Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

Genomics Data Federation through Global Alliance for Genomics and Health Standards

Development and Implementation of the GA4GH Beacon Protocol

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

Genomics has seen massive and ongoing changes in technology

15 February 2001

nature

www.nature.com

the **human** genome

Nuclear fission
Five-dimensional energy landscapes

Seafloor spreading
The view from under the Arctic ice

Career prospects
Sequence creates new opportunities

naturejobs
genomics special

200+ Genomic Data Initiatives Globally

Clinical/Genomic Medicine

Research

National

Cohorts

How Many Genomes?

RESEARCH

HEALTHCARE

60M individuals
132.5M sequences

CLINICAL TRIALS

2.7-3M individuals

COHORTS

140M individuals

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

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

GENOMICS

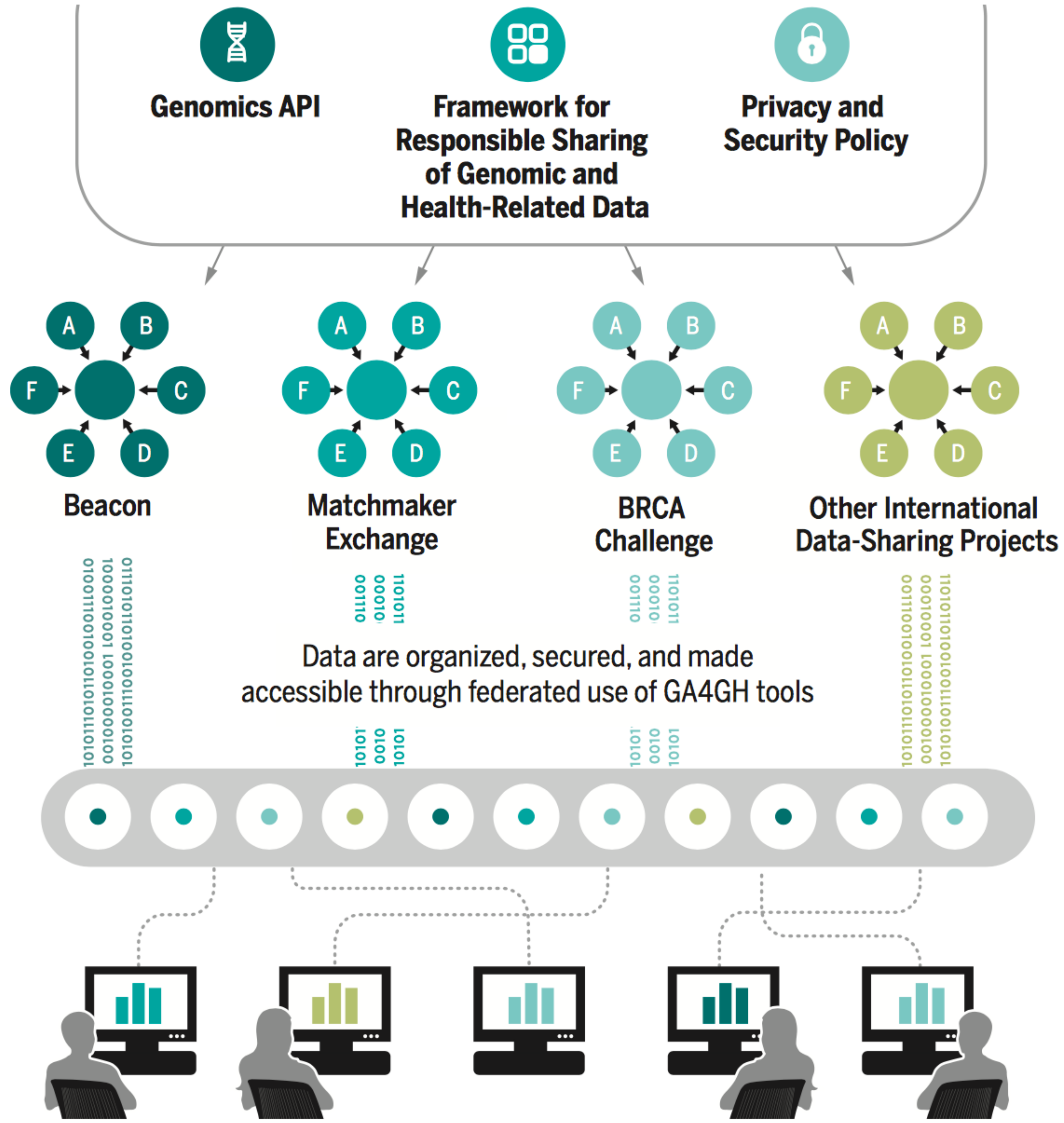
A federated ecosystem for sharing genomic, clinical data

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

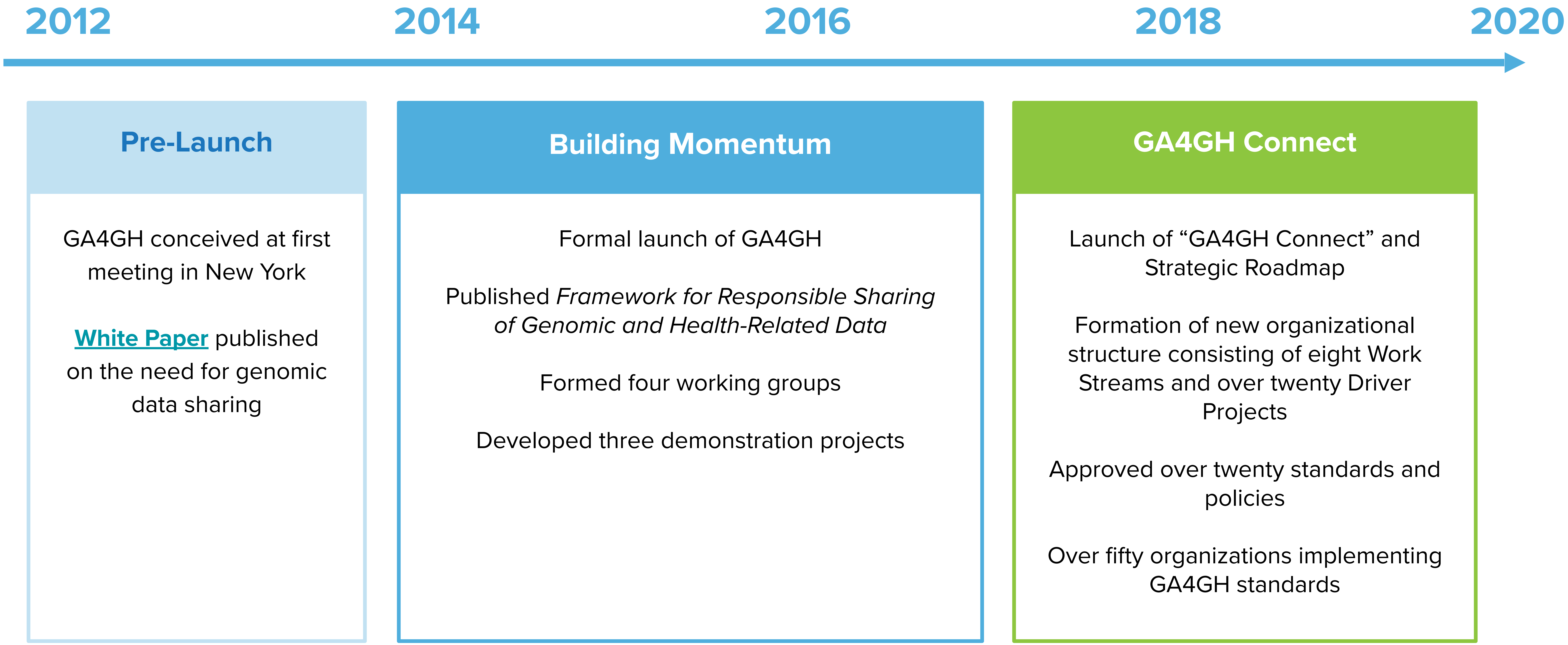
A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.


Genomics API
Framework for Responsible Sharing of Genomic and Health-Related Data
Privacy and Security Policy
A
B
C
D
E
F
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

GA4GH Timeline

Global Alliance for Genomics & Health



2012

2014

2016

2018

2020

Pre-Launch

GA4GH conceived at first meeting in New York

White Paper published on the need for genomic data sharing

Building Momentum

Formal launch of GA4GH

Published *Framework for Responsible Sharing of Genomic and Health-Related Data*

Formed four working groups

Developed three demonstration projects

GA4GH Connect

Launch of "GA4GH Connect" and Strategic Roadmap

Formation of new organizational structure consisting of eight Work Streams and over twenty Driver Projects

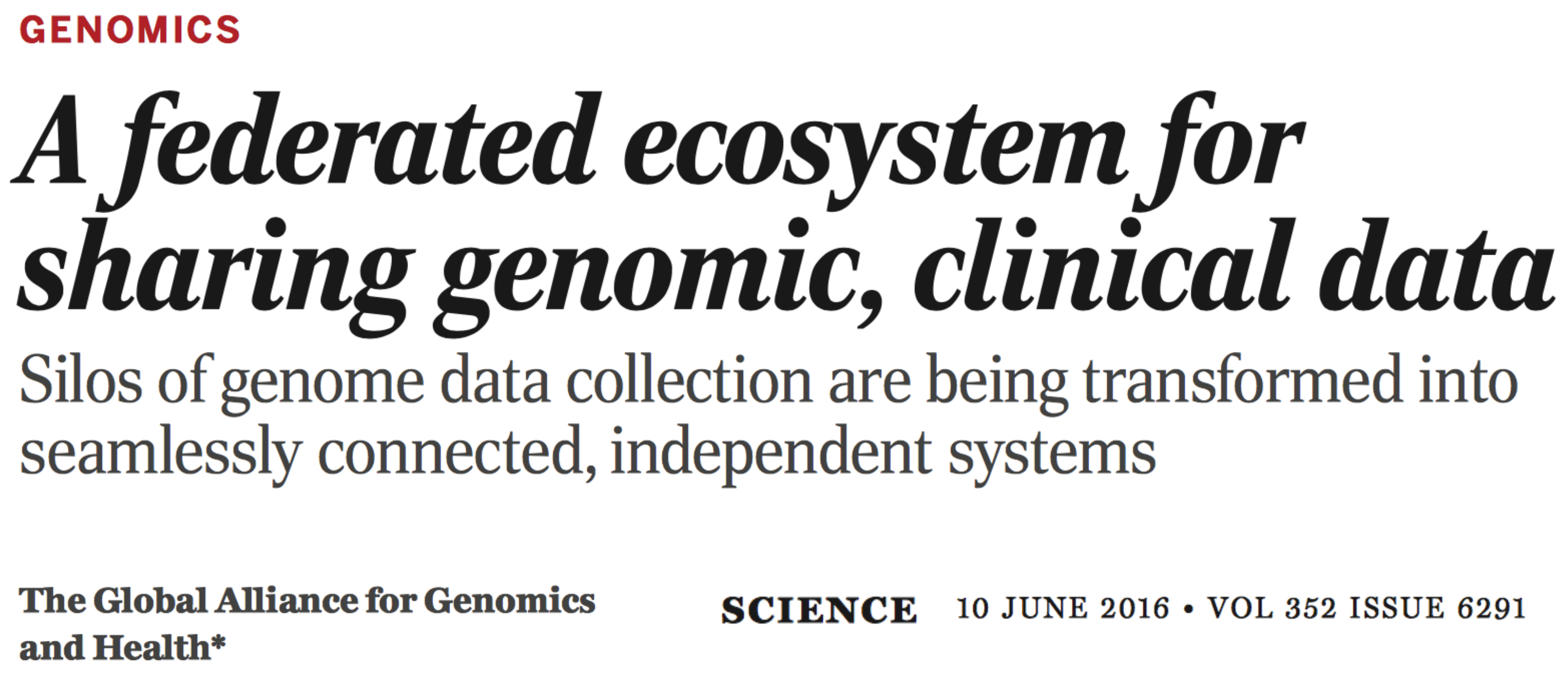
Approved over twenty standards and policies

Over fifty organizations implementing GA4GH standards

The Global Alliance for Genomics and Health

Making genomic data accessible for research and health

- January 2013 - 50 participants from eight countries
- June 2013 - White Paper, over next year signed by 70 “founding” member institutions (e.g. SIB, UZH)
- March 2014 - Working group meeting in Hinxton & 1st plenary in London
- October 2014 - Plenary meeting, San Diego; interaction with ASHG meeting
- June 2015 - 3rd Plenary meeting, Leiden
- September 2015 - GA4GH at ASHG, Baltimore
- October 2015 - DWG / New York Genome Centre
- April 2016 - Global Workshop @ ICHG 2016, Kyoto
- October 2016 - 4th Plenary Meeting, Vancouver
- May 2017 - Strategy retreat, Hinxton
- October 2017 - 5th plenary, Orlando
- May 2018 - Vancouver
- October 2018 - 6th plenary, Basel
- May 2019 - GA4GH Connect, Hinxton
- October 2019 - 7th Plenary, Boston
- October 2020 - Virtual Plenary, June 2021 - Virtual Connect ...
- October 2021 - Virtual Plenary ...
- September 2022 - 10th Plenary, Barcelona



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*

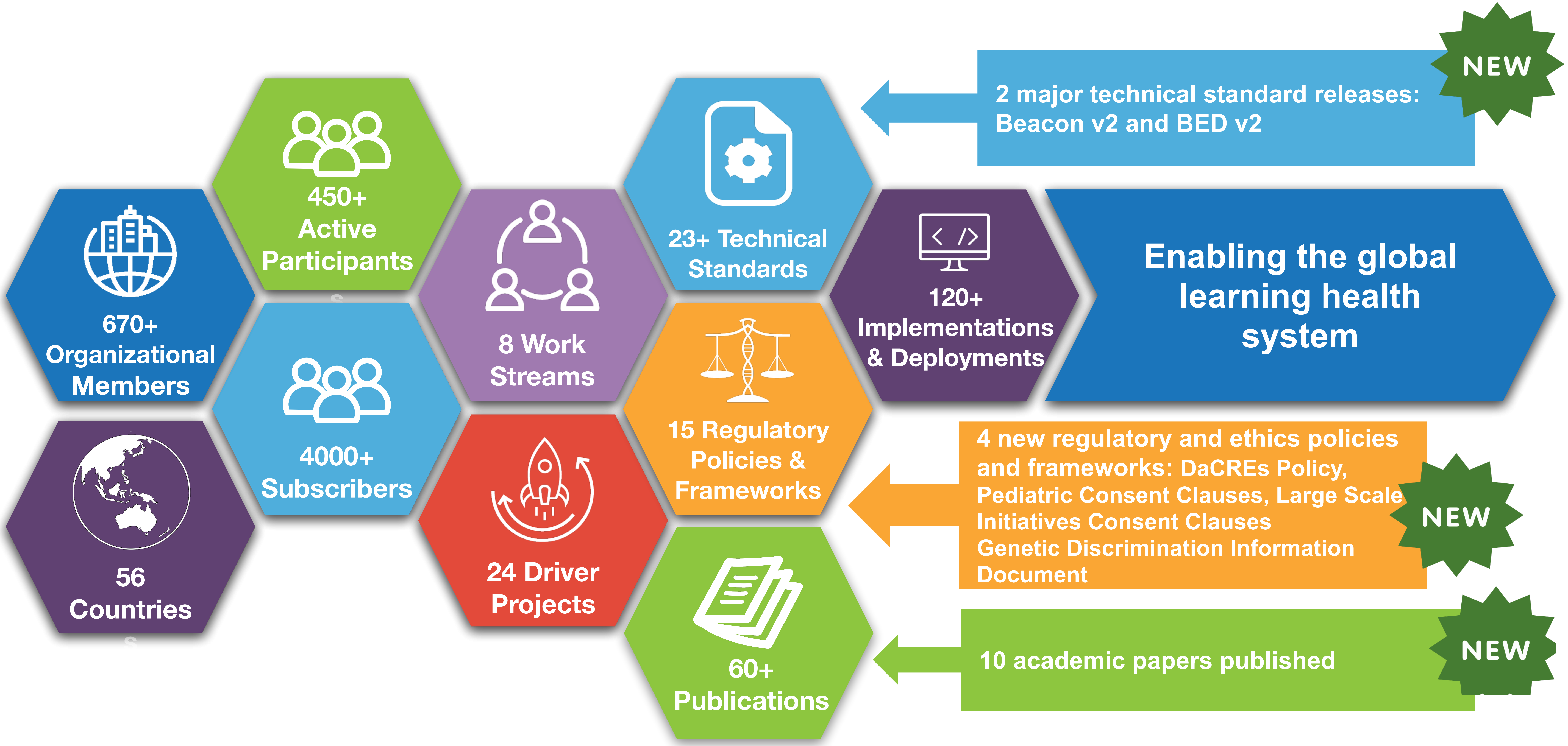
SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

22 SEPTEMBER 2022 | BARCELONA, SPAIN

GA4GH 10th Plenary

Global Alliance for Genomics & Health

The GA4GH ecosystem and outputs



670+ Organizational Members

56 Countries

450+ Active Participants

4000+ Subscribers

8 Work Streams

24 Driver Projects

23+ Technical Standards

15 Regulatory Policies & Frameworks

60+ Publications

120+ Implementations & Deployments

Enabling the global learning health system

NEW

2 major technical standard releases: Beacon v2 and BED v2

NEW

4 new regulatory and ethics policies and frameworks: DaCREs Policy, Pediatric Consent Clauses, Large Scale Initiatives Consent Clauses
Genetic Discrimination Information Document

NEW

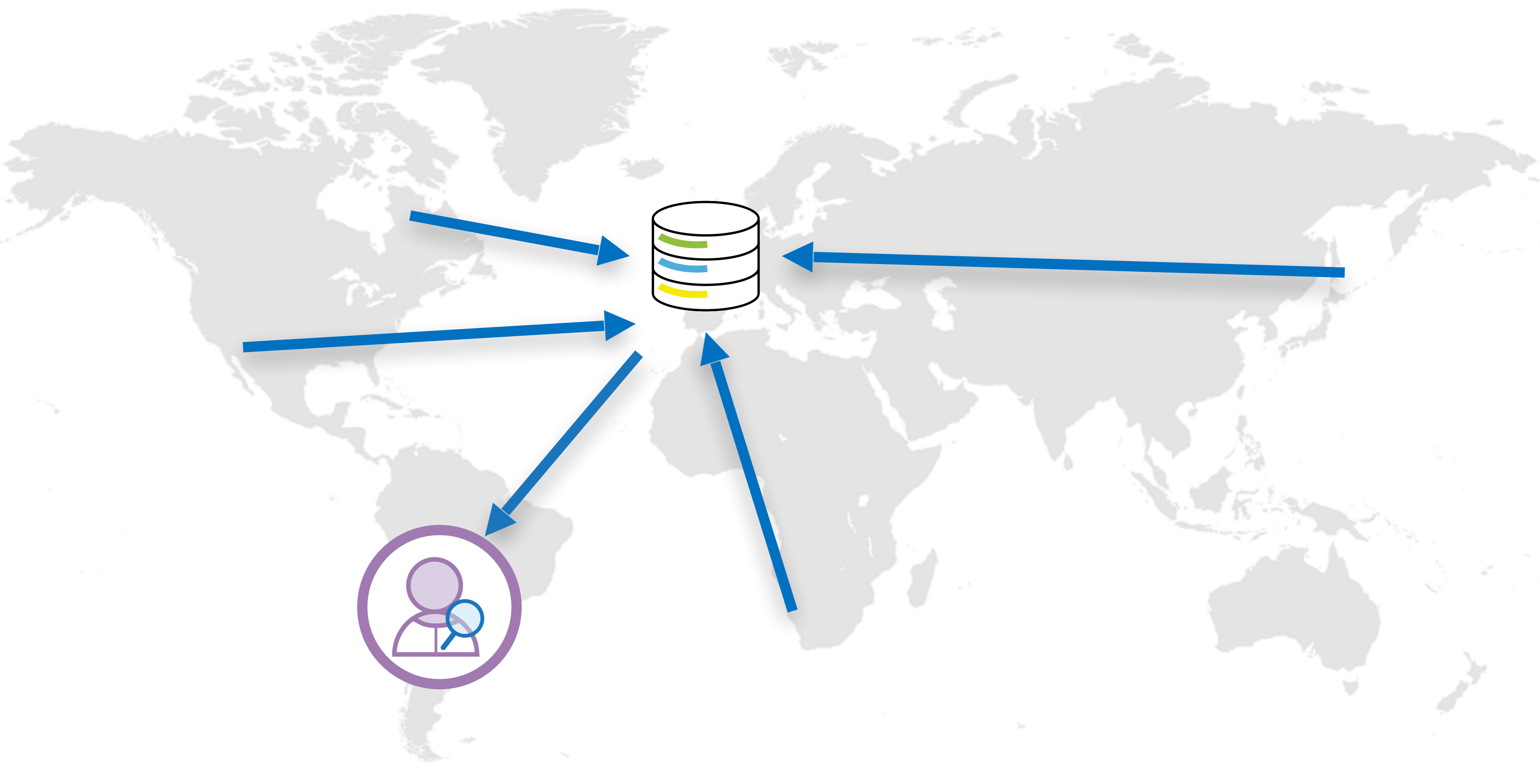
10 academic papers published

Federation

Global Alliance
for Genomics & Health

Central Database

Basic research consented for data sharing

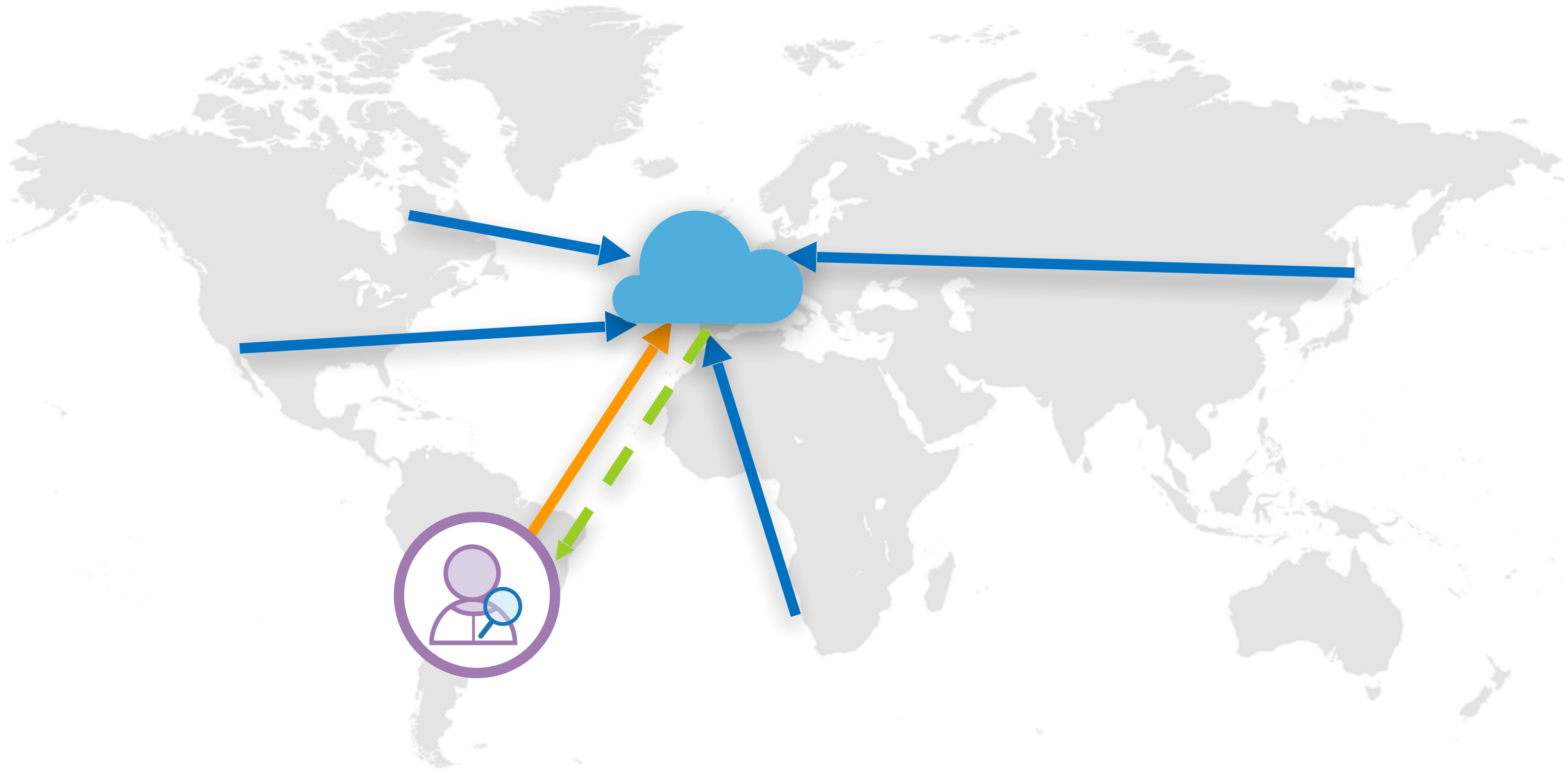


Aggregate data globally

Download, analyze locally

Secure Cloud

Large scale research datasets

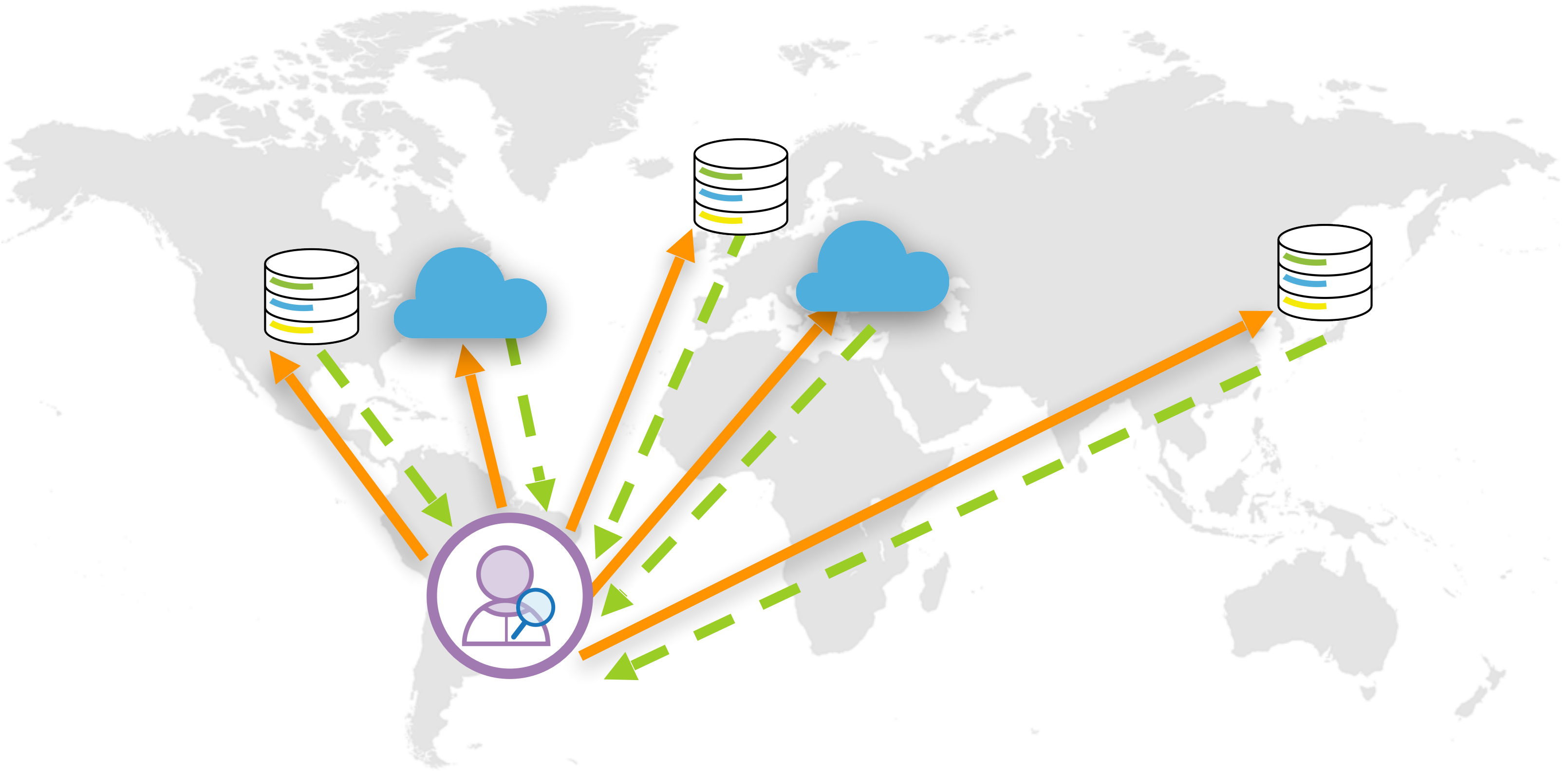


Aggregate data globally

Analyze centrally in secure cloud

Federated Approach

Connecting national genomics initiatives



Host data locally

Analyze data remotely and collate results

User

→ Data transmission

→ Data Visiting

- - - ▸ Results sent back to user

Federation: a solution for data analysis

No data copying or transfer

Data can remain in original jurisdiction

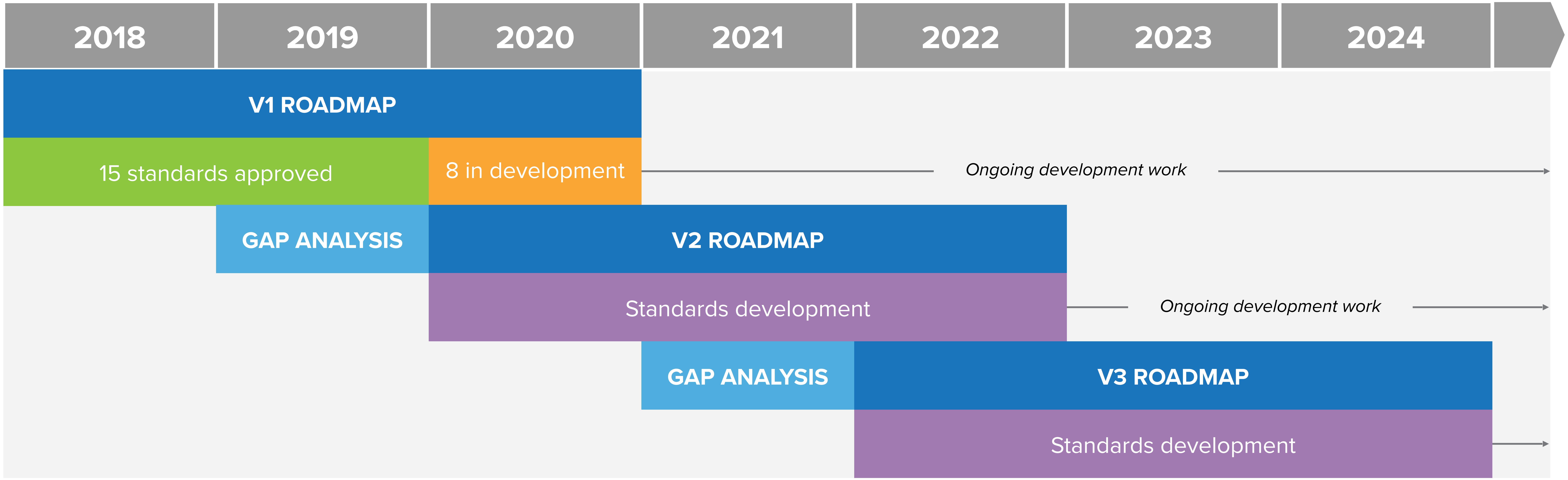
Ownership and access control retained

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

GA4GH Standards Development

GA4GH Roadmap Development Process



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

Strategic Roadmap Alignment to F.A.I.R. Principles

Findable

- Beacon API
- Data Use Ontology
- refget API
- Search API
- Service Registry Prototype
- Tool Registry Service (TRS) API

