

Data Resources, Sharing, Discovery in Biomedical Genetics and Cancer Genomics

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



Universität
Zürich^{UZH}



Swiss Institute of
Bioinformatics



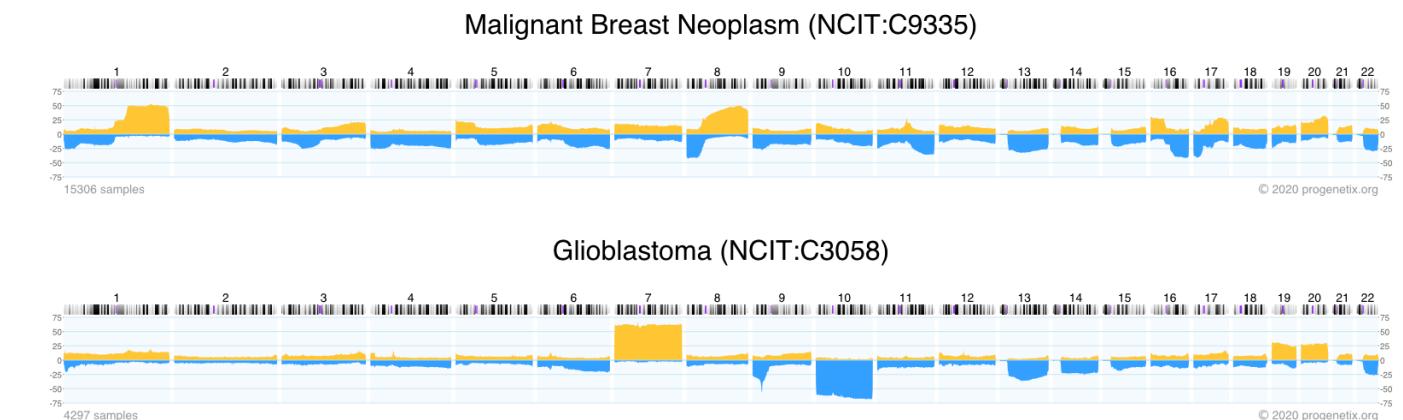
Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.



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"

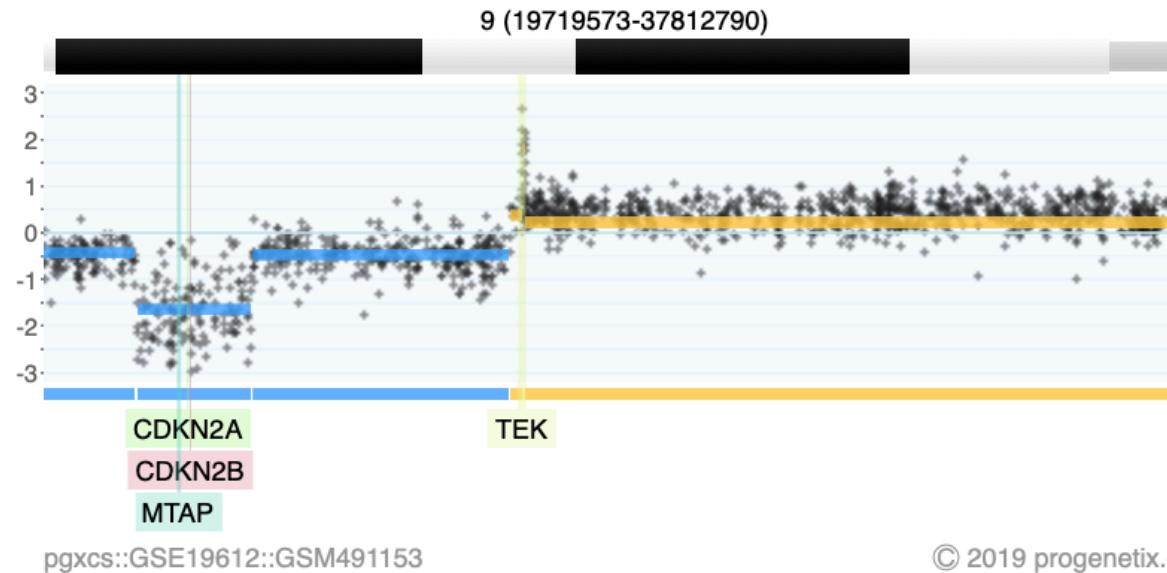


Theoretical Cytogenetics and Oncogenomics

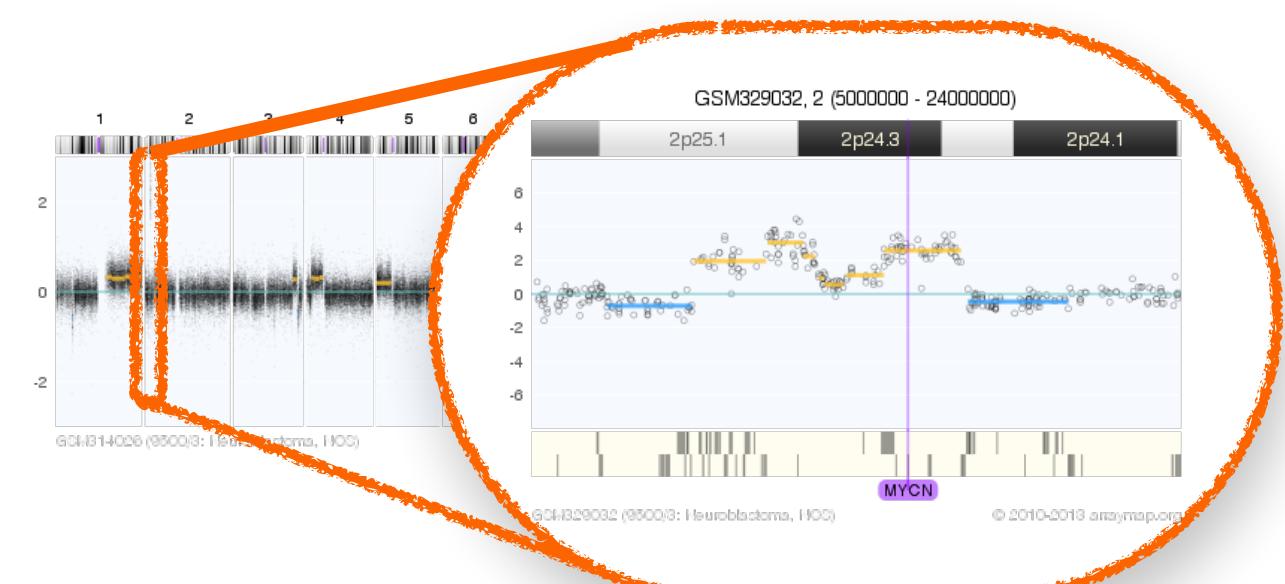
Research | Methods | Standards

Genomic Imbalances in Cancer - Copy Number Variations (CNV)

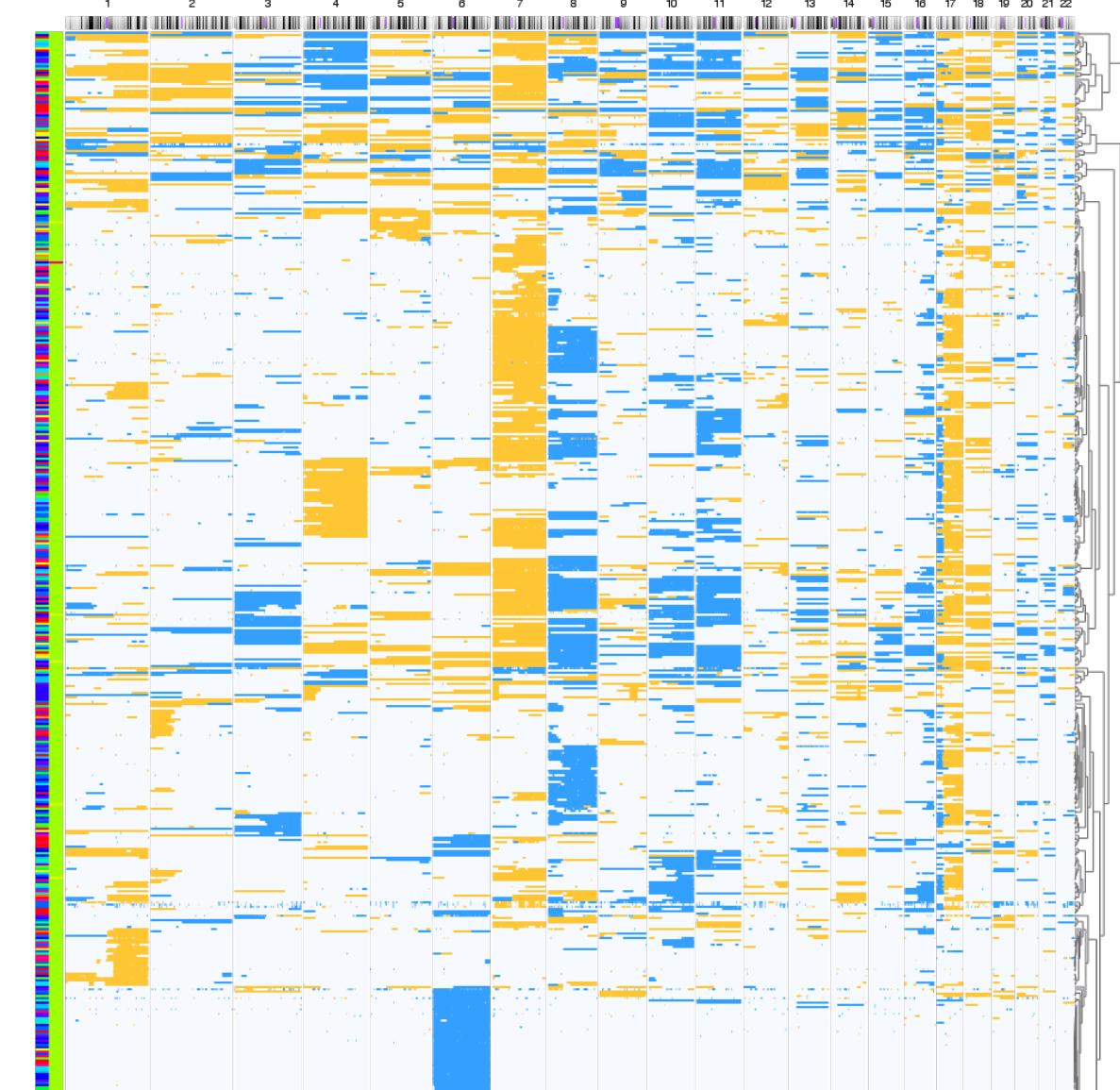
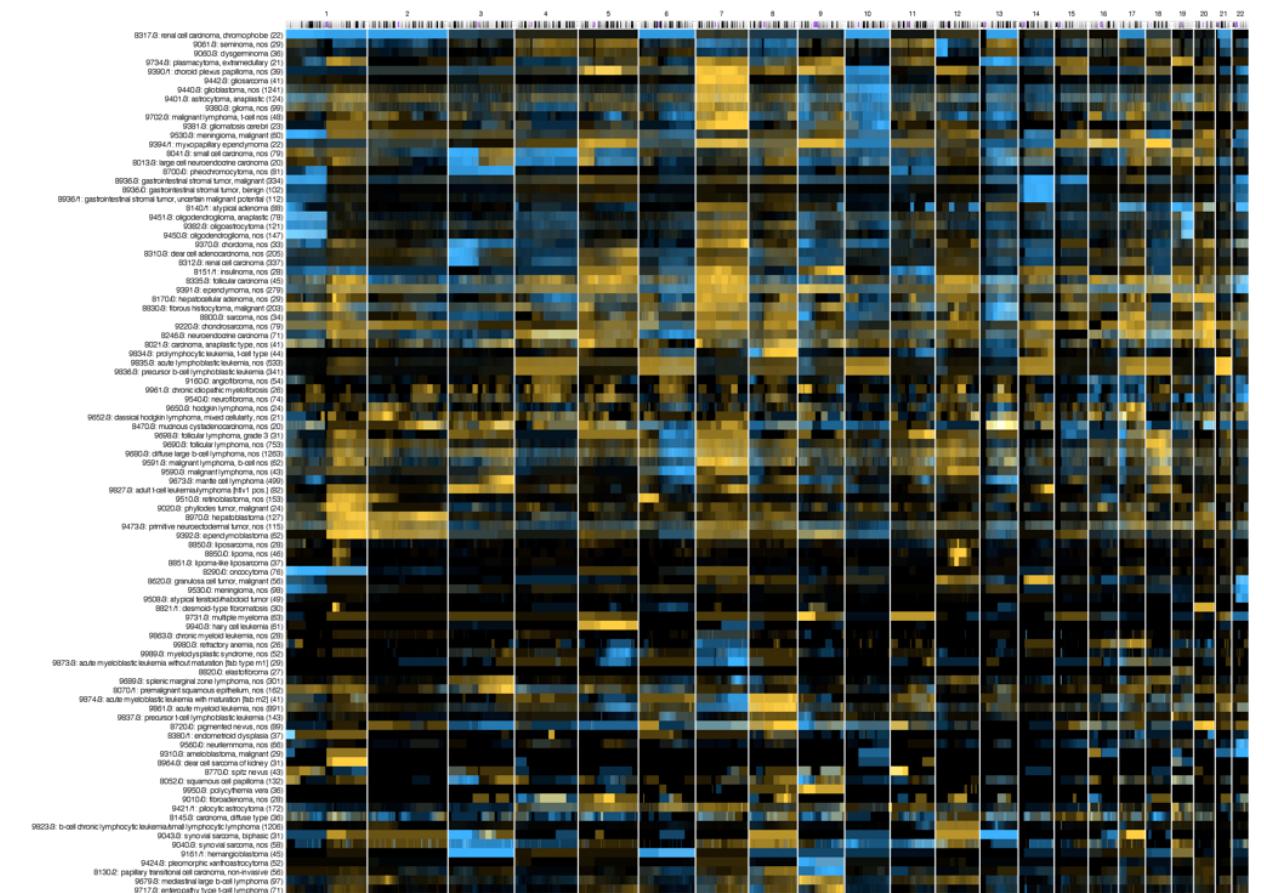
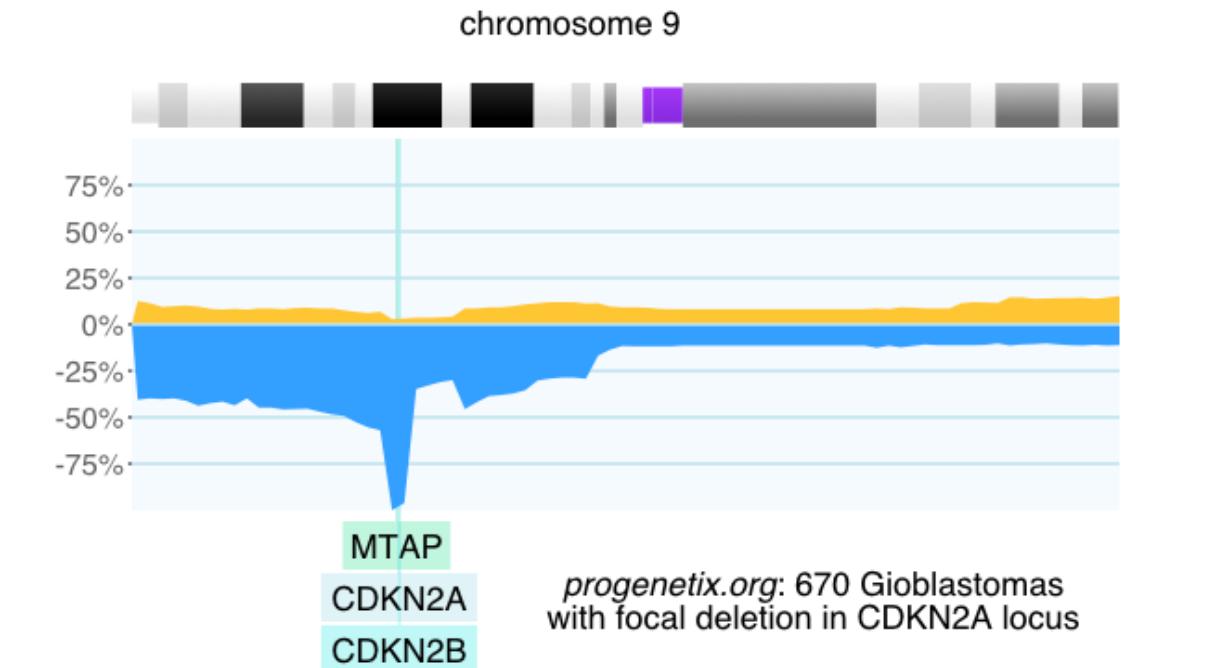
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)



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

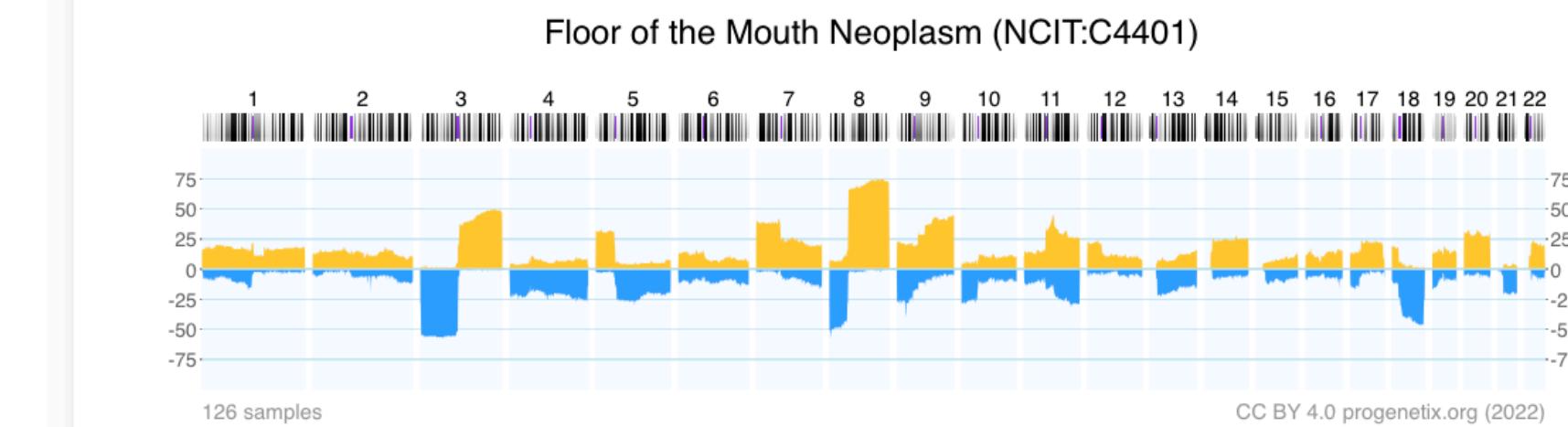
Beacon⁺

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.



