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

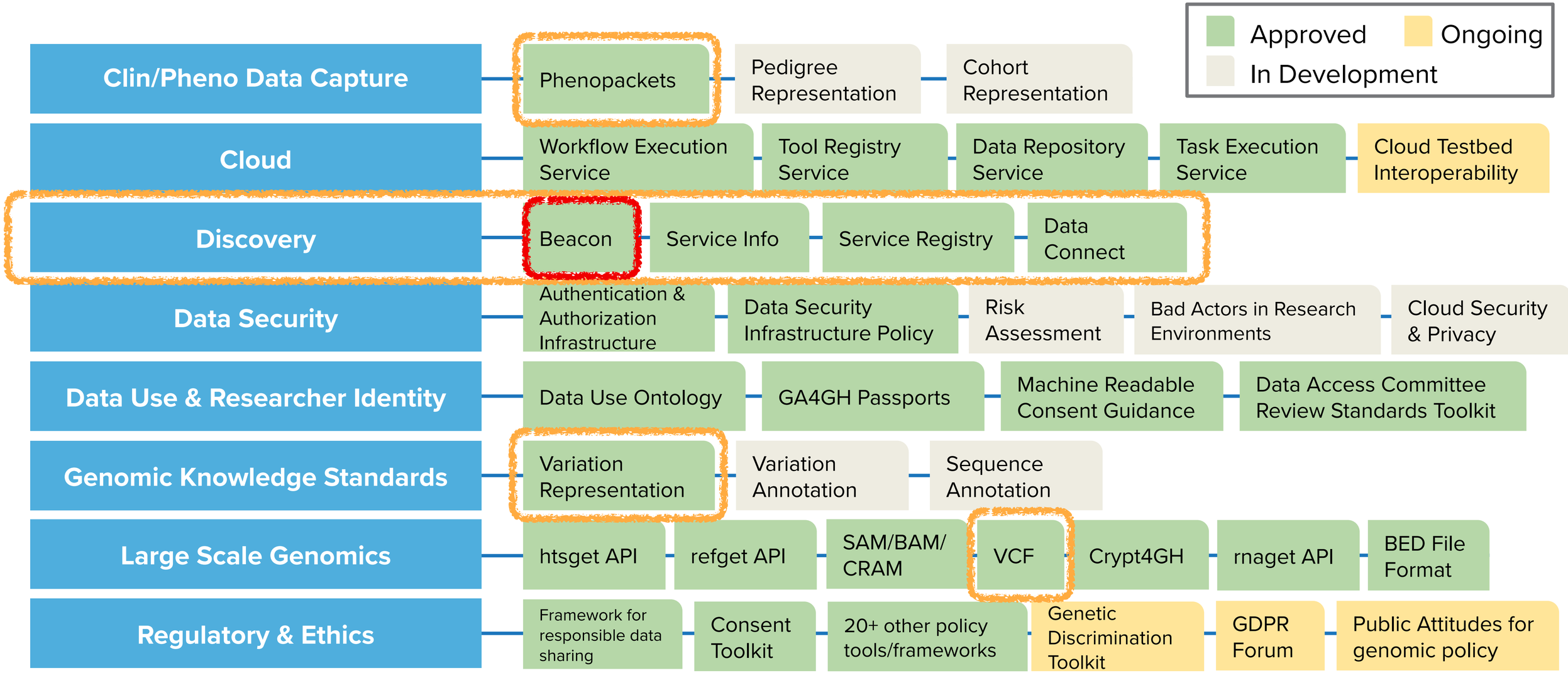
OP

Technology

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

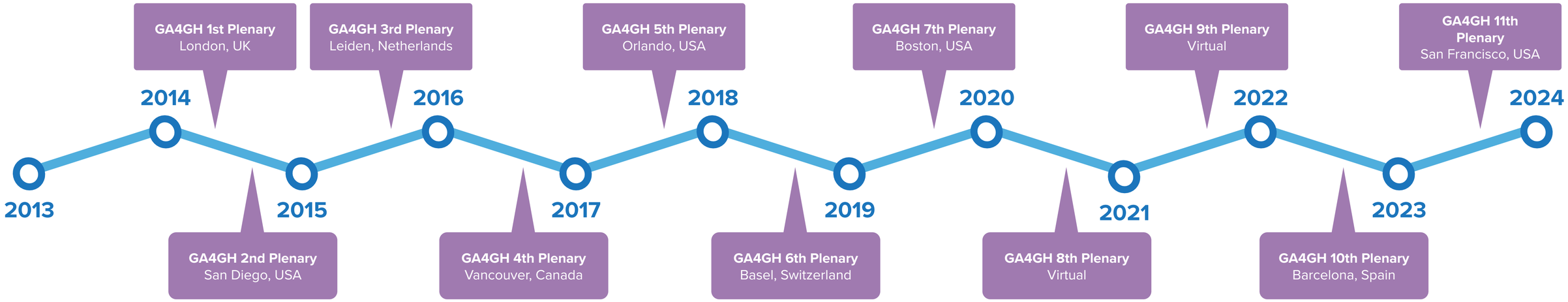
Alex H. Wagner,[1,2,25,*] Lawrence Babb,[3,*] Gil Alterovitz,[4,5] Michael Baudis,[6] Matthew Brush,[7] Daniel L. Cameron,[8,9]
Melissa Cline,[10] Malachi Griffith,[11] Obi L. Griffith,[11] Sarah E. Hunt,[12] David Kreda,[13] Jennifer M. Lee,[14] Stephanie Li,[15]
Javier Lopez,[16] Eric Moyer,[17] Tristan 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,*]

Overview of GA4GH standards and frameworks



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

GA4GH timeline

Global Alliance for Genomics & Health

GA4GH 1st Plenary London, UK

GA4GH 2nd Plenary San Diego, USA

GA4GH 3rd Plenary Leiden, Netherlands

GA4GH 4th Plenary Vancouver, Canada

GA4GH 5th Plenary Orlando, USA

GA4GH 6th Plenary Basel, Switzerland

GA4GH 7th Plenary Boston, USA

GA4GH 8th Plenary Virtual

GA4GH 9th Plenary Virtual

GA4GH 10th Plenary Barcelona, Spain

GA4GH 11th Plenary San Francisco, USA

2013 — 2014 — 2021 — 2022 — 2023 — 2024

Beacon Proposal and Conceptual Discussions

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

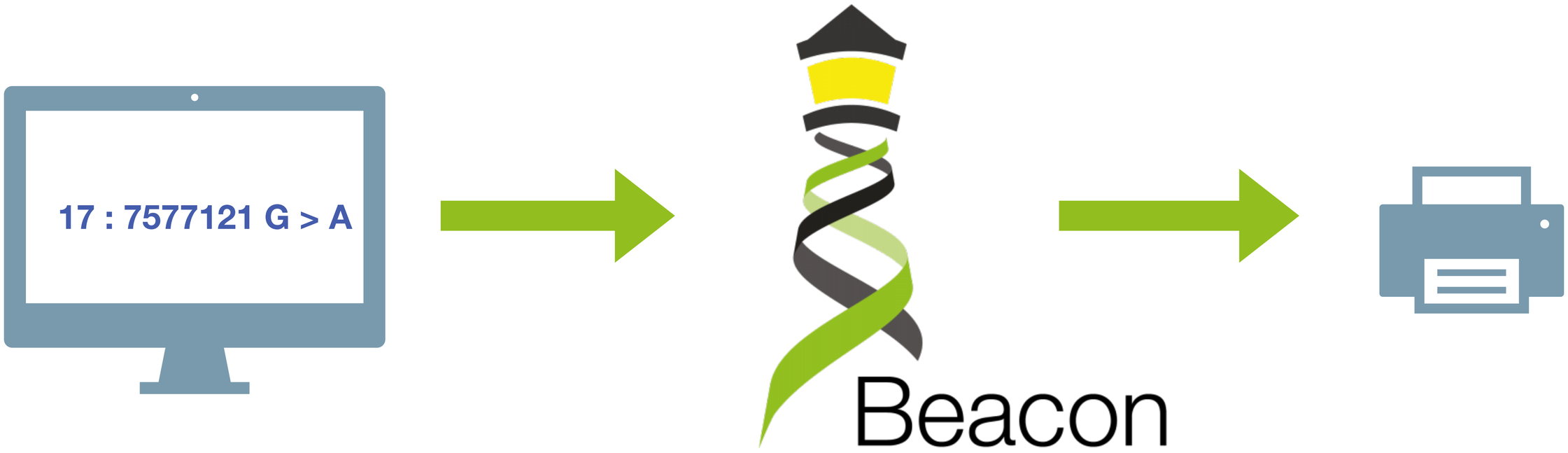
Gap analysis

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



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

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

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

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

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

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

Global Alliance for Genomics & Health

Introduction

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

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

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

Utility

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

A number of more useful first versions have been suggested.

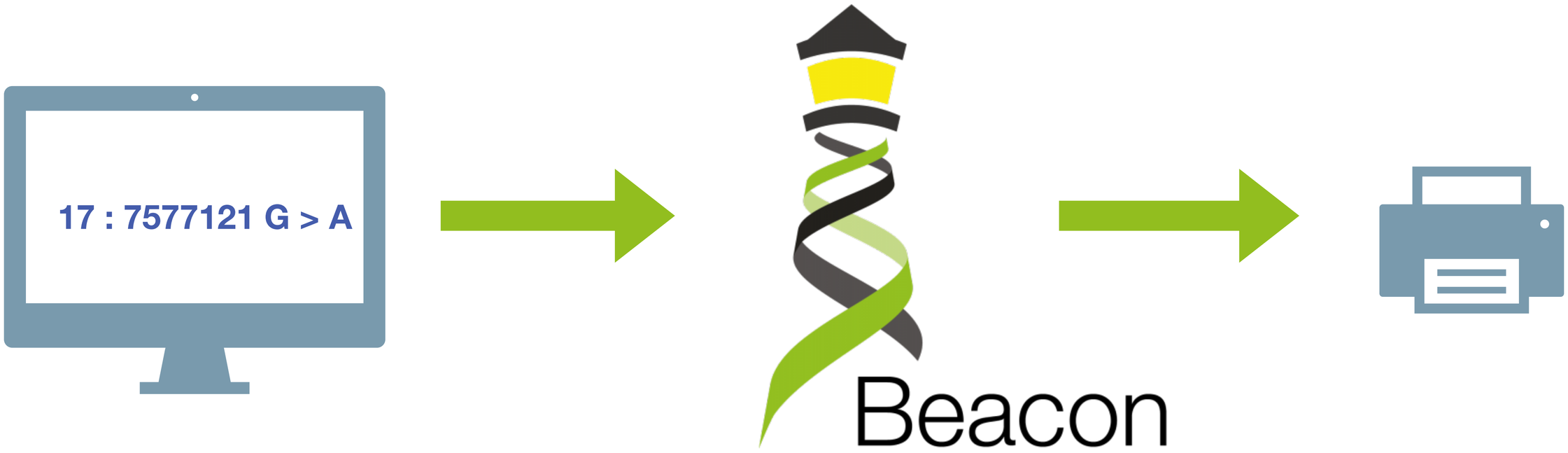
1. Provide ***frequencies of all alleles*** at that point
2. Ask for all alleles seen in a gene ***region*** (and more elaborate versions of this)
3. Other more complicated queries

