



# **ELIXIR & GA4GH - Beacons**

API , Implementations & Concepts, Engagement

## ***Beyond Basic Variant Discovery***



Global Alliance  
for Genomics & Health

# ELIXIR Beacon Project

- Driver project on GA4GH roadmap
- aligns with Discovery Work Stream
- strong impact on GA4GH developments as a concrete, funded project

The screenshot shows the 'Driver Projects' section of the GA4GH website. It features a red circular icon with a white rocket ship. Below it, the text 'Driver Projects' is displayed. A detailed description follows: 'GA4GH Driver Projects are real-world genomic data initiatives that help guide our development efforts and pilot our tools. Stakeholders around the globe advocate, mandate, implement, and use our frameworks and standards in local contexts.' To the right, there is a box for the 'ELIXIR Beacon' project, which includes the ELIXIR logo, the URL [www.elixir-europe.org](http://www.elixir-europe.org), the location 'Europe', and the 'Champions' names: Serena Scollen, Ilkka Lappalainen, Michael Baudis.

## Beacon forward



- structural variations** (DUP, DEL) in addition to SNV
- ... more structural queries (translocations/fusions...)
- filters** (phenotypes, datasets, metadata...)
- layered authentication system using **ELIXIR AAI**
- quantitative responses
- descriptive responses
- Beacon queries as entry for **data handover** (outside Beacon protocol)
- Ubiquitous **deployment** (e.g. throughout ELIXIR network)





This example shows the query for CNV deletion variants overlapping the CDKN2A gene's coding region with at least a single base, but limited to "focal" hits (here i.e. <= ~4Mbp in size). The query is against the arrayMap collection and can be modified e.g. through changing the position parameters or data source.

CNV Example   SNV Range Example   SNV Example   BND Example

**Dataset\*** arraymap

**Reference name\*** 9

**Genome Assembly\*** GRCh38 / hg38

**(structural) variantType** DEL (Deletion)

**Gene Coordinates** CDKN2A

**Start min Position\*** 18000000

**Start max Position** 21975098

**End min Position** 21967753

**End max Position** 26000000

**Bio-ontology** no selection  
icdom-94403: Glioblastoma, NOS  
icdom-94423: Gliosarcoma (9)  
icdot-C00-C14+: Lip, oral cavity  
icdot-C01+: Base of tongue (41)  
icdot-C01.9: Base of tongue, NO

**Biosample Type** neoplastic sample

**Beacon Query**

Response

There were no previous searches yet. Please, perform a query by using the form above.

# Beacon API 2019

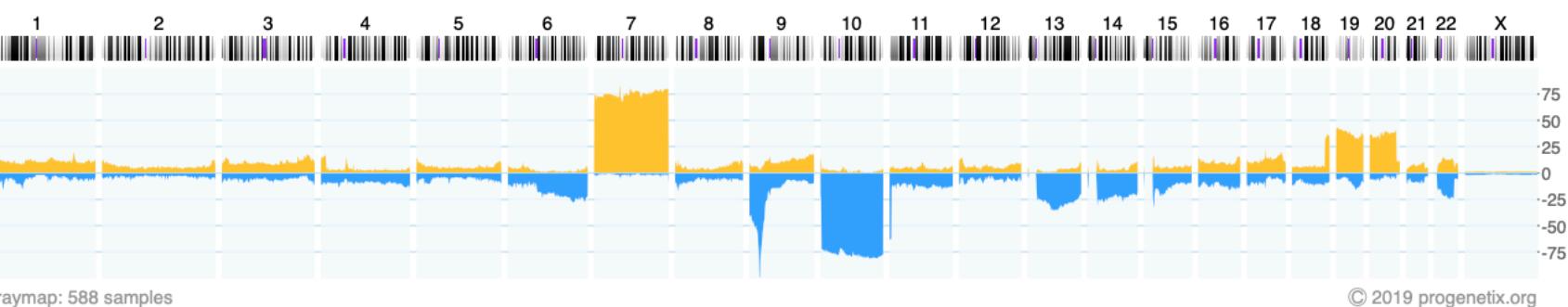
- ✓ Handover
- ✓ Filters
- ✓ Range Queries

