Beacon v2 and Beyond
The Standard for Data *Discovery* and Data *Sharing* in Biomedical Genomics

Michael Baudis
Professor of Bioinformatics
University of Zürich
Swiss Institute of Bioinformatics **SIB**
Member GA4GH Strategic Leadership Committee
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.

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

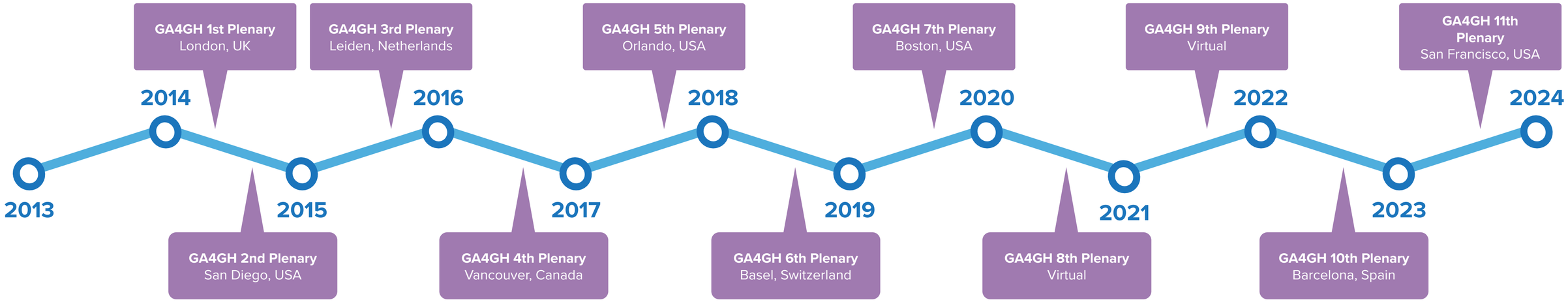
Technology

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,*]

GA4GH timeline

Global Alliance for Genomics & Health



Pre-launch

73 partners sign a letter of intent to form an alliance

Building momentum

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

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

GA4GH CONNECT

Launch of **GA4GH Connect** and Strategic Roadmap

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

Gap analysis

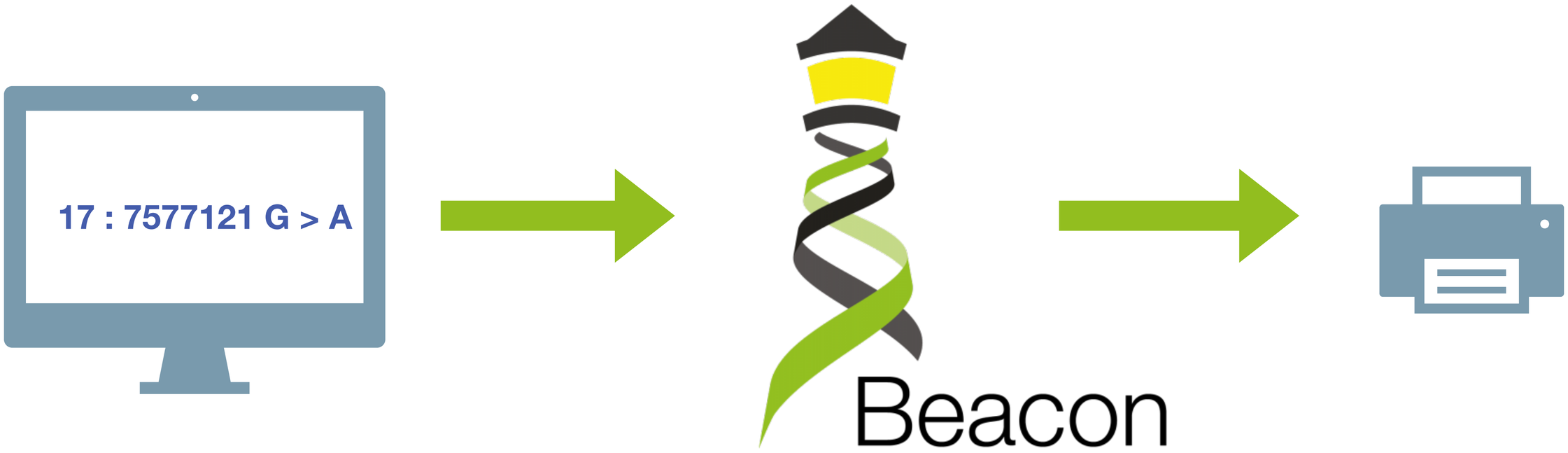
Gap analysis identifies three community imperatives

- Interoperability and alignment
- Implementation support
- Engaging with healthcar and clinical standards

Strategic Refresh

Strategic refresh introduces updates to GA4GH to better meet the three community imperatives

	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 @ **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 • Beacon API v2 Kick off • adopting "handover" concept	• new Beacon website (March)
2019	• ELIXIR Beacon Network		• Beacon publication at Nature Biotechnology
2020		• "**Scouts**" teams working on different aspects - filters, genomic variants, compliance ...	
2021		• **framework** + **models** concept implemented • range and bracket queries, variant length • starting of GA4GH review process	• Phenopackets v2 approved
2022		• changes in default model, aligning with Phenopackets and VRS • unified beacon-v2 code & docs repository • **Beacon v2 approved** at April GA4GH Connect	• *docs.genomebeacons.org*



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.

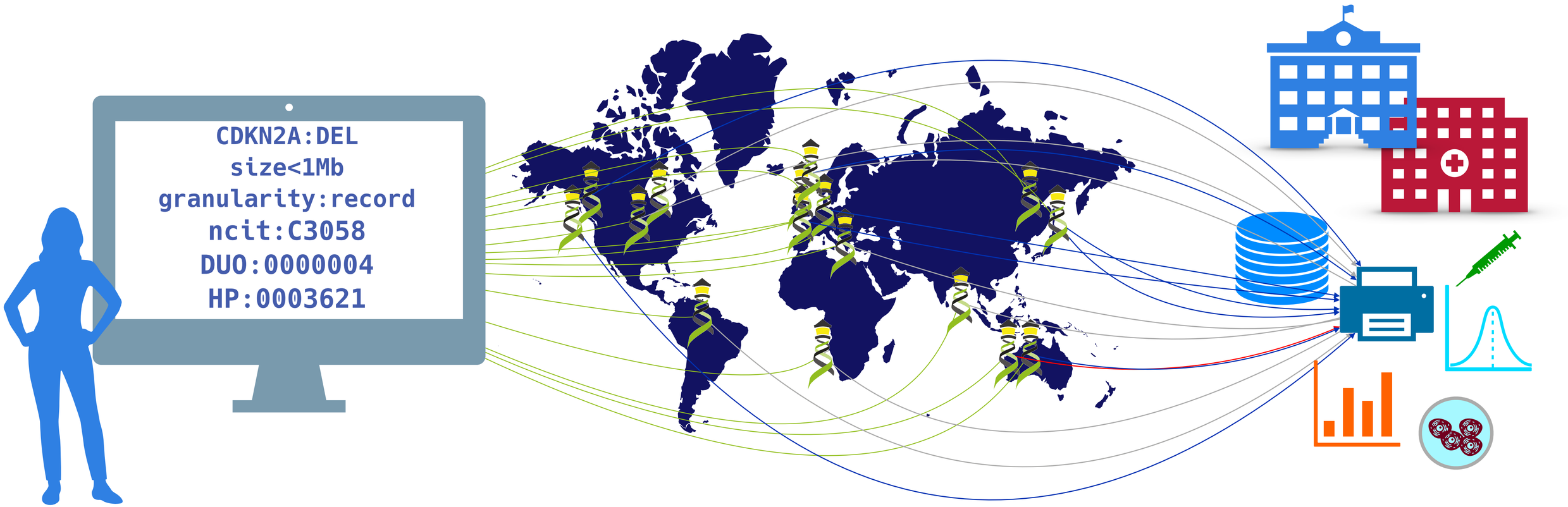
A number of more useful first versions have been suggested.

