

Data Discovery in Biomedical Genomics and Cancer Research

Implementing a New Paradigm

Michael Baudis

Professor of Bioinformatics

University of Zürich

Swiss Institute of Bioinformatics **SIB**

Member GA4GH Strategic Leadership Committee

GA4GH Workstream Co-lead *DISCOVERY*

Co-lead ELIXIR Beacon API Development

Co-lead ELIXIR hCNV Community



Universität
Zürich^{UZH}



SIB
Swiss Institute of
Bioinformatics



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

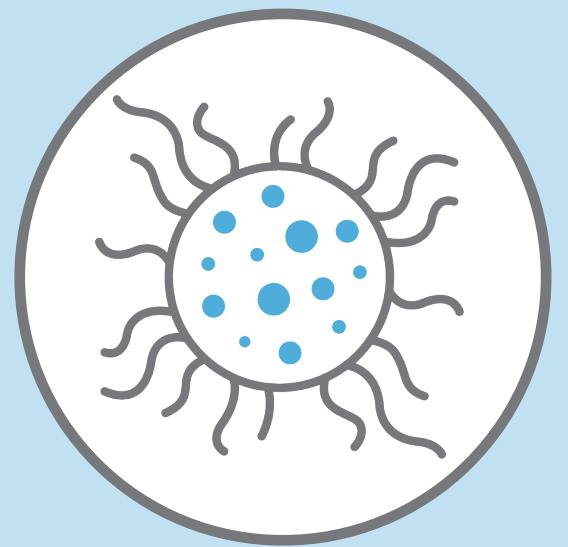


Genomics
has seen
massive and
ongoing
changes in
technology

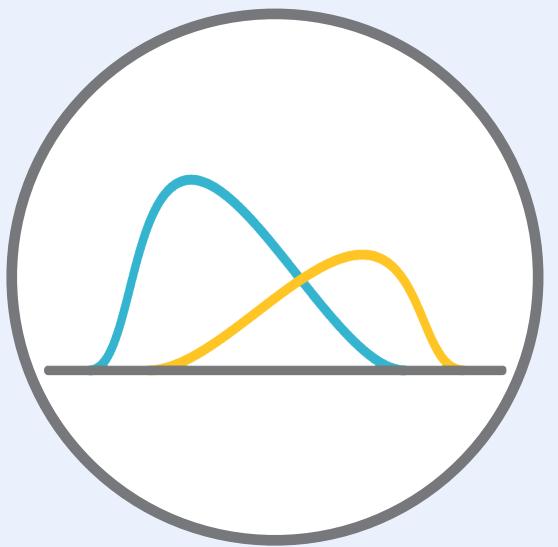




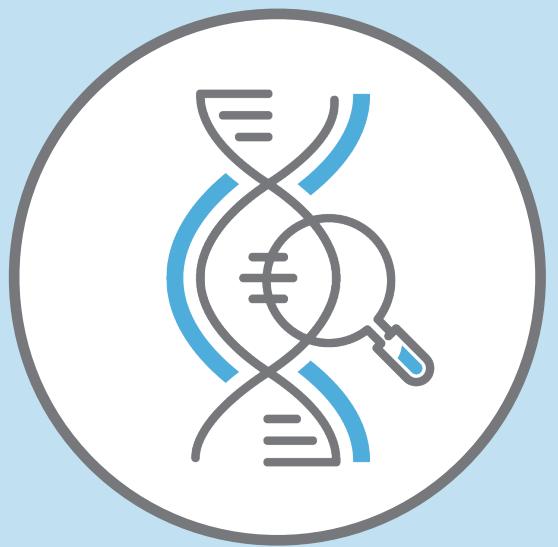
Global Genomic Data Sharing Can...



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



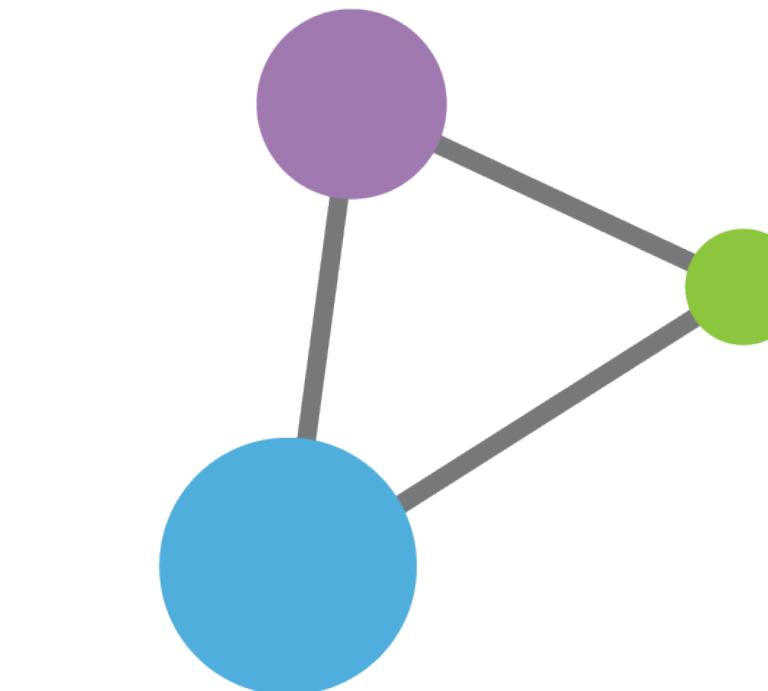
Centralized Genomic Knowledge Bases



Data Commons
Trusted, controlled repository of multiple datasets

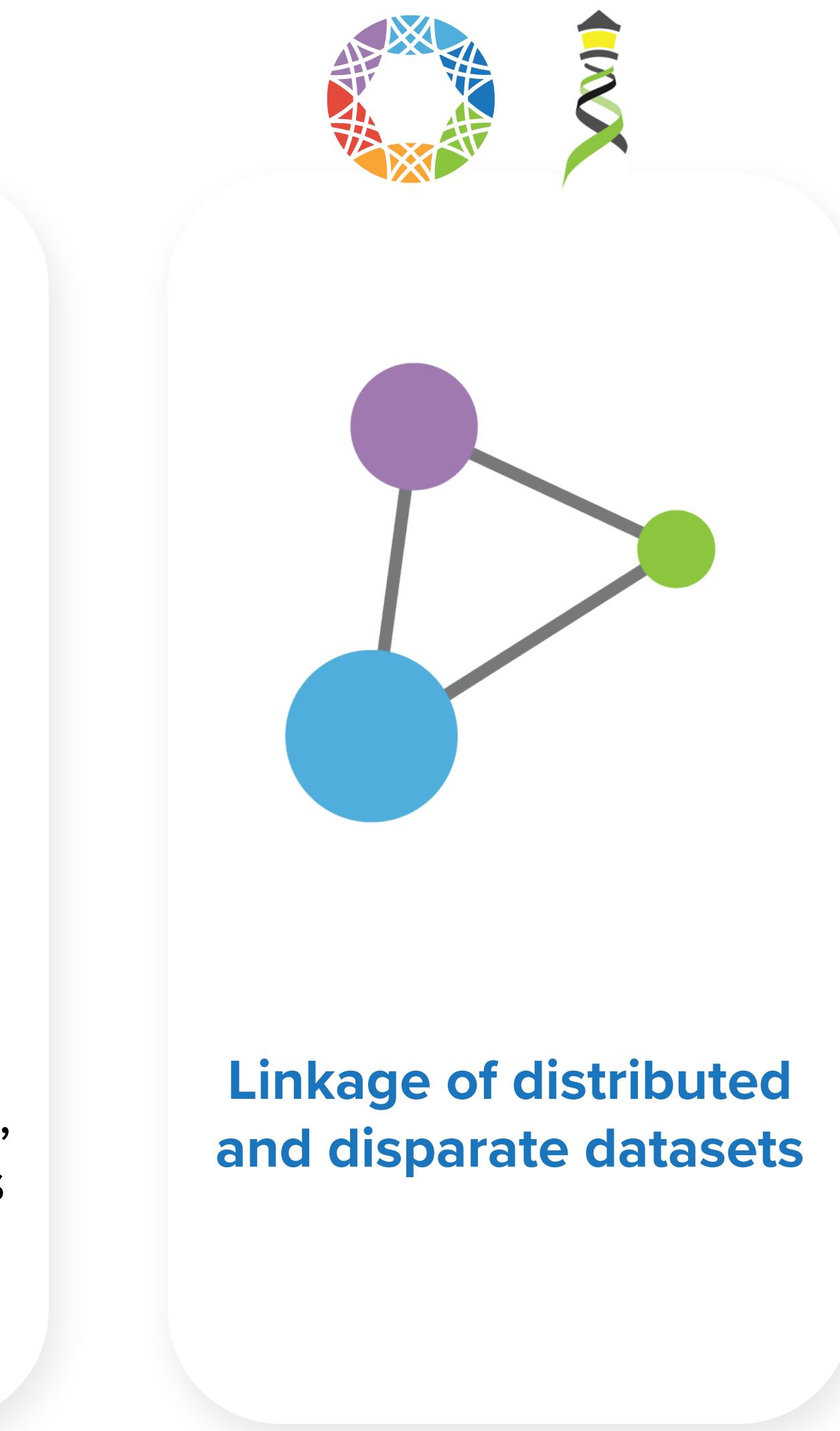
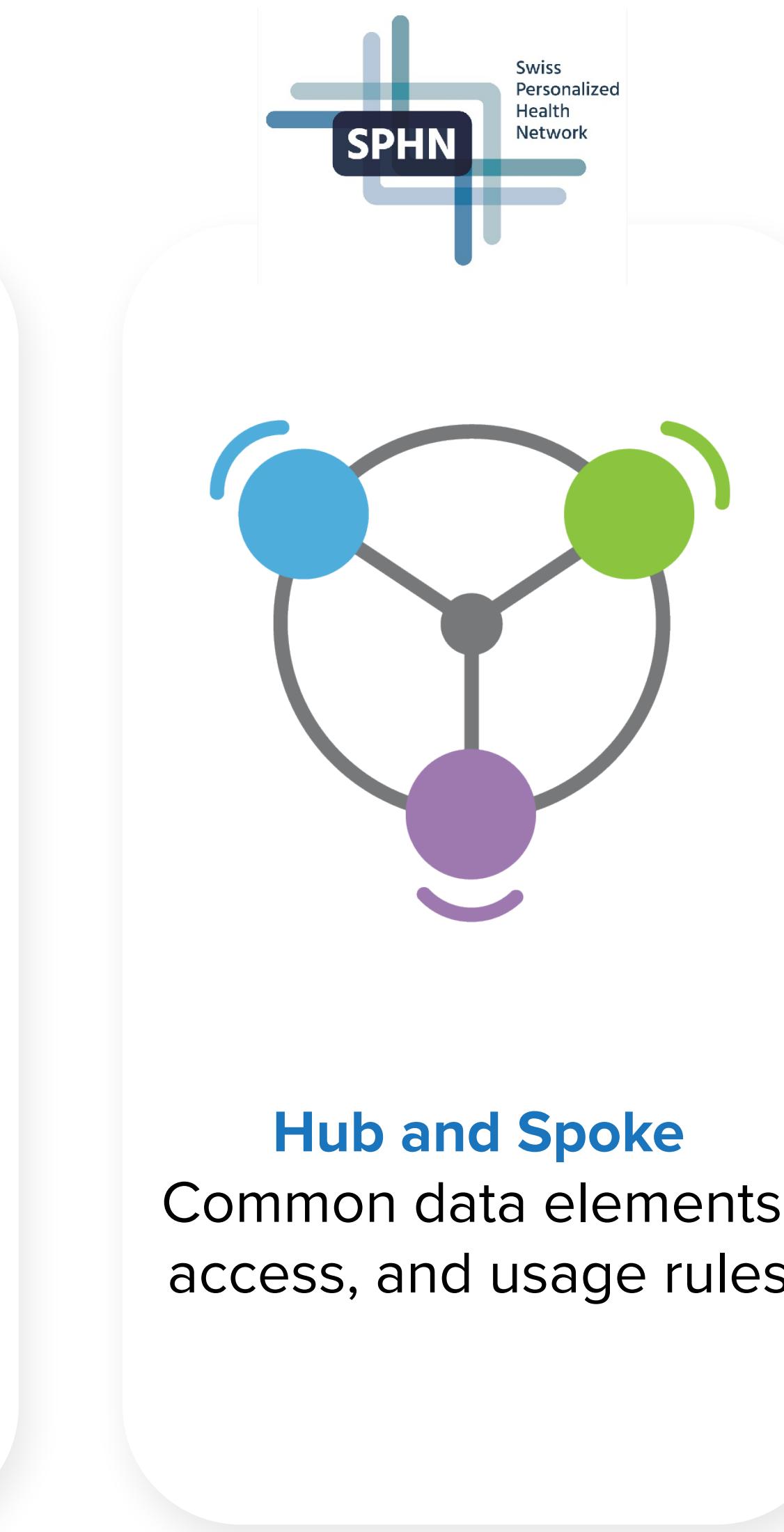
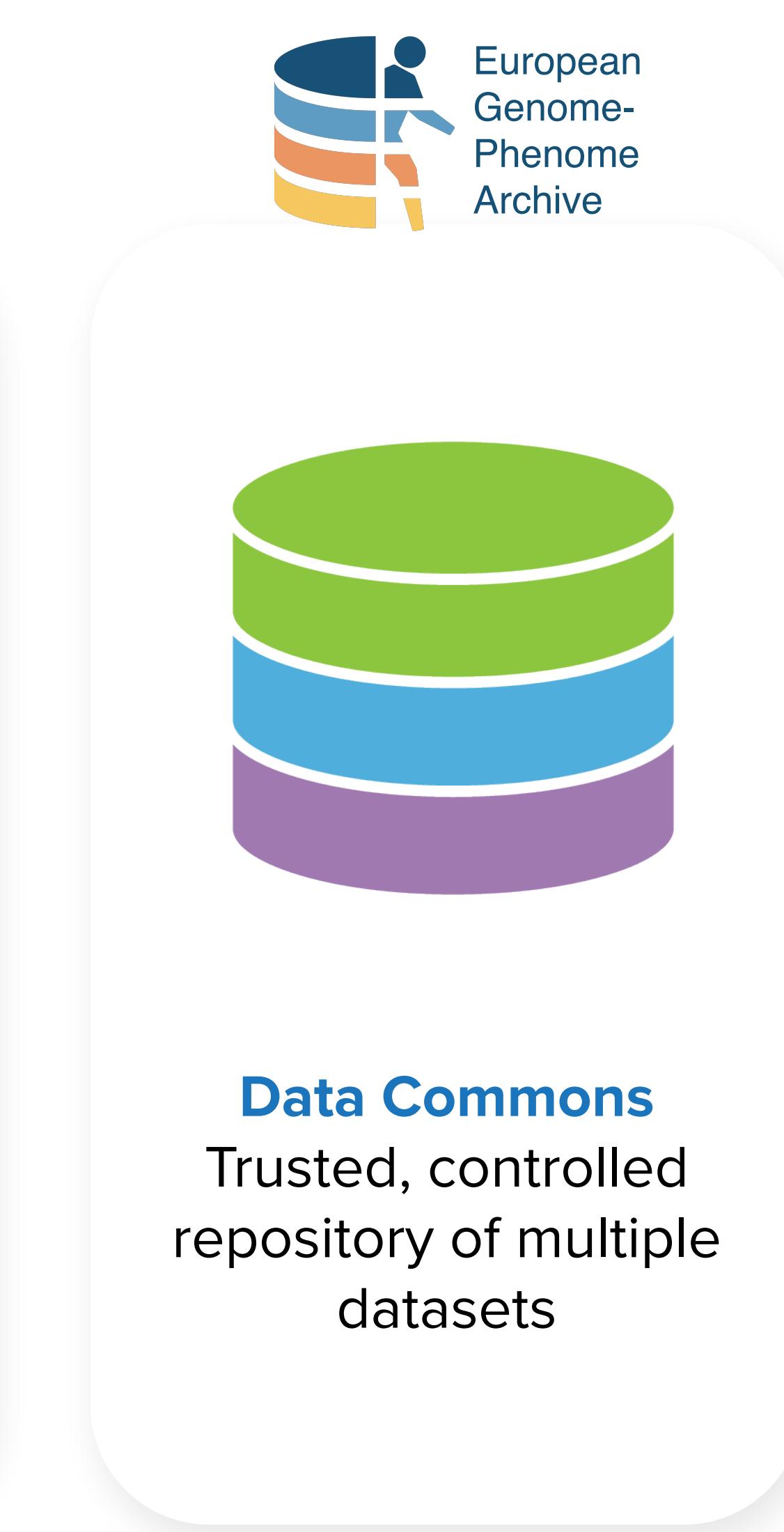
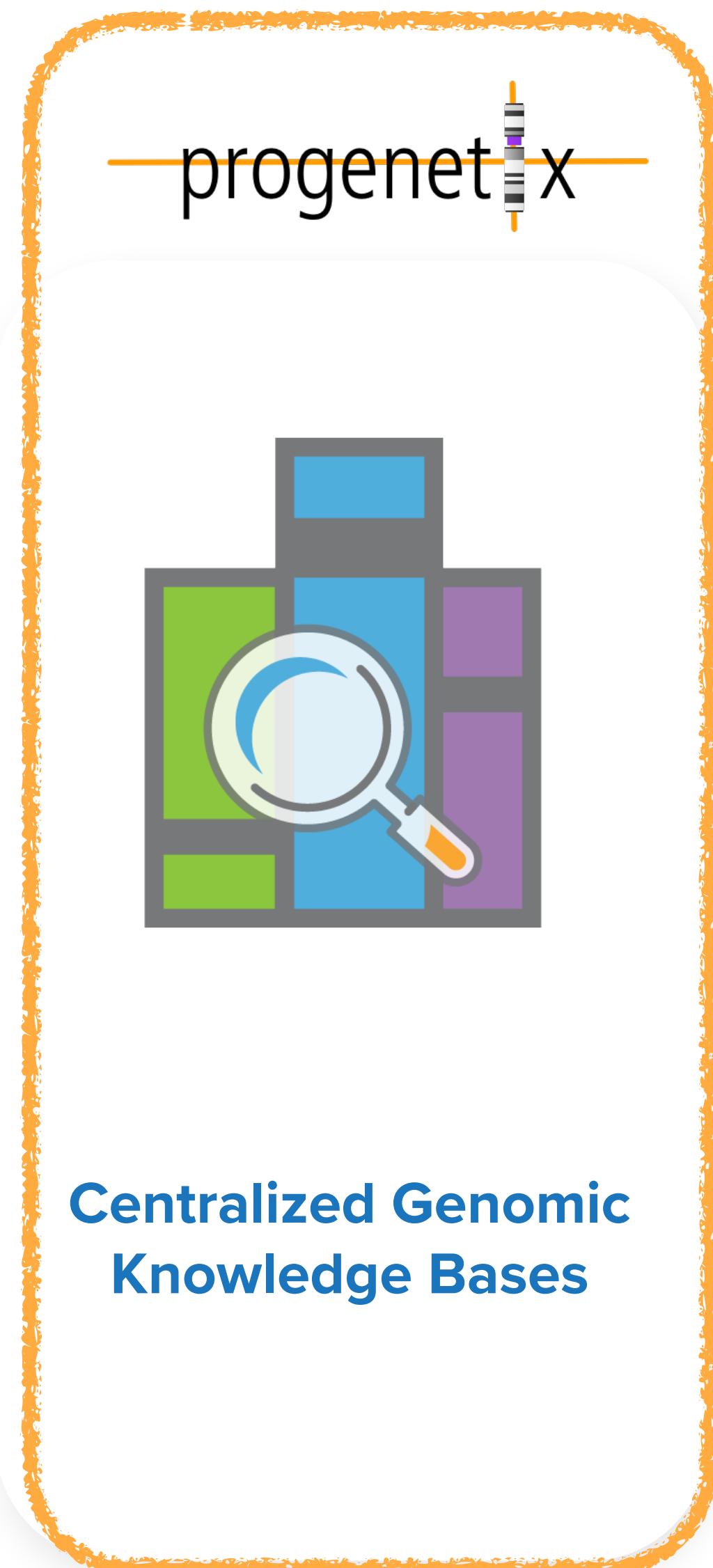


Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

Different Approaches to Data Sharing

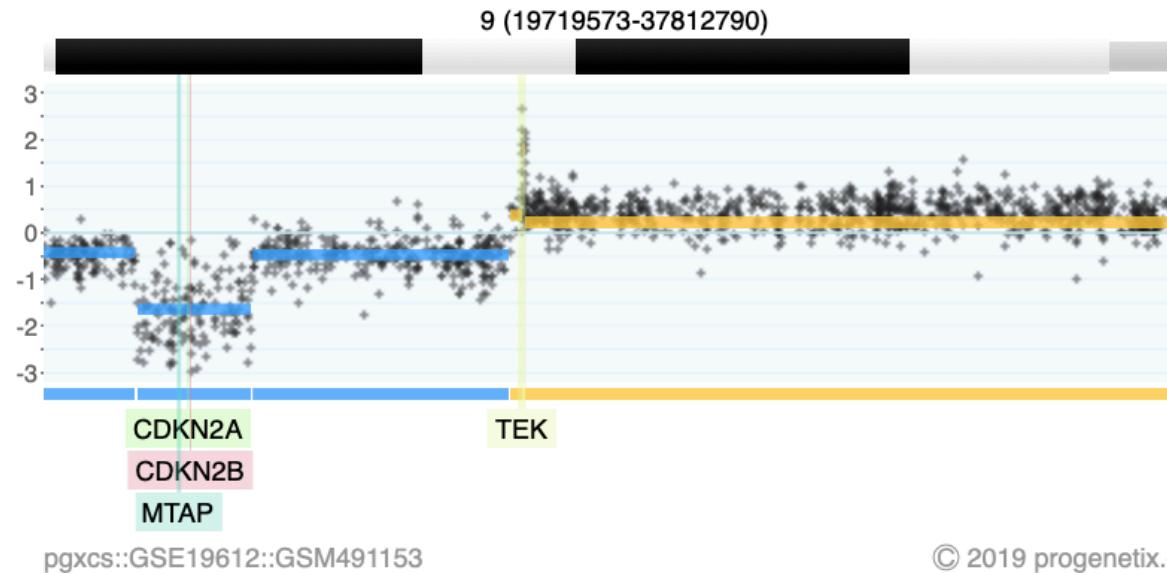


Theoretical Cytogenetics and Oncogenomics

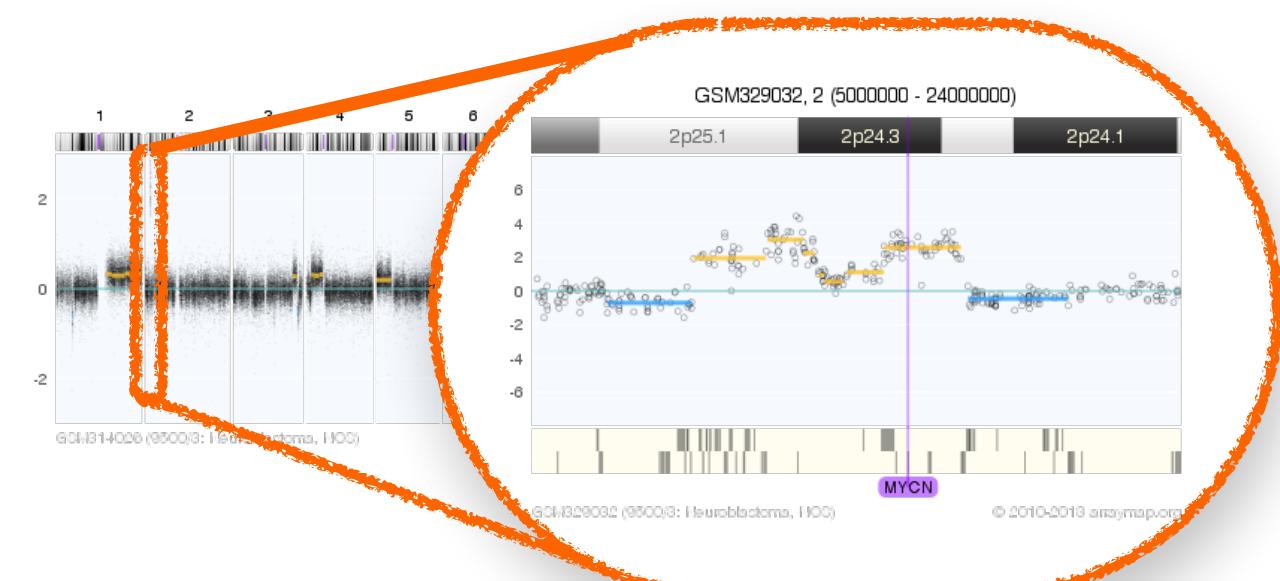
Research | Methods | Standards

Genomic Imbalances in Cancer - Copy Number Variations (CNV)

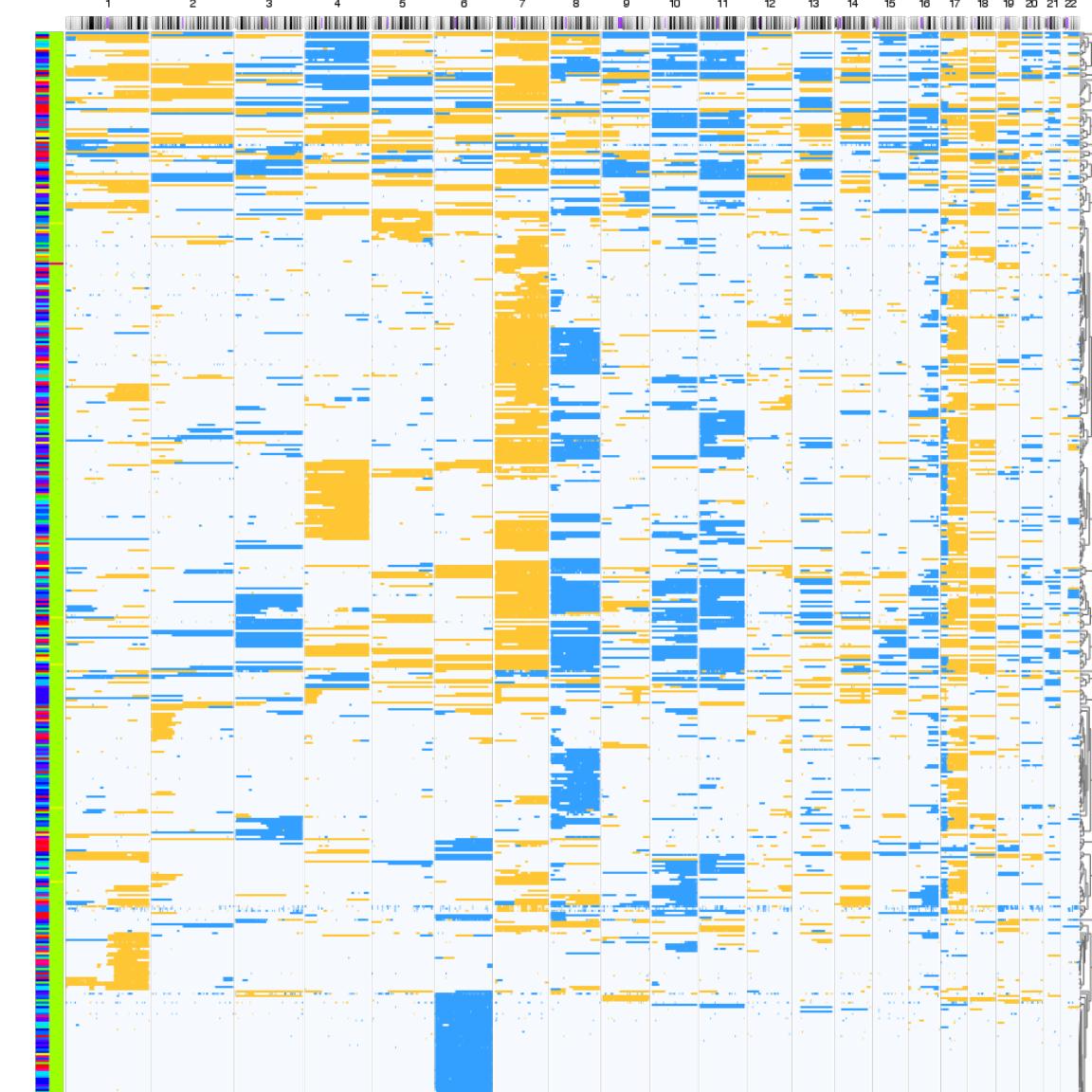
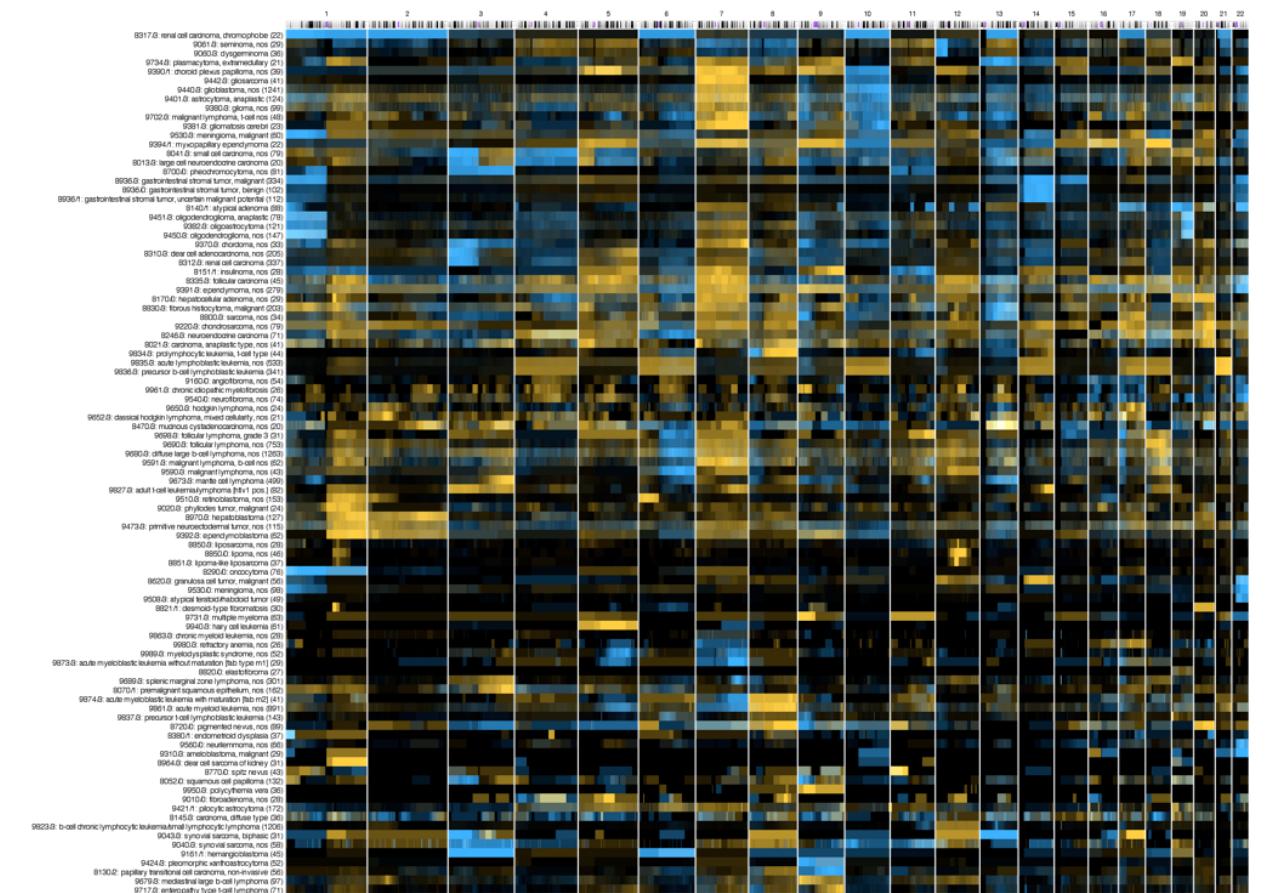
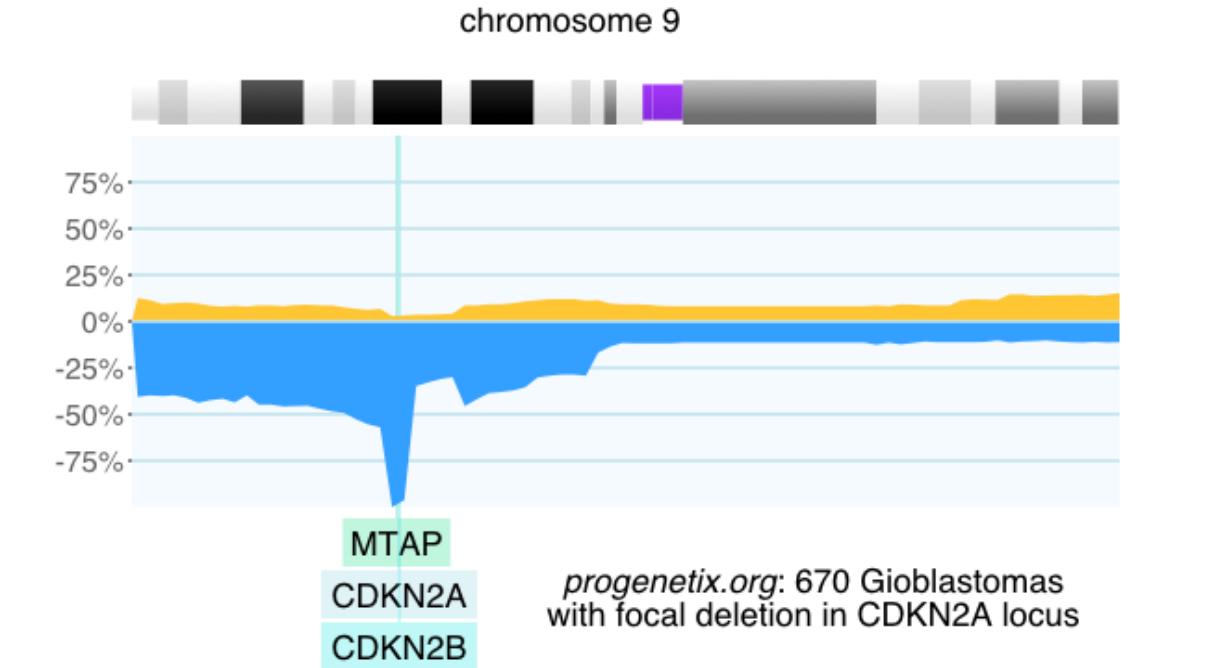
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)



