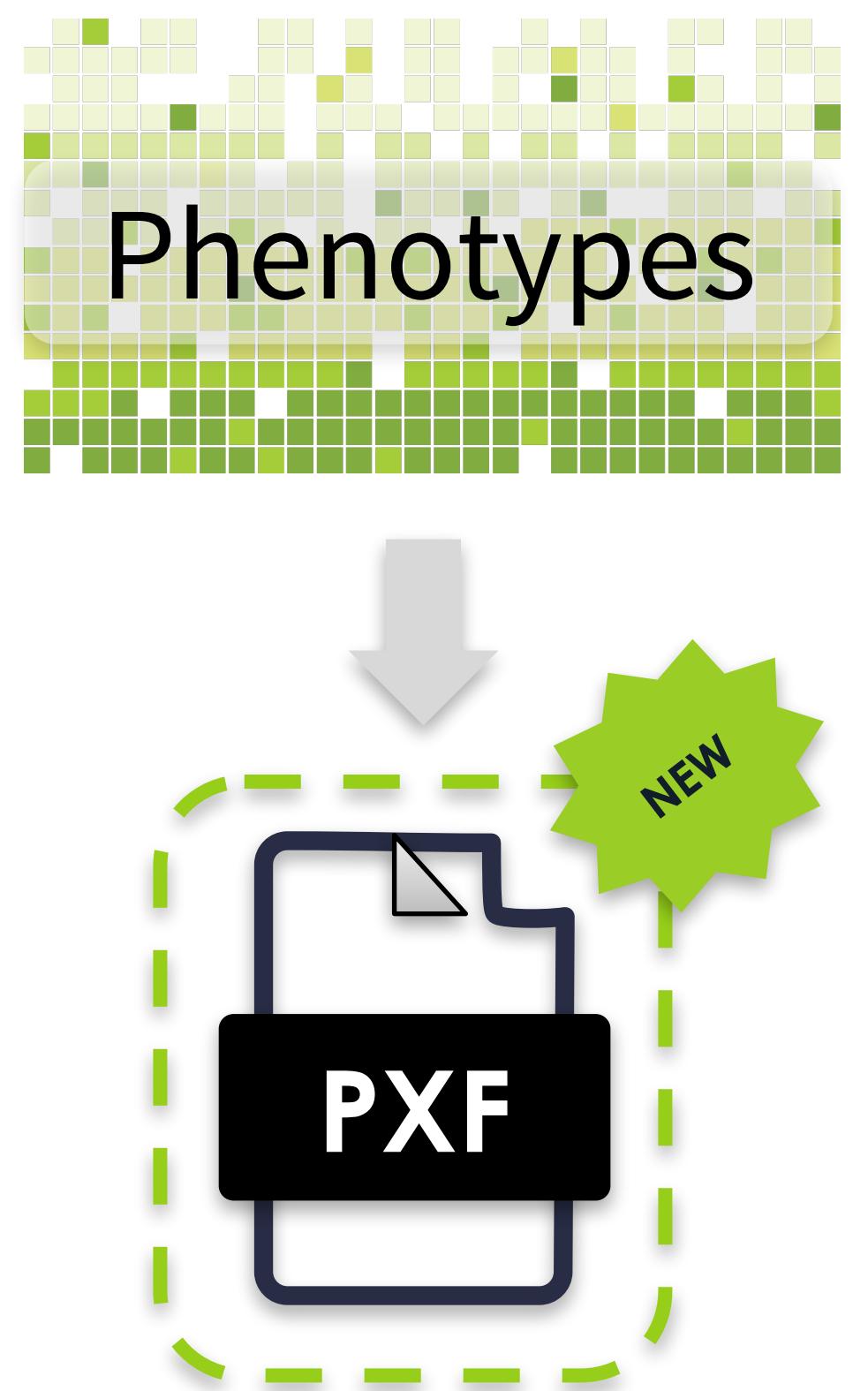
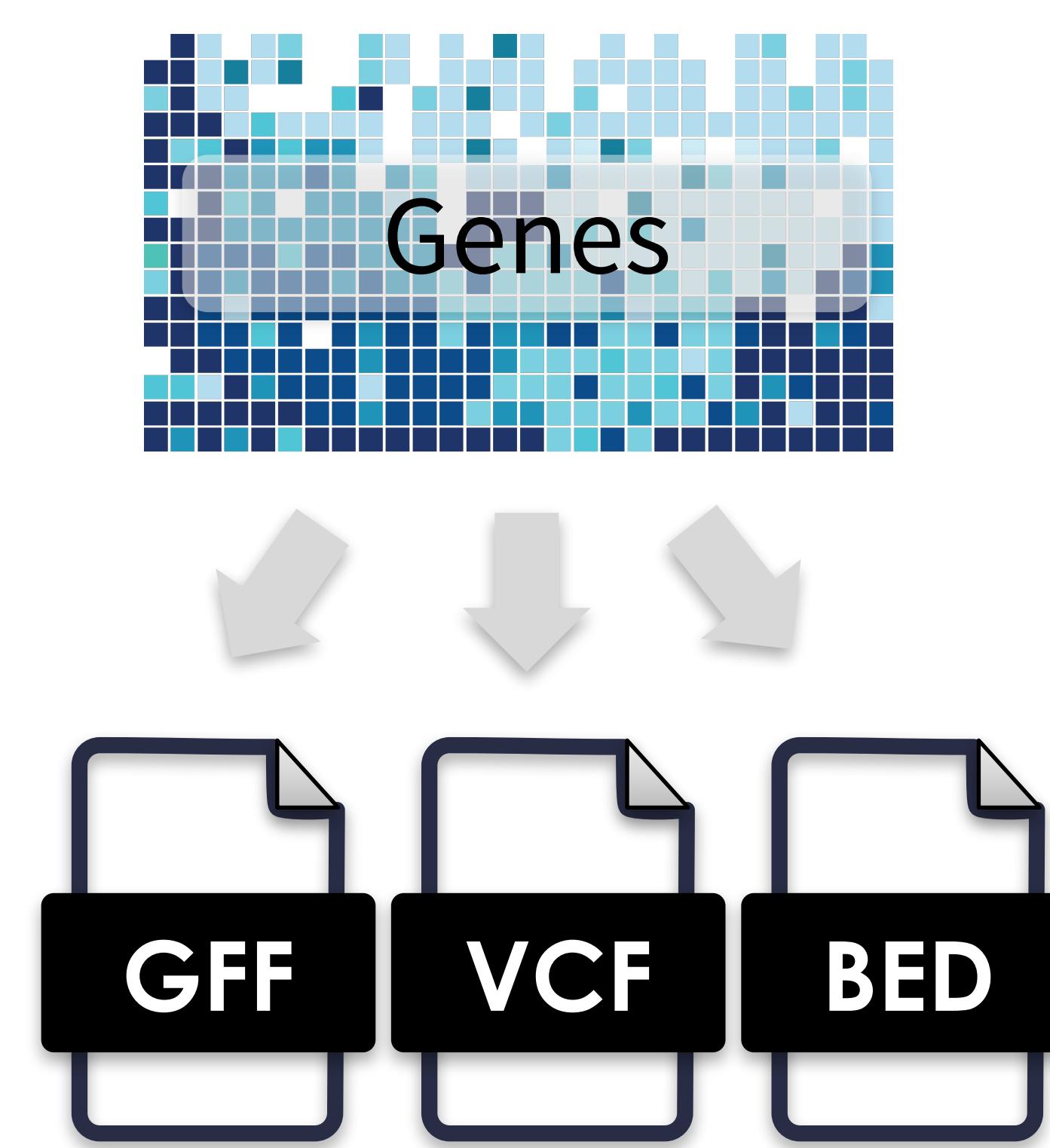
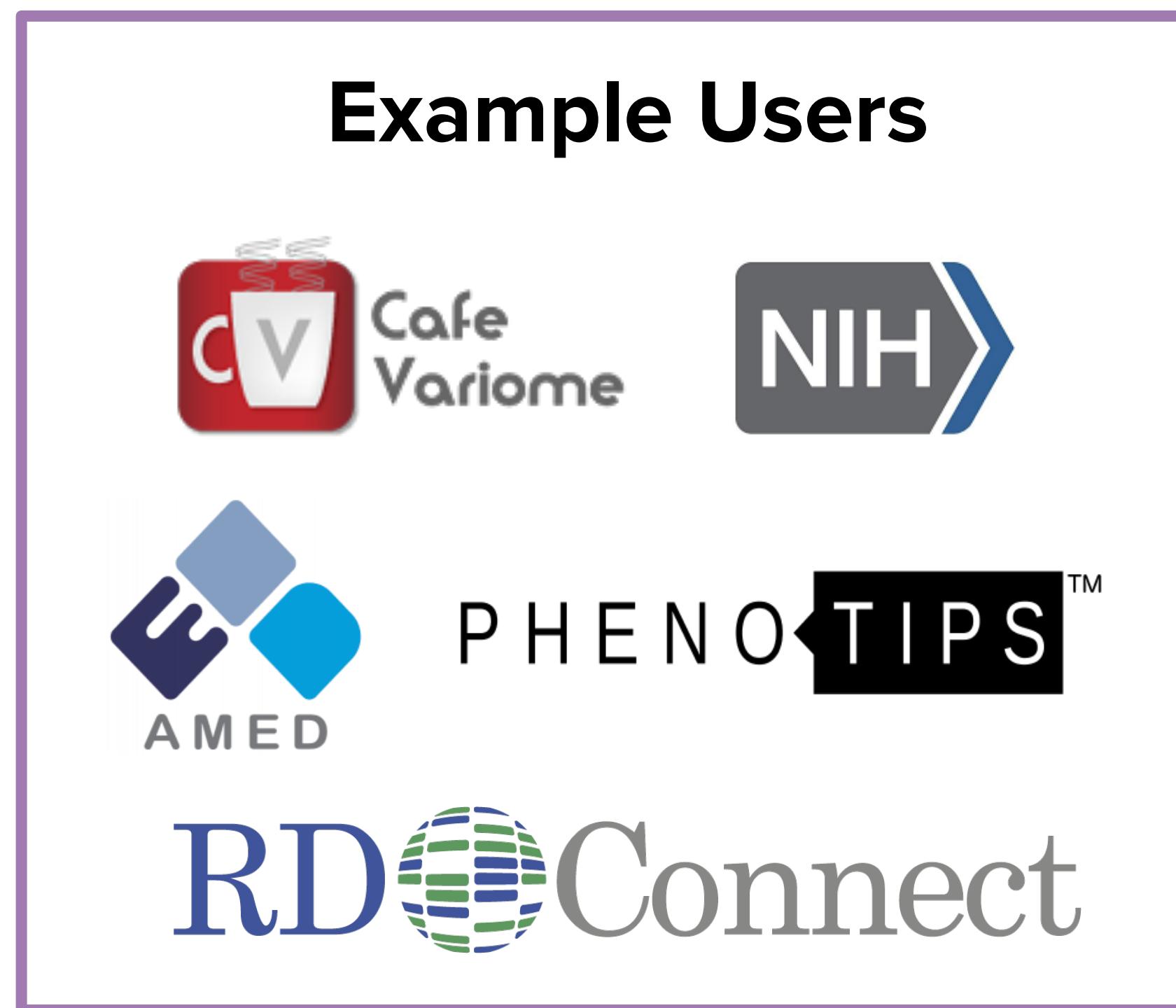
Overview of GA4GH standards and frameworks



Phenopackets v2

Phenopackets is a standard schema for sharing phenotypic information.

Approved: June 24, 2021



VCF/BCF

The Variant Call Format (VCF) specifies the format of a text file used in bioinformatics for storing gene sequence variations. The Binary Call Format (BCF) is the Binary equivalent, smaller and more efficient to process.

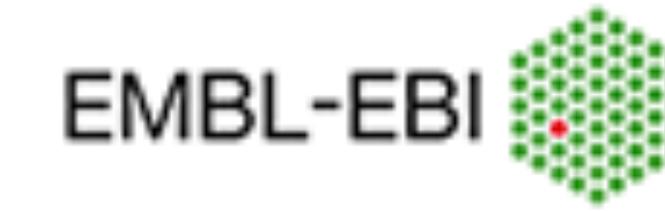
Software Libraries: [htslib](#) | [htsjdk](#)

Tools: [Samtools](#) | [BCFtools](#)

Databases: [European Variation Archive \(EVA\)](#) | [dbGAP](#) | [dbSNP](#) | [1000 Genomes Projects / IGSR](#)

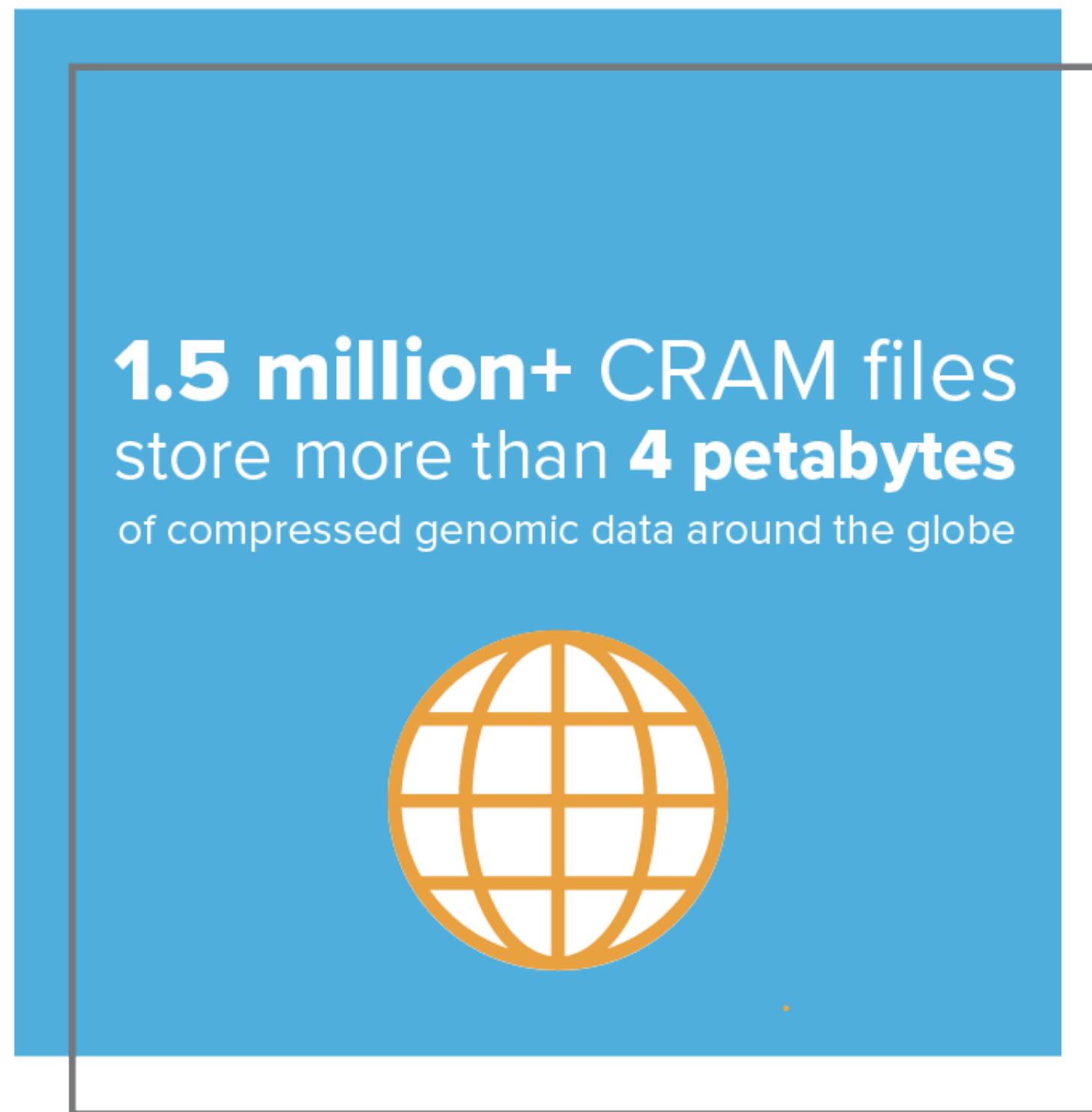
Genome Browsers: [ENSEMBL](#) | [JBrowse](#) | [UCSC Genome Browser](#)