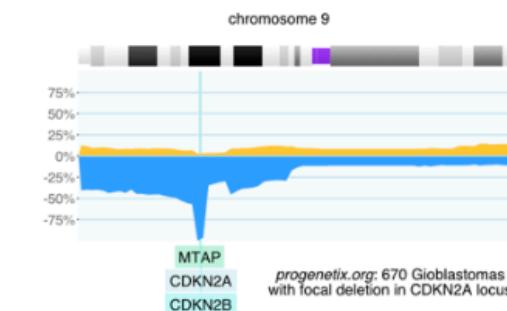
[Download SVG](#) | [Go to NCIT:C4401](#) | [Download CNV Frequencies](#)

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. NCI, 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 S

Head and Neck Squamous Cell Carcinoma (NCIT:C34447)

Subset Type

- NCI Thesaurus OBO Edition NCIT:C34447 ↗

Sample Counts

- 2061 samples
- 57 direct NCIT:C34447 code matches
- 200 CNV analyses
 - Download CNV frequencies ↗

Search Samples

Select NCIT:C34447 samples in the [Search Form](#)

Raw Data (click to show/hide)

Download SVG | Go to NCIT:C34447 | Download CNV Frequencies

- NCIT:C6958: Astrocytic Tumor (5882 samples, 5896 CNV profiles)
- NCIT:C6960: Oligodendroglial Tumor (703 samples, 703 CNV profiles)
- NCIT:C8501: Brain Stem Glioma (2 samples, 2 CNV profiles)
- NCIT:C3716: Primitive Neuroectodermal T... (2213 samples, 2214 CNV profiles)
- NCIT:C4747: Glioneuronal and Neuronal Tumors (89 samples, 89 CNV profiles)
- NCIT:C6965: Pineal Parenchymal Cell Neoplasm (51 samples, 51 CNV profiles)

Cancer Genomics Reference Resource

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



Universität
Zürich UZH

—progenetix—



Swiss Institute of
Bioinformatics

Edit Query

Assembly: GRCh38 chro: refseq:NC_000009.12 Start: 21500001-21975098

End: 21967753-22500000 Type: EFO:0030067 Filters: NCIT:C3058

progenetix

Matched Samples: 657

Retrieved Samples:

Variants: 276

Calls: 659

UCSC region ↗

Variants in UCSC ↗

Dataset Responses (JSON) ↗

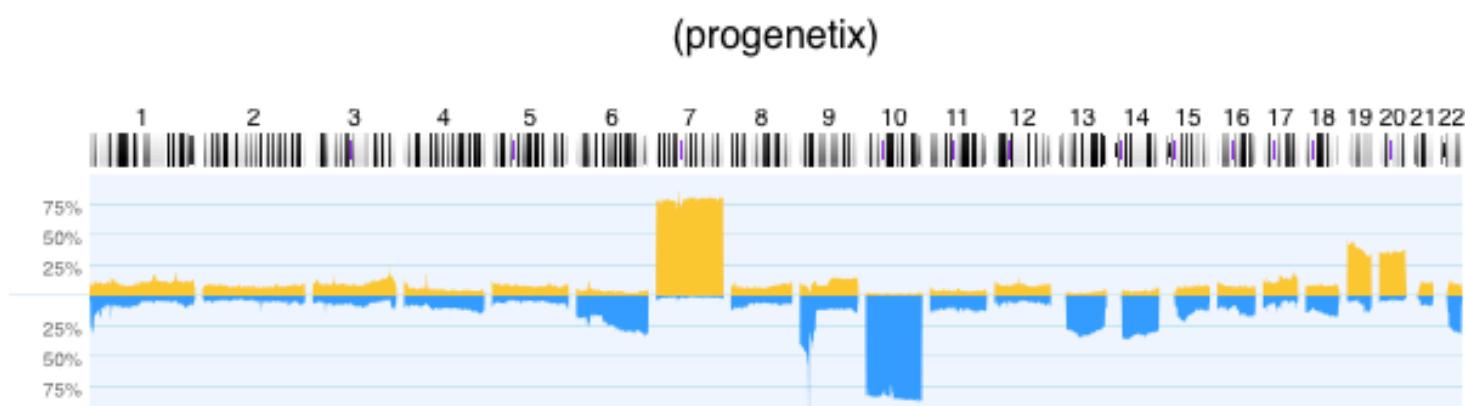
Visualization options

Results

Biosamples

Biosamples Map

Variants



© CC-BY 2001 - 2023 progenetix.org

Reload histogram in new window ↗

Matched Subset Codes	Subset Samples	Matched Samples	Subset Match Frequencies
pgx:icdot-C71.4	4	1	0.250
pgx:icdom-94403	4286	653	0.152
NCIT:C3058	4370	653	0.149
pgx:icdot-C71.1	14	2	0.143
pgx:icdot-C71.9	7204	640	0.089
NCIT:C3796	84	4	0.048
pgx:icdom-94423	84	4	0.048
pgx:icdot-C71.0	1714	14	0.008

Download Sample Data (TSV)

1-657 ↗

Download Sample Data (JSON)

1-657 ↗

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

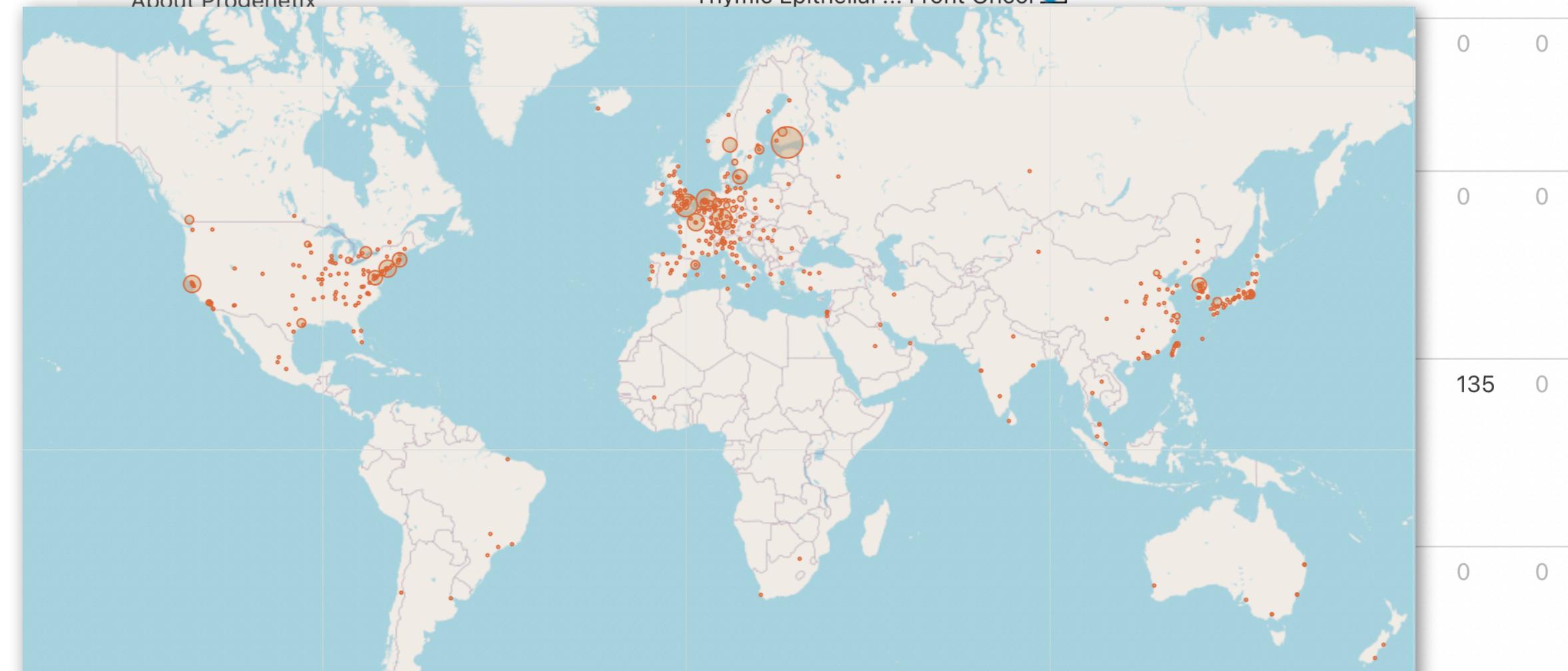
Progenetix Publication Collection

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](#) [↗](#).

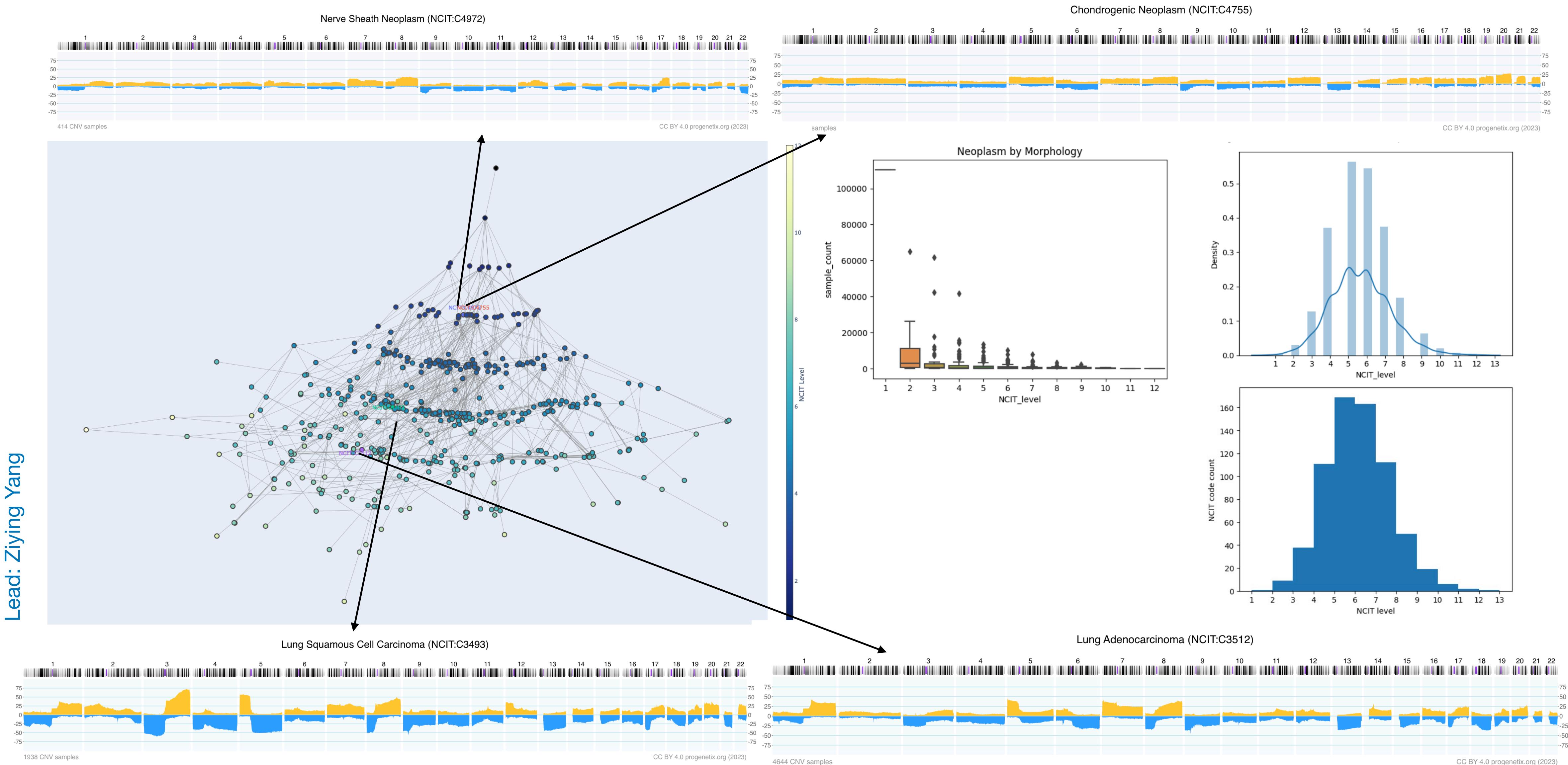
Publications (3349)		Samples				
id i ▾	Publication	cCGH	aCGH	WES	WGS	pgx
PMID:34604048	Dai J, Jiang M, He K, Wang H, Chen P et al. (2021) DNA Damage Response and Repair Gene Alterations Increase Tumor Mutational Burden and ... <i>Front Oncol</i>	0	0	122	0	0
PMID:34573430	Juhari WKW, Ahmad Amin Noordin 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.

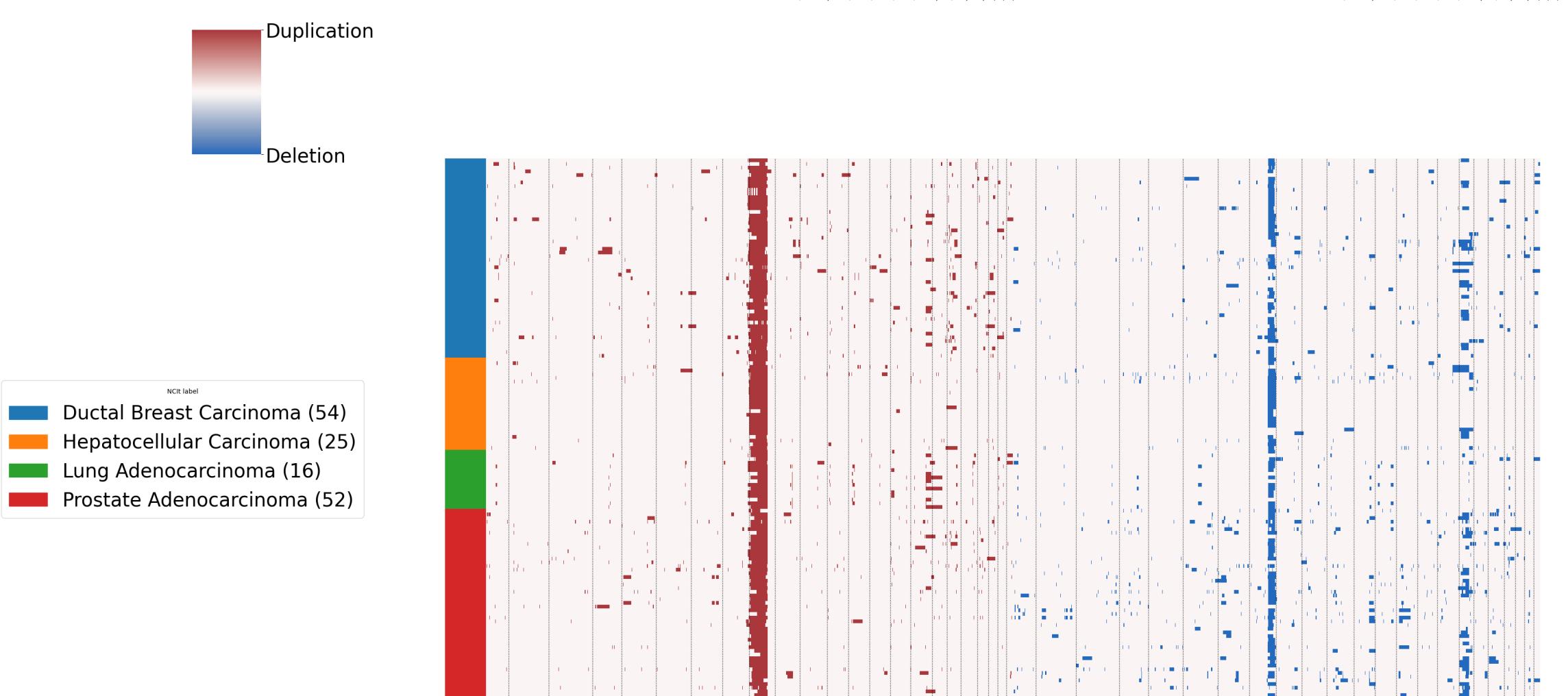
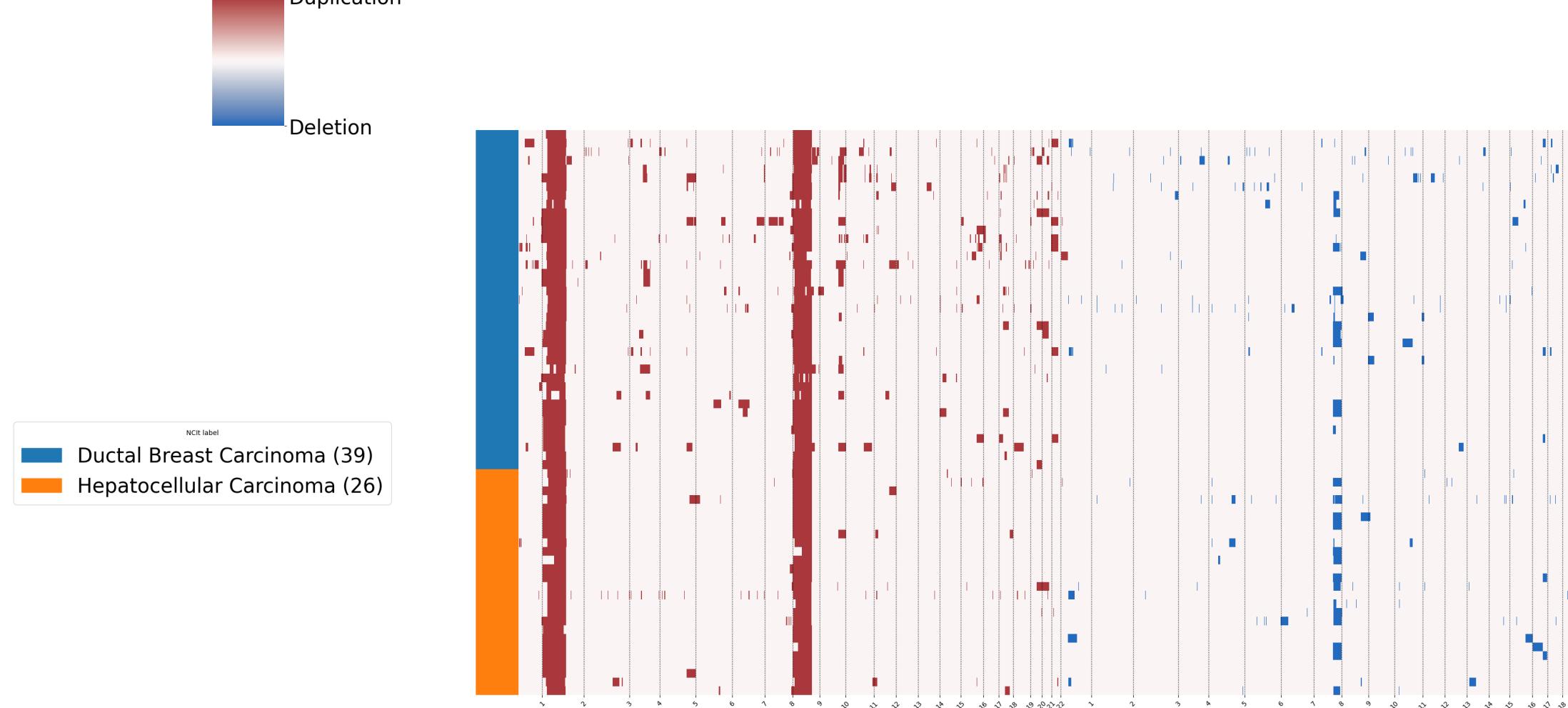
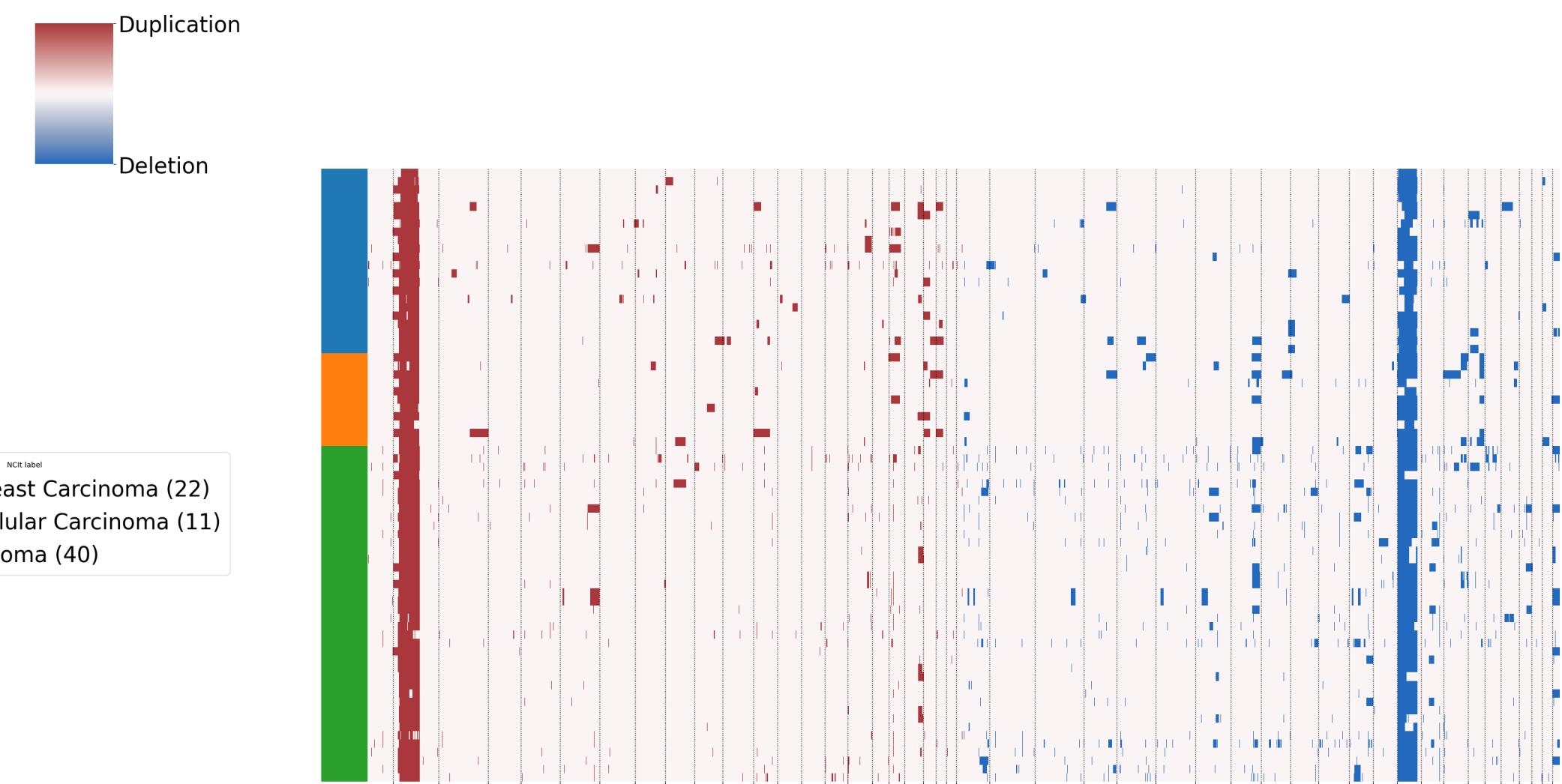
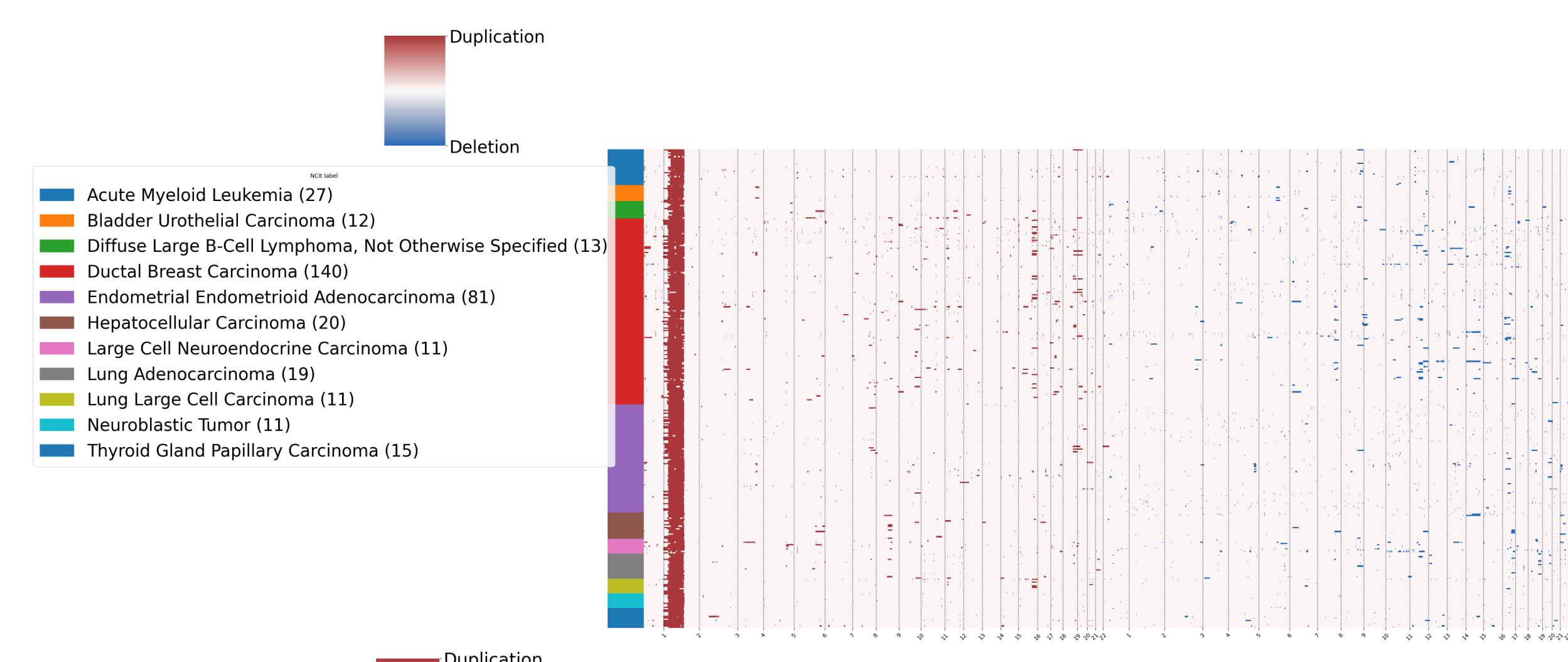
CNV profiles heterogeneity vs cancer classification

