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

Universität
Zürich^UZH

Genomic Data Sharing Standard Development with GA4GH and ELIXIR

Opportunities and Pitfalls in Federated Data Discovery

Michael Baudis @ DMLS Lecture Series 2024-02-27

SIB
Swiss Institute of
Bioinformatics

elixir

Theoretical Cytogenetics and Oncogenomics

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

Curators

Data ~~Parasites~~

Bioinformatics & Bioinformaticians are ...

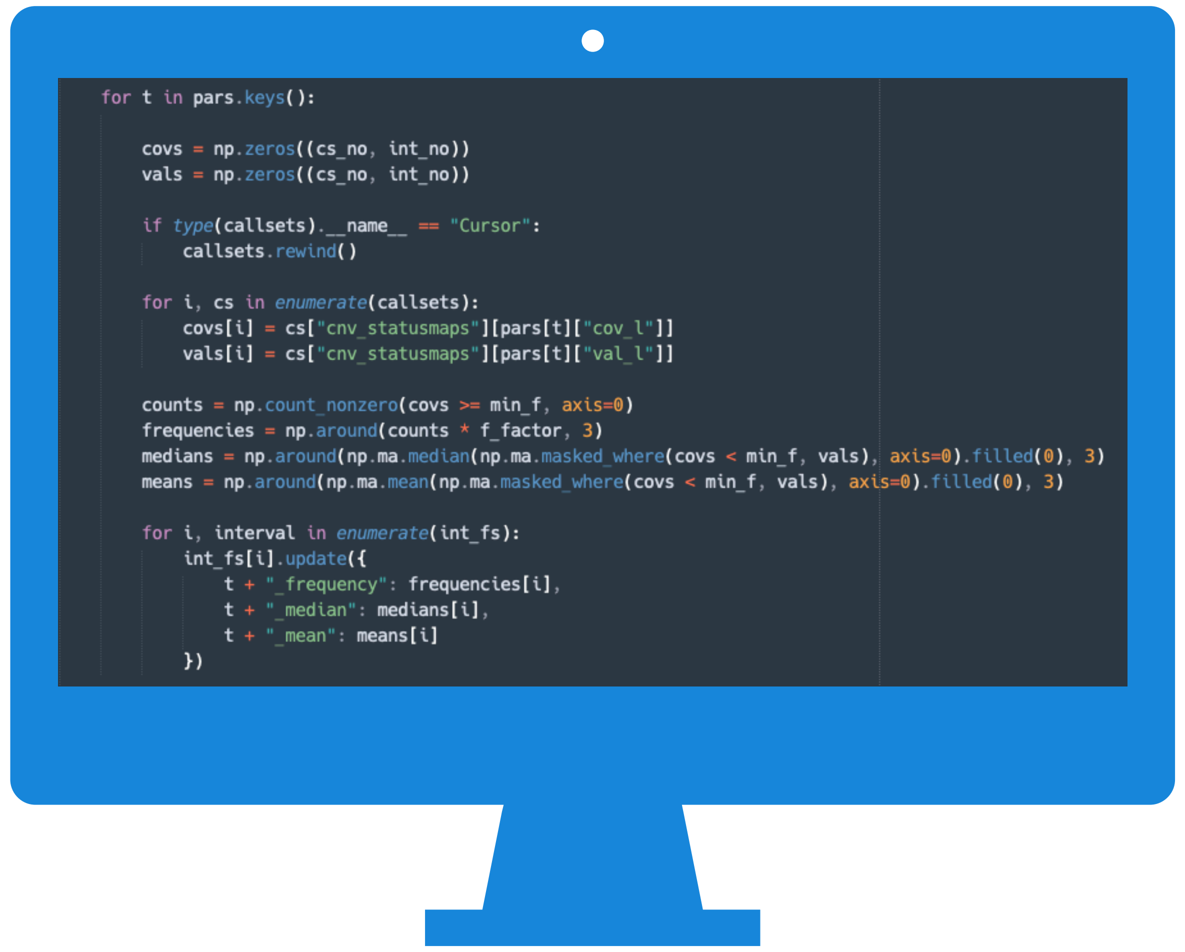
Bioinformatician	Bio**informatician**
strong biological knowledge	sufficient biological background
provides hypothesis and / or dataset	provides statistical, analysis methods
sufficient statistical and **computational** expertise to correctly use bioinformatics tools & develop workflows (scripting ...)	**sufficient biological** or **medical** background to understand problems presented and identify pitfalls and hidden biases arising from data generation
expert **user** of informatics tools	**developer** of informatics tools
may get a Nobel	may get rich

Bioinformatics & Bioinformaticians are ...

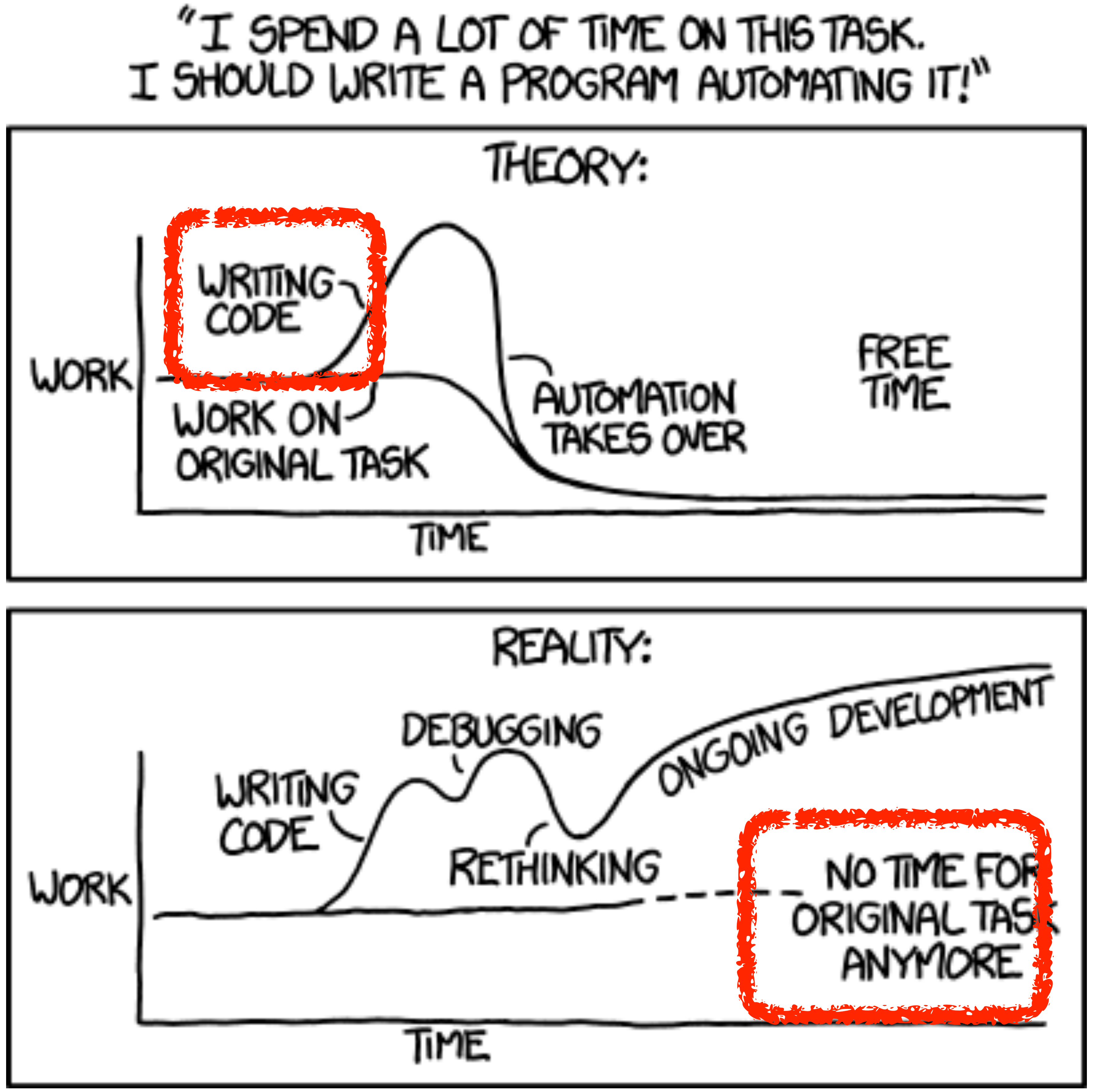
Bioinformatician	Bio**informatician**
strong biological knowledge	sufficient biological background
provides hypothesis and / or dataset	provides statistical, analysis methods
sufficient statistical and **computational** expertise to correctly use bioinformatics tools & develop workflows (scripting ...)	**sufficient biological** or **medical** background to understand problems presented and identify pitfalls and hidden biases arising from data generation
expert **user** of informatics tools	**developer** of informatics tools
may get a Nobel	may get rich

flux

{BioInformaticsScience}

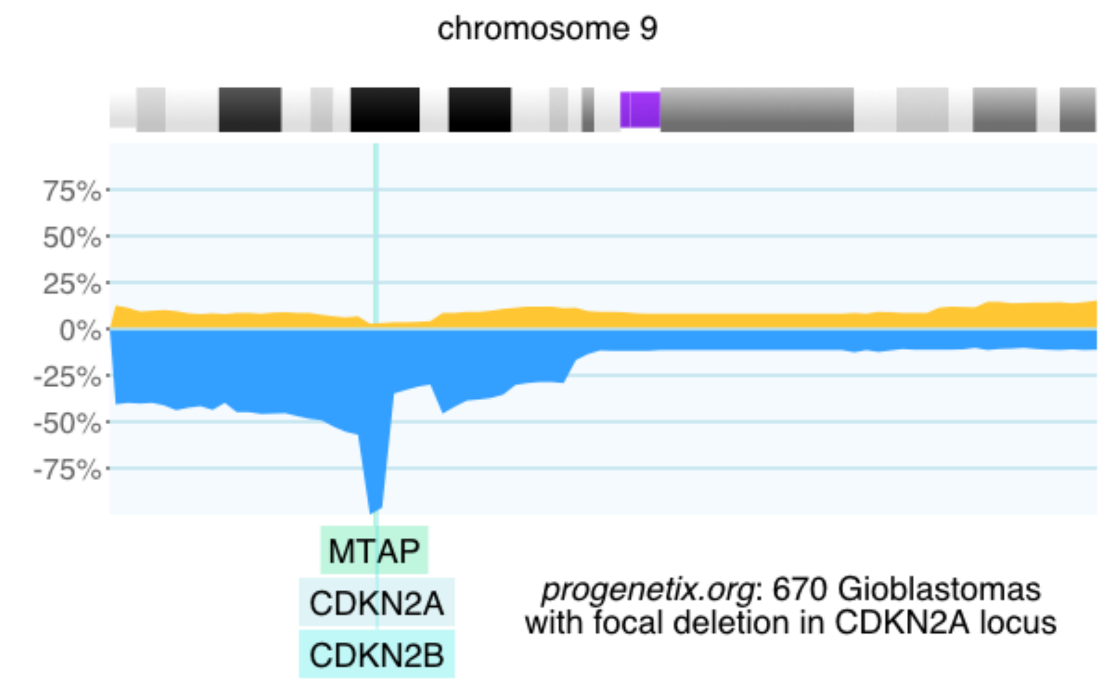


```
for t in pars.keys():

    covs = np.zeros((cs_no, int_no))
    vals = np.zeros((cs_no, int_no))

    if type(callsets).__name__ == "Cursor":
        callsets.rewind()

    for i, cs in enumerate(callsets):
        covs[i] = cs["cnv_statusmaps"][pars[t]["cov_l"]]
        vals[i] = cs["cnv_statusmaps"][pars[t]["val_l"]]

    counts = np.count_nonzero(covs >= min_f, axis=0)
    frequencies = np.around(counts * f_factor, 3)
    medians = np.around(np.ma.median(np.ma.masked_where(covs < min_f, vals), axis=0).filled(0), 3)
    means = np.around(np.ma.mean(np.ma.masked_where(covs < min_f, vals), axis=0).filled(0), 3)

    for i, interval in enumerate(int_fs):
        int_fs[i].update({
            t + "_frequency": frequencies[i],
            t + "_median": medians[i],
            t + "_mean": means[i]
        })
```



Randall Munroe - XKCD
https://xkcd.com/1319/

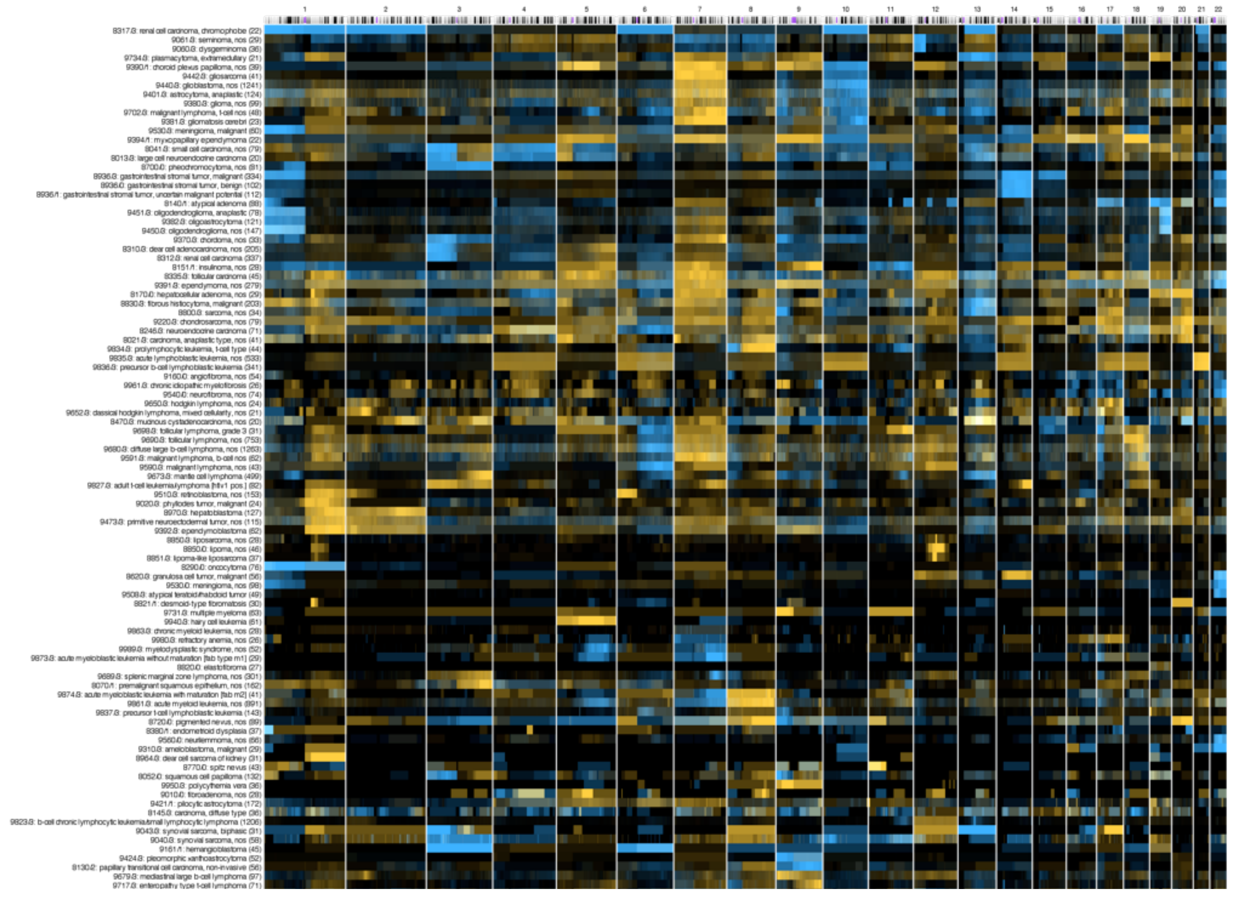
Theoretical Cytogenetics and Oncogenomics Research | Methods | Standards

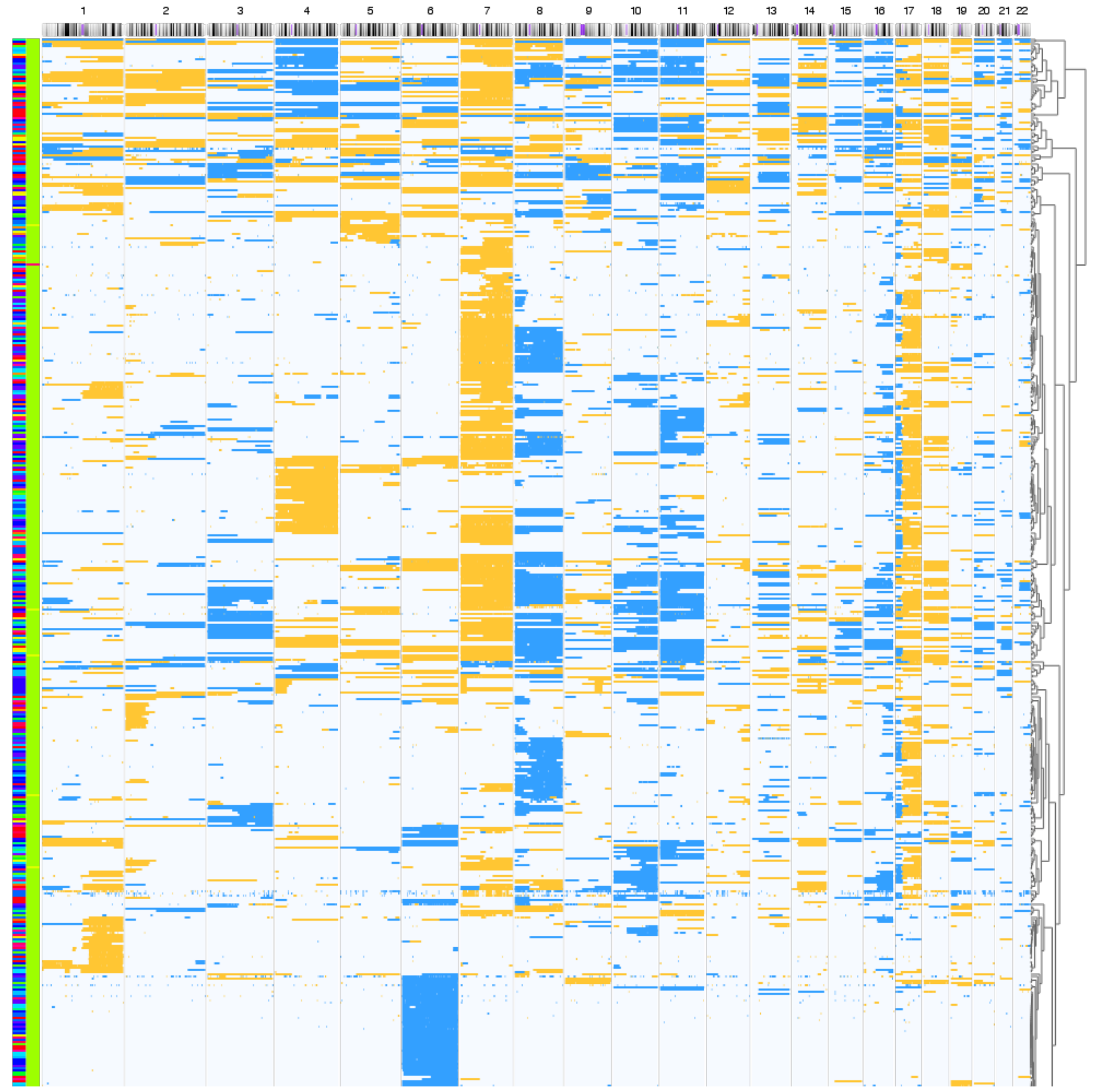


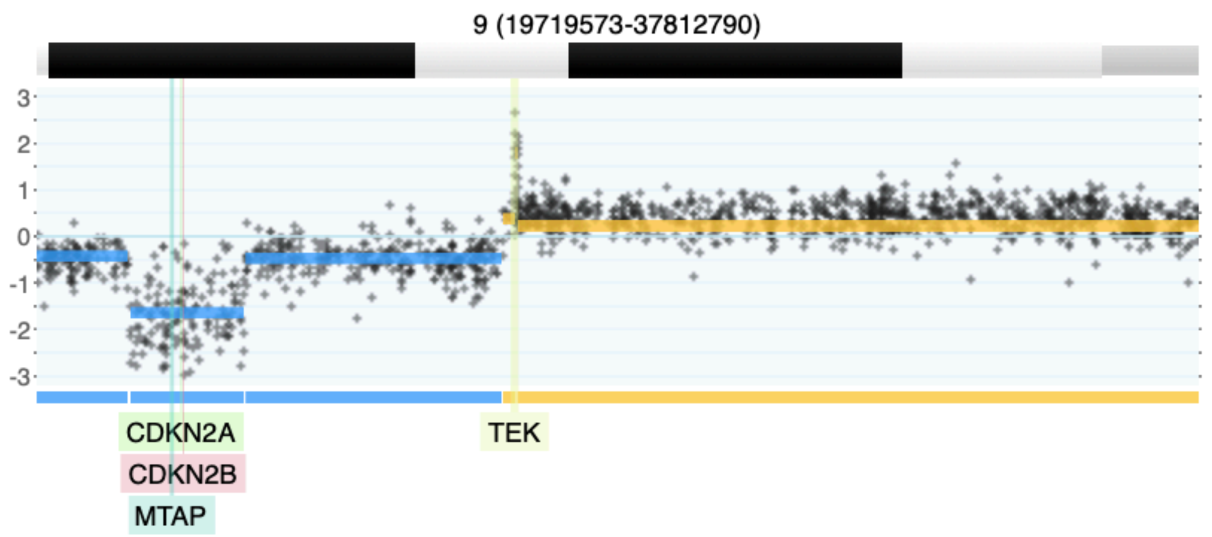
progenetix.org: 670 Gioblastomas with focal deletion in CDKN2A locus

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)



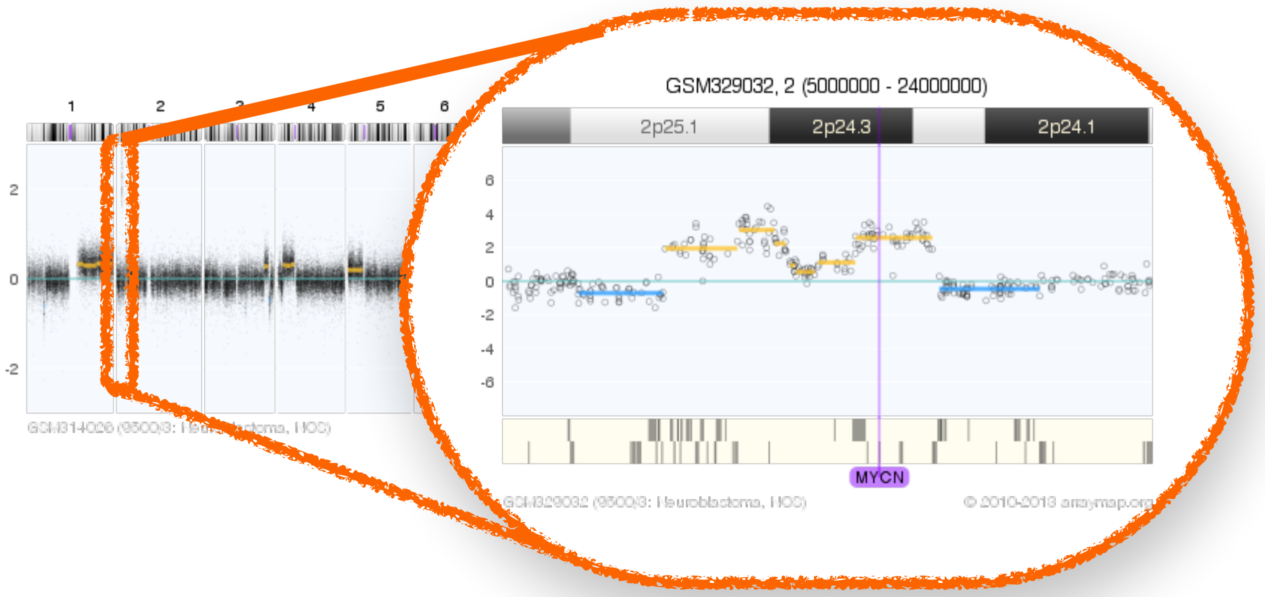




pgxcs::GSE19612::GSM491153

© 2019 progenetix.org

2-event, homozygous deletion in a Glioblastoma



MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

progenetix.org

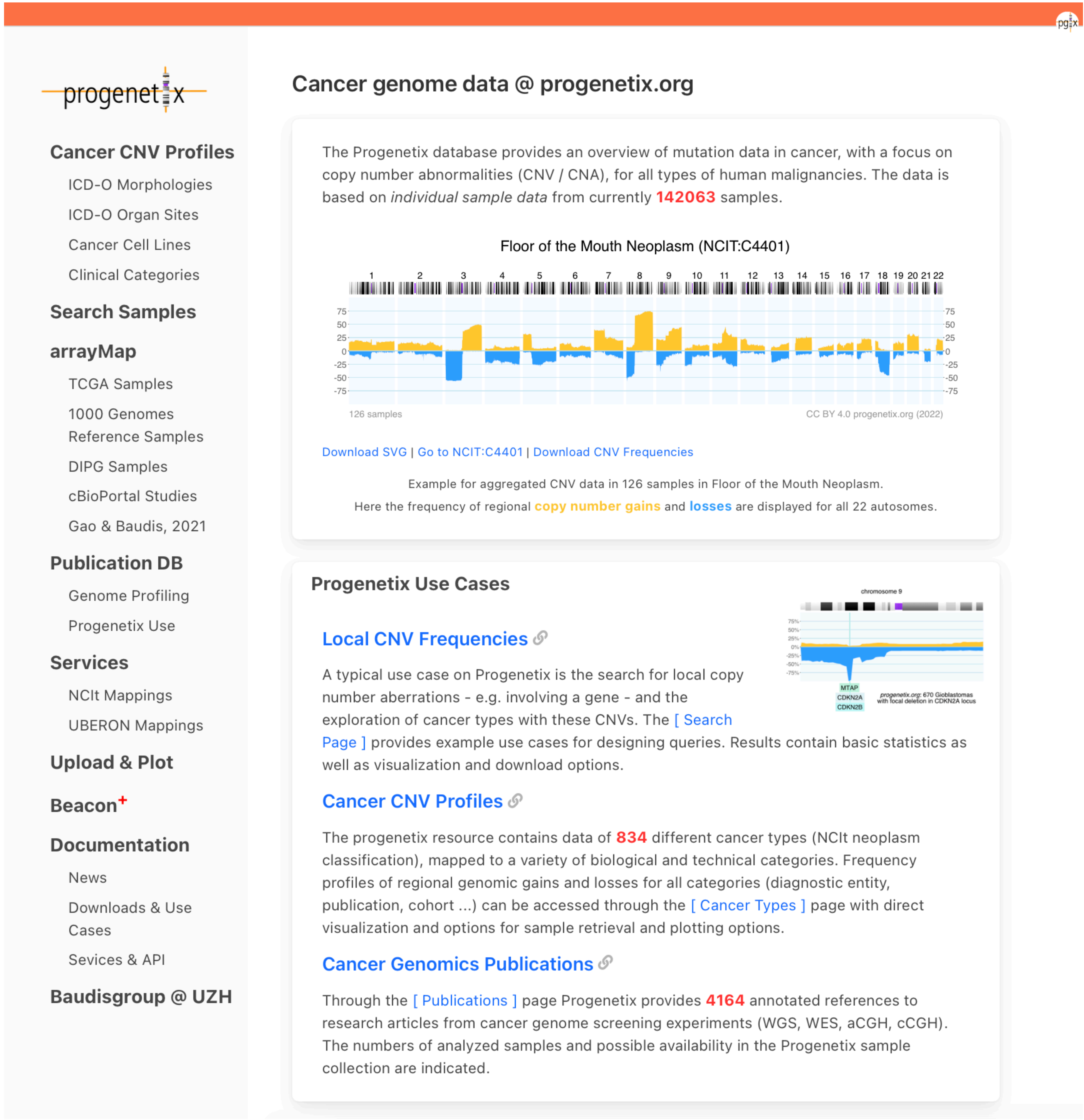
Cancer Genomics Reference Resource

- ***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

Universität Zürich^UZH

progenetix

SIB Swiss Institute of Bioinformatics



progenetix

Cancer CNV Profiles

- ICD-O Morphologies
- ICD-O Organ Sites
- Cancer Cell Lines
- Clinical Categories

Search Samples

arrayMap

- TCGA Samples
- 1000 Genomes Reference Samples
- DIPG Samples
- cBioPortal Studies
- Gao & Baudis, 2021

Publication DB

- Genome Profiling
- Progenetix Use

Services

- NCIt Mappings
- UBERON Mappings

Upload & Plot

Beacon+

Documentation

- News
- Downloads & Use Cases
- Sevices & API

Baudisgroup @ UZH

Cancer genome data @ progenetix.org

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

Floor of the Mouth Neoplasm (NCIT:C4401)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
75
50
25
0
-25
-50
-75
126 samples
CC BY 4.0 progenetix.org (2022)

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

chromosome 9
75%
50%
25%
0%
-25%
-50%
-75%
MTAP
CDKN2A
CDKN2B
progenetix.org: 670 Gioblastomas with focal deletion in CDKN2A locus

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.

progenetix.org

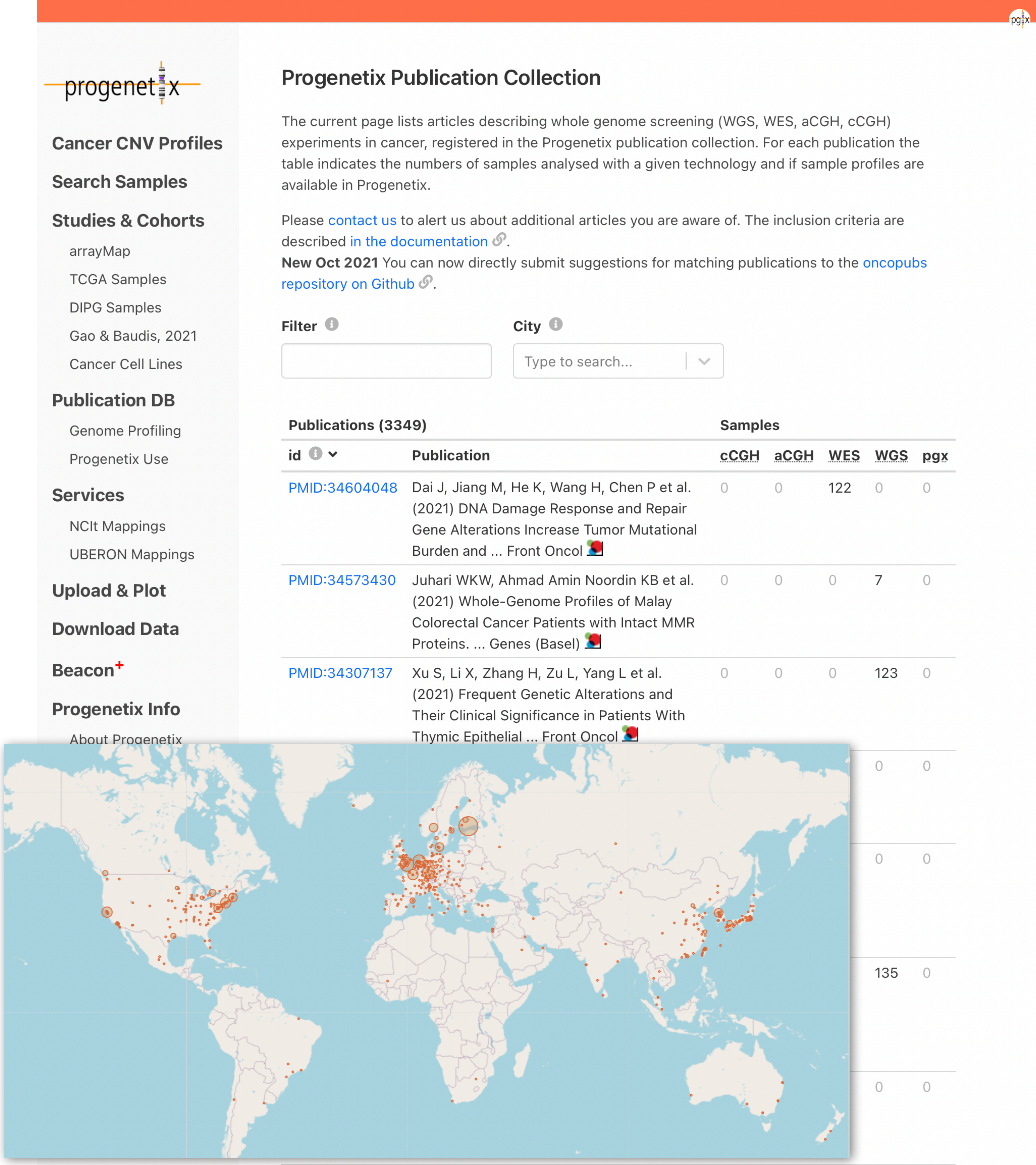
Cancer Genomics Reference Resource

- ***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

Universität Zürich UZH

progenetix

SIB Swiss Institute of Bioinformatics



Progenetix Publication Collection

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

Please contact us to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation.
New Oct 2021 You can now directly submit suggestions for matching publications to the oncopubs repository on Github.

Filter | City: Type to search...

Publications (3349)		Samples				
id	Publication	cCGH	aCGH	WES	WGS	pgx
PMID:34604048	Dai J, Jiang M, He K, Wang H, Chen P et al. (2021) DNA Damage Response and Repair Gene Alterations Increase Tumor Mutational Burden and ... Front Oncol	0	0	122	0	0
PMID:34573430	Juhari WKW, Ahmad Amin Noordin KB et al. (2021) Whole-Genome Profiles of Malay Colorectal Cancer Patients with Intact MMR Proteins. ... Genes (Basel)	0	0	0	7	0
PMID:34307137	Xu S, Li X, Zhang H, Zu L, Yang L et al. (2021) Frequent Genetic Alterations and Their Clinical Significance in Patients With Thymic Epithelial ... Front Oncol	0	0	0	123	0
					0	0
					0	0
					135	0
					0	0

Cancer Cell Lines

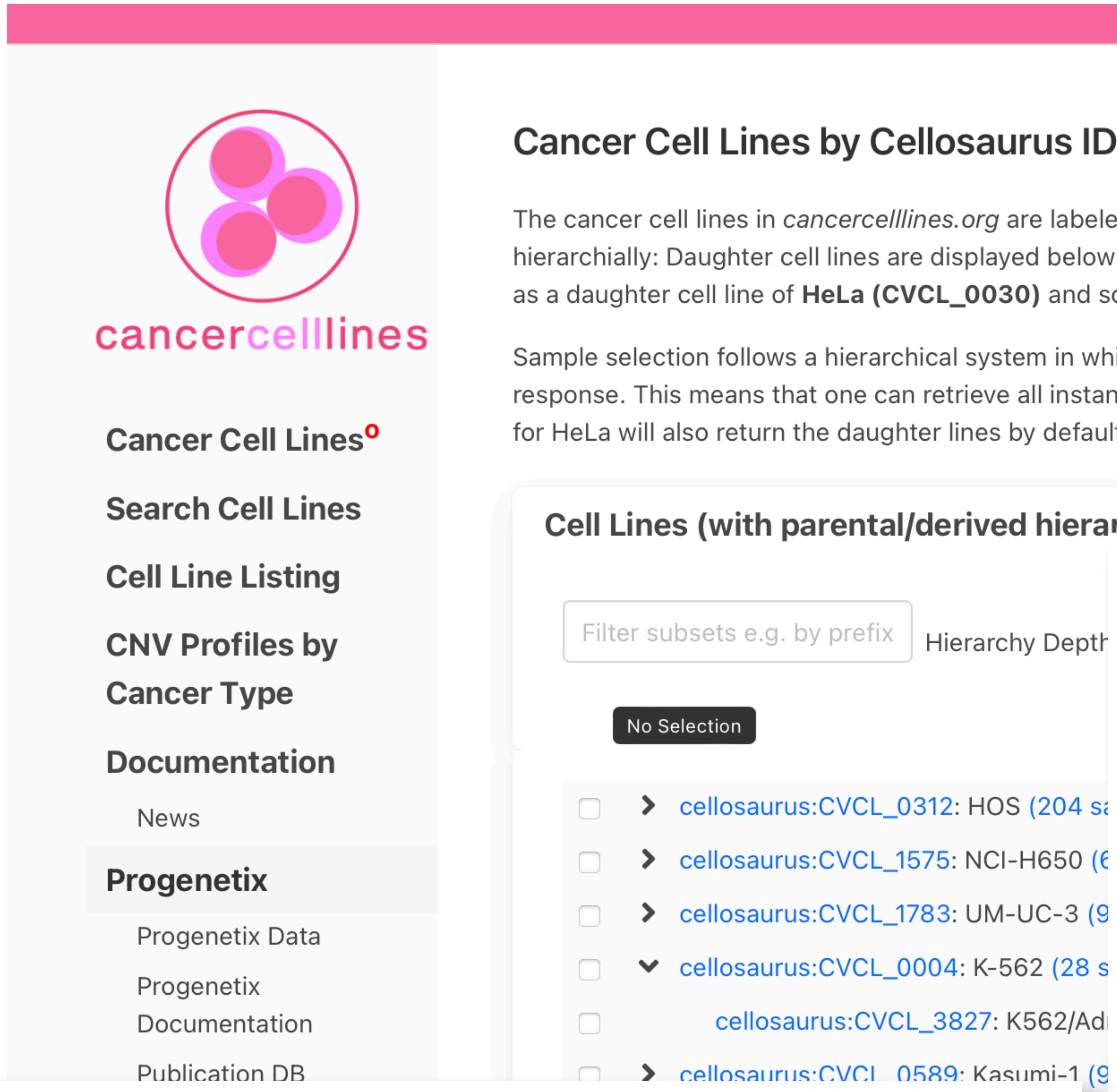
Cancer Genomics Reference Resource

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

➡ integration in data federation approaches

cancercelllines.org

Lead: Rahel Paloots



cancercelllines

Cancer Cell Lines°
Search Cell Lines
Cell Line Listing
CNV Profiles by Cancer Type
Documentation
News
Progenetix
Progenetix Data
Progenetix Documentation
Publication DB

Cancer Cell Lines by Cellosaurus ID

The cancer cell lines in *cancercelllines.org* are labeled by th
hierarchially: Daughter cell lines are displayed below the pri
as a daughter cell line of **HeLa (CVCL_0030)** and so forth.

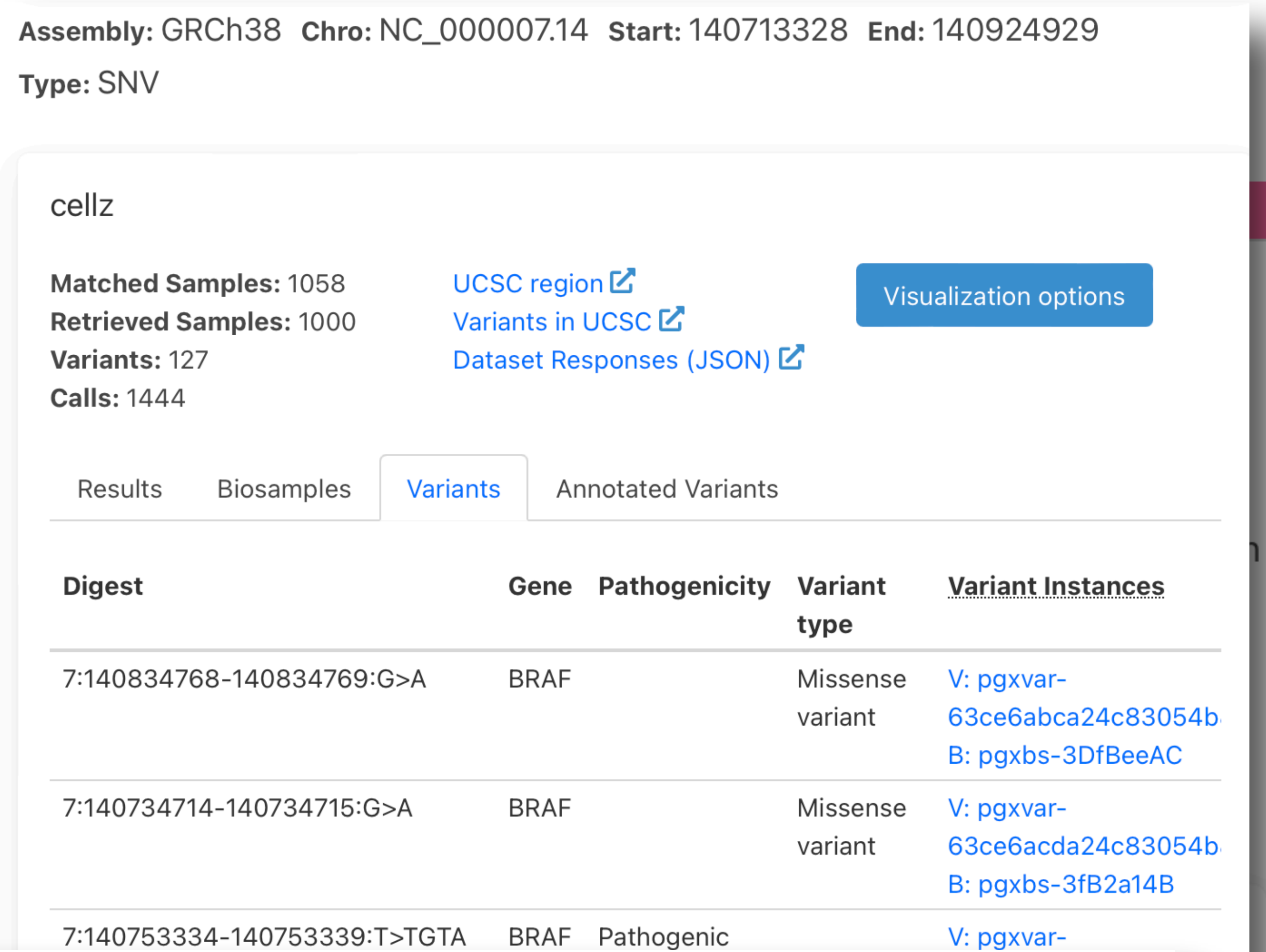
Sample selection follows a hierarchical system in which sam
response. This means that one can retrieve all instances anc
for HeLa will also return the daughter lines by default - but c

Cell Lines (with parental/derived hierarchies

Filter subsets e.g. by prefix | Hierarchy Depth

No Selection

- □ cellosaurus:CVCL_0312: HOS (204 sa
- □ cellosaurus:CVCL_1575: NCI-H650 (6
- □ cellosaurus:CVCL_1783: UM-UC-3 (9
- □ cellosaurus:CVCL_0004: K-562 (28 s
    - □ cellosaurus:CVCL_3827: K562/Ad
- □ cellosaurus:CVCL_0589: Kasumi-1 (9



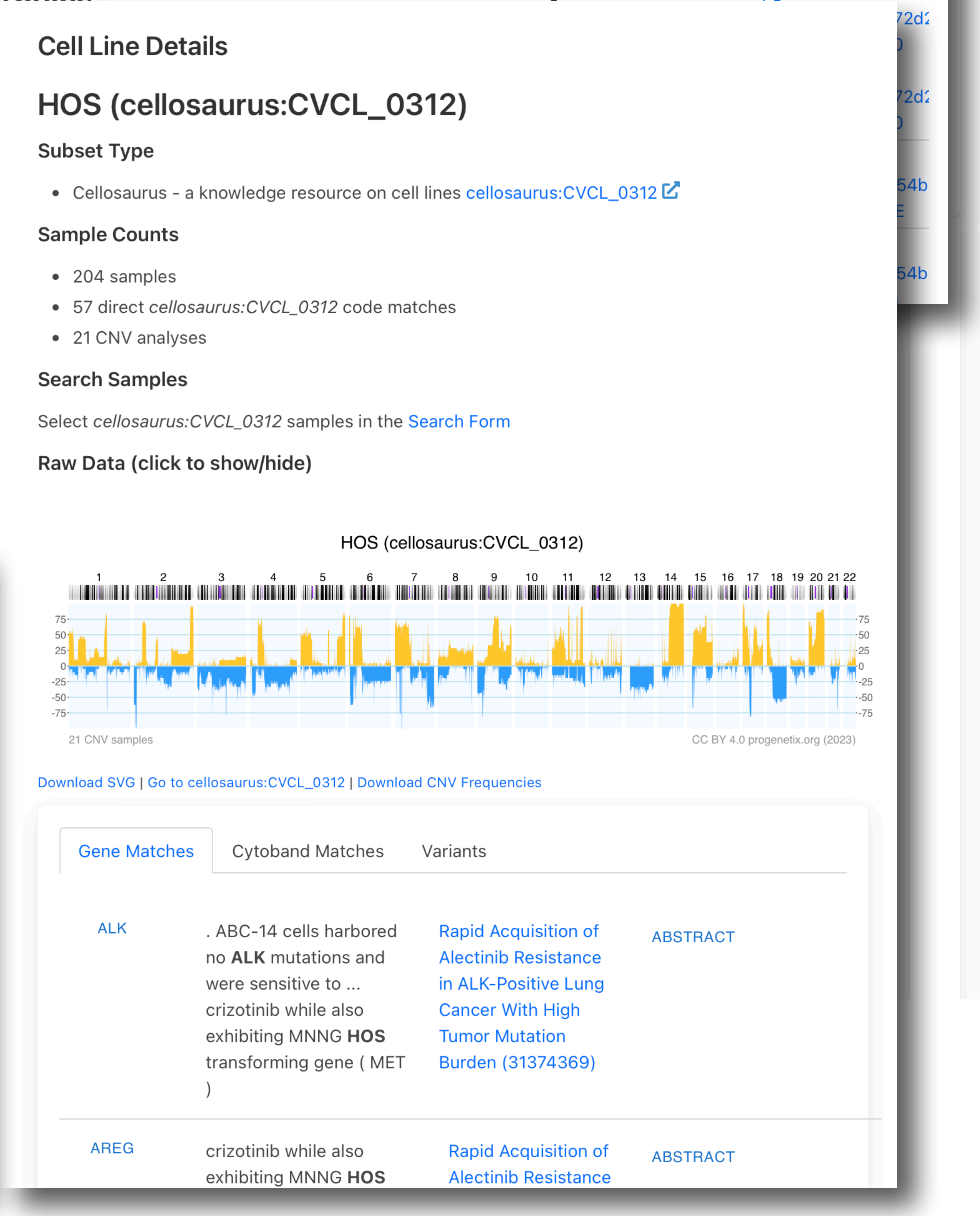
Assembly: GRCh38 **Chro:** NC_000007.14 **Start:** 140713328 **End:** 140924929
Type: SNV

cellz

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

UCSC region
Variants in UCSC
Dataset Responses (JSON)

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



Cell Line Details

HOS (cellosaurus:CVCL_0312)