**Example
Users**

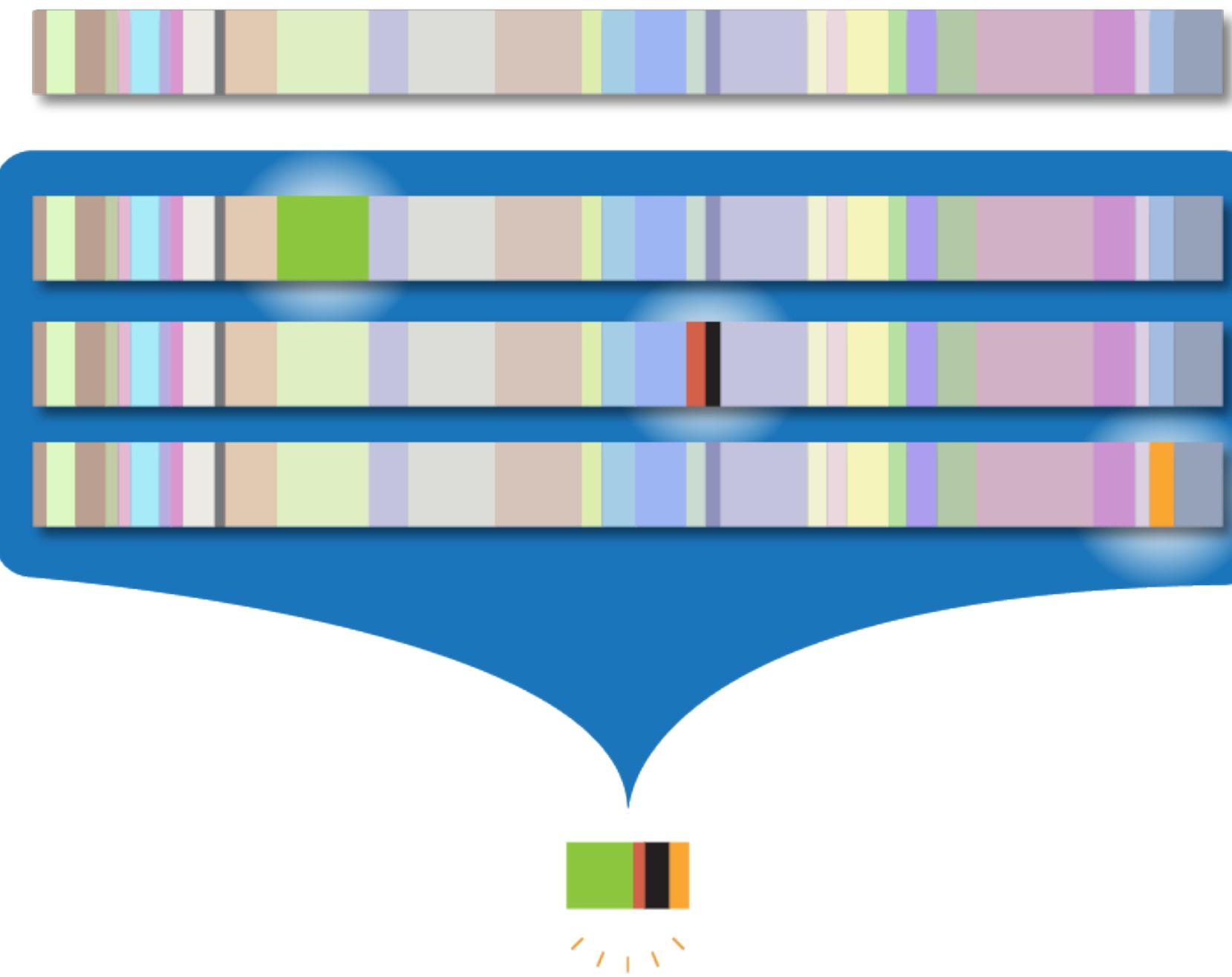


CRAM

CRAM is a file format for storing compressed genomic data. To make files small and efficient, the algorithm compresses information by only storing the parts that are different from the reference human genome.



CRAM compresses data by only storing the difference.



Genomics England implements GA4GH API to provide secure access to genomic data for the NHS

Genomics England has implemented the standard GA4GH API to provide secure access to genomic data for the NHS Genomes Program and the Genomic Medicine Service.



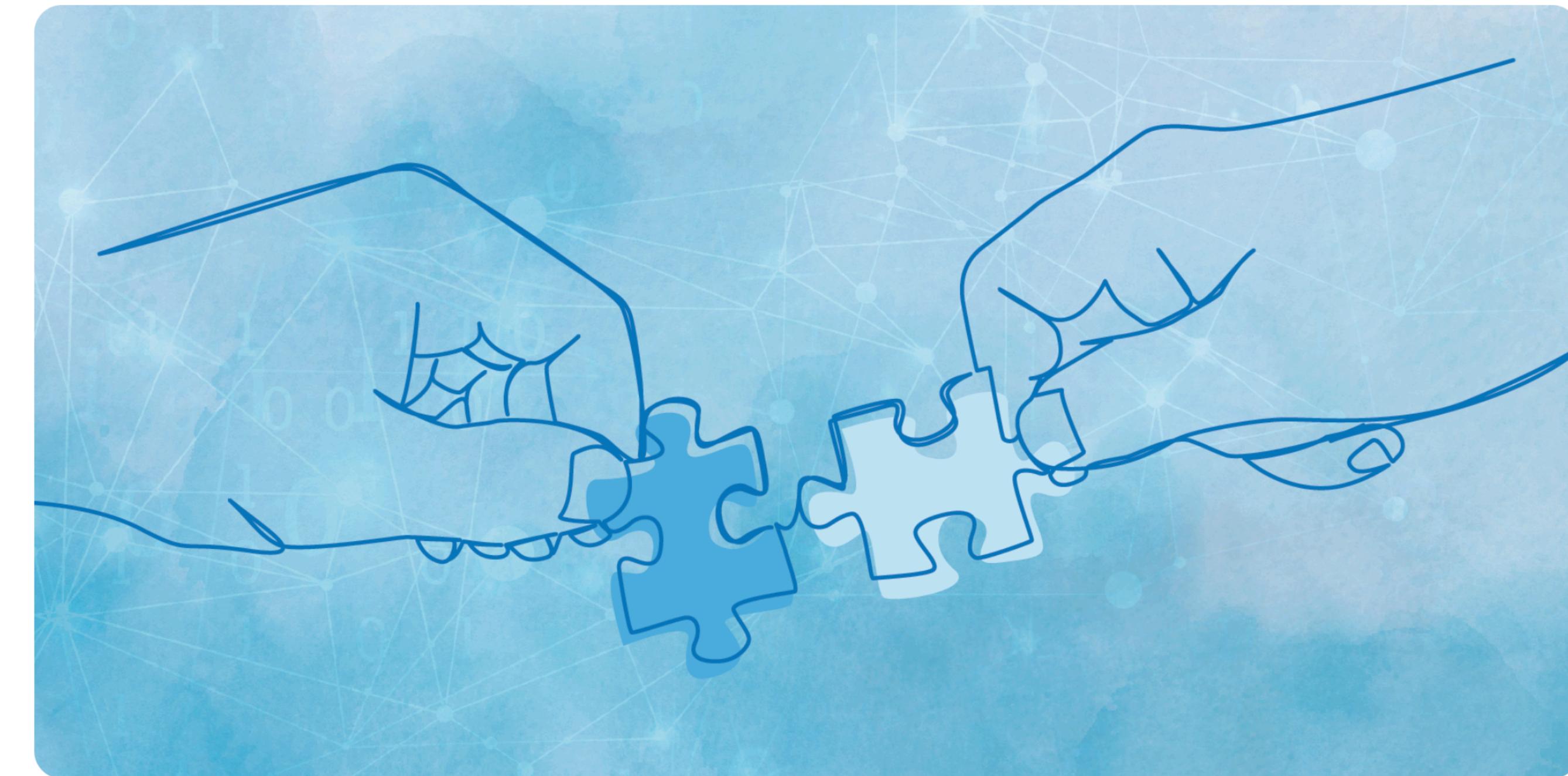
News

14 Feb 2024



NIH and GA4GH commit to ongoing collaboration

NIH and GA4GH strengthen their partnership to expand responsible data use for the benefit of human health through a Memorandum of Agreement.



The United States National Institutes of Health (NIH) Office of Data Science Strategy (ODSS) and the Global Alliance for Genomics and Health (GA4GH) have announced a strategic collaboration in the form of a Memorandum of Agreement. This partnership aims to bolster the development of technology standards, tools, and policy frameworks to support responsible sharing of genomic and related health data on a global scale.

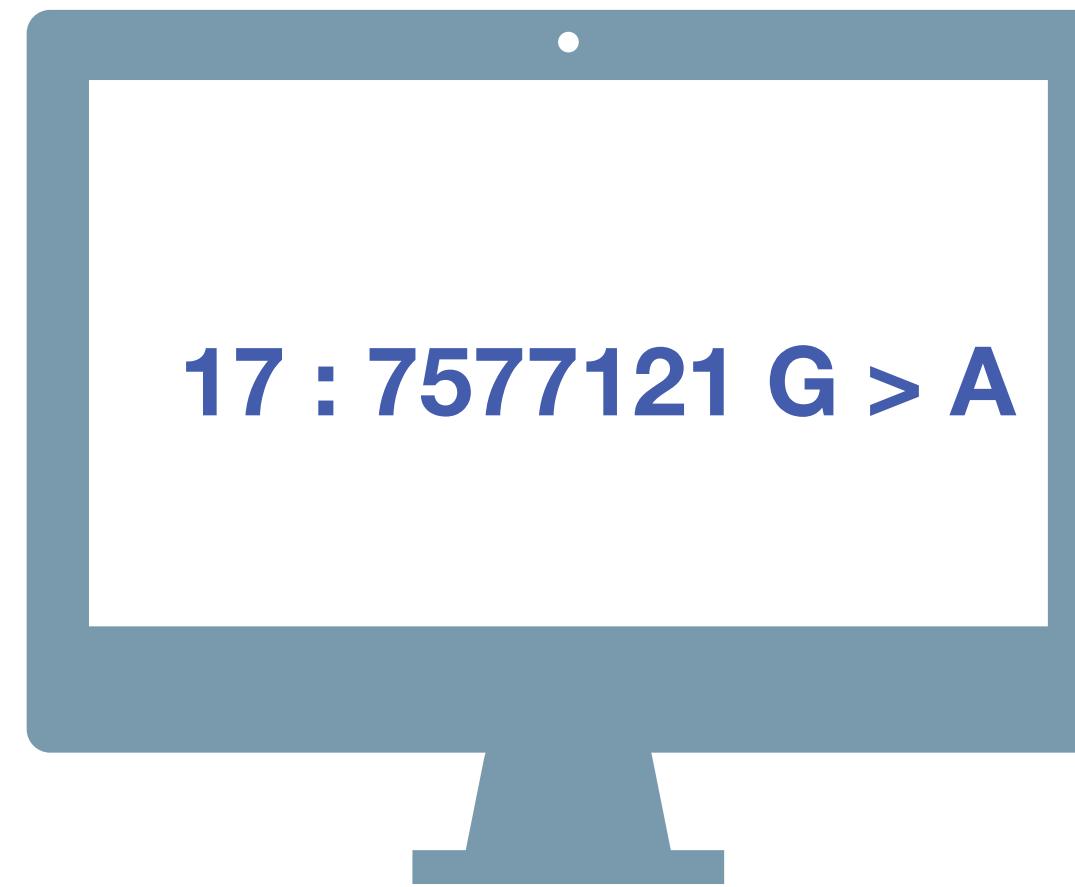


Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.



The GA4GH Beacon Protocol

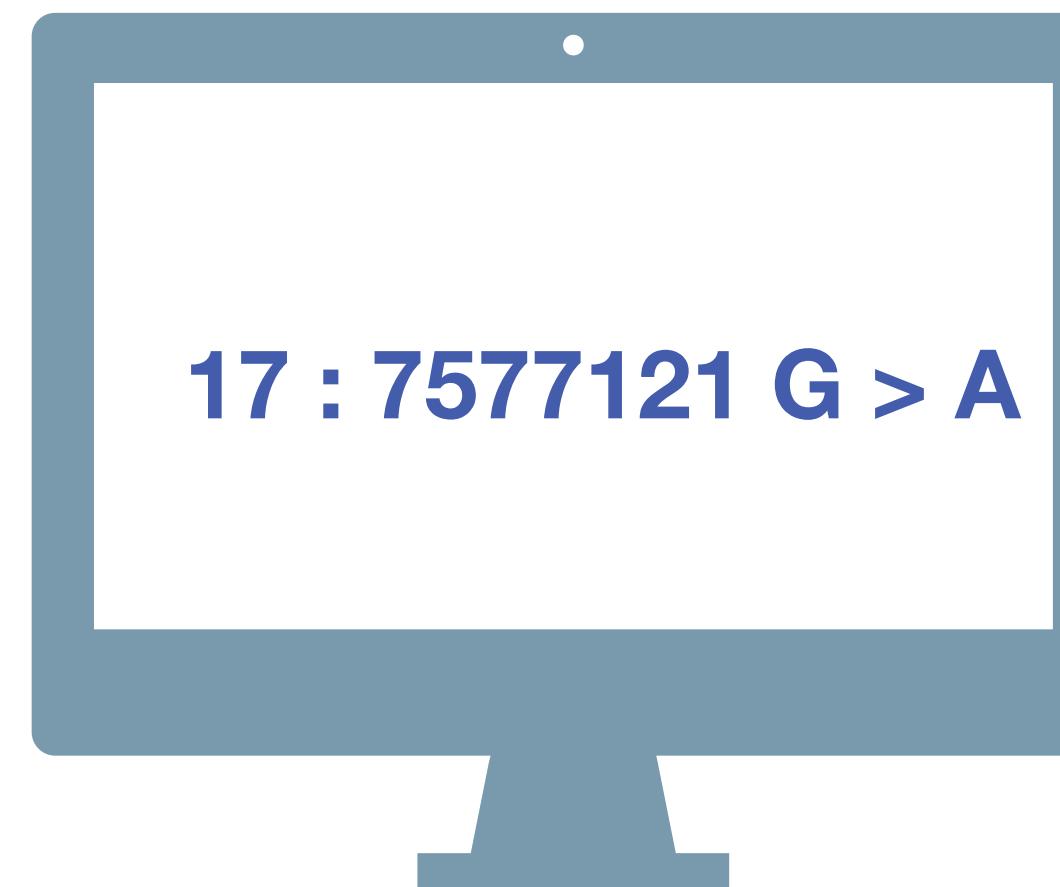
Federating Genomic Discoveries



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**.

Introduction

... I proposed a challenge application for all those wishing to seriously engage in *international* data sharing for human genomics. ...

1. Provide a public web service
2. Which accepts a query of the form “Do you have any genomes with an “A” at position 100,735 on chromosome 3?”
3. And responds with one of “Yes” or “No” ...

“Beacon” because ... people have been scanning the universe of human research for *signs of willing participants in far reaching data sharing*, but ... it has remained a dark and quiet place. The hope of this challenge is to 1) *trigger the issues* blocking groups ... in way that isn’t masked by the ... complexities of the science, fully functional interfaces, and real issues of privacy, and to 2) in *short order* ... see *real beacons of measurable signal* ... from *at least some sites* ... Once your “GABeacon” is shining, you can start to take the *next steps to add functionality* to it, and *finding the other groups* ... following their GABeacons.

Utility

Some have argued that this simple example is not “useful” so nobody would build it. Of course it is not the first priority for this application to be scientifically useful. ...intended to provide a *low bar for the first step of real ... engagement*. ... there is some utility in ...locating a rare allele in your data, ... not zero.

A number of more useful first versions have been suggested.

1. Provide *frequencies of all alleles* at that point
2. Ask for all alleles seen in a gene *region* (and more elaborate versions of this)
3. Other more complicated queries



"I would personally recommend all those be held for
version 2, when the beacon becomes a service."
Jim Ostell, 2014

Implementation

1. Specifying the chromosome ... The interface needs to specify the *accession.version* of a chromosome, or *build number*...
2. Return values ... right to *refuse* to answer without it being an error ... DOS *attack* ... or because ...especially *sensitive*...
3. Real time response ... Some sites suggest that it would be necessary to have a “*phone home*” response ...

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

Beacon v2 Development

- 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

2022

Related ...