Response								
Dataset	Assembly	Chro	Position Start Range	Ref Alt Type	Bio Query	Variants Calls Samples	f <sub>alleles</sub>	Response Context
arraymap	GRCh38	9	18000000 - 21975098 21967753 - 26000000	*	icdom-94403 EFO:0009656	588 588 588	0.0081	[H>O] Biosamples [H>O] Callsets Variants [H>O] CNV Histogram [H>O] Progenetix Interface [H>O] Variants

```

variant_type: "DEL"
callset_id: "pgxcs:GSE13021:GSM326195"
variantset_id: "AM_VS_GRCH38"
biosample_id: "PGX_AM_BS_GSM326195"
end:
  0: 21968713
info:
  cnv_value: -0.3552
  cnv_length: 194772
start:
  0: 21773941
digest: "9:21773941-21968713:DEL"
reference_name: "9"

```



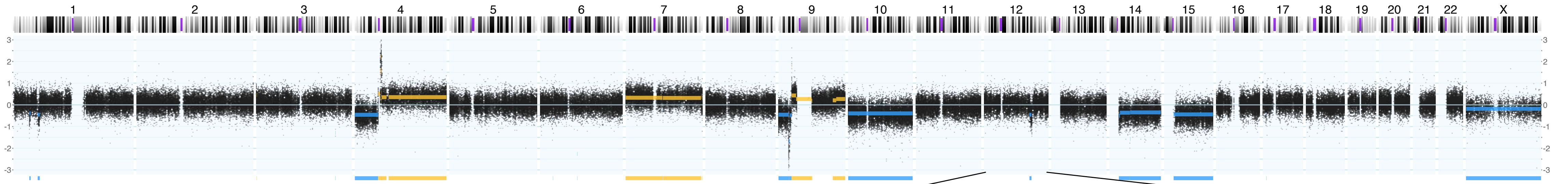
```

individual_id: "PGX_IND_GSM326195"
provenance:
  material:
    type:
      label: "neoplastic sample"
      id: "EFO:0009656"
      description: "glioblastoma [xenograft]"
    geo:
      city: "Washington"
      longitude: -89.41
      label: "Washington, United States"
      precision: "city"
      latitude: 40.7
      country: "United States"
    age_at_collection: {}
  biocharacteristics:
    0:
      description: "glioblastoma [xenograft]"
      type:
        id: "icdot-C71.9"
        label: "Brain, NOS"
    1:
      description: "glioblastoma [xenograft]"
      type:
        label: "Glioblastoma, NOS"
        id: "icdom-94403"
    2:
      type:
        label: "Glioblastoma"
        id: "ncit:C3058"
        description: "glioblastoma [xenograft]"
  data_use_conditions:
    id: "DUO:000004"
    label: "no restriction"
  external_references:
    0:
      relation: "denotes"
      type:
        id: "geo:GSE13021"
        label: ""
        description: "geo:gse"
    1:
    2:
    3:
      id: "PGX_AM_BS_GSM326195"
      description: "glioblastoma [xenograft]"
      info: {}
      project_id: "GSE13021"

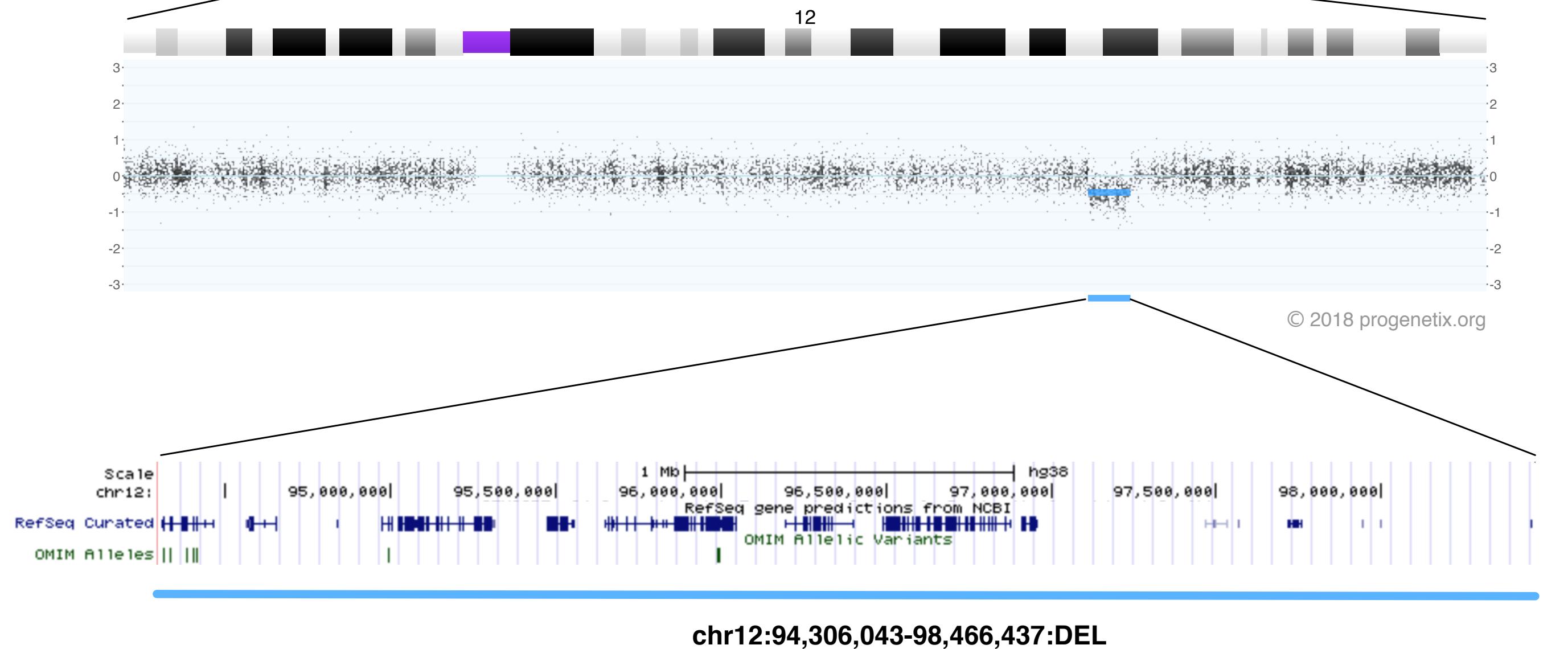
```