Correspondance of genomic profiles to NCIT cancer hierarchy



Example Use of Progenetix Data

Inter-tumoral CNV pattern similarity



Mostly Carcinoma and Adenocarcinoma in different organs

Cancer Cell Lines

Cancer Genomics Reference Resource

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

→ integration in data federation approaches

cancercelllines.org

Lead: Rahel Paloots



Cold
Spring
Harbor
Laboratory

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

cancercelllines.org - a Novel Resource for Genomic Variants in Cancer Cell Lines

Rahel Paloots, Michael Baudis

doi: <https://doi.org/10.1101/2023.12.12.571281>

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



Cancer Cell Lines

Search Cell Lines

Cell Line Listing

CNV Profiles by
Cancer Type

Documentation

News

Progenetix

Progenetix Data

Progenetix

Documentation

Publication DB

Cancer Cell Lines by Cellosaurus ID

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

Sample selection follows a hierarchical system in which samples can be retrieved at any level. For example, selecting "HOS" for HeLa will also return the daughter lines by default - but can also be used to retrieve all samples for HOS.

Cell Lines (with parental/derived hierarchies)

Filter subsets e.g. by prefix: Hierarchy Depth:

- > cellosaurus:CVCL_0312: HOS (204 samples)
- > cellosaurus:CVCL_1575: NCI-H650 (6 samples)
- > cellosaurus:CVCL_1783: UM-UC-3 (9 samples)
- > cellosaurus:CVCL_0004: K-562 (28 samples)
- cellosaurus:CVCL_3827: K562/Adenocarcinoma (1 sample)
- > cellosaurus:CVCL_0589: Kasumi-1 (9 samples)

Cell Line Details

HOS (cellosaurus:CVCL_0312)

Subset Type

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

Sample Counts

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

Search Samples

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

Raw Data (click to show/hide)

HOS (cellosaurus:CVCL_0312)

21 CNV samples

CC BY 4.0 progenetix.org (2023)

Download SVG | Go to cellosaurus:CVCL_0312 | Download CNV Frequencies

Gene Matches

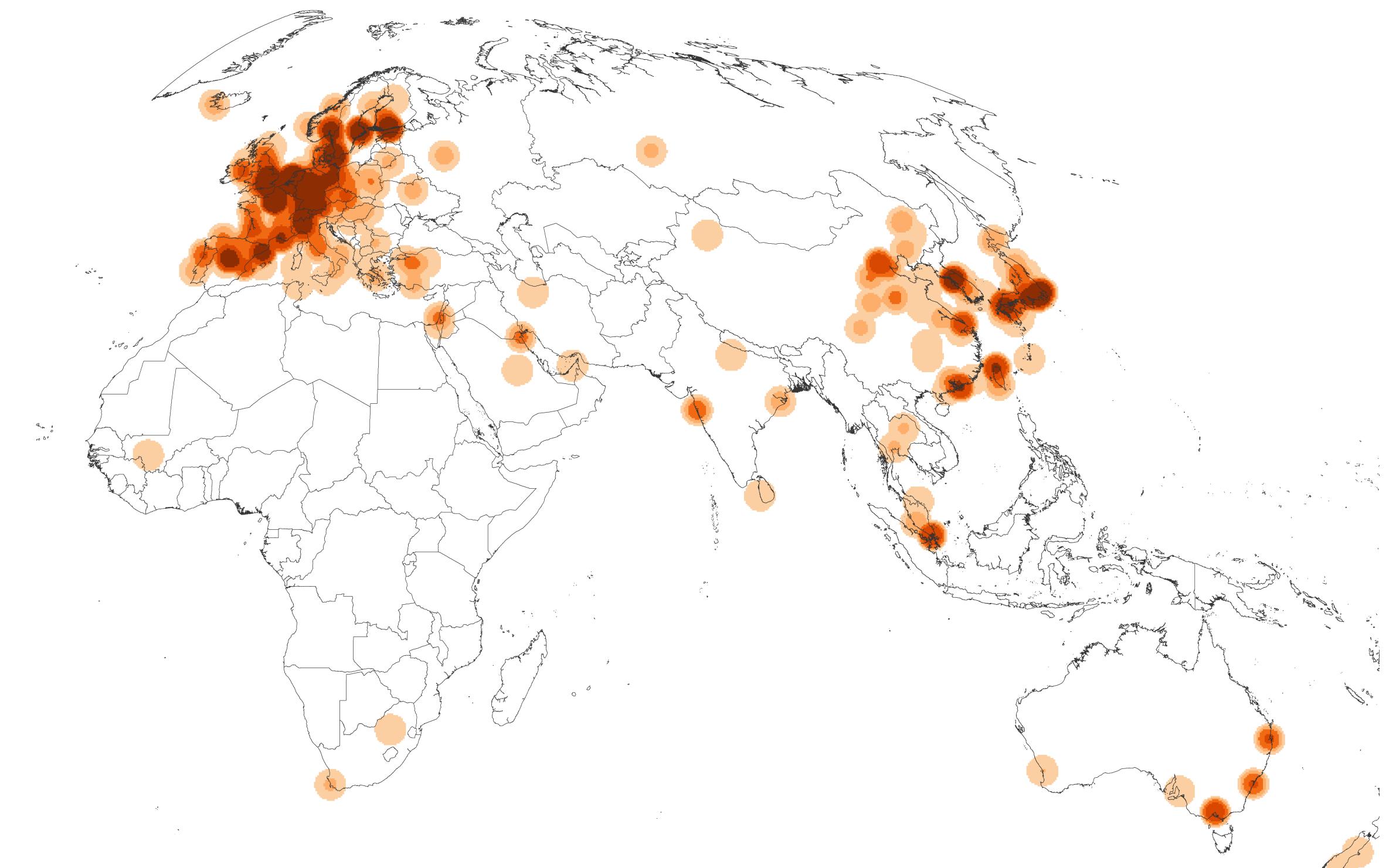
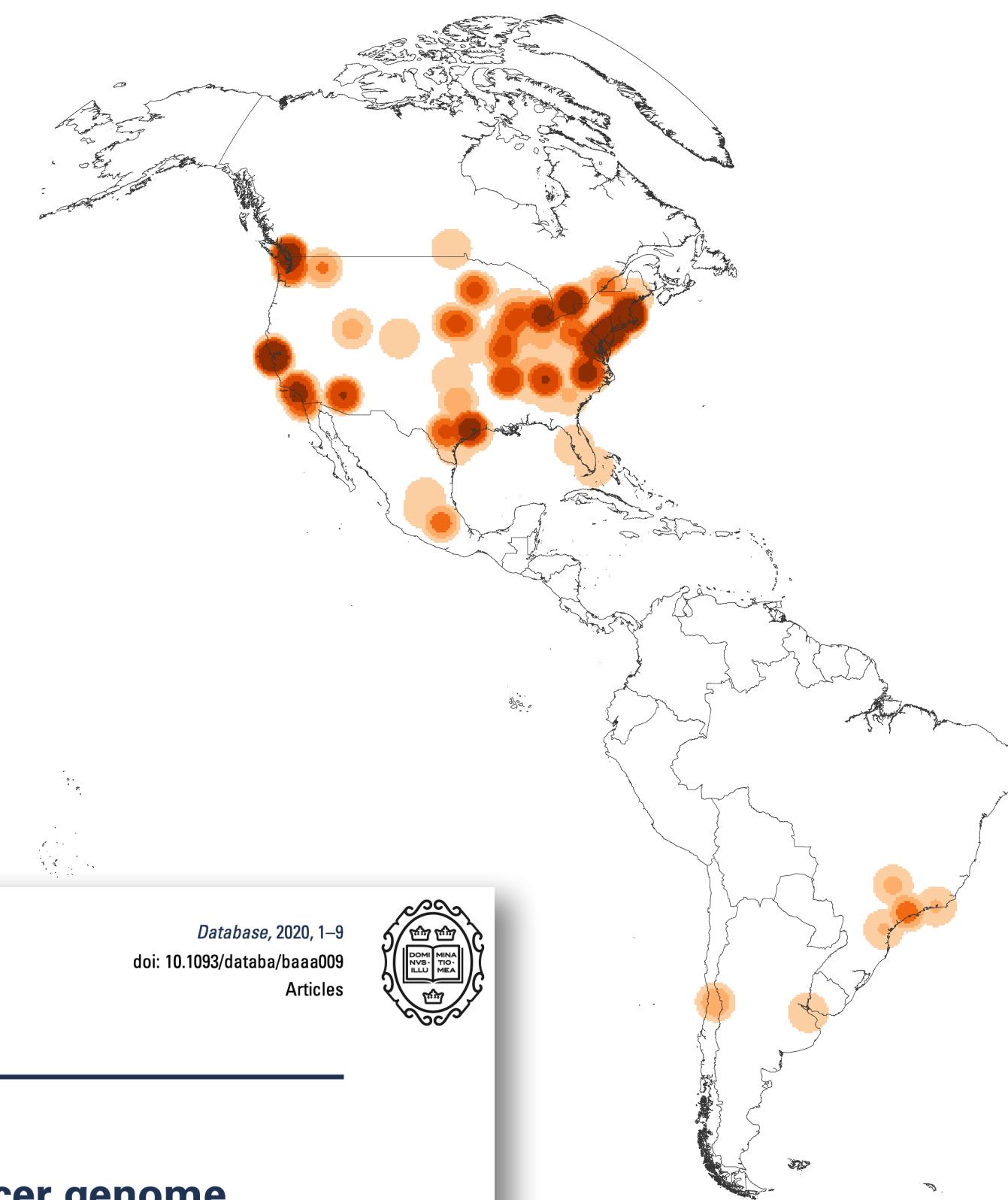
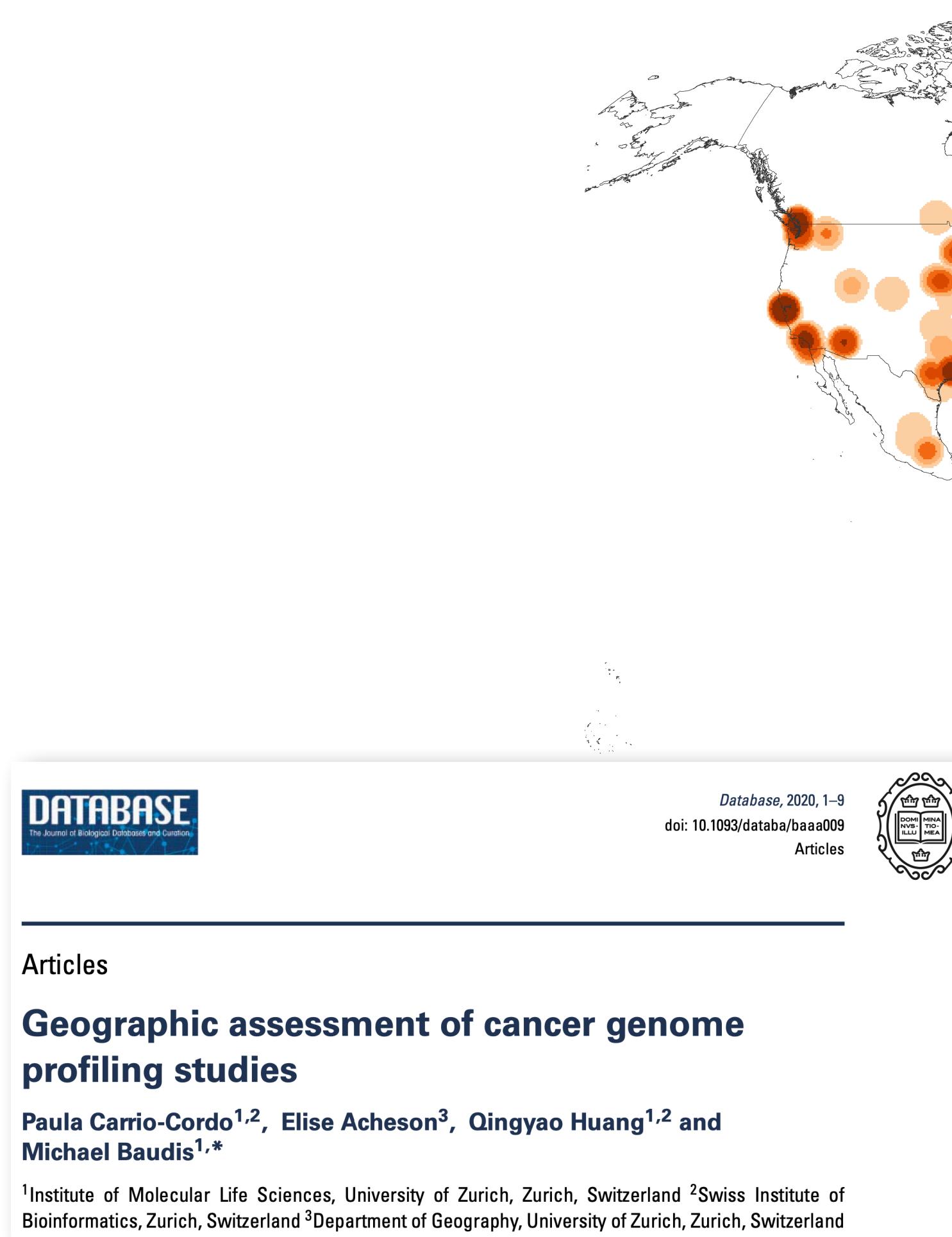
Gene	Description	Abstract
ALK	ABC-14 cells harbored no ALK mutations and were sensitive to ... crizotinib while also exhibiting MNNG HOS transforming gene (MET)	Rapid Acquisition of Alectinib Resistance in ALK-Positive Lung Cancer With High Tumor Mutation Burden (31374369)
AREG	crizotinib while also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

Cytoband Matches

Variants

Where Does Cancer Genomic Data Come From?