Subset Type

- Cellosaurus - a knowledge resource on cell lines cellosaurus:CVCL_0312

Sample Counts

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

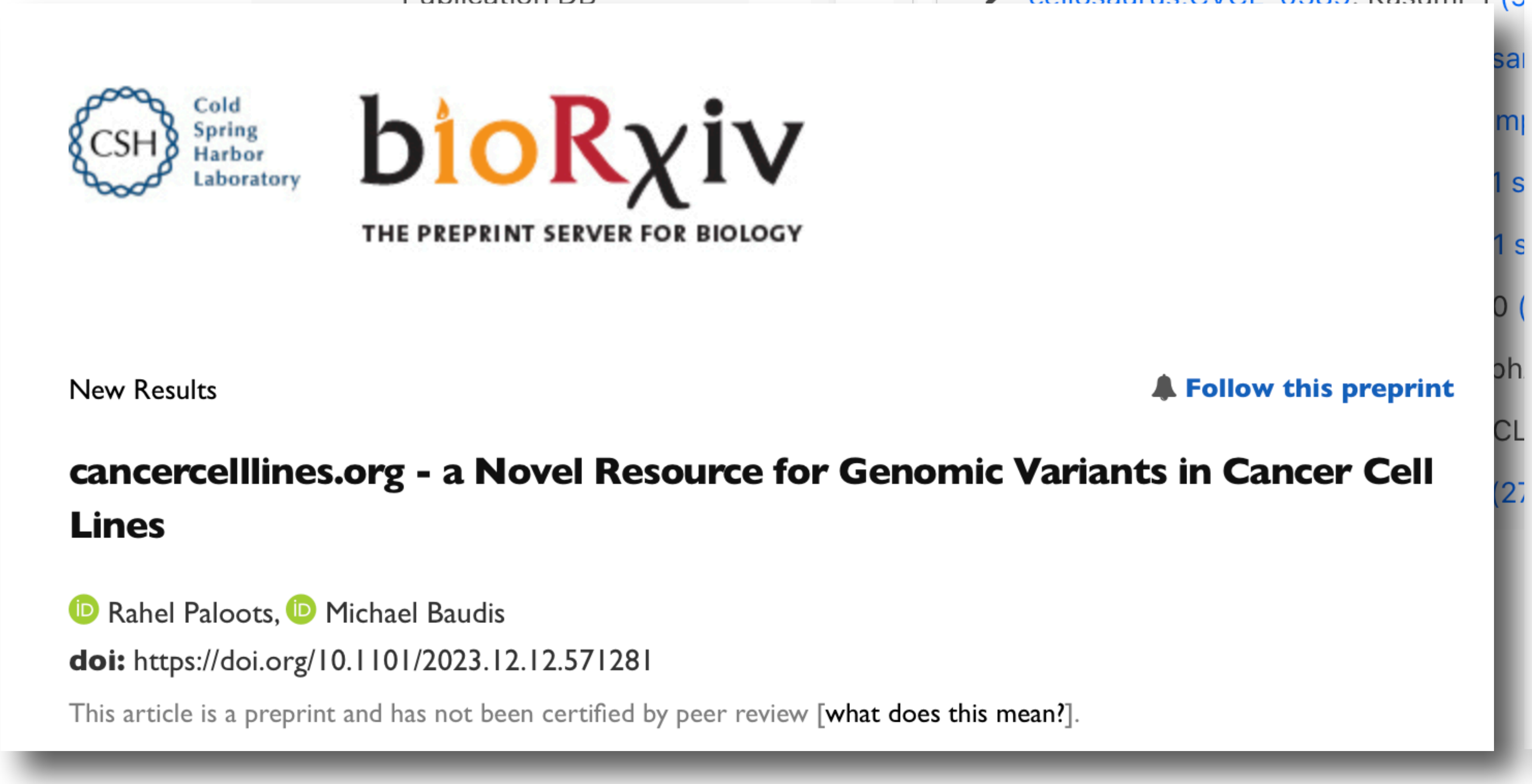
Search Samples

Select *cellosaurus:CVCL_0312* samples in the Search Form

Raw Data (click to show/hide)

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

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



CSH Cold Spring Harbor Laboratory

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

Follow this preprint

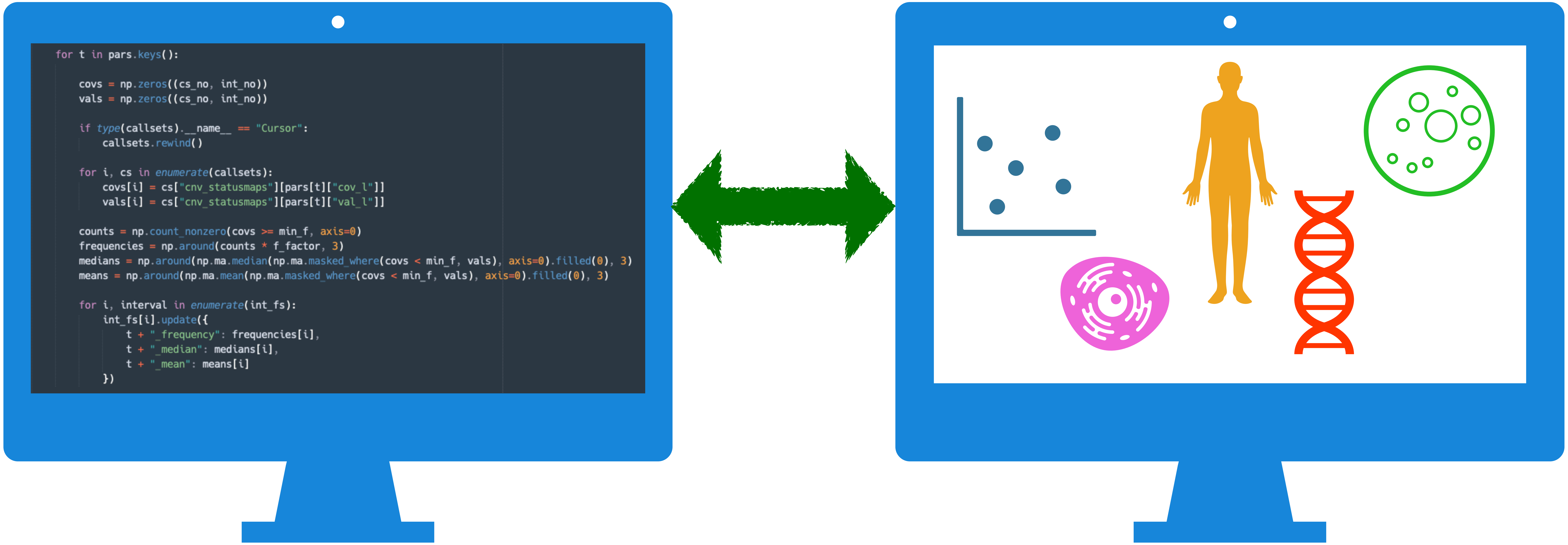
cancercelllines.org - a Novel Resource for Genomic Variants in Cancer Cell Lines

Rahel Paloots, Michael Baudis

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

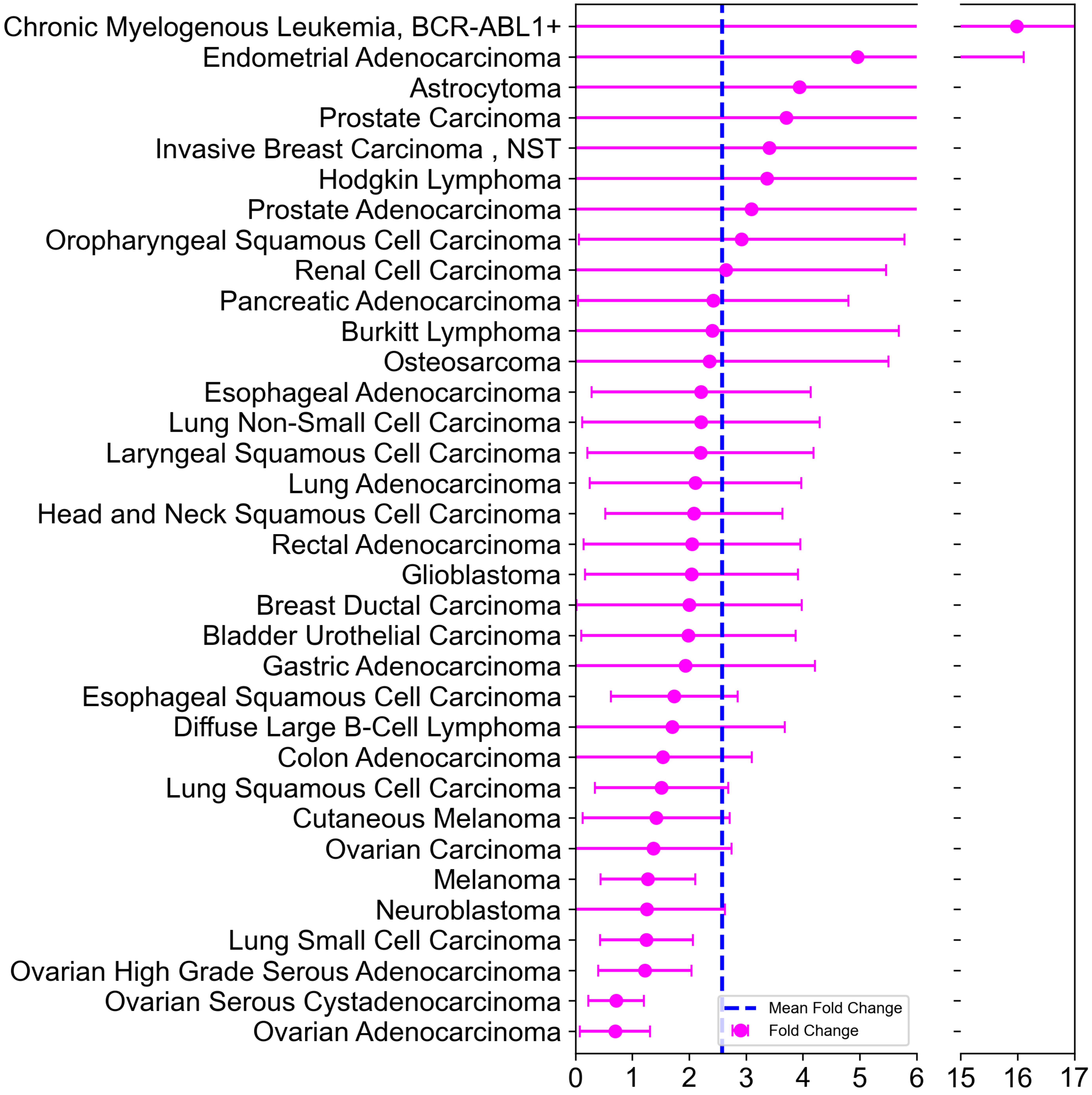
This article is a preprint and has not been certified by peer review [what does this mean?].

{BioInformaticsScience}



```
for t in pars.keys():

    covs = np.zeros((cs_no, int_no))
    vals = np.zeros((cs_no, int_no))

    if type(callsets).__name__ == "Cursor":
        callsets.rewind()

    for i, cs in enumerate(callsets):
        covs[i] = cs["cnv_statusmaps"][pars[t]["cov_l"]]
        vals[i] = cs["cnv_statusmaps"][pars[t]["val_l"]]

    counts = np.count_nonzero(covs >= min_f, axis=0)
    frequencies = np.around(counts * f_factor, 3)
    medians = np.around(np.ma.median(np.ma.masked_where(covs < min_f, vals), axis=0).filled(0), 3)
    means = np.around(np.ma.mean(np.ma.masked_where(covs < min_f, vals), axis=0).filled(0), 3)

    for i, interval in enumerate(int_fs):
        int_fs[i].update({
            t + "_frequency": frequencies[i],
            t + "_median": medians[i],
            t + "_mean": means[i]
        })
```

Higher level of CNV coverage of the genomes of cancer cell lines compared to their origins

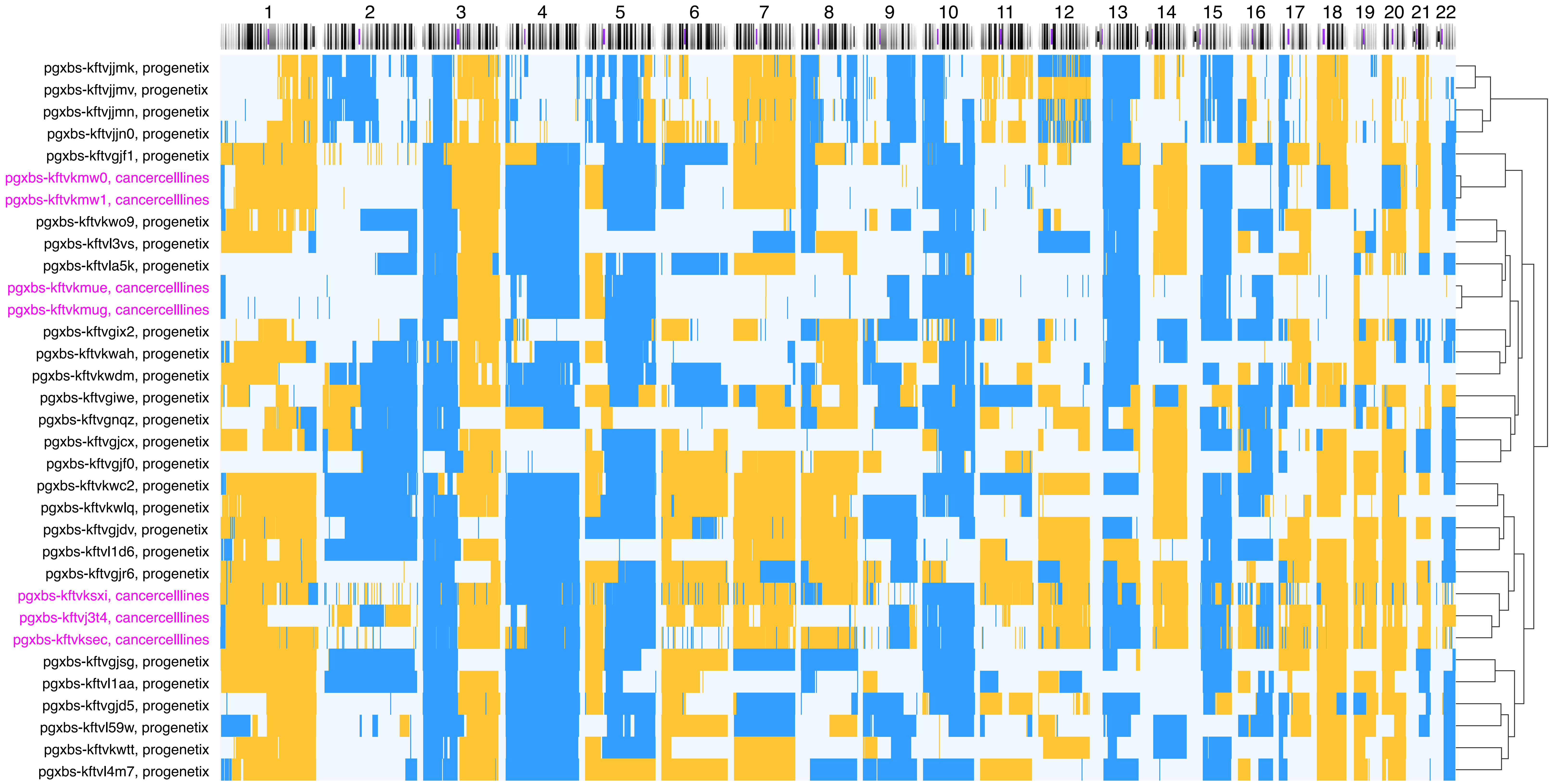


Fold changes between genome CNV coverages of cell lines and tumors

Lead: Rahel Paloots

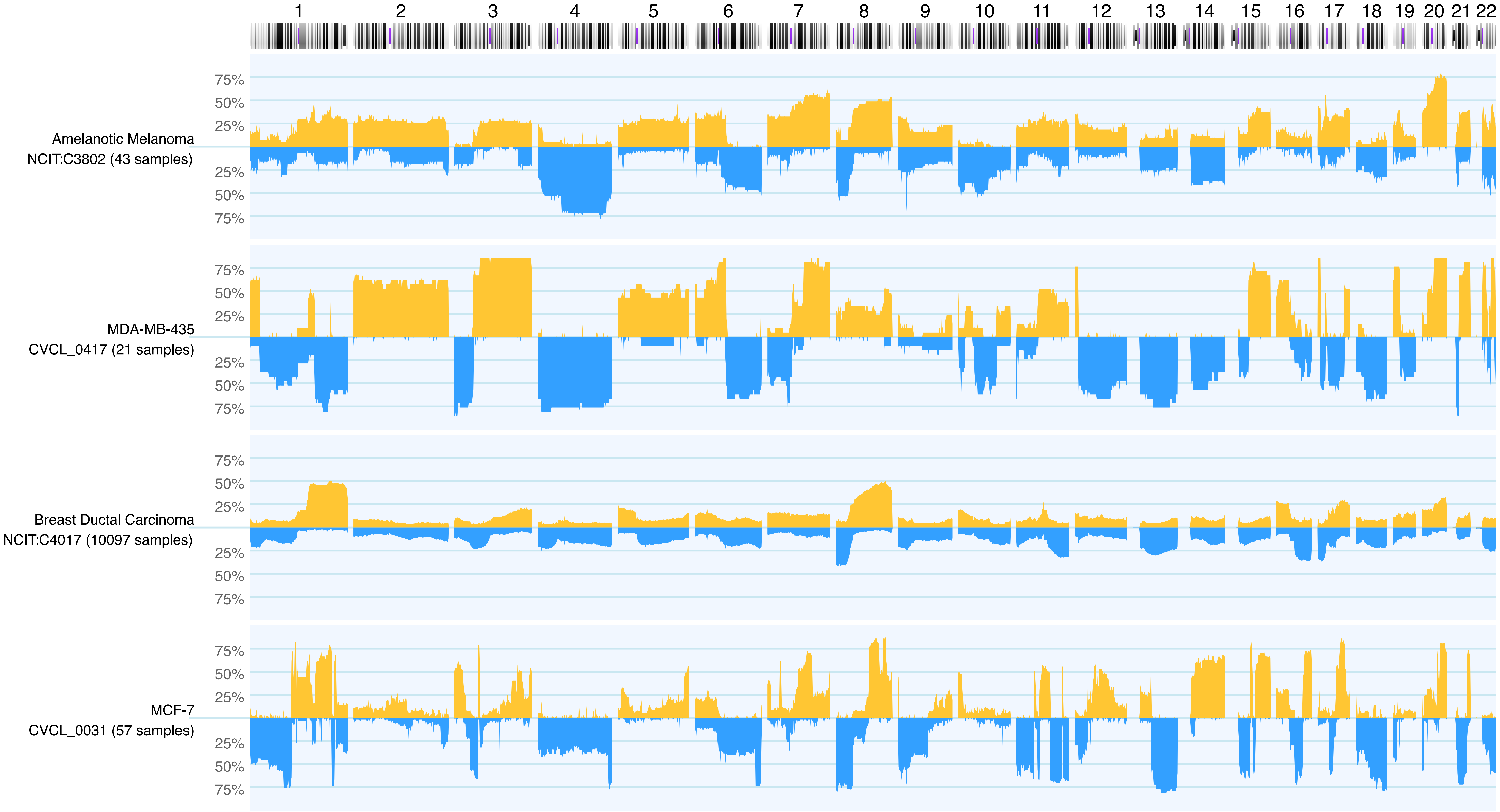
Tumor subpopulations can be matched with highly similar cell lines

- Lung Small Cell Carcinoma Subpopulation
- Cell Lines:
  - CVCL_1140: COR-L279
  - CVCL_1455: NCI-H1105
  - CVCL_1527: NCI-H2107



Lead: Rahel Paloots

Tumor subpopulations can be matched with highly similar cell lines?!

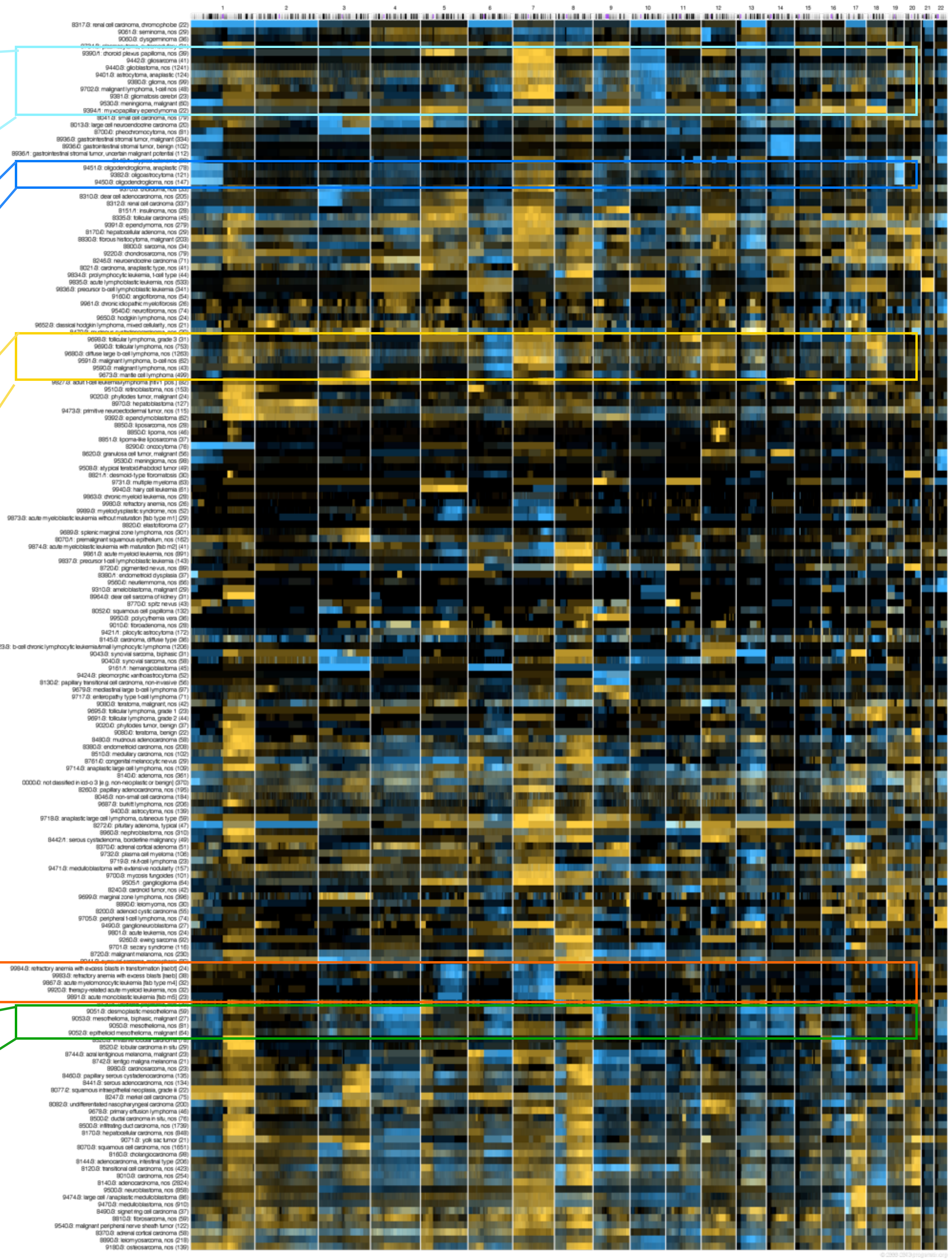


Lead: Rahel Paloots

Somatic Mutations In Cancer: Patterns

Making the case for genomic classifications

Some related cancer entities show similar copy number profiles



9390/1: choroid plexus papilloma, nos (39)
9442/3: gliosarcoma (41)
9440/3: glioblastoma, nos (1241)
9401/3: astrocytoma, anaplastic (124)
9380/3: glioma, nos (99)
9702/3: malignant lymphoma, t-cell nos (48)
9381/3: gliomatosis cerebri (23)
9530/3: meningioma, malignant (60)
9394/1: myxopapillary ependymoma (22)

9451/3: oligodendroglioma, anaplastic (78)
9382/3: oligoastrocytoma (121)
9450/3: oligodendroglioma, nos (147)

9698/3: follicular lymphoma, grade 3 (31)
9690/3: follicular lymphoma, nos (753)
9680/3: diffuse large b-cell lymphoma, nos (1263)
9591/3: malignant lymphoma, b-cell nos (62)
9590/3: malignant lymphoma, nos (43)
9673/3: mantle cell lymphoma (499)

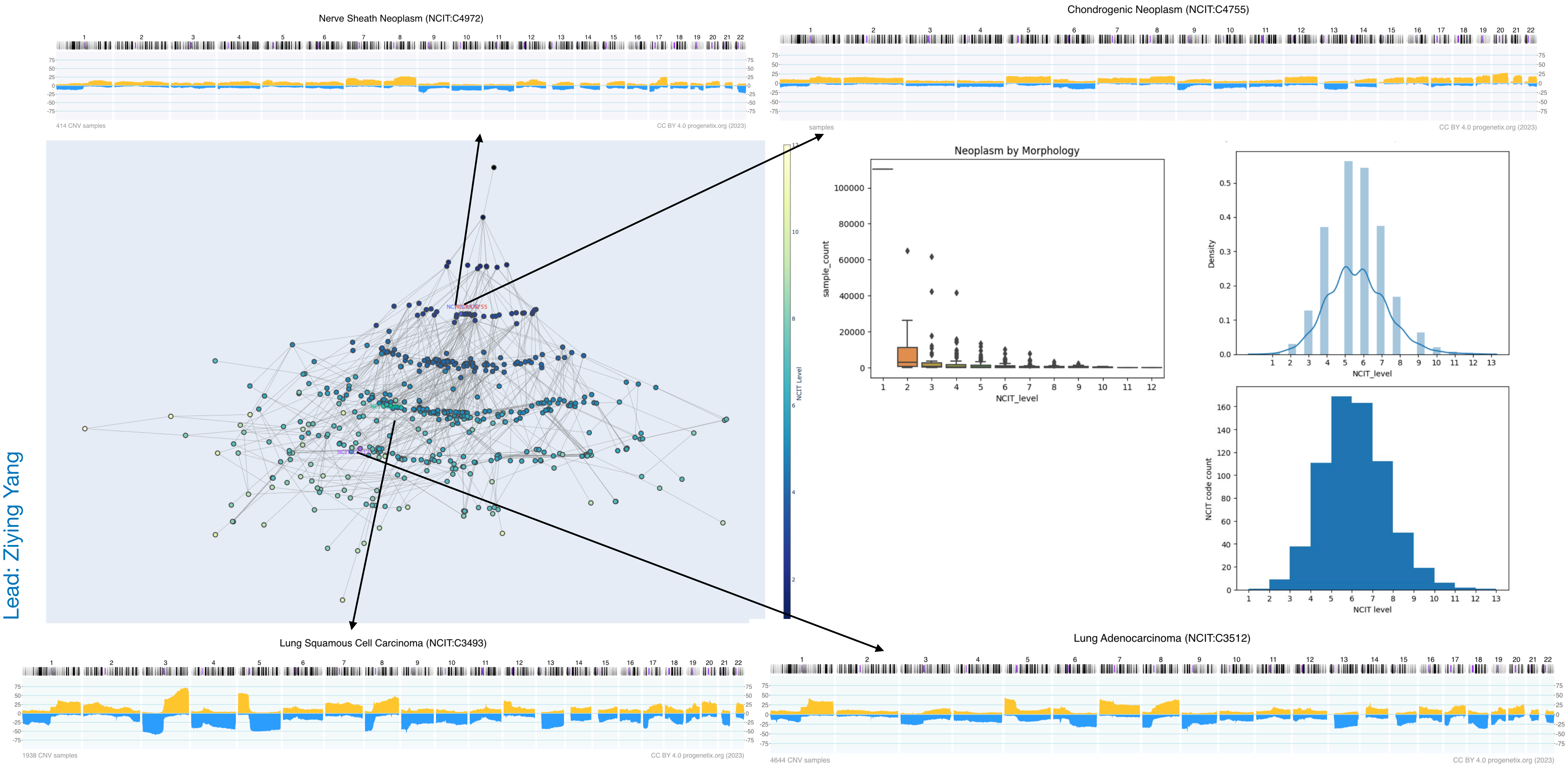
9984/3: refractory anemia with excess blasts in transformation [raebt] (24)
9983/3: refractory anemia with excess blasts [raeb] (38)
9867/3: acute myelomonocytic leukemia [fab type m4] (32)
9920/3: therapy-related acute myeloid leukemia, nos (32)
9891/3: acute monoblastic leukemia [fab m5] (23)

9051/3: desmoplastic mesothelioma (59)
9053/3: mesothelioma, biphasic, malignant (27)
9050/3: mesothelioma, nos (81)
9052/3: epithelioid mesothelioma, malignant (64)

arrayMap

CNV profiles heterogeneity vs cancer classification

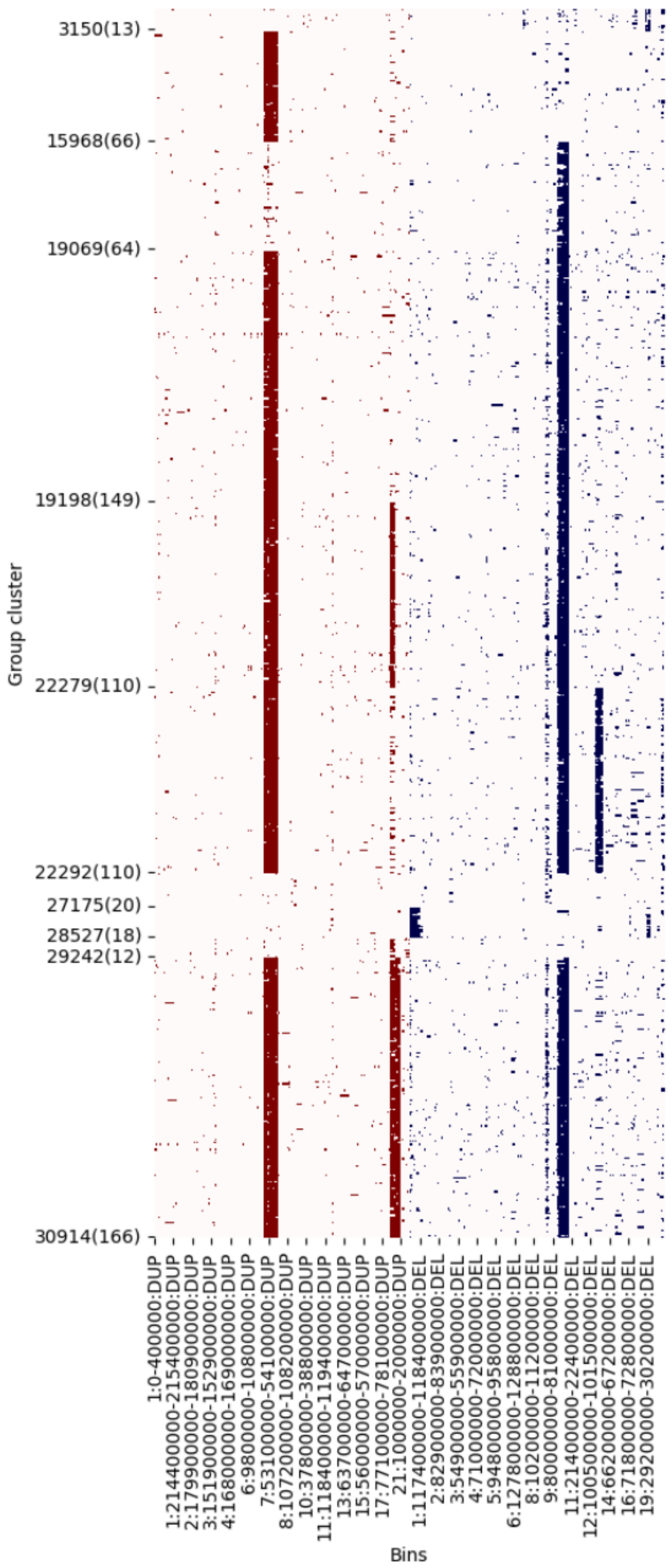
Correspondance of genomic profiles to NCIT cancer hierarchy

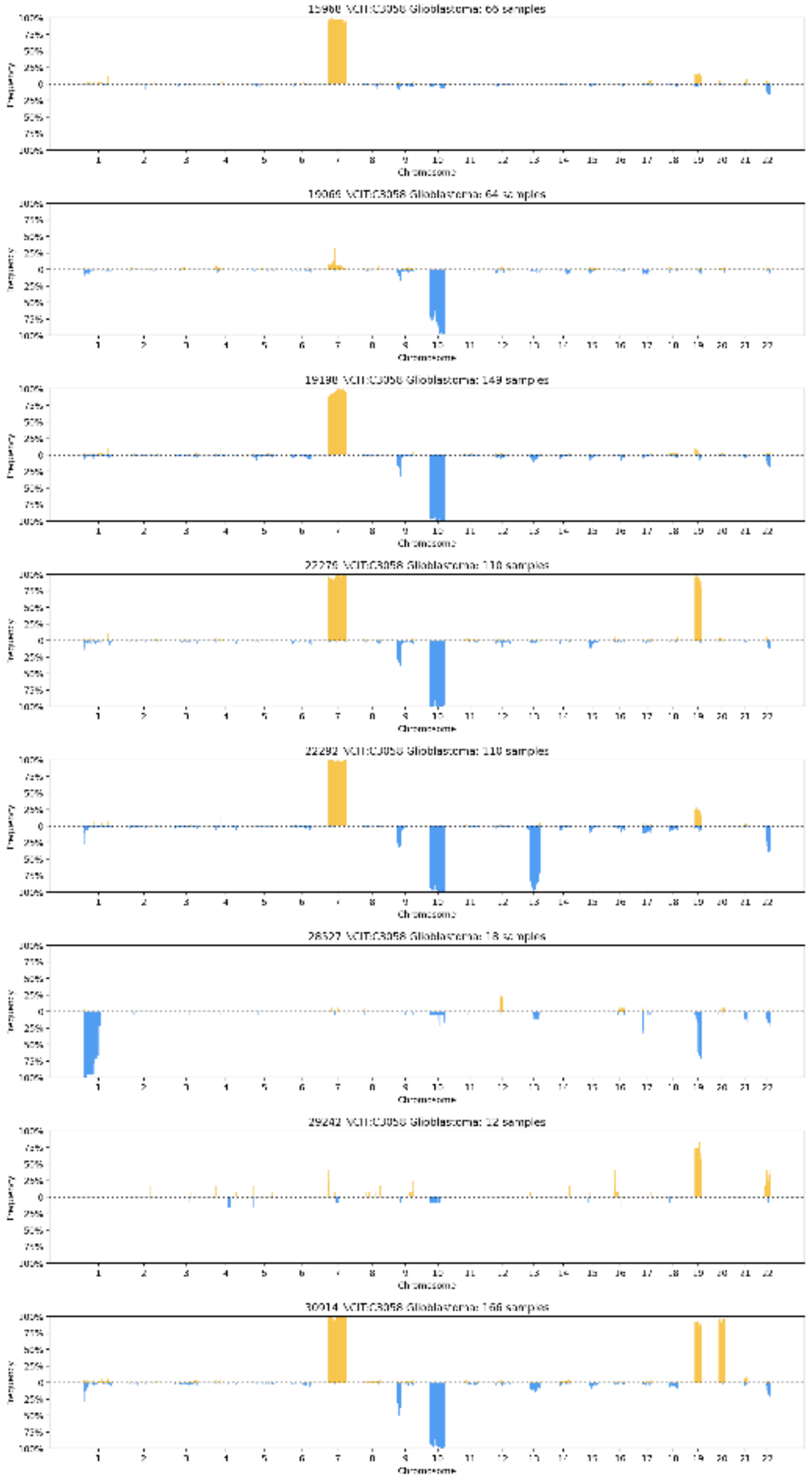


Lead: Ziying Yang

Results

Entity CNV heterogeneity: Glioblastoma



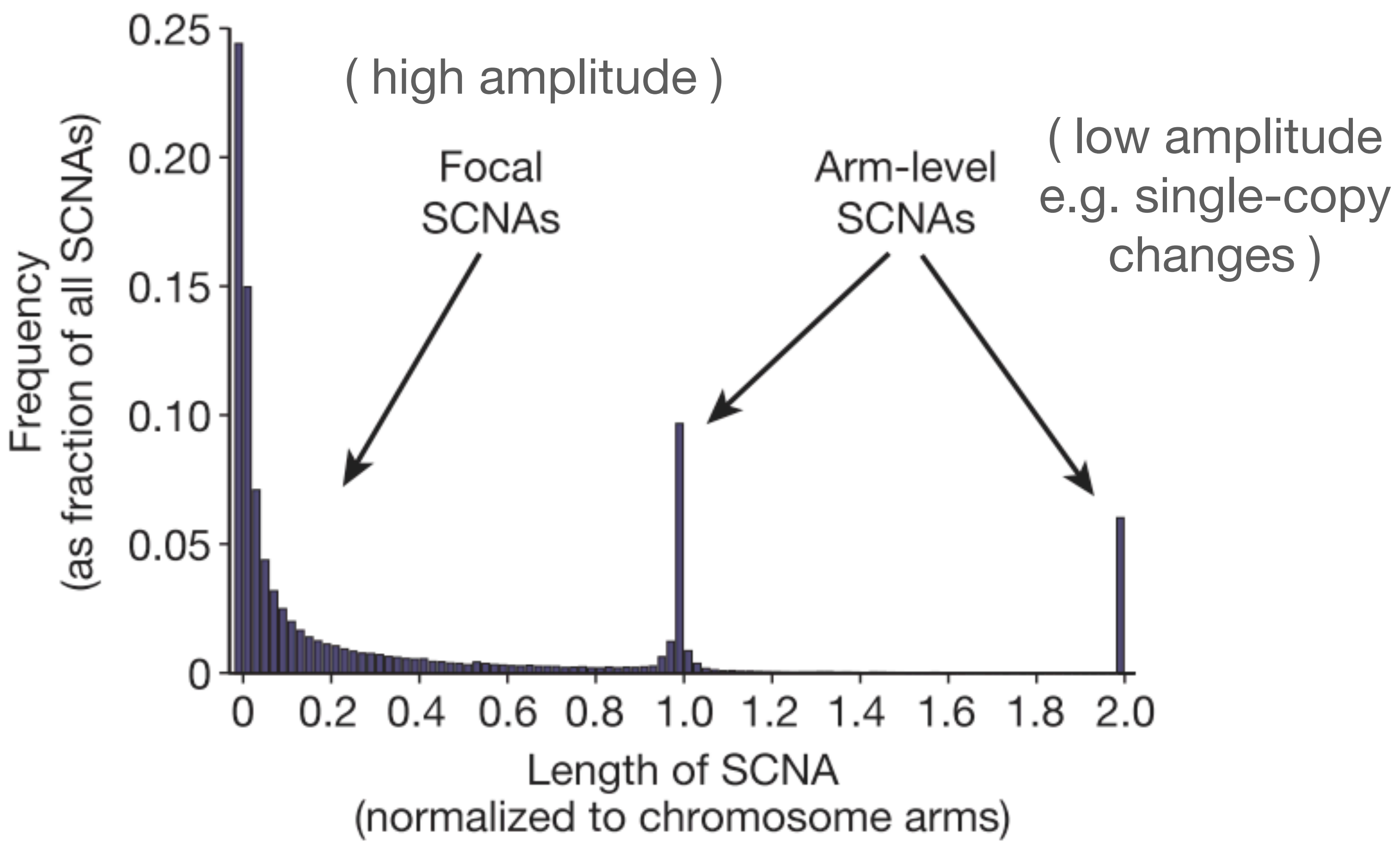


group cluster	CNV features
15968	Dup 7
19069	Del 10
19198	Dup 7, Del 10
22279	Dup 7, Del 10, Dup 19
22292	Dup 7, Del 10, Del 13
28527	Del 1p, Del 19q
29242	Dup 19
30914	Dup 7, Del 10, Dup 19, Dup 20

Lead: Ziying Yang

CNV Categorization

different levels of CNV

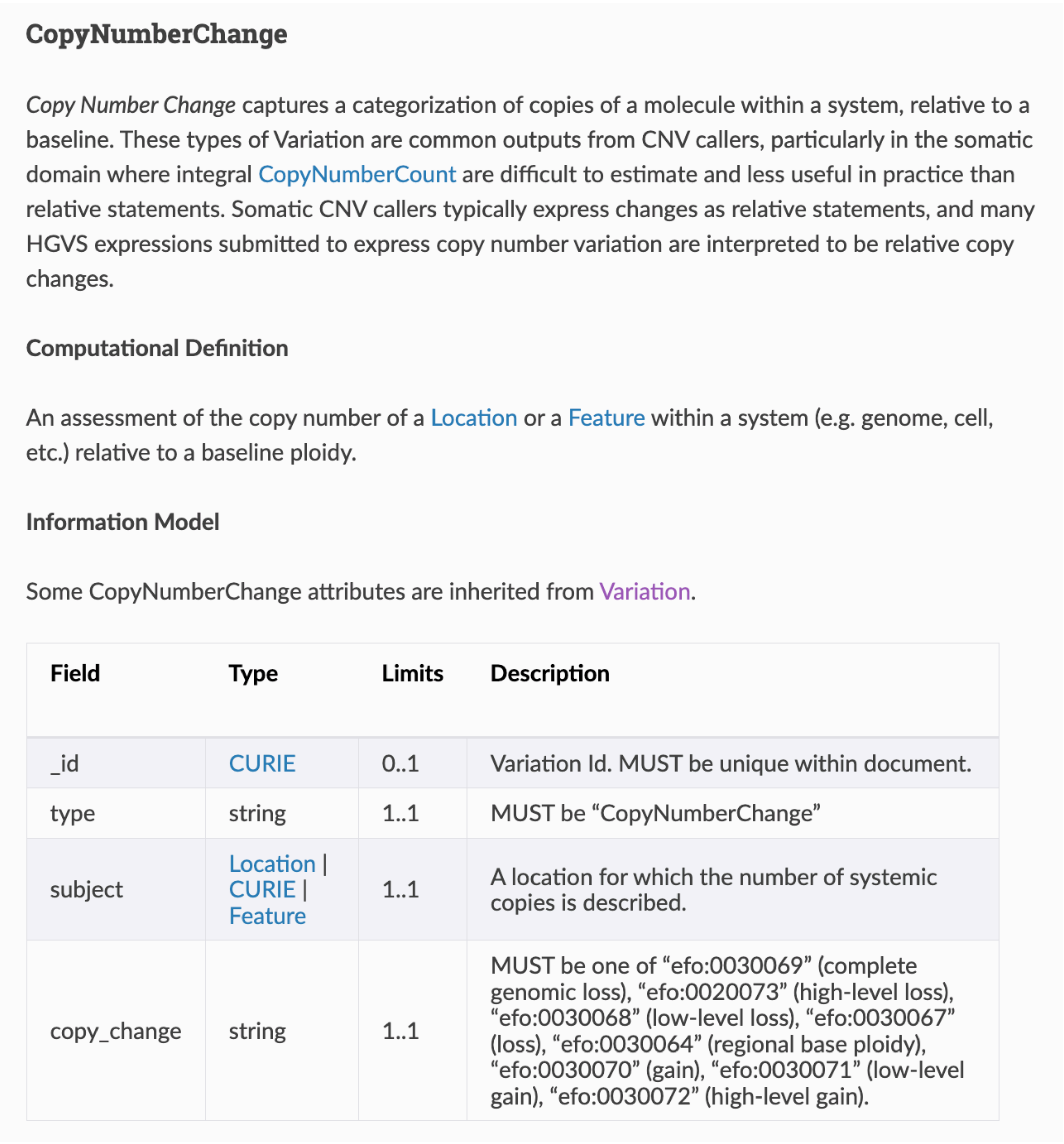


Rameen et al 2010 Nature

GA4GH Variation Representation Specification

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.



CopyNumberChange

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

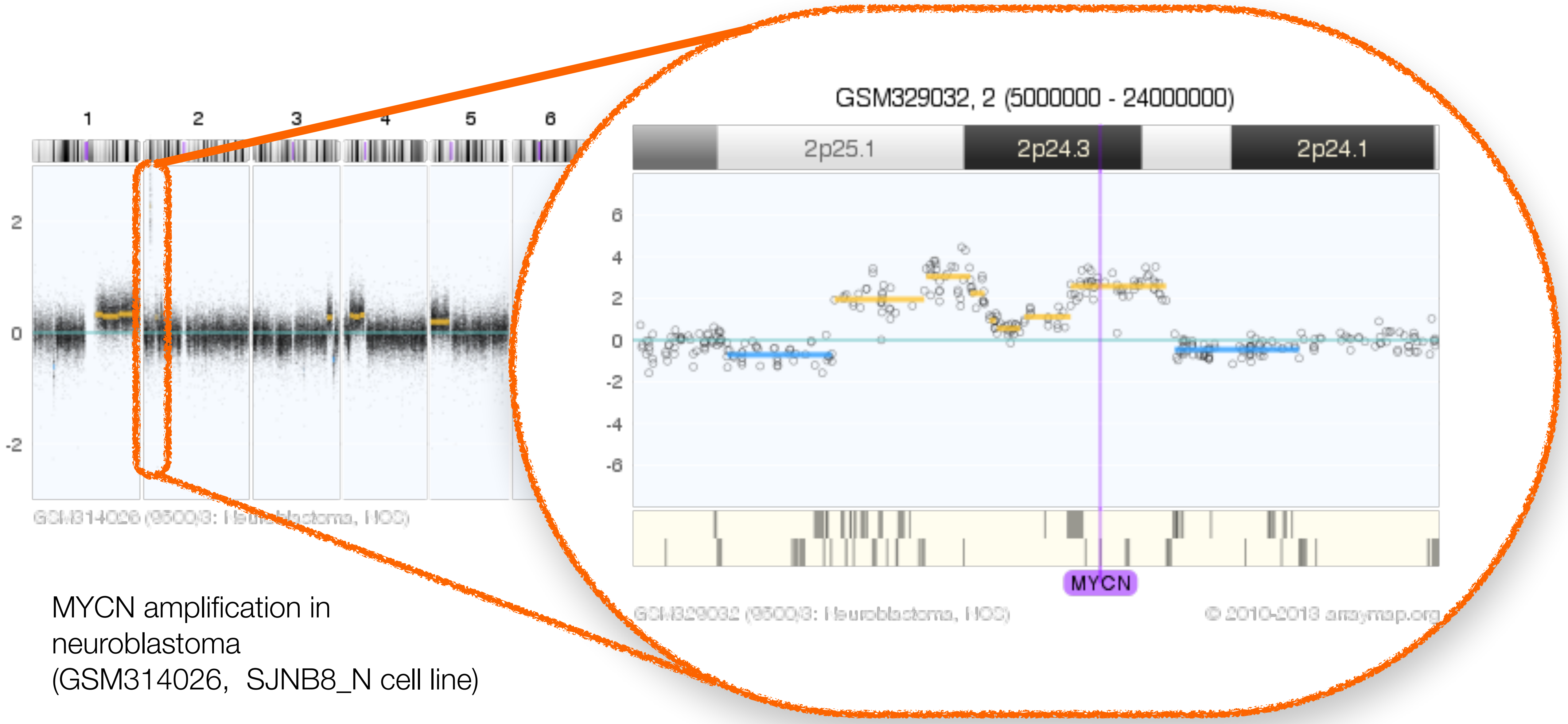
Computational Definition

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

Information Model

Some CopyNumberChange attributes are inherited from Variation.