2-event, homozygous deletion in a Glioblastoma

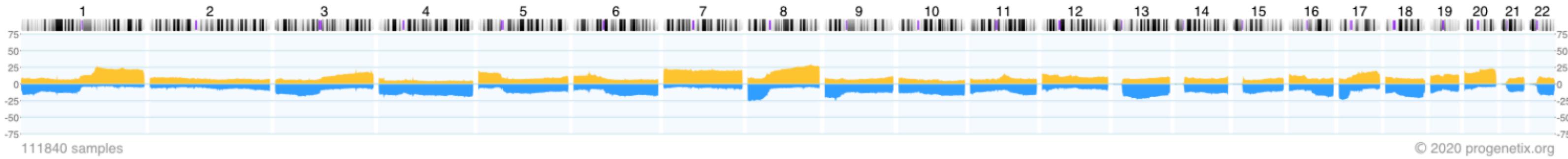


MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

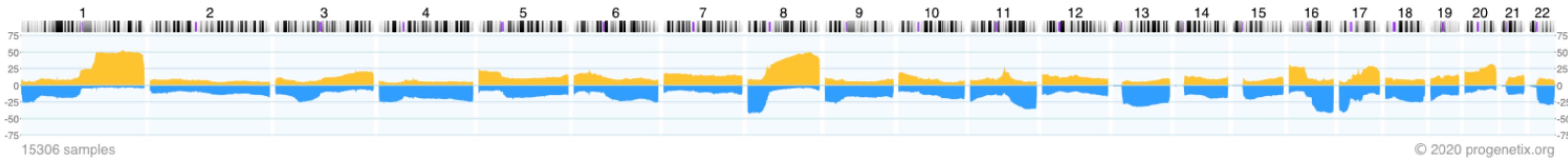


Somatic CNV in Cancer

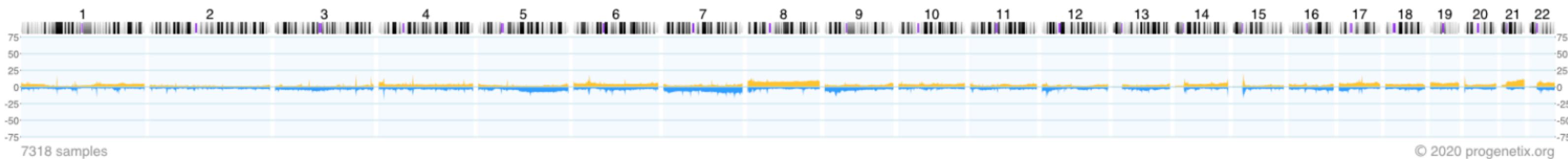
Progenetix: Regional CNV Frequencies in 111'840 Neoplasm (NCIT:C3262)



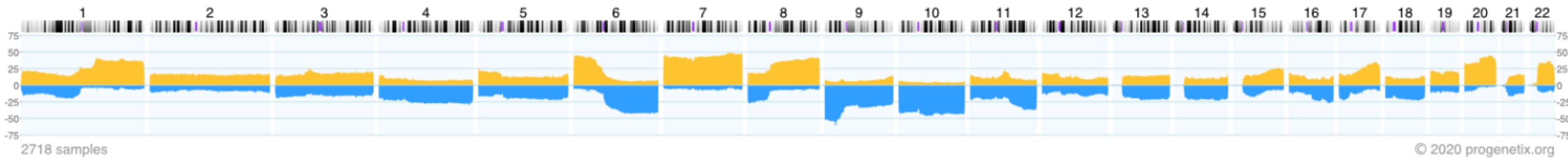
Malignant Breast Neoplasm (NCIT:C9335)



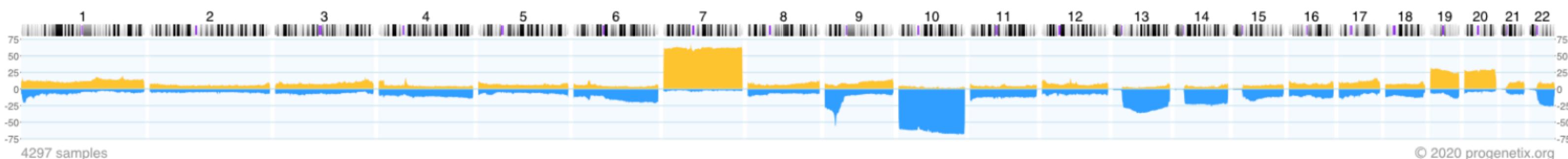
Acute Leukemia (NCIT:C9300)



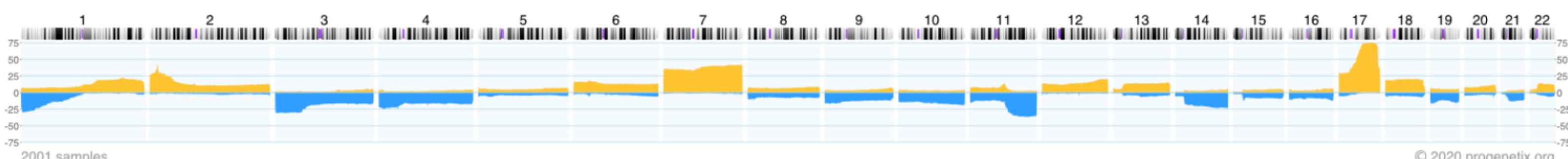
Melanoma (NCIT:C3224)



Glioblastoma (NCIT:C3058)



Neuroblastoma (NCIT:C3270)



Cancer Genomics Reference Resource

- **open** resource for oncogenomic profiles
- over **140'000 cancer CNV profiles**
- SNV data for some series (e.g. TCGA)
- more than **900 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



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⁺

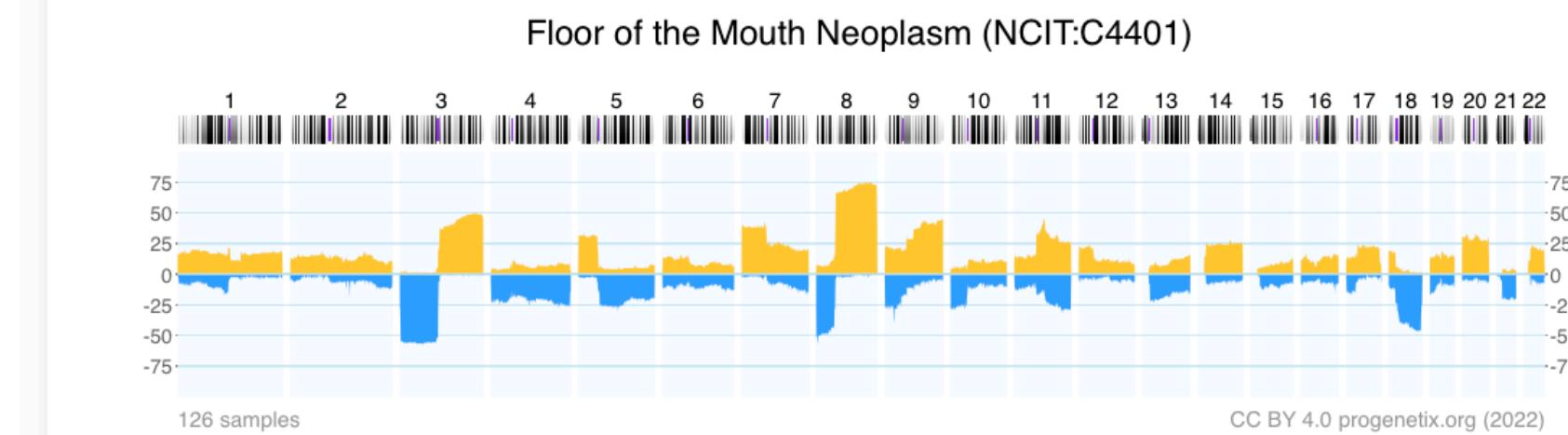
Documentation

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.



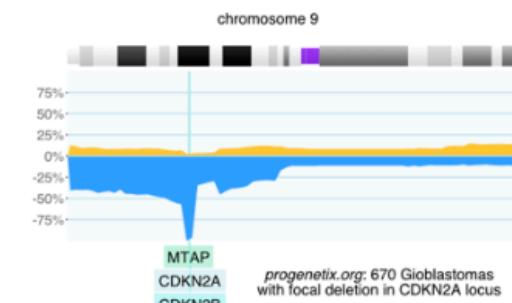
[Download SVG](#) | [Go to NCIT:C4401](#) | [Download CNV Frequencies](#)

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 **140'000 cancer CNV profiles**
- SNV data for some series (e.g. TCGA)
- more than **900 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

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 ↗

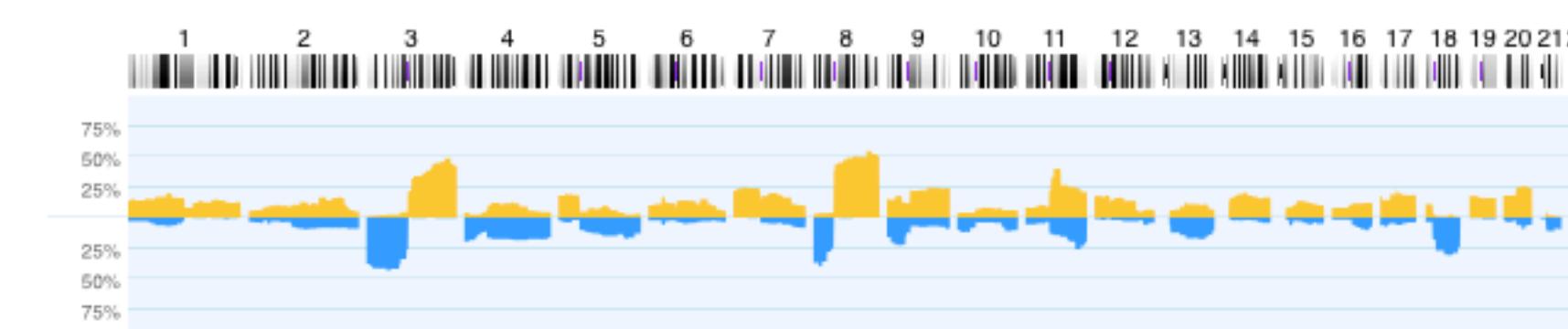
Sample Counts

- 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

© CC-BY 2001 - 2024 progenetix.org

- 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)



Universität
Zürich UZH

progenetix



Swiss Institute of
Bioinformatics

Cancer Genomics Reference Resource

- **open** resource for oncogenomic profiles
- over **140'000 cancer CNV profiles**
- SNV data for some series (e.g. TCGA)
- more than **900 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



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) ↗

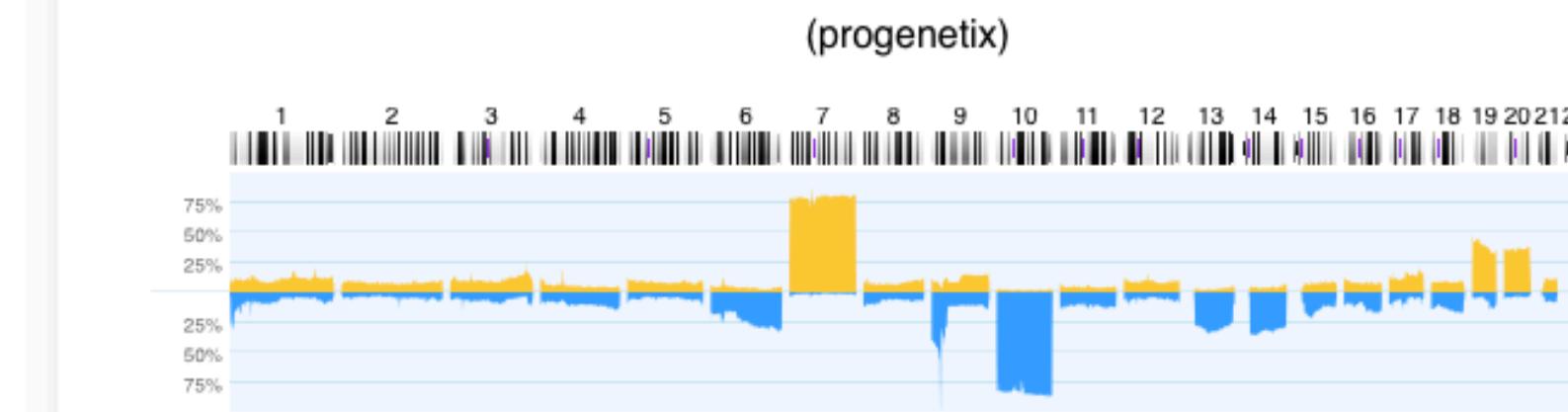
Visualization options

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)

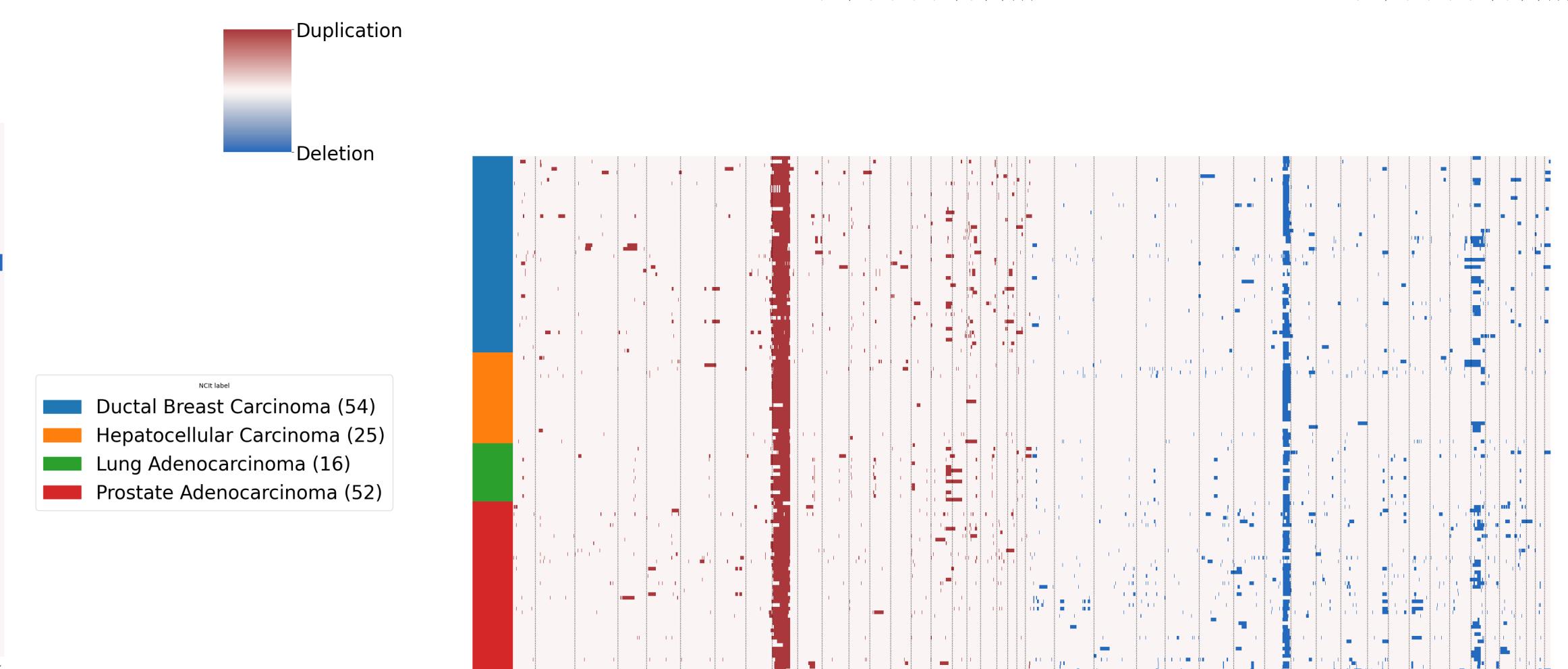
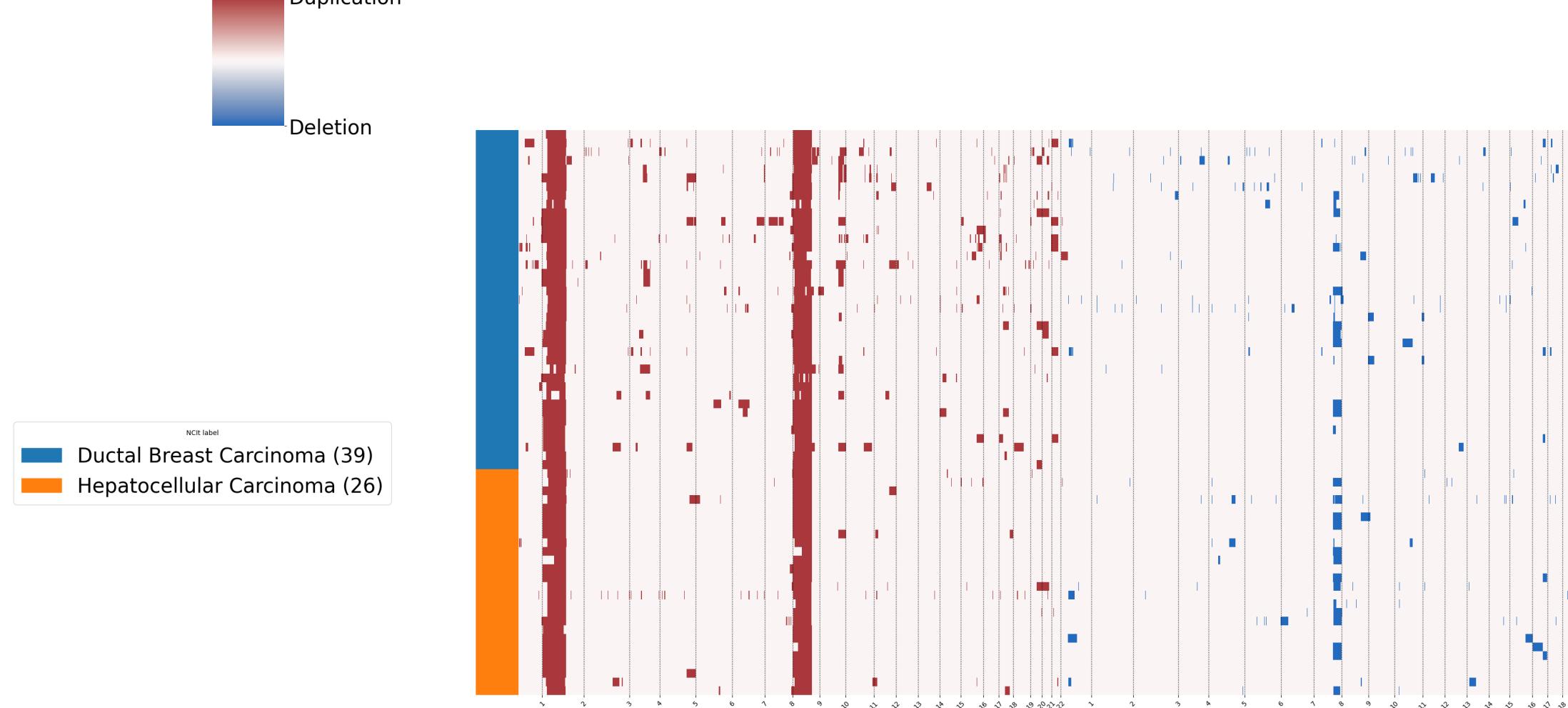
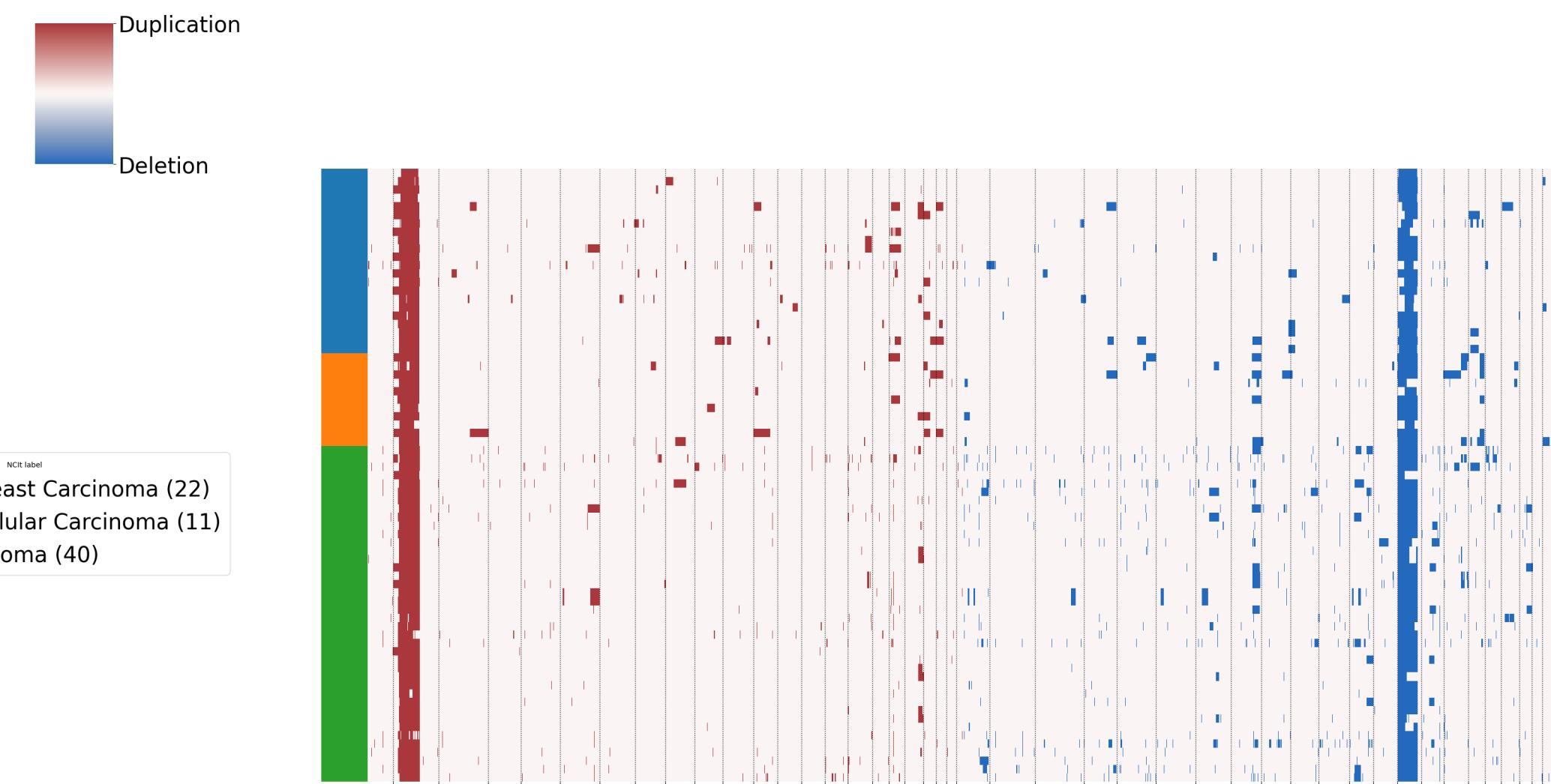
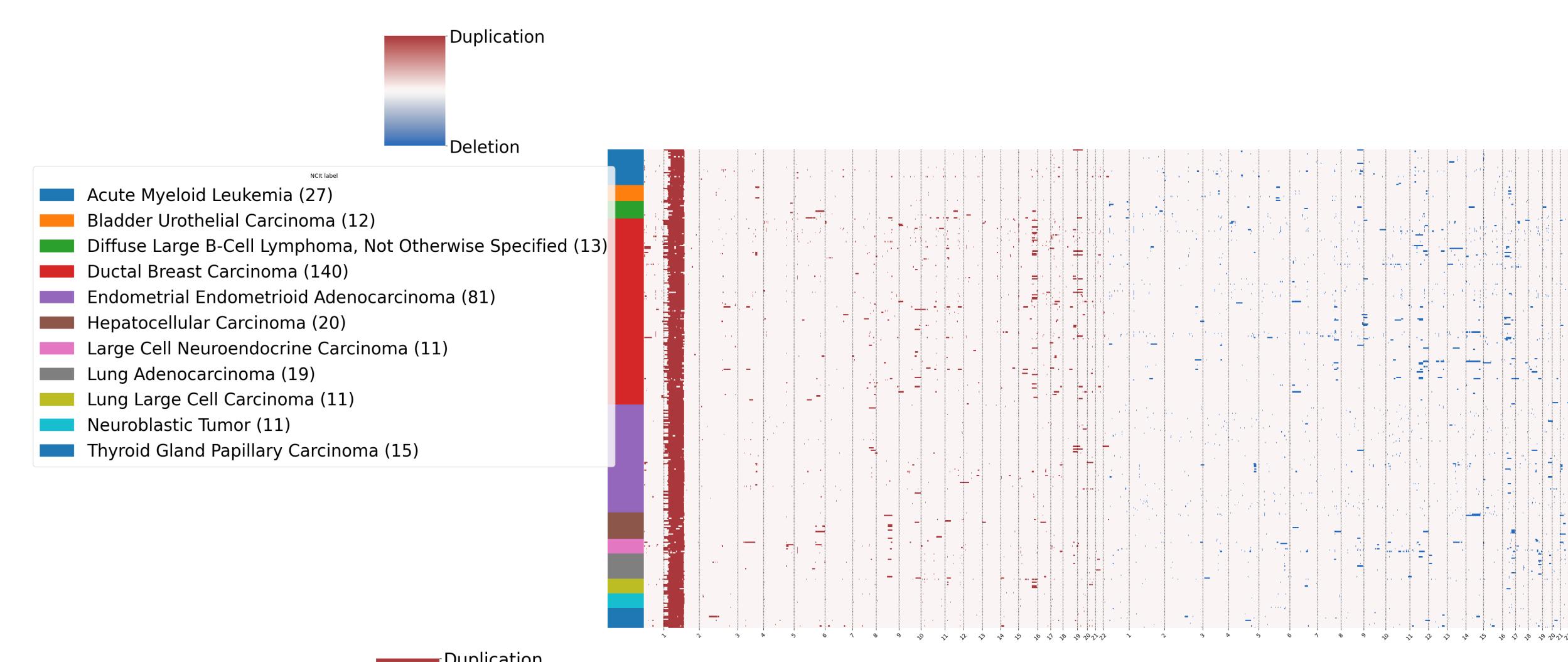
1-657 ↗

Download Sample Data (JSON)

1-657 ↗

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 are retrieved based on the selected cell line. For example, selecting "HOS" for HeLa will also return the daughter lines by default - but can be filtered to only return the primary cell line.