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*** …



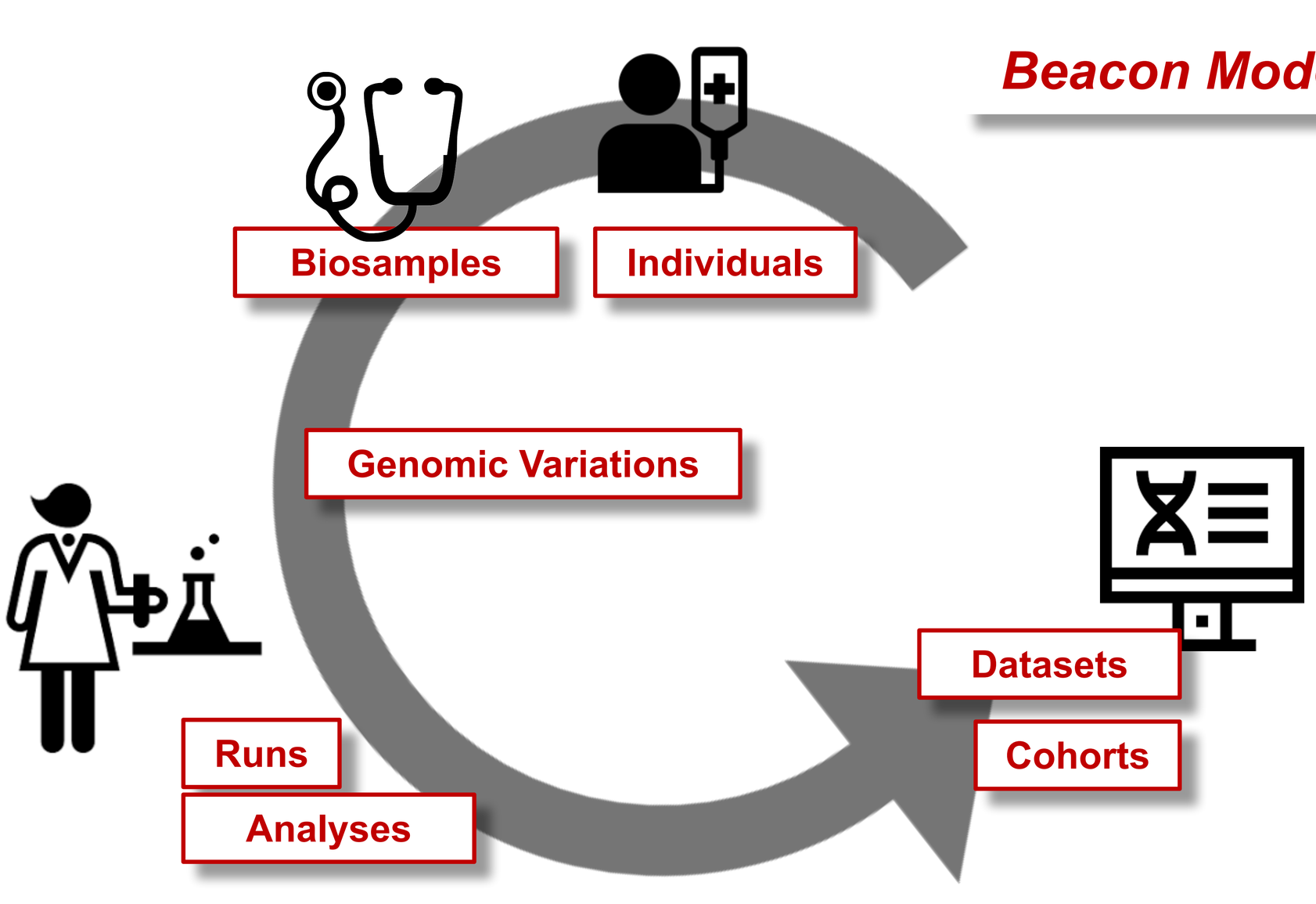
Can you provide data about focal deletions in CDKN2A in Glioblastomas from juvenile patients with unrestricted access?

Beacon *v2* API

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

Beacon v2

docs.genomebeacons.org

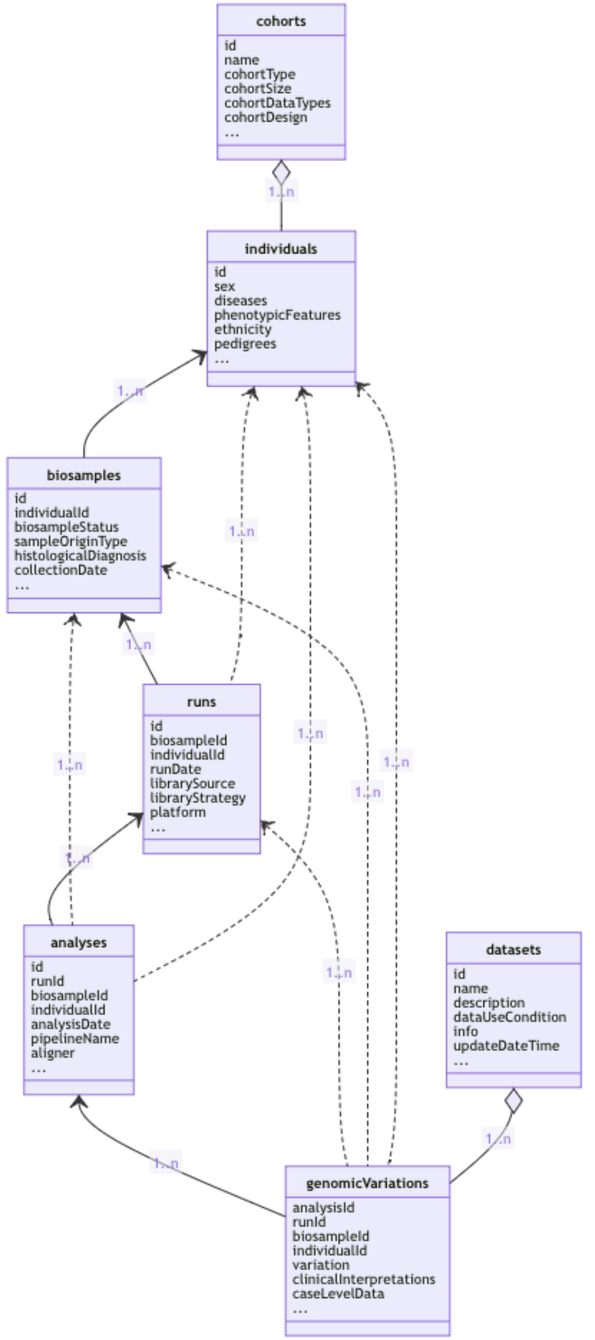


Beacon Model

Beacon Framework (protocol)

Beacon Default v2 Model

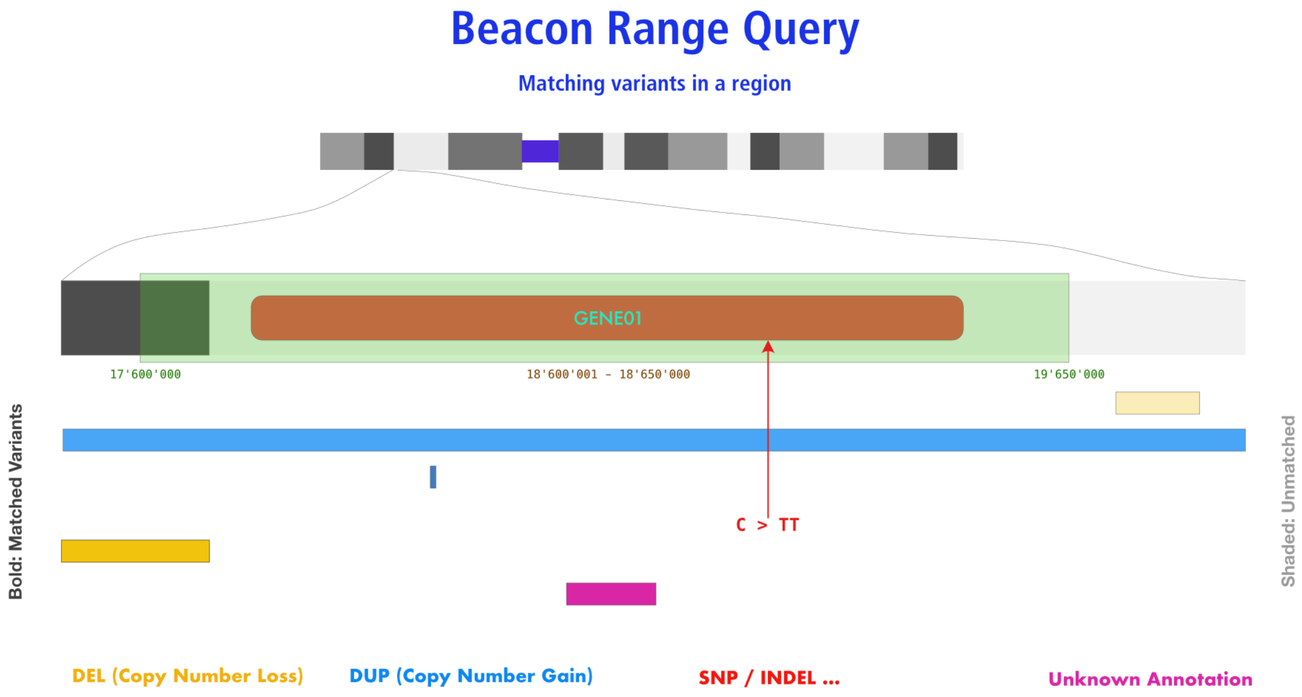
- The Beacon ***framework*** describes the overall structure of the API requests, responses, parameters, the common components, etc.
- Beacon ***models*** describe the set of concepts included in a Beacon, like individual or biosample, and also the relationships between them.
- Besides logical concepts, 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