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

Implementation

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

	Beacon 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*

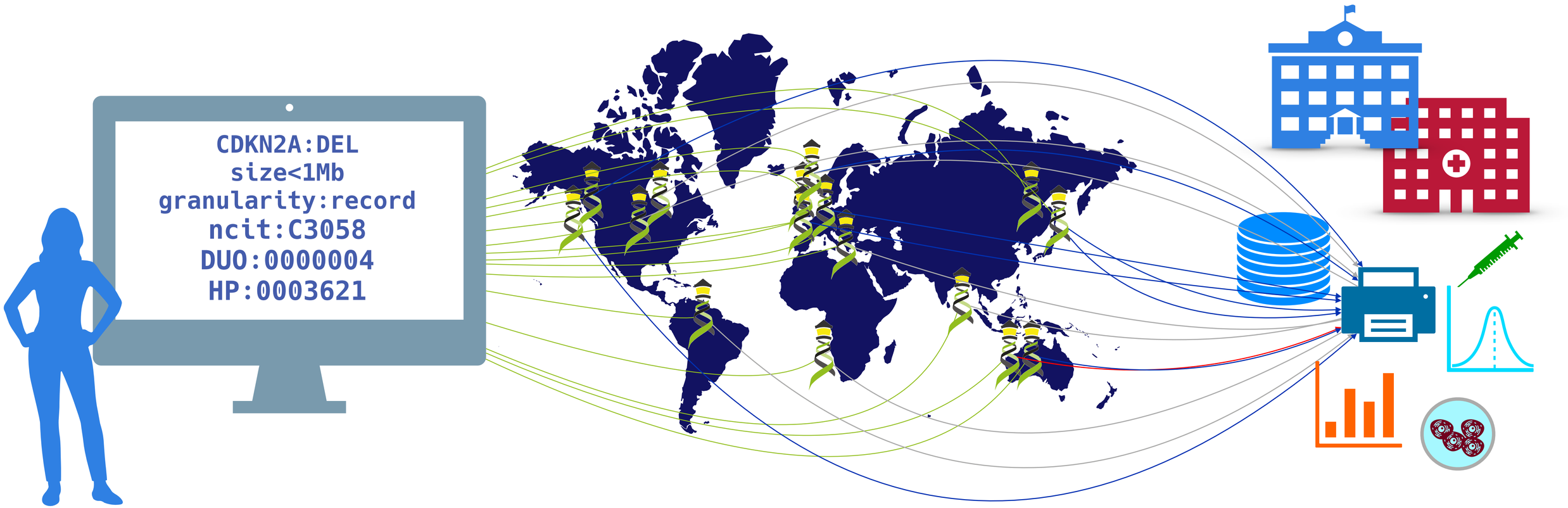


A ***Beacon*** answers a query for a specific genome variant against individual or aggregate genome collections

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

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

YES | **NO** | \0



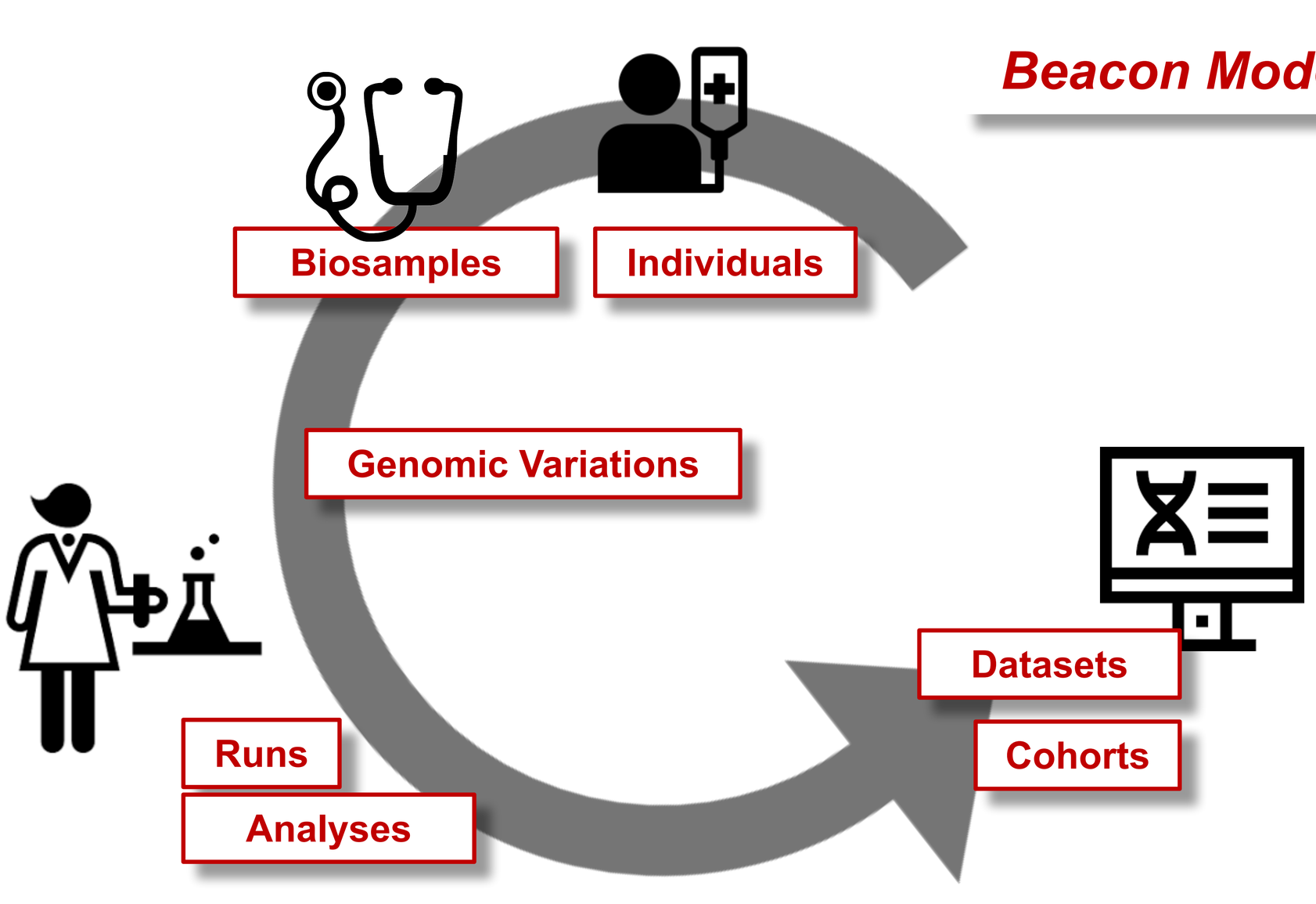
Can you provide data about focal deletions in CDKN2A in Glioblastomas from juvenile patients with unrestricted access?

Beacon API

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

Beacon v2

docs.genomebeacons.org

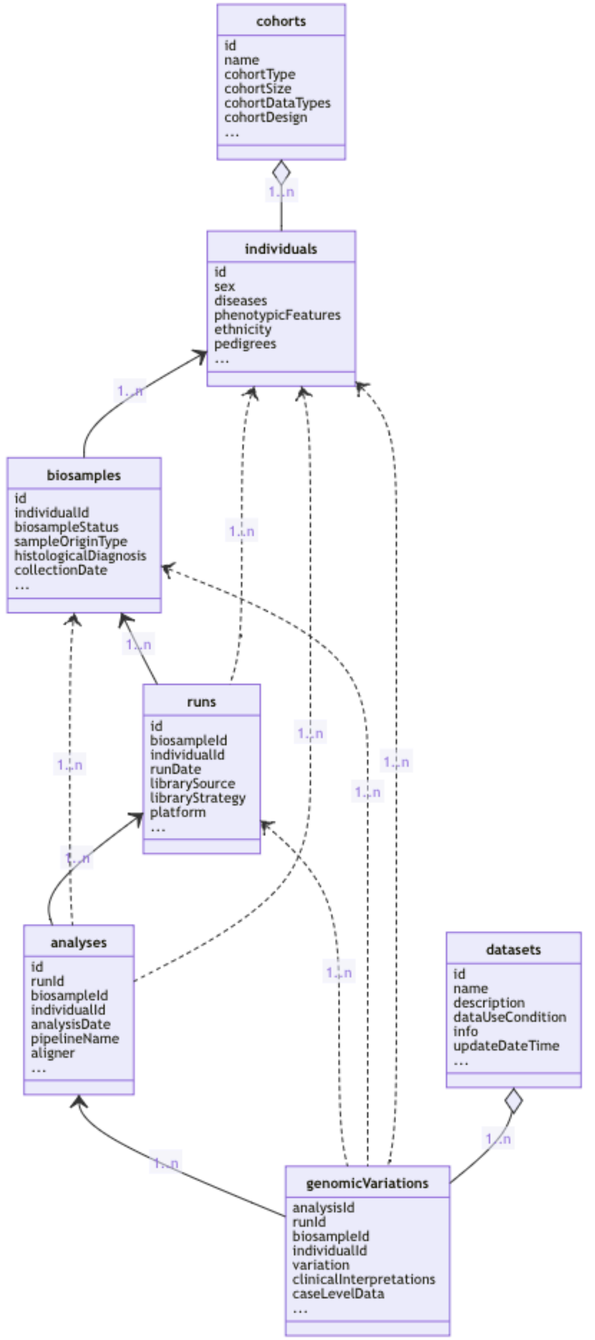


Beacon Model

Beacon Framework (protocol)