- ELIXIR starts Beacon project support

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

- new Beacon website (March)

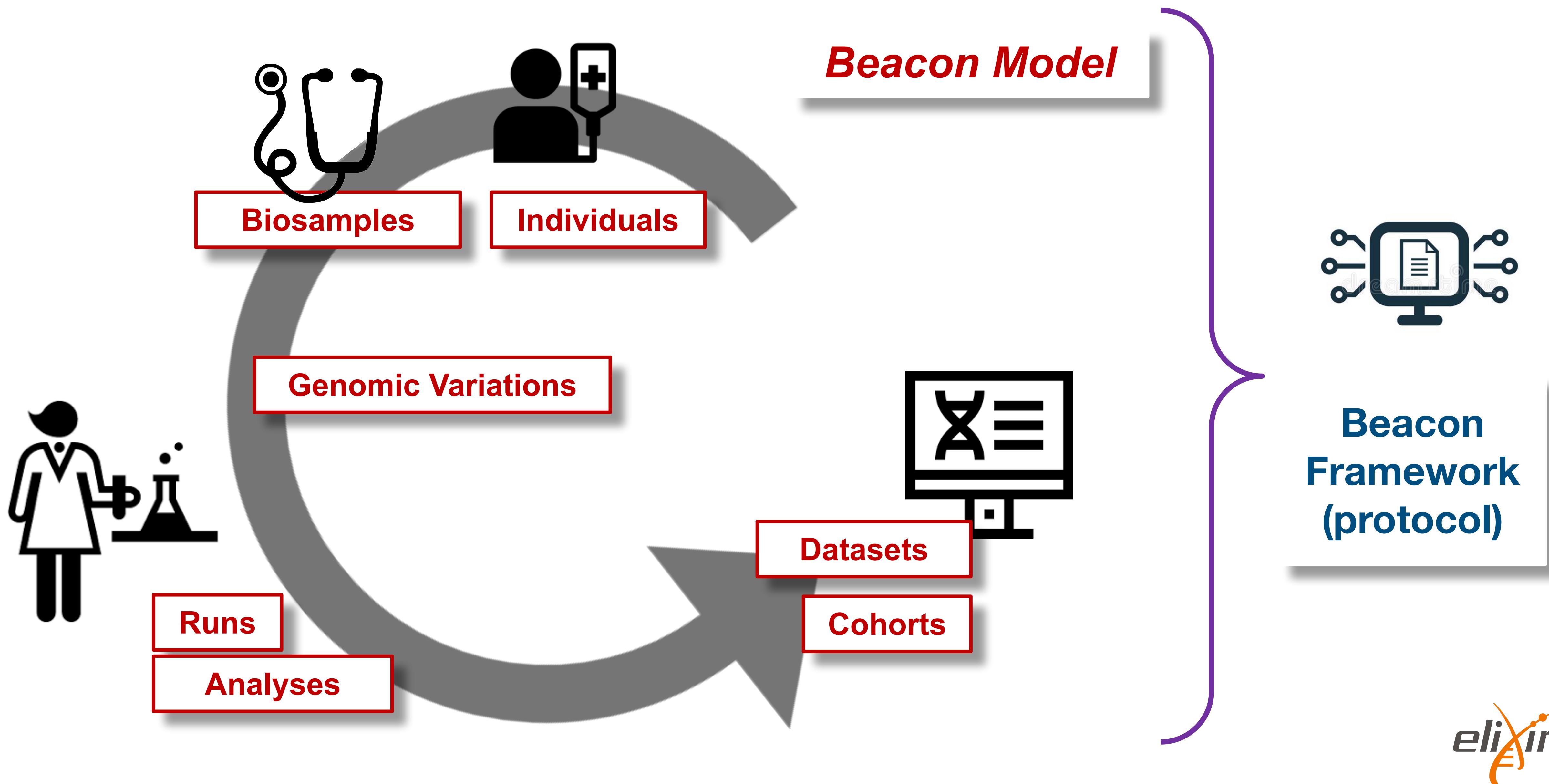
- Beacon publication at Nature Biotechnology

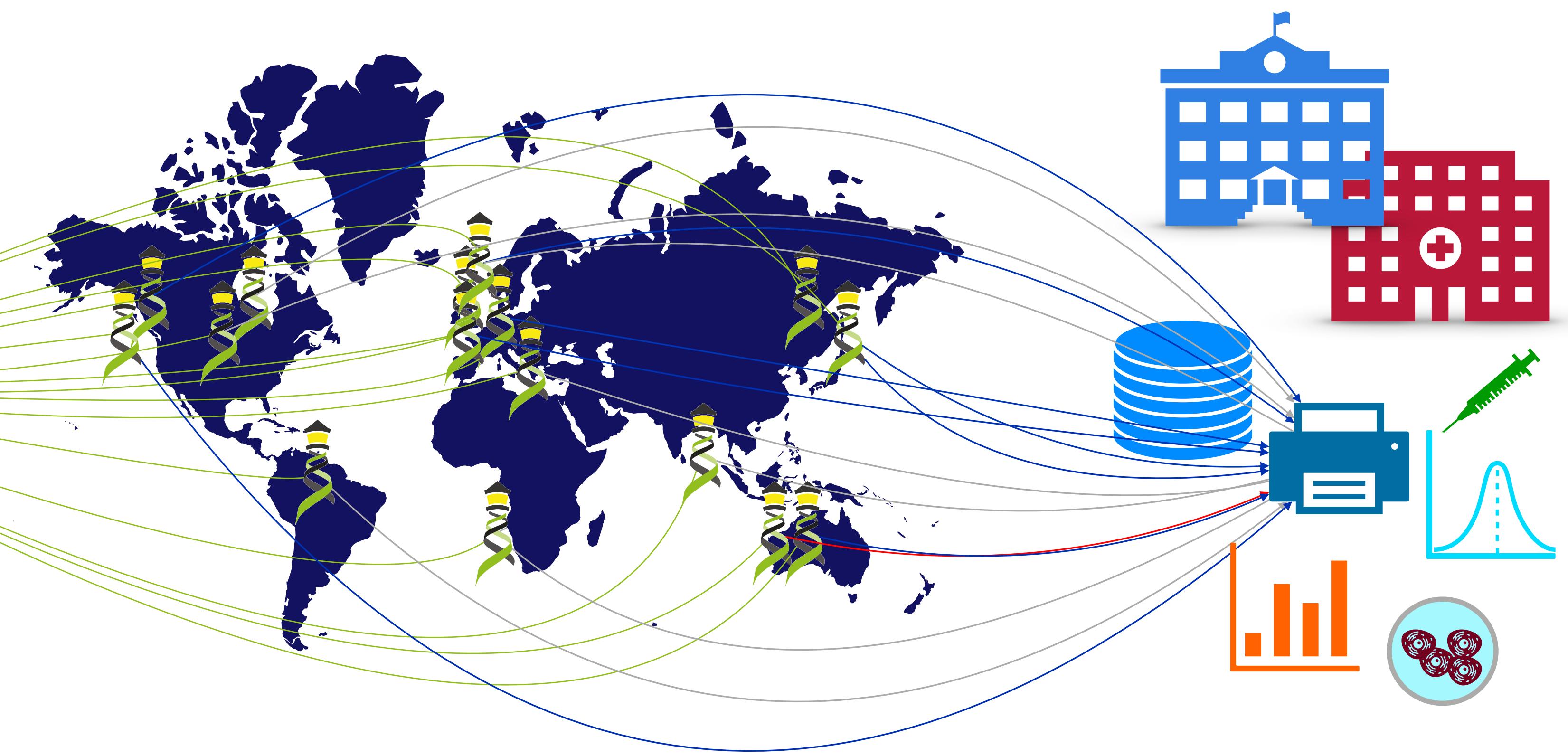
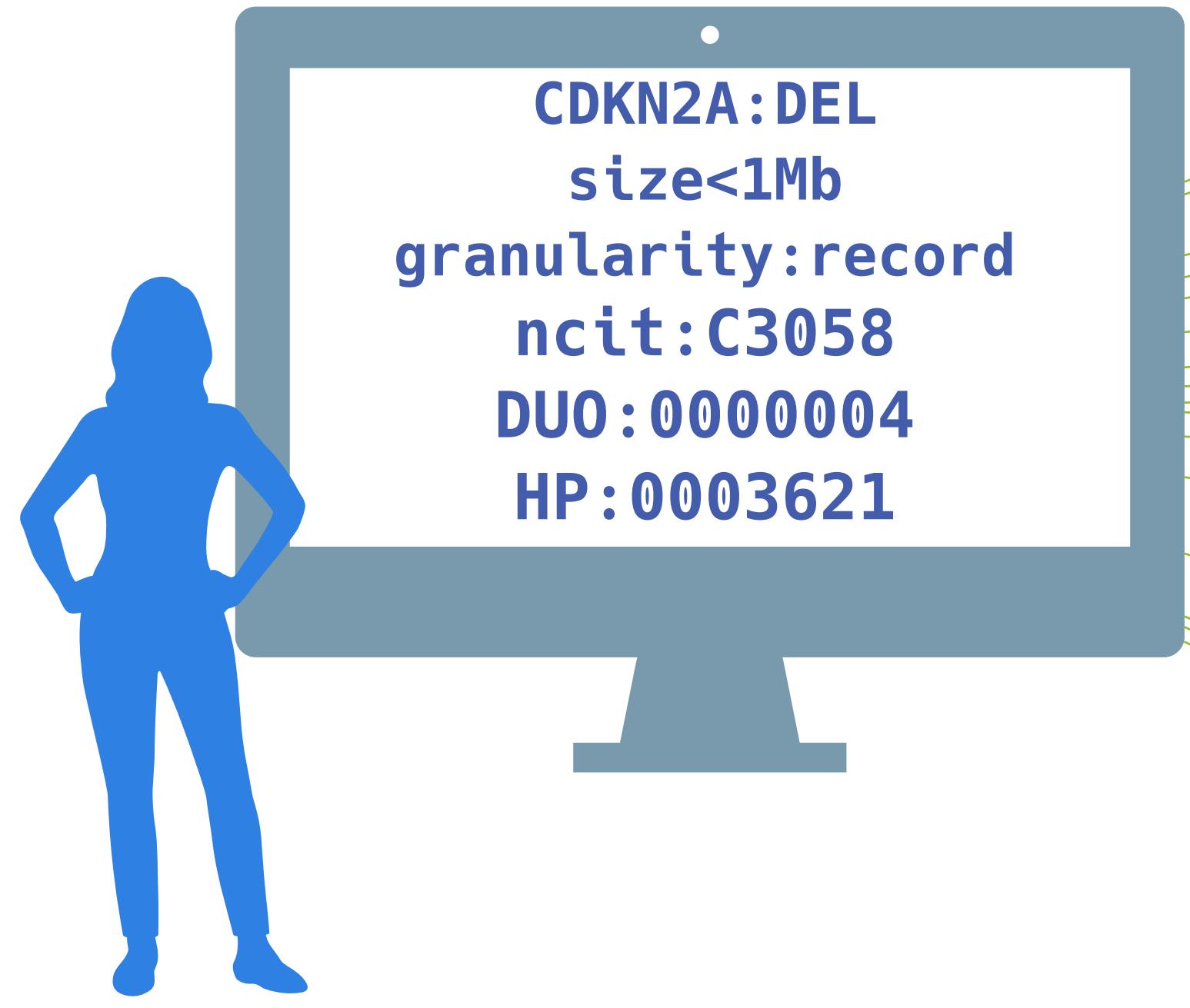
- Phenopackets v2 approved

- docs.genomebeacons.org

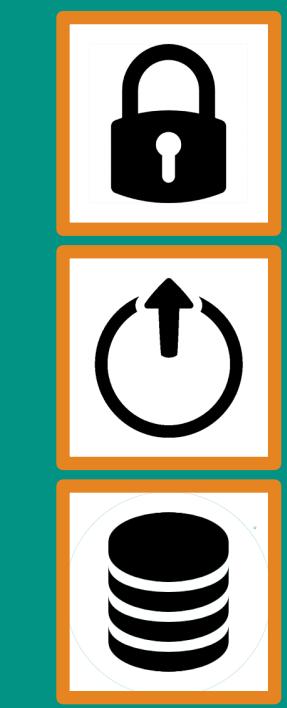
Beacon v2

docs.genomebeacons.org





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

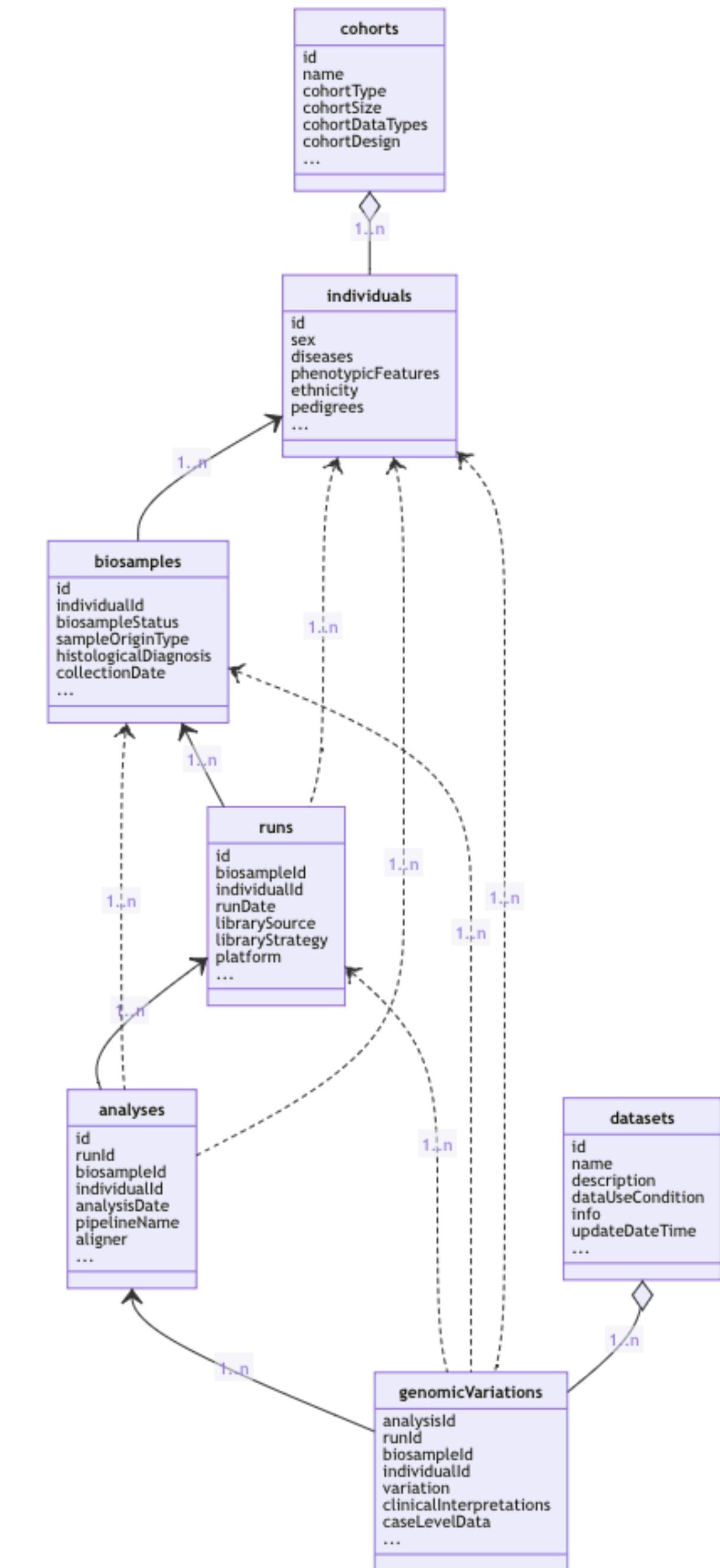


Beacon v2 API

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

Beacon Default v2 Model

- The Beacon **framework** describes the overall structure of the API requests, responses, parameters, the common components, etc.
- Beacon **models** describe the set of concepts included in a Beacon, like individual or biosample, and also the relationships between them.
- Besides logical concepts, the Beacon **models** represent the schemas for data delivery in “record” granularity
- Beacon explicitly allows the use of other models besides its *version specific default*.
- Adherence to a shared model empowers federation
- Use of the framework w/ different models extends adoption



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

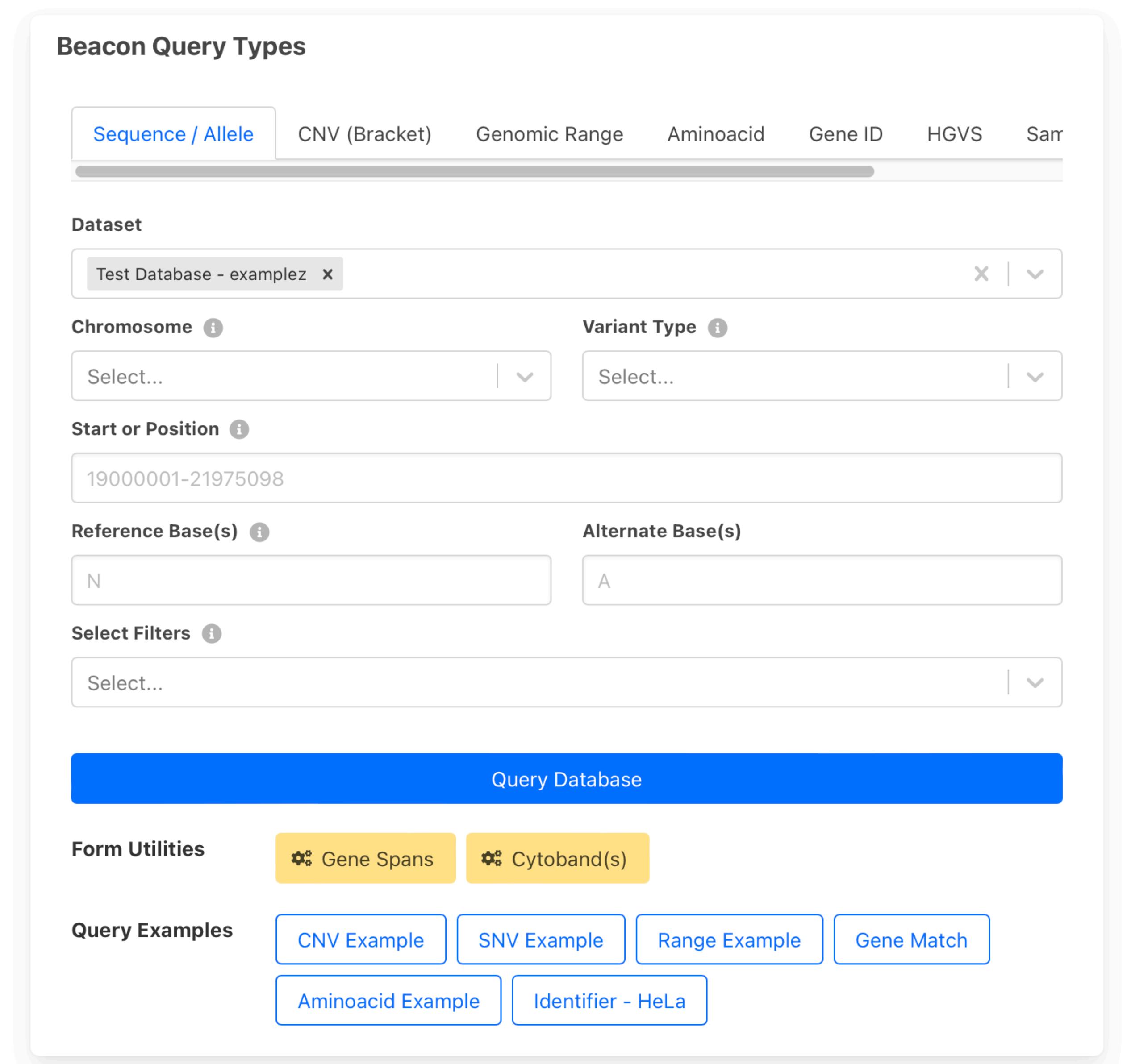
Reference Base(s) i Alternate Base(s)
N A

Select Filters i
Select...

