



Universität  
Zürich<sup>UZH</sup>



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

# Opening Cancer Genomics

## Deploying the GA4GH Beacon protocol

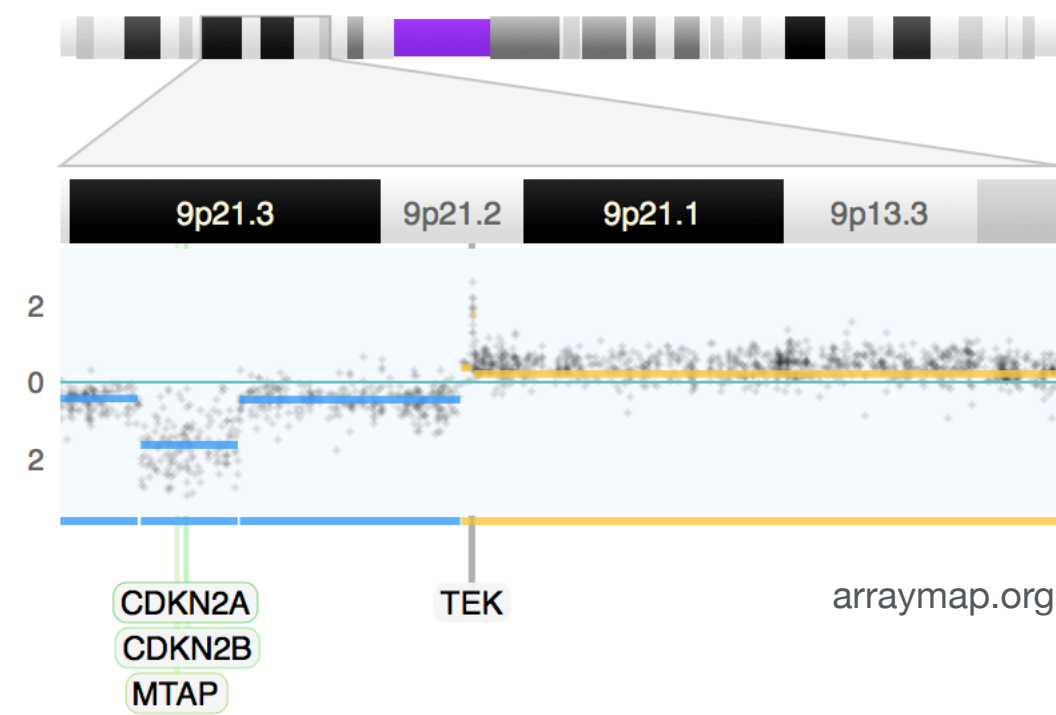
Michael Baudis @ ORD Cancer Care Zürich Project Workshop





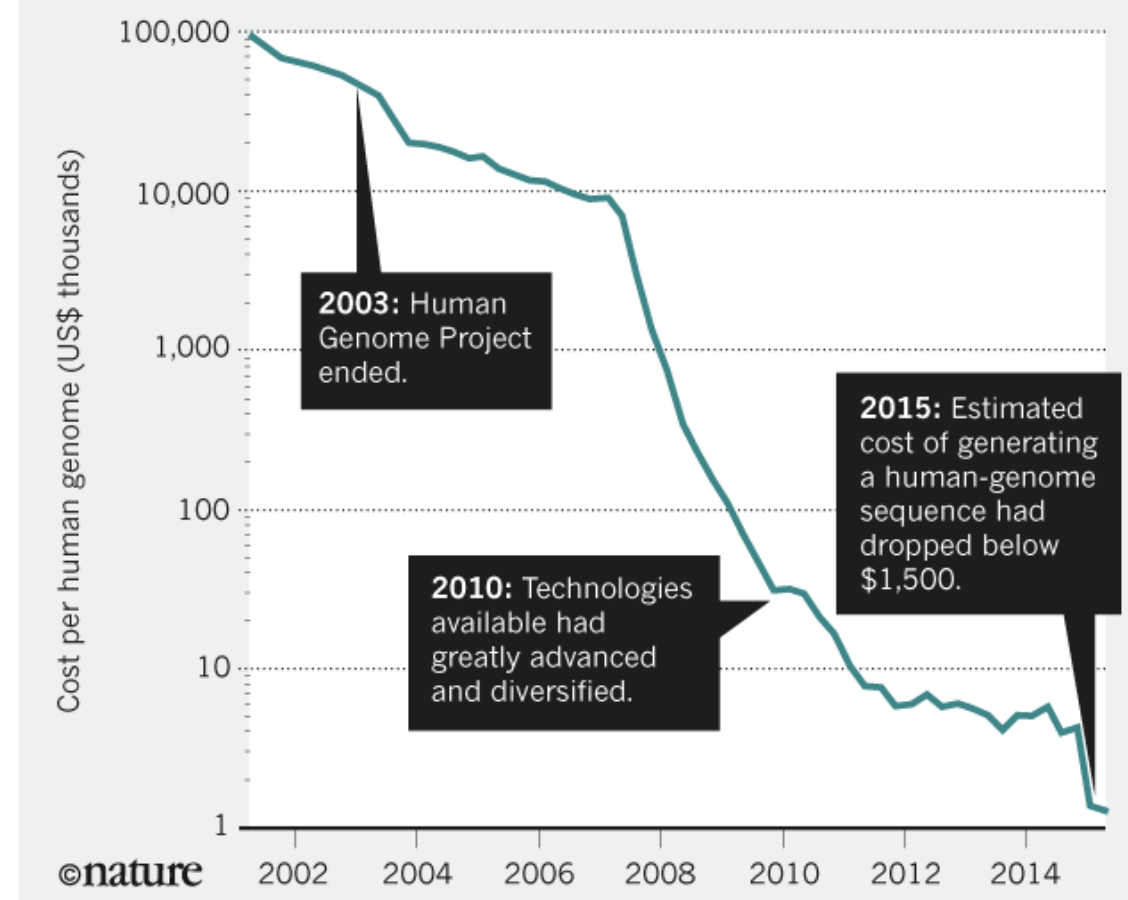
# Genome screening at the core of “Personalised Health”

- ▶ **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- ▶ The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**
- ▶ **Standardized data formats** and **exchange protocols** are needed to connect these resources throughout the world, for reciprocal, international **data sharing** and **biocuration** efforts
- ▶ Our work @ UZH:
  - ▶ **cancer** genome repositories
  - ▶ **biocuration**
  - ▶ **protocols & formats**

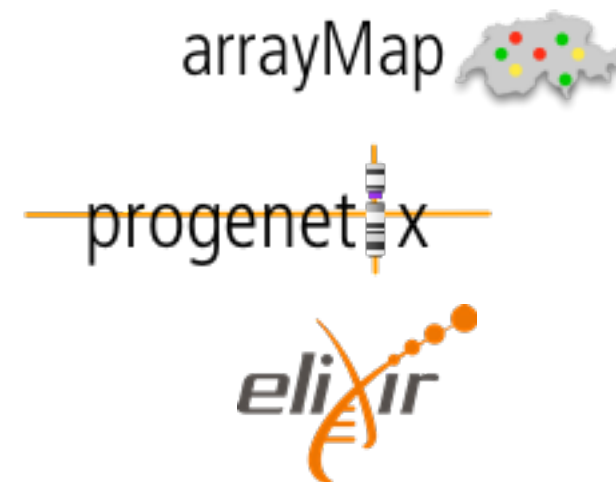
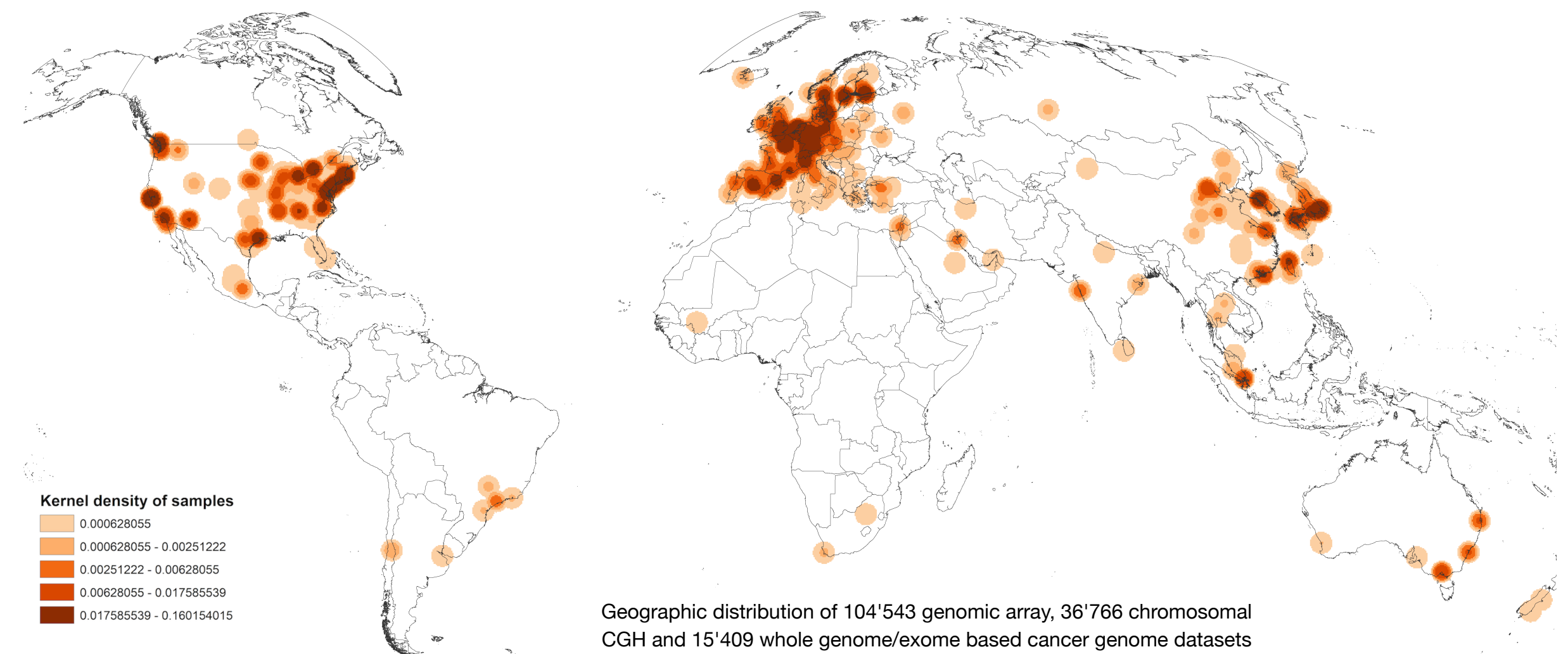


## BETTER, CHEAPER, FASTER

The cost of DNA sequencing has dropped dramatically over the past decade, enabling many more applications.



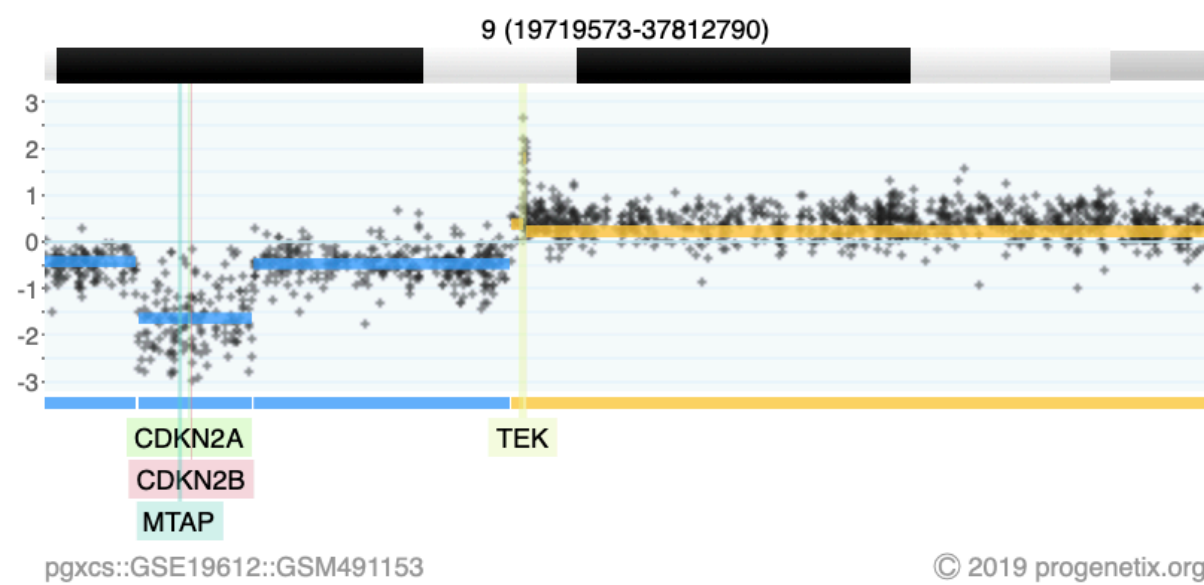
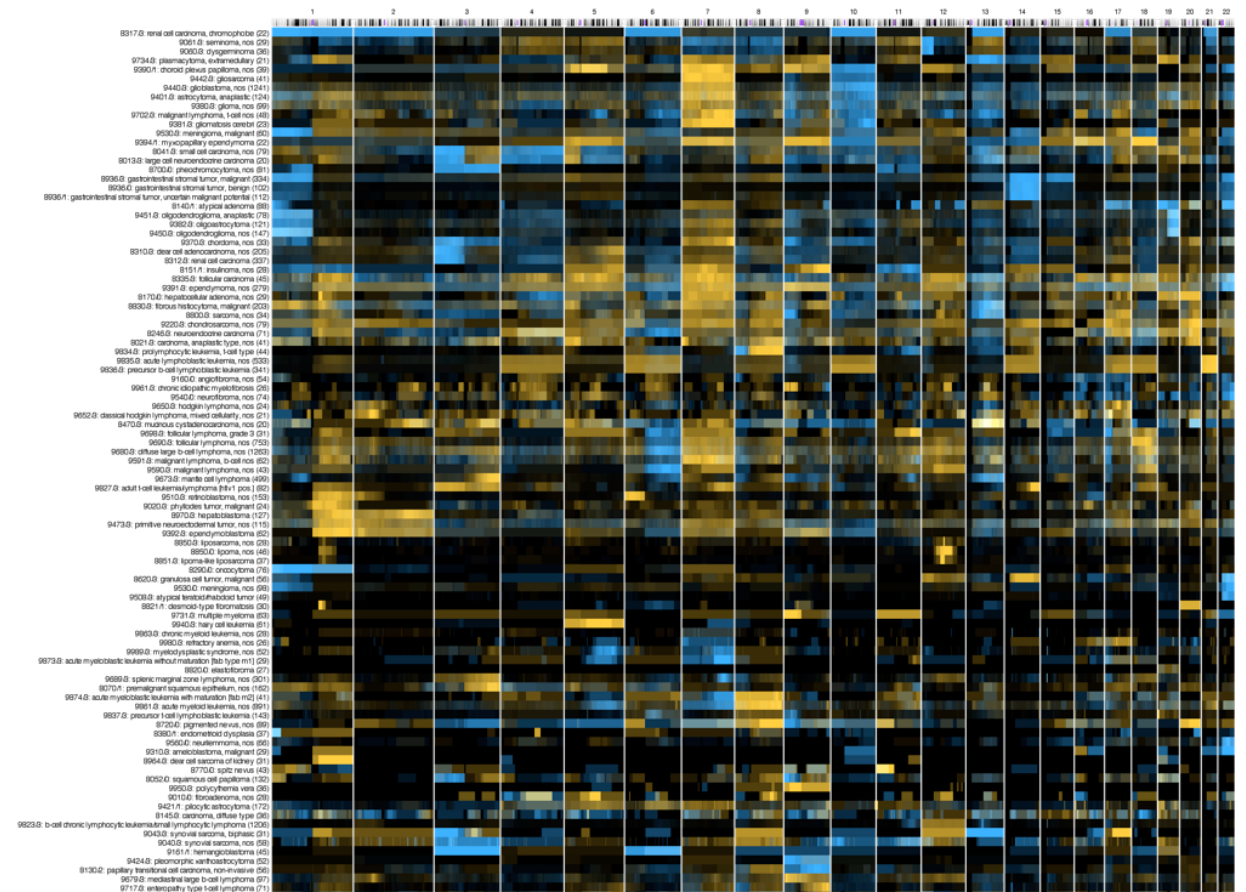
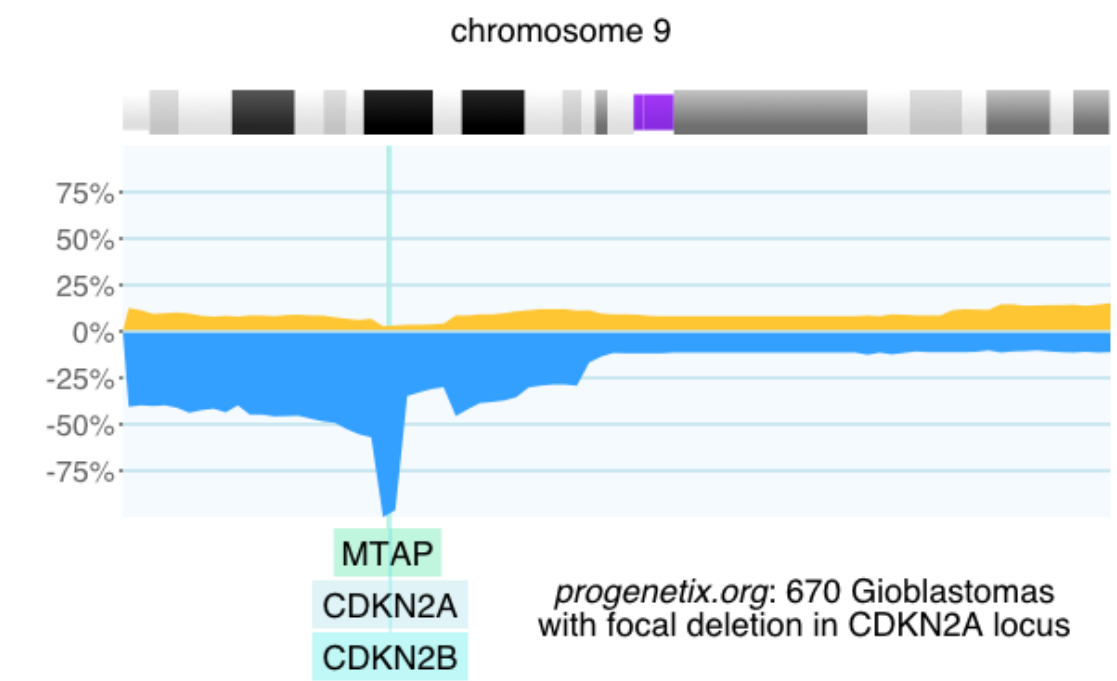
The future of DNA sequencing. Eric D. Green, Edward M. Rubin & Maynard V. Olson. Nature; 11 October 2017 (News & Views)