Beacon Default v2 Model

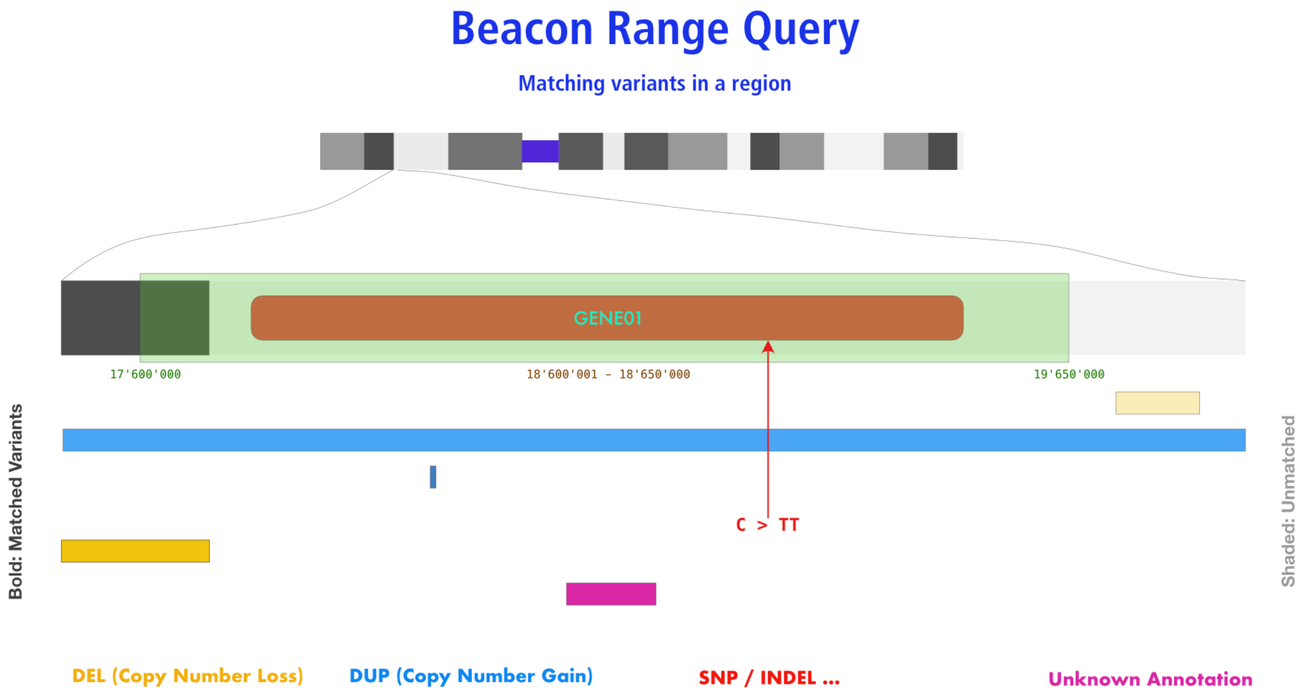
- The Beacon ***framework*** describes the overall structure of the API requests, responses, parameters, the common components, etc.
- Beacon ***models*** describe the set of concepts included in a Beacon, like individual or biosample, and also the relationships between them.
- Besides logical concepts,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