Accessible

- Authentication and Authorization Infrastructure
- Data Repository Service (DRS) API
- Data Use Ontology
- GA4GH Passports

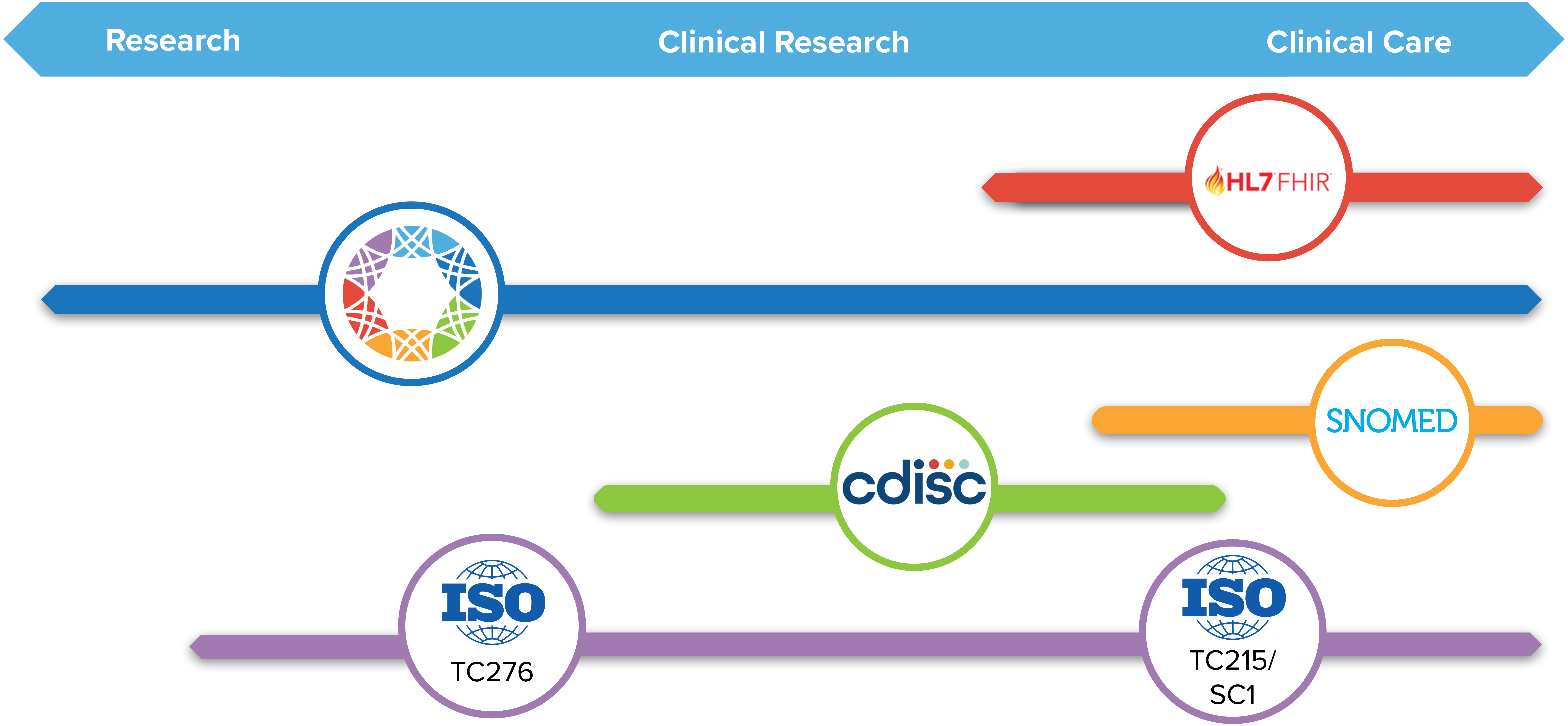
Interoperable

- Phenopackets/FIHR
- Pedigree Representation
- Genetic variant file formats
- Read file formats
- RNAget API
- Crypt4GH
- Variant Annotation
- Variant Representation
- Task Execution Service (TES) API
- Testbed interoperability demonstration
- Tool Registry Service (TRS) API
- Workflow Execution Service (WES) API

Reusable

- htsget streaming API
- refget API
- Variant Annotation
- Workflow Execution Service (WES) API
- Testbed interoperability demonstration

Alignment with Other Standards Organizations



Global Alliance
for Genomics & Health

GA4GH Work Streams

Clinical and Phenotypic Data Capture

Cloud

Data Use and Researcher Identities

Discovery

Genomic Knowledge Standards

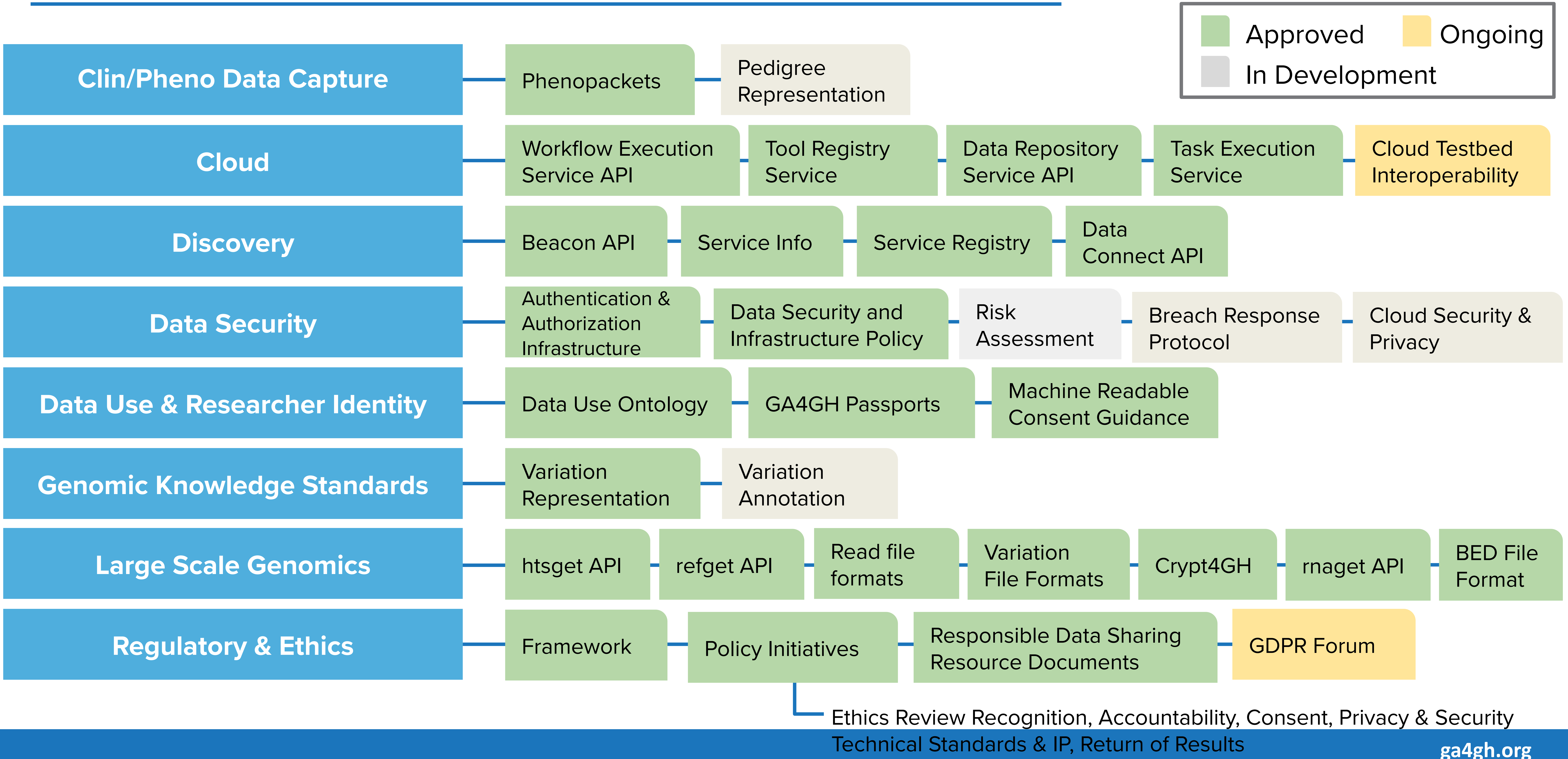
Large Scale Genomics

Regulatory and Ethics

Data Security

GA4GH 2020-2022 Strategic Roadmap

Global Alliance
for Genomics & Health



🟩 Approved 🟨 Ongoing ⬜ In Development

Clin/Pheno Data Capture	🟩 Phenopackets – ⬜ Pedigree Representation
Cloud	🟩 Workflow Execution Service API – 🟩 Tool Registry Service – 🟩 Data Repository Service API – 🟩 Task Execution Service – 🟨 Cloud Testbed Interoperability
Discovery	🟩 Beacon API – 🟩 Service Info – 🟩 Service Registry – 🟩 Data Connect API
Data Security	🟩 Authentication & Authorization Infrastructure – 🟩 Data Security and Infrastructure Policy – ⬜ Risk Assessment – ⬜ Breach Response Protocol – ⬜ Cloud Security & Privacy
Data Use & Researcher Identity	🟩 Data Use Ontology – 🟩 GA4GH Passports – 🟩 Machine Readable Consent Guidance
Genomic Knowledge Standards	🟩 Variation Representation – ⬜ Variation Annotation
Large Scale Genomics	🟩 htsget API – 🟩 refget API – 🟩 Read file formats – 🟩 Variation File Formats – 🟩 Crypt4GH – 🟩 rnaget API – 🟩 BED File Format
Regulatory & Ethics	🟩 Framework – 🟩 Policy Initiatives – 🟩 Responsible Data Sharing Resource Documents – 🟨 GDPR Forum

Policy Initiatives: Ethics Review Recognition, Accountability, Consent, Privacy & Security Technical Standards & IP, Return of Results

GA4GH 2020-2022 Strategic Roadmap

Global Alliance
for Genomics & Health

🟩 Approved 🟨 Ongoing
⬜ In Development

Clin/Pheno Data Capture	🟩 Phenopackets	⬜ Pedigree Representation					
Cloud	🟩 Workflow Execution Service API	🟩 Tool Registry Service	🟩 Data Repository Service API	🟩 Task Execution Service	🟨 Cloud Testbed Interoperability		
Discovery	🟩 Beacon API	🟩 Service Info	🟩 Service Registry	🟩 Data Connect API			
Data Security	🟩 Authentication & Authorization Infrastructure	🟩 Data Security and Infrastructure Policy	⬜ Risk Assessment	⬜ Breach Response Protocol	⬜ Cloud Security & Privacy		
Data Use & Researcher Identity	🟩 Data Use Ontology	🟩 GA4GH Passports	🟩 Machine Readable Consent Guidance				
Genomic Knowledge Standards	🟩 Variation Representation	⬜ Variation Annotation					
Large Scale Genomics	🟩 htsget API	🟩 refget API	🟩 Read file formats	🟩 Variation File Formats	🟩 Crypt4GH	🟩 rnaget API	🟩 BED File Format
Regulatory & Ethics	🟩 Framework	🟩 Policy Initiatives	🟩 Responsible Data Sharing Resource Documents	🟨 GDPR Forum			

Ethics Review Recognition, Accountability, Consent, Privacy & Security Technical Standards & IP, Return of Results

Support clinical adoption of genomics through information models and standards for describing and exchanging clinical phenotypes.

Proposed Solution

Standardize exchange formats for representing clinical data and describing clinical phenotypes.

GA4GH Deliverables

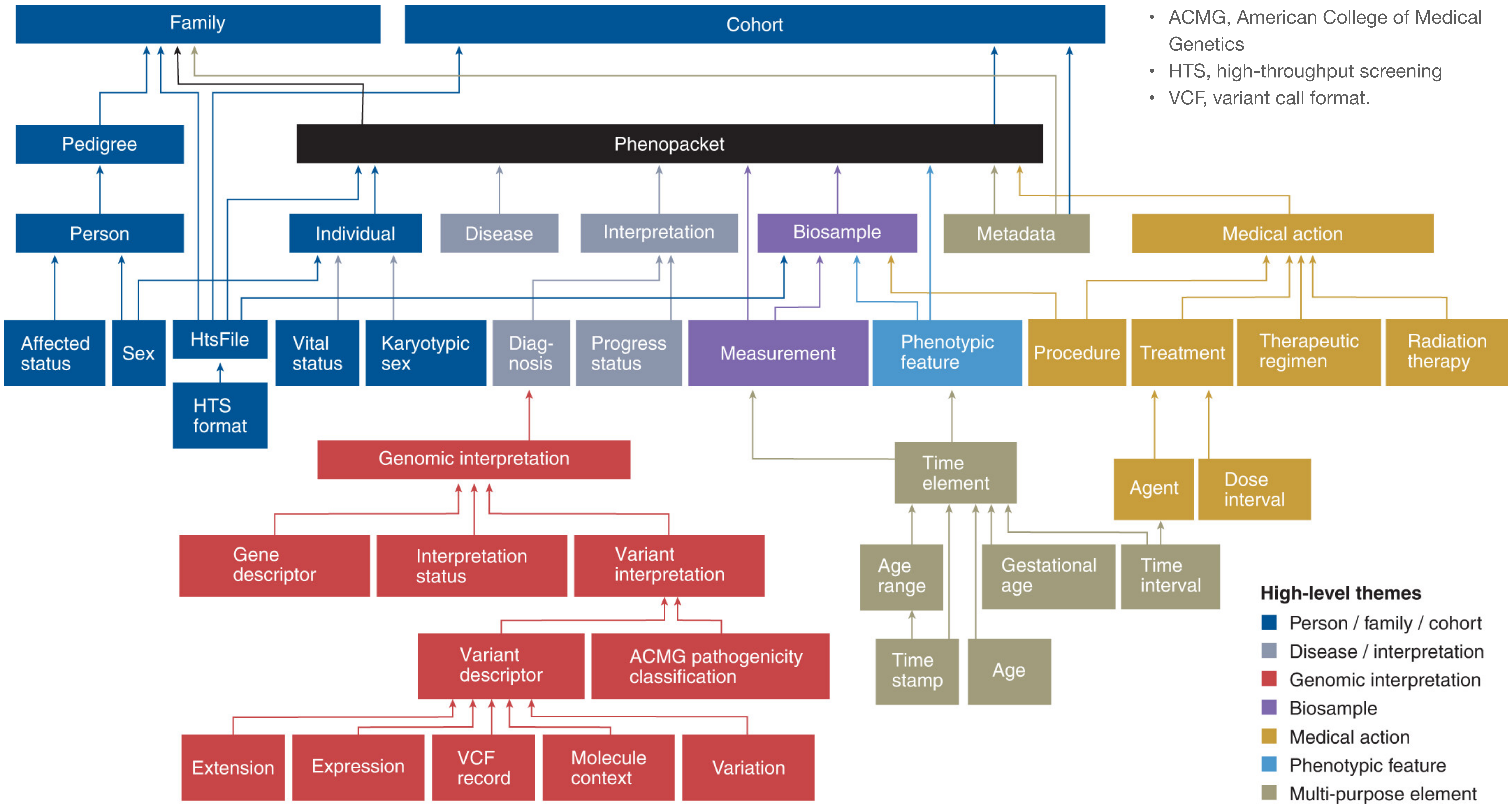
- Phenopackets
- Pedigree
- New project: Cohort Representation

The GA4GH Phenopackets v2 Standard
A Computable Representation of Clinical Data

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

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

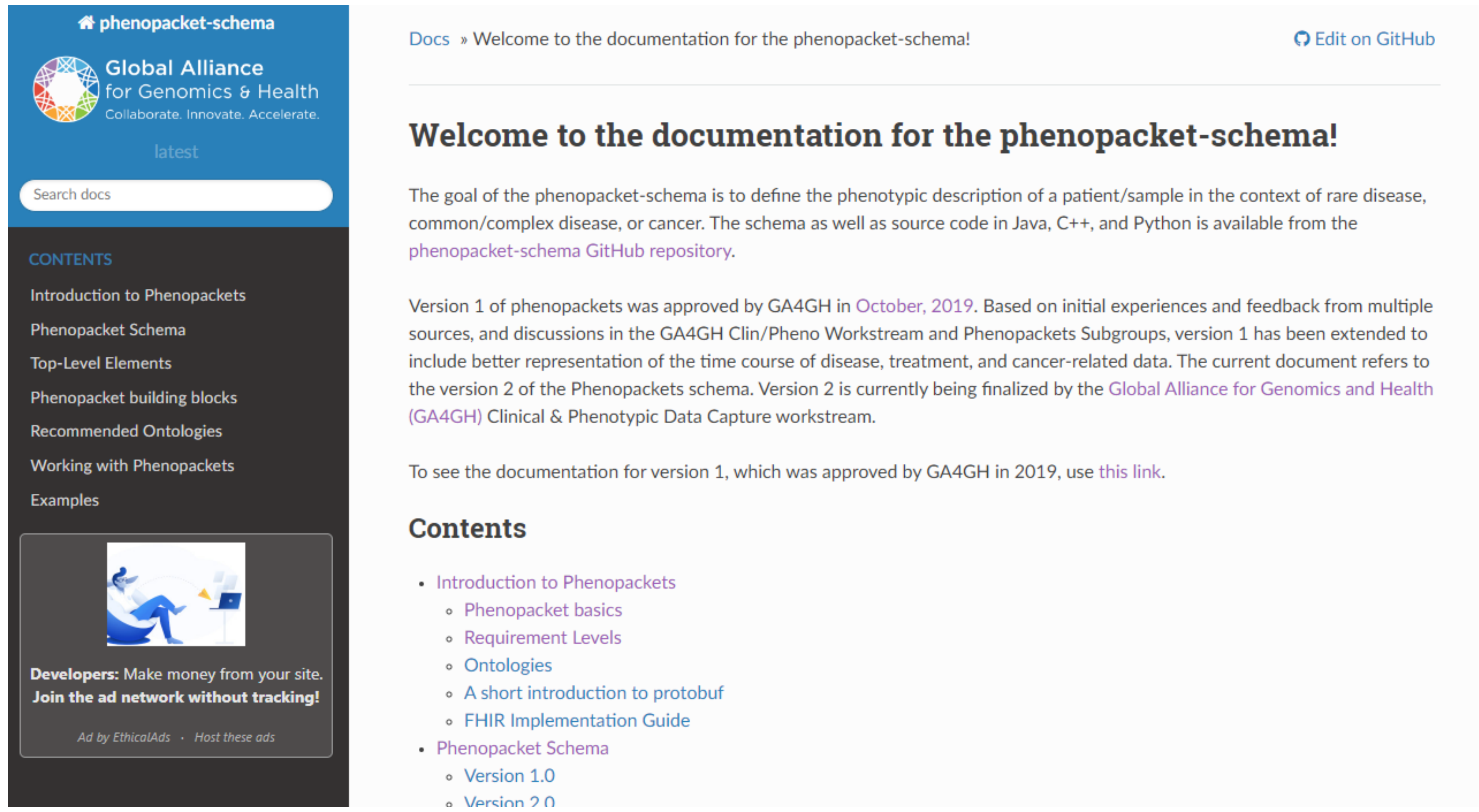
Jacobsen JOB, Baudis M, Baynam GS, Beckmann JS, Beltran S, Buske OJ, Callahan TJ, *et al.* 2022.
The GA4GH Phenopacket Schema Defines a Computable Representation of Clinical Data.
Nature Biotechnology 40 (6): 817–20.



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

Phenopackets Available via GA4GH and ISO

GA4GH


phenopacket-schema
Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.
latest
Search docs
CONTENTS
Introduction to Phenopackets
Phenopacket Schema
Top-Level Elements
Phenopacket building blocks
Recommended Ontologies
Working with Phenopackets
Examples
Developers: Make money from your site. Join the ad network without tracking!
Ad by EthicalAds · Host these ads
Docs » Welcome to the documentation for the phenopacket-schema!
Edit on GitHub

Welcome to the documentation for the phenopacket-schema!

The goal of the phenopacket-schema is to define the phenotypic description of a patient/sample in the context of rare disease, common/complex disease, or cancer. The schema as well as source code in Java, C++, and Python is available from the phenopacket-schema GitHub repository.

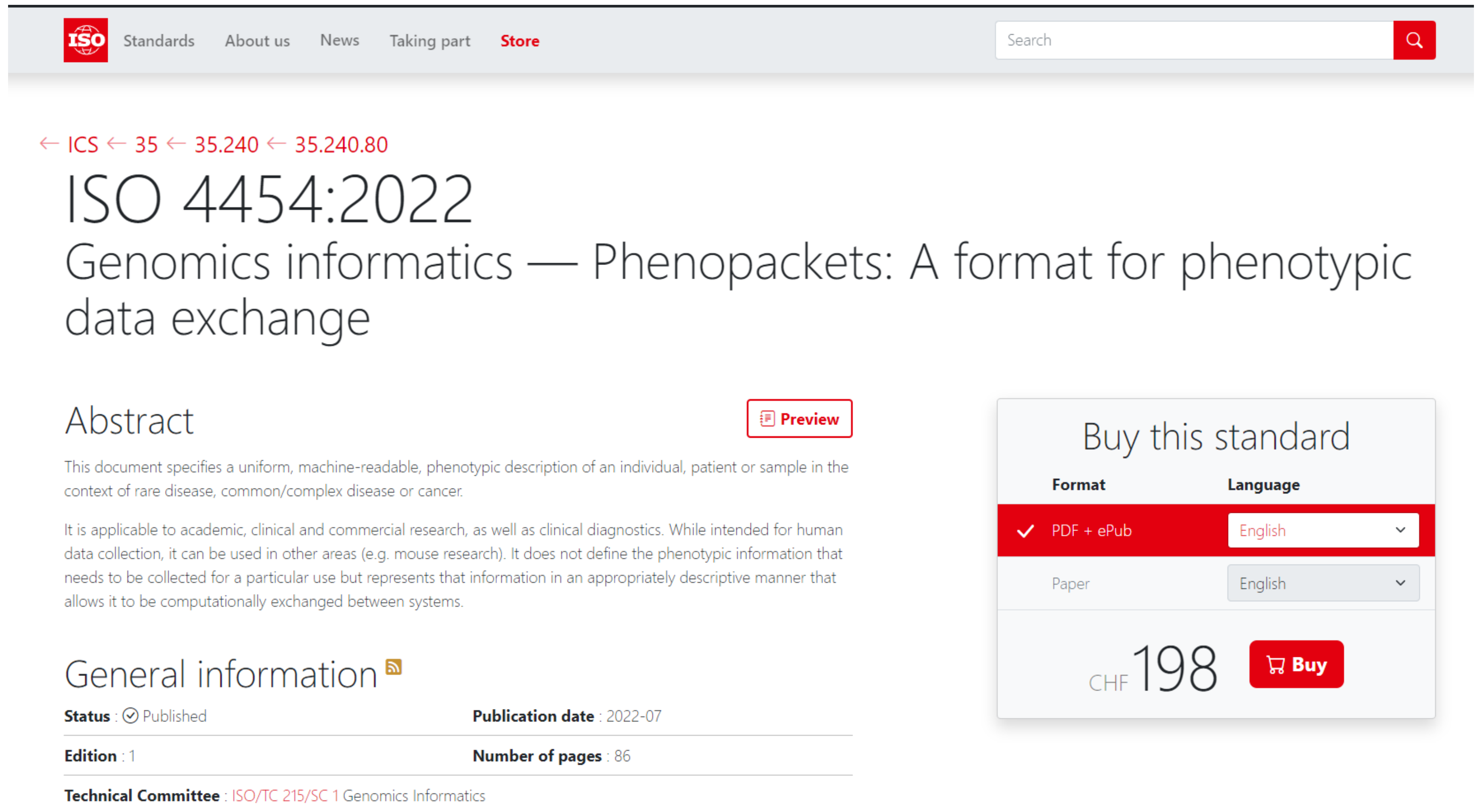
Version 1 of phenopackets was approved by GA4GH in October, 2019. Based on initial experiences and feedback from multiple sources, and discussions in the GA4GH Clin/Pheno Workstream and Phenopackets Subgroups, version 1 has been extended to include better representation of the time course of disease, treatment, and cancer-related data. The current document refers to the version 2 of the Phenopackets schema. Version 2 is currently being finalized by the Global Alliance for Genomics and Health (GA4GH) Clinical & Phenotypic Data Capture workstream.

To see the documentation for version 1, which was approved by GA4GH in 2019, use this link.

Contents

https://bit.ly/PhenopacketsDocs

ISO



ISO 4454:2022

Genomics informatics — Phenopackets: A format for phenotypic data exchange

Abstract

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

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

General information

Status : ✓ Published **Publication date** : 2022-07

Edition : 1 **Number of pages** : 86

Technical Committee : ISO/TC 215/SC 1 Genomics Informatics

Buy this standard

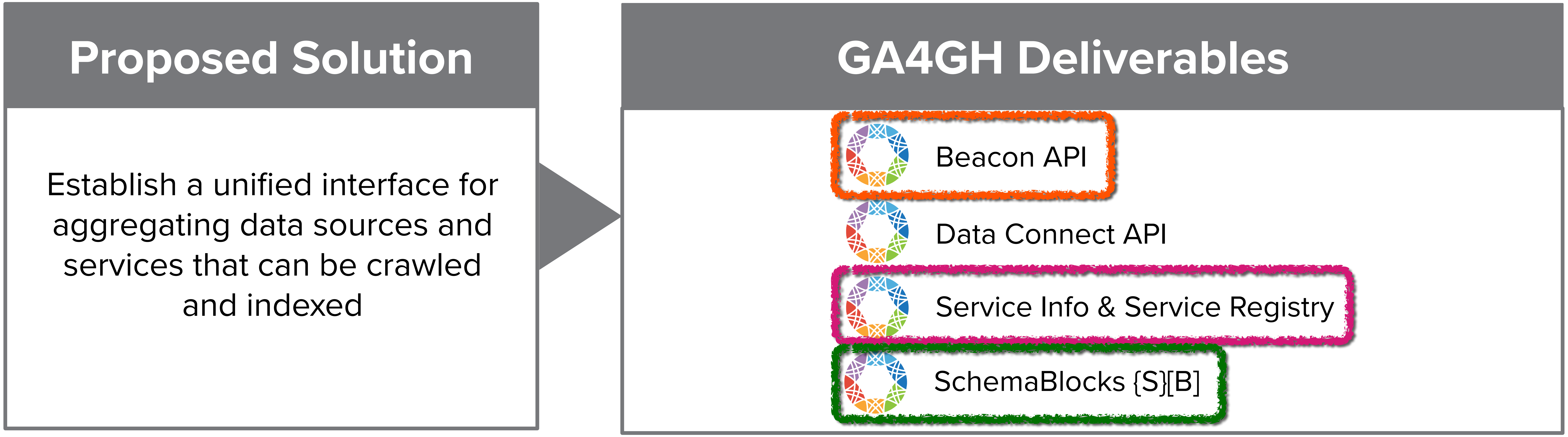
Format	Language
✓ PDF + ePub	English
Paper	English

CHF 198

https://www.iso.org/standard/79991.html

Discovery Work Stream

Facilitate the discovery and utilization of data sources and services



Proposed Solution

Establish a unified interface for aggregating data sources and services that can be crawled and indexed

GA4GH Deliverables

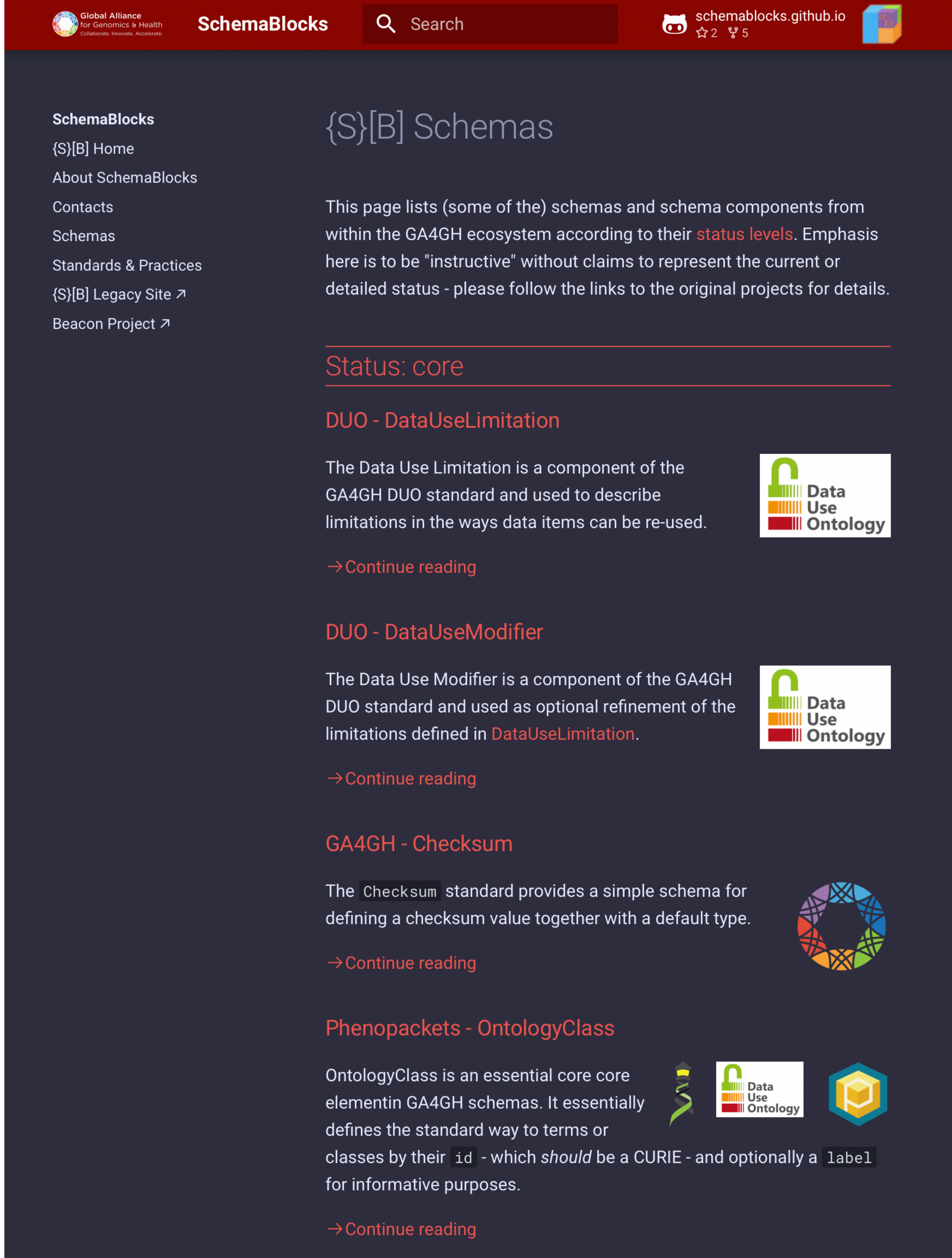
- Beacon API
- Data Connect API
- Service Info & Service Registry
- SchemaBlocks {S}[B]

GA4GH {S}[B]

SchemaBlocks

- “cross-workstreams, cross-drivers” initiative to document GA4GH object standards and prototypes, data formats and semantics
- documentation and implementation examples provided by GA4GH members
- no attempt to develop a rigid, complete data schema
- object vocabulary and semantics for a large range of developments
- currently not “authoritative GA4GH recommendations”
- GA4GH roadmap as element in "TASC"

schemablocks.org



Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.
SchemaBlocks
Search
schemablocks.github.io
☆2
5

SchemaBlocks

{S}[B] Home

About SchemaBlocks

Contacts

Schemas

Standards & Practices

{S}[B] Legacy Site ↗

Beacon Project ↗

{S}[B] Schemas

This page lists (some of the) schemas and schema components from within the GA4GH ecosystem according to their status levels. Emphasis here is to be "instructive" without claims to represent the current or detailed status - please follow the links to the original projects for details.