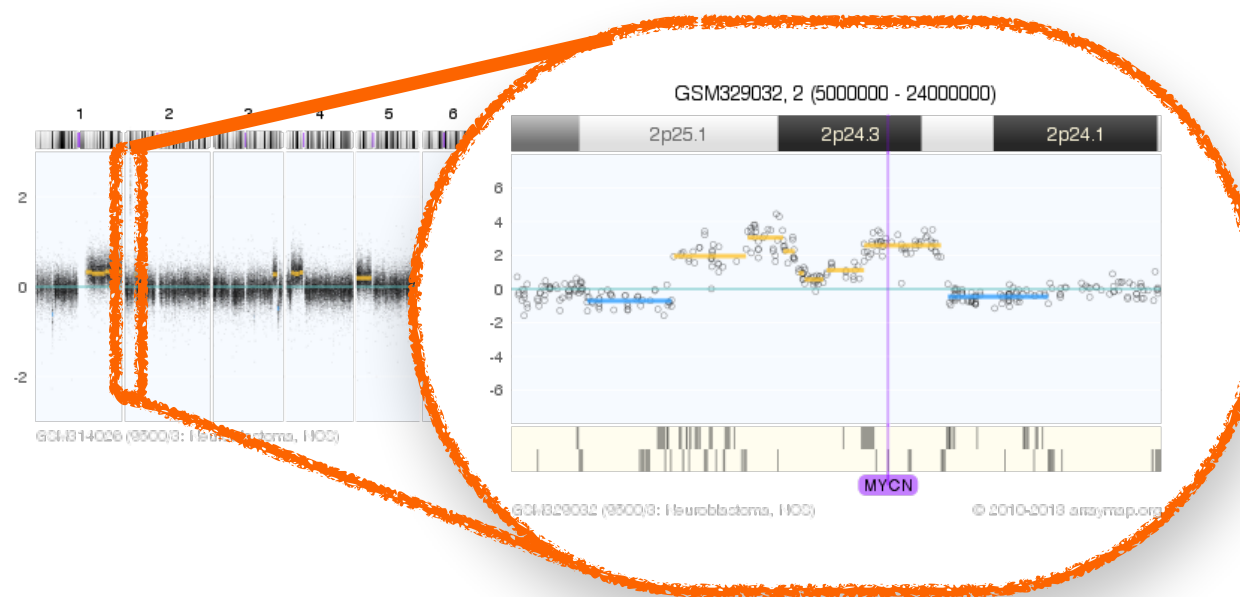
# Theoretical Cytogenetics and Oncogenomics Research | Methods | Standards

## Genomic Imbalances in Cancer - Copy Number Variations (CNV)

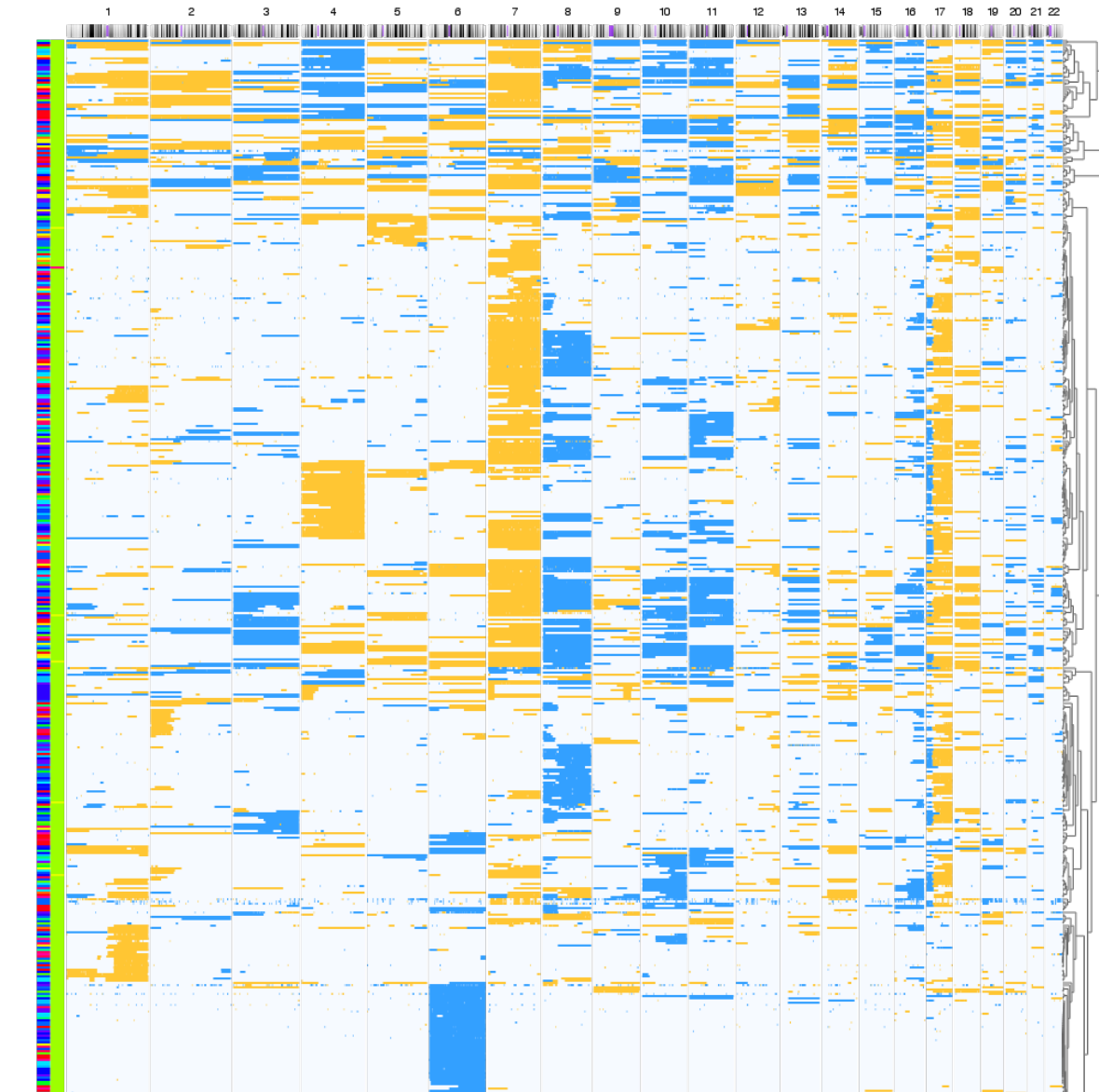
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations (losses, gains)**
- Epigenetic changes (e.g. DNA methylation abnormalities)



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<sup>+</sup>

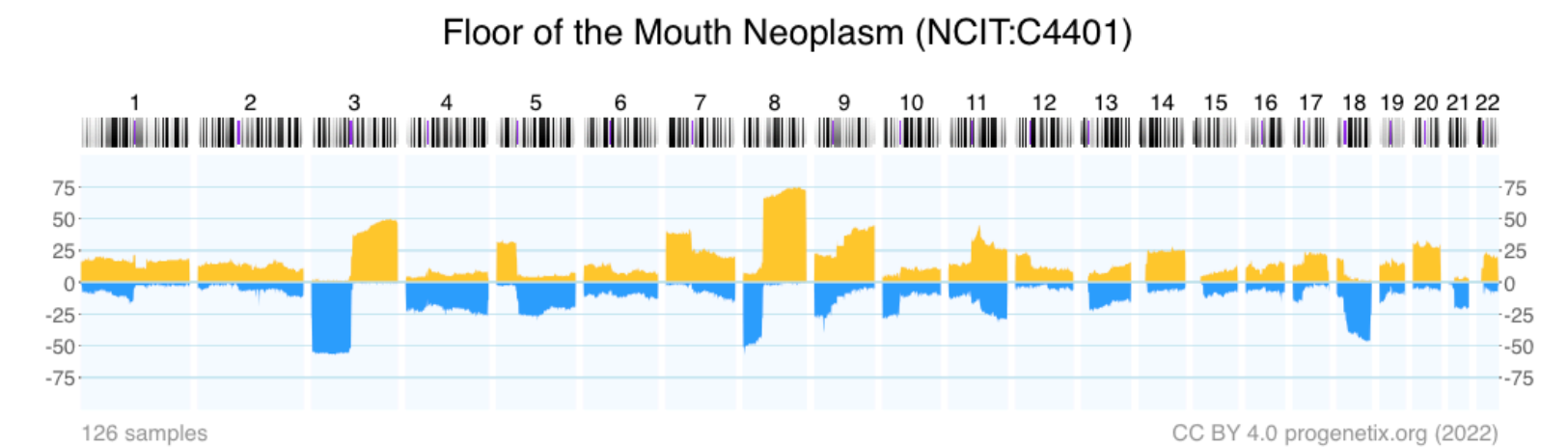
### Documentation

News  
Downloads & Use  
Cases  
Services & 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](#)

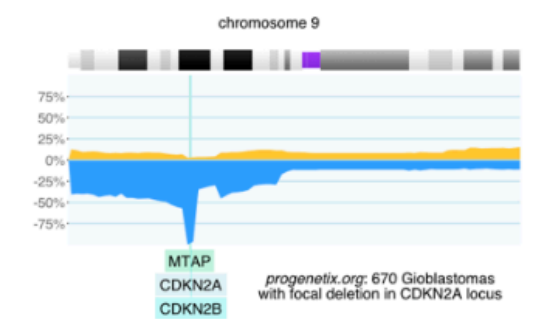
Example for aggregated CNV data in 126 samples in Floor of the Mouth Neoplasm.

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. 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 Types by National Cancer Institute NCIt 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)



© CC-BY 2001 - 2024 progenetix.org

[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 \(51 samples, 51 CNV profiles\)](#)

# progenetix.org

## Cancer Genomics Reference Resource

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



Edit Query

Assembly: GRCh38 Chro: refseq:NC\_000009.12 Start: 21500001-21975098  
End: 21967753-22500000 Type: EFO:0030067 Filters: NCIT:C3058

progenetix

Matched Samples: 657

Retrieved Samples:

Variants: 276

Calls: 659

[UCSC region](#)

[Variants in UCSC](#)

[Dataset Responses \(JSON\)](#)

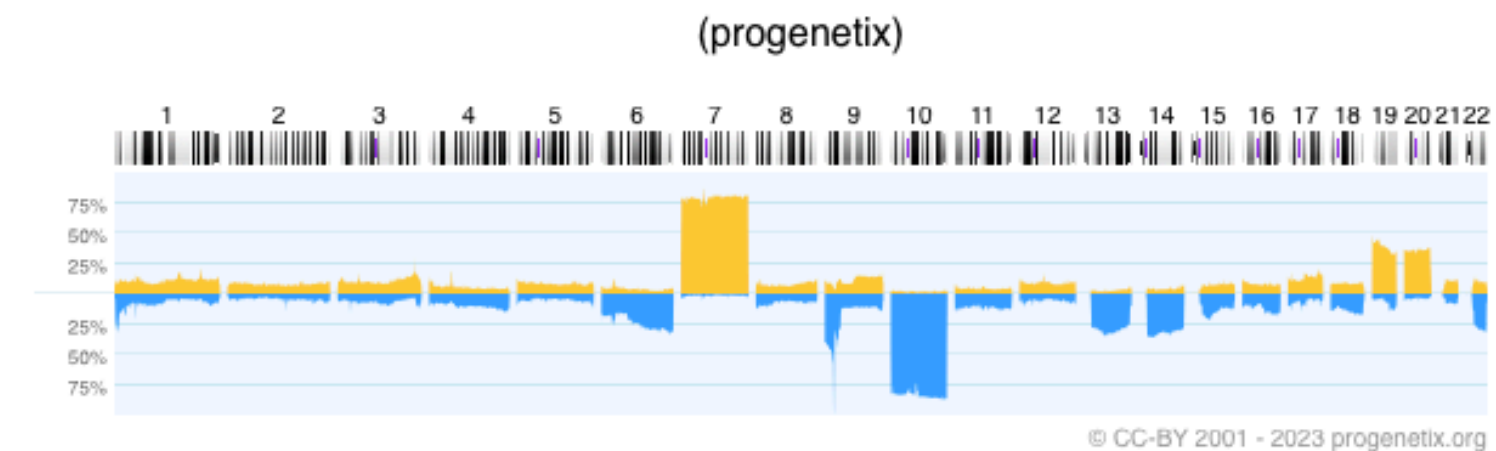
Visualization options

Results

Biosamples

Biosamples Map

Variants



[Reload histogram in new window](#)

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

Download Sample Data (TSV)

1-657

Download Sample Data (JSON)

1-657

# 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

cancerellines.org

Lead: Rahel Paloots

The screenshot shows the cancerellines.org website. At the top is the logo, a pink circle with three smaller circles inside. Below it is the text 'cancerellines'. A navigation menu on the left includes: Cancer Cell Lines, Search Cell Lines, Cell Line Listing, CNV Profiles by Cancer Type, Documentation, Progenetix, and Publication DB. The main content area is titled 'Cancer Cell Lines by Cellosaurus ID' and contains a search filter 'Filter subsets e.g. by prefix' and a 'Hierarchy Depth' dropdown. Below the filter is a list of cell lines with expandable arrows and their sample counts, such as 'cellosaurus:CVCL\_0312: HOS (204 samples)'.

### Cancer Cell Lines by Cellosaurus ID

The cancer cell lines in *cancerellines.org* are labeled by the hierarchially: Daughter cell lines are displayed below the parent as a daughter cell line of **HeLa (CVCL\_0030)** and so forth.

Sample selection follows a hierarchical system in which same response. This means that one can retrieve all instances and for HeLa will also return the daughter lines by default - but c

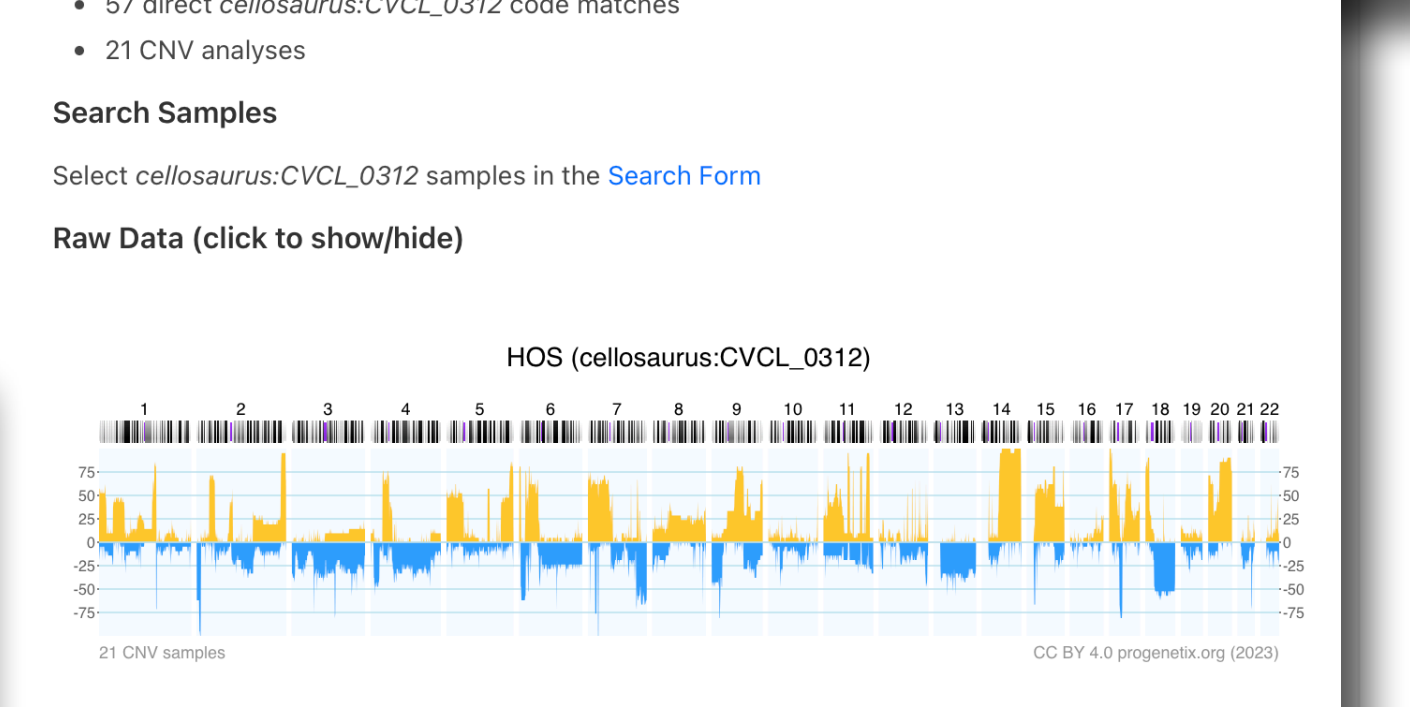
### Cell Lines (with parental/derived hierarchies)

This screenshot shows a filtered list of cell lines. Each entry includes a checkbox, an expandable arrow, the cell line ID, name, and sample count. Examples include:
 

- 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
- cellosaurus:CVCL\_0589: Kasumi-1 (9 samples)

This screenshot shows the UCSC Genome Browser interface for a specific cell line. At the top, it displays 'Assembly: GRCh38 Chro: NC\_000007.14 Start: 140713328 End: 140924929 Type: SNV'. Below this, it shows 'Matched Samples: 1058', 'Retrieved Samples: 1000', 'Variants: 127', and 'Calls: 1444'. There are links for 'UCSC region', 'Variants in UCSC', and 'Dataset Responses (JSON)'. A 'Visualization options' button is also present. The main table shows variant details with columns for 'Digest', 'Gene', 'Pathogenicity', 'Variant type', and 'Variant instances'. The table lists three variants, all of which are missense variants in the BRAF gene.