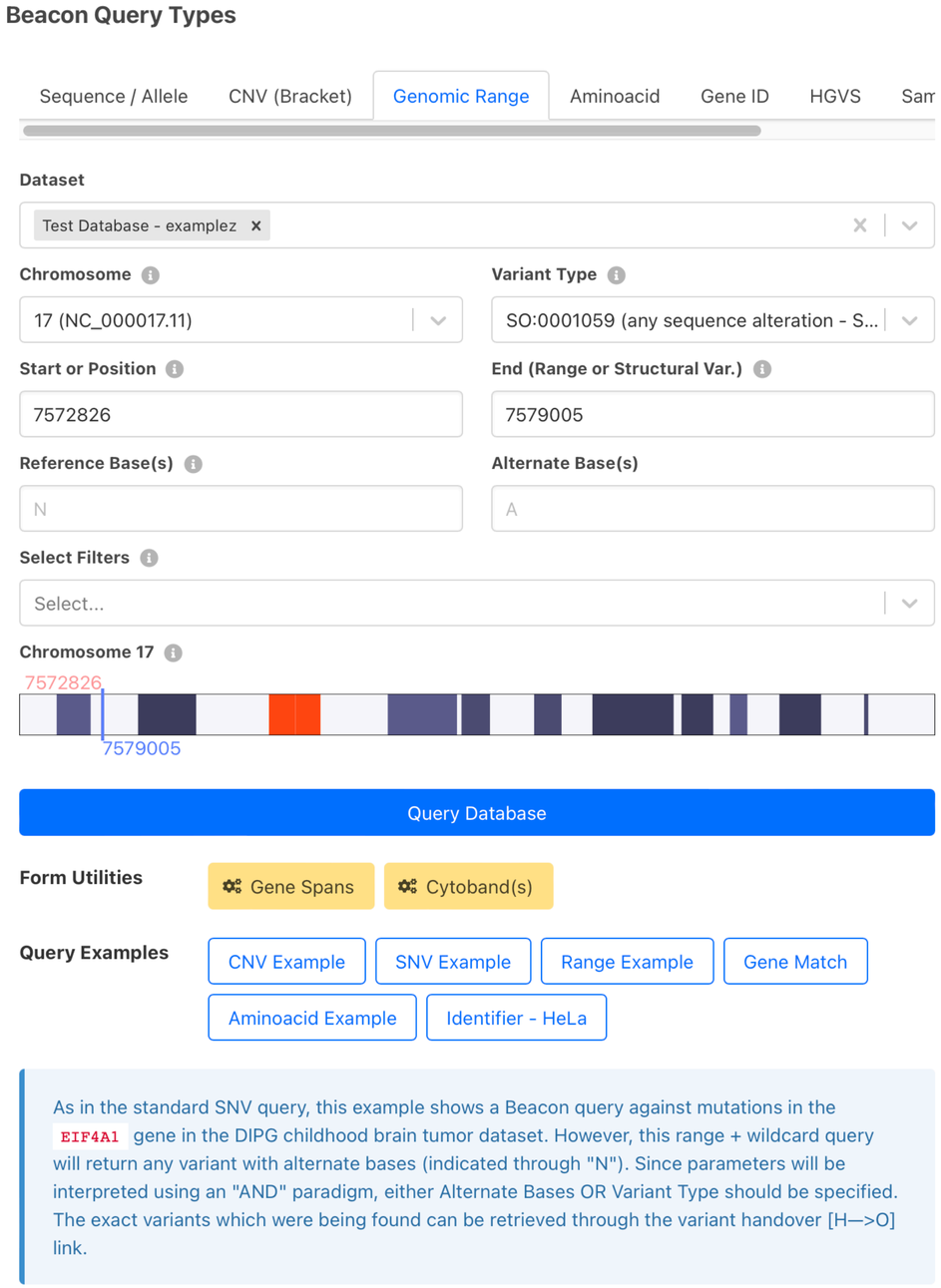


Variation Queries

Range ("anything goes") Request

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

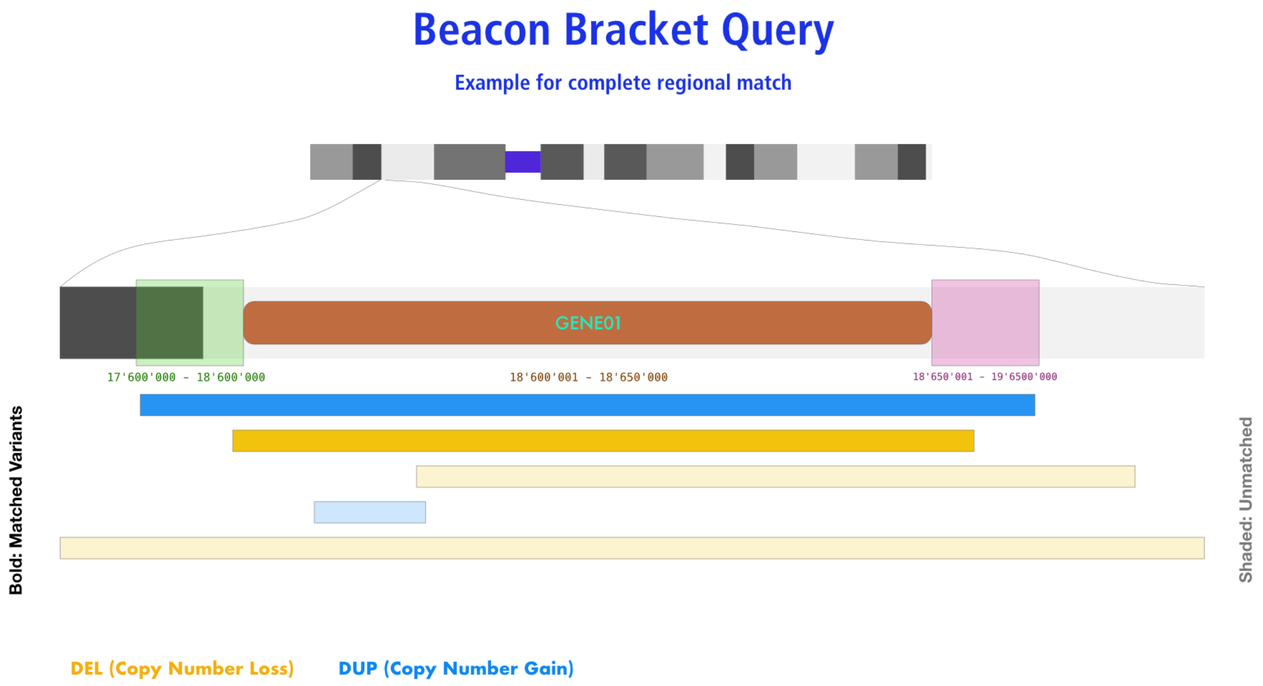


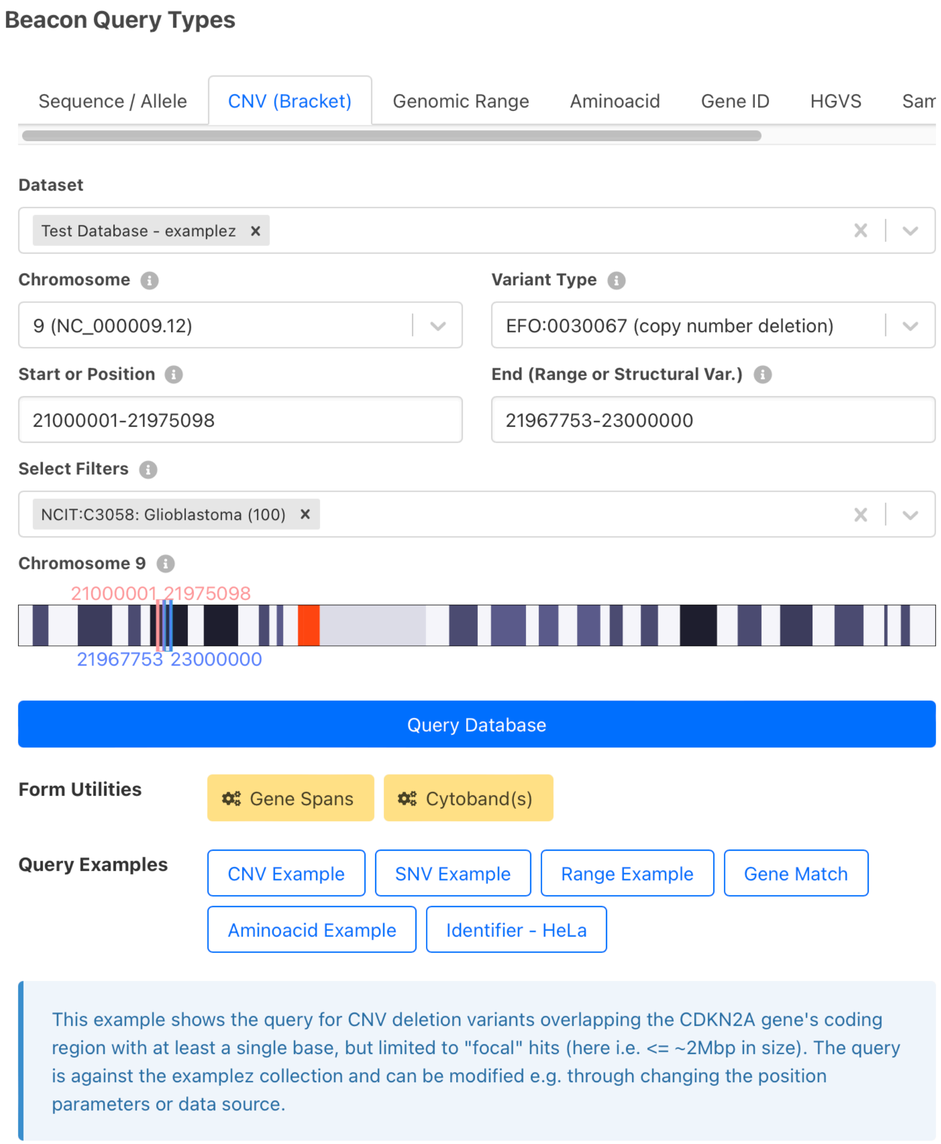


Variation Queries

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

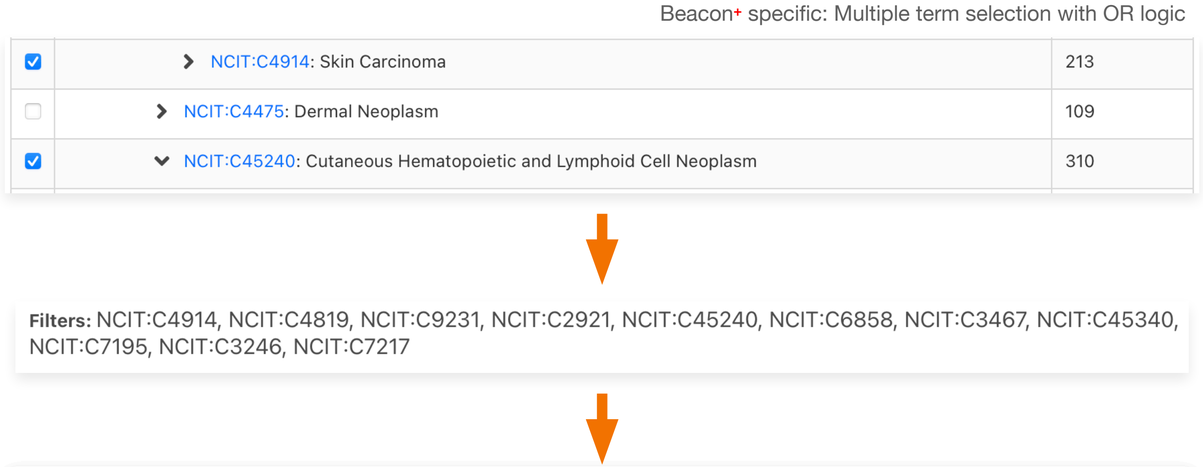
Beacon v2 Filters

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

ELIXIR Beacon Network

WP4 Beacon Network

Leads:

Jordi Rambla (ES)
Michael Baudis (CH)
Jaakko Leinonen (FI)

Objectives

This Work Package will

- Maintain an **operational Beacon Network service**, with ELIXIR lead participation, including a service level target for availability and user service and incident response.
- Deliver transparent and responsive governance structures that appropriately represent stakeholders, delineate strategy, and **react to user feedback**.
- Provide an approach to change management which ensures **ongoing development** of the service through other activities is integrated into the live service with minimum impact on existing users and dependencies.

Beacon as Core Genomics Discovery API

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

- originally GA4GH proposal for proof of feasibility and community engagement
  - then limited functionality API standard to answer "do you have this variant" queries
- rapid uptake w/ ~90 beacons in 2016; ~200 datasets
- ELIXIR support since 2016:
  - GA4GH Beacon v1 standard in 2018
  - ELIXIR Beacon v1 network implementation
  - **GA4GH Beacon v2 standard in 2022**

A ***Beacon*** answers a query for a specific genome variant against individual or aggregate genome collections
YES | **NO** | \0

- massive community interest and many worldwide initiatives to "beaconize" resources, both for wider **data discovery** and **data exchange**
- Beacon is an open standard supporting data *aggregation* but major **functionality boost** comes from **rk strategies** with a level of **harmonization**

ELIXIR Beacon Network Plans & Challenges

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

- ELIXIR Beacon Network has been *prototyped* in a previous IS
- This CoS will:
  - Contribute to the costs of running a service in production
  - Solidify the governing bodies
    - Strategic
    - Operational
  - While keeping the running costs to a minimum
  - Allowing for take over by other partners if ever necessary
- Challenges
  - Relies on partner willingness to run the service
    - Only a part of the operational costs are covered
    - Additional funds should be guaranteed by the partners
  - Funding for development or **software maintenance** is not included, but these activities are **clearly necessary**

cnag
European Genomic Data Infrastructure
progenetix
NBIS
GHGA
THE UNIVERSITY OF MELBOURNE
health RI
enabling data driven health & life sciences
cancercelllines
UNIVERSITY OF LEICESTER
BioData.pt
elixir

ELIXIR Beacon Network Summary

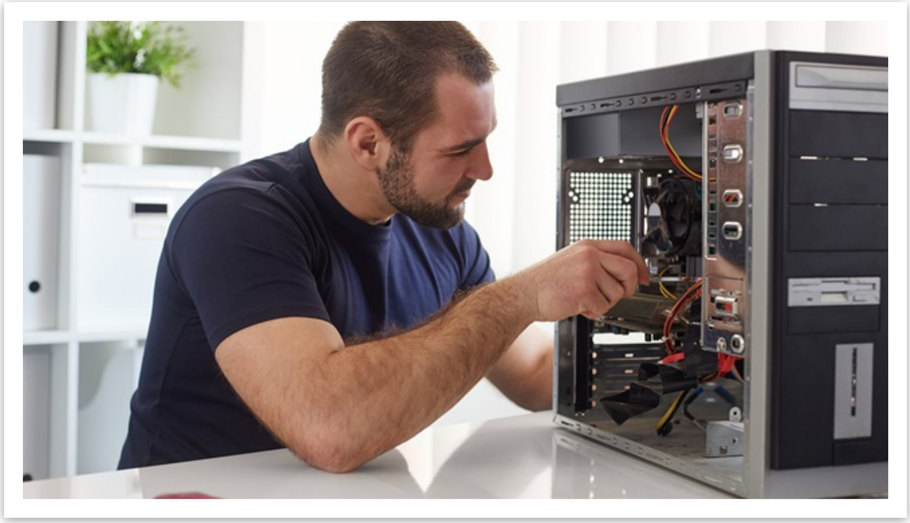
- ELIXIR Beacon Network in alignment with core stakeholders and external contributors to provide a working service with maximum capability profile
  - as GA4GH standard external input but also requirements to be accommodated
- importance of partner engagement since stakeholders (e.g. GDI …) currently lack ***live*** multi-node data scenarios but have a projected need
- high relevance of further ***protocol development*** to accommodate use cases **within** protocol specifications
  - RNAseq, pathogens, imaging …
  - massive opportunity to network people and data across domains through shared framework
- limited resources for "engage and ingest" work which is relevant for the global reach and - conversely - recognition and use of this ELIXIR standard and service

Beacon is a core ELIXIR asset. The Beacon Network provides a unique infrastructure service and is an attractor for cross-domain engagement

	"Aggregator Summaries Response"	Boolean/Count Response	Resultsets Response	"Aggregator Records Response"
Meta	requestedGranularity: boolean/count	requestedGranularity: boolean/count	requestedGranularity: record pagination	requestedGranularity: record pagination
Summary	exists numTotalResults **maximumGranularity** **minimumGranularity** ...	exists (numTotalResults)	exists numTotalResults ...	exists numTotalResults **maximumGranularity** **minimumGranularity** ...
Response	resultSets - id **beaconId** **infoUrl** exists resultsCount - id **beaconId** **infoUrl** exists - id **beaconId** ...	**Beacon v2 Aggregator / Network Responses**	resultSets - id setType exists resultsCount results [] - id setType exists resultsCount results [] - id setType ...	resultSets - id **beaconId** **infoUrl** setType exists resultsCount returnedGranularity ... results [] - id **beaconId** **infoUrl** exists ...

