Data Discovery in Biomedical Genomics
Time for a New Paradigm

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

Universität Zürich[UZH]

Swiss Institute of Bioinformatics

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

Genomics has seen massive and ongoing changes in technology

200+ genomic data initiatives globally

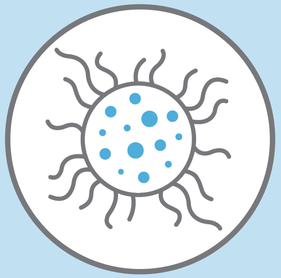
Clinical/Genomic Medicine

Research

National

Cohorts

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

progenetix

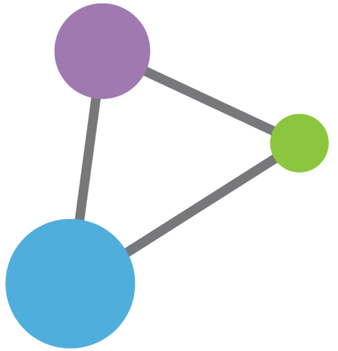
European Genome-Phenome Archive

SPHN Swiss Personalized Health Network

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

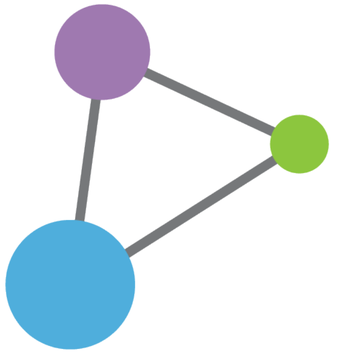
Centralized Genomic Knowledge Bases

European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

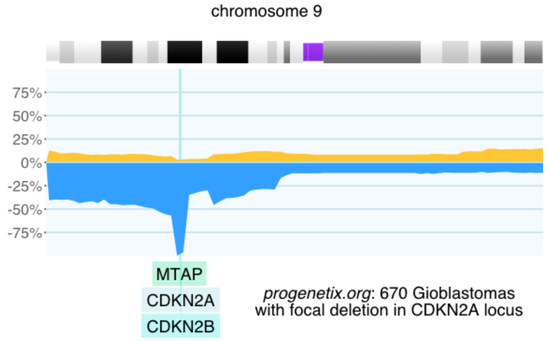
SPHN Swiss Personalized Health Network

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



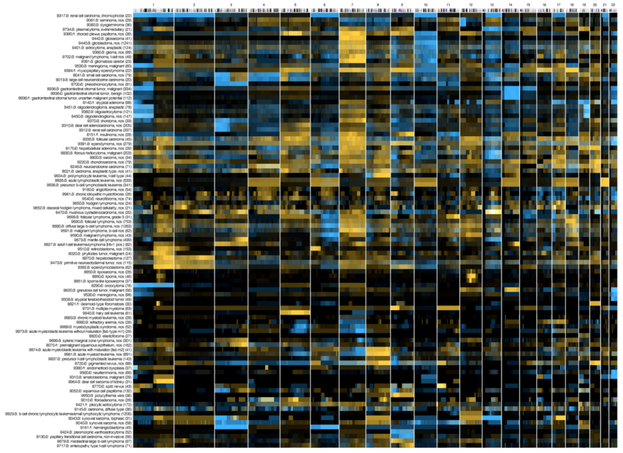
Linkage of distributed and disparate datasets

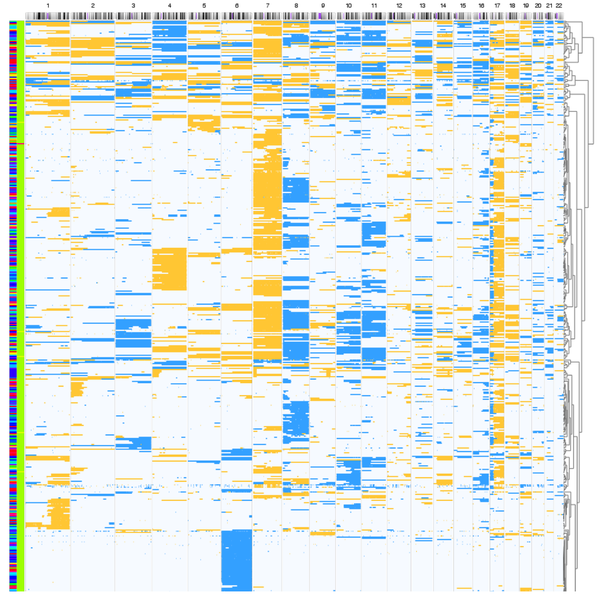
Theoretical Cytogenetics and Oncogenomics Research | Methods | Standards

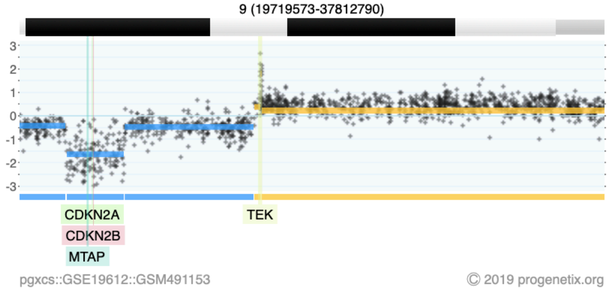


Genomic Imbalances in Cancer - Copy Number Variations (CNV)

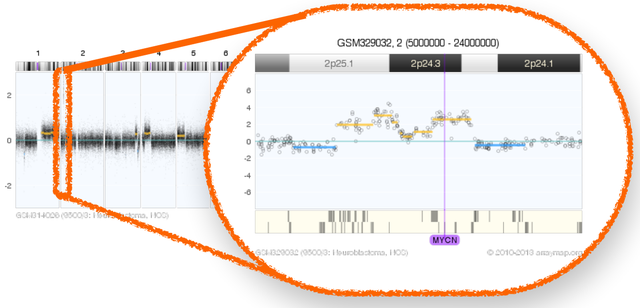
- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)







2-event, homozygous deletion in a Glioblastoma



MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

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]

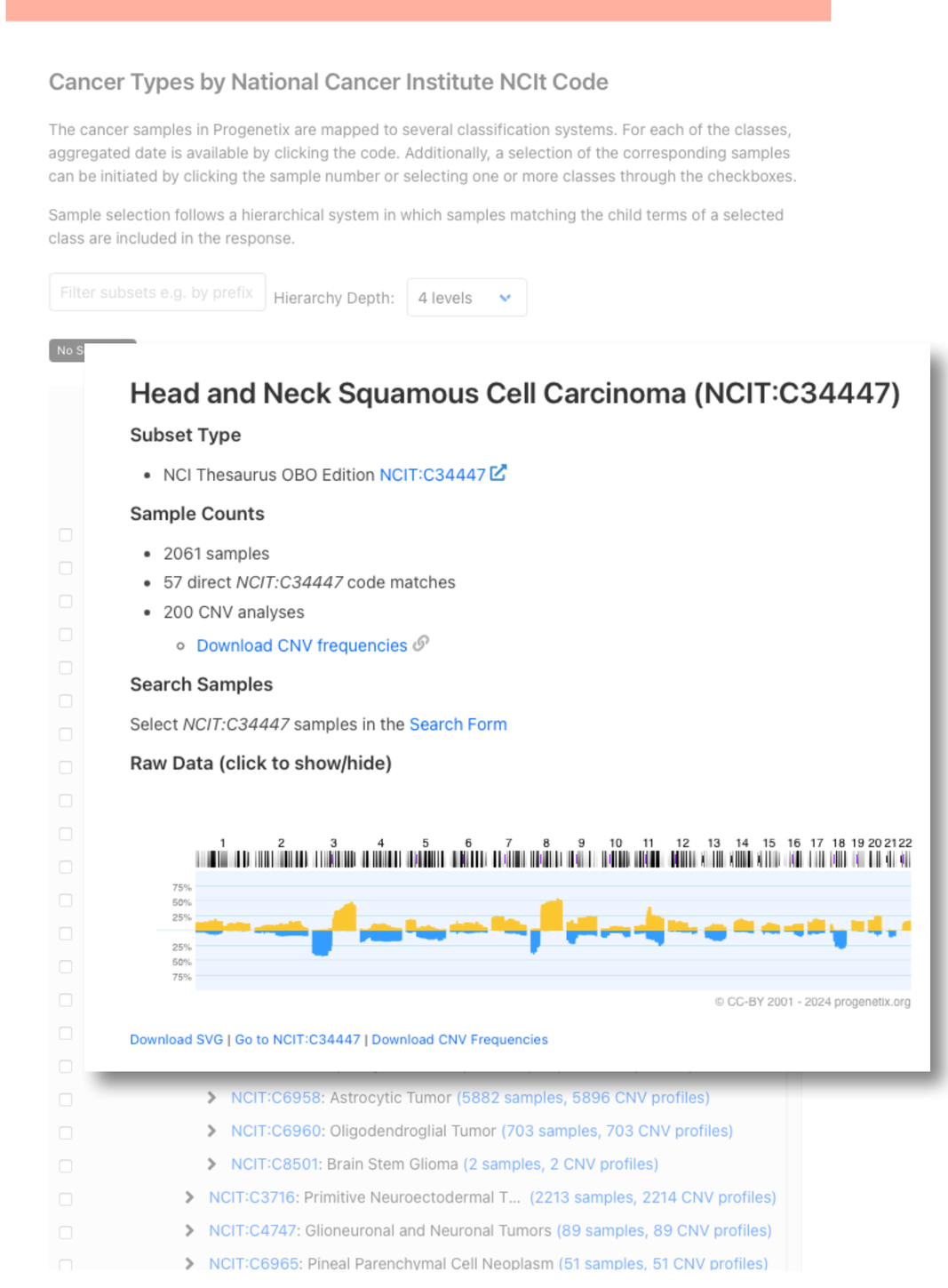
progenetix

SIB Swiss Institute of Bioinformatics

progenetix.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]