This screenshot shows the 'Cell Line Details' page for HOS (cellosaurus:CVCL\_0312). It includes a 'Subset Type' section with a link to the cell line resource. The 'Sample Counts' section lists: 204 samples, 57 direct cellosaurus:CVCL\_0312 code matches, and 21 CNV analyses. The 'Search Samples' section has a link to the search form. The 'Raw Data (click to show/hide)' section is partially visible.

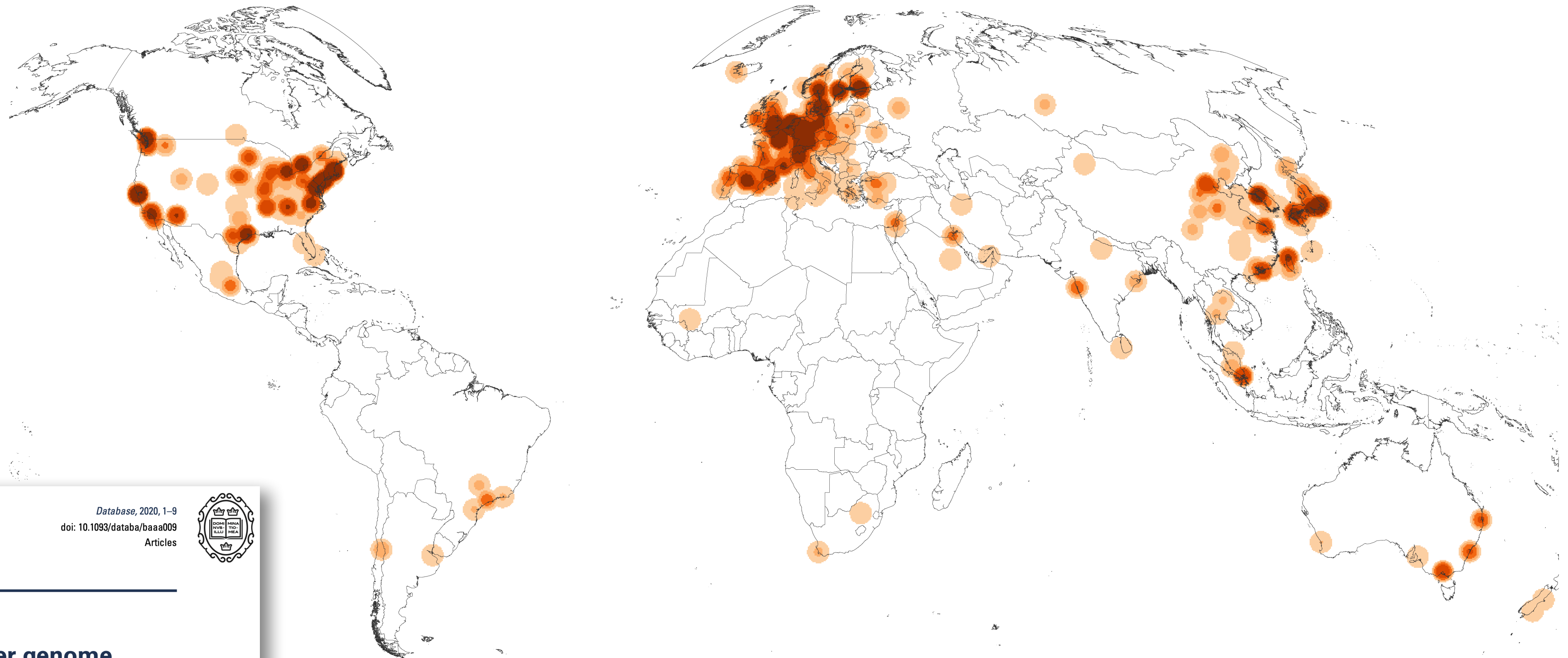


This screenshot shows the 'Gene Matches' section. It lists two genes: ALK and AREG. For each gene, there is a brief description of its association with the cell line and a link to an abstract. For example, for ALK, it mentions 'ABC-14 cells harbored no ALK mutations and were sensitive to ... crizotinib while also exhibiting MNNG HOS transforming gene ( MET )' and links to an abstract about 'Rapid Acquisition of Alectinib Resistance in ALK-Positive Lung Cancer With High Tumor Mutation Burden (31374369)'.

This is a bioRxiv preprint banner. It features the logos for CSH Cold Spring Harbor Laboratory, bioRxiv (The Preprint Server for Biology), and the DATABASE (The Journal of Biological Databases and Curation). The text reads: 'New Results Follow this preprint cancerellines.org - a Novel Resource for Genomic Variants in Cancer Cell Lines'. It lists the authors: Rahel Paloots and Michael Baudis, and provides the DOI: https://doi.org/10.1101/2023.12.12.571281. A disclaimer at the bottom states: 'This article is a preprint and has not been certified by peer review [what does this mean?]'.


# Where Does Cancer Genomic Data Come From?

## Geographic bias in published cancer genome profiling studies



**DATABASE**  
The Journal of Biological Databases and Curation

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




---

Articles

**Geographic assessment of cancer genome profiling studies**

Paula Carrio-Cordo<sup>1,2</sup>, Elise Acheson<sup>3</sup>, Qingyao Huang<sup>1,2</sup> and Michael Baudis<sup>1,\*</sup>

<sup>1</sup>Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland <sup>2</sup>Swiss Institute of Bioinformatics, Zurich, Switzerland <sup>3</sup>Department of Geography, University of Zurich, Zurich, Switzerland

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





# Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

## GENOMICS

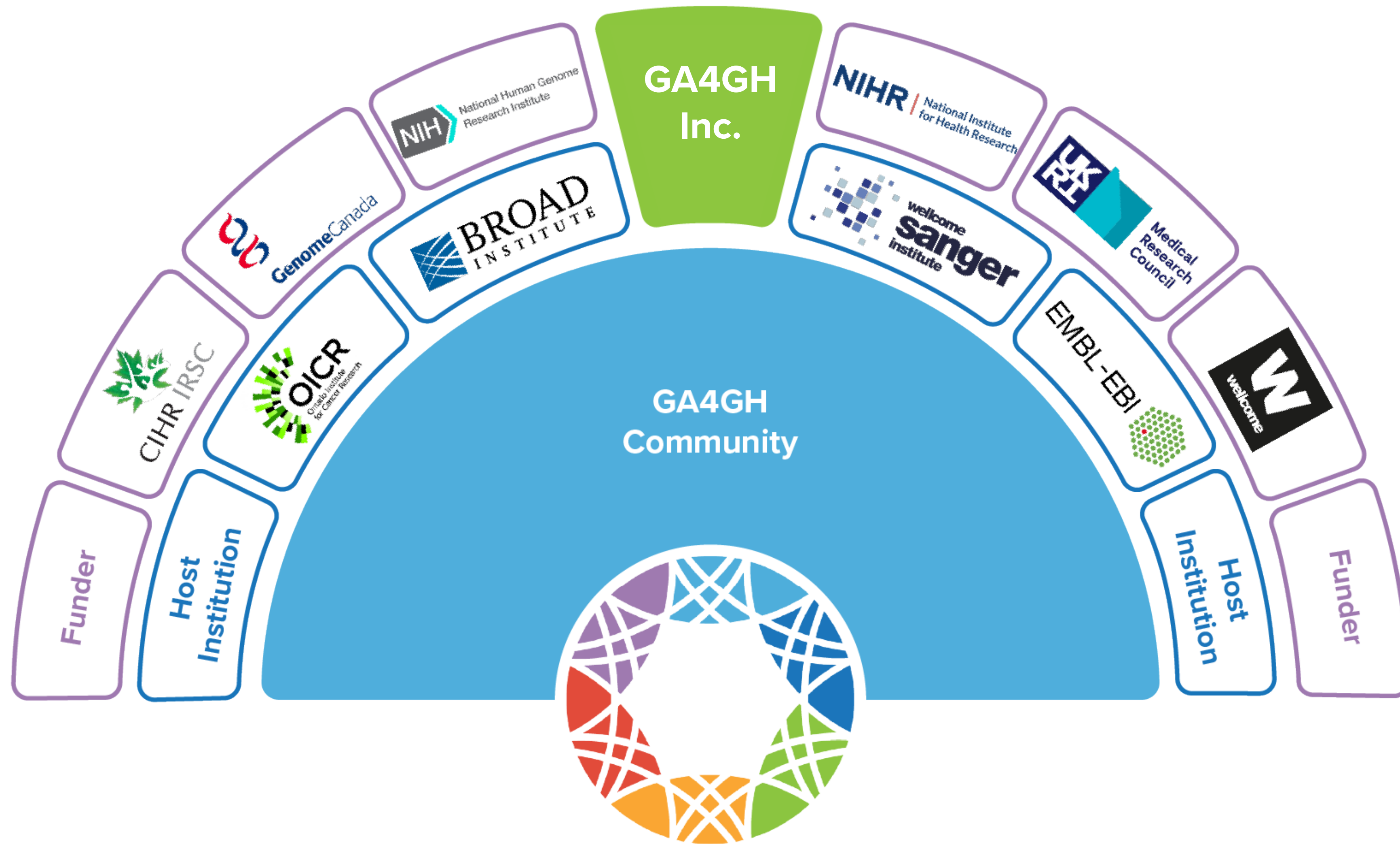
### *A federated ecosystem for sharing genomic, clinical data*