Beacon v2 deployment

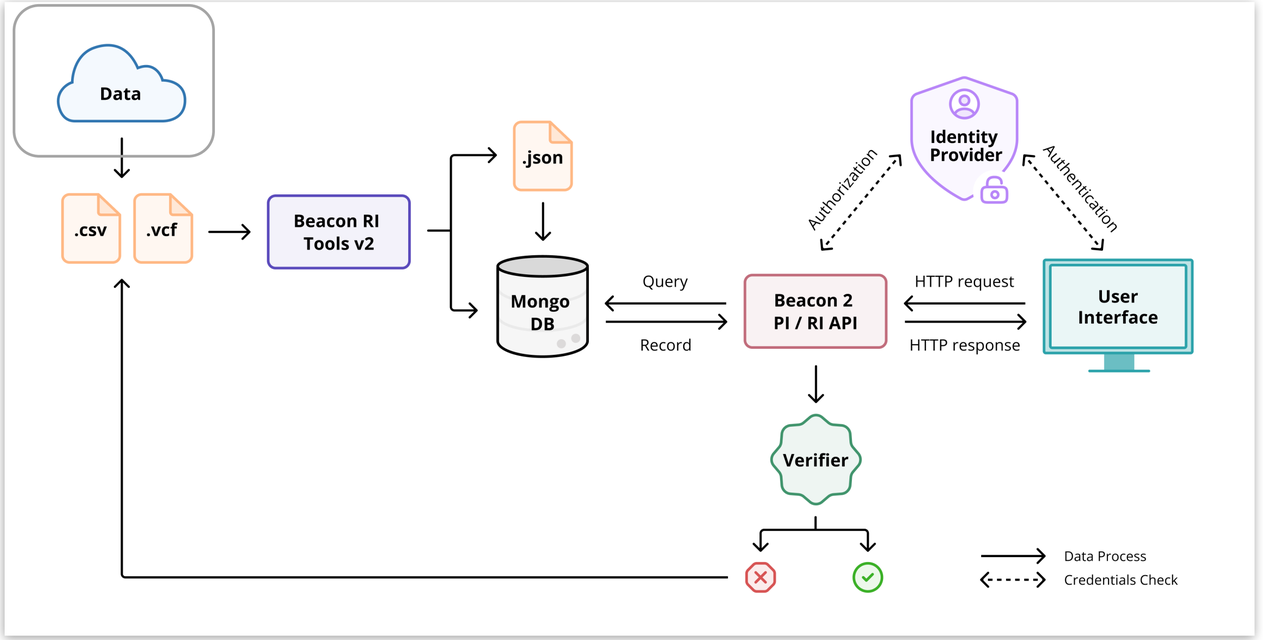
Build it yourself



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 Beacon+ Stack

progenetix

- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
  - `pubmed:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703…`
  - precomputed frequencies per collection informative e.g. in form autofills
- *querybuffer* stores id values of all entities matched by a query and provides the corresponding `accessid` for **handover** generation
- complete query aggregation; i.e. individual queries are run against the corresponding entities and ids are intersected
  - ➡ retrieval of any entity, e.g. all individuals which have queried variants analyzed on a given platform
  - ➡ allows multi-variant queries, i.e. all bio samples or individuals which had matches of all of the individual variant queries

React

APACHE HTTP SERVER PROJECT

python™

mongoDB

Entity collections

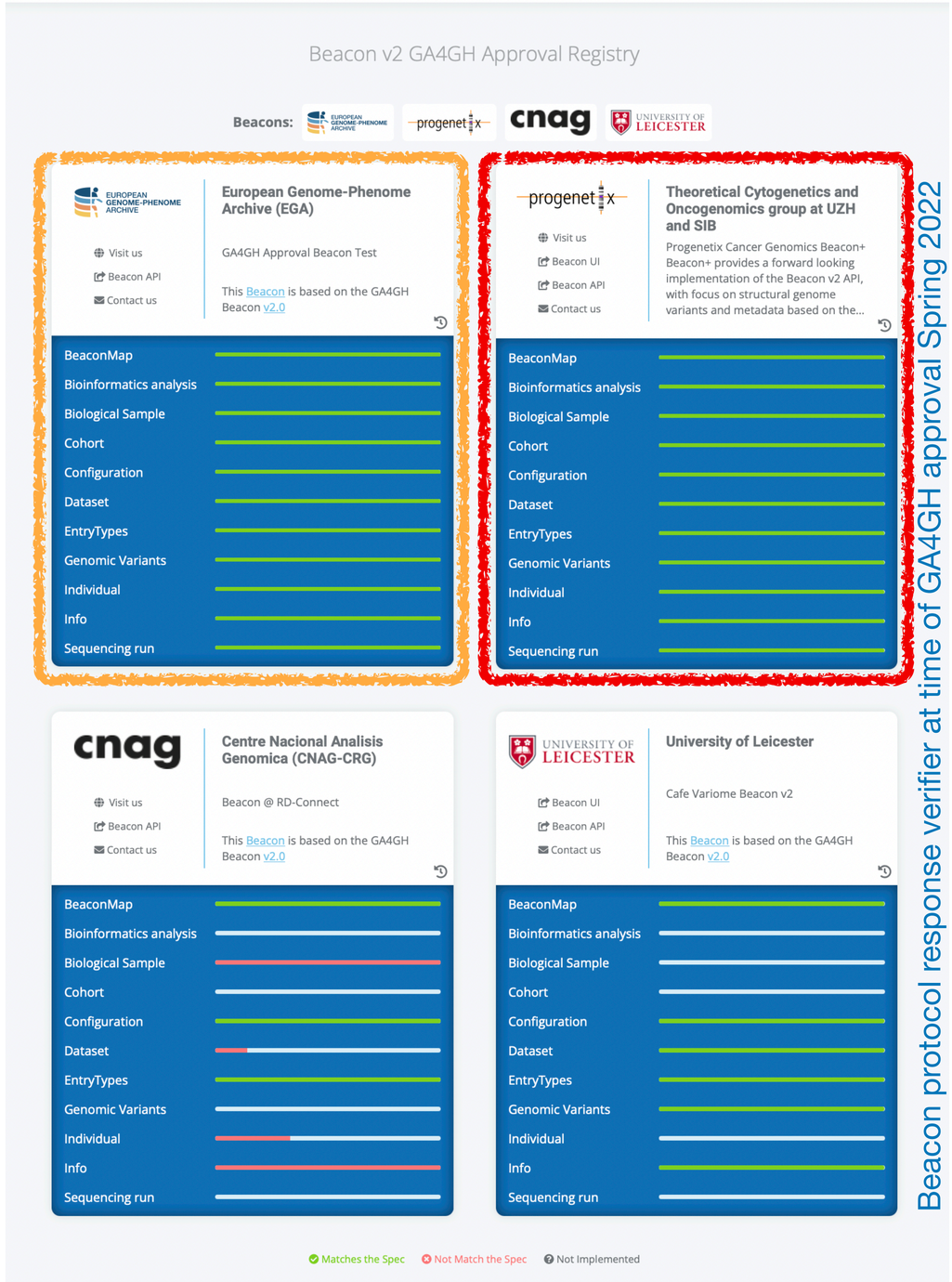
Utility collections

github.com/progenetix/bycon GitHub

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

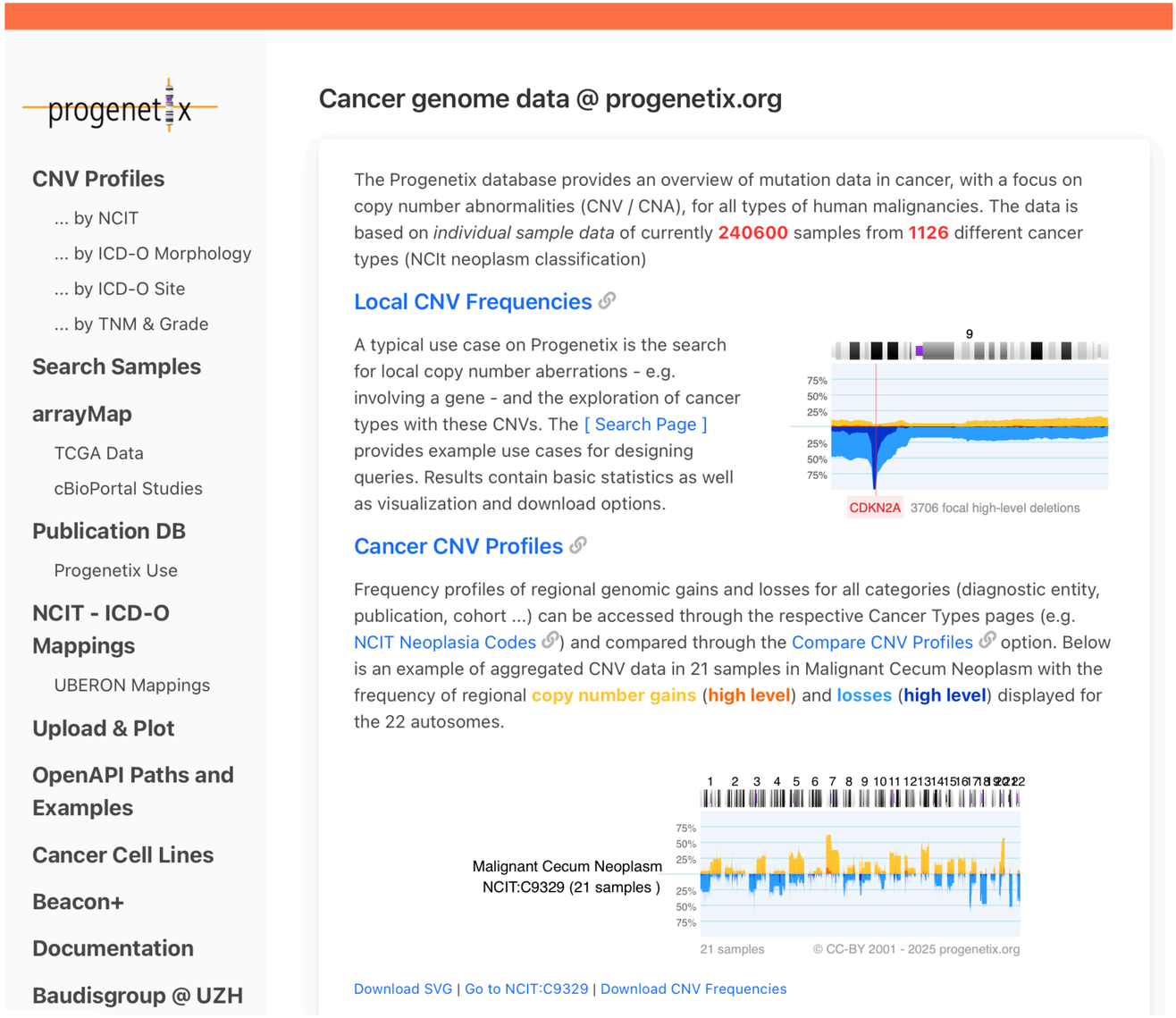
progenetix

elixir

progenetix.org

Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **240'000 cancer CNV profiles**
- SNV data for some series (e.g. TCGA)
- more than **1100 diagnostic types**
- inclusion of reference datasets (e.g. TCGA, GENIE, cBioPortal)
- 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

