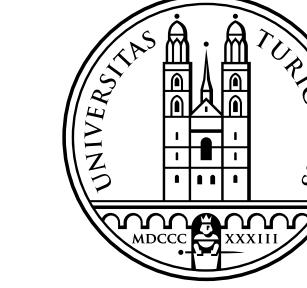




**Global Alliance**  
for Genomics & Health  
Collaborate. Innovate. Accelerate.



Swiss Institute of  
Bioinformatics



Universität  
**Zürich**<sup>UZH</sup>

# Reference Resources and Standards Development for Biomedical Genomics and Cancer Research



**Michael Baudis**

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



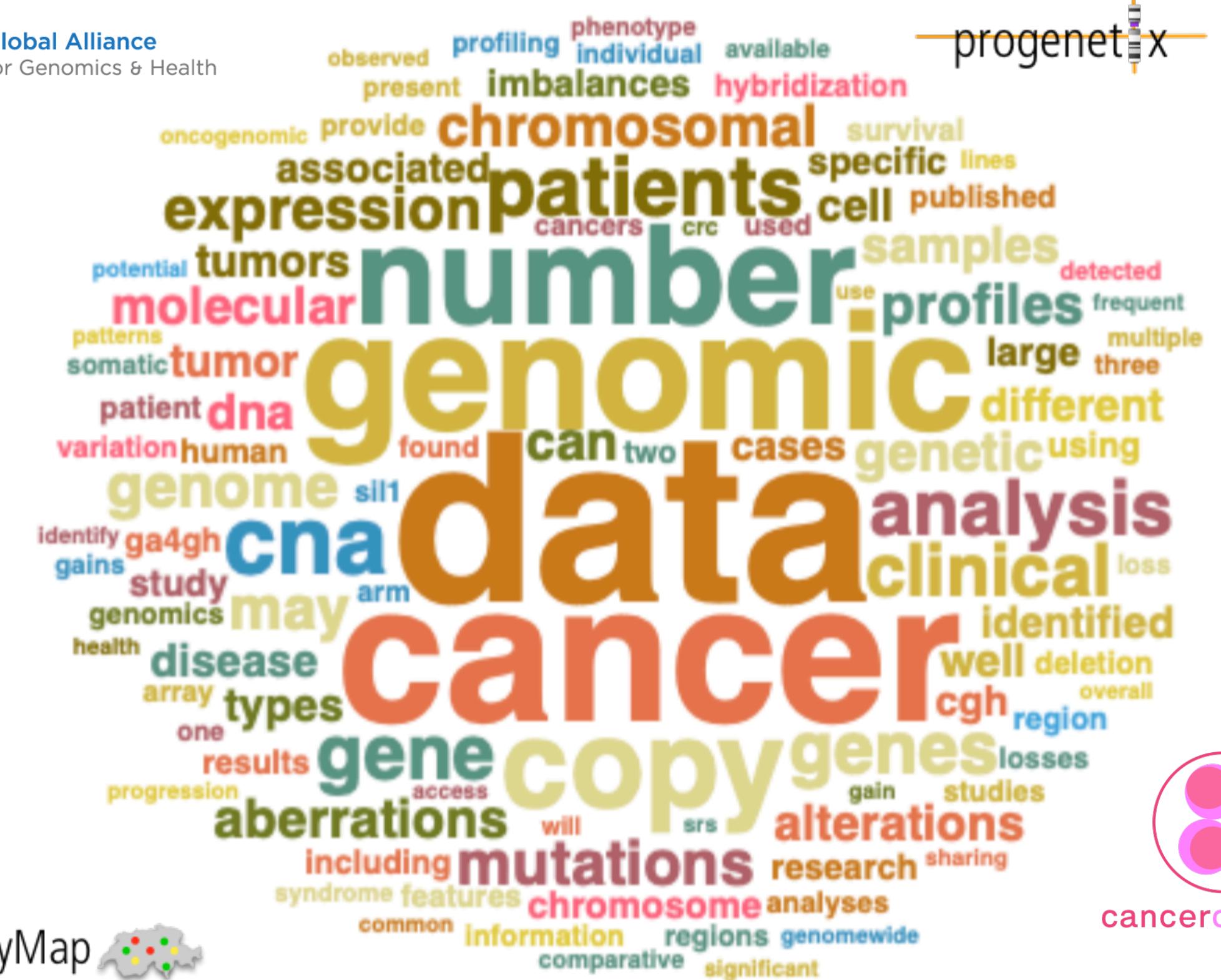


## Theoretical Cytogenetics and Oncogenomics

Our work @ UZH:

- **cancer** genome repositories
- biocuration
- protocols & formats

Curators  
~~Data Parasites~~



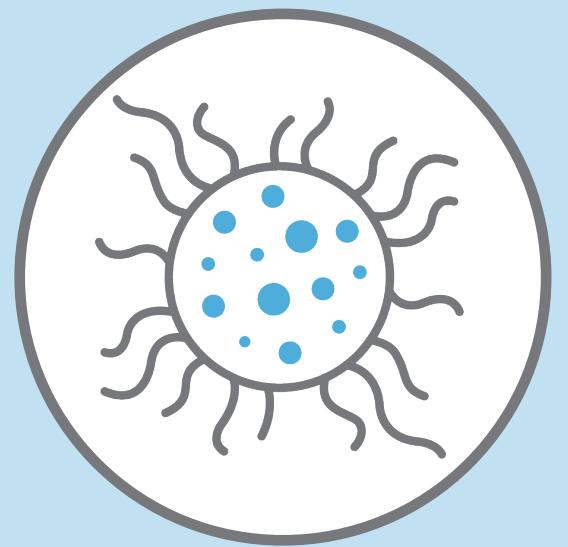


## Genome screening at the core of “Personalised Health”

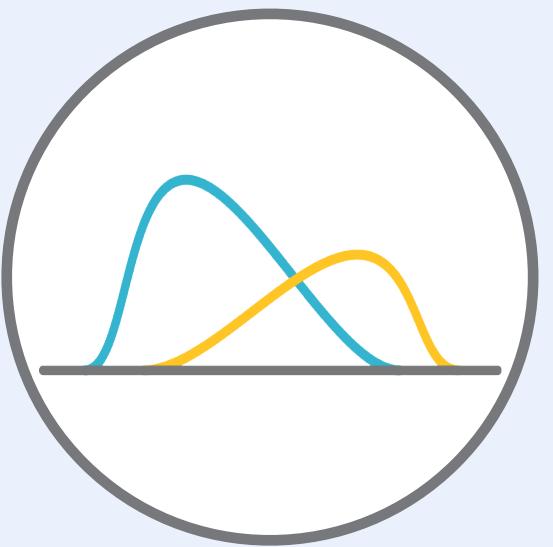
- ▶ **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- ▶ The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**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