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

# Organization

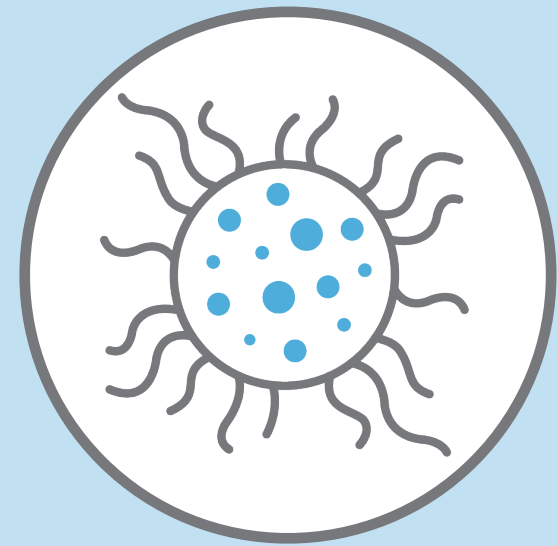


Global Alliance  
for Genomics & Health

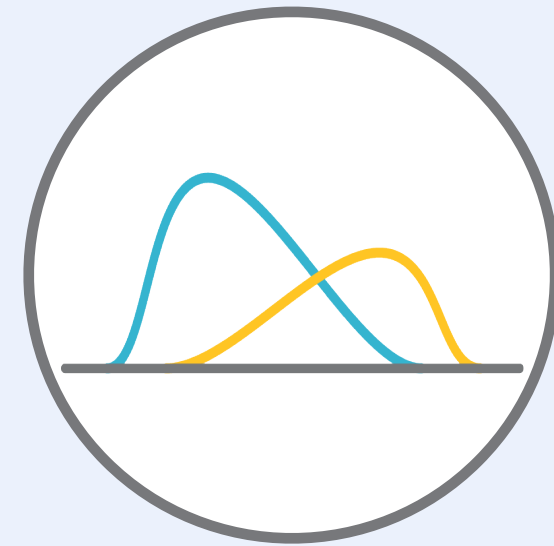




# Global Genomic Data Sharing Can...



Demonstrate  
patterns in health  
& disease



Increase statistical  
significance of  
analyses



Lead to  
“stronger” variant  
interpretations



Increase  
accurate  
diagnosis



Advance  
precision  
medicine

# Different Approaches to Data Sharing



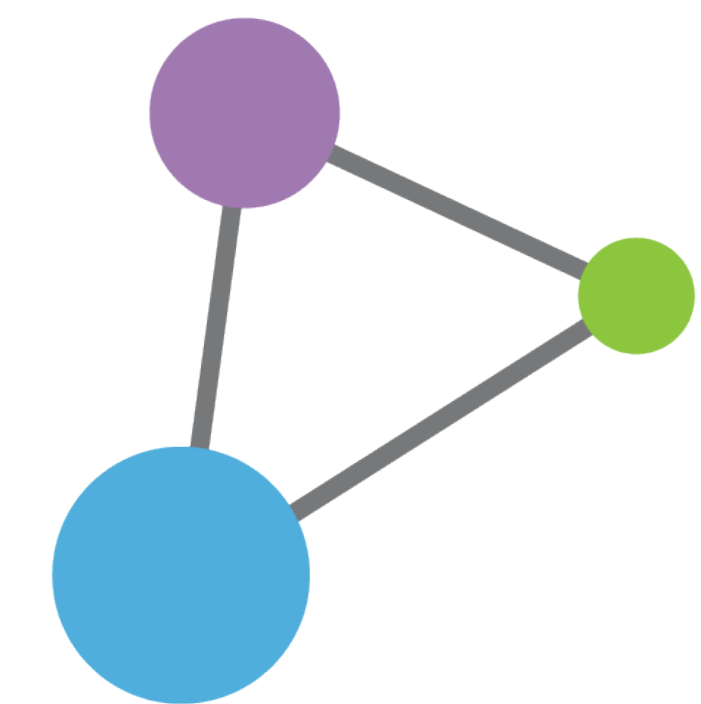
**Centralized Genomic Knowledge Bases**



**Data Commons**  
Trusted, controlled repository of multiple datasets



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



**Linkage of distributed and disparate datasets**

# Different Approaches to Data Sharing

progenetix



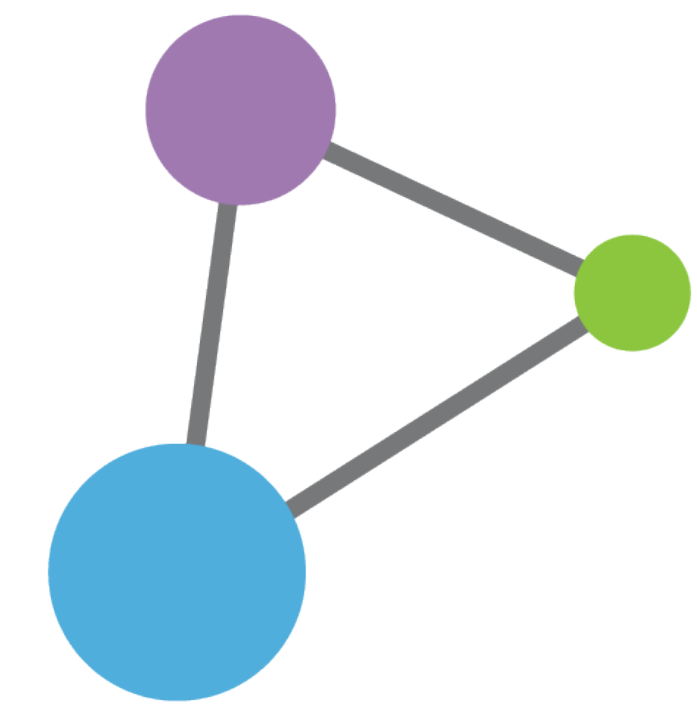
**Centralized Genomic Knowledge Bases**



**Data Commons**  
Trusted, controlled repository of multiple datasets



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



**Linkage of distributed and disparate datasets**

# Different Approaches to Data Sharing



**Centralized Genomic Knowledge Bases**



**Data Commons**  
Trusted, controlled repository of multiple datasets



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



**Linkage of distributed and disparate datasets**

# 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

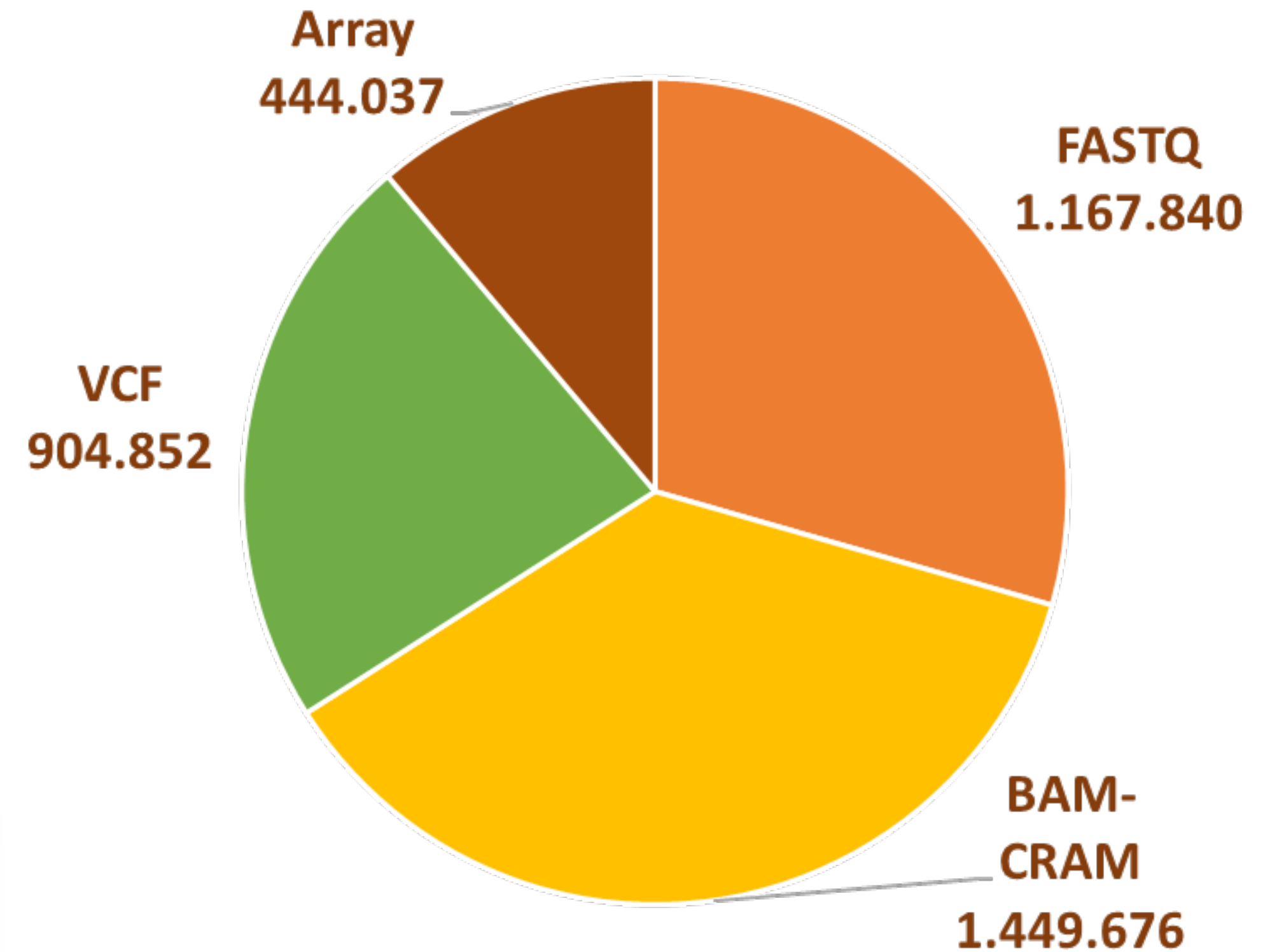


- EGA “owns” nothing; data controllers tell who is authorized to access **their** datasets
- EGA admins provide smooth “all or nothing” data sharing process

The screenshot shows the EGA DAC management interface. The top part displays 'My DACs - EGAC50000000005 - Requests' with 'EuCanImage DAC' and a description 'This is a DAC for EuCanImage data'. Below this is a search bar and a list of requests, including one from 'Dr Teresa Garcia Lezana'. The bottom part shows the 'History' tab with a table of requests:

Date	Requester	Dataset	DAC Admin/Member	Actions
18 August 2022	gemma.milla@crg.eu	EGAD500000000032	Dr Lauren A Fromont	
17 August 2022	Dr Teresa Garcia Lezana	EGAD500000000033	Dr Teresa Garcia Lezana	revoke permission <input type="checkbox"/>
16 August 2022	Dr Teresa Garcia Lezana	EGAD500000000032	Dr Lauren A Fromont	revoke permission <input type="checkbox"/>

## # Files



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



# Different Approaches to Data Sharing



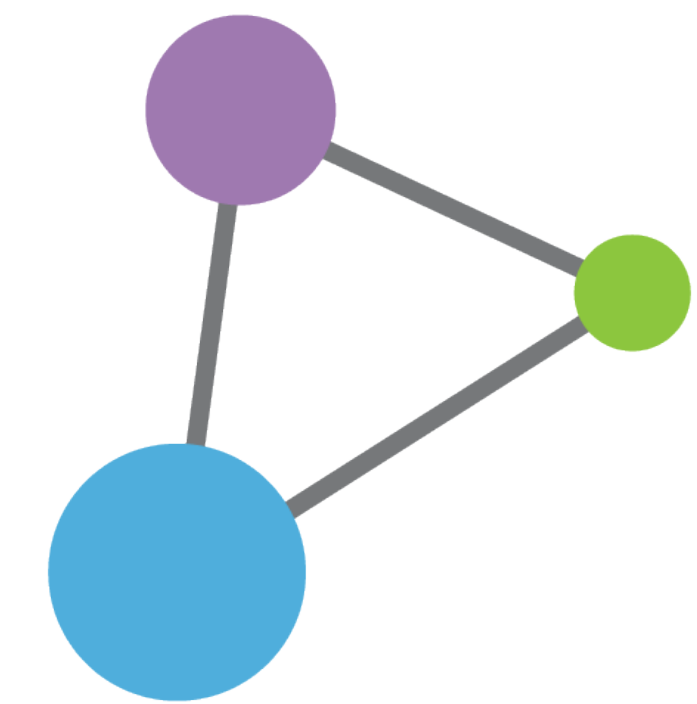
**Centralized Genomic Knowledge Bases**



**Data Commons**  
Trusted, controlled repository of multiple datasets

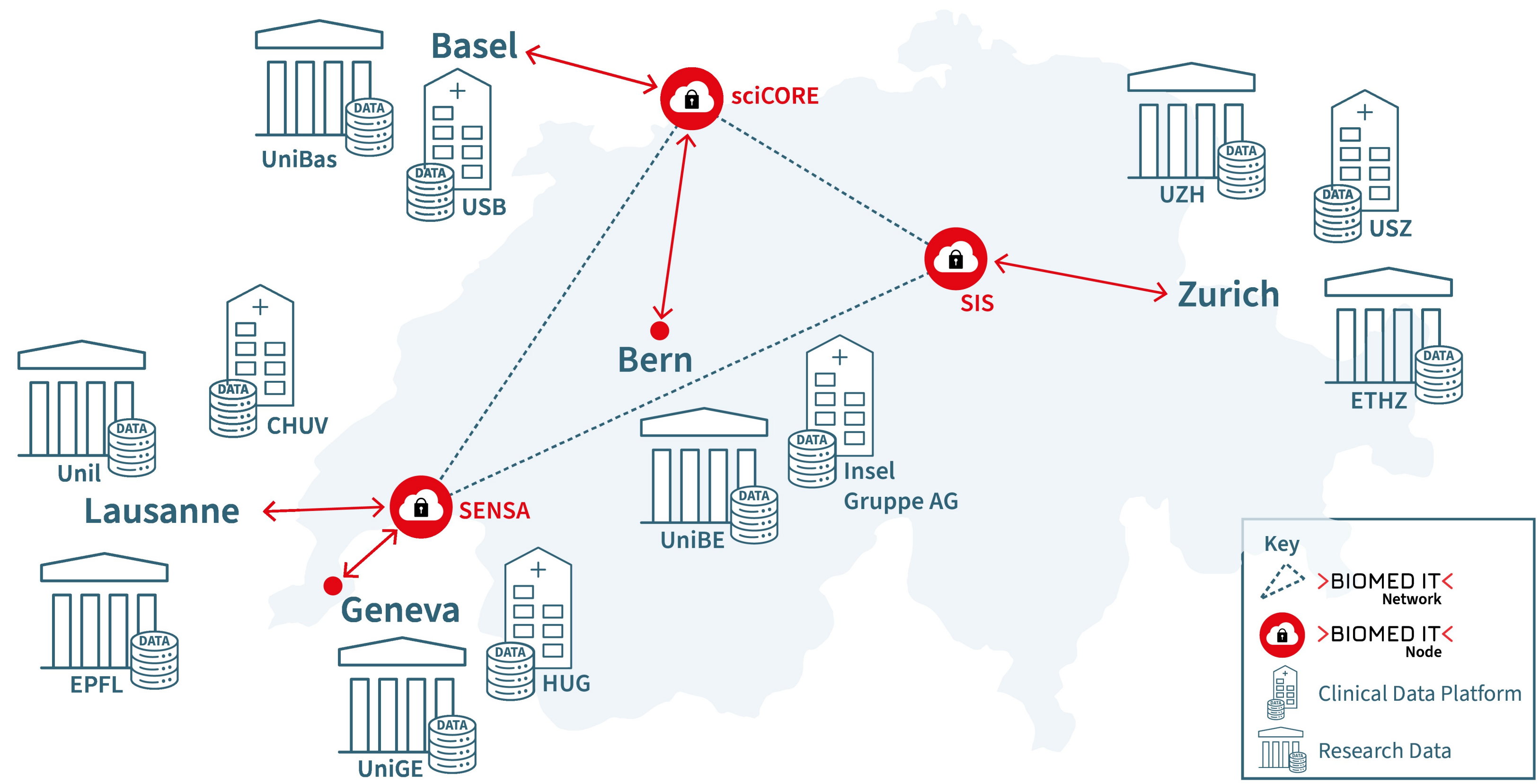


**Hub and Spoke**  
Common data elements, access, and usage rules



**Linkage of distributed and disparate datasets**

# The Swiss Personalized Health Network



**Key**

- >BIOMED IT< Network
- >BIOMED IT< Node
- Clinical Data Platform
- Research Data

**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

**FN-SNF**  
FONDS NATIONAL SUISSE  
SCHWEIZERISCHER NATIONALFONDS  
FONDO NAZIONALE SVIZZERO  
SWISS NATIONAL SCIENCE FOUNDATION

**THE LOOP ZÜRICH**  
MEDICAL RESEARCH CENTER

**Personalized Health Alliance**  
Basel-Zurich

**SWISS BIOBANKING PLATFORM**

**SAKK**  
WE BRING PROGRESS TO CANCER CARE

**SCTO**

**SSPH+**  
SWISS SCHOOL OF PUBLIC HEALTH

life sciences cluster basel

**swissuniversities**

**Universitäre Medizin Schweiz**  
Médecine Universitaire Suisse



# Different Approaches to Data Sharing



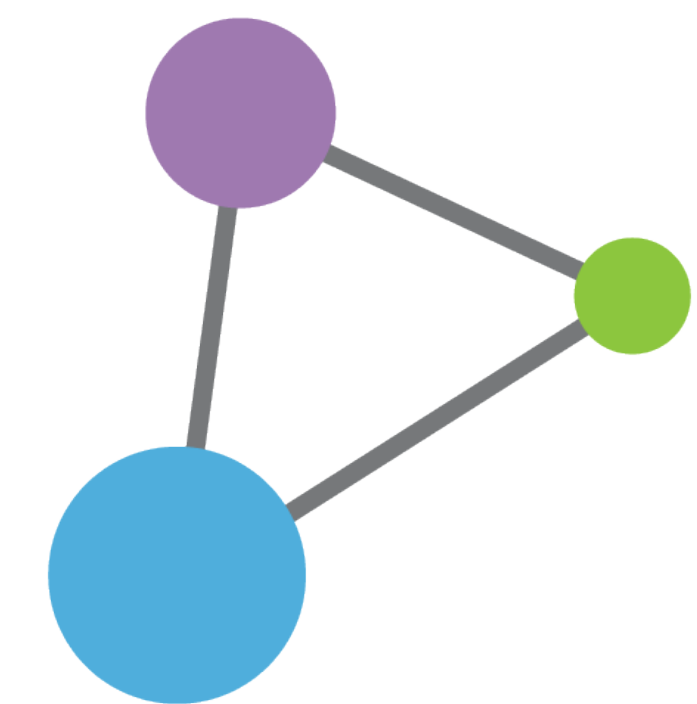
**Centralized Genomic Knowledge Bases**



**Data Commons**  
Trusted, controlled repository of multiple datasets



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



**Linkage of distributed and disparate datasets**

**Federation**



Commentary

**International federation of genomic medicine databases using GA4GH standards**

Adrian Thorogood,<sup>1,2,\*</sup> Heidi L. Rehm,<sup>3,4</sup> Peter Goodhand,<sup>5,6</sup> Angela J.H. Page,<sup>4,5</sup> Yann Joly,<sup>2</sup> Michael Baudis,<sup>7</sup> Jordi Rambla,<sup>8,9</sup> Arcadi Navarro,<sup>8,10,11,12</sup> Tommi H. Nyronen,<sup>13,14</sup> Mikael Linden,<sup>13,14</sup> Edward S. Dove,<sup>15</sup> Marc Fiume,<sup>16</sup> Michael Brudno,<sup>17</sup> Melissa S. Cline,<sup>18</sup> and Ewan Birney<sup>19</sup>

INFORMATICS

**Beacon v2 and Beacon networks: federated data discovery in biomedicine**

Jordi Rambla<sup>1,2</sup> | Michael Baudis<sup>3</sup> | Roberto Ariosio<sup>1</sup> | Tim Beck<sup>4</sup> |  
Lauren A. Fromont<sup>1</sup> | Arcadi Navarro<sup>1,5,6,7</sup> | Rahel Paloots<sup>3</sup> |  
Manuel Rueda<sup>1</sup> | Gary Saunders<sup>8</sup> | Babita Singh<sup>1</sup> | John D. Spalding<sup>9</sup> |  
Juha Törnroos<sup>9</sup> | Claudia Vasallo<sup>1</sup> | Colin D. Veal<sup>4</sup> | Anthony J. Brookes<sup>10</sup>

Perspective

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

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

(Author list continued on next page)