Beacon Security

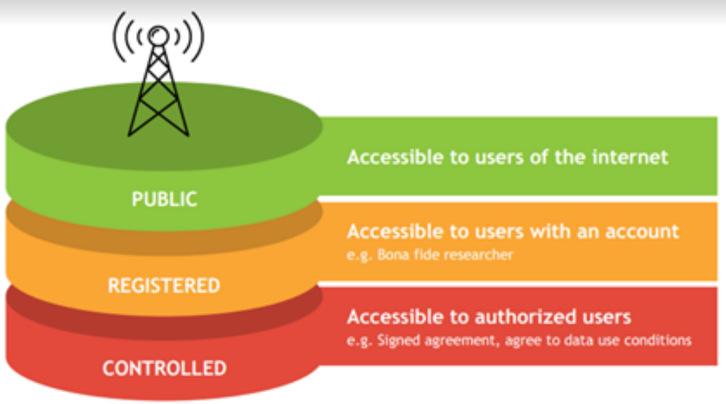
Making Beacons Biomedical - Beacon v2

- 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

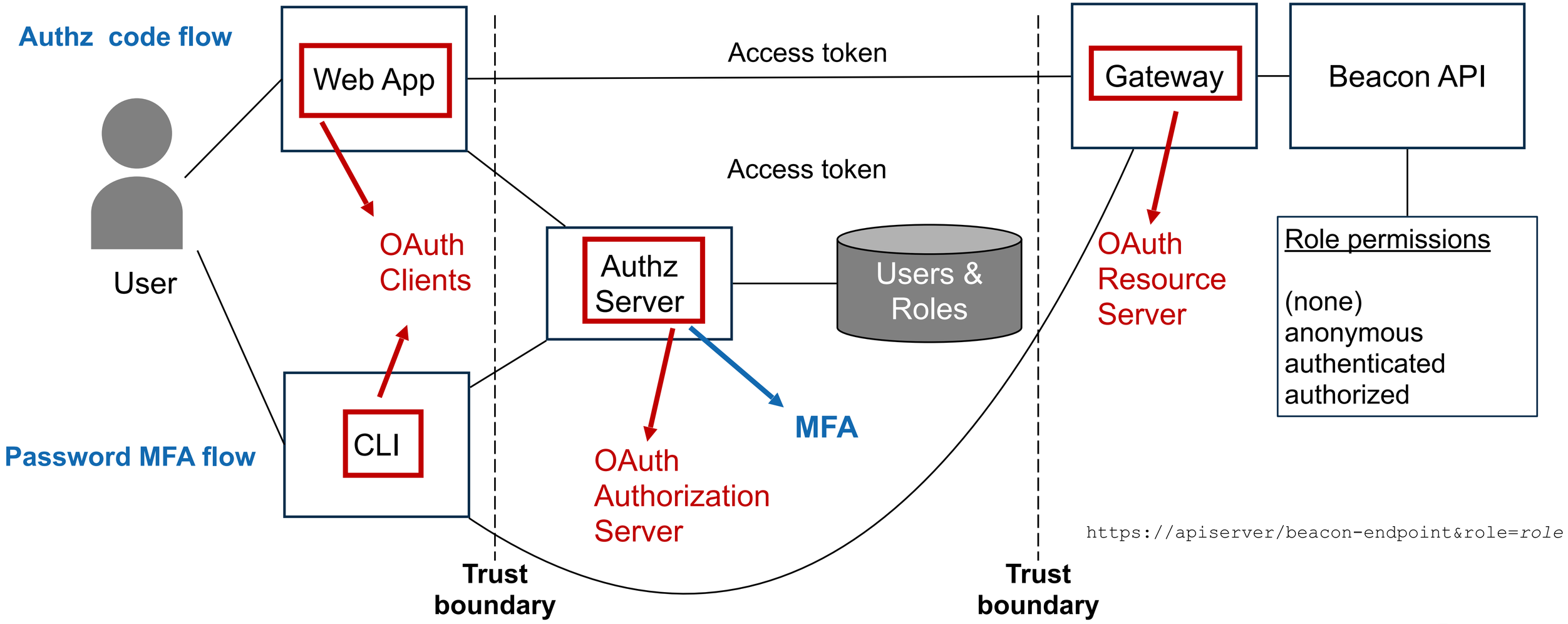
Security by Design ... if Implemented in the Environment



- 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

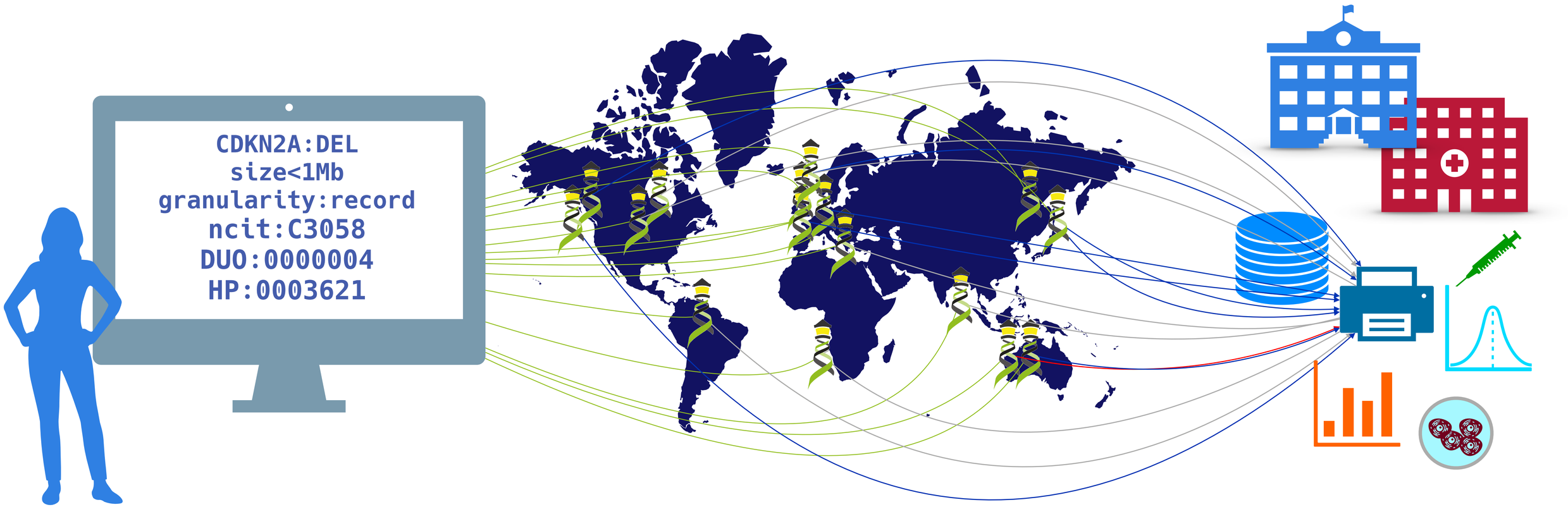
Common diseases

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

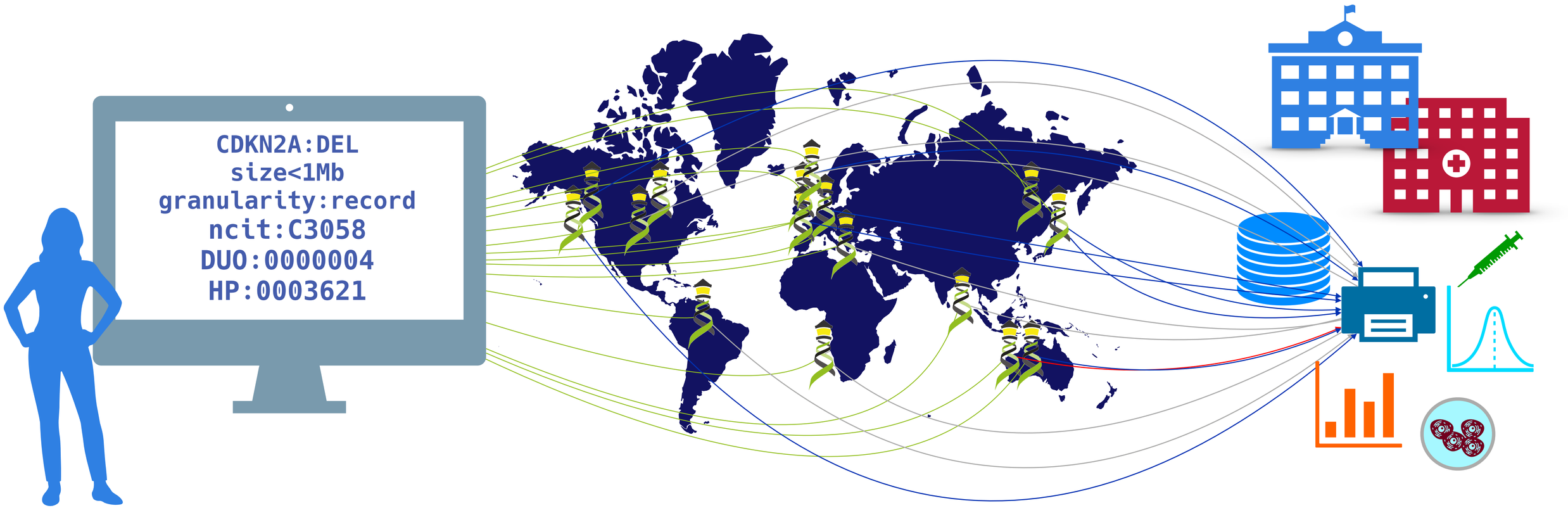
EUROPEAN GENOME-PHENOME ARCHIVE



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

Beacon API

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



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

But... Can you really do that with the Beacon default model? Let's explore...