Status: core

DUO - DataUseLimitation

The Data Use Limitation is a component of the GA4GH DUO standard and used to describe limitations in the ways data items can be re-used.

→Continue reading

DUO - DataUseModifier

The Data Use Modifier is a component of the GA4GH DUO standard and used as optional refinement of the limitations defined in DataUseLimitation.

→Continue reading

GA4GH - Checksum

The `Checksum` standard provides a simple schema for defining a checksum value together with a default type.

→Continue reading

Phenopackets - OntologyClass

OntologyClass is an essential core core elementin GA4GH schemas. It essentially defines the standard way to terms or classes by their `id` - which *should* be a CURIE - and optionally a `label` for informative purposes.

→Continue reading

Service Info and Service Registry APIs

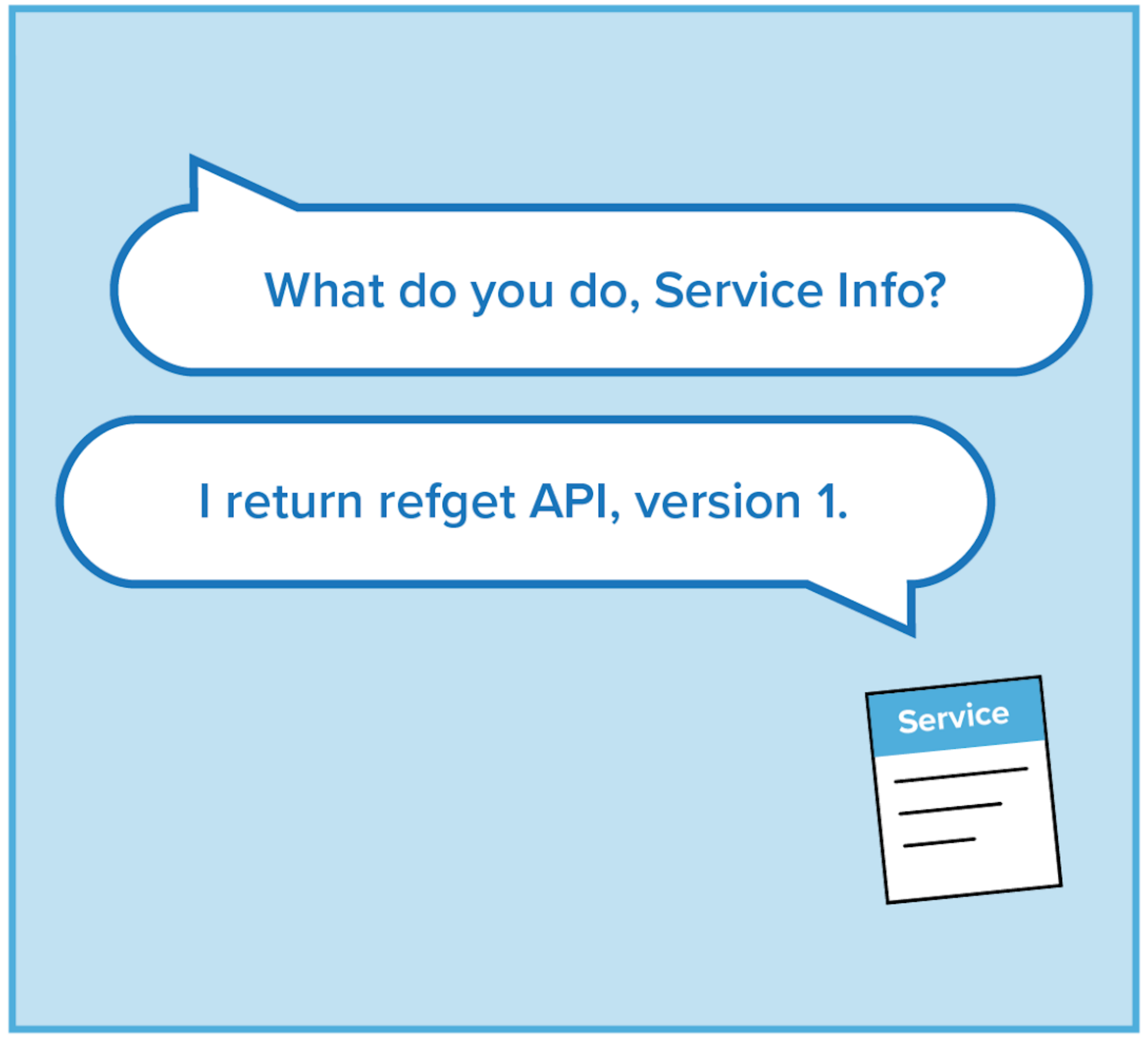
The Service Info and Registry APIs are minimalistic, light-weight APIs that provide a standard format for describing and listing genomics web services along with their metadata.

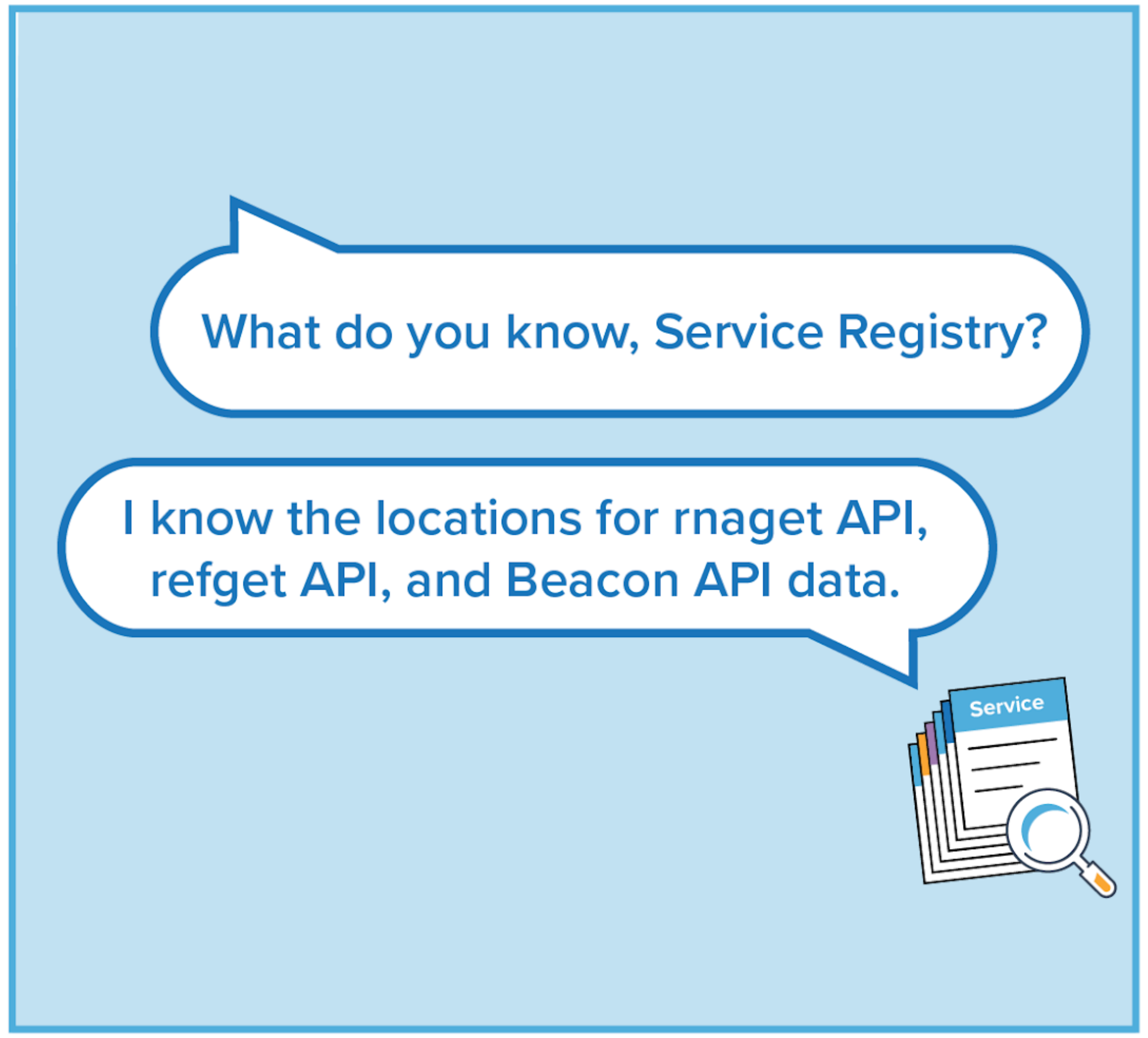
Approved: January 22, 2020

Example Users

elixir

AUTISM SHARING INITIATIVE


What do you do, Service Info?
I return refget API, version 1.
Service


What do you know, Service Registry?
I know the locations for rnaget API, refget API, and Beacon API data.
Service

The GA4GH Beacon Protocol

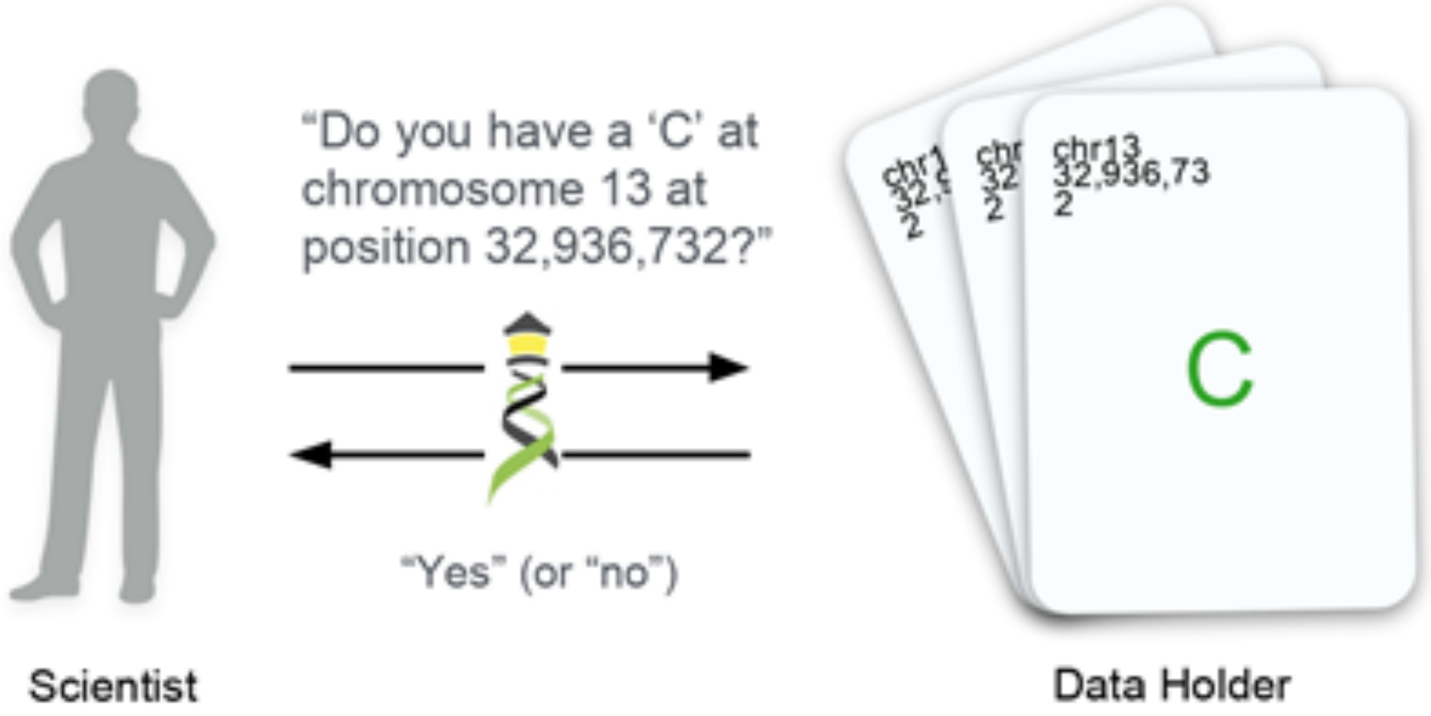
A GA4GH standard for genomics data discovery (and exchange)

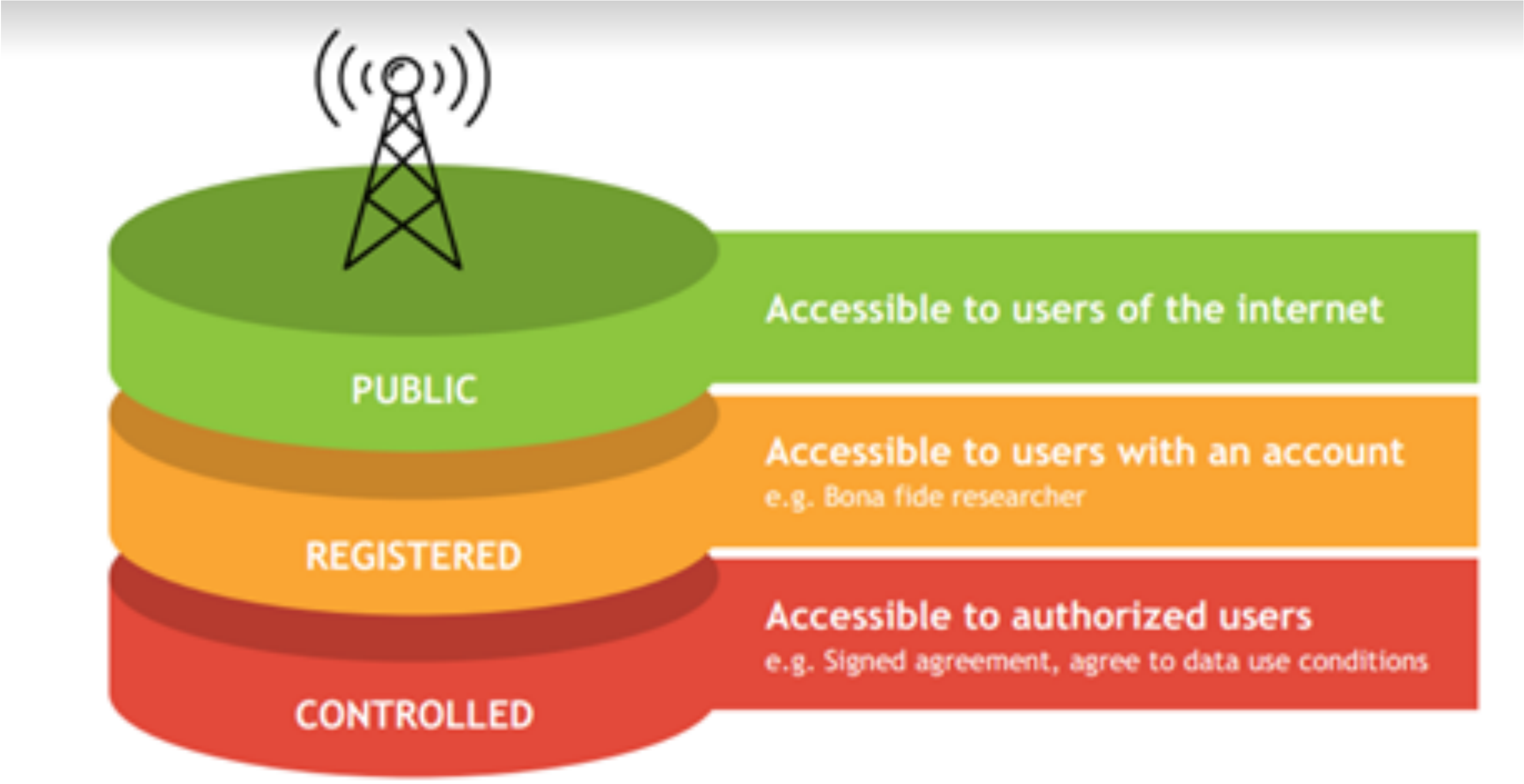
Beacon API v1

Global Alliance for Genomics & Health

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

Approved: October 3, 2018


"Do you have a 'C' at chromosome 13 at position 32,936,732?"
"Yes" (or "no")
chr13 32,936,73 2
C
Scientist
Data Holder


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

Example Users

EMBL-EBI

Australian Genomics

SciLifeLab

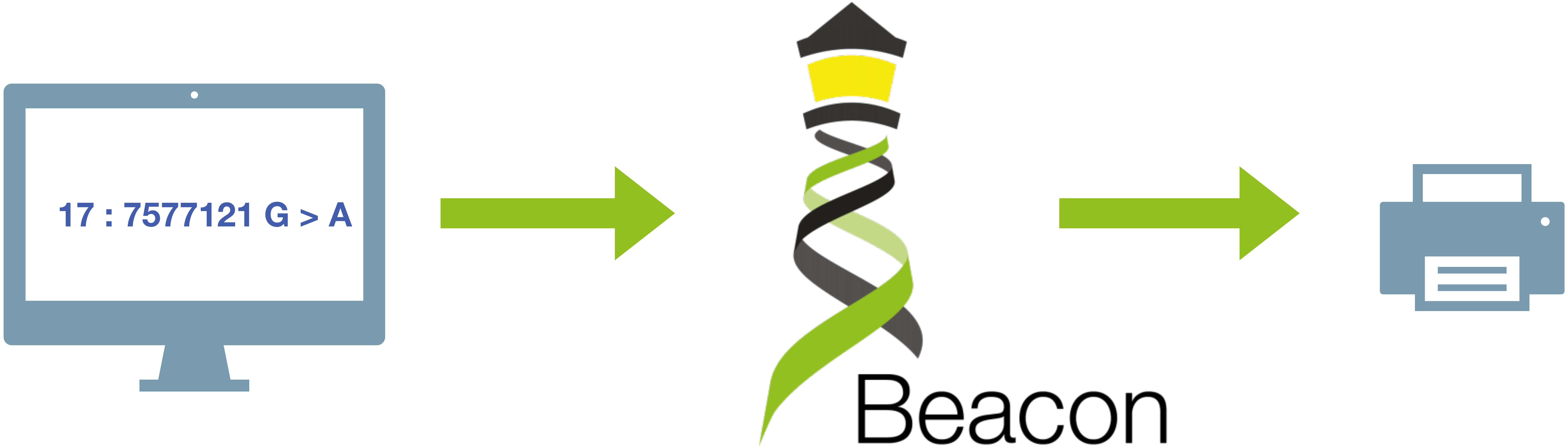
elixir

University of California Santa Cruz

Broad Institute

European Genome-Phenome Archive

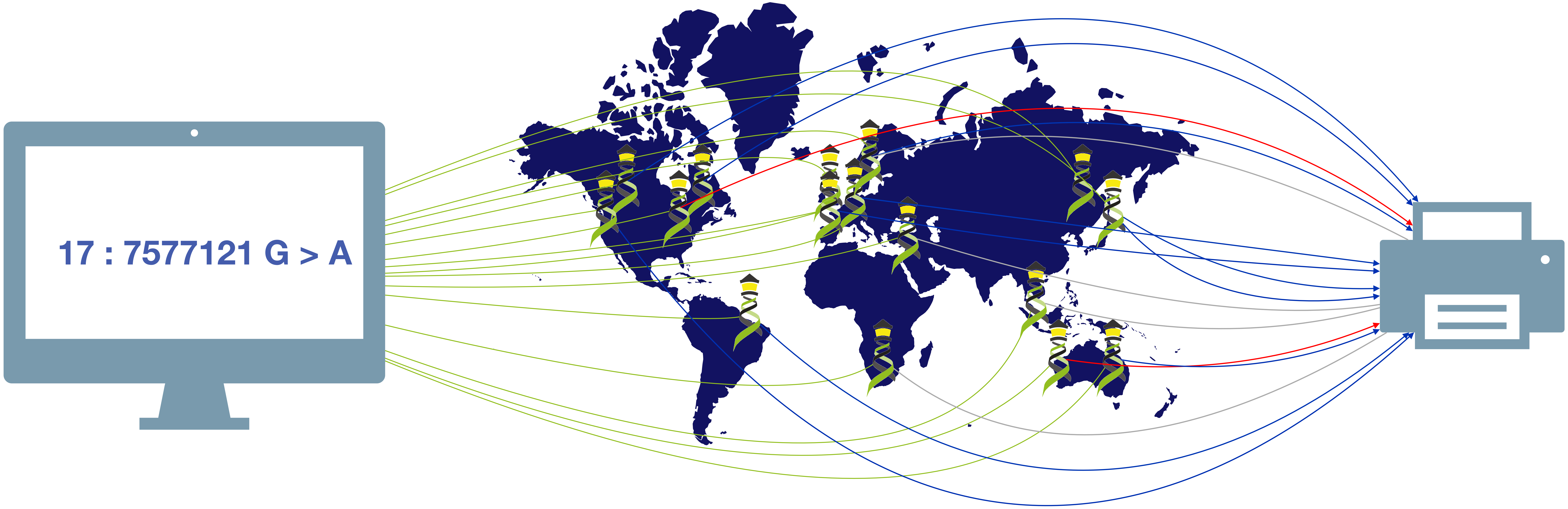
International Cancer Genome Consortium



17 : 7577121 G > A

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*** …

ELIXIR - Making Beacons Biomedical

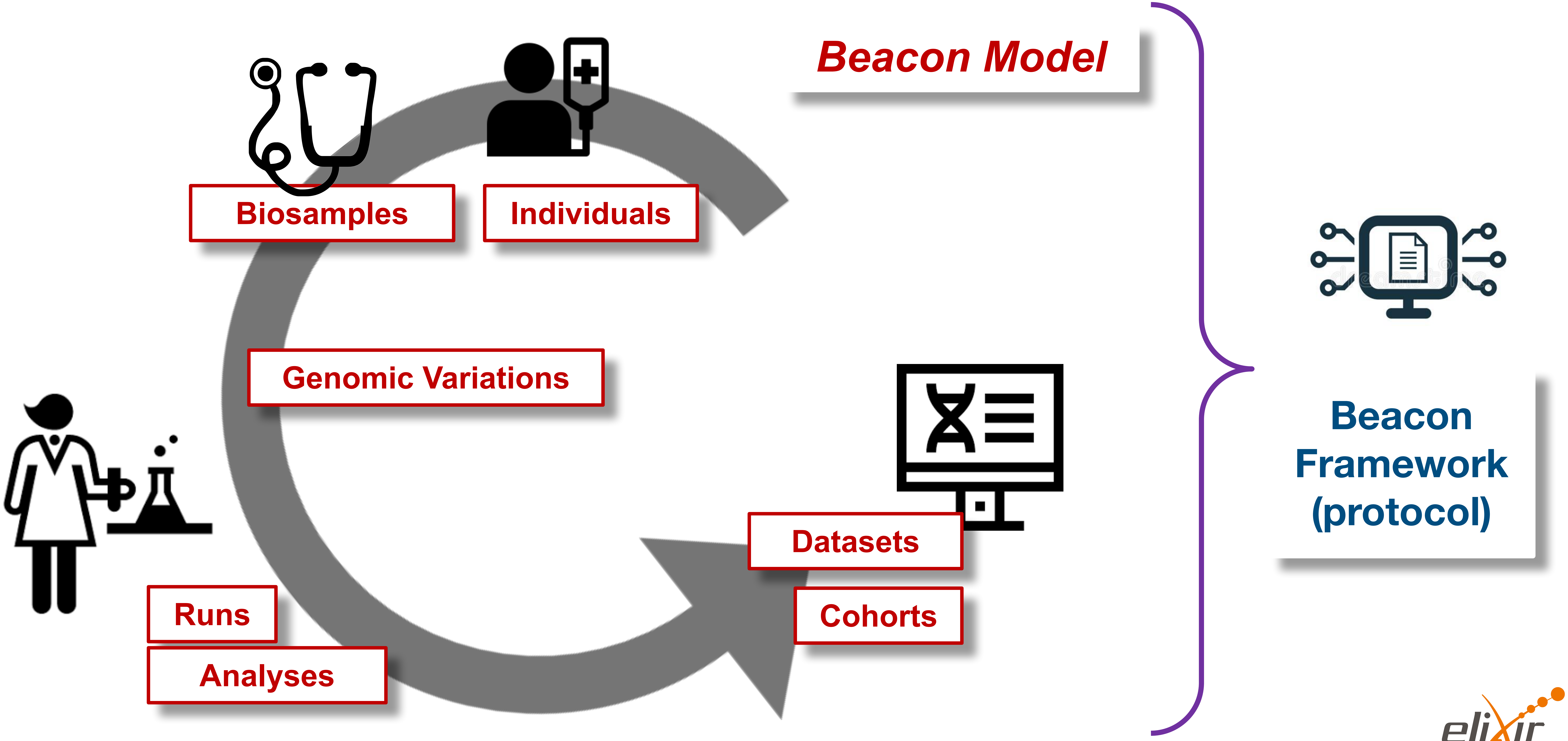
- Authentication to enable non-aggregate, patient derived datasets
  - ELIXIR AAI with compatibility to other providers (OAuth...)
- Scoping queries through "biodata" parameters
- Extending the queries towards clinically ubiquitous variant formats
  - cytogenetic annotations, named variants, variant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

2016 and beyond ...