Technology

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

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

# A New Paradigm for Data Sharing

FROM



Data Copying



TO



Data Visiting

# A New Paradigm for Data Sharing

FROM



Data Copying

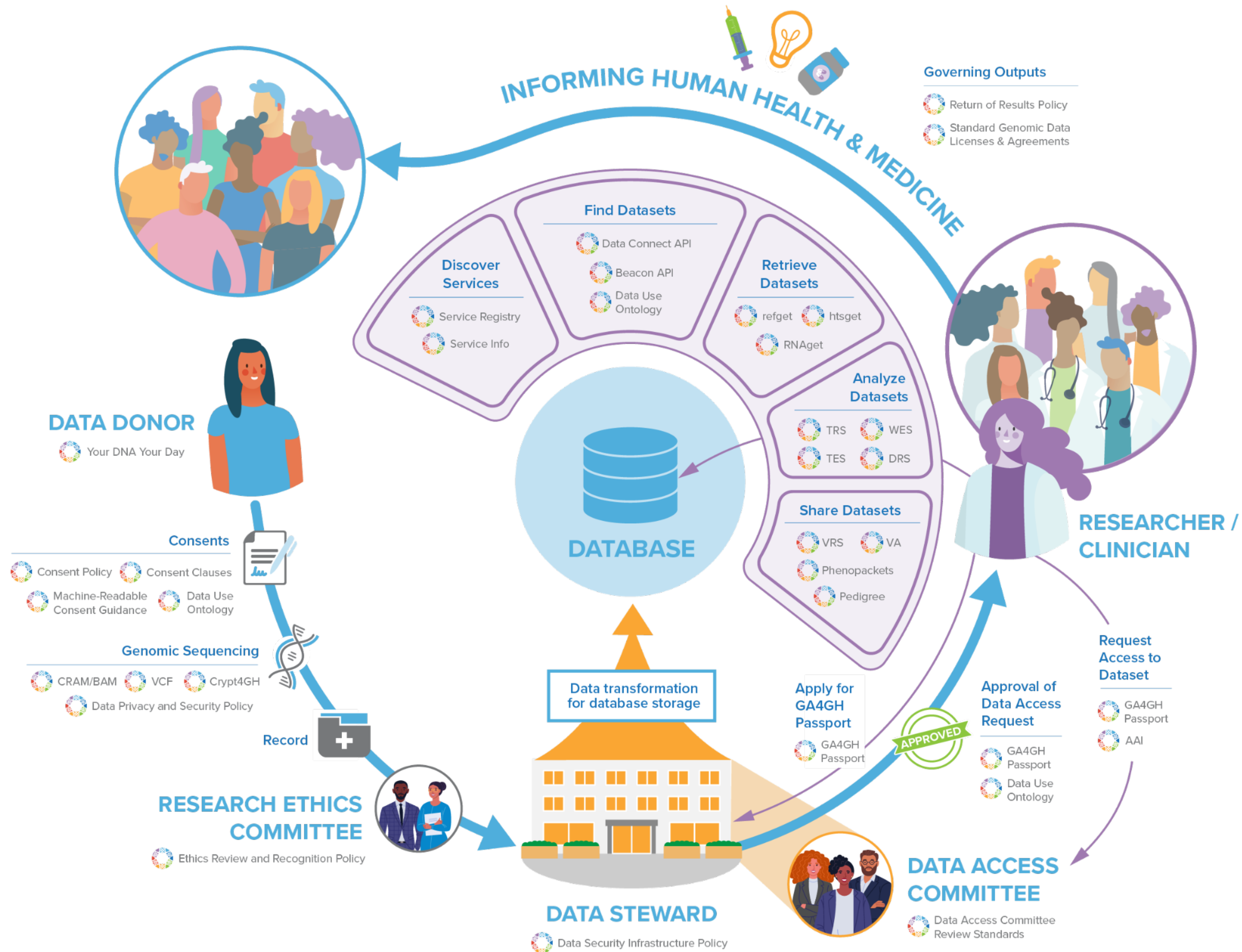
STANDARDS



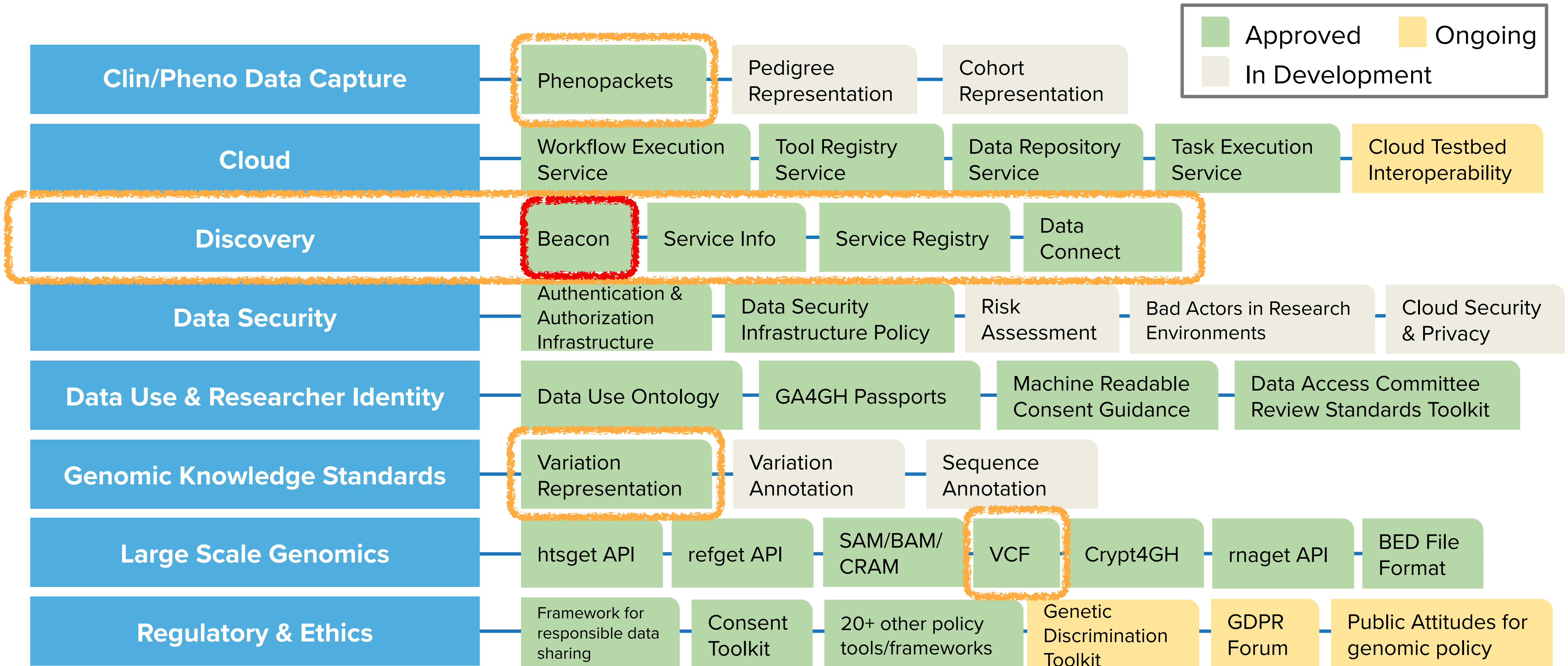
TO



Data Visiting



# Overview of GA4GH standards and frameworks





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

Genomics England has implemented the standard GA4GH API for the Genomes Program and the Genomic Medicine Service.



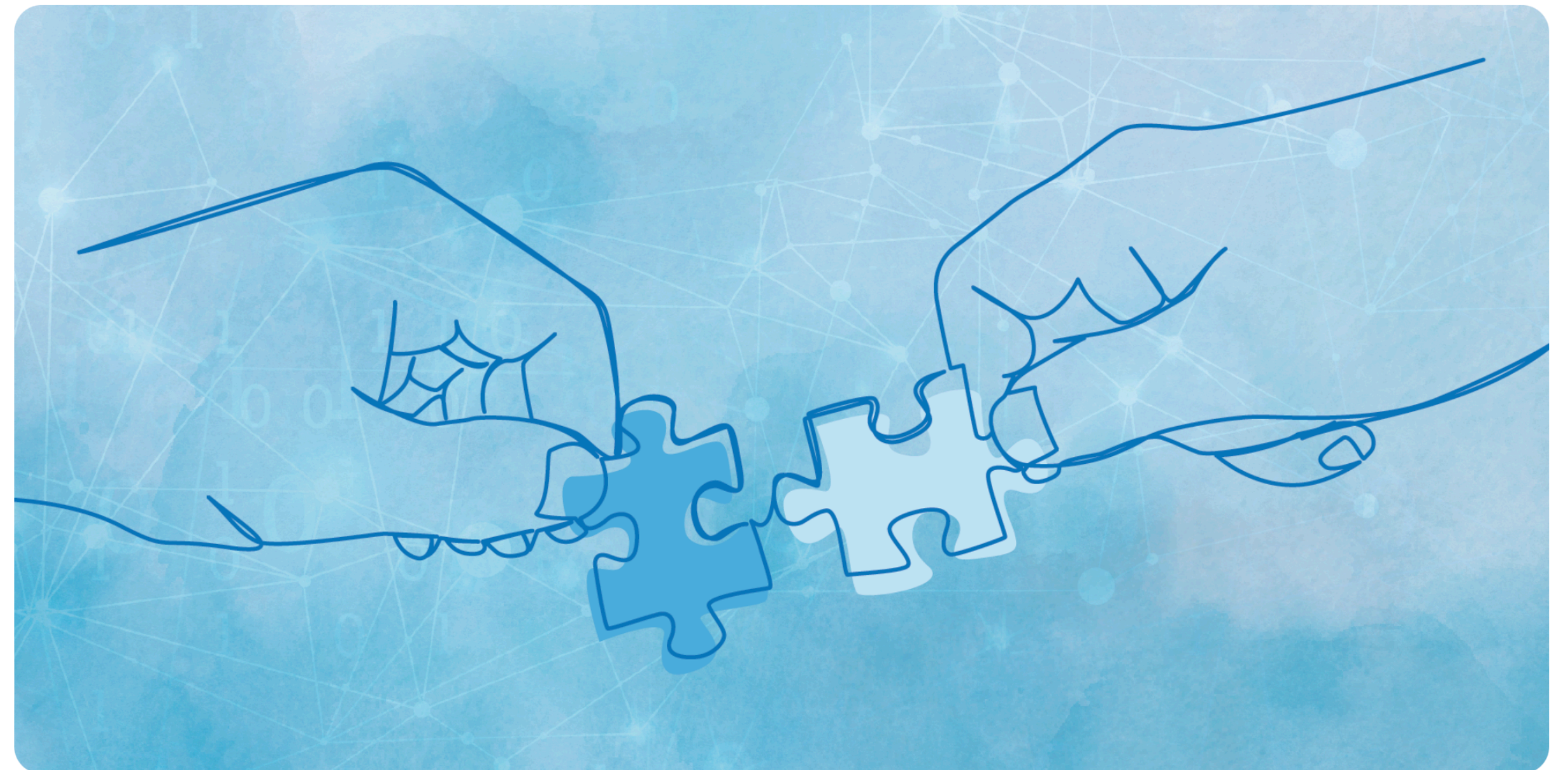
### News

14 Feb 2024



## NIH and GA4GH commit to ongoing collaboration

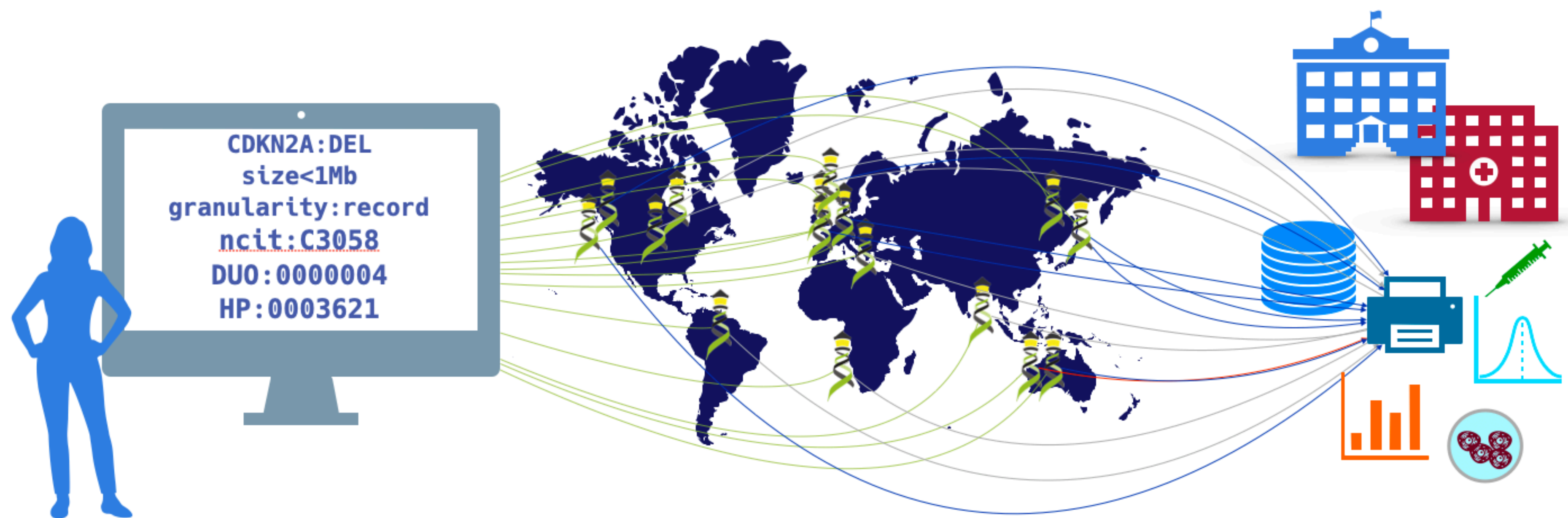
NIH and GA4GH strengthen their partnership to expand responsible data use for the benefit of human health through a Memorandum of Agreement.



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



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



# The GA4GH Beacon Protocol

**Federating Genomic Discoveries**



Beacon



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

**YES** | **NO** | \0

## Introduction

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

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

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

## Utility

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

A number of more useful first versions have been suggested.

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

“I would personally recommend all those be held for **version 2**, when the beacon becomes a service.”

Jim Ostell, 2014

## Implementation

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



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 Labora...

Brazilian Initiative on ...

BRCA Exchange

Broad Institute

Centre for Genomic R...

Centro Nacional de A...

Curoverse

EMBL European Biol...

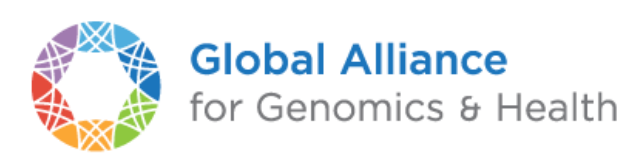
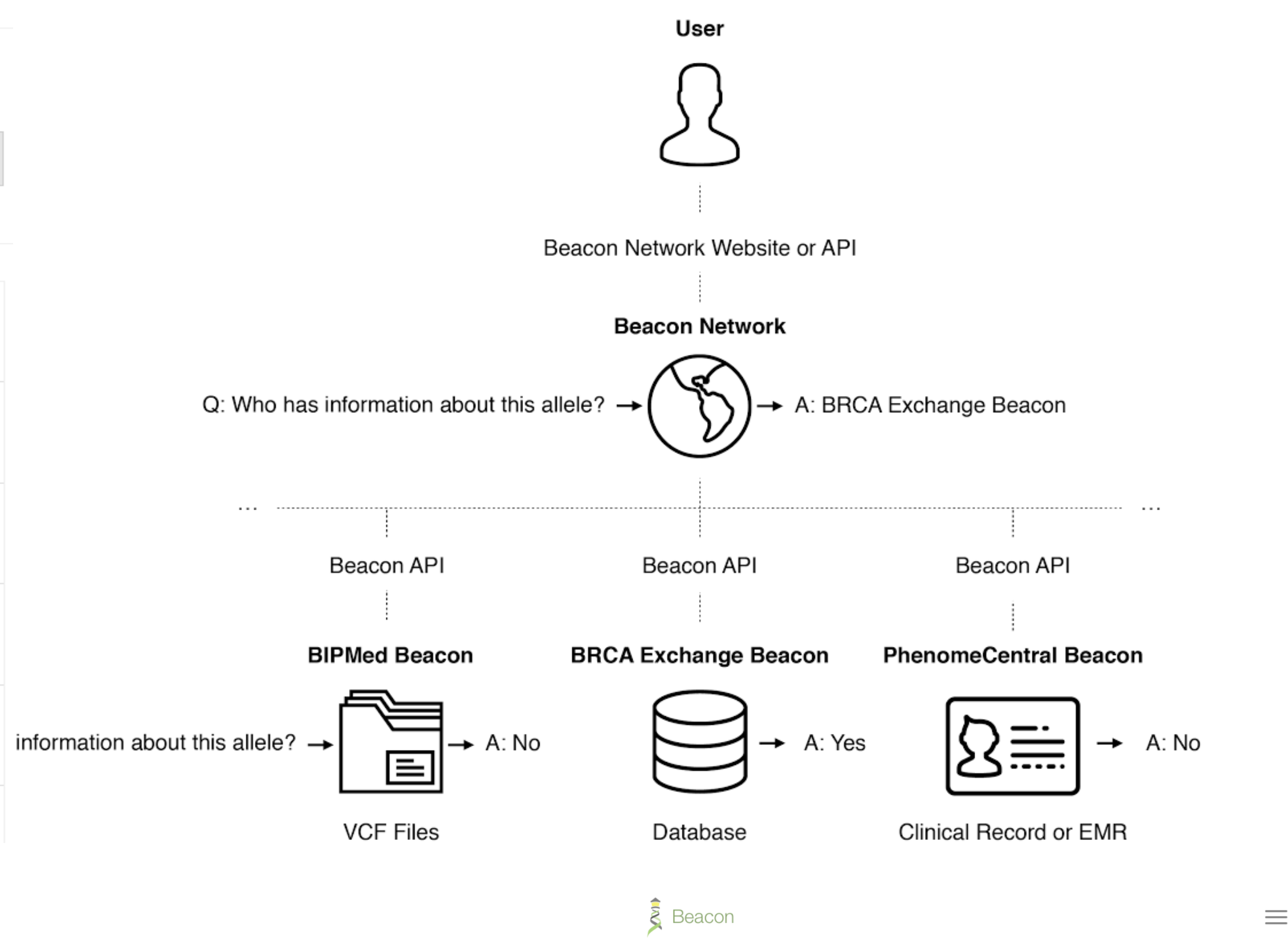
Global Alliance for G...

Google

Institute for Systems ...

Instituto Nacional de ...

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

## Beacon v2 Development

## Related ...

2014

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

2015

- beacon-network.org aggregator created by DNASTack

2016

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

2017

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

2018

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

2019

- ELIXIR Beacon Network

2020

- Beacon 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

2021

- framework + models concept implemented
- range and bracket queries, variant length parameters
- starting of GA4GH review process

2022

- further changes esp. in default model, aligning with Phenopackets and VRS
- unified beacon-v2 code & docs repository
- Beacon v2 approved at Apr GA4GH Connect

- ELIXIR starts Beacon project support

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

- new Beacon website (March)

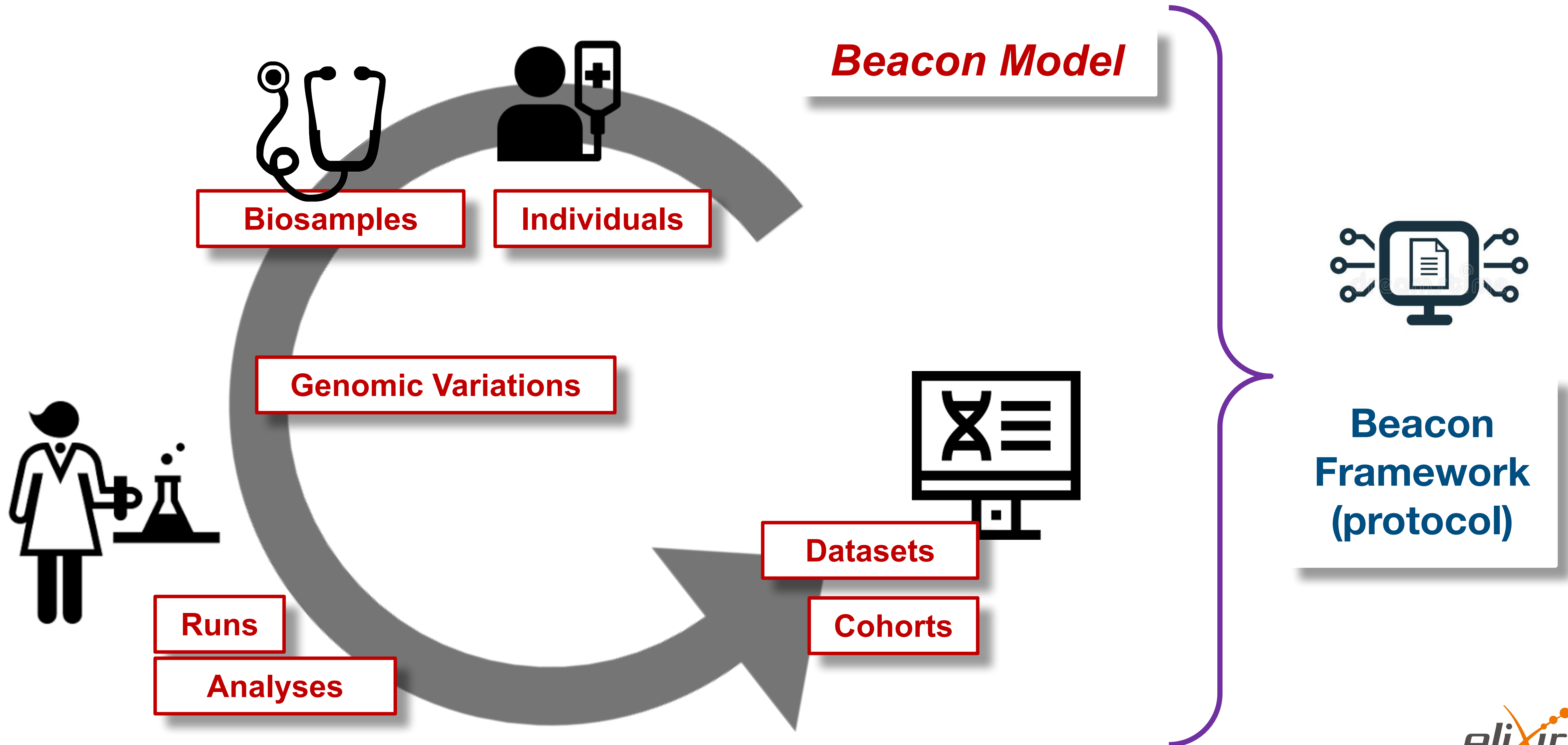
- Beacon publication at Nature Biotechnology