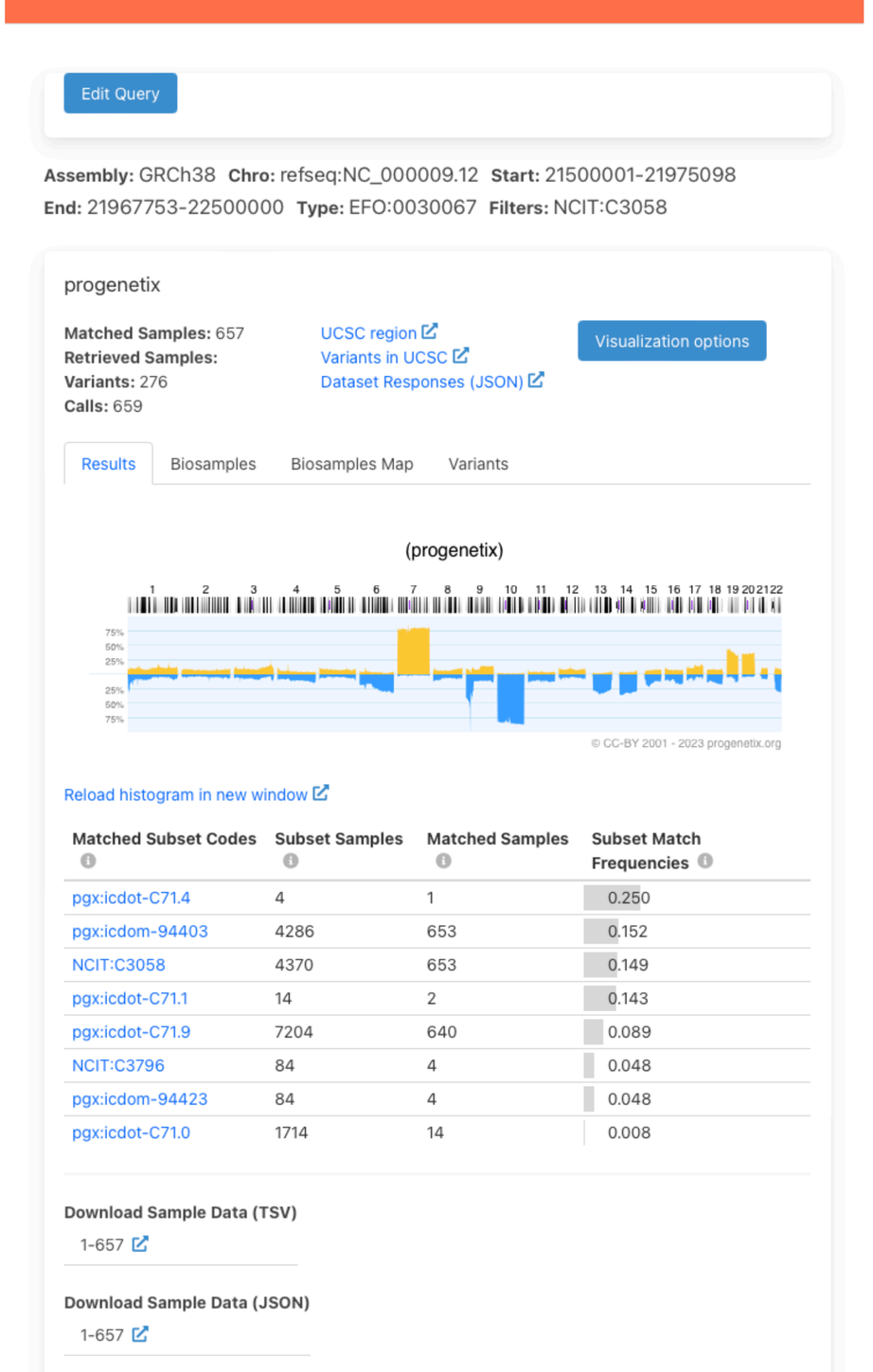
progenetix

SIB Swiss Institute of Bioinformatics

progenetix.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

Cancer Cell Lines

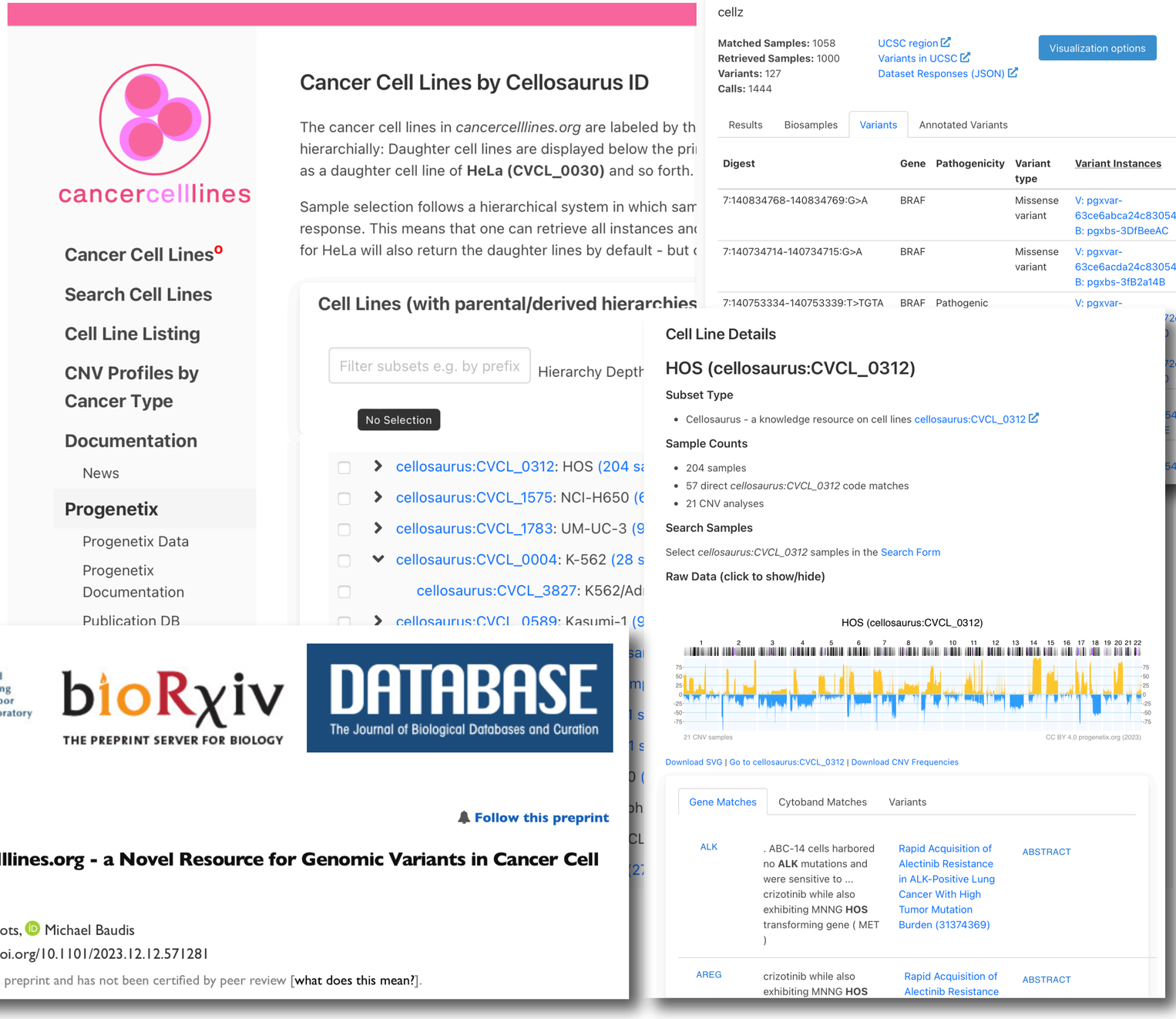
Cancer Genomics Reference Resource

- 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



Different Approaches to Data Sharing

progenetix

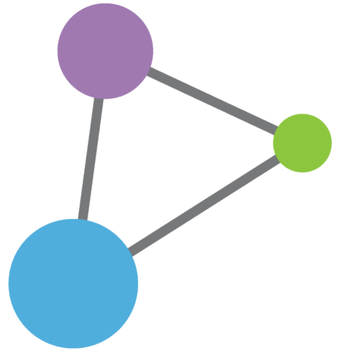
Centralized Genomic Knowledge Bases

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

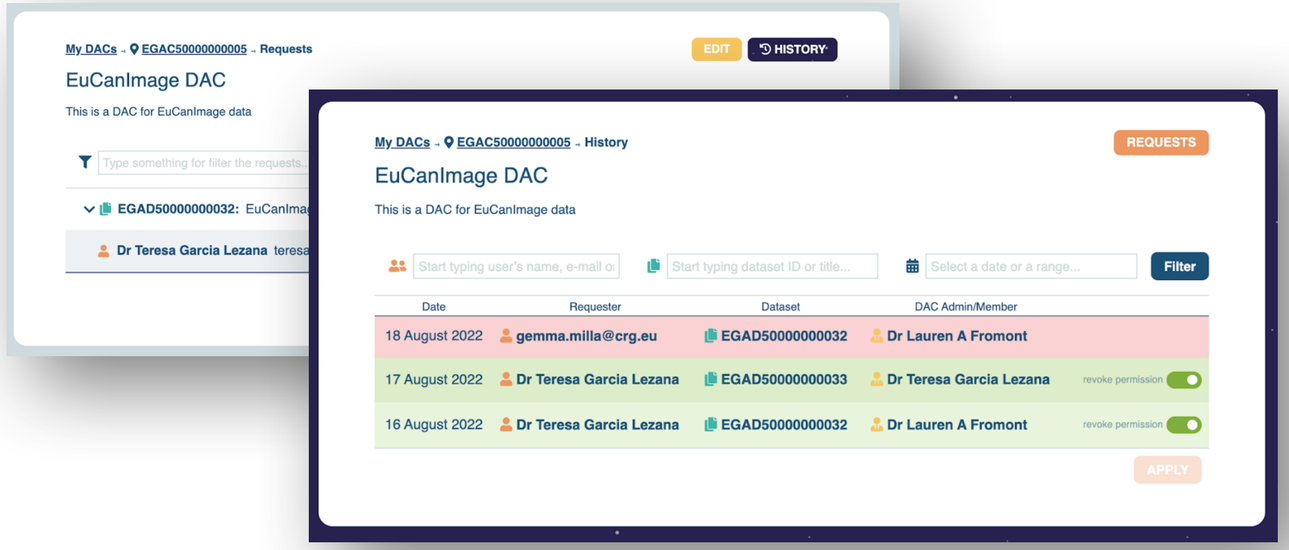
DRIVER PROJECT

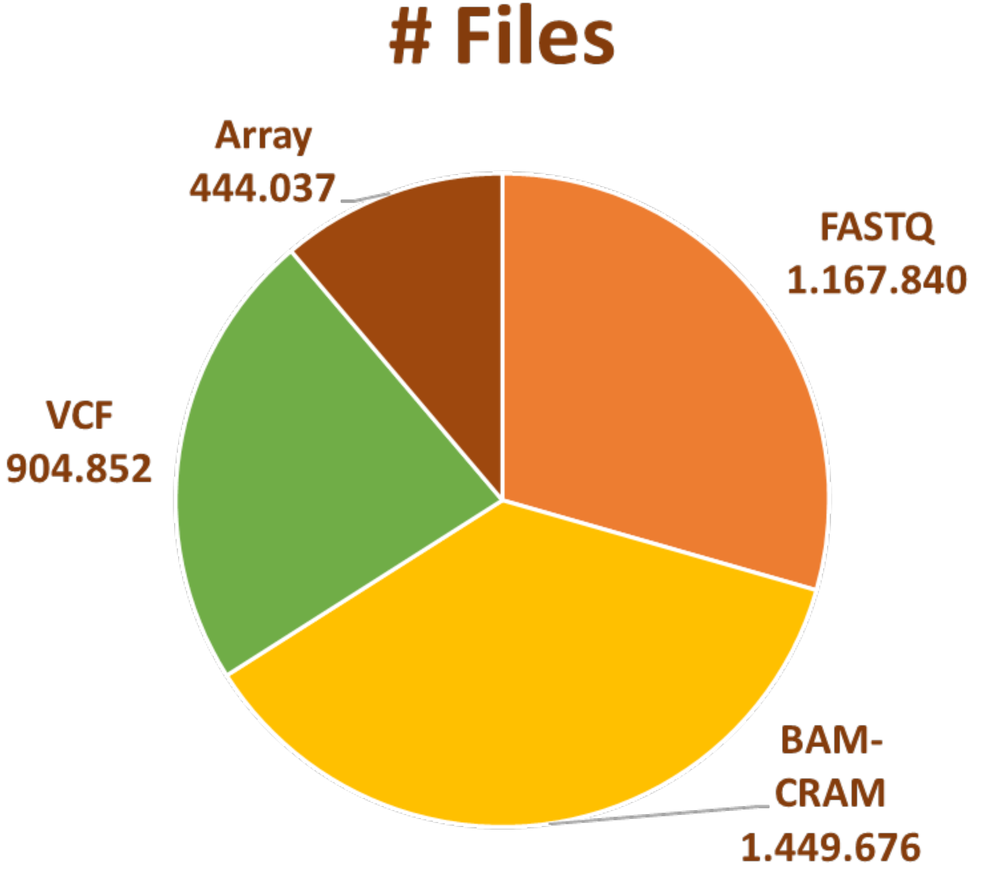
Slide: adapted from Jordi Rambla@ GA4GH 2023