	Beacon v1 Development	Beacon v2 Development	Related ...
2014	GA4GH founding event; Jim Ostell proposes Beacon concept with "more features… version 2"		
2015	• beacon-network.org aggregator created by DNAstack • Beacon v0.3 release		• ELIXIR starts Beacon project support
2016	• 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	• Beacon hackathon Stockholm; settling on "filters" • Barcelona goes Zurich developers meeting	• new Beacon website (March)
2019	• ELIXIR Beacon Network	• 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 • starting of GA4GH review process	
2022		• further changes esp. in default model, aligning with Phenopackets and VRS • unified beacon-v2 code & docs repository • **Beacon v2 approved at Apr GA4GH Connect**	• *docs.genomebeacons.org*

Beacon v2

docs.genomebeacons.org



Beacon Model

Biosamples

Individuals

Genomic Variations

Runs

Analyses

Datasets

Cohorts

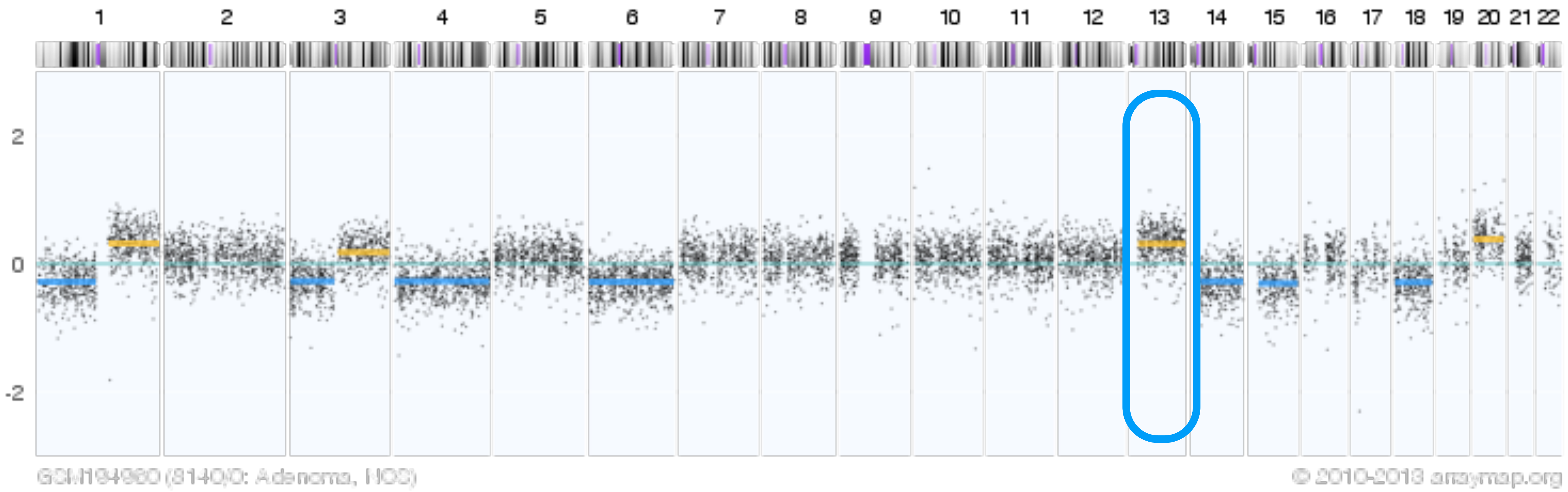
Beacon Framework (protocol)

Progenetix and GA4GH Beacon

Implementation driven development of a GA4GH standard

progenetix

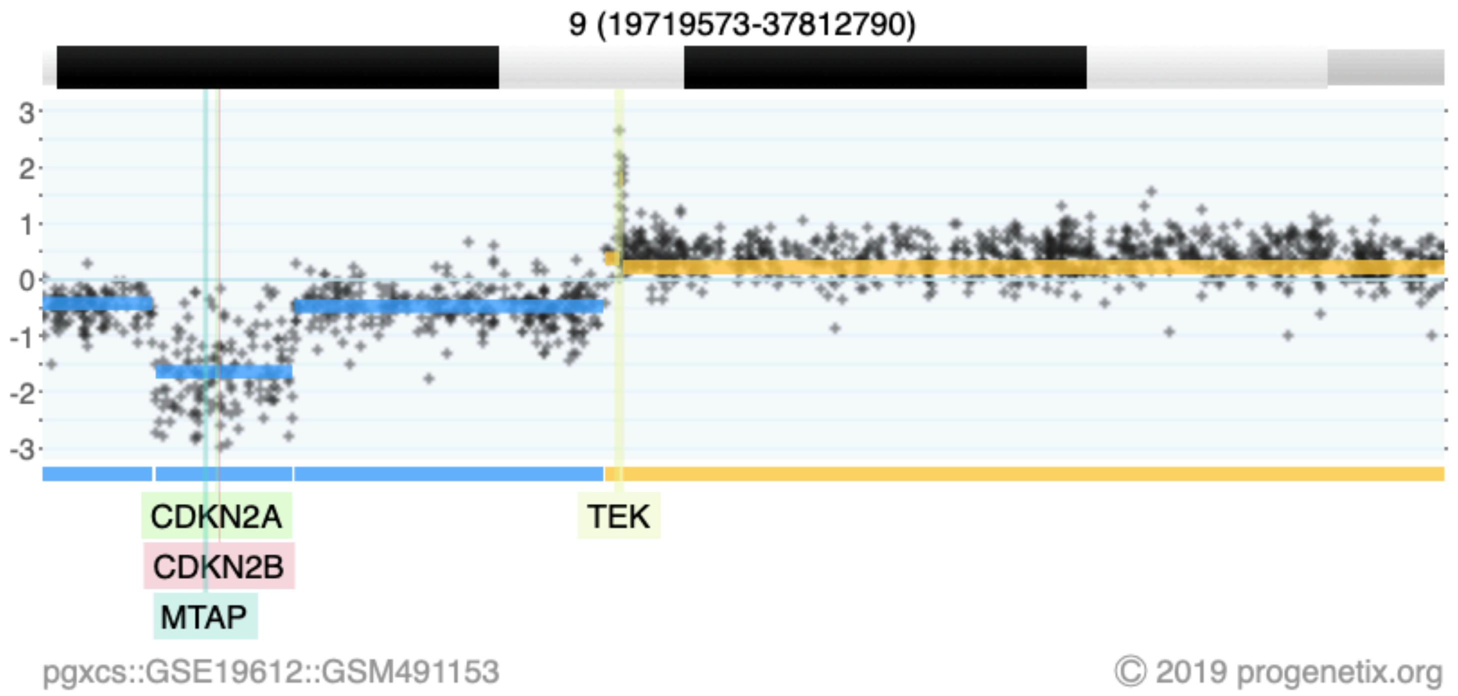
Somatic Copy Number Variations



GSM194980 (8140/0: Adenoma, NOS)

© 2010-2013 arraymap.org

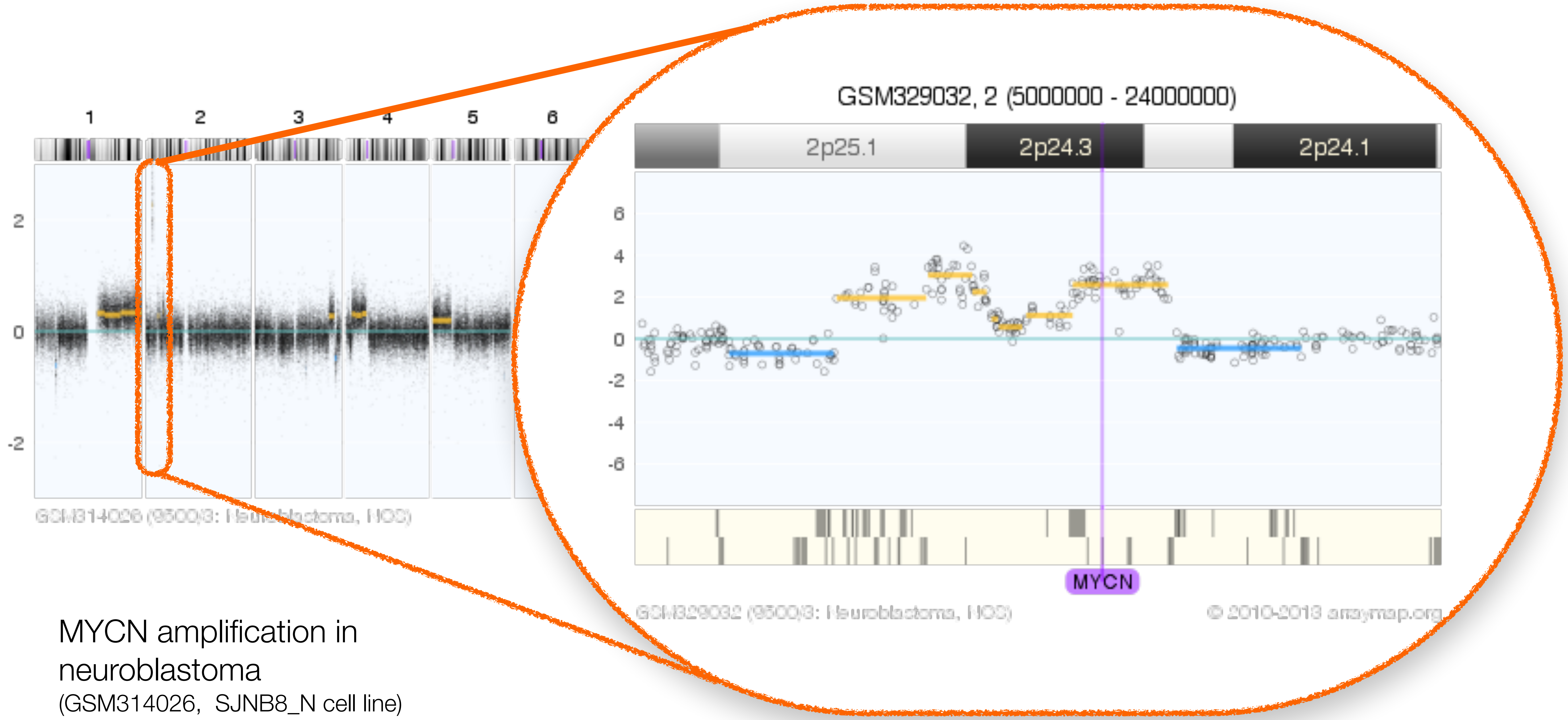
Gain of chromosome arm 13q in colorectal carcinoma



pgxcs::GSE19612::GSM491153

© 2019 progenetix.org

2-event, homozygous deletion in a Glioblastoma



GSM314026 (9500/3: Neuroblastoma, NOS)

GSM329032 (9500/3: Neuroblastoma, NOS)

© 2010-2013 arraymap.org

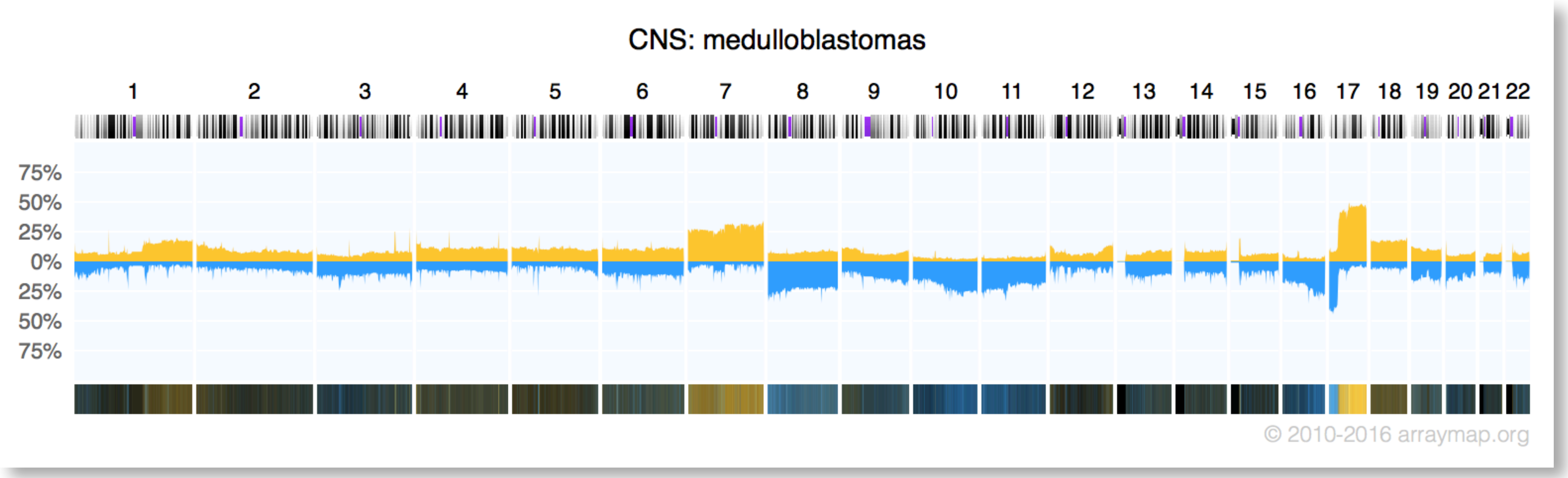
MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

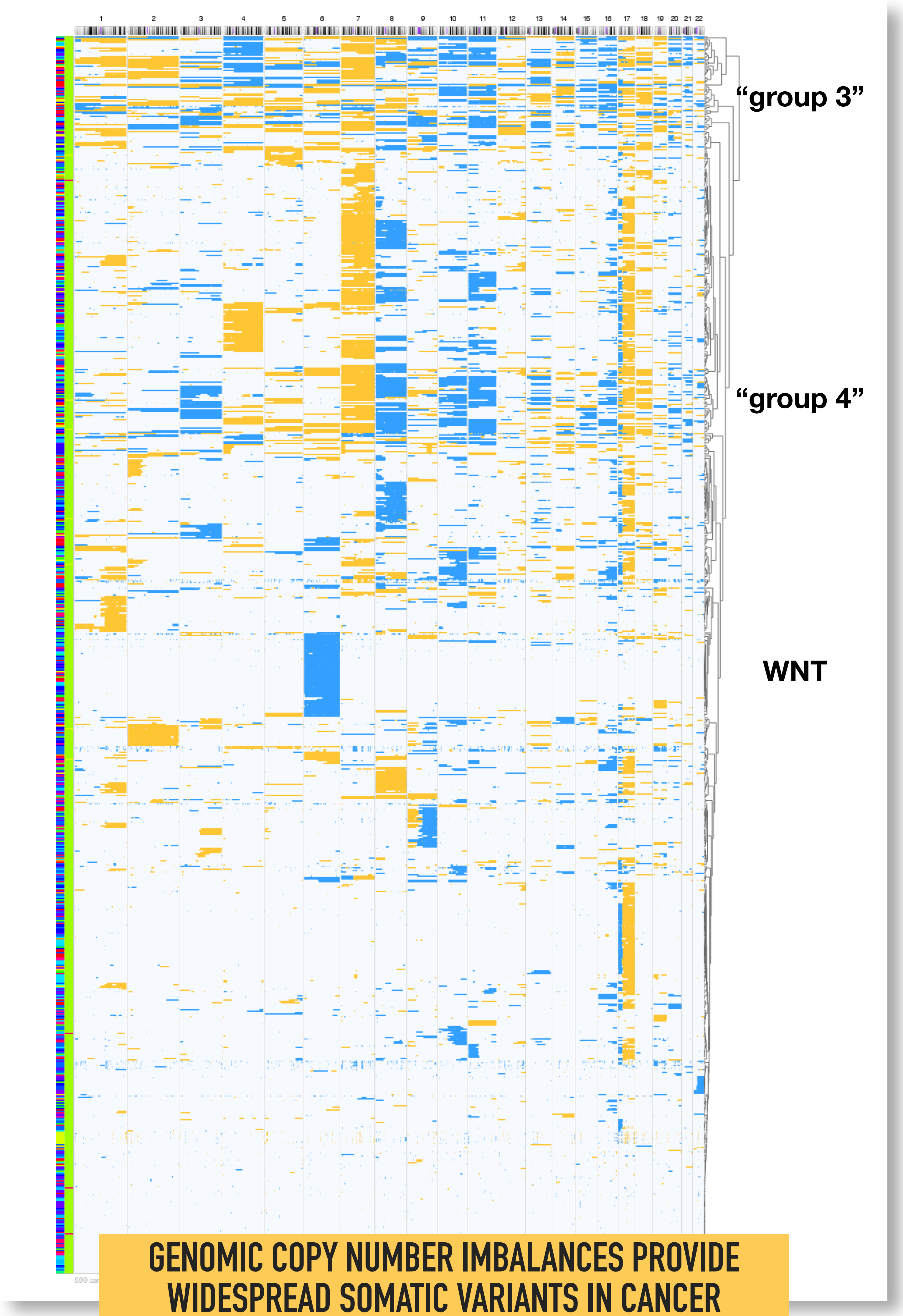
low level/**high level** copy number alterations (CNAs)

Somatic CNVs In Cancer

Recurrent mutation patterns

How can those patterns be used for classification and determination of biological mechanisms?





GENOMIC COPY NUMBER IMBALANCES PROVIDE WIDESPREAD SOMATIC VARIANTS IN CANCER

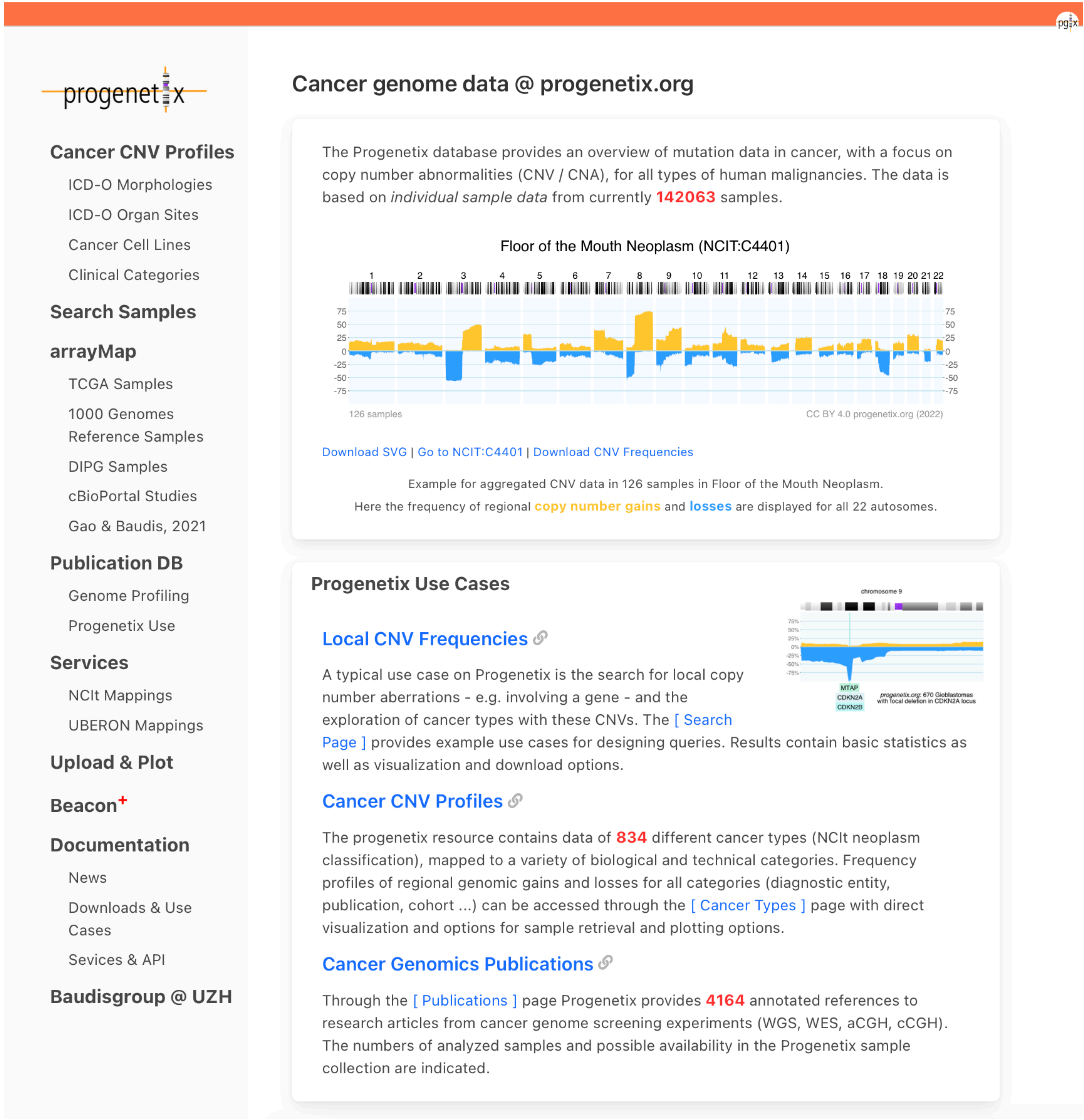
A genomic copy number histogram for malignant medulloblastomas, the most frequent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional relevance and clinical correlation.

progenetix

Progenetix in 2022

Cancer Genomics Reference Resource

- open resource for curated oncogenomic profiles
- >116'000 cancer CNV profiles, from >800 types
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core biosample and technical metadata where accessible (TNM, sex, survival ...)
- publication database and code mapping services



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
- Services & API

Baudisgroup @ UZH

Cancer genome data @ progenetix.org

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

Floor of the Mouth Neoplasm (NCIT:C4401)

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

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

Local CNV Frequencies

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

Cancer CNV Profiles

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

Cancer Genomics Publications

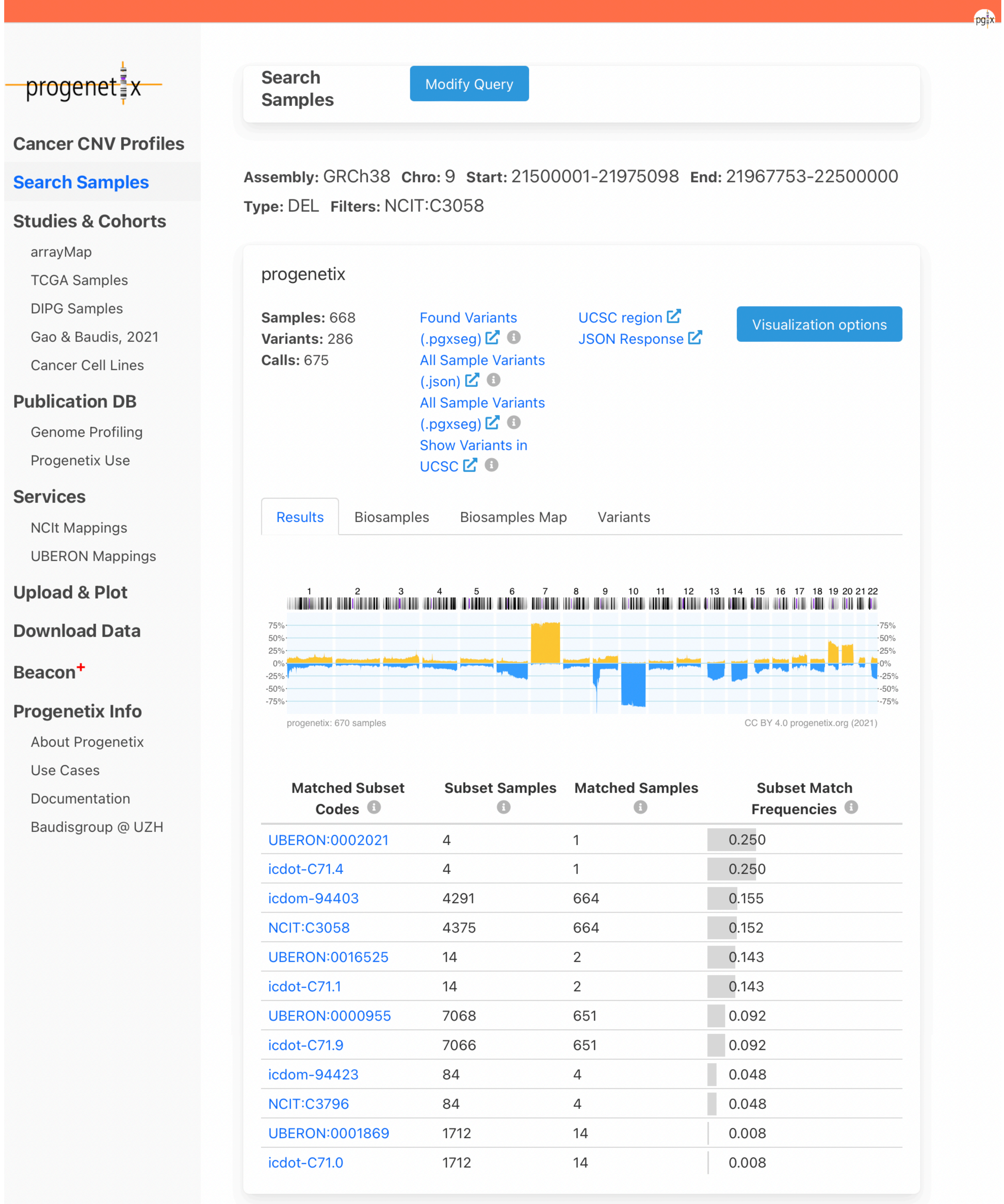
Through the [Publications] page Progenetix provides **4164** annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.

progenetix

Progenetix in 2022

Cancer Genomics Reference Resource

- open resource for curated oncogenomic profiles
- >116'000 cancer CNV profiles, from >800 types
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core biosample and technical metadata where accessible (TNM, sex, survival ...)
- publication database and code mapping services



Assembly: GRCh38 **Chro:** 9 **Start:** 21500001-21975098 **End:** 21967753-22500000
Type: DEL **Filters:** NCIT:C3058

progenetix

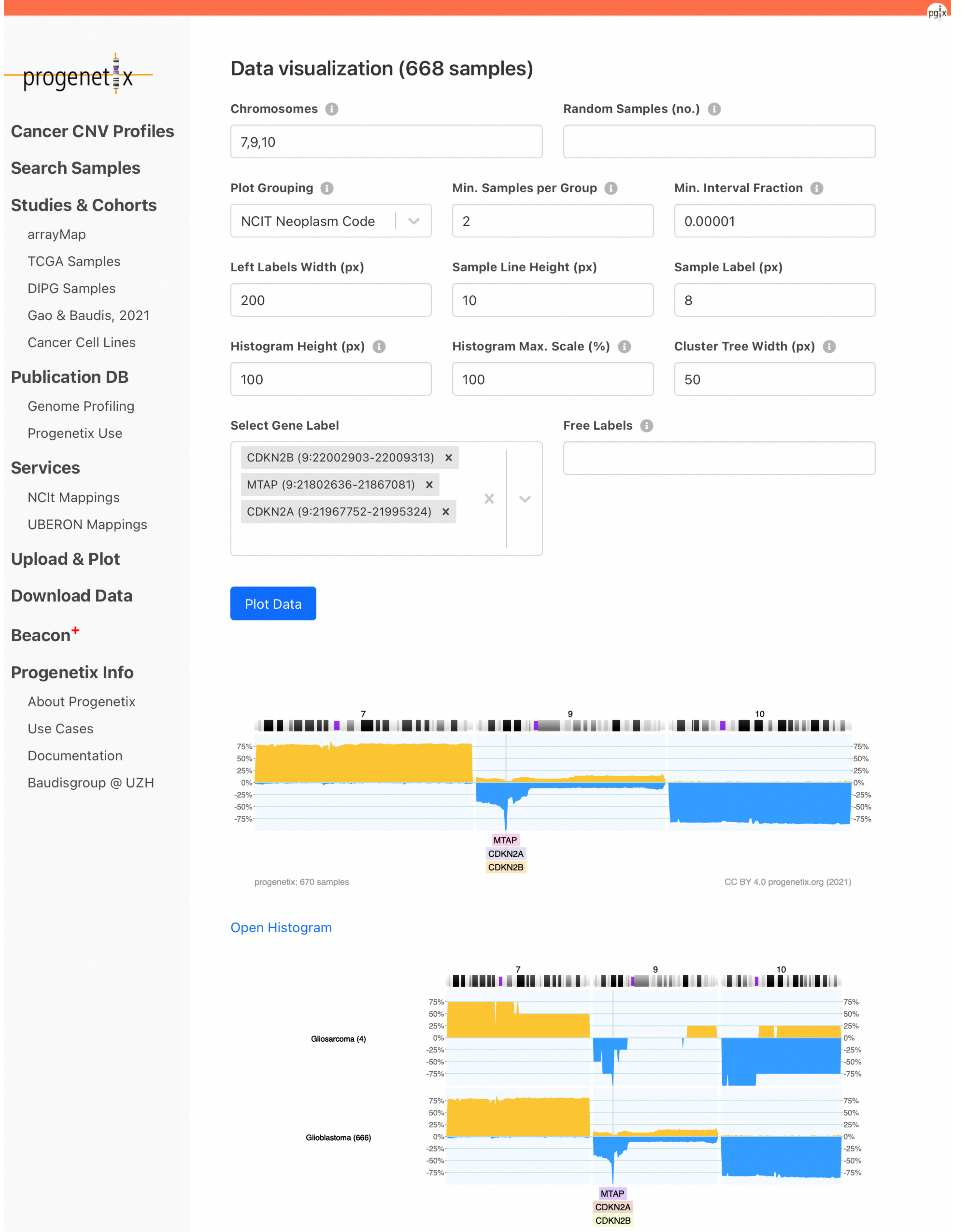
Samples: 668
Variants: 286
Calls: 675

Matched Subset Codes	Subset Samples	Matched Samples	Subset Match Frequencies
UBERON:0002021	4	1	0.250
icdot-C71.4	4	1	0.250
icdom-94403	4291	664	0.155
NCIT:C3058	4375	664	0.152
UBERON:0016525	14	2	0.143
icdot-C71.1	14	2	0.143
UBERON:0000955	7068	651	0.092
icdot-C71.9	7066	651	0.092
icdom-94423	84	4	0.048
NCIT:C3796	84	4	0.048
UBERON:0001869	1712	14	0.008
icdot-C71.0	1712	14	0.008

Progenetix in 2022

Cancer Genomics Reference Resource

- open resource for curated oncogenomic profiles
- >116'000 cancer CNV profiles, from >800 types
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core biosample and technical metadata where accessible (TNM, sex, survival ...)
- publication database and code mapping services



Data visualization (668 samples)

Chromosomes: 7,9,10

Random Samples (no.):

Plot Grouping: NCIT Neoplasm Code

Min. Samples per Group: 2

Min. Interval Fraction: 0.00001

Left Labels Width (px): 200

Sample Line Height (px): 10

Sample Label (px): 8

Histogram Height (px): 100

Histogram Max. Scale (%): 100

Cluster Tree Width (px): 50

Select Gene Label: CDKN2B (9:22002903-22009313) × MTAP (9:21802636-21867081) × CDKN2A (9:21967752-21995324) ×

Free Labels:

progenetix

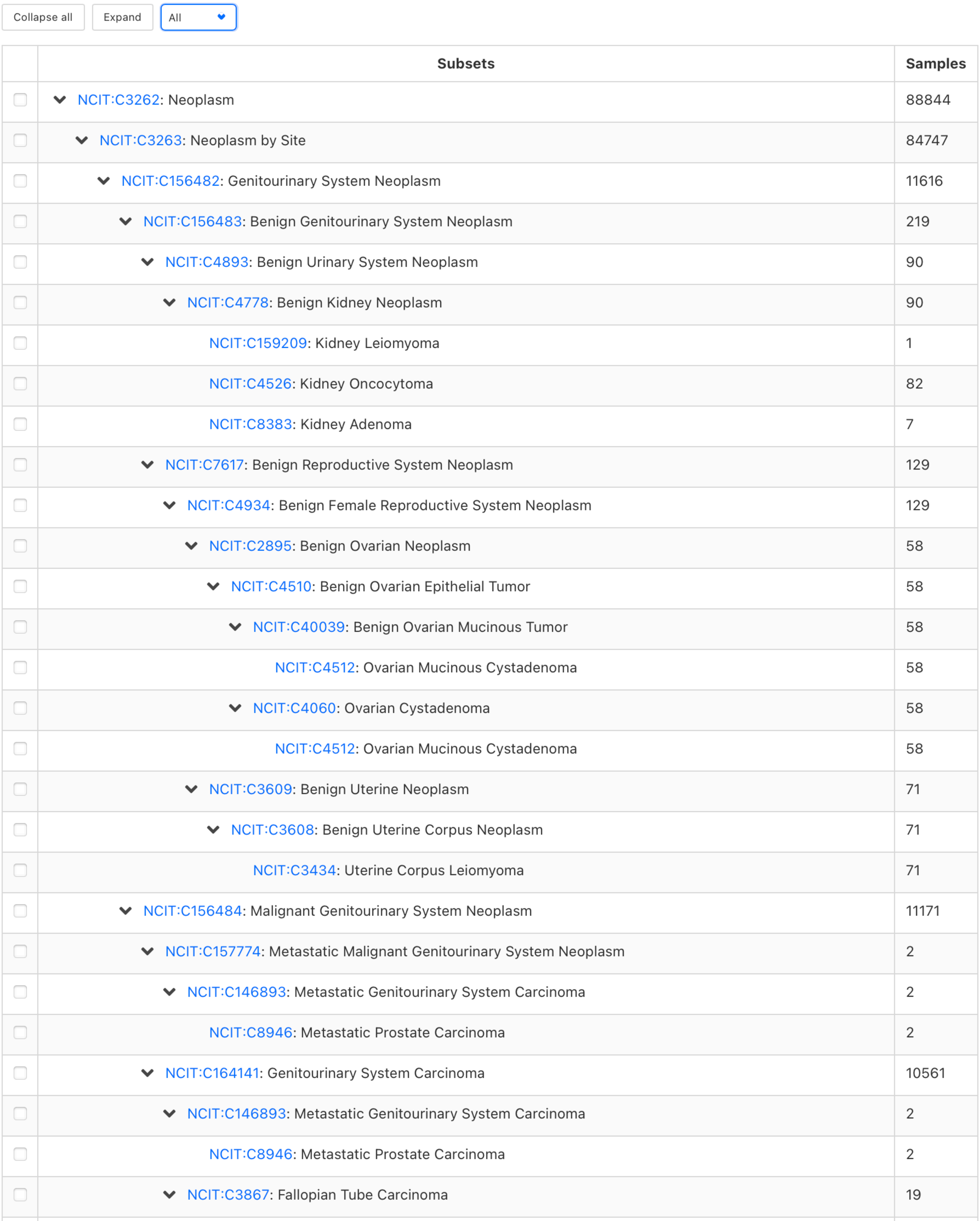
DX Ontologies

Example: Hierarchical NCIt Neoplasm Core replaces heterogeneous primary annotations