Geographic bias in published cancer genome profiling studies



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

Different Approaches to Data Sharing



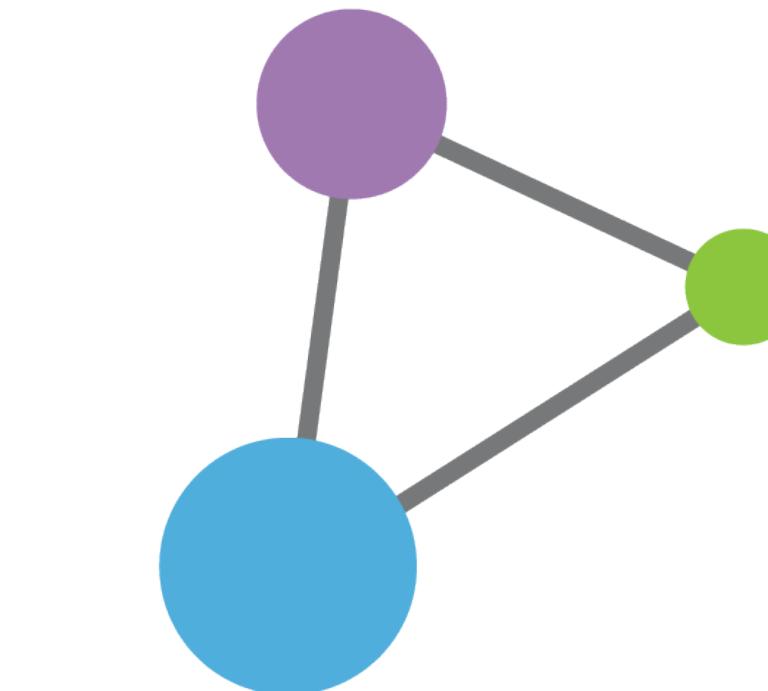
Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets

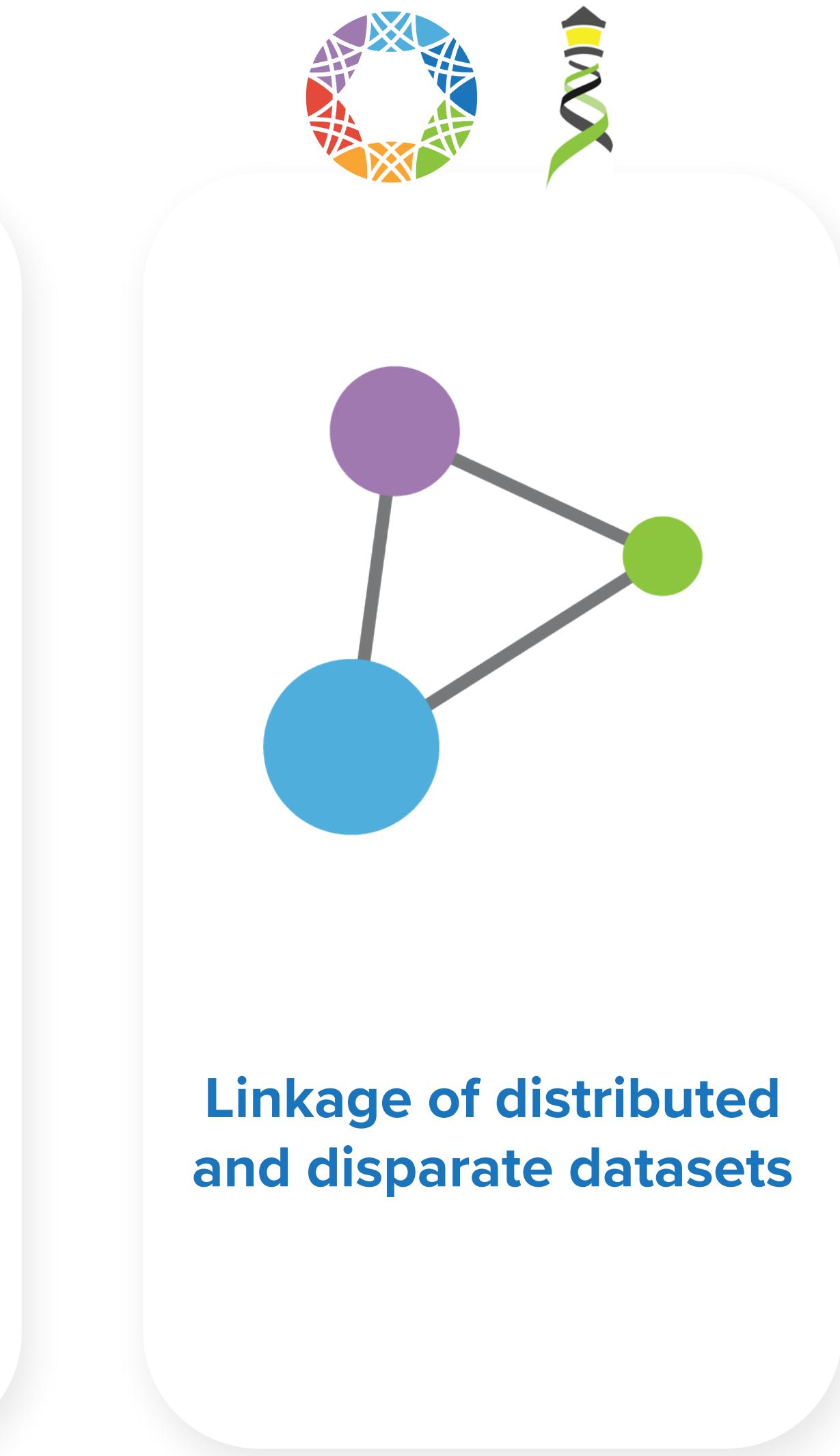
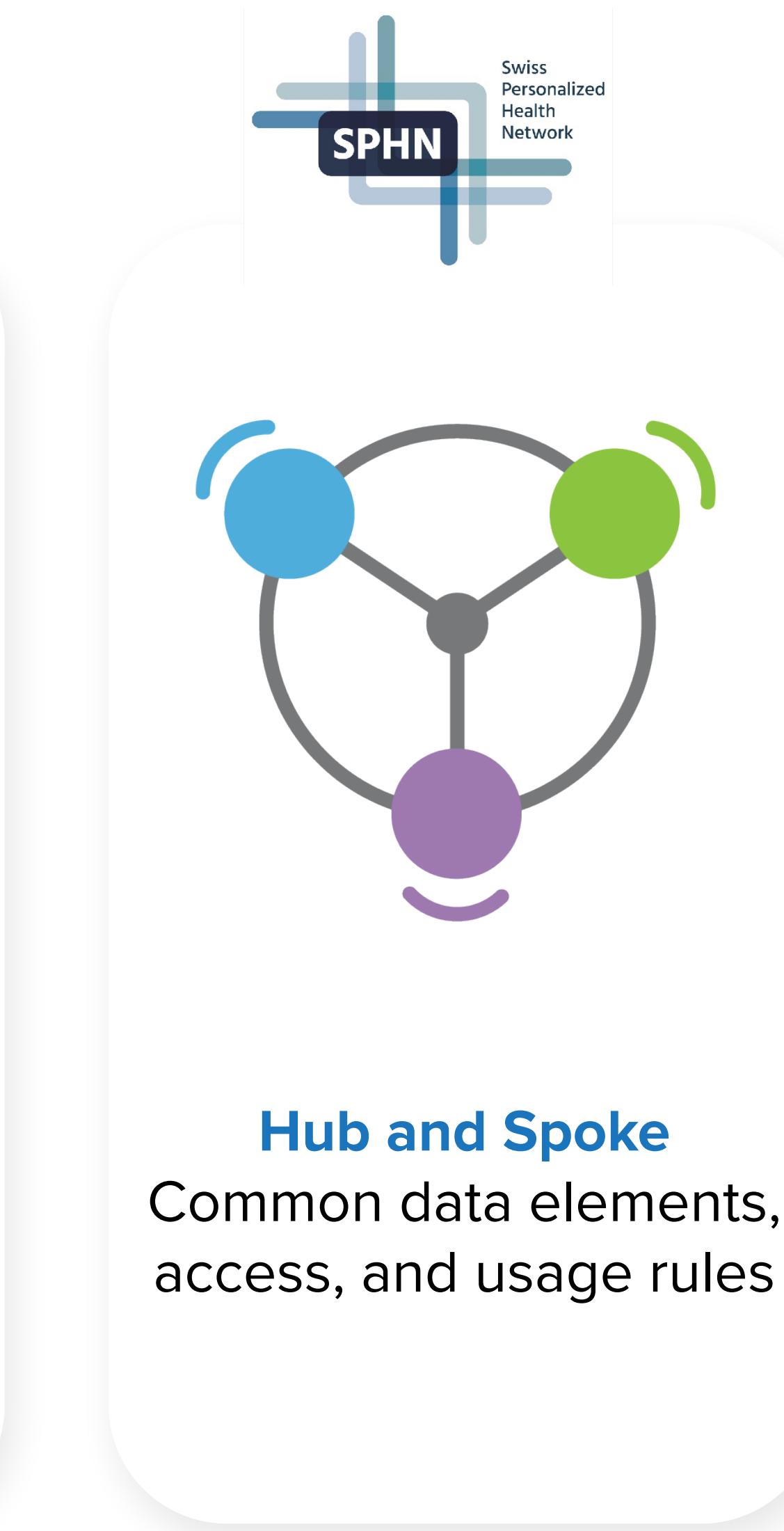
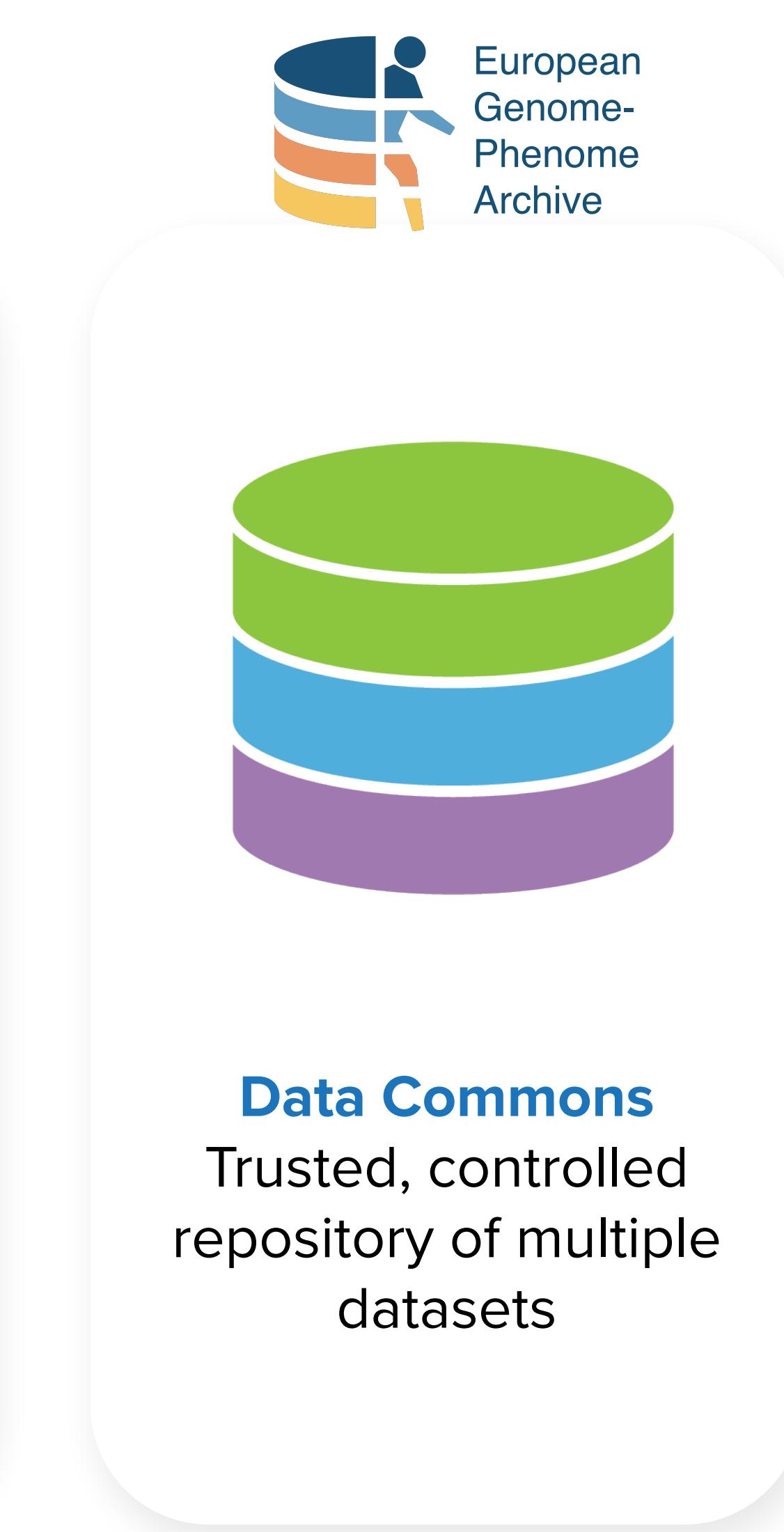


Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

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*”)



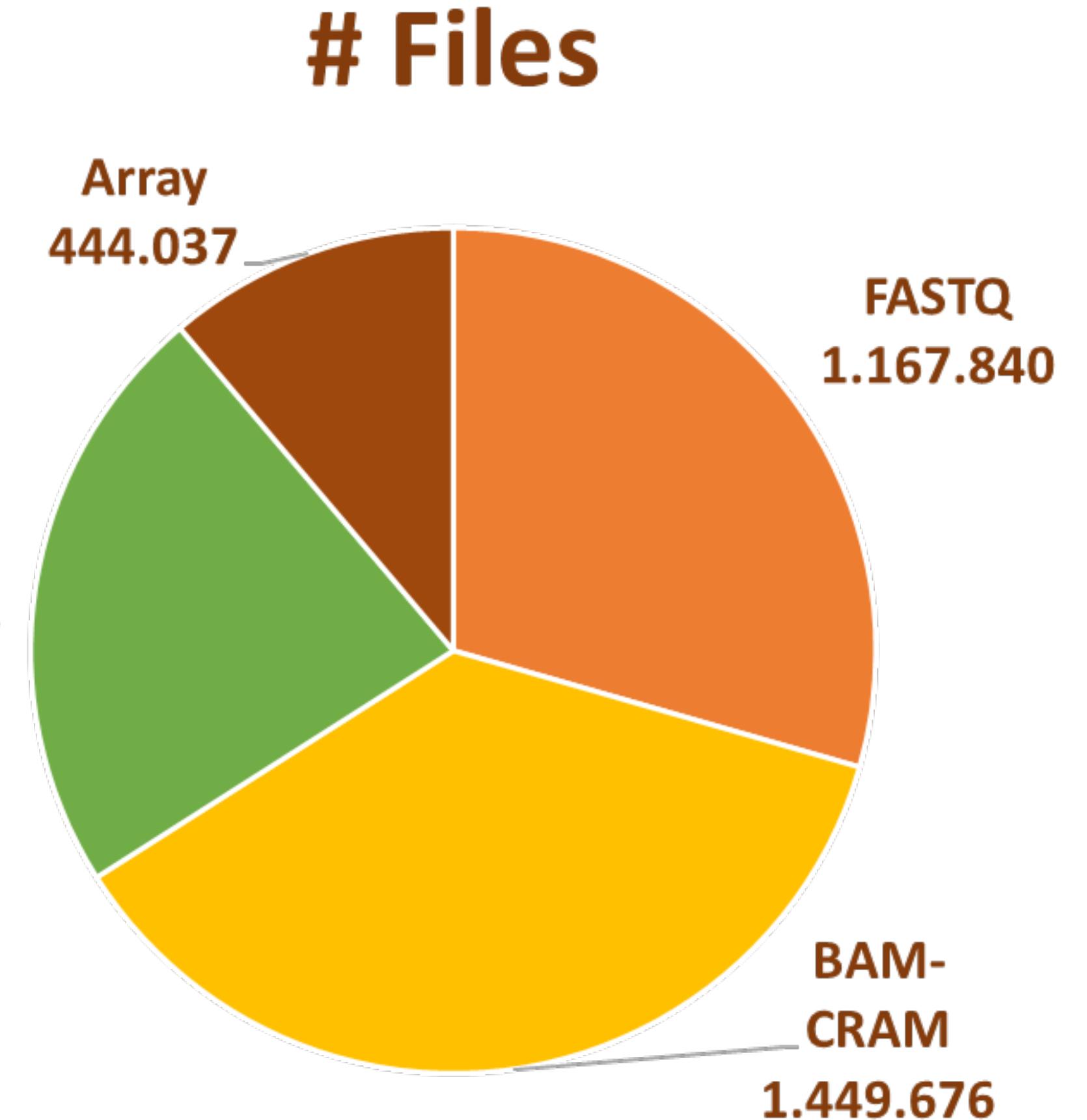
The EGA



- 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 shows 'My DACs - EGAC5000000005 - Requests' and 'HISTORY'. Below this, it says 'EuCanImage DAC' and 'This is a DAC for EuCanImage data'. A search bar says 'Type something for filter the requests...'. It lists three requests from 'Dr Teresa Garcia Lezana':

- 18 August 2022: Requester gemma.milla@crg.eu, Dataset EGAD5000000032, DAC Admin/Member Dr Lauren A Fromont
- 17 August 2022: Requester Dr Teresa Garcia Lezana, Dataset EGAD5000000033, DAC Admin/Member Dr Teresa Garcia Lezana (with a 'revoke permission' button)
- 16 August 2022: Requester Dr Teresa Garcia Lezana, Dataset EGAD5000000032, DAC Admin/Member Dr Lauren A Fromont (with a 'revoke permission' button)

An 'EDIT' button is at the top right, and an 'APPLY' button is at the bottom right of the request list.

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

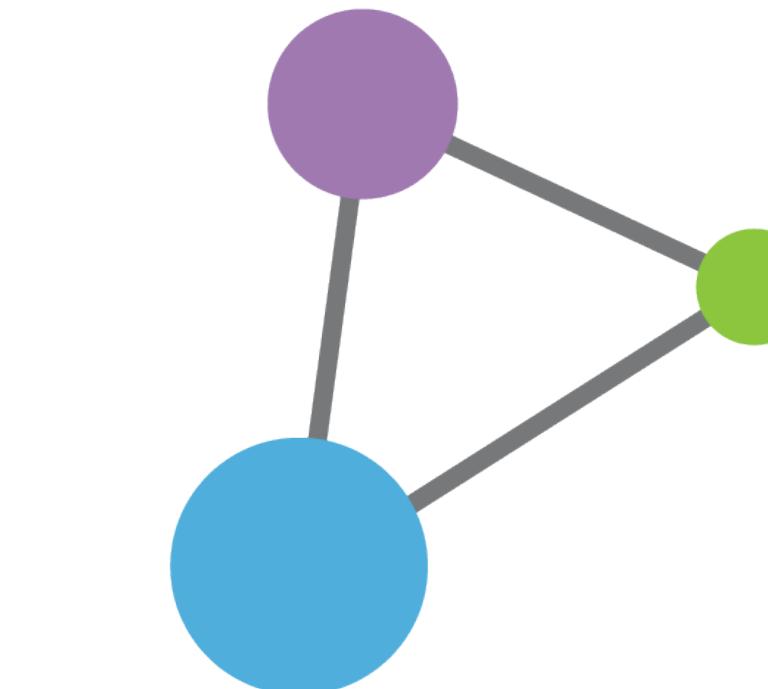
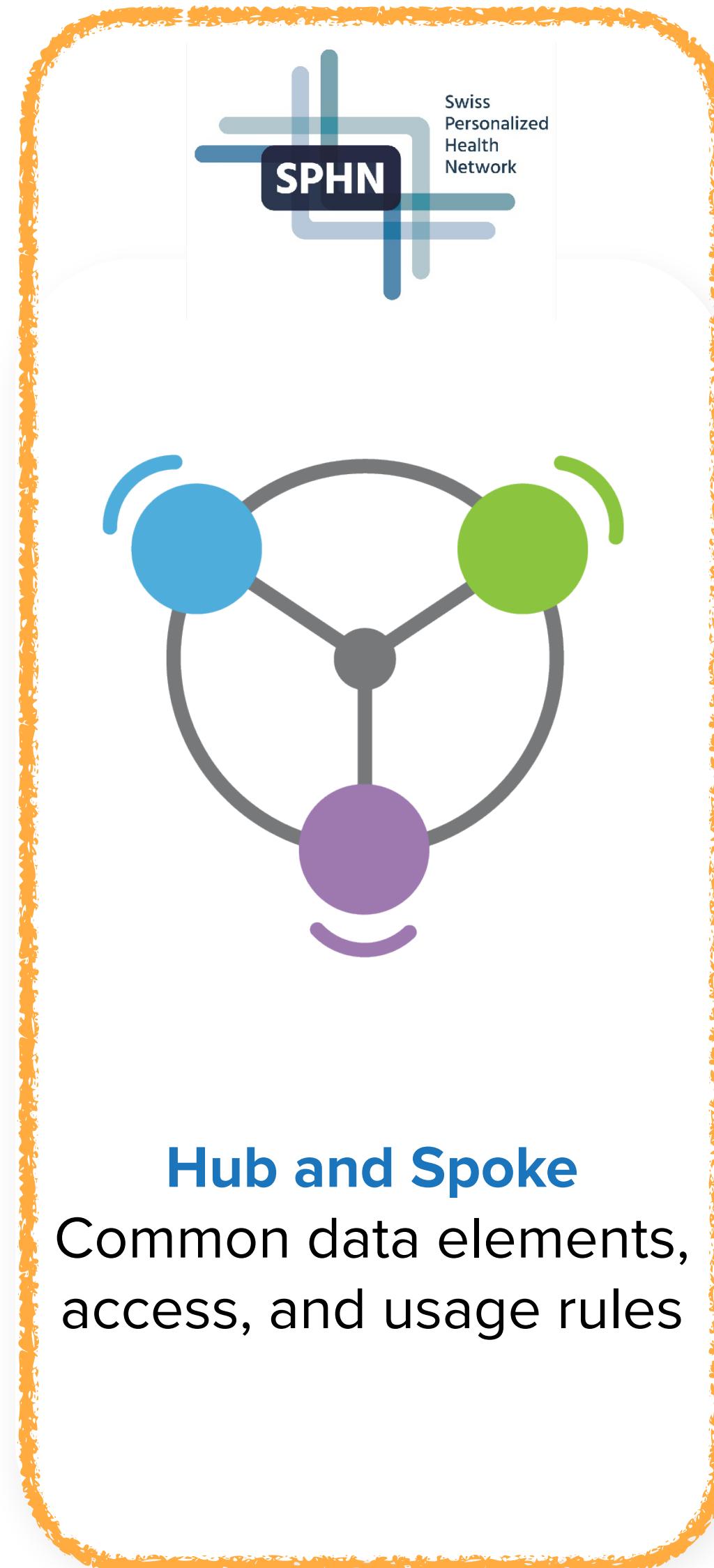
Different Approaches to Data Sharing



Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets



Linkage of distributed and disparate datasets

Different Approaches to Data Sharing



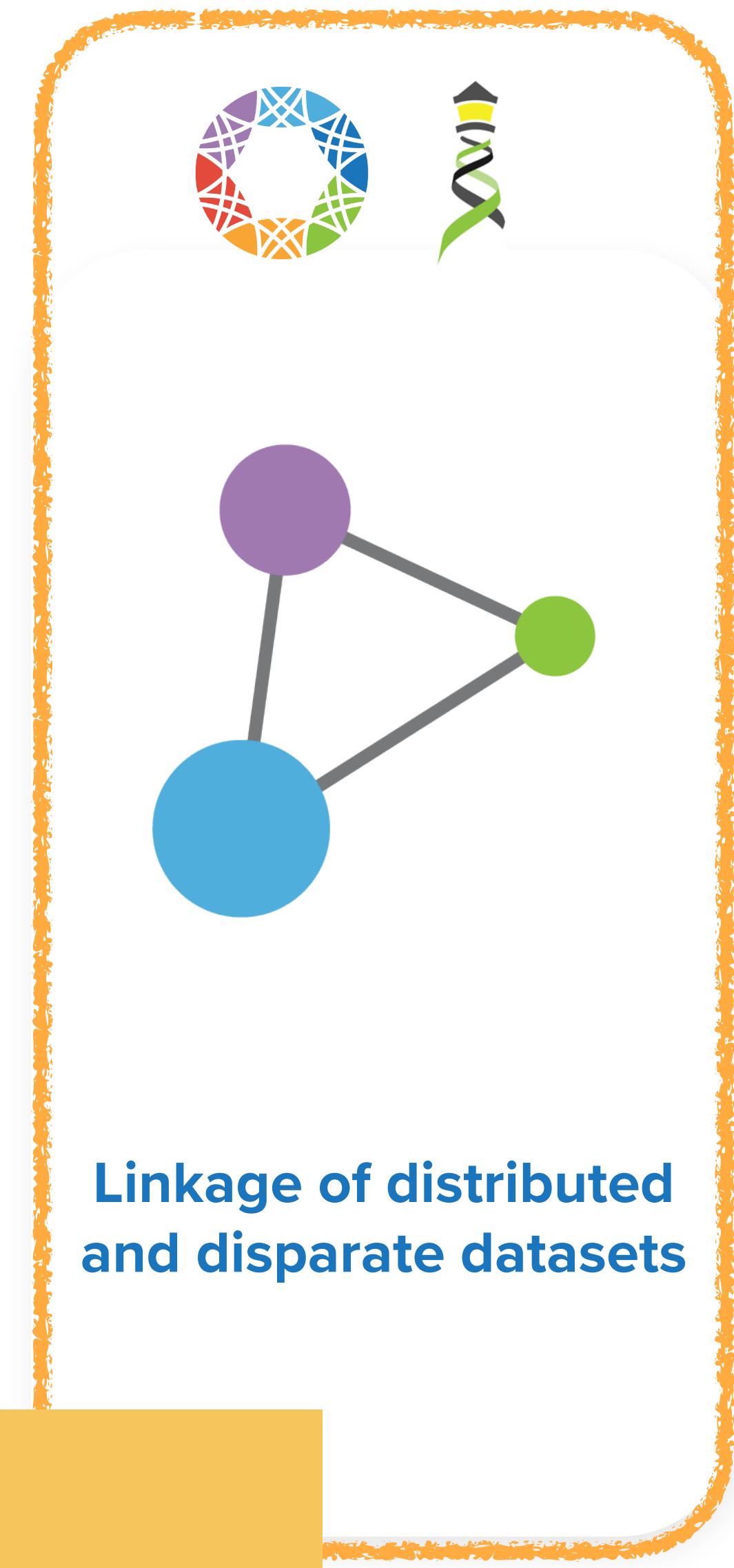
Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets



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



Linkage of distributed and disparate datasets

Federation



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

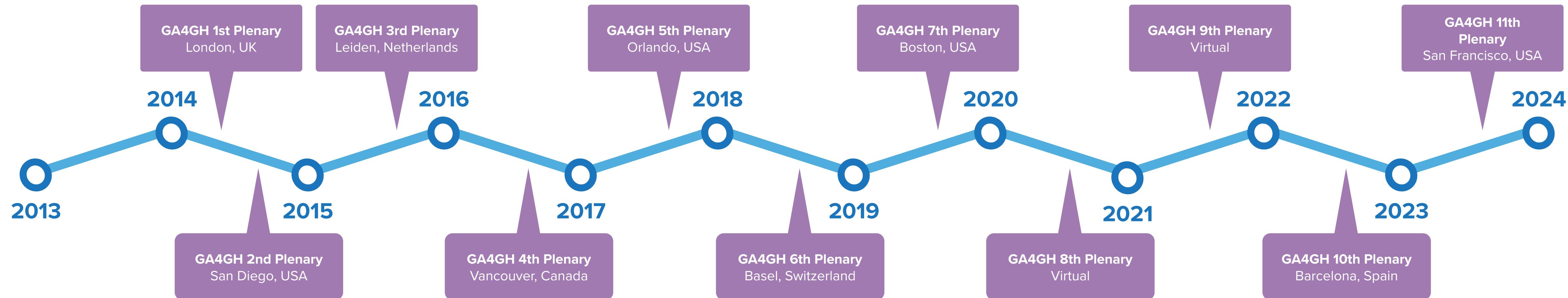
The Global Alliance for Genomics
and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

GA4GH timeline



Global Alliance
for Genomics & Health



Pre-launch	Building momentum	GA4GH Connect	Gap analysis	Strategic Refresh
 <p>73 partners sign a letter of intent to form an alliance</p>	 <p>Global Alliance for Genomics & Health Collaborate. Innovate. Accelerate.</p> <p>Formal launch of GA4GH</p> <p>Published <i>Framework for Responsible Sharing of Genomic and Health-Related Data</i></p> <p>Formed four working groups</p> <p>Developed three demonstration projects</p>	 <p>Launch of GA4GH Connect and Strategic Roadmap</p> <p>Formation of new organizational structure consisting of eight Work Streams and over twenty Driver Projects</p>	<p>Gap analysis identifies three community imperatives</p> <ul style="list-style-type: none"> Interoperability and alignment Implementation support Engaging with healthcare and clinical standards	 <p>Strategic refresh introduces updates to GA4GH to better meet the three community imperatives</p>

Our funders, partners, and Driver Projects

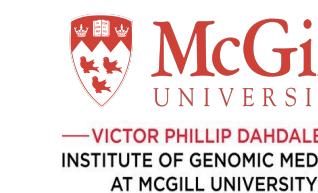


Global Alliance
for Genomics & Health

Core Funders



Host Institutions



Supporting Funders



Assigned Expert Funders/Employers



Strategic Partner



GDI is funded by the European Commission under the Digital Europe Programme under grant agreement number 101081813 and through co-funding from participating Member States.

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

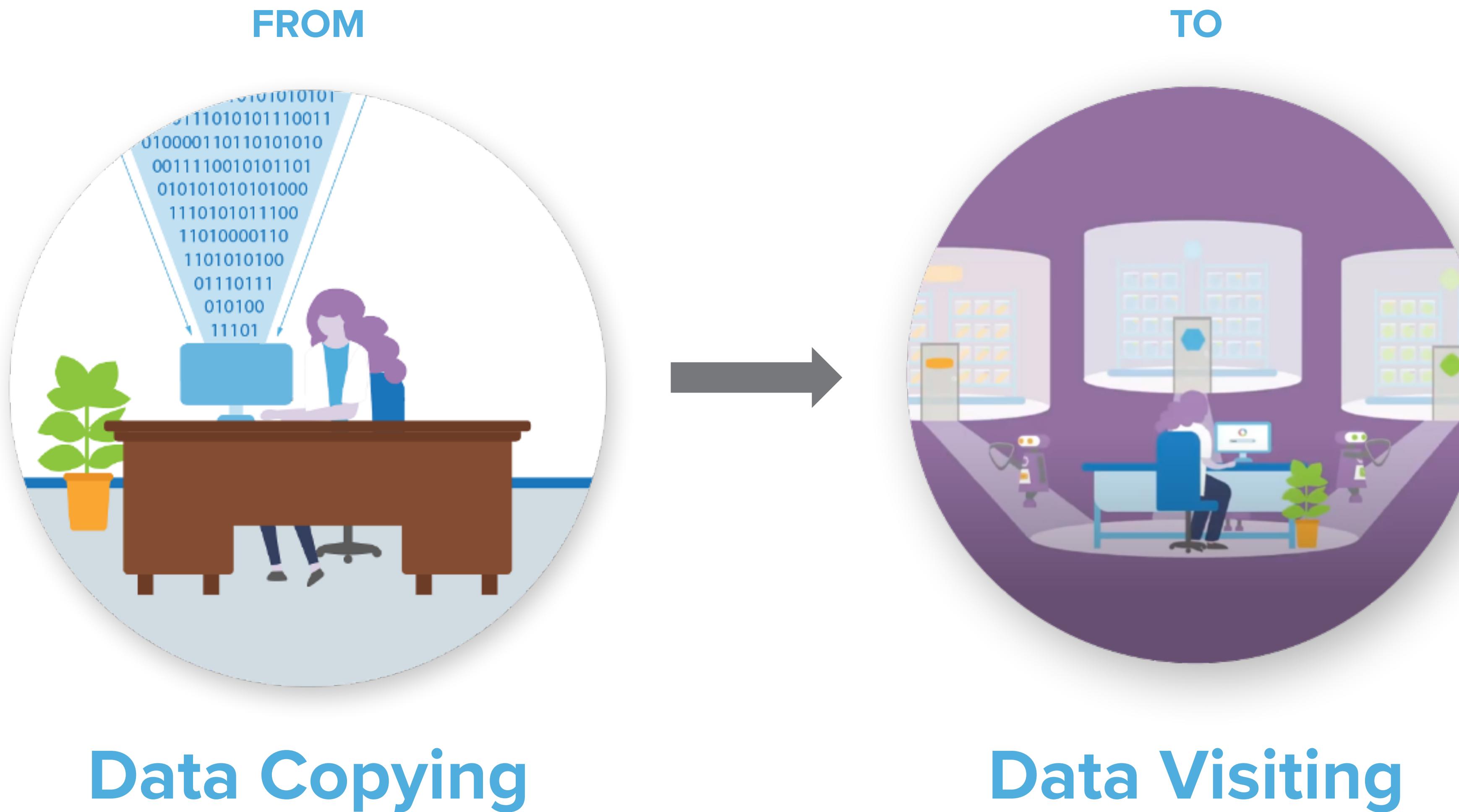
Perspective

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

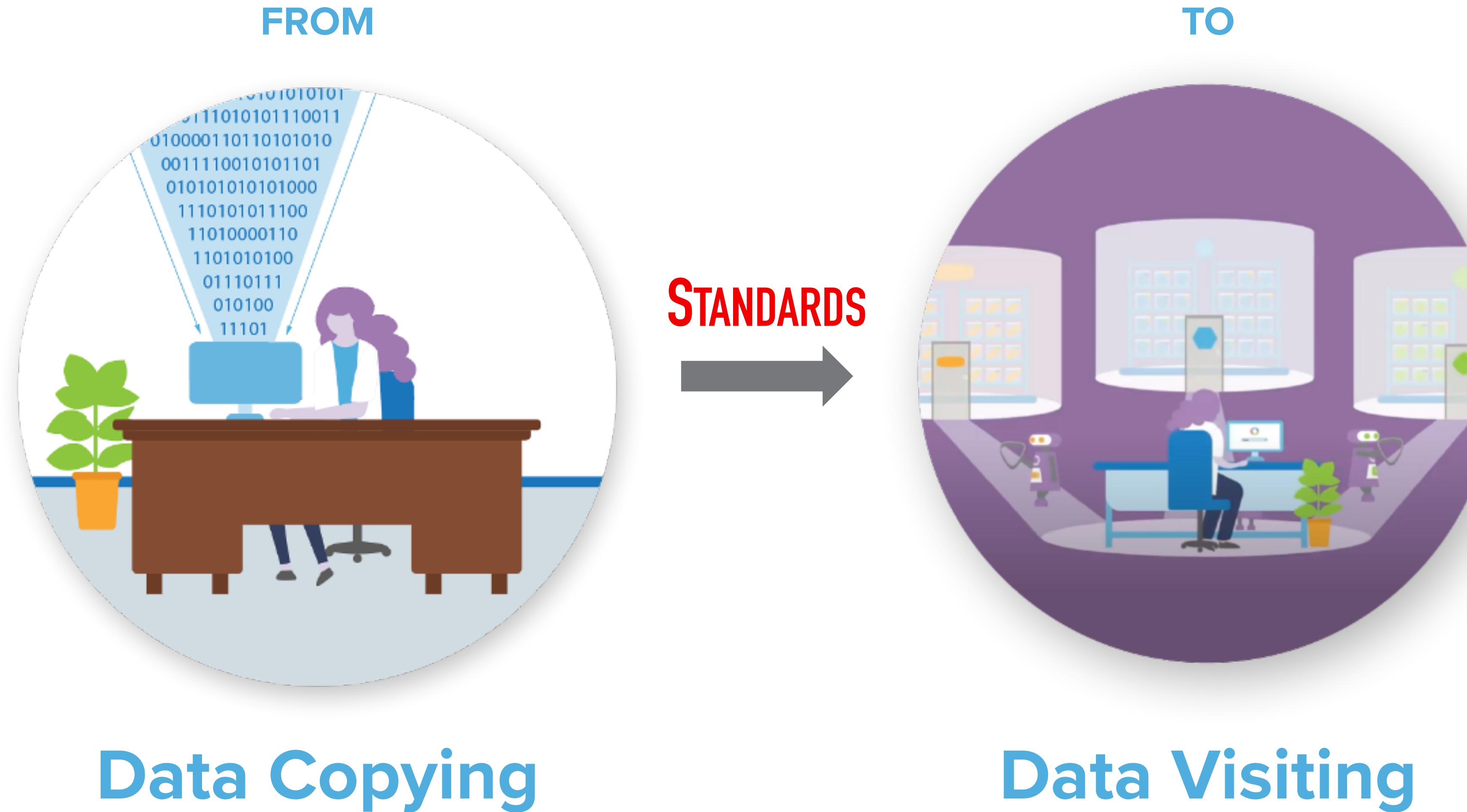
Heidi L. Rehm,^{1,2,47} Angela J.H. Page,^{1,3,*} Lindsay Smith,^{3,4} Jeremy B. Adams,^{3,4} Gil Alterovitz,^{5,47} Lawrence J. Babb,¹ Maxmillian P. Barkley,⁶ Michael Baudis,^{7,8} Michael J.S. Beauvais,^{3,9} Tim Beck,¹⁰ Jacques S. Beckmann,¹¹ Sergi Beltran,^{12,13,14} David Bernick,¹ Alexander Bernier,⁹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

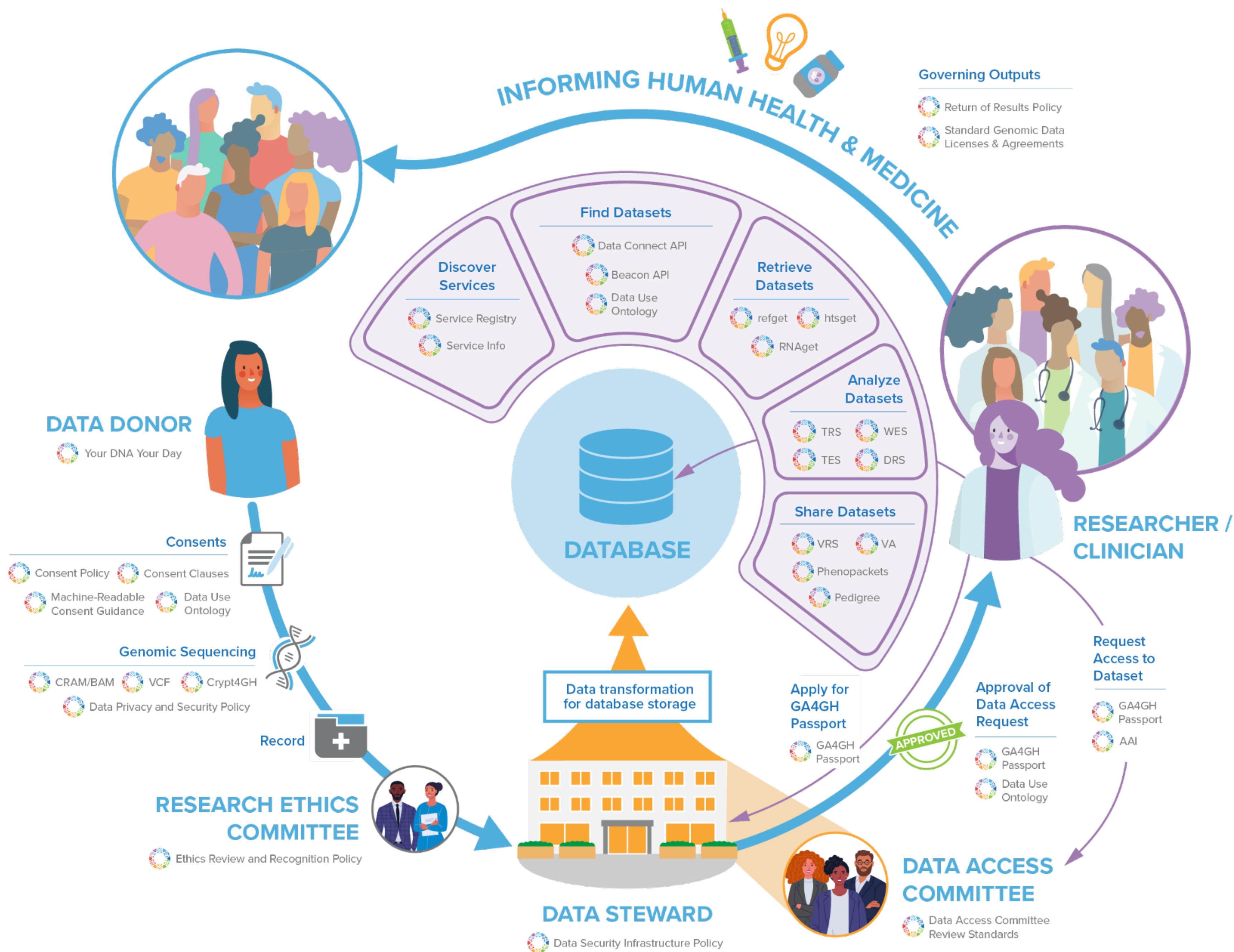
(Author list continued on next page)