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





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

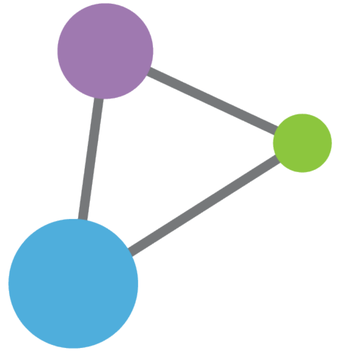
Centralized Genomic Knowledge Bases

European Genome-Phenome Archive

Data Commons
Trusted, controlled repository of multiple datasets

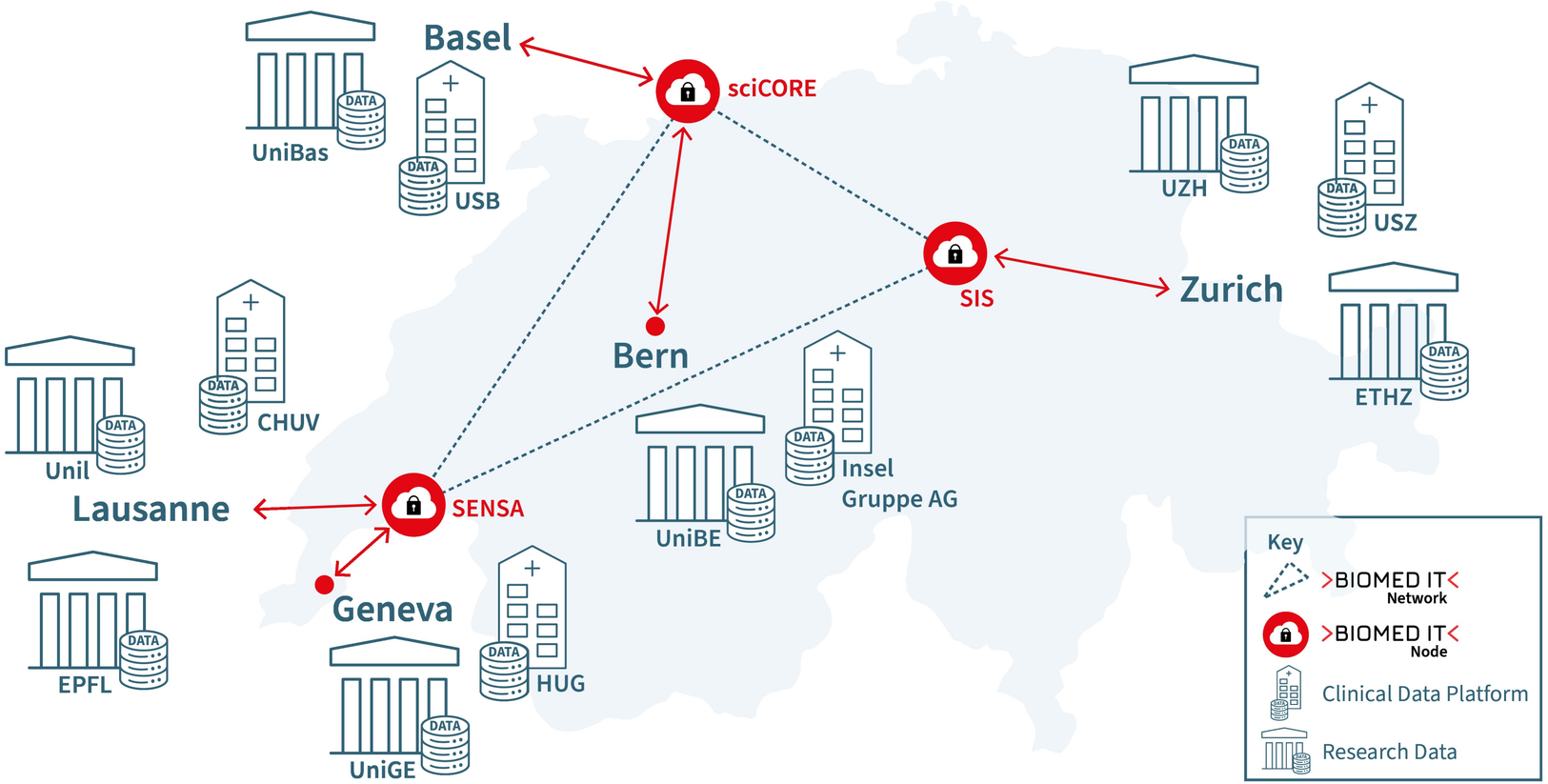
SPHN Swiss Personalized Health Network

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



Linkage of distributed and disparate datasets

The Swiss Personalized Health Network



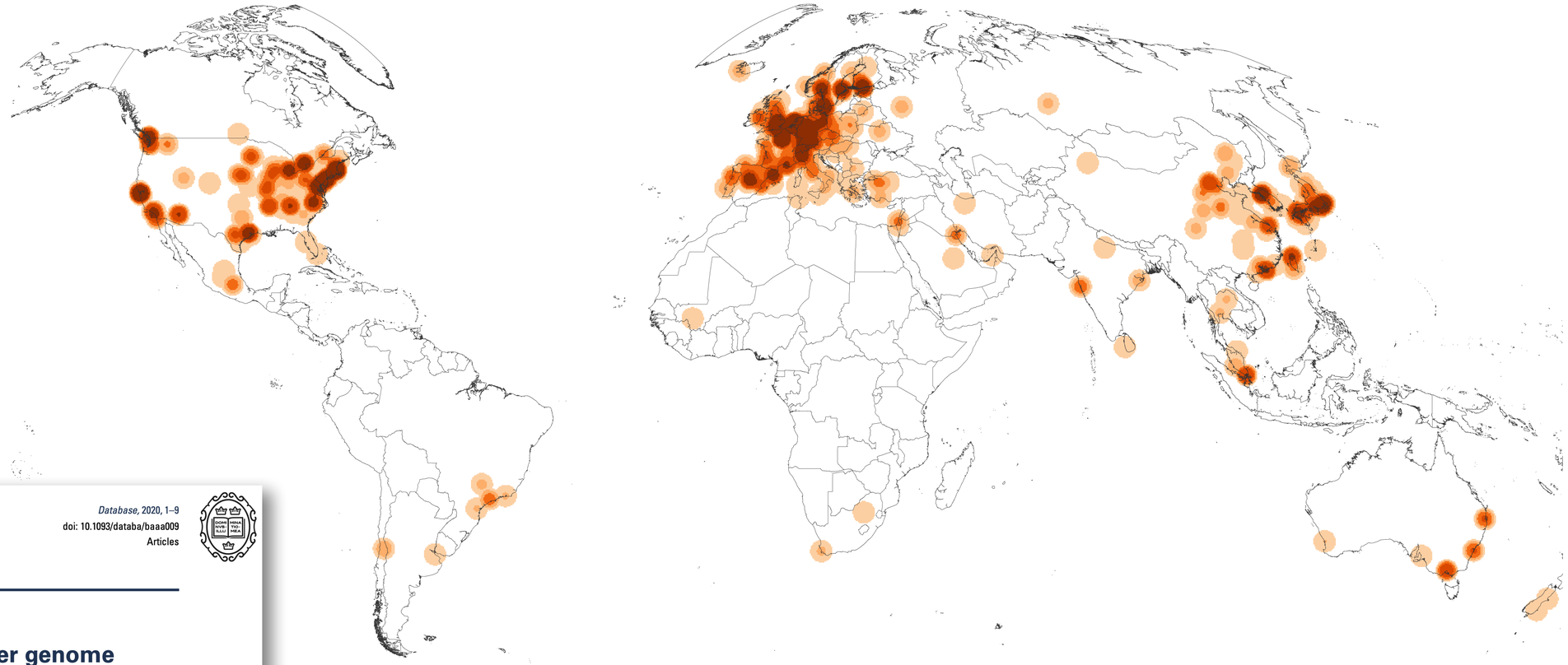
Personalized Health Informatics Group

SPHN Data Coordination Center (DCC)
BioMedIT Network

Strategic Focus Area
Personalized Health and Related Technologies

Where Does Cancer Genomic Data Come From?

Geographic bias in published cancer genome profiling studies



DATABASE
The Journal of Biological Databases and Curation

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

Articles

Geographic assessment of cancer genome profiling studies

Paula Carrio-Cordo[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.

Different Approaches to Data Sharing

progenetix

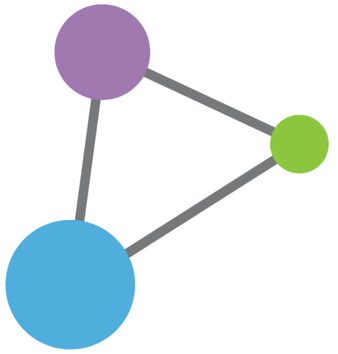
European Genome-Phenome Archive

SPHN Swiss Personalized Health Network

Centralized Genomic Knowledge Bases

Data Commons
Trusted, controlled repository of multiple datasets

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



Linkage of distributed and disparate datasets

Federation

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

GENOMICS

A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

Alignment with other standards organizations

Global Alliance for Genomics & Health

Research → Clinical research → Clinical care

Our funders, partners, and Driver Projects

Global Alliance for Genomics & Health

Core Funders

CIHR IRSC Canadian Institutes of Health Research / Instituts de recherche en santé du Canada

UKRI Medical Research Council

GenomeCanada

NIH National Human Genome Research Institute

NIHR | National Institute for Health Research

Wellcome

Host Institutions

OICR Ontario Institute for Cancer Research

McGill University — Victor Phillip Dahdaleh Institute of Genomic Medicine at McGill University

Broad Institute

Wellcome Sanger Institute

EMBL-EBI

Supporting Funders

abbvie

aws

DNAStack

illumina

Ontario Brain Institute / Institut ontarien du cerveau

Roche

Assigned Expert Funders/Employers

NIH National Human Genome Research Institute

University of Cape Town / iYunivesithi yaseKapa · Universiteit van Kaapstad

Strategic Partner

elixir

Driver Projects

All of Us Research Program

Autism Speaks

ClinGen Clinical Genome Resource

Matchmaker Exchange

epi share

Genomics England

VICC

elixir beacon network

NIH National Cancer Institute

monarch initiative

NHLBI BioData Catalyst®

Human Pangenome Project

RIFGC

NCPI

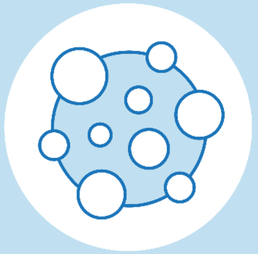
eosc cancer

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

Driver Project areas of focus

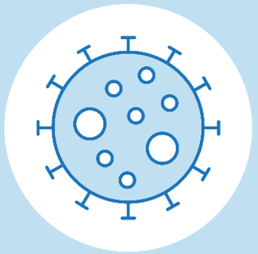
Global Alliance
for Genomics & Health

DRIVER PROJECT AREAS OF FOCUS



17 Driver Projects
focus on **cancer**

19 Driver Projects
focus on **rare disease**



12 Driver Projects
focus on **infectious disease**

15 Driver Projects
focus on **common disease**

TYPES OF DRIVER PROJECTS

23 Driver Projects
in **academic research**

19 Driver Projects
in **clinical research**

14 Driver Projects
in **industry**

7 Driver Projects
in **government**

GA4GH Driver Projects

Global Alliance for Genomics & Health

All of Us Research Program

Autism Sharing Initiative

European Joint Programme - RD

ClinGen

CanDIG

European Genome-phenome Archive

H3Africa

Genomics England

Australian Genomics

Variant Interpretation for Cancer Consortium

Human Cell Atlas

ELIXIR Beacon Project

ELIXIR Cloud & AAI

NCI CRDC

Matchmaker Exchange

Epishare

ICGC ARGO

Monarch Initiative

NHLBI BioData Catalyst®

Biomedical Research Hub

imCORE®

Human Pangenome Project

International Precision Child Health Partnership

Repository of the International Fetal Genomics Consortium

NIH Cloud Platform Interoperability effort

Qatar Genome Program

EOSC4Cancer

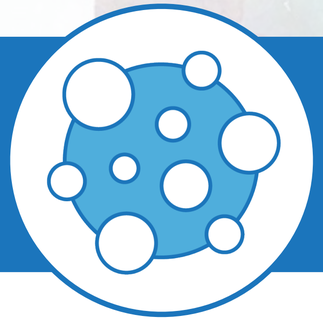
Genomic Data Infrastructure

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

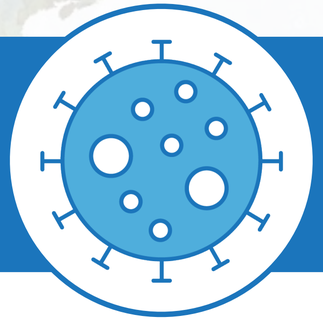
GA4GH Communities of Interest

Global Alliance for Genomics & Health

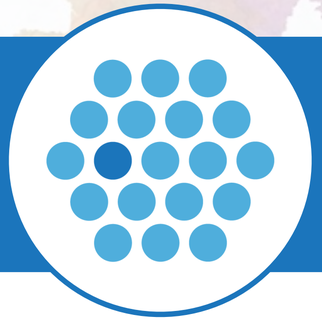
Domain-specific groups promoting global cooperation, data sharing and collaborative research through identifying the need for new standards, and implementing existing GA4GH standards.



Cancer

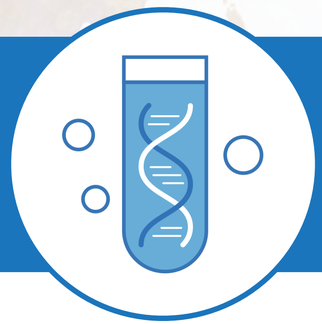


Infectious Disease



Rare Disease

Neuroscience



Clinical Laboratory

Opportunities for collaboration

Learn from others' implementations

Develop a study group

Propose a GIF project

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

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[4]

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]

Cell Genomics

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

OP

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,*]