- heterogeneous and inconsistent diagnostic annotations are common in clinical reports and research studies ("text", ICD-10, ICD-O 3, OncoTree, domain-specific)
- highly **variable granularity** of annotations as major road block for large scale data integration
  - "Colorectal Cancer" or "Rectal Mucinous Adenoca."
- initiatives and services such as **Phenopackets**, MONDO, OXO ... rely on and/or provide mappings to hierarchical ontologies
- consistent CURIEs were instrumental in the development & testing of the Beacon v2 "Filters" paradigm
  - final "Filtering Term" object dev. by Tim Beck, U. of Leicester

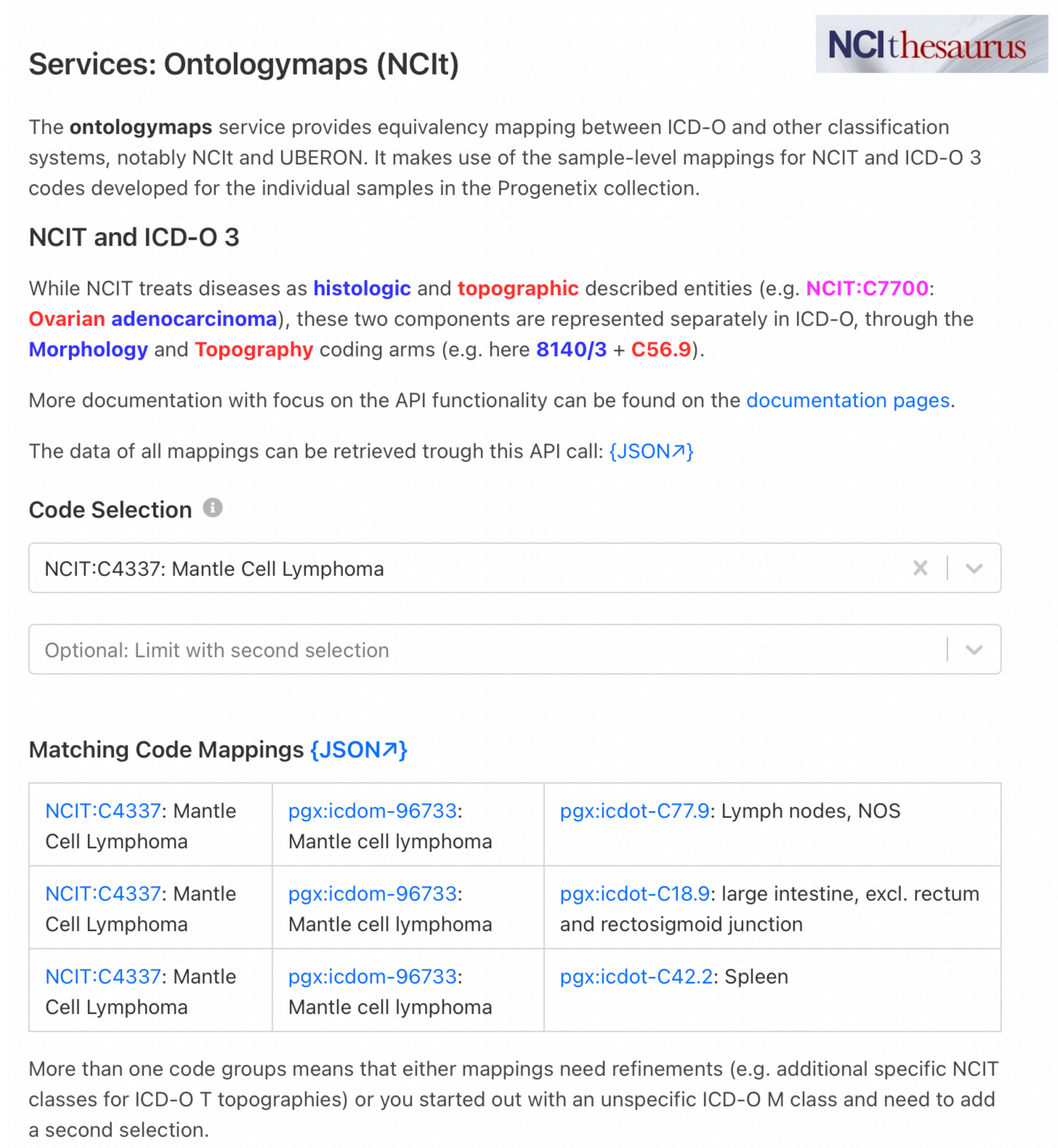
progenetix

NCIt Neoplasm Core coded display (excerpt) for samples in the Progenetix cancer genome data resource allows sample selection on multiple hierarchy levels →



	Subsets	Samples
□	NCIT:C3262: Neoplasm	88844
□	NCIT:C3263: Neoplasm by Site	84747
□	NCIT:C156482: Genitourinary System Neoplasm	11616
□	NCIT:C156483: Benign Genitourinary System Neoplasm	219
□	NCIT:C4893: Benign Urinary System Neoplasm	90
□	NCIT:C4778: Benign Kidney Neoplasm	90
□	NCIT:C159209: Kidney Leiomyoma	1
□	NCIT:C4526: Kidney Oncocytoma	82
□	NCIT:C8383: Kidney Adenoma	7
□	NCIT:C7617: Benign Reproductive System Neoplasm	129
□	NCIT:C4934: Benign Female Reproductive System Neoplasm	129
□	NCIT:C2895: Benign Ovarian Neoplasm	58
□	NCIT:C4510: Benign Ovarian Epithelial Tumor	58
□	NCIT:C40039: Benign Ovarian Mucinous Tumor	58
□	NCIT:C4512: Ovarian Mucinous Cystadenoma	58
□	NCIT:C4060: Ovarian Cystadenoma	58
□	NCIT:C4512: Ovarian Mucinous Cystadenoma	58
□	NCIT:C3609: Benign Uterine Neoplasm	71
□	NCIT:C3608: Benign Uterine Corpus Neoplasm	71
□	NCIT:C3434: Uterine Corpus Leiomyoma	71
□	NCIT:C156484: Malignant Genitourinary System Neoplasm	11171
□	NCIT:C157774: Metastatic Malignant Genitourinary System Neoplasm	2
□	NCIT:C146893: Metastatic Genitourinary System Carcinoma	2
□	NCIT:C8946: Metastatic Prostate Carcinoma	2
□	NCIT:C164141: Genitourinary System Carcinoma	10561
□	NCIT:C146893: Metastatic Genitourinary System Carcinoma	2
□	NCIT:C8946: Metastatic Prostate Carcinoma	2
□	NCIT:C3867: Fallopian Tube Carcinoma	19

Ontologies and Classifications



NCIthesaurus

Services: Ontologymaps (NCIt)

The **ontologymaps** service provides equivalency mapping between ICD-O and other classification systems, notably NCIt and UBERON. It makes use of the sample-level mappings for NCIT and ICD-O 3 codes developed for the individual samples in the Progenetix collection.

NCIT and ICD-O 3

While NCIT treats diseases as **histologic** and **topographic** described entities (e.g. **NCIT:C7700**: **Ovarian adenocarcinoma**), these two components are represented separately in ICD-O, through the **Morphology** and **Topography** coding arms (e.g. here **8140/3** + **C56.9**).

More documentation with focus on the API functionality can be found on the documentation pages.

The data of all mappings can be retrieved trough this API call: {JSON↗}

Code Selection

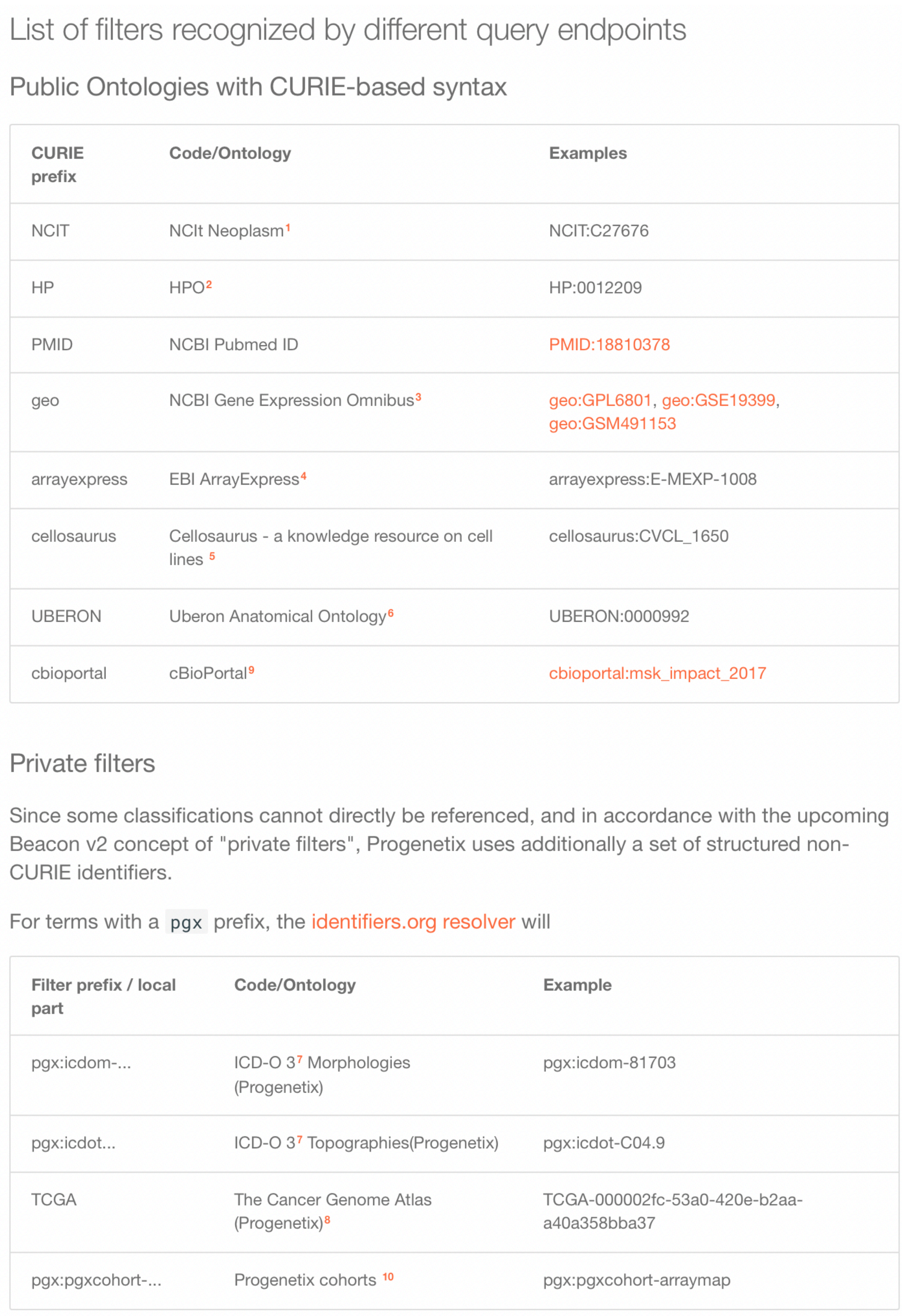
NCIT:C4337: Mantle Cell Lymphoma

Optional: Limit with second selection

Matching Code Mappings {JSON↗}

NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C77.9: Lymph nodes, NOS
NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C18.9: large intestine, excl. rectum and rectosigmoid junction
NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C42.2: Spleen

More than one code groups means that either mappings need refinements (e.g. additional specific NCIT classes for ICD-O T topographies) or you started out with an unspecific ICD-O M class and need to add a second selection.



List of filters recognized by different query endpoints

Public Ontologies with CURIE-based syntax

CURIE prefix	Code/Ontology	Examples
NCIT	NCIt Neoplasm[1]	NCIT:C27676
HP	HPO[2]	HP:0012209
PMID	NCBI Pubmed ID	PMID:18810378
geo	NCBI Gene Expression Omnibus[3]	geo:GPL6801, geo:GSE19399, geo:GSM491153
arrayexpress	EBI ArrayExpress[4]	arrayexpress:E-MEXP-1008
cellosaurus	Cellosaurus - a knowledge resource on cell lines [5]	cellosaurus:CVCL_1650
UBERON	Uberon Anatomical Ontology[6]	UBERON:0000992
cbioportal	cBioPortal[9]	cbioportal:msk_impact_2017

Private filters

Since some classifications cannot directly be referenced, and in accordance with the upcoming Beacon v2 concept of "private filters", Progenetix uses additionally a set of structured non-CURIE identifiers.

For terms with a `pgx` prefix, the identifiers.org resolver will

Filter prefix / local part	Code/Ontology	Example
pgx:icdom-...	ICD-O 3[7] Morphologies (Progenetix)	pgx:icdom-81703
pgx:icdot...	ICD-O 3[7] Topographies(Progenetix)	pgx:icdot-C04.9
TCGA	The Cancer Genome Atlas (Progenetix)[8]	TCGA-000002fc-53a0-420e-b2aa-a40a358bba37
pgx:pgxcohort-...	Progenetix cohorts [10]	pgx:pgxcohort-arraymap

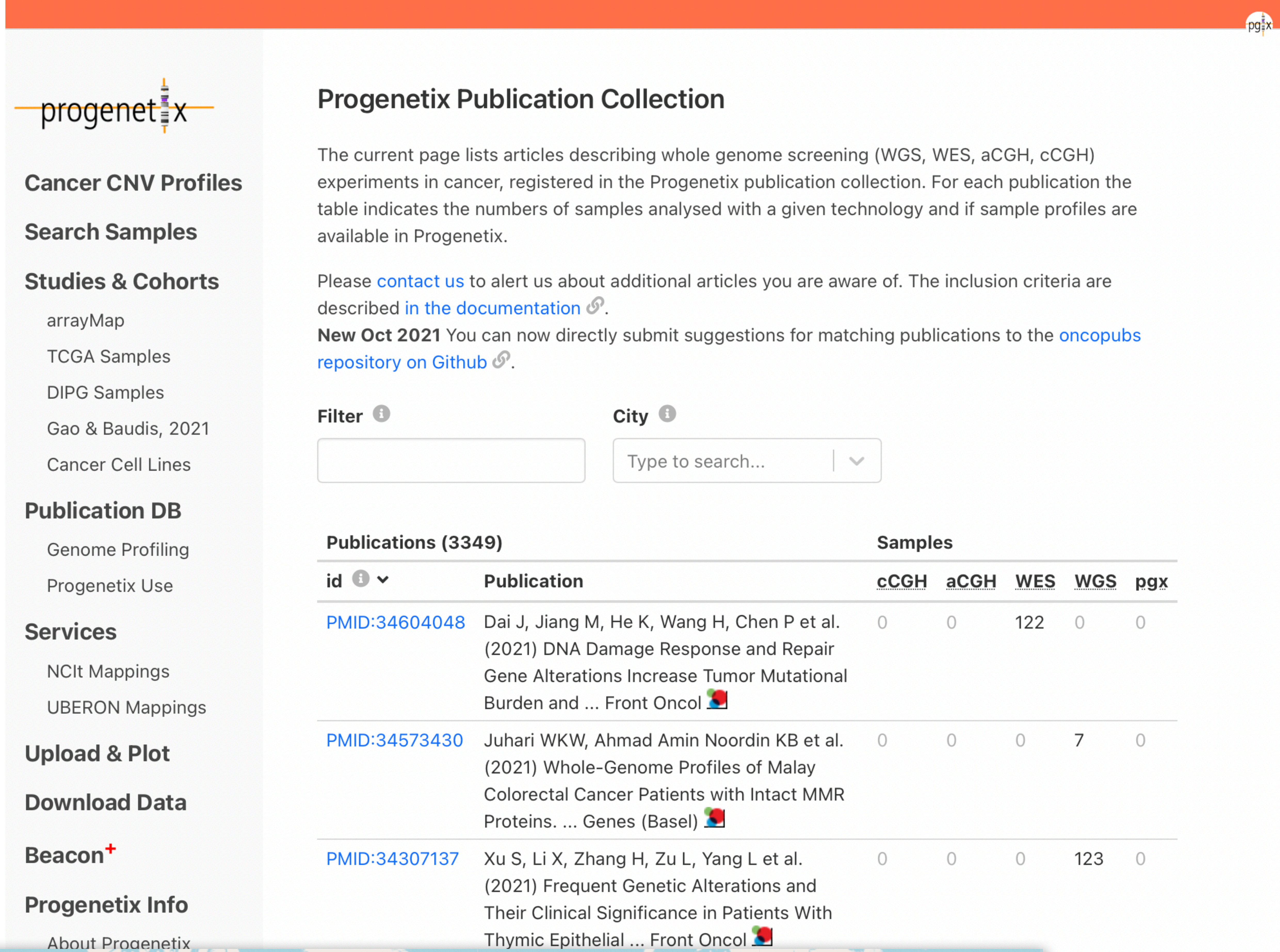
In Progenetix all cancer diagnoses are coded to both NCIt neoplasm codes and ICD-O 3 Morphology + Topography combinations. The matched mappings are provided as lookup-service since neither an official ICD-O ontology nor such a "disease defined by ICD-O M+T" concept is codified anywhere.

Progenetix in 2022

Cancer Genomics Reference Resource

- open resource for curated oncogenomic profiles
- >116'000 cancer CNV profiles, from >800 types
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core biosample and technical metadata where accessible (TNM, sex, survival ...)
- publication database and code mapping services

progenetix



Progenetix Publication Collection

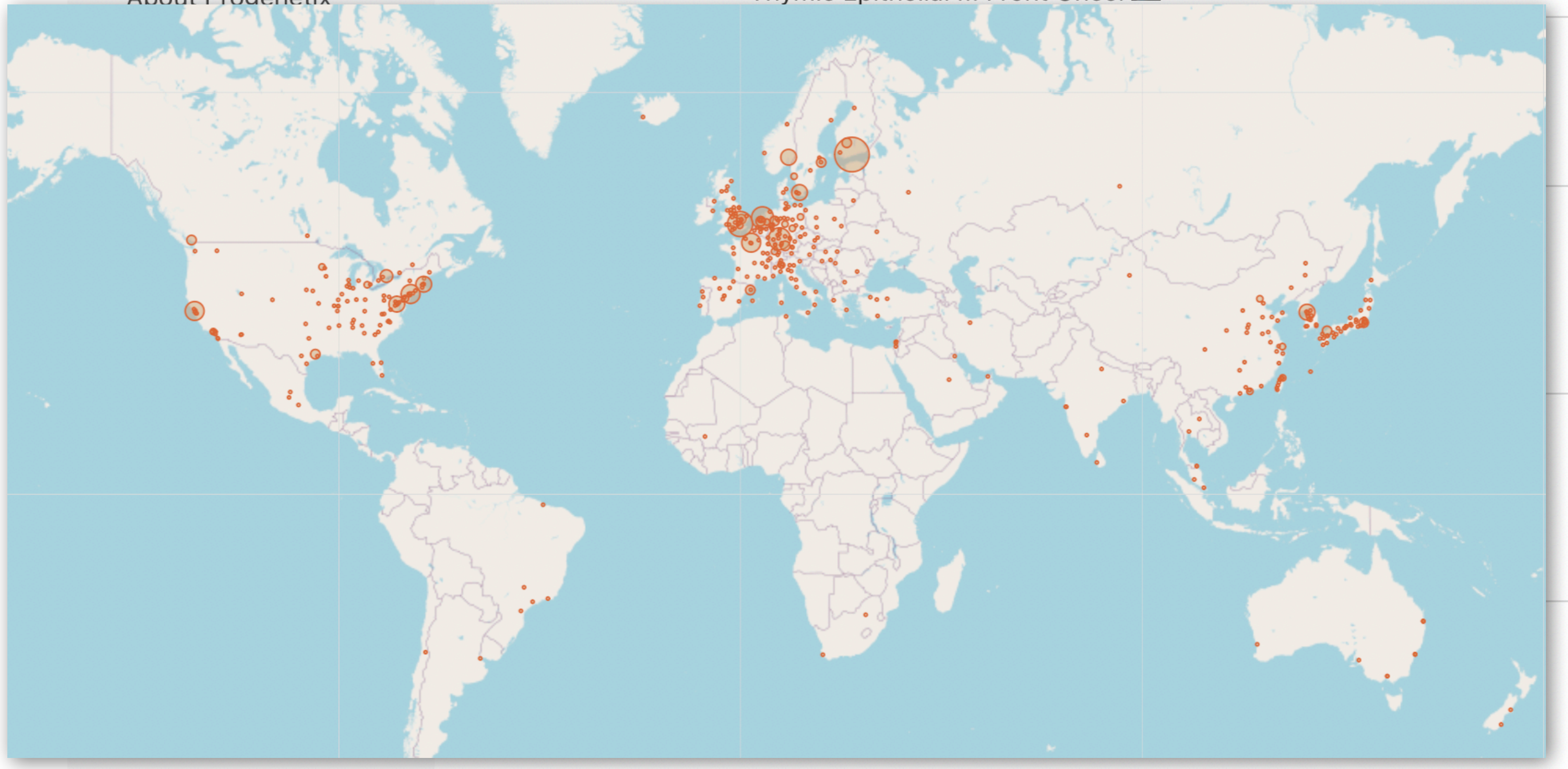
The current page lists articles describing whole genome screening (WGS, WES, aCGH, cCGH) experiments in cancer, registered in the Progenetix publication collection. For each publication the table indicates the numbers of samples analysed with a given technology and if sample profiles are available in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

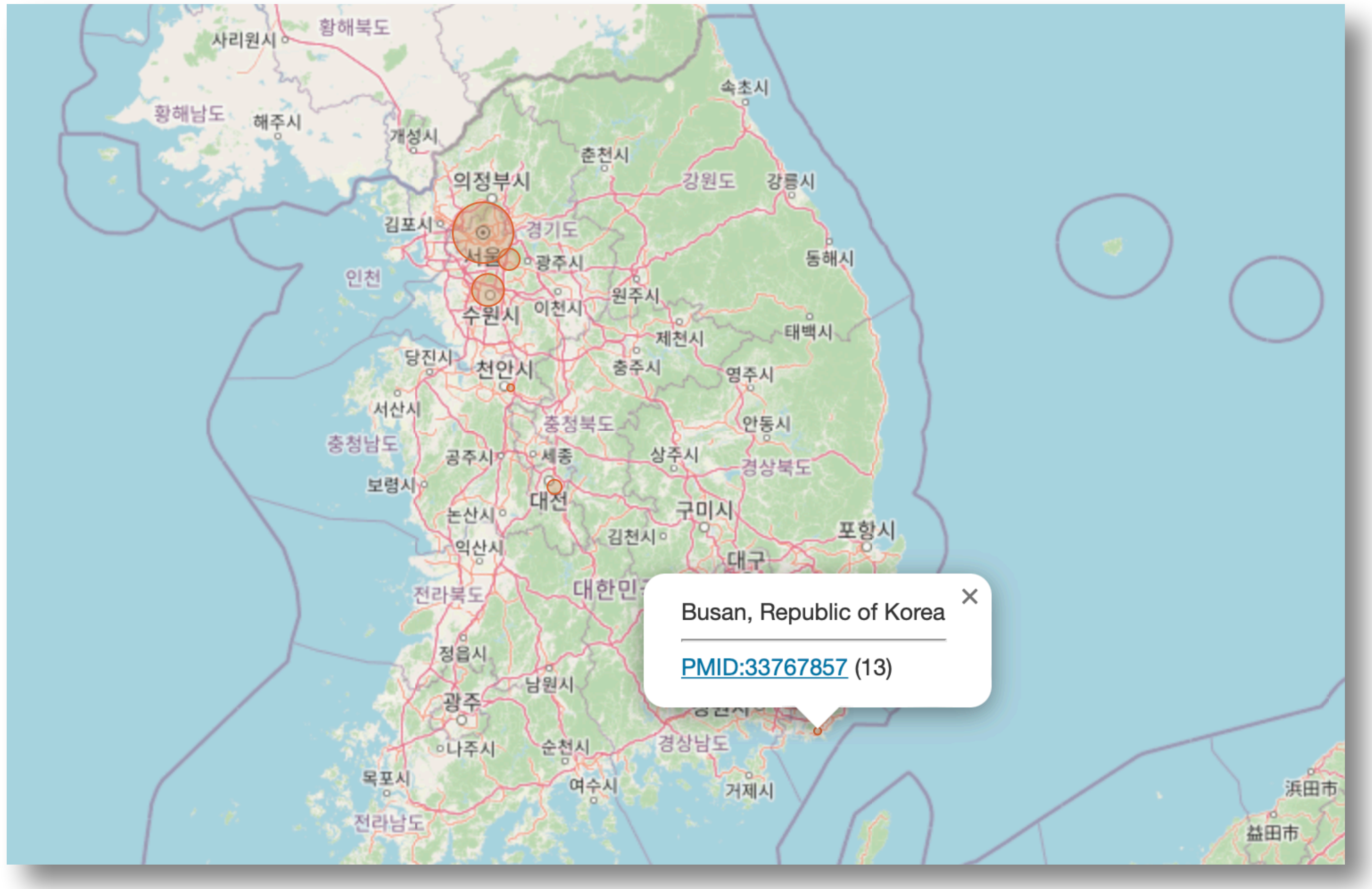
Publications (3349)		Samples				
id	Publication	cCGH	aCGH	WES	WGS	pgx
PMID:34604048	Dai J, Jiang M, He K, Wang H, Chen P et al. (2021) DNA Damage Response and Repair Gene Alterations Increase Tumor Mutational Burden and ... Front Oncol	0	0	122	0	0
PMID:34573430	Juhari WKW, Ahmad Amin Noordin KB et al. (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



Service: Publications

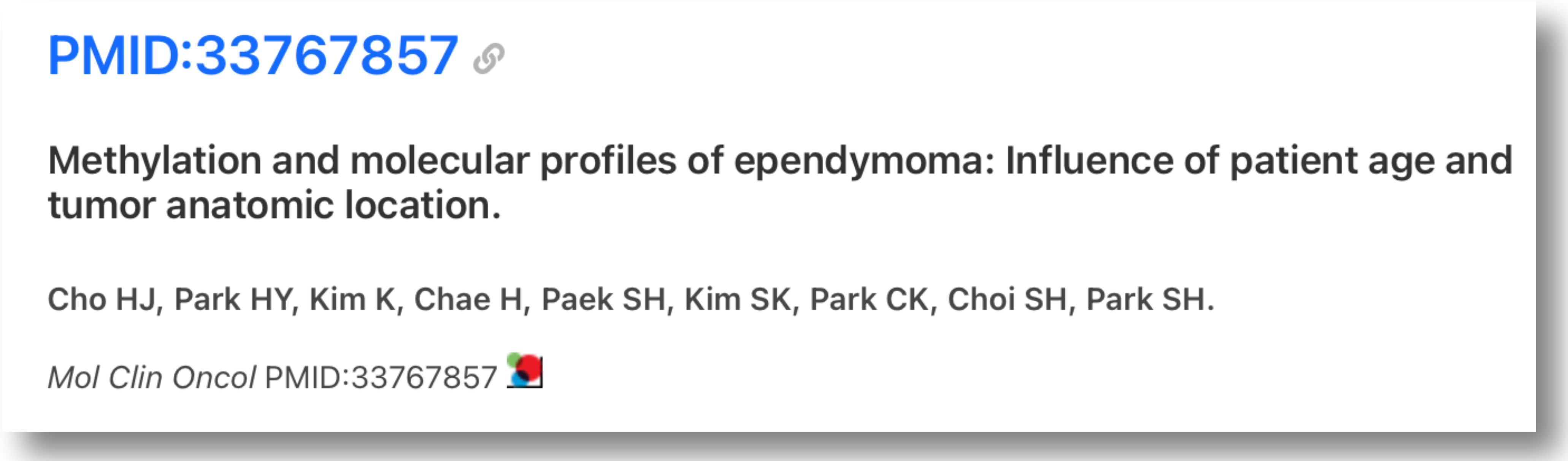
Location Mapping for Statistics and Discovery...

- all publications are tagged for "best fit" geographic origin in order
  1. specific sample origin
  2. processing laboratory
  3. corresponding author
- enables searches for e.g. "all publications or samples in HCC from 2000km around Taipeh"
- handy utility for discovering locally performed research, partners...