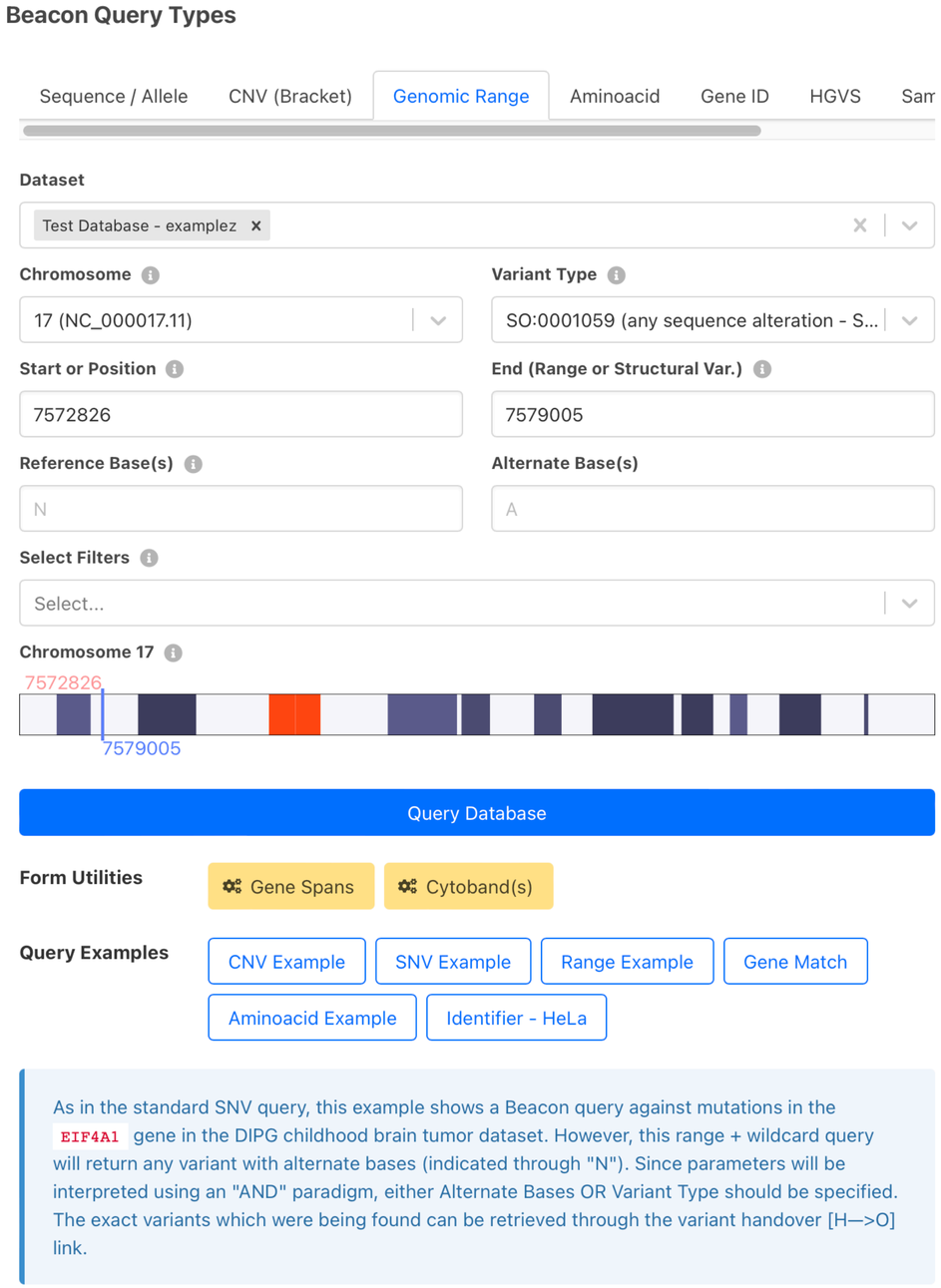


Variation Queries

Range ("anything goes") Request

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

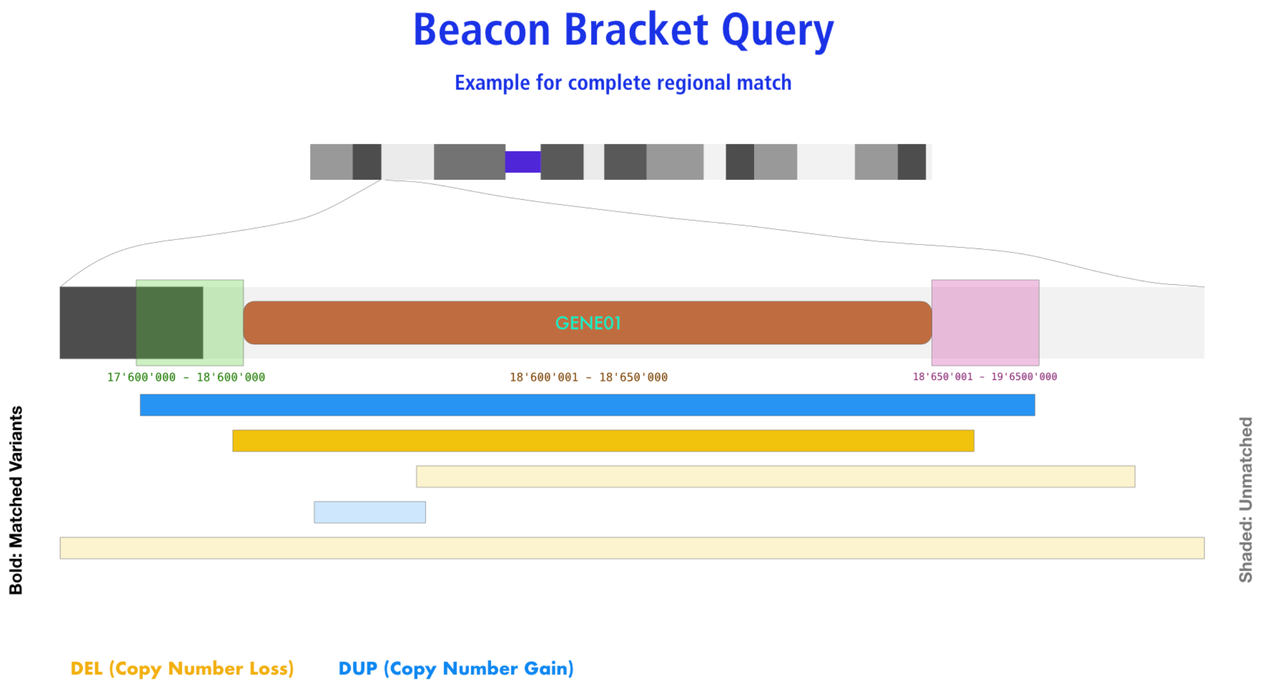


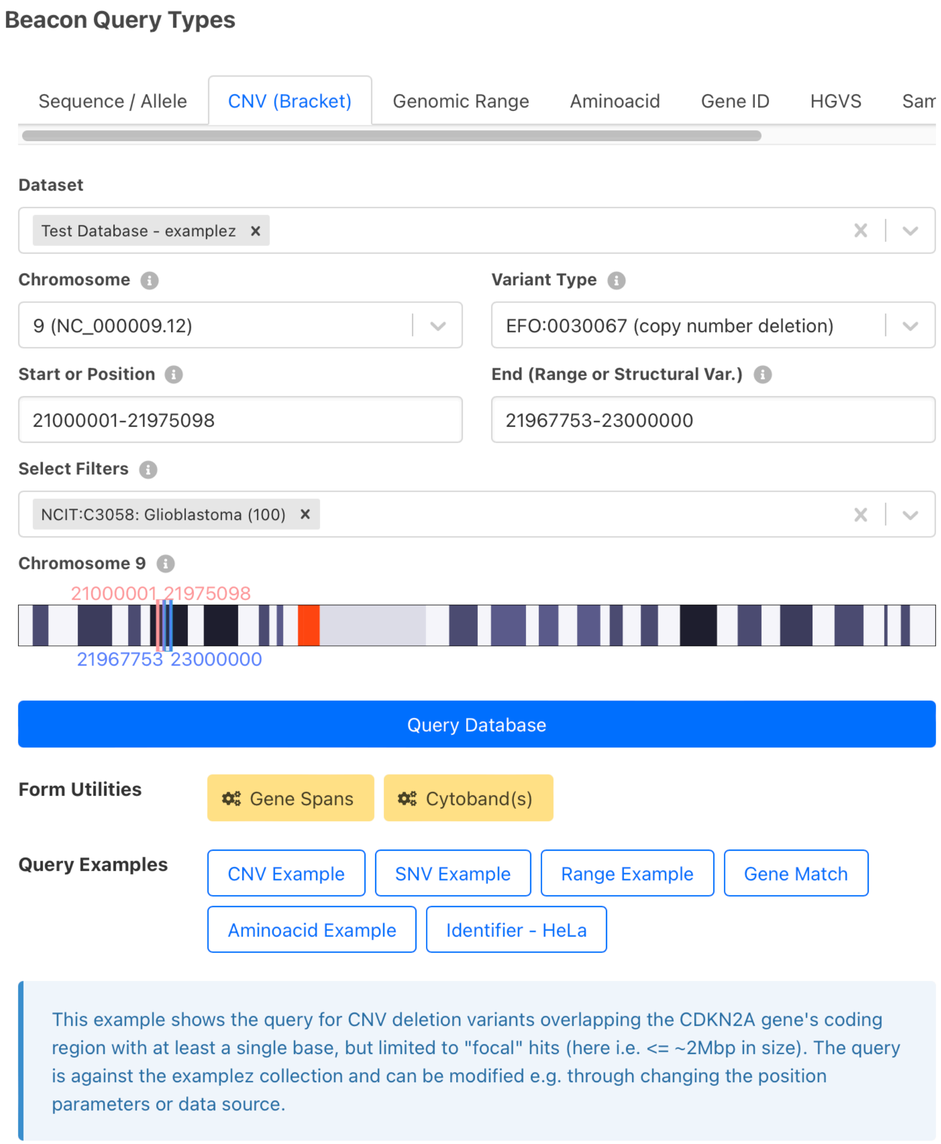


Variation Queries

Bracket ("CNV") Query

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





Standards Development & Implementation: CNV Terms

in computational (file/schema) formats

```
- EFO:0030064
- EFO:0030067
    |- EFO:0030068
    \- EFO:0020073
         \- EFO:0030069
- EFO:0030070
    |- EFO:0030071
    \- EFO:0030072
```

GA4GH VRS1.3+	Beacon v2	VCF v4.4	SO
EFO:0030070 gain	DUP or **EFO:0030070**	DUP SVCLAIM=D	SO:0001742 copy_number_gain
EFO:0030071 low-level gain	DUP or **EFO:0030071**	DUP SVCLAIM=D	SO:0001742 copy_number_gain
EFO:0030072 high-level gain	DUP or **EFO:0030072**	DUP SVCLAIM=D	SO:0001742 copy_number_gain
EFO:0030072 high-level gain	DUP or EFO:0030073	DUP SVCLAIM=D	SO:0001742 copy_number_gain
EFO:0030067 loss	DEL or **EFO:0030067**	DEL SVCLAIM=D	SO:0001743 copy_number_loss
EFO:0030068 low-level loss	DEL or **EFO:0030068**	DEL SVCLAIM=D	SO:0001743 copy_number_loss
EFO:0020073 high-level loss	DEL or **EFO:0020073**	DEL SVCLAIM=D	SO:0001743 copy_number_loss
EFO:0030069 complete genomic loss	DEL or **EFO:0030069**	DEL SVCLAIM=D	SO:0001743 copy_number_loss

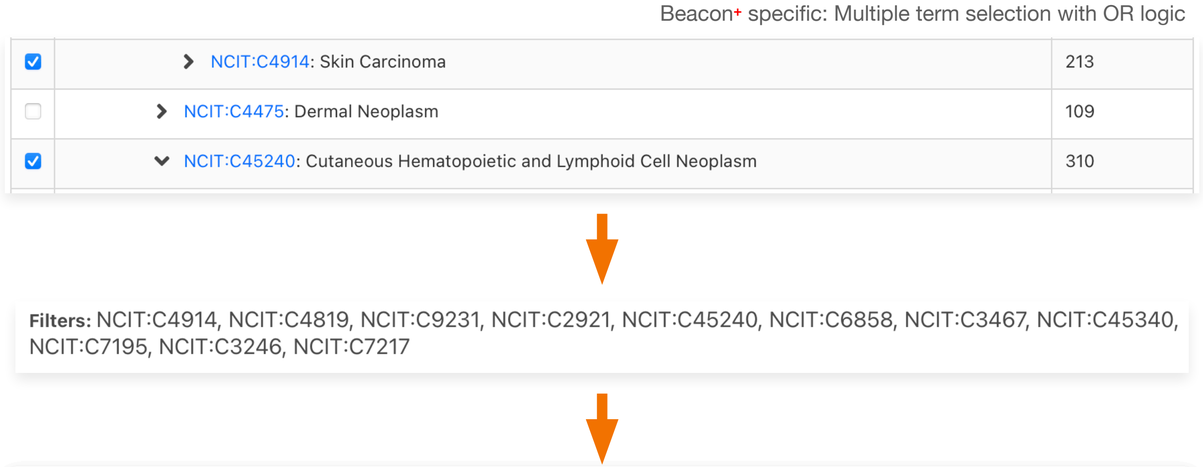
Beacon v2 Filters