Request Components

Deparsing the Beacon v2 Example

- query against genomic variations, no matter how they are stored
- copy number deletion, as indicated through the VCF symbolic allele `DEL` expression
- a combination of `genId` (server side gene data) OR
- a range query and `variantMaxLength`, or positional (`start`, `end`)
- a filter for the Glioblastoma diagnosis, as NCIT term NCIT:C3058
- as an HPO term for "juvenile" HP:0003621
- full data access as per DUO:0000004

Request Components

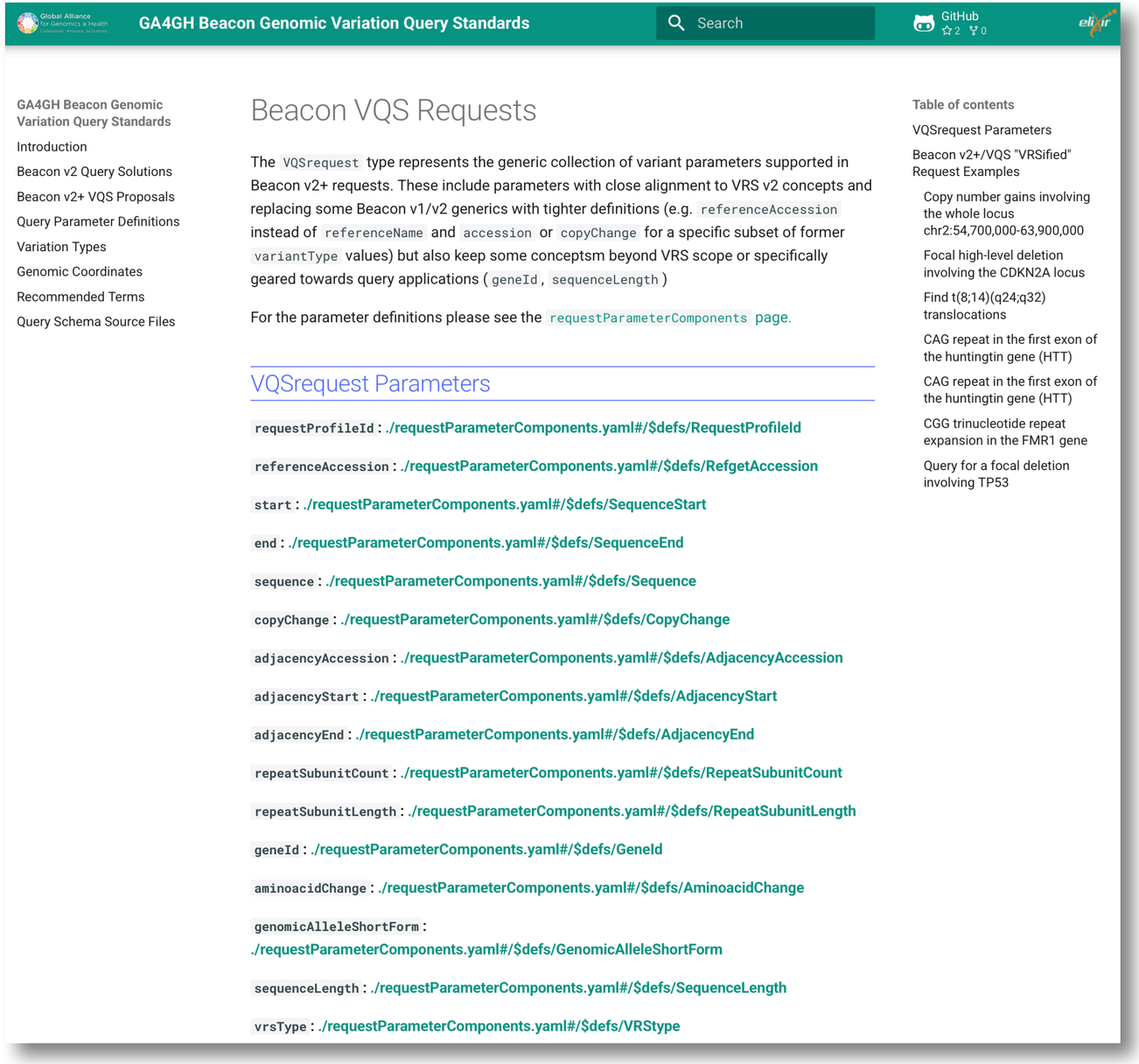
Deparsing the Beacon v2 Example

- Where can these parameters be applied in the current default model?
  - variant parameters `/g_variants`
  - histologicalDiagnosis `/biosamples`
  - dataUseConditions `/dataset`
  - ageOfOnset term `/individual`
- Such a request requires the application of filtering terms to other entities than the ones indicated by their entry type
- Variant parameters always need **aggregation**
- ... this does not prohibit simple implementations

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…
    - re-use of parameters where clear (e.g. **`sequenceLength`** instead of **`variantMinLength`** + **`variantMaxLength`**)



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

VQS - Variant Query Standard

VRS aligned typed queries

- Typed queries
  - ➡ query schemas with defined set of (required and optional) parameters
    - can be verified
    - profile ids can be advertised by beacons
- VRS aligned
  - ➡ explicit reference to VRS types
  - ➡ … but differ in (some) parameter use since query NE representation
- Expanding library
  - ➡ adjacency, repeats...

```yaml
VQScopyChangeRequest:
  description: |-
    A typical Beacon v2.n request for copy number vari
    approximate positions for CNV start and end regior
    `Range` type. The `copyChange` parameter indicates
    genomic copy number (pls. refer to the class defir
  type: object
  properties:
    requestProfile:
      const: VQScopyChangeRequest
    referenceAccession:
      $ref: "./requestParameterComponents.yaml#/$defs/
    startRange:
      $ref: "./requestParameterComponents.yaml#/$defs/
    endRange:
      $ref: "./requestParameterComponents.yaml#/$defs/
    copyChange:
      $ref: "./requestParameterComponents.yaml#/$defs/
    sequenceLength:
      $ref: "./requestParameterComponents.yaml#/$defs/
    vrsType:
      const: CopyNumberChange
  required:
    - requestProfile
    - referenceAccession
    - startRange
    - endRange
    - copyChange
```

```yaml
requestProfile: VQScopyChangeRequest
referenceAccession: refseq:NC_000002.12
start:
  - 21000001
  - 21975098
end:
  - 21967753
  - 23000000
copyChange: EFO:0020073
vrsType: CopyNumberChange
```

```yaml
requestProfile: VQSadjacencyRequest
referenceAccession: refseq:NC_000008.11
start: 116700000
end: 145138636
adjacencyAccession: refseq:NC_000014.9
adjacencyStart: 89300000
adjacencyEnd: 107043718
vrsType: Adjacency
```

```yaml
requestProfile: VQSsequenceRepeatRequest
geneId: HTT
repeatSubunitLength: 3
sequenceLength:
  - 105
  - 750
vrsType: ReferenceLengthExpression
```

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

```yaml
VQSadjacencyRequest:
  description: |-
    A typical Beacon v2.n request for sequence adjacency queries, e.g. for
    the retrieval of chromosomal translocation events or sequence fusions.

    TODO: In VRS v2 there is an implicit sequence directionality from the use
    of either start or end parameters for either side of the adjacency. This
    might be problematic on the query side where in many instances just the
    approximate position of the (fused) breakpoints maight be of interest.

    This might need additional clarification (e.g. use of `startRange` or
    `endRange`, `adjacencyStartRange` and `adjecencyEndRange` parameters to
    indicate direction dependent matching).
  type: object
  properties:
    requestProfile:
      const: VQSadjacencyRequest
    referenceAccession:
      $ref: "./requestParameterComponents.yaml#/$defs/RefgetAccession"
    sequenceRange:
      $ref: "./requestParameterComponents.yaml#/$defs/Range"
    adjacencyAccession:
      $ref: "./requestParameterComponents.yaml#/$defs/AdjacencyAccession"
    adjacencyRange:
      $ref: "./requestParameterComponents.yaml#/$defs/Range"
    vrsType:
      const: Adjacency
  required:
    - requestProfile
    - referenceAccession
    - sequenceRange
    - adjacencyAccession
    - adjacencyRange
    - vrsType
```

```yaml
examples:
  VQSadjacency_01:
    description: |-
      ### Find  t(8;14)(q24;q32) translocations

      #### Solution for `VQSrequest` using genomic ranges (`VQSadjacencyRequest`)

      This is a query for translocations between the MYC and IgH loci, where the
      breakpoints are loosely defined through there well known cytogenetic bands.
      The query here follows the VRS adjacency model. In contrast to the VRS
      representational model, here:


      - VRS uses an array of 2 genomic locations while Beacon names the location
        parameters individually (using "adjacency..." for the second partner)
      - VRS explicitely encodes directionality by using either `start` or `end`
        position parameters (integers or ranges) while this query example creates
        non-directional ranges on both sides since directionality might not be known,
        the storage system might not encode this or all options could be of interest

    request:
      requestProfile: VQSadjacencyRequest
      referenceAccession: refseq:NC_000008.11
      start: 116700000
      end: 145138636
      adjacencyAccession: refseq:NC_000014.9
      adjacencyStart: 89300000
      adjacencyEnd: 107043718
      vrsType: Adjacency
```

Variant Query Standard

VRS aligned typed queries - Open Questions…

- Parameter Zoo?
  - ➡ Should we be explicit in parameters themselves
    - **`startRange`** vs. **`start`** and “requires 2 pos. in context of profile”
- Level of VRSification?
  - ➡ Queries don’t necessarily correspond to VRS objects (polymorphic matches) - is the use of VRS vocabularies appropriate?

```yaml
VQScopyChangeRequest:
  description: |-
    A typical Beacon v2.n request for copy number variations (CNVs) queries
    approximate positions for CNV start and end regions through use of the
    `Range` type. The `copyChange` parameter indicates the relative change in
    genomic copy number (pls. refer to the class definition.)
  type: object
  properties:
    requestProfile:
      const: VQScopyChangeRequest
    referenceAccession:
      $ref: "./requestParameterComponents.yaml#/$defs/RefgetAc
    startRange:
      $ref: "./requestParameterComponents.yaml#/$defs/Range"
    endRange:
      $ref: "./requestParameterComponents.yaml#/$defs/Range"
    copyChange:
      $ref: "./requestParameterComponents.yaml#/$defs/CopyCha
    sequenceLength:
      $ref: "./requestParameterComponents.yaml#/$defs/Sequence
    vrsType:
      const: CopyNumberChange
  required:
    - requestProfile
    - referenceAccession
    - startRange
    - endRange
    - copyChange
```

```
requestProfile : VQScopyChangeRequest
referenceAccession : refseq:NC_000002.12
start :
  - 21000001
  - 21975098
end :
  - 21967753
  - 23000000
copyChange : EFO:0020073
vrsType : CopyNumberChange
```

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

Beacon v2 Variant Requests

Mix & Match?