Query Database

Form Utilities Gene Spans Cytoband(s)

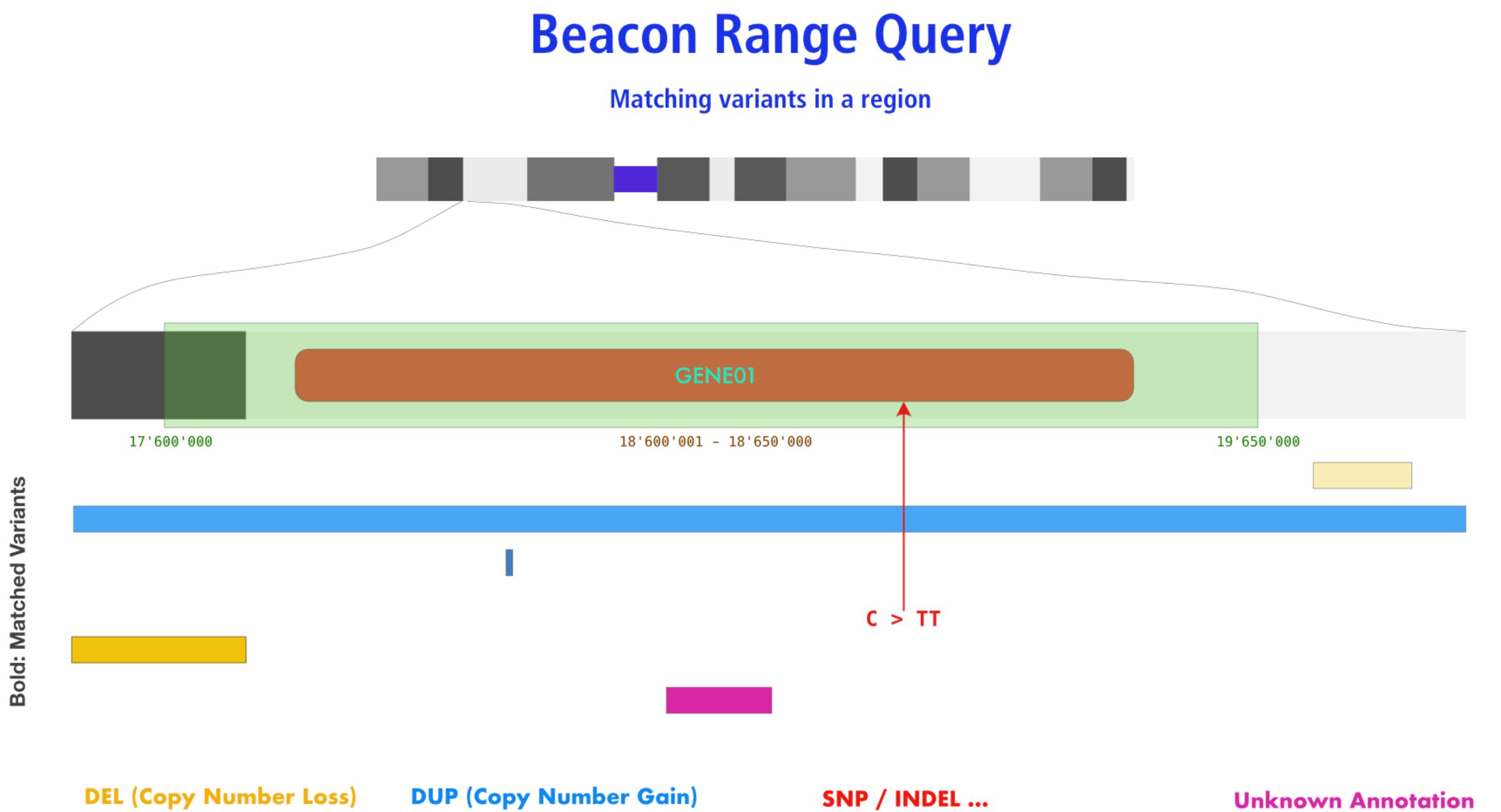
Query Examples CNV Example SNV Example Range Example Gene Match
Aminoacid Example Identifier - HeLa



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

17 (NC_000017.11)

Variant Type

SO:0001059 (any sequence alteration - S...)

Start or Position

7572826

End (Range or Structural Var.)

7579005

Reference Base(s)

N

Alternate Base(s)

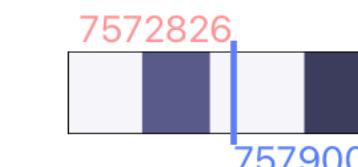
A

Select Filters

Select...

Chromosome 17

7572826
7579005



Query Database

Form Utilities

Gene Spans Cytoband(s)

Query Examples

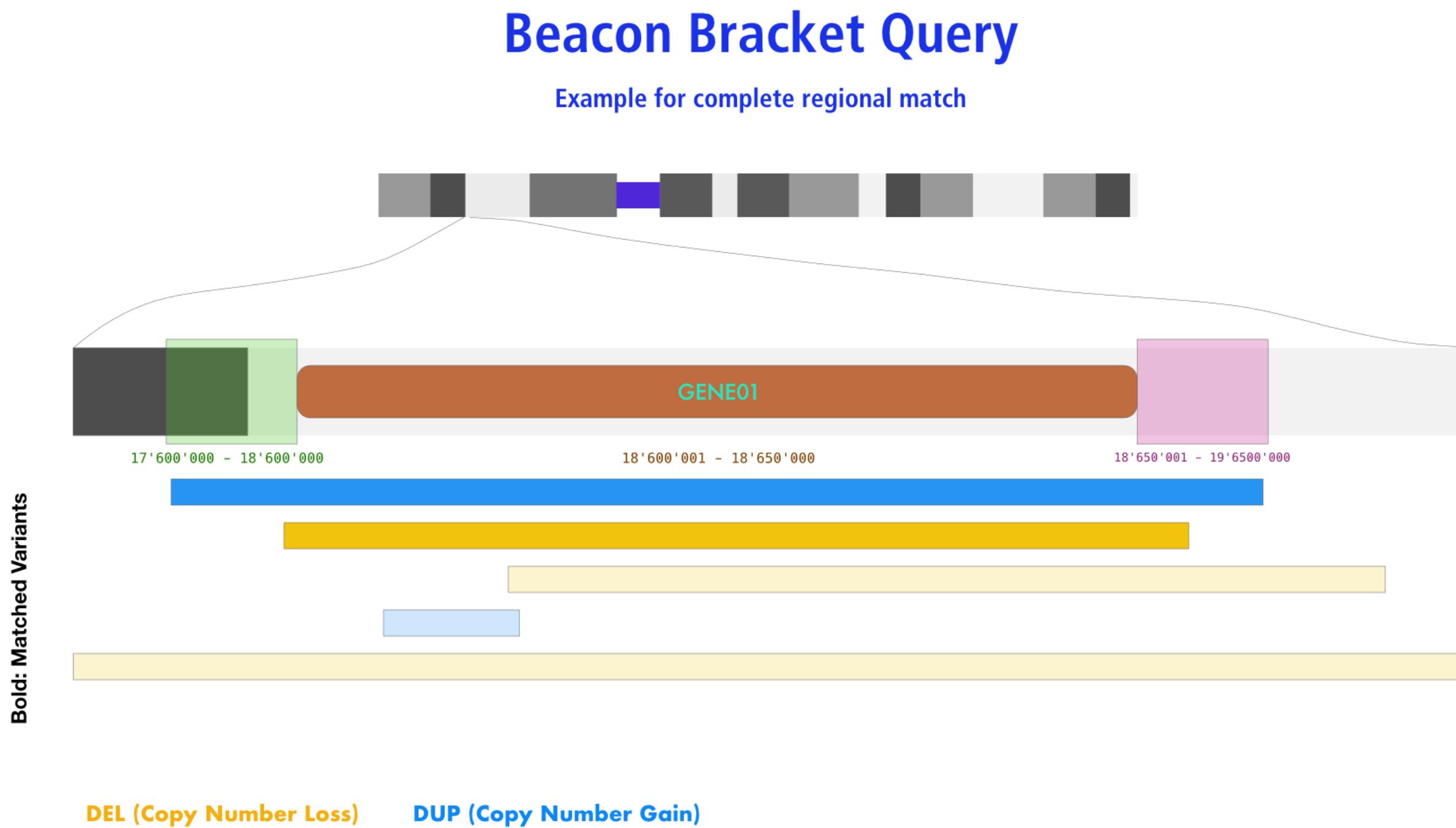
CNV Example SNV Example Range Example Gene Match
Aminoacid Example Identifier - HeLa

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

Beacon Queries

Bracket ("CNV") Query

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



Beacon Query Types

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

Dataset

Test Database - examplez x | v

Chromosome

9 (NC_000009.12) | v

Variant Type

EFO:0030067 (copy number deletion) | v

Start or Position

21000001-21975098

End (Range or Structural Var.)

21967753-23000000

Select Filters

NCIT:C3058: Glioblastoma (100) x | v

Chromosome 9

21000001-21975098



Query Database

Form Utilities

Gene Spans

Cytoband(s)

Query Examples

CNV Example

SNV Example

Range Example

Gene Match

Aminoacid Example

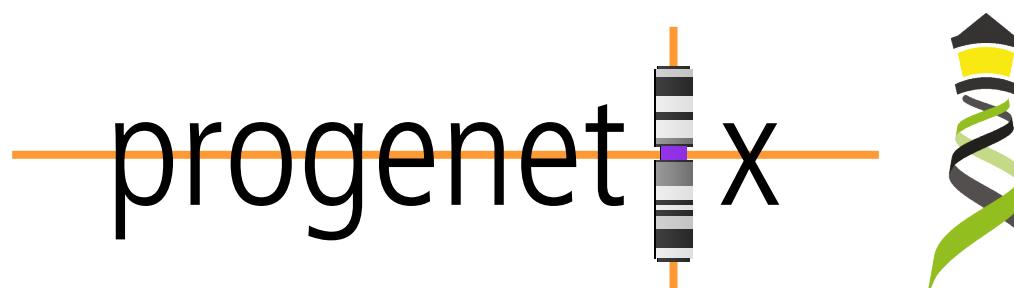
Identifier - HeLa

This example shows the query for CNV deletion variants overlapping the CDKN2A gene's coding region with at least a single base, but limited to "focal" hits (here i.e. <= ~2Mbp in size). The query is against the examplez collection and can be modified e.g. through changing the position parameters or data source.

Beacon 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](#)

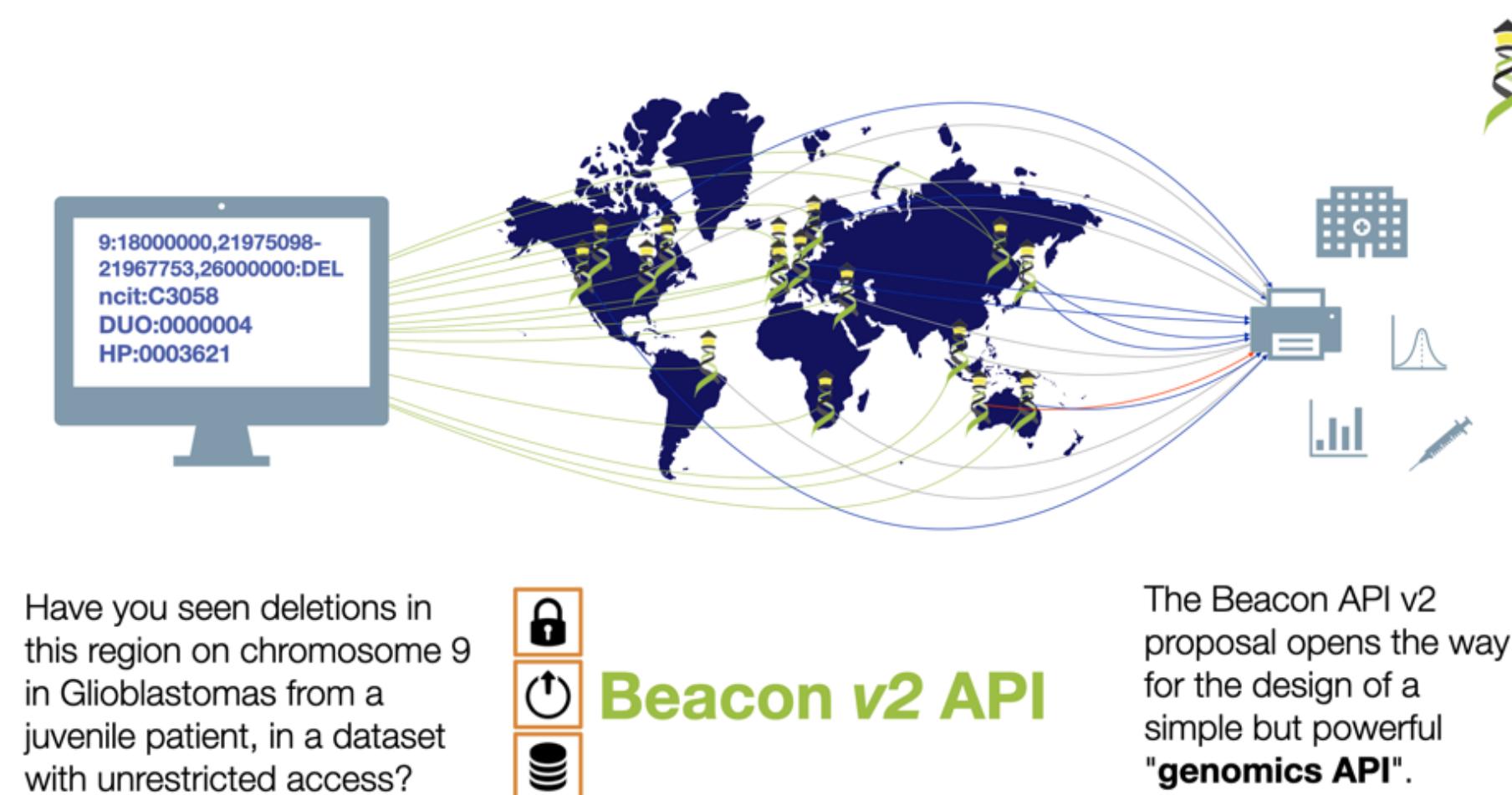
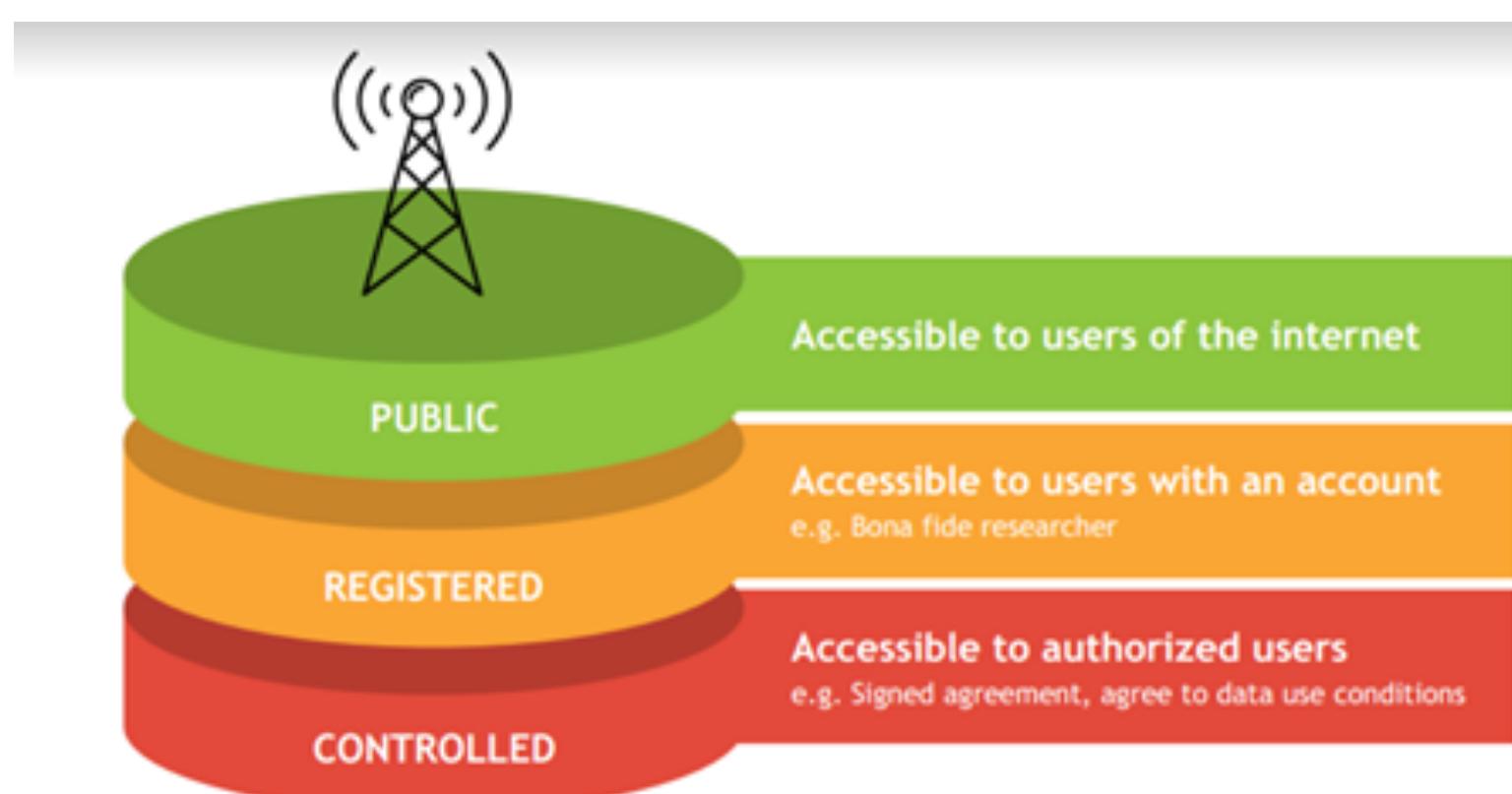
Results [Biosamples](#)