A New Paradigm for Data Sharing

FROM

TO

Data Copying

Data Visiting

A New Paradigm for Data Sharing

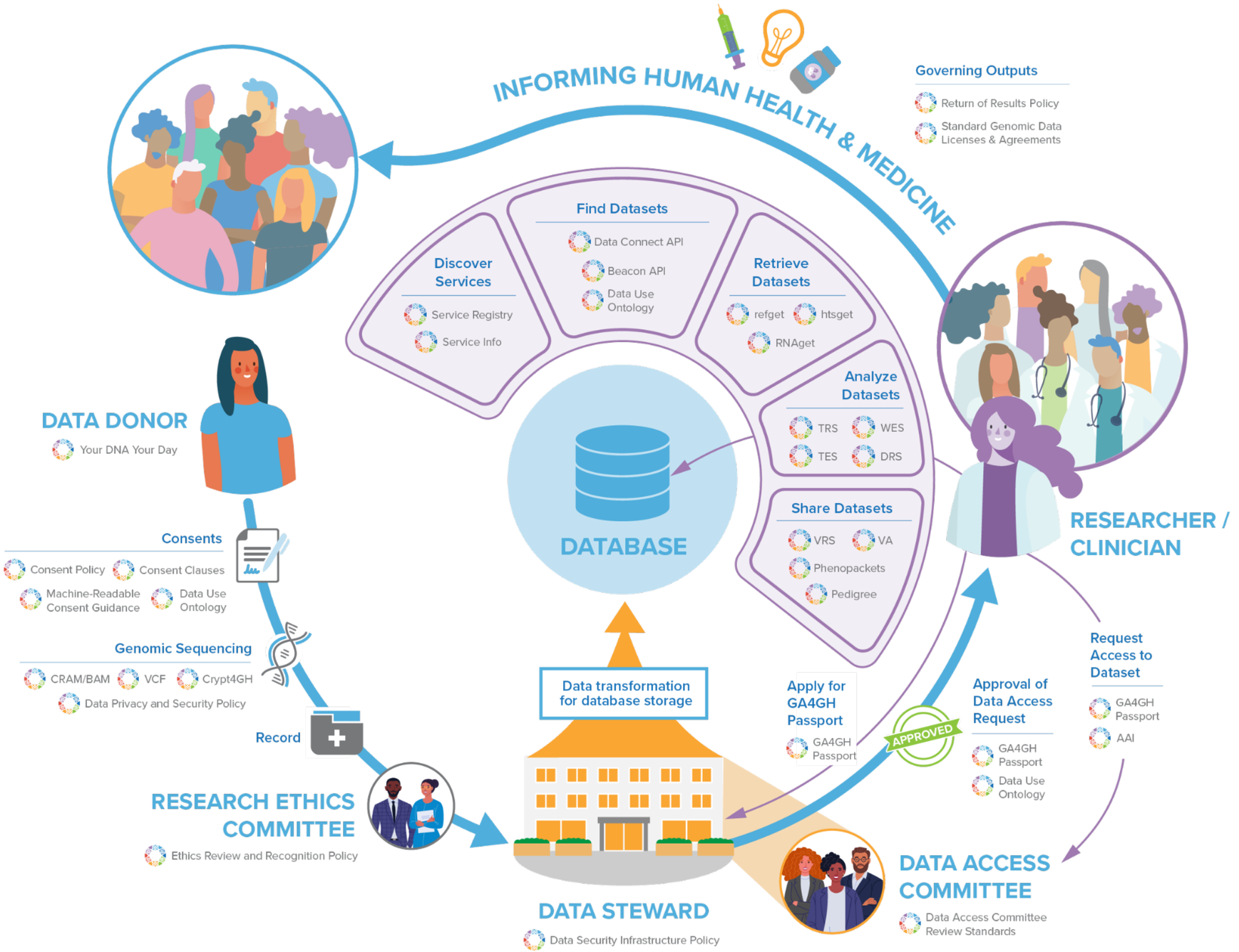
FROM

STANDARDS

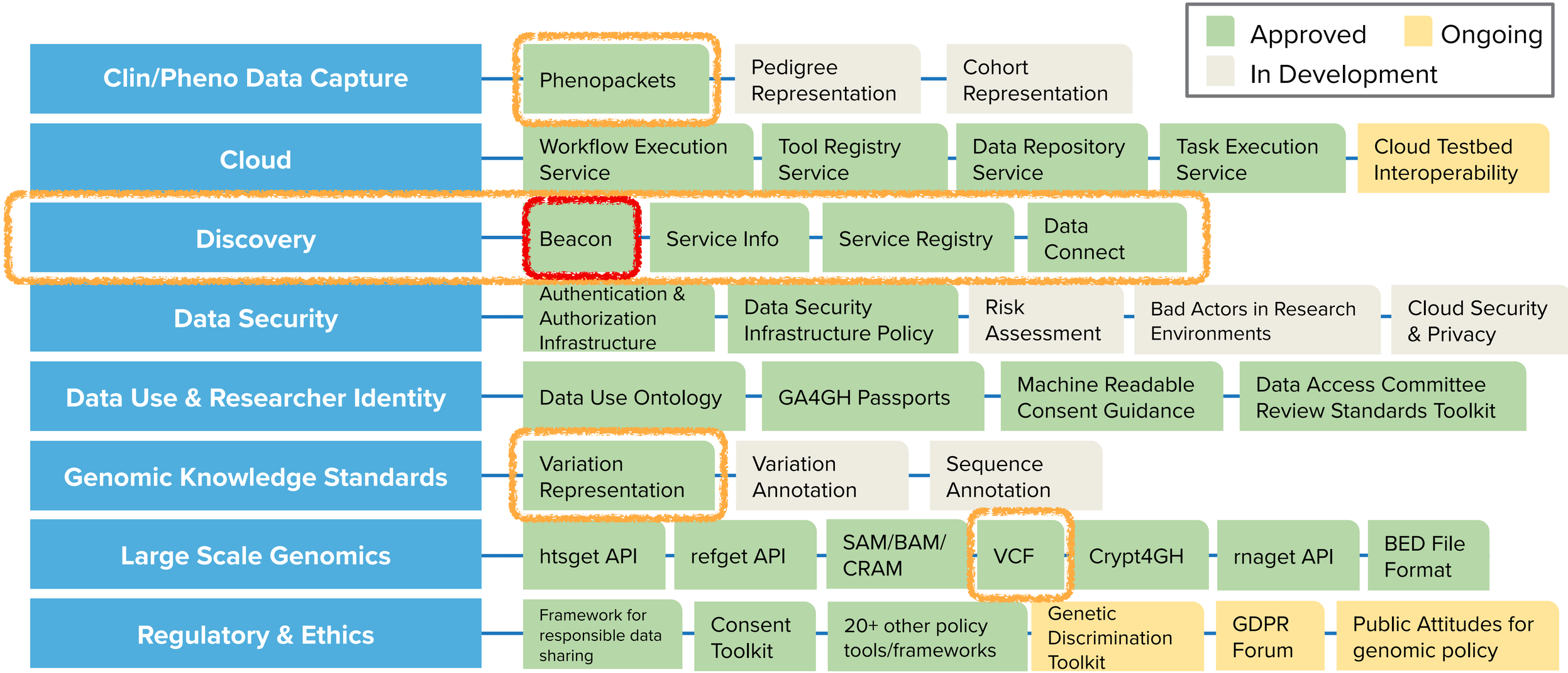
TO

Data Copying

Data Visiting

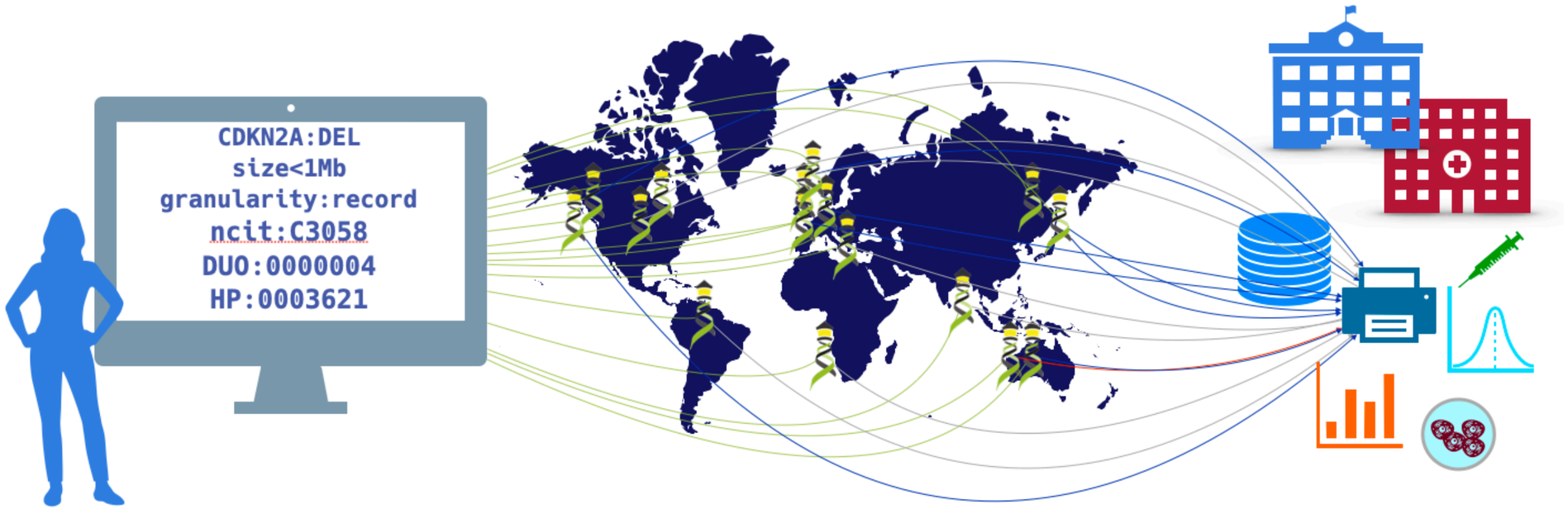

INFORMING HUMAN HEALTH & MEDICINE
Governing Outputs
Return of Results Policy
Standard Genomic Data Licenses & Agreements
Find Datasets
Data Connect API
Beacon API
Data Use Ontology
Discover Services
Service Registry
Service Info
Retrieve Datasets
refget
htsget
RNAget
Analyze Datasets
TRS
WES
TES
DRS
Share Datasets
VRS
VA
Phenopackets
Pedigree
DATABASE
DATA DONOR
Your DNA Your Day
RESEARCHER / CLINICIAN
Consents
Consent Policy
Consent Clauses
Machine-Readable Consent Guidance
Data Use Ontology
Genomic Sequencing
CRAM/BAM
VCF
Crypt4GH
Data Privacy and Security Policy
Record
Data transformation for database storage
Apply for GA4GH Passport
GA4GH Passport
APPROVED
Approval of Data Access Request
GA4GH Passport
Data Use Ontology
Request Access to Dataset
GA4GH Passport
AAI
RESEARCH ETHICS COMMITTEE
Ethics Review and Recognition Policy
DATA STEWARD
Data Security Infrastructure Policy
DATA ACCESS COMMITTEE
Data Access Committee Review Standards