A New Paradigm for Data Sharing

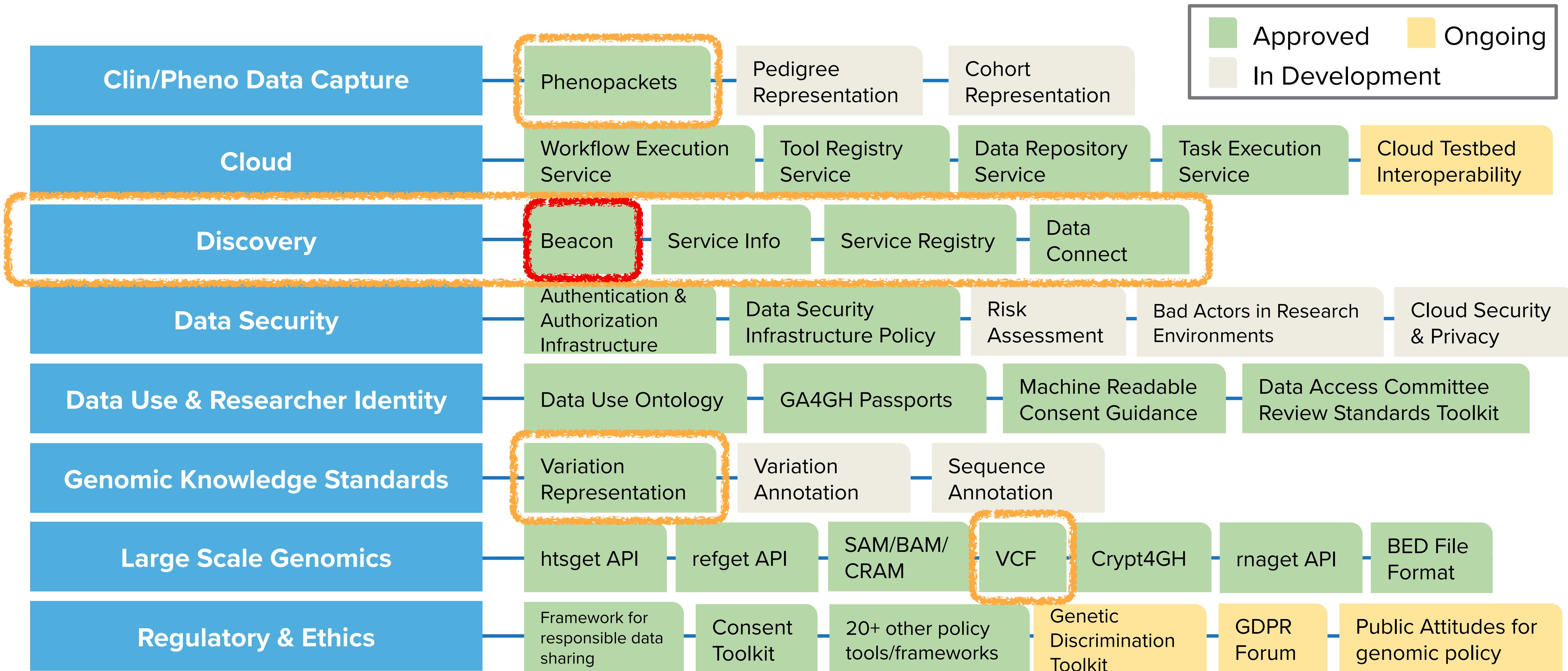


A New Paradigm for Data Sharing



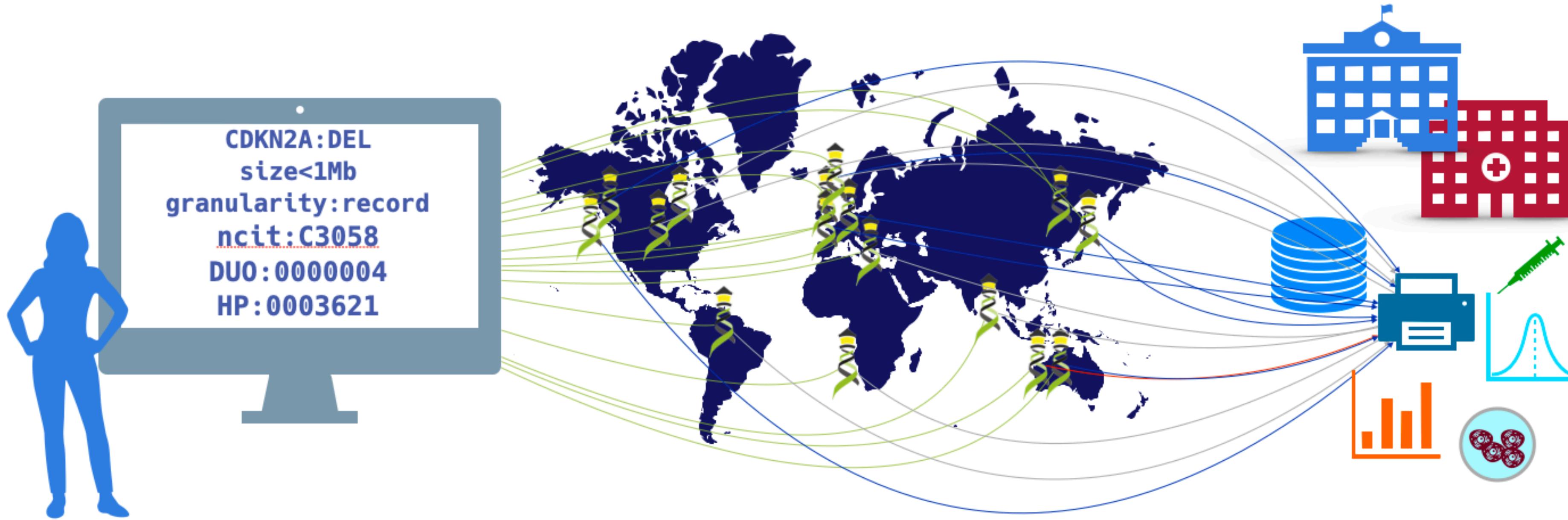


Overview of GA4GH standards and frameworks



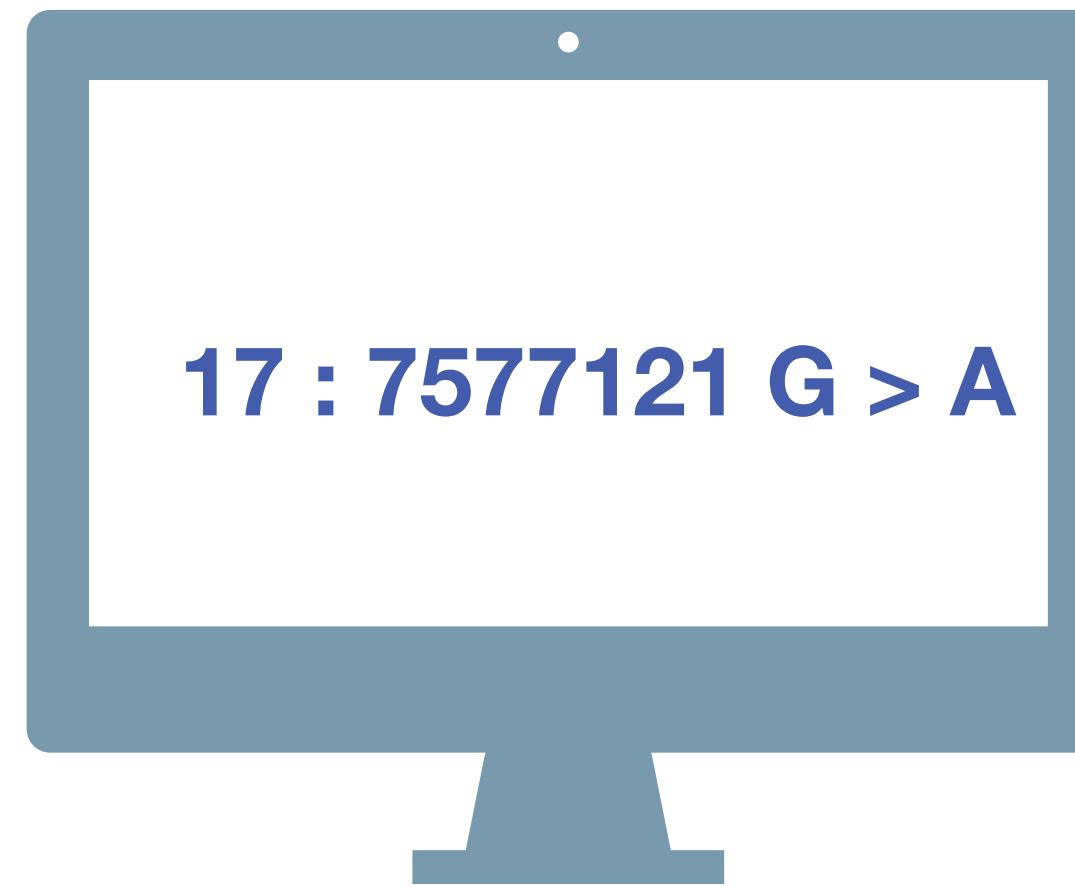


Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.



The GA4GH Beacon Protocol

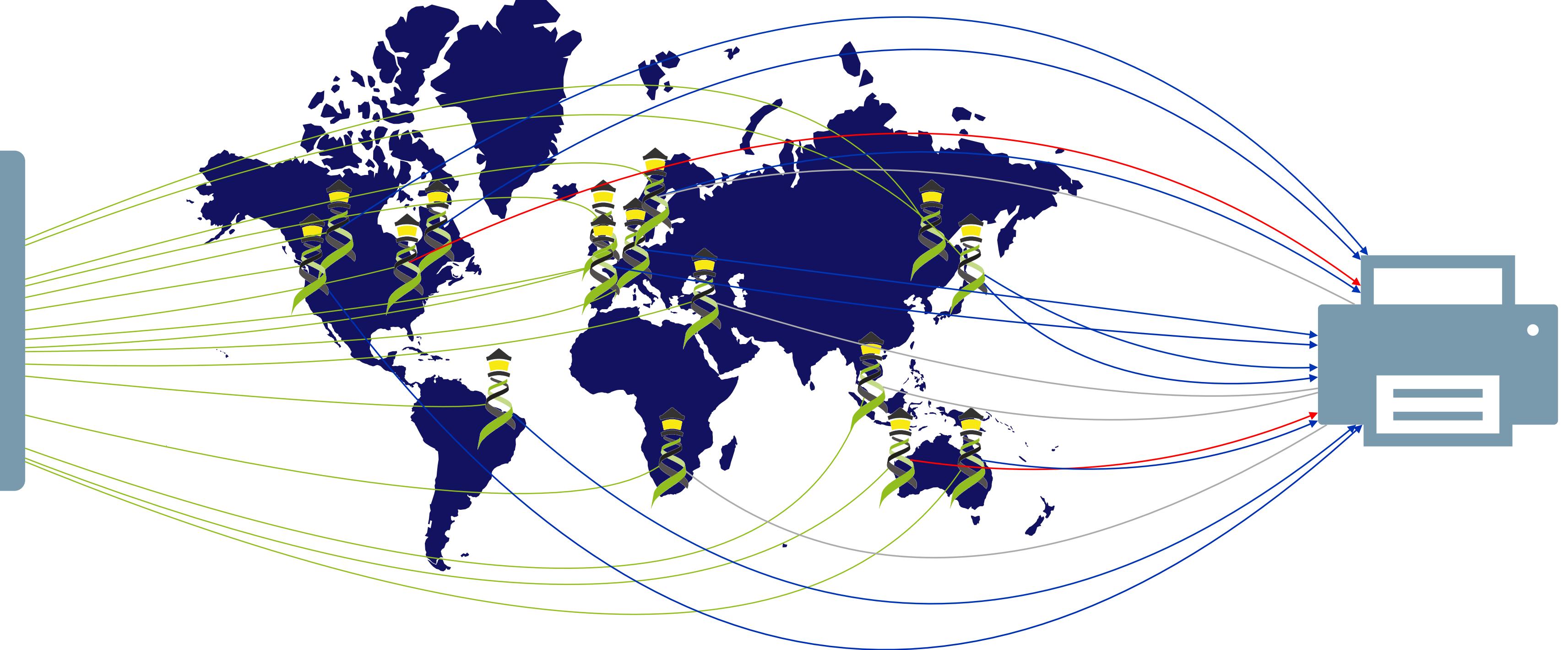
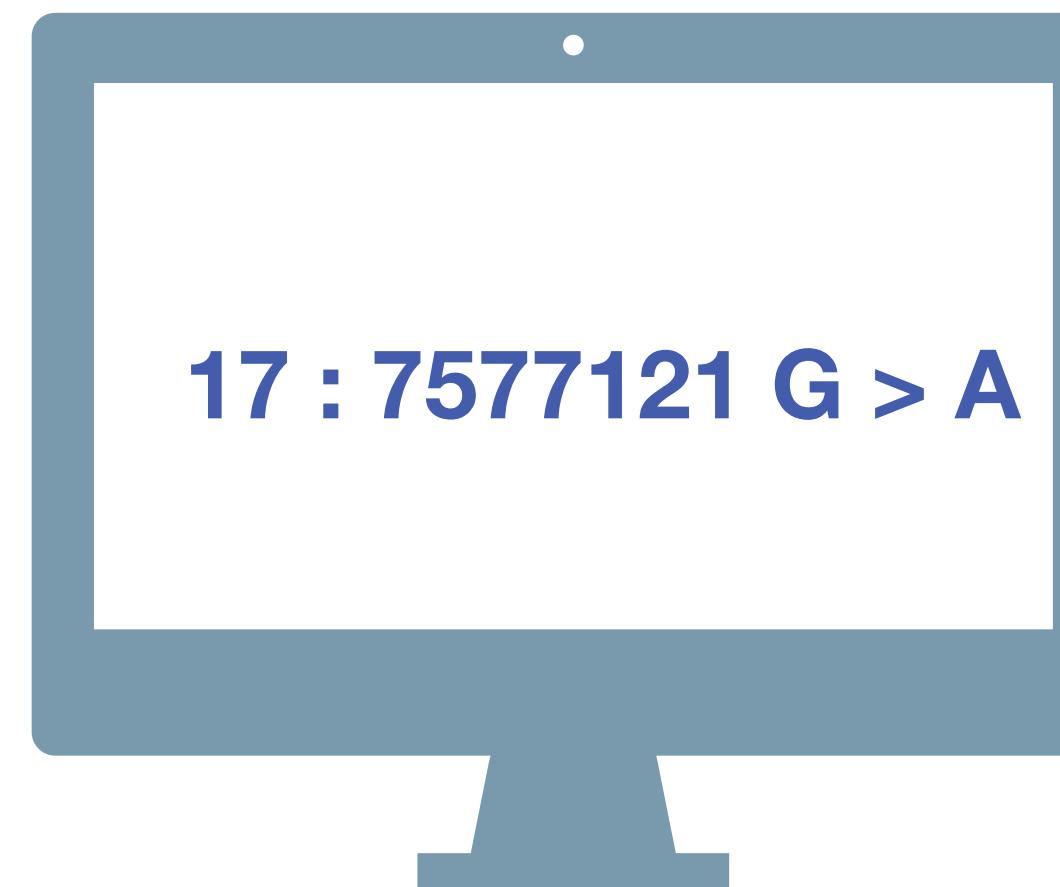
Federating Genomic Discoveries



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 Project in 2016

An open web service that tests the willingness of international sites to share genetic data.



Beacon Network

Search Beacons

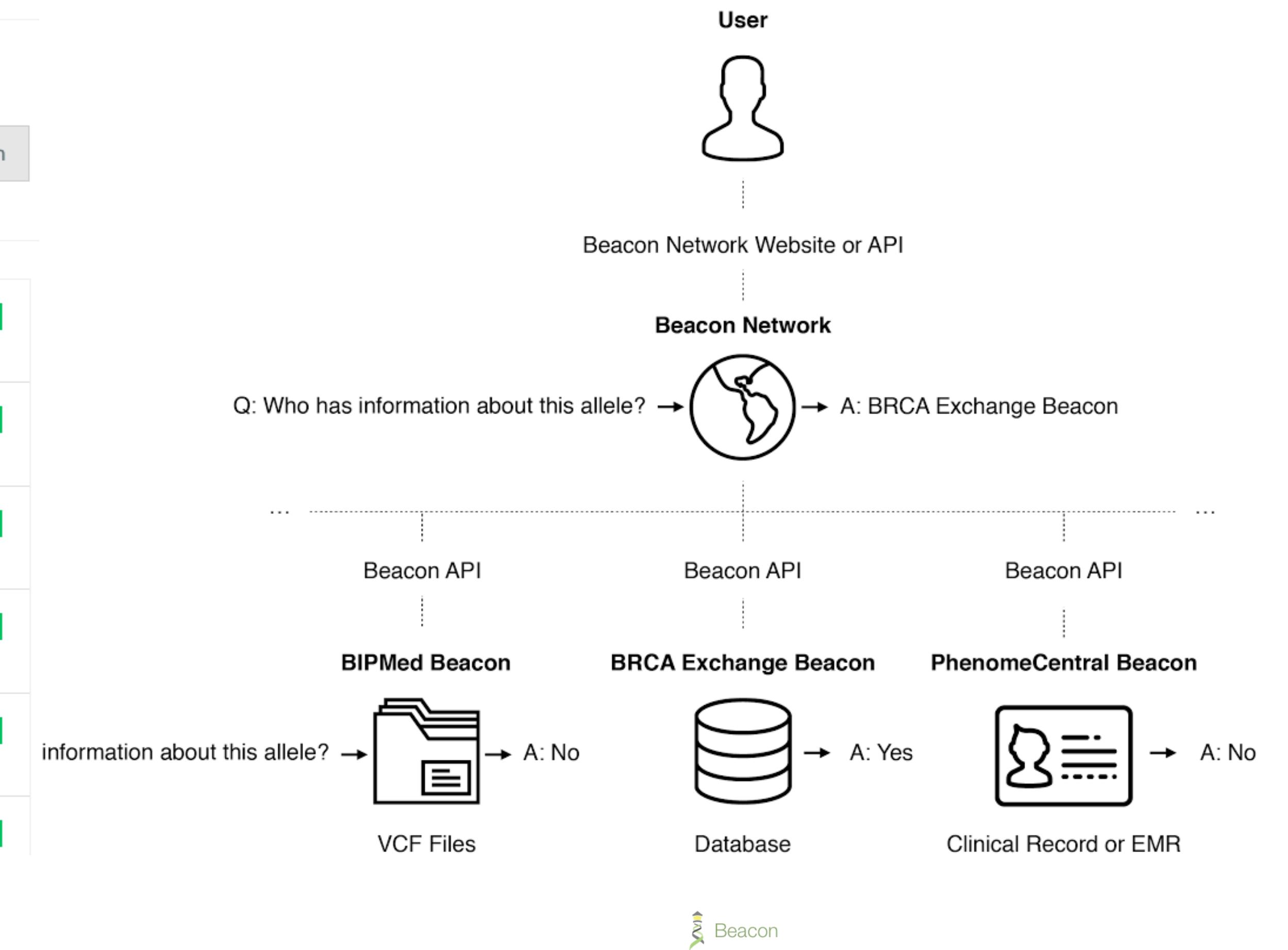
Search all beacons for allele

GRCh37 ▾ 10:118969015 C / CT Search

Response All None
 Found 16
 Not Found 27
 Not Applicable 22

Organization All None
 AMPLab, UC Berkeley
 BGI
 BioReference Laborato...
 Brazilian Initiative on ...
 BRCA Exchange
 Broad Institute
 Centre for Genomic R...
 Centro Nacional de A...
 Curoverse
 EMBL European Bio...
 Global Alliance for G...
 Google
 Institute for Systems ...
 Instituto Nacional de ...