# 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



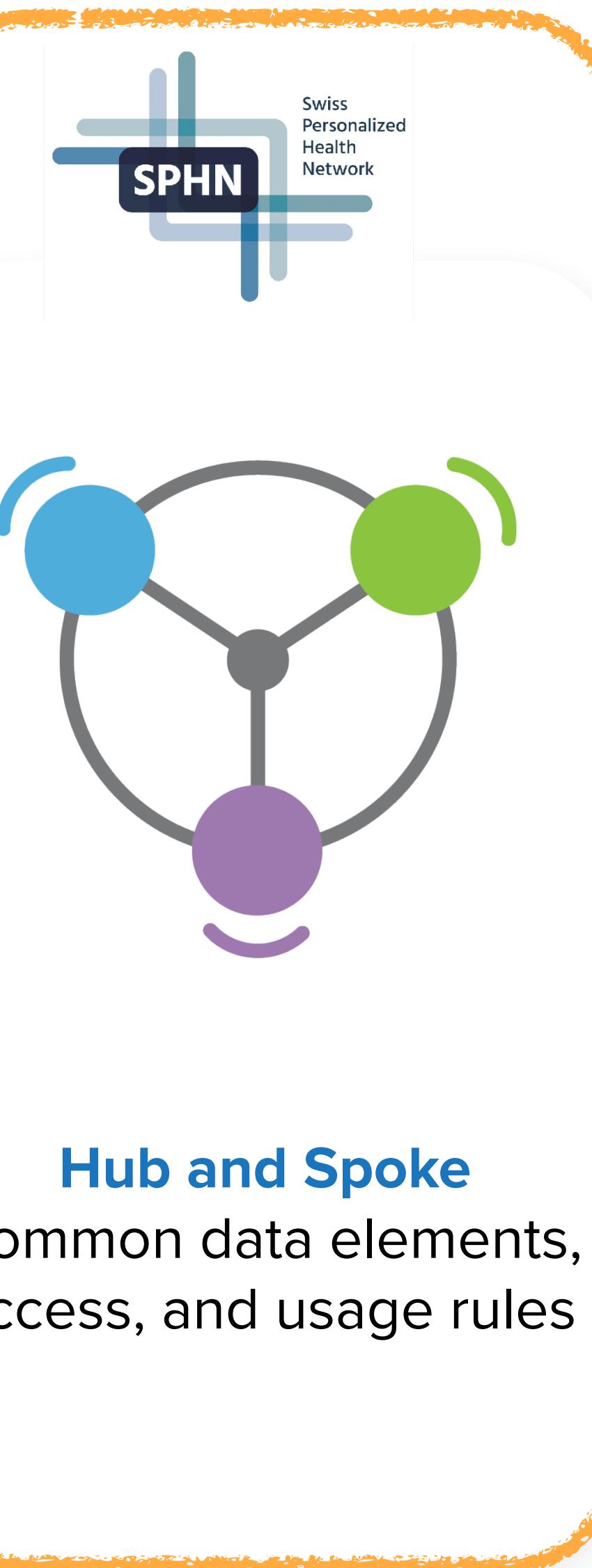
**Centralized Genomic Knowledge Bases**



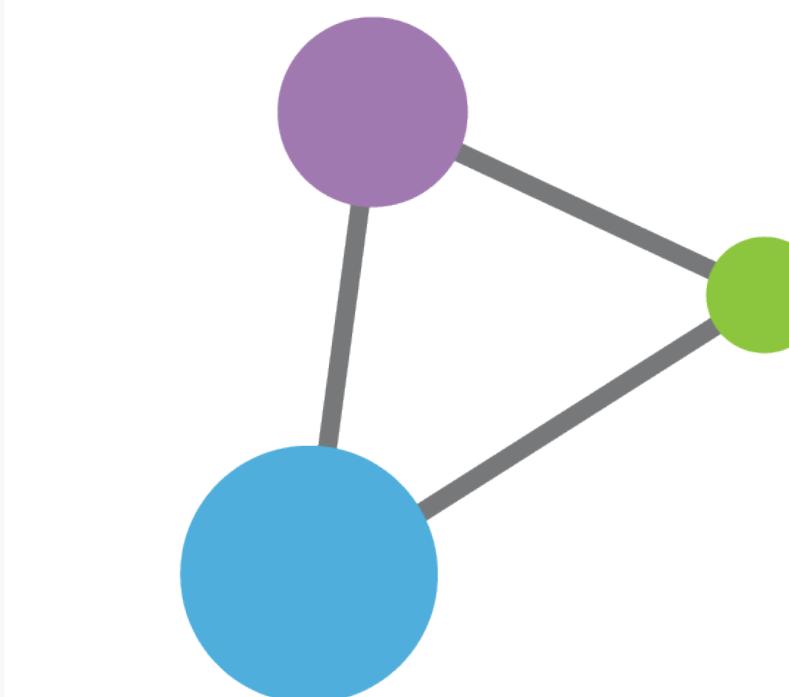
European  
Genome-  
Phenome  
Archive



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



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



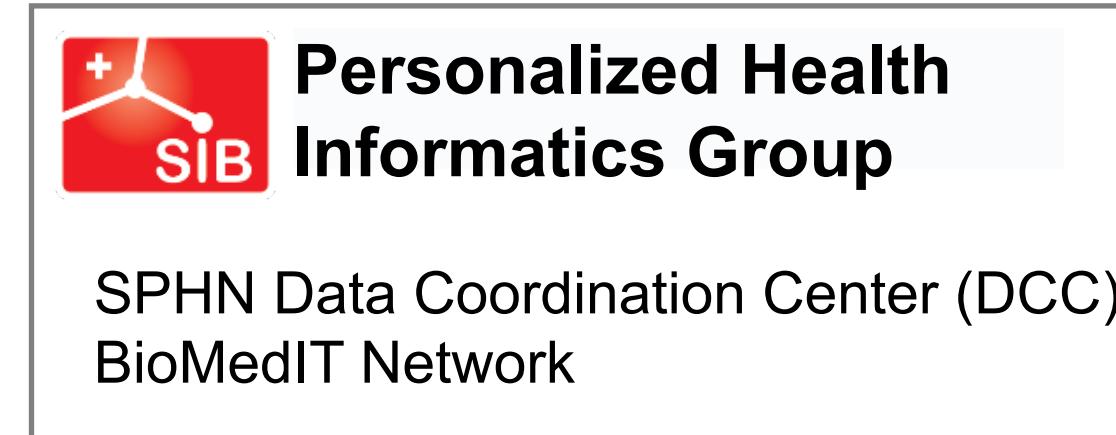
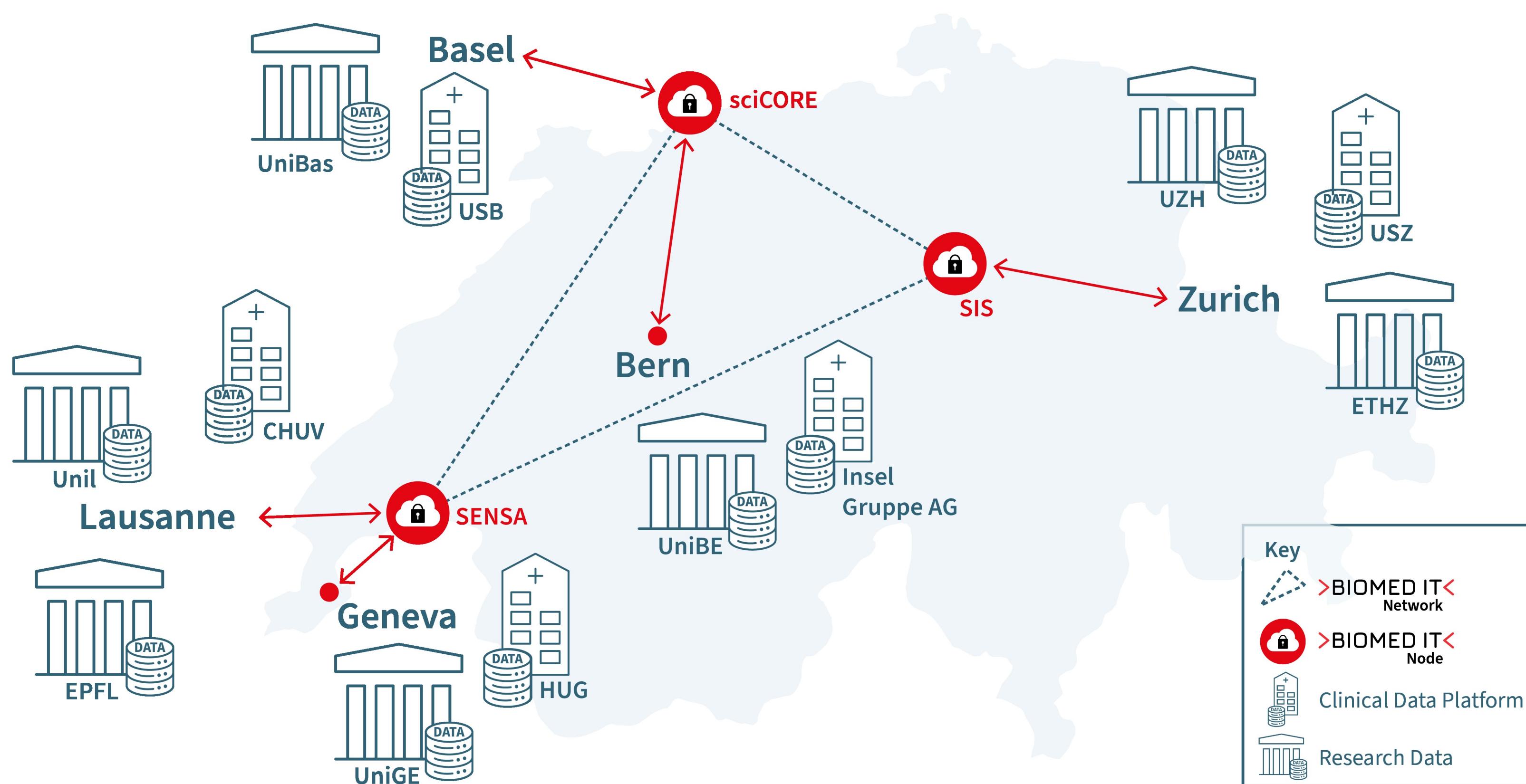
**Linkage of distributed and disparate datasets**

# The Swiss Personalized Health Network (SPHN)

## Creation of a scalable and sustainable data-enabling environment

- Including routine health data, molecular / omics data, registry data, clinical research data, and other health-related data types
  - Research infrastructure initiative funded 2017-2024 by the Swiss Government with CHF135 million;
  - Operating under a common Ethical Framework and one Information Security Policy, incl. the setup of a Trusted Research Environment
  - Foreseen consolidation of data coordination efforts with CHF 21 million 2025-2028
- Enable institutions to responsibly share interoperable health data
- Enable researchers to access, integrate, and analyze data

# The Swiss Personalized Health Network



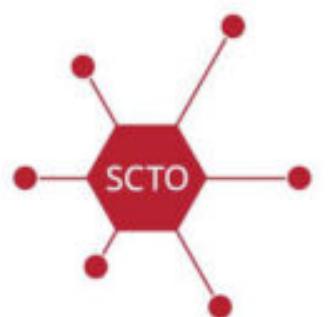
**swissuniversities**



**ehealthsuisse**



**Personalized Health Alliance**  
Basel-Zurich



**life sciences  
cluster** basel



# SPHN infrastructures, services and support



**UH Clinical Data Platforms**

SPHN Connector tool: Easy on-boarding of data providers

**Data discoverability and reuse**

Swiss cohort findability  
**mælstrom**  
Federated data exploration  
**TUNE INSIGHT**  
SPHN Metadata catalogue

FOPH Diagnosis	discharge-diagnosis given respecting the rules of FOPH and used for building the DRGs, e.g. K35 ICD10 acute appendicitis
Healthcare Encounter	an interaction between an individual and a specific unit or service of a healthcare provider, institute, e.g. emergency, intensive care unit, for the purpose of providing healthcare service(s)
Heart Rate	frequency of the heart beats, i.e. the number of time a heart beats per unit of time
ICD-Diagnosis	ICD diagnosis
Lab Result	laboratory analysis transmitted

**BioMedIT Network**

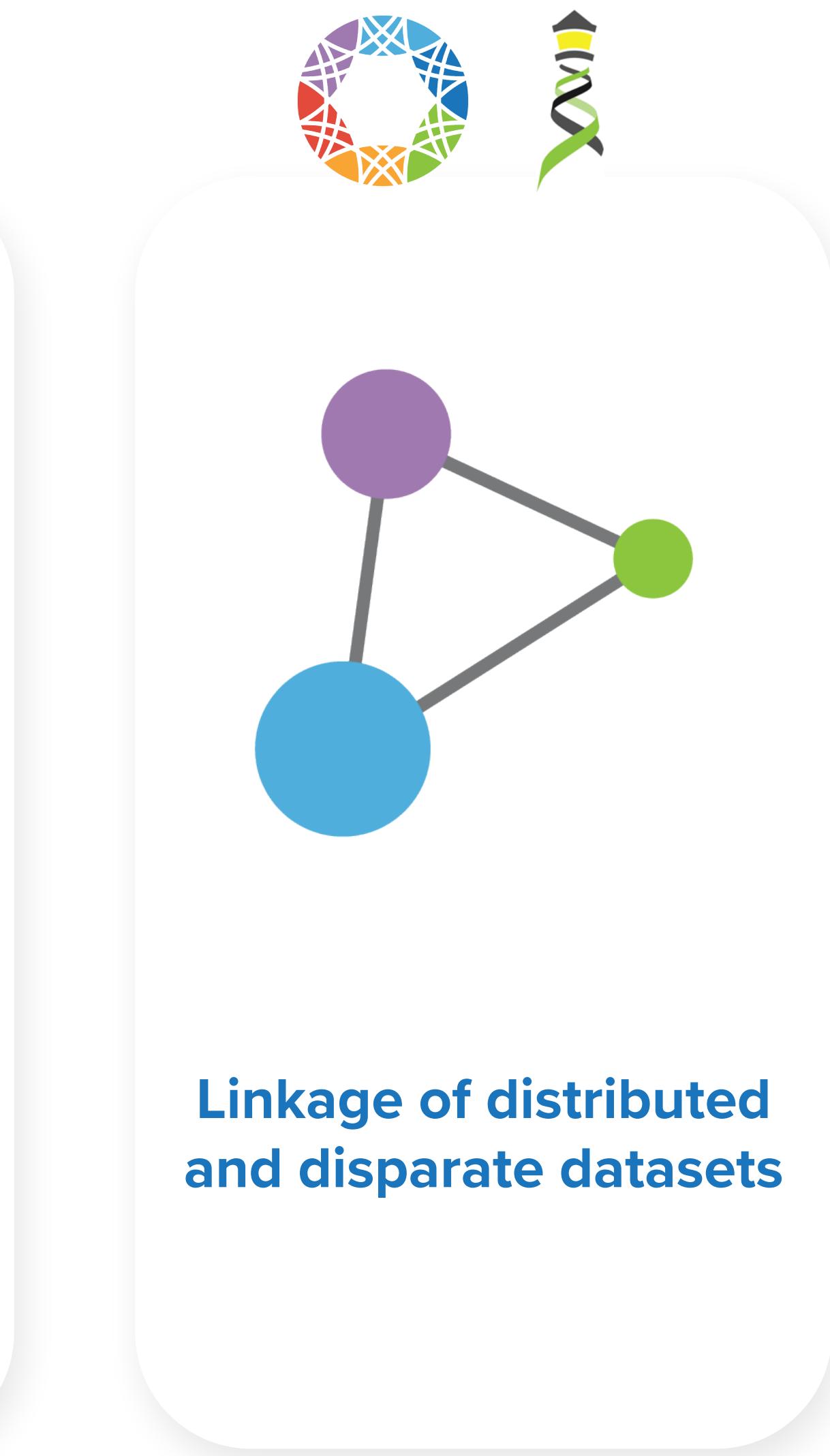
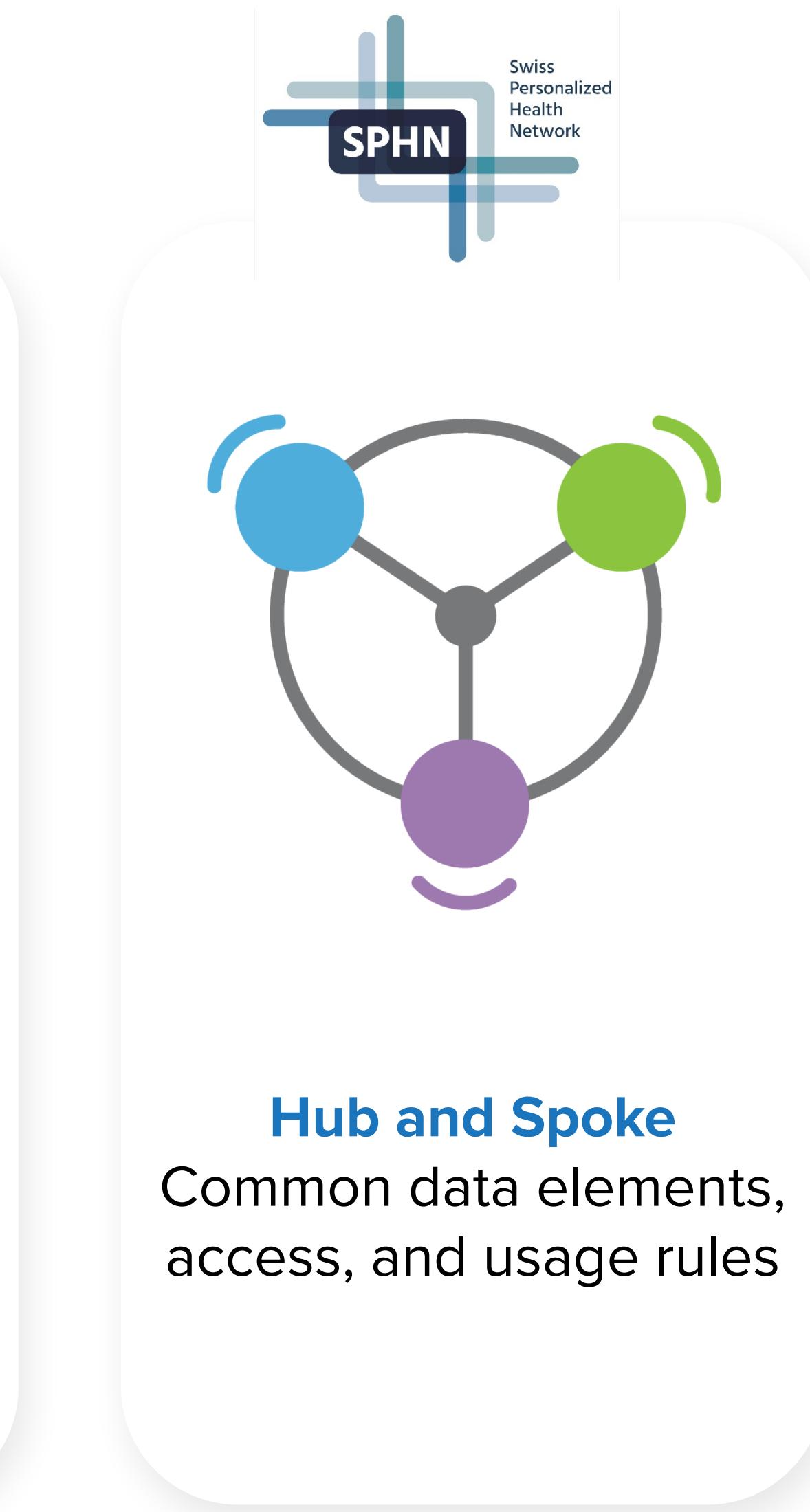
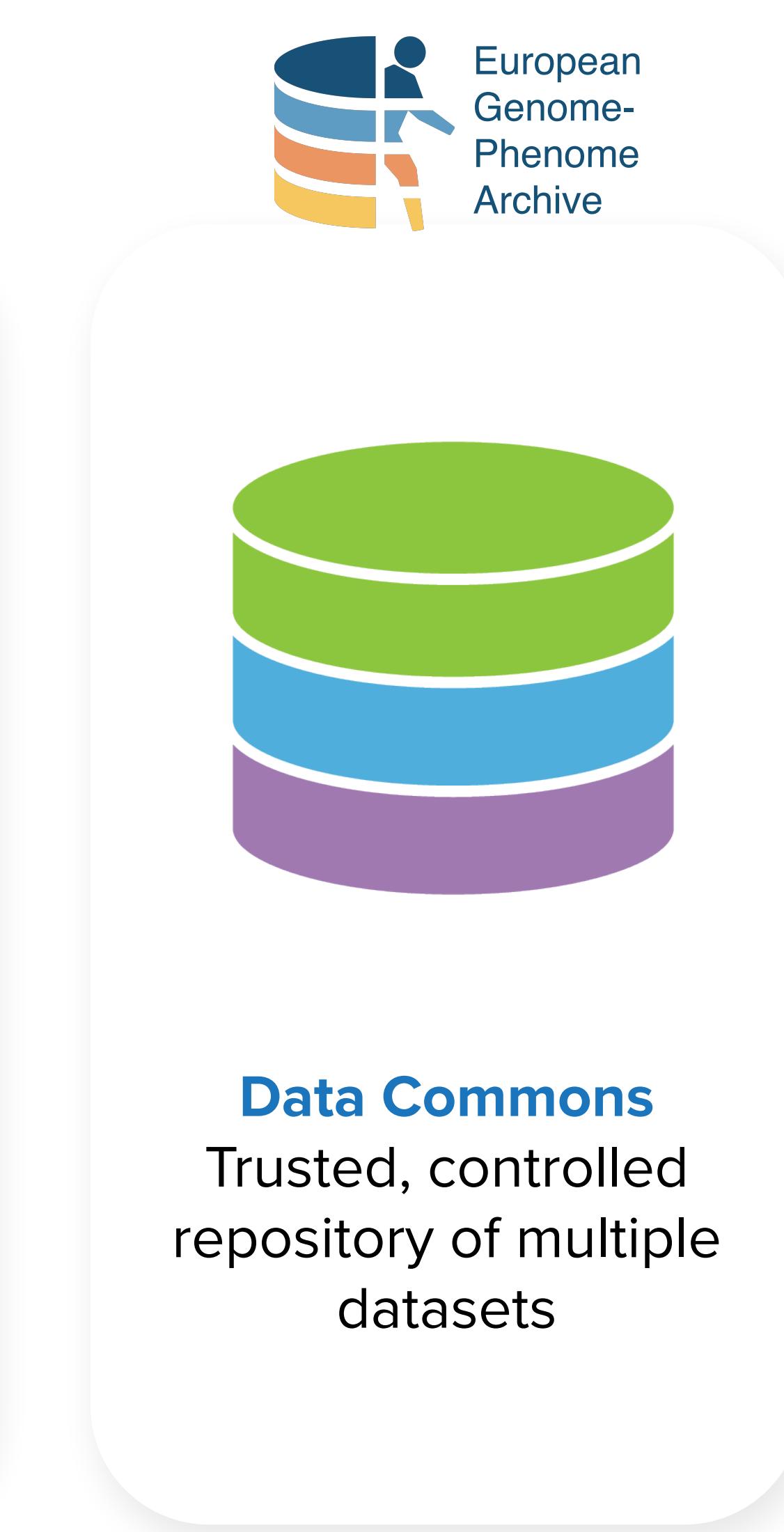
**SPHN Interoperability Framework**