Overview of GA4GH standards and frameworks



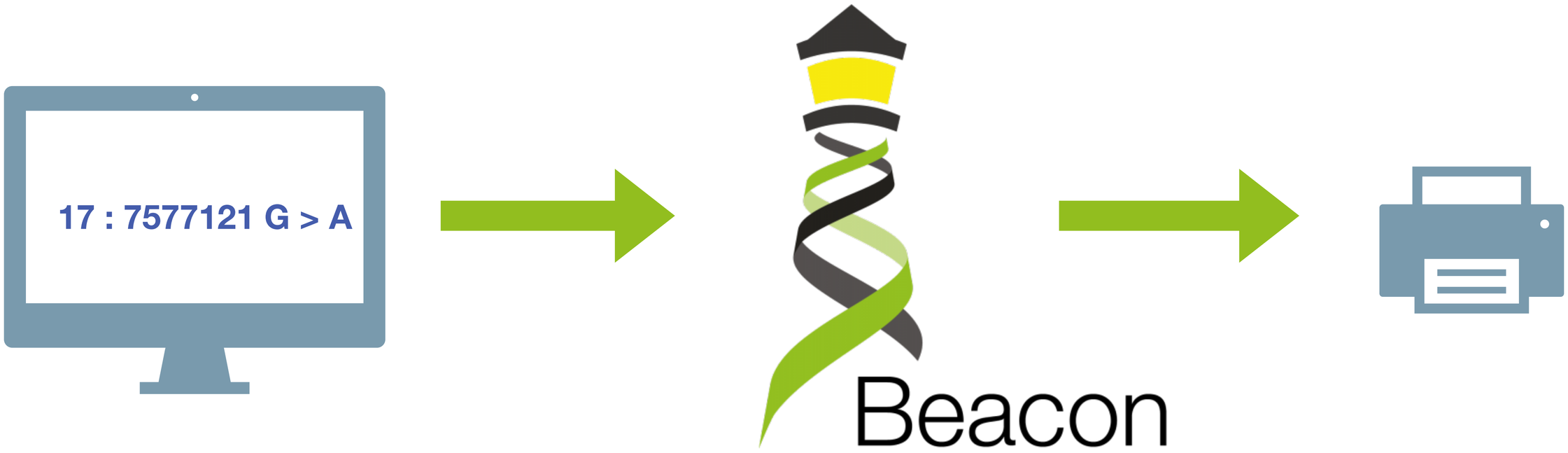
Global Alliance
for Genomics & Health

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**

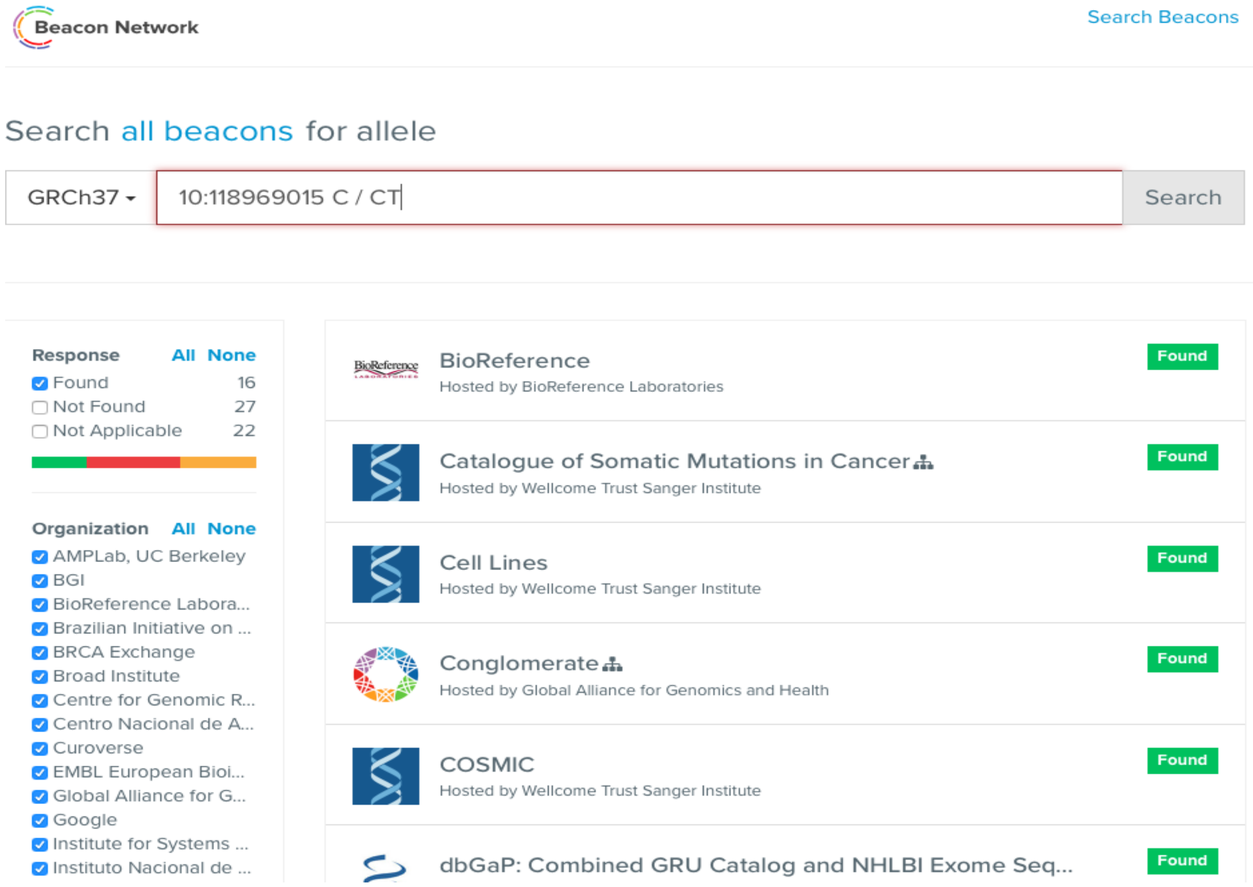
Have you seen this variant? It came up in my patient and we don't know if this is a common SNP or worth following up.

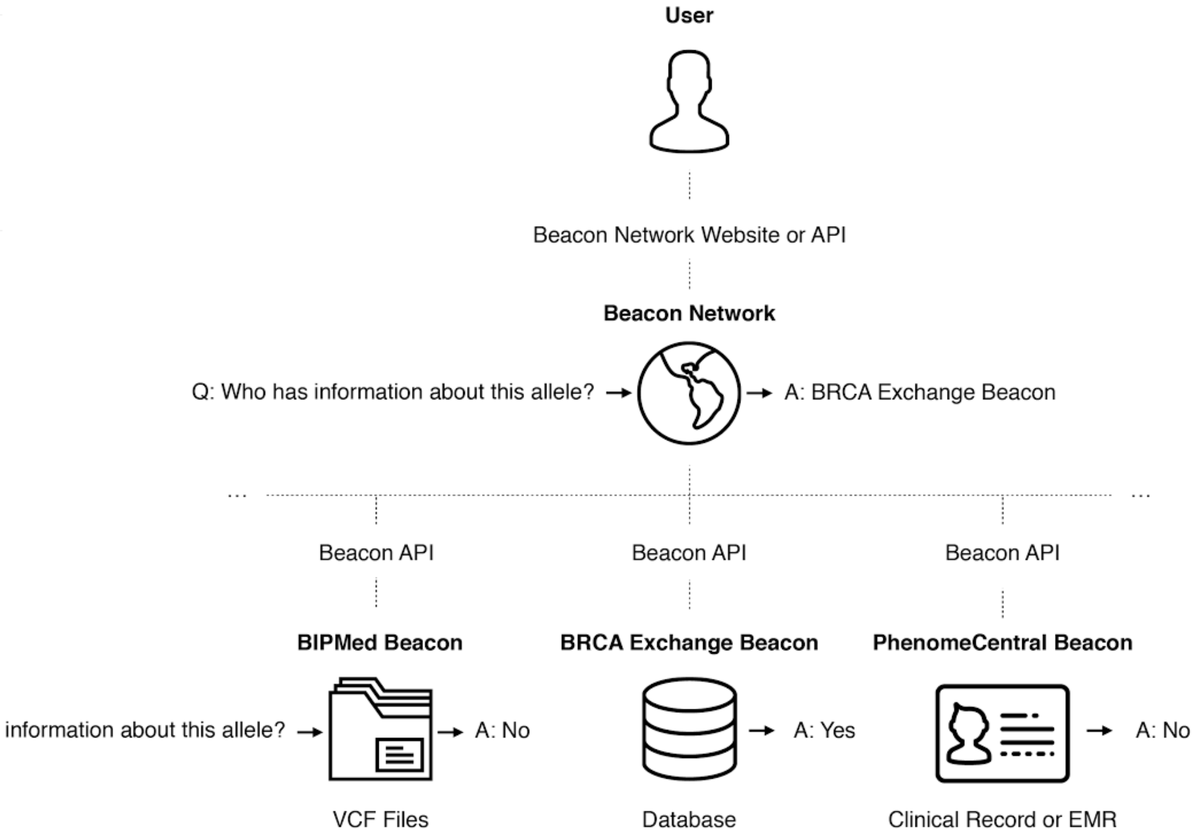
A Beacon network federates *genome variant queries* across databases that support the ***Beacon API***

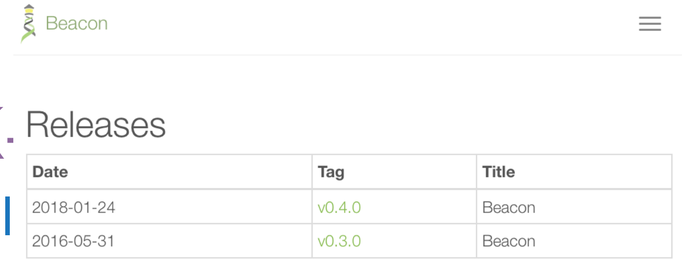
Here: The variant has been found in **few** resources, and those are from **disease** specific **collections**.

Beacon Project in 2016

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







Global Alliance for Genomics & Health

35+ Organizations

90+ Beacons

200+ Datasets

100K+

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

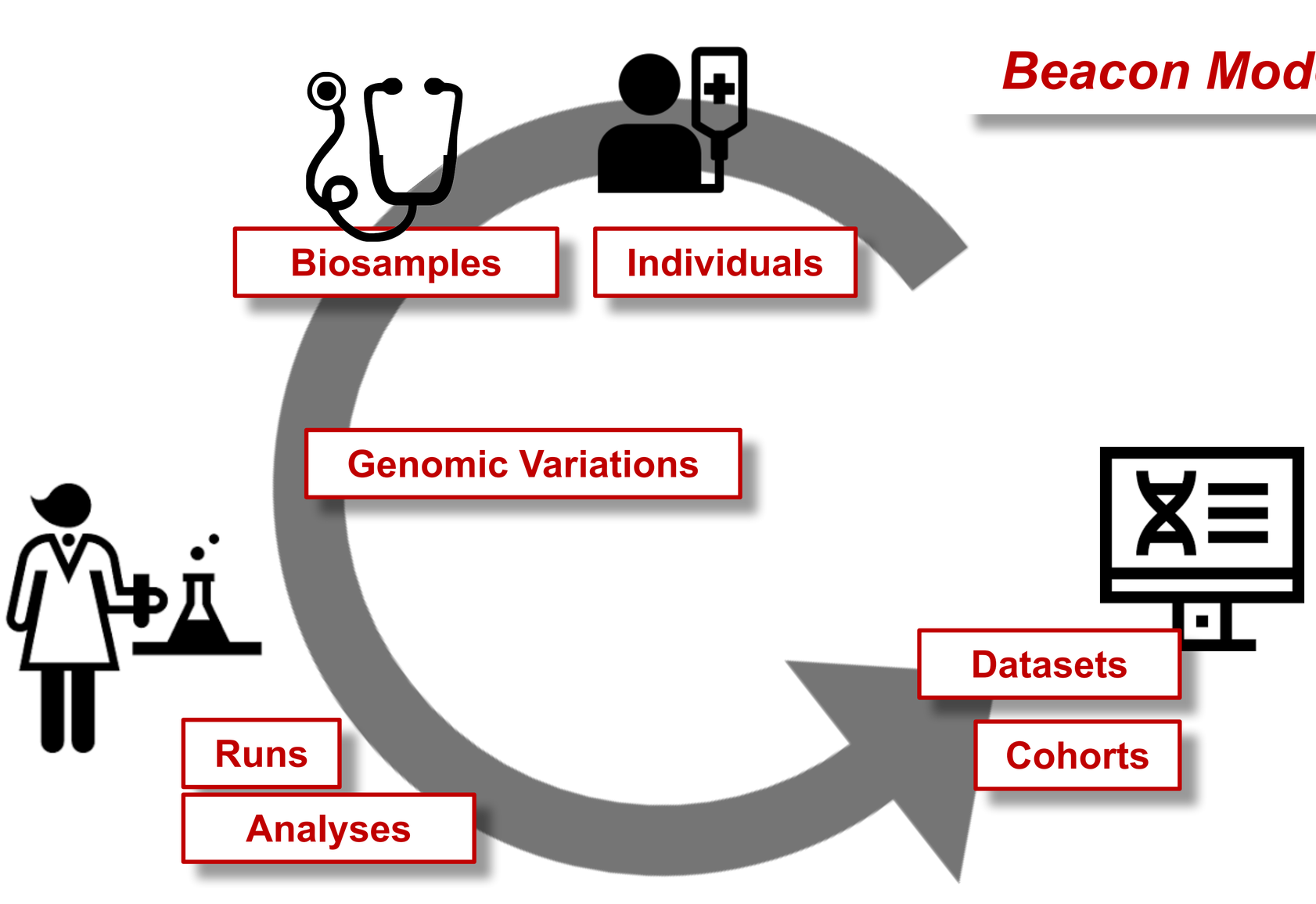
2022

- 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*

Beacon v2

docs.genomebeacons.org

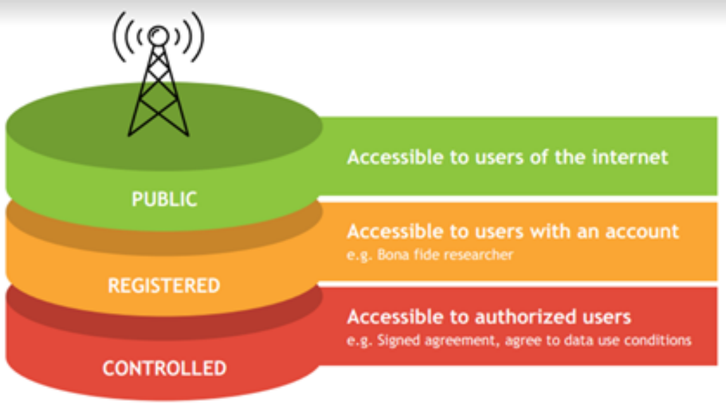


Beacon Model

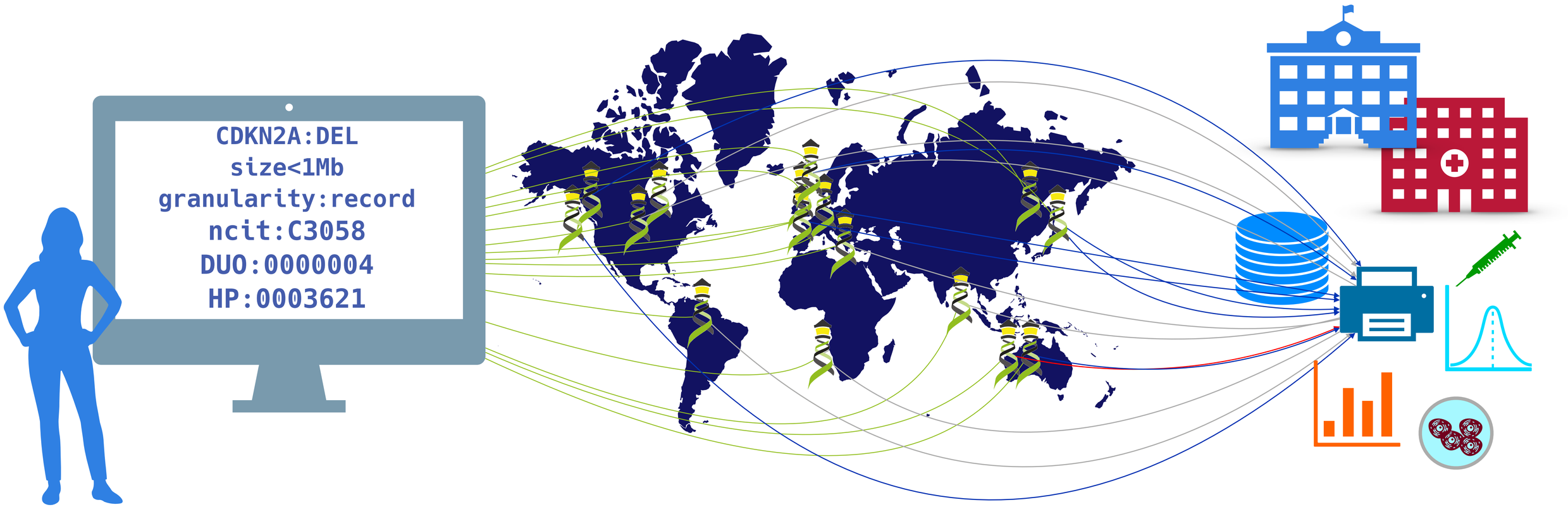
Beacon Framework (protocol)