BioReference	Hosted by BioReference Laboratories	Found
Catalogue of Somatic Mutations in Cancer	Hosted by Wellcome Trust Sanger Institute	Found
Cell Lines	Hosted by Wellcome Trust Sanger Institute	Found
Conglomerate	Hosted by Global Alliance for Genomics and Health	Found
COSMIC	Hosted by Wellcome Trust Sanger Institute	Found
dbGaP: Combined GRU Catalog and NHLBI Exome Seq...		Found



35+ Organizations 90+ Beacons 200+ Datasets

100K+ Releases

Date	Tag	Title
2018-01-24	v0.4.0	Beacon
2016-05-31	v0.3.0	Beacon

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

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)

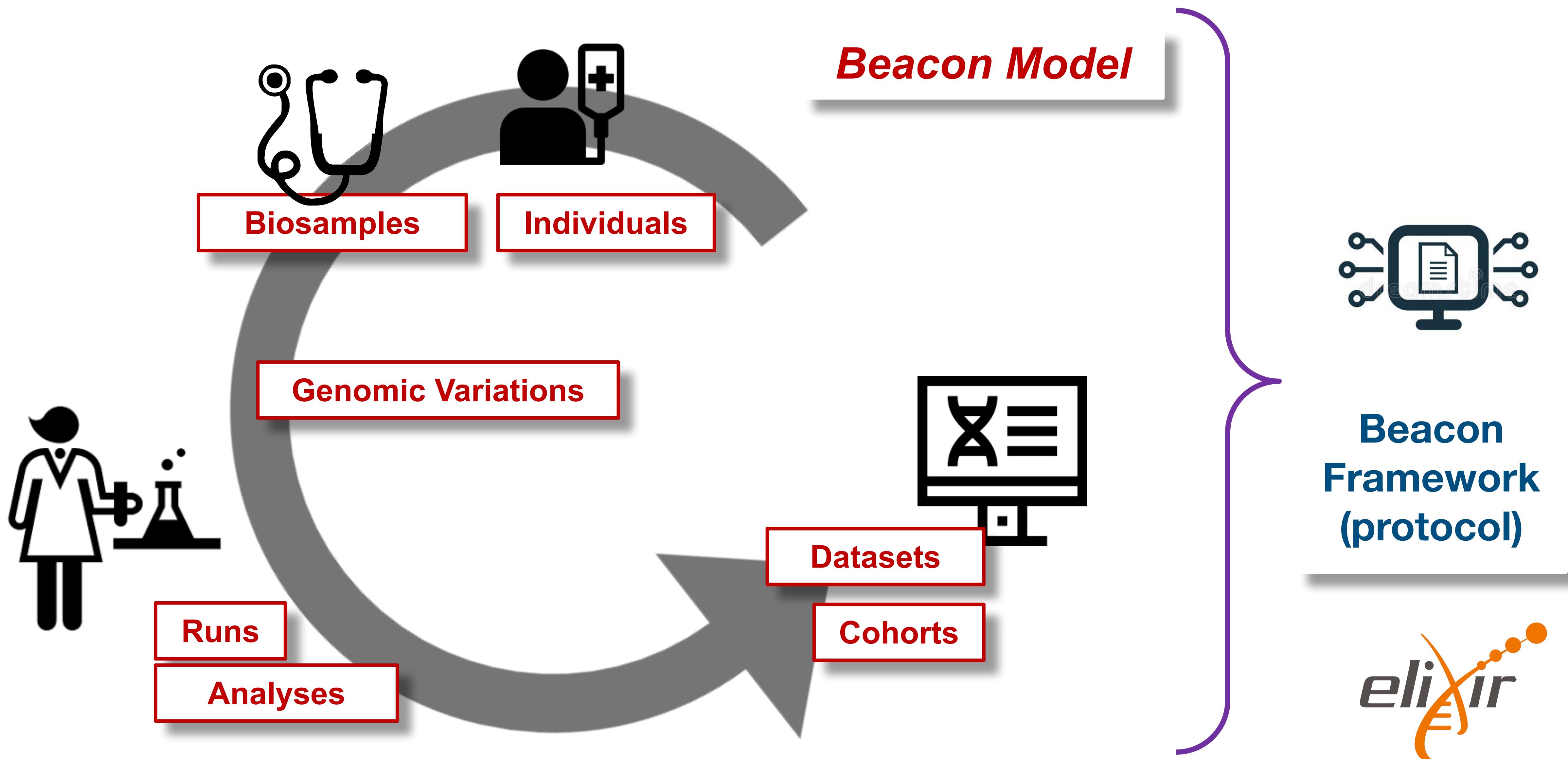
- Beacon publication at Nature Biotechnology

- Phenopackets v2 approved

- docs.genomebeacons.org

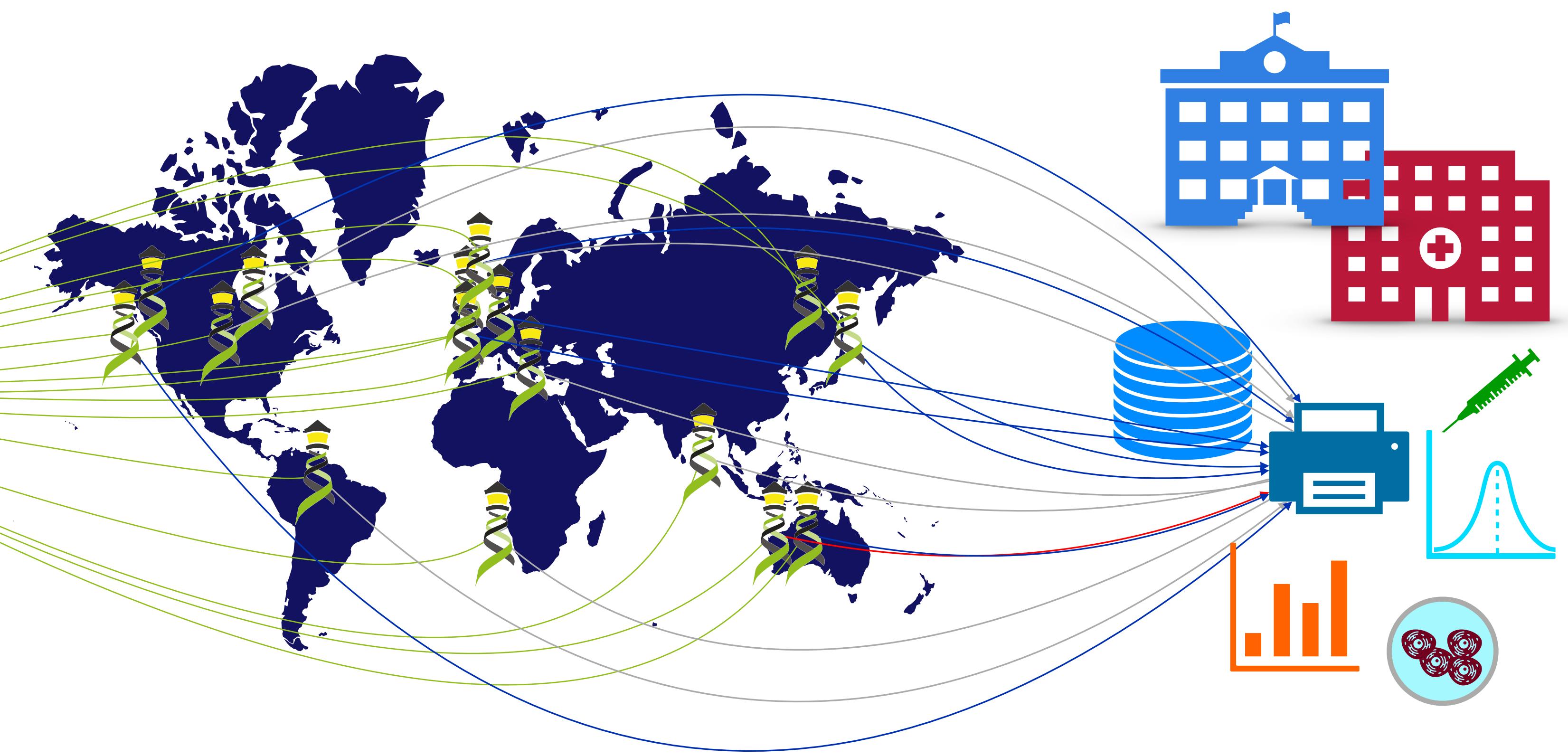
Beacon v2

docs.genomebeacons.org

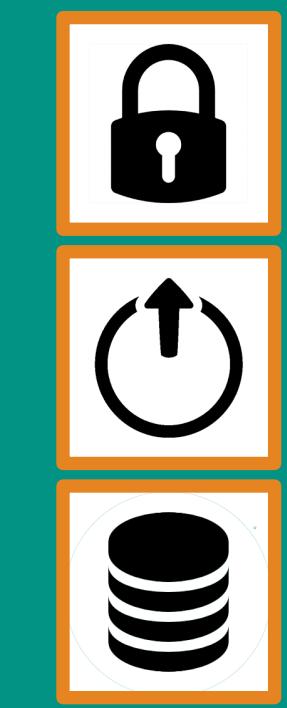


**Beacon
Framework
(protocol)**





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



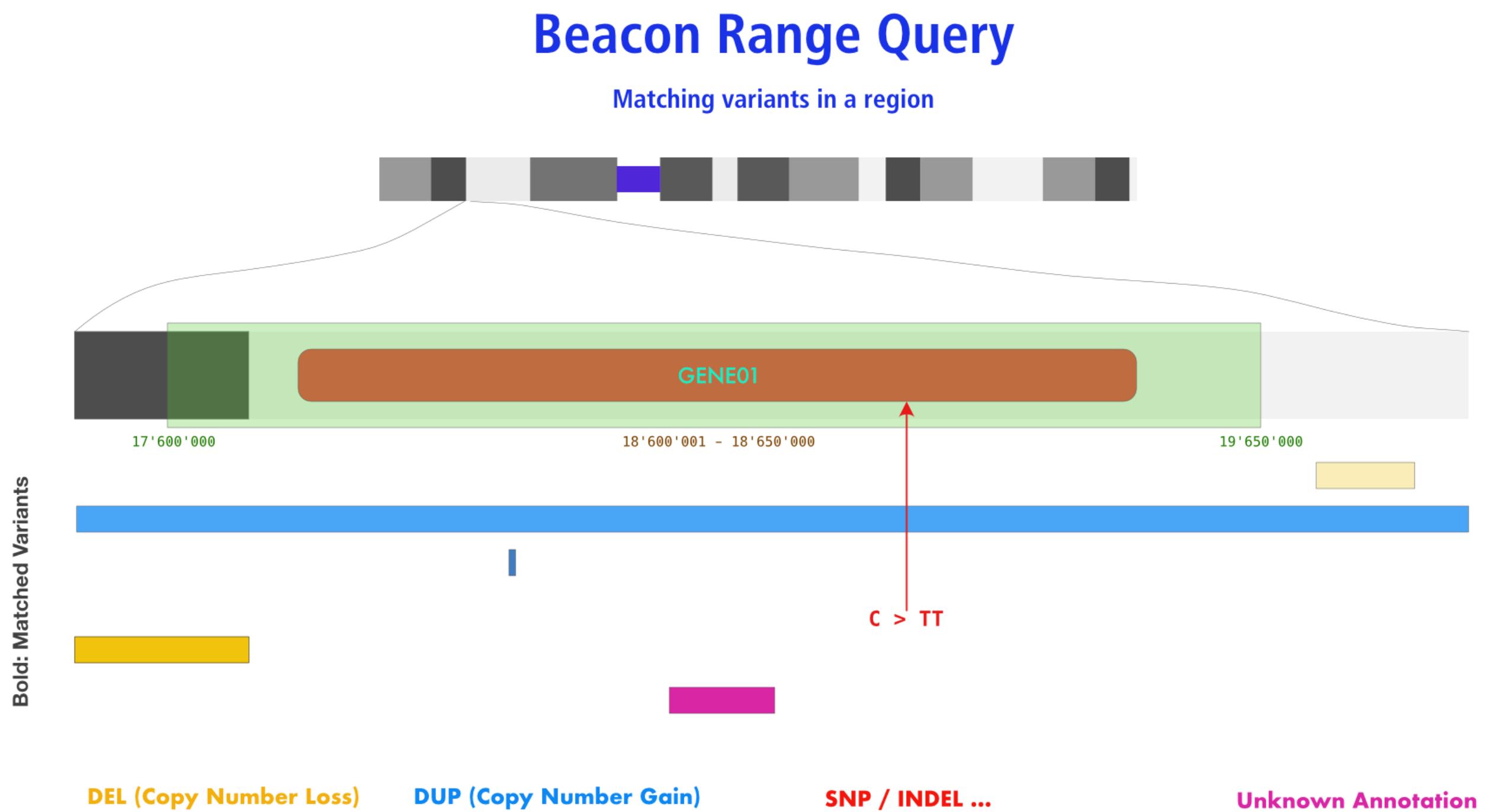
Beacon v2 API

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

Variation 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

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: Chromosome 17

Query Database

Form Utilities: Gene Spans, Cytoband(s)

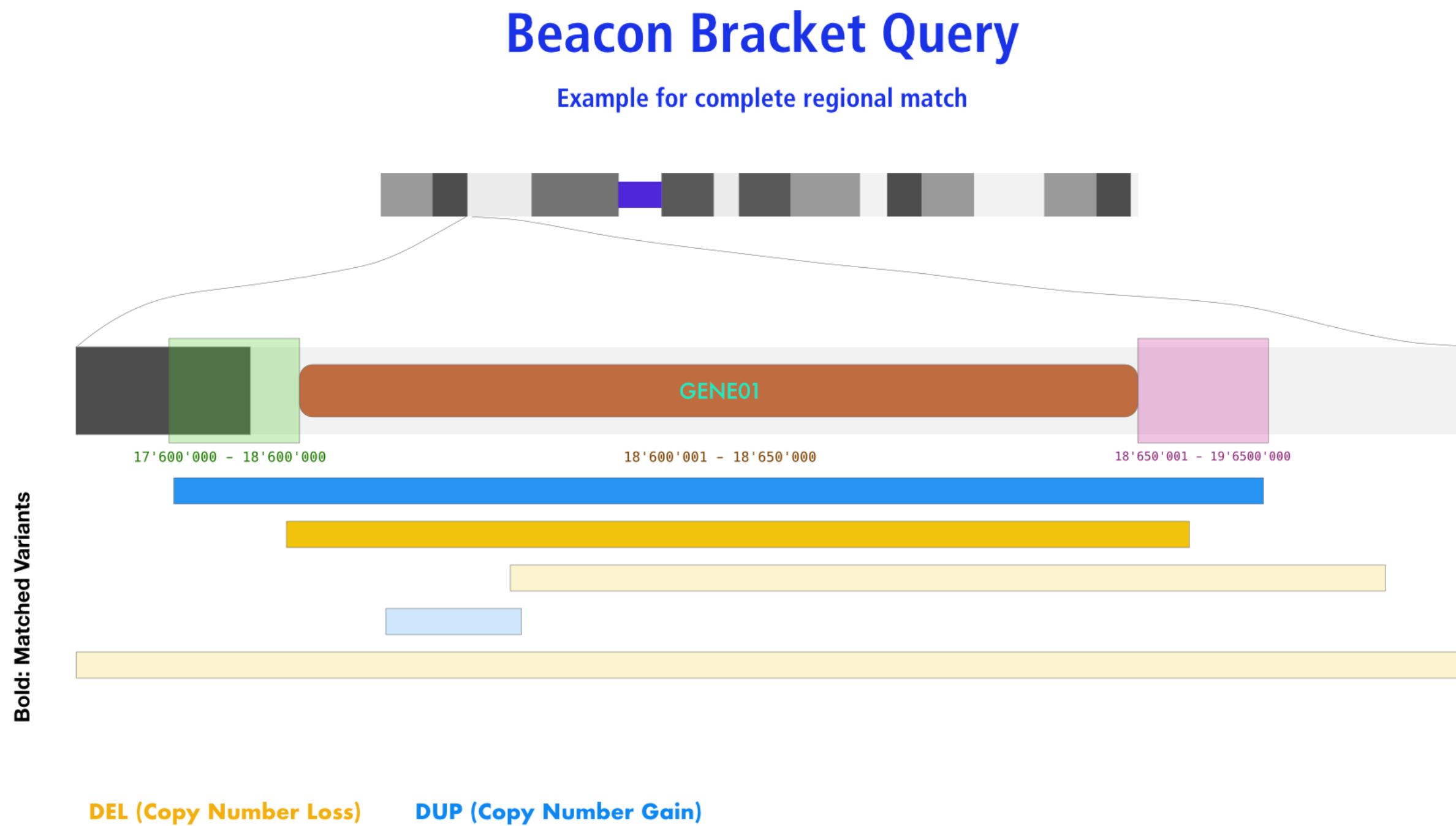
Query Examples: CNV Example, SNV Example, Range Example, Gene Match, Aminoacid Example, Identifier - HeLa

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.

Variation 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 i

9 (NC_000009.12) | ▼

Variant Type i

EFO:0030067 (copy number deletion) | ▼

Start or Position i

21000001-21975098

End (Range or Structural Var.) i

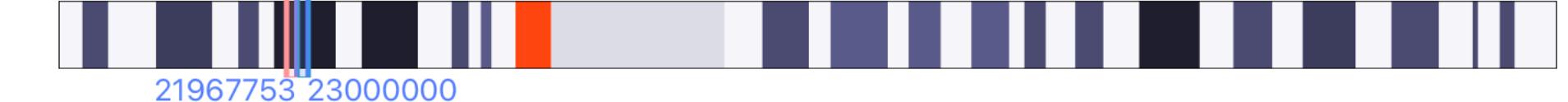
21967753-23000000

Select Filters i

NCIT:C3058: Glioblastoma (100) X | ▼

Chromosome 9 i

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 neoplasm core)

- Beacon v2 relies heavily on "filters"
 - ontology term / CURIE
 - alphanumeric
 - custom
 - 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 falleles: 0 Callsets Variants ↗ UCSC region ↗ Calls: 0 Legacy Interface ↗ Samples: 523 [Show JSON Response](#)

Results Biosamples

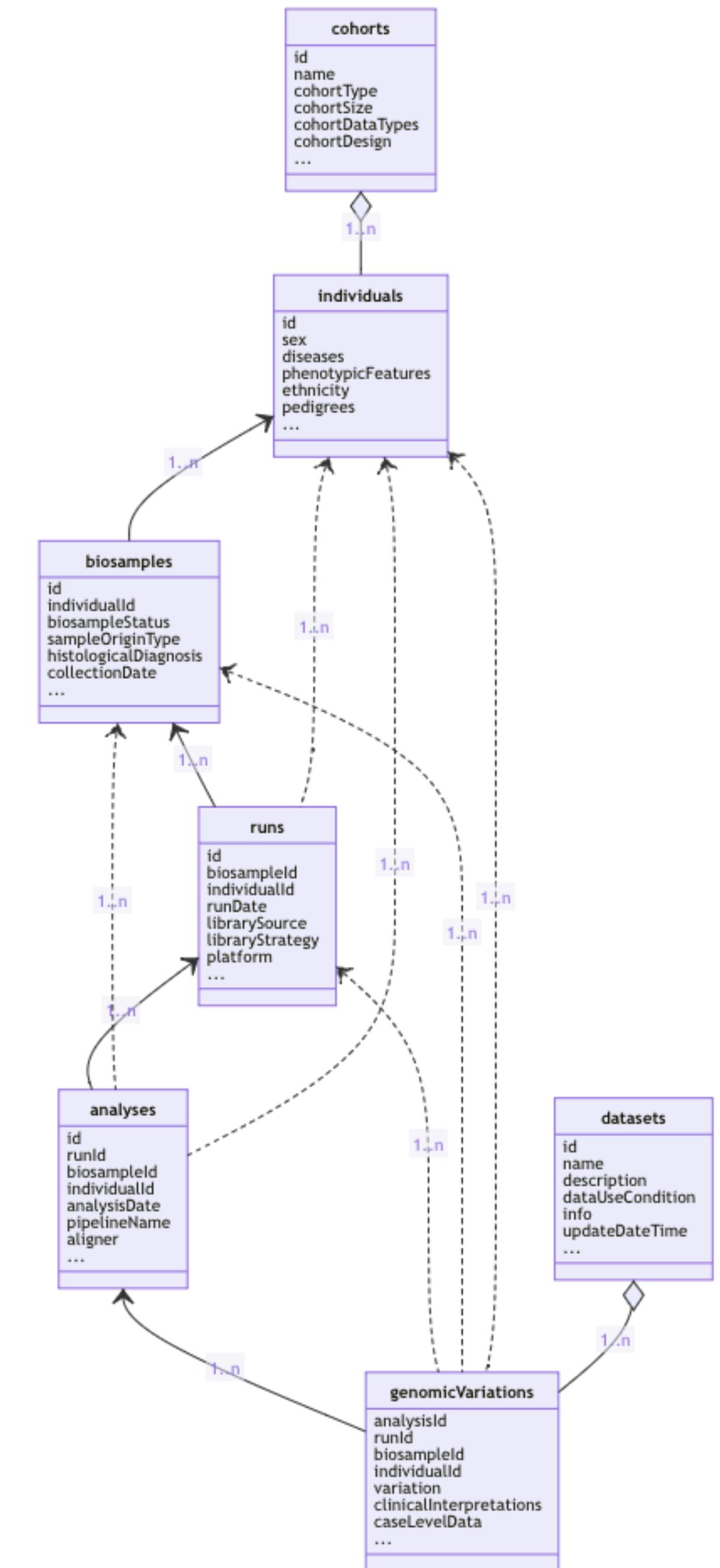
Id	Description	Classifications	Identifiers	DEL	DUP	CNV
PGX_AM_BS_MCC01	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.116	0.104	0.22
PGX_AM_BS_MCC02	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma 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

« < > »