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



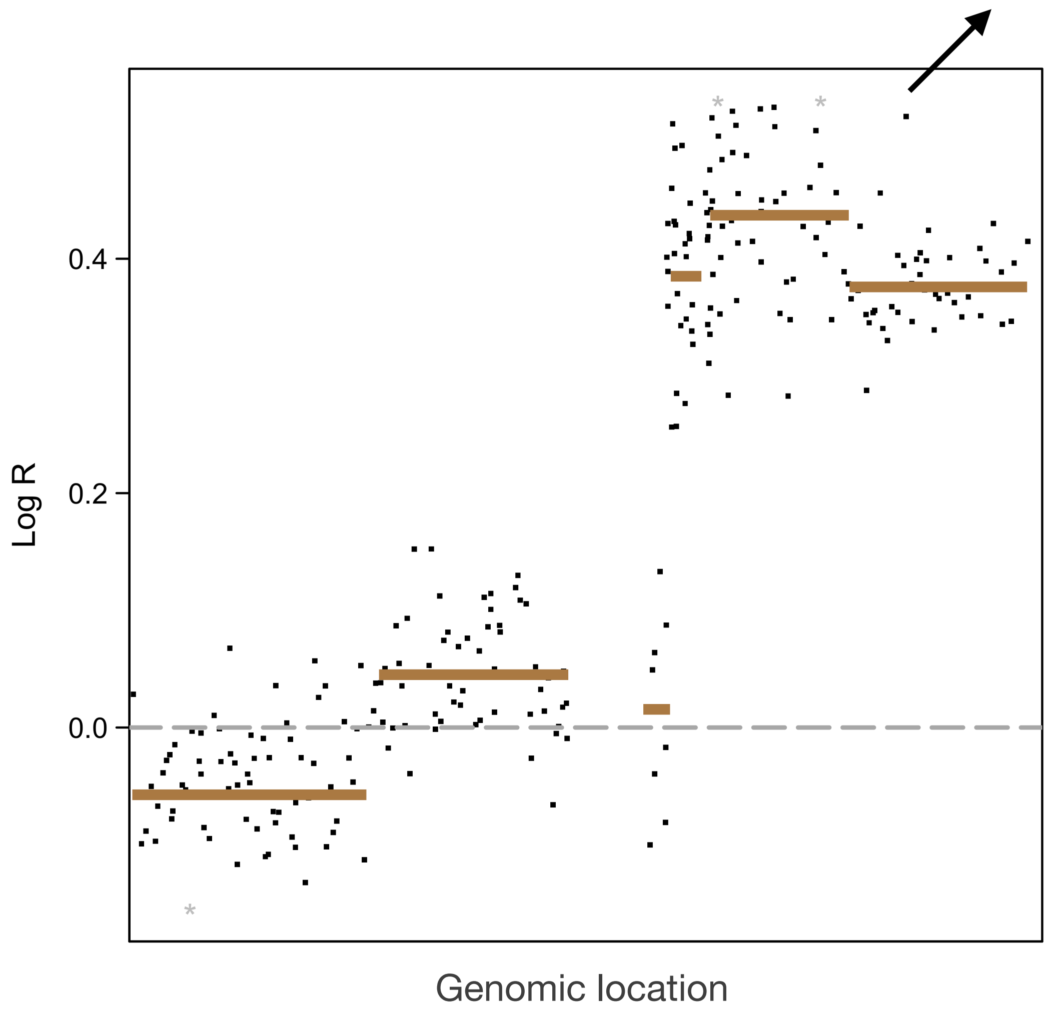
MYCN amplification in neuroblastoma (GSM314026, SJNB8_N cell line)

Lead: Hangjia Zhao

labelSeg

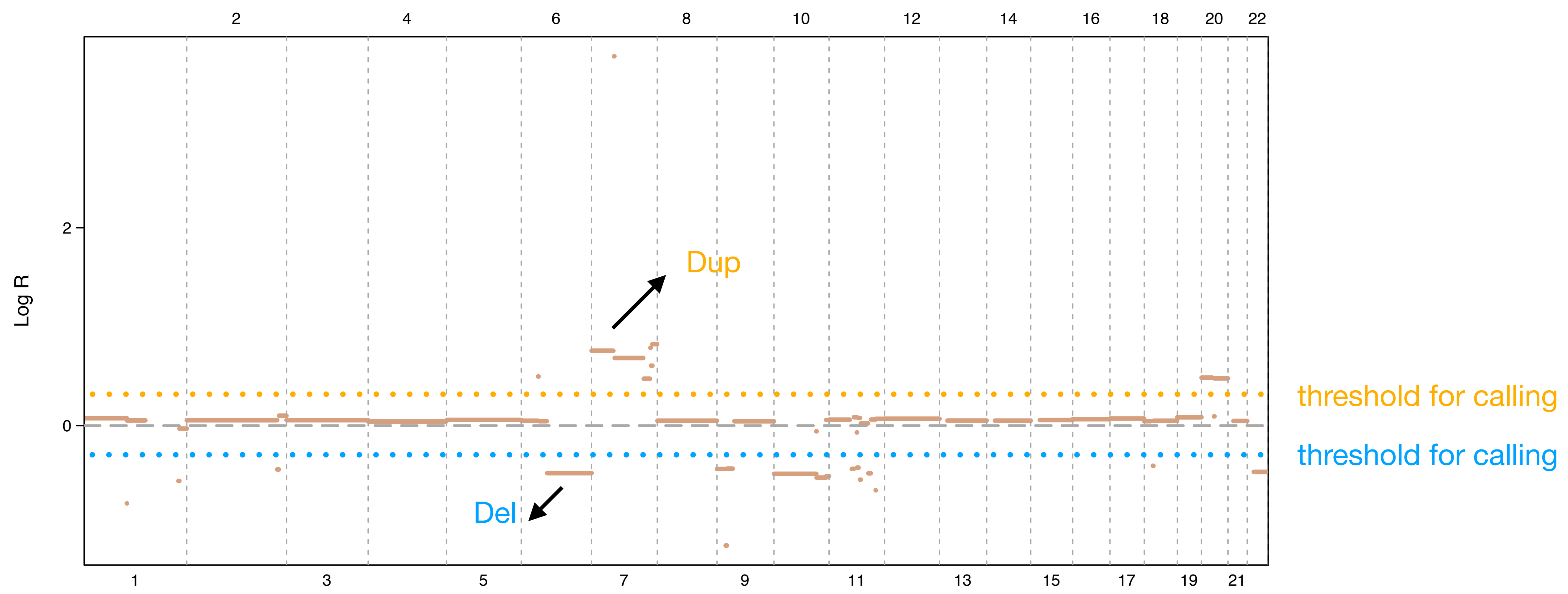
segment annotation for tumor copy number variation profiles

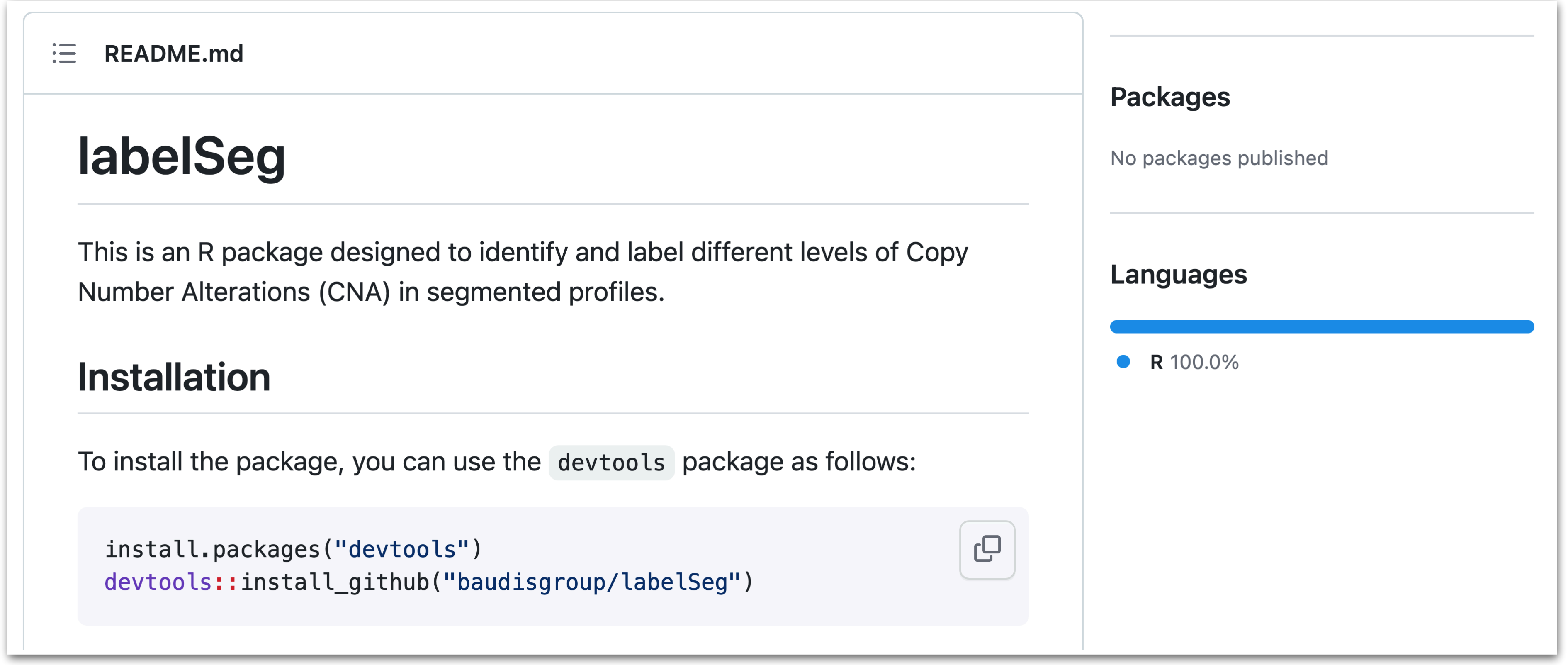
Signal from probes in microarray or from reads in NGS



Segmentation

a step to split the chromosomes into regions of equal copy number that accounts for the noise in the data.





README.md

labelSeg

This is an R package designed to identify and label different levels of Copy Number Alterations (CNA) in segmented profiles.

Installation

To install the package, you can use the `devtools` package as follows:

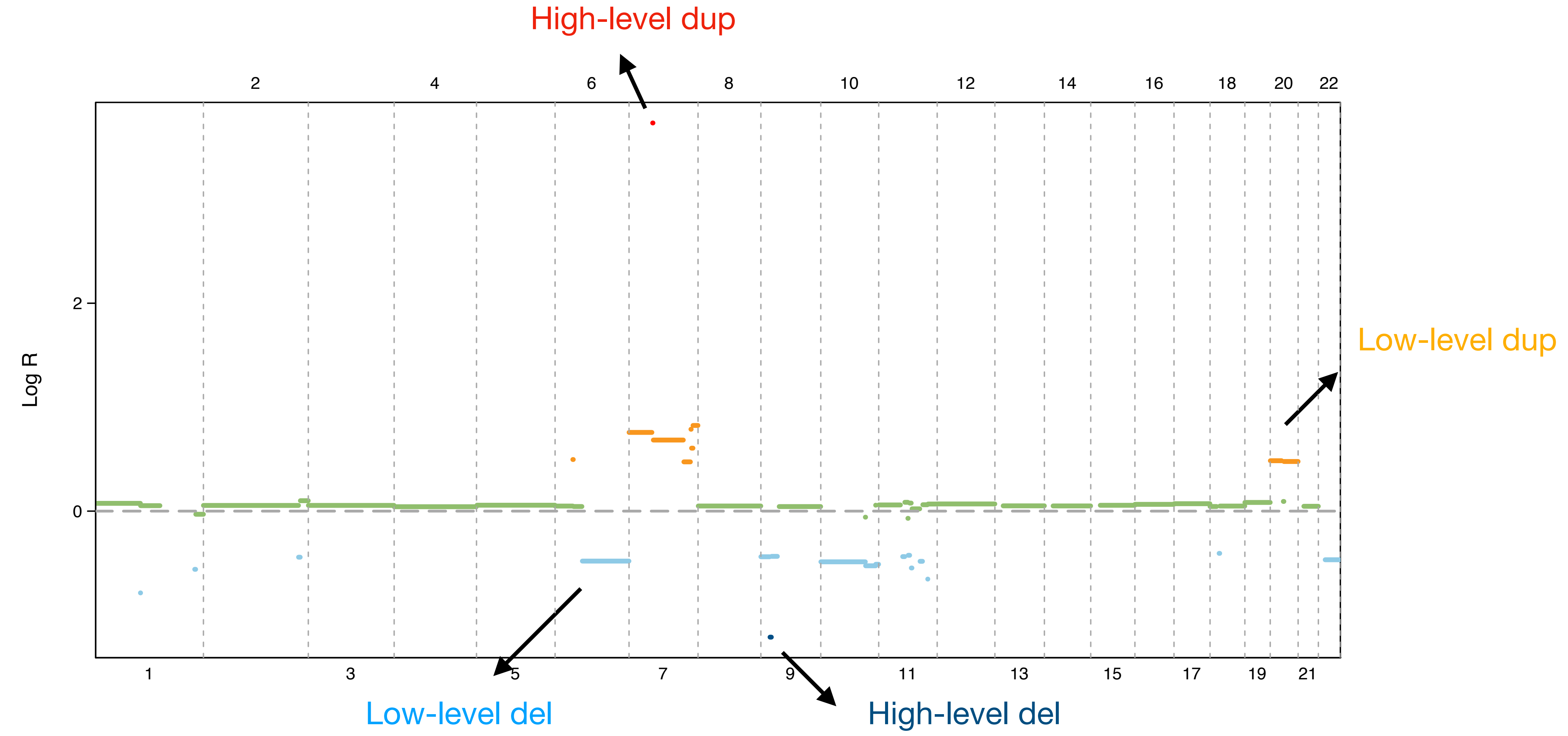
```
install.packages("devtools")
devtools::install_github("baudisgroup/labelSeg")
```

Packages

No packages published

Languages

● R 100.0%

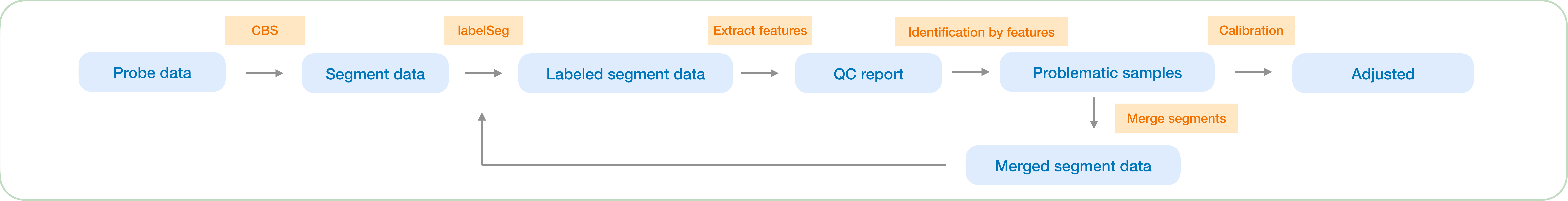


Lead: Hangjia Zhao

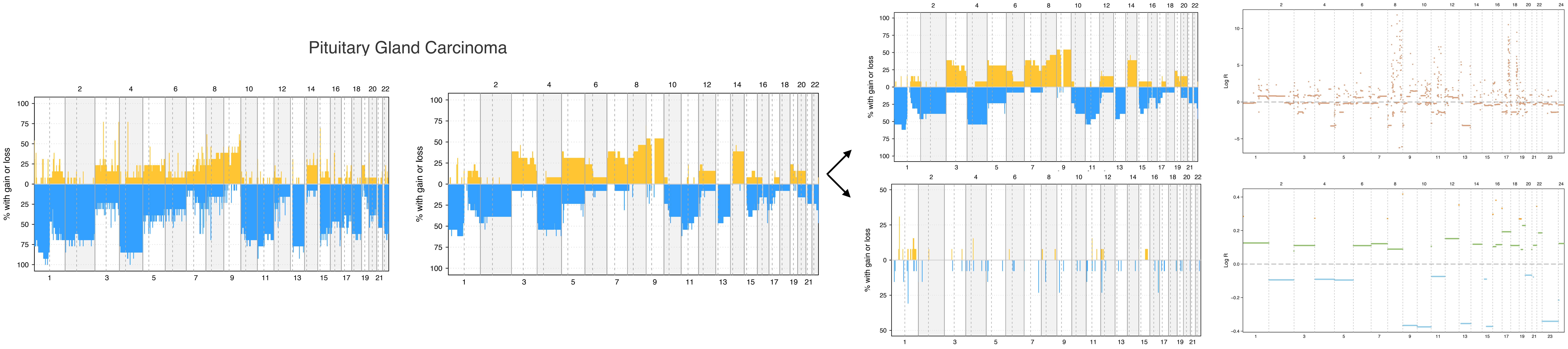
Pipeline Development

improve CNV calling in large numbers of heterogeneous cancer samples

nextflow



Pituitary Gland Carcinoma



Lead: Hangjia Zhao

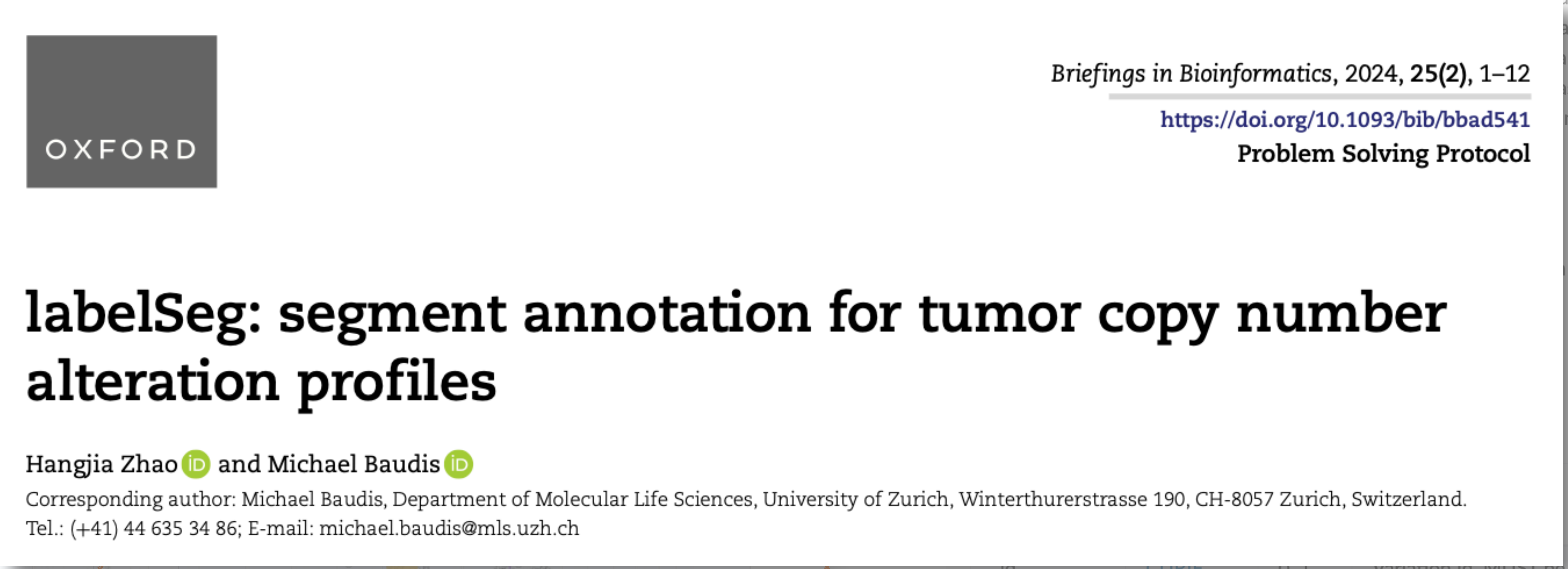
CNV Categorization

different levels of CNV

GA4GH Variation Representation Specification

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

a
Frequency (as fraction of all SCNAs)
0.25
0.20
0.15
0.10
0.05
0
0
0.2
(high amplitude)
Focal
Arm-level
(low amplitude



OXFORD

Briefings in Bioinformatics, 2024, **25(2)**, 1–12

https://doi.org/10.1093/bib/bbad541

Problem Solving Protocol

labelSeg: segment annotation for tumor copy number alteration profiles

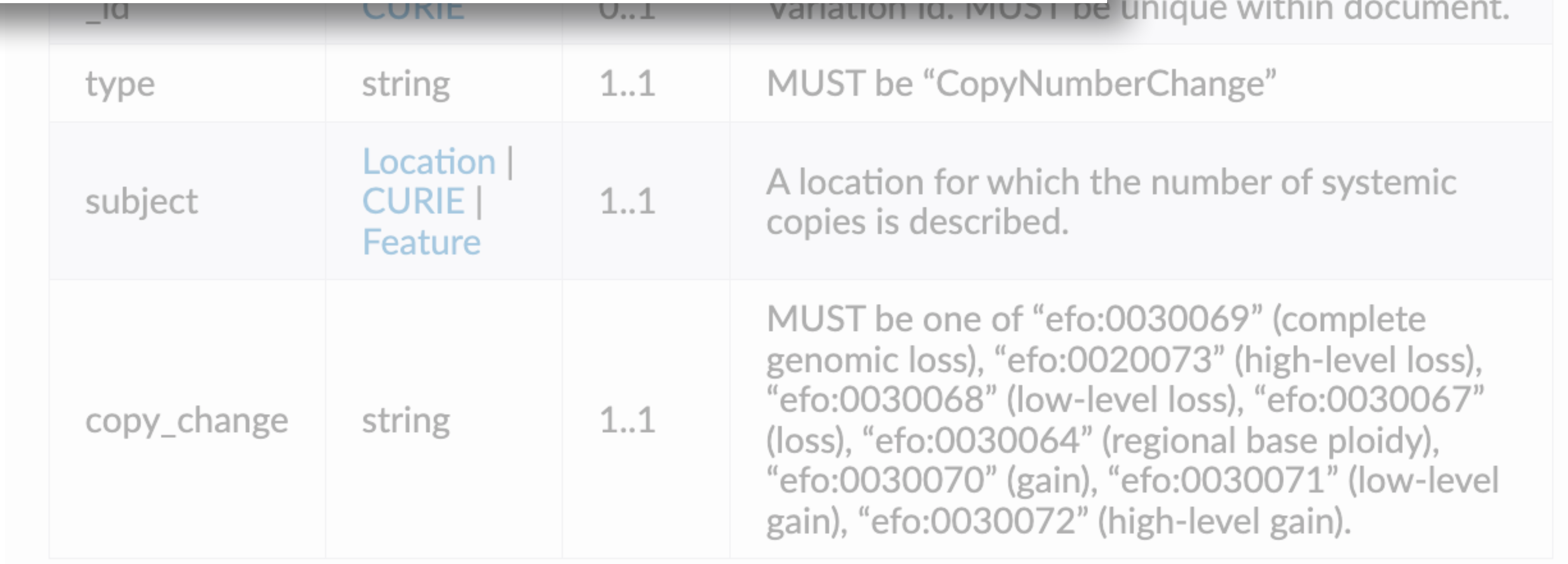
Hangjia Zhao and Michael Baudis

Corresponding author: Michael Baudis, Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland. Tel.: (+41) 44 635 34 86; E-mail: michael.baudis@mls.uzh.ch

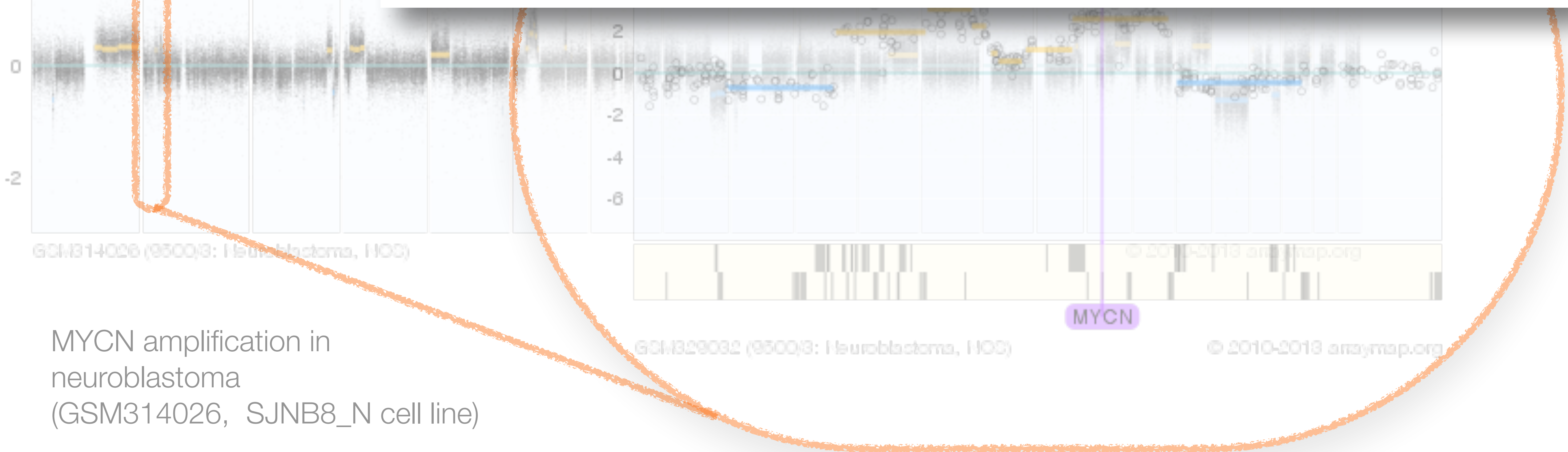
CopyNumberChange

Copy Number Change captures a categorization of copies of a molecule within a system, relative to a
...llers, particularly in the somatic
...nd less useful in practice than
...s relative statements, and many
...nterpreted to be relative copy

... a system (e.g. genome, cell,



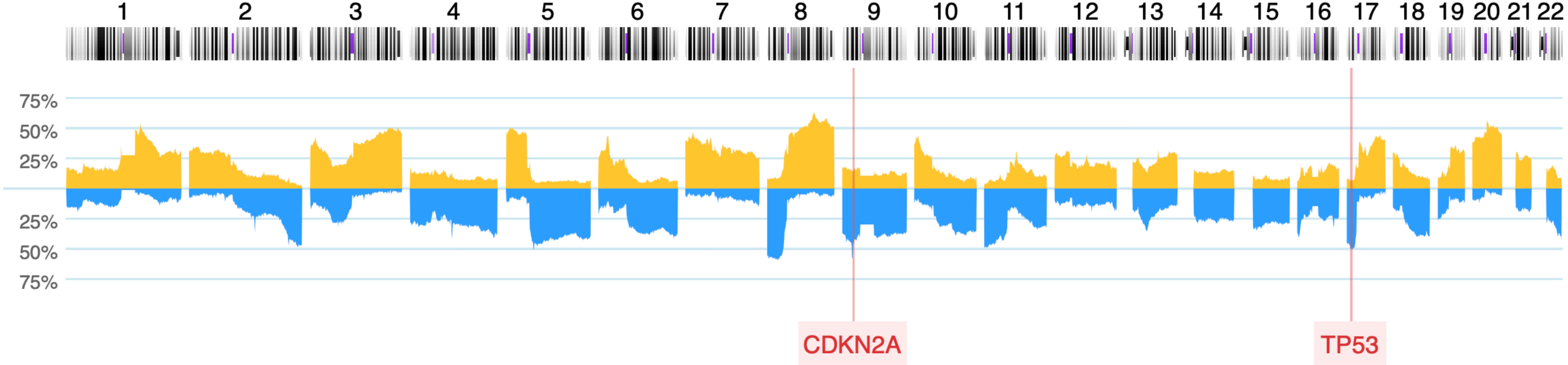
_id	CURIE	0..1	Variation Id. MUST be unique within document.
type	string	1..1	MUST be "CopyNumberChange"
subject	Location | CURIE | Feature	1..1	A location for which the number of systemic copies is described.
copy_change	string	1..1	MUST be one of "efo:0030069" (complete genomic loss), "efo:0020073"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

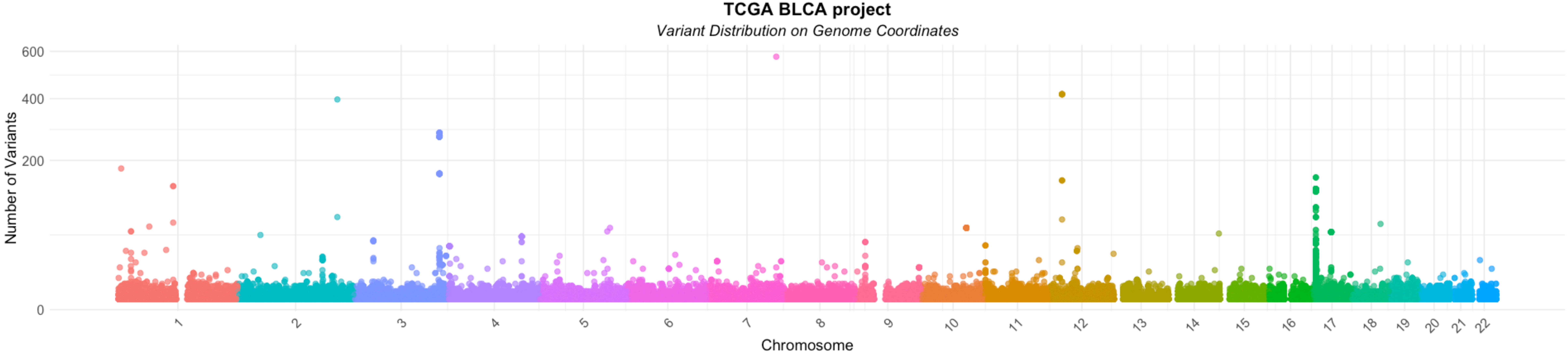


MYCN amplification in neuroblastoma (GSM314026, SJNB8_N cell line)

Lead: Hangjia Zhao

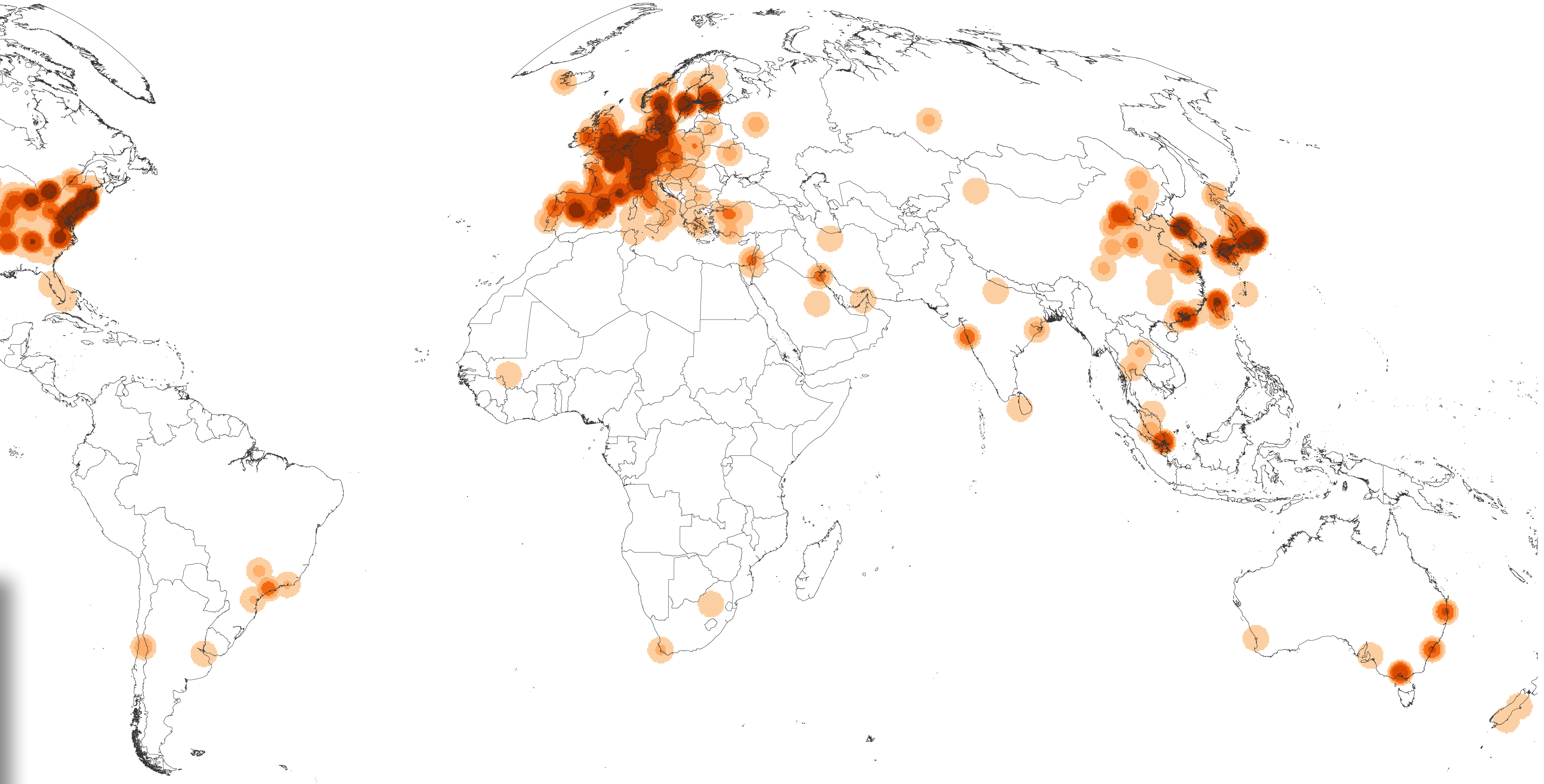
TCGA BLCA project (pgx:TCGA.BLCA)

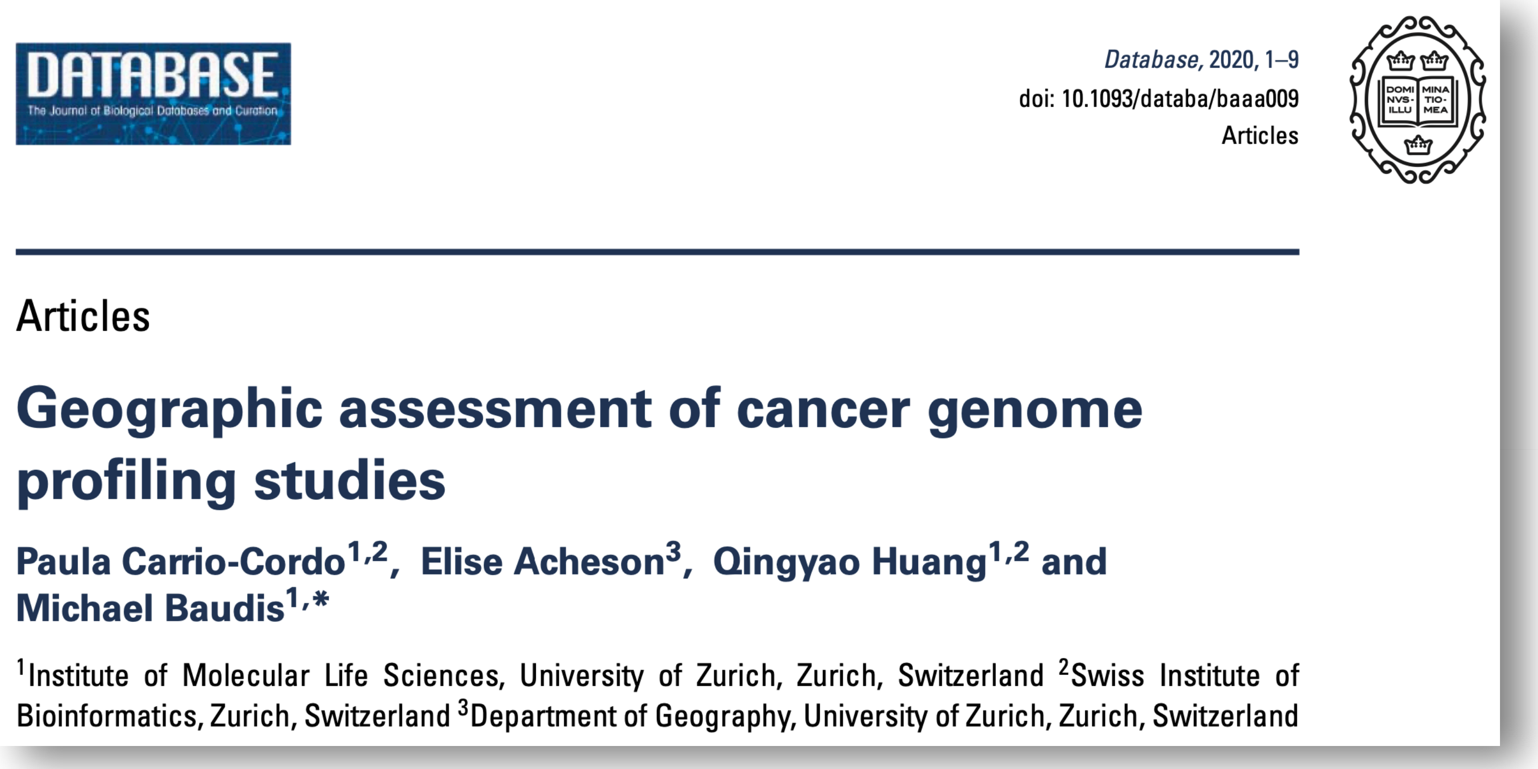




Where does Genomic Data Come From?

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

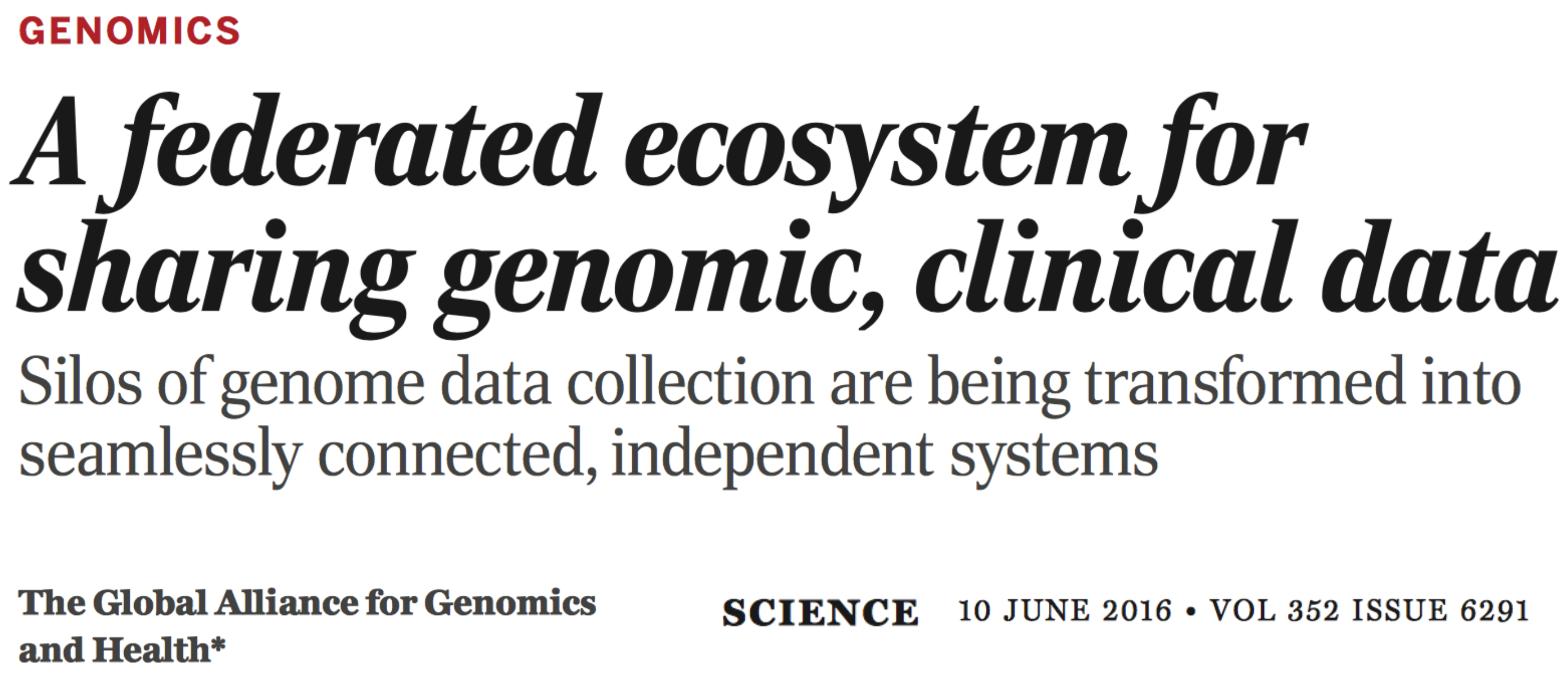
Paula Carrio-Cordo[1,2], Elise Acheson[3], Qingyao Huang[1,2] and Michael Baudis[1,*]

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland [2]Swiss Institute of Bioinformatics, Zurich, Switzerland [3]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

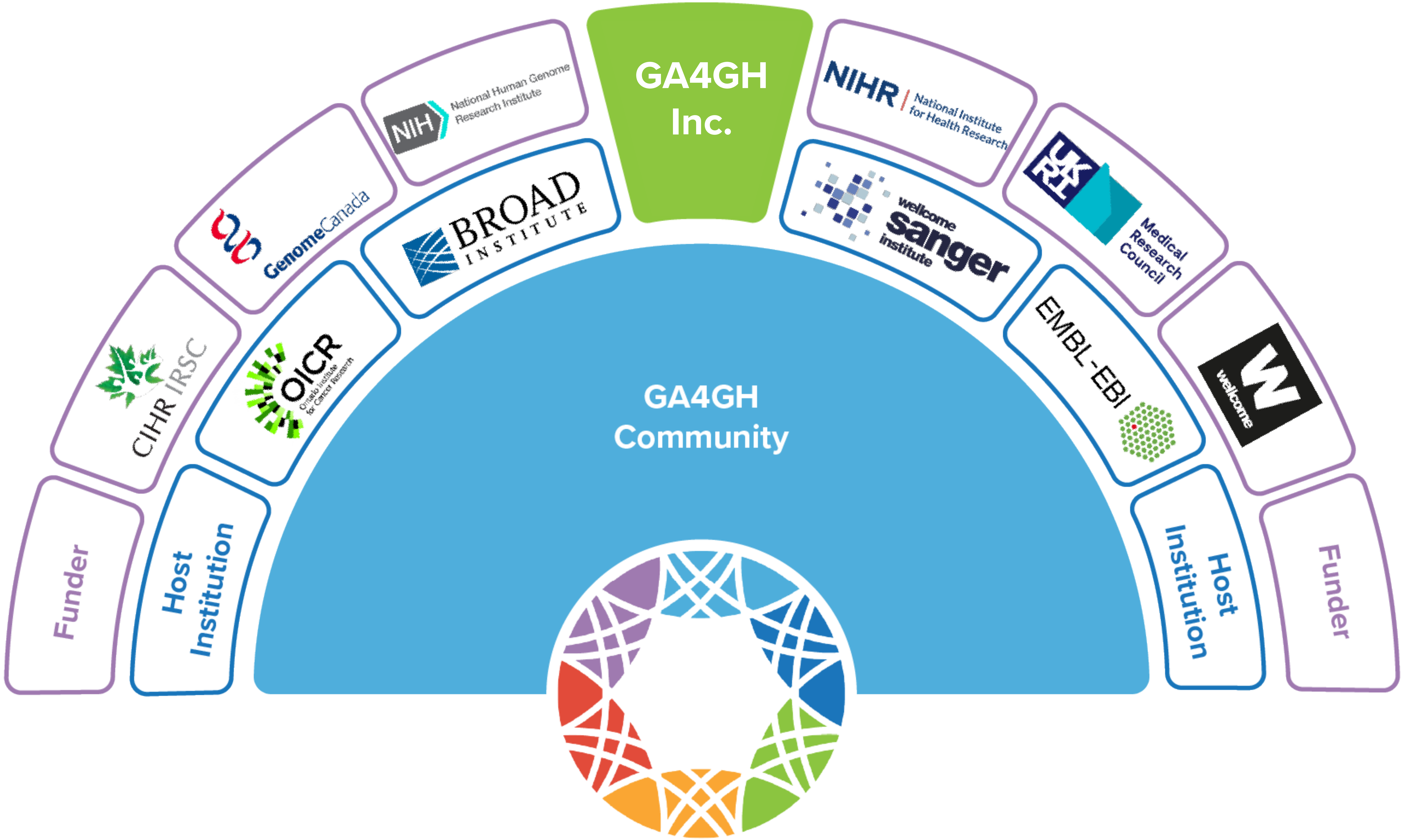
Silos of genome data collection are being transformed into seamlessly connected, independent systems

The Global Alliance for Genomics and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

Organization

Global Alliance
for Genomics & Health



GA4GH Inc.