- parameters allow positional and some identifier/classification based queries
  - ➡ genomic positions, sequences, variant types
  - ➡ no definition of allowed combinations so strange options possible…
    - genome assembly + versioned reference
  - ➡ patterns by convention/documentation
    - single start, end => range
    - 2 start, 2 end => bracket/CNV-style

g_variant Parameters

`assemblyId` : ./requestParameterComponents.yaml#/$defs/Assembly

`referenceName` : ./requestParameterComponents.yaml#/$defs/RefSeqId

`referenceBases` : ./requestParameterComponents.yaml#/$defs/ReferenceBases

`alternateBases` : ./requestParameterComponents.yaml#/$defs/AlternateBases

`variantType` : ./requestParameterComponents.yaml#/$defs/VariantType

`start` : ./requestParameterComponents.yaml#/$defs/Start

`end` : ./requestParameterComponents.yaml#/$defs/End

`geneId` : ./requestParameterComponents.yaml#/$defs/GeneId

`aminoacidChange` : ./requestParameterComponents.yaml#/$defs/AminoacidChange

`genomicAlleleShortForm` : ./requestParameterComponents.yaml#/$defs/GenomicAlleleShortForm

`variantMinLength` : ./requestParameterComponents.yaml#/$defs/VariantMinLength

`variantMaxLength` : ./requestParameterComponents.yaml#/$defs/VariantMaxLength

Aggregation / Summaries

Data Summaries

Proposal for representation of summary results inside Beacon responses

- summary results (probably less confusing than "aggregated ...") offer a way for beacons to
  - de-couple information about the resource's data content from samples and individuals
  - highlight relevant features and data
  - support performant front-ends/ dashboards
- Simple summaries can be defined through filtering terms with counts
- *distributions* for multi-value filters and custom types

```
SummaryResultsInstance:
  type: object
   properties:
     id:
       $ref: ../../requests/summaryTerms.yaml#/$defs/AggregationTerm
     label:
       type: string
     entity:
       description: >-
         Entity for which the observations were reported...
       type: string
     description:
      type: string
     count:
      type: integer
     distribution:
       $ref: "#/$defs/Distribution"

   required:
     - id
     - oneOf:
       - count
       - distribution

 Distribution:
   description: >-
     Distribution of results for this aggregation te
     ways to represent the distribution especially f
     average, median...).
   type: object
   properties:
     items:
      type: array
       items:
         $ref: "#/$defs/DistributionItem"
     values:
      type: array
     distincts:
      type: array
   required:
     - oneOf:
       - values
       - distincts
       - items
   additionalItems: True

 DistributionItem:
  type: object
```

```
examples:
  -
    - id: NCIT:C3359
      label: Rhabdomyosarcoma
      entity: biosample
      count: 42
    - id: NCIT:C3360
      label: Osteosarcoma
      entity: biosample
      count: 7
    - id: ageAtDiagnosis
      label: age at diagnosis
      entity: individual
      distribution:
        distincts:
          - P2Y4M
          - P3Y2M
          - P17Y9M
    - id: sex
      label: genotypic sex
      entity: individual
      distribution:
        items:
          - male: 13
          - female: 36
```

Data Summaries

Proposal for representation of summary results inside Beacon responses

- summary results (probably less confusing than "aggregated ...") offer a way for beacons to
  - de-couple information about the resource's data content from samples and individuals
  - highlight relevant features and data
  - support performant front-ends/ dashboards
- Simple summaries can be defined through filtering terms with counts
- *distributions* for multi-value filters and custom types

Example for **custom** summary: Binned CNV frequencies for pgx:icdom-81703 (Hepatocellular Carcinoma).

https://staging.progenetix.org/beacon/analyses/?filters=pgx:icdom-81703&summaryTerms=cnvfrequencies&requestedGranularity=count


```
"meta": {
        "apiVersion": "v2.2.0-beaconplus",
        "beaconId": "org.progenetix",
        "receivedRequestSummary": {
            "datasetIds": ["progenetix"],
            "filters": ["pgx:icdom-81703"],
...

"response": {
    "resultSets": [
        {
            "id": "progenetix",
            "responseEntityId": "analysis",
            "exists": true,
            "resultsCount": 2909,
            "summaryResults": [
                {
                    "entity": "analysis",
                    "id": "cnvfrequencies",
                    "count": 2909,
                    "description": "Binned CNV frequencies for 2909 matched analyses",
                    "distribution": {
                        "items": [
                            {
                                "cytobands": "1p36.33",
                                "end": 400000,
                                "gainFrequency": 2.613,
                                "gainHlfrequency": 0.378,
                                "id": "1p:000000000-000400000",
                                "lossFrequency": 6.188,
                                "lossHlfrequency": 0.206,
                                "start": 0
                            },
                            {
                                "cytobands": "1p36.33",
                                "end": 1400000,
                                "gainFrequency": 3.816,
                                "gainHlfrequency": 0.481,
                                "id": "1p:000400000-001400000",
                                "lossFrequency": 9.694,
                            ...
```

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

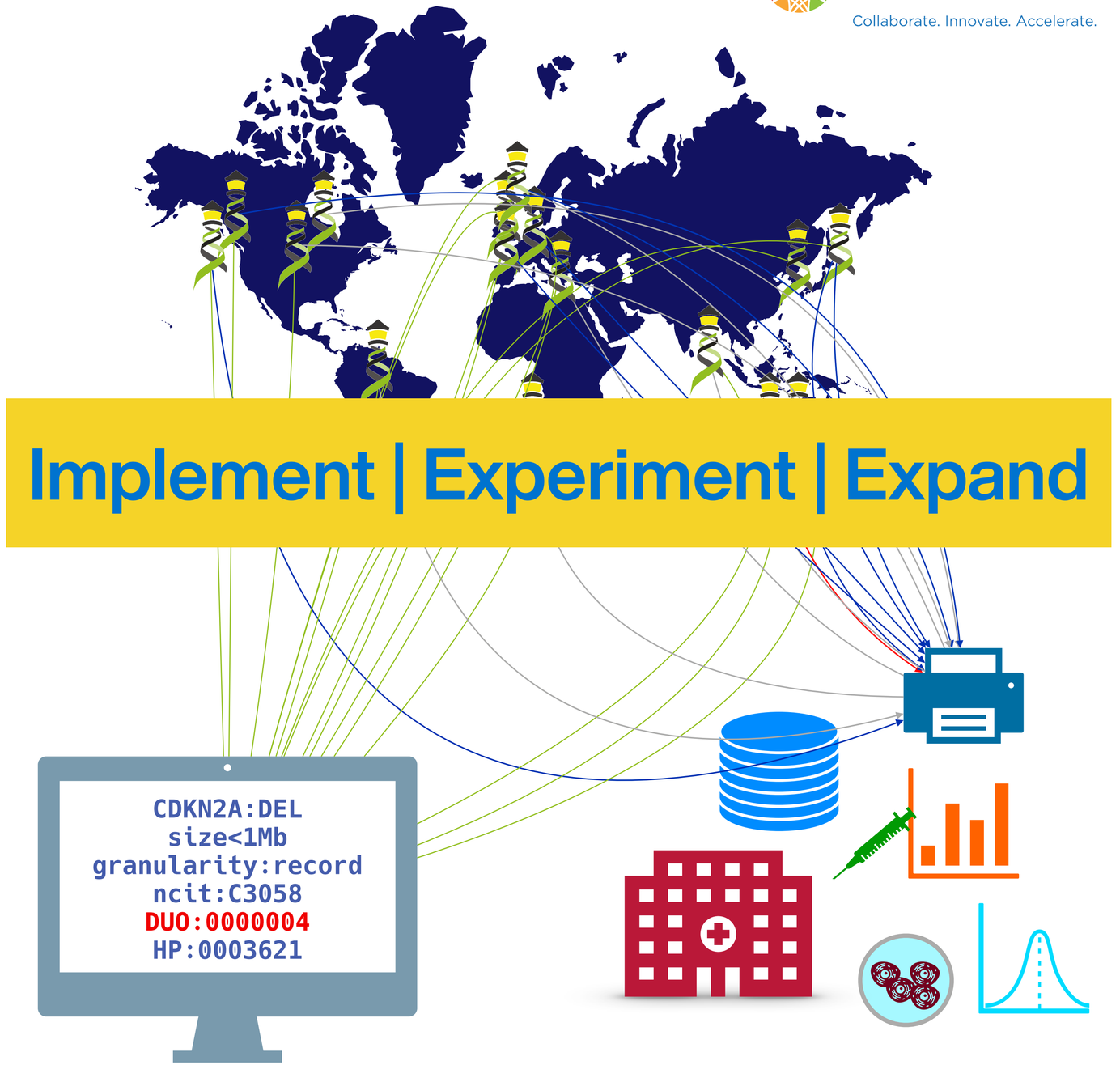
What Can You Do?

- 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

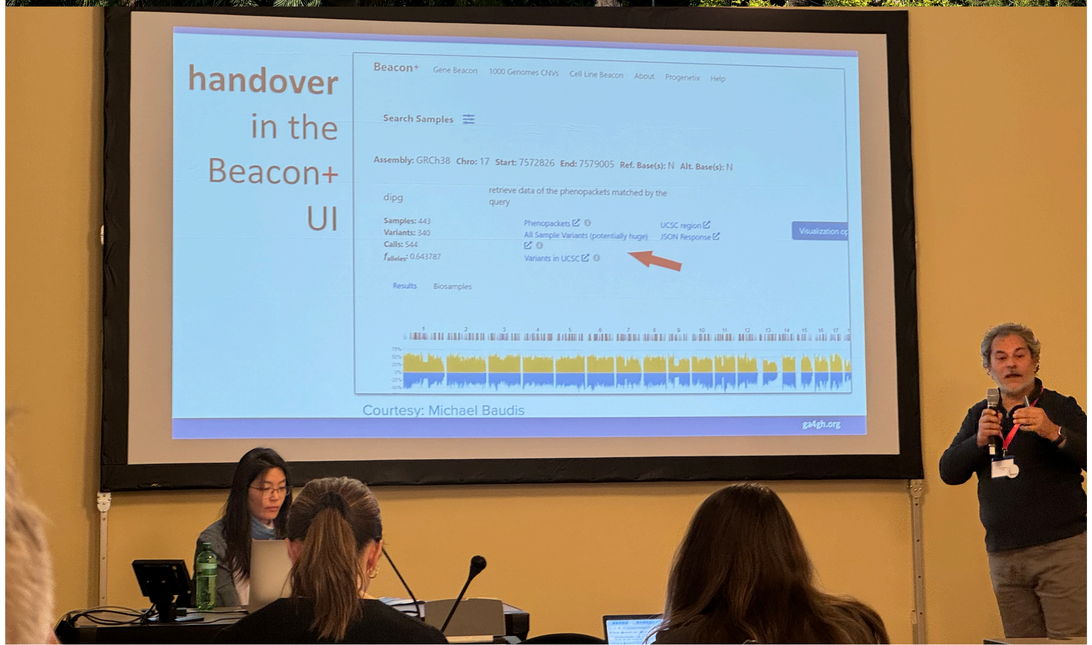
Registration Open Now

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