**Legal Framework**

**Outreach and Training**

- International alignment and collaborations
- Responsible use of health-related data for research
- FAIR Health data for research
- Information Security and privacy

# Different Approaches to Data Sharing



## 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

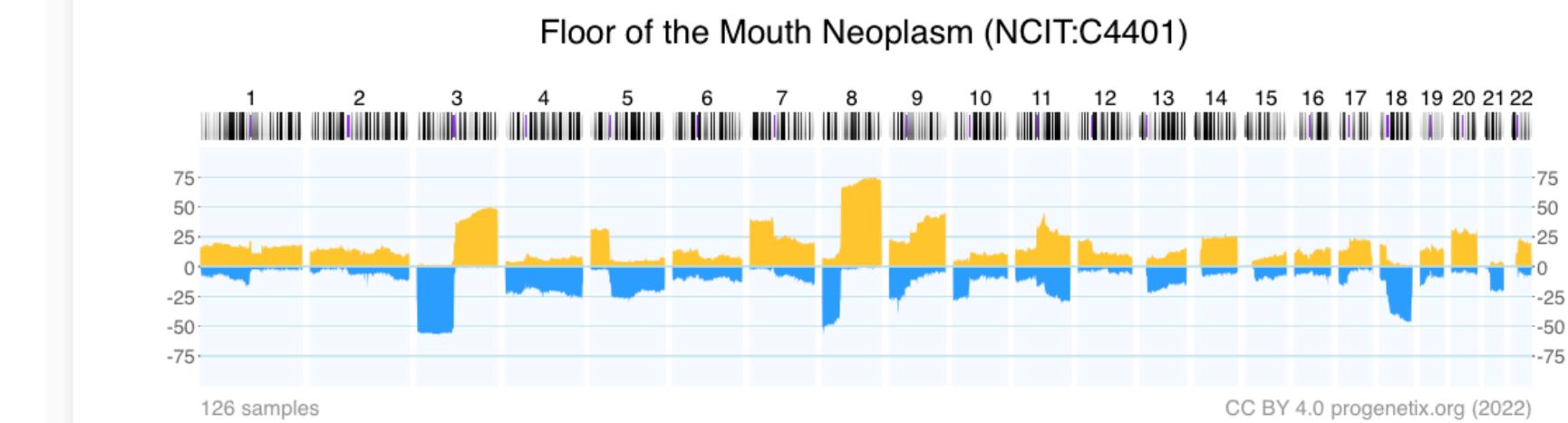
### Beacon<sup>+</sup>

Documentation  
News  
Downloads & Use  
Cases  
Sevices & API

### Baudisgroup @ UZH

## Cancer genome data @ progenetix.org

The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* from currently **142063** samples.



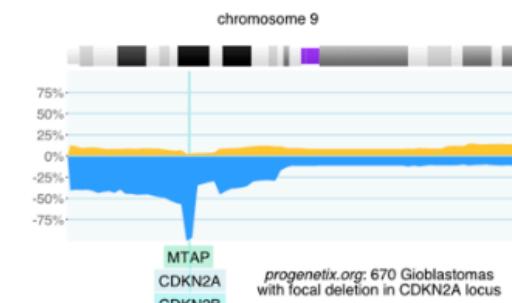
[Download SVG](#) | [Go to NCIT:C4401](#) | [Download CNV Frequencies](#)

Example for aggregated CNV data in 126 samples in Floor of the Mouth Neoplasm.  
Here the frequency of regional **copy number gains** and **losses** are displayed for all 22 autosomes.

### Progenetix Use Cases

#### Local CNV Frequencies

A typical use case on Progenetix is the search for local copy number aberrations - e.g. involving a gene - and the exploration of cancer types with these CNVs. The [\[ Search Page \]](#) provides example use cases for designing queries. Results contain basic statistics as well as visualization and download options.



#### Cancer CNV Profiles

The progenetix resource contains data of **834** different cancer types (NCIt neoplasm classification), mapped to a variety of biological and technical categories. Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the [\[ Cancer Types \]](#) page with direct visualization and options for sample retrieval and plotting options.

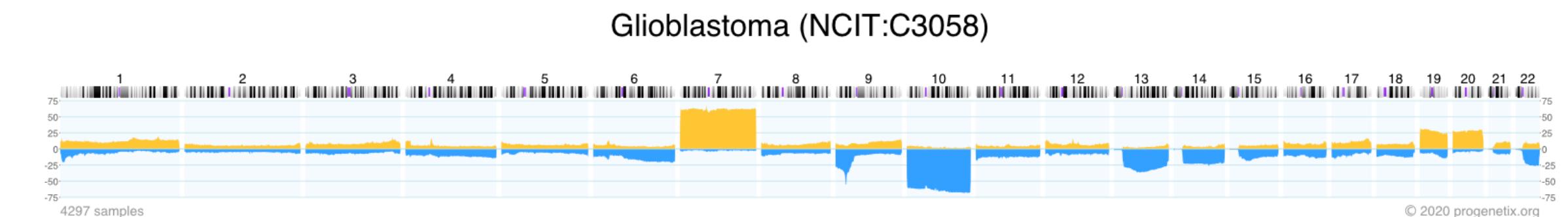
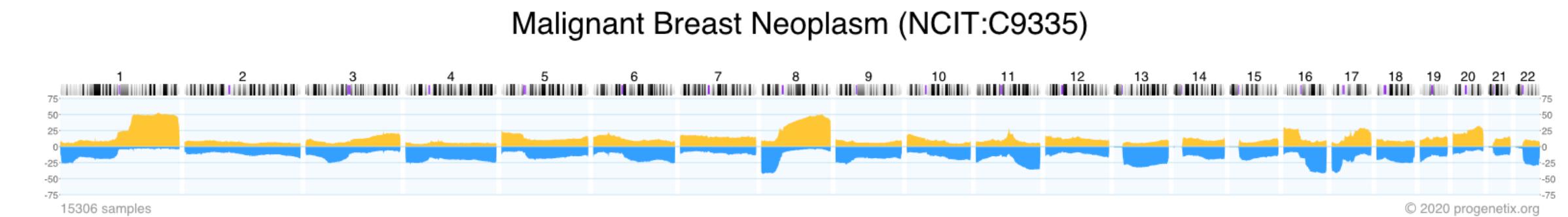
#### Cancer Genomics Publications

Through the [\[ Publications \]](#) page Progenetix provides **4164** annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.