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

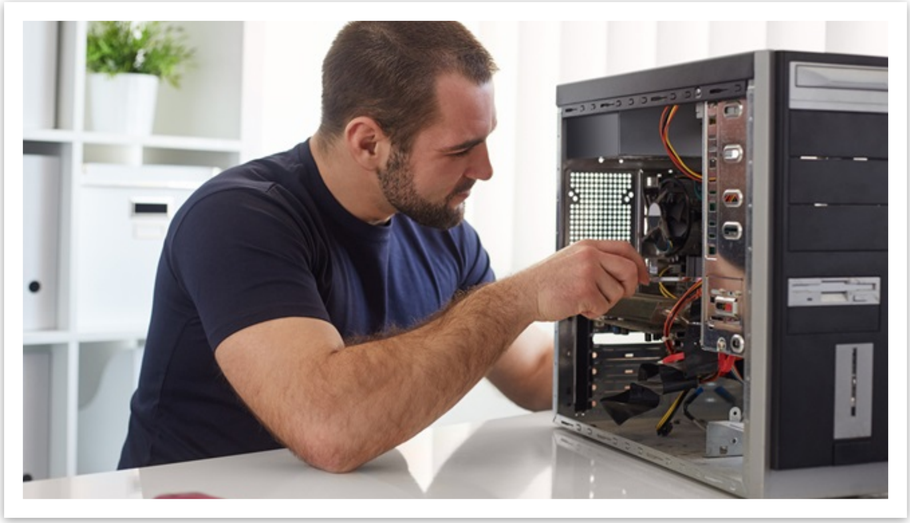
Begriffsbestimmung

The right expressions help to conceptualize...

- **Beacon**: The protocol/API, with framework and default model
- **beacon**: Implementation of Beacon
  - ➡ using the Beacon v2 framework & supporting at minimum boolean responses
  - ➡ suggested support of Beacon v2 default model but can choose other
- Beacon **Aggregator**: service distributes queries to beacons and aggregates responses into a single Beacon response
  - ➡ potential to liftover genomes, remap filtering terms, translate between protocol versions...
  - ➡ entry point to or potentially itself node in a ...
- Beacon **Network**: Set of beacons wit shared entry point for distributed queries and aggregated response delivery
  - ➡ "true" beacon networks should have managed aspects - scope, term use...
  - ➡ networks may combine mixes of internal (protected, rich data, additional extensions...) and external interfaces

Beacon v2 deployment

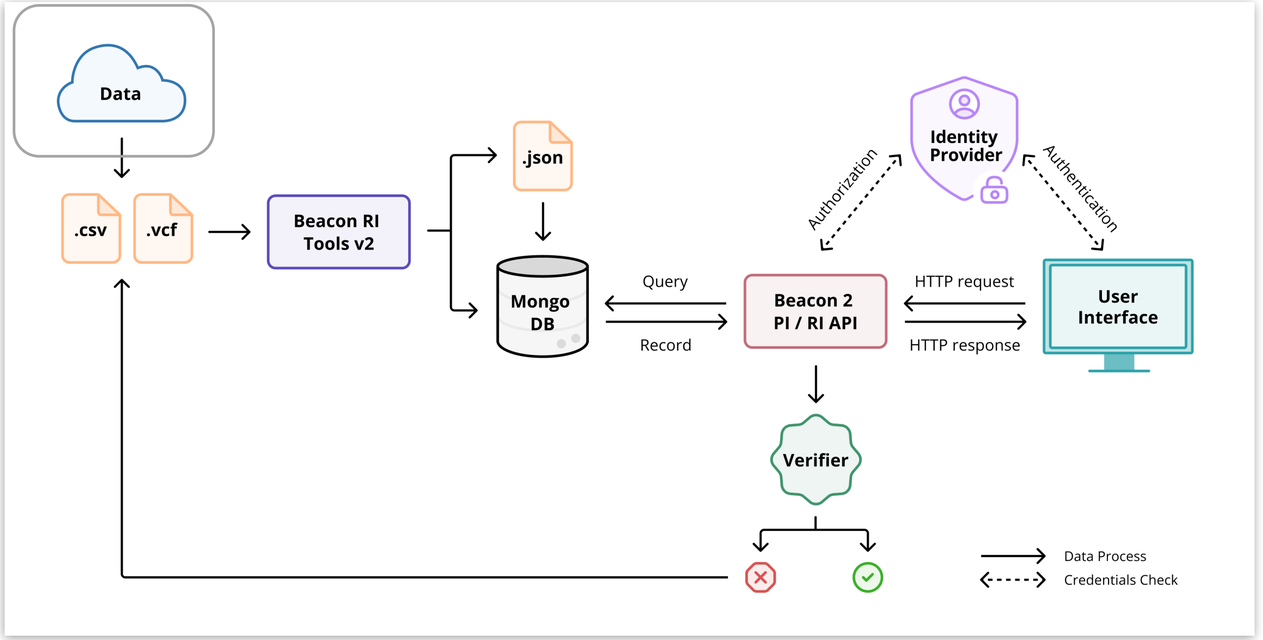
Build it yourself



GitHub

Beacon v2 API
https://github.com/ga4gh-beacon/beacon-v2

Toolkit for production environments



Beacon v2 Production Implementation (released Oct 2024)
https://github.com/ga4gh-beacon/beacon-v2

https://www.choice.com.au/electronics-and-technology/computers/desktop-and-laptop-computers/articles/is-building-your-own-pc-worth-it