- Phenopackets v2 approved

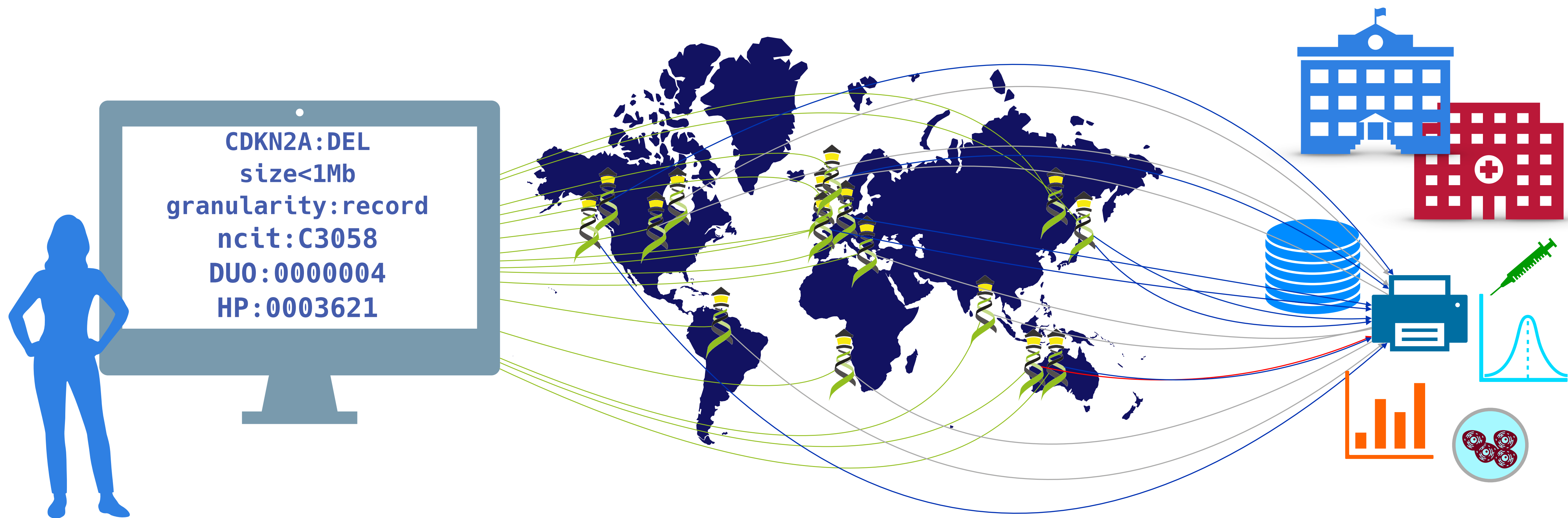
- [docs.genomebeacons.org](https://docs.genomebeacons.org)

# Beacon v2

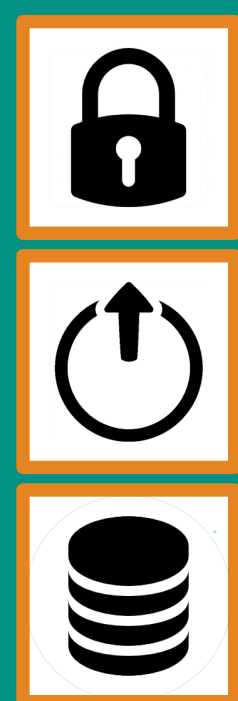
[docs.genomebeacons.org](https://docs.genomebeacons.org)







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

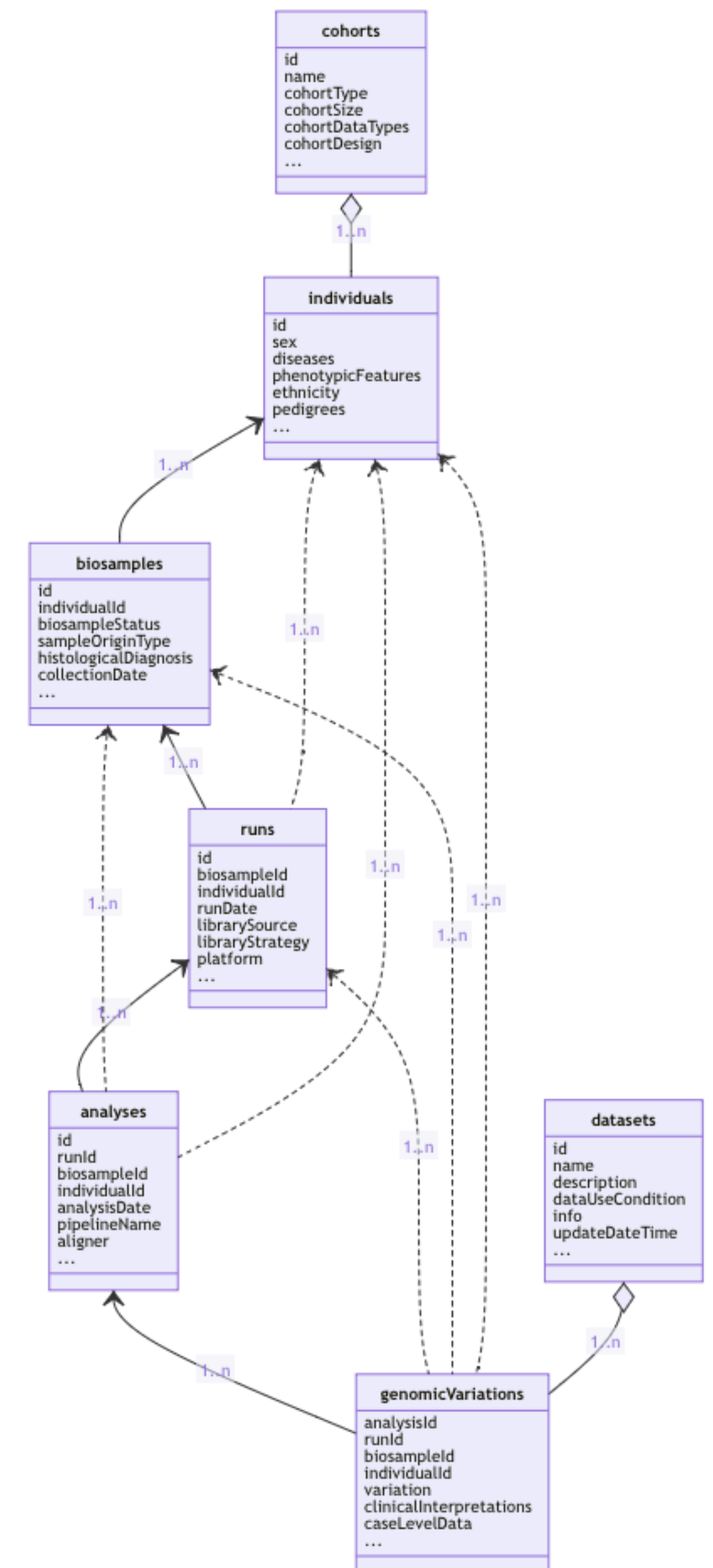


## Beacon v2 API

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

# Beacon Default v2 Model

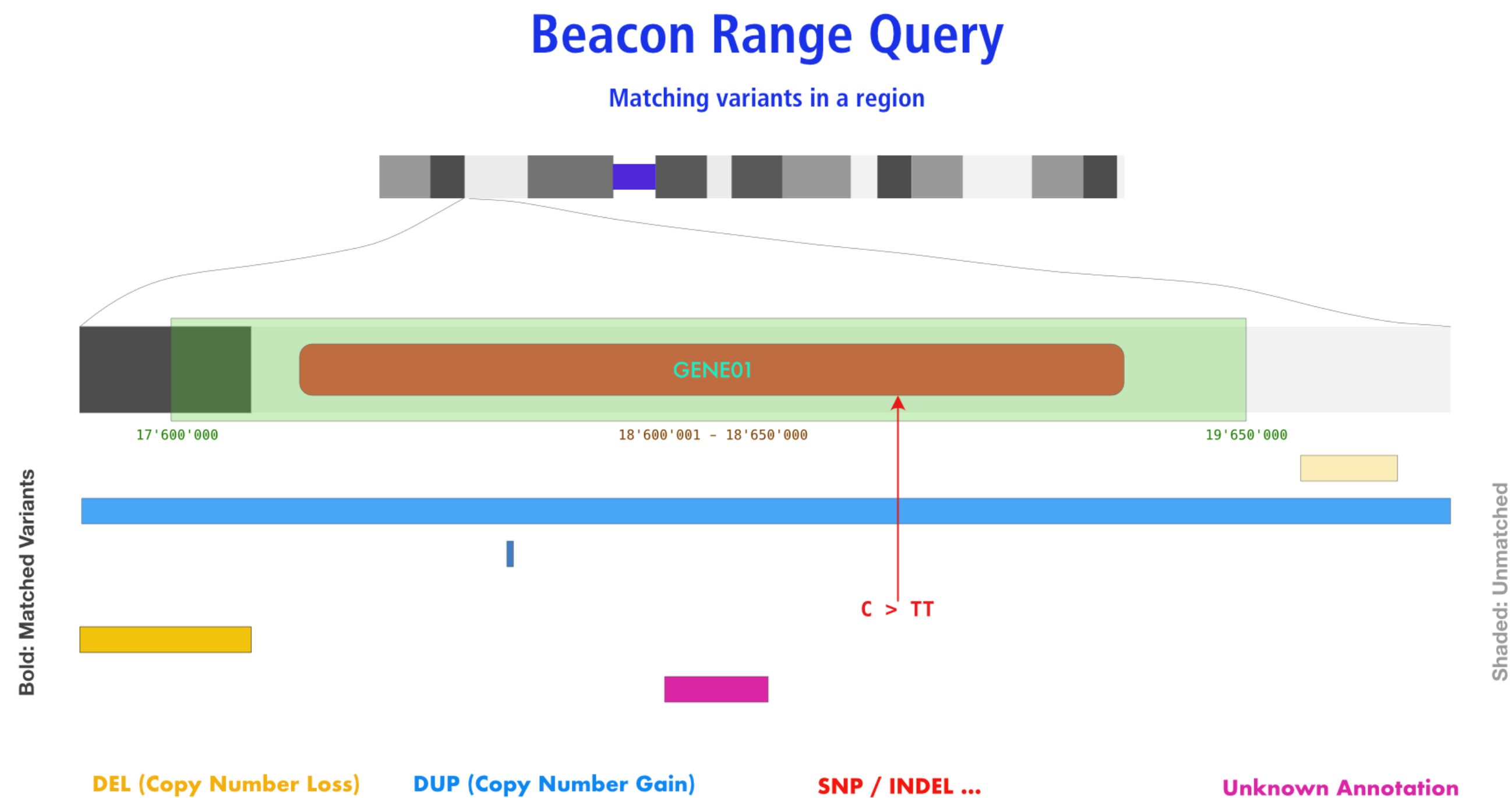
- The Beacon **framework** describes the overall structure of the API requests, responses, parameters, the common components, etc.
- Beacon **models** describe the set of concepts included in a Beacon, like individual or biosample, and also the relationships between them.
- Besides logical concepts,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



# 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 Sarr

### Dataset

Test Database - exemplez x

### Chromosome

17 (NC\_000017.11)

### Variant Type

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

### Start or Position

7572826

### End (Range or Structural Var.)

7579005

### Reference Base(s)

N

### Alternate Base(s)

A

### Select Filters

Select...

### Chromosome 17

7572826

7579005

Query Database

### Form Utilities

Gene Spans

Cytoband(s)

### Query Examples

CNV Example

SNV Example

Range Example

Gene Match

Aminoacid Example

Identifier - HeLa

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

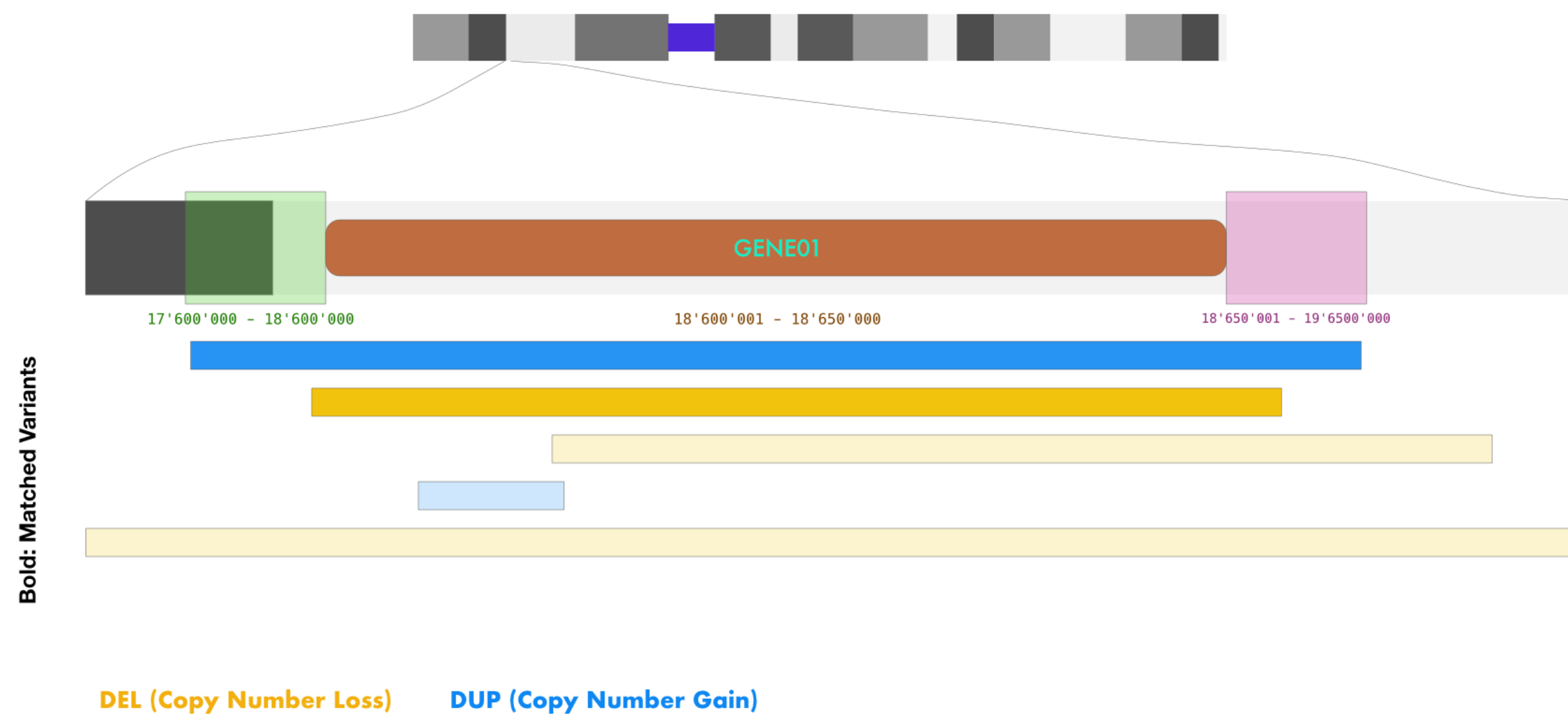
# Variation Queries

## Bracket ("CNV") Query

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

### Beacon Bracket Query

Example for complete regional match



### Beacon Query Types

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

---

**Dataset**

Test Database - examplez x | v

**Chromosome** i   **Variant Type** i

9 (NC\_000009.12) | v   EFO:0030067 (copy number deletion) | v

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

21000001-21975098   21967753-23000000

**Select Filters** i

NCIT:C3058: Glioblastoma (100) x | v

**Chromosome 9** i

21000001 21975098

21967753 23000000

**Query Database**

**Form Utilities**   **Gene Spans**   **Cytoband(s)**

**Query Examples**   **CNV Example**   **SNV Example**   **Range Example**   **Gene Match**

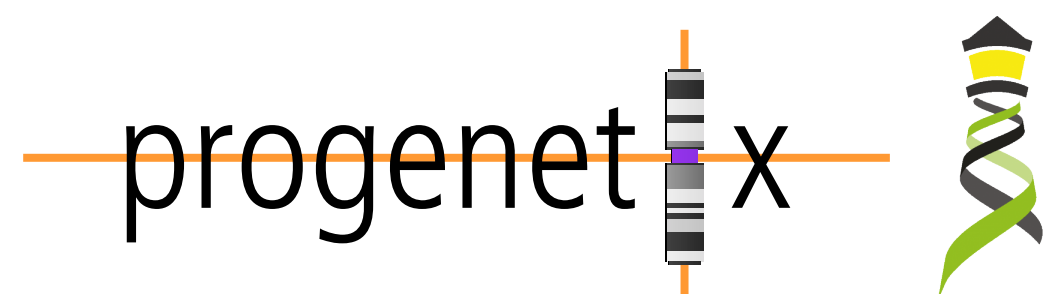
**Aminoacid Example**   **Identifier - HeLa**