Cell Lines (with parental/derived hierarchies)

Filter subsets e.g. by prefix

Hierarchy Depth

No Selection

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

Cell Line Details

HOS (cellosaurus:CVCL_0312)

Subset Type

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

Sample Counts

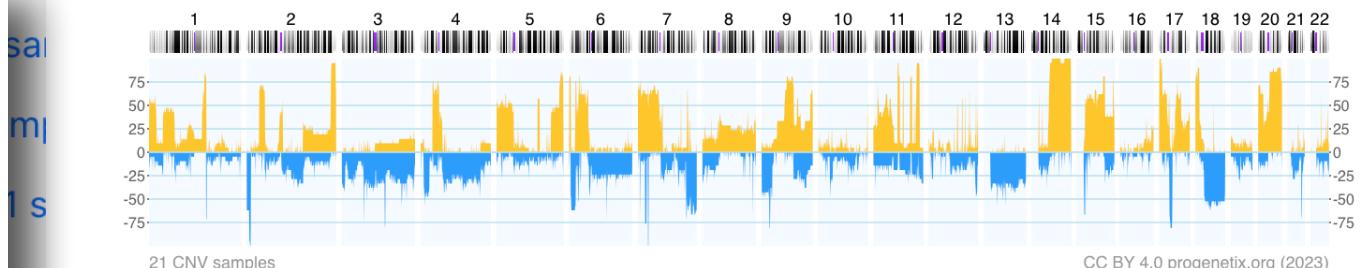
- 204 samples
- 57 direct cellosaurus:CVCL_0312 code matches
- 21 CNV analyses

Search Samples

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

Raw Data (click to show/hide)

HOS (cellosaurus:CVCL_0312)



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

DATABASE
The Journal of Biological Databases and Curation

Follow this preprint

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

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-	72d

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

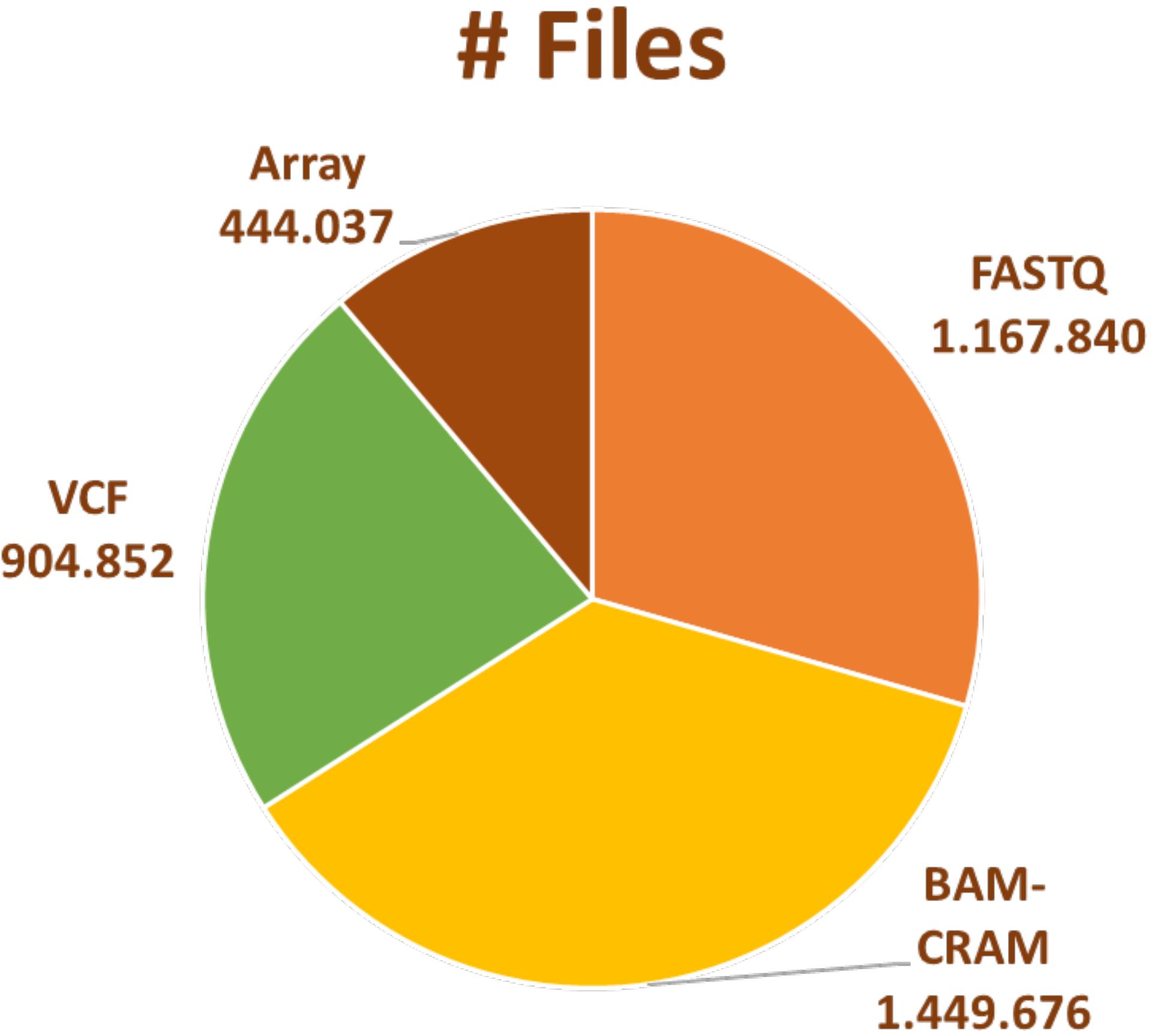


- EGA “owns” nothing; data controllers tell who is authorized to access ***their*** datasets
- EGA admins provide smooth “all or nothing” data sharing process

A screenshot of the EGA DAC interface. At the top, it says "My DACs - EGAC5000000005 - Requests". Below that, it shows "EuCanImage DAC" and "This is a DAC for EuCanImage data". On the right, there's an "EDIT" button and a "HISTORY" button. The main area is titled "My DACs - EGAC5000000005 - History" and "EuCanImage DAC". It says "This is a DAC for EuCanImage data". There's a search bar with placeholder text "Type something for filter the requests...". Below the search bar, there's a table with columns: Date, Requester, Dataset, and DAC Admin/Member. The table contains three rows of data:

Date	Requester	Dataset	DAC Admin/Member
18 August 2022	gemma.milla@crg.eu	EGAD5000000032	Dr Lauren A Fromont
17 August 2022	Dr Teresa Garcia Lezana	EGAD5000000033	Dr Teresa Garcia Lezana
16 August 2022	Dr Teresa Garcia Lezana	EGAD5000000032	Dr Lauren A Fromont

Each row has a "revoke permission" toggle switch next to the DAC Admin/Member column. At the bottom of the table, there's an "APPLY" button.



4,328 Studies released

10,470 Datasets

2,309 Data Access Committees

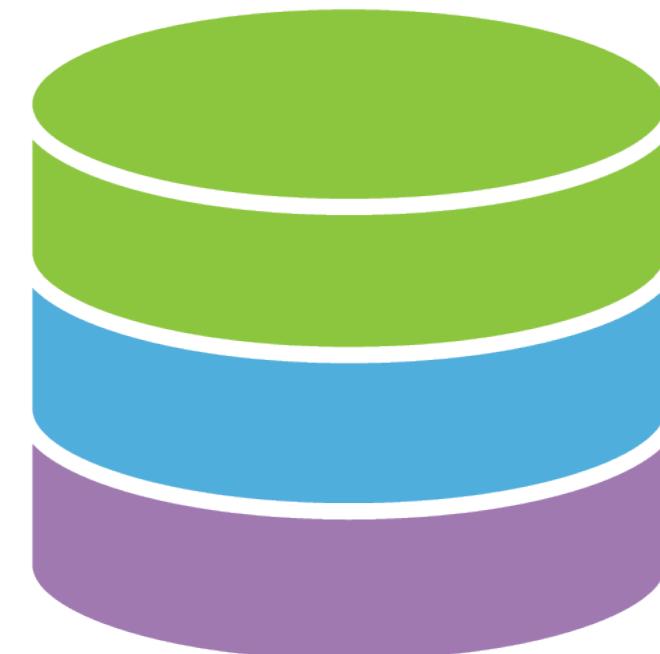
Different Approaches to Data Sharing



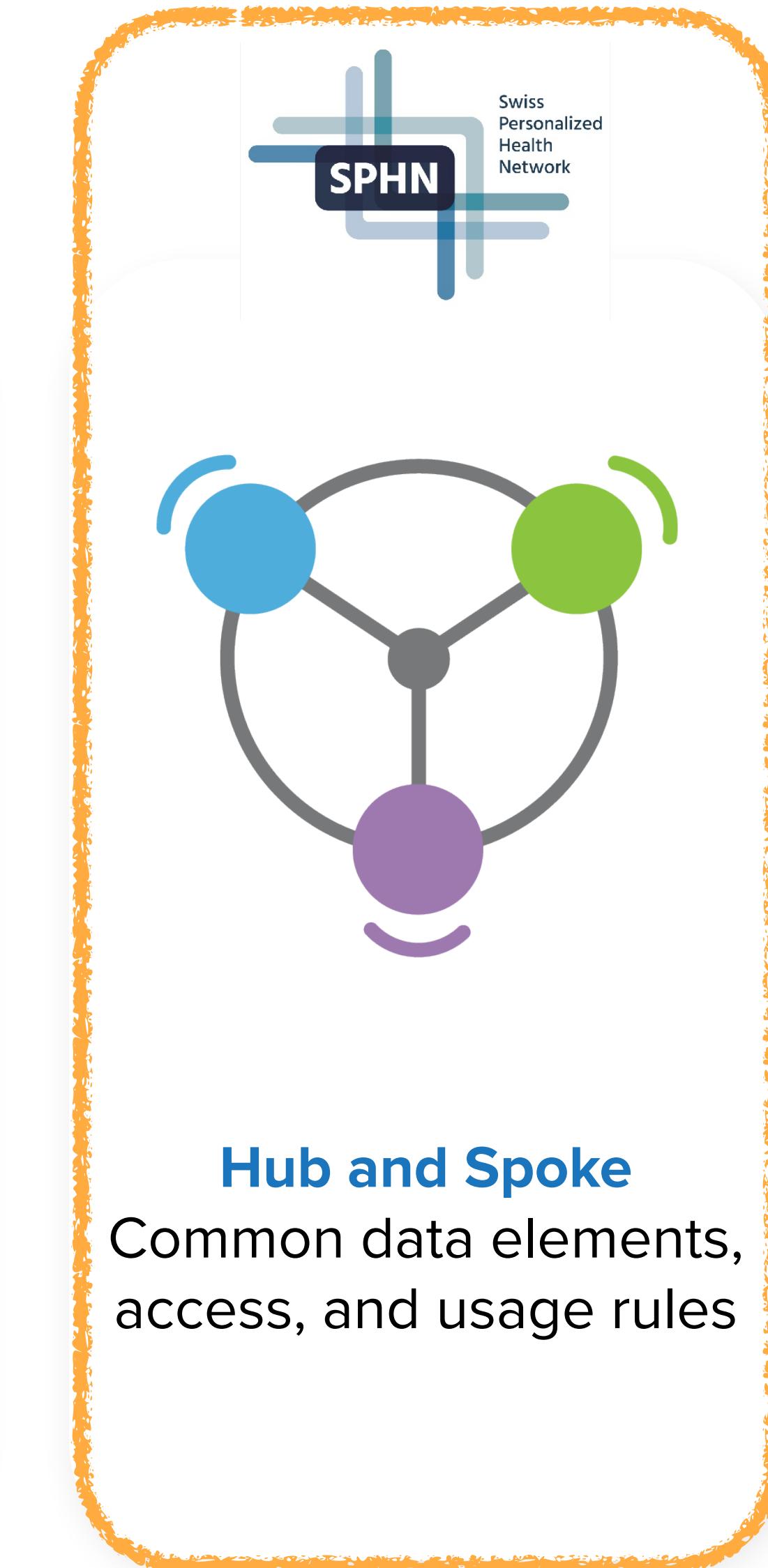
Centralized Genomic Knowledge Bases



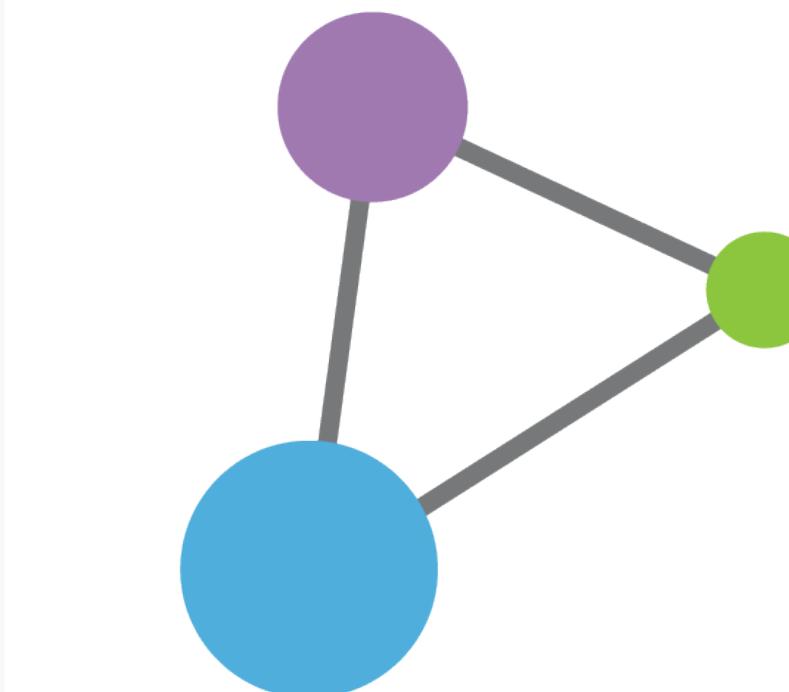
European
Genome-
Phenome
Archive



Data Commons
Trusted, controlled
repository of multiple
datasets

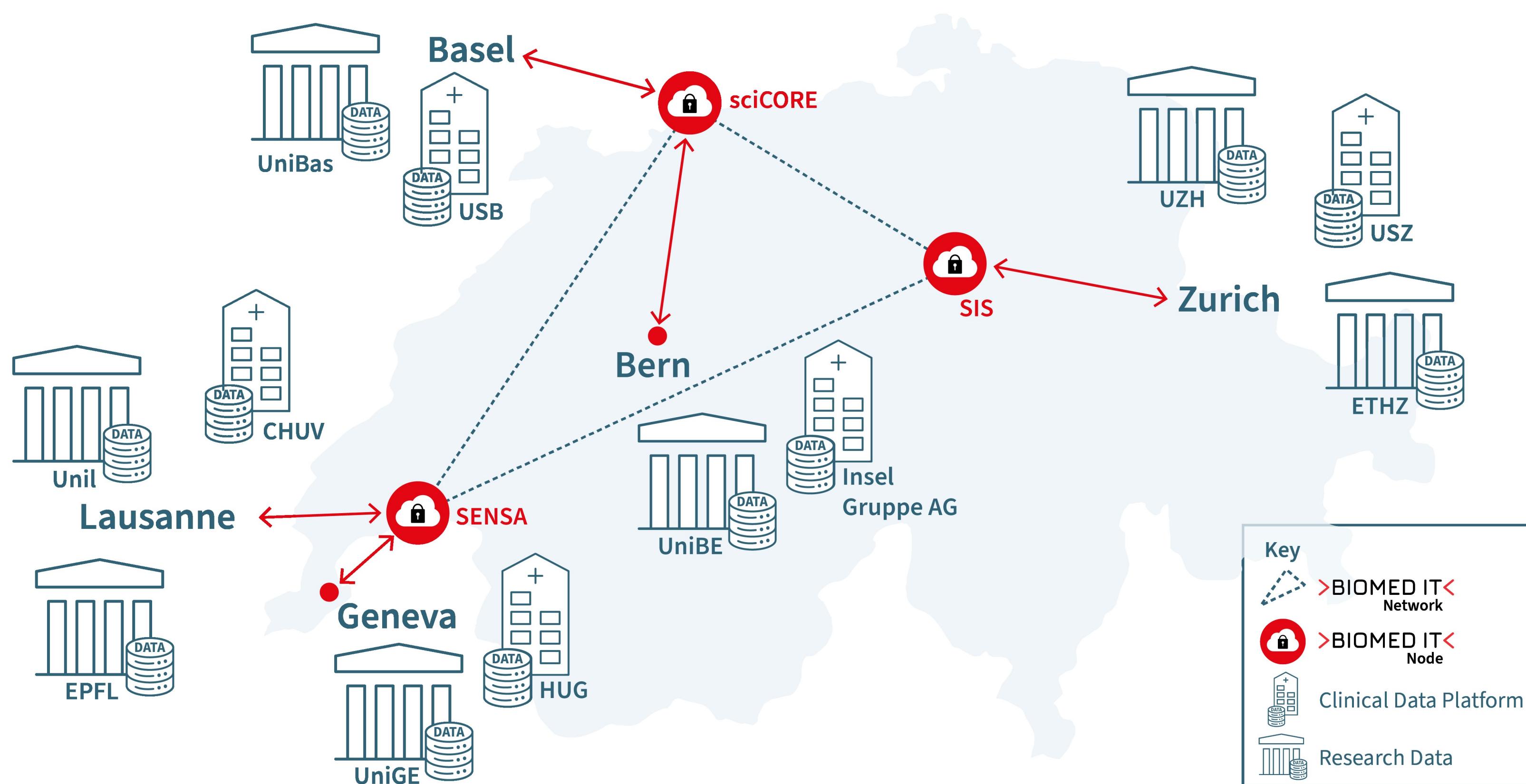


Hub and Spoke
Common data elements,
access, and usage rules



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
SCHWEIZERISCHE 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

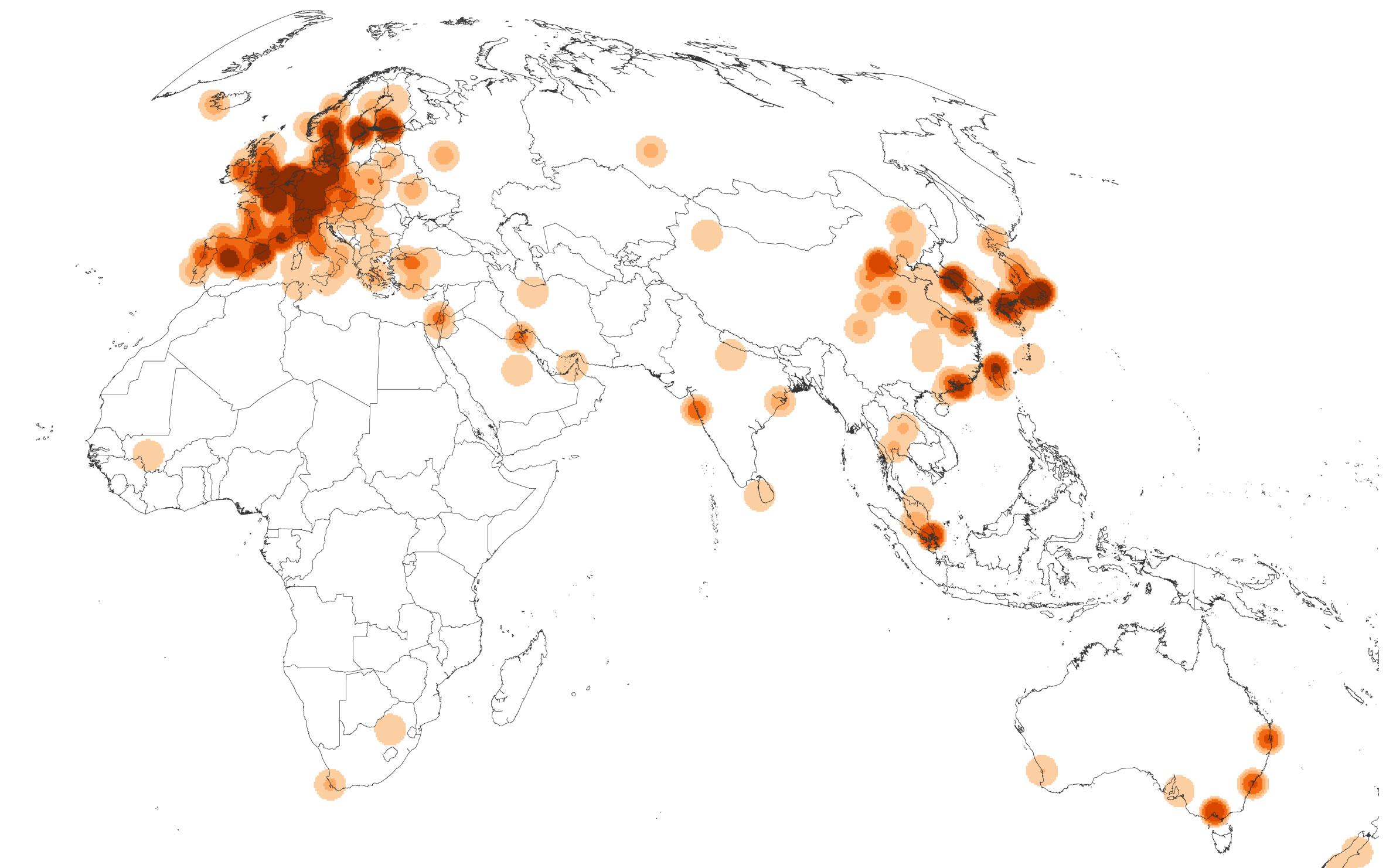
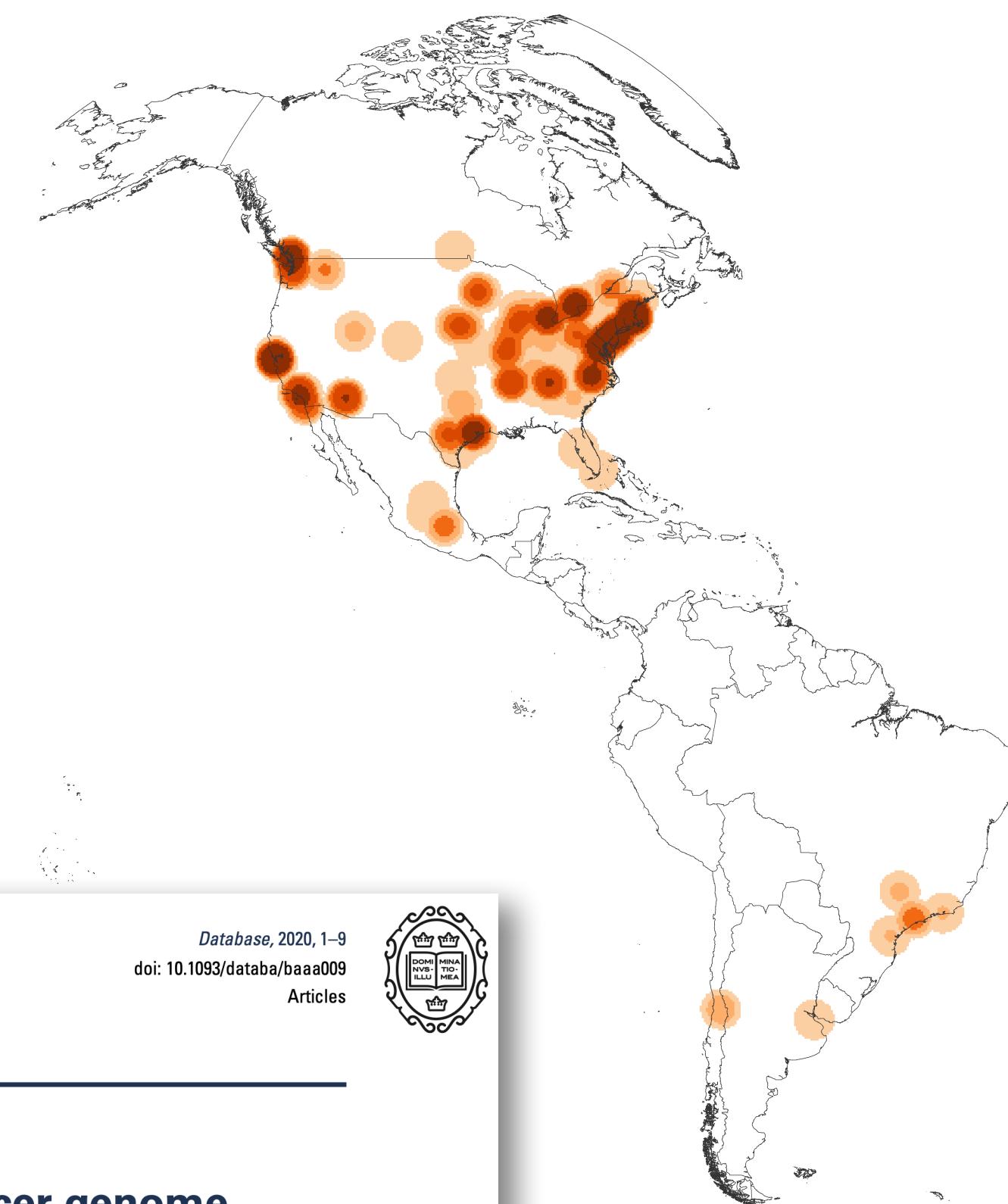
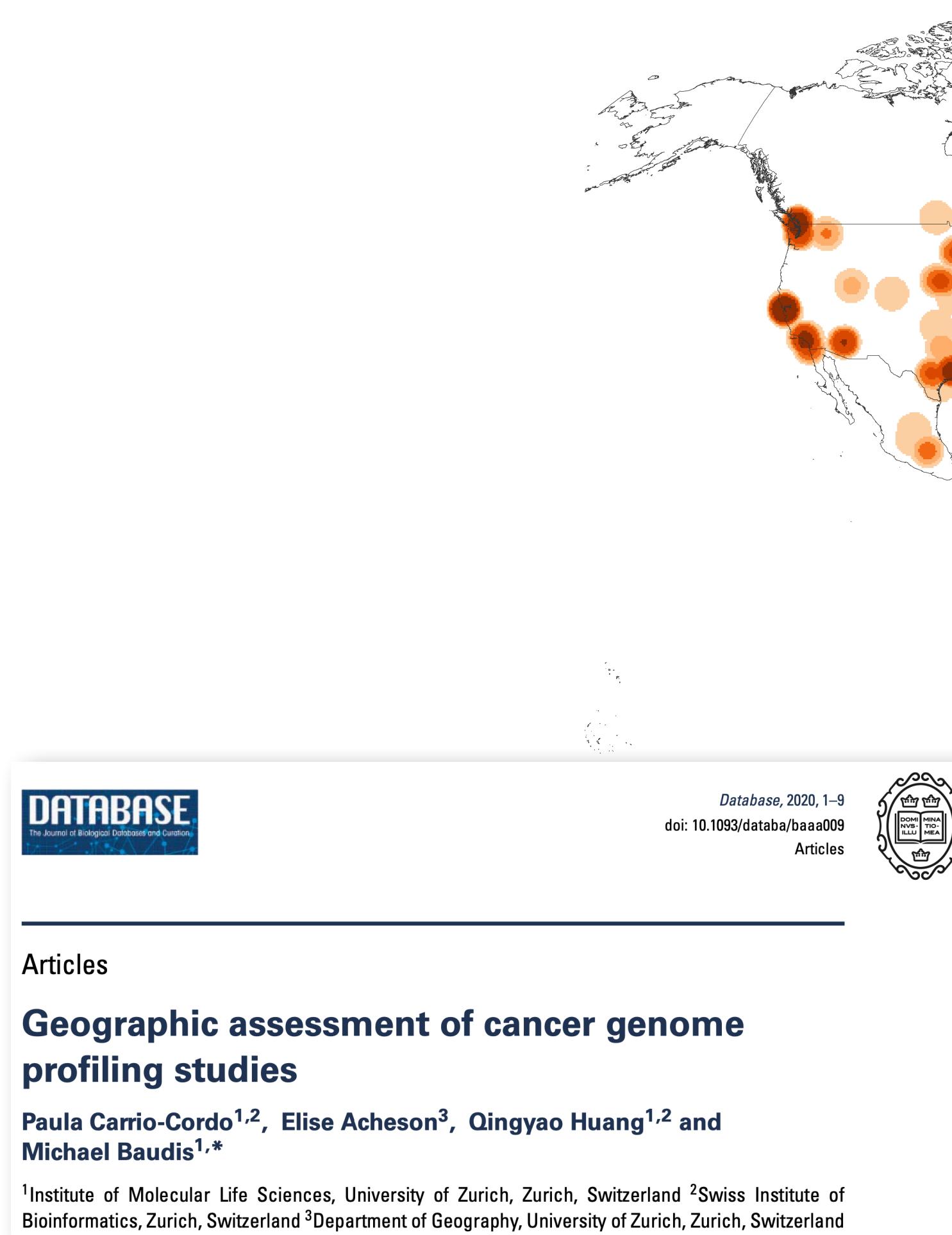
 **Université Medizin Schweiz
Médecine Universitaire Suisse**

swissuniversities



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.