GA4GH Community

Funder

Host Institution

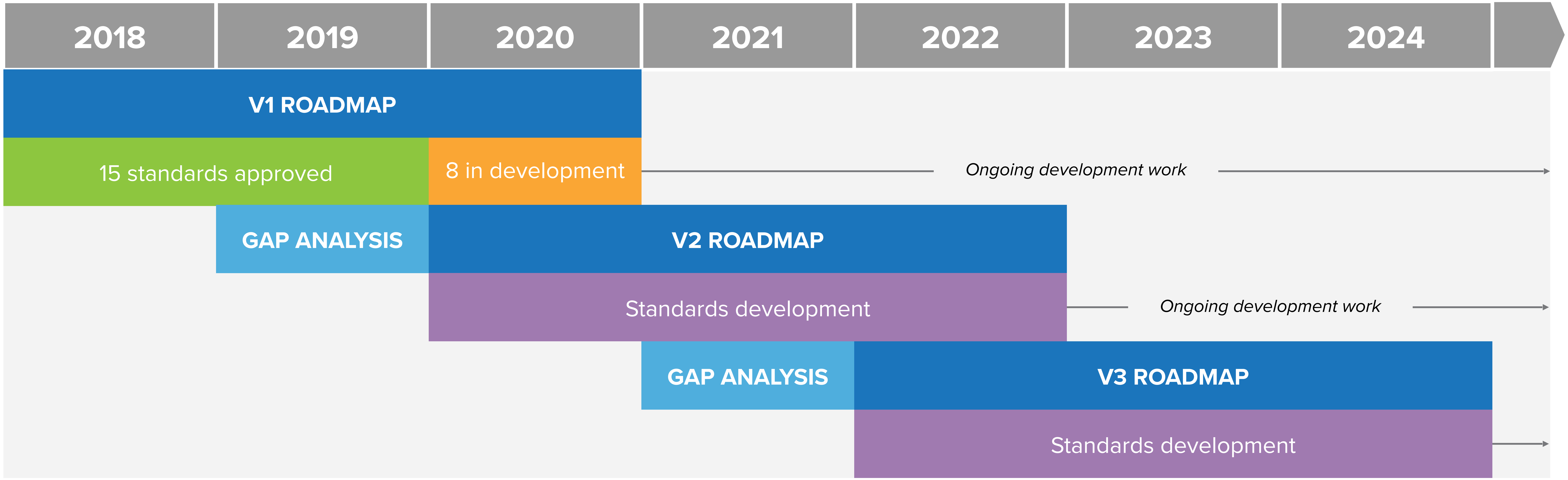
Host Institution

Funder

GenomeCanada
NIH National Human Genome Research Institute
NIHR | National Institute for Health Research
UKRI Medical Research Council
CIHR IRSC
wellcome
BROAD INSTITUTE
wellcome sanger institute
OICR Ontario Institute for Cancer Research
EMBL-EBI

GA4GH Roadmap Development Process

Global Alliance
for Genomics & Health



2018 2019 2020 2021 2022 2023 2024

V1 ROADMAP

15 standards approved

8 in development

Ongoing development work

GAP ANALYSIS

V2 ROADMAP

Standards development

Ongoing development work

GAP ANALYSIS

V3 ROADMAP

Standards development

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

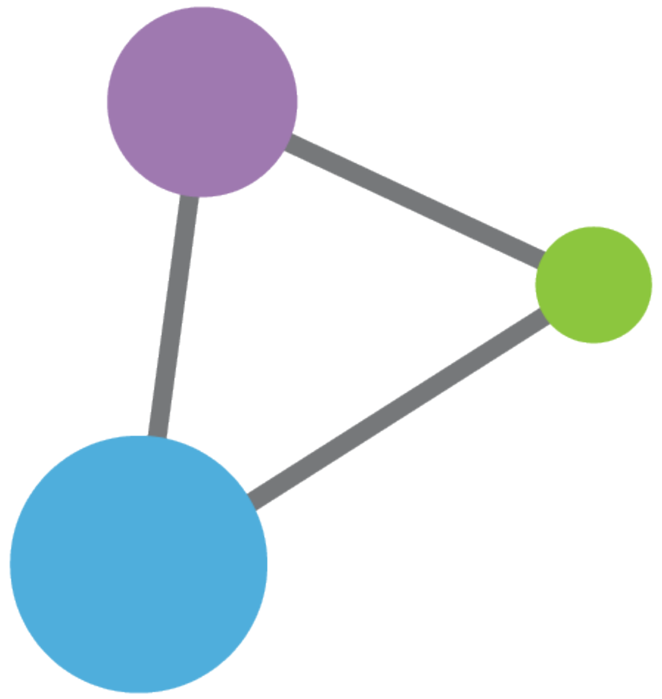
Centralized Genomic Knowledge Bases

European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network

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



Linkage of distributed and disparate datasets

Different Approaches to Data Sharing

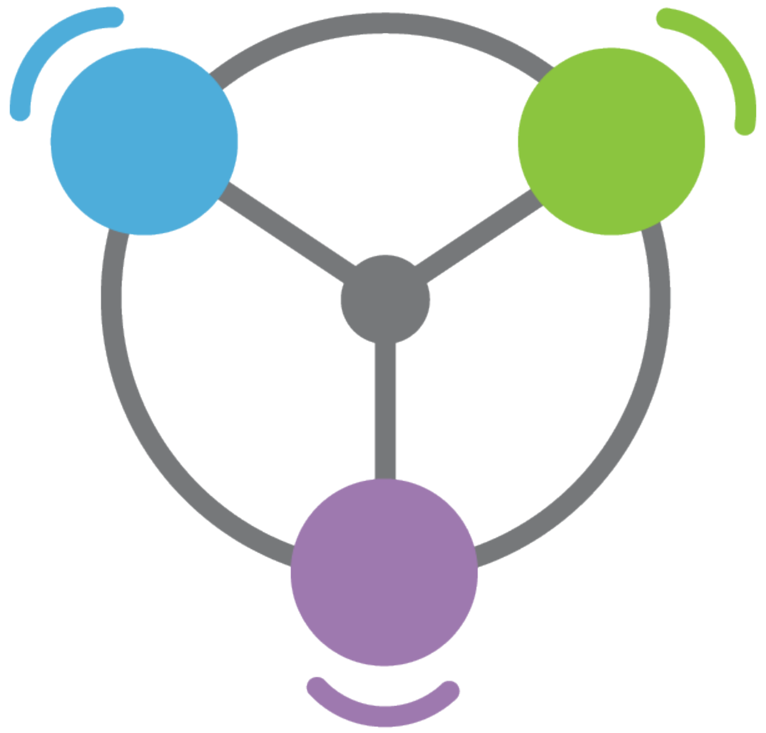
progenetix

Centralized Genomic Knowledge Bases

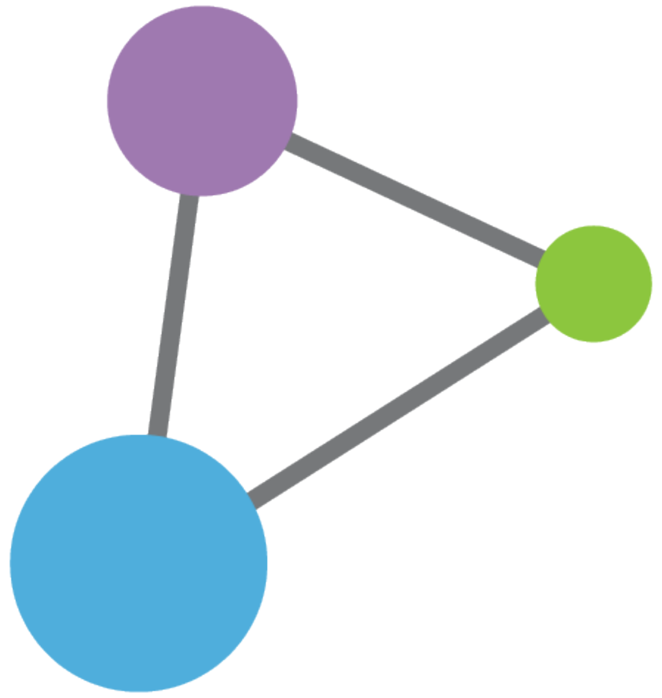
European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

SPHN
Swiss Personalized Health Network



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



Linkage of distributed and disparate datasets

Different Approaches to Data Sharing

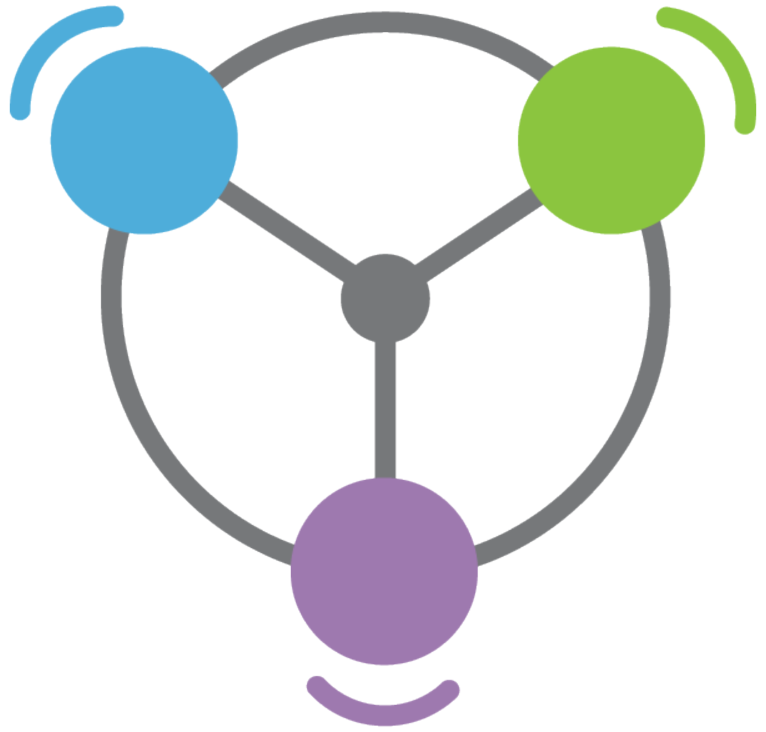
progenetix

Centralized Genomic Knowledge Bases

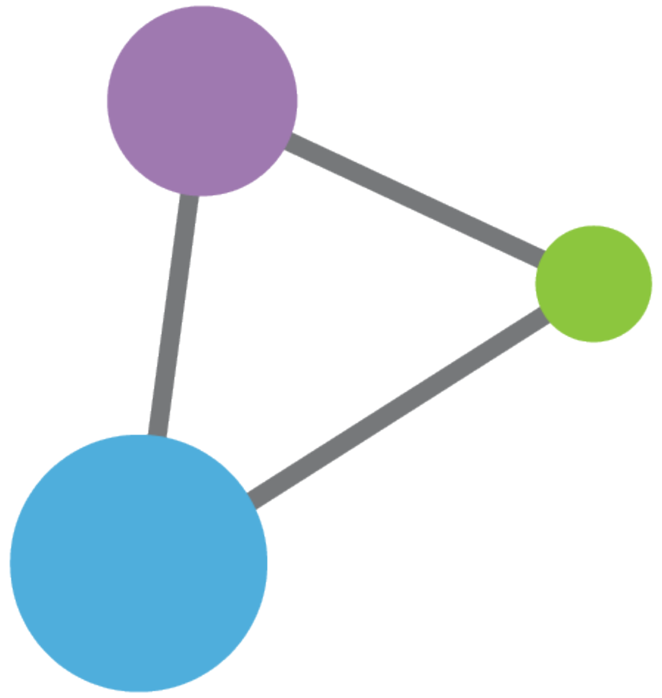
European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network



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



Linkage of distributed and disparate datasets

The EGA

European Genome-Phenome Archive

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*”)

EUROPEAN GENOME-PHENOME ARCHIVE

CRG Centre for Genomic Regulation

EMBL-EBI

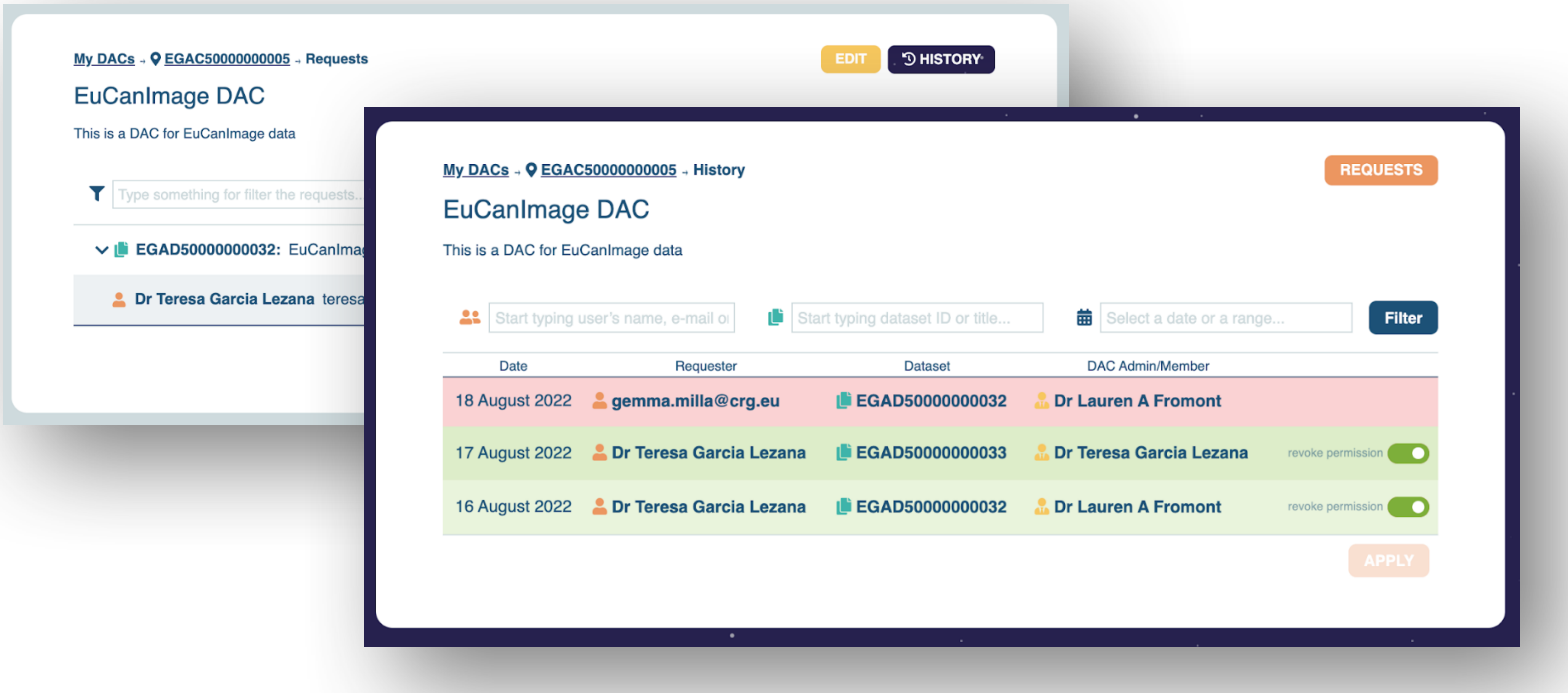
Global Alliance for Genomics & Health
DRIVER PROJECT

Slide: adapted from Jordi Rambla@ GA4GH 2023

The EGA

European Genome-Phenome Archive

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



My DACs › EGAC50000000005 › Requests

EDIT HISTORY

EuCanImage DAC

This is a DAC for EuCanImage data

Type something for filter the requests...

EGAD50000000032: EuCanIma

Dr Teresa Garcia Lezana teresa

My DACs › EGAC50000000005 › History

REQUESTS

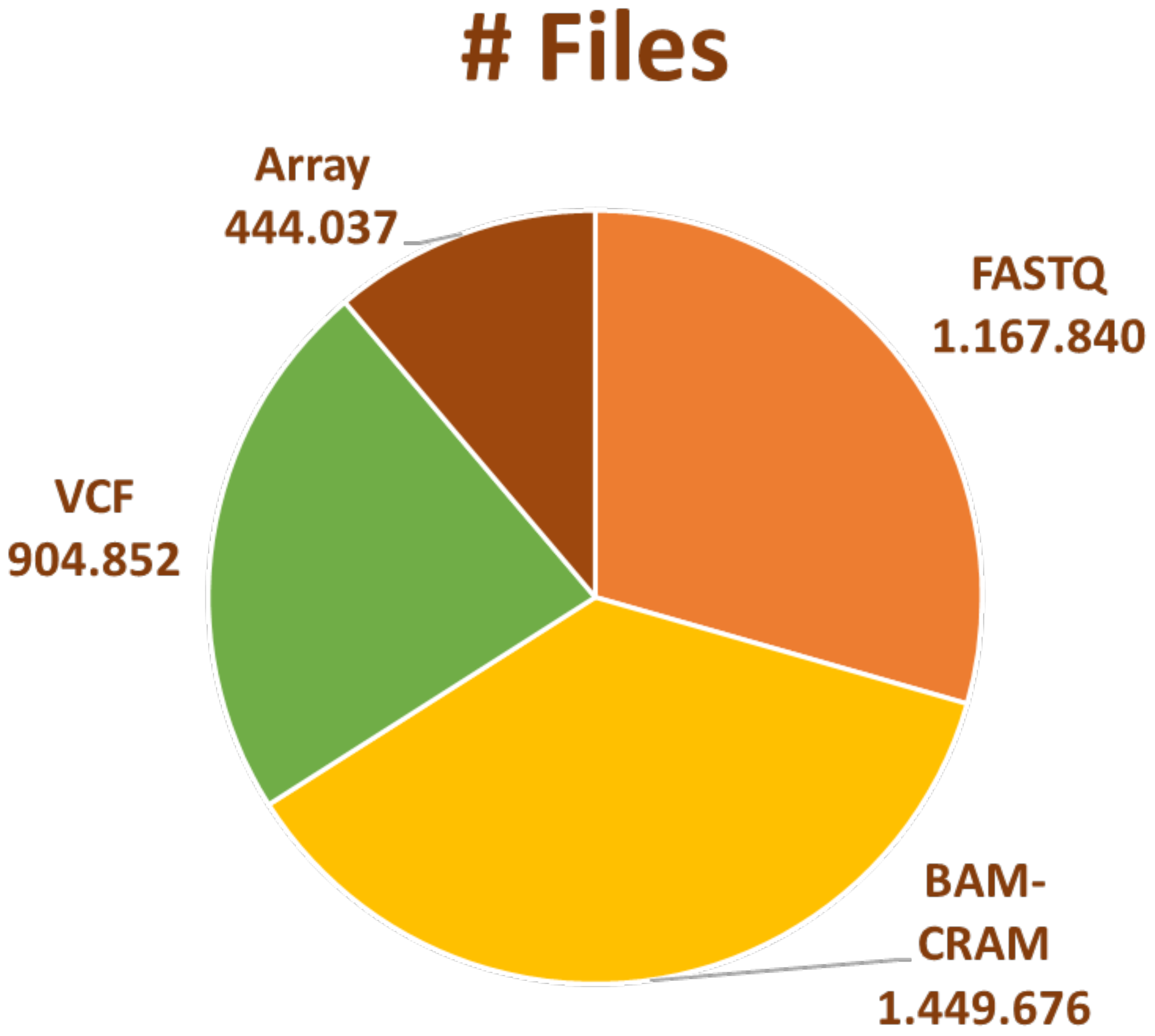
EuCanImage DAC

This is a DAC for EuCanImage data

Start typing user's name, e-mail o | Start typing dataset ID or title... | Select a date or a range... | Filter

Date	Requester	Dataset	DAC Admin/Member	
18 August 2022	gemma.milla@crg.eu	EGAD50000000032	Dr Lauren A Fromont	
17 August 2022	Dr Teresa Garcia Lezana	EGAD50000000033	Dr Teresa Garcia Lezana	revoke permission
16 August 2022	Dr Teresa Garcia Lezana	EGAD50000000032	Dr Lauren A Fromont	revoke permission

APPLY



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

Slide: adapted from Jordi Rambla@ GA4GH 2023

Different Approaches to Data Sharing

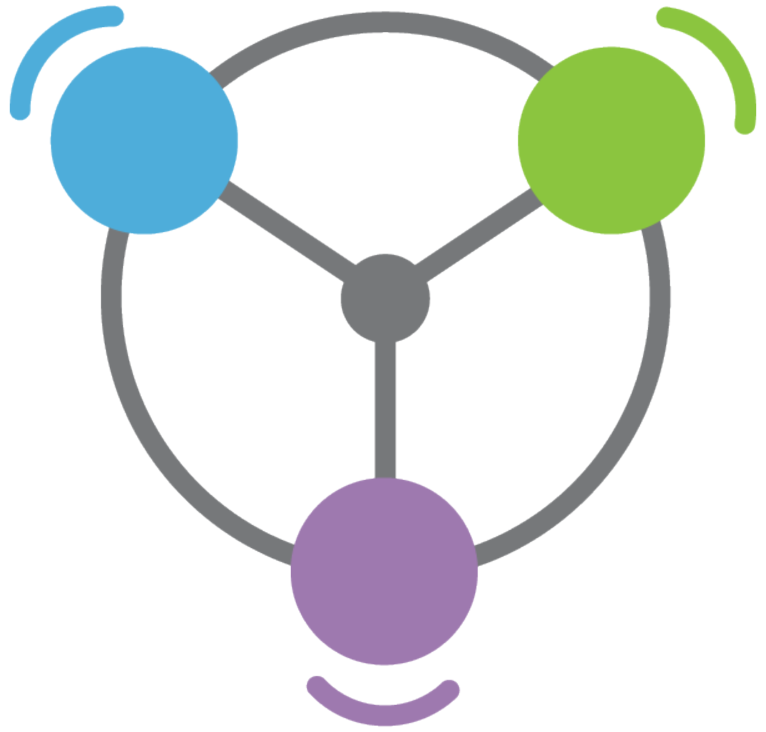
progenetix

Centralized Genomic Knowledge Bases

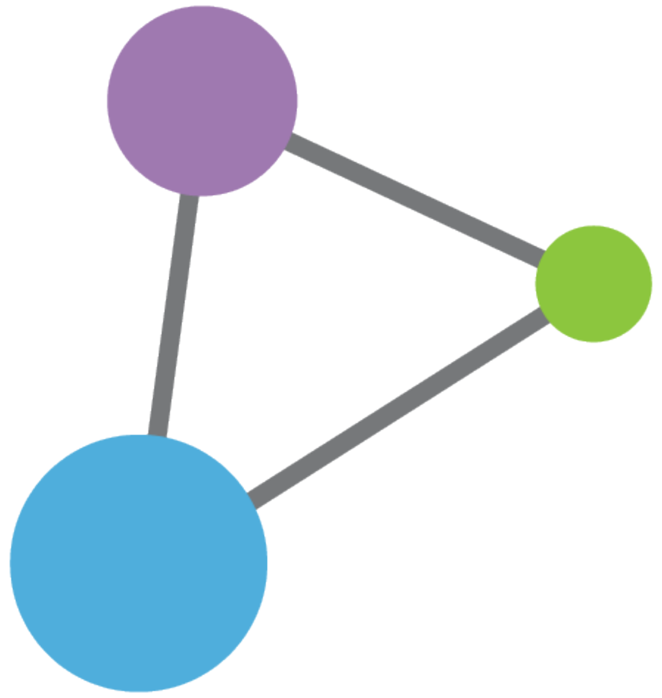
European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network



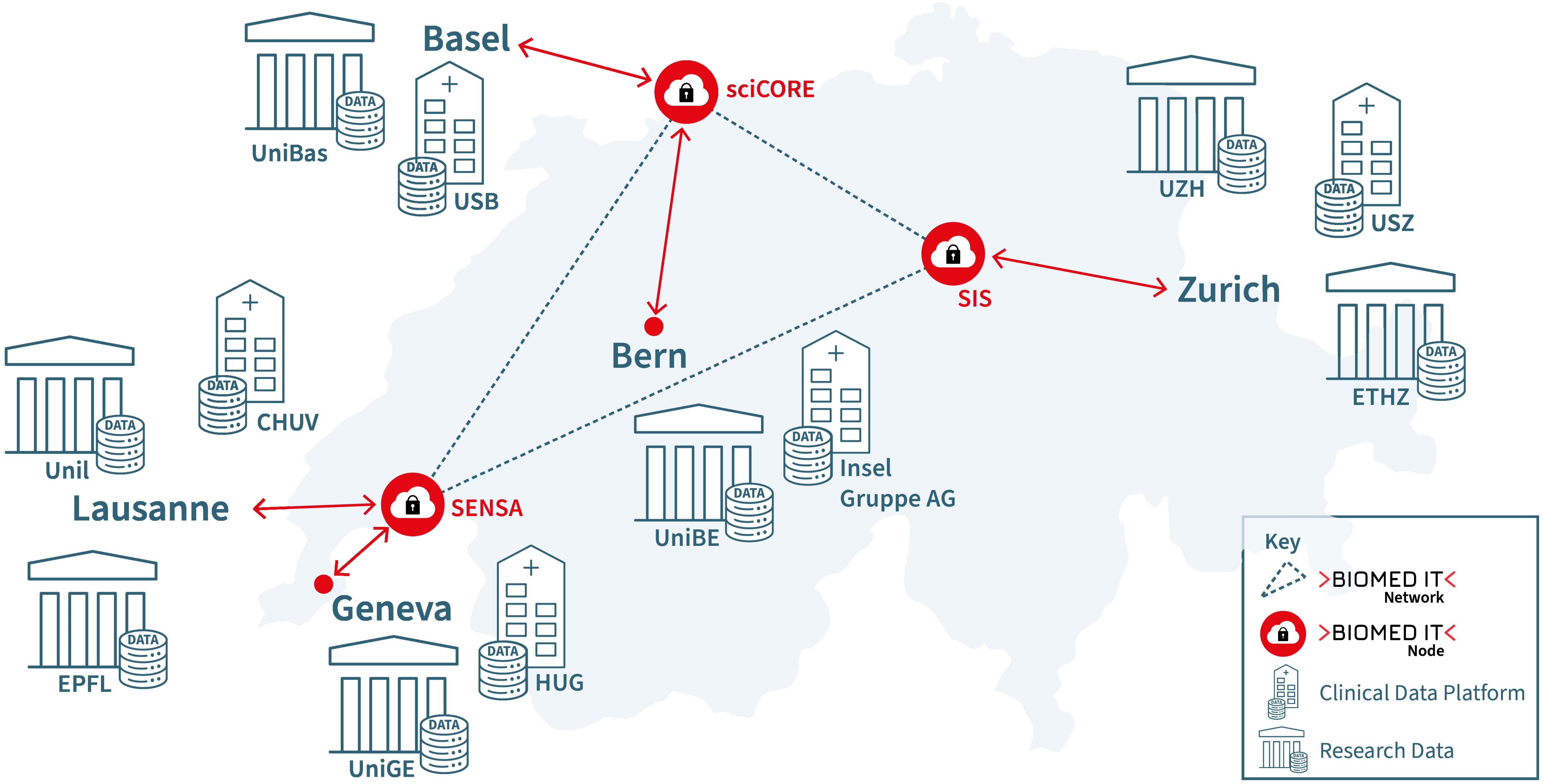
Hub and Spoke
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

The Swiss Personalized Health Network

SPHN Swiss Personalized Health Network



Strategic Focus Area
Personalized Health and Related Technologies

ehealthsuisse

THE LOOP ZURICH MEDICAL RESEARCH CENTER

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

Personalized Health Alliance Basel-Zurich

Swiss Biobanking Platform

SAKK

SCTO

SSPH+ Swiss School of Public Health

life sciences cluster basel

swissuniversities

Universitäre Medizin Schweiz
Médecine Universitaire Suisse

SIB **Personalized Health Informatics Group**

SPHN Data Coordination Center (DCC)
BioMedIT Network

University Hospital Basel

USZ Universitäts Spital Zürich

HUG Hôpitaux Universitaires Genève

CHUV Centre hospitalier universitaire vaudois

INSELSPITAL
Universitätsspital Bern
Hopital universitaire de Berne
Bern University Hospital

Different Approaches to Data Sharing

progenetix

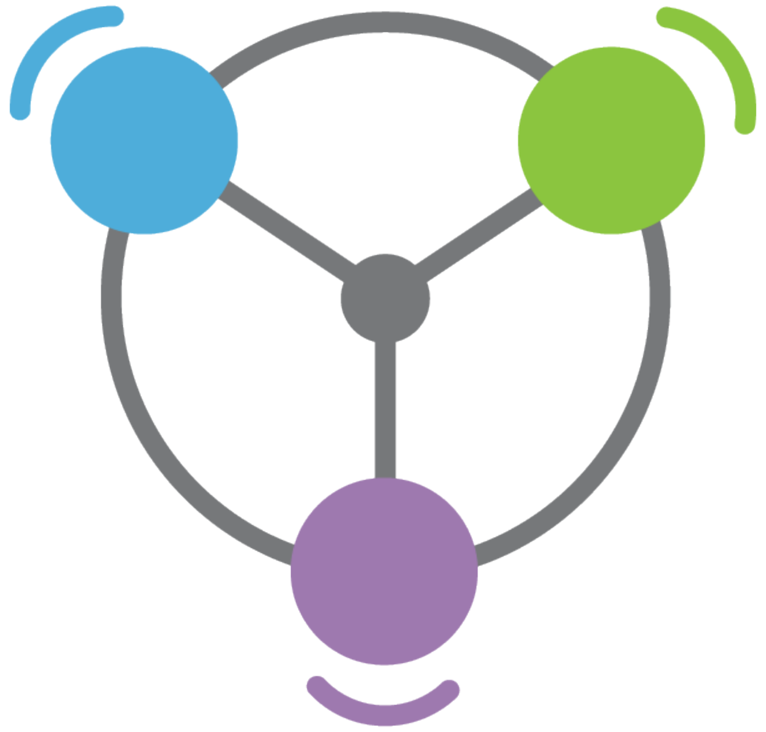
Centralized Genomic Knowledge Bases

European Genome-Phenome Archive

Data Commons

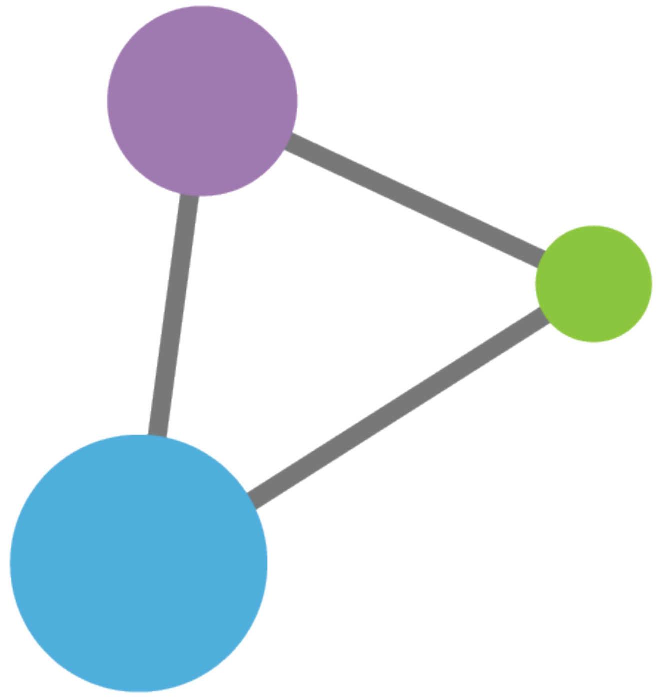
Trusted, controlled repository of multiple datasets

SPHN Swiss Personalized Health Network



Hub and Spoke

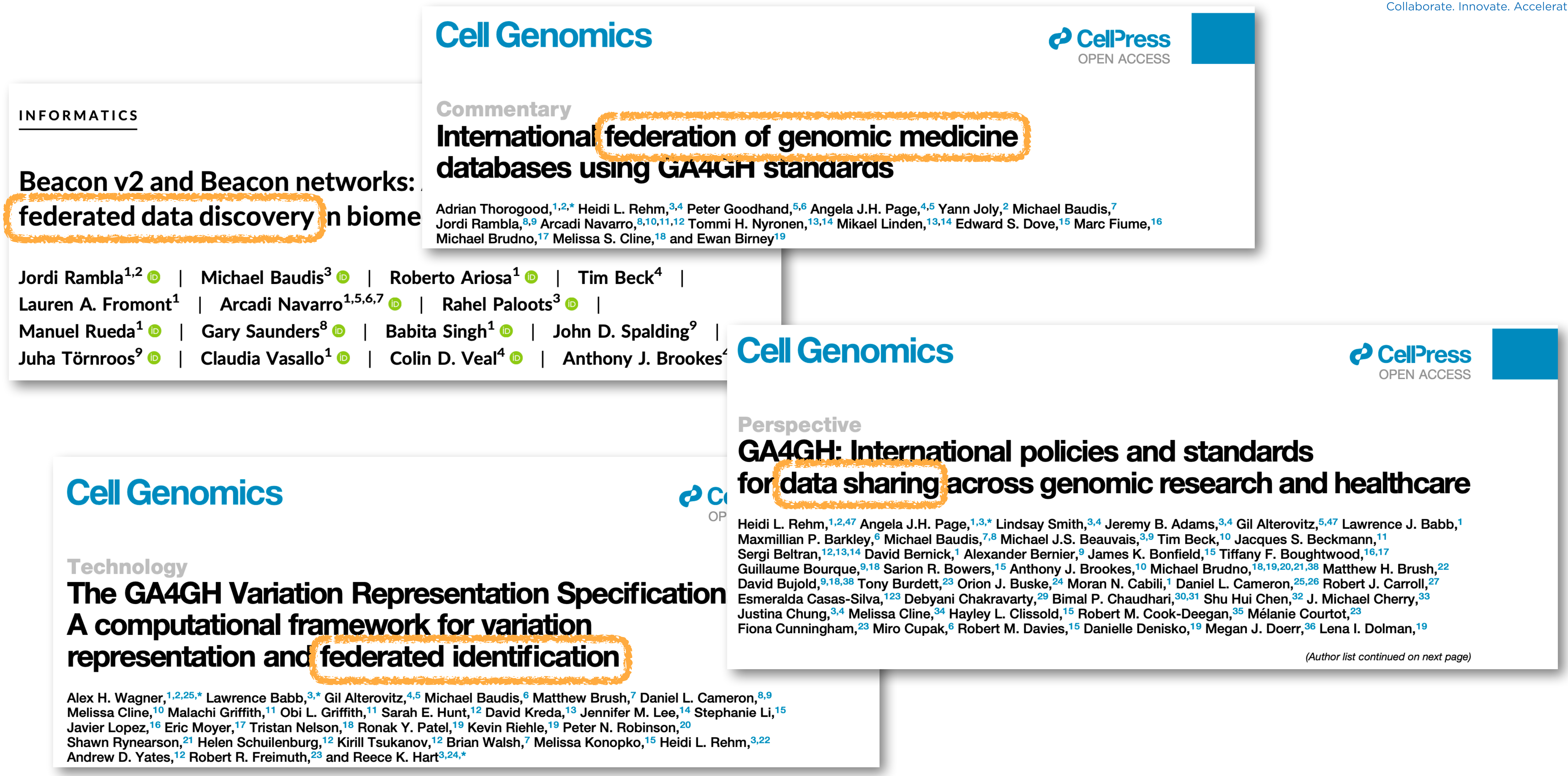
Common data elements, access, and usage rules



Linkage of distributed and disparate datasets

Federation

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



Cell Genomics

CellPress
OPEN ACCESS

Commentary

International federation of genomic medicine databases using GA4GH standards

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

INFORMATICS

Beacon v2 and Beacon networks:
federated data discovery n biome

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

Cell Genomics

CellPress
OPEN ACCESS

Perspective

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

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

(Author list continued on next page)

Cell Genomics

Technology

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

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

GENOMICS

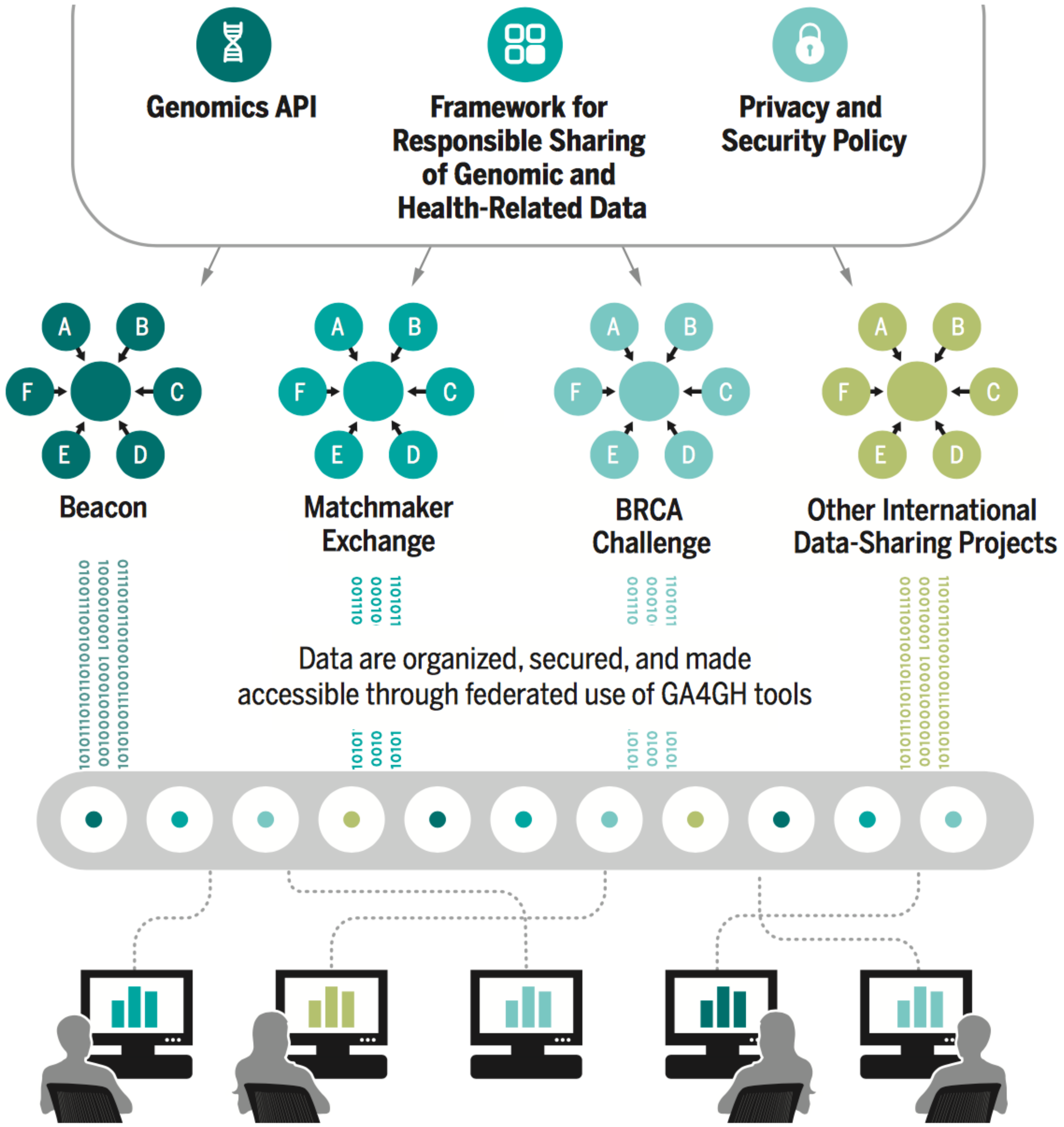
A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.


Genomics API

Framework for Responsible Sharing of Genomic and Health-Related Data

Privacy and Security Policy

Beacon

Matchmaker Exchange

BRCA Challenge

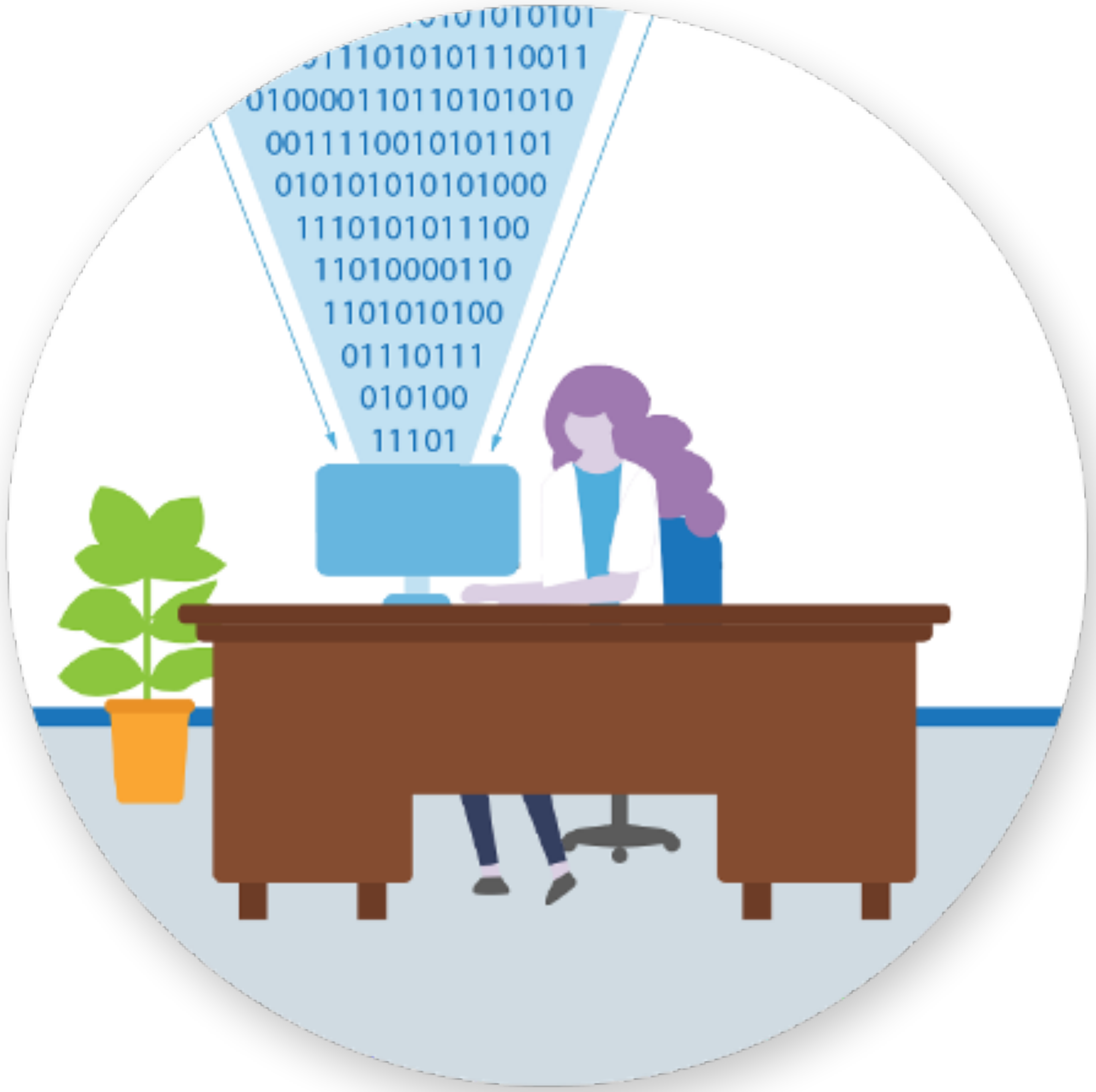
Other International Data-Sharing Projects

Data are organized, secured, and made accessible through federated use of GA4GH tools

Global Alliance for Genomics & Health

A New Paradigm for Data Sharing

FROM



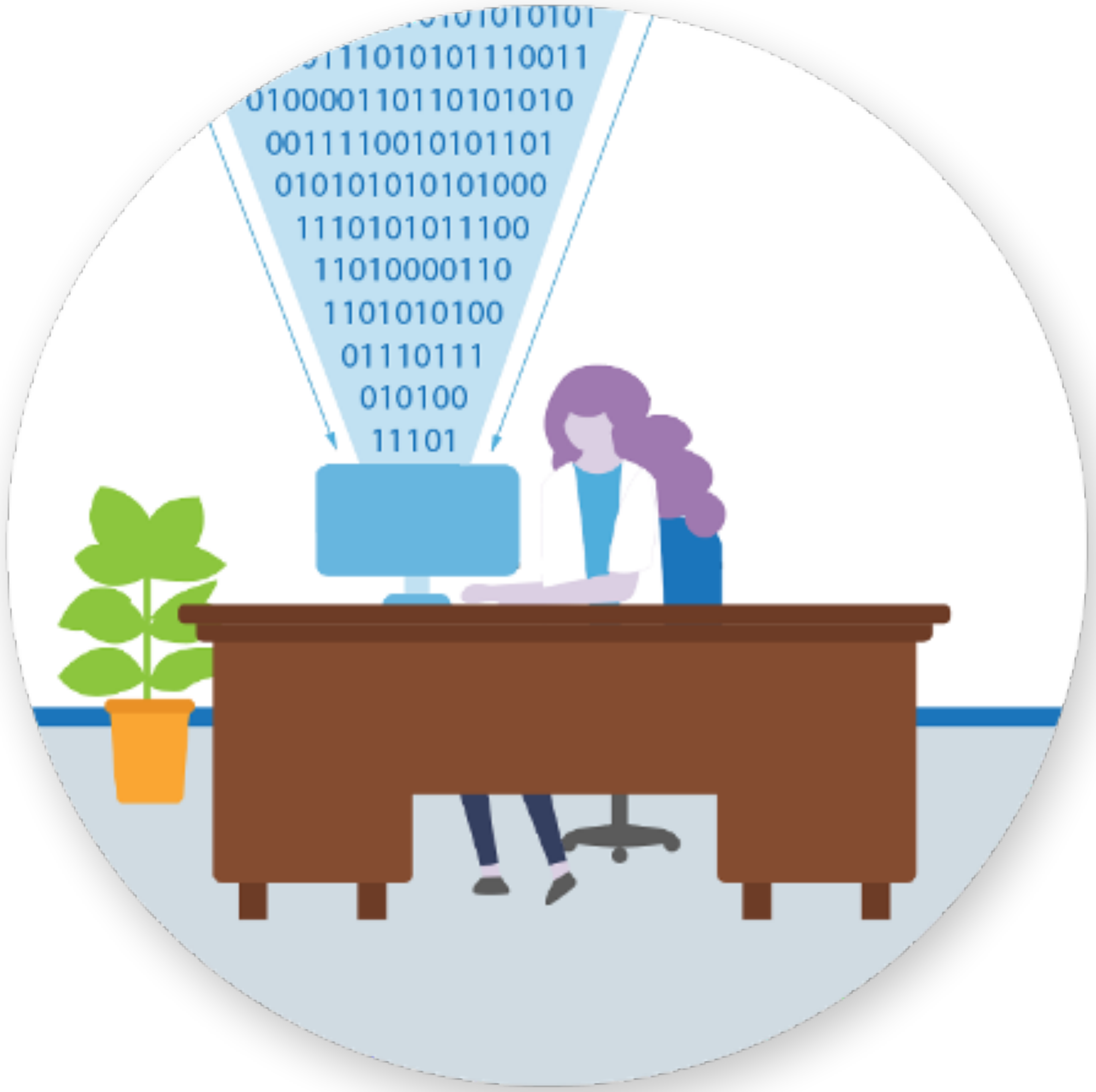
Data Copying

TO

Data Visiting

A New Paradigm for Data Sharing

FROM

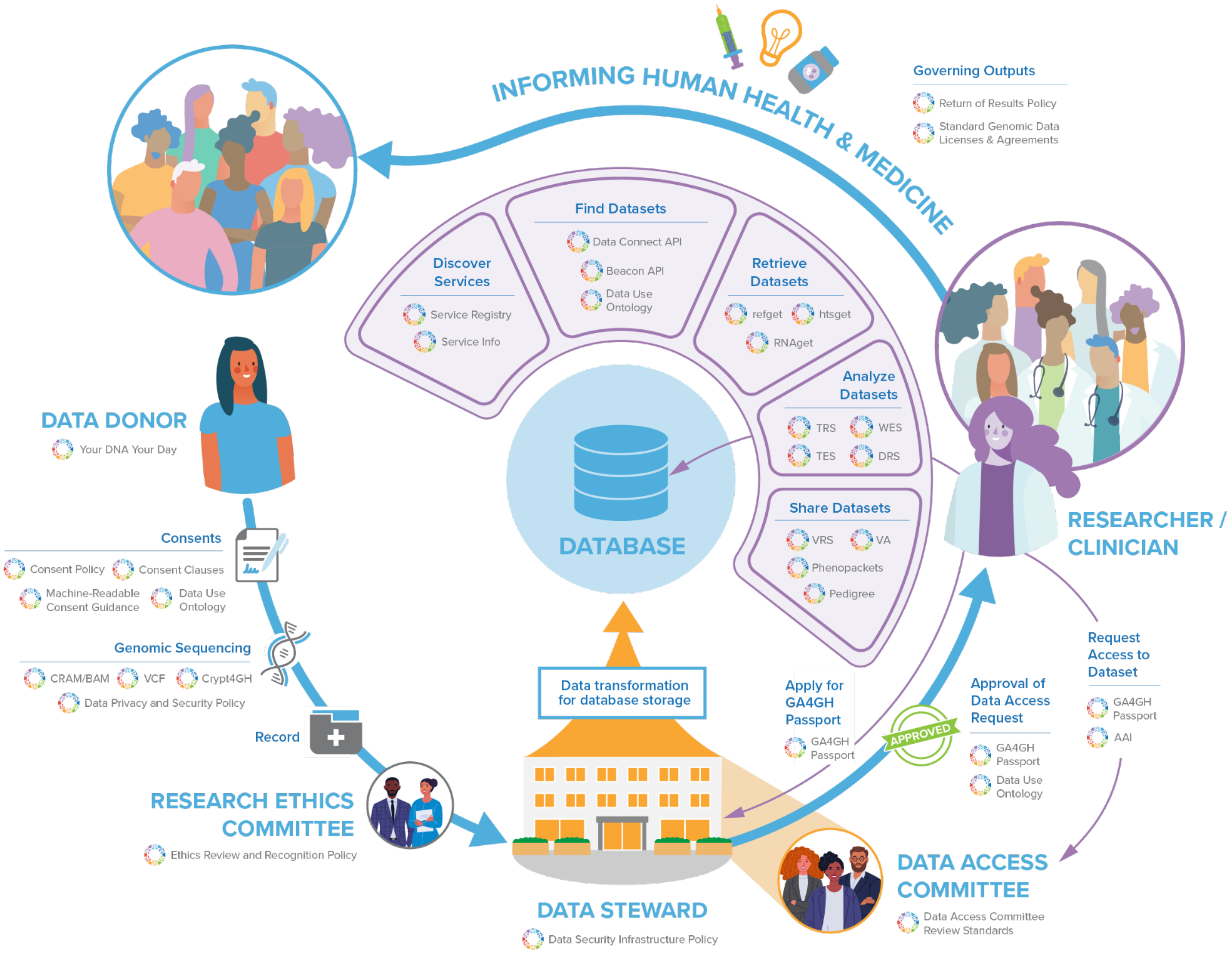


Data Copying

STANDARDS →

TO

Data Visiting



INFORMING HUMAN HEALTH & MEDICINE

Governing Outputs

- Return of Results Policy
- Standard Genomic Data Licenses & Agreements

DATA DONOR

- Your DNA Your Day

Consents

- Consent Policy
- Consent Clauses
- Machine-Readable Consent Guidance
- Data Use Ontology