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 NCIt 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](#)   [Show JSON Response](#)

Calls: 0   [Legacy Interface](#)

Samples: 523

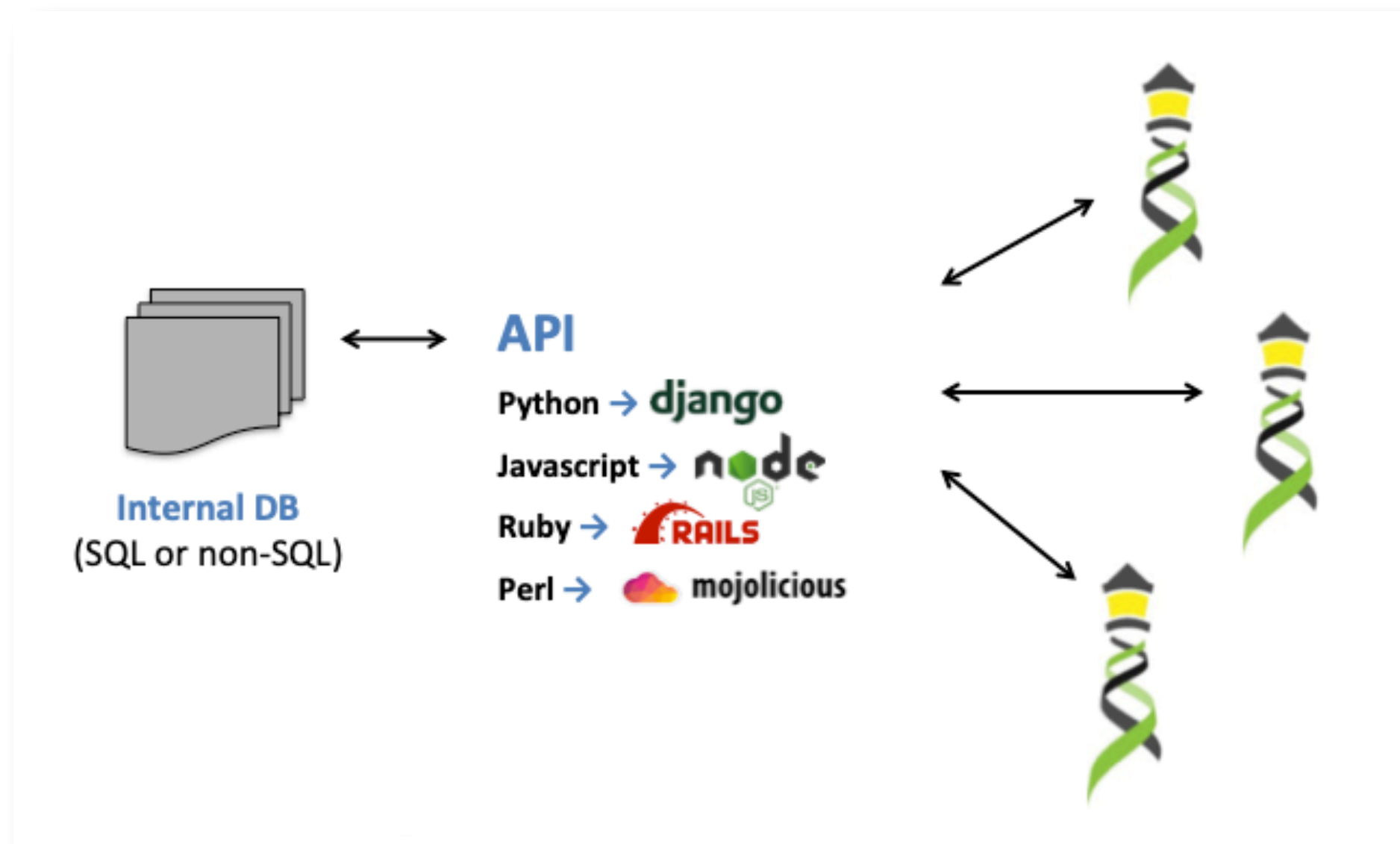
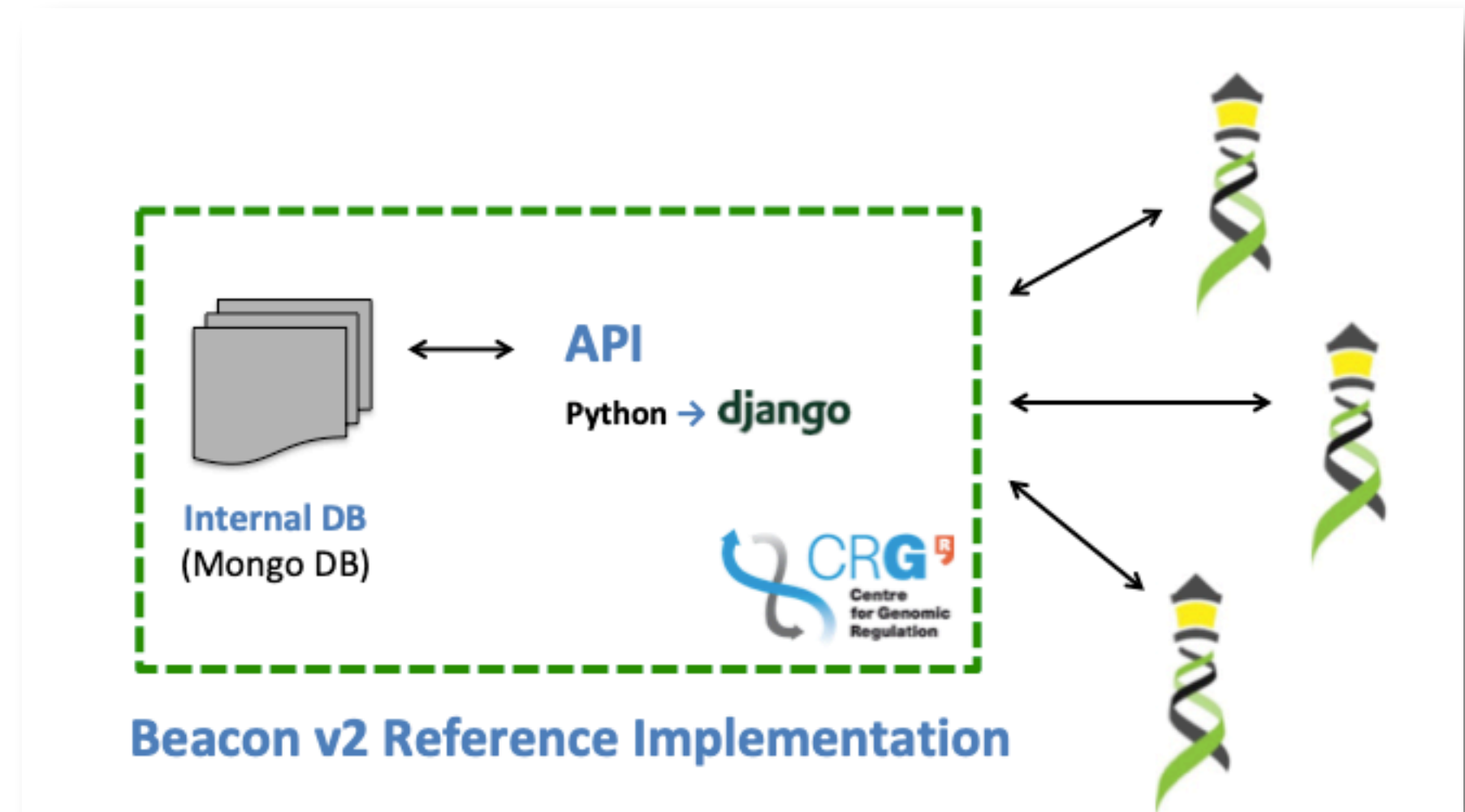
Results   **Biosamples**

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

Page 1 of 105

# Implementing Beacon v2

... its just code ㄟ( ͜ʖ)ㄟ



## Progenetix Stack

progenetix

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

React

APACHE  
HTTP SERVER PROJECT

python

```

_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: "5bab578b727983b2e8ca99e",
source_collection: "variants",
source_db: "progenetix",
source_key: "id",
target_collection: "variants",
target_count: 607,
target_key: "id",
target_values: [
  ObjectId("5bab578b727983b2e8ca99e"),
  ObjectId("5bab578b727983b2e8ca99e")
]
    
```

mongoDB

Entity collections

Utility collections

# *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 v2 GA4GH Approval Registry

Beacons:  EUROPEAN GENOME-PHENOME ARCHIVE  -progenetix-  cnag  UNIVERSITY OF LEICESTER

 EUROPEAN GENOME-PHENOME ARCHIVE


**European Genome-Phenome Archive (EGA)**

GA4GH Approval Beacon Test

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

Visit us  
Beacon API  
Contact us

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


 -progenetix-

**Theoretical Cytogenetics and Oncogenomics group at UZH and SIB**

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

Visit us  
Beacon UI  
Beacon API  
Contact us

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

 cnag


**Centre Nacional Analisis Genomica (CNAG-CRG)**

Beacon @ RD-Connect

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

Visit us  
Beacon API  
Contact us

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

 UNIVERSITY OF LEICESTER

**University of Leicester**

Cafe Variome Beacon v2

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

Beacon UI  
Beacon API  
Contact us

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

Beacon protocol response verifier at time of GA4GH approval Spring 2022

Matches the Spec Not Match the Spec Not Implemented



# bycon based Progenetix Stack

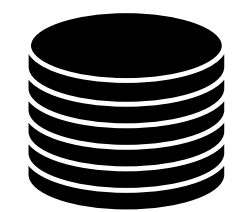


- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
  - biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads...
- the complete middleware / CGI stack is provided through the **bycon** package
  - schemas, query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

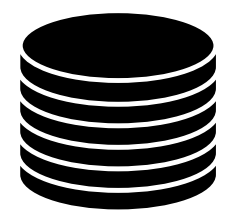


- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
  - PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703...
- *querybuffer* stores id values of all entities matched by a query and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

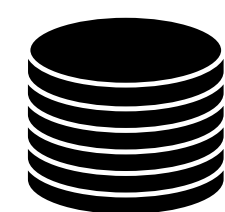
```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e0ca99e"),
  ObjectId("5bab578d727983b2e0cb505")
]
```



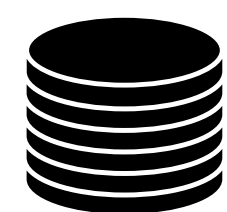
variants



analyses

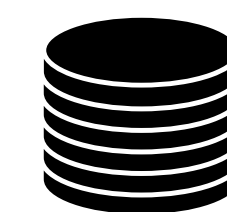


biosamples

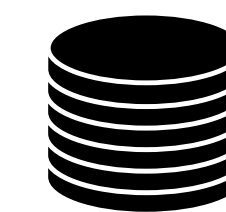


individuals

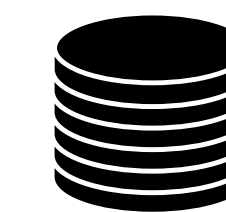
Entity collections



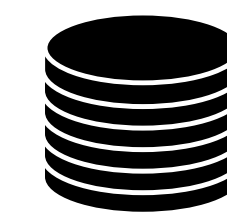
collations



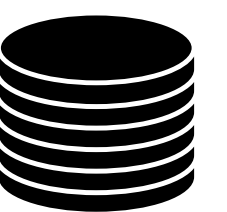
geolocs



genespans



publications




qBuffer

Utility collections

# Beacon v2 Conformity and Extensions in *bycon*

Putting the + into Beacon ...

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

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



progenetix / byconaut

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

[bycon.progenetix.org](https://bycon.progenetix.org)  
[github.com/progenetix/bycon/](https://github.com/progenetix/bycon/)

byconaut Public

main 2 branches

mbaudis get\_plot\_parameters

- bin
- docs
- exports
- imports
- local
- rsrc
- services
- tmp
- .gitignore
- LICENSE
- README.md
- \_\_init\_\_.py
- install.py
- install.yaml
- mkdocs.yaml

progenetix / beaconplus-web

Code Pull requests Actions Projects Security Insights Settings

beaconplus-web Public

forked from progenetix/progenetix-web

main 1 branch 0 tags

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

mbaudis code cleaning, no feature changes

- .github/workflows cleanup
- docs still first implementation clean-up
- extra documentation
- public graphic refinement
- src code cleaning, no feature changes
- .babelrc Simplify query generation and add
- .env.development first working version
- .env.local first working version
- .env.production env
- .env.staging env
- .eslintrc.json BioSubsetsPage perf optimisations

bycon Public

progenetix / bycon

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

bycon Public

main 4 branches 25 tags

mbaudis 1.3.6 852 commits

.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

[beaconplus.progenetix.org](https://beaconplus.progenetix.org)  
[.../progenetix/beaconplus-web/](https://github.com/progenetix/beaconplus-web/)

[bycon.progenetix.org](https://bycon.progenetix.org)  
[github.com/progenetix/bycon/](https://github.com/progenetix/bycon/)

### 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](#)

# pgxRpi

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

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

Bioconductor

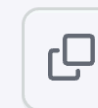
README.md

### pgxRpi

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

You can install this package from GitHub using:

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



For accessing metadata of biosamples/individuals, or learning more about filters, get started from the vignette [Introduction\\_1\\_loadmetadata](#).

For accessing CNV variant data, get started from this vignette [Introduction\\_2\\_loadvariants](#).

For accessing CNV frequency data, get started from this vignette [Introduction\\_3\\_loadfrequency](#).

For processing local pgxseg files, get started from this vignette [Introduction\\_4\\_process\\_pgxseg](#).

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

### pgxRpi

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

DOI: [10.18129/B9.bioc.pgxRpi](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 at uzh.ch>

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

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

# Beacon Security



Beacon



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

**YES** | **NO** | \0



# Genome *Beacons* Compromise Security?

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

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

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

OCT 29  
2015

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

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

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



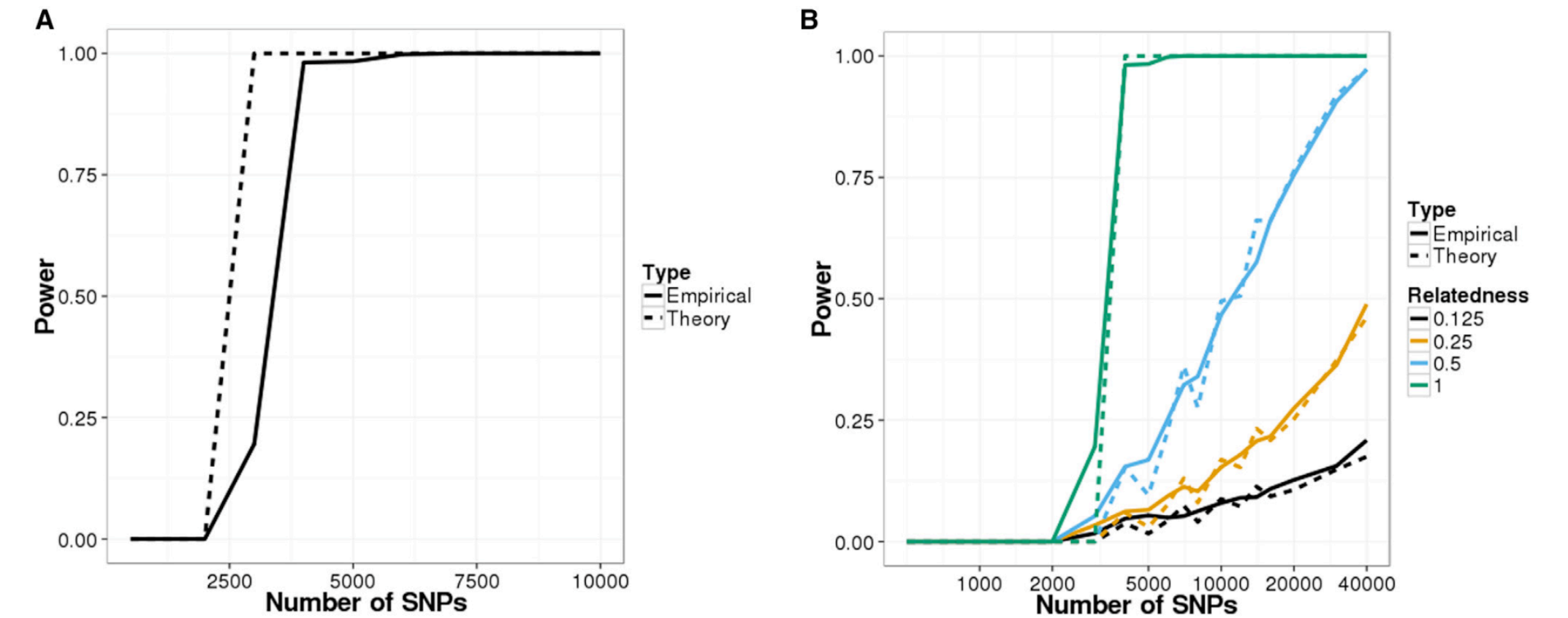
Stanford researchers are working with the Global Alliance for Genomics and Health to make genomic information in the Beacon Project more secure.  
*Science photo/Shutterstock*