200+ genomic data initiatives globally

Clinical/Genomic
Medicine



Research



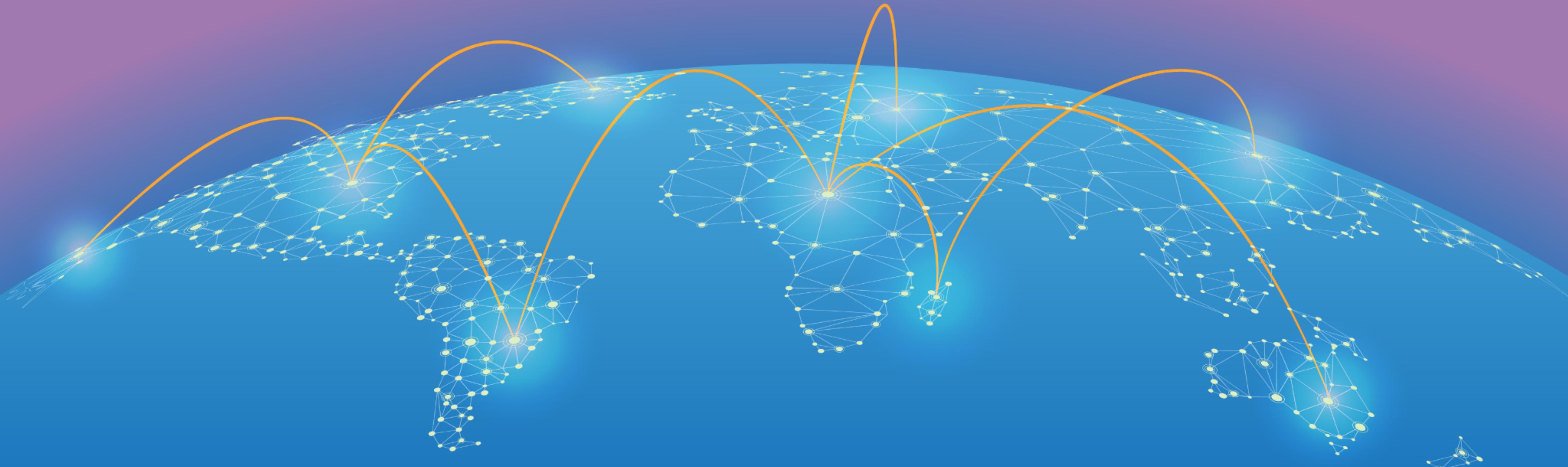
National



Cohorts



Since data is distributed globally, we need interoperable standards to answer research questions



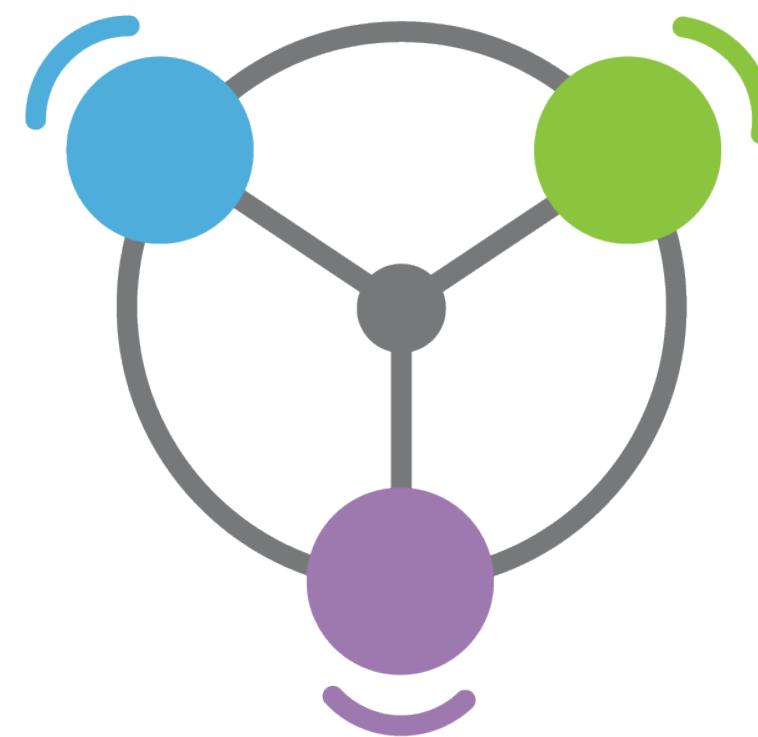
Different Approaches to Data Sharing



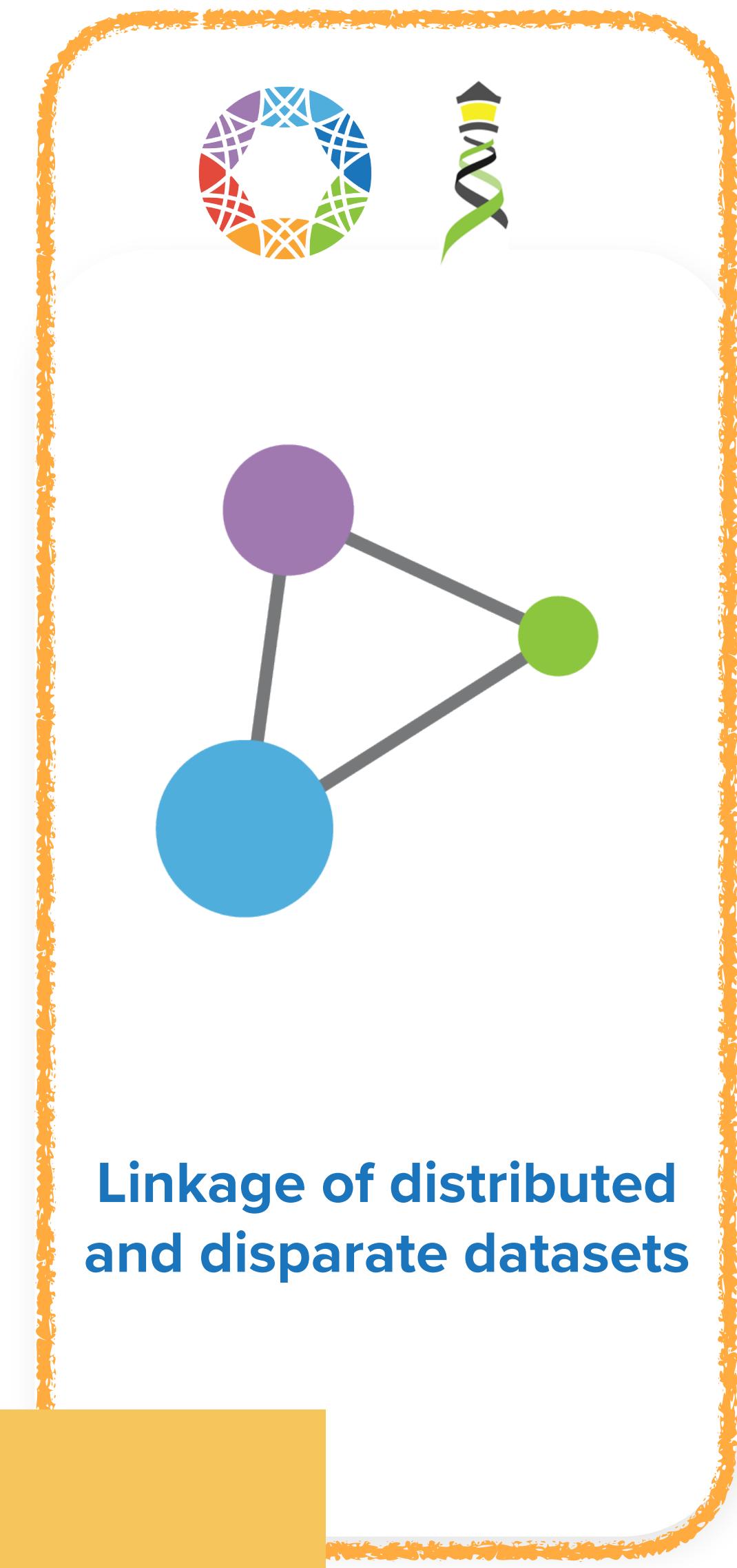
Centralized Genomic Knowledge Bases



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*

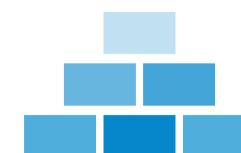
SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291



Framework for Responsible Sharing of Genomics and Health-Related Data

ga4gh.org/framework

Translated into
14 languages



FOUNDATIONAL PRINCIPLES

- Respect Individuals, Families and Communities
- Advance Research and Scientific Knowledge
- Promote Health, Wellbeing and the Fair Distribution of Benefits
- Foster Trust, Integrity and Reciprocity



AIMS OF THE FRAMEWORK

- Foster responsible data sharing
- Protect and promote the welfare, rights, and interests of groups and individuals who donate their data
- Provide benchmarks for accountability
- Establish a framework for greater international data sharing, cooperation, collaboration, and governance

Universal Declaration
of Human Rights (1948)

27(1)

“The Right
to Science”

27(2)

“The Right
to Recognition”

Strategic Leadership Committee

Executive team

**HEIDI REHM**

MASSACHUSETTS GENERAL HOSPITAL

BROAD INSTITUTE OF MIT AND HARVARD

Chair

Driver Project Champion for: [Clinical Genome Resource \(ClinGen\)](#) | [Matchmaker Exchange](#)Community lead for: [Clinical Genomics Laboratory Community](#)**EWAN BIRNEY**

EUROPEAN MOLECULAR BIOLOGY LABORATORY (EMBL) | EMBL'S EUROPEAN BIOINFORMATICS INSTITUTE (EBI)

Chair Emeritus

**KATHRYN NORTH**

MURDOCH CHILDREN'S RESEARCH INSTITUTE | AUSTRALIAN GENOMICS

Vice-Chair

Driver Project Champion for: [Australian Genomics](#) | [International Precision Child Health Partnership \(IPCHIP\)](#)**PETER GOODHAND**

ONTARIO INSTITUTE FOR CANCER

RESEARCH (OICR)

Chief Executive Officer | President, GA4GH Inc.

**ANGELA PAGE**

BROAD INSTITUTE OF MIT AND HARVARD

Director of Strategy and Engagement | Secretary, GA4GH Inc.

**ANDY YATES**

EMBL'S EUROPEAN BIOINFORMATICS INSTITUTE (EBI)

Interim Chief Standards Officer

Product lead for: [refget](#)**MICHAEL BAUDIS**

UNIVERSITY OF ZURICH

Work Stream lead for: [Discovery Work Stream](#)
Product lead for: [Beacon](#)**TIFFANY BOUGHTWOOD**

AUSTRALIAN GENOMICS

Product lead for: [Machine Readable Consent Guidance \(MRCG\)](#)**MÉLANIE COURTOT**

ONTARIO INSTITUTE FOR CANCER

RESEARCH (OICR)

Driver Project Champion for: [Pan-Canadian Genome Library \(PCGL\)](#)Work Stream lead for: [Clinical & Phenotypic Data Capture \(Clin/Pheno\) Work Stream](#)Product lead for: [Data Use Ontology \(DUO\)](#)**ROBERT FREIMUTH**

MAYO CLINIC

Driver Project Champion for: [All of Us Research](#)**DAVID GLAZER**

VERITLY

**OLIVER HOFMANN**

UNIVERSITY OF MELBOURNE CENTRE FOR CANCER RESEARCH

**YANN JOLY**

CENTRE OF GENOMICS AND POLICY

Driver Project Champion for: [EpisShare](#) | [Pan-Canadian Genome Library \(PCGL\)](#)Work Stream lead for: [Regulatory & Ethics Work Stream \(REWS\)](#)
Product lead for: [Clinical Data Sharing and Consent](#) | [Genetic Discrimination Toolkit](#)**AUGUSTO RENDON**

GENOMICS ENGLAND

Driver Project Champion for: [Genomics England](#)**SERENA SCOLLEN**

ELIXIR

Strategic Partner Champion for: [ELIXIR](#)**HEIDI SOFIA**

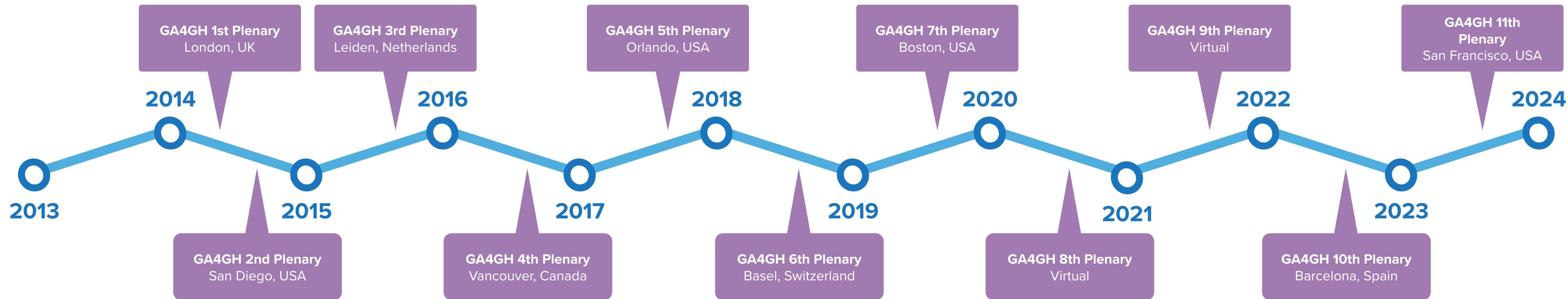
NIH NATIONAL HUMAN GENOME RESEARCH INSTITUTE (NHGRI)



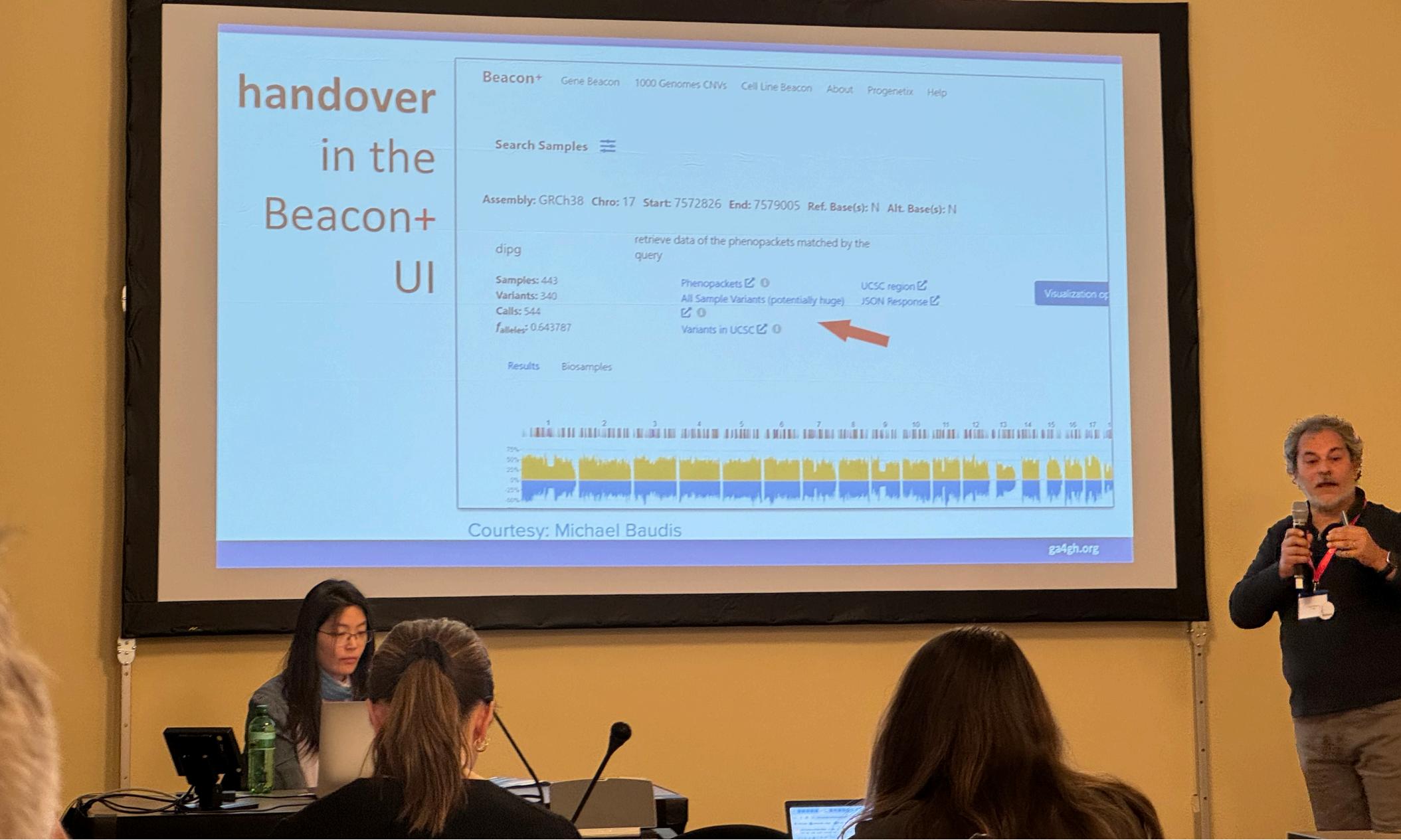
GA4GH timeline



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>



The Global Alliance for Genomics and Health (GA4GH) gathered for the 2024 [April Connect meeting](#) in Ascona, Switzerland and online from 21 to 24 April. The GA4GH Connect meetings provide an opportunity for contributors to advance the GA4GH Road Map, showcase GA4GH standards and policies in action, and gather feedback on product development and community needs. The meeting brought together 103 in-person attendees and 312 virtual attendees for updates from Work Streams and Driver Projects, breakout sessions, and themed events.



Our funders, partners, and Driver Projects

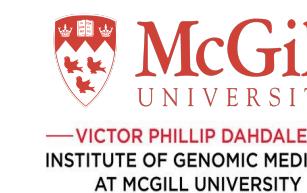


Global Alliance
for Genomics & Health

Core Funders



Host Institutions



Assigned Expert Funders/Employers



Strategic Partner

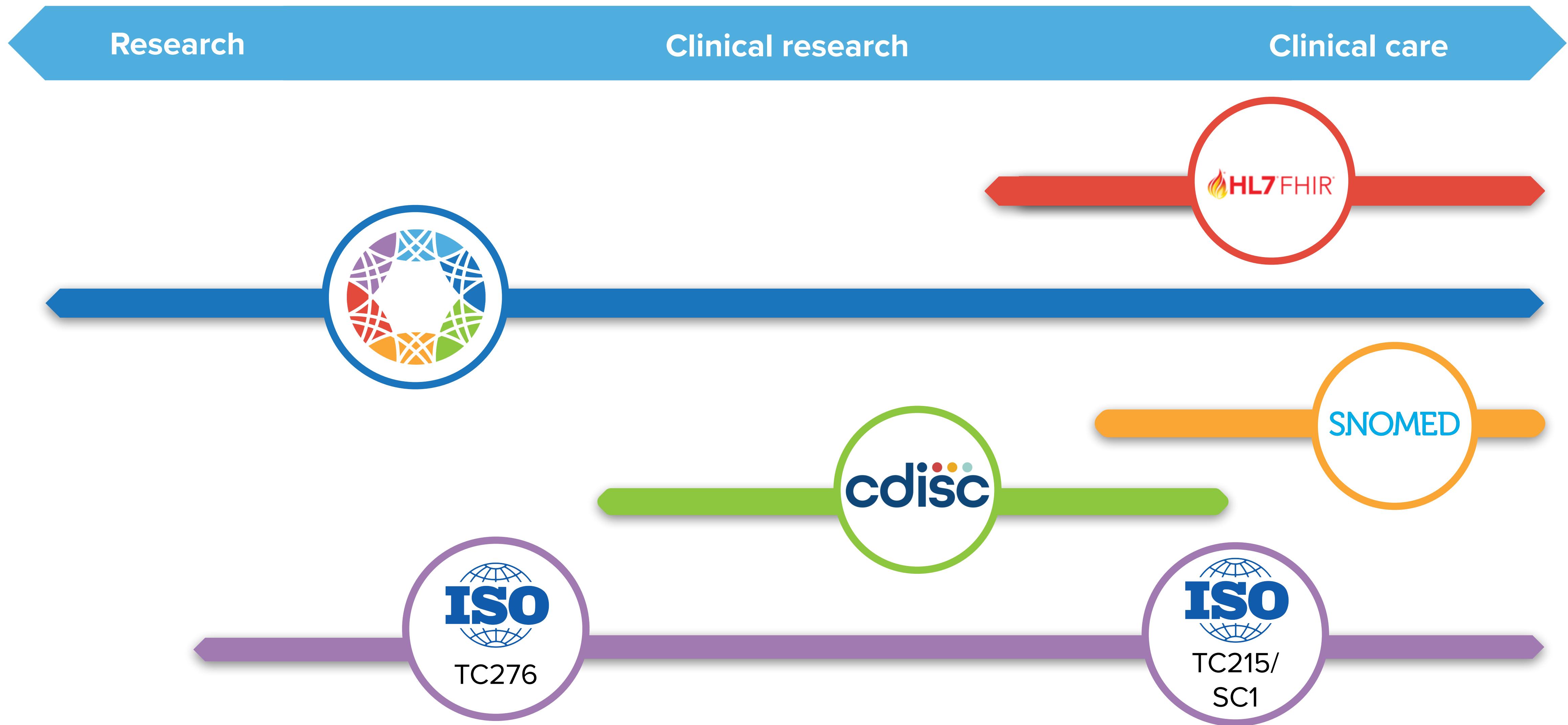


GDI is funded by the European Commission under the Digital Europe Programme under grant agreement number 101081813 and through co-funding from participating Member States.

Alignment with other standards organizations



Global Alliance
for Genomics & Health



Host Institutions



Global Alliance
for Genomics & Health



Toronto, Canada

OICR is a collaborative research institute that conducts and enables high-impact translational cancer research.



Hinxton, UK

The Wellcome Sanger Institute is a world leader in genome research delivering insights into human and pathogen biology.



Cambridge, USA

The Broad Institute seeks to narrow the gap between new biological insights and impact for patients by fulfilling the promise of genomic medicine.



Hinxton, UK

EMBL-EBI provides the infrastructure needed to share data openly in life sciences to make discoveries that benefit humankind.



VICTOR PHILLIP DAHDALEH
INSTITUTE OF GENOMIC MEDICINE
AT MCGILL UNIVERSITY

Montreal, Canada

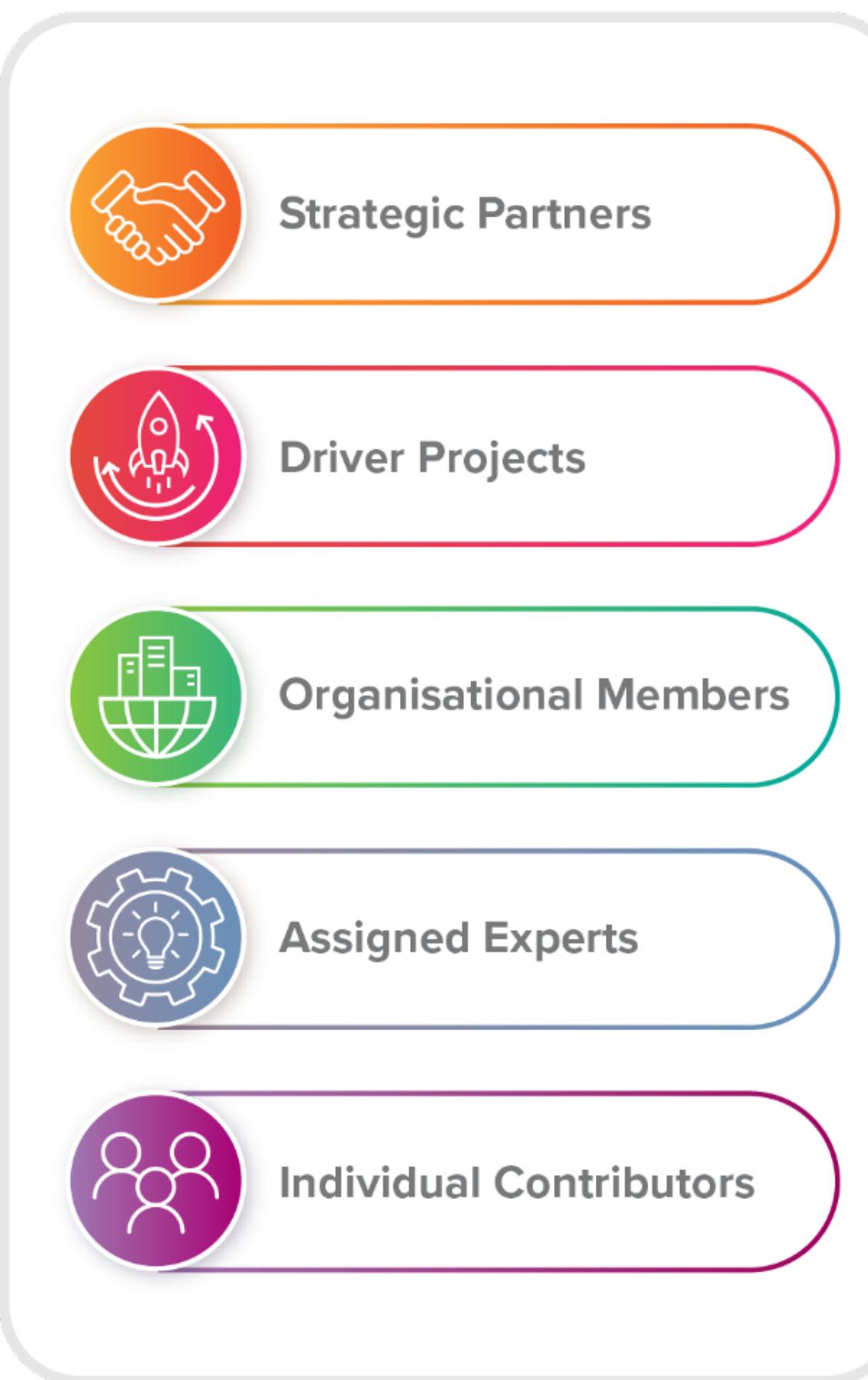
The Victor Phillip Dahdaleh Institute of Genomic Medicine applies genomic innovation to pave the way towards a healthier, more sustainable, and informed future.

How GA4GH works



Global Alliance
for Genomics & Health

Here's our community...



...and here's what we do.