Genomic Sequencing

- CRAM/BAM
- VCF
- Crypt4GH
- Data Privacy and Security Policy

Record

RESEARCH ETHICS COMMITTEE

- Ethics Review and Recognition Policy

DATABASE

Data transformation for database storage

DATA STEWARD

- Data Security Infrastructure Policy

Discover Services

- Service Registry
- Service Info

Find Datasets

- Data Connect API
- Beacon API
- Data Use Ontology

Retrieve Datasets

- refget
- htsget
- RNAget

Analyze Datasets

- TRS
- WES
- TES
- DRS

Share Datasets

- VRS
- VA
- Phenopackets
- Pedigree

RESEARCHER / CLINICIAN

Request Access to Dataset

- GA4GH Passport
- AAI

Approval of Data Access Request

- GA4GH Passport
- Data Use Ontology

APPROVED

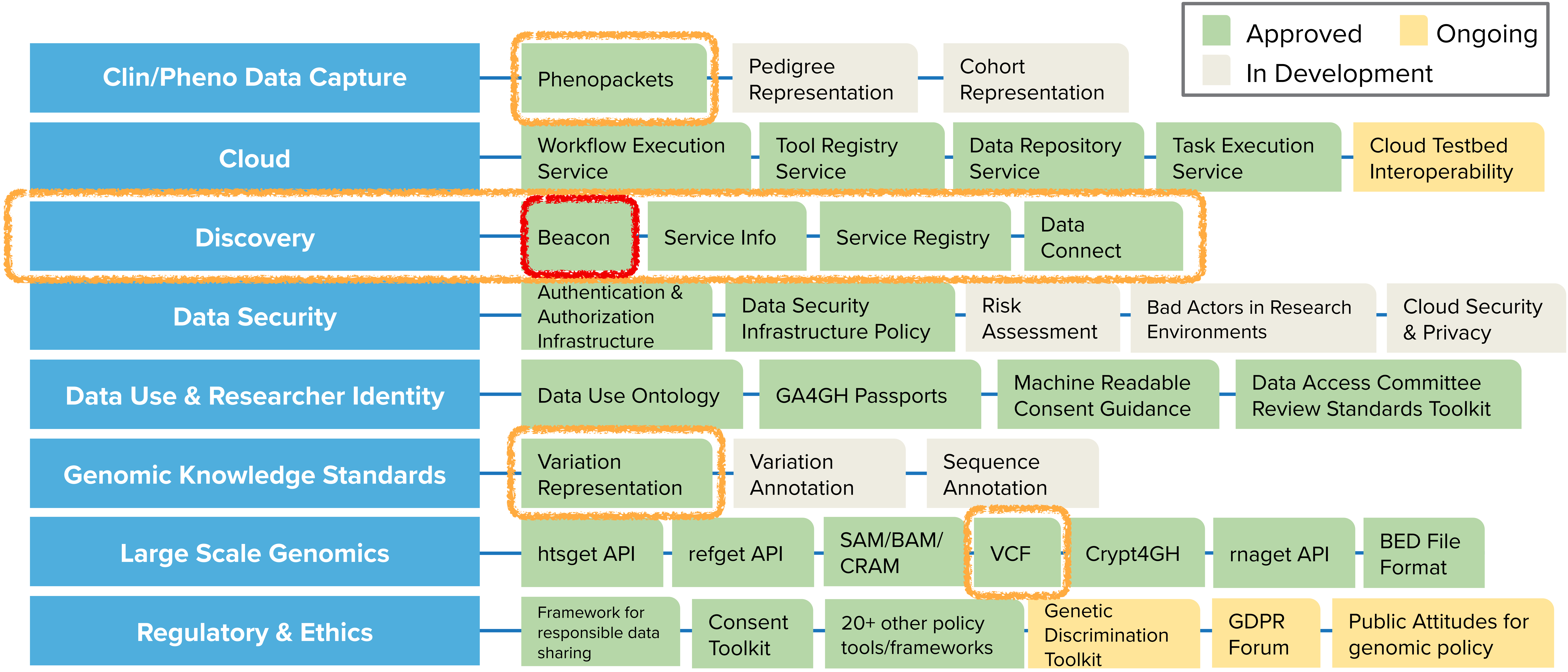
Apply for GA4GH Passport

- GA4GH Passport

DATA ACCESS COMMITTEE

- Data Access Committee Review Standards

Overview of GA4GH standards and frameworks



🟩 Approved 🟨 Ongoing
⬜ In Development

Clin/Pheno Data Capture	🟩 Phenopackets	⬜ Pedigree Representation	⬜ Cohort Representation				
Cloud	🟩 Workflow Execution Service	🟩 Tool Registry Service	🟩 Data Repository Service	🟩 Task Execution Service	🟨 Cloud Testbed Interoperability		
Discovery	🟩 Beacon	🟩 Service Info	🟩 Service Registry	🟩 Data Connect			
Data Security	🟩 Authentication & Authorization Infrastructure	🟩 Data Security Infrastructure Policy	⬜ Risk Assessment	⬜ Bad Actors in Research Environments	⬜ Cloud Security & Privacy		
Data Use & Researcher Identity	🟩 Data Use Ontology	🟩 GA4GH Passports	🟩 Machine Readable Consent Guidance	🟩 Data Access Committee Review Standards Toolkit			
Genomic Knowledge Standards	🟩 Variation Representation	⬜ Variation Annotation	⬜ Sequence Annotation				
Large Scale Genomics	🟩 htsget API	🟩 refget API	🟩 SAM/BAM/ CRAM	🟩 VCF	🟩 Crypt4GH	🟩 rnaget API	🟩 BED File Format
Regulatory & Ethics	🟩 Framework for responsible data sharing	🟩 Consent Toolkit	🟩 20+ other policy tools/frameworks	🟨 Genetic Discrimination Toolkit	🟨 GDPR Forum	🟨 Public Attitudes for genomic policy	

Phenopackets v2

Phenopackets is a standard schema for sharing phenotypic information.

Approved: June 24, 2021

Example Users

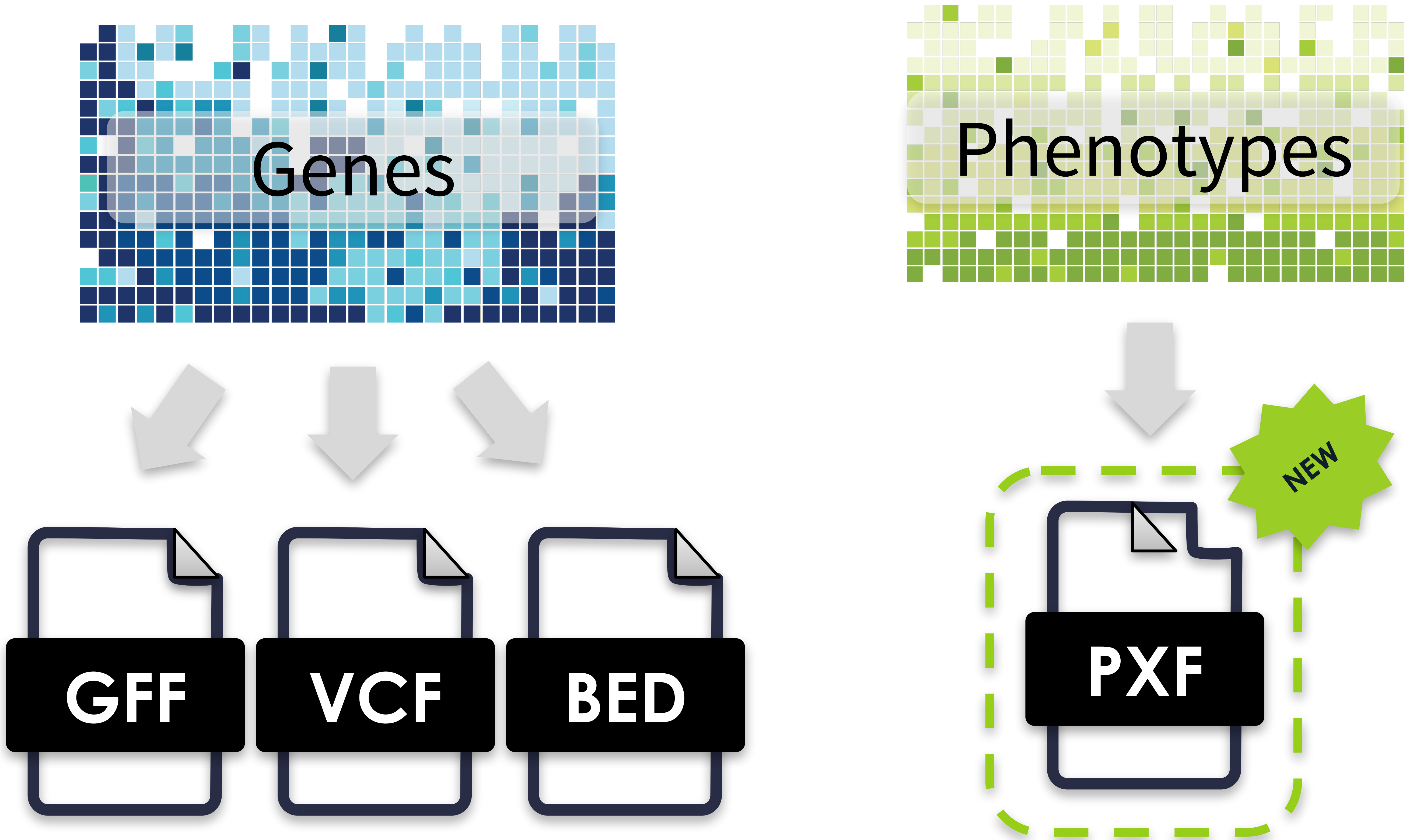
Cafe Variome

NIH

AMED

PHENOTIPS™

RD Connect



VCF/BCF

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

Software Libraries: htslib | htsjdk

Tools: Samtools | BCFtools

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

Genome Browsers: ENSEMBL | JBrowse | UCSC Genome Browser

Example Users

All of Us RESEARCH PROGRAM

BROAD INSTITUTE

elixir

EMBL-EBI

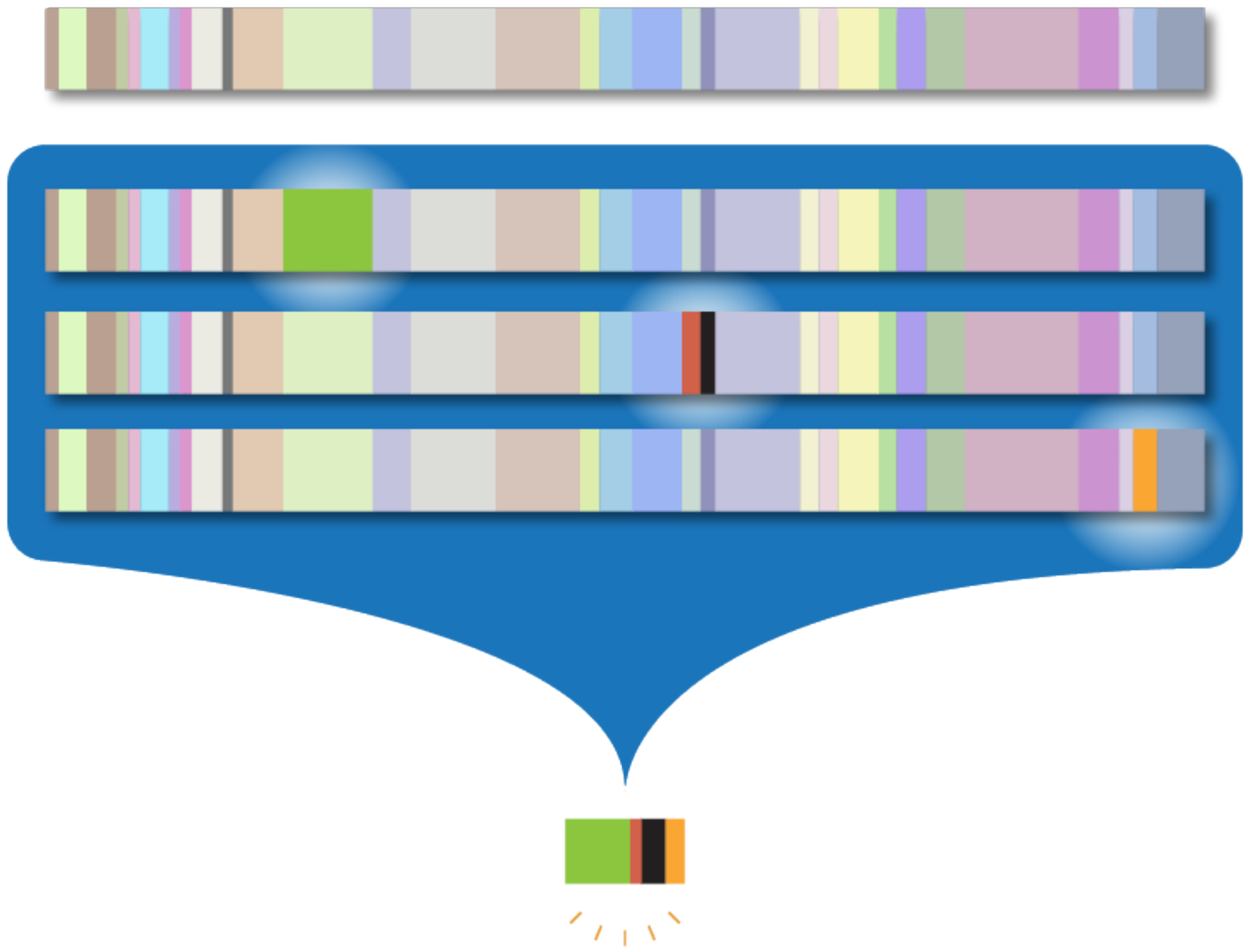
Genomics england

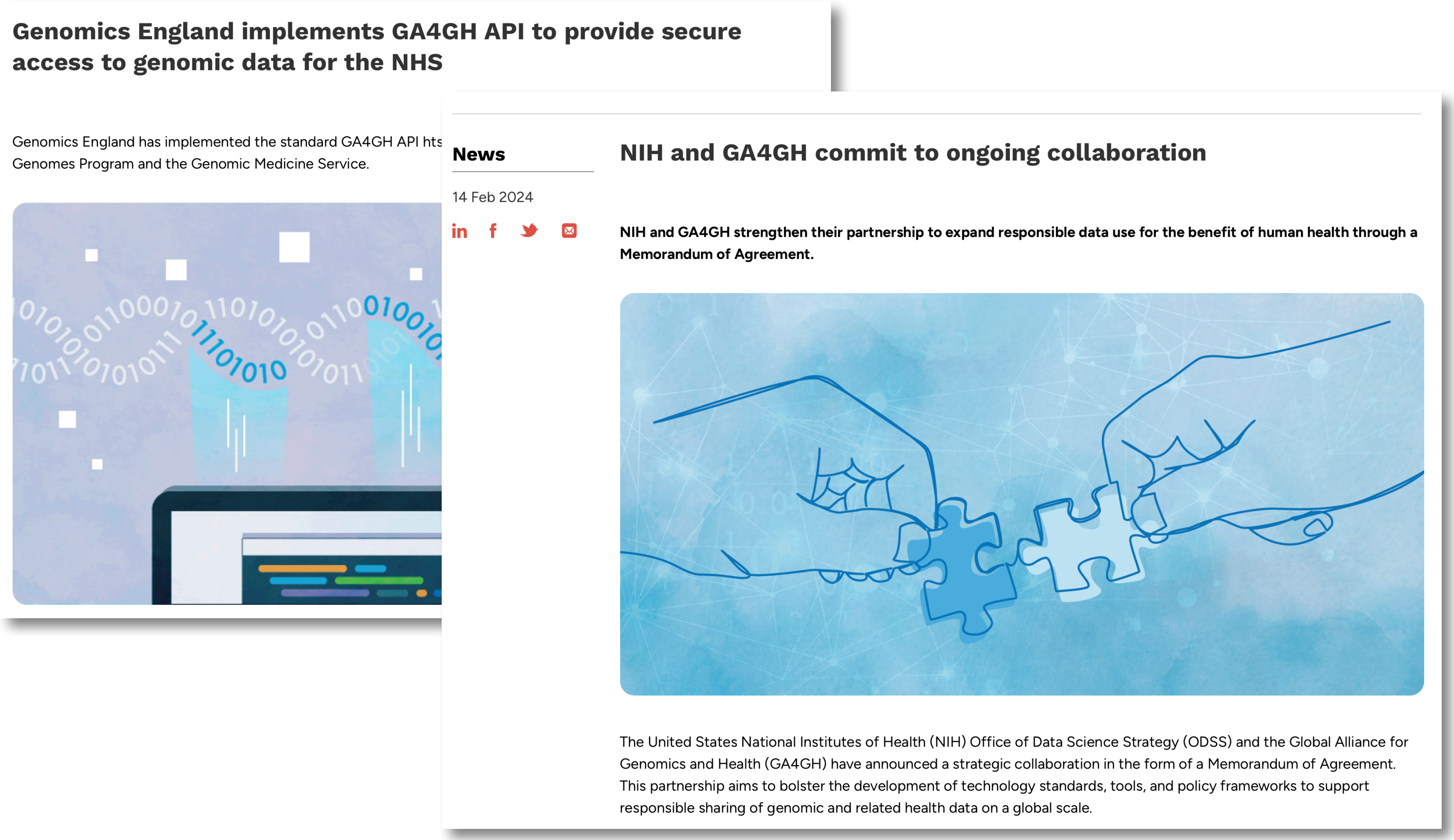
CRAM

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

1.5 million+ CRAM files
store more than **4 petabytes**
of compressed genomic data around the globe

CRAM compresses data by only storing the difference.





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

Genomics England has implemented the standard GA4GH API hts

Genomes Program and the Genomic Medicine Service.

News

14 Feb 2024

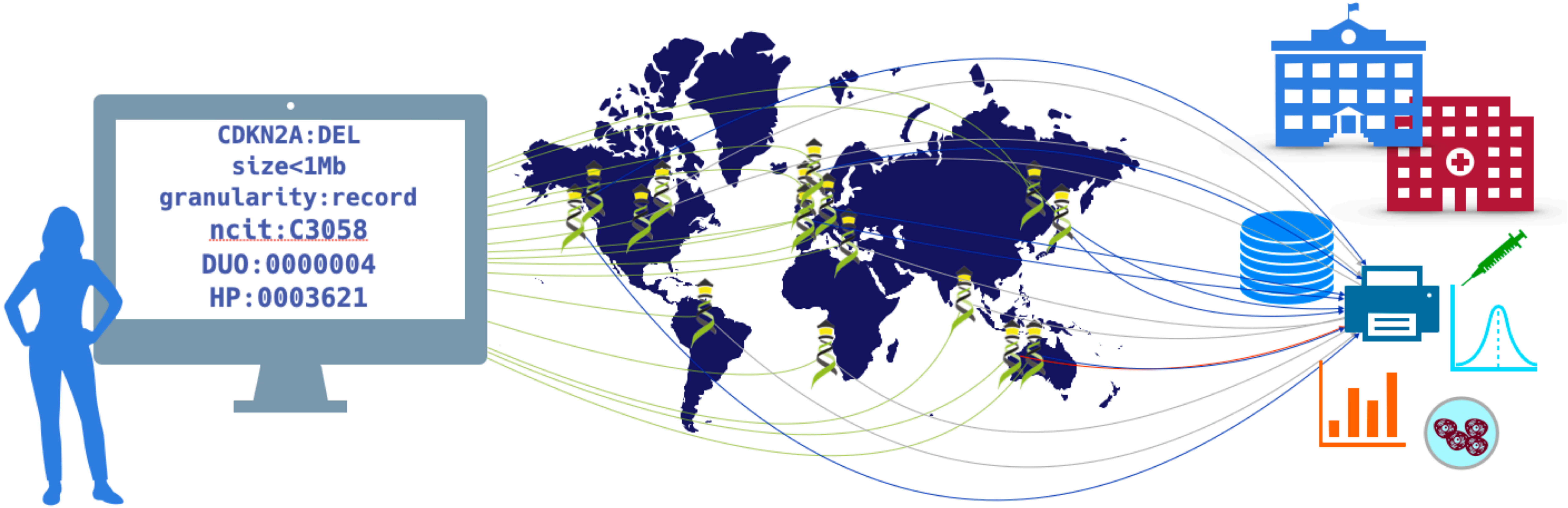
NIH and GA4GH commit to ongoing collaboration

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.

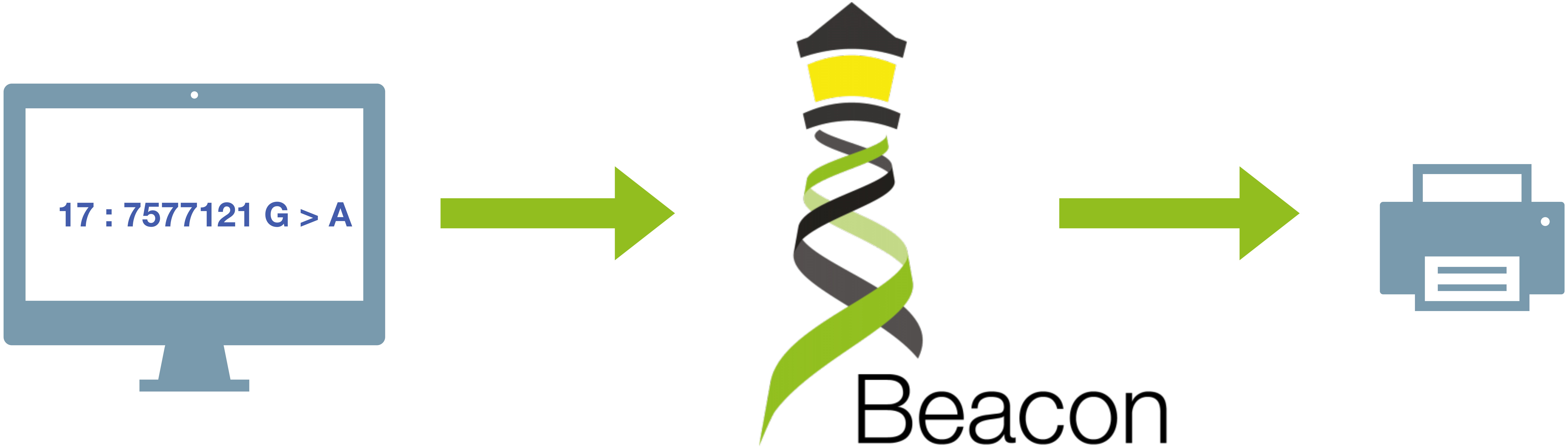
Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.



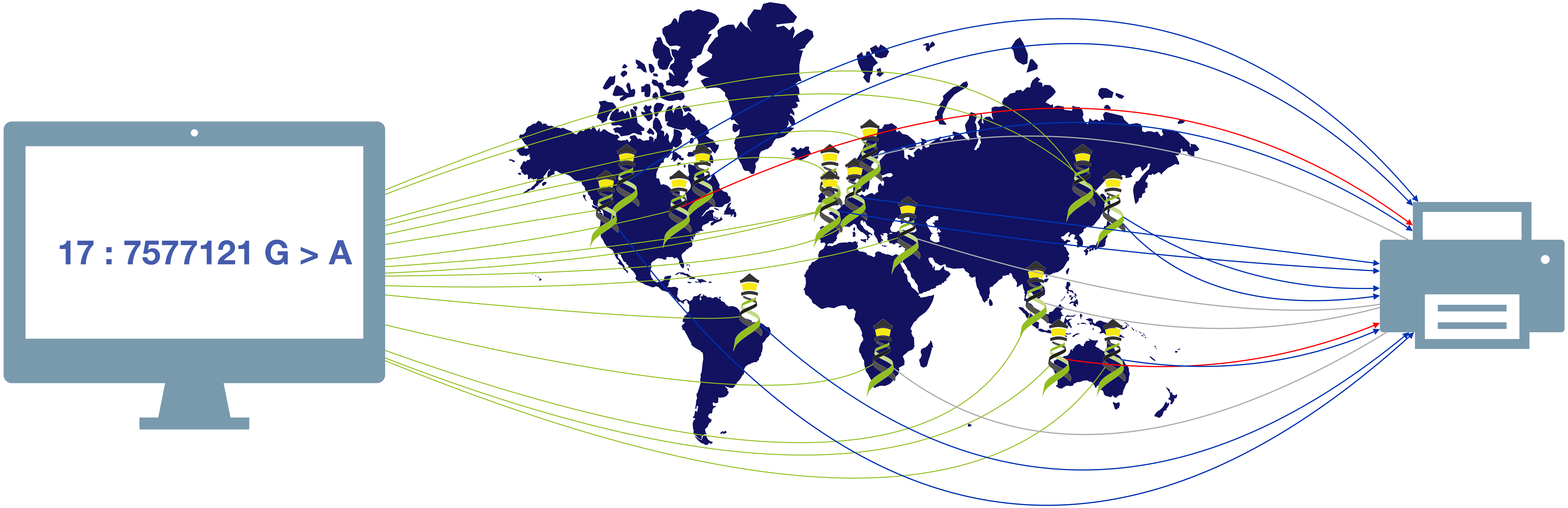
The GA4GH Beacon Protocol

Federating Genomic Discoveries



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

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



Have you seen this variant? It came up in my patient and we don't know if this is a common SNP or worth following up.

A Beacon network federates *genome variant queries* across databases that support the ***Beacon API***

Here: The variant has been found in **few** resources, and those are from **disease** specific **collections**.

Global Alliance “Beacon” - Jim Ostell, NCBI, March 7, 2014

Global Alliance for Genomics & Health

Introduction

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

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

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

Utility

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

A number of more useful first versions have been suggested.

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

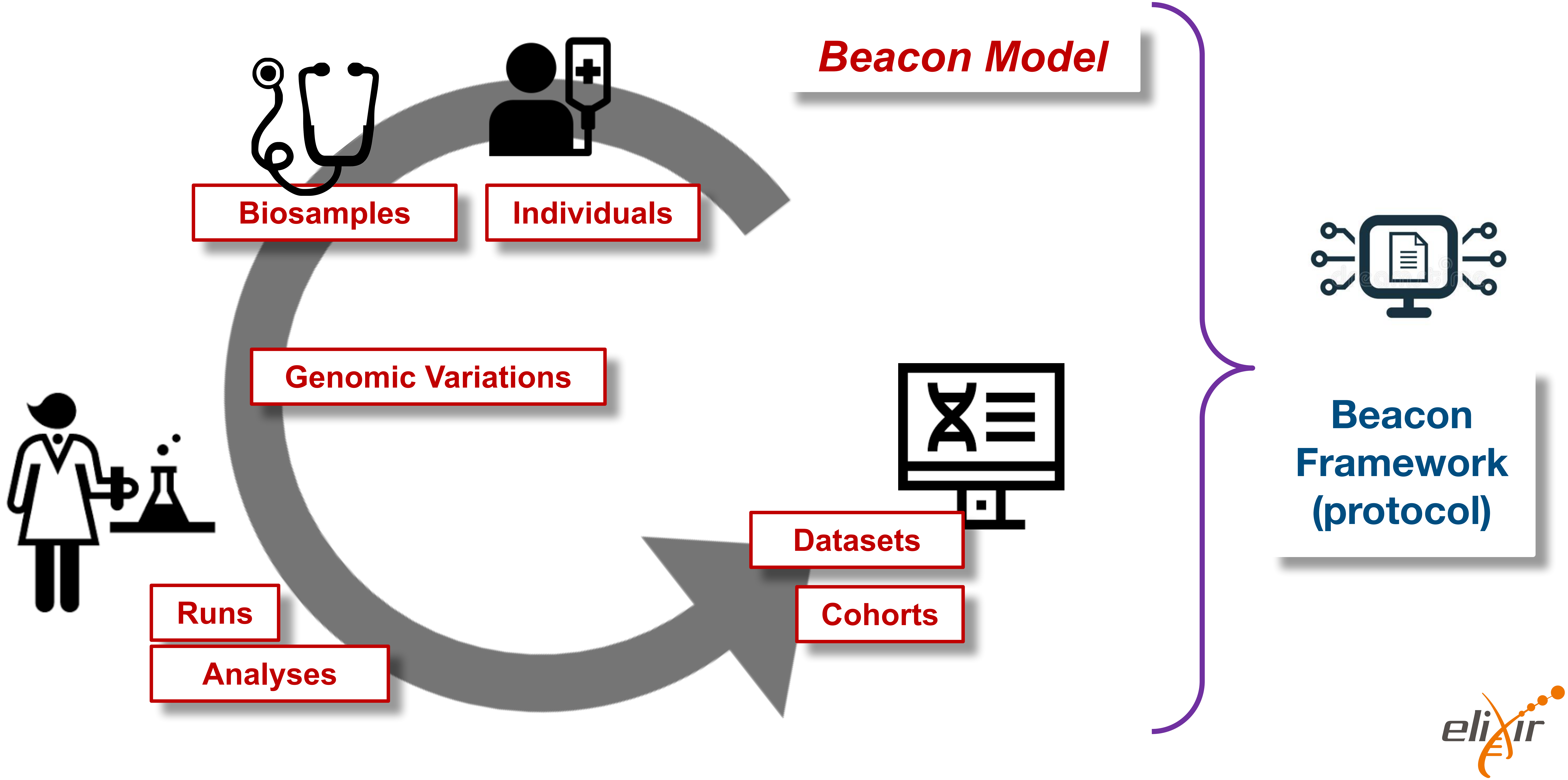
"I would personally recommend all those be held for **version 2**, when the beacon becomes a service."
Jim Ostell, 2014

Implementation

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

Beacon v2

docs.genomebeacons.org



Beacon Model

Biosamples

Individuals

Genomic Variations

Runs

Analyses

Datasets

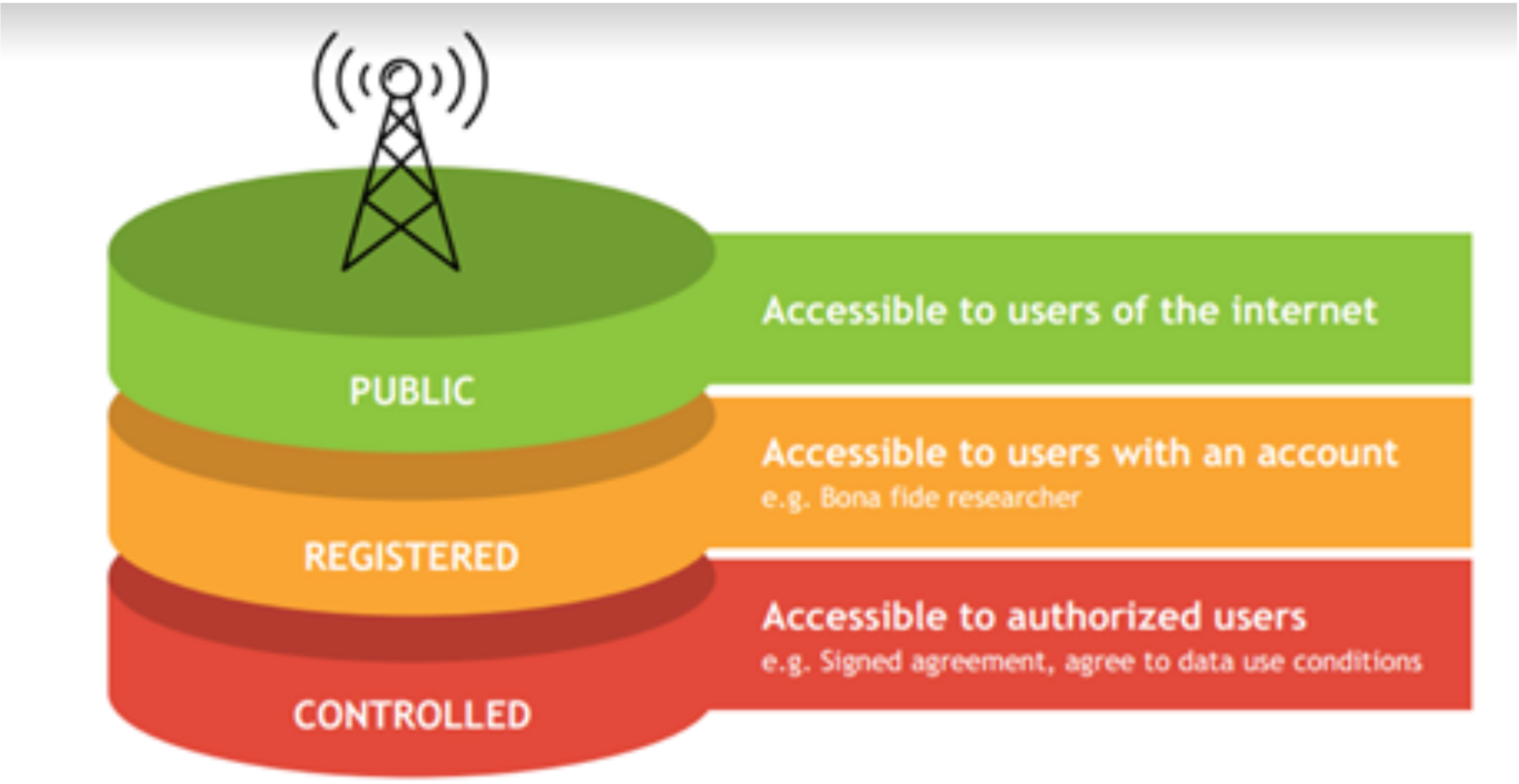
Cohorts

Beacon Framework (protocol)

Beacon API v2

The Beacon API can be implemented as a web-accessible service that users may query for information about a specific allele.

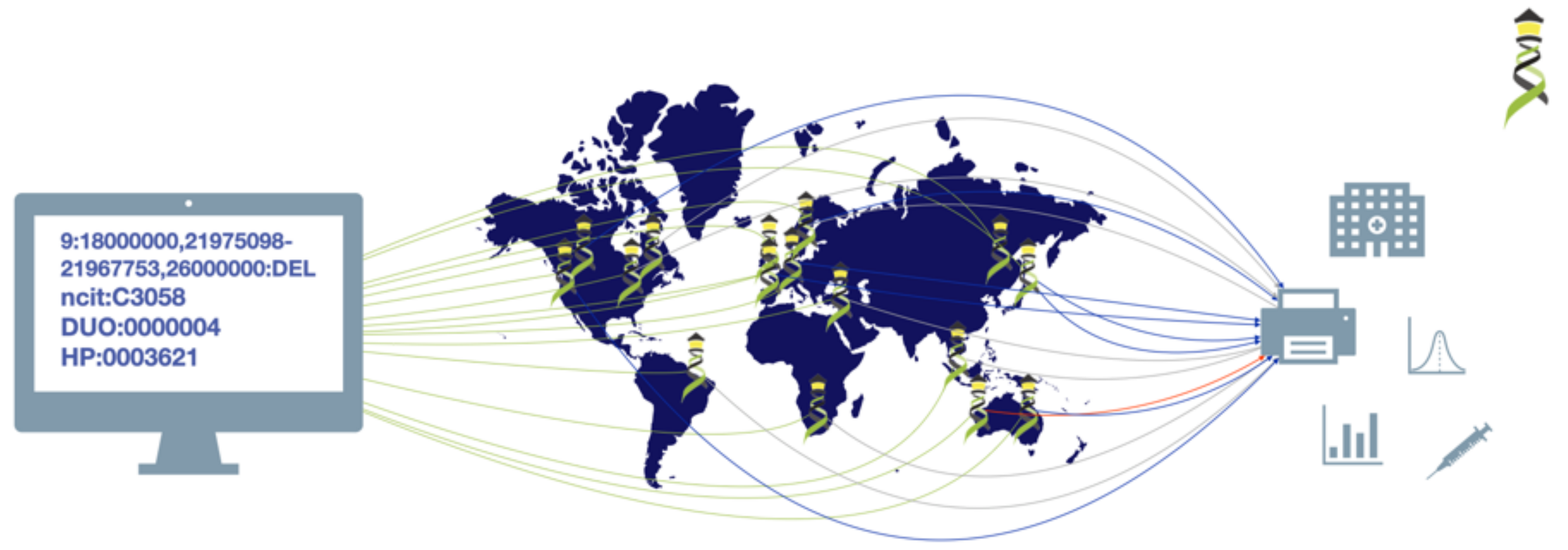
Approved: April 21, 2022



PUBLIC – Accessible to users of the internet

REGISTERED – Accessible to users with an account
e.g. Bona fide researcher

CONTROLLED – Accessible to authorized users
e.g. Signed agreement, agree to data use conditions



9:18000000,21975098-21967753,26000000:DEL
ncit:C3058
DUO:0000004
HP:0003621

Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

Beacon v2 API

The Beacon API v2 proposal opens the way for the design of a simple but powerful "**genomics API**".

Example Users

EMBL-EBI

Australian Genomics

SciLifeLab

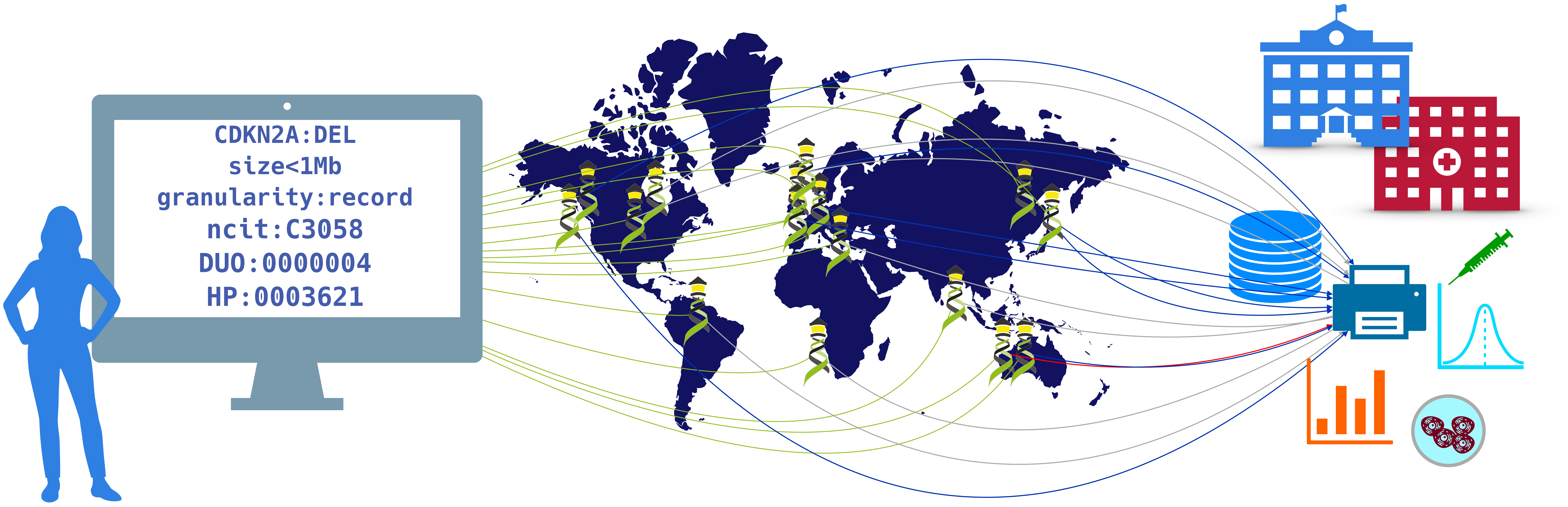
elixir

University of California Santa Cruz

Broad Institute

European Genome-Phenome Archive

International Cancer Genome Consortium



CDKN2A:DEL
size<1Mb
granularity:record
ncit:C3058
DUO:0000004
HP:0003621

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

Universität Zürich^UZH

Global Alliance for Genomics & Health

Progenetix and GA4GH Beacon

Implementation driven development of a GA4GH standard

progenetix

	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		• ELIXIR starts Beacon project support
2016	• Beacon v0.3 release • work on queries for structural variants (brackets for fuzzy start and end parameters...)	• Beacon+ concept implemented on progenetix.org • concepts from GA4GH Metadata (ontologies…) • entity-scoped query parameters ("individual.age")	
2017	• OpenAPI implementation • integrating CNV parameters (e.g. "startMin, statMax")	• Beacon+ demos "handover" concept	• GA4GH re-structuring (workstreams...) • Beacon part of Discovery WS
2018	• Beacon v0.4 release in January; feature release for GA4GH approval process • GA4GH Beacon v1 approved at Oct plenary		• new Beacon website (March)
2019	• ELIXIR Beacon Network	• Beacon hackathon Stockholm; settling on "filters" • Barcelona goes Zurich developers meeting • Beacon API v2 Kick off • adopting "handover" concept	• Beacon publication at Nature Biotechnology
2020		• "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	• Phenopackets v2 approved
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	• *docs.genomebeacons.org*

Progenetix & Beacon

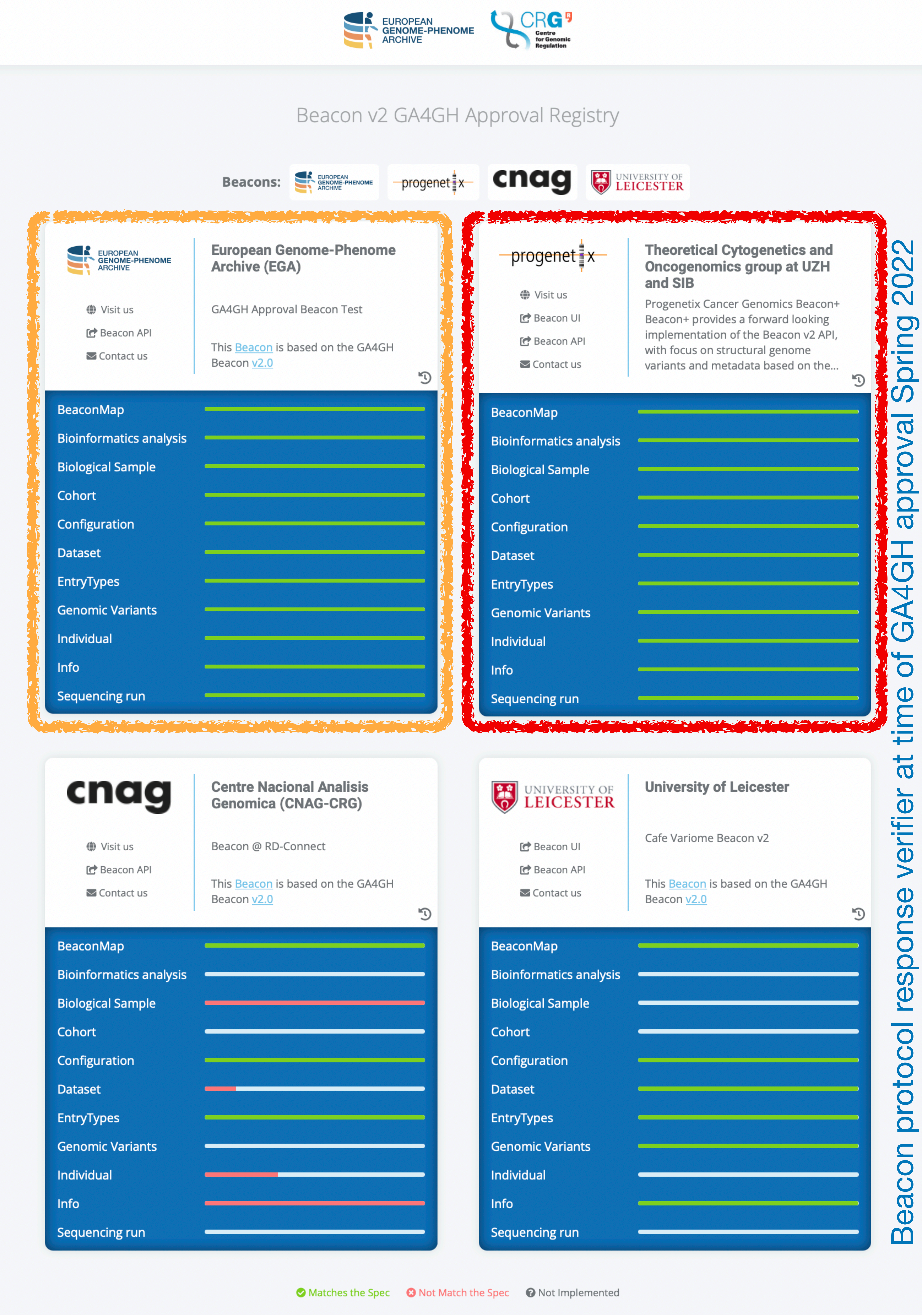
Implementation driven standards development

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

Global Alliance for Genomics & Health

progenetix

elixir



EUROPEAN GENOME-PHENOME ARCHIVE

CRG Centre for Genomic Regulation

Beacon v2 GA4GH Approval Registry

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

European Genome-Phenome Archive (EGA)

Visit us · Beacon API · Contact us

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

Visit us · Beacon UI · Beacon API · Contact us

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...