# IDENTIFICATION OF INDIVIDUALS FROM MIXED COLLECTIONS USING RARE ALLELES

## Privacy Risks from Genomic Data-Sharing Beacons

Suyash S. Shringarpure<sup>1,\*</sup> and Carlos D. Bustamante<sup>1,\*</sup>

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



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

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





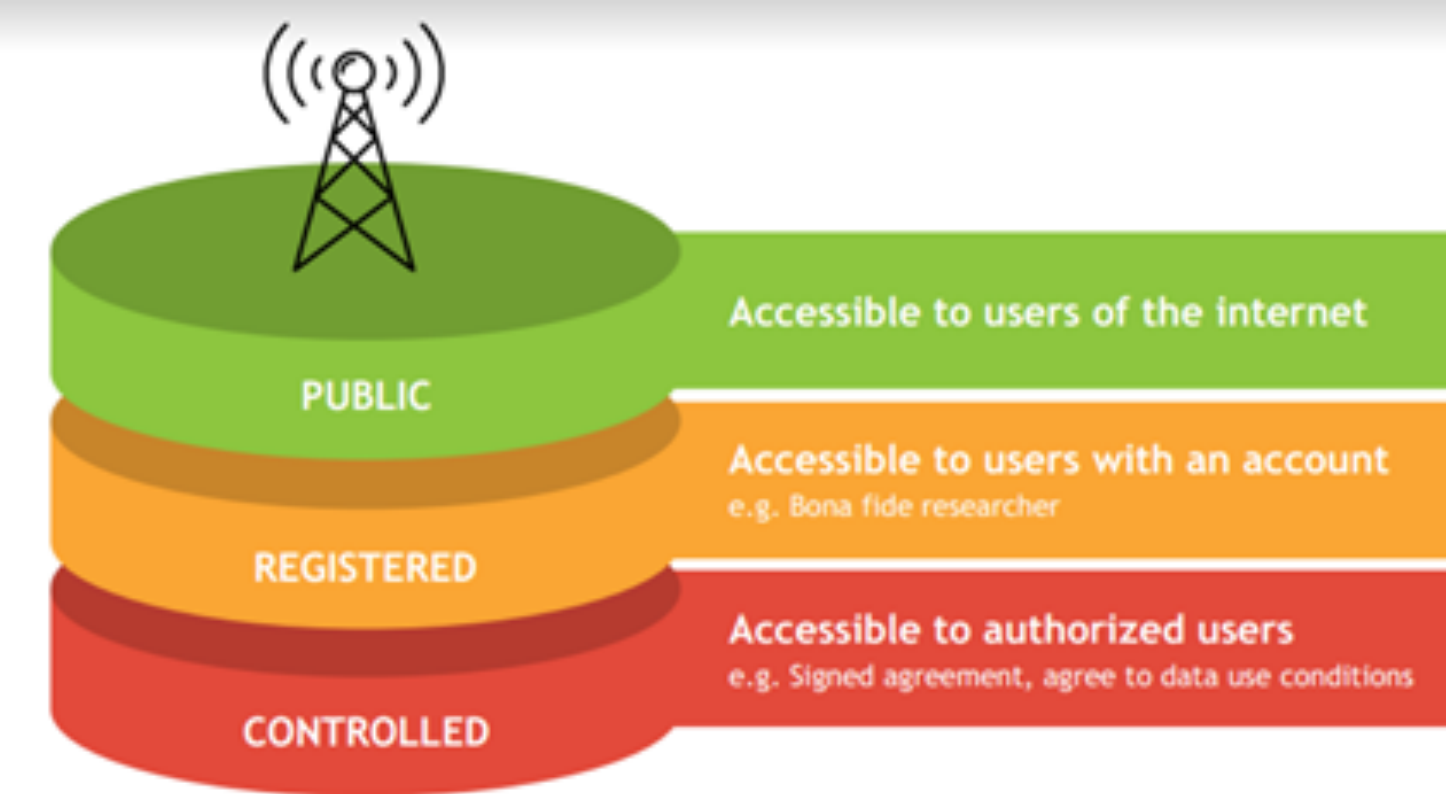
# Making Beacons Biomedical - Beacon v2

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

Definitely breaks the  
"Relative Security  
by Design"  
Concept!

# Beacon Security

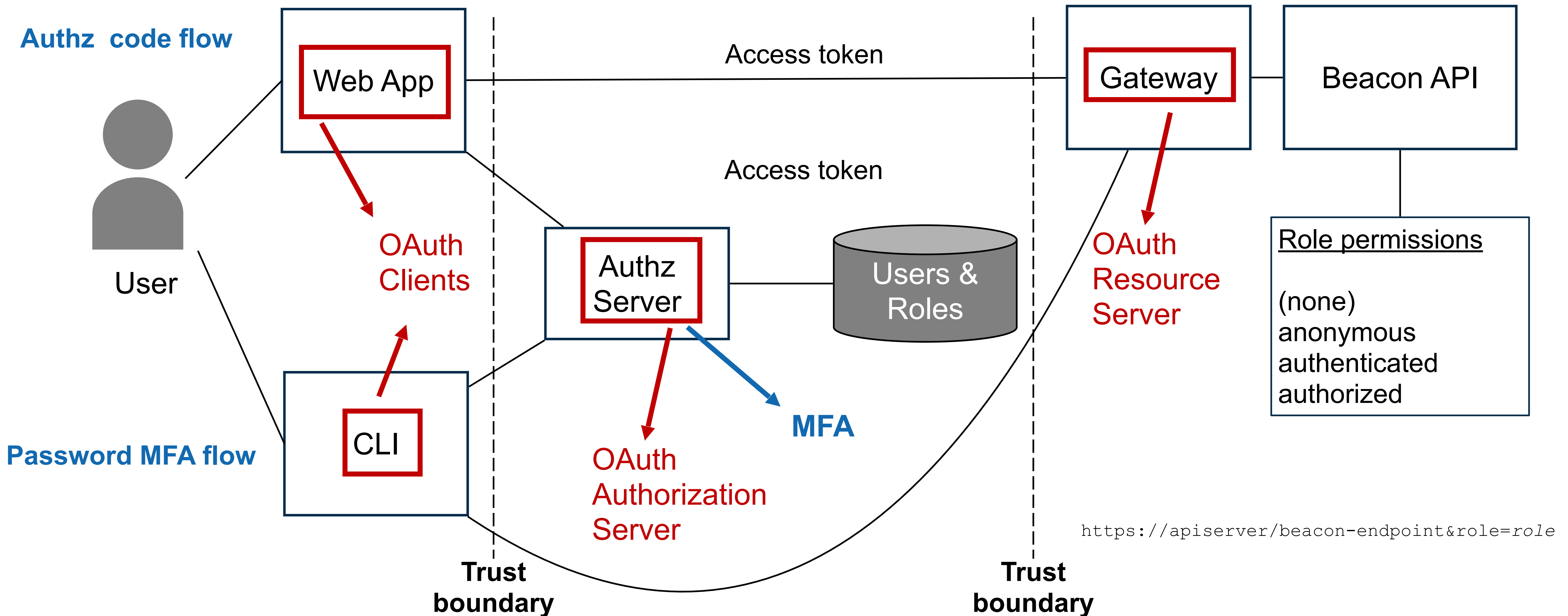
## Security by Design ... if Implemented in the Environment



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

# Architecture

## Running the *bycon* stack in a secure environment



# Architecture

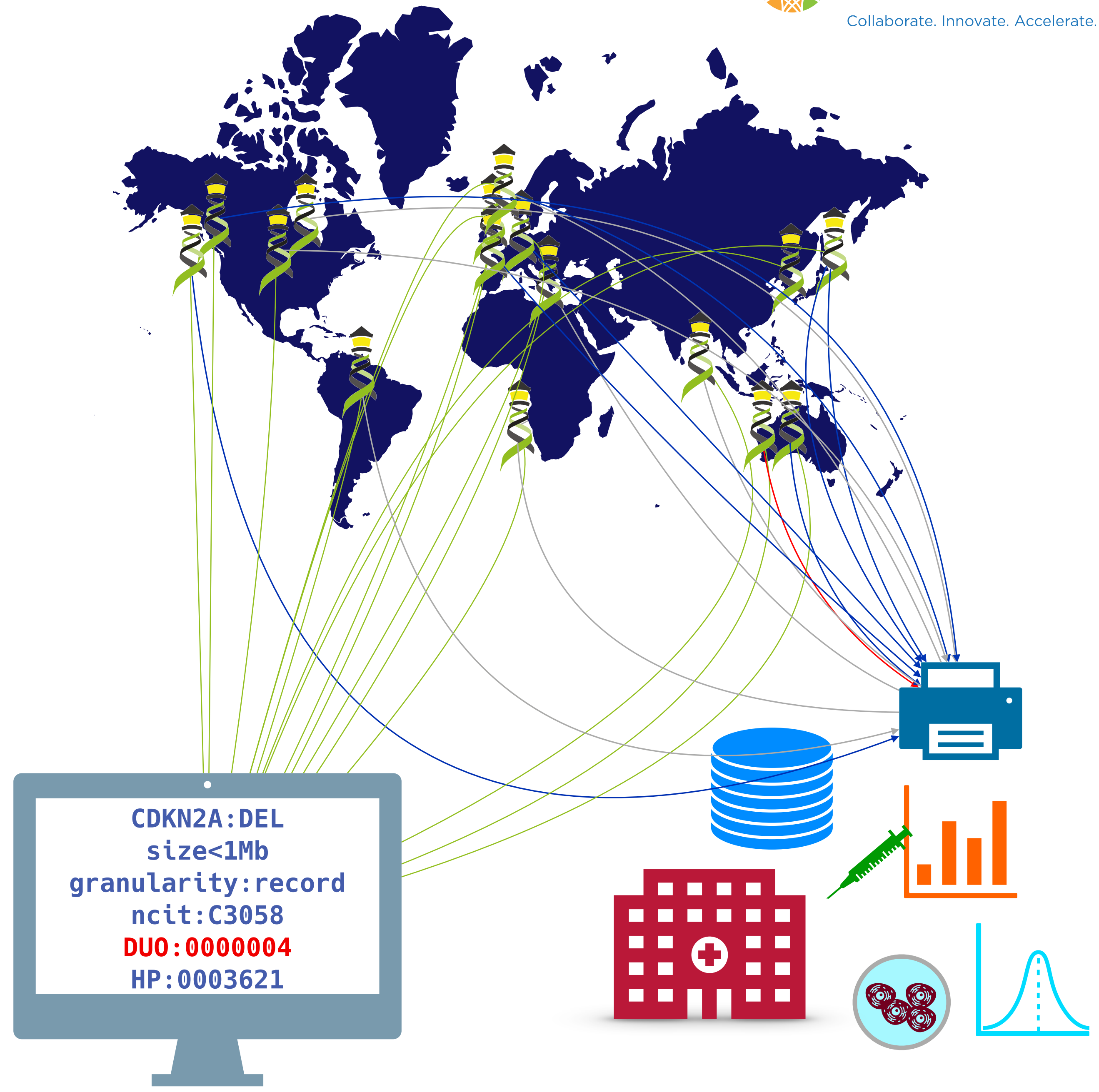
## Running the *bycon* stack in a secure environment

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

# What Can You Do?

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

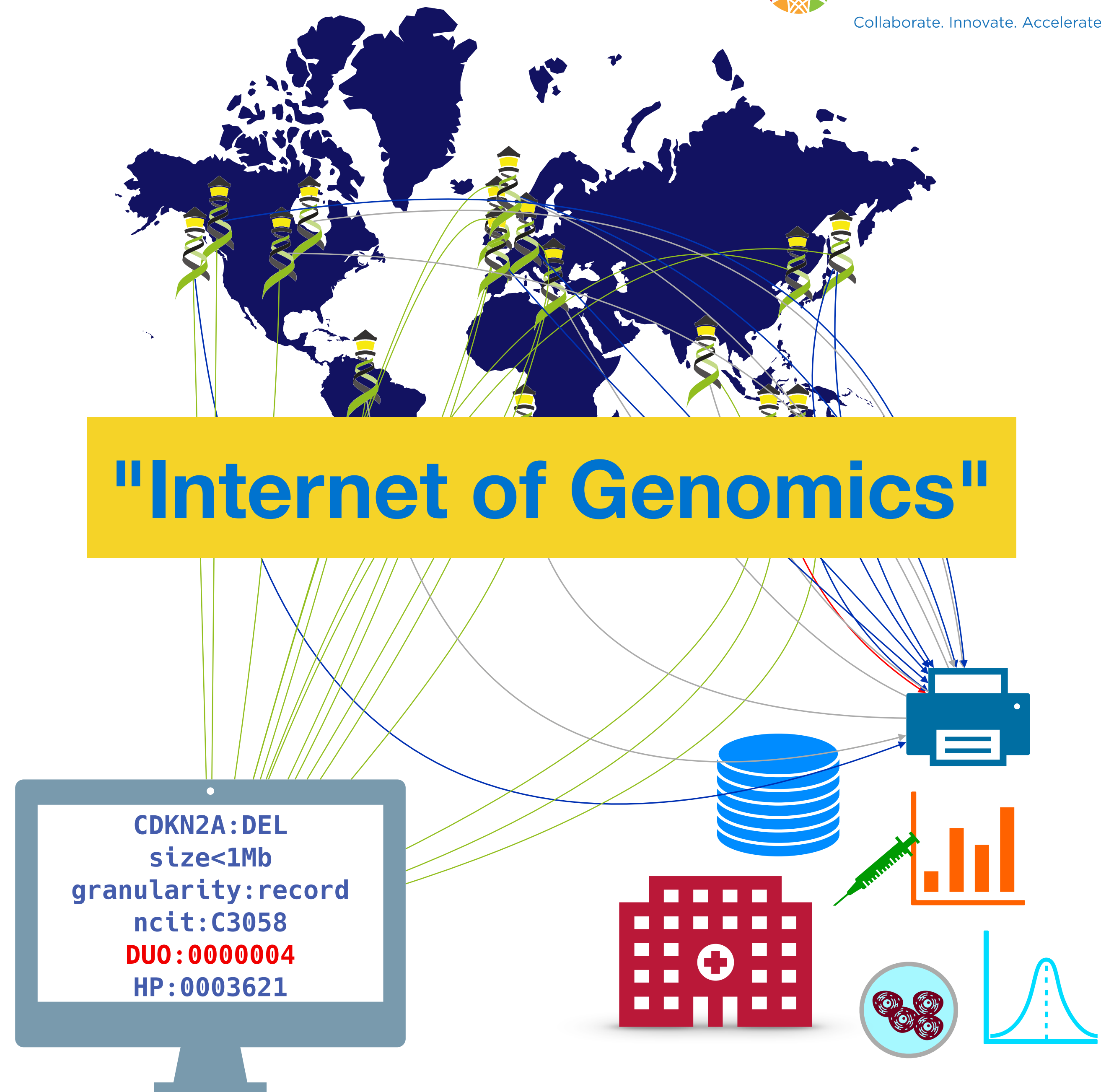
 **Collaborate!**



# What Can You Do?

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

➔ **Collaborate!**





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



Juha Törnroos  
 Teemu Kataja  
 Ilkka Lappalainen  
 Dylan Spalding



**Tony Brookes**  
**Tim Beck**  
 Colin Veal  
 Tom Shorter



**Michael Baudis**  
**Rahel Paloots**  
**Hangjia Zhao**  
**Ziying Yang**  
 Bo Gao  
 Qingyao Huang



**Augusto Rendon**  
**Ignacio Medina**  
 Javier López  
 Jacobo Coll  
 Antonio Rueda



centre nacional d'anàlisi genòmica  
 centro nacional de análisis genómico  
**Sergi Beltran**  
 Carles Hernandez



Institut national  
 de la santé et de la recherche médicale  
 David Salgado



**Salvador Capella**  
 Dmitry Repchevski  
 JM Fernández



**Laura Furlong**  
 Janet Piñero



**Serena Scollen**  
 Gary Saunders  
 Giselle Kerry  
 David Lloyd



**Nicola Mulder**  
 Mamana  
 Mbiyavanga  
 Ziyaad Parker



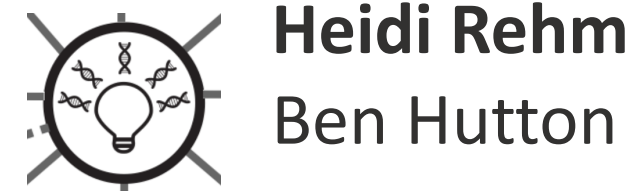
**David Torrents**  
**Dean Hartley**



**Joaquin Dopazo**  
 Javier Pérez  
 J.L. Fernández  
 Gema Roldan



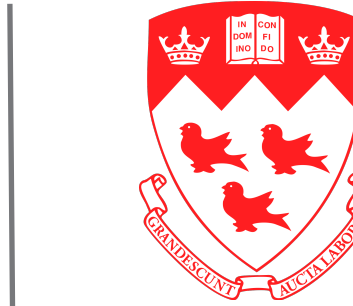
**Thomas Keane**  
 Melanie Courtot  
 Jonathan Dursi



**Heidi Rehm**  
 Ben Hutton



**Toshiaki Katayama**



**Stephane Dyke**



**Marc Fiume**  
 Miro Cupak



**BRCA EXCHANGE**  
**Melissa Cline**



EMBL-EBI  
**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





Universität  
Zürich<sup>UZH</sup>



Swiss Institute of  
Bioinformatics