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™

- *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"),
```

mongoDB

variants analyses biosamples individuals

collations geolocs genespans publications qBuffer

Entity collections

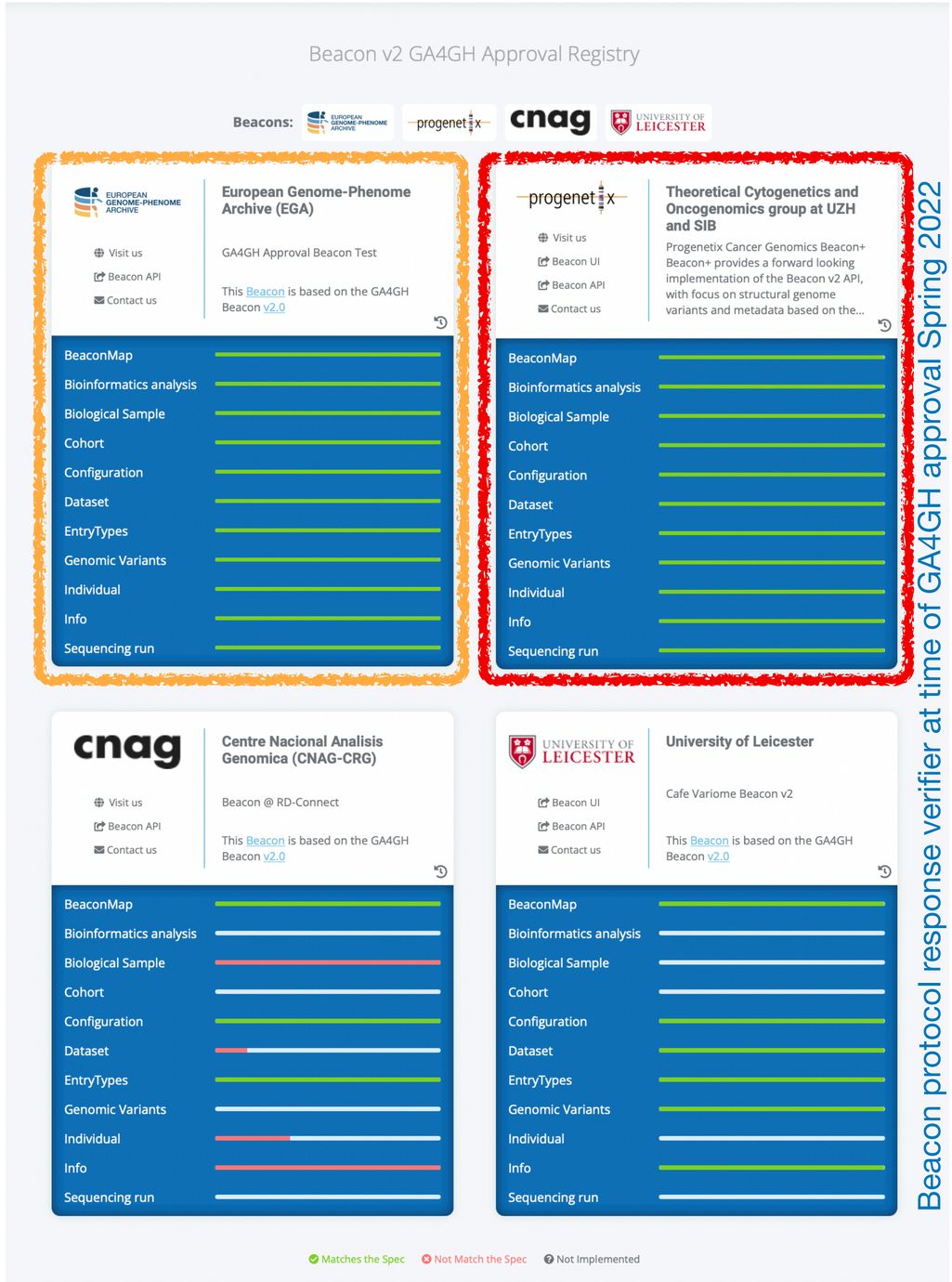
GitHub github.com/progenetix/bycon

Utility collections

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

progenetix.org

Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **140'000 cancer CNV profiles**
- SNV data for some series (e.g. TCGA)
- more than **900 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

Universität Zürich[UZH]

progenetix

SIB Swiss Institute of Bioinformatics

progenetix

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

Search Samples

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

Publication DB
- Genome Profiling
- Progenetix Use

Services
- NCIt Mappings
- UBERON Mappings

Upload & Plot

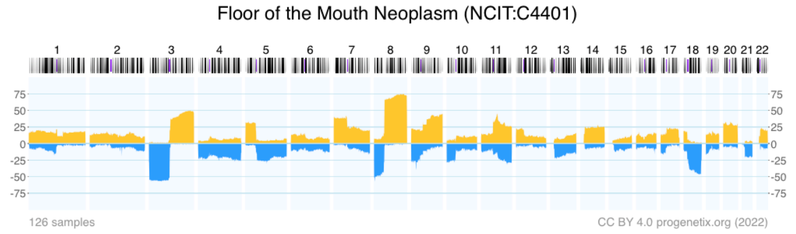
Beacon+

Documentation
- News
- Downloads & Use Cases
- Sevices & API

Baudisgroup @ UZH

Cancer genome data @ progenetix.org

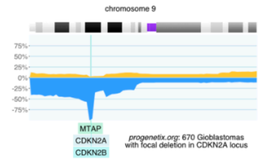
The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* from currently **142063** samples.



Download SVG | Go to NCIT:C4401 | Download CNV Frequencies

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

Progenetix Use Cases



Local CNV Frequencies

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

Cancer CNV Profiles

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

Cancer Genomics Publications

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

Beacon+: Phenopackets

Testing alternative response schemas...

https://progenetix.org/beacon/phenopackets/pgxind-kftx26j0

- 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

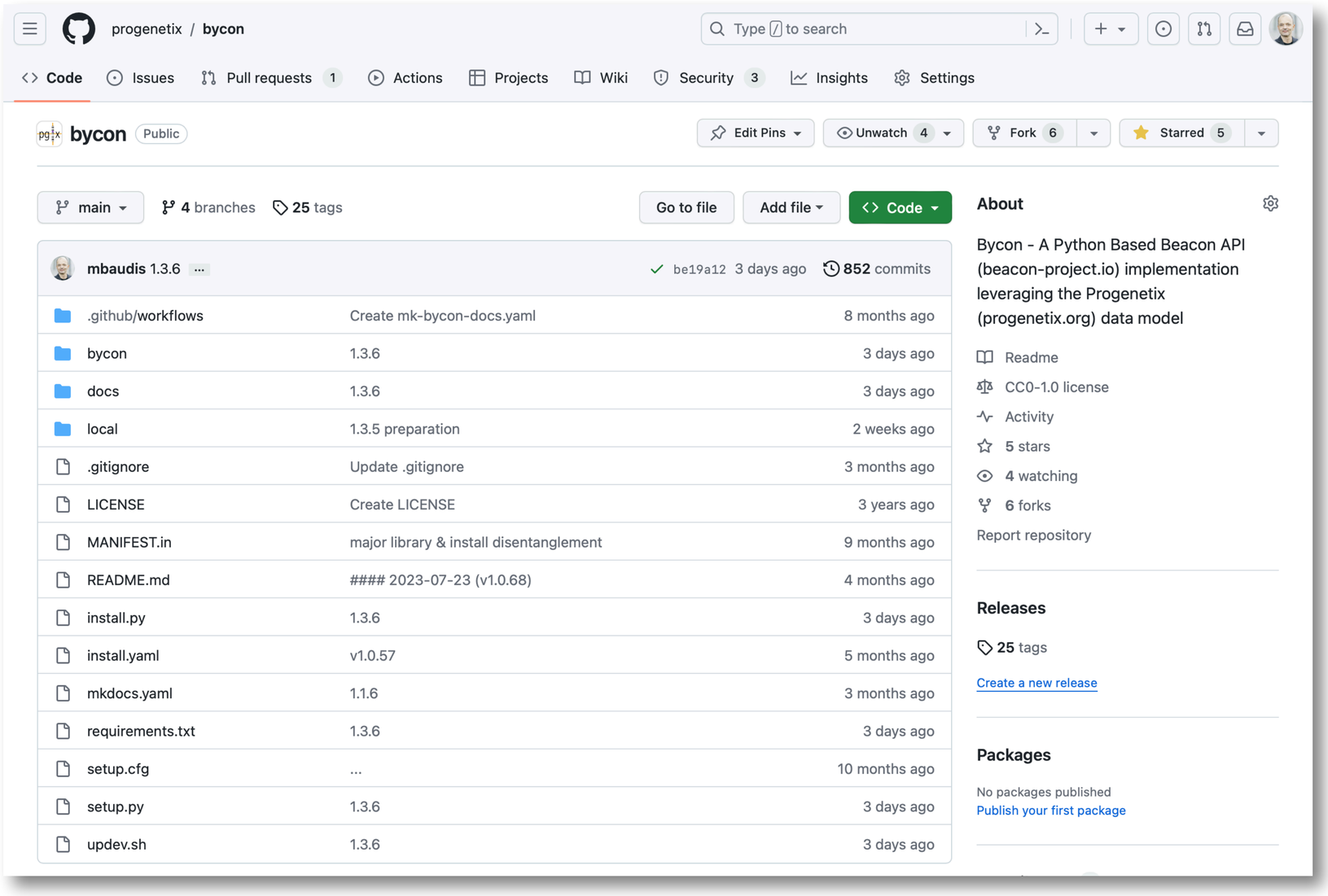