10 new GA4GH Driver Projects in 2024



Global Alliance
for Genomics & Health



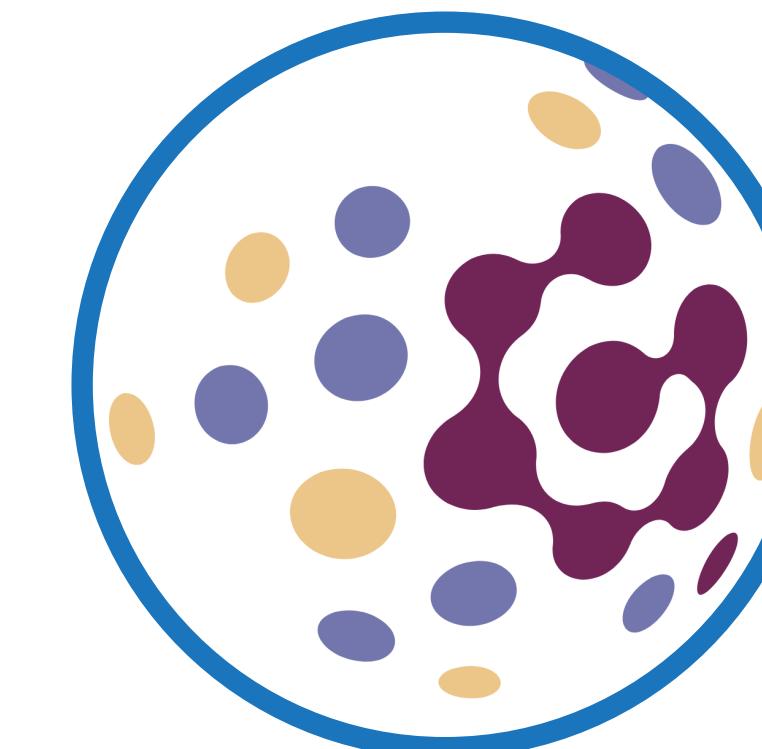
Biomedical Research Hub



imCORE®



EOSC4Cancer



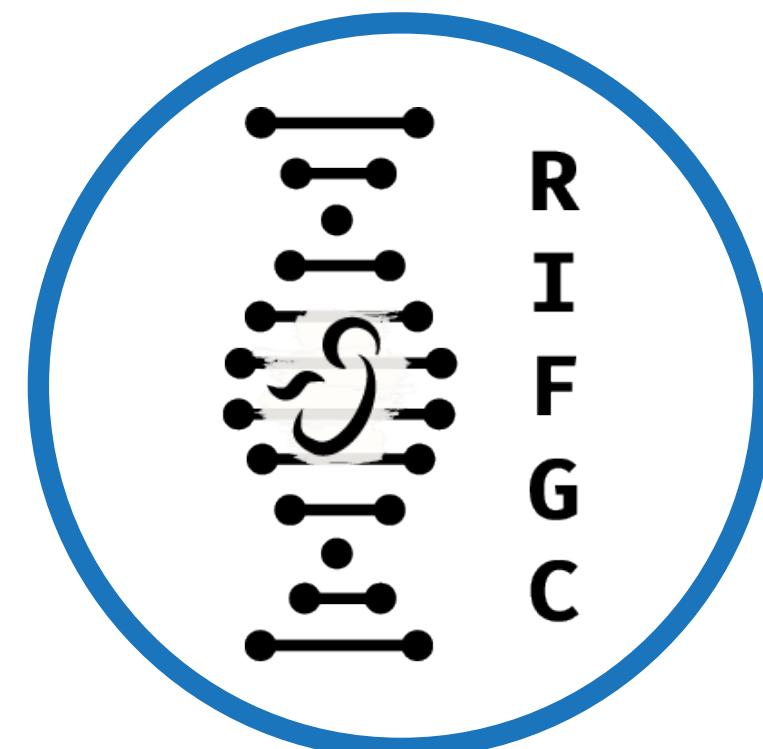
Genomic Data Infrastructure



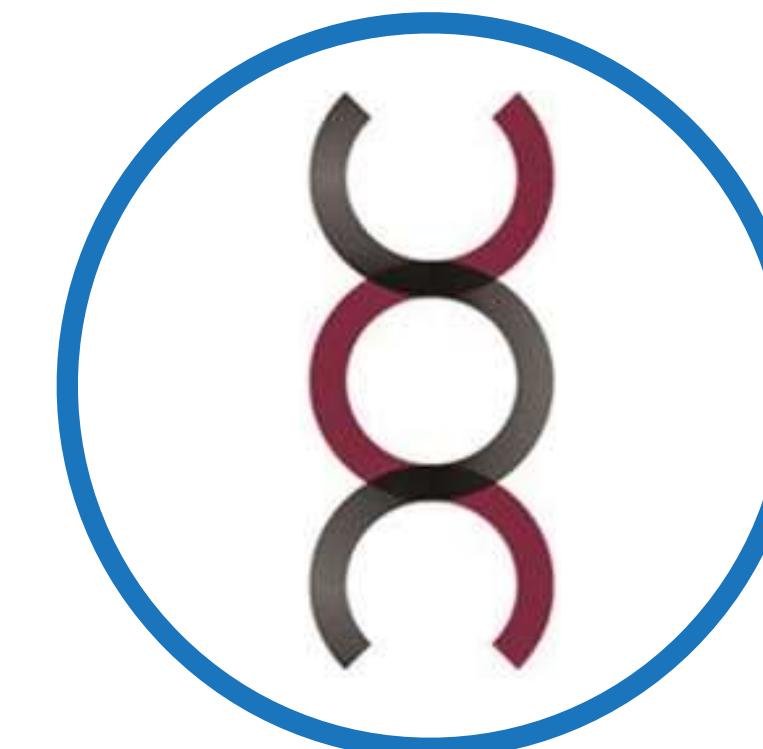
HUMAN
PANGENOME
PROJECT



International Precision
Child Health Partnership



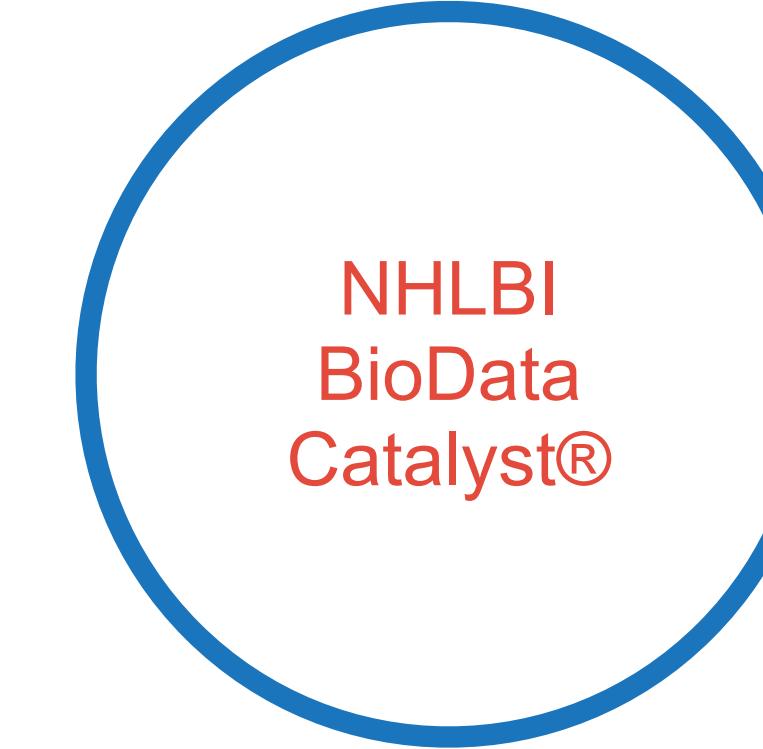
Repository of the International
Fetal Genomics Consortium



Qatar Genome Program



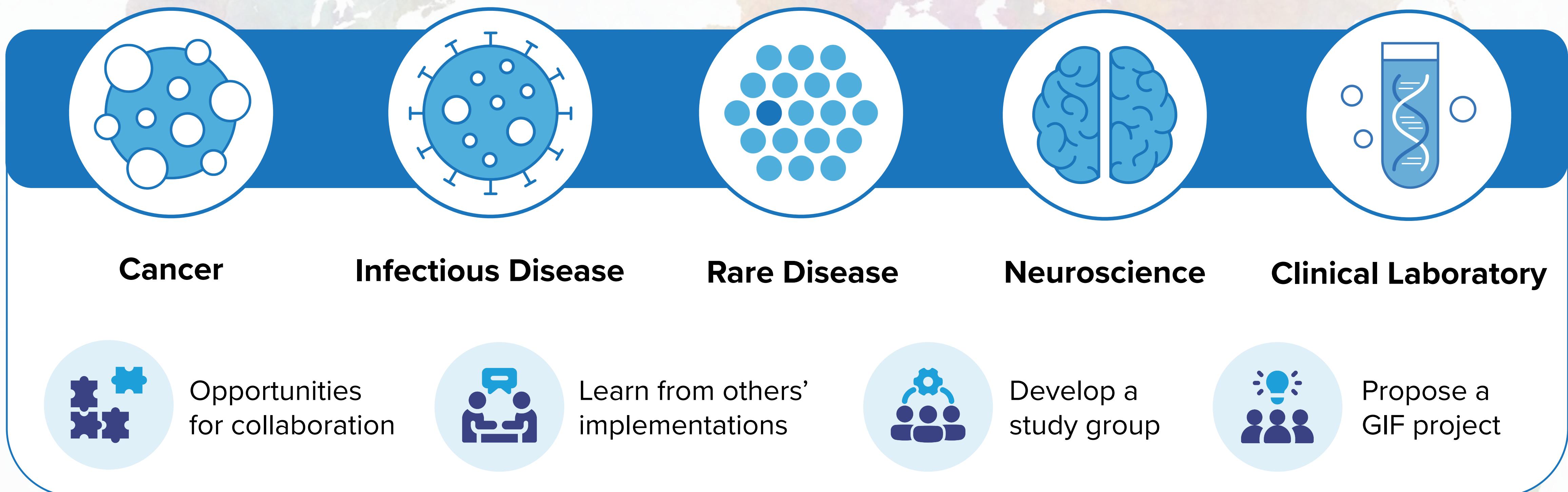
NIH Cloud Platform
Interoperability effort

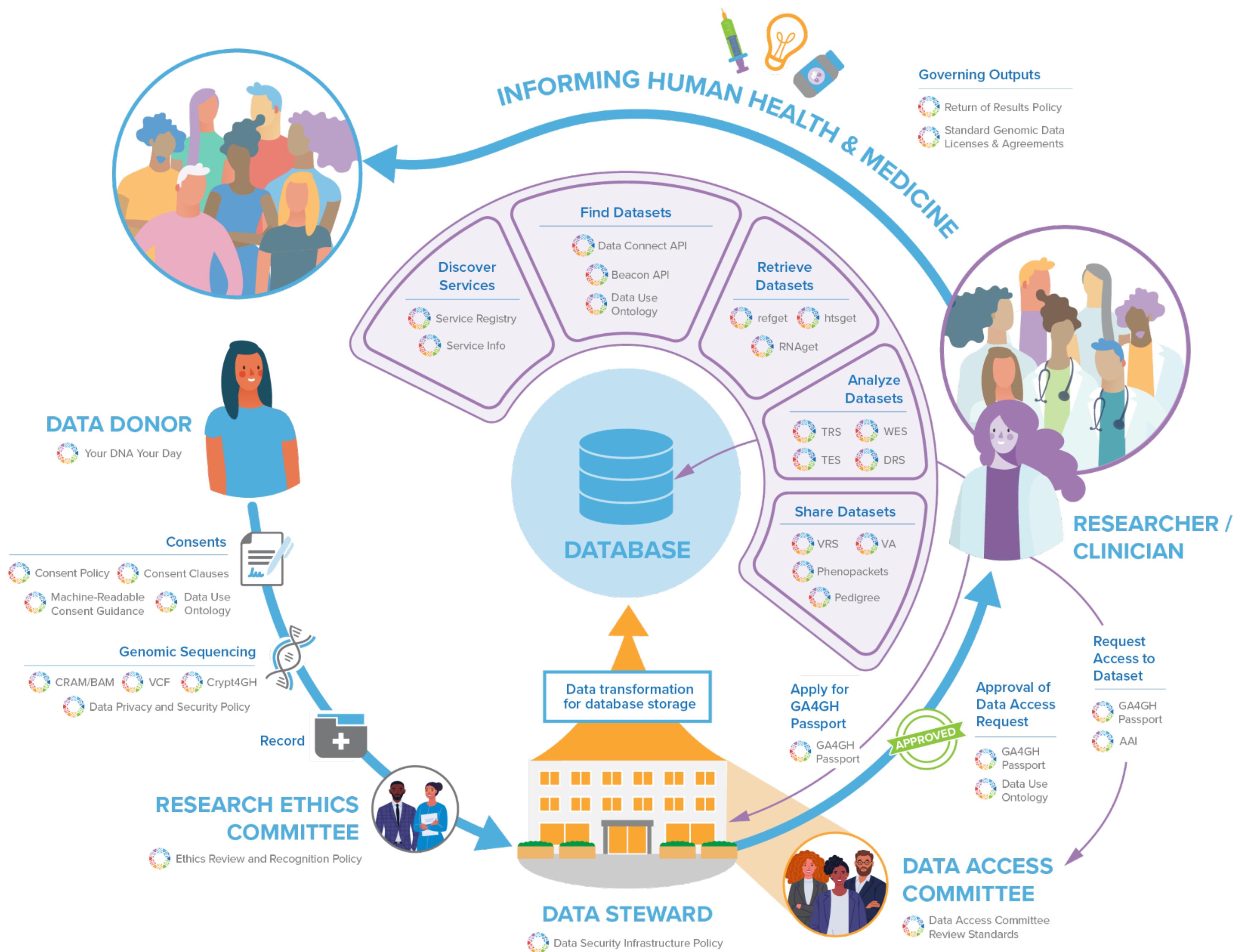


NHLBI BioData
Catalyst®

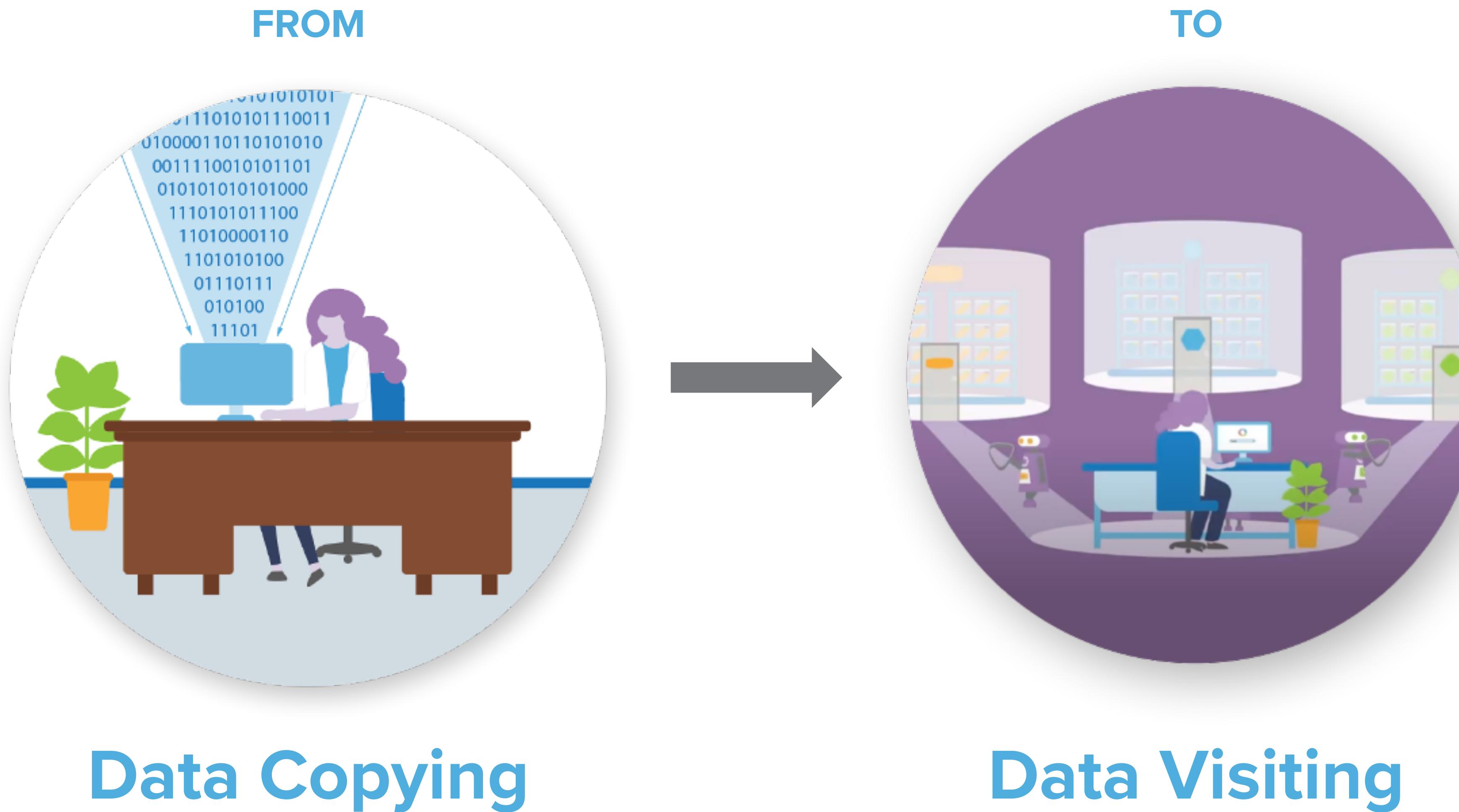


Domain-specific groups promoting global cooperation, data sharing and collaborative research through identifying the need for new standards, and implementing existing GA4GH standards.

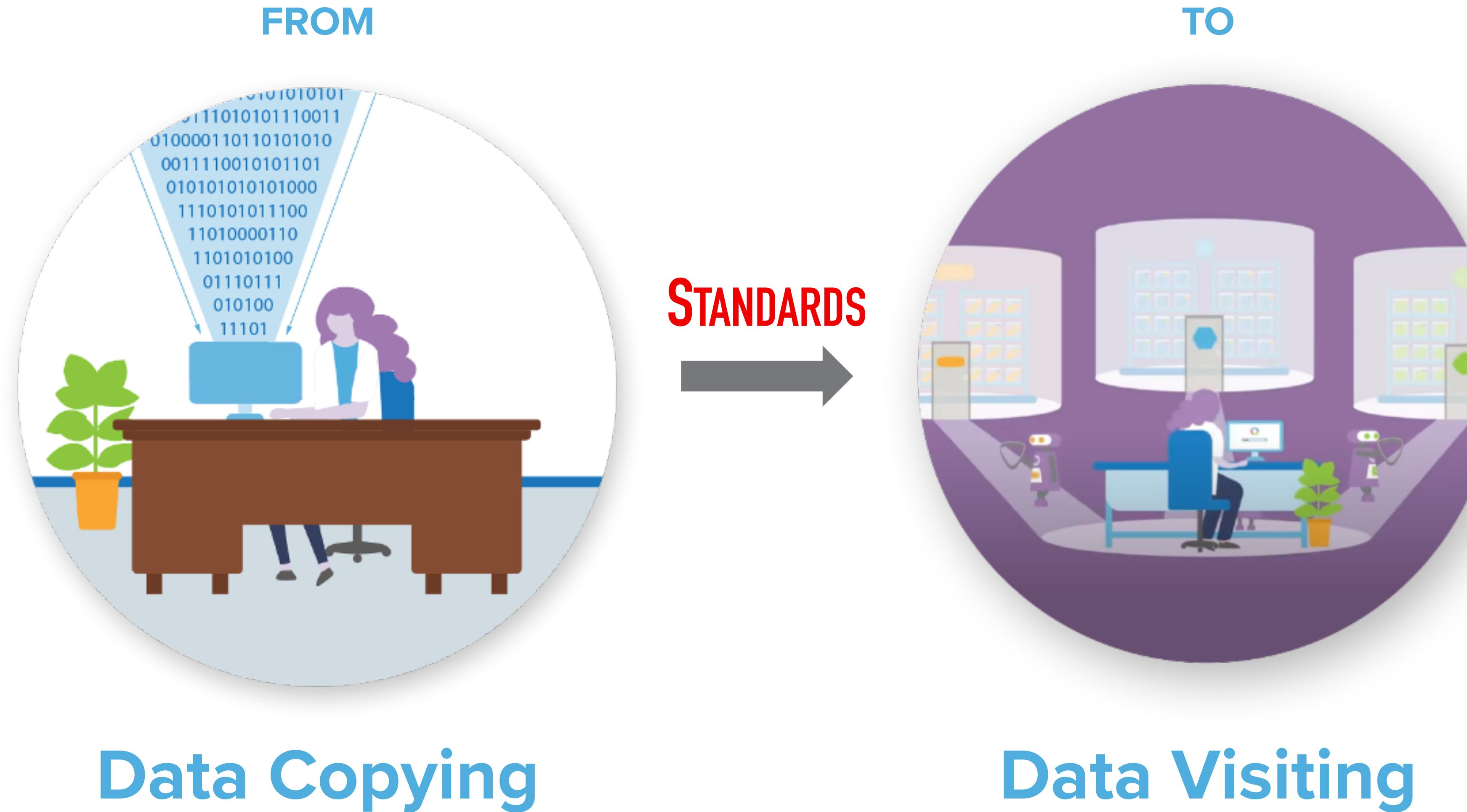




A New Paradigm for Data Sharing



A New Paradigm for Data Sharing



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

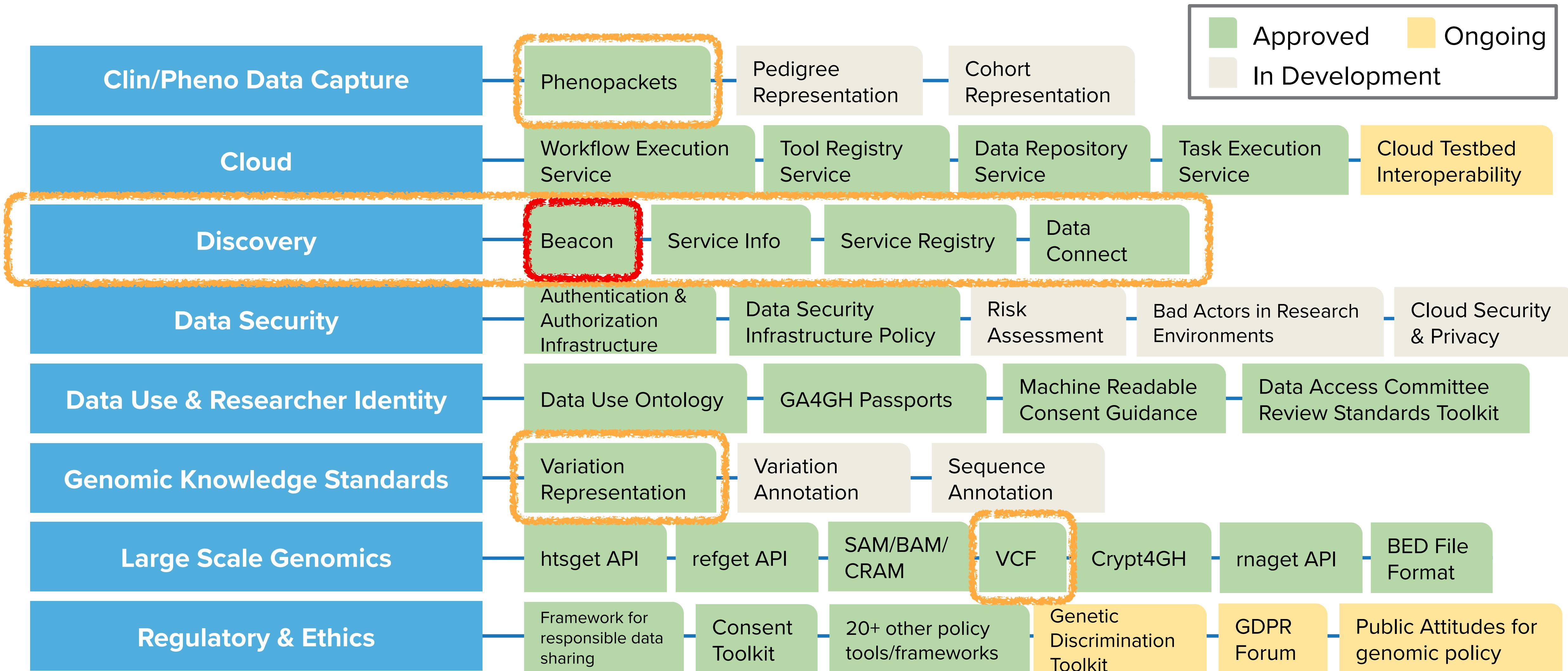
Perspective

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)

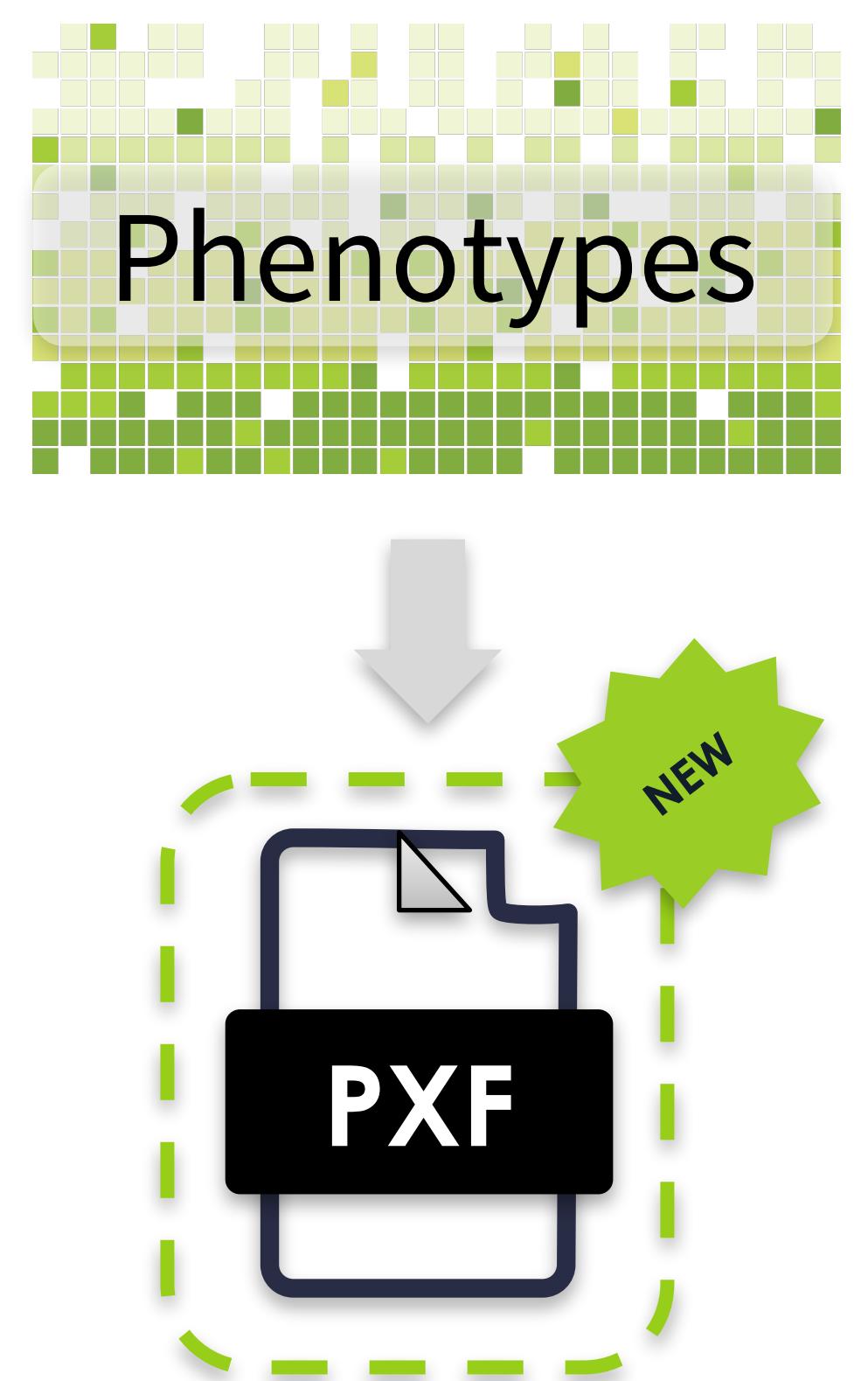
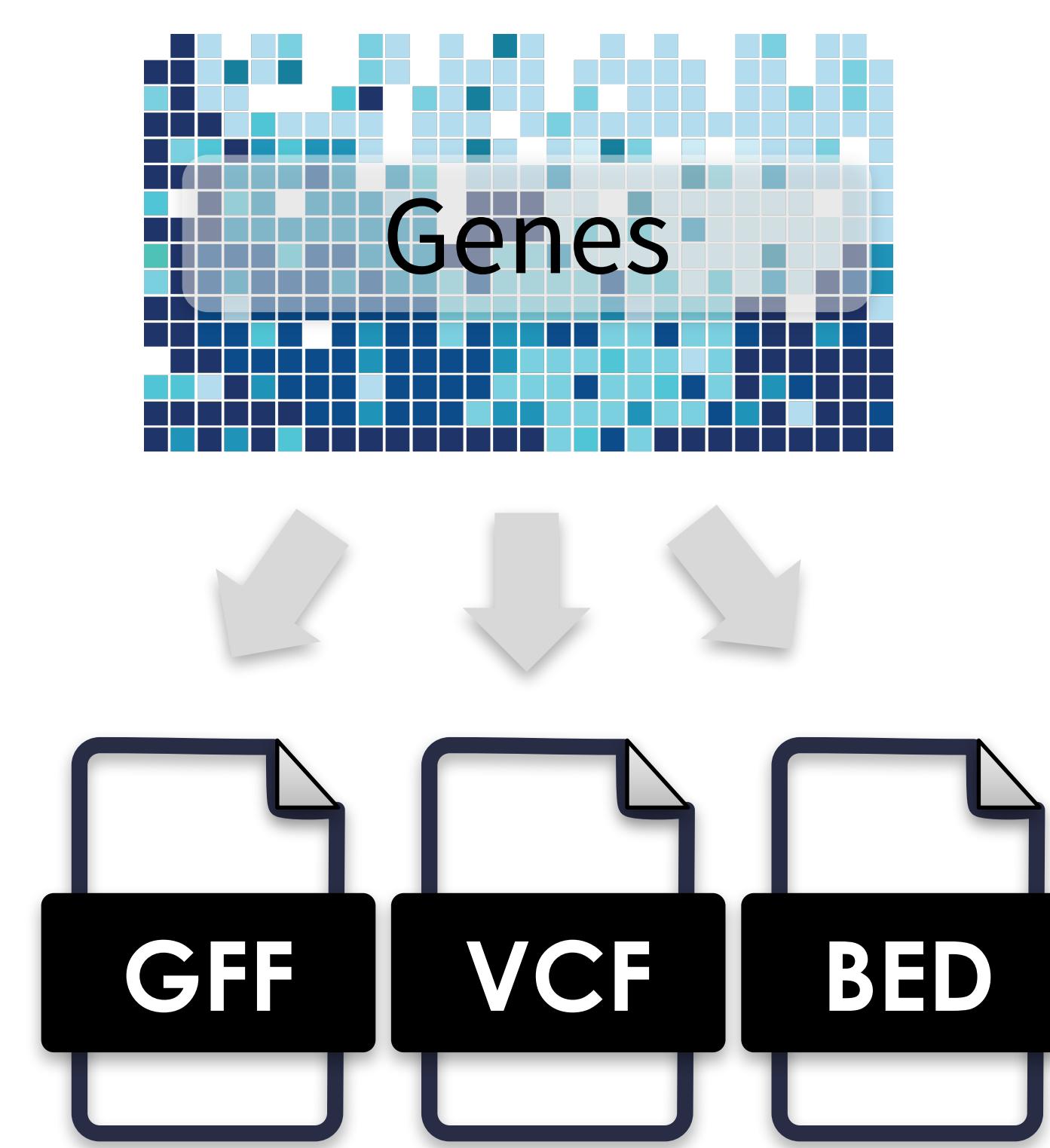
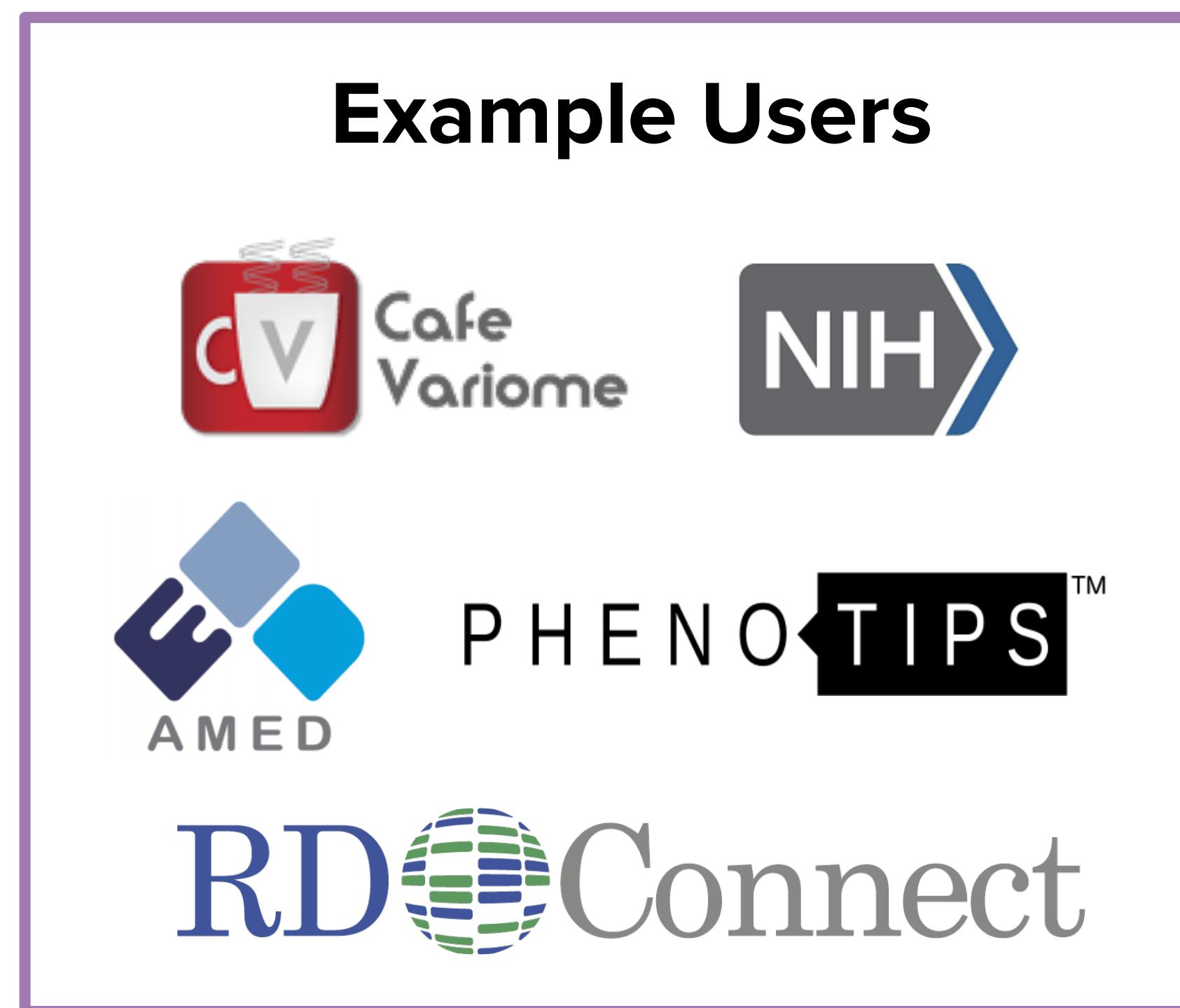
Overview of GA4GH standards and frameworks



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.