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

Centre Nacional Analisis Genomica (CNAG-CRG)

Visit us · Beacon API · Contact us

Beacon @ RD-Connect

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	⬜

University of Leicester

Beacon UI · Beacon API · Contact us

Cafe Variome Beacon v2

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	⬜

✓ Matches the Spec · ✗ Not Match the Spec · ? Not Implemented

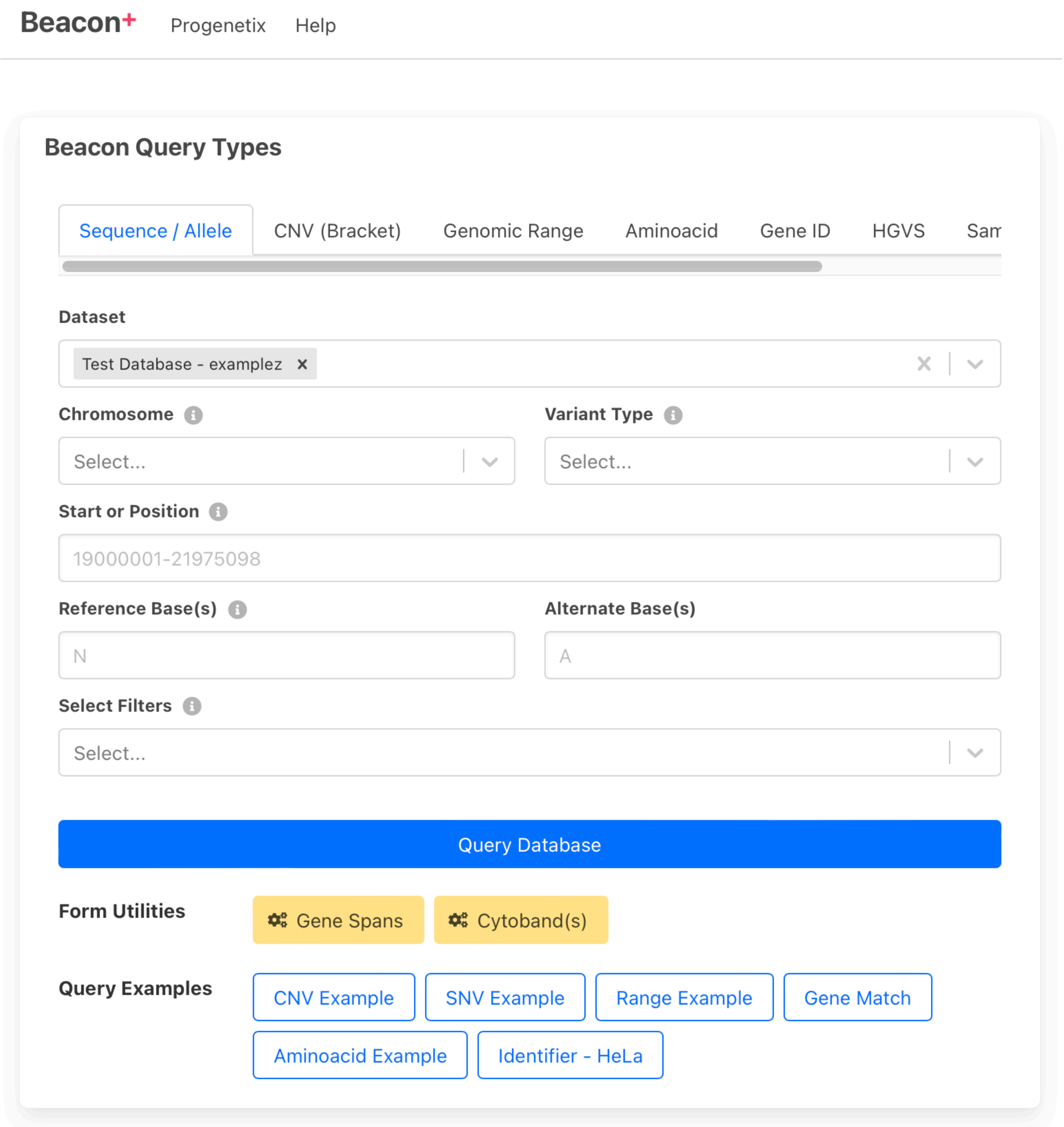
Beacon protocol response verifier at time of GA4GH approval Spring 2022

Beacon Queries

Implementation of Current Options

- (so far) the Beacon model does not define explicit query types
- disambiguation of parameters is left to implementers
- implicit query types:
  - ➡ allele/sequence query
  - ➡ range query, w/ or w/o additional parameters
  - ➡ bracket query (e.g. sized CNVs)
  - ➡ aminoacid, HGVS, gene

beaconplus.progenetix.org



Beacon Query Types

Dataset: Test Database - examplez

Chromosome: Select...

Variant Type: Select...

Start or Position: 19000001-21975098

Reference Base(s): N

Alternate Base(s): A

Select Filters: Select...

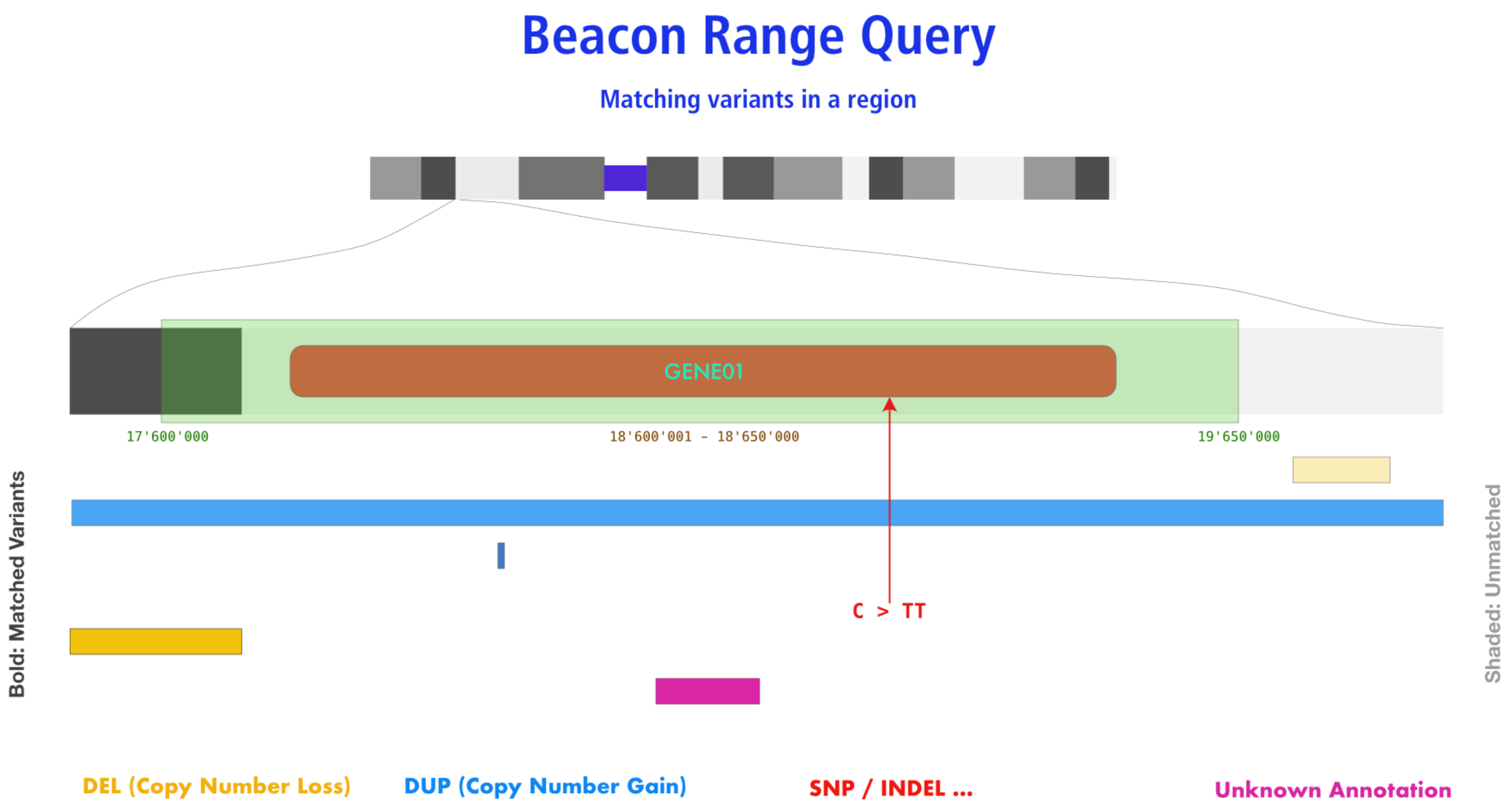
Form Utilities

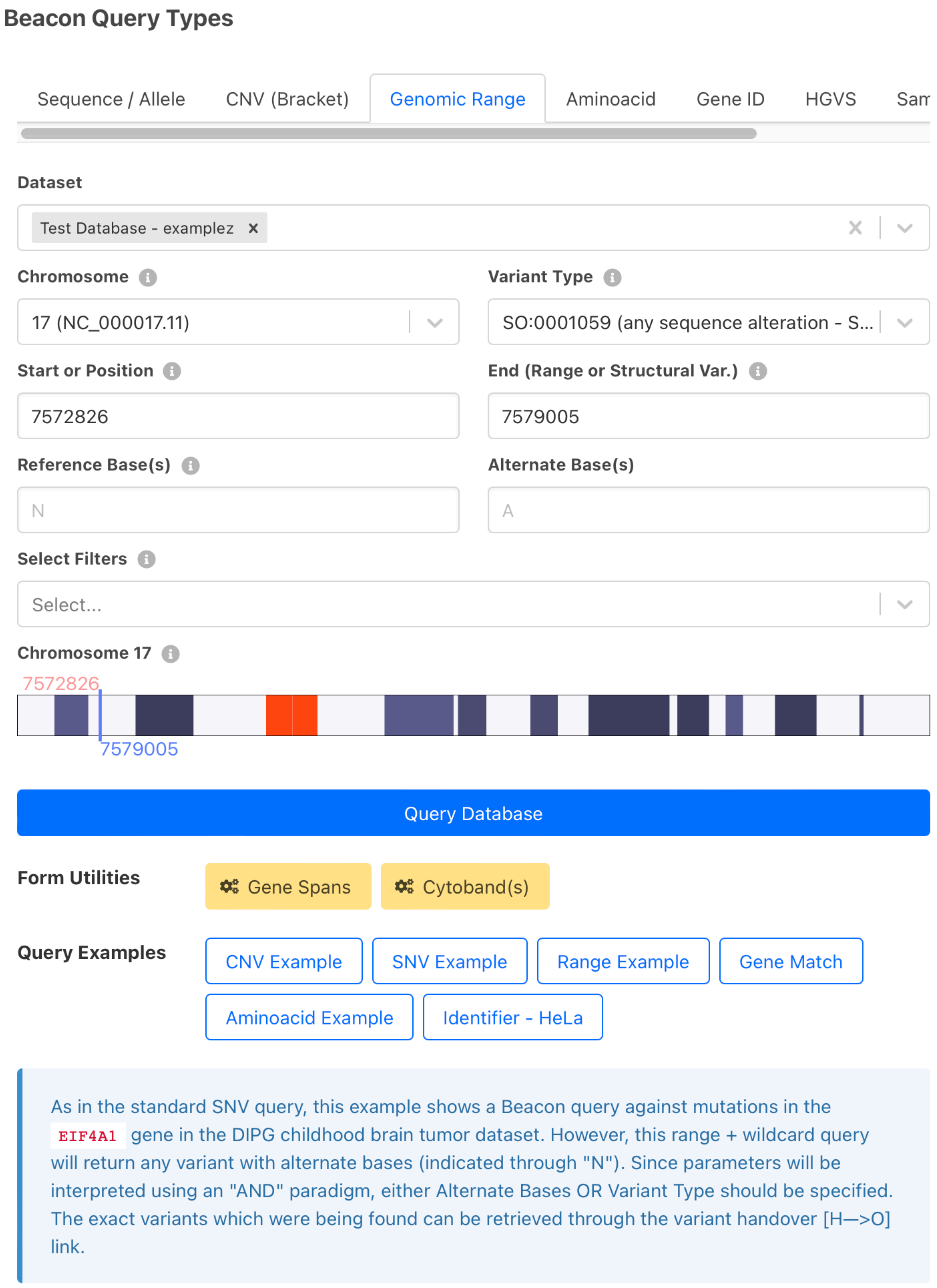
Query Examples

Beacon Queries

Range ("anything goes") Request

- defined through the use of 1 start, 1 end
- any variant... but can be limited by type etc.





Beacon Query Types

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

Dataset

Test Database - examplez ×

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

Query Database

Form Utilities: Gene Spans | Cytoband(s)

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

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

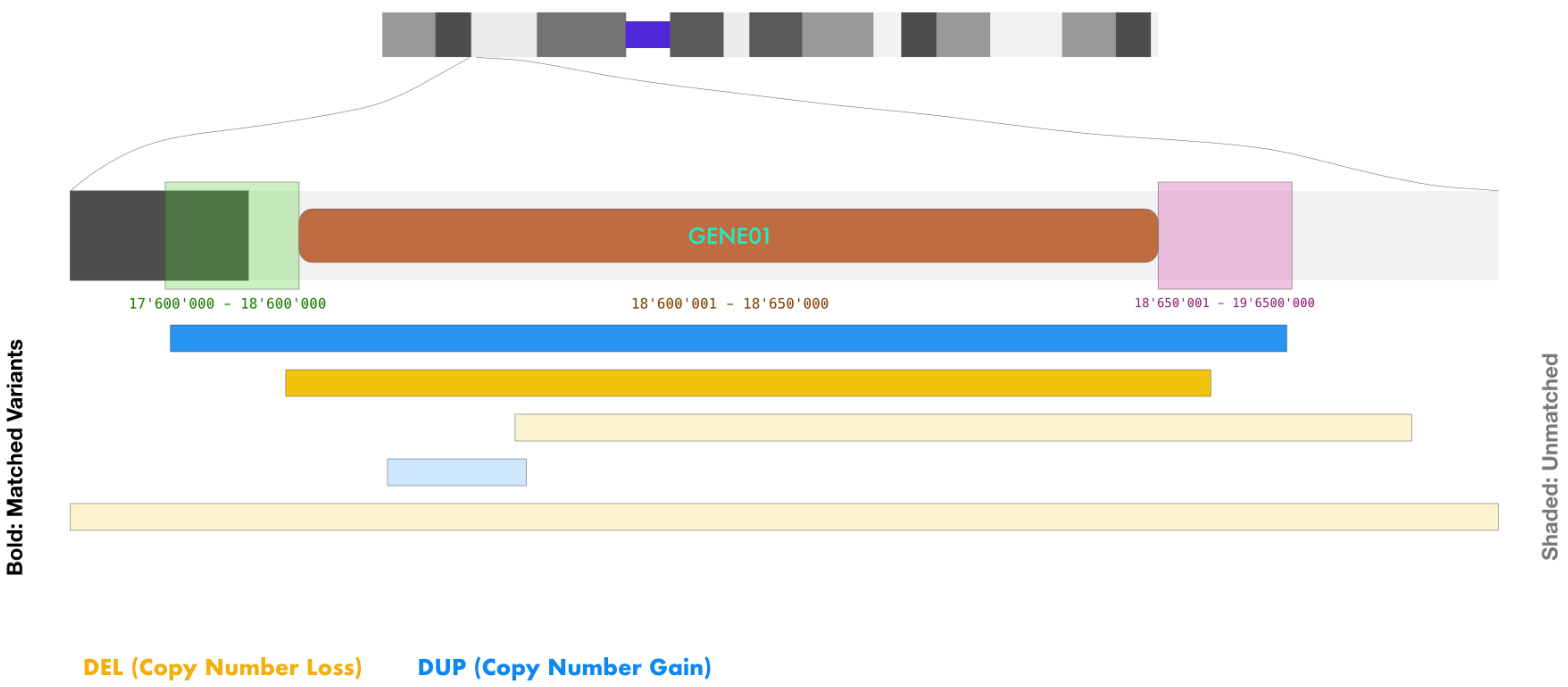
Beacon Queries

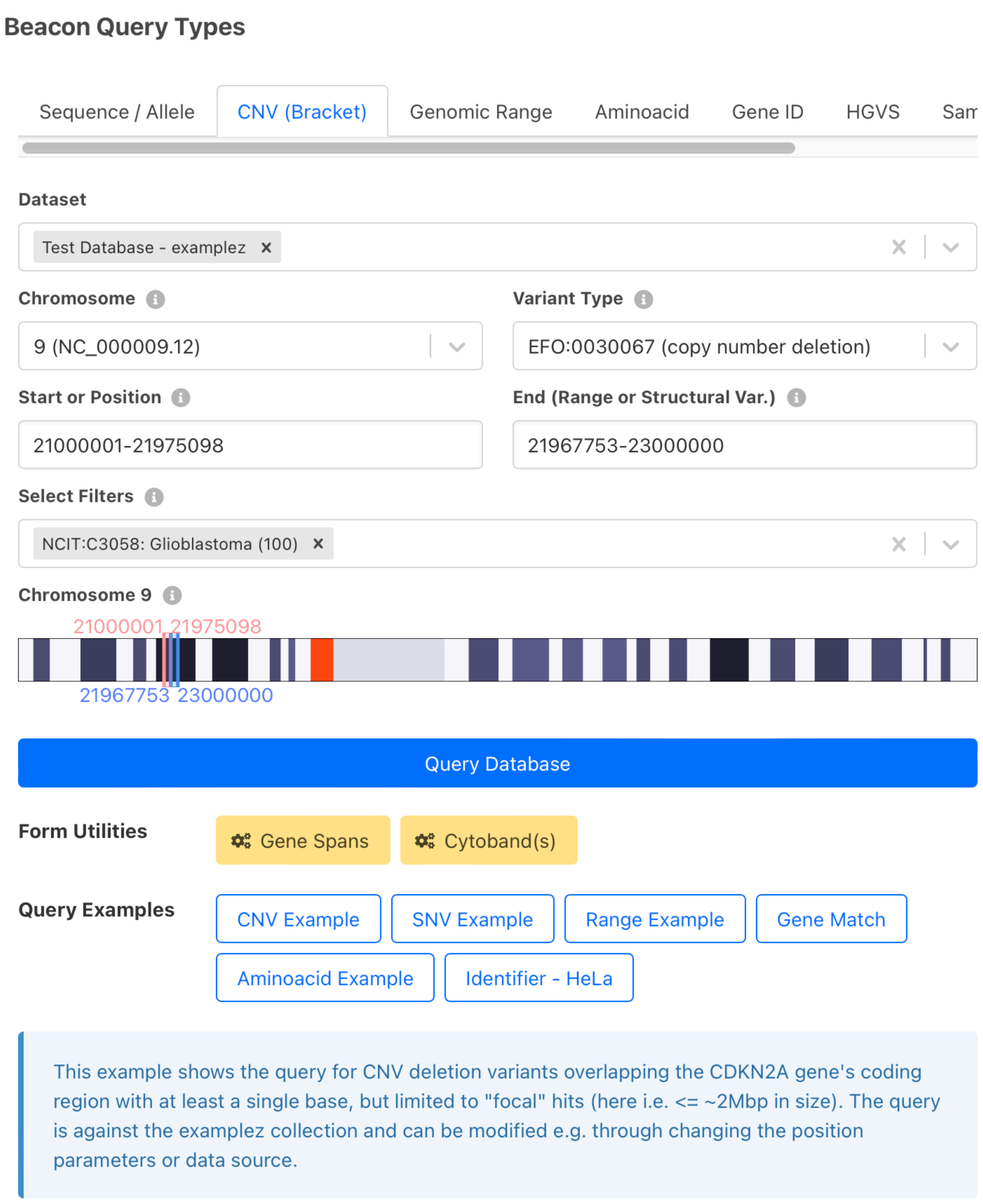
Bracket ("CNV") Query

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

Beacon Bracket Query

Example for complete regional match





Beacon Query Types

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

Dataset	Test Database - examplez ×
Chromosome	9 (NC_000009.12)
Variant Type	EFO:0030067 (copy number deletion)
Start or Position	21000001-21975098
End (Range or Structural Var.)	21967753-23000000
Select Filters	NCIT:C3058: Glioblastoma (100) ×

Chromosome 9

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.

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

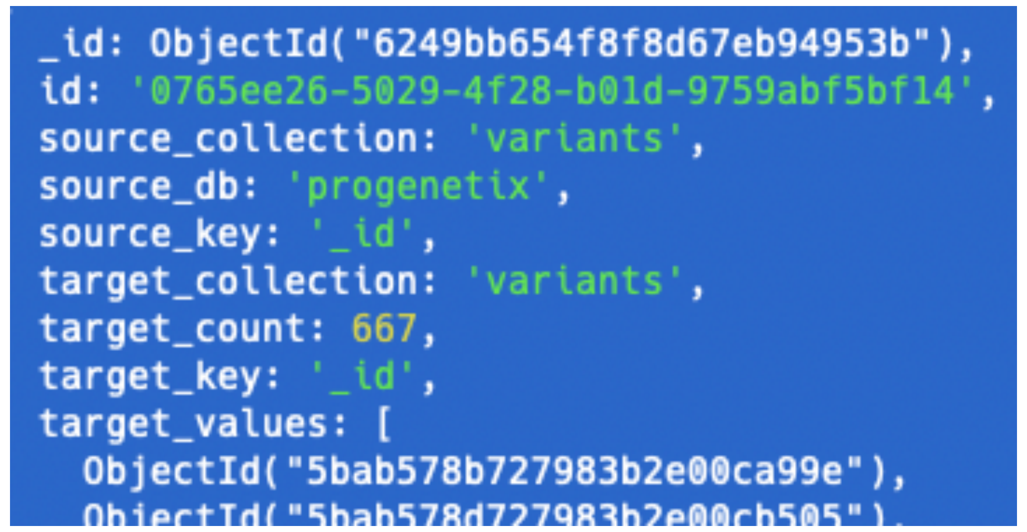
React

APACHE HTTP SERVER PROJECT

python™

mongoDB

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



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

variants analyses biosamples individuals

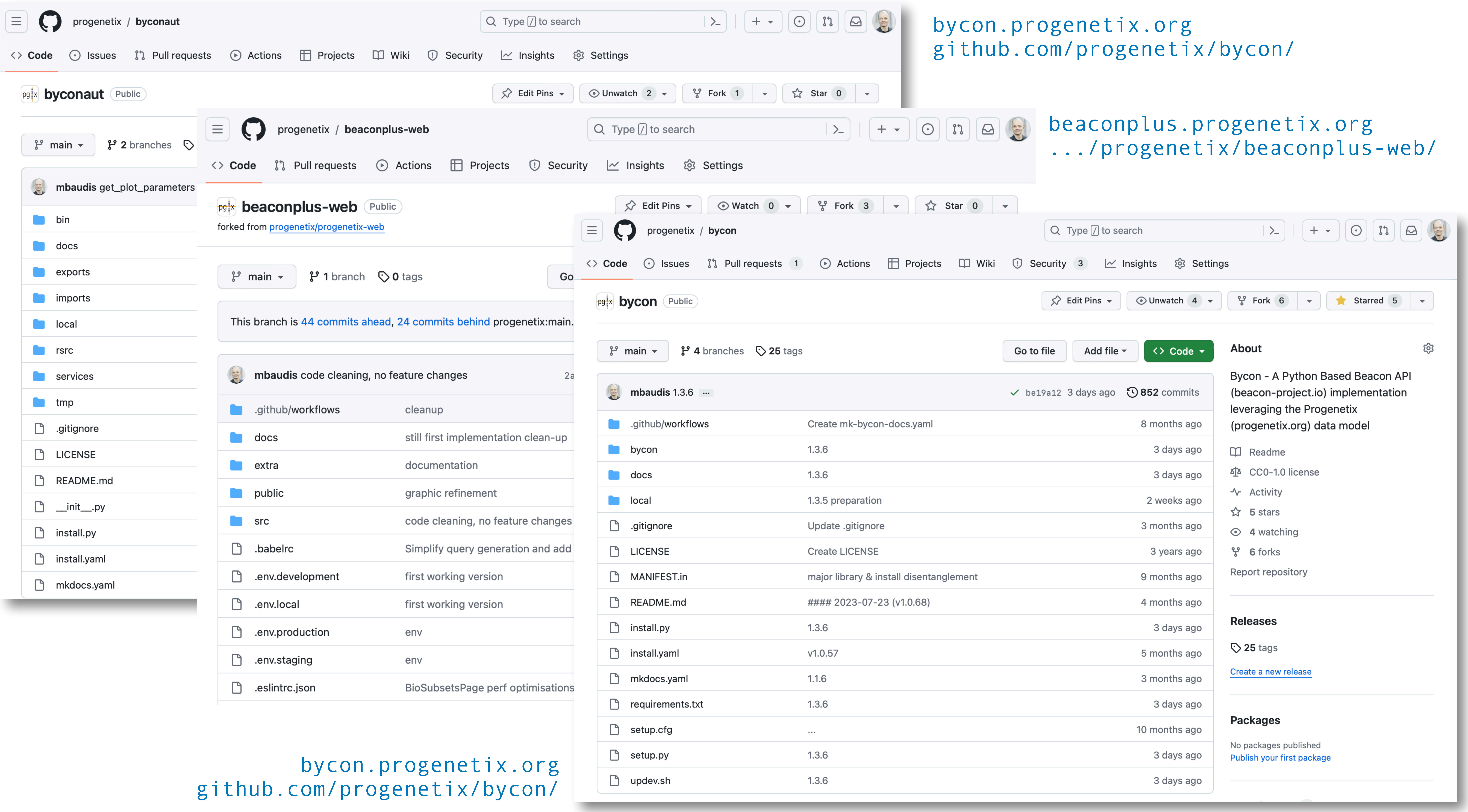
Entity collections

collations geolocs genespans publications qBuffer

Utility collections

bycon.progenetix.org
github.com/progenetix/bycon/

beaconplus.progenetix.org
.../progenetix/beaconplus-web/



mbaudis get_plot_parameters

bin
docs
exports
imports
local
rsrc
services
tmp
.gitignore
LICENSE
README.md
__init__.py
install.py
install.yaml
mkdocs.yaml

forked from progenetix/progenetix-web

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

mbaudis code cleaning, no feature changes 2a

.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

mbaudis 1.3.6 ... ✓ be19a12 3 days ago 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

About

Bycon - A Python Based Beacon API (beacon-project.io) implementation leveraging the Progenetix (progenetix.org) data model

Readme
CC0-1.0 license
Activity
5 stars
4 watching
6 forks
Report repository

Releases

25 tags

Create a new release

Packages

No packages published
Publish your first package

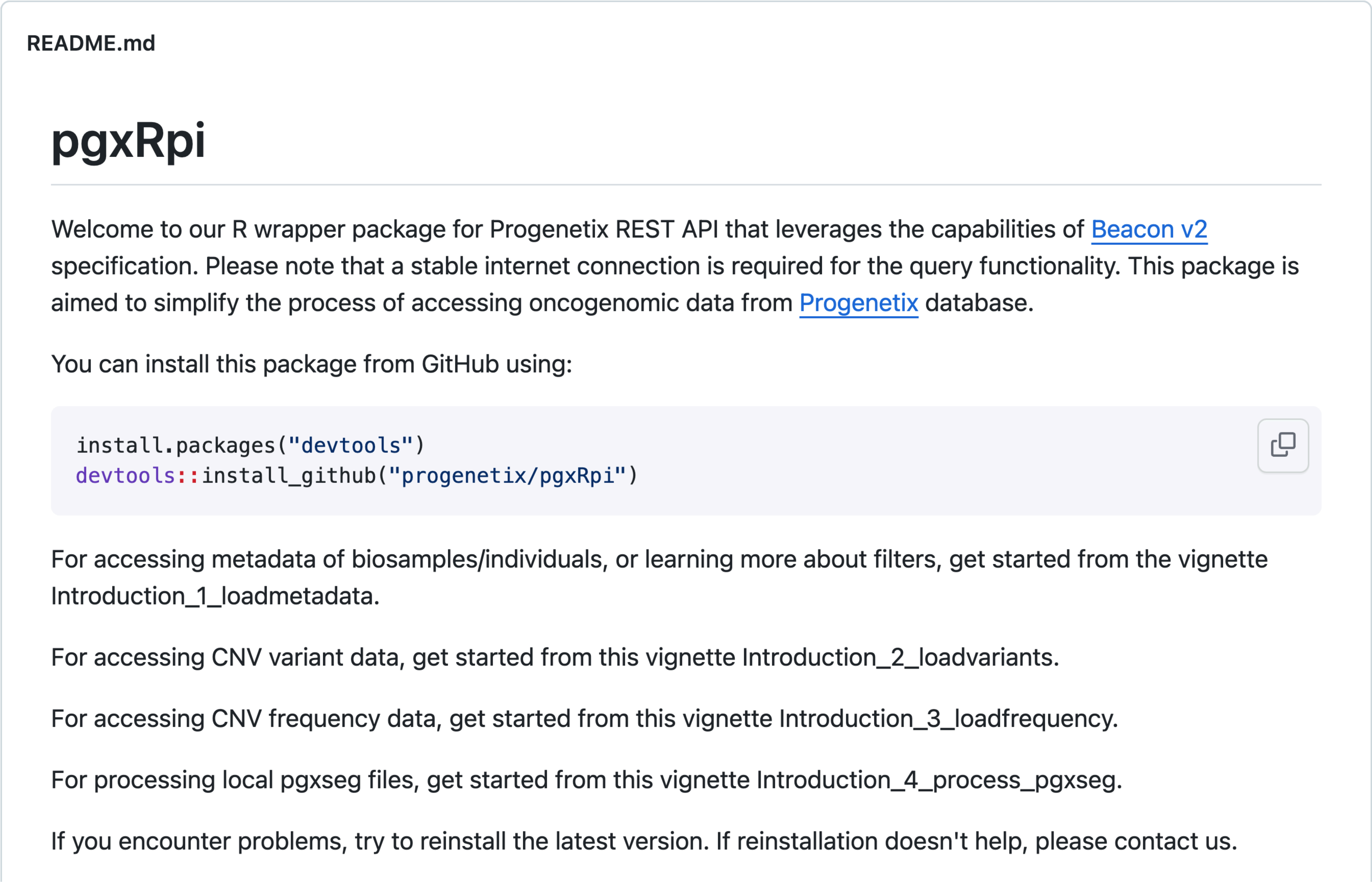
bycon.progenetix.org
github.com/progenetix/bycon/

pgxRpi

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

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

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")
```

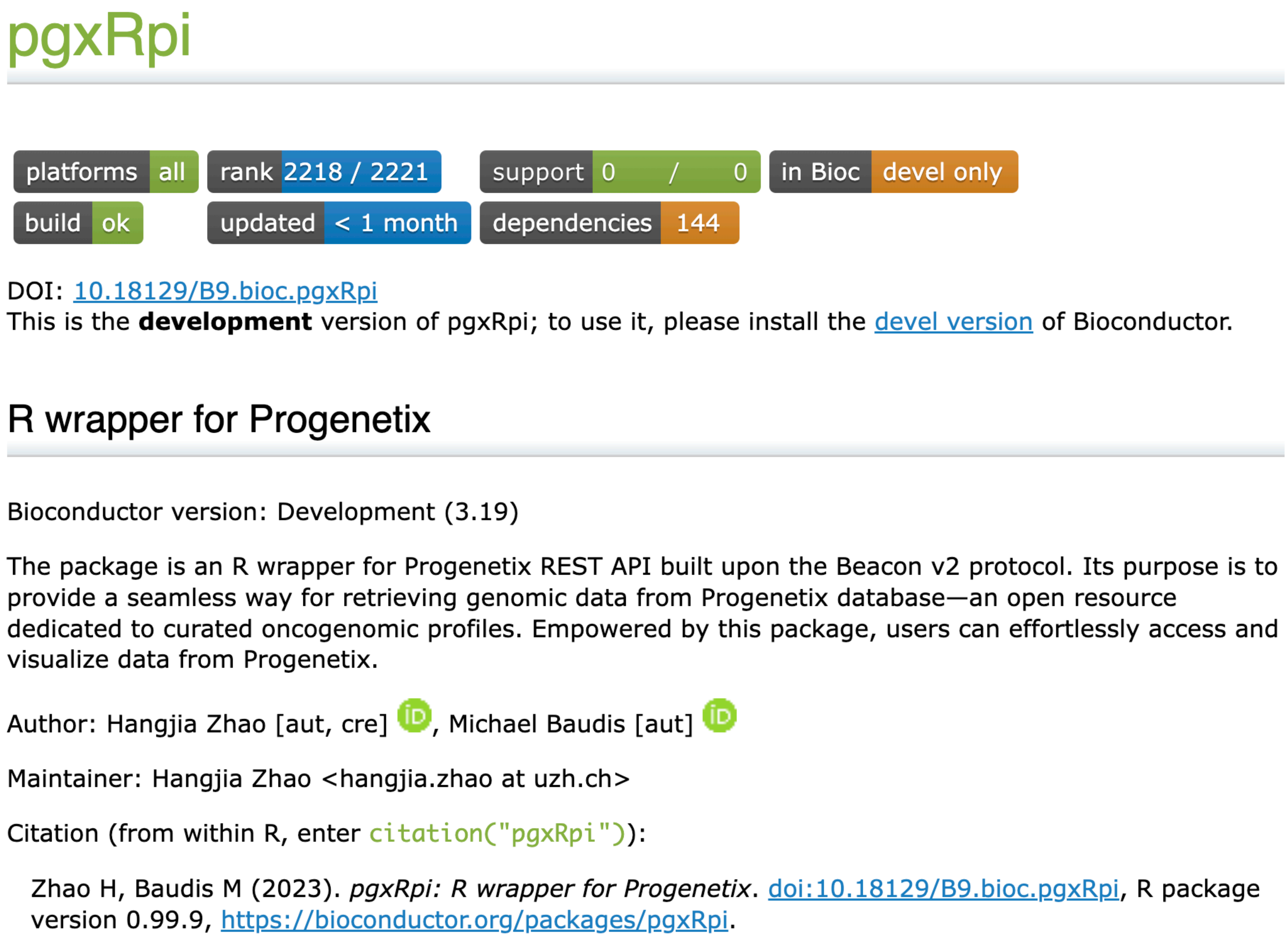
For accessing metadata of biosamples/individuals, or learning more about filters, get started from the vignette Introduction_1_loadmetadata.

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
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, R package version 0.99.9, https://bioconductor.org/packages/pgxRpi.

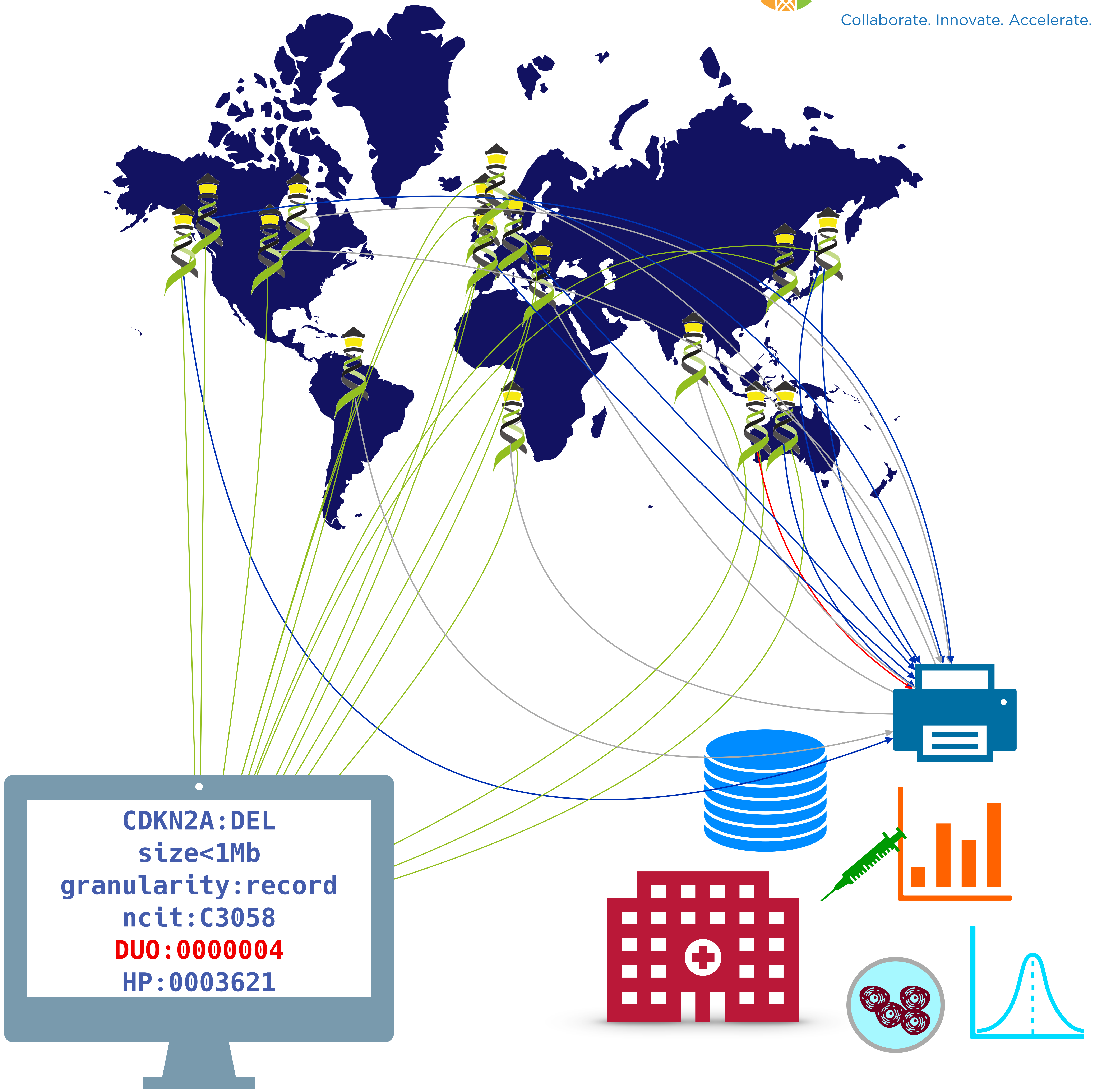
Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

What Can You Do?

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

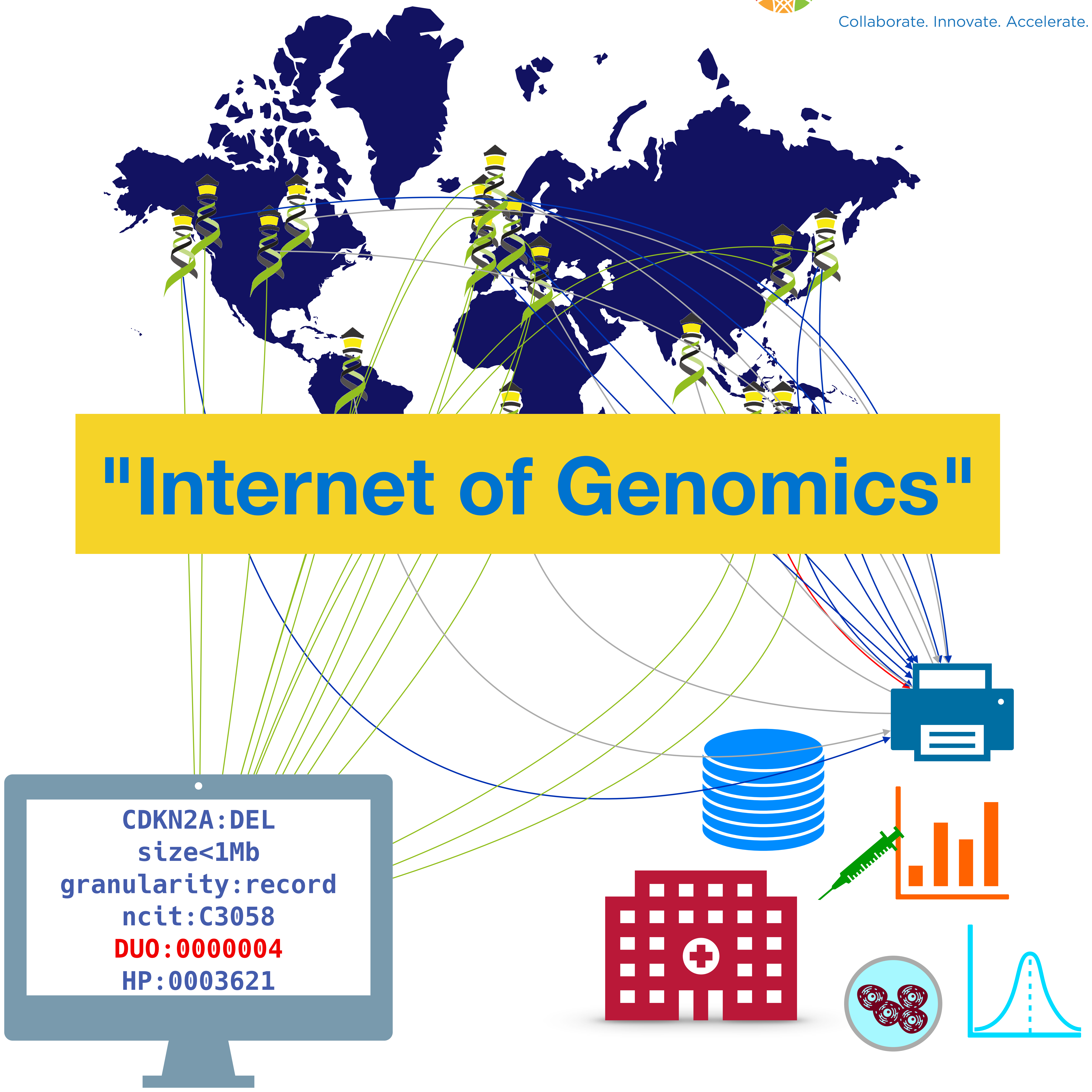
➡ **Collaborate!**



What Can You Do?

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

➡ **Collaborate!**



"Internet of Genomics"