```
"id": "pgxpxf-kftx3tl5",
"metaData": {
  "phenopacketSchemaVersion": "v2",
  "resources": [
    {
      "id": "NCIT",
      "iriPrefix": "http://purl.obolibrary.org/obo/NCIT_",
      "name": "NCIt Plus Neoplasm Core",
      "namespacePrefix": "NCIT",
      "url": "http://purl.obolibrary.org/obo/ncit/neoplasm-core.c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

Looking for implementers and contributors

- containerization
- data I/O ...
- standard library integration (VRSification of variants...)



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

pgxRpi: an R/Bioconductor package

Client for Accessing Beaconized Data

pgxRpi

This is the **development** version of pgxRpi; for the stable release version, see pgxRpi.

R wrapper for Progenetix

platforms all | rank 2178 / 2266 | support 0 / 0 | in Bioc < 6 months | build unknown | updated < 1 month | dependencies 137

DOI: 10.18129/B9.bioc.pgxRpi

Bioconductor version: Development (3.20)

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>

- **Query and export variants**

https://progenetix.org/beacon/biosamples/pgxbs-kftvh94d/g_variants

```
> variants <- pgxLoader(type="variant",biosample_id="pgxbs-kftvh94d")
```

- **Query metadata of biosamples and individuals by filters (e.g. NCIt, PMID)**

http://progenetix.org/services/sampletable/?filters=NCIT:C3697

```
> biosamples <- pgxLoader(type="biosample",filters="NCIT:C3697")
```

- **Query and visualize CNV frequency by filters**

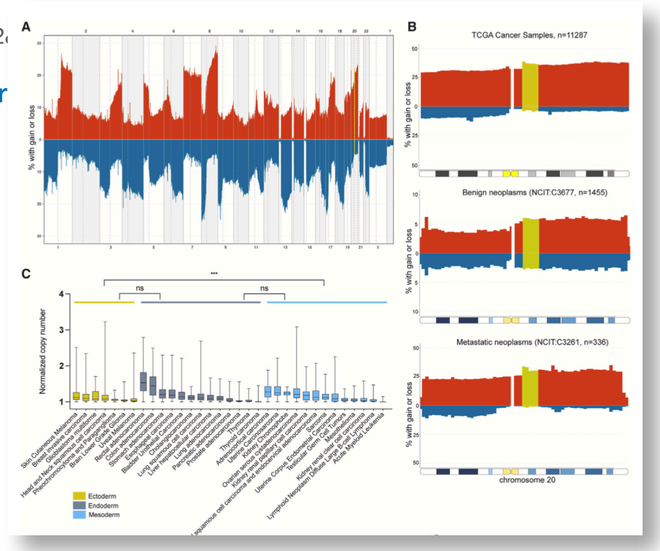
http://www.progenetix.org/services/intervalFrequencies/?filters=NCIT:C3512

```
> freq <- pgxLoader(type="frequency",output="pgxfreq",filter
> pgxFreqplot(freq)
```

- **Process local .pgxseg files**

```
> info <- pgxSegprocess(file=file, show_KM_plot = T,
return_seg = T,  return_metadata = T, return_frequency = T)
```

Use case: 2024 article using Progenetix' ***pgxRpi*** to retrieve & visualize 117'587 cancer CNV profiles for a study into pluripotent stem cells' genomics



Stem Cell Reports

Review

ISSCR

OPEN ACCESS

Gains of 20q11.21 in human pluripotent stem cells: Insights from cancer research

Nuša Krivec,[1,2] Manjusha S. Ghosh,[1,2] and Claudia Spits[1,2,*]

[1]Research Group Reproduction and Genetics, Faculty of Medicine and Pharmacy, Vrije Universiteit Brussel, Brussels, Laarbeeklaan 103, 1090 Brussels, Belgium

[2]These authors contributed equally

*Correspondence: claudia.spits@vub.be

https://doi.org/10.1016/j.stemcr.2023.11.013

Figure 2. Copy-number alterations of human chromosome 20q11.21 in cancers

(A) Aggregated copy-number variation (CNV) data of 117,587 neoplasms (NCIT: C3262) from the Progenetix database (Huang et al., 2021) were plotted using R library pgxRpi. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079-35871578 is marked in moss green. NCIT, National Cancer Institute Thesaurus.

(B) Top to bottom: Aggregated CNV data of 11,287 TCGA cancer samples, 336 metastatic neoplasms (NCIT: C3261), and 1,455 benign neoplasms (NCIT: C3677) from the Progenetix database (Huang et al., 2021), respectively, were plotted using R library pgxRpi. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079–35871578 is marked in moss green.

Beacon Security

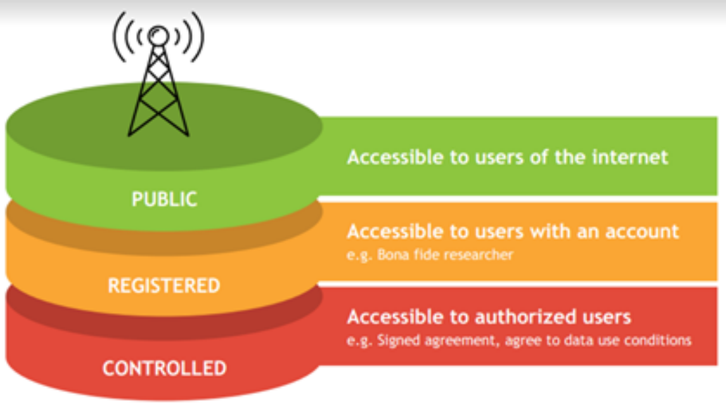
Making Beacons Biomedical - Beacon v2

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

Definitely breaks the "Relative Security by Design" Concept!

Beacon Security

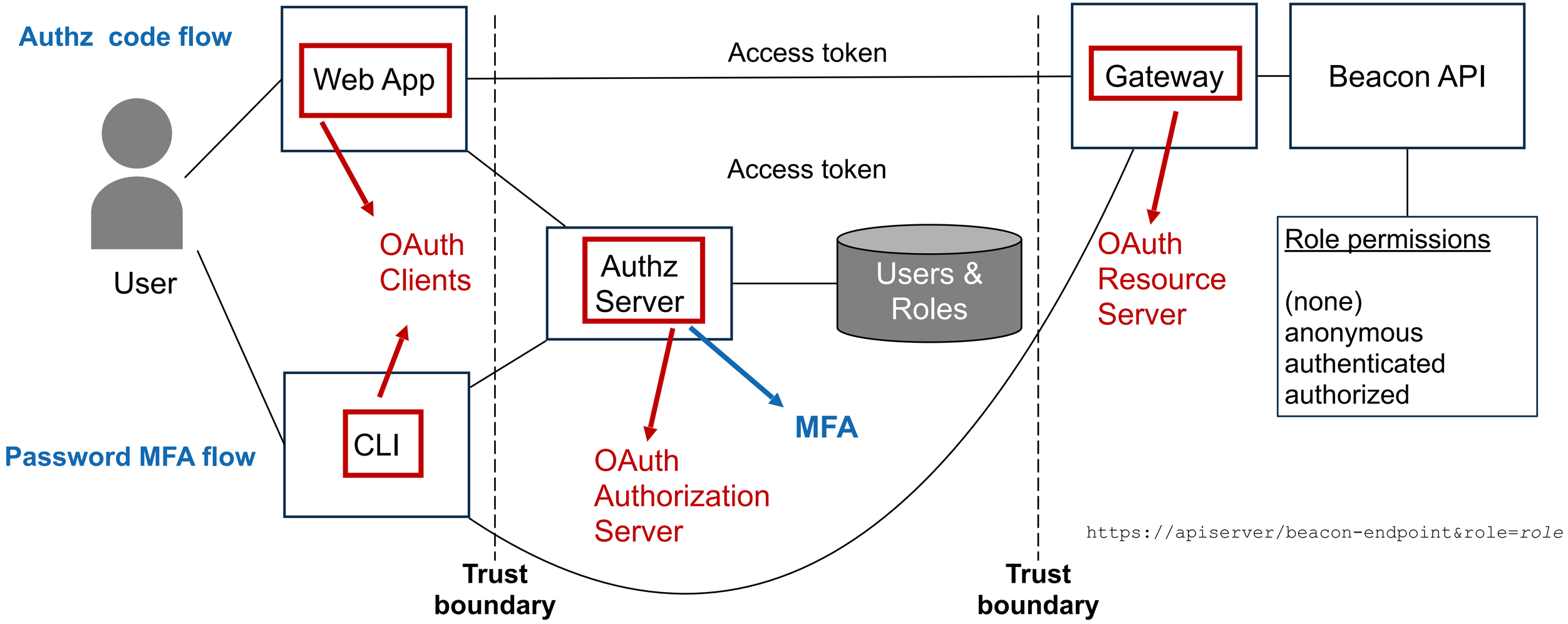
Security by Design ... if Implemented in the Environment