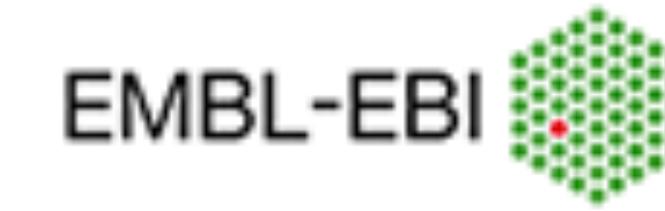
Software Libraries: [htslib](#) | [htsjdk](#)

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

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

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

Example Users





Enable downstream analysis of genomic data by developing common APIs for exchanging reference genomic information.

Proposed Solution

Develop consistency in exchanging information related to genomic sequence, annotation, and associated metadata/provenance.

GA4GH Deliverables



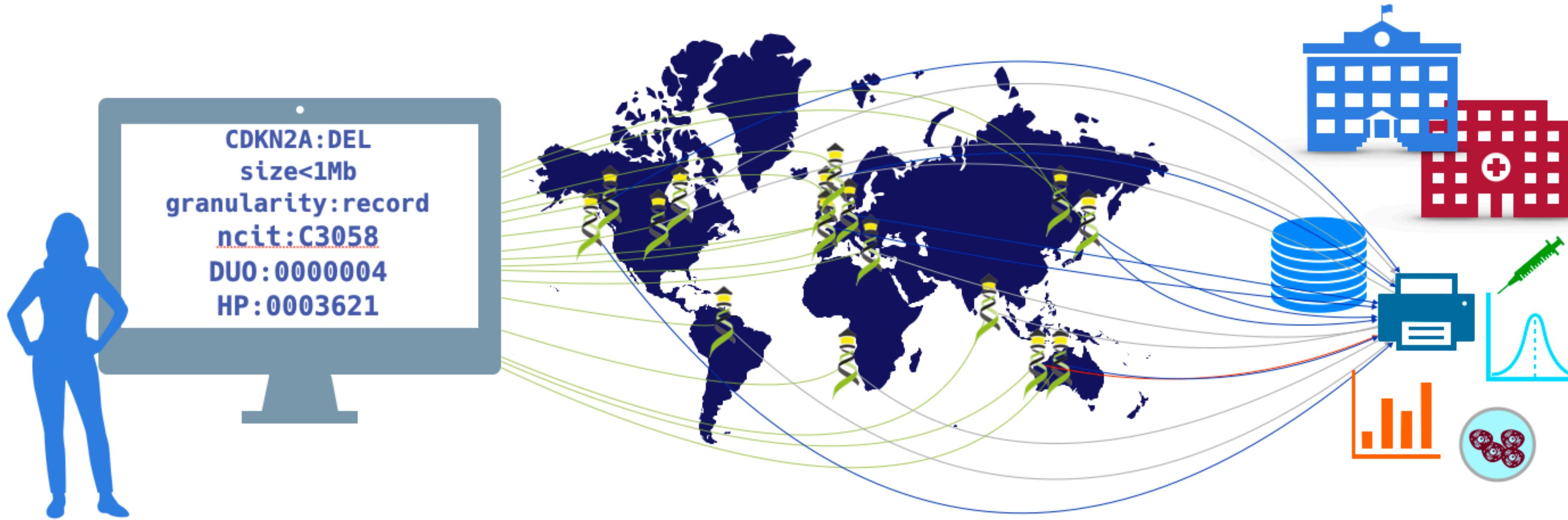
Variation Representation Specification (VRS)
& VRSATILE



Variant Annotation (VA)

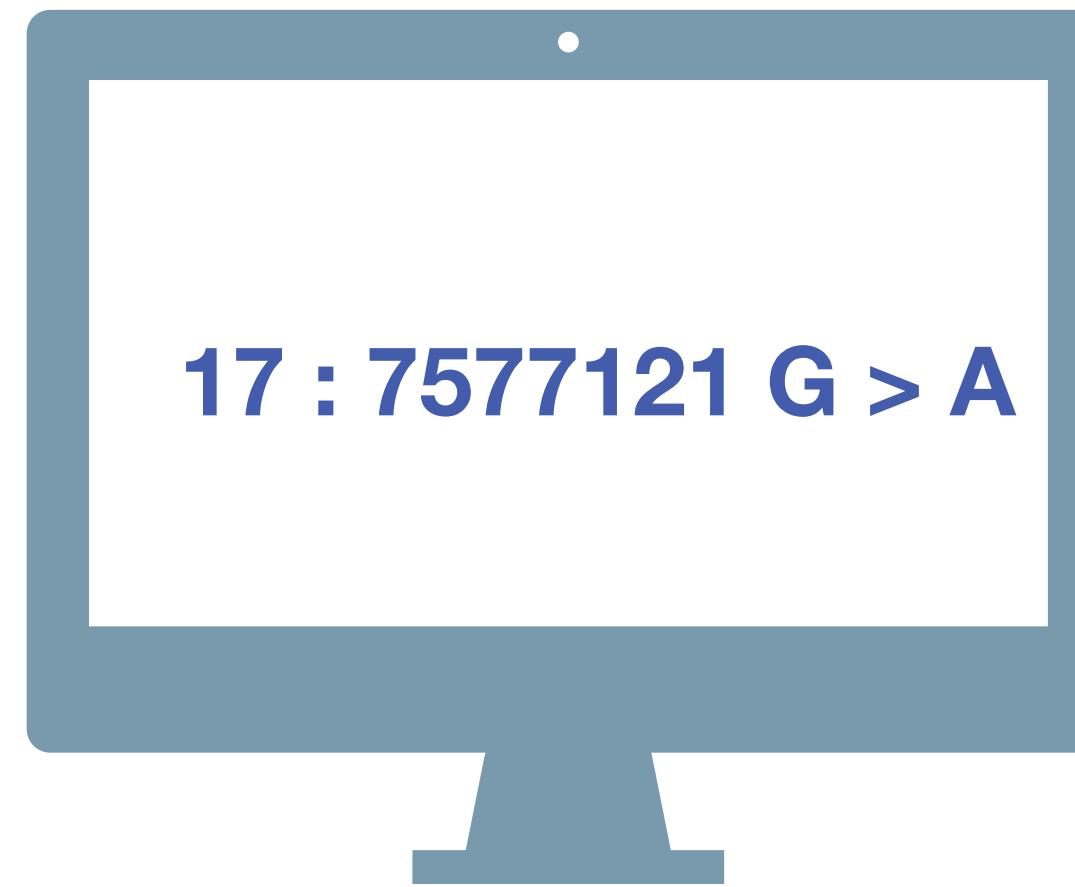


Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.



The GA4GH Beacon Protocol

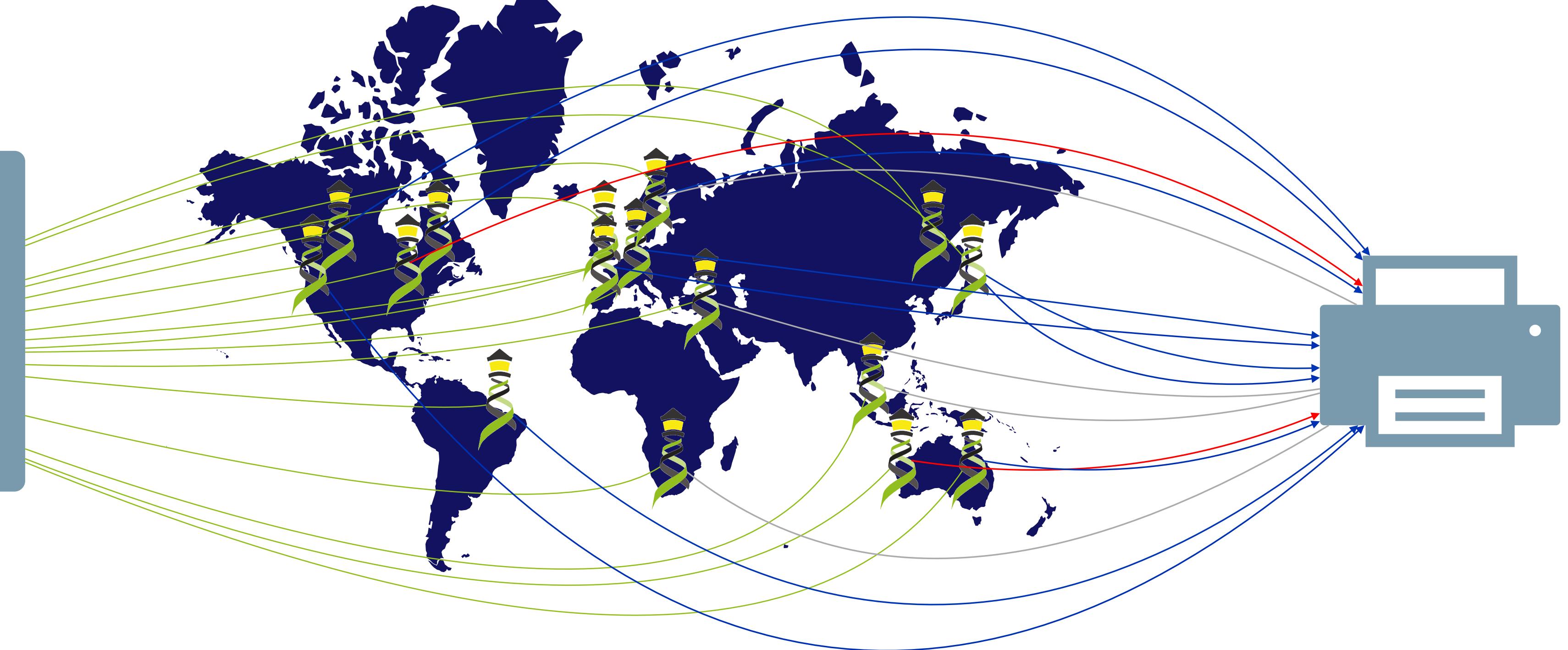
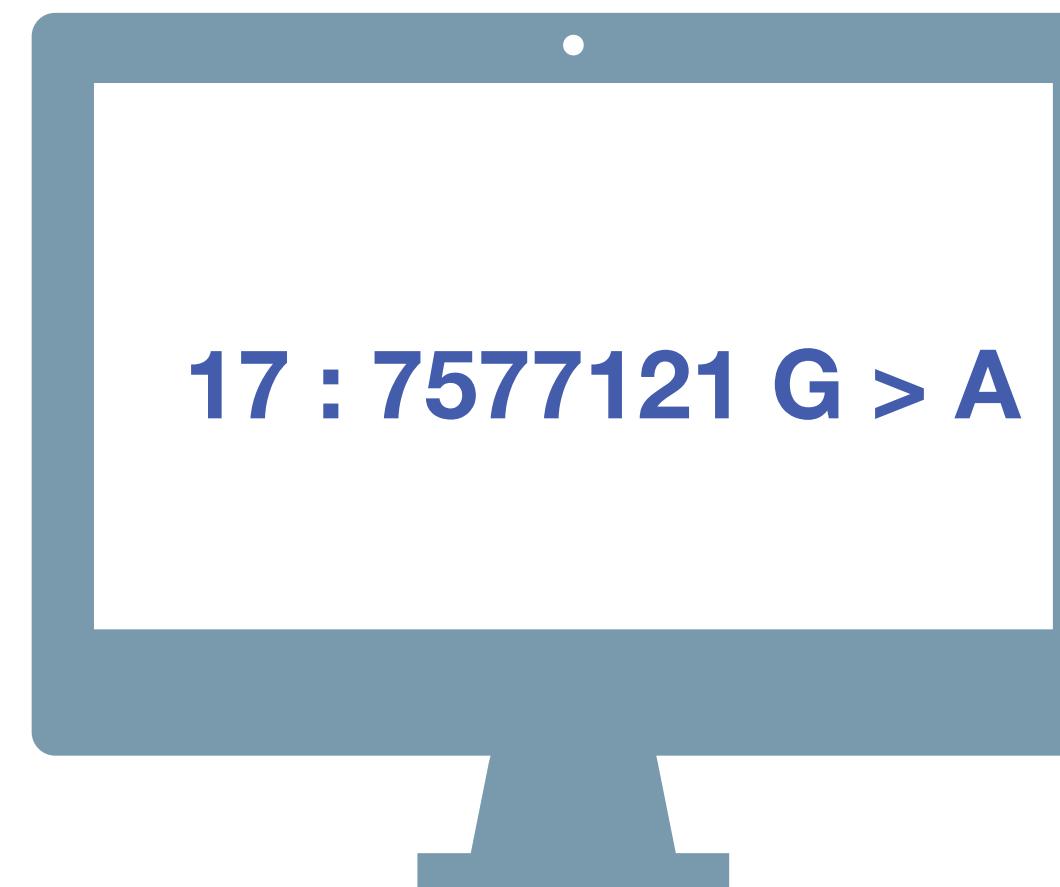
Federating Genomic Discoveries



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 ...

Response	All	None
<input checked="" type="checkbox"/> Found	16	
<input type="checkbox"/> Not Found	27	
<input type="checkbox"/> Not Applicable	22	

BioReference BioReference Hosted by BioReference Laboratories Found

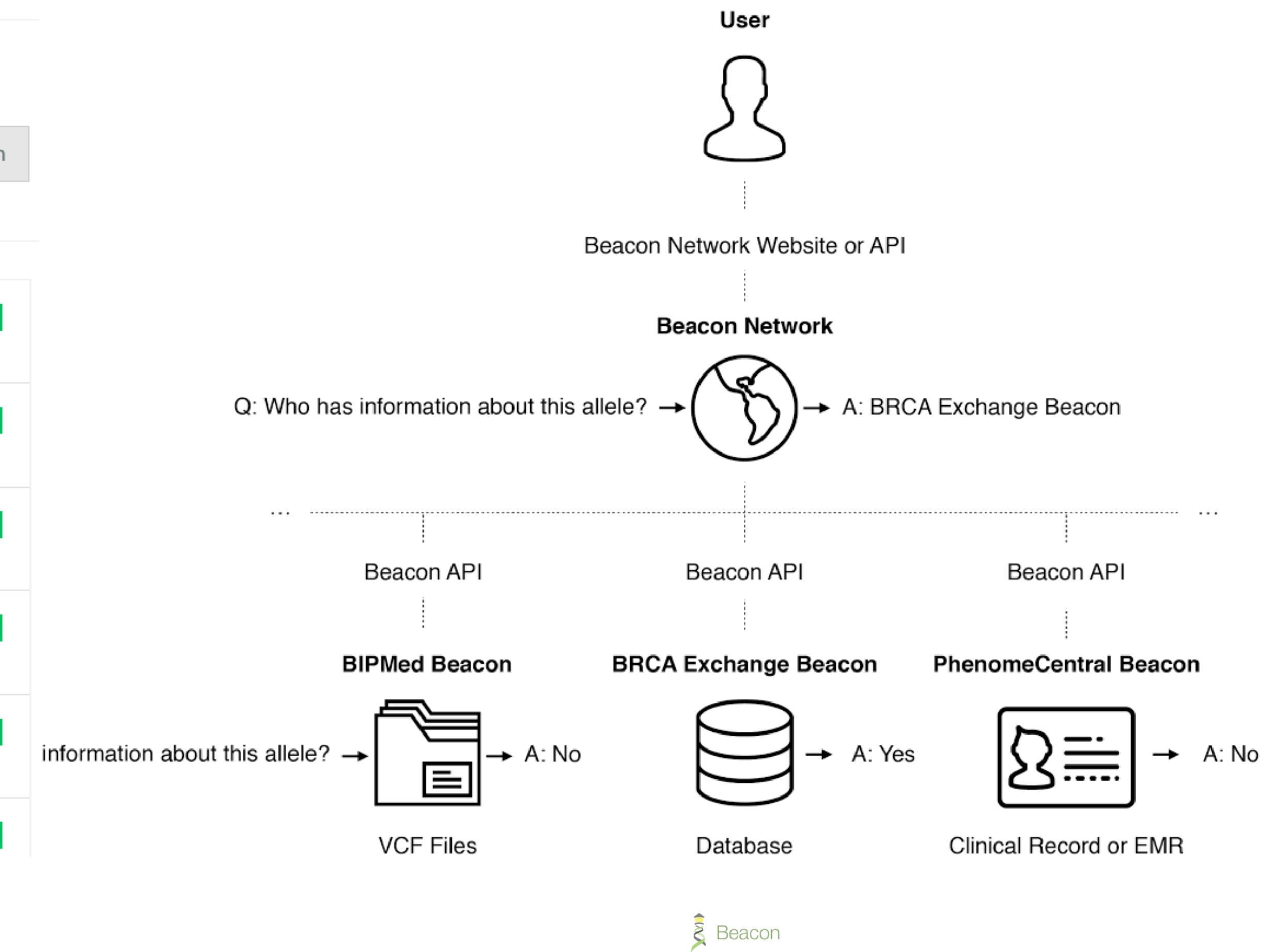
Catalogue of Somatic Mutations in Cancer Catalogue of Somatic Mutations in Cancer Hosted by Wellcome Trust Sanger Institute Found

Cell Lines Cell Lines Hosted by Wellcome Trust Sanger Institute Found

Conglomerate Conglomerate Hosted by Global Alliance for Genomics and Health Found

COSMIC COSMIC Hosted by Wellcome Trust Sanger Institute Found

dbGaP: Combined GRU Catalog and NHLBI Exome Seq... dbGaP: Combined GRU Catalog and NHLBI Exome Seq... Found



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)

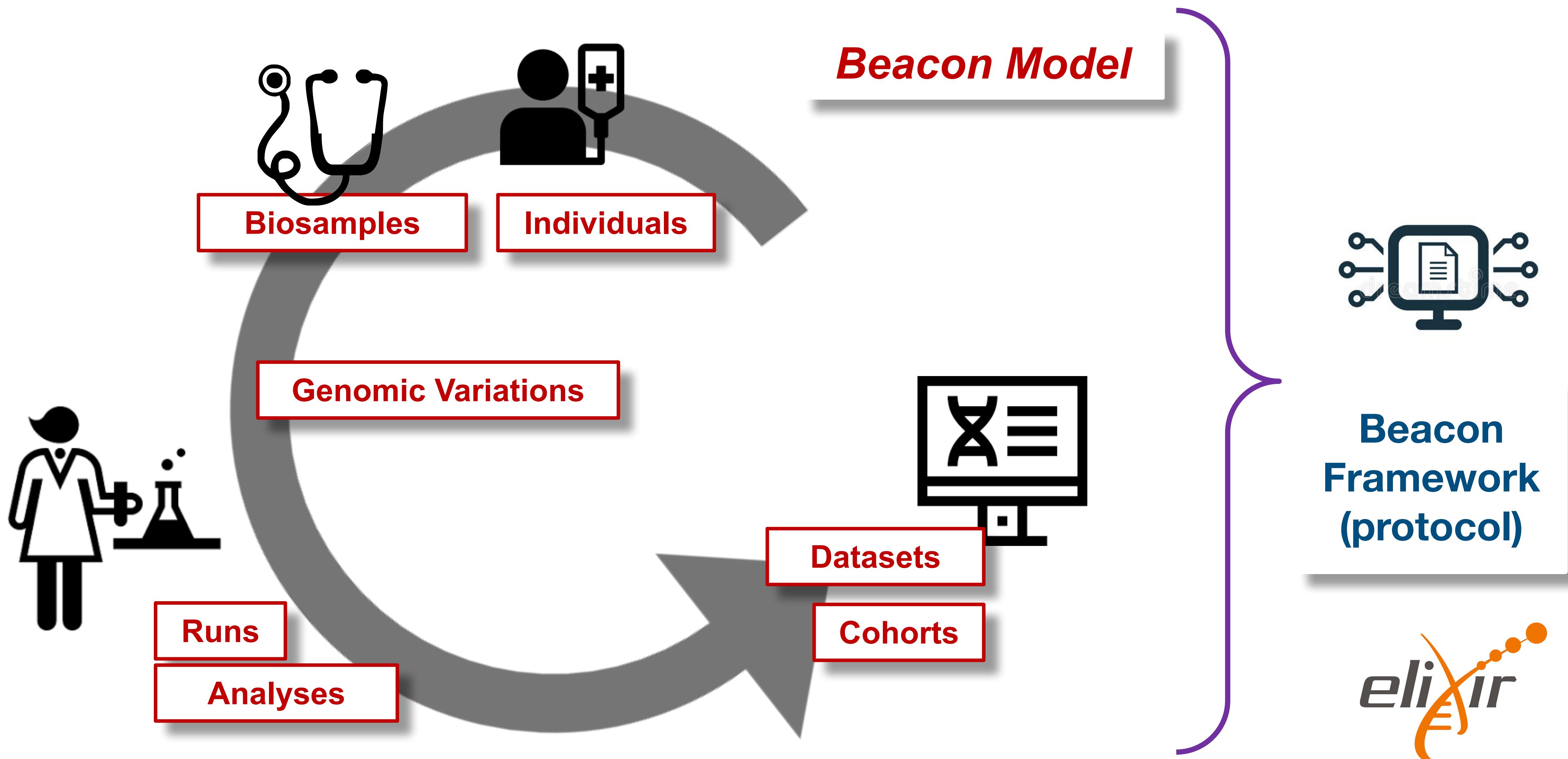
- Beacon publication at Nature Biotechnology