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 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



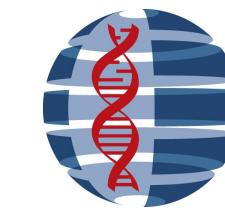
Example Users



UNIVERSITY OF CALIFORNIA
SANTA CRUZ

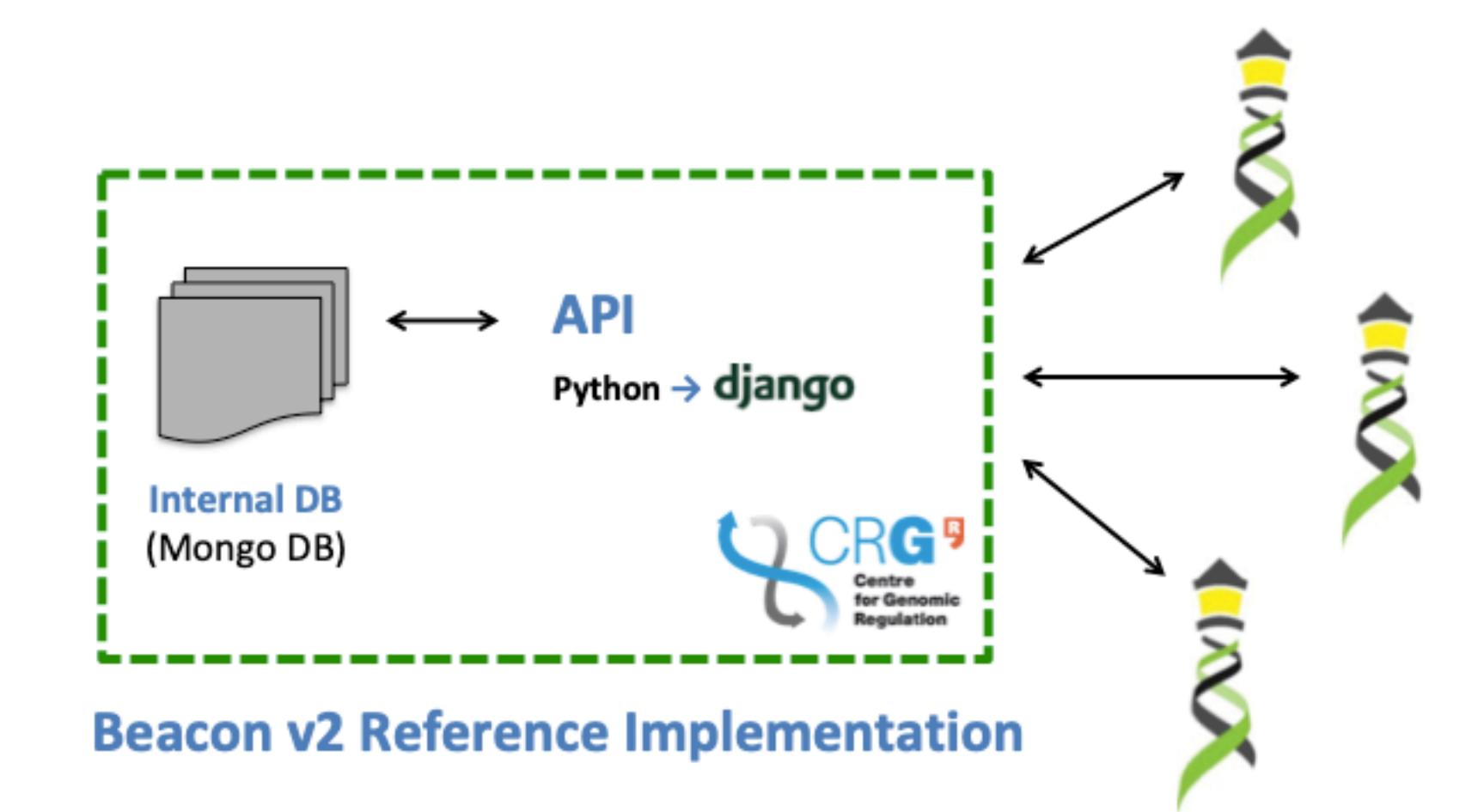
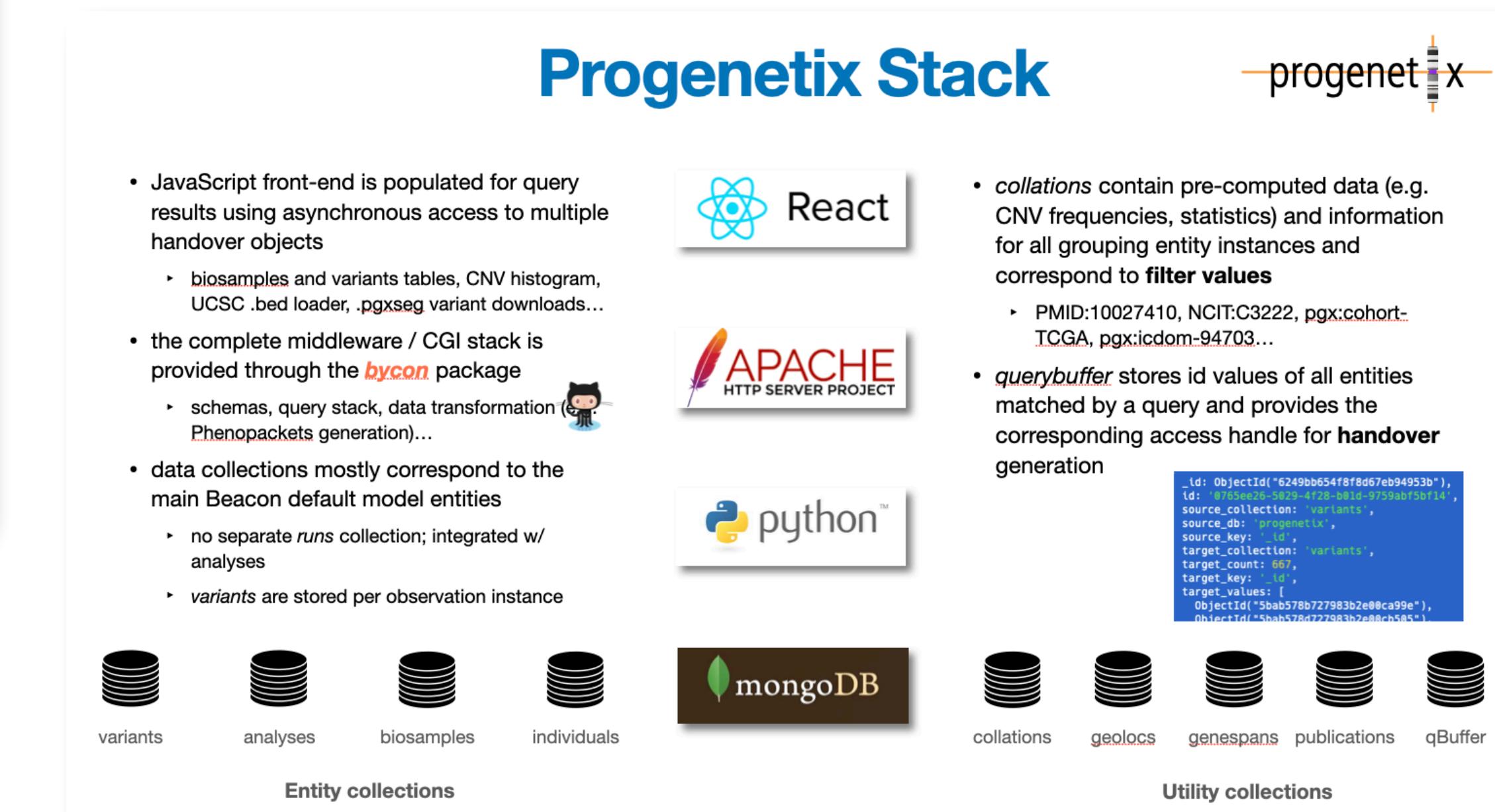
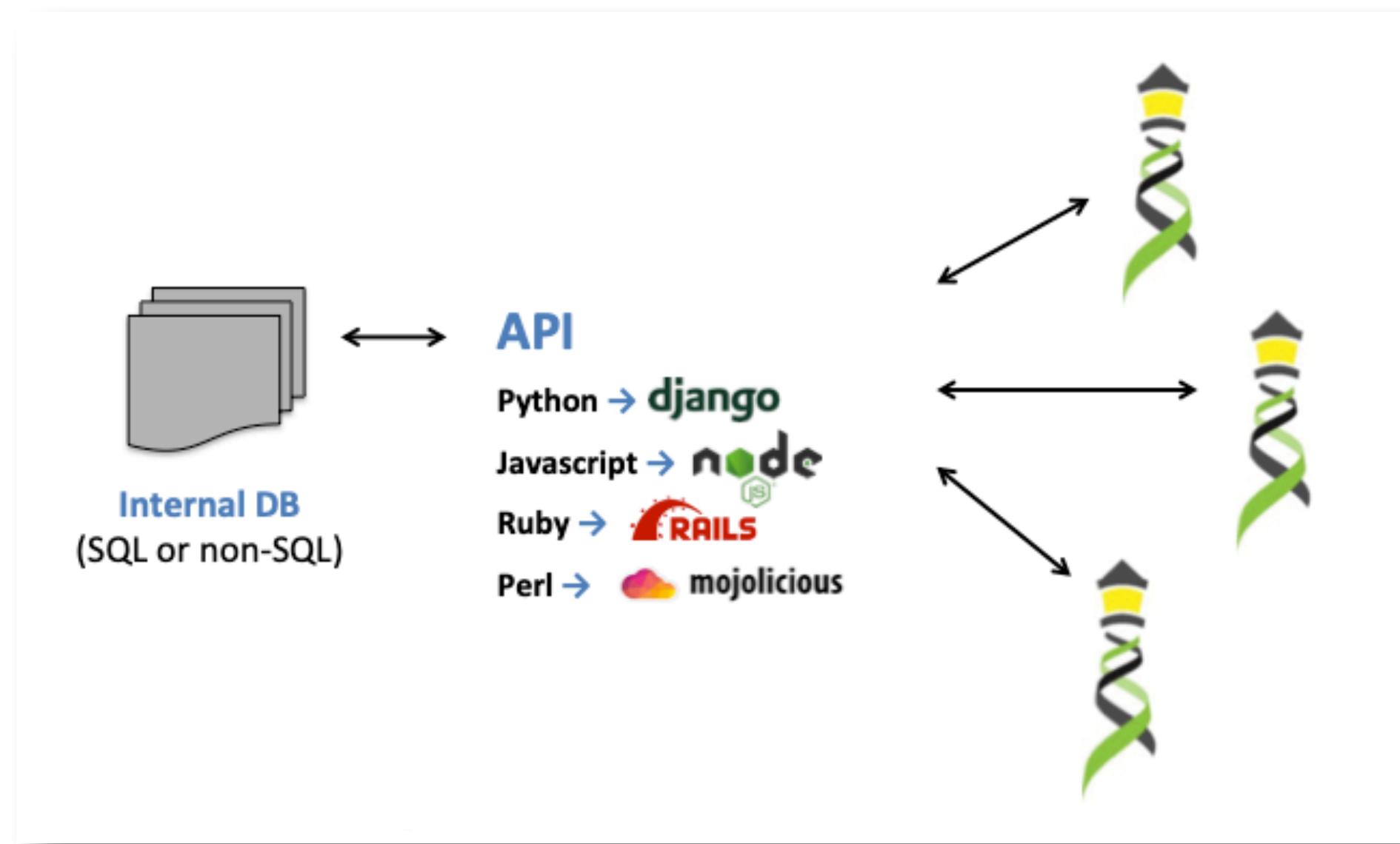


SciLifeLab



Implementing Beacon v2

... its just code _(_ツ)_/





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

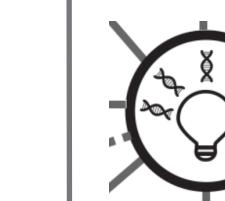
Gema Roldan



Thomas Keane

Melanie Courtot

Jonathan Dursi



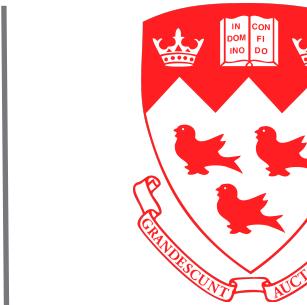
Heidi Rehm

Ben Hutton



Toshiaki Katayama

GEM Japan

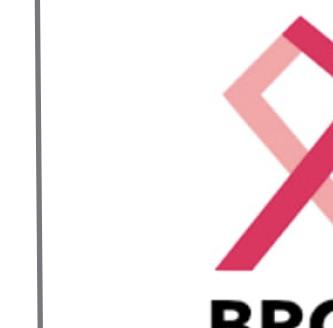


Stephane Dyke



Marc Fiume

Miro Cupak



Melissa Cline



Diana Lemos



GA4GH Phenopackets

Peter Robinson
Jules Jacobsen



Alex Wagner
Reece Hart

Beacon PRC

Alex Wagner
Jonathan Dursi
Mamana Mbiyavanga
Alice Mann
Neerjah Skantharajah



The Beacon team through the ages

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

Beacon protocol response verifier at time of GA4GH approval Spring 2022

Beacon v2 GA4GH Approval Registry

Beacons: European Genome-Phenome Archive | progenetix | cnag | University of Leicester

Category	EGA	progenetix	Theoretical Cytogenetics and Oncogenomics group at UZH and SIB
BeaconMap	Green	Green	Green
Bioinformatics analysis	Green	Green	Green
Biological Sample	Green	Green	Green
Cohort	Green	Green	Green
Configuration	Green	Green	Green
Dataset	Green	Green	Green
EntryTypes	Green	Green	Green
Genomic Variants	Green	Green	Green
Individual	Green	Green	Green
Info	Green	Green	Green
Sequencing run	Green	Green	Green

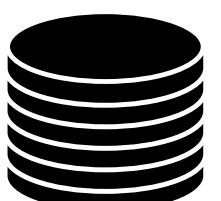
Category	cnag	Centre Nacional Analisis Genomica (CNAG-CRG)	University of Leicester
BeaconMap	Green	Green	Green
Bioinformatics analysis	White	White	White
Biological Sample	Red	Red	White
Cohort	White	White	White
Configuration	Green	Green	White
Dataset	Red	White	White
EntryTypes	Green	Green	White
Genomic Variants	White	White	White
Individual	Red	Red	White
Info	White	White	White
Sequencing run	White	White	White