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

Architecture

Running the *bycon* stack in a secure environment



Architecture

Running the *bycon* stack in a secure environment

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

Beacon as a global standard

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

Beacon Scouts

Real-world needs

Cancer

Common diseases

Rare Diseases

…

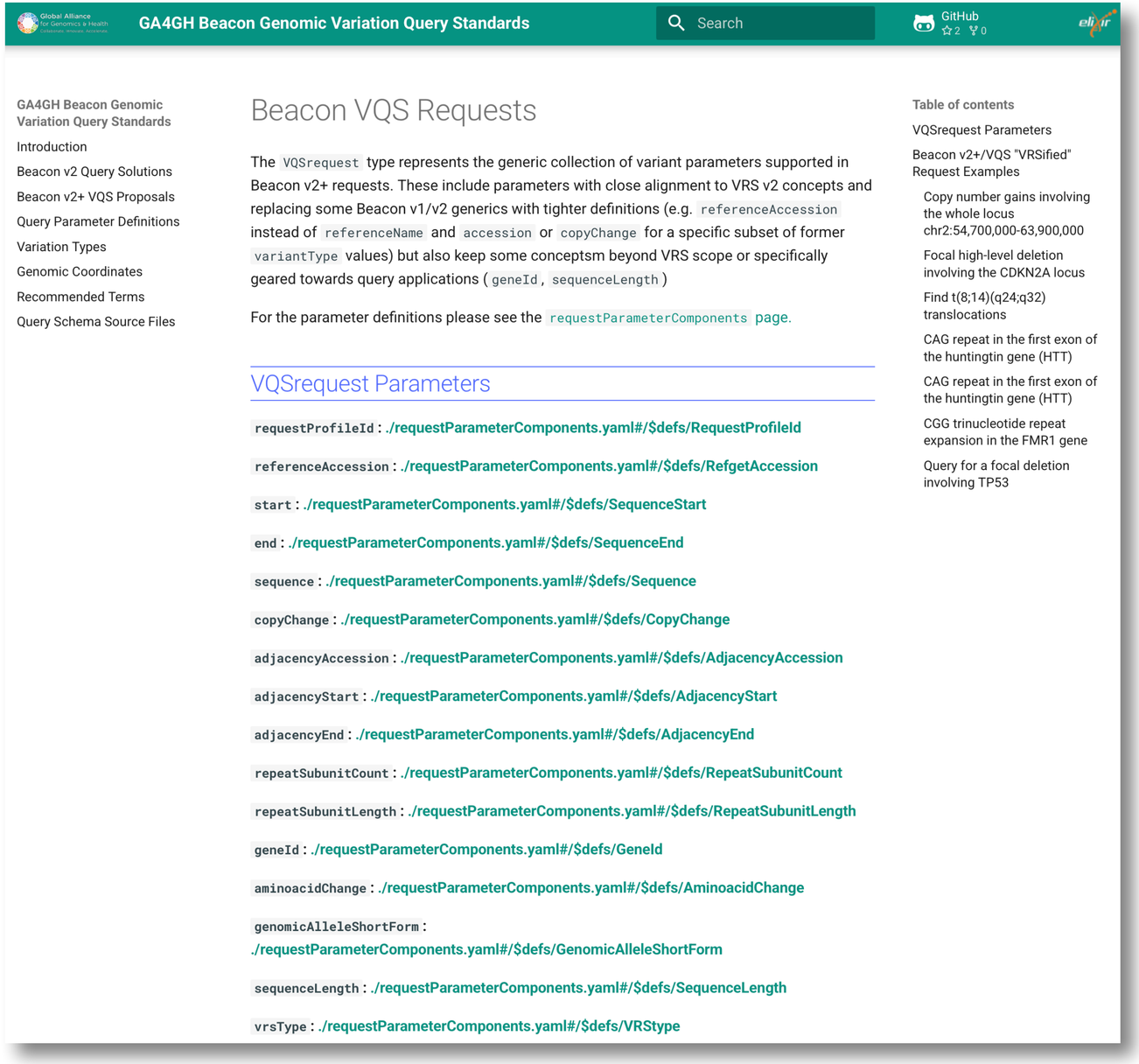
- **Beacon Filters** – improve current filter solutions
- **Beacon Cohorts** – develop aggregated request and response (e.g. counts by sex and age)
- **Beacon Variants** – expand specification to cover new use cases and typed queries
- **Beacon Dev** – improve API (cleaning code, GitHub issues)
- **Beacon Matchmaking** – implementation in matchmaking use cases

EUROPEAN GENOME-PHENOME ARCHIVE

Beacon Scouts

Finding the Paths to Beacon's Future

- Genomic Variation Scouts
  - ➡ extension to the query model based on assessed needs
    - fusions/breakpoints, cytogenetic annotations, repeats, categorical variants...
  - ➡ adoption of evolving VRS... standards for variant representation
    - adjacency, repeats...



https://genomebeacons.org/variant-query-types/variant-scouts-home/

Progenetix Cancer Genomics Beacon+ v2.1.0-beaconplus OAS 3.0

/api

Beacon+ provides a forward looking implementation of the Beacon v2 API, with focus on structural genome variants and metadata based on the cancer and reference genome profiling data in the Progenetix resource (https://progenetix.org).

This page presents a prototype for an OpenAPI (Swagger) definition for the GA4GH Beacon API. The definitions are generated from the `entity_defaults` and `argument_definitions` in the bycon project. The complete. Please be aware that the whole capabilities of the project cannot be represented solely through the OpenAPI definitions and also involve features such as filtering terms logic and result aggregation entities. Additionally, the bycon project implements a number of data services beyong Beacon standards which again are only partially covered here.

`bycon` and Data Aggregation

The Beacon standard implements a REST style syntax - e.g. consistent id-based document retrieval for entities indicated through their framework provide full data aggregation; *i.e.* queries with parameters against **any** of the main data entities (g_variants, runs, analyses, intersection of the query results at the level of the response entity.

[Beacon v2 API] | [Example: Bv2 CNV / bracket] | [Proposal: VQS CNV / bracket] | [Proposal: VQS CNV by Gene ID] | [Proposa

Contact the developer

Servers

https://progenetix.org

Beacon

- GET /beacon/info Get info entries
- GET /beacon/datasets Get dataset entries
- GET /beacon/cohorts Get cohort entries
- GET /beacon/g_variants Get genomicVariant entries
- GET /beacon/g_variants/{id} Get genomicVariant entries
- GET /beacon/g_variants/{id}/analyses Get analysis entries
- GET /beacon/g_variants/{id}/biosamples Get biosample entries
- GET /beacon/g_variants/{id}/individuals Get individual entries
- GET /beacon/analyses Get analysis entries
- GET /beacon/analyses/{id} Get analysis entries
- GET /beacon/analyses/{id}/g_variants Get genomicVariant entries
- GET /beacon/analyses/{id}/biosamples Get biosample entries

Bv2minimalAlleleRequest

GET /beacon/g_variants Get genomicVariant entr

Get genomicVariant entries

Parameters

Name	Description
referenceName string (query)	Examples: Chromosome 17 17
start array<integer> (query)	Examples: Base position on chromosome 17 7577120 Add integer item
alternateBases string (query)	Examples: An `A` allele at the specified posit A
referenceBases string (query)	Examples: A reference `G` allele at the specif G
skip integer (query)	skip
limit integer (query)	limit
requestedGranularity string (query)	Examples: The minimal boolean response boolean