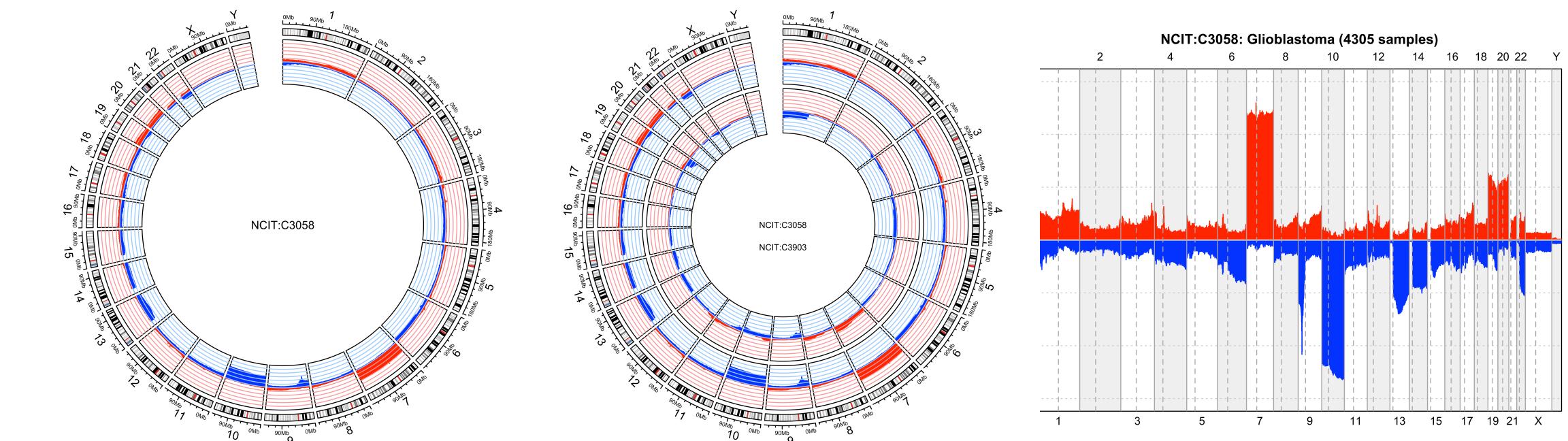
## Cancer Genomics Reference Resource

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

### Regional CNV Frequencies for >800 Cancer Types

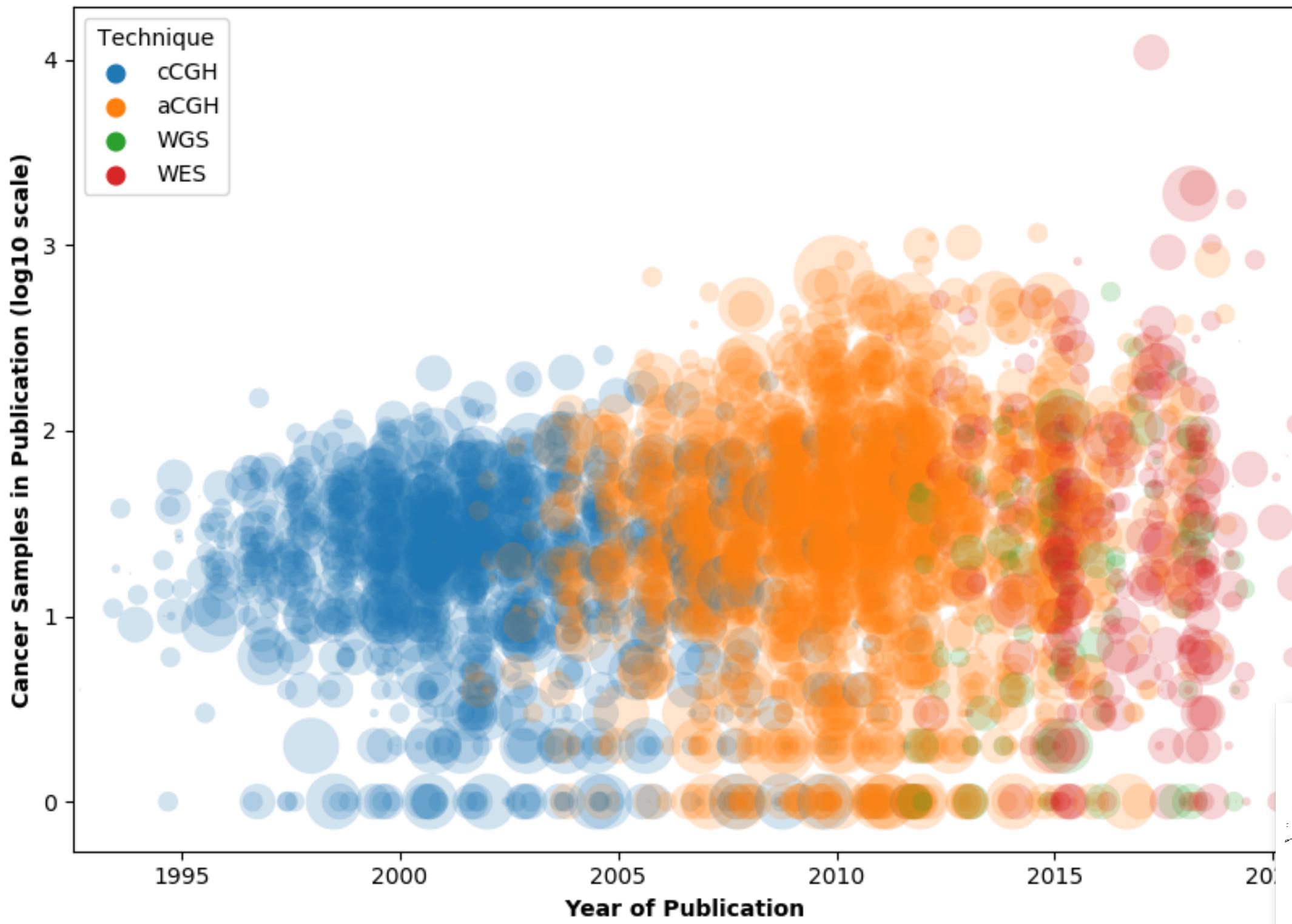


### Progenetix R API using Beacon handover objects



Visualization of CNV features using the pgxRpi R package. Aggregated CNV data for cancer types displayed using Circos or frequency plots in a local R environment. The R package relies on the Beacon v2 API to communicate with Progenetix.

## Number of tumor samples for each publication across the years



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

Database, 2020, 1–9  
doi: 10.1093/database/baa009  
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

progenetix

Cancer CNV Profiles

Search Samples

Studies & Cohorts

arrayMap

TCGA Samples

DIPG Samples

Gao & Baudis, 2021

Cancer Cell Lines

Publication DB

Services

NCIt Mappings

## Progenetix Publication Collection

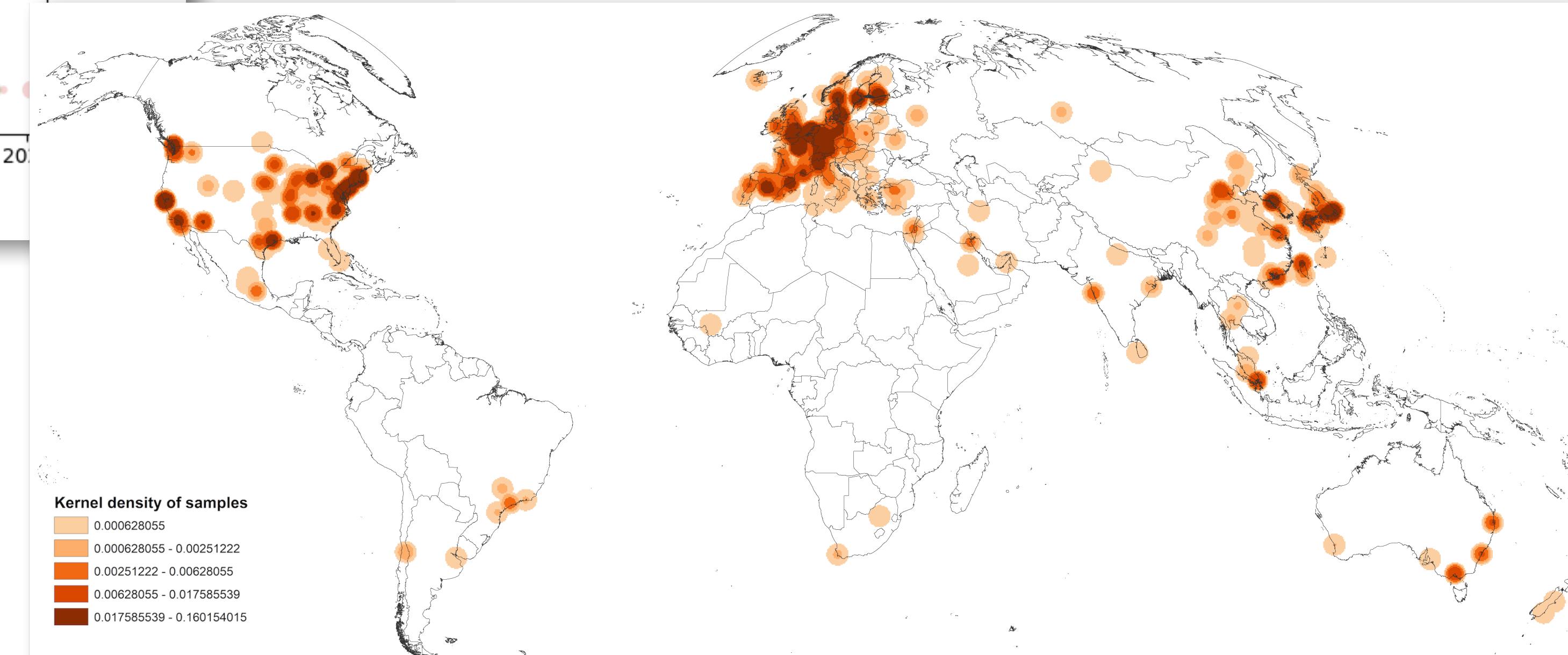
The current page lists articles describing whole genome screening (WGS, WES, aCGH, cCGH) experiments in cancer, registered in the Progenetix publication collection. For each publication the table indicates the numbers of samples analysed with a given technology and if sample profiles are available in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

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

Filter [i](#)

City [i](#)

id <a href="#">i</a> ▾	Publication	Samples				
		cCGH	aCGH	WES	WGS	pgx
PMID:34103027	Peng G, Chai H, Ji W, Lu Y, Wu S et al. (2021) Correlating genomic copy number alterations	0	79	0	0	0



Map of the geographic distribution (by 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 in Progenetix.

# The Progenetix oncogenomic resource in 2021

Qingyao Huang<sup>1,2</sup>, Paula Carrio-Cordo<sup>1,2</sup>, Bo Gao<sup>1,2</sup>, Rahel Paloots<sup>1,2</sup> and Michael Baudis<sup>1,2,\*</sup>

<sup>1</sup>Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland

<sup>2</sup>Swiss Institute of Bioinformatics, Winterthurerstrasse 190, Zurich 8057, Switzerland

\*Corresponding author: Tel: +41 44 635 34 86; Email: [michael.baudis@mls.uzh.ch](mailto:michael.baudis@mls.uzh.ch)

Citation details: Huang, Q., Carrio-Cordo, P., Gao, B. et al. The Progenetix oncogenomic resource in 2021. *Database* (2021) Vol. 2021: article ID baab043; DOI: <https://doi.org/10.1093/database/baab043>