Busan, Republic of Korea

PMID:33767857 (13)



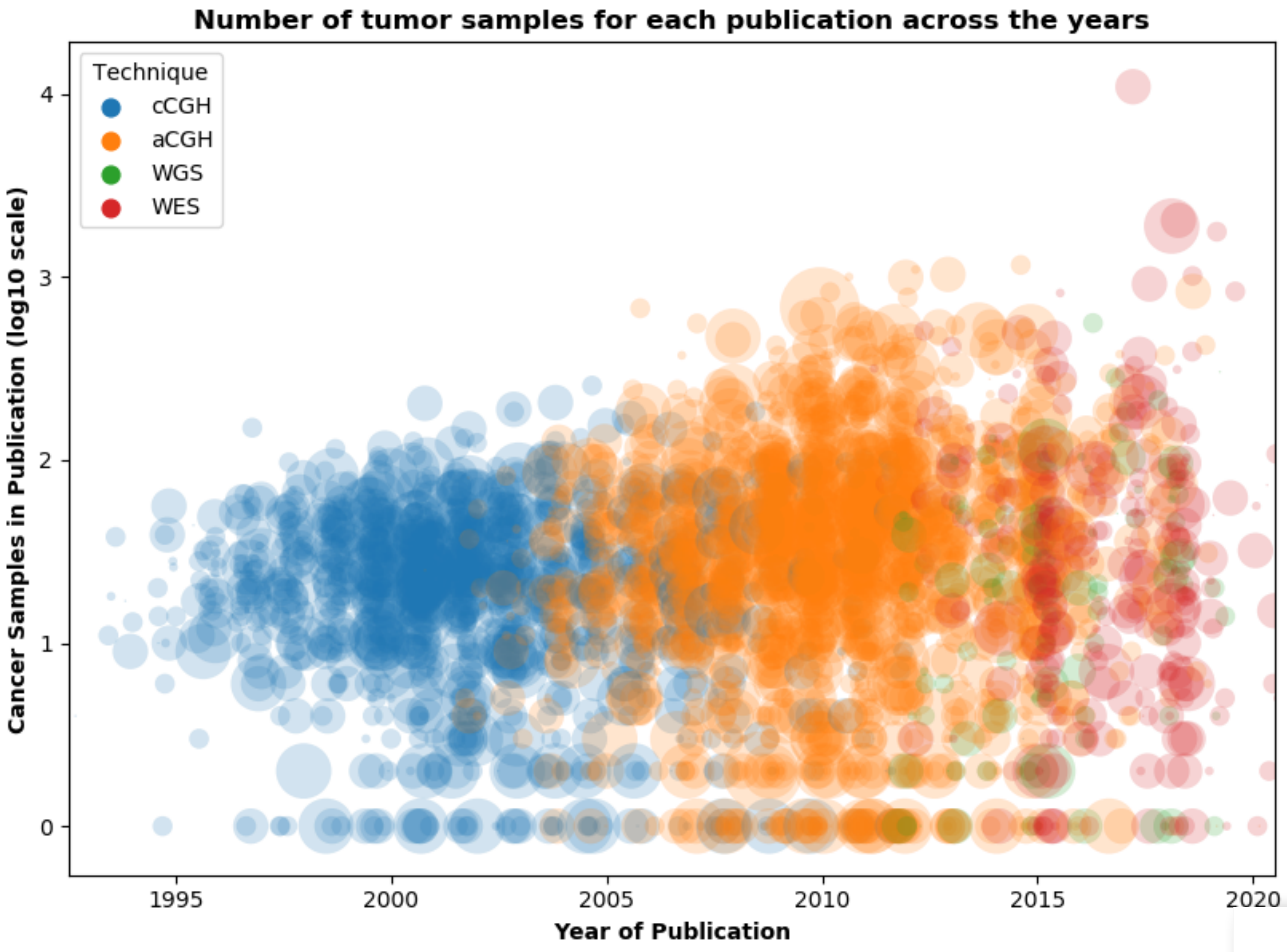
PMID:33767857

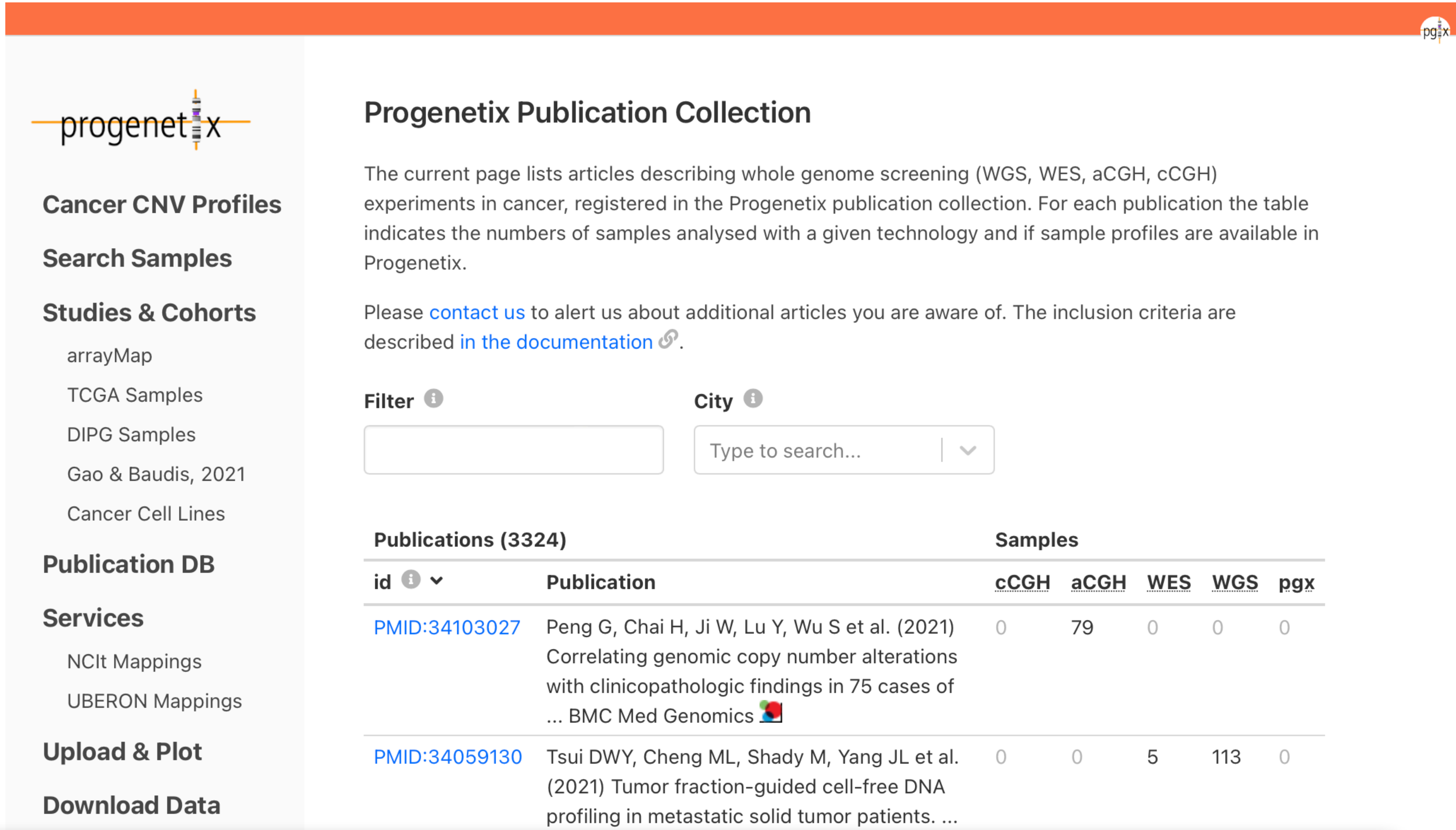
Methylation and molecular profiles of ependymoma: Influence of patient age and tumor anatomic location.

Cho HJ, Park HY, Kim K, Chae H, Paek SH, Kim SK, Park CK, Choi SH, Park SH.

Mol Clin Oncol PMID:33767857

`http://progenetix.org/services/publications/?filters=genomes:>0&ISO3166alpha2=kr&output=map`





Progenetix Publication Collection

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

Please contact us to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation.

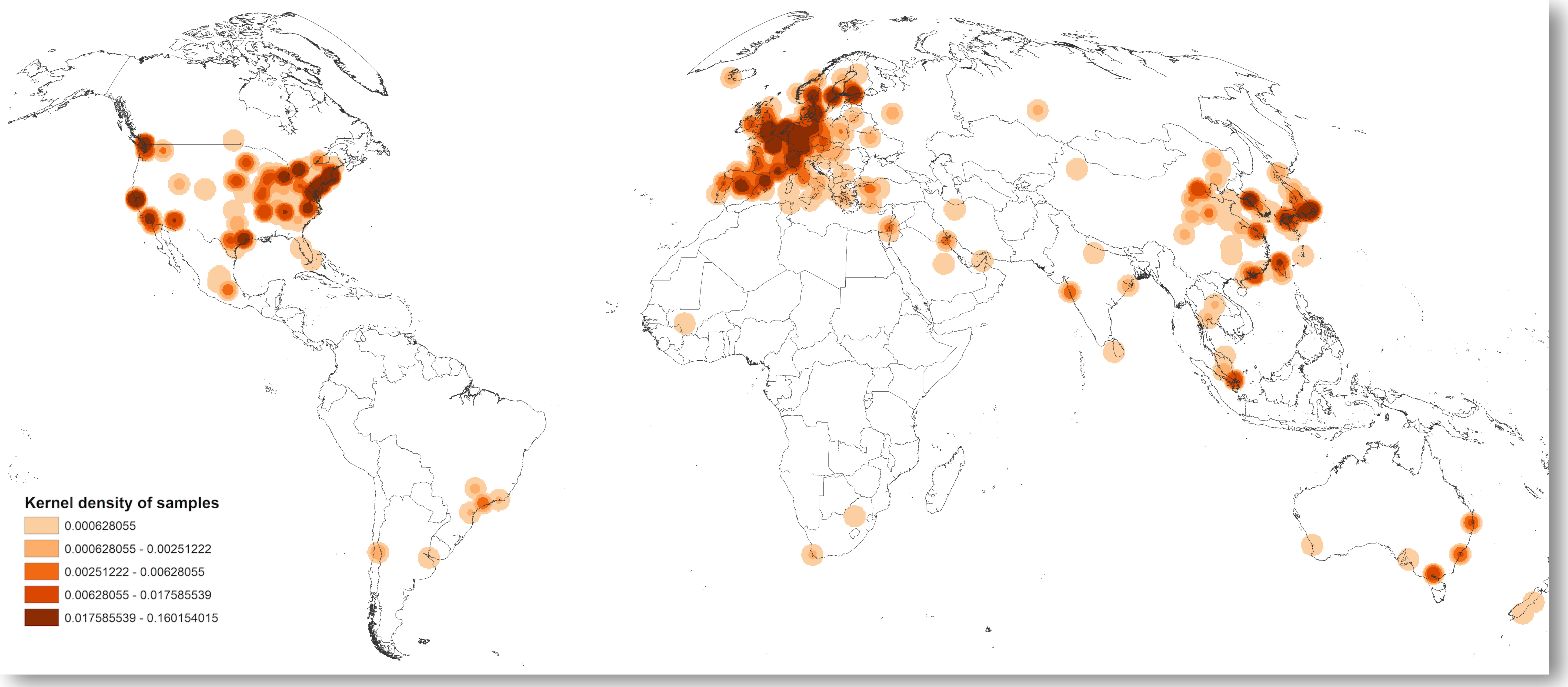
Filter

City: Type to search...

Publications (3324)		Samples				
id	Publication	cCGH	aCGH	WES	WGS	pgx
PMID:34103027	Peng G, Chai H, Ji W, Lu Y, Wu S et al. (2021) Correlating genomic copy number alterations with clinicopathologic findings in 75 cases of ... BMC Med Genomics	0	79	0	0	0
PMID:34059130	Tsui DWY, Cheng ML, Shady M, Yang JL et al. (2021) Tumor fraction-guided cell-free DNA profiling in metastatic solid tumor patients. ...	0	0	5	113	0

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.



progenetix

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

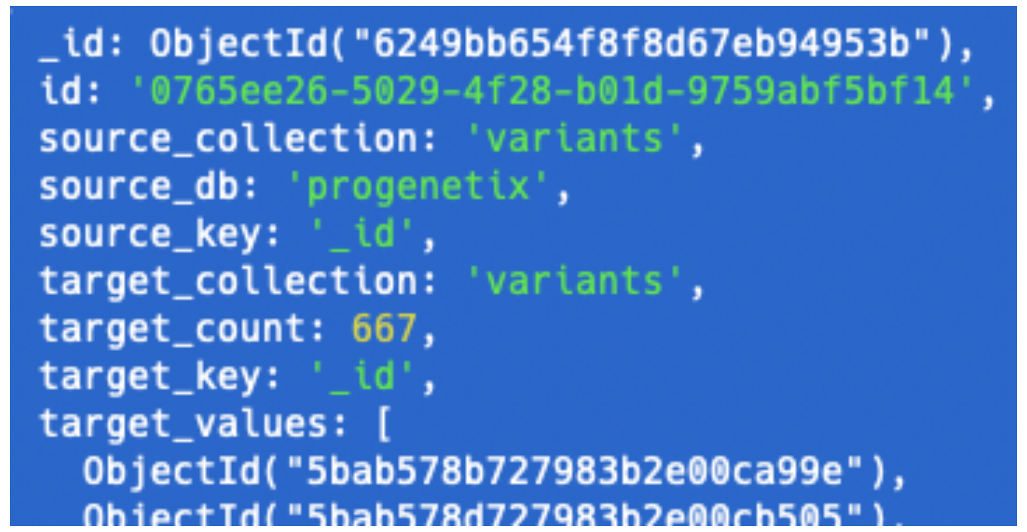
React

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



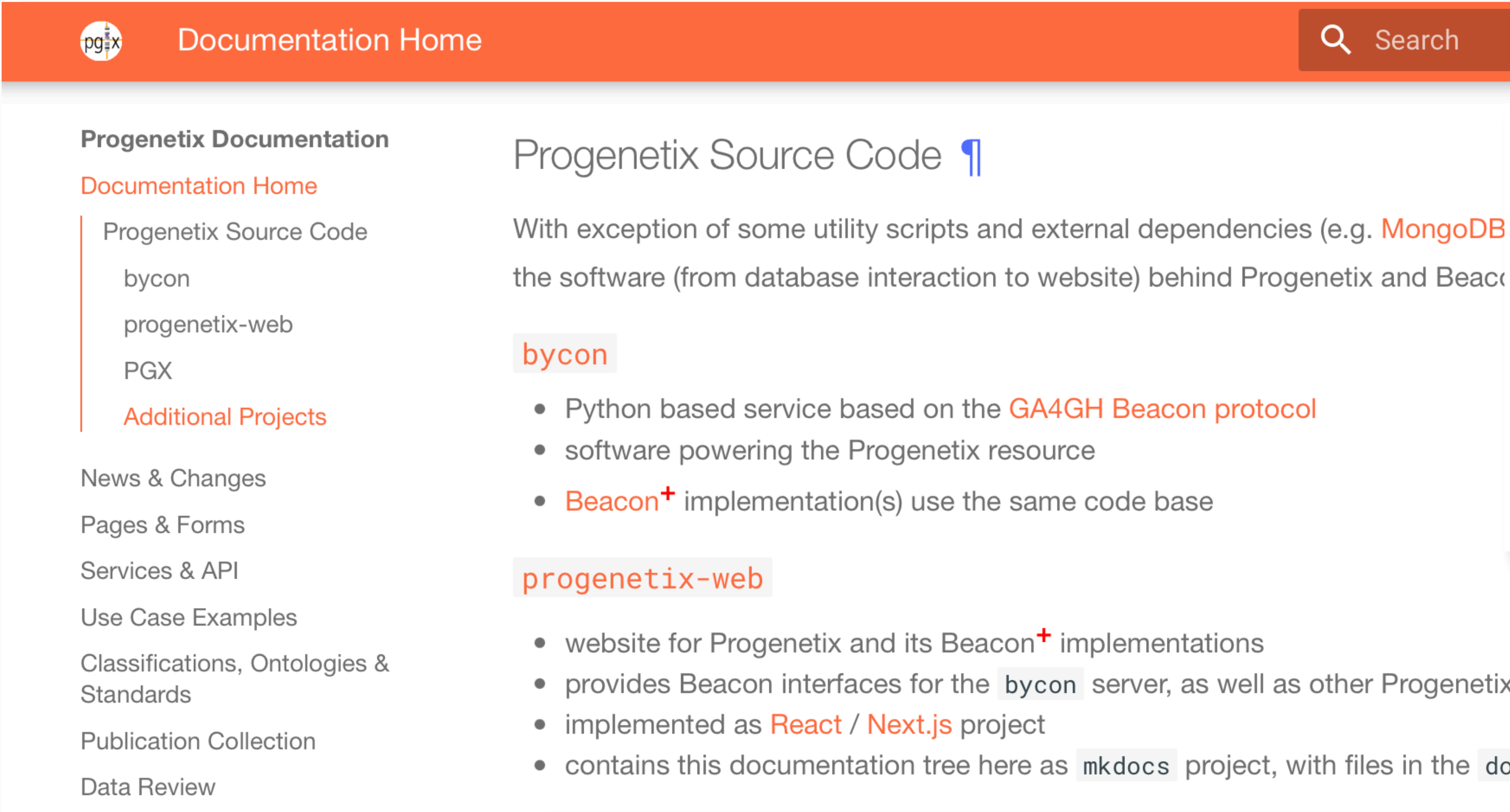
```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
  ObjectId("5bab578d727983b2e00cb505"),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

variants analyses biosamples individuals

Entity collections

collations geolocs genespans publications qBuffer

Utility collections



Progenetix Documentation

- Documentation Home
  - Progenetix Source Code
    - bycon
    - progenetix-web
    - PGX
    - Additional Projects
- News & Changes
- Pages & Forms
- Services & API
- Use Case Examples
- Classifications, Ontologies & Standards
- Publication Collection
- Data Review
- Beacon+ & bycon
- Technical Notes
- Progenetix Data
- Baudisgroup @ UZH

Progenetix Source Code ¶

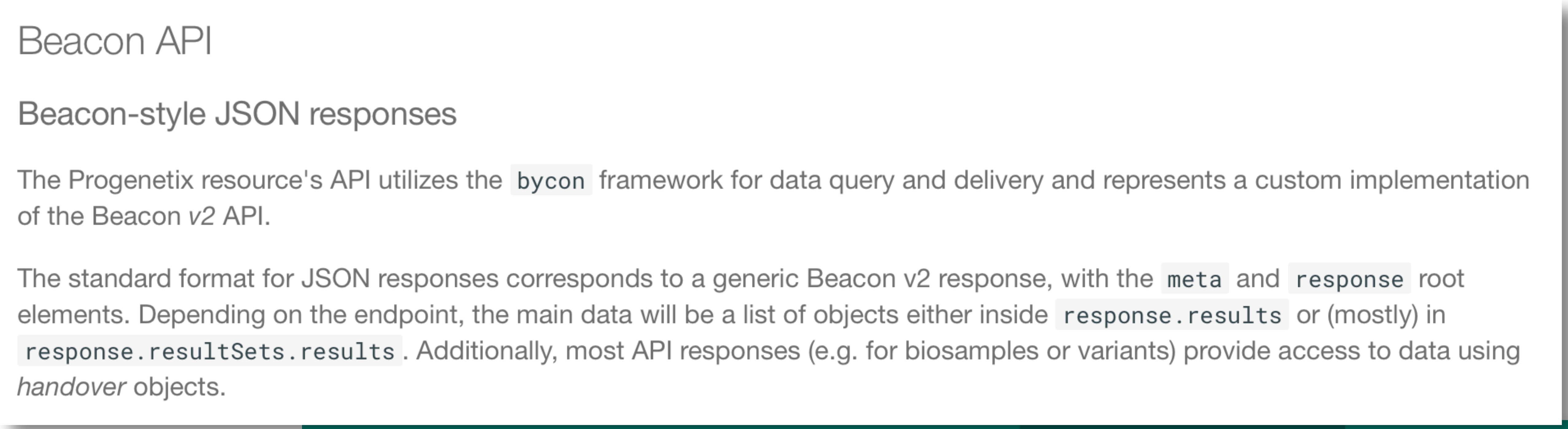
With exception of some utility scripts and external dependencies (e.g. MongoDB
the software (from database interaction to website) behind Progenetix and Beac

`bycon`

- Python based service based on the GA4GH Beacon protocol
- software powering the Progenetix resource
- Beacon+ implementation(s) use the same code base

`progenetix-web`

- website for Progenetix and its Beacon+ implementations
- provides Beacon interfaces for the `bycon` server, as well as other Progenetix sevices (e.g. the publicat
- implemented as React / Next.js project
- contains this documentation tree here as `mkdocs` project, with files in the `docs` directory

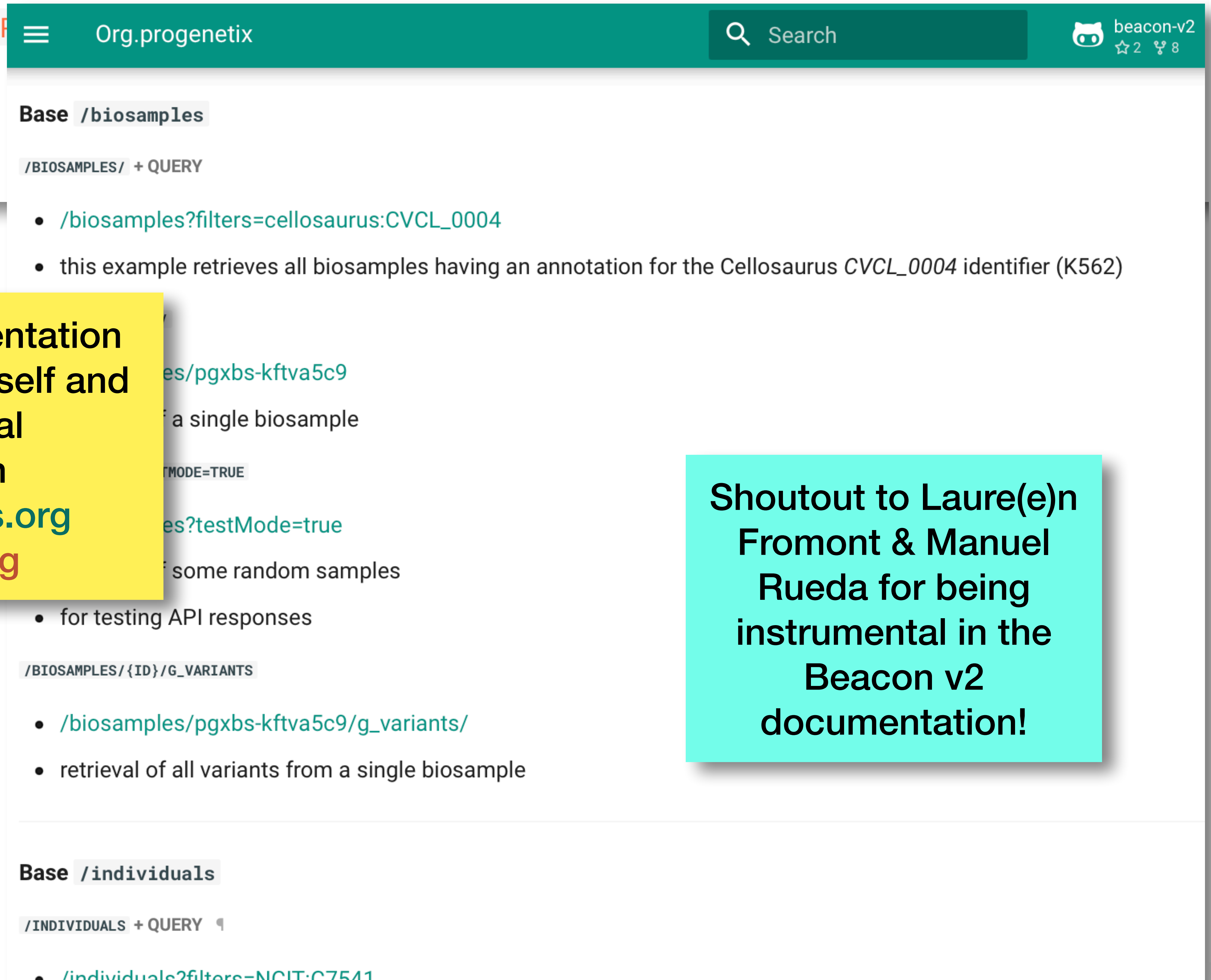


Beacon API

Beacon-style JSON responses

The Progenetix resource's API utilizes the `bycon` framework for data query and delivery and represents a custom implementation of the Beacon *v2* API.

The standard format for JSON responses corresponds to a generic Beacon v2 response, with the `meta` and `response` root elements. Depending on the endpoint, the main data will be a list of objects either inside `response.results` or (mostly) in `response.resultSets.results`. Additionally, most API responses (e.g. for biosamples or variants) provide access to data using *handover* objects.



Base `/biosamples`

`/BIOSAMPLES/` + QUERY

- /biosamples?filters=cellosaurus:CVCL_0004
- this example retrieves all biosamples having an annotation for the Cellosaurus *CVCL_0004* identifier (K562)

es/pgxbs-kftva5c9

a single biosample

MODE=TRUE

es?testMode=true

some random samples

- for testing API responses

`/BIOSAMPLES/{ID}/G_VARIANTS`

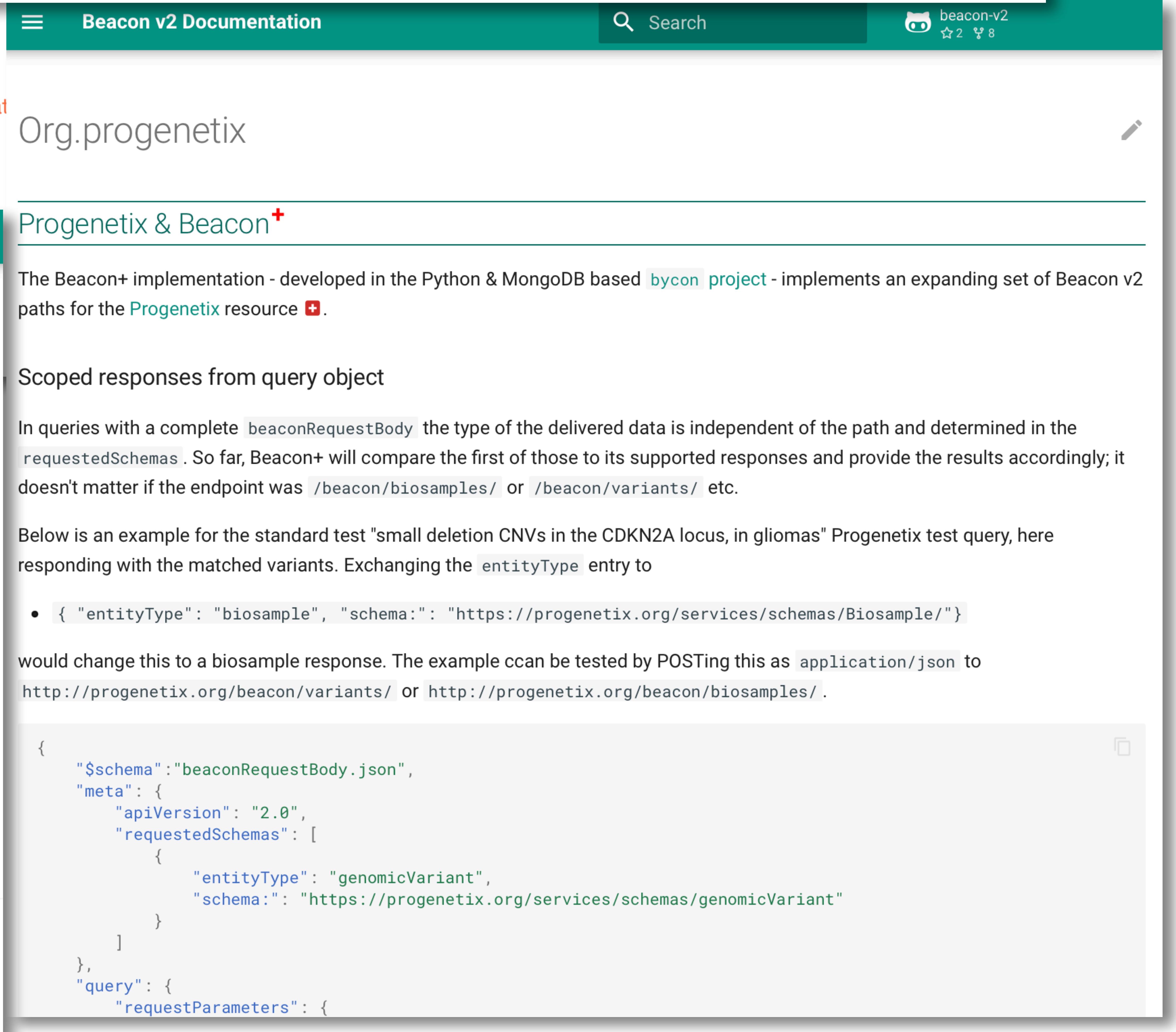
- /biosamples/pgxbs-kftva5c9/g_variants/
- retrieval of all variants from a single biosample

Base `/individuals`

`/INDIVIDUALS` + QUERY ¶

- /individuals?filters=NCIT:C7541

Rapidly evolving documentation of both the Beacon API itself and its use and technical implementation on
docs.genomebeacons.org
docs.progenetix.org



Org.progenetix

Progenetix & Beacon+

The Beacon+ implementation - developed in the Python & MongoDB based `bycon` project - implements an expanding set of Beacon v2 paths for the Progenetix resource.

Scoped responses from query object

In queries with a complete `beaconRequestBody` the type of the delivered data is independent of the path and determined in the `requestedSchemas`. So far, Beacon+ will compare the first of those to its supported responses and provide the results accordingly; it doesn't matter if the endpoint was `/beacon/biosamples/` or `/beacon/variants/` etc.

Below is an example for the standard test "small deletion CNVs in the CDKN2A locus, in gliomas" Progenetix test query, here responding with the matched variants. Exchanging the `entityType` entry to

- `{ "entityType": "biosample", "schema:": "https://progenetix.org/services/schemas/Biosample/"}`

would change this to a biosample response. The example ccan be tested by POSTing this as `application/json` to `http://progenetix.org/beacon/variants/` or `http://progenetix.org/beacon/biosamples/`.