Green: Matches the Spec | Red: Not Match the Spec | White: Not Implemented

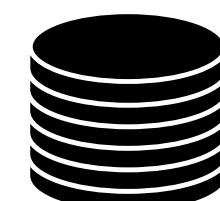
Progenetix Stack



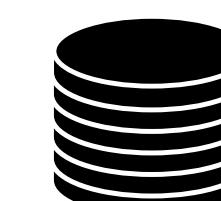
- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
 - biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads...
- the complete middleware / CGI stack is provided through the **bycon** package
 - schemas, query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



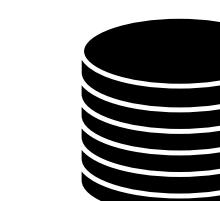
variants



analyses



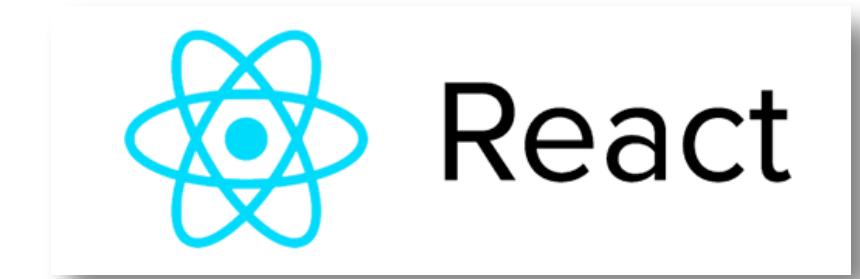
biosamples



individuals

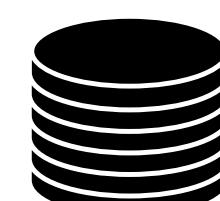


Entity collections

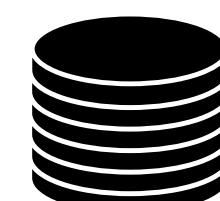


- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
 - PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703...
- *querybuffer* stores id values of all entities matched by a query and provides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

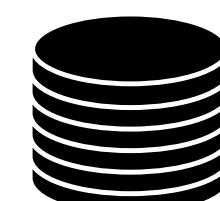
```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
  ObjectId("5bab578d727983b2e00cb505")]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



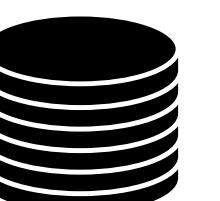
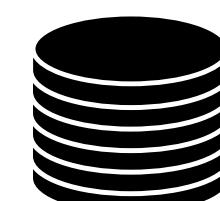
collations



geolocs



genespans publications



Utility collections

Beacon v2 Conformity and Extensions in *bycon*

Putting the ⁺ into Beacon ...

- support & use of standard Beacon v2 PUT & GET variant queries, filters and meta parameters
 - variant parameters, genelid, lengths, EFO & VCF CNV types, pagination
 - widespread, self-scoping filter use for bio-, technical- and id parameters with switch for descending terms use (globally or per term if using POST)
- **extensive use of handovers**
 - asynchronous delivery of e.g. variant and sample data, data plots
- ⁺ optional use of OR logic for filter combinations (global)
- ⁺ extension of query parameters
 - **geographic queries** incl. \$geonear and use of GeoJSON in schemas
- ↗ ↘ ↛ ↚ no implementation of authentication on this open dataset

bycon provides a number of additional services and output formats which are initiated over the /services path or provided as request parameters and are not considered Beacon extensions (though they follow the syntax where possible).



Beacon⁺: Phenopackets

Testing alternative response schemas...

<http://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/phenopackets>

- the v2 default schemas are mostly aligned w/ Phenopackets v2
- creating phenopackets can be done mostly by re-wrapping of Beacon entities (individual, biosample)
- variants can be included through file resource URLs; in Beacon⁺ this is done through *ad hoc* handover URIs

```

    "id": "pgpxpf-kftx3tl5",
    "metaData": {
      "phenopacketSchemaVersion": "v2",
      "resources": [
        {
          "id": "NCIT",
          "iriPrefix": "http://purl.obolibrary.org/obo/NCIT_",
          "name": "NCIt Plus Neoplasm Core",
          "namespacePrefix": "NCIT",
          "url": "http://purl.obolibrary.org/obo/ncit/neoplasm-core.c",
          "version": "2022-04-01"
        },
        {
          "subject": {
            "dataUseConditions": {
              "id": "DUO:0000004",
              "label": "no restriction"
            },
            "diseases": [
              {
                "clinicalTnmFinding": [],
                "diseaseCode": {
                  "id": "NCIT:C3099",
                  "label": "Hepatocellular Carcinoma"
                },
                "onset": {
                  "age": "P48Y9M26D"
                },
                "stage": {
                  "id": "NCIT:C27966",
                  "label": "Stage I"
                }
              }
            ],
            "id": "pgxind-kftx3tl5",
            "sex": {
              "id": "PATO:0020001",
              "label": "male genotypic sex"
            },
            "updated": "2018-12-04 14:53:11.674000",
            "vitalStatus": {
              "status": "UNKNOWN_STATUS"
            }
          }
        }
      ],
      "biosamples": [
        {
          "biosampleStatus": {
            "id": "EFO:0009656",
            "label": "neoplastic sample"
          },
          "dataUseConditions": {
            "id": "DUO:0000004",
            "label": "no restriction"
          },
          "description": "Primary Tumor",
          "externalReferences": [
            {
              "id": "pgx:TCGA-0004d251-3f70-4395-b175-c94c2f5b1b81",
              "label": "TCGA case_id"
            },
            {
              "id": "pgx:TCGA-TCGA-DD-AAVP",
              "label": "TCGA submitter_id"
            },
            {
              "id": "pgx:TCGA-9259e9ee-7279-4b62-8512-509cb705029c",
              "label": "TCGA sample_id"
            },
            {
              "id": "pgx:TCGA-LIHC",
              "label": "TCGA LIHC project"
            }
          ],
          "files": [
            {
              "fileAttributes": {
                "fileFormat": "pgxseg",
                "genomeAssembly": "GRCh38"
              },
              "uri": "https://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/variants/?output=pgxseg"
            }
          ],
          "histologicalDiagnosis": {
            "id": "NCIT:C3099",
            "label": "Hepatocellular Carcinoma"
          },
          "id": "pgxbs-kftvhyvb",
          "individualId": "pgxind-kftx3tl5",
          "pathologicalStage": {
            "id": "NCIT:C27966",
            "label": "Stage I"
          },
          "sampledTissue": {
            "id": "UBERON:0002107",
            "label": "liver"
          },
          "timeOfCollection": {
            "age": "P48Y9M26D"
          }
        }
      ]
    }
  
```

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] 

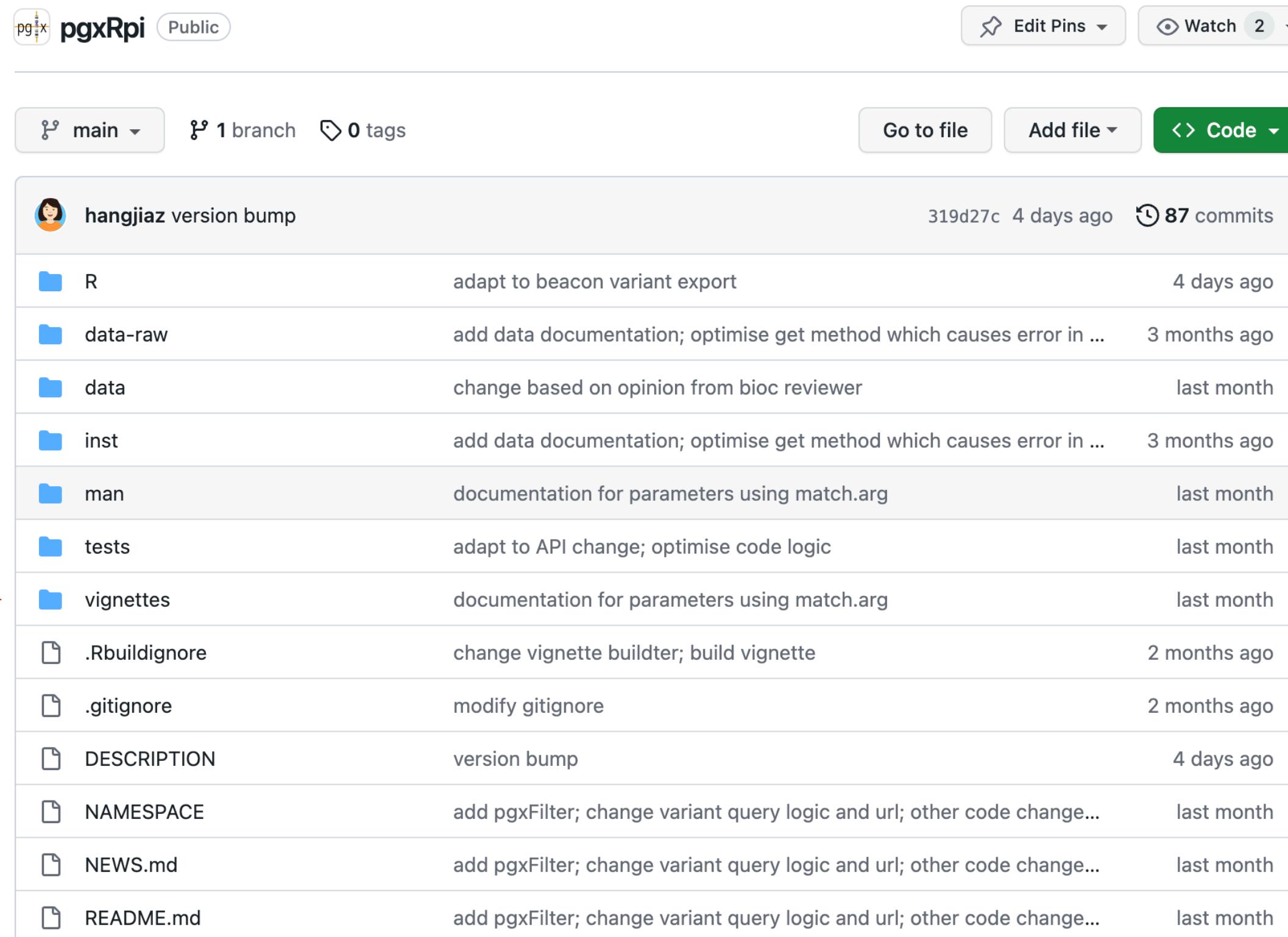
Maintainer: Hangjia Zhao <hangjia.zhao@uzh.ch>

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'. At the top, there's a navigation bar with 'Edit Pins', 'Watch 2', and a 'Code' button. Below the navigation, it shows 'main' branch, 1 branch, 0 tags, 'Go to file', 'Add file', and another 'Code' button. The main content is a list of commits:

- hangjiaz version bump (319d27c, 4 days ago, 87 commits)
 - 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 pgxFILTER; change variant query logic and url; other code change... (last month)
 - NEWS.md: add pgxFILTER; change variant query logic and url; other code change... (last month)
 - README.md: add pgxFILTER; 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 `pgxFILTER` function helps access available filters used in Progenetix. Here is the example use:

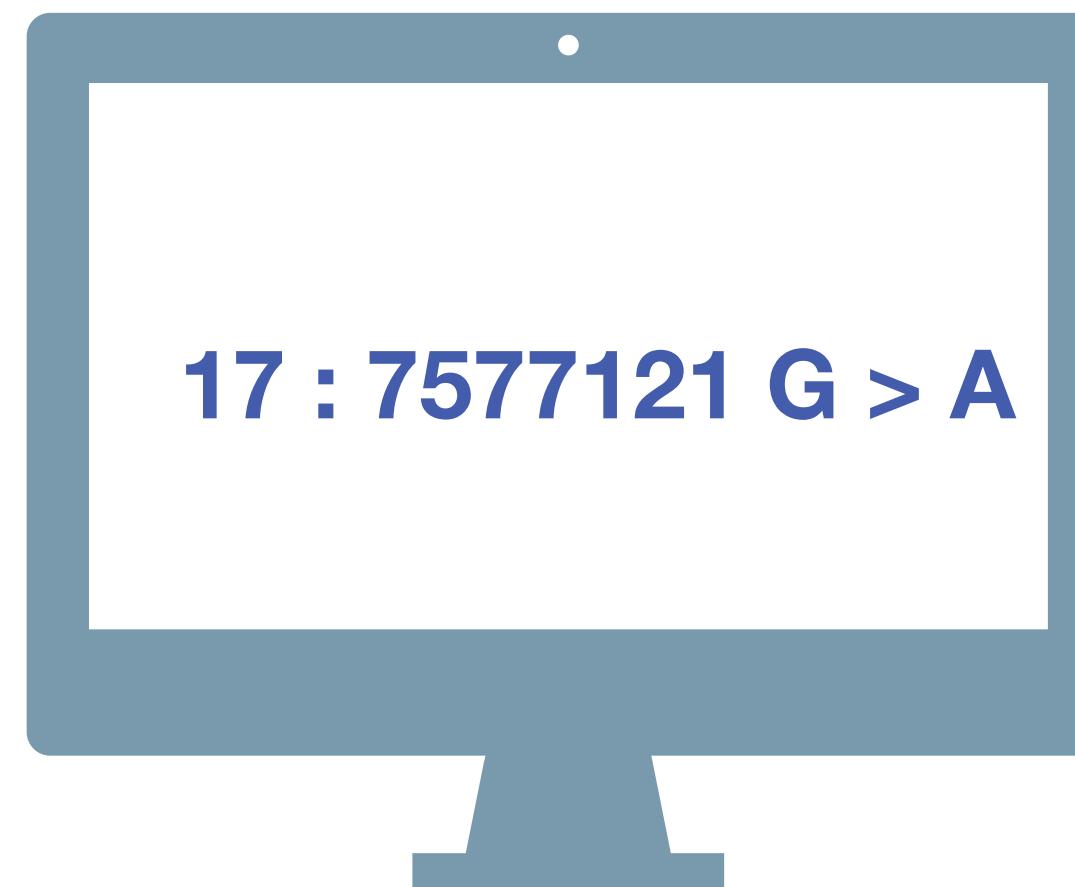
```
# access all filters
all_filters <- pgxFILTER()
# get all prefix
all_prefix <- pgxFILTER(return_all_prefix = TRUE)
# access specific filters based on prefix
ncit_filters <- pgxFILTER(prefix="NCIT")
head(ncit_filters)
#> [1] "NCIT:C28076" "NCIT:C18000" "NCIT:C14158" "NCIT:C14161" "NCIT:C28077"
#>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

Genomic Data & Privacy

How about Beacon?



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

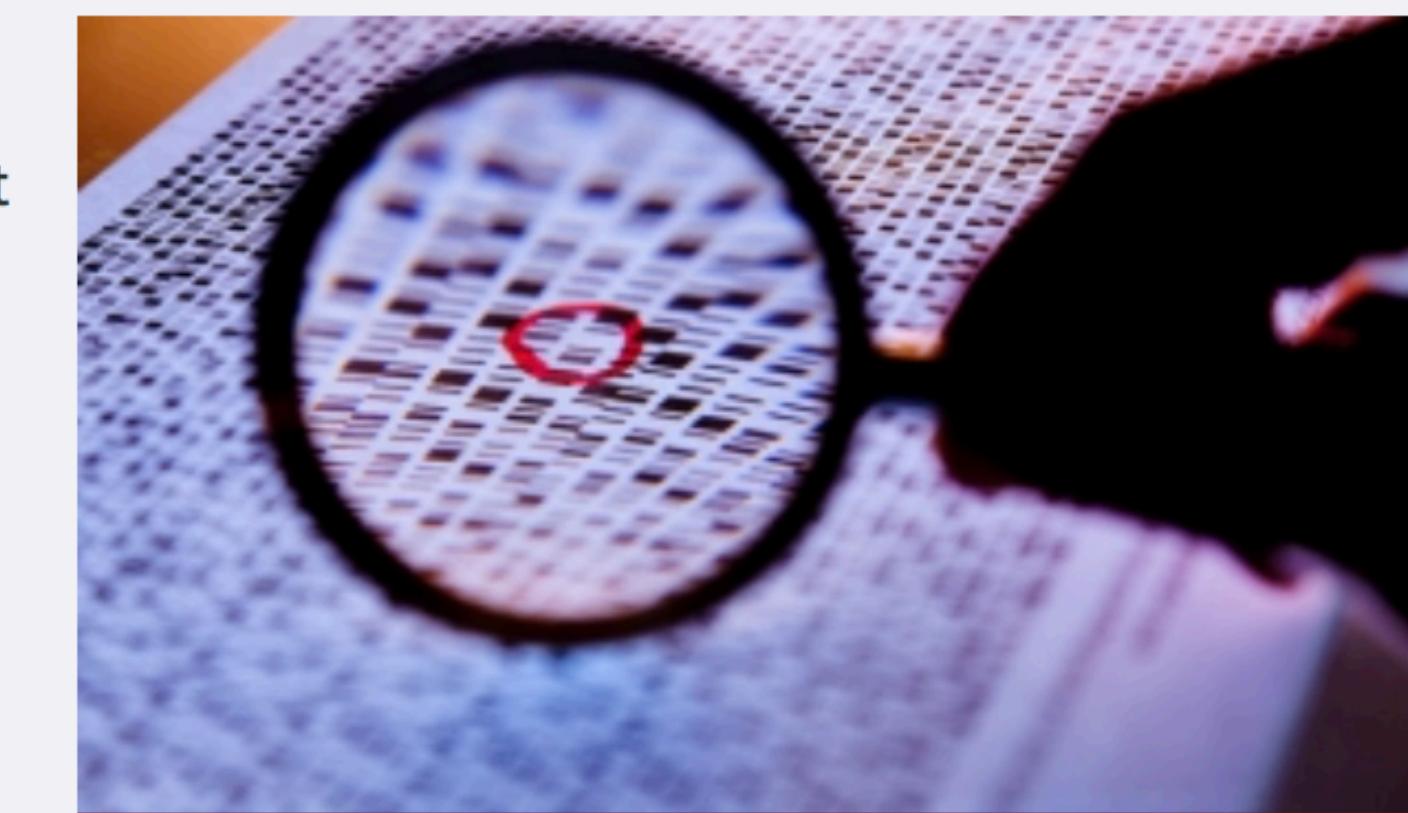
Hackers with access to a person's genome might find out if that genome is in an international network of disease databases.