## Abstract

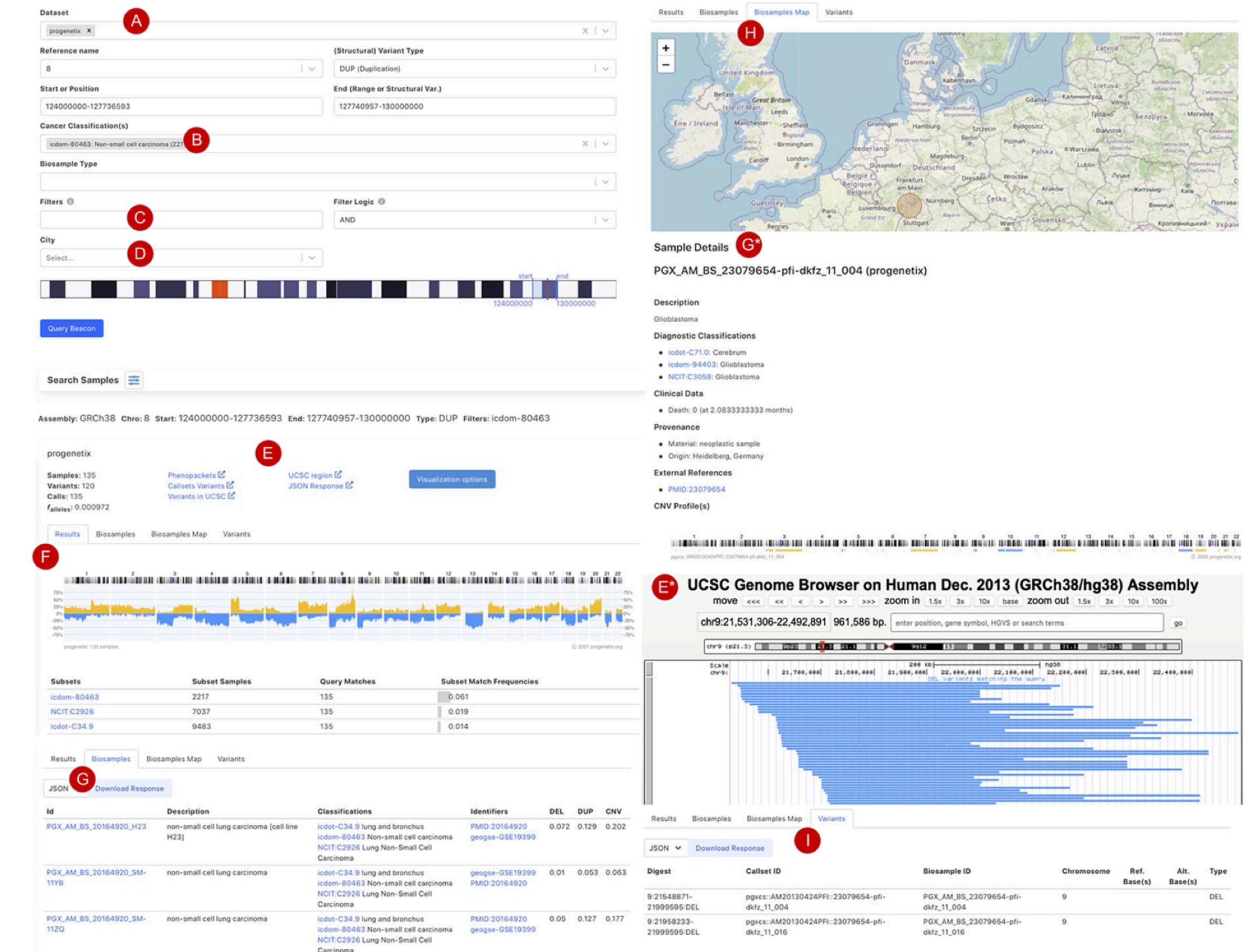
In cancer, copy number aberrations (CNAs) represent a type of nearly ubiquitous and frequently extensive structural genome variations. To disentangle the molecular mechanisms underlying tumorigenesis as well as identify and characterize molecular subtypes, the comparative and meta-analysis of large genomic variant collections can be of immense importance. Over the last decades, cancer genomic profiling projects have resulted in a large amount of somatic genome variation profiles, however segregated in a multitude of individual studies and datasets. The Progenetix project, initiated in 2001, curates individual cancer CNA profiles and associated metadata from published oncogenomic studies and data repositories with the aim to empower integrative analyses spanning all different cancer biologies. During the last few years, the fields of genomics and cancer research have seen significant advancement in terms of molecular genetics technology, disease concepts, data standard harmonization as well as data availability, in an increasingly structured and systematic manner. For the Progenetix resource, continuous data integration, curation and maintenance have resulted in the most comprehensive representation of cancer genome CNA profiling data with 138 663 (including 115 357 tumor) copy number variation (CNV) profiles. In this article, we report a 4.5-fold increase in sample number since 2013, improvements in data quality, ontology representation with a CNV landscape summary over 51 distinctive National Cancer Institute Thesaurus cancer terms as well as updates in database schemas, and data access including new web front-end and programmatic data access.

Database URL: [progenetix.org](http://progenetix.org)

**Table 1.** Statistics of samples from various data resources

Data source	GEO	ArrayExpress	cBioPortal	TCGA	Total
No. of studies	898	51	38	33	1939
No. of samples	63 568	4351	19 712	22 142	138 663
Tumor	52 090	3887	19 712	11 090	115 357
Normal	11 478	464	0	11 052	23 306
Classifications					
ICD-O (Topography)	100	54	88	157	209
ICD-O (Morphology)	246	908	265	140	491
NCIt	346	148	422	182	788
Collections					
Individuals	63 568	4351	19 712	10 995	127 549
Biosamples	63 568	4351	19 712	22 142	138 663
Callsets <sup>a</sup>	63 568	4351	19 712	22 376	138 930
Variants	5 514 126	118 4170	1 778 096	2 654 065	10 716 093

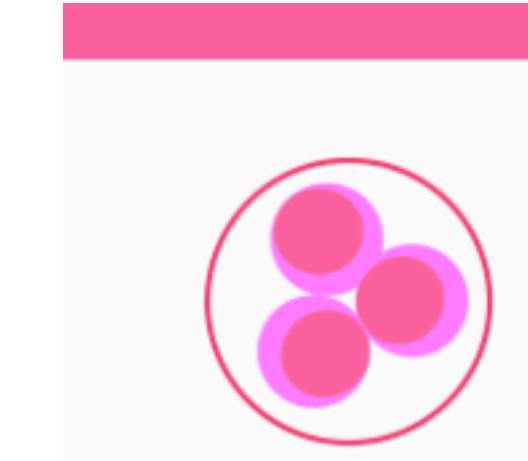
<sup>a</sup>set of variants from one genotyping experiment; ICD-O, International Classification of Diseases for Oncology; NCIt, National Cancer Institute Thesaurus.



**Figure 3.** Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range. This example queries for a continuous, focal duplication covering the complete MYC gene's coding region with  $\leq 6$  Mb in size. A: Filter for dataset; B: filter for cancer classification (NCIt and ICD-O-3 ontology terms available); C: additional filter, e.g. Cellosaurus; D: additional filter for geographic location; E: external link to UCSC browser to view the alignment of matched variants; F: cancer type classification sorted by frequency of the matched biosamples present in the subset; G: list of matched biosamples with description, statistics and reference. More detailed biosample information can be viewed through 'id' link to the sample detail page; H: matched variants with reference to biosamples can be downloaded in json or csv format.

## Cancer Cell Line Genomics Resource

- cancertcellines.org built on Progenetix platform
- includes over **5600 cell line CNV profiles**
- cancer cell line variants (SNV, INDELs ...) for **16178 cell lines** from 400 different disease classifications
- mapped to to *Cellosaurus*
- hierarchical representation ("derived from" ...)
- SNVs mapped from ClinVar with variant severity and disease ontologies
- CCLE per cell line include variant effect
- CNV profiles allow temporal stability estimates and tumor type similarity matching



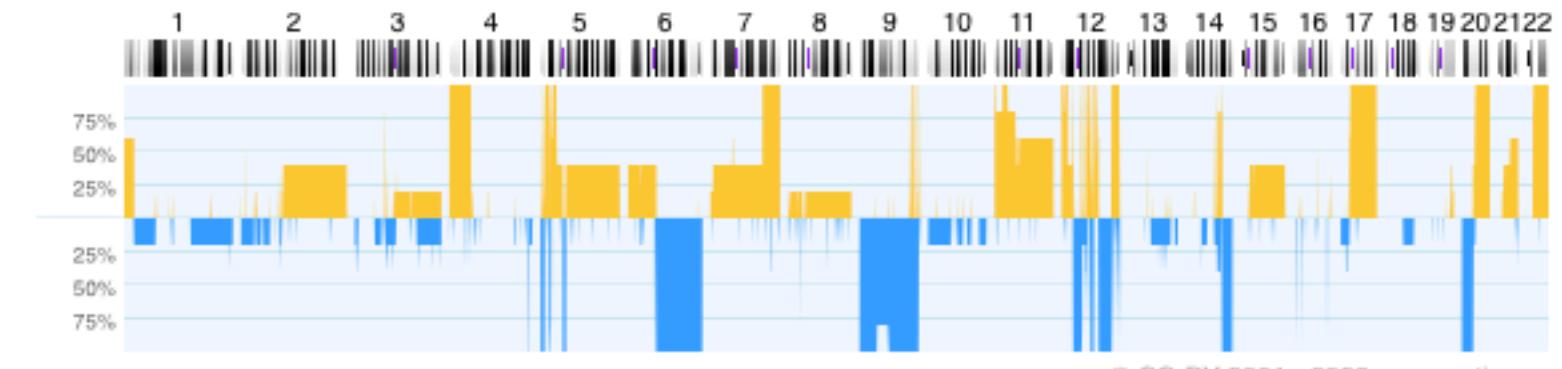
The logo for cancertcellines.org features a circular icon with three pink circles of varying sizes, resembling chromosomes or cells. Below the icon, the word "cancercelllines" is written in a lowercase, sans-serif font.

[Cancer Cell Lines<sup>o</sup>](#)  
[Cell Line Listing](#)  
[Search Cell Lines](#)  
[CNV Profiles by Cancer Type](#)  
[NCIT Codes](#)  
[ICD-O 3 Morphologies](#)  
[Documentation](#)  
[Progenetix](#)  
[Progenetix Data](#)  
[Progenetix Documentation](#)  
[Baudisgroup @ UZH](#)

### Cancer Cell Line Genomics

The cancertcellines.org genomic information resource contains genome profiling data, somatic mutation information and associated metadata for thousands of human cancer cell lines. It has its origins in genomic copy number variation (CNV) profiling data of cell lines originally collected as part of the more than 100'000 individual datasets in the Progenetix<sup>o</sup> oncogenomic resource. However, by providing genome mapped, annotated data for many types of genomic mutations, together with CNV profiles for a subset of the overall more than 16'000 cell lines, cancertcellines.org provides a unique entry point for the comparative analysis of genomic variants in cell lines as well as for the exploration of related publications.

SK-MEL-1 (cellosaurus:CVCL\_0068)



© CC-BY 2001 - 2023 progenetix.org

**Cell Line Data CNV Frequency Plot** The CNV histogram above represents CNV data from a randomly selected set of samples – either instances of a common cell line or with a shared diagnosis. In this example the frequencies of regional gains and losses in 5 samples from cellosaurus:CVCL\_0068 (SK-MEL-1) are on display.

[Download SVG](#) | [Go to cellosaurus:CVCL\\_0068](#) | [Download CNV Frequencies](#)

In cancertcellines.org genomic variation data collected from a variety of external resources and from original data (re-) analyses has been mapped to GRCh38 genome coordinates and is queryable using the Beacon v2 API<sup>o</sup>. The resource contains data of **16340** individual cancer cell lines from **382** different cancer types (NCIt neoplasm classification).

A large amount of the cancer cell line data has been collected based on annotations and pointers from [Cellosaurus](#), a reference knowledge resource on cell lines.

#### Citation

- cancertcellines.org: **Cancer cell line oncogenomic online resource** (2023)
- Huang Q, Carrio-Cordo P, Gao B, Paloots R, Baudis M. (2021): **The Progenetix oncogenomic resource in 2021**. *Database (Oxford)*. 2021 Jul 17

# Maintaining some Standards

## CNV Term Use in Computational (File/Schema) Formats



EFO	Beacon	VCF	SO	GA4GH VRS ⇒ VRS proposal <sup>1</sup>	Notes
EFO:0030070 copy number gain	DUP <sup>2</sup> or EFO:0030070	DUP	SO:0001742 copy_number_gain	low-level gain (implicit) ⇒ EFO:0030070 copy number gain	a sequence alteration whereby the copy number of a given genomic region is greater than the reference sequence
EFO:0030071 low-level copy number gain	DUP <sup>2</sup> or EFO:0030071	DUP	SO:0001742 copy_number_gain	low-level gain ⇒ EFO:0030071 low-level copy number gain	
EFO:0030072 high-level copy number gain	DUP <sup>2</sup> or EFO:0030072	DUP	SO:0001742 copy_number_gain	high-level gain ⇒ EFO:0030072 high-level copy number gain	commonly but not consistently used for >=5 copies on a bi-allelic genome region
EFO:0030073 focal genome amplification	DUP <sup>2</sup> or EFO:0030073	DUP	SO:0001742 copy_number_gain	high-level gain ⇒ EFO:0030073 focal genome amplification	commonly but not consistently used for >=5 copies on a bi-allelic genome region, of limited size (operationally max. 1-5Mb)
EFO:0030067 copy number loss	DEL <sup>2</sup> or EFO:0030067	DEL	SO:0001743 copy_number_loss	partial loss (implicit) ⇒ EFO:0030067 copy number loss	a sequence alteration whereby the copy number of a given genomic region is smaller than the reference sequence
EFO:0030068 low-level copy number loss	DEL <sup>2</sup> or EFO:0030068	DEL	SO:0001743 copy_number_loss	partial loss ⇒ EFO:0030068 low-level copy number loss	
EFO:0020073 high-level copy number loss	DEL <sup>2</sup> or EFO:0020073	DEL	SO:0001743 copy_number_loss	partial loss ⇒ EFO:0020073 high-level copy number loss	a loss of several copies; also used in cases where a complete genomic deletion cannot be asserted
EFO:0030069 complete genomic deletion	DEL <sup>2</sup> or EFO:0030069	DEL	SO:0001743 copy_number_loss	complete loss ⇒ EFO:0030069 complete genomic deletion	complete genomic deletion (e.g. homozygous deletion on a bi-allelic genome region)

Hangjia Zhao  
Michael Baudis  
(& the VRS group!)