```
{
    "$schema":"beaconRequestBody.json",
    "meta": {
        "apiVersion": "2.0",
        "requestedSchemas": [
            {
                "entityType": "genomicVariant",
                "schema:": "https://progenetix.org/services/schemas/genomicVariant"
            }
        ]
    },
    "query": {
        "requestParameters": {
```

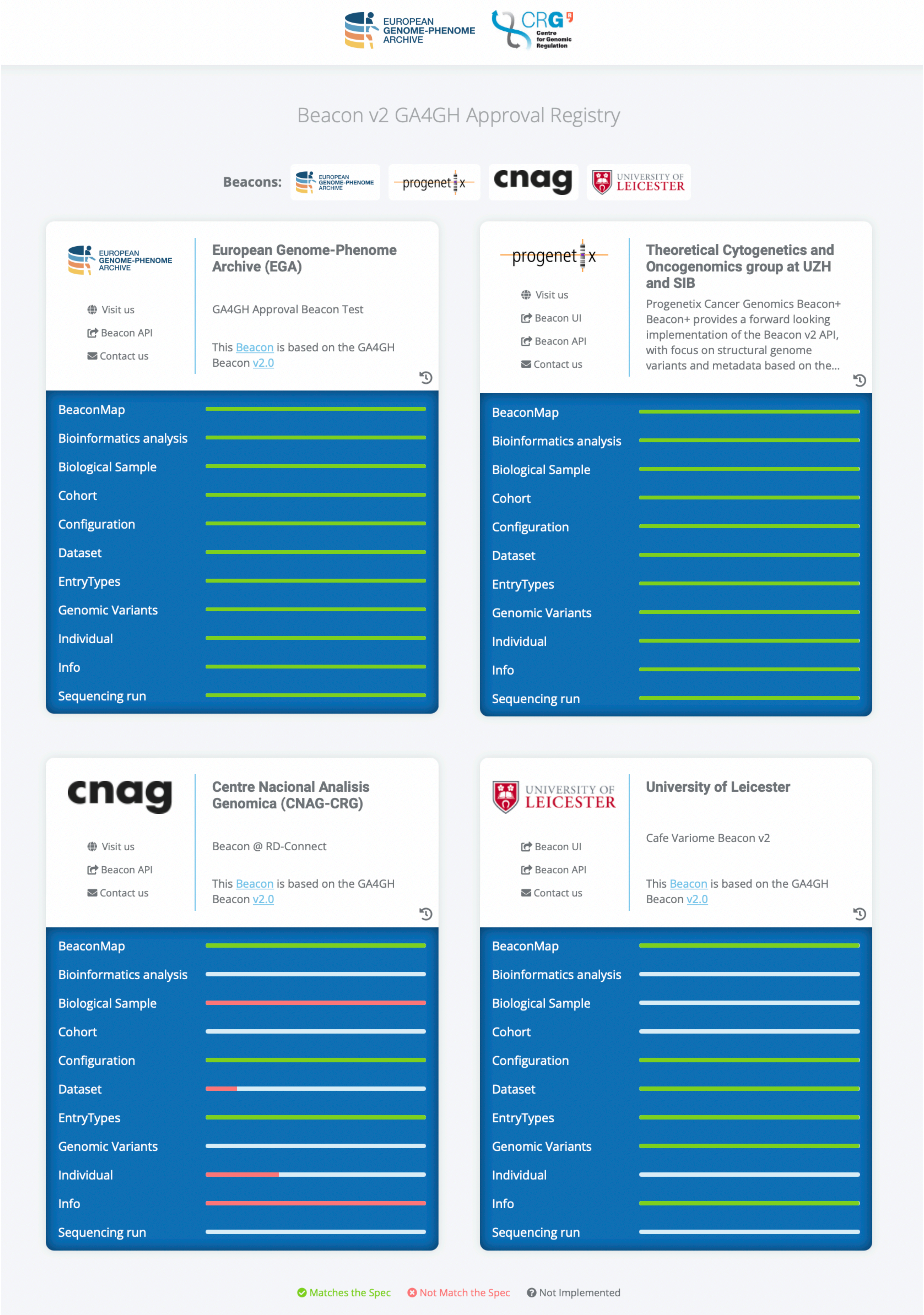
Shoutout to Laure(e)n Fromont & Manuel Rueda for being instrumental in the Beacon v2 documentation!

Onboarding

Demonstrating Compliance

- Progenetix Beacon+ has served as implementation driver since 2016
- Beacon v2 as service with protocol-driven registries for federation
- GA4GH approved Beacon v2 in April 2022

Global Alliance for Genomics & Health



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

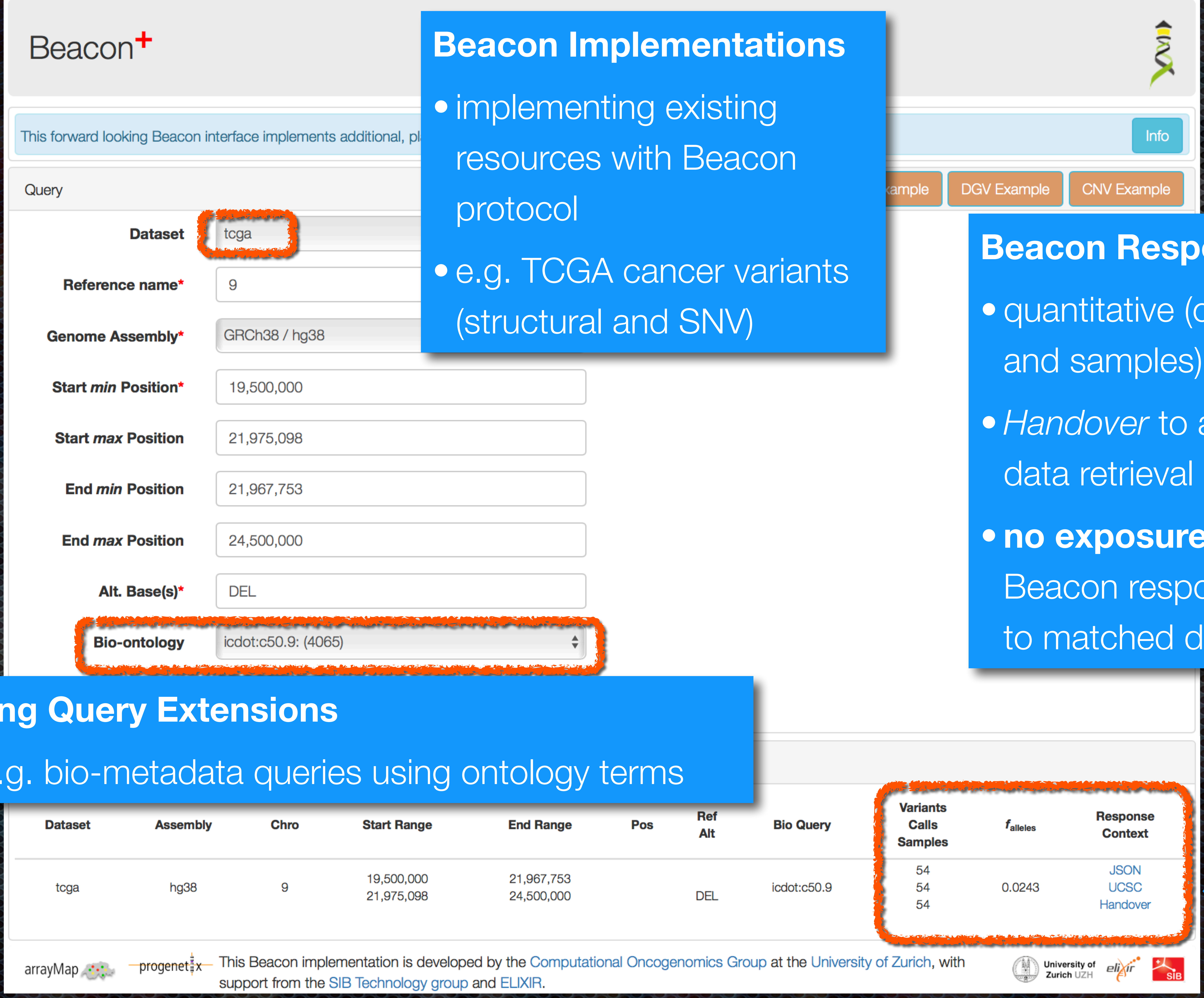
This Beacon is based on the GA4GH Beacon v2.0

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 v1 Development	Beacon v2 Development	Related ...
2014	GA4GH founding event; Jim Ostell proposes Beacon concept with "more features... version 2"		
2015	• beacon-network.org aggregator created by DNAstack • Beacon v0.3 release	• Beacon^{+} concept implemented on progenetix.org	• ELIXIR starts Beacon project support
2016	• work on queries for structural variants (brackets for fuzzy start and end parameters...)	• 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	• Beacon hackathon Stockholm; settling on "filters" • Barcelona goes Zurich developers meeting	• new Beacon website (March)
2019	• ELIXIR Beacon Network	• 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 • starting of GA4GH review process	
2022		• further changes esp. in default model, aligning with Phenopackets and VRS • unified beacon-v2 code & docs repository • **Beacon v2 approved at Apr GA4GH Connect**	• *docs.genomebeacons.org*

beacon.progenetix.org



Beacon+

This forward looking Beacon interface implements additional, pl

Info

Query

xample DGV Example CNV Example

Dataset	tcga
Reference name*	9
Genome Assembly*	GRCh38 / hg38
Start *min* Position*	19,500,000
Start *max* Position	21,975,098
End *min* Position	21,967,753
End *max* Position	24,500,000
Alt. Base(s)*	DEL
Bio-ontology	icdot:c50.9: (4065)

Dataset	Assembly	Chro	Start Range	End Range	Pos	Ref Alt	Bio Query	Variants Calls Samples	$f_{alleles}$	Response Context
tcga	hg38	9	19,500,000 21,975,098	21,967,753 24,500,000		DEL	icdot:c50.9	54 54 54	0.0243	JSON UCSC Handover

arrayMap progenetix This Beacon implementation is developed by the Computational Oncogenomics Group at the University of Zurich, with support from the SIB Technology group and ELIXIR.

Beacon Implementations

- implementing existing resources with Beacon protocol
- e.g. TCGA cancer variants (structural and SNV)

Beacon Response

- quantitative (counts for variants, callsets and samples)
- *Handover* to authentication system for data retrieval
- **no exposure** of data beyond standard Beacon response and additional pointer to matched data

Prototyping Query Extensions

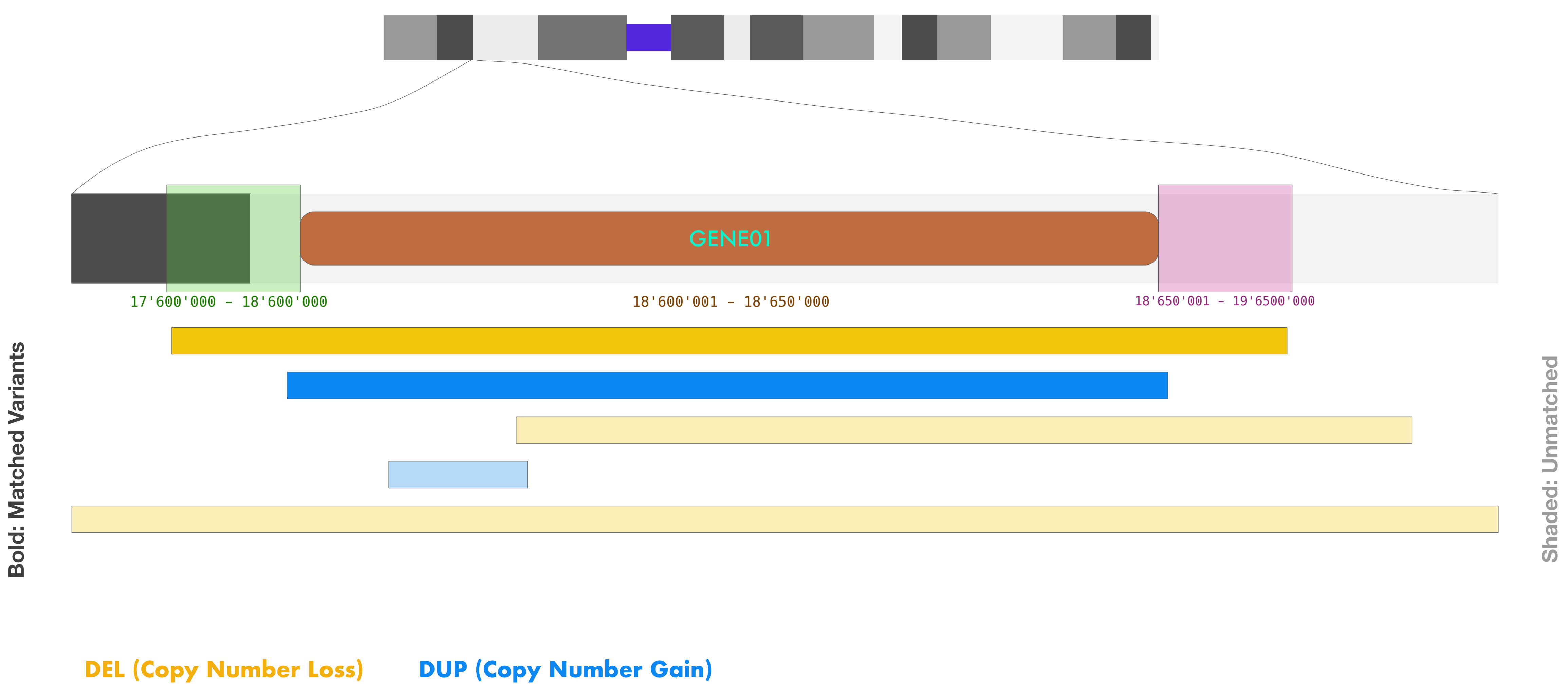
- testing e.g. bio-metadata queries using ontology terms

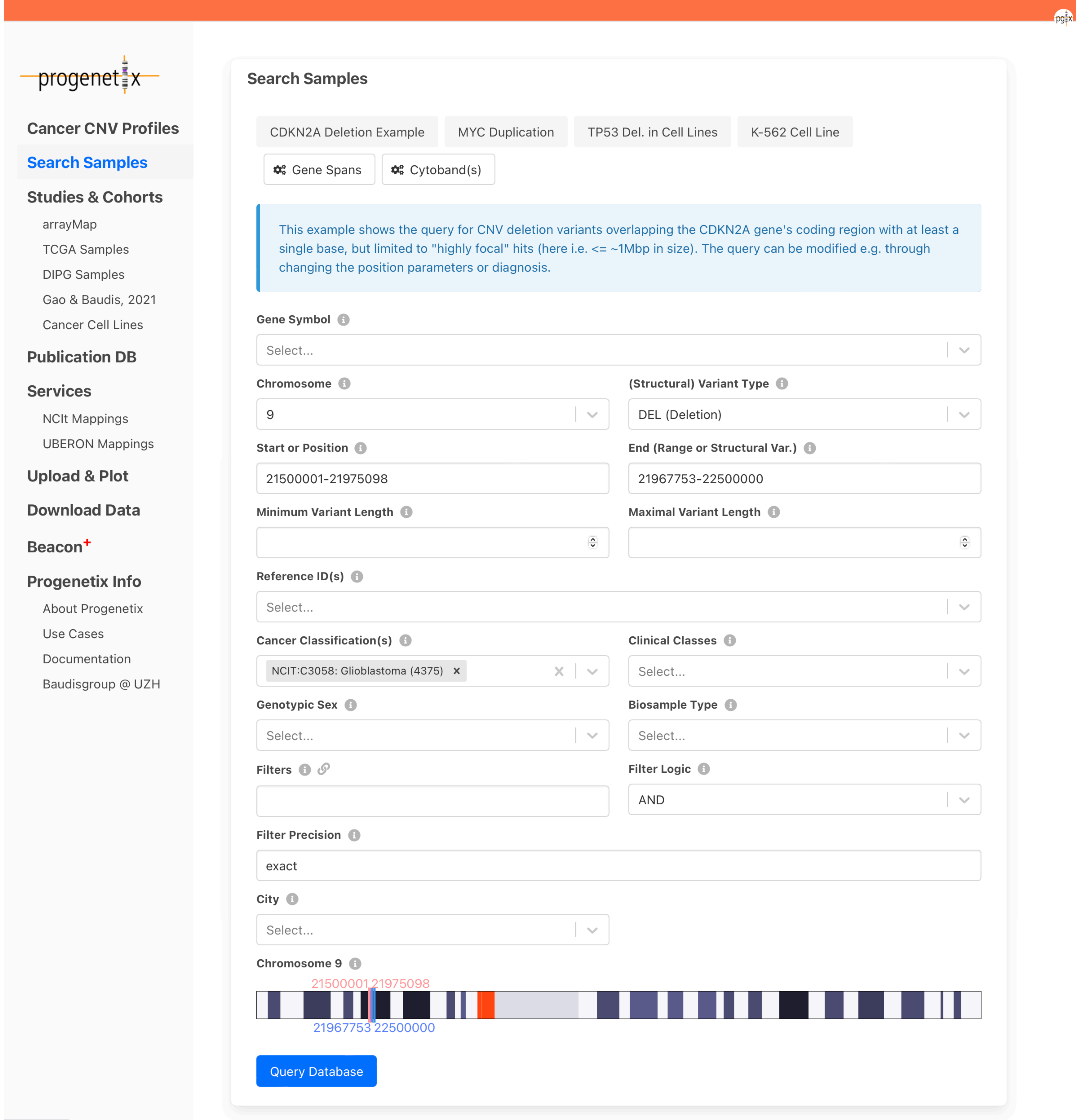
Progenetix in 2022

Variant and Metadata for Sample Discovery

- positional queries for genomic variants using the **GA4GH Beacon protocol**
- metadata queries (diagnoses, identifiers, clinical classes ...) using **Beacon "filters"**

Genome Bracket Query
(full match)





Search Samples

CDKN2A Deletion Example | MYC Duplication | TP53 Del. in Cell Lines | K-562 Cell Line

Gene Spans | Cytoband(s)

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 "highly focal" hits (here i.e. <= ~1Mbp in size). The query can be modified e.g. through changing the position parameters or diagnosis.

Gene Symbol: Select...

Chromosome: 9

(Structural) Variant Type: DEL (Deletion)

Start or Position: 21500001-21975098

End (Range or Structural Var.): 21967753-22500000

Minimum Variant Length:

Maximal Variant Length:

Reference ID(s): Select...

Cancer Classification(s): NCIT:C3058: Glioblastoma (4375)

Clinical Classes: Select...

Genotypic Sex: Select...

Biosample Type: Select...

Filters:

Filter Logic: AND

Filter Precision: exact

City: Select...

Chromosome 9

21500001 21975098

21967753 22500000

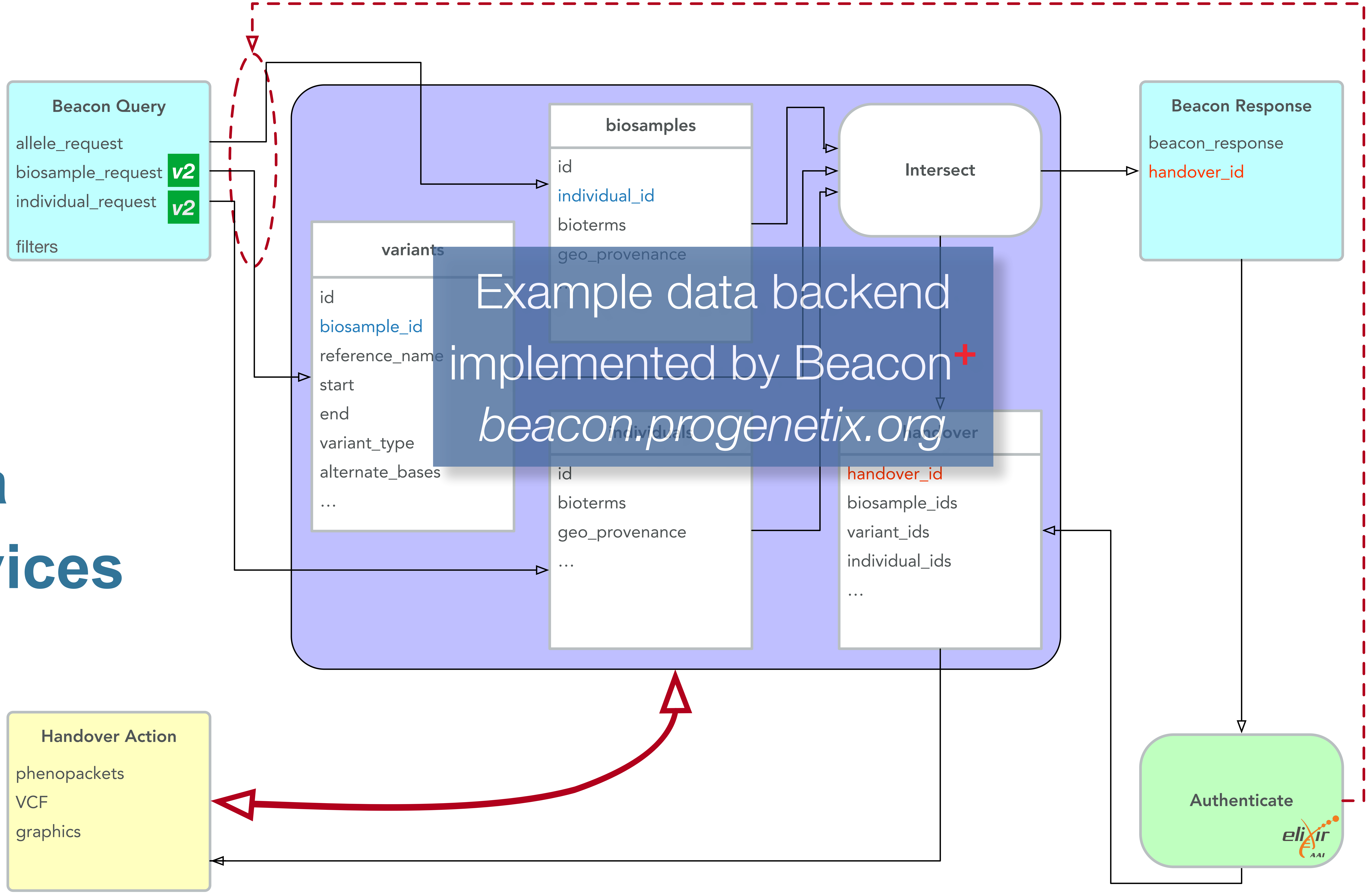
Query Database

Filter types

- Ontology Filters
  - **Hierarchical ontology query** is assumed by default, where the requested ontology filter(s) and all descendent terms are queried
  - **Exact term match** requests, where descendent terms are excluded, are supported
  - **Semantic similarity queries** for entities that are associated with terms that are similar to the requested filters are supported by Beacon 2.0
    - Agnostic to the semantic similarity model used by a Beacon
    - Relative similarity thresholds of *high*, *medium* and *low*
- Numeric Filters … (using equality and relational operators)
- Alphanumeric Filters … (e.g. string matches)

Beacon & Handover

Beacons v1.1 supports **data delivery services**



Example data backend implemented by Beacon+
beacon.progenetix.org

Beacon I/O

Handover

Authentication

Beacon v2 Conformity and Extensions in Progenetix

Putting the + into Beacon ...

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

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

pgxRpi

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

Beacon Path: Retrieve variants by biosample id(s)

```
https://progenetix.org/beacon/g_variants/
?biosampleIds=pgxbs-kftvh94d,pgxbs-kftvh94g,pgxbs-kftvh972
&output=pgxseg
```

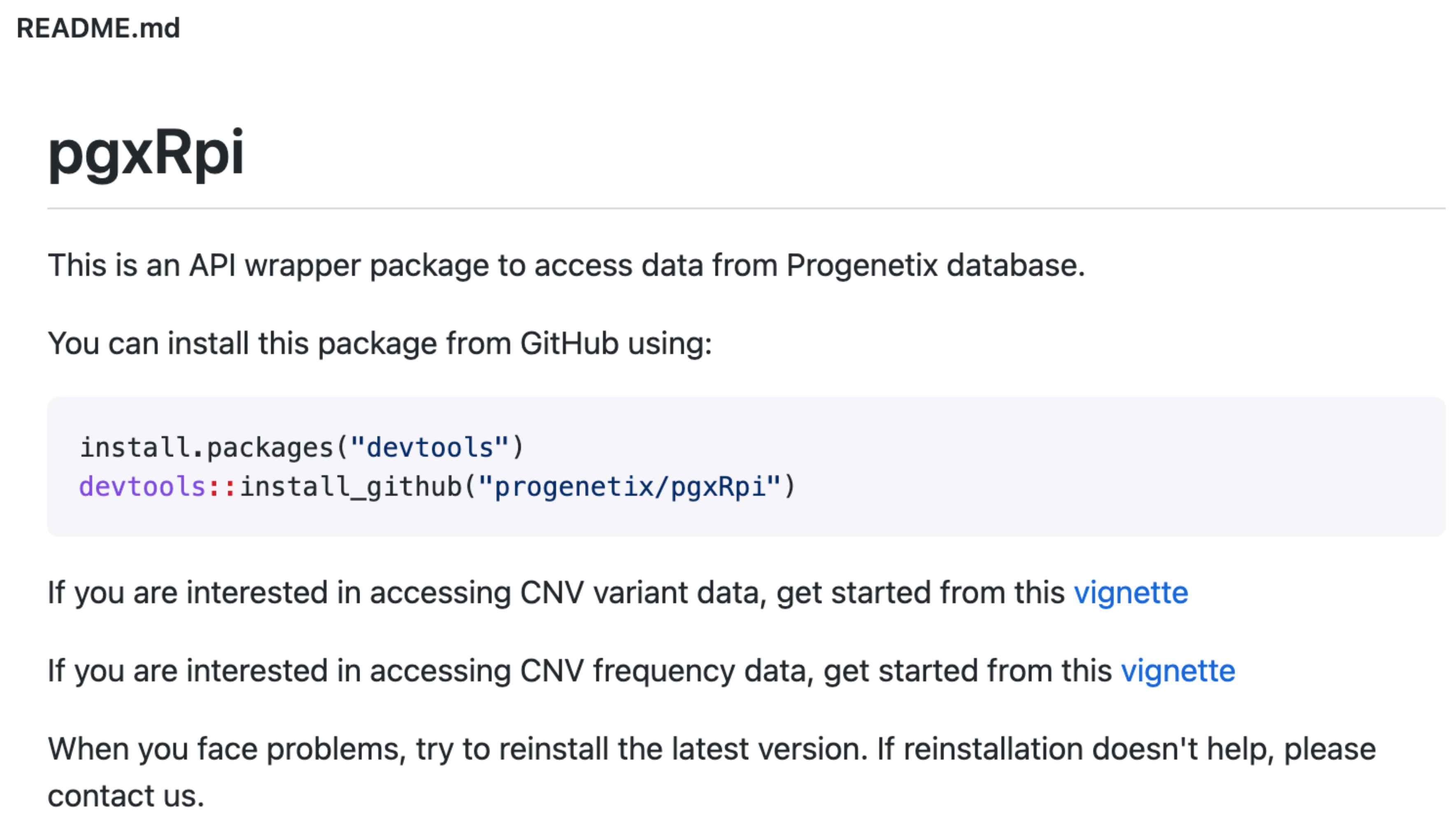
Beacon Path: Get biosamples by filter(s)

```
http://progenetix.org/beacon/biosamples/
?filters=NCIT:C3697&output=datatable
```

Service Path: Retrieve CNV frequencies by filter(s)

```
http://www.progenetix.org/services/intervalFrequencies/
?id=NCIT:C4323&output=pgxseg
```

Author: Hangjia Zhao | @hangjiaz



README.md

pgxRpi

This is an API wrapper package to access data from Progenetix database.

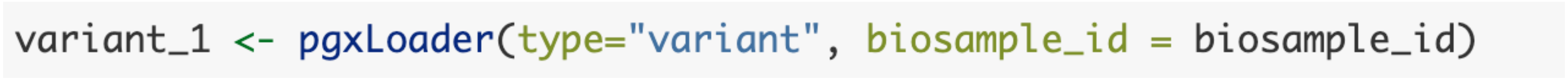
You can install this package from GitHub using:

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

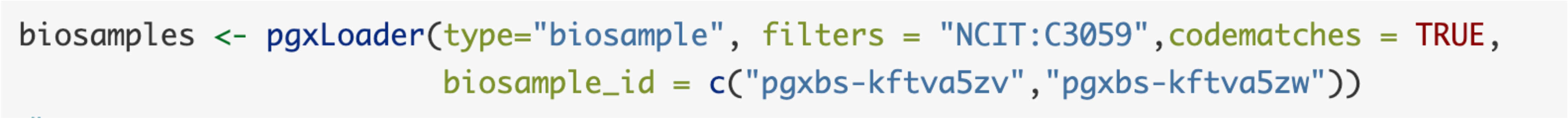
If you are interested in accessing CNV variant data, get started from this vignette

If you are interested in accessing CNV frequency data, get started from this vignette

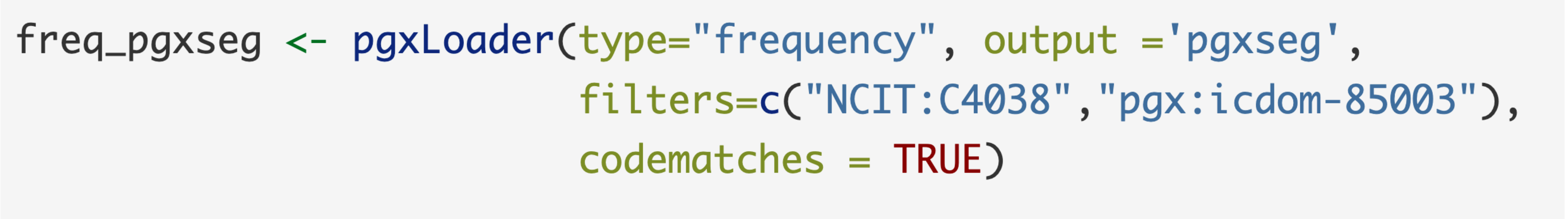
When you face problems, try to reinstall the latest version. If reinstallation doesn't help, please contact us.



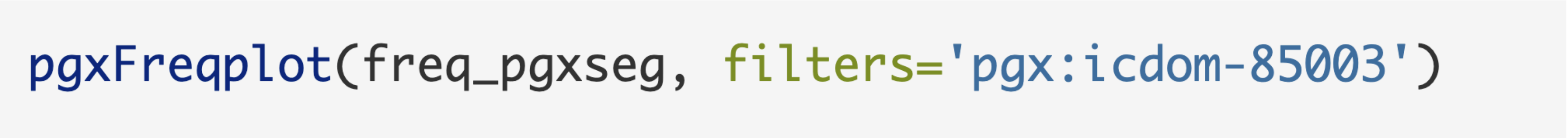
```
variant_1 <- pgxLoader(type="variant", biosample_id = biosample_id)
```



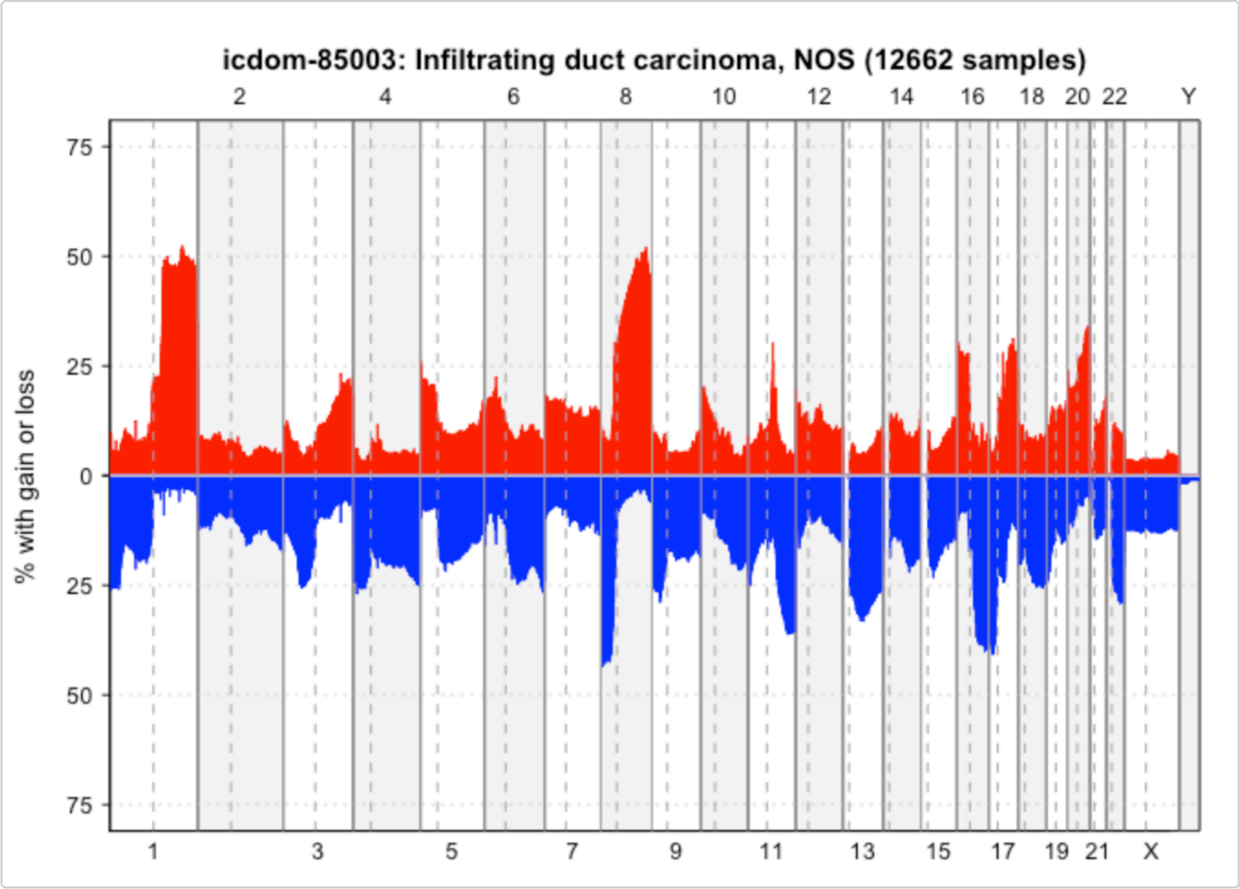
```
biosamples <- pgxLoader(type="biosample", filters = "NCIT:C3059",codematches = TRUE,
                        biosample_id = c("pgxbs-kftva5zv","pgxbs-kftva5zw"))
```