Beacon Security

Security by Design ... if Implemented in the Environment



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



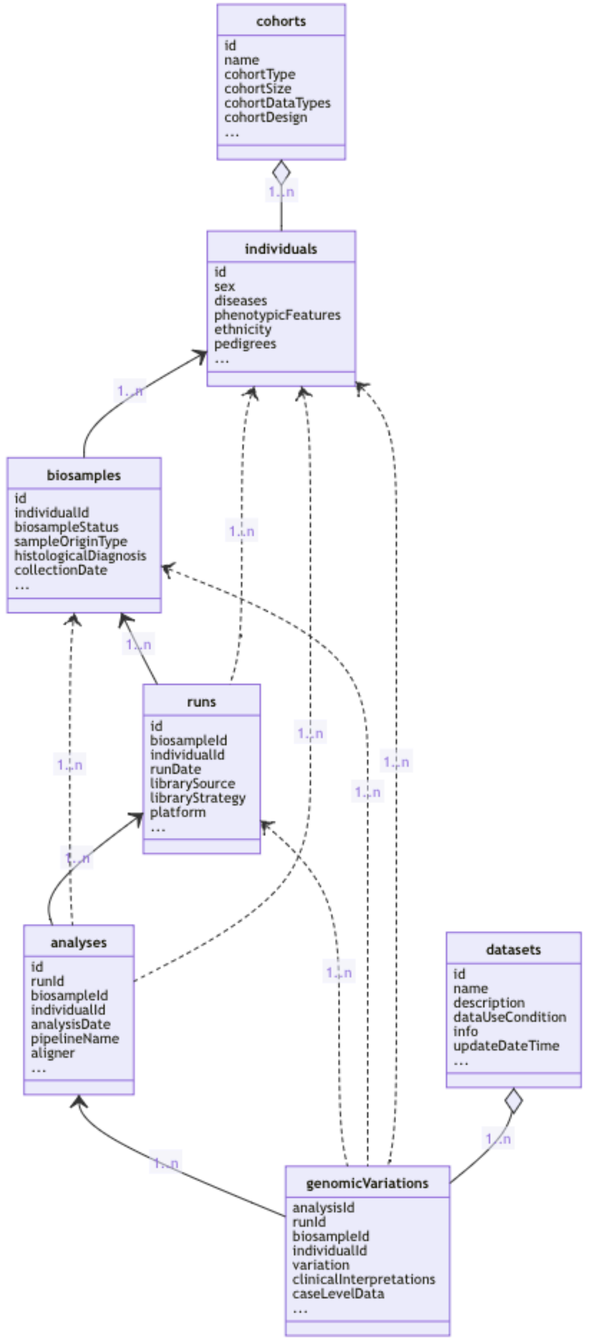
Can you provide data about focal deletions in CDKN2A in Glioblastomas from juvenile patients with unrestricted access?

Beacon *v2* API

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

Beacon Default v2 Model

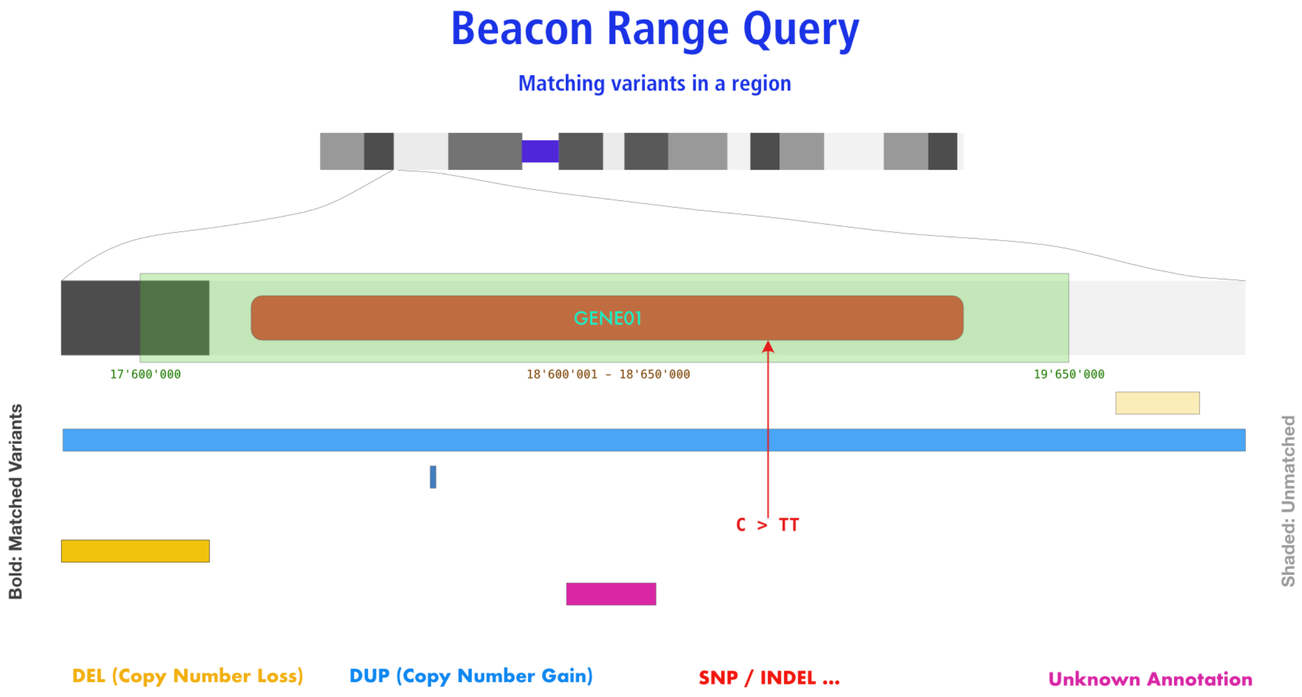
- The Beacon ***framework*** describes the overall structure of the API requests, responses, parameters, the common components, etc.
- Beacon ***models*** describe the set of concepts included in a Beacon, like individual or biosample, and also the relationships between them.
- Besides logical concepts,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