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.

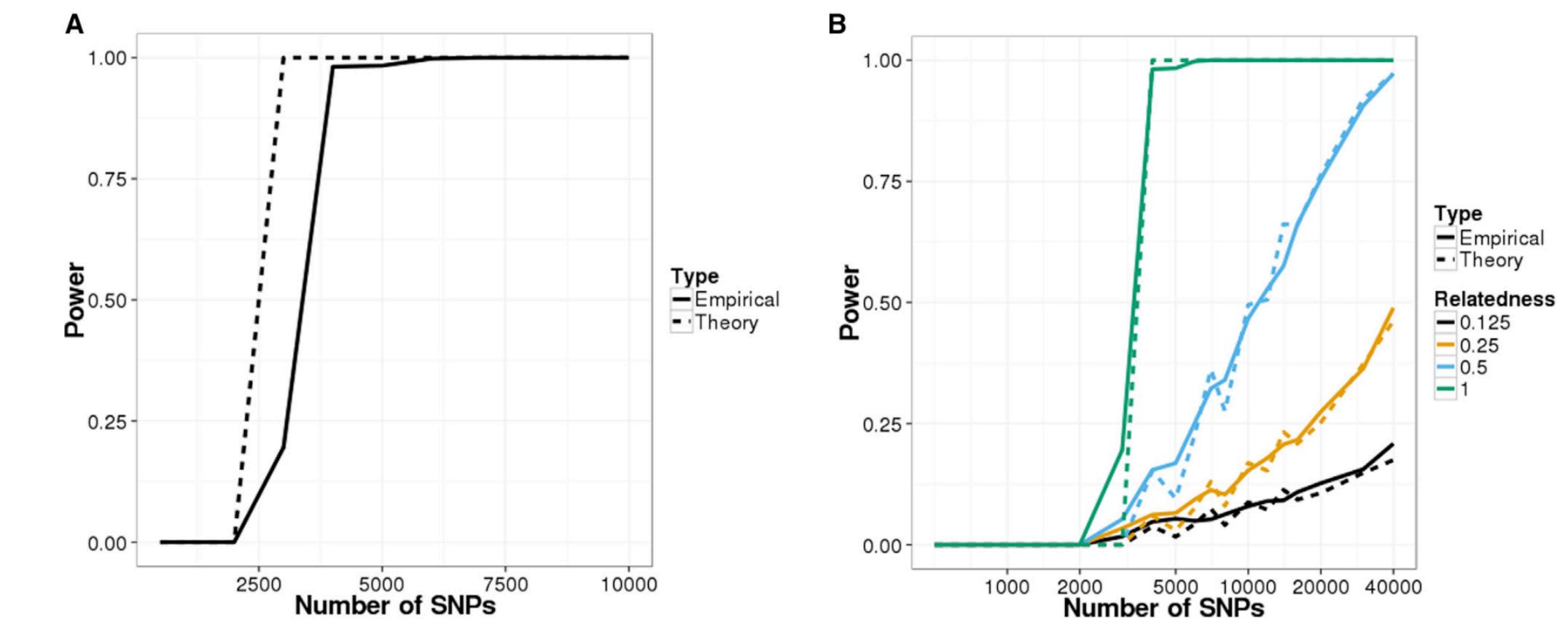


Figure 1. Power of Re-identification Attacks on Beacons Constructed with Simulated Data
Power curves for the likelihood-ratio test (LRT) on (A) a simulated beacon with 1,000 individuals and (B) detecting relatives in the simulated beacon. The false-positive rate was set to 0.05 for all scenarios.

- ▶ 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



Making Beacons Biomedical - Beacon v2

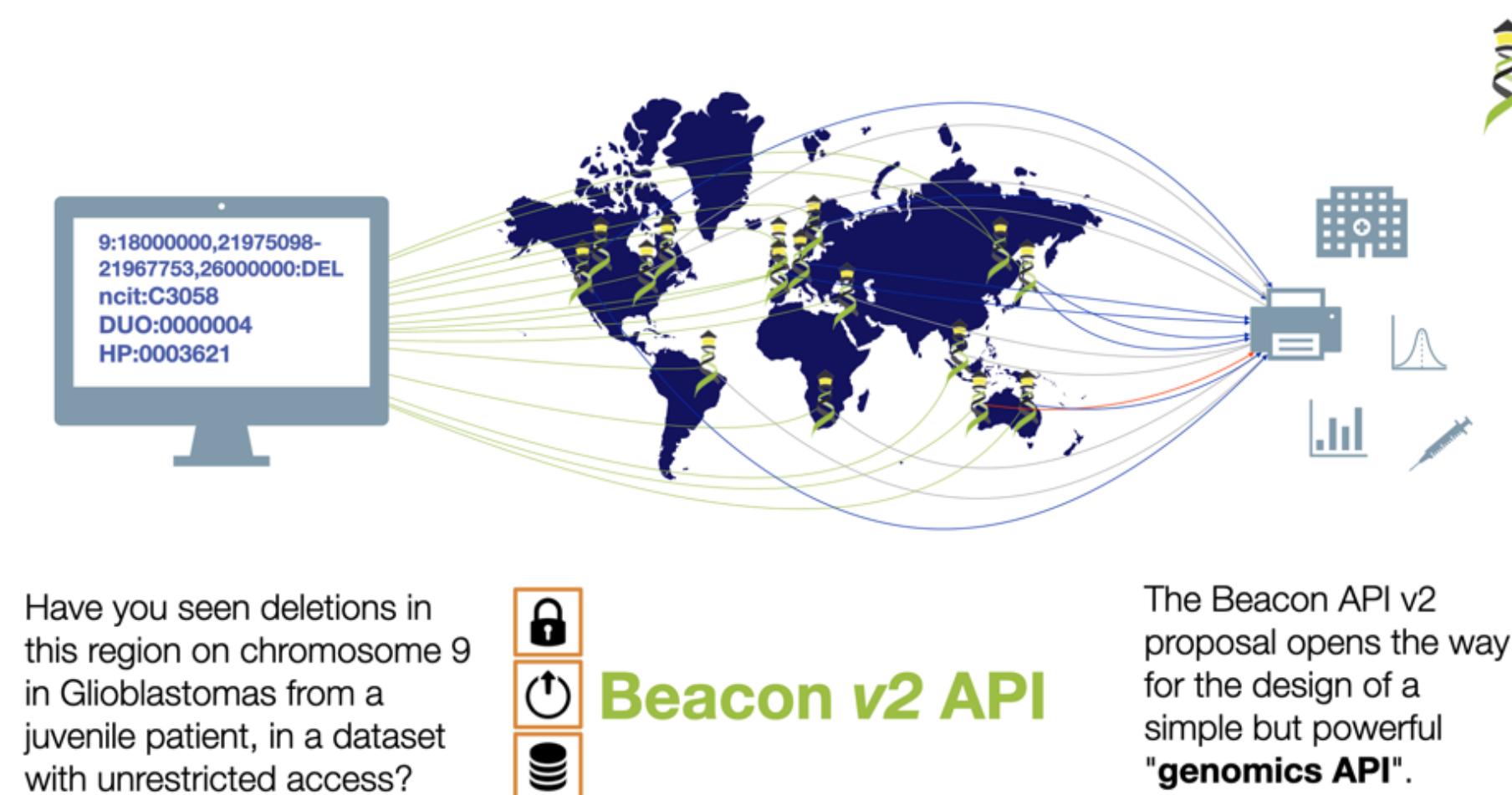
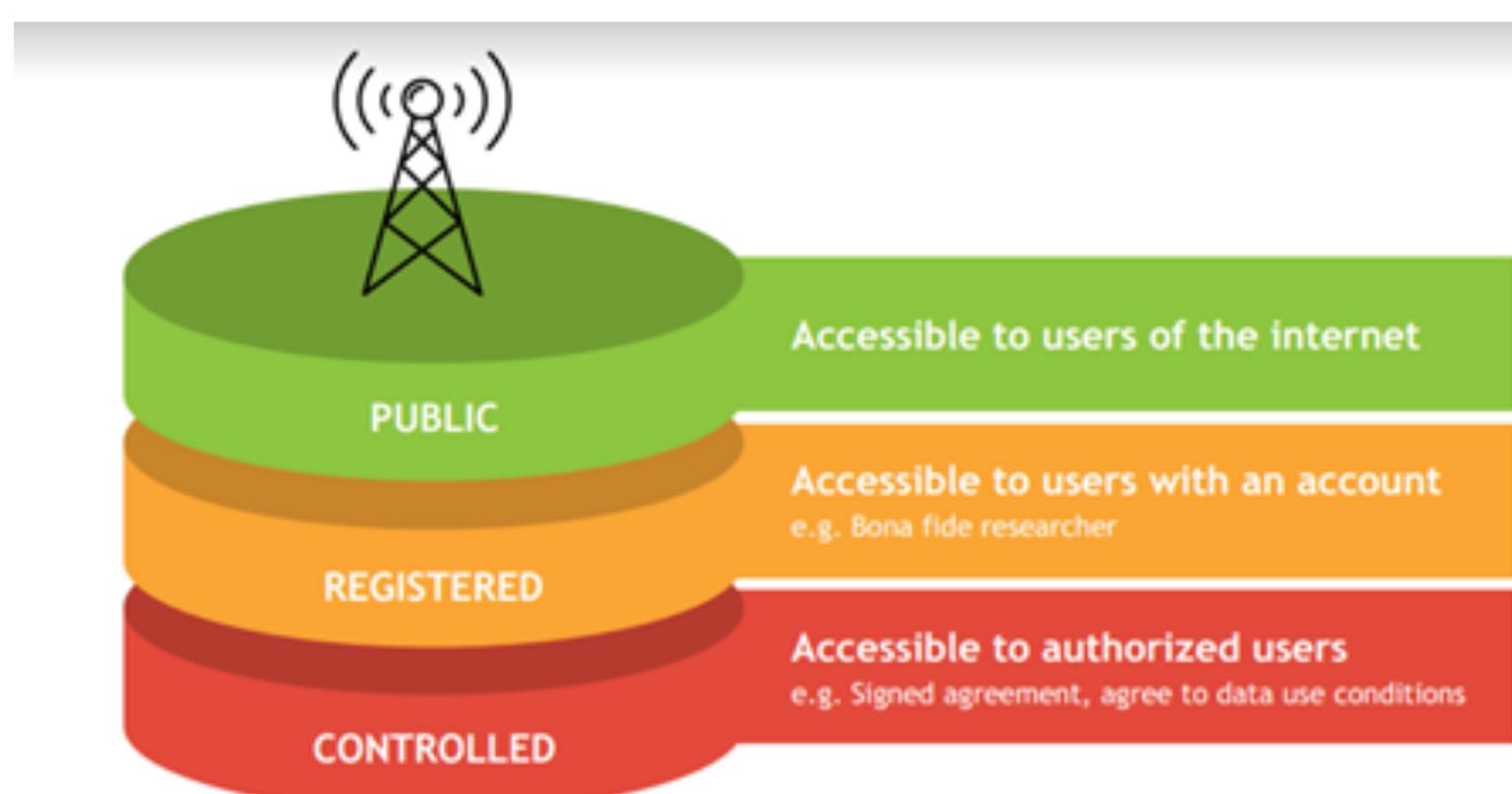
- Scoping queries through "biodata" parameters
- Extending the queries towards clinically ubiquitous variant formats
 - ▶ cytogenetic annotations, named variants, variant effects
- Beacon queries as entry for **data delivery**
 - ▶ Beacon v2 permissive to respond with variety of data types
 - Phenopackets, biosample data, cohort information ...
 - ▶ handover to stream and download using htsgt, VCF, EHRs
- Interacting with EHR standards
 - ▶ FHIR translations for queries and handover ...
- Beacons as part of local, secure environments
- Authentication to enable non-aggregate, patient derived datasets
 - ▶ ELIXIR AAI with compatibility to other providers (OAuth...)

Definitely breaks the
"Relative Security
by Design"
Concept!