```
CDKN2A:DEL
size<1Mb
granularity:record
ncit:C3058
DUO:0000004
HP:0003621
```

The Beacon team through the ages

EUROPEAN GENOME-PHENOME ARCHIVE

CRG Centre for Genomic Regulation

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

CSC

Juha Törnroos
Teemu Kataja
Ilkka Lappalainen
Dylan Spalding

UNIVERSITY OF LEICESTER

Cafe Variome Central

Tony Brookes
Tim Beck
Colin Veal
Tom Shorter

SPHN Swiss Personalized Health Network

University of Zurich UZH

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

Genomics england

Augusto Rendon
Ignacio Medina
Javier López
Jacobo Coll
Antonio Rueda

cnag centre nacional d'anàlisi genòmica centro nacional de análisis genómico

Sergi Beltran
Carles Hernandez

Inserm Institut national de la santé et de la recherche médicale

David Salgado

BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación

Salvador Capella
Dmitry Repchevski
JM Fernández

DisGeNET

Laura Furlong
Janet Piñero

elixir

B1MG

Serena Scollen
Gary Saunders
Giselle Kerry
David Lloyd

H3Africa Human Heredity & Health in Africa

Nicola Mulder
Mamana
Mbiyavanga
Ziyaad Parker

EU CAN Can.

David
Torrents

AUTISM SPEAKS

Dean Hartley

JUNTA DE ANDALUCIA

Fundación Progreso y Salud
CONSEJERÍA DE SALUD

Joaquin Dopazo
Javier Pérez
J.L. Fernández
Gema Roldan

CINECA

Thomas Keane
Melanie Courtot
Jonathan Dursi

Heidi Rehm
Ben Hutton

Toshiaki
Katayama

GEM Japan

Stephane Dyke

DNASTACK

Marc Fiume
Miro Cupak

BRCA EXCHANGE

Melissa Cline

ENA

EMBL-EBI

Diana Lemos

EUROPEAN JOINT PROGRAMME RARE DISEASES

VICC Variant Interpretation for Cancer Consortium

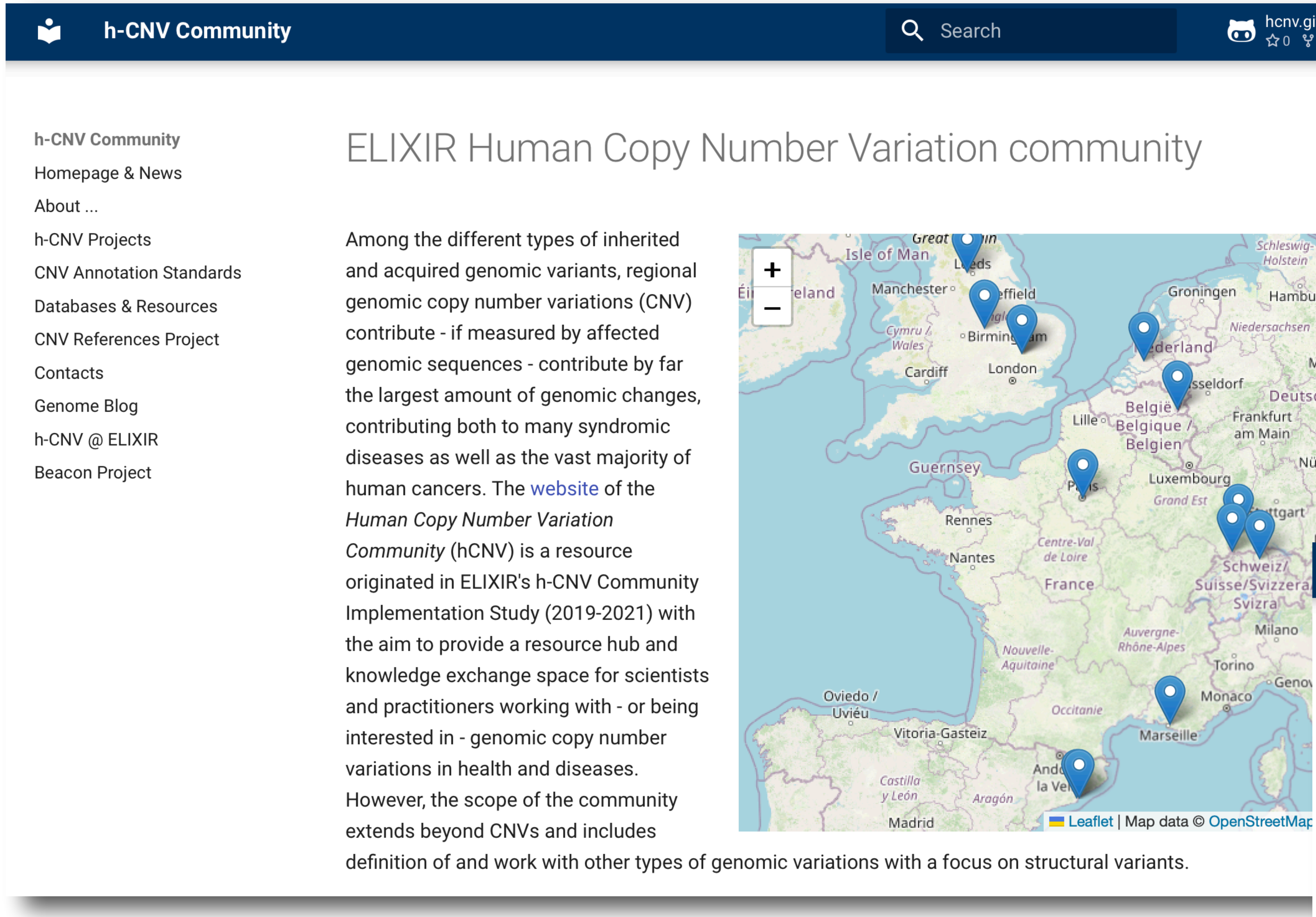
GA4GH Phenopackets
Peter Robinson
Jules Jacobsen

GA4GH VRS
Alex Wagner
Reece Hart

Beacon PRC
Alex Wagner
Jonathan Dursi
Mamana Mbiyavanga

Alice Mann
Neerjah Skantharajah

elixir



h-CNV Community

Homepage & News

About ...

h-CNV Projects

CNV Annotation Standards

Databases & Resources

CNV References Project

Contacts

Genome Blog

h-CNV @ ELIXIR

Beacon Project

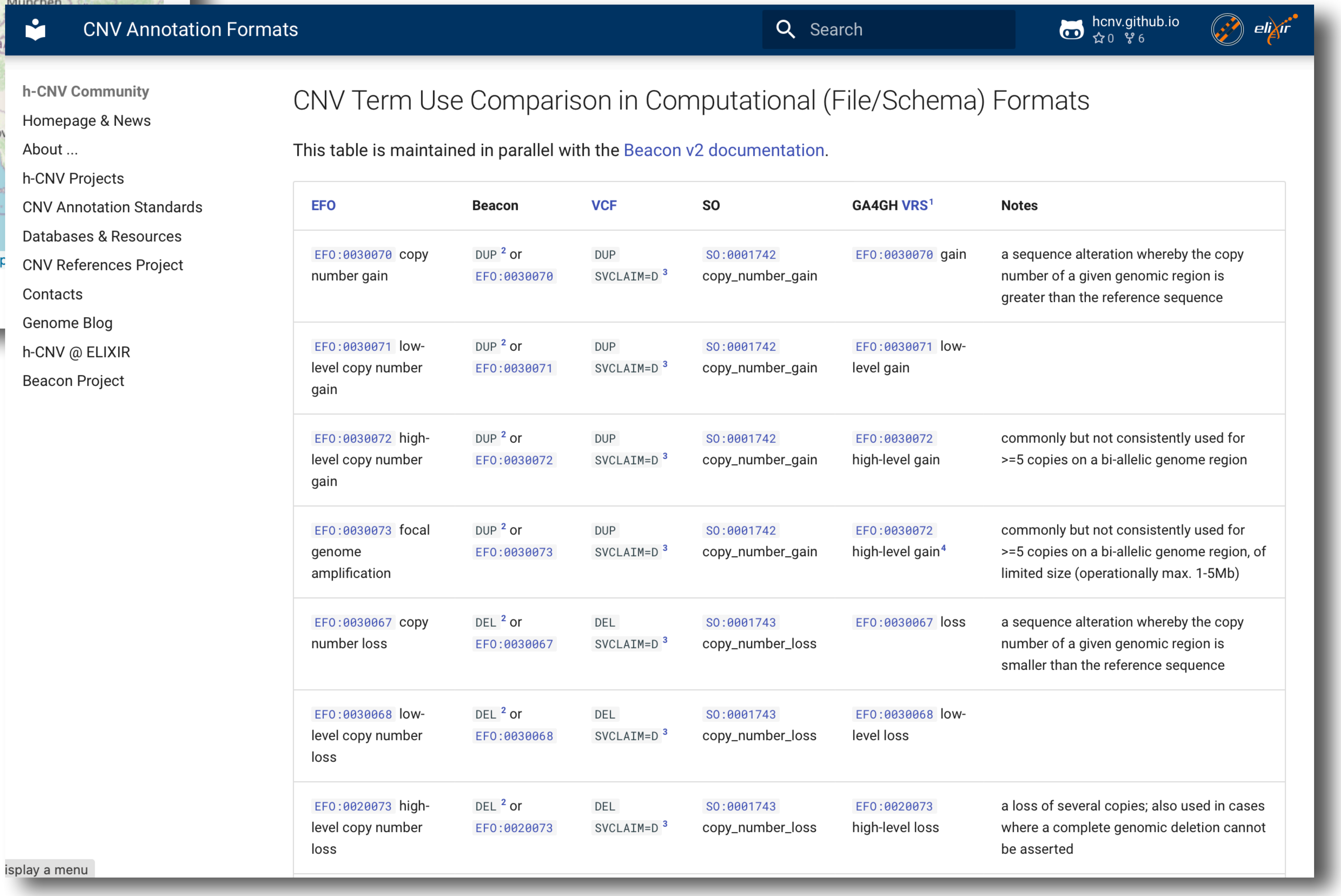
ELIXIR Human Copy Number Variation community

Among the different types of inherited and acquired genomic variants, regional genomic copy number variations (CNV) contribute - if measured by affected genomic sequences - contribute by far the largest amount of genomic changes, contributing both to many syndromic diseases as well as the vast majority of human cancers. The website of the *Human Copy Number Variation Community* (hCNV) is a resource originated in ELIXIR's h-CNV Community Implementation Study (2019-2021) with the aim to provide a resource hub and knowledge exchange space for scientists and practitioners working with - or being interested in - genomic copy number variations in health and diseases. However, the scope of the community extends beyond CNVs and includes definition of and work with other types of genomic variations with a focus on structural variants.

ELIXIR hCNV Community

https://cnvar.org/

elixir



h-CNV Community

Homepage & News

About ...

h-CNV Projects

CNV Annotation Standards

Databases & Resources

CNV References Project

Contacts

Genome Blog

h-CNV @ ELIXIR

Beacon Project

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

This table is maintained in parallel with the Beacon v2 documentation.

EFO	Beacon	VCF	SO	GA4GH VRS[1]	Notes
EFO:0030070 copy number gain	DUP [2] or EFO:0030070	DUP SVCLAIM=D [3]	SO:0001742 copy_number_gain	EFO:0030070 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 [2] or EFO:0030071	DUP SVCLAIM=D [3]	SO:0001742 copy_number_gain	EFO:0030071 low-level gain	
EFO:0030072 high-level copy number gain	DUP [2] or EFO:0030072	DUP SVCLAIM=D [3]	SO:0001742 copy_number_gain	EFO:0030072 high-level gain	commonly but not consistently used for >=5 copies on a bi-allelic genome region
EFO:0030073 focal genome amplification	DUP [2] or EFO:0030073	DUP SVCLAIM=D [3]	SO:0001742 copy_number_gain	EFO:0030072 high-level gain[4]	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 [2] or EFO:0030067	DEL SVCLAIM=D [3]	SO:0001743 copy_number_loss	EFO:0030067 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 [2] or EFO:0030068	DEL SVCLAIM=D [3]	SO:0001743 copy_number_loss	EFO:0030068 low-level loss	
EFO:0020073 high-level copy number loss	DEL [2] or EFO:0020073	DEL SVCLAIM=D [3]	SO:0001743 copy_number_loss	EFO:0020073 high-level loss	a loss of several copies; also used in cases where a complete genomic deletion cannot be asserted

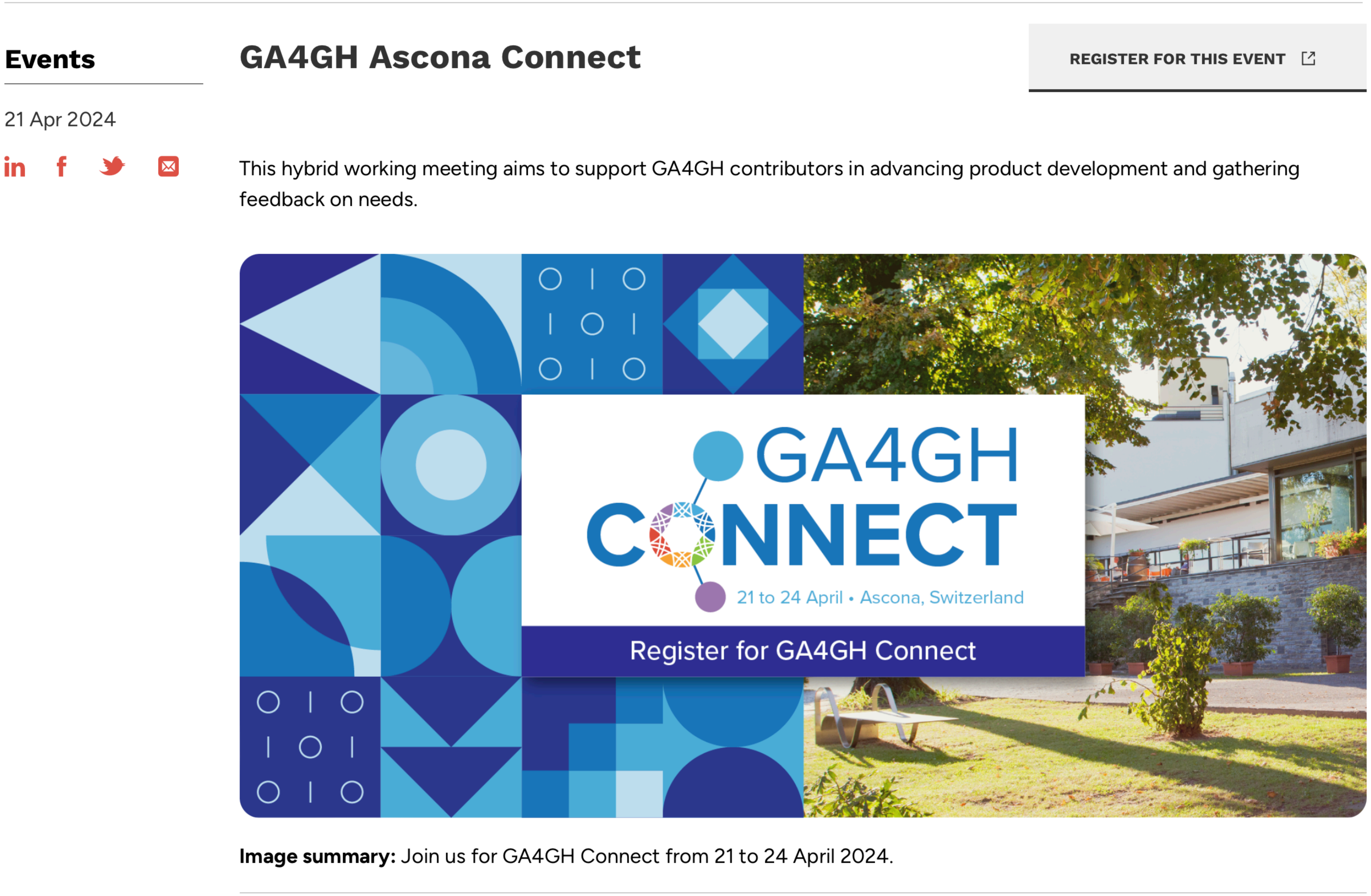
ELIXIR hCNV Community

www.elixir-europe.org

CNV Term Use Comparison

in computational (file/schema) formats

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



Events

21 Apr 2024

GA4GH Ascona Connect

REGISTER FOR THIS EVENT

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

GA4GH
CONNECT
21 to 24 April • Ascona, Switzerland
Register for GA4GH Connect

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

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

Genomic Data & Privacy

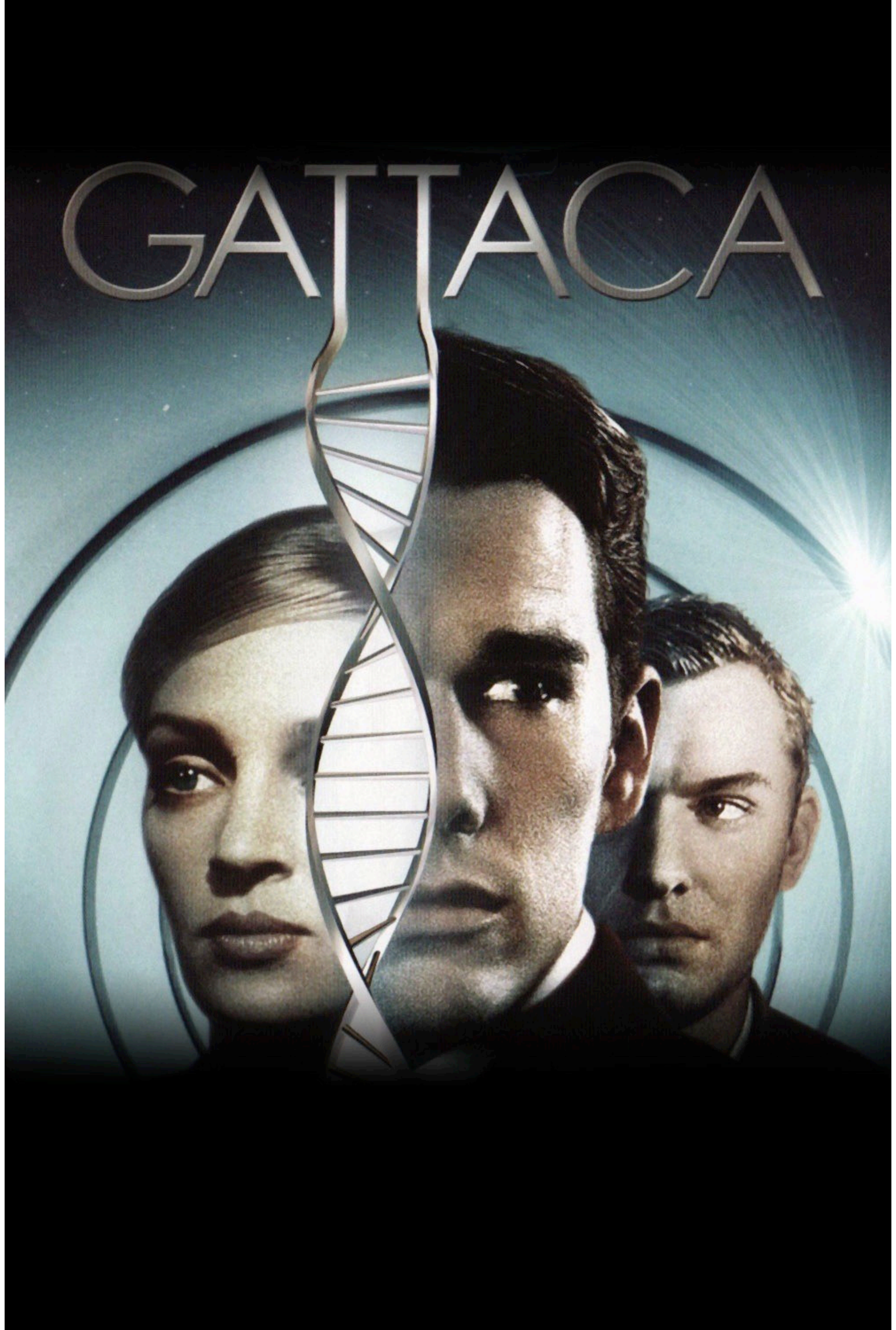
Risks & opportunities

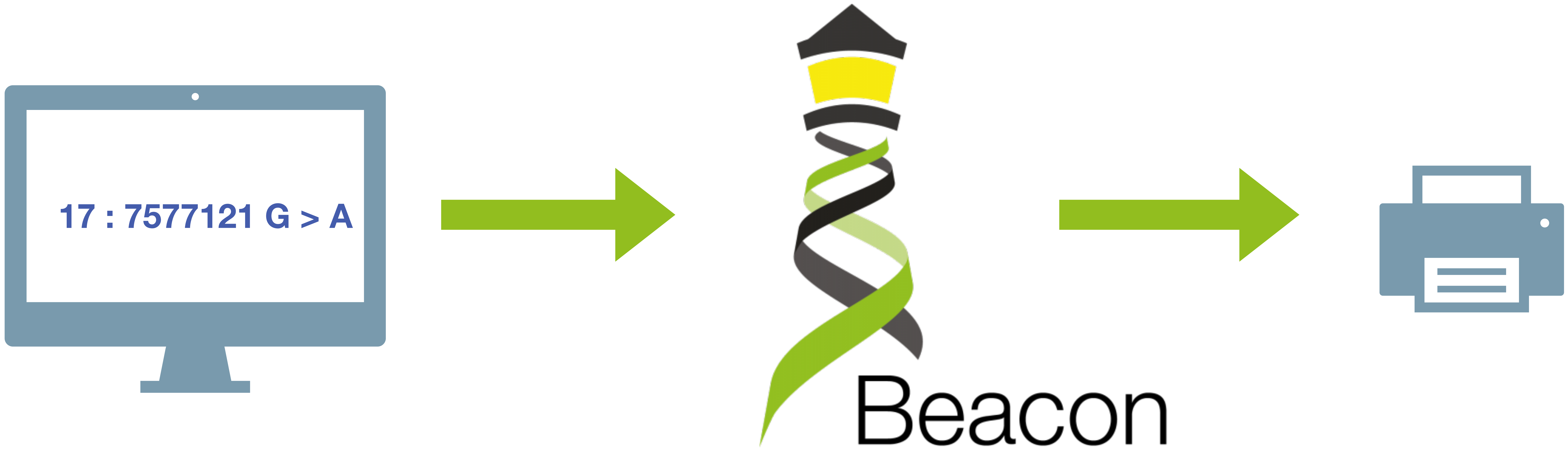
Gattaca (1997)

A genetically inferior man assumes the identity of a superior one in order to pursue his lifelong dream of space travel.

- genetic determinism
  - main character has been determined to be unsuitable for complex jobs based on genetic analysis
- genetic identification
  - the use of genetic sampling for personal identification is daily routine

With information from https://www.imdb.com/title/tt0119177/



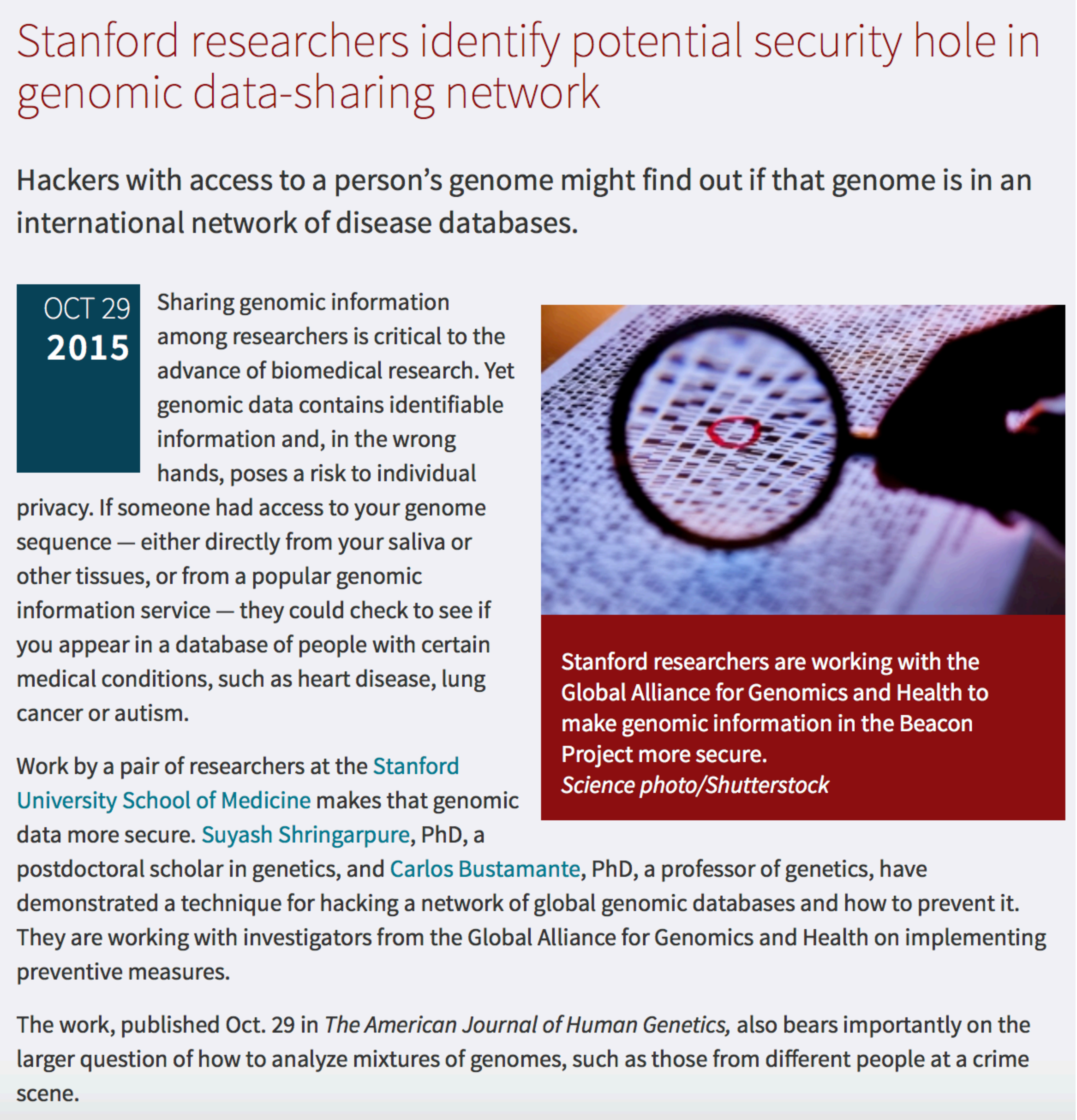


A ***Beacon*** answers a query for a specific genome variant against individual or aggregate genome collections

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

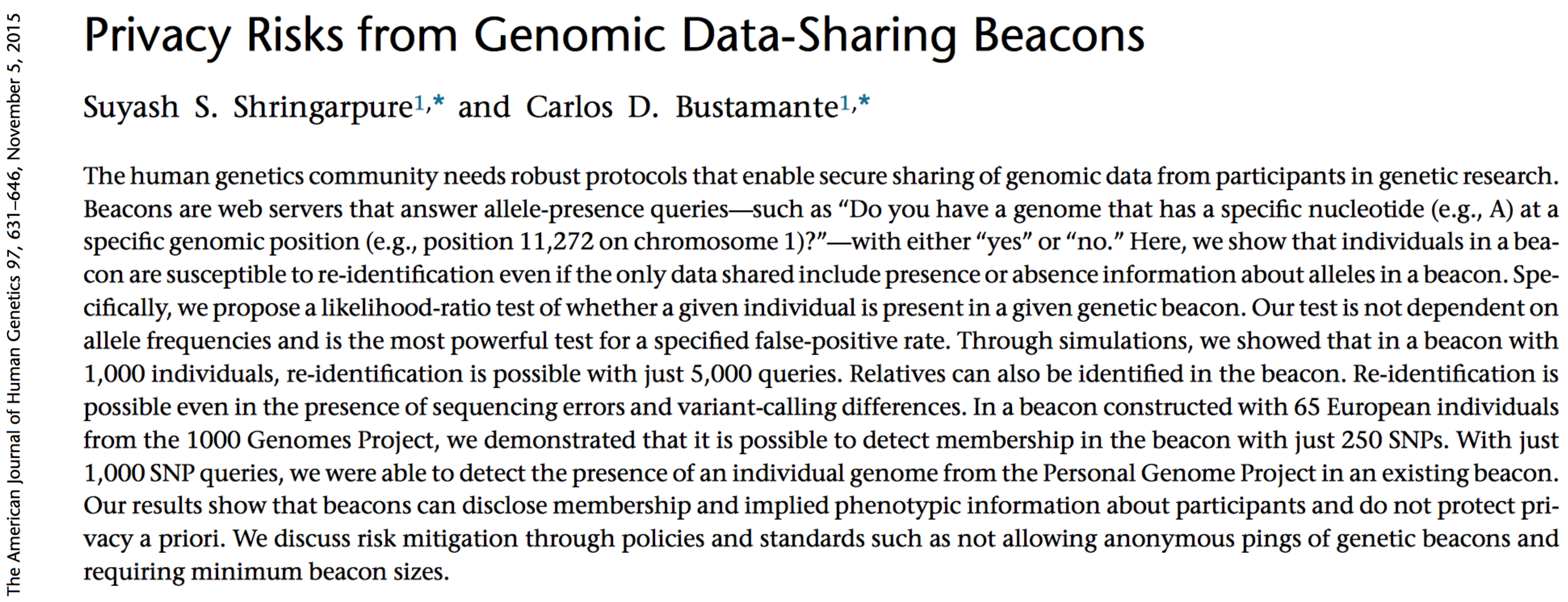
Sharing genomic information among researchers is critical to the advance of biomedical research. Yet genomic data contains identifiable information and, in the wrong hands, poses a risk to individual privacy. If someone had access to your genome sequence — either directly from your saliva or other tissues, 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.

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

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.

IDENTIFICATION OF INDIVIDUALS FROM MIXED COLLECTIONS USING RARE ALLELES



The American Journal of Human Genetics 97, 631–646, November 5, 2015

Privacy Risks from Genomic Data-Sharing Beacons

Suyash S. Shringarpure[1,*] and Carlos D. Bustamante[1,*]

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

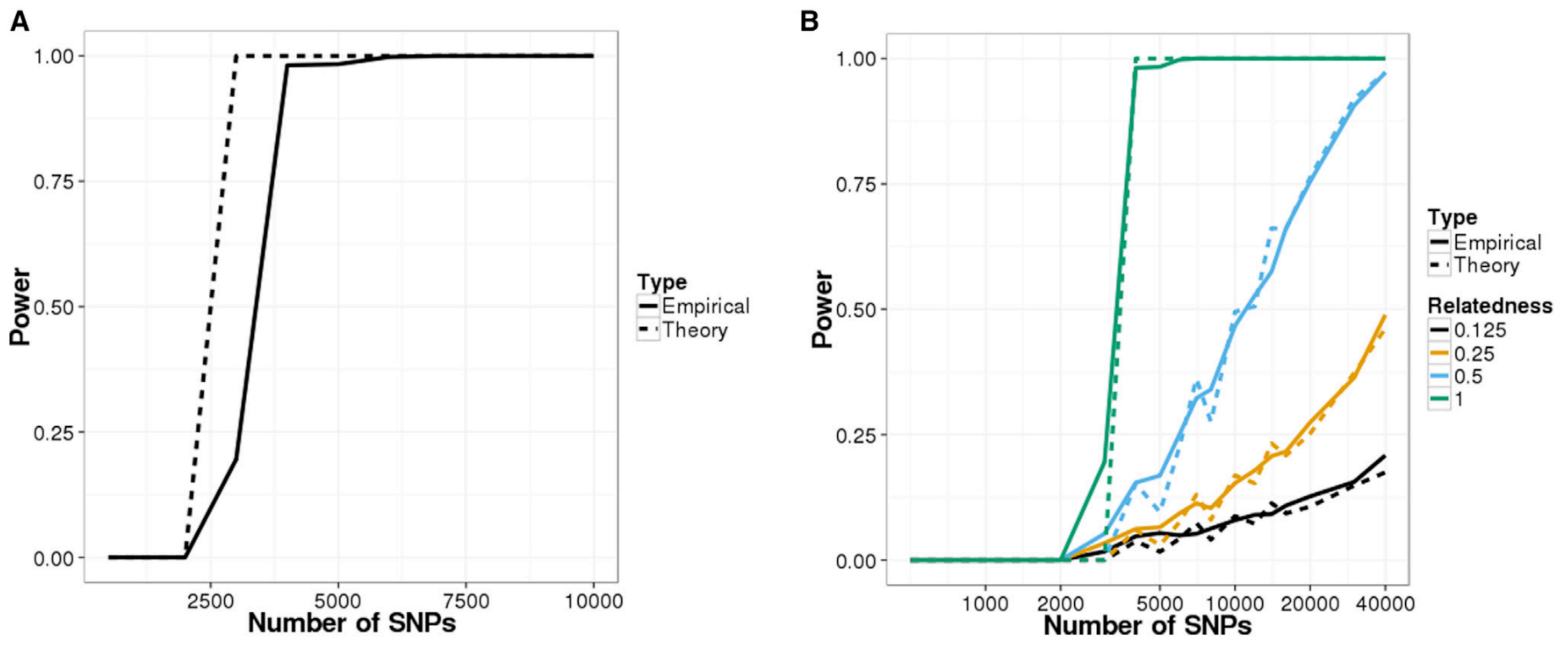


Figure 1. Power of Re-identification Attacks on Beacons Constructed with Simulated Data
Power curves for the likelihood-ratio test (LRT) on (A) a simulated beacon with 1,000 individuals and (B) detecting relatives in the simulated beacon. The false-positive rate was set to 0.05 for all scenarios.

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

Direct to Consumer DNA Analyses

Population Background, Family Trees, Traits & Disease Risks...

Enorme Ersparnisse

Letzte Chance DNA-Weihnachtsaktion

Nur

39 CHF ~~89 CHF~~

MyHeritage

Entdecken Ihre Wurz

Erweitern Sie Ihren Stamm Sie Verwandte, und durchf historische Dokumente mit KOSTENLOSEN Testversio

By the numbers

2006
The year we set out to make DNA more accessible and meaningful for all.

12M+
The number of DNA kits we've sold in that time.

Best DNA test kits on sale for Cyber Monday 2023

By Maren Estrada · Published Nov 26th, 2023 11:11AM EST

Image: Maren Estrada for BGR

If you buy through a BGR link, we may earn an affiliate commission, h

23andMe · OUR SERVICE · LEARN · SIGN IN · REGISTER K

Your DNA is amazing!

Behi point being

We hea
Deciphering most excitin our lifetime.

And you should be able to access, benefit from the endlessly interest things your genetics can tell you. H to explore your DNA is up to you. W we'd like to be the first to say, "Welcome to you®."

Jacqueline	100%
European	50.1%
British & Irish	39.7%
French & German	7.0%
Broadly Northwestern European	3.2%
Scandinavian	0.2%
East Asian & Indigenous American	49.9%
Vietnamese	46.3%
Indonesian, Thai & Myanma	1.5%
Chinese	0.5%

ancestry · GENEALOGY · DNA · FREE TRIAL · SIGN IN · EN

What would you like to learn about your family history?
Select all that apply

□ Details about my ancestors
□ My origins
□ I'm not really sure what I can discover on Ancestry

Skip · Next

Dismiss

Bring the generations together with a gift from Ancestry®.

Build a family tree and uncover your story.
Start a free trial

Save up to $50* on DNA offers for a limited time.
Start saving

*Ends 31 Dec 2022. Terms apply. Pricing for U.S. customers only.

“We’re an information economy. They teach you that in school. What they don’t tell you is that it’s impossible to move, to live, to operate at any level without leaving traces, bits, seemingly meaningless fragments of personal information. Fragments that can be retrieved, amplified . . .”

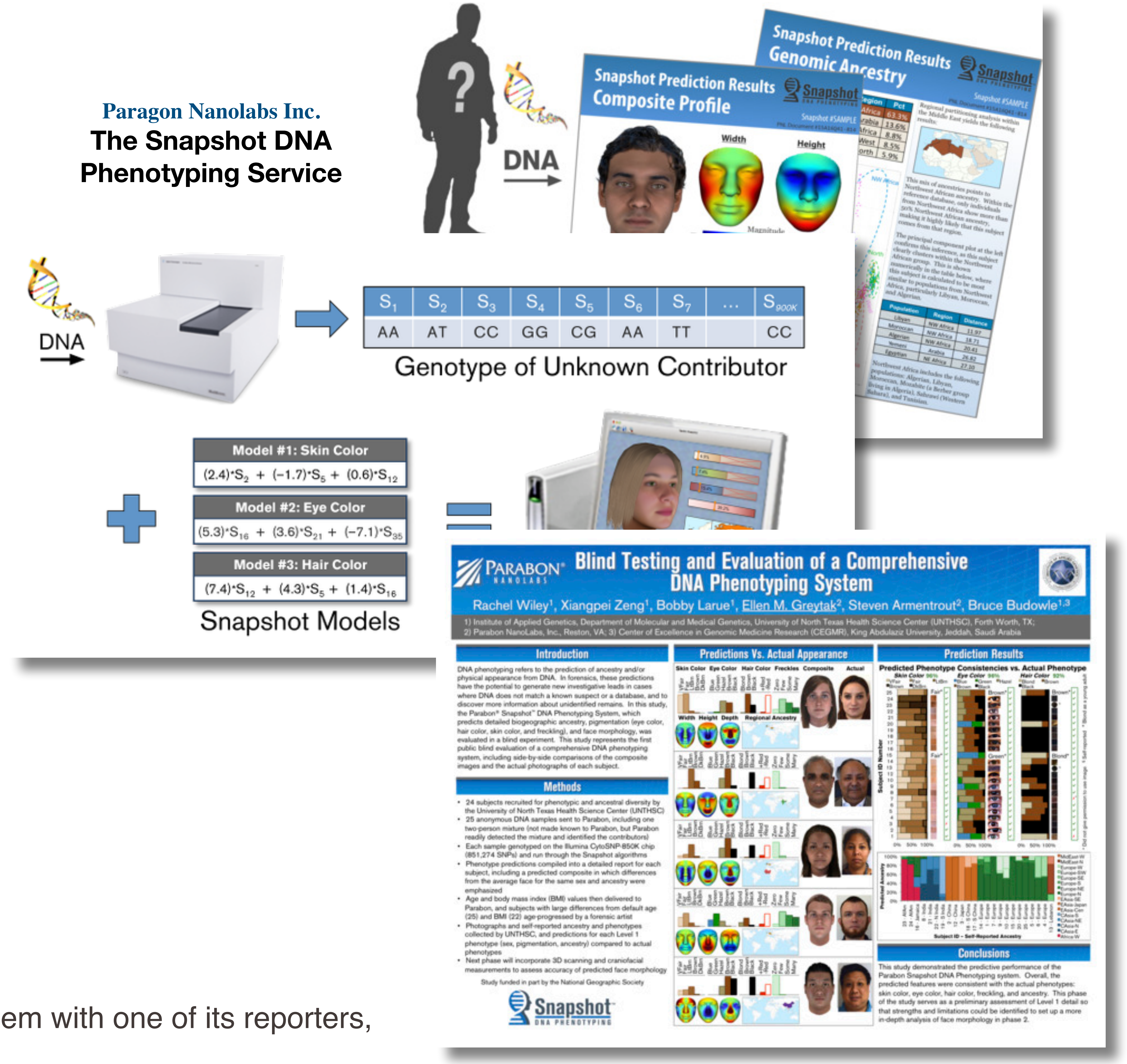
–William Gibson in "Johnny Mnemonic" (1986)

Phenotyping from DNA

From DNA to "Wanted" Posters?

- association of genomic variants with phenotypic data collection
- while hair, eye color are easy targets not useful for relevant phenotypic features especially if large environmental component
- huge biases based on input/collection data
- Belgium and Germany do not allow forensic DNA phenotyping
- Switzerland: Bundesrat decision on 2020-12-04 to allow phenotyping for law enforcement purposes

Paragon Nanolabs Inc.
The Snapshot DNA Phenotyping Service


DNA
Snapshot Prediction Results
Composite Profile
Snapshot Prediction Results
Genomic Ancestry
Snapshot
Width
Height
DNA
S_1 S_2 S_3 S_4 S_5 S_6 S_7 … S_{900K}
AA AT CC GG CG AA TT CC
Genotype of Unknown Contributor
Model #1: Skin Color
$(2.4) \cdot S_2 + (-1.7) \cdot S_5 + (0.6) \cdot S_{12}$
Model #2: Eye Color
$(5.3) \cdot S_{16} + (3.6) \cdot S_{21} + (-7.1) \cdot S_{35}$
Model #3: Hair Color
$(7.4) \cdot S_{12} + (4.3) \cdot S_5 + (1.4) \cdot S_{16}$
Snapshot Models
Predictions Vs. Actual Appearance
Skin Color
Eye Color
Hair Color
Freckles
Composite
Actual
Width
Height
Depth
Regional Ancestry
Prediction Results
Predicted Phenotype Consistencies vs. Actual Phenotype
Subject ID Number
Subject ID – Self-Reported Ancestry
Snapshot
DNA PHENOTYPING

PARABON NANOLABS **Blind Testing and Evaluation of a Comprehensive DNA Phenotyping System**

Rachel Wiley[1], Xiangpei Zeng[1], Bobby Larue[1], Ellen M. Greytak[2], Steven Armentrout[2], Bruce Budowle[1,3]

1) Institute of Applied Genetics, Department of Molecular and Medical Genetics, University of North Texas Health Science Center (UNTHSC), Forth Worth, TX; 2) Parabon NanoLabs, Inc., Reston, VA; 3) Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

Introduction

DNA phenotyping refers to the prediction of ancestry and/or physical appearance from DNA. In forensics, these predictions have the potential to generate new investigative leads in cases where DNA does not match a known suspect or a database, and to discover more information about unidentified remains. In this study, the Parabon® Snapshot™ DNA Phenotyping System, which predicts detailed biogeographic ancestry, pigmentation (eye color, hair color, skin color, and freckling), and face morphology, was evaluated in a blind experiment. This study represents the first public blind evaluation of a comprehensive DNA phenotyping system, including side-by-side comparisons of the composite images and the actual photographs of each subject.

Methods

- 24 subjects recruited for phenotypic and ancestral diversity by the University of North Texas Health Science Center (UNTHSC)
- 25 anonymous DNA samples sent to Parabon, including one two-person mixture (not made known to Parabon, but Parabon readily detected the mixture and identified the contributors)
- Each sample genotyped on the Illumina CytoSNP-850K chip (851,274 SNPs) and run through the Snapshot algorithms
- Phenotype predictions compiled into a detailed report for each subject, including a predicted composite in which differences from the average face for the same sex and ancestry were emphasized
- Age and body mass index (BMI) values then delivered to Parabon, and subjects with large differences from default age (25) and BMI (22) age-progressed by a forensic artist