- Phenopackets v2 approved

- docs.genomebeacons.org

Beacon v2

docs.genomebeacons.org

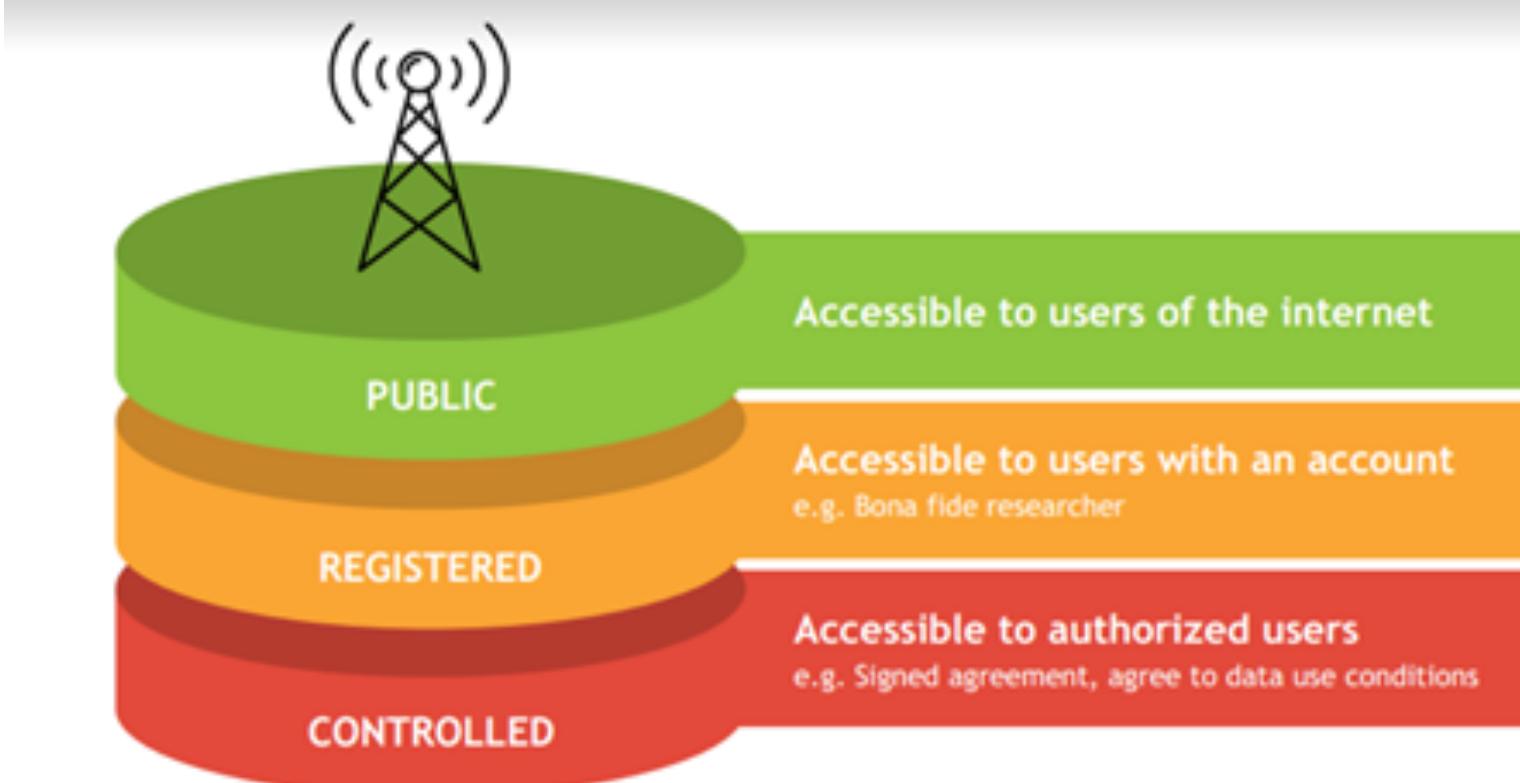


**Beacon
Framework
(protocol)**

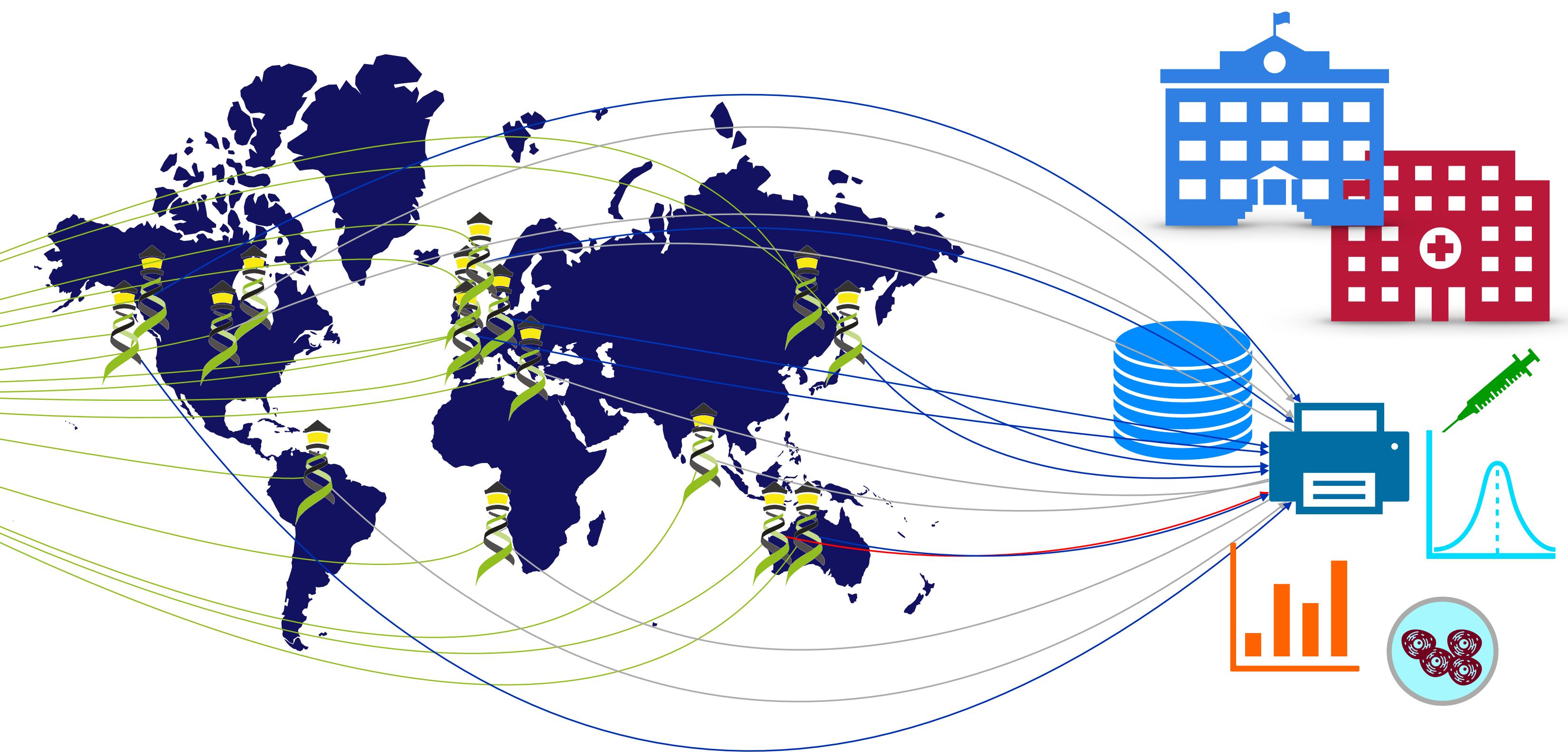
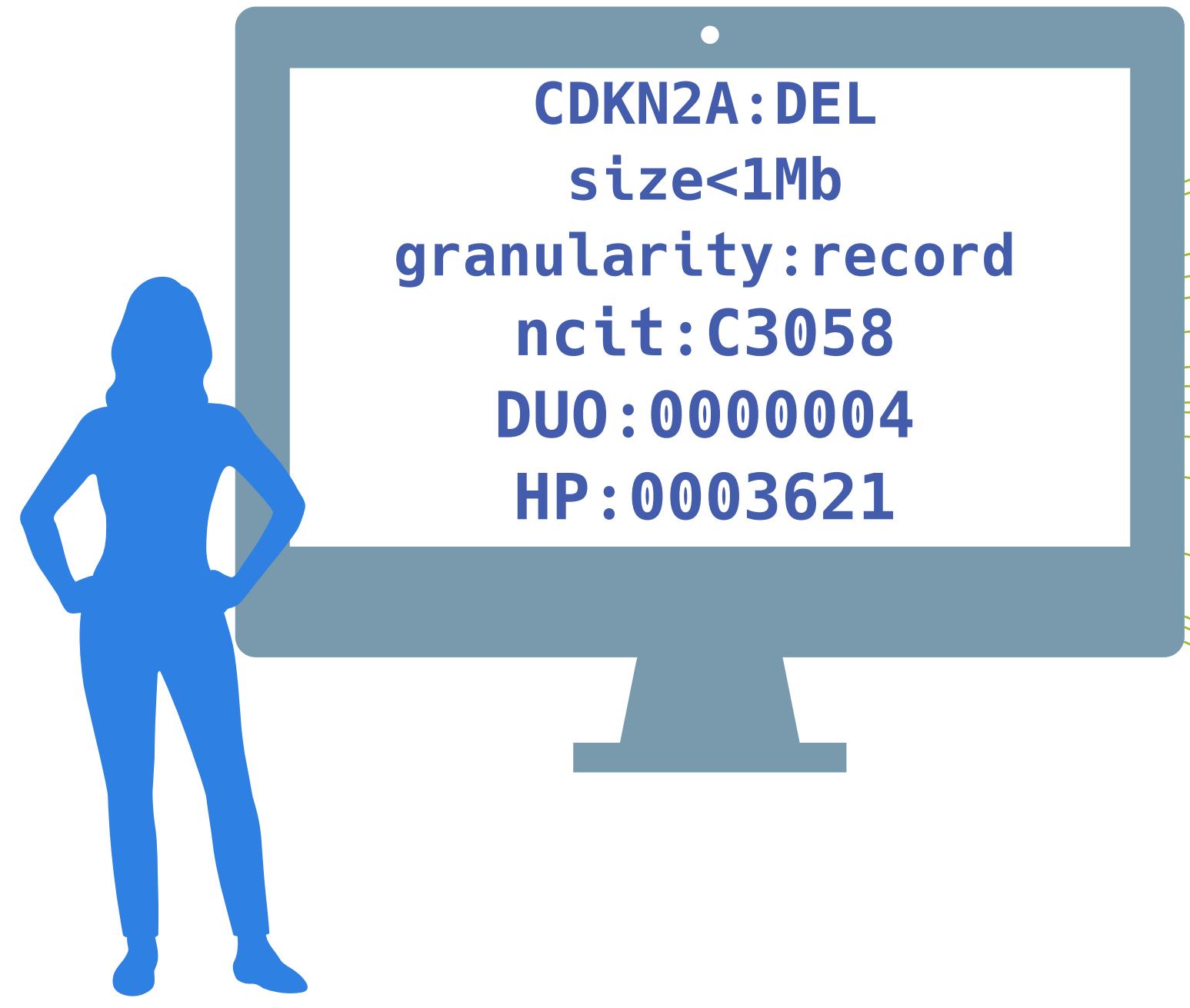


Beacon Security

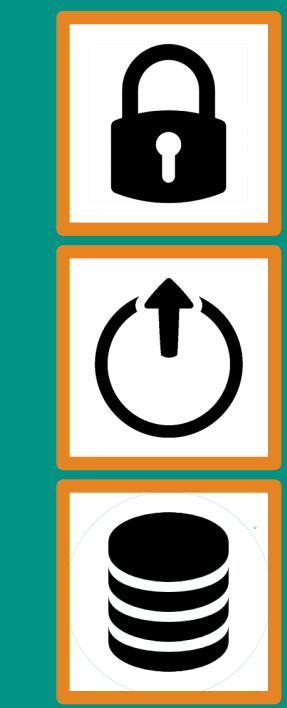
Security by Design ... if Implemented in the Environment



- 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



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



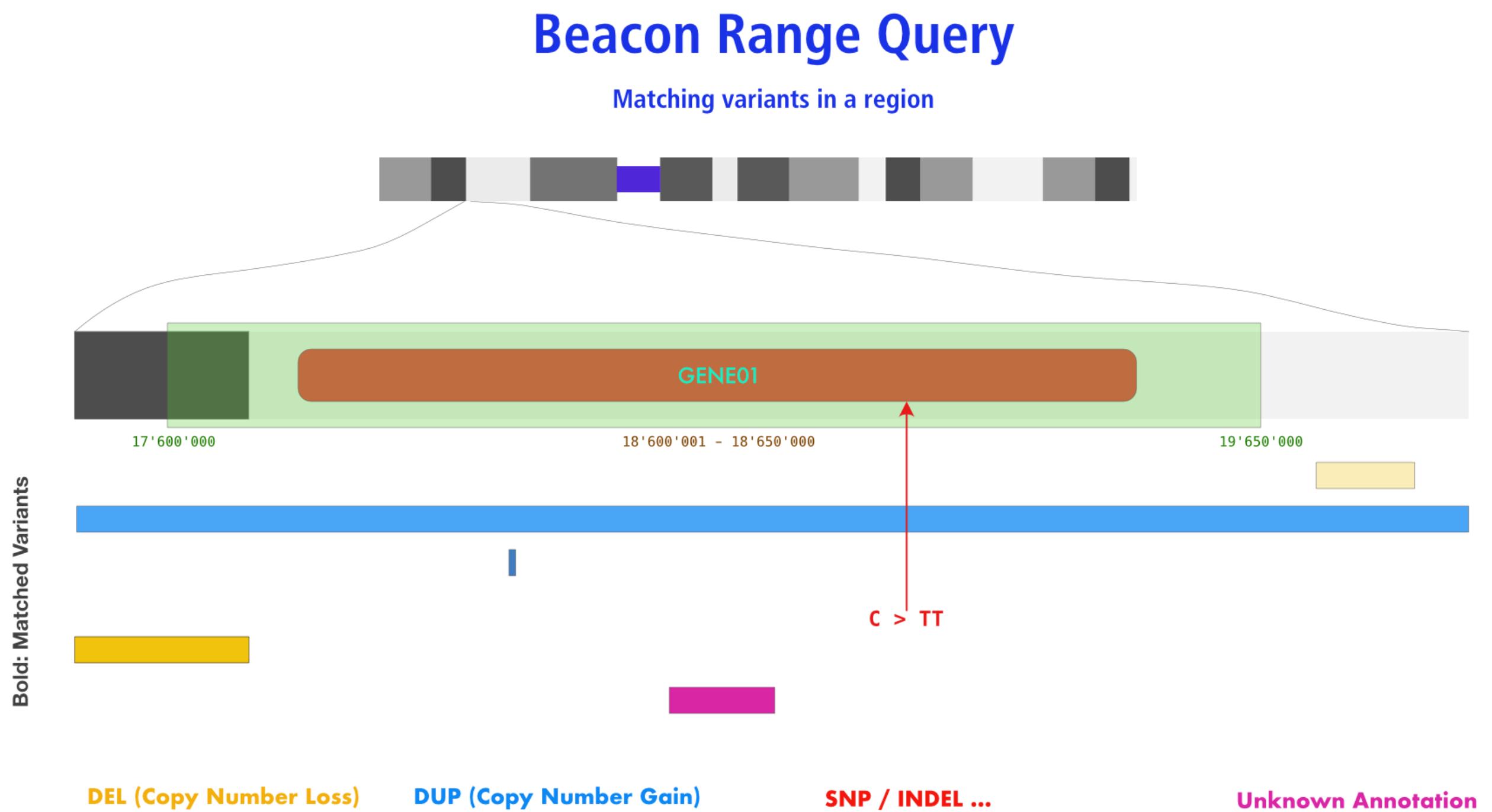
Beacon **v2** API

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)

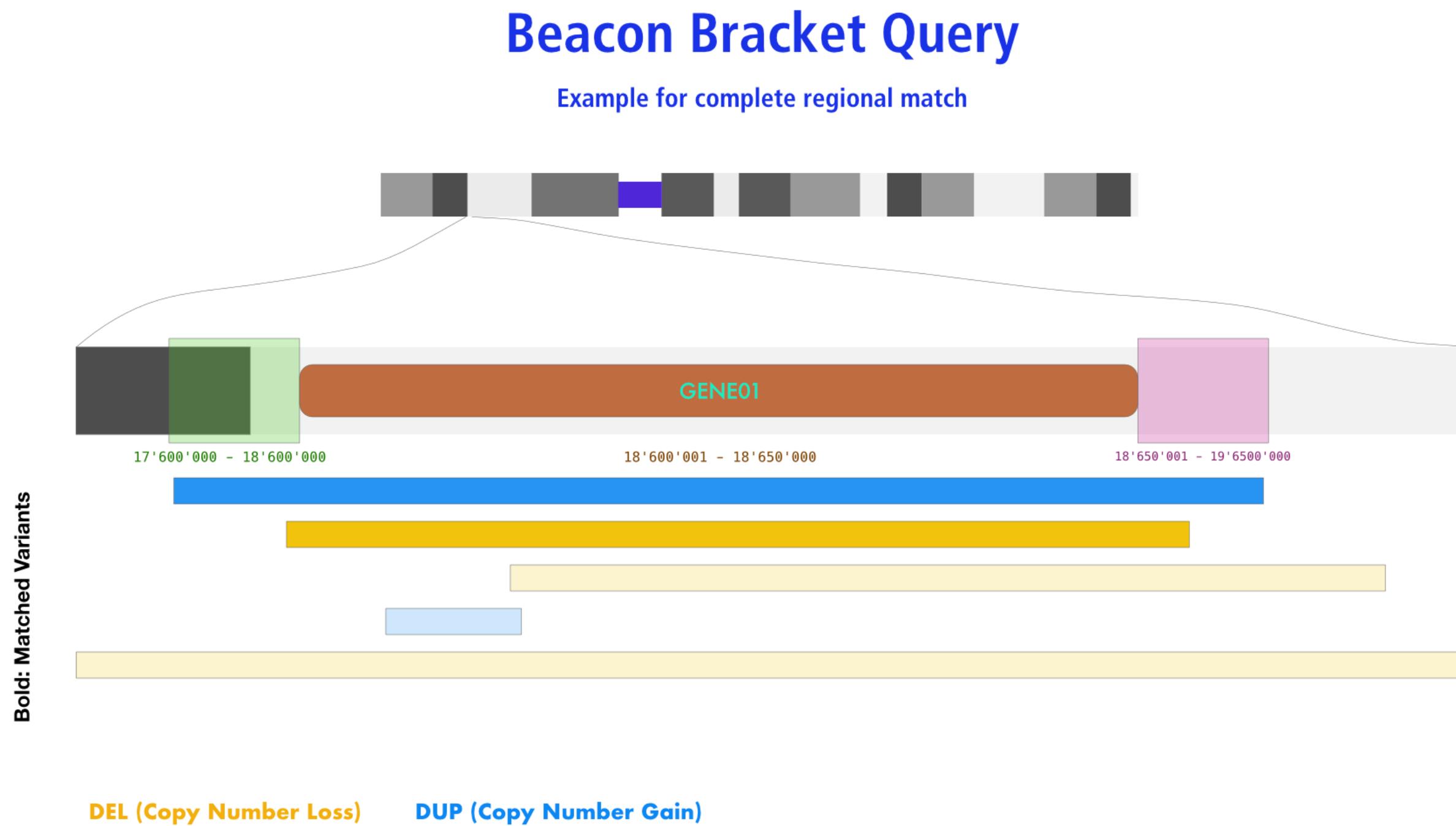
Query Examples: CNV Example, SNV Example, Range Example, Gene Match, Aminoacid Example, Identifier - HeLa

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. $\leq \sim 2\text{Mbp}$ in size). The query is against the examplez collection and can be modified e.g. through changing the position parameters or data source.

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

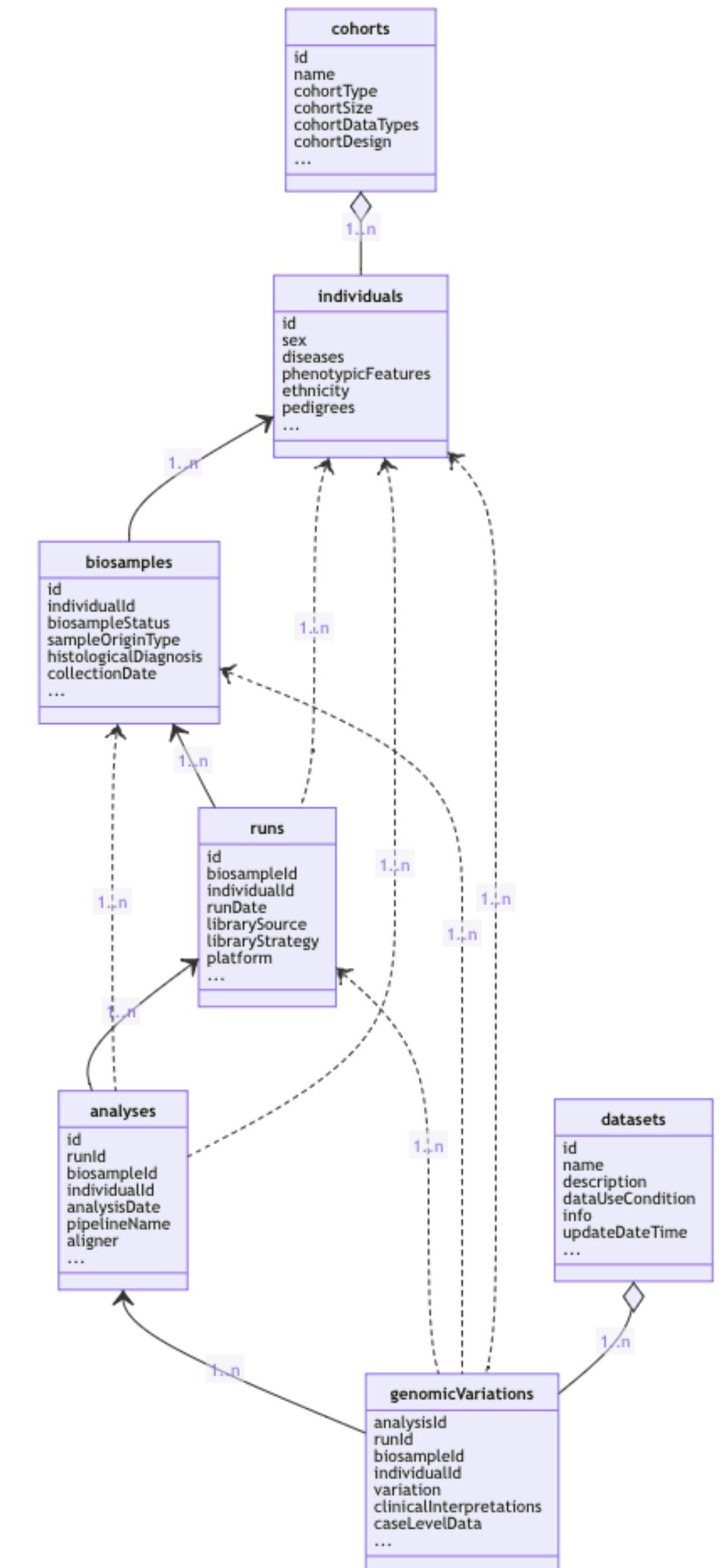
<input checked="" type="checkbox"/>	> NCIT:C4914: Skin Carcinoma	213
<input type="checkbox"/>	> NCIT:C4475: Dermal Neoplasm	109
<input checked="" type="checkbox"/>	> NCIT:C45240: Cutaneous Hematopoietic and Lymphoid Cell Neoplasm	310

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	$f_{alleles}$: 0	Callsets	Variants	UCSC region	Legacy Interface	 Show JSON Response	
Calls: 0							
Samples: 523							
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			

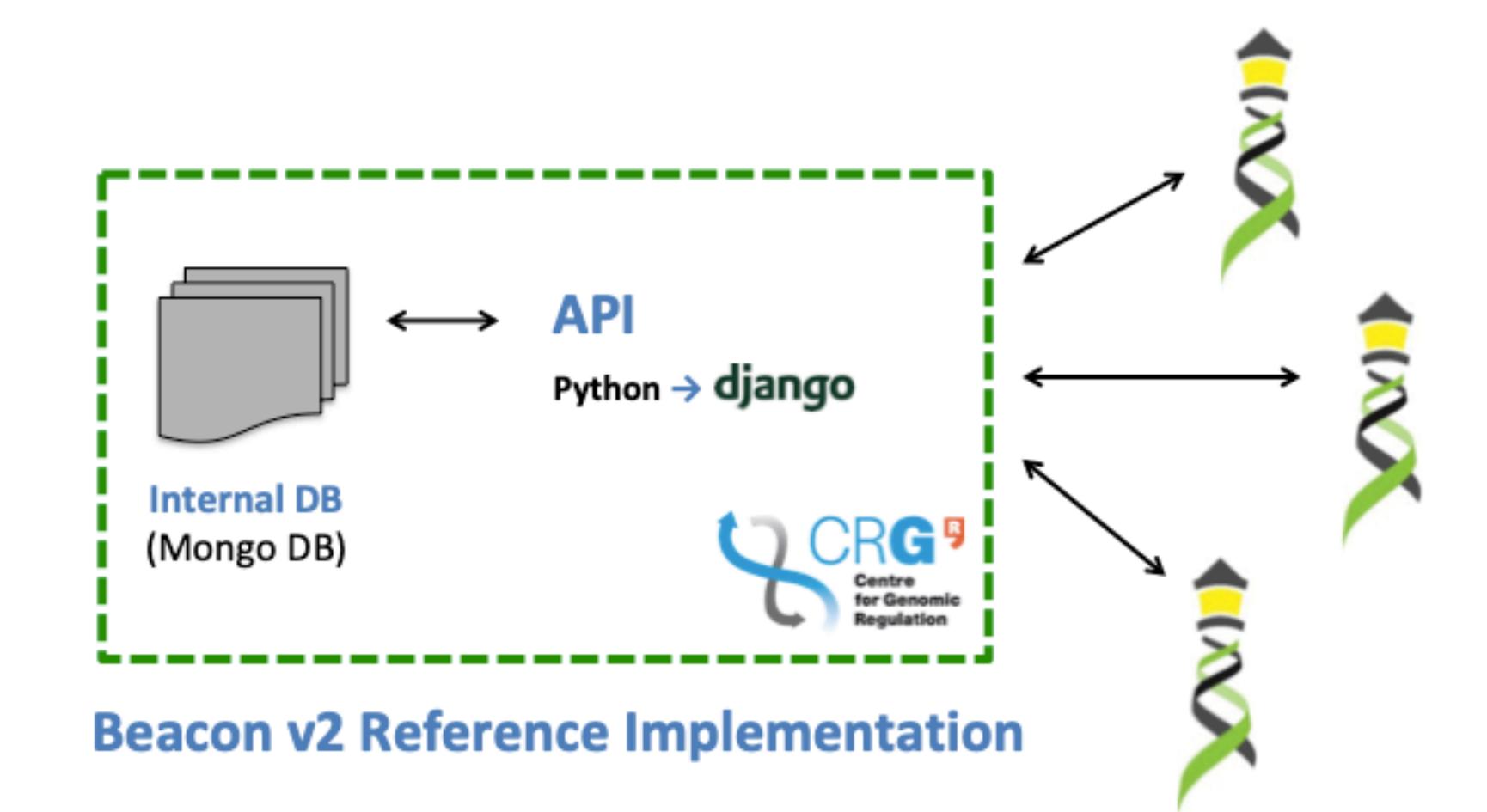
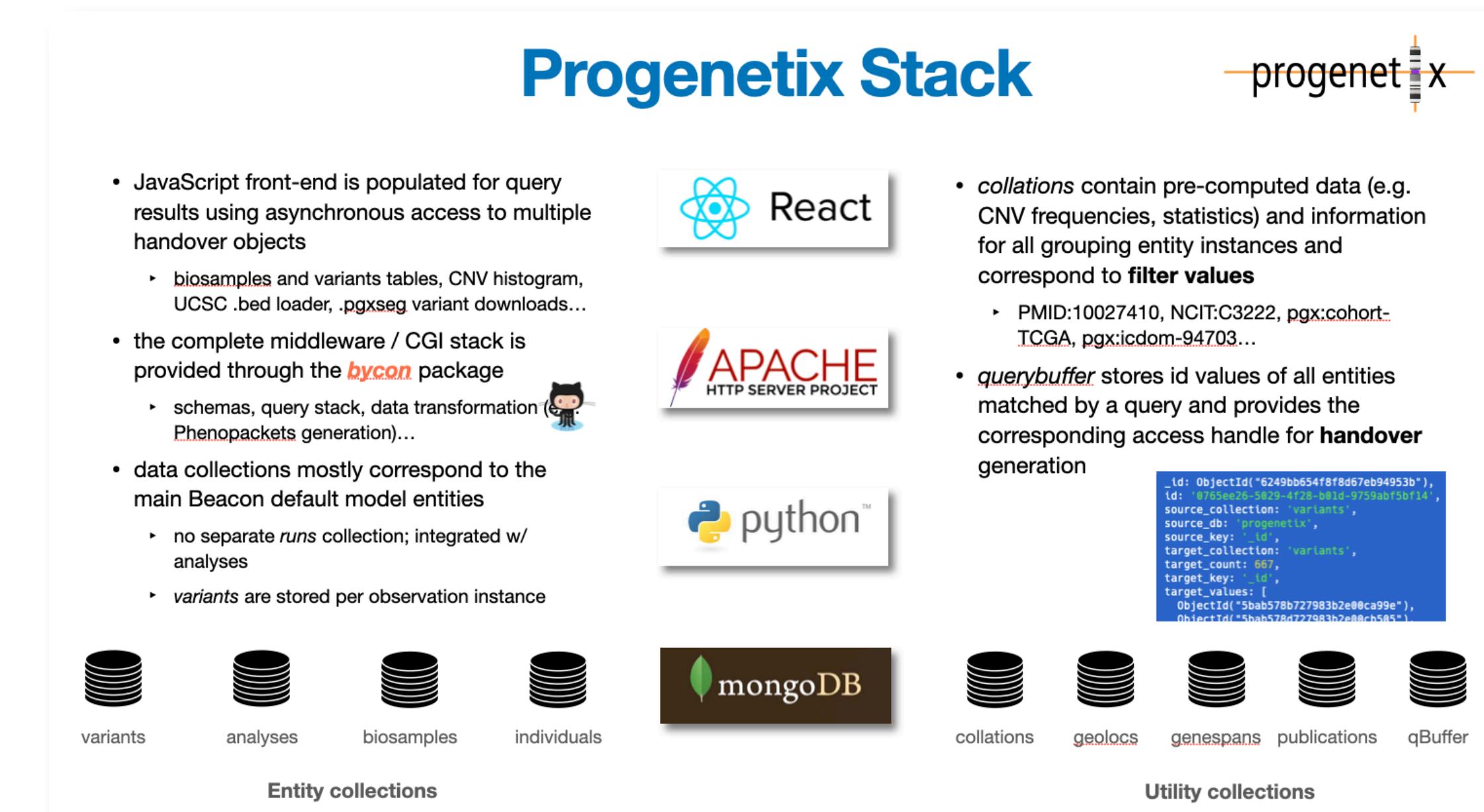
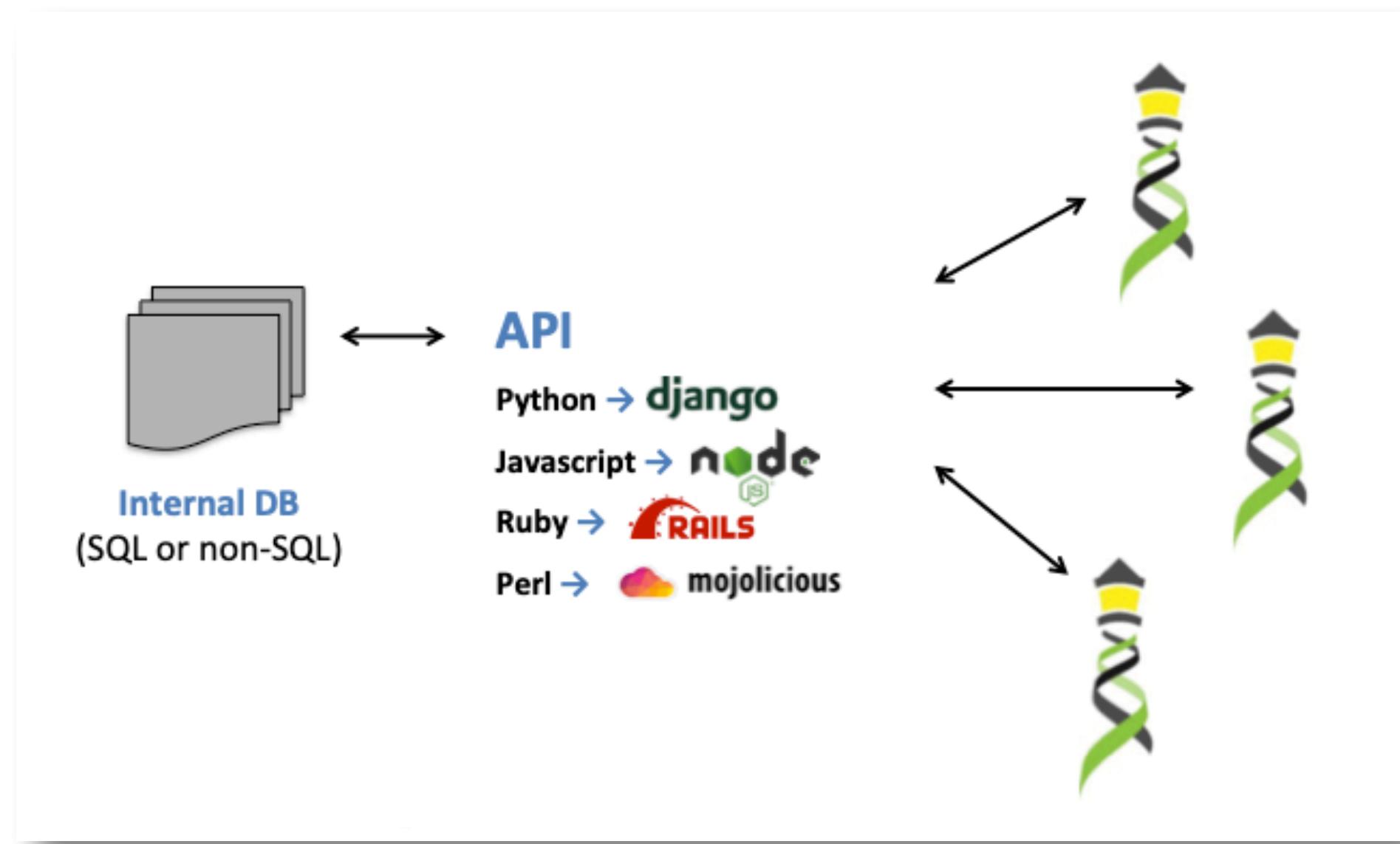
Beacon Default v2 Model