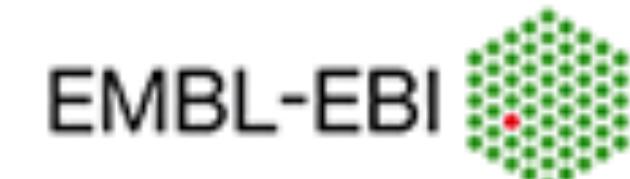
Beacon API v2

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



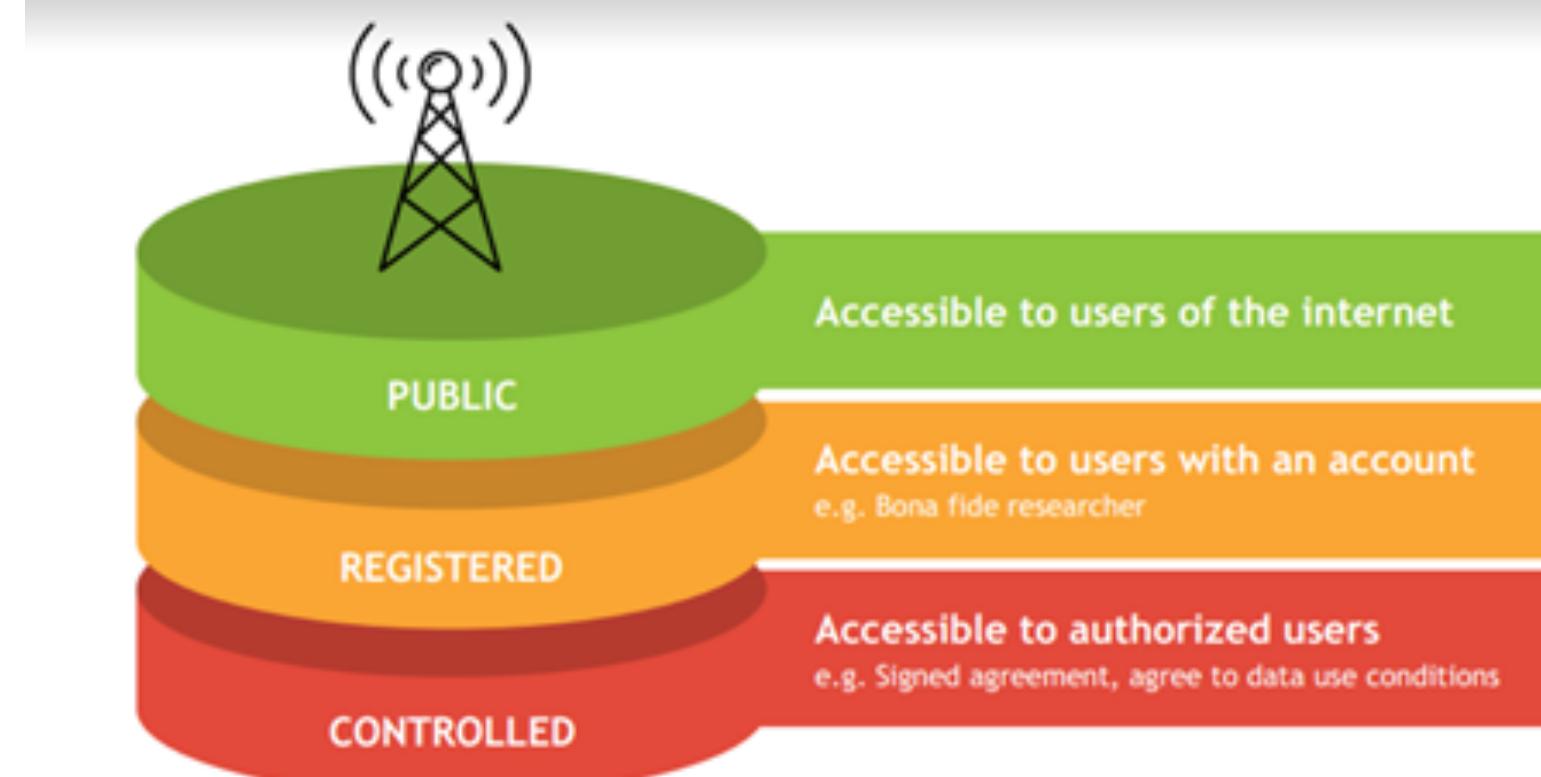
Example Users



Beacon Security

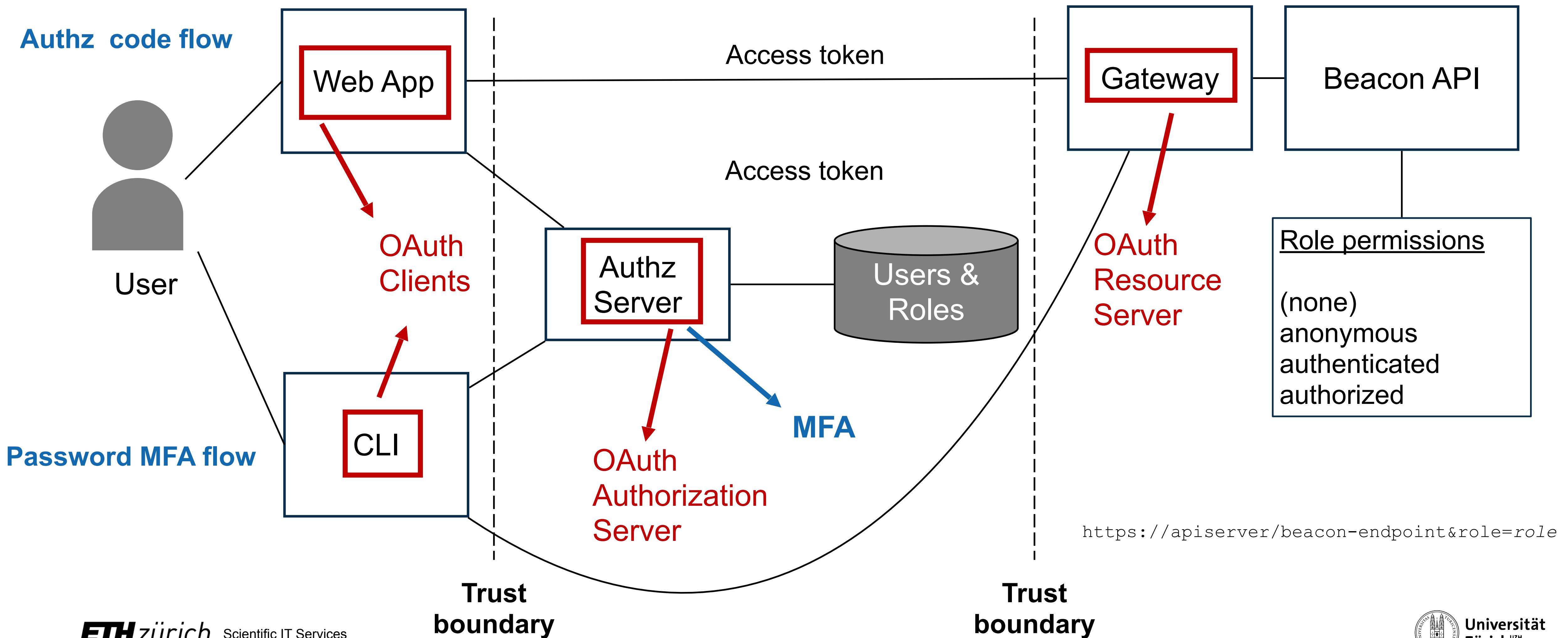
Security by Design ... if Implemented in the Environment

- the beacon API specification does not implement explicit security (e.g. checking user authentication and authorization)
- the framework implements different levels of response granularity which can be mapped to authorization levels (**boolean** / **count** / **record** level responses)
- implementations can have beacons running in secure environments with a **gatekeeper** service managing authentication and authorization levels, and potentially can filter responses for escalated levels
- the backend can implement additional access reduction, on a user <-> dataset level if needed



Architecture

Running the *bycon* stack in a secure environment



Architecture

Running the *bycon* stack in a secure environment

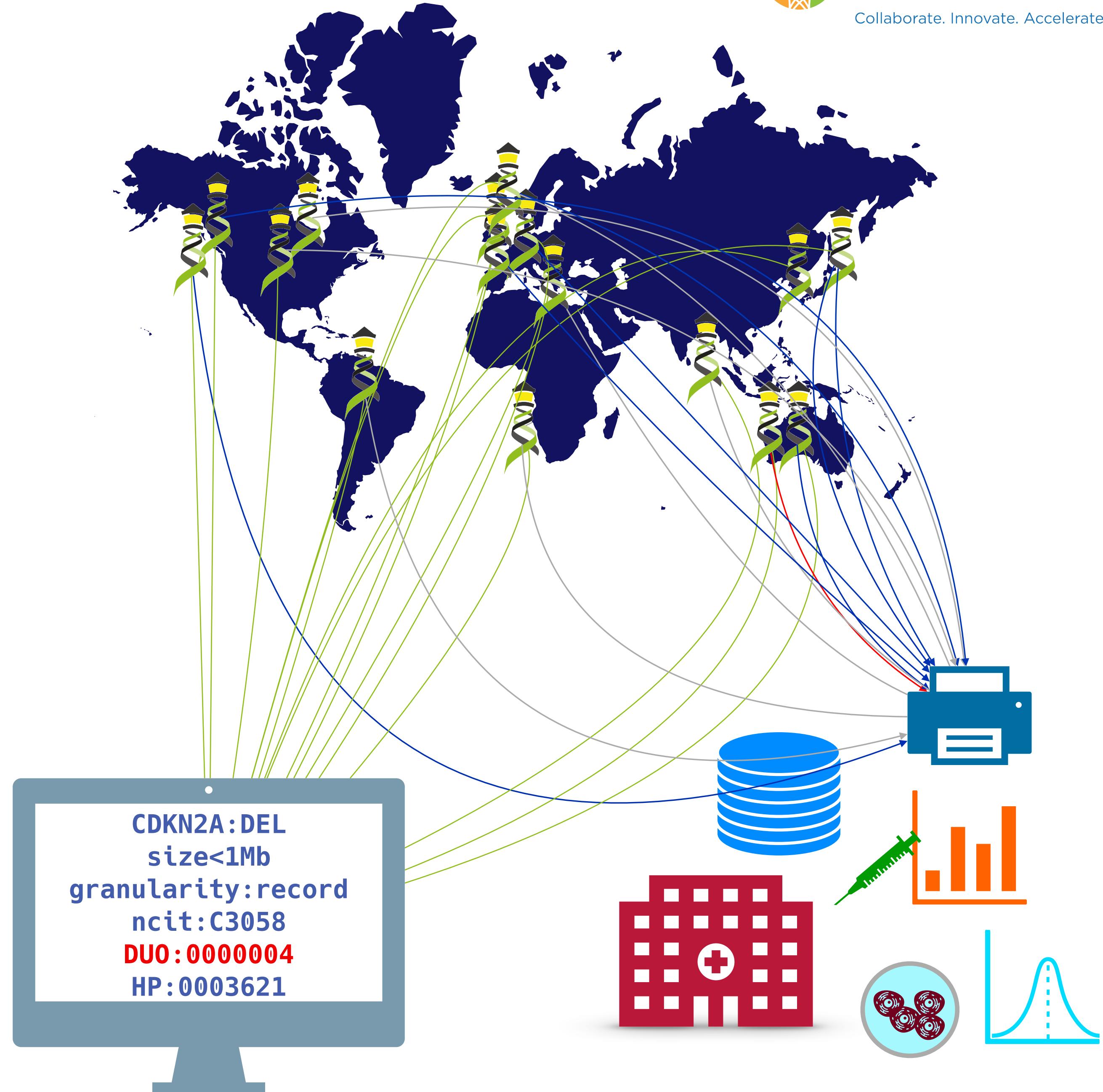
- The **Beacon API** implementation stack (e.g. bycon) is authentication procedure agnostic; i.e. it just accepts that a user has been authenticated and passed the general authorization gatekeeping
- The **Beacon API** server and the **Gateway** reside in a single VM, with only the **Gateway**'s port exposed (with TLS). Beacon's port is not exposed by the VM and can only be reached through the **Gateway**
- The **Authentication Server** can run on the same or separate VM; needs a database with user accounts.
- The **Web Client** can be in the same VM or a separate one.
- Separate **Gateways** (e.g. university firewall vs. public) can be configured to modify different roles, e.g. the public gateway may turn registered roles into anonymous, regardless of whether the user has registered status
- Users can write their own clients (web / command line) which are registered with the **Authorization Server** and are issued with a Client ID and Client Secret to use against the **Authorization Server**.

What Can You Do?

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



Collaborate!



Events

GA4GH Ascona Connect

[REGISTER FOR THIS EVENT](#) 

21 Apr 2024



This hybrid working meeting aims to support GA4GH contributors in advancing product development and gathering feedback on needs.



Image summary: Join us for GA4GH Connect from 21 to 24 April 2024.

<https://www.ga4gh.org/event/ga4gh-connect-2/>

Friday, 19 April 2024

10:00 - 14:15 **GA4GH Connect satellite workshop: Beacon implementers**

Sunday, 21 April 2024

10:30 - 12:00	VCF v4.5: scalability and methylation	GA4GH Implementation Forum (GIF)	REWS general
12:00 - 13:00	Lunch		
13:00 - 14:30	Driver Project workshop: data harmonisation	Opportunities and obligations in conducting responsible genomics research: role-based perspectives	GA4GH Infect Community of
14:30 - 15:00	Break		
15:00 - 16:30	Federated variant level matching	Data Model and Schema Consensus (DaMaSC) session 1	GKS Connect 2024 release workshop

Monday, 22 April 2024

09:00 - 10:30	Opening session (plenary session)		
10:30 - 12:00	Phenopackets for querying Beacon	Driver Project workshop: artificial intelligence and machine learning	Ethical final Access Clause
12:00 - 13:00	Lunch		
13:00 - 14:30	Welcome to the Cat-VRS	Experiments Metadata Standard	GA4GH Connect 2024
14:30 - 15:00	Break		
15:00 - 16:30	Passports in production	Driver Project workshop: data discovery	Command line interface for GA4GH environments enhancement with confidential computing

Tuesday, 23 April 2024

09:00 - 10:30	Around the world in one query	Unveiling GA4GH's five-year Equity, Diversity and Inclusion (EDI) Strategic Plan
09:00 - 12:00	DRS v1.5: key features and plans for 2024	
10:30 - 12:00	Technical Alignment Subcommittee (TASC) meeting	Beacon cohorts and aggregator
12:00 - 13:00	Lunch	
13:00 - 14:30	Beacon variants	Data visiting
14:30 - 15:00	Break	Sequence Annotation (SA) model hackathon (v0.2)
18:00	Cocktail reception at Hotel Eden Roc	

Wednesday, 24 April 2024

09:00 - 10:30	Implementation of Beacon in cancer use cases		
09:00 - 12:00	GKS implementation forum		
10:30 - 12:00	Beacon resolved: a session for development		
12:00 - 13:00	Lunch		
13:00 - 14:30	Federated cohort building		GKS road map: scope and priorities
14:30 - 15:00	Break		
15:00 - 16:30	Data Model and Schema Consensus (DaMaSC) session 2		Beacon filter solutions
	Driver Project workshop: multi-site geography collaboration and governance		

Beacon Queries

Missing or ill defined options

- Translocations
 - in principle possible (start bracket with "referenceName" and end bracket with "mateName" but not yet documented / battle tested)
- Functional elements?
- Cytoband queries?
- Exon hits beyond specifying individual ones by sequence
- Tandem dups ...

→ **Beacon & Genomic Variations
Scout Team**

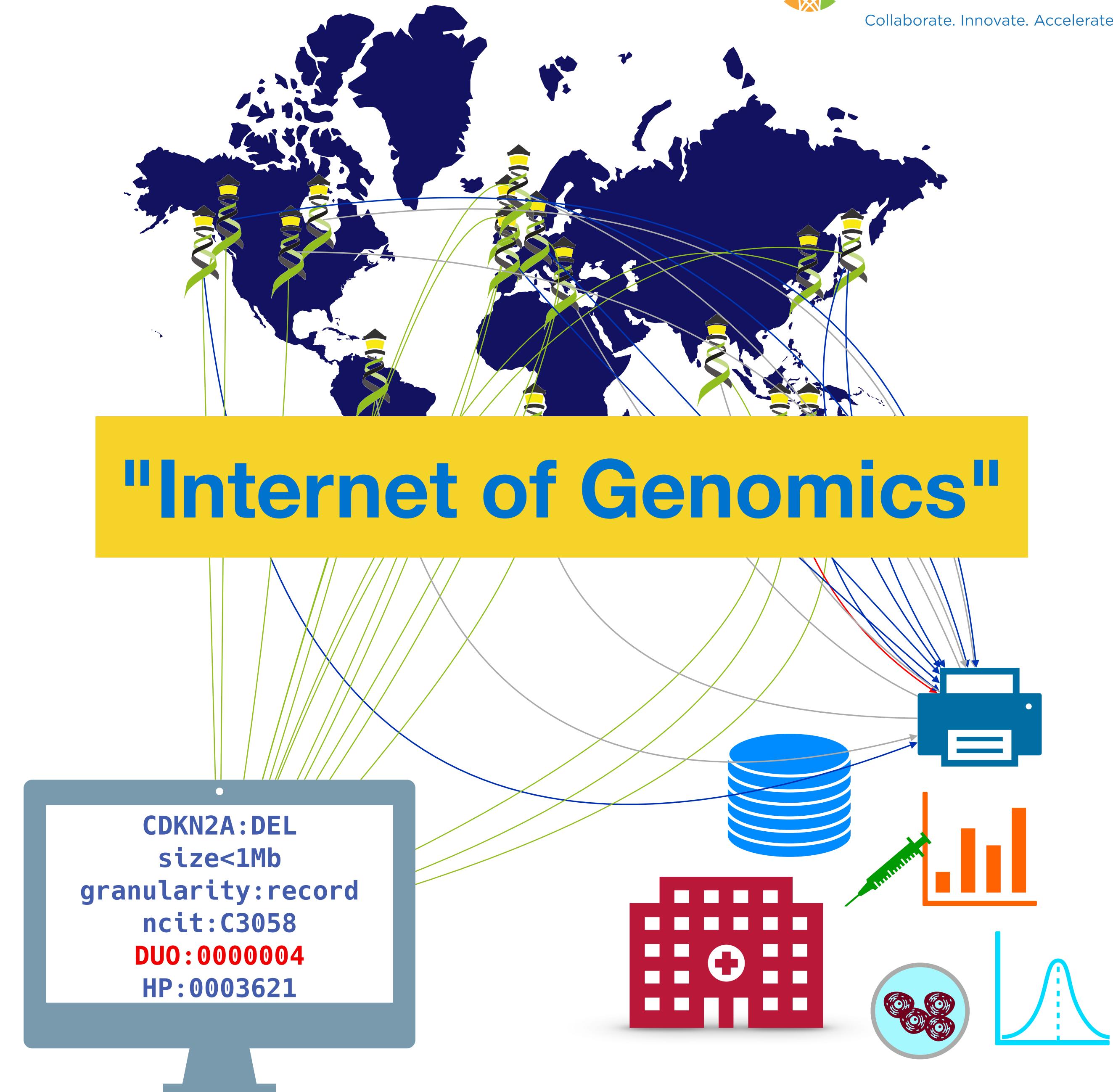
The screenshot shows the Beacon+ query interface. At the top, there is a navigation bar with the logo 'Beacon+' followed by 'Progenetix' and 'Help'. Below the navigation bar is a header titled 'Beacon Query Types' with several tabs: 'Sequence / Allele' (which is selected), 'CNV (Bracket)', 'Genomic Range', 'Aminoacid', 'Gene ID', 'HGVS', and 'Sam'. Underneath the tabs is a section titled 'Dataset' containing a dropdown menu set to 'Test Database - examplez'. The main query area contains several input fields: 'Chromosome' (with a dropdown menu labeled 'Select...'), 'Variant Type' (with a dropdown menu labeled 'Select...'), 'Start or Position' (containing the value '19000001-21975098'), 'Reference Base(s)' (containing 'N'), 'Alternate Base(s)' (containing 'A'), and 'Select Filters' (with a dropdown menu labeled 'Select...'). At the bottom of the query area is a large blue button labeled 'Query Database'. Below the query area are two sections: 'Form Utilities' containing buttons for 'Gene Spans' and 'Cytoband(s)', and 'Query Examples' containing links to 'CNV Example', 'SNV Example', 'Range Example', 'Gene Match', 'Aminoacid Example', and 'Identifier - HeLa'.

What Can You Do?

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



Collaborate!





Universität
Zürich^{UZH}



Swiss Institute of
Bioinformatics





Universität
Zürich^{UZH}



Swiss Institute of
Bioinformatics

elixir