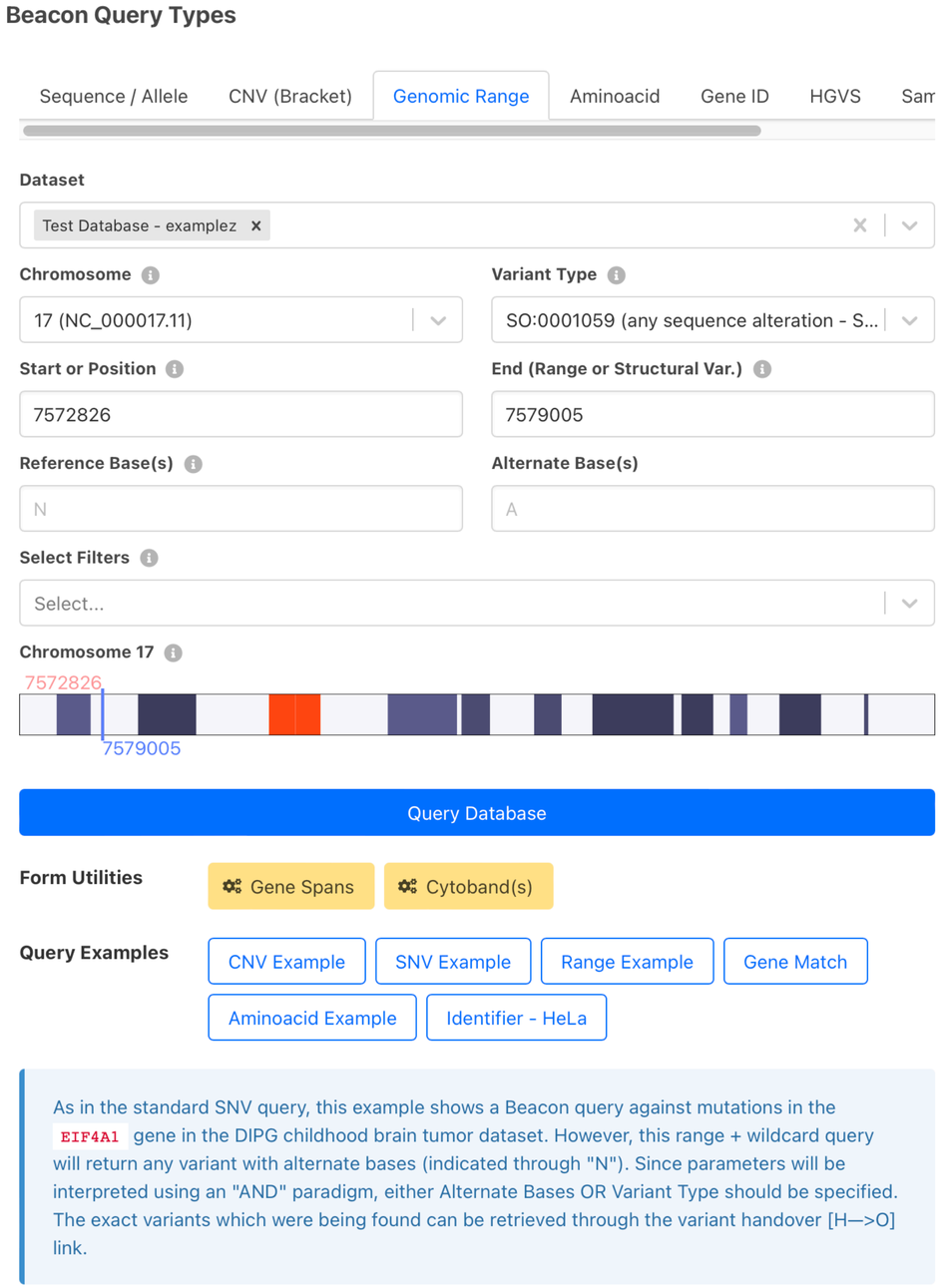


Variation Queries

Range ("anything goes") Request

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

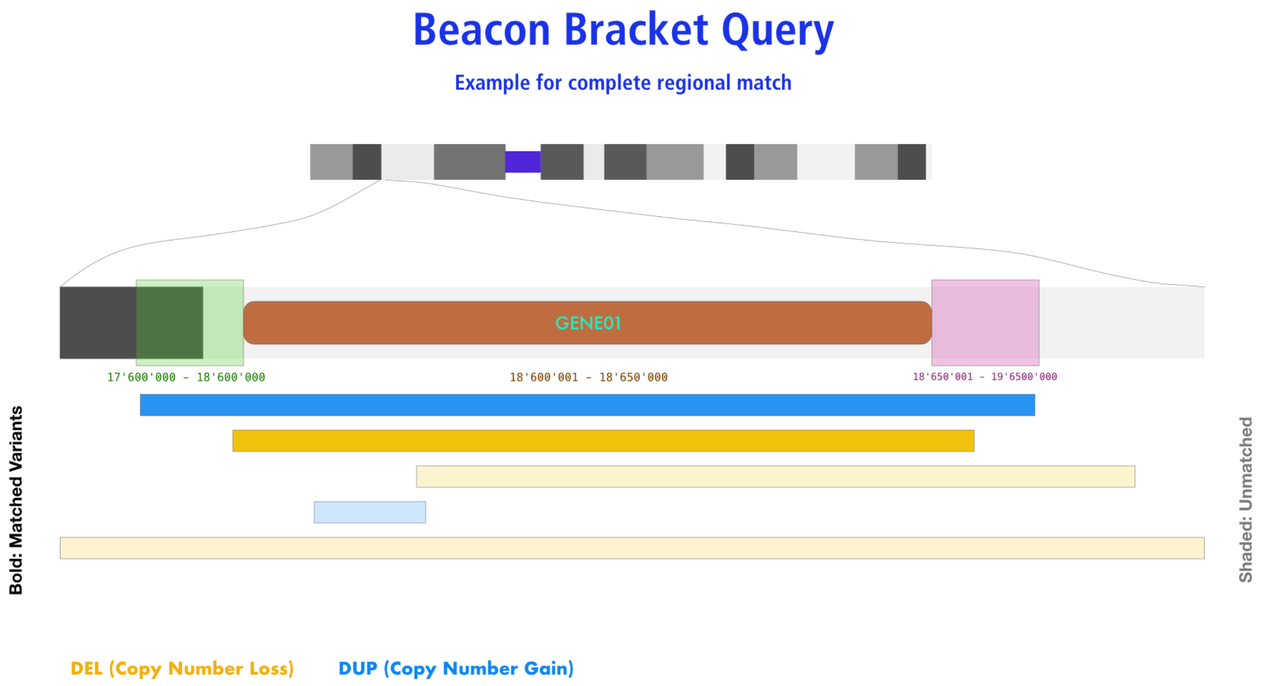


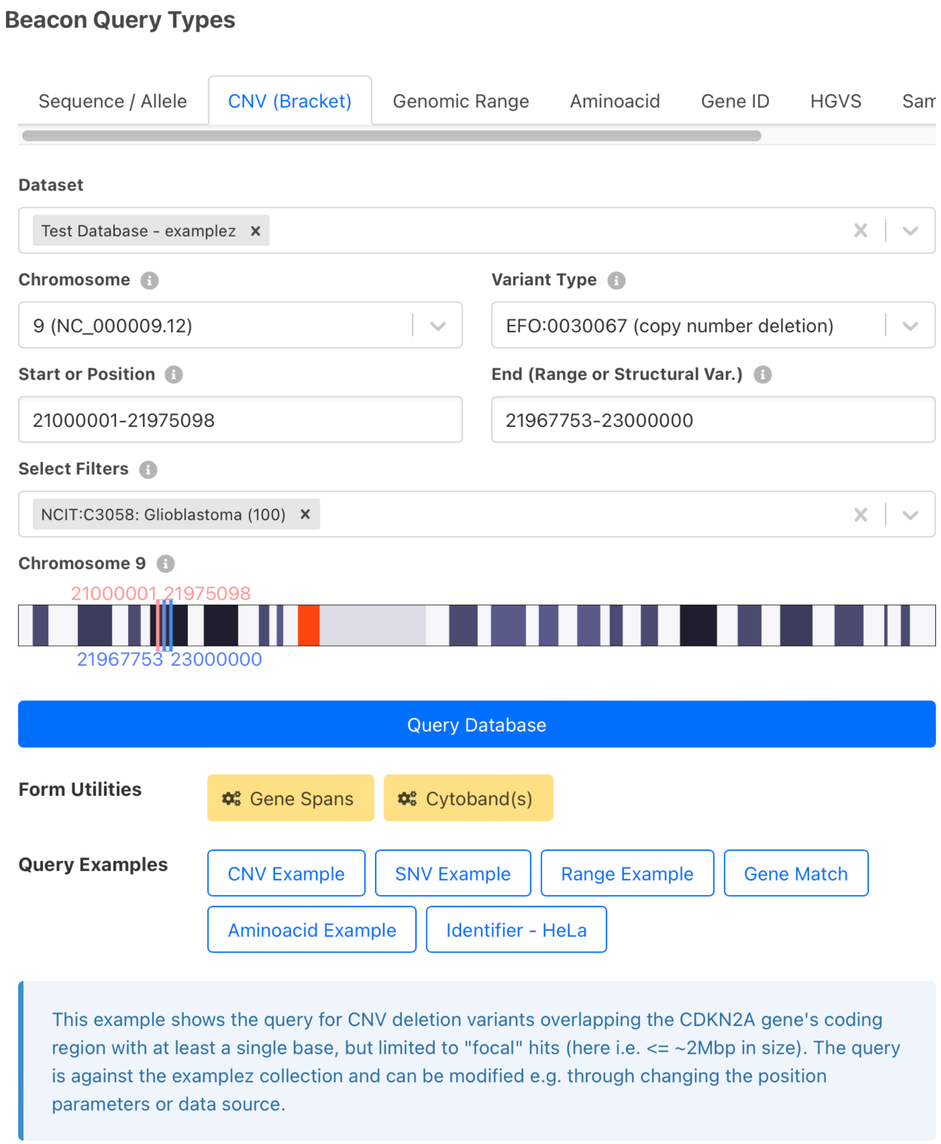


Variation Queries

Bracket ("CNV") Query

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





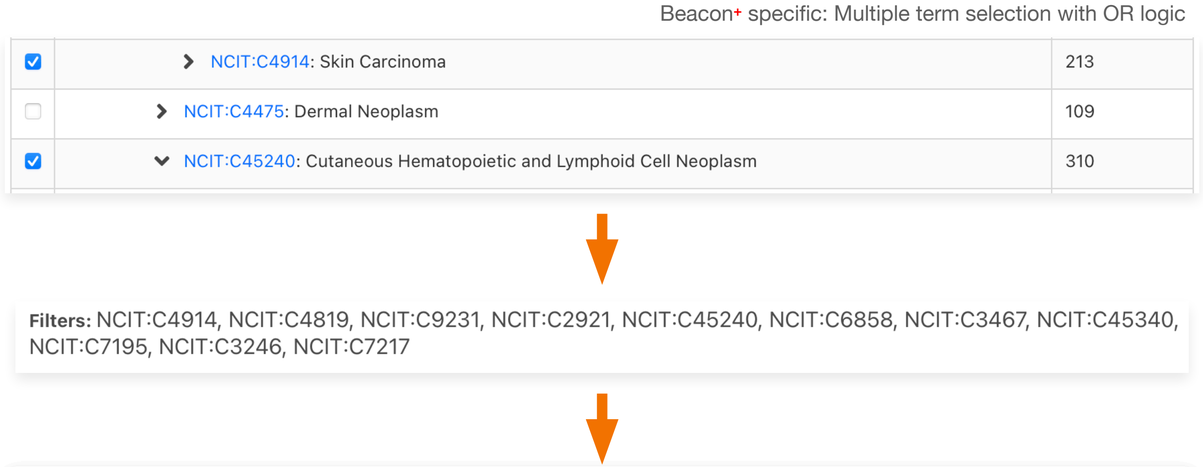
Beacon v2 Filters

Example: Use of hierarchical classification systems (here NCIt neoplasm core)

- Beacon v2 relies heavily on "filters"
  - ontology term / CURIE
  - alphanumeric
  - custom
- Beacon v2 "filters" assumes inclusion of child terms when using hierarchical classifications

  ➡ implicit *OR* with otherwise assumed *AND*
- implementation of hierarchical annotations overcomes some limitations of "fuzzy" disease annotations

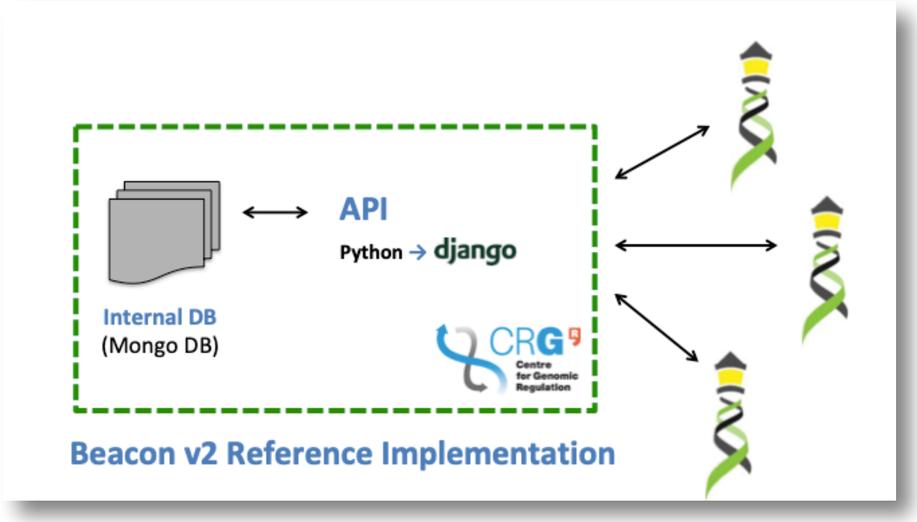
Beacon+ specific: Multiple term selection with OR logic

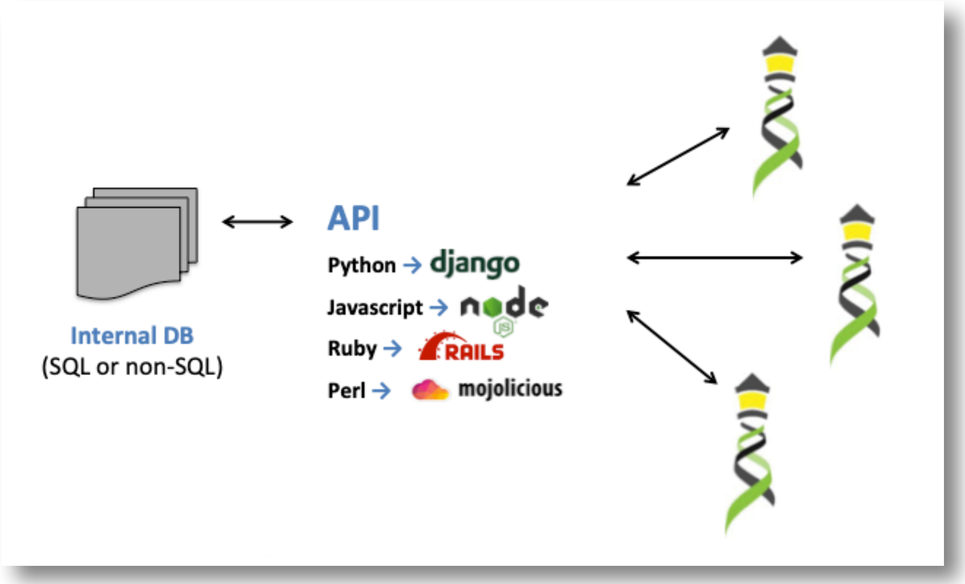


Id	Description	Classifications	Identifiers	DEL	DUP	CNV
PGX_AM_BS_MCC01	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.116	0.104	0.22
PGX_AM_BS_MCC02	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma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