Page 1 of 105

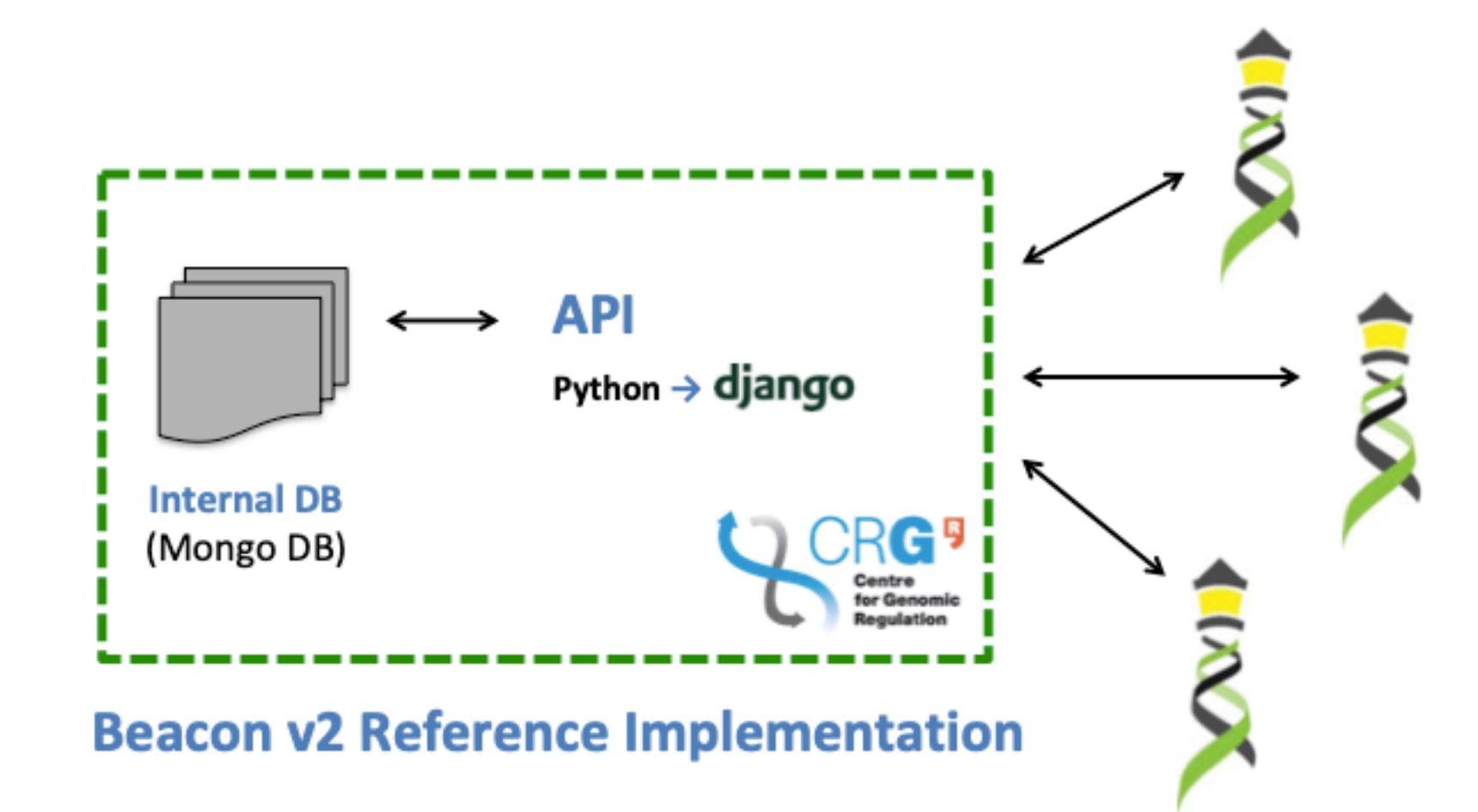
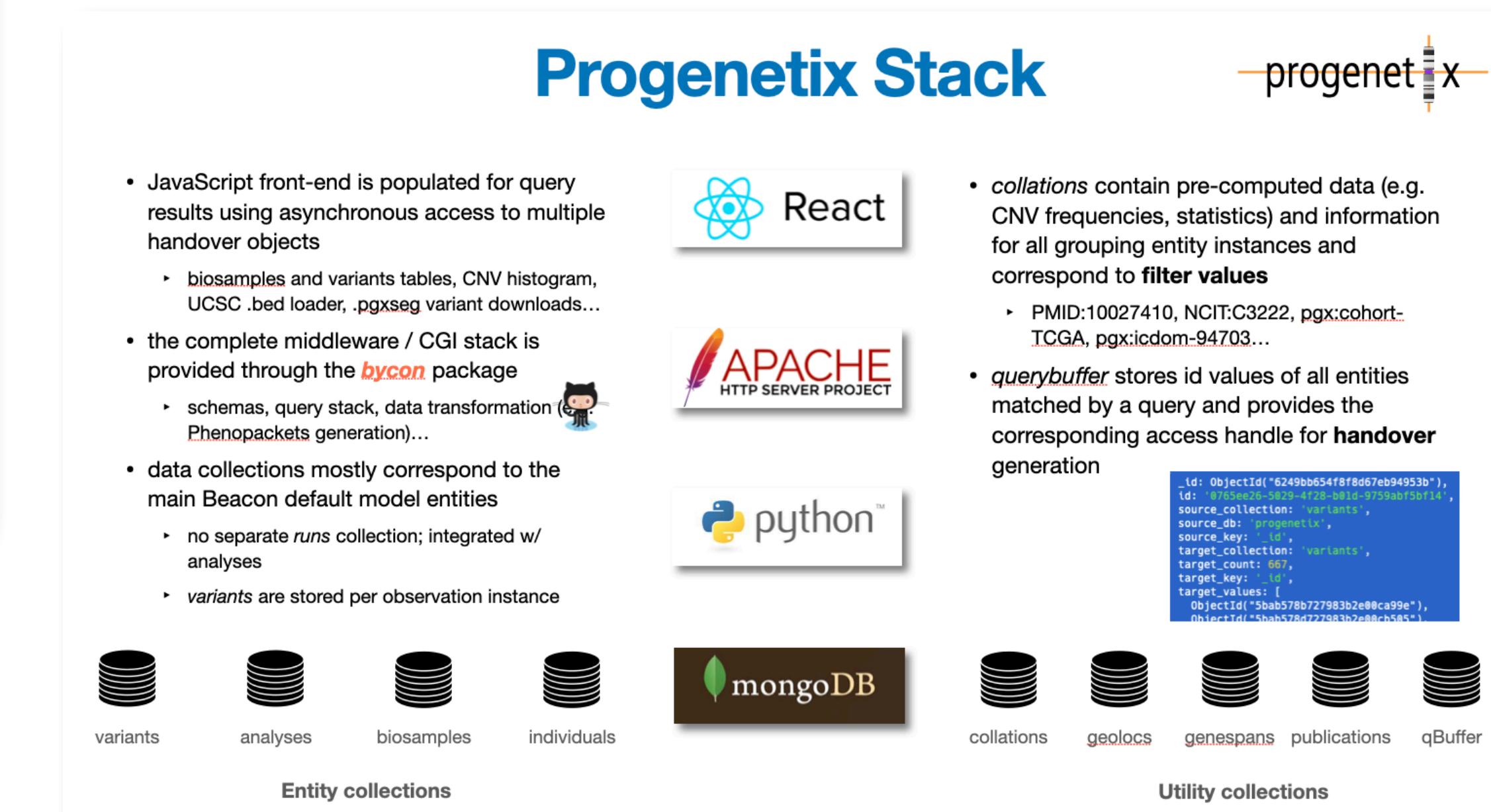
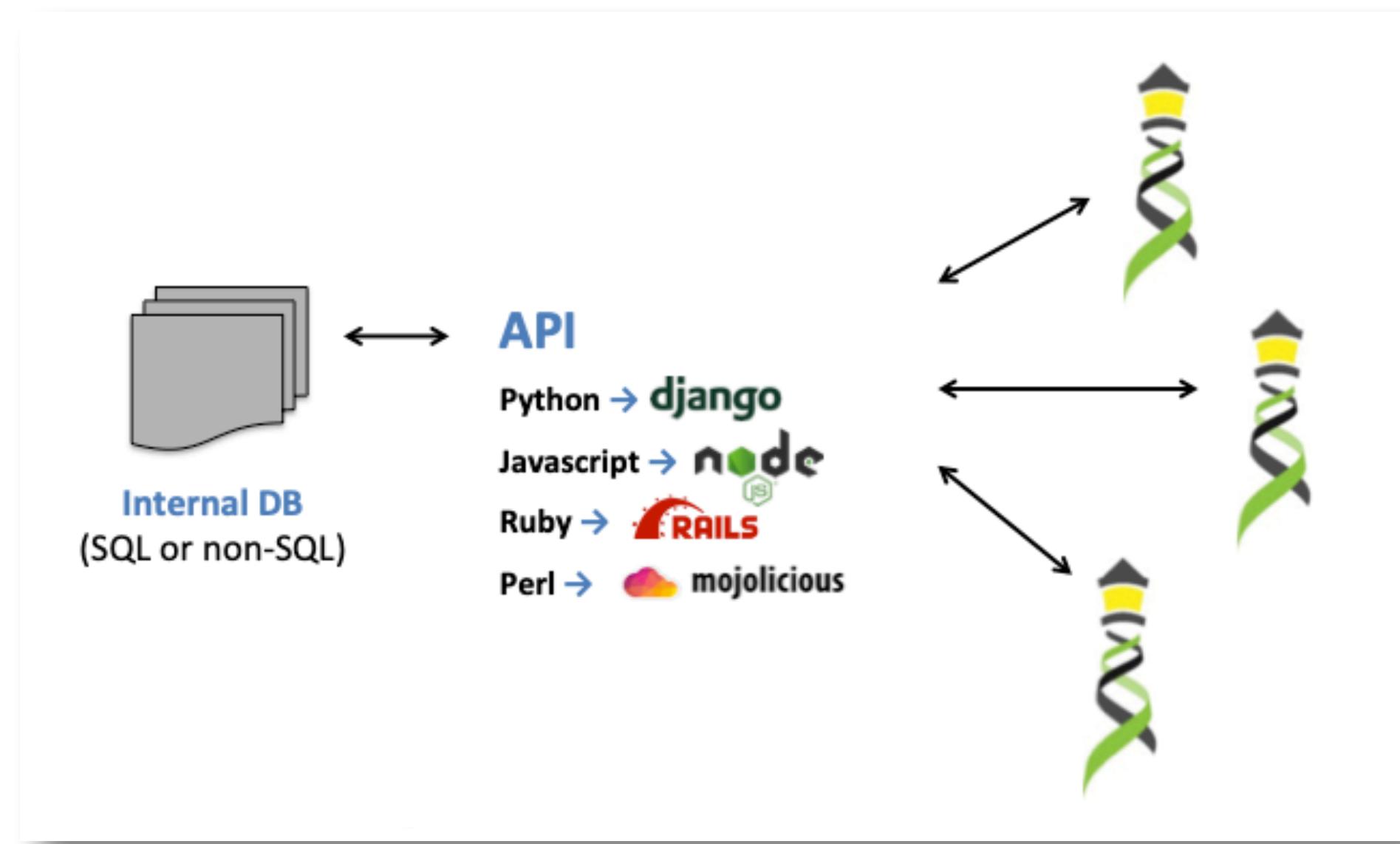
Beacon Default v2 Model

- 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



Implementing Beacon v2

... its just code _(_ツ)_/



bycon for GA4GH Beacon

Implementation driven development of a GA4GH standard

bycon Beacon

Implementation driven standards development

- Progenetix' Beacon+ has served as implementation driver since 2016
- the *bycon* package is used to prototype advanced Beacon features such as
 - structural variant queries
 - data handovers
 - Phenopackets integration
 - variant co-occurrences
 - ...

Beacon protocol response verifier at time of GA4GH approval Spring 2022

Beacon v2 GA4GH Approval Registry

Beacons:    

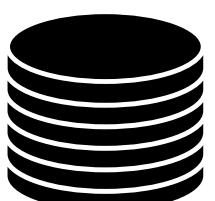
Category	EGA	progenetix	cnag	University of Leicester
BeaconMap	Green	Green	Green	Green
Bioinformatics analysis	Green	Green	Green	Green
Biological Sample	Green	Red	Red	Green
Cohort	Green	Green	Green	Green
Configuration	Green	Green	Green	Green
Dataset	Green	Red	Red	Green
EntryTypes	Green	Green	Green	Green
Genomic Variants	Green	Green	Green	Green
Individual	Green	Red	Red	Green
Info	Green	Red	Red	Green
Sequencing run	Green	Green	Green	Green

Legend:  Matches the Spec  Not Match the Spec  Not Implemented

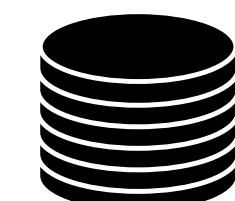
bycon based Progenetix Stack



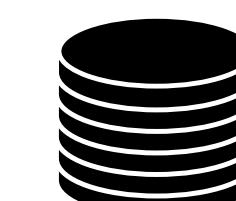
- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
 - ▶ biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads...
- the complete middleware / CGI stack is provided through the *bycon* package
 - ▶ schemas, query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



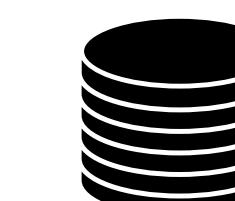
variants



analyses



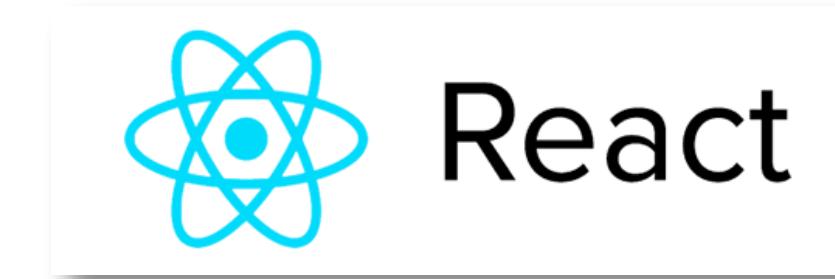
biosamples



individuals

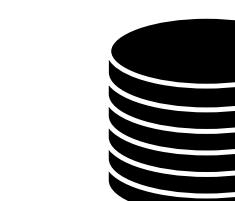


Entity collections

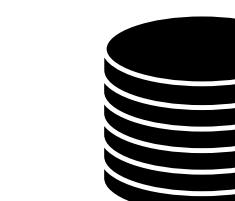


- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
 - ▶ PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703...
- *querybuffer* stores id values of all entities matched by a query and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

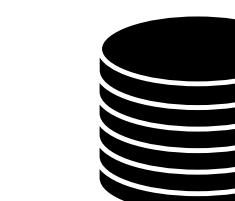
```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
  ObjectId("5bab578d727983b2e00cb505")]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



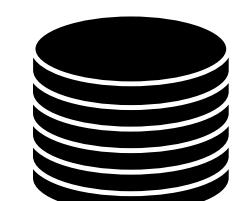
collations



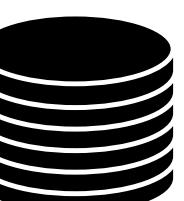
geolocs



genespans



publications



qBuffer

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] 

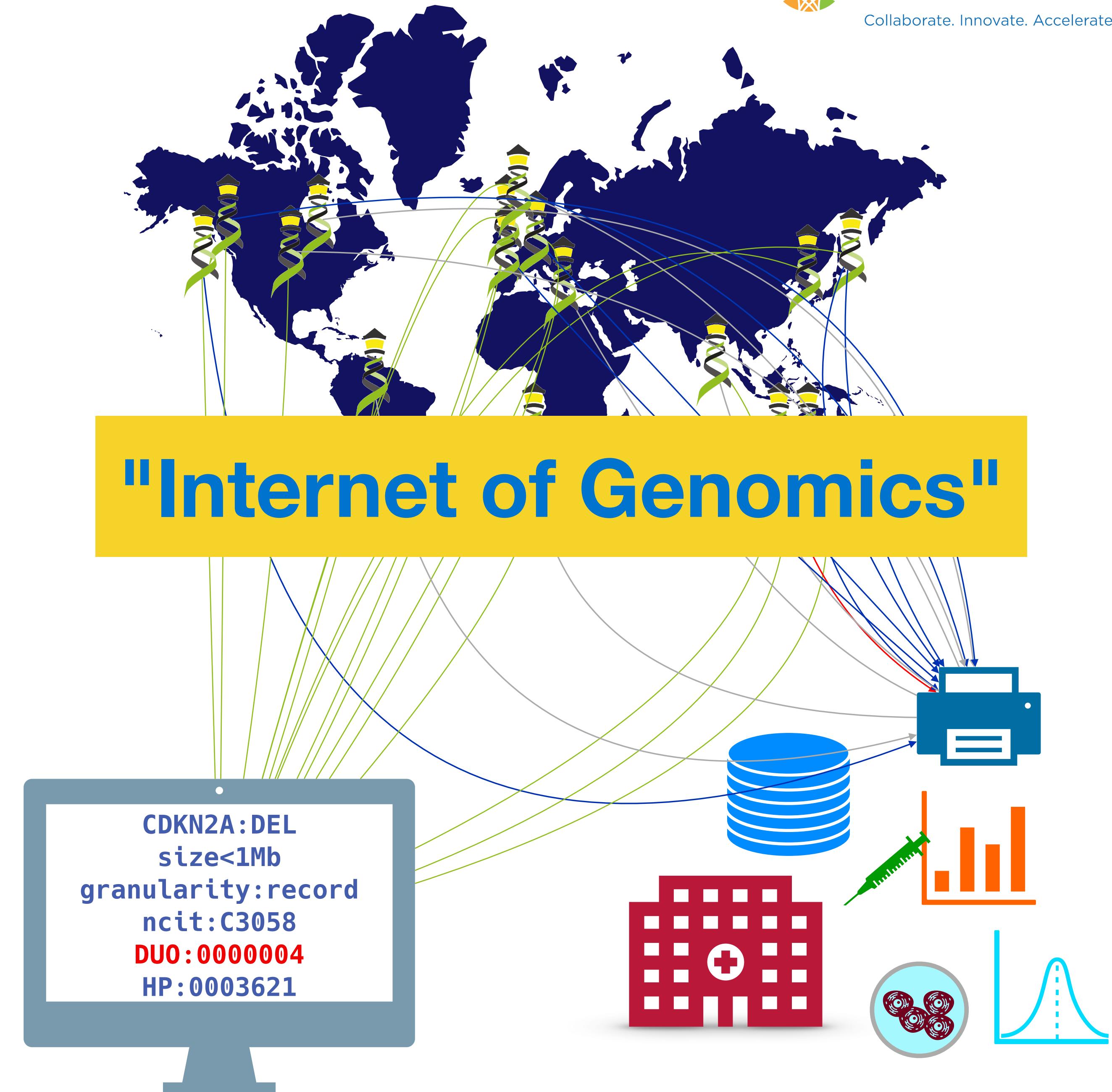
Maintainer: Hangjia Zhao <hangjia.zhao@uzh.ch>

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>.

What Can You Do?

- find a way to make your (patients') **data discoverable** - through adding *at least* the relevant metadata to national or project centric repositories
- use 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 project





University of
Zurich UZH



Swiss Institute of
Bioinformatics

Michael Baudis

Hangjia Zhao

Ziying Yang

Ramon Benitez

Brito

Rahel Paloots

Bo Gao

Qingyao Huang



Jordi Rambla

Arcadi Navarro

Roberto Ariosa

Manuel Rueda

Lauren Fromont

Mauricio Moldes

Claudia Vasallo

Babita Singh

Sabela de la Torre

Fred Haziza



Tony Brookes

Tim Beck

Colin Veal

Tom Shorter



Juha Törnroos

Teemu Kataja

Ikkka Lappalainen

Dylan Spalding



Augusto Rendon

Ignacio Medina

Javier López

Jacobo Coll

Antonio Rueda



centre nacional d'anàlisi genòmica
centro nacional de análisis genómico

Sergi Beltran

Carles Hernandez



Institut national
de la santé et de la recherche médicale

David Salgado



Barcelona Supercomputing Center

Centro Nacional de Supercomputación

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

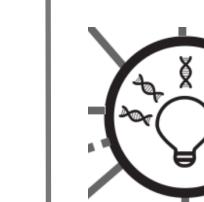


CINECA

Thomas Keane

Melanie Courtot

Jonathan Dursi



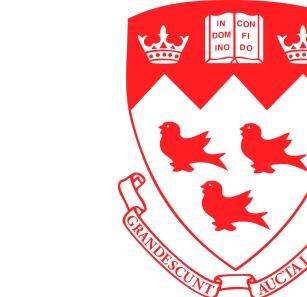
Heidi Rehm

Ben Hutton



Toshiaki

Katayama



Stephane Dyke

DNA STACK

Marc Fiume

Miro Cupak



Melissa Cline



EMBL-EBI



Diana Lemos



VICC Variant Interpretation for Cancer Consortium

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



Universität
Zürich^{UZH}



Swiss Institute of
Bioinformatics