<https://cnvar.org/resources/CNV-annotation-standards/>

# 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**



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

# **GA4GH Standards for Federated Genomic Data Discovery**



## INFORMATICS

### Beacon v2 and Beacon networks: A “lingua franca” for federated data discovery in biomedical genomics

#### 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>

Jordi Rambla<sup>1,2</sup> | Michael Baudis<sup>3</sup> | Roberto Ariosa<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>4</sup>

# Cell Genomics

## 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>

# Cell Genomics

## Perspective

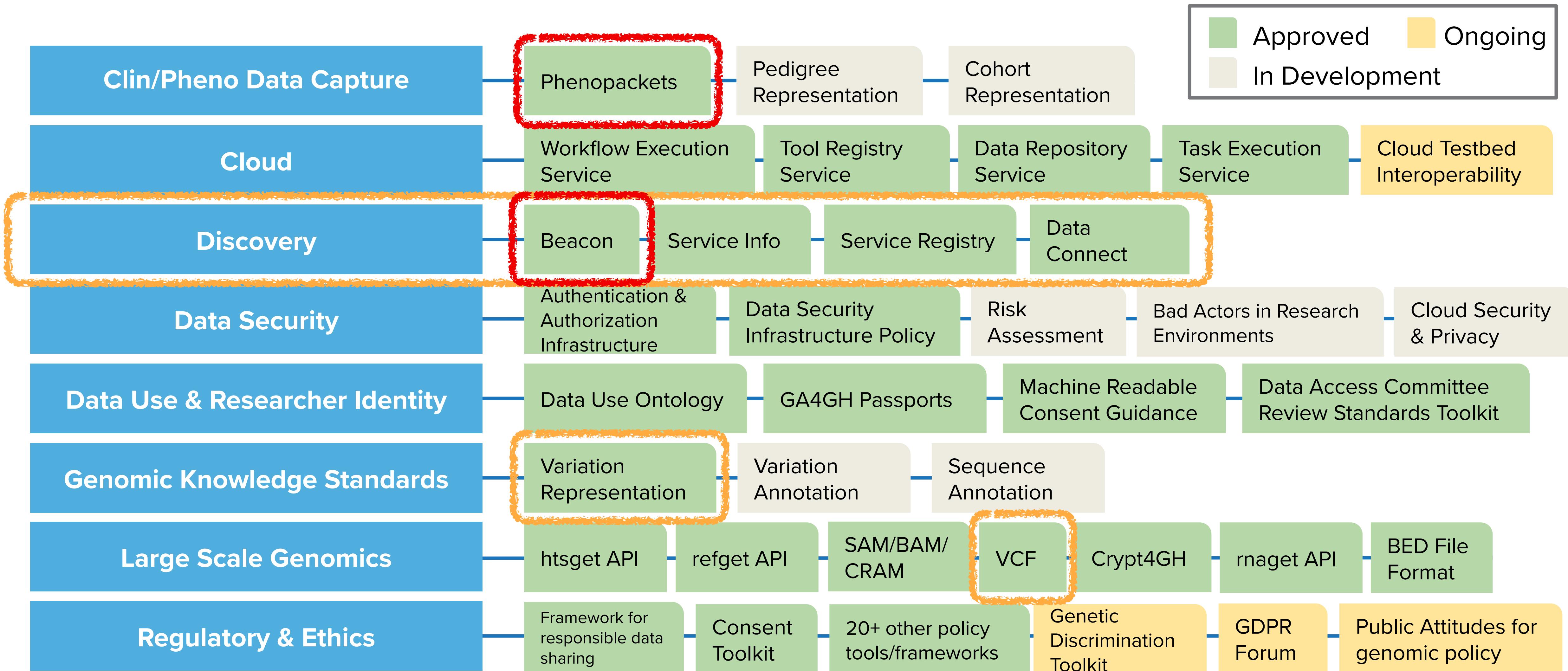
### GA4GH: International policies and standards for data sharing across genomic research and healthcare

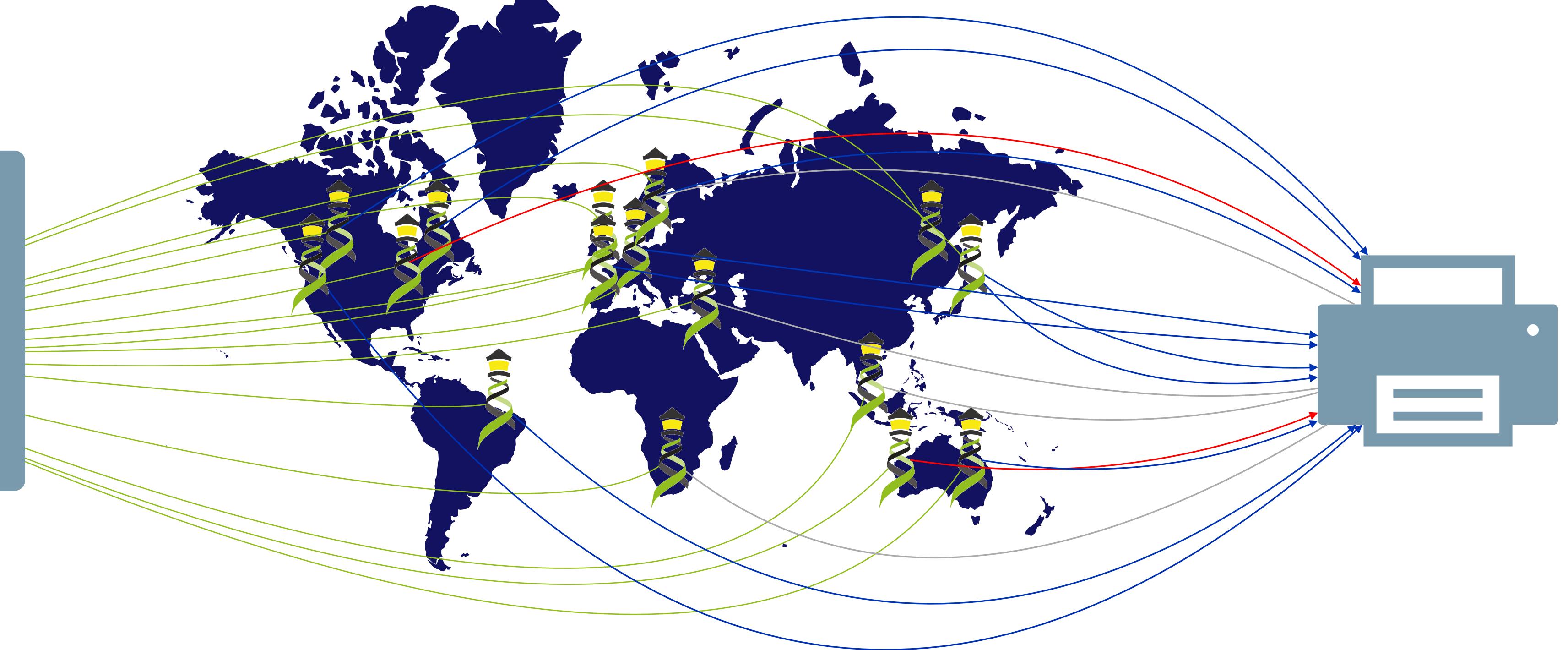
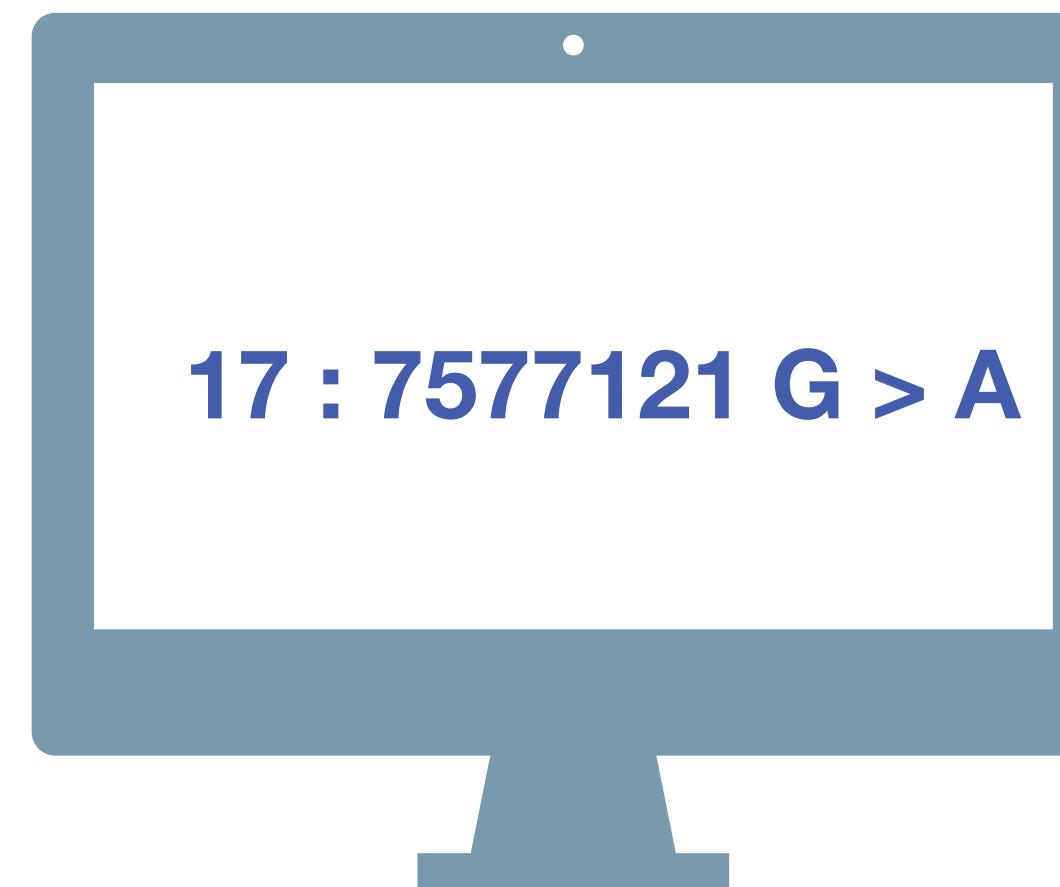
Heidi L. Rehm,<sup>1,2,47</sup> Angela J.H. Page,<sup>1,3,\*</sup> Lindsay Smith,<sup>3,4</sup> Jeremy B. Adams,<sup>3,4</sup> Gil Alterovitz,<sup>5,47</sup> Lawrence J. Babb,<sup>1</sup> Maxmillian P. Barkley,<sup>6</sup> Michael Baudis,<sup>7,8</sup> Michael J.S. Beauvais,<sup>3,9</sup> Tim Beck,<sup>10</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>12,3</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 Courtoot,<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)

### The GA4GH Phenopacket schema defines a computable representation of clinical data

# Overview of GA4GH standards and frameworks





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 v1 Development

2014      GA4GH founding event; Jim Ostell proposes Beacon concept with "more features... version 2"

2015      • beacon-network.org aggregator created by DNAstack

• Beacon v0.3 release

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

• OpenAPI implementation

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

• Beacon v0.4 release in January; feature release for GA4GH approval process

• GA4GH Beacon v1 approved at Oct plenary

## Beacon v2 Development

## Related ...

- ELIXIR starts Beacon project support

- GA4GH re-structuring (workstreams...)