Implementing Beacon v2

... its just code ¯_(ツ)_/¯





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: '876eed6-5829-4f28-b01d-9759a6f55f14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 567,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
```


variants | analyses | biosamples | individuals

Entity collections

collations | geolocs | genespans | publications | qBuffer

Utility collections

Universität Zürich[UZH]

Global Alliance for Genomics & Health

bycon for GA4GH Beacon

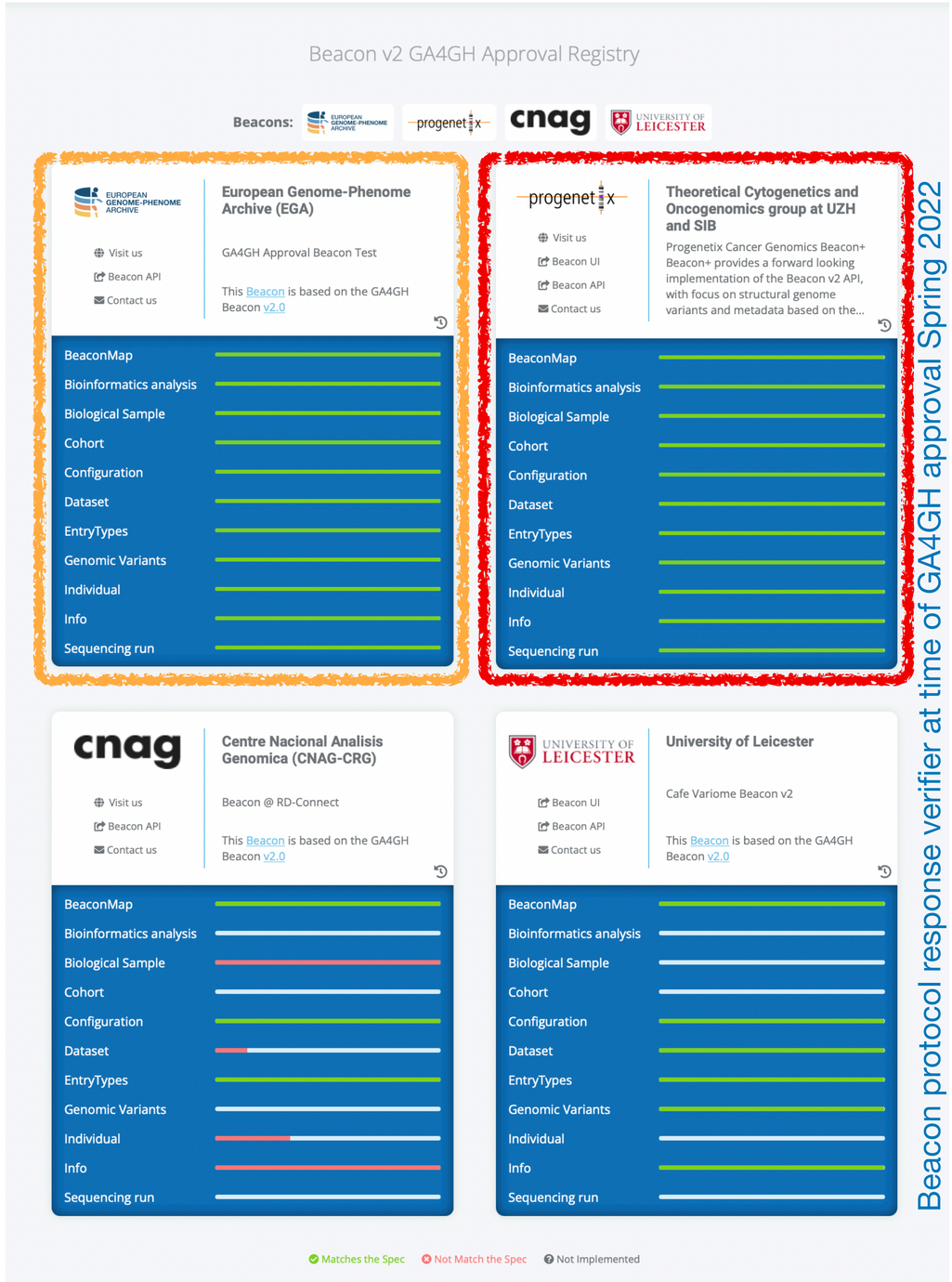
Implementation driven development of a GA4GH standard

progenetix

bycon Beacon

Implementation driven standards development

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



Global Alliance for Genomics & Health

progenetix

elixir

bycon based 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™

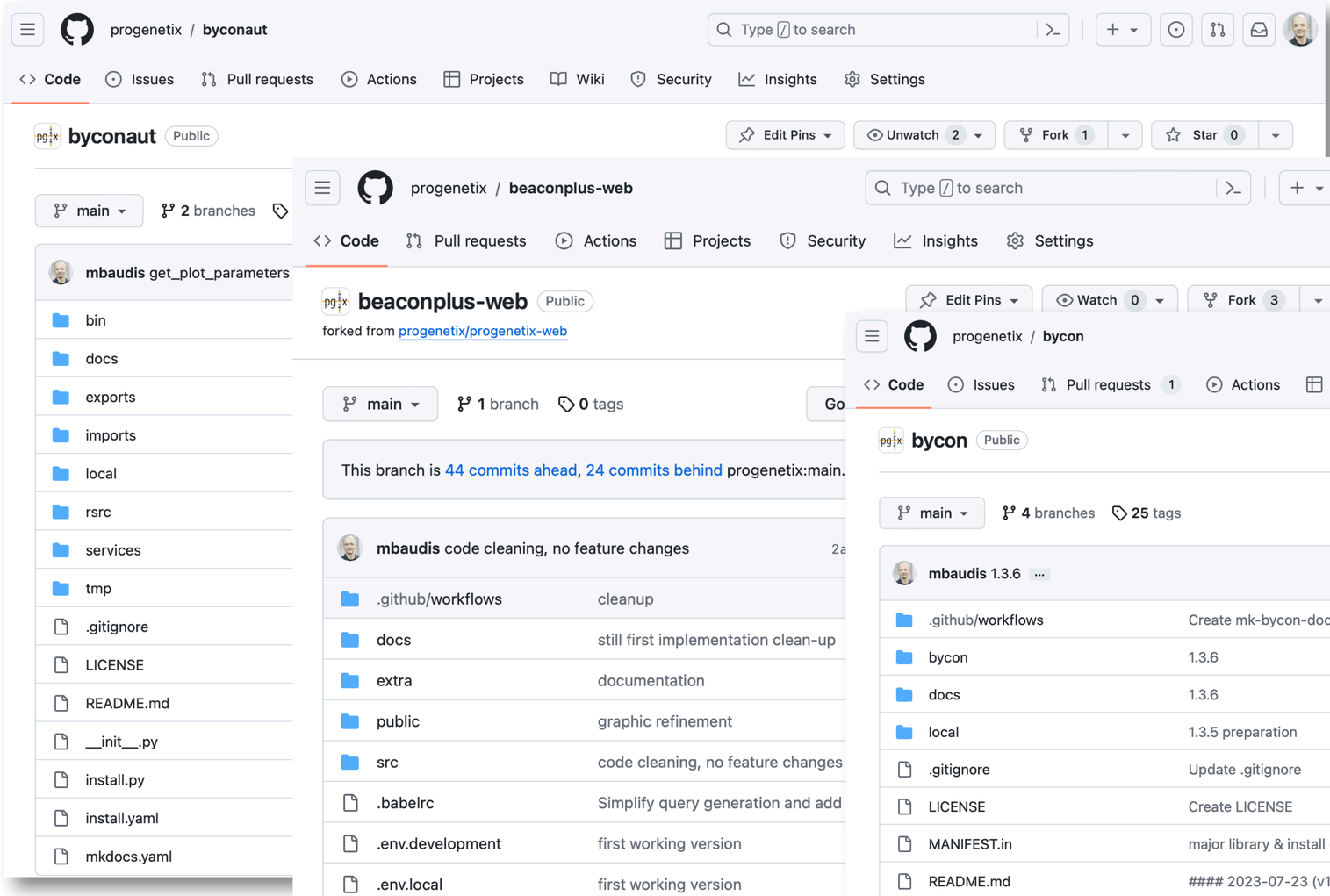
mongoDB

- *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"),
```

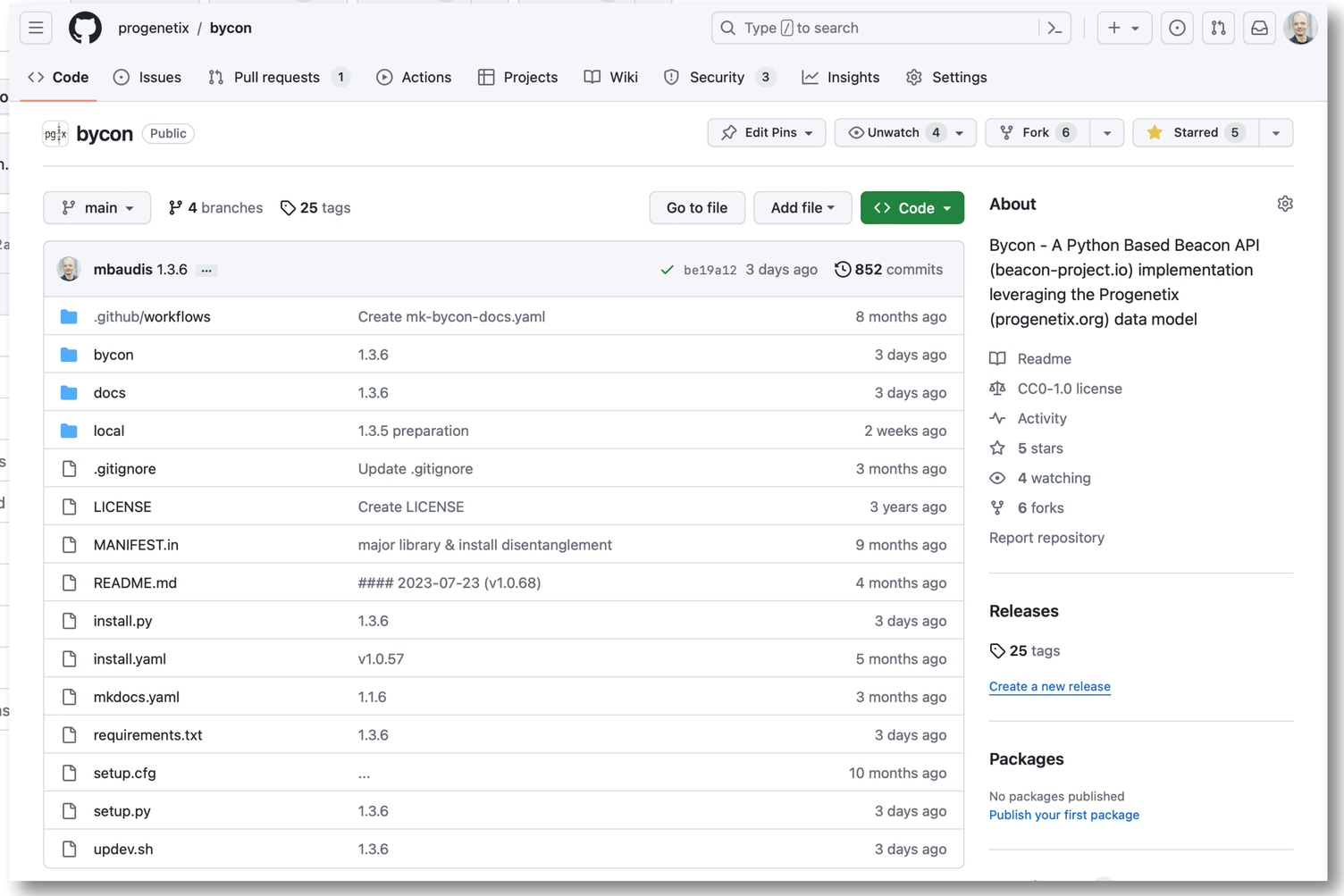
Entity collections

Utility collections



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

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



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?

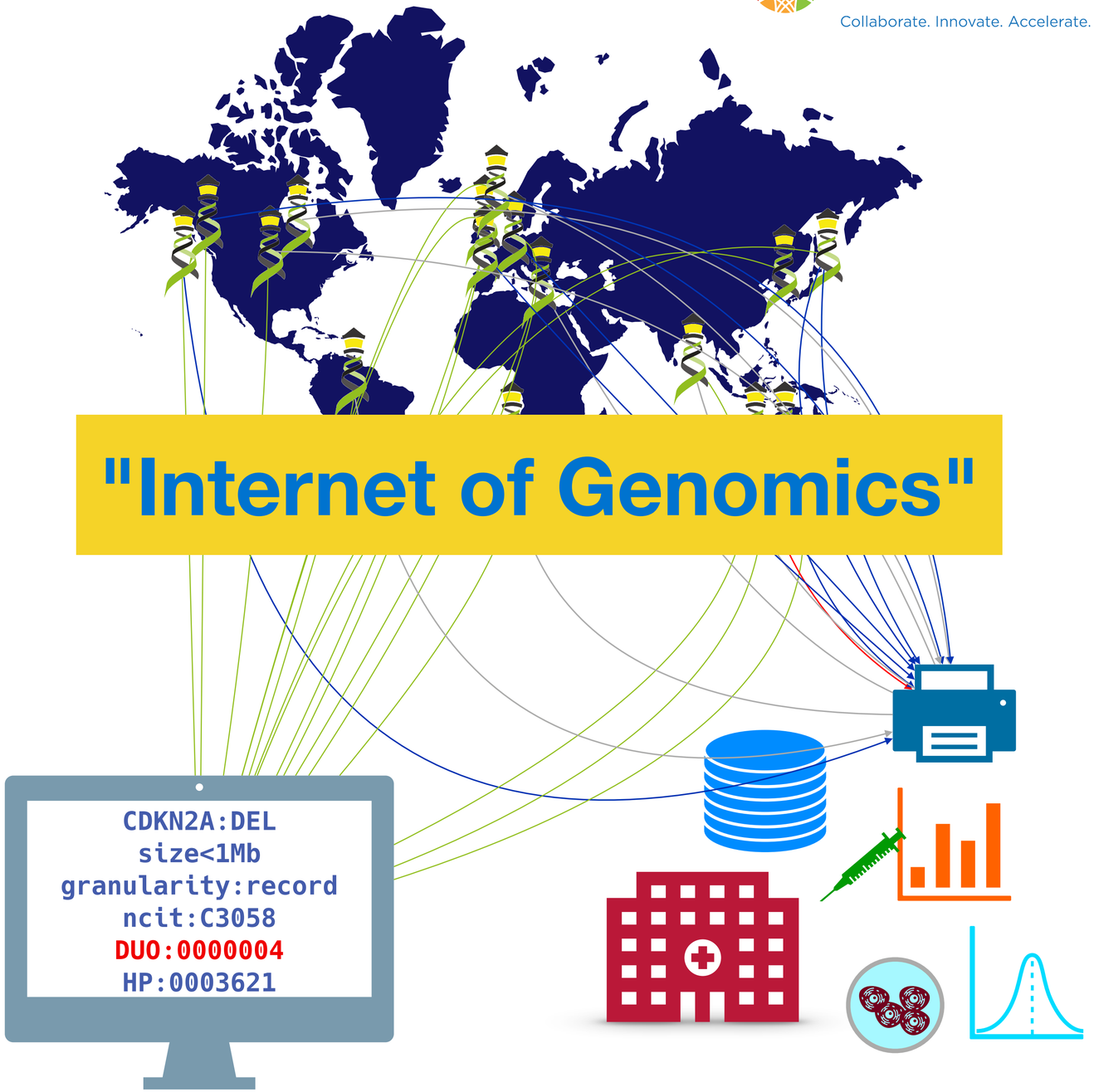
- Patient provided data is valuable - but only if it can be discovered
- Doctors are only curators and stewards of information collected from their patients
- Rare diseases: identify and learn from related cases & help patients to find a community
- Cancer: Learn from data clusters emerging from large collections and transversal analyses

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

What Can You Do?

- find a way to make your (patients') **data discoverable** - through adding *at least* the relevant metadata to national or project centric repositories
- use forward looking consent and data protection models (**ORD** principle "*as secure as necessary, as open as possible*")
- **support** and/or get involved with international **data standards** efforts and projects

➡ **Collaborate!**



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."

The Beacon team through the ages

University of Zurich[UZH]

SIB Swiss Institute of Bioinformatics

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

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

UNIVERSITY OF LEICESTER

Cafe Variome Central

Tony Brookes
Tim Beck
Colin Veal
Tom Shorter

CSC

Juha Törnroos
Teemu Kataja
Ilkka Lappalainen
Dylan Spalding

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

Universität Zürich[UZH]

SIB Swiss Institute of Bioinformatics

elixir