```
freq_pgxseg <- pgxLoader(type="frequency", output ='pgxseg',
                         filters=c("NCIT:C4038","pgx:icdom-85003"),
                         codematches = TRUE)
```



```
pgxFreqplot(freq_pgxseg, filters='pgx:icdom-85003')
```

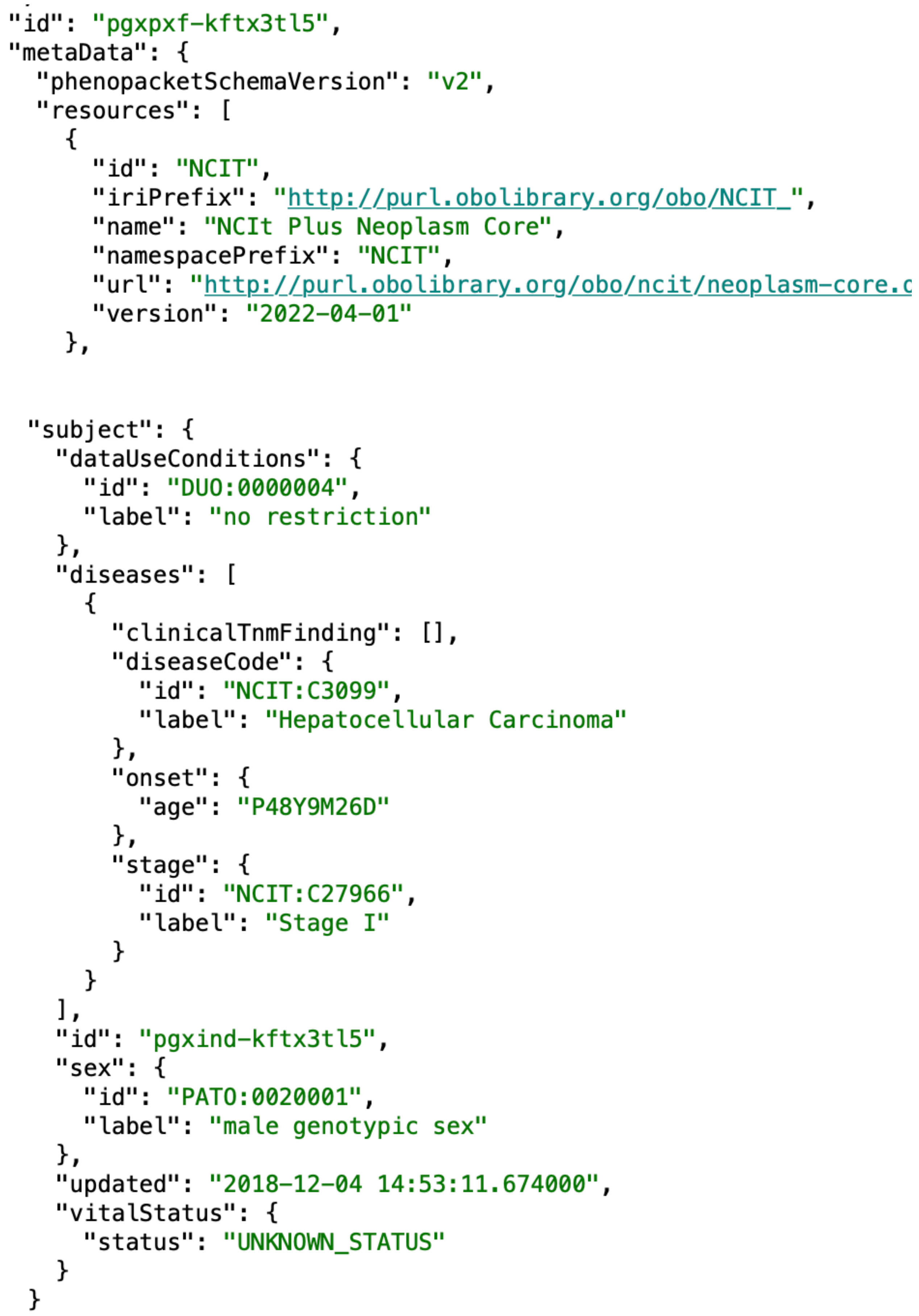


Beacon+: Phenopackets

Testing alternative response schemas...

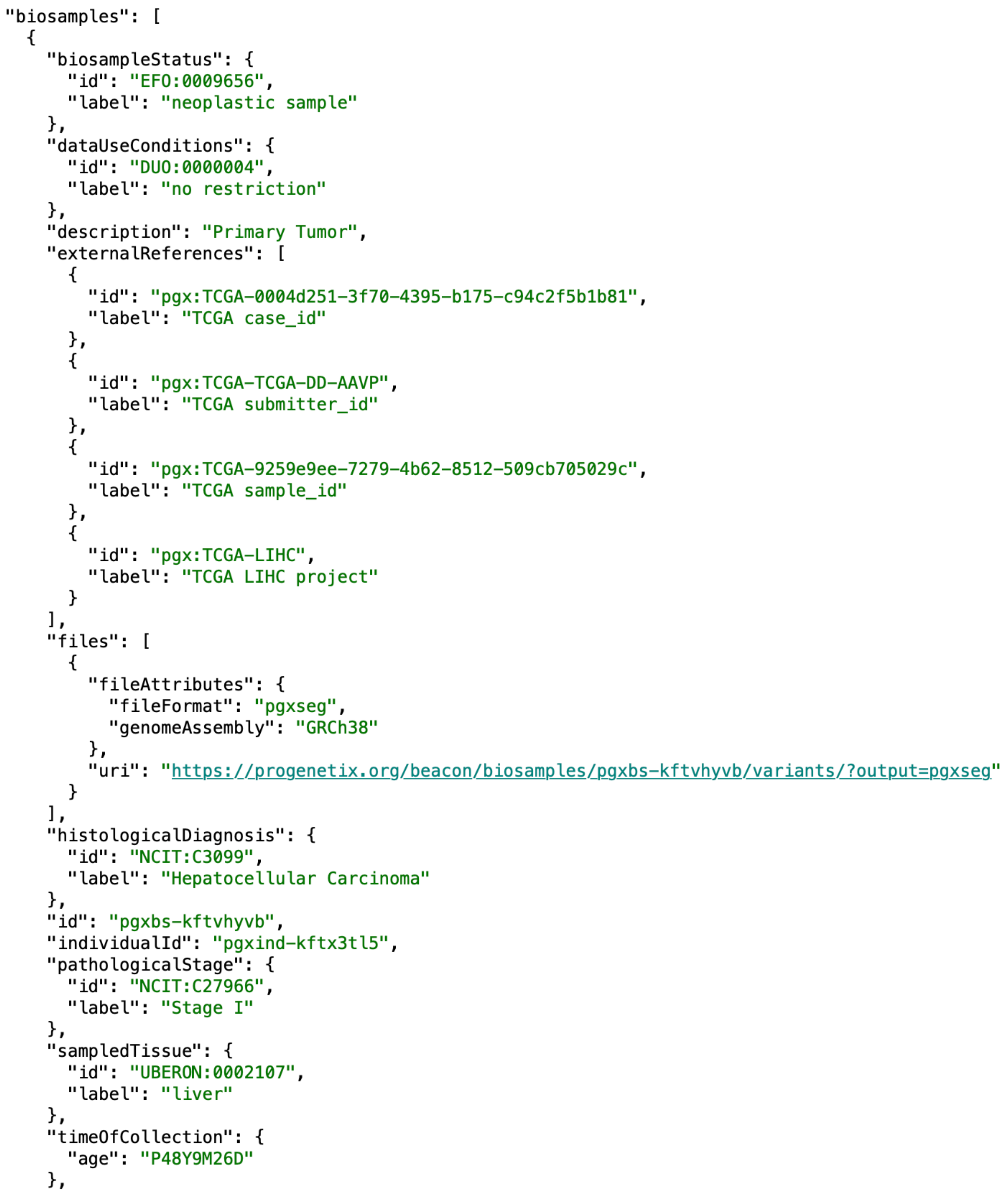
`http://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/phenopackets`

- the v2 default schemas are mostly aligned w/ Phenopackets v2
- creating phenopackets can be done mostly by re-wrapping of Beacon entities (individual, biosample)
- variants can be included through file resource URLs; in Beacon+ this is done through *ad hoc* handover URIs



```
"id": "pgxpxf-kftx3tl5",
"metaData": {
  "phenopacketSchemaVersion": "v2",
  "resources": [
    {
      "id": "NCIT",
      "iriPrefix": "http://purl.obolibrary.org/obo/NCIT_",
      "name": "NCIt Plus Neoplasm Core",
      "namespacePrefix": "NCIT",
      "url": "http://purl.obolibrary.org/obo/ncit/neoplasm-core.c
      "version": "2022-04-01"
    },


 "subject": {
   "dataUseConditions": {
     "id": "DUO:0000004",
     "label": "no restriction"
   },
   "diseases": [
     {
       "clinicalTnmFinding": [],
       "diseaseCode": {
         "id": "NCIT:C3099",
         "label": "Hepatocellular Carcinoma"
       },
       "onset": {
         "age": "P48Y9M26D"
       },
       "stage": {
         "id": "NCIT:C27966",
         "label": "Stage I"
       }
     }
   ],
   "id": "pgxind-kftx3tl5",
   "sex": {
     "id": "PATO:0020001",
     "label": "male genotypic sex"
   },
   "updated": "2018-12-04 14:53:11.674000",
   "vitalStatus": {
     "status": "UNKNOWN_STATUS"
   }
 }
```



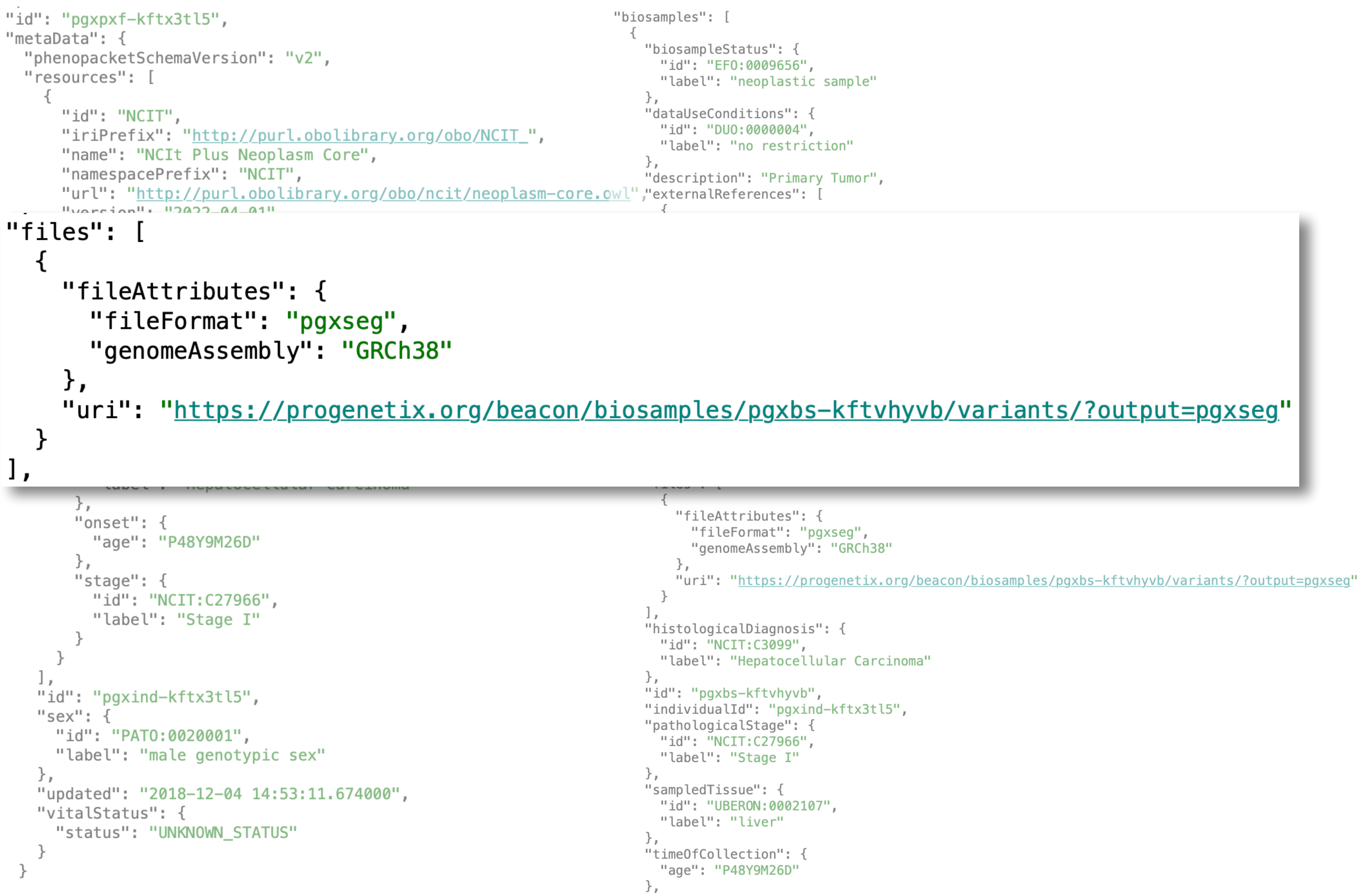
```
"biosamples": [
  {
    "biosampleStatus": {
      "id": "EFO:0009656",
      "label": "neoplastic sample"
    },
    "dataUseConditions": {
      "id": "DUO:0000004",
      "label": "no restriction"
    },
    "description": "Primary Tumor",
    "externalReferences": [
      {
        "id": "pgx:TCGA-0004d251-3f70-4395-b175-c94c2f5b1b81",
        "label": "TCGA case_id"
      },
      {
        "id": "pgx:TCGA-TCGA-DD-AAVP",
        "label": "TCGA submitter_id"
      },
      {
        "id": "pgx:TCGA-9259e9ee-7279-4b62-8512-509cb705029c",
        "label": "TCGA sample_id"
      },
      {
        "id": "pgx:TCGA-LIHC",
        "label": "TCGA LIHC project"
      }
    ],
    "files": [
      {
        "fileAttributes": {
          "fileFormat": "pgxseg",
          "genomeAssembly": "GRCh38"
        },
        "uri": "https://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/variants/?output=pgxseg"
      }
    ],
    "histologicalDiagnosis": {
      "id": "NCIT:C3099",
      "label": "Hepatocellular Carcinoma"
    },
    "id": "pgxbs-kftvhyvb",
    "individualId": "pgxind-kftx3tl5",
    "pathologicalStage": {
      "id": "NCIT:C27966",
      "label": "Stage I"
    },
    "sampledTissue": {
      "id": "UBERON:0002107",
      "label": "liver"
    },
    "timeOfCollection": {
      "age": "P48Y9M26D"
    },
```

Beacon+: Phenopackets

Testing alternative response schemas...

`http://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/phenopackets`

- the v2 default schemas are mostly aligned w/ Phenopackets v2
- creating phenopackets can be done mostly by re-wrapping of Beacon entities (individual, biosample)
- variants can be included through file resource URLs; in Beacon+ this is done through *ad hoc* handover URIs



```
"id": "pgxpxf-kftx3tl5",
"metaData": {
  "phenopacketSchemaVersion": "v2",
  "resources": [
    {
      "id": "NCIT",
      "iriPrefix": "http://purl.obolibrary.org/obo/NCIT_",
      "name": "NCIt Plus Neoplasm Core",
      "namespacePrefix": "NCIT",
      "url": "http://purl.obolibrary.org/obo/ncit/neoplasm-core.owl",
      "version": "2022-04-01"
```

```
"biosamples": [
  {
    "biosampleStatus": {
      "id": "EFO:0009656",
      "label": "neoplastic sample"
    },
    "dataUseConditions": {
      "id": "DUO:0000004",
      "label": "no restriction"
    },
    "description": "Primary Tumor",
    "externalReferences": [
      {
```

```
"files": [
  {
    "fileAttributes": {
      "fileFormat": "pgxseg",
      "genomeAssembly": "GRCh38"
    },
    "uri": "https://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/variants/?output=pgxseg"
  }
],
```

```
        "label": "Hepatocellular Carcinoma"
      },
      "onset": {
        "age": "P48Y9M26D"
      },
      "stage": {
        "id": "NCIT:C27966",
        "label": "Stage I"
      }
    }
  ],
  "id": "pgxind-kftx3tl5",
  "sex": {
    "id": "PATO:0020001",
    "label": "male genotypic sex"
  },
  "updated": "2018-12-04 14:53:11.674000",
  "vitalStatus": {
    "status": "UNKNOWN_STATUS"
  }
}
```

```
    {
      "fileAttributes": {
        "fileFormat": "pgxseg",
        "genomeAssembly": "GRCh38"
      },
      "uri": "https://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/variants/?output=pgxseg"
    }
  ],
  "histologicalDiagnosis": {
    "id": "NCIT:C3099",
    "label": "Hepatocellular Carcinoma"
  },
  "id": "pgxbs-kftvhyvb",
  "individualId": "pgxind-kftx3tl5",
  "pathologicalStage": {
    "id": "NCIT:C27966",
    "label": "Stage I"
  },
  "sampledTissue": {
    "id": "UBERON:0002107",
    "label": "liver"
  },
  "timeOfCollection": {
    "age": "P48Y9M26D"
  },
```

Progenetix & Beacon+

A cancer genomics reference resource powered by GA4GH standards

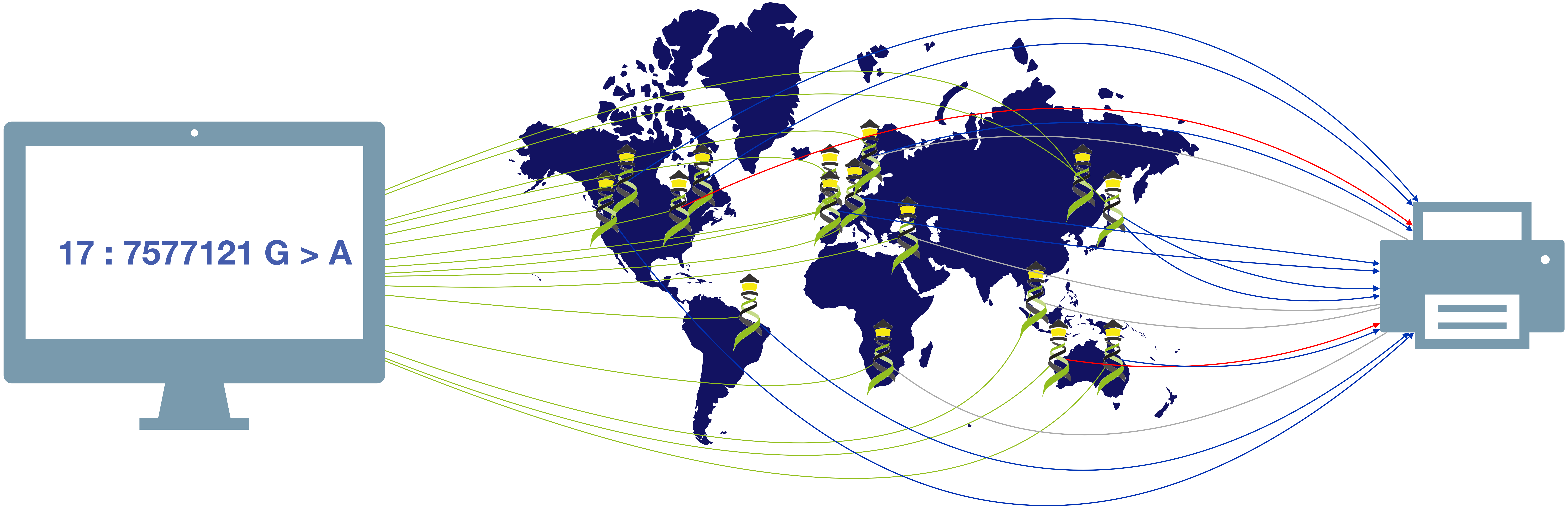
- Copy number variations constitute a complex, exciting and still poorly understood research topic in cancer and rare disease genomics
- Progenetix is the largest public resource for CNV in cancers (and increasingly reference genomes)
- The complexity of inherited and somatic genomic variations requires data access beyond individual resources => **Federated Data Access**
- The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization with focus on genomic data sharing
- Beacon v2 is the main GA4GH data discovery and sharing protocol, developed with support from the European bioinformatics organization ELIXIR
- Progenetix serves as a testbed for the early implementation of GA4GH standards such as Beacon extensions, Phenopackets and VRS

Beacon's v2.n Future?

Some proposals for a stepwise Beacon protocol extension

- Query language expansion, e.g. Boolean options for chaining filters

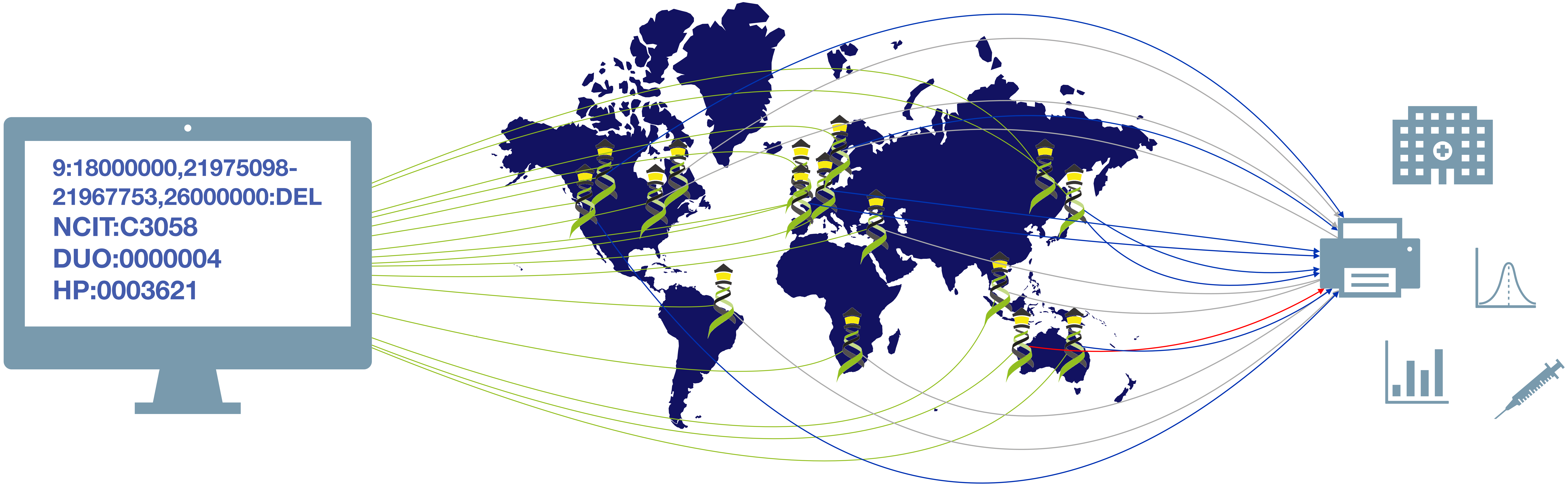
  ➡use of heterogeneous/alternative annotations within and across resources
- **Phenopackets** support as a (the?) default format for biodata export
- **Phenopackets** as **request** documents
- Focus on service & **resource discovery**
- **ELIXIR Beacon Network**, including translations for federated queries to Beacon and Beacon-like resources



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**.

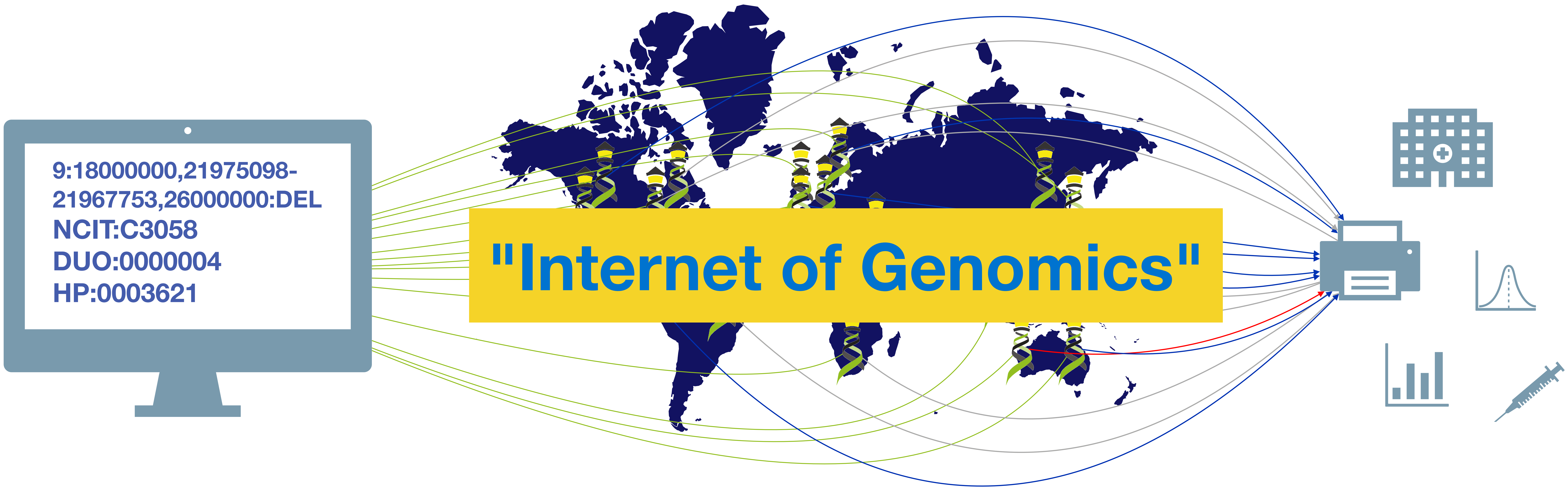


```
9:18000000,21975098-
21967753,26000000:DEL
NCIT:C3058
DUO:0000004
HP:0003621
```

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**".



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

"Internet of Genomics"

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**".

The Beacon team

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
Bo Gao

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

GEM Japan

Toshiaki
Katayama

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 teams —-

GA4GH Phenopackets
Peter Robinson
Jules Jacobsen

GA4GH VRS
Alex Wagner
Reece Hart

Beacon PRC
Alex Wagner
Jonathan Dursi
Mamana Mbiyavanga

Alice Mann
Neerjah Skantharajah

Get Involved! Visit GA4GH.ORG

Join a Work Stream!

Contact secretariat@ga4gh.org

Become an Organizational Member
ga4gh.org/members

Subscribe to GA4GH Updates
ga4gh.org/subscribe

progenetix

Progenetix Needs & Offers

What we have ...

- ✓ collection of >4000 articles assessed for scope
    - training set for NLP & search engine generation
- ✓ cancer specific ontologies with cross-mappings (ICD-O vs. NCIt) based on >100k samples
    - existing service API
- ✓ metadata ontology mappings for some 10k samples, with varying coverage for grade / stage / survival / ...
- ✓ CNV profiles for >110k samples, >700 entities with disease codes and metadata
- ✓ cell line CNV profiles together with mapped variants with clinical evidences

What we're working on...

- ➡ (semi-)automated detection of additional articles for scope (genome screening technologies, cancer samples, geographies)
- ➡ generation of a complete ICD-O terminology tree with NCIT (?) correspondence
    - improved service API & publication
- ➡ improved annotations using smarter source (article, annotation files) pre-/processing
- ➡ correlation between individual profiles, profile heterogeneity and external parameters
- ➡ relation between cell lines and native tumor types, with consideration of non-CNV parameters and publication use

Universität Zürich UZH

info.baudisgroup.org