- 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	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	University of Leicester
BeaconMap	Green	Green
Bioinformatics analysis	White	White
Biological Sample	Red	White
Cohort	White	White
Configuration	Green	White
Dataset	Red	White
EntryTypes	Green	White
Genomic Variants	White	White
Individual	Red	White
Info	White	White
Sequencing run	White	White

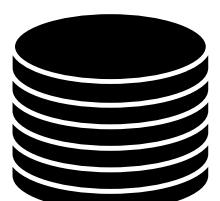
Green: Matches the Spec, Red: Not Match the Spec, White: Not Implemented



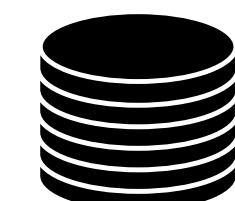
bycon based Progenetix Stack



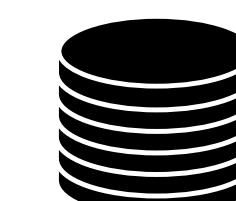
- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
 - ▶ biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads...
- the complete middleware / CGI stack is provided through the *bycon* package
 - ▶ schemas, query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



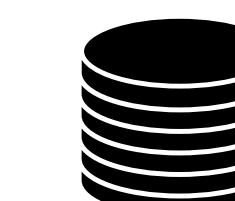
variants



analyses



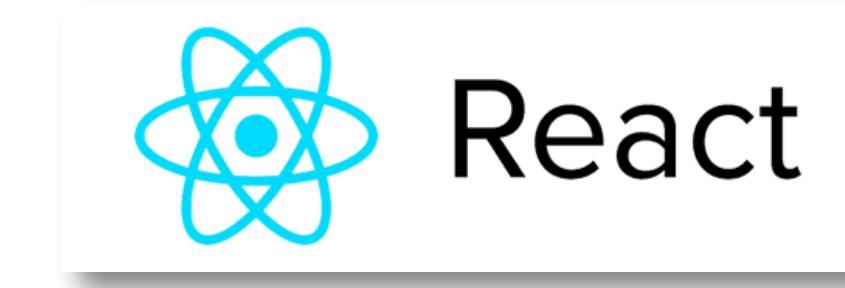
biosamples



individuals

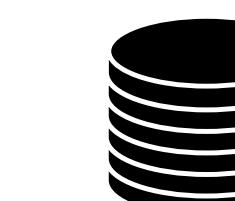


Entity collections

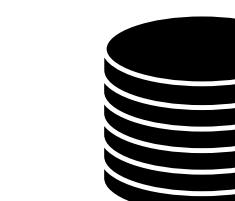


- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
 - ▶ PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703...
- *querybuffer* stores id values of all entities matched by a query and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

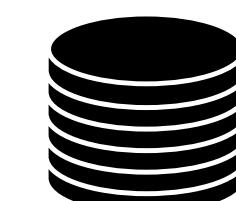
```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
  ObjectId("5bab578d727983b2e00cb505")]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



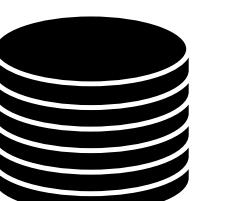
collations



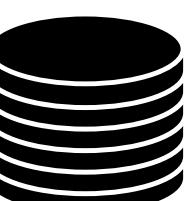
geolocs



genespans



publications



qBuffer

Utility collections

Beacon⁺: Phenopackets

Testing alternative response schemas...

<http://progenetix.org/beacon/phenopackets/pgxind-kftx26j0>

- 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/pgxbss-kftvhvzb/variants/?output=pgxseg"
          }
        ],
        "histologicalDiagnosis": {
          "id": "NCIT:C3099",
          "label": "Hepatocellular Carcinoma"
        },
        "id": "pgxbss-kftvhvzb",
        "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'. The repository is public, has 1 branch, and 0 tags. The main branch has 87 commits from user 'hangjiaz' made 4 days ago. The commits are listed in a table:

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

2 Retrieve metadata of samples

2.1 Relevant parameters

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

2.2 Search by filters

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

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

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

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

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

Client for Accessing Progenetix

pgxRpi: an R/Bioconductor package

- **Query and export variants**

https://progenetix.org/beacon/biosamples/pgxbs-kftvh94d/g_variants

```
> variants <- pgxLoader(type="variant",biosample_id="pgxbs-kftvh94d")
```

- **Query metadata of biosamples and individuals by filters (e.g. NCIt, PMID)**

<http://progenetix.org/services/sampletable/?filters=NCIT:C3697>

```
> biosamples <- pgxLoader(type="biosample",filters="NCIT:C3697")
```

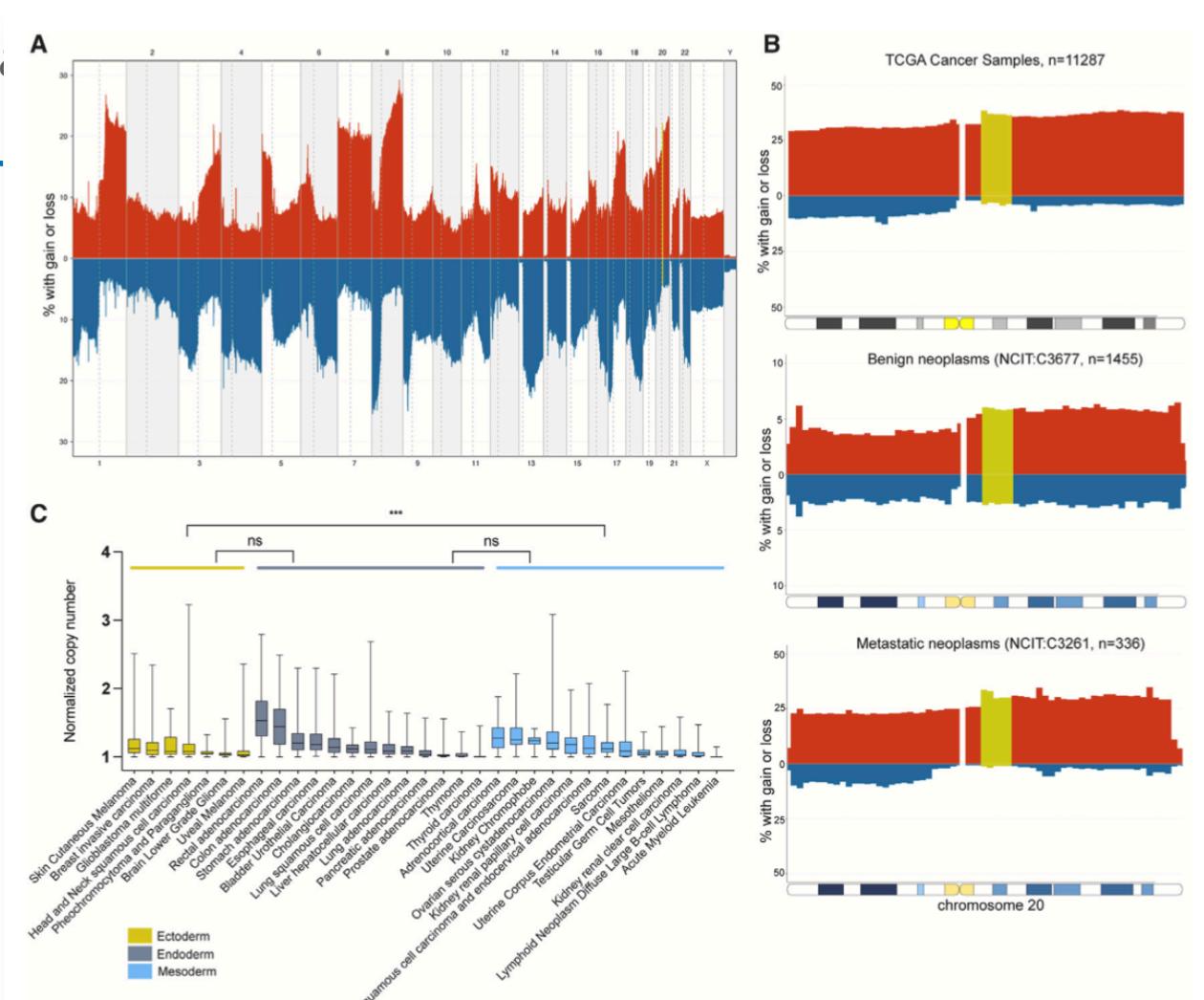
- **Query and visualize CNV frequency by filters**

<http://www.progenetix.org/services/intervalFrequencies/?filters=NCIT:C35126>

```
> freq <- pgxLoader(type="frequency",output="pgxfreq",filter  
> pgxFreqplot(freq)
```

- **Process local .pgxseg files**

```
> info <- pgxSegprocess(file=file, show_KM_plot = T,  
return_seg = T, return_metadata = T, return_frequency = T)
```



pgxRpi

This is the **development** version of pgxRpi; for the stable release version, see [pgxRpi](#).

R wrapper for Progenetix

platforms all rank 2178 / 2266 support 0 / 0 in Bioc < 6 months build unknown updated < 1 month dependencies 137
DOI: [10.18129/B9.bioc.pgxRpi](https://doi.org/10.18129/B9.bioc.pgxRpi)

Bioconductor version: Development (3.20)

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>

Use case: 2024 article using Progenetix' **pgxRpi** to retrieve & visualize 117'587 cancer CNV profiles for a study into pluripotent stem cells' genomics

Stem Cell Reports Review



OPEN ACCESS

Gains of 20q11.21 in human pluripotent stem cells: Insights from cancer research

Nuša Krivec,^{1,2} Manjusha S. Ghosh,^{1,2} and Claudia Spits^{1,2,*}

¹Research Group Reproduction and Genetics, Faculty of Medicine and Pharmacy, Vrije Universiteit Brussel, Brussels, Laarbeeklaan 103, 1090 Brussels, Belgium

²These authors contributed equally

*Correspondence: claudia.spits@vub.be

<https://doi.org/10.1016/j.stemcr.2023.11.013>

Figure 2. Copy-number alterations of human chromosome 20q11.21 in cancers

(A) Aggregated copy-number variation (CNV) data of 117,587 neoplasms (NCIT: C3262) from the Progenetix database (Huang et al., 2021) were plotted using R library pgxRpi. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079–35871578 is marked in moss green. NCIT, National Cancer Institute Thesaurus.

(B) Top to bottom: Aggregated CNV data of 11,287 TCGA cancer samples, 336 metastatic neoplasms (NCIT: C3261), and 1,455 benign neoplasms (NCIT: C3677) from the Progenetix database (Huang et al., 2021), respectively, were plotted using R library pgxRpi. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079–35871578 is marked in moss green.

CNV Term Use Comparison

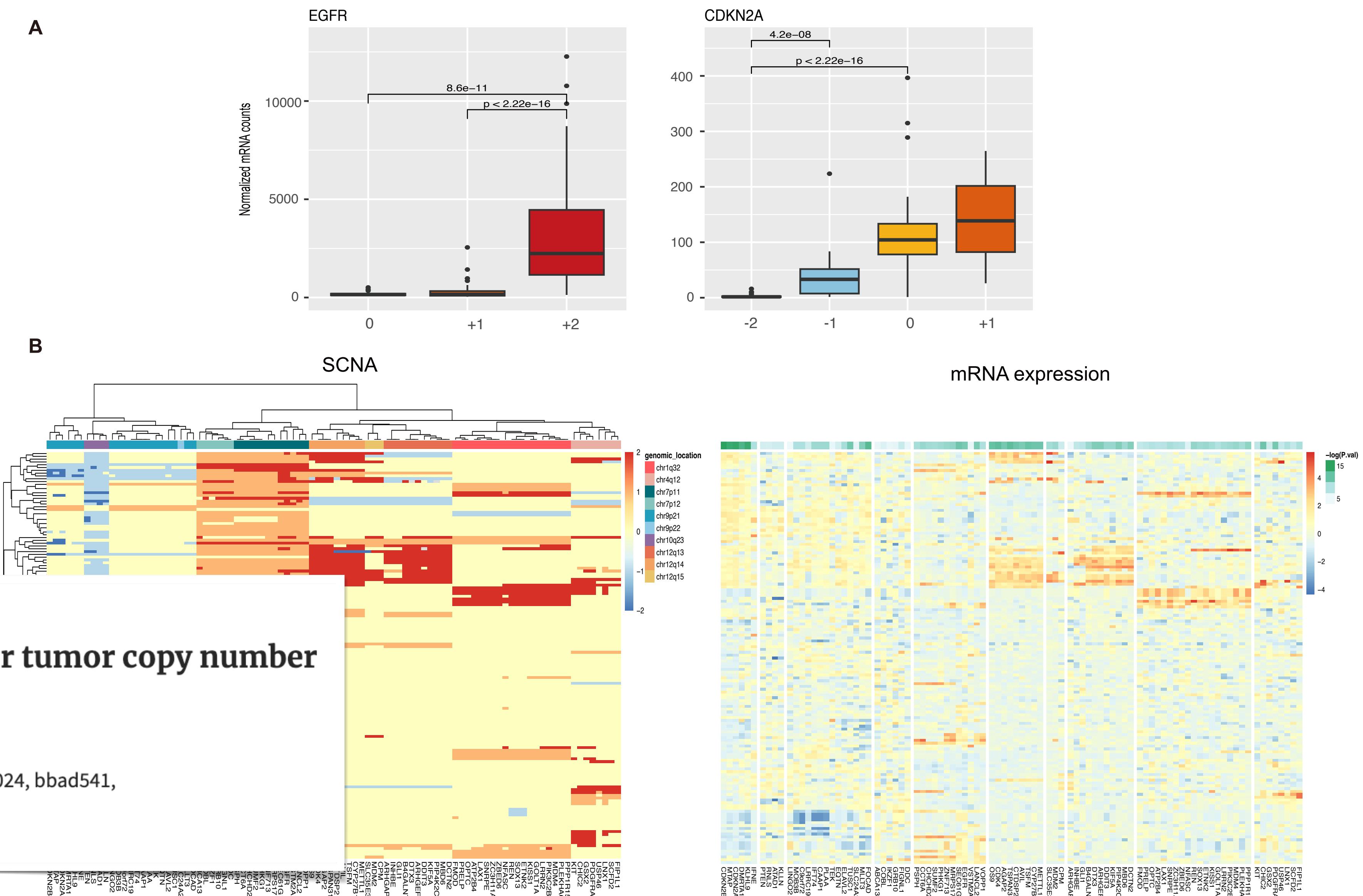
in computational (file/schema) formats

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

labelSeg

Application

Strong relationship between copy-number dosage and messenger RNA expression



JOURNAL ARTICLE

labelSeg: segment annotation for tumor copy number alteration profiles

Hangjia Zhao, Michael Baudis

Briefings in Bioinformatics, Volume 25, Issue 2, March 2024, bbad54
<https://doi.org/10.1093/bib/bbad541>

Published: 31 January 2024 [Article history](#)



University of
Zurich UZH



Swiss Institute of
Bioinformatics

Michael Baudis

Hangjia Zhao

Ziying Yang

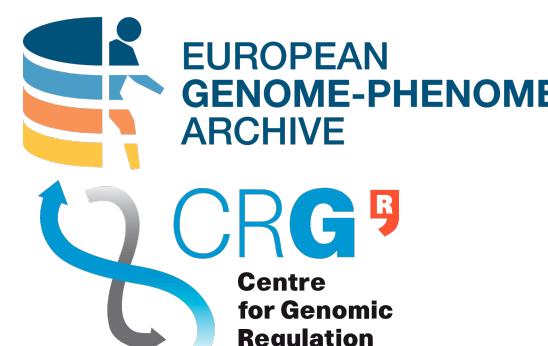
Ramon Benitez

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

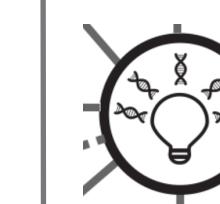


CINECA

Thomas Keane

Melanie Courtot

Jonathan Dursi



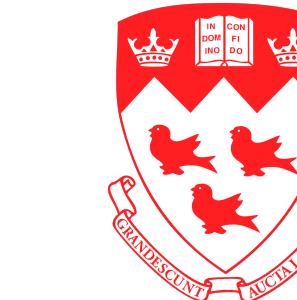
Heidi Rehm

Ben Hutton



Toshiaki

Katayama



Stephane Dyke



Marc Fiume

Miro Cupak



Melissa Cline



Diana Lemos



GA4GH Phenopackets

Peter Robinson

Jules Jacobsen



GA4GH VRS

Alex Wagner
Reece Hart

Beacon PRC

Alex Wagner

Jonathan Dursi

Mamana Mbiyavanga

Alice Mann

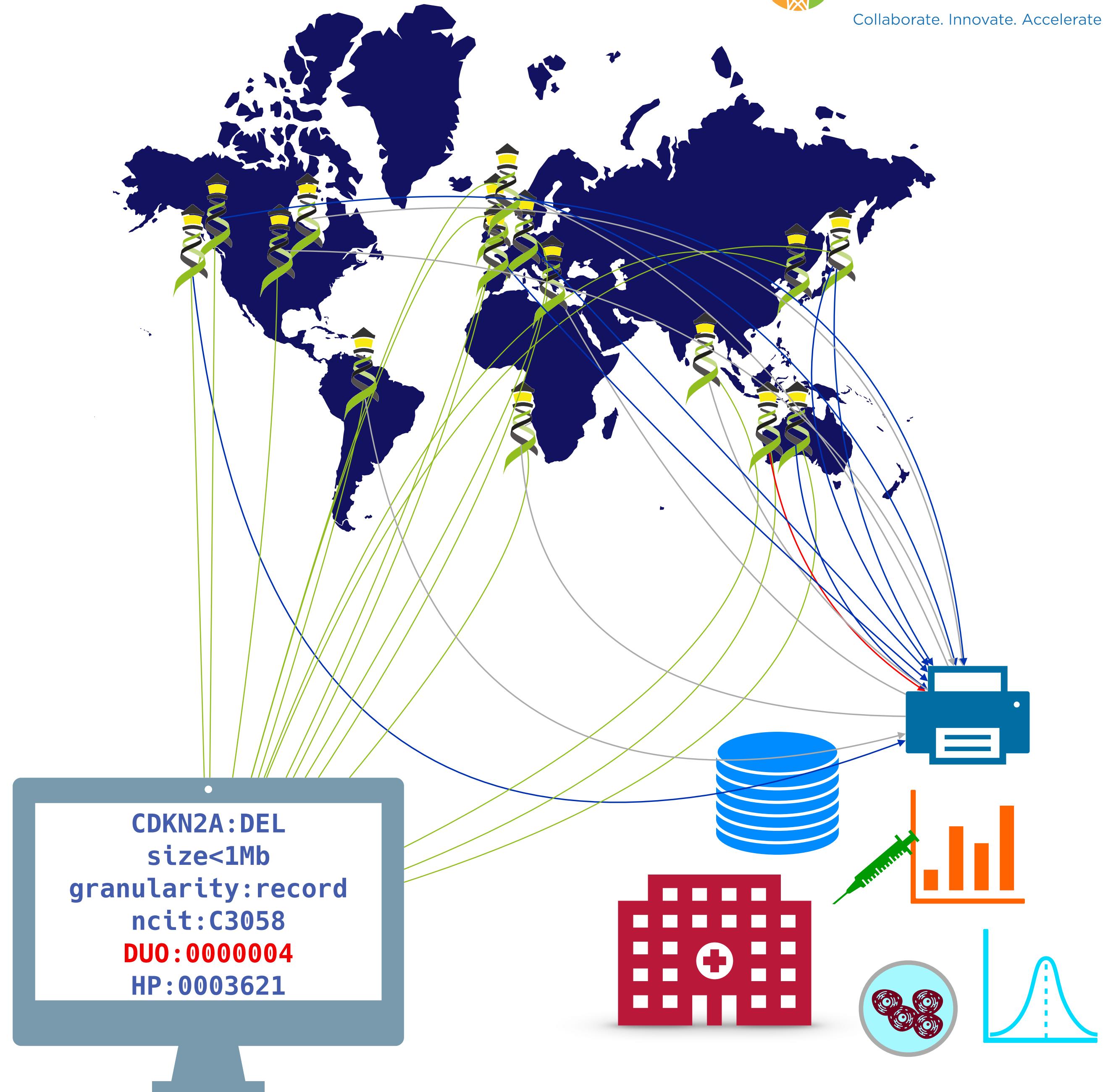
Neerjah Skantharajah



The Beacon team through the ages

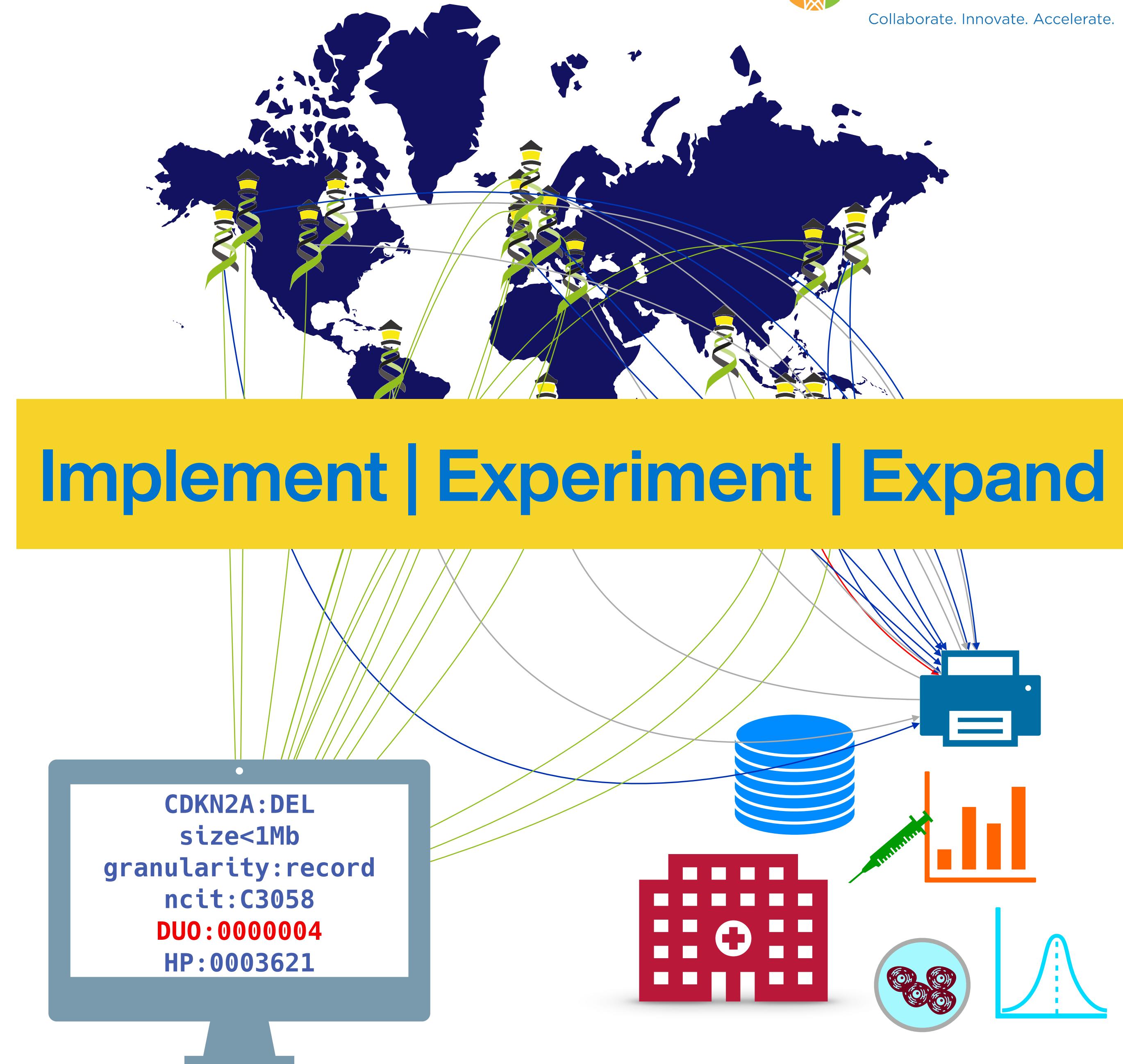
What Can You Do?

- 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



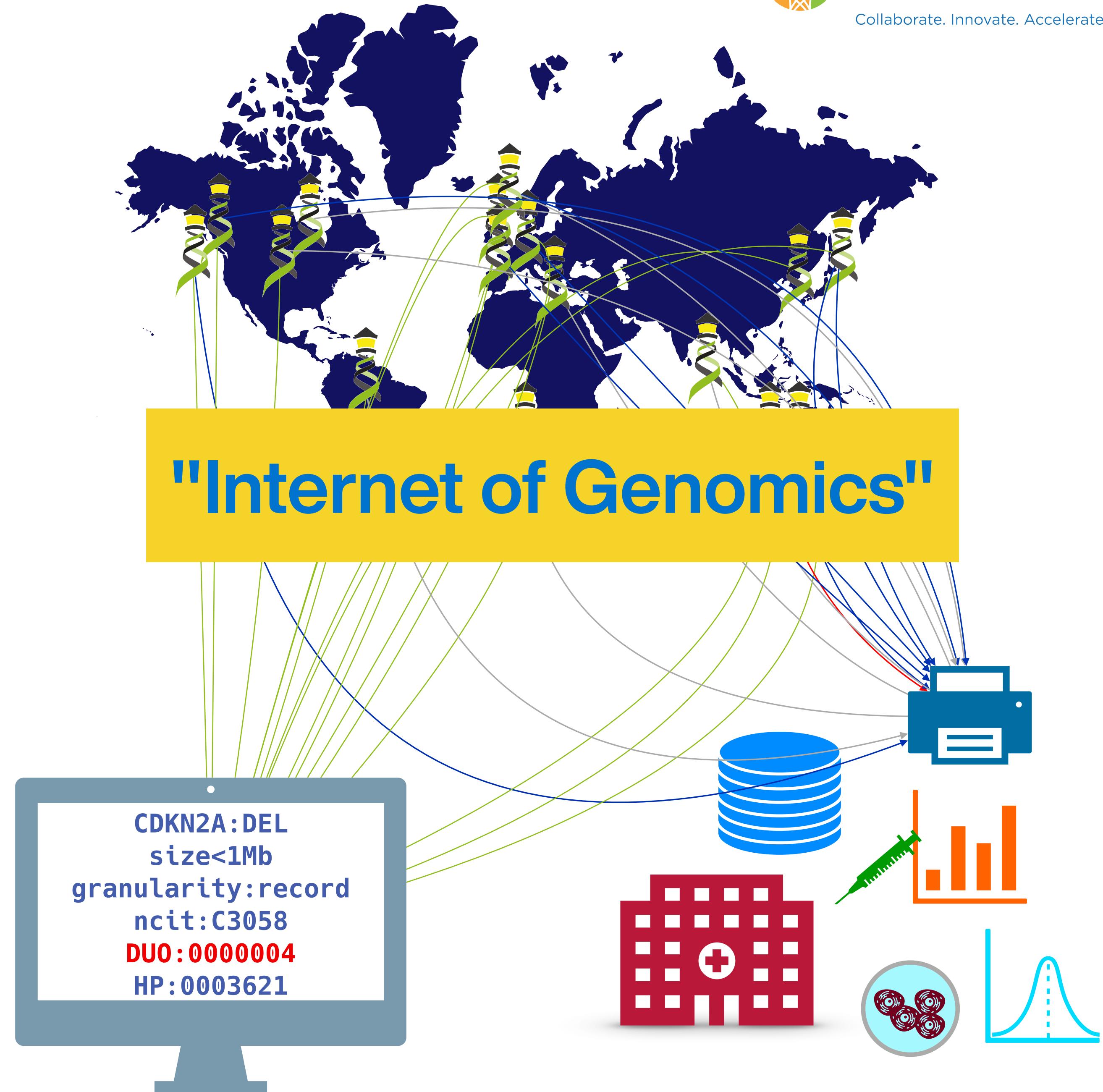
Beacon for Genomic Discovery Proxies

- Feature beacons for privacy protecting data discovery
 - privacy protection through aggregated data, cohorts
 - alternative is "horizontal gatekeeping": separate Beacons for discovery of e.g. genomic and phenotypic data and interleaving by data owner upon request



Beacon for Genomic Discovery Proxies

- Feature beacons for privacy protecting data discovery
 - privacy protection through aggregated data, cohorts
 - alternative is "horizontal gatekeeping": separate Beacons for discovery of e.g. genomic and phenotypic data and interleaving by data owner upon request



Get Involved! Visit GA4GH.ORG



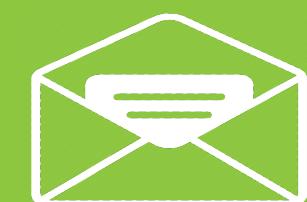
Global Alliance
for Genomics & Health

Join a Work Stream!

Contact secretariat@ga4gh.org



Become an Organisational Member
ga4gh.org/members



Subscribe to GA4GH Updates
ga4gh.org/subscribe

Save the dates!



Global Alliance
for Genomics & Health

April Connect 2025

1 to 4 April 2025

Broad Institute, Cambridge, USA

[Registration Open Now](#)



13th Plenary

6 to 10 October

UKK, Uppsala, Sweden

Registration Opening Soon





Universität
Zürich^{UZH}



Swiss Institute of
Bioinformatics