- Beacon part of Discovery WS

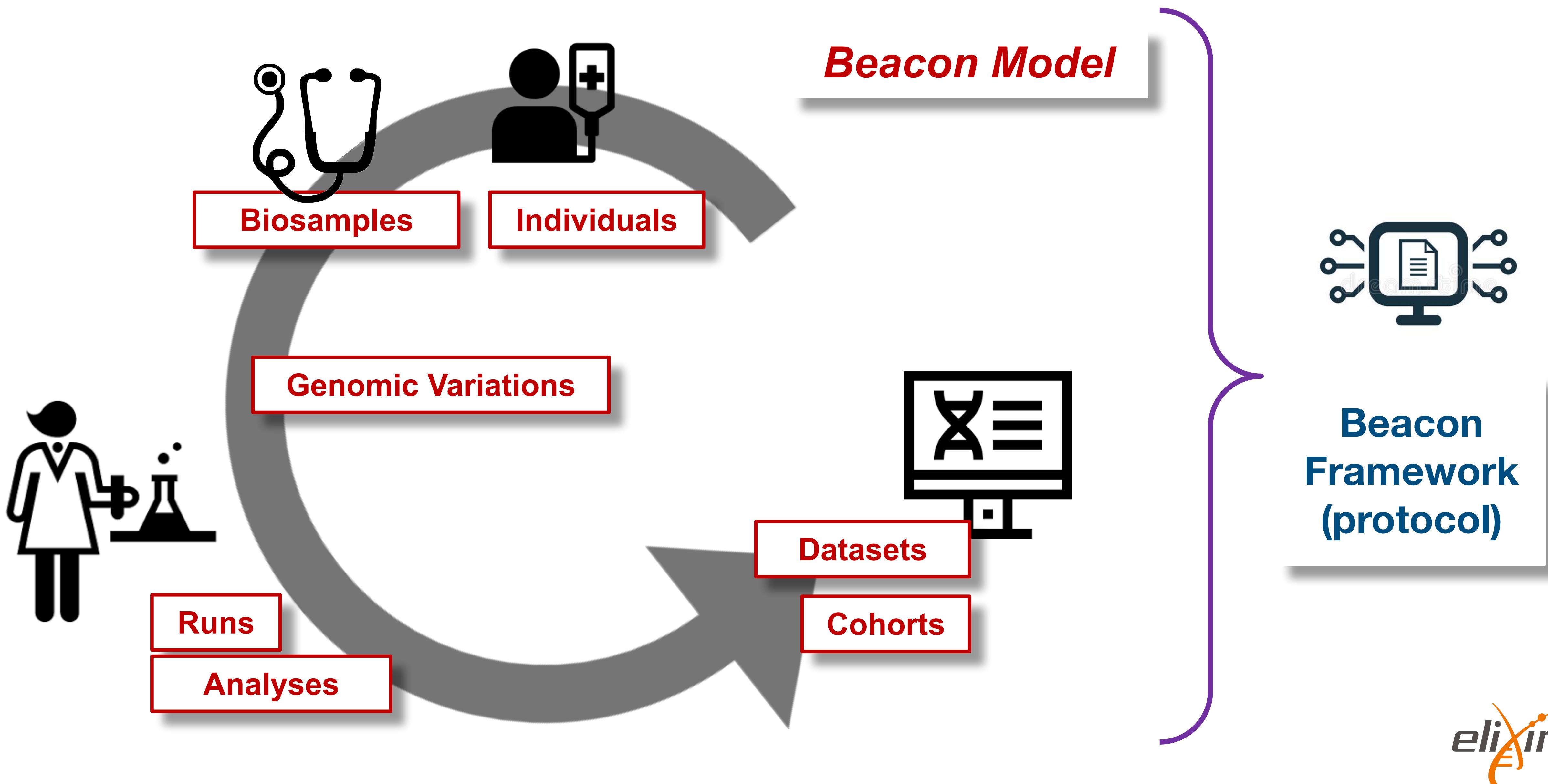
- new Beacon website (March)

- Beacon publication at Nature Biotechnology

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

# Beacon v2

docs.genomebeacons.org



# Progenetix & Beacon

## Implementation driven standards development

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

Beacon v2 GA4GH Approval Registry

Beacons:    

 European Genome-Phenome Archive (EGA)

GA4GH Approval Beacon Test

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

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

 Theoretical Cytogenetics and Oncogenomics group at UZH and SIB

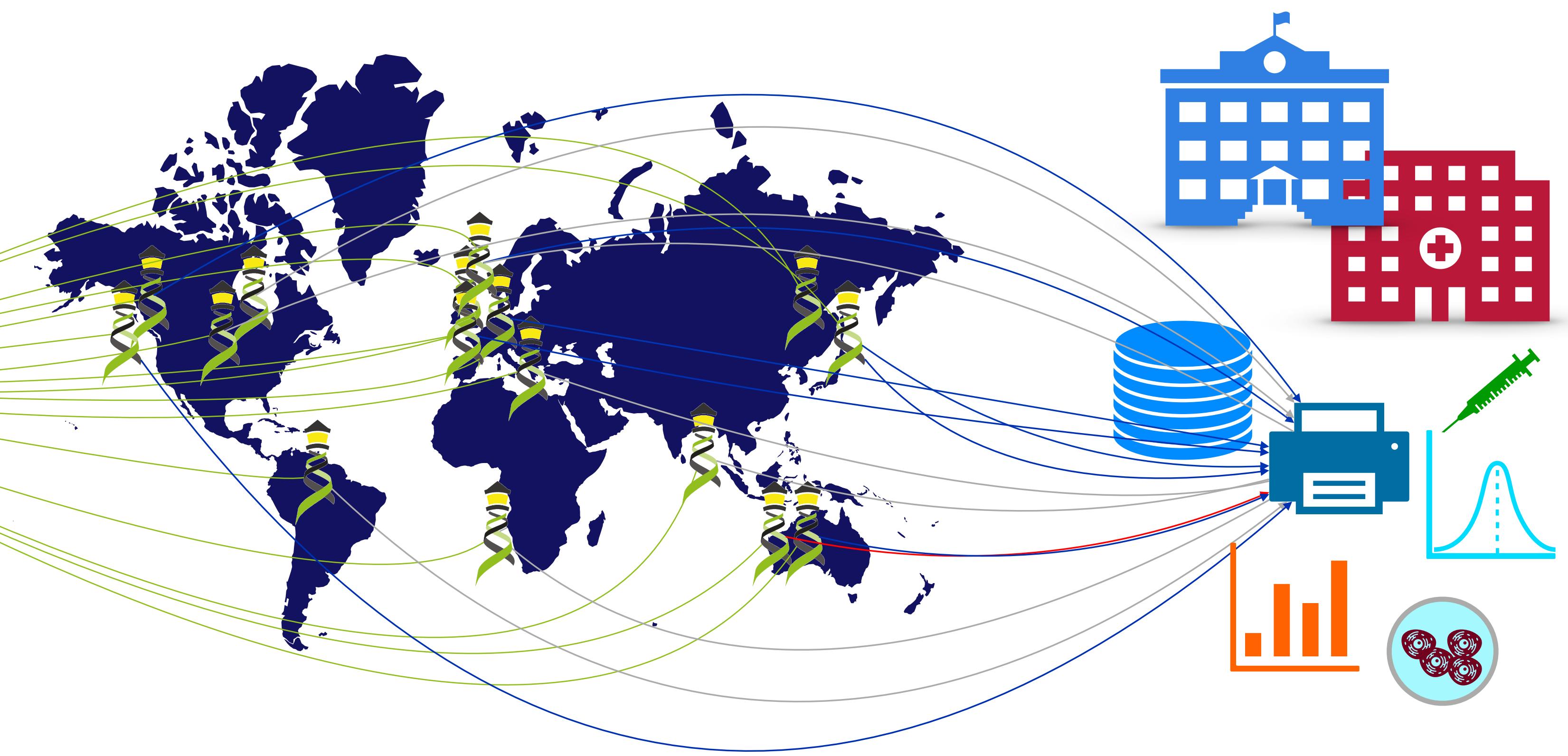
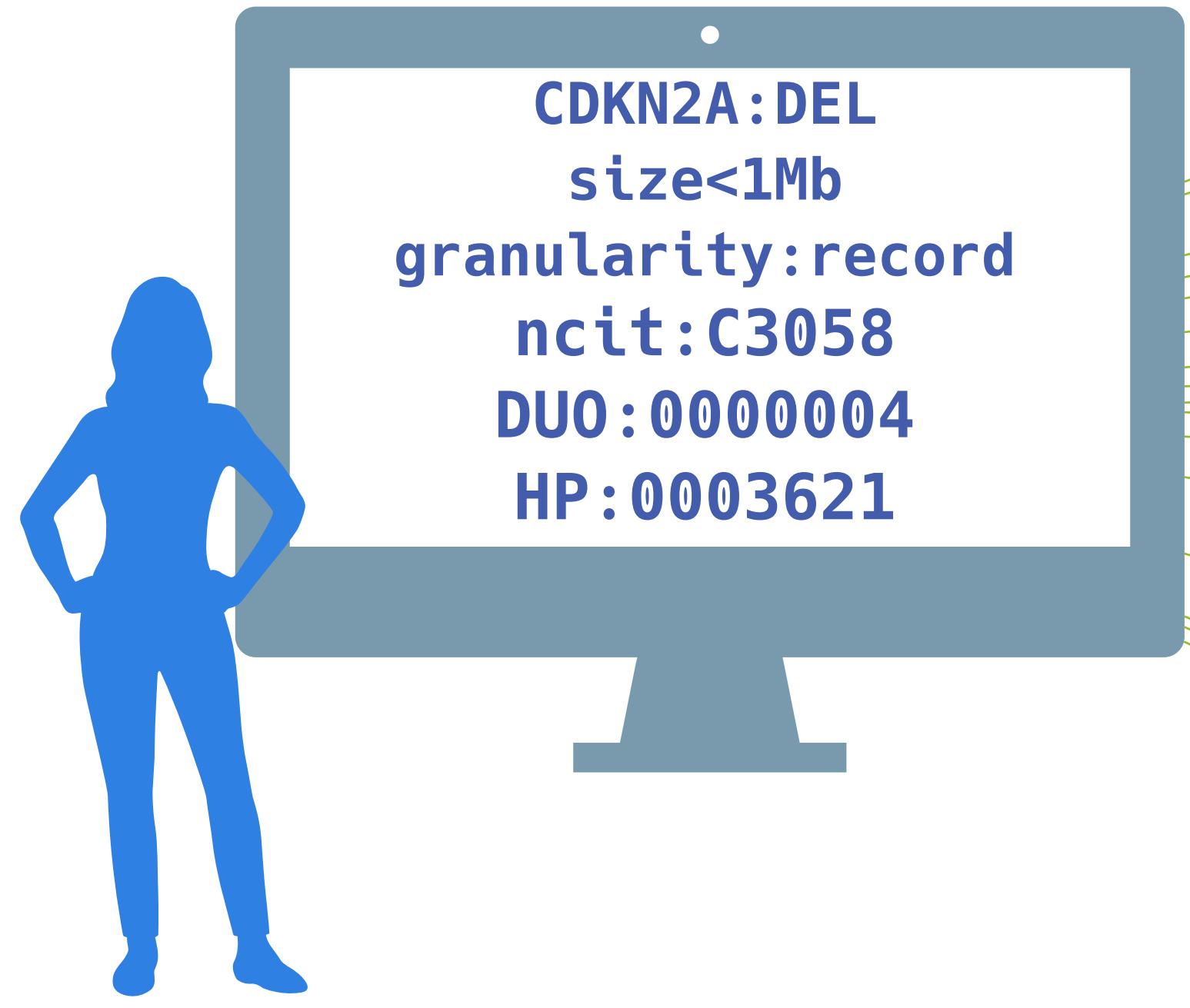
Progenetix Cancer Genomics Beacon+ provides a forward looking implementation of the Beacon v2 API, with focus on structural genome variants and metadata based on the...