- Photographs and self-reported ancestry and phenotypes collected by UNTHSC, and predictions for each Level 1 phenotype (sex, pigmentation, ancestry) compared to actual phenotypes
- Next phase will incorporate 3D scanning and craniofacial measurements to assess accuracy of predicted face morphology

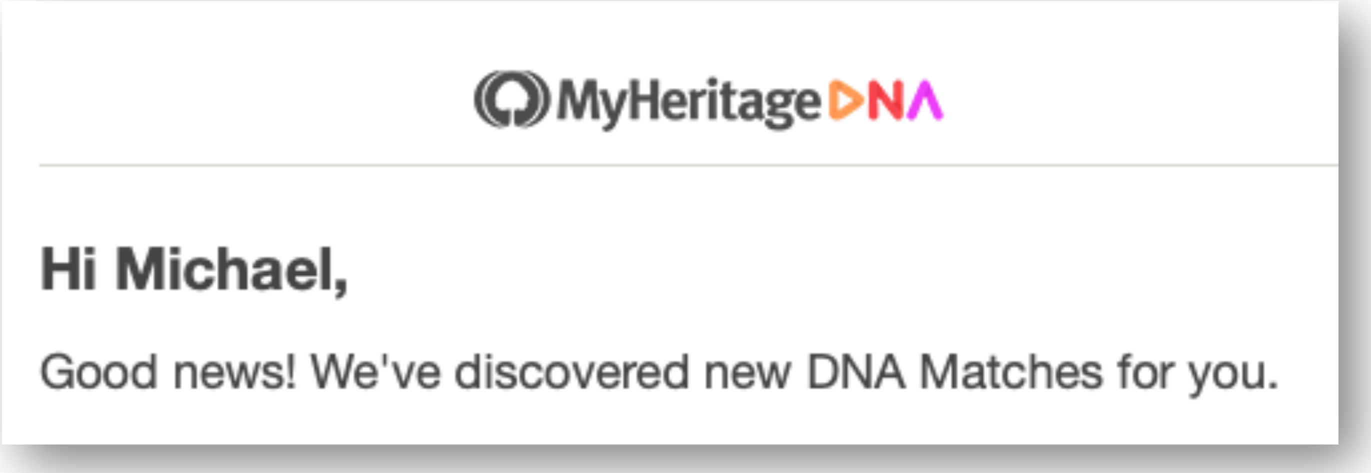
Study funded in part by the National Geographic Society

Conclusions

This study demonstrated the predictive performance of the Parabon Snapshot DNA Phenotyping system. Overall, the predicted features were consistent with the actual phenotypes: skin color, eye color, hair color, freckling, and ancestry. This phase of the study serves as a preliminary assessment of Level 1 detail so that strengths and limitations could be identified to set up a more in-depth analysis of face morphology in phase 2.

"When the New York Times ran an informal test of the Parabon system with one of its reporters, it failed badly." (ACLU.org)

https://snapshot.parabon-nanolabs.com/phenotyping



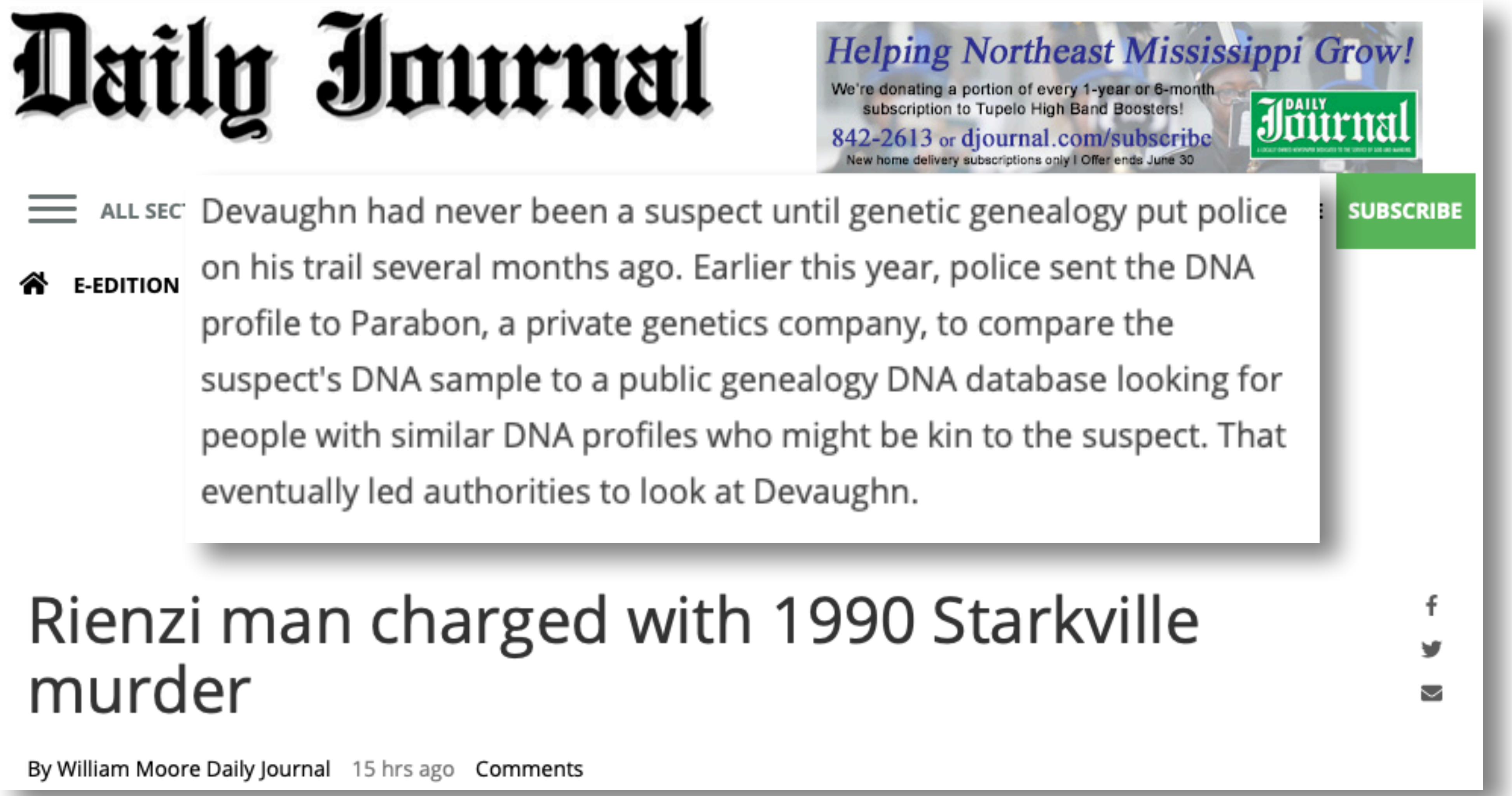
MyHeritage DNA

Hi Michael,

Good news! We've discovered new DNA Matches for you.

Long-Range Familial Searches

- Commercial, "Direct to Customer" DNA analyses are provided through independent sites and such affiliated to genealogy services (MyHeritage, Ancestry.com, 23andMe...)
- Genealogy sites identify individuals with matching haplotype blocks & provide a prediction about degree of genetic relation
- Law enforcement agencies (and who else?!) can send individual SNP profiles (e.g. recovered from evidence many years after a crime) using a *Jane Doe* identity, to identify relatives of the suspect - **long range familial search**



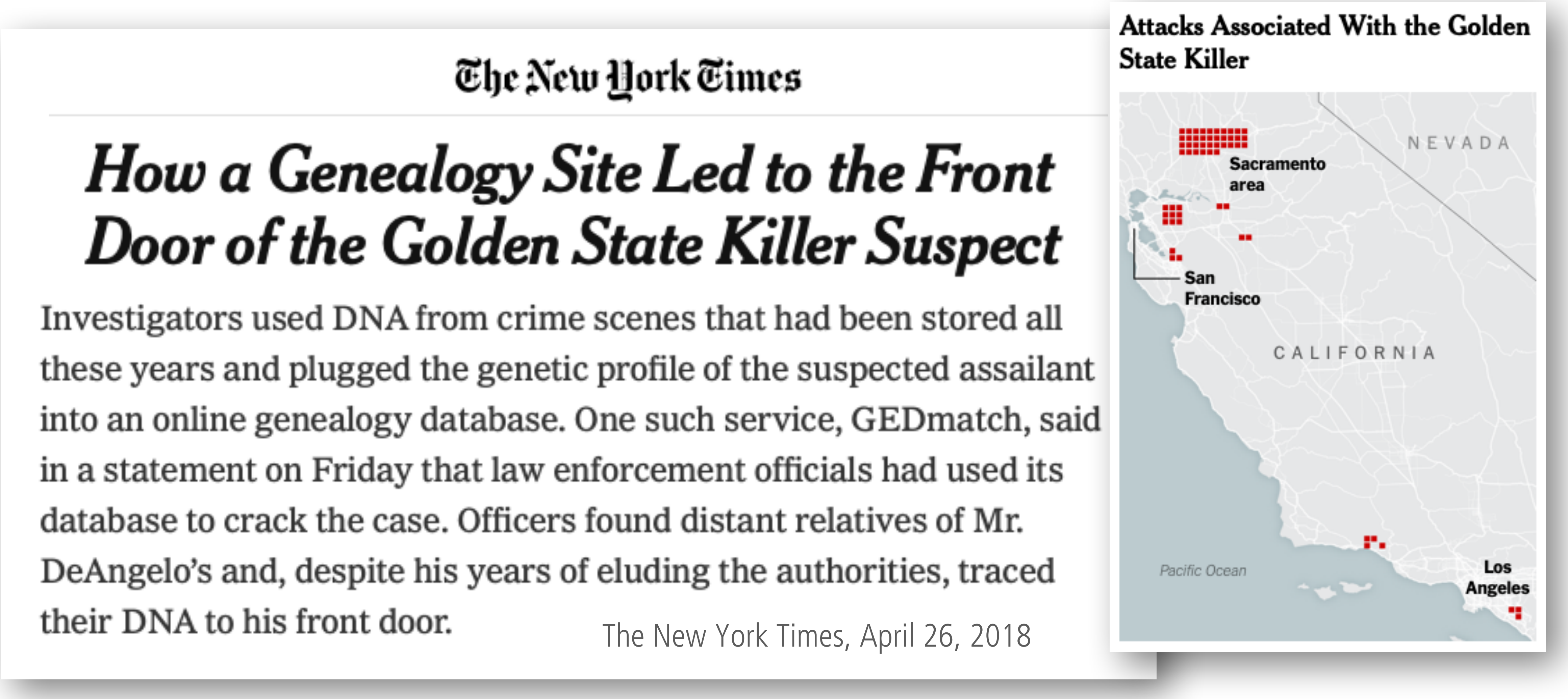
Daily Journal

Devaughn had never been a suspect until genetic genealogy put police on his trail several months ago. Earlier this year, police sent the DNA profile to Parabon, a private genetics company, to compare the suspect's DNA sample to a public genealogy DNA database looking for people with similar DNA profiles who might be kin to the suspect. That eventually led authorities to look at Devaughn.

Rienzi man charged with 1990 Starkville murder

By William Moore Daily Journal 15 hrs ago Comments

© Copyright 2018 Daily Journal, 1242 S Green St Tupelo, MS



The New York Times

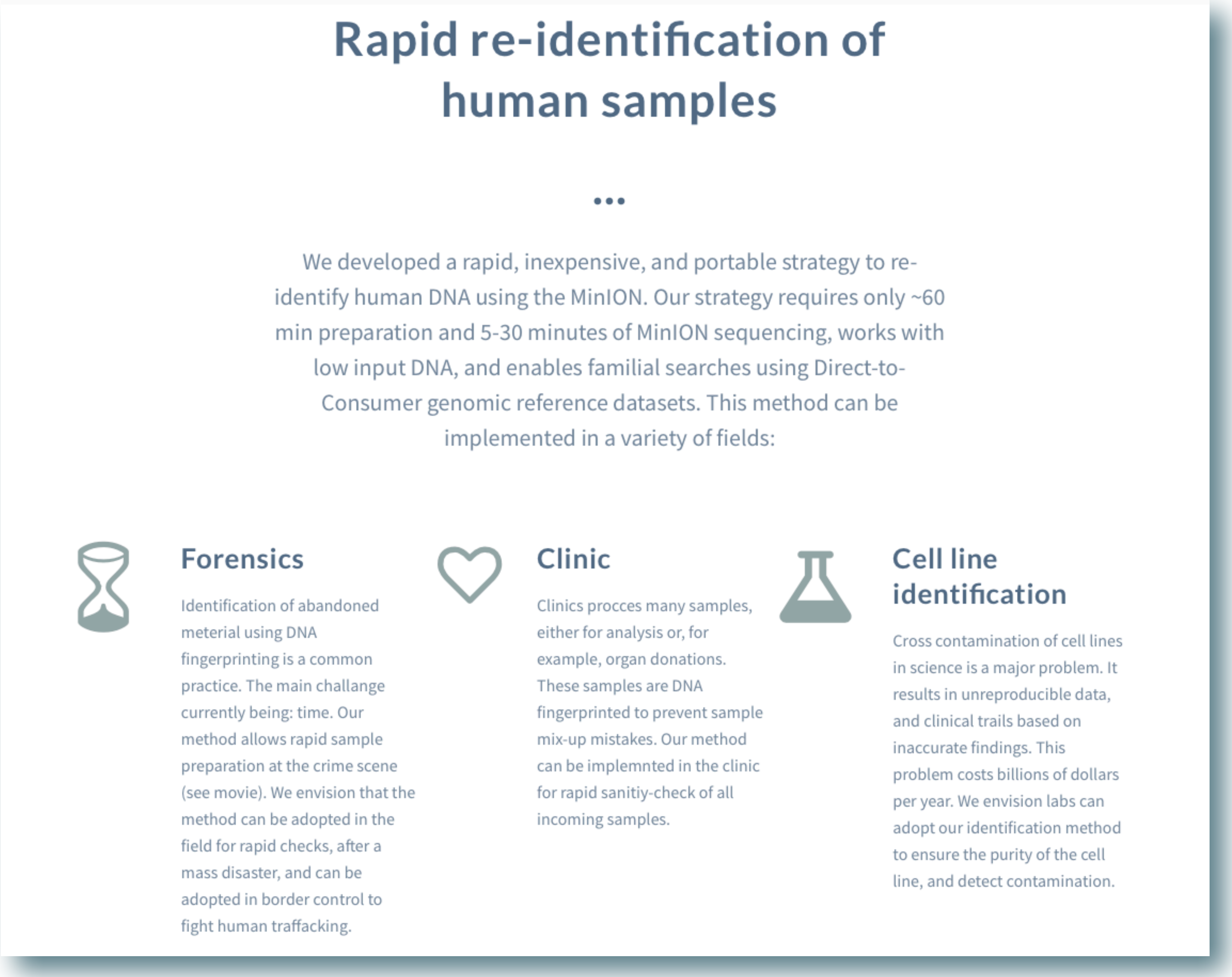
How a Genealogy Site Led to the Front Door of the Golden State Killer Suspect

Investigators used DNA from crime scenes that had been stored all these years and plugged the genetic profile of the suspected assailant into an online genealogy database. One such service, GEDmatch, said in a statement on Friday that law enforcement officials had used its database to crack the case. Officers found distant relatives of Mr. DeAngelo's and, despite his years of eluding the authorities, traced their DNA to his front door.

The New York Times, April 26, 2018

Attacks Associated With the Golden State Killer

But genotyping itself is for professional labs, right?



Rapid re-identification of human samples

We developed a rapid, inexpensive, and portable strategy to re-identify human DNA using the MinION. Our strategy requires only ~60 min preparation and 5-30 minutes of MinION sequencing, works with low input DNA, and enables familial searches using Direct-to-Consumer genomic reference datasets. This method can be implemented in a variety of fields:

Forensics

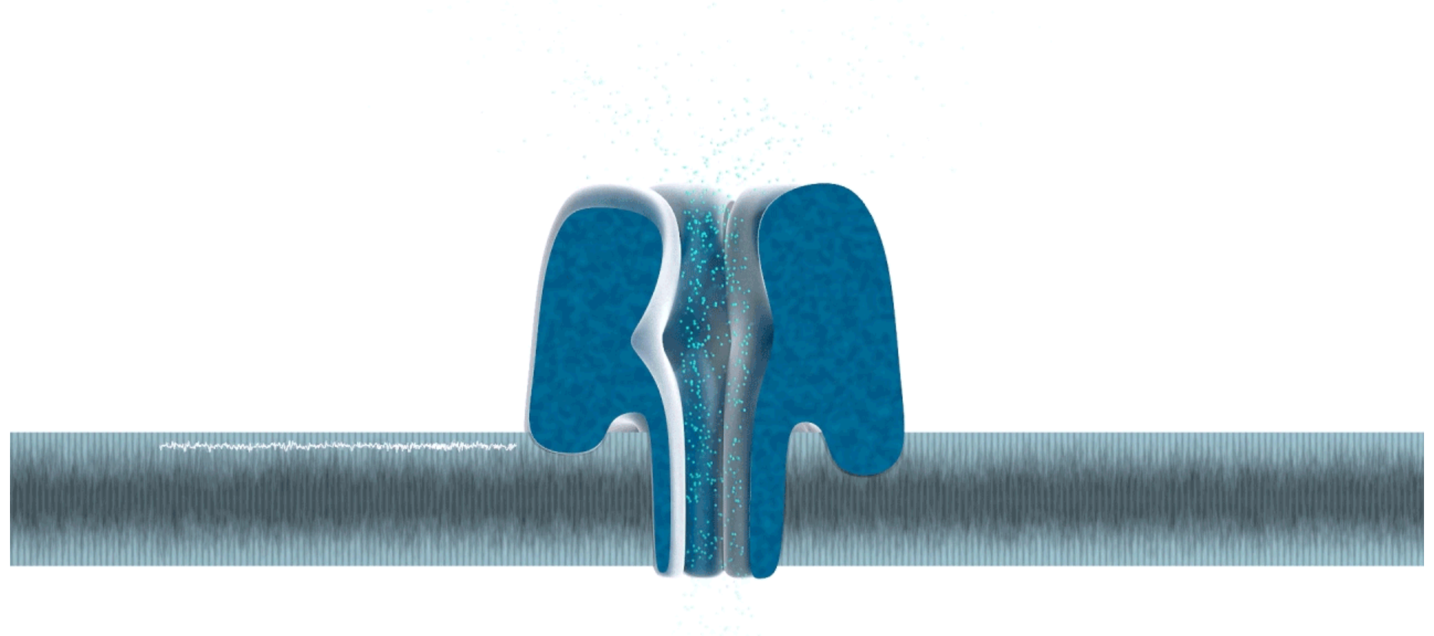
Identification of abandoned meterial using DNA fingerprinting is a common practice. The main challange currently being: time. Our method allows rapid sample preparation at the crime scene (see movie). We envision that the method can be adopted in the field for rapid checks, after a mass disaster, and can be adopted in border control to fight human traffacking.

Clinic

Clinics procces many samples, either for analysis or, for example, organ donations. These samples are DNA fingerprinted to prevent sample mix-up mistakes. Our method can be implemnted in the clinic for rapid sanitiy-check of all incoming samples.

Cell line identification

Cross contamination of cell lines in science is a major problem. It results in unreproducible data, and clinical trails based on inaccurate findings. This problem costs billions of dollars per year. We envision labs can adopt our identification method to ensure the purity of the cell line, and detect contamination.

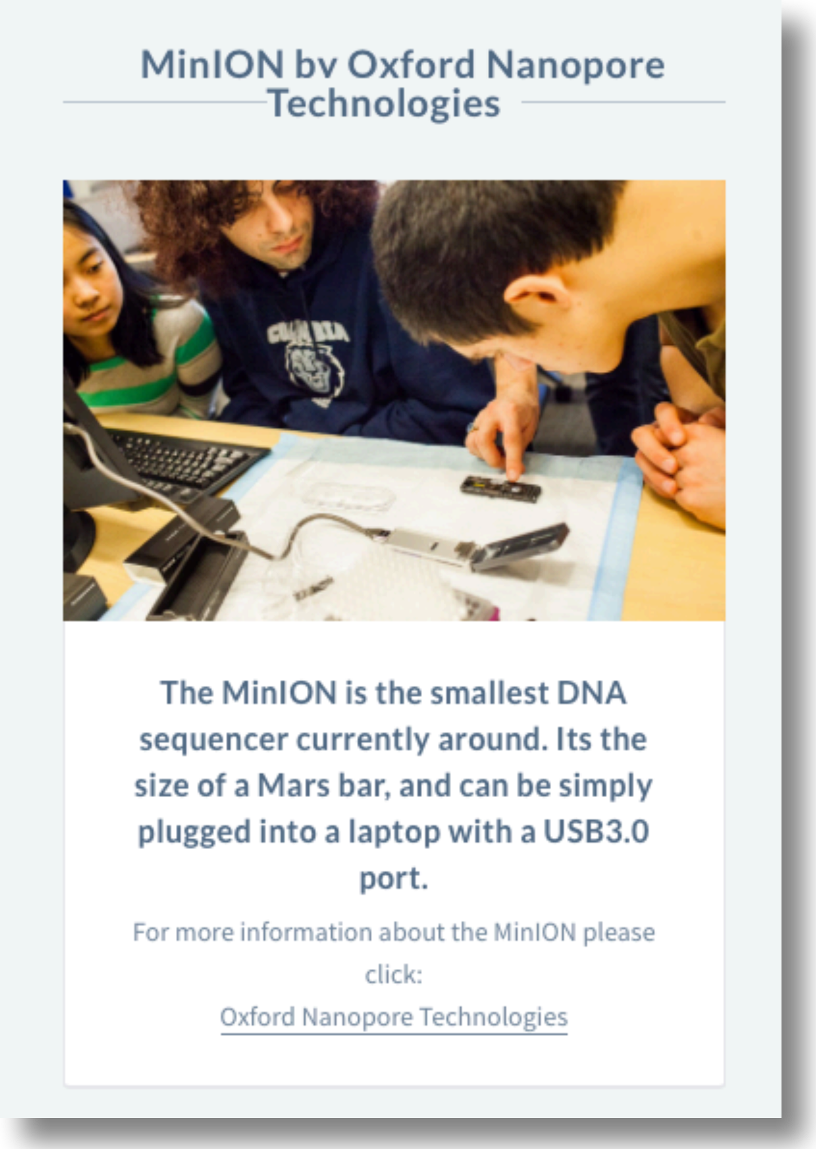


The MinION (Oxford Nanopore)
Source: Sophie Zaaijer

DEMOCRATIZING DNA FINGERPRINTING

Sophie Zaaijer, Assaf Gordon, Robert Piccone, Daniel Speyer, Yaniv Erlich, 2016

ddf.teamerlich.org

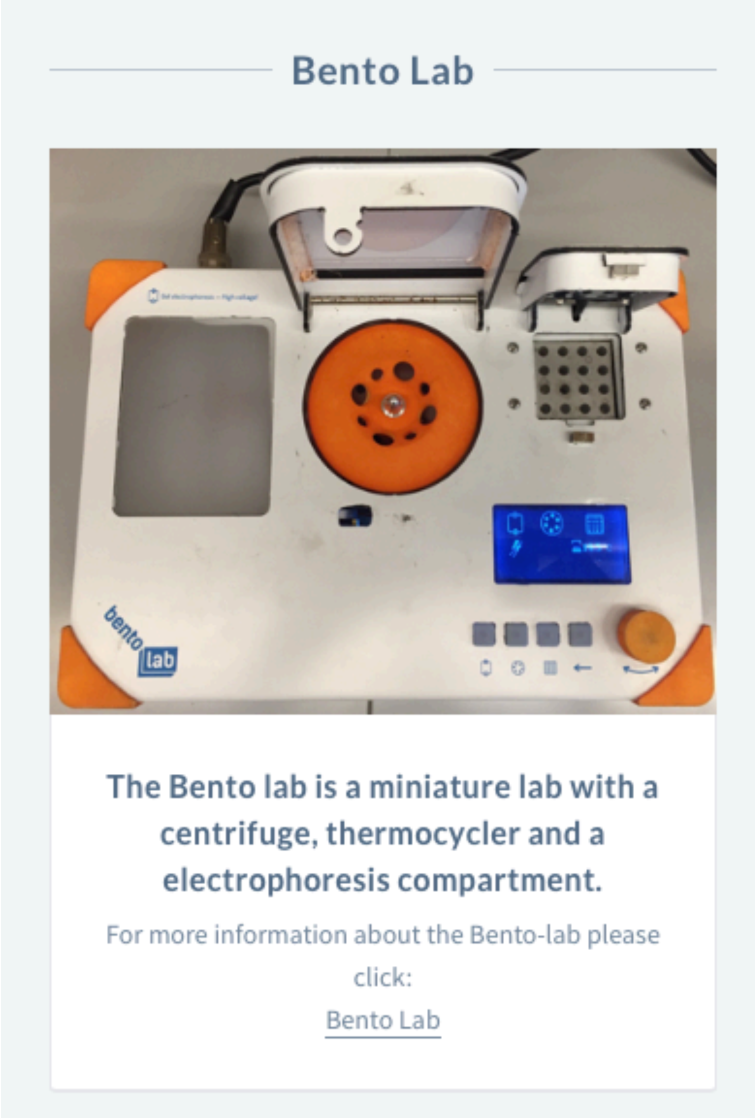


MinION by Oxford Nanopore Technologies

The MinION is the smallest DNA sequencer currently around. Its the size of a Mars bar, and can be simply plugged into a laptop with a USB3.0 port.

For more information about the MinION please click:

Oxford Nanopore Technologies

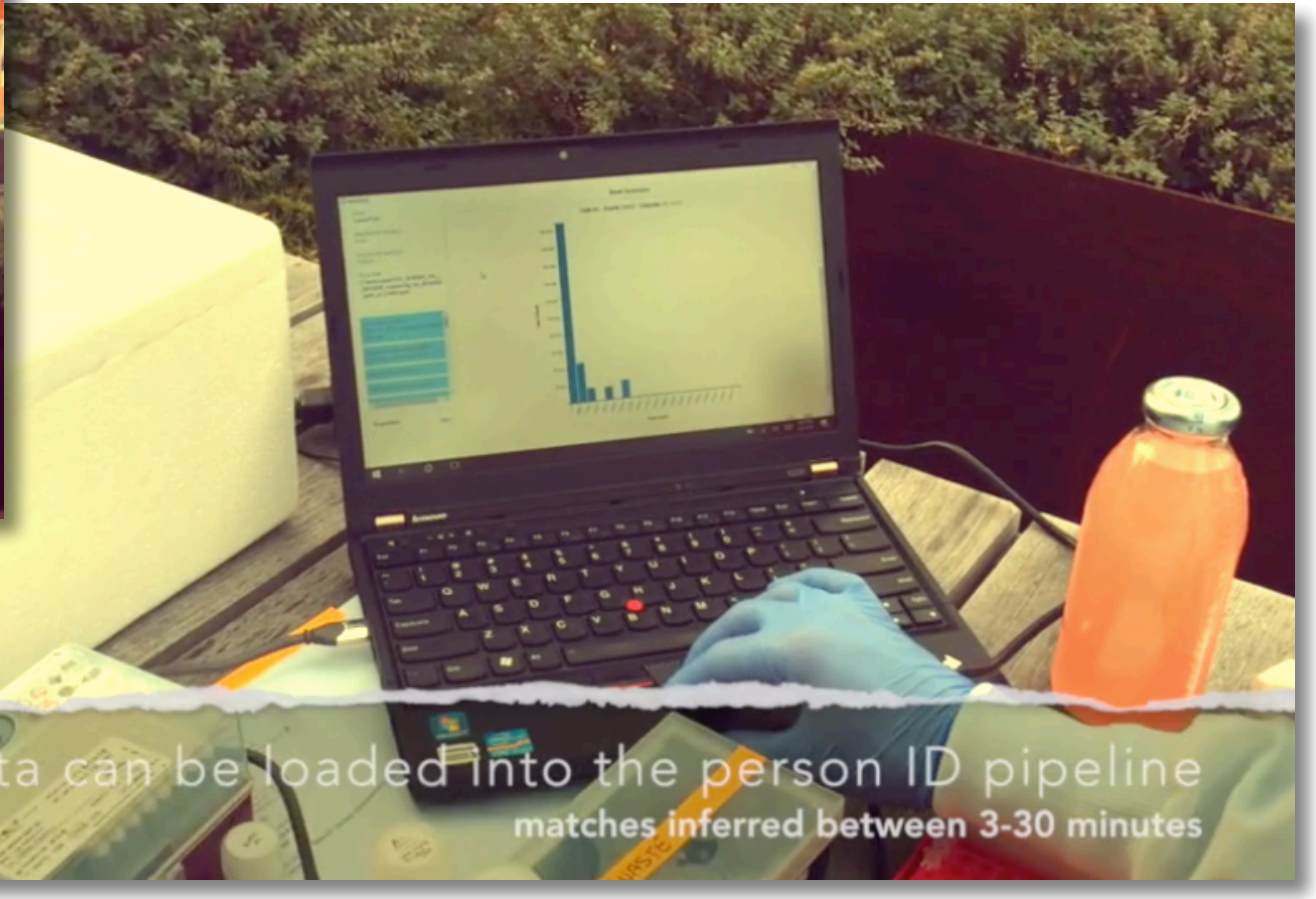


Bento Lab

The Bento lab is a miniature lab with a centrifuge, thermocycler and a electrophoresis compartment.

For more information about the Bento-lab please click:

Bento Lab


Data can be loaded into the person ID pipeline
matches inferred between 3-30 minutes

DNA sequencing for identification/fingerprinting soon "commodity" technology (in contrast with technological/data challenges in "precision medicine")

Generalkonsent

BENEFIT

BLOCKCHAIN

HEALTH

PRIVACY

SECURITY

CONSENT

ACCESS

Right to Research

HACKERS

LAWS

Genetic
Information
Nondiscrimination
Act

Health
Insurance
Portability and
Accountability
Act

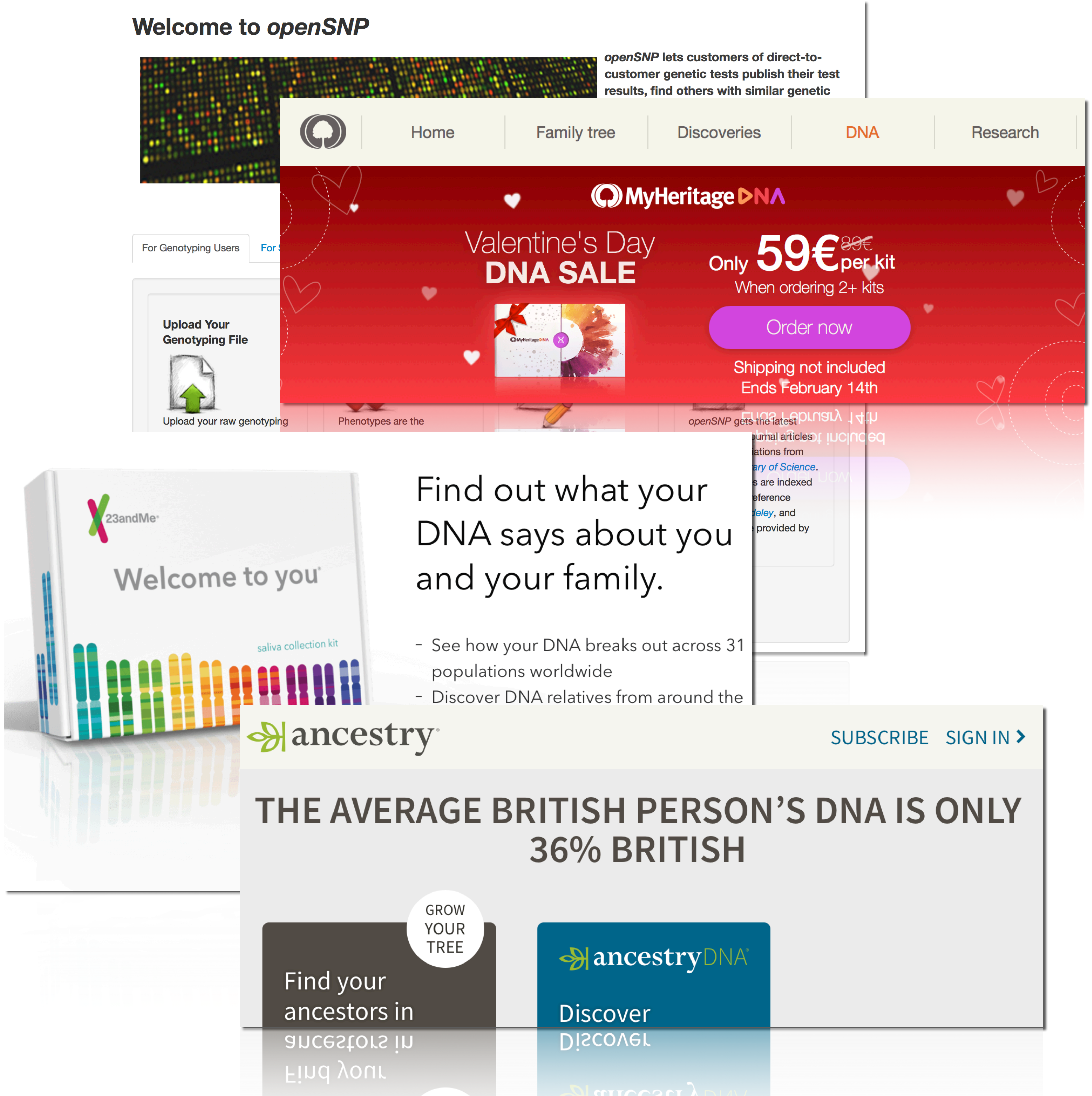
SAFETY

CRYPTOGRAPHY

Share *YOUR* Genome data?

- The Beacon concept - balanced approach for accessing genome variant data from internationally distributed resources
- However: Genome data has the inherent “risk” of being identified and linked to a person

Solutions from Technology or Society? Discourse!



Welcome to *openSNP*

openSNP lets customers of direct-to-customer genetic tests publish their test results, find others with similar genetic

Upload Your Genotyping File

Upload your raw genotyping

Phenotypes are the

openSNP gets the latest

urnal articles

ations from

ary of Science.

s are indexed

eference

eley, and

provided by

Valentine's Day DNA SALE

Only 59€ ~~89€~~ per kit

When ordering 2+ kits

Shipping not included

Ends February 14th

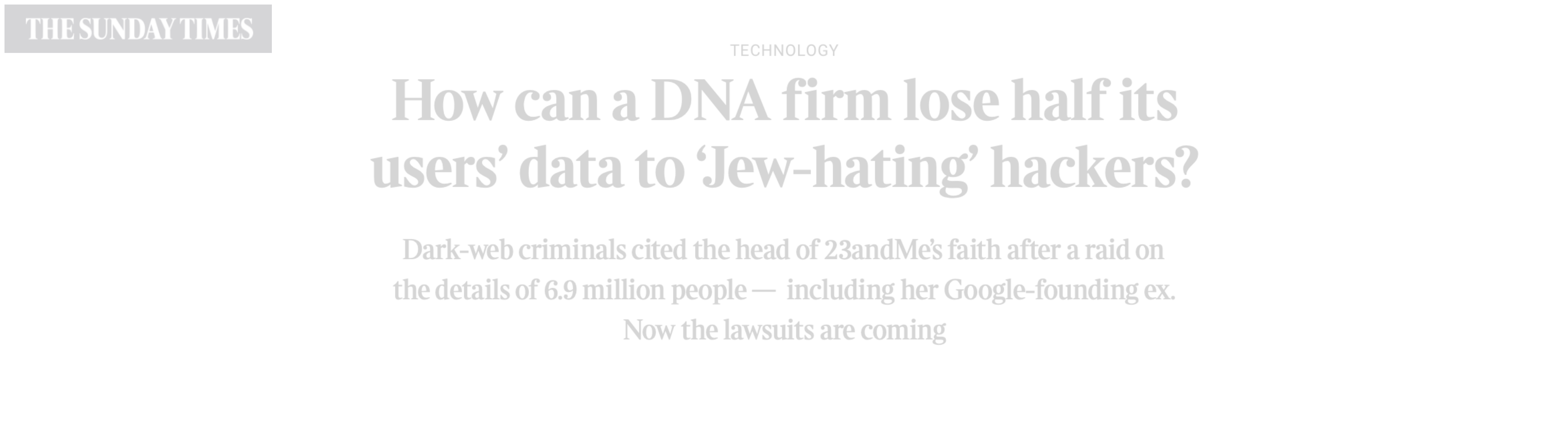
Find out what your DNA says about you and your family.

- See how your DNA breaks out across 31 populations worldwide
- Discover DNA relatives from around the

THE AVERAGE BRITISH PERSON’S DNA IS ONLY 36% BRITISH

Find your ancestors in

Discover

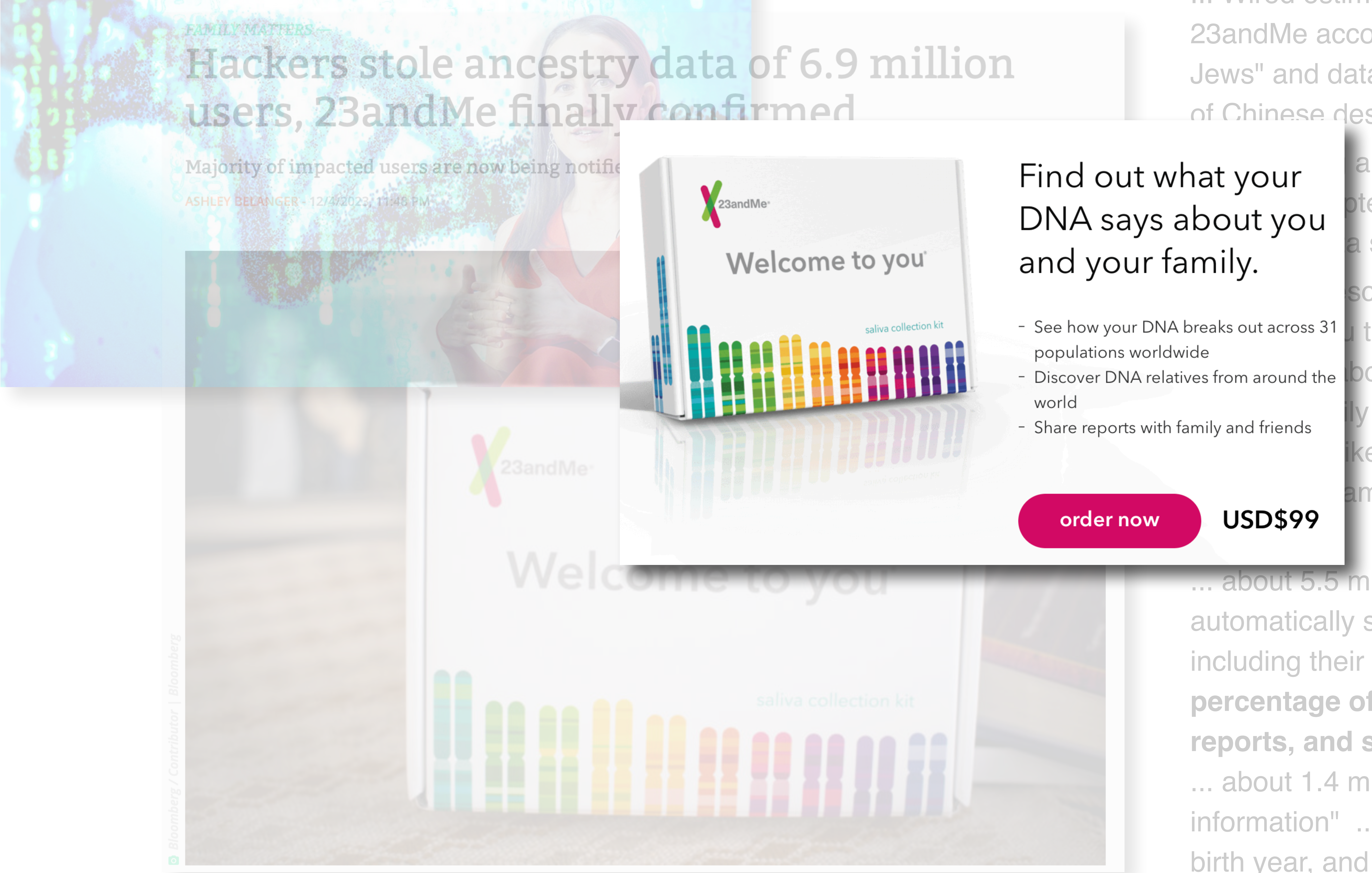


THE SUNDAY TIMES

TECHNOLOGY

How can a DNA firm lose half its users' data to 'Jew-hating' hackers?

Dark-web criminals cited the head of 23andMe's faith after a raid on the details of 6.9 million people — including her Google-founding ex. Now the lawsuits are coming



FAMILY MATTERS —

Hackers stole ancestry data of 6.9 million users, 23andMe finally confirmed

Majority of impacted users are now being notifie

ASHLEY BELANGER - 12/4/2023, 11:48 PM

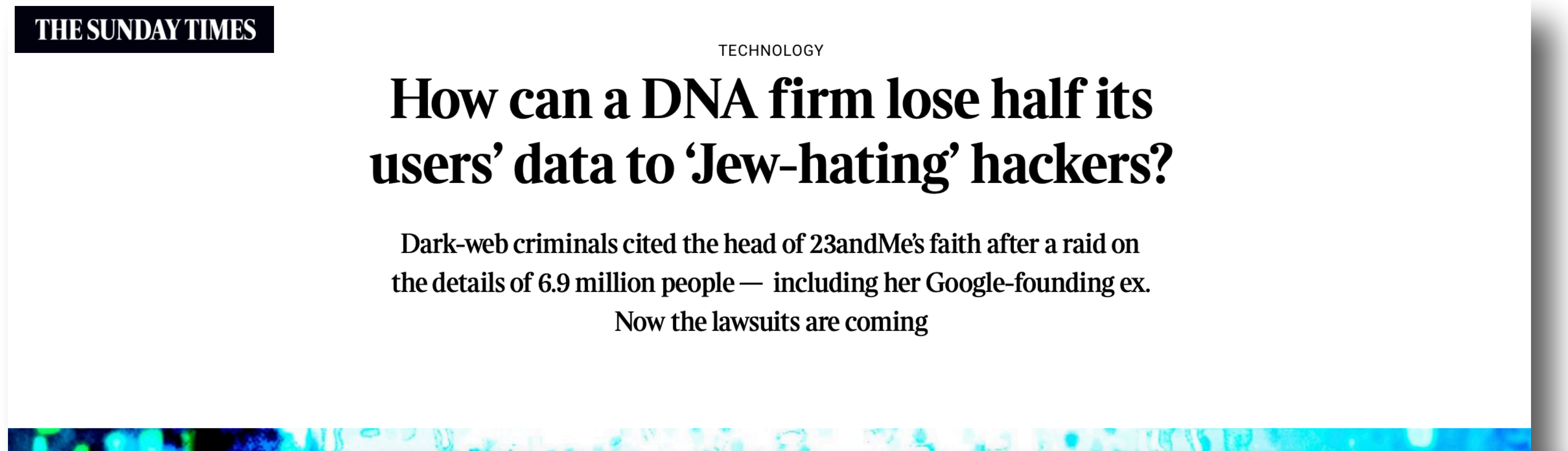
It has now been confirmed that an additional **6.9 million 23andMe users had ancestry data stolen** after hackers accessed thousands of accounts by likely reusing previously leaked passwords.

... Wired estimated that "at least a million data points from 23andMe accounts" that were "exclusively about Ashkenazi Jews" and data points from "hundreds of thousands of users of Chinese descent" seemed to be exposed.

a spokesperson to confirm that two groups of
pted into the **DNA Relatives feature** had their
a stolen.

scribes the DNA Relatives feature as ...
u to find and connect with genetic relatives and
bout your family." By **opting in**, users hope to
ly members by **willingly** giving others access to
ike their birth year, current location, and
ames and birth locations. Users can opt out at

... about 5.5 million users, was hacked after opting in to automatically sharing information with DNA Relatives, including their "**name, birth year, relationship labels, the percentage of DNA shared with relatives, ancestry reports, and self-reported location**," TechCrunch reported. ... about 1.4 million users, shared "Family Tree profile information" ... including display names, relationship labels, birth year, and self-reported location, TechCrunch reported.

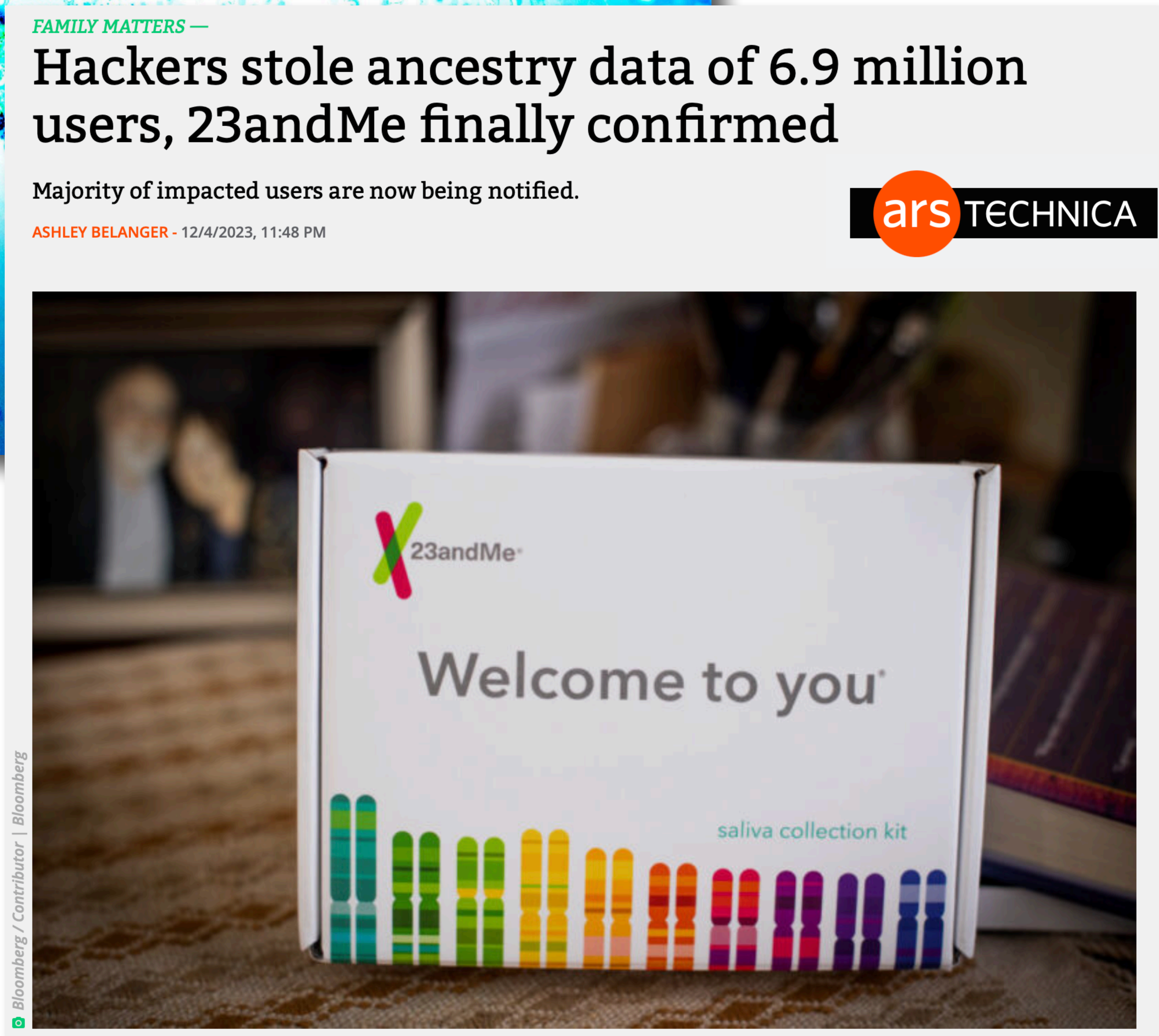


THE SUNDAY TIMES

TECHNOLOGY

How can a DNA firm lose half its users' data to 'Jew-hating' hackers?

Dark-web criminals cited the head of 23andMe's faith after a raid on the details of 6.9 million people — including her Google-founding ex. Now the lawsuits are coming



FAMILY MATTERS —

Hackers stole ancestry data of 6.9 million users, 23andMe finally confirmed

Majority of impacted users are now being notified.

ASHLEY BELANGER - 12/4/2023, 11:48 PM

ars TECHNICA

Bloomberg / Contributor | Bloomberg

It has now been confirmed that an additional **6.9 million 23andMe users had ancestry data stolen** after hackers accessed thousands of accounts by likely reusing previously leaked passwords.

... Wired estimated that "at least a million data points from 23andMe accounts" that were "exclusively about Ashkenazi Jews" and data points from "hundreds of thousands of users of Chinese descent" seemed to be exposed.

... prompting a spokesperson to confirm that two groups of users who opted into the **DNA Relatives feature** had their personal data stolen.

23andMe describes the DNA Relatives feature as ... "allowing you to find and connect with genetic relatives and learn more about your family." By **opting in**, users hope to find lost family members by **willingly** giving others access to information like their birth year, current location, and ancestors' names and birth locations. Users can opt out at any time ...

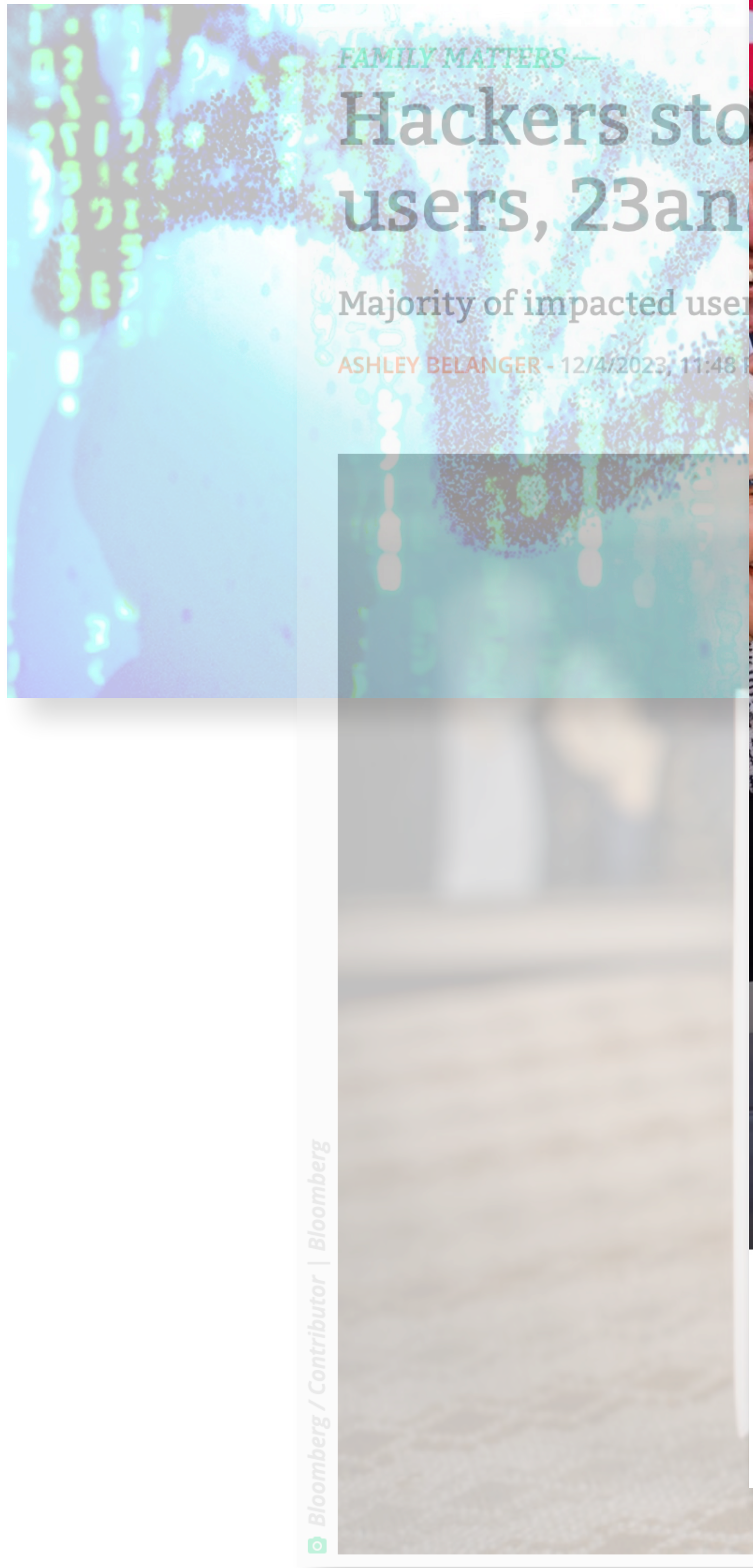
... about 5.5 million users, was hacked after opting in to automatically sharing information with DNA Relatives, including their "**name, birth year, relationship labels, the percentage of DNA shared with relatives, ancestry reports, and self-reported location**," TechCrunch reported. ... about 1.4 million users, shared "Family Tree profile information" ... including display names, relationship labels, birth year, and self-reported location, TechCrunch reported.

THE SUNDAY TIMES

TECHNOLOGY

How can a DNA fir
users' data to 'Jew-ha

Dark-web criminals cited the head of 23a
the details of 6.9 million people — includi
Now the lawsuits are



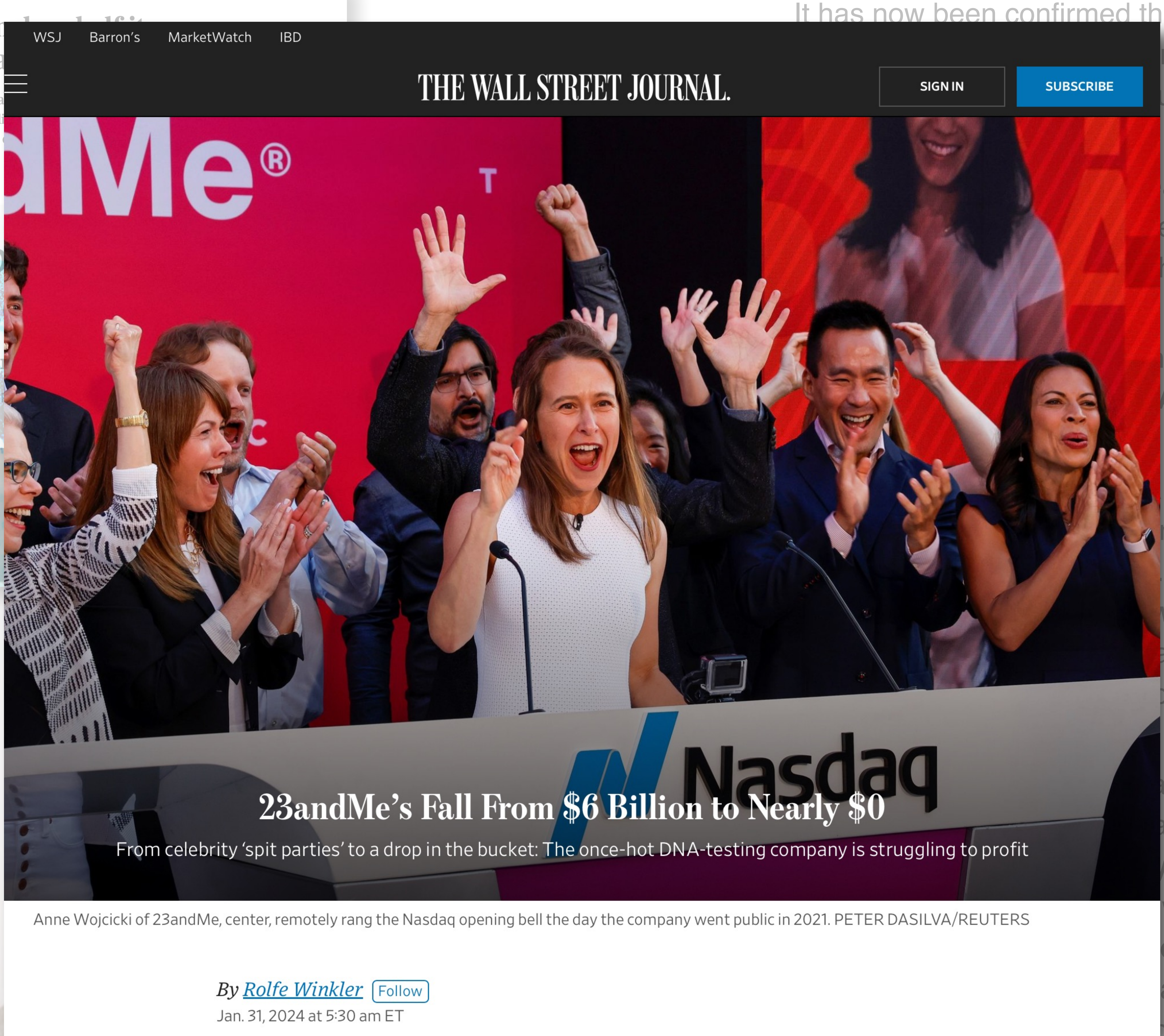
FAMILY MATTERS —

Hackers sto
users, 23an

Majority of impacted use

ASHLEY BELANGER - 12/4/2023, 11:48

Bloomberg / Contributor | Bloomberg



23andMe's Fall From $6 Billion to Nearly $0

From celebrity 'spit parties' to a drop in the bucket: The once-hot DNA-testing company is struggling to profit

Anne Wojcicki of 23andMe, center, remotely rang the Nasdaq opening bell the day the company went public in 2021. PETER DASILVA/REUTERS

By Rolfe Winkler Follow
Jan. 31, 2024 at 5:30 am ET

It has now been confirmed that an additional **6.9 million**
y data stolen after hackers
unts by likely reusing previously

ast a million data points from
e "exclusively about Ashkenazi
hundreds of thousands of users
to be exposed.

to confirm that two groups of
A Relatives feature had their

Relatives feature as ...
nect with genetic relatives and
" By **opting in**, users hope to
willingly giving others access to
ar, current location, and
ocations. Users can opt out at

s hacked after opting in to
ation with DNA Relatives,
year, relationship labels, the
with relatives, ancestry
ocation," TechCrunch reported.
ared "Family Tree profile
lay names, relationship labels,
birth year, and self-reported location, TechCrunch reported.

Universal Declaration of Human Rights (1948)

27(1) "The Right to Science"

"Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and **to share in scientific advancement and its benefits.**"

27(2) "The Right to Recognition"

"Everyone has the right to the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author."

Universität Zürich^UZH

SIB

Swiss Institute of Bioinformatics

elixir