Bv2cnvbracketquery

GET /beacon/g_variants Get genomicVariant entries

Get genomicVariant entries

Parameters

Name	Description
filters array<string> (query)	Examples: Glioblastoma NCIT:C3058 Add string item
referenceName string (query)	Examples: Chromosome 9 (GRCh38) refseq:NC_000009.12
start array<integer> (query)	Examples: Range for start of CNV involving CDKN2A 21000001 21975098 Add integer item
end array<integer> (query)	Examples: Range for end of CNV involving CDKN2A 21967753 23000000 Add integer item
variantType string (query)	Examples: High-level copy number loss EFO:0020073

VQSadjacencyRequest

GET /beacon/g_variants Get genomicVariant entries

GET /beacon/analyses Get analysis entries

GET /beacon/biosamples Get biosample entries

Get biosample entries

Parameters

Name	Description
filters array<string> (query)	Examples: Malignant lymphoma, NOS (ICD-O 3 code 9680/3) pgx:icdom-95903 Add string item
referenceAccession string (query)	Examples: RefSeq ID for Chromosome 8 (GRCh38) refseq:NC_000008.11
breakpointRange array<integer> (query)	Examples: Range for band q24 on chromosome 8 116700000 145138636 Add integer item
adjacencyAccession string (query)	Examples: RefSeq ID for Chromosome 14 (GRCh38) refseq:NC_000014.9
adjacencyRange array<integer> (query)	Examples: Range for band q32 on chromosome 14 89300000 107043718 Add integer item
vrsType string (query)	Examples: Adjacency Adjacency

a menu

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

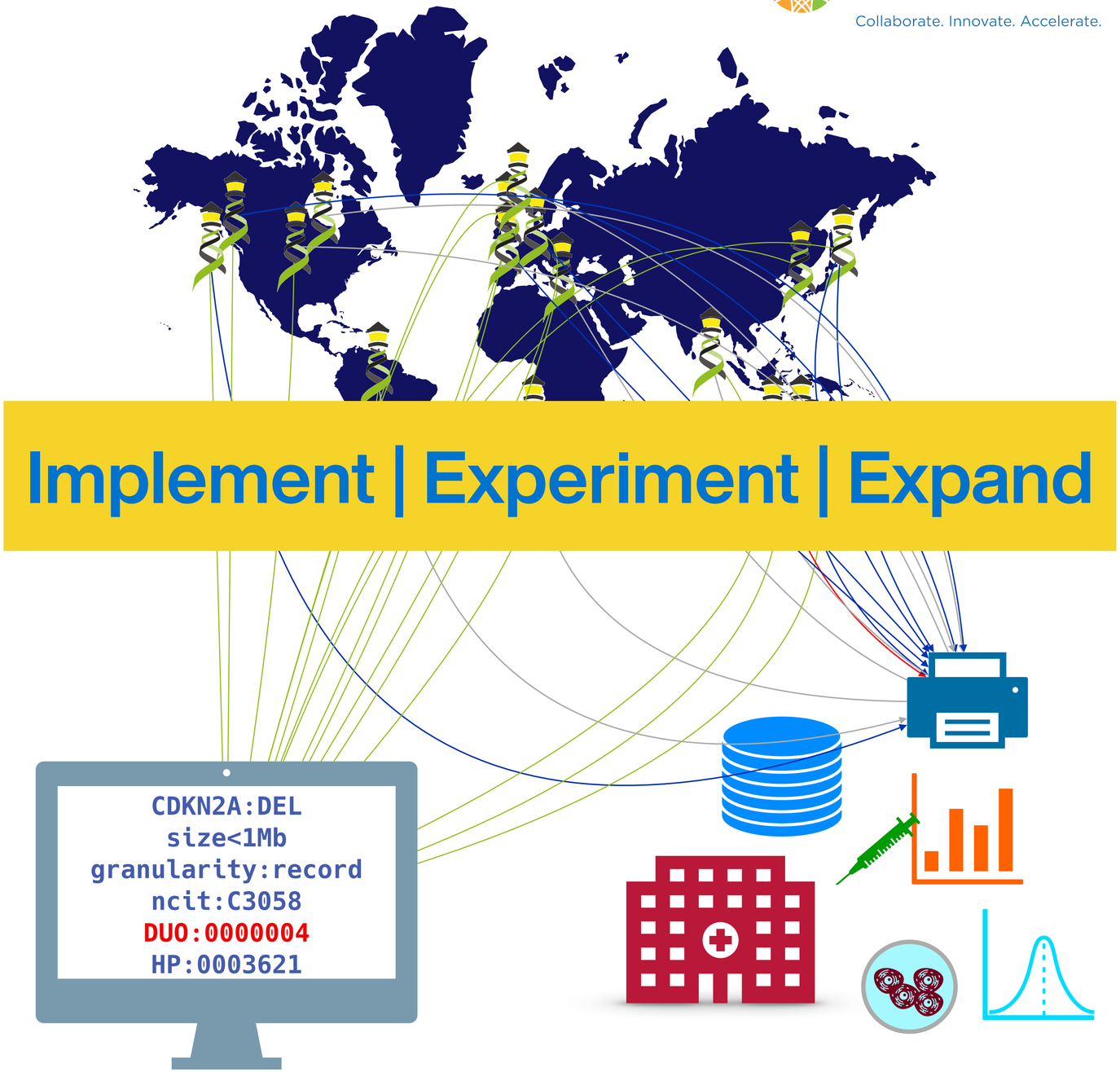
What Can You Do?

- 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

Beacon for Genomic Discovery Proxies

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

- Feature beacons for privacy protecting data discovery
  - ➡ privacy protection through aggregated data, cohorts
  - ➡ alternative is "**horizontal gatekeeping**": separate Beacons for ***discovery*** of e.g. genomic and phenotypic data and ***data delivery*** upon request / authentication
  - ➡ We'd love to help launching your beacon (especially as a **bycon**...)



Save the dates!

Global Alliance for Genomics & Health

April Connect 2025

1 to 4 April 2025

Broad Institute, Cambridge, USA

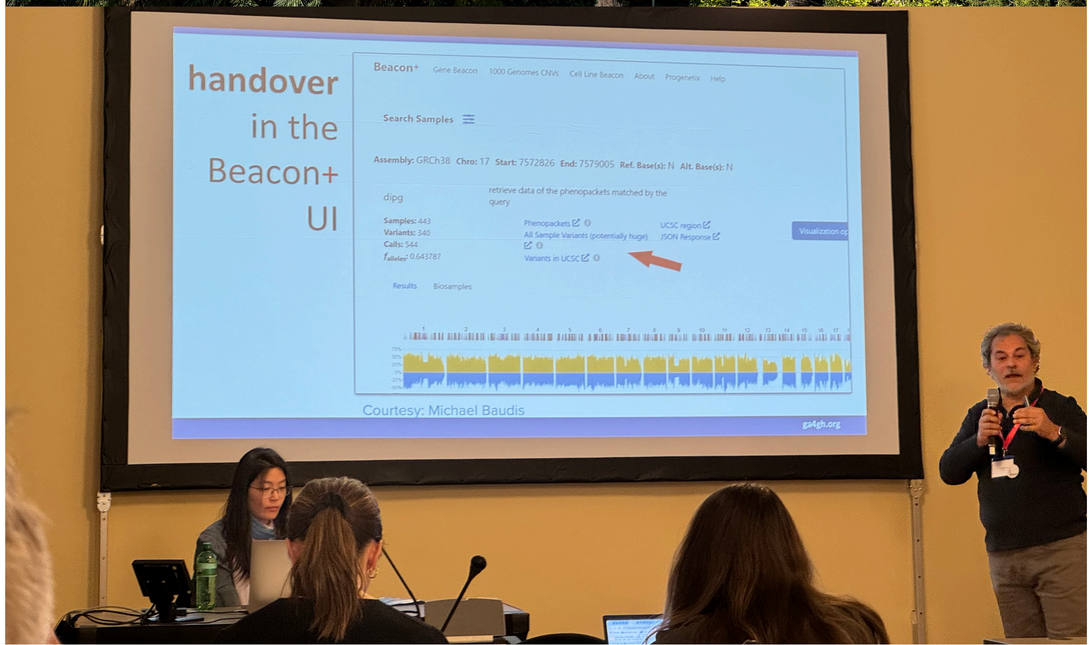
Registration Open Now

13th Plenary

6 to 10 October

UKK, Uppsala, Sweden

Registration Opening Soon



The Global Alliance for Genomics and Health (GA4GH) gathered for the 2024 **April Connect meeting** in Ascona, Switzerland and online from 21 to 24 April. The GA4GH Connect meetings provide an opportunity for contributors to advance the GA4GH Road Map, showcase GA4GH standards and policies in action, and gather feedback on product development and community needs. The meeting brought together 103 in-person attendees and 312 virtual attendees for updates from Work Streams and Driver Projects, breakout sessions, and themed events.

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
Liina Nagirnaja
Claudia Vasallo
Babita Singh
Sabela de la Torre
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