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

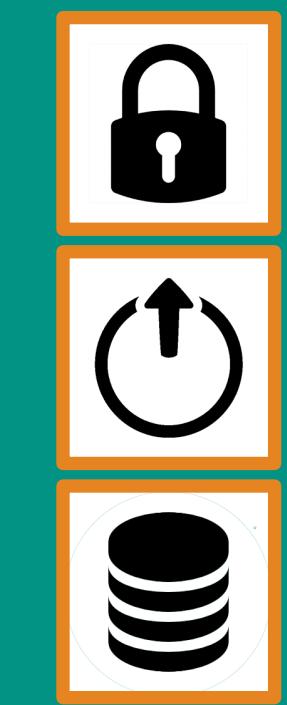
Beacon protocol response verifier at time of GA4GH approval Spring 2022

 Centre Nacional Analisis Genomica (CNAG-CRG)	University of Leicester
Beacon @ RD-Connect	Cafe Variome Beacon v2
This <a href="#">Beacon</a> is based on the GA4GH Beacon <a href="#">v2.0</a>	This <a href="#">Beacon</a> is based on the GA4GH Beacon <a href="#">v2.0</a>
BeaconMap	██████████
Bioinformatics analysis	██████████
Biological Sample	██████████
Cohort	██████████
Configuration	██████████
Dataset	██████████
EntryTypes	██████████
Genomic Variants	██████████
Individual	██████████
Info	██████████
Sequencing run	██████████

✓ Matches the Spec   ✗ Not Match the Spec   ● Not Implemented



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

# The GA4GH Phenopackets v2 Standard

## A Computable Representation of Clinical Data

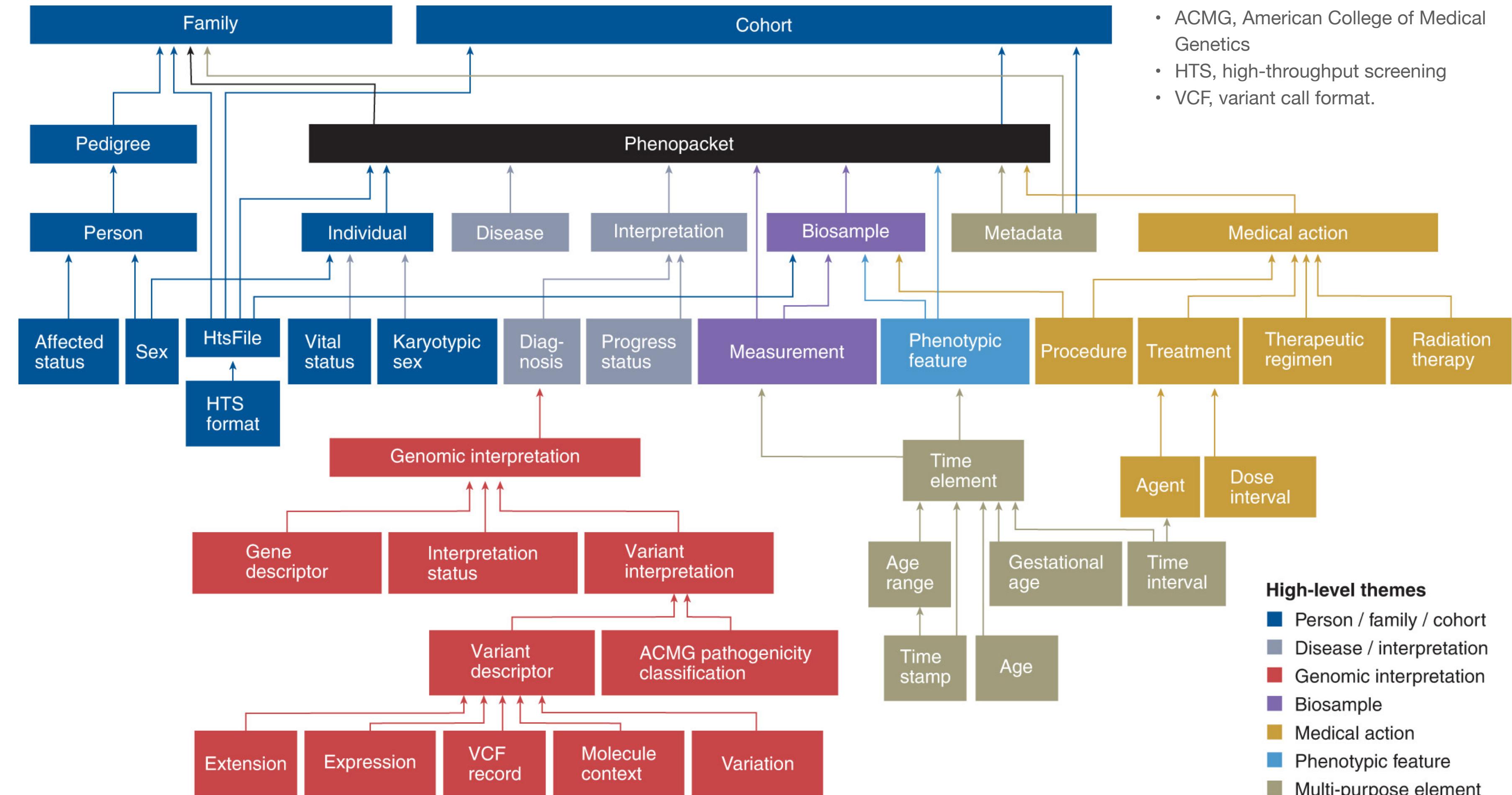


The GA4GH Phenopacket schema consists of several optional elements, each containing information about a certain topic, such as phenotype, variant or pedigree. An element can contain other elements, which allows a hierarchical representation of data.

For instance, Phenopacket contains elements of type *Individual*, *PhenotypicFeature*, *Biosample* and so on. Individual elements can therefore be regarded as **building blocks** of larger structures.

Jacobsen JOB, Baudis M, Baynam GS, Beckmann JS, Beltran S, Buske OJ, Callahan TJ, et al. 2022.

**The GA4GH Phenopacket Schema Defines a Computable Representation of Clinical Data.**  
*Nature Biotechnology* 40 (6): 817–20.



# The GA4GH Phenopackets v2 Standard

## A Computable Representation of Clinical Data



The GA4GH Phenopacket schema

consists of a set of elements, such as pedigree, other elements, and hierarchical structures. For instance, it contains elements like Individual, Biosample, and Cohort, which are regarded as the building blocks of larger structures.

ISO 4454:2022 Genomics informatics — Phenopackets: A format for phenotypic data exchange

**Abstract**

This document specifies a uniform, machine-readable, phenotypic description of an individual, patient or sample in the context of rare disease, common/complex disease or cancer.

It is applicable to academic, clinical and commercial research, as well as clinical diagnostics. While intended for human data collection, it can be used in other areas (e.g. mouse research). It does not define the phenotypic information that needs to be collected for a particular use but represents that information in an appropriately descriptive manner that allows it to be computationally exchanged between systems.

**General information**

Status: Published | Publication date: 2022-07 | Edition: 1 | Number of pages: 86 | Technical Committee: ISO/TC 215/SC 1 Genomics Informatics | Nature Biotechnology 40 (6): 817–20.

**Buy this standard**

Format: PDF + ePub | Language: English | Price: CHF 198 | Buy

High-level themes:

- Person / family / cohort
- Disease / interpretation
- Genomic interpretation
- Biosample
- Medical action
- Phenotypic feature
- Multi-purpose element

- ACMG, American College of Medical Genetics
- HTS, high-throughput screening
- VCF, variant call format.

Co-hosted by  
Smart Health  
Standards Forum, ISO  
TC215 SC1 Korean  
mirror committee

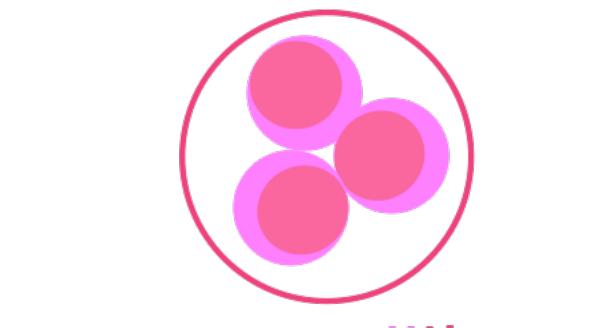
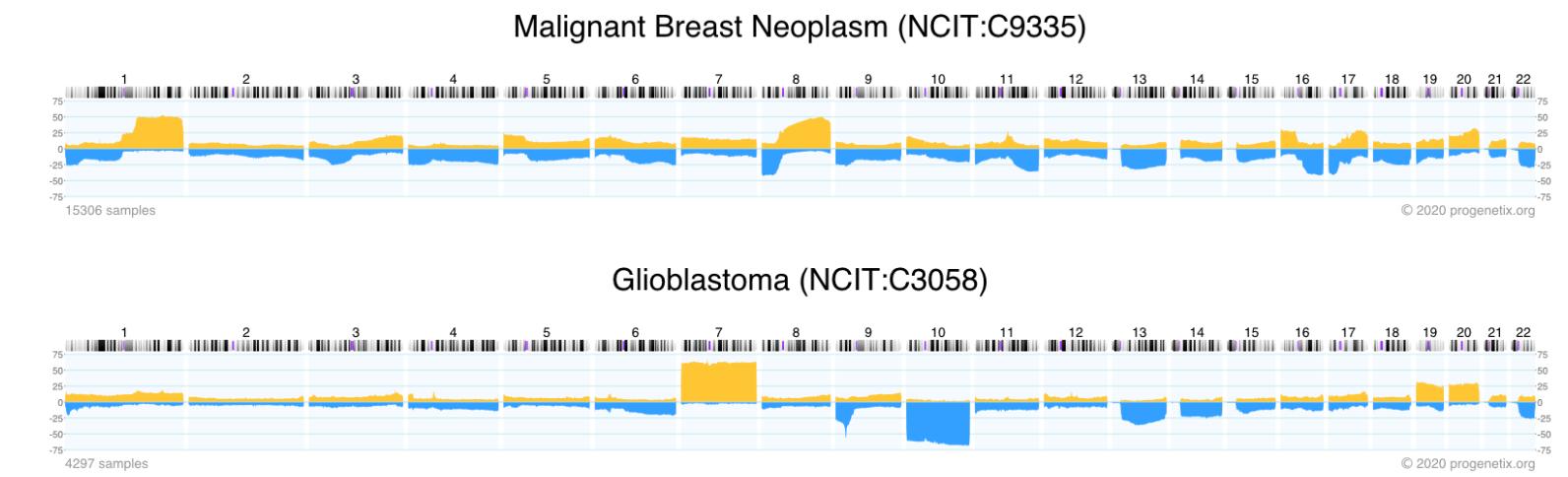
- Person / family / cohort
- Disease / interpretation
- Genomic interpretation
- Biosample
- Medical action
- Phenotypic feature
- Multi-purpose element

# Theoretical Cytogenetics and Oncogenomics @baudisgroup

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



Universität  
Zürich<sup>UZH</sup>

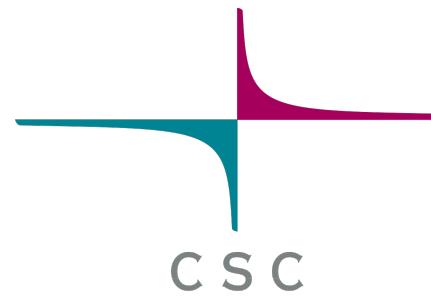


cancer cell lines





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



Juha Törnroos  
Teemu Kataja  
Ikkka Lappalainen  
Dylan Spalding



Tony Brookes  
Tim Beck  
Colin Veal  
Tom Shorter



Michael Baudis  
Rahel Paloots  
Hangjia Zhao  
Ziying Yang  
Bo Gao



Augusto Rendon  
Ignacio Medina  
Javier López  
Jacobo Coll  
Antonio Rueda



centre nacional d'anàlisi genòmica  
centro nacional de análisis genómico



David Salgado



Salvador Capella  
Dmitry Repchevski  
JM Fernández



Laura Furlong  
Janet Piñero



Sergi Beltran  
Carles Hernandez  
  
Serena Scollen  
Gary Saunders  
Giselle Kerry  
David Lloyd



Nicola Mulder  
Mamana  
Mbiyavanga  
Ziyaad Parker



David Torrents  
Dean Hartley



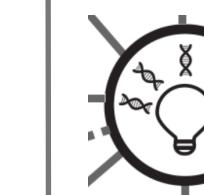
Heidi Rehm  
Ben Hutton



Toshiaki  
Katayama



Fundación Progreso y Salud  
CONSEJERÍA DE SALUD



ENAS



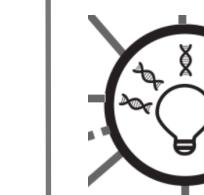
Diana Lemos



Joaquin Dopazo  
Javier Pérez  
J.L. Fernández  
Gema Roldan



Thomas Keane  
Melanie Courtot  
Jonathan Dursi



Heidi Rehm  
Ben Hutton



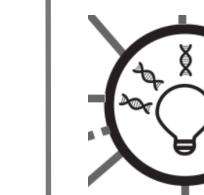
Toshiaki  
Katayama



Stephane Dyke  
DNA STACK  
Marc Fiume  
Miro Cupak



Thomas Keane  
Melanie Courtot  
Jonathan Dursi



Heidi Rehm  
Ben Hutton



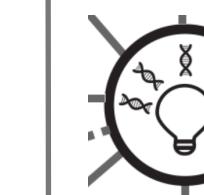
Toshiaki  
Katayama



GA4GH Phenopackets  
Peter Robinson  
Jules Jacobsen



Thomas Keane  
Melanie Courtot  
Jonathan Dursi



Heidi Rehm  
Ben Hutton



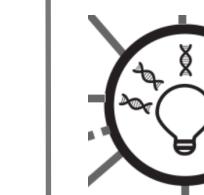
Toshiaki  
Katayama



GA4GH VRS  
Alex Wagner  
Reece Hart



Thomas Keane  
Melanie Courtot  
Jonathan Dursi



Heidi Rehm  
Ben Hutton



Toshiaki  
Katayama



GA4GH PRC  
Alex Wagner  
Jonathan Dursi  
Mamana Mbiyavanga  
Alice Mann  
Neerjah Skantharajah



# The Beacon team through the ages



Universität  
Zürich<sup>UZH</sup>



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



Swiss Institute of  
Bioinformatics