GSM491153



- Beacon+ **range queries** allow the definition of a genome region of interest, containing a specified variant (or other mappable feature)
- “fuzzy” matching of region ends is essential for features without base specific positions
- current Beacon implementation addresses CNV (<DUP>, <DEL>), as are specified in VCF && GA4GH variant schema



start\_min: 94,000,000  
start\_max: 94,500,000  
variant\_type: “**BND**”

reference\_name: “9”  
variant\_type: “**DEL**”



end\_min: 98,200,000  
end\_max: 98,700,000  
variant\_type: “**BND**”



# Filters

- adding non-variant query elements to aBeacon query
- using a single type of attribute ("filters")
- mapping filters to the corresponding data elements based on their prefixes
- heavily relying on ontology classes expressed as CURIEs
- allowing for "private classes/filters", i.e. non-standard extensions for richer queries in private context

```
filter_prefix_mappings:  
  icdom:  
    parameter: 'biocharacteristics.type.id'  
  icdot:  
    parameter: 'biocharacteristics.type.id'  
  ncit:  
    parameter: 'biocharacteristics.type.id'  
  HPO:  
    parameter: 'biocharacteristics.type.id'  
  pubmed:  
    parameter: 'external_references.type.id'  
  geo:  
    parameter: 'external_references.type.id'  
  snmi:  
    parameter: 'external_references.type.id'  
  cellosaurus:  
    parameter: 'external_references.type.id'  
  tcga:  
    parameter: 'external_references.type.id'  
  EFO:  
    parameter: 'provenance.material.type.id'  
  city:  
    parameter: 'provenance.geo.city'  
    remove_prefix: true  
  country:  
    parameter: 'provenance.geo.country'  
    remove_prefix: true  
  wes:  
    parameter: 'counts.wes'  
    remove_prefix: true  
  wgs:  
    parameter: 'counts.wgs'  
    remove_prefix: true  
  ccgh:  
    parameter: 'counts.ccgh'  
    remove_prefix: true  
  acgh:  
    parameter: 'counts.acgh'  
    remove_prefix: true  
  genomes:  
    parameter: 'counts.genomes'  
    remove_prefix: true  
  ngs:  
    parameter: 'counts.ngs'  
    remove_prefix: true
```

# Beacon *Flavours*

---



- Standard Beacon implementations report on allelic variants, based on querying collections of aggregate or sample mapped genomic variant data
- This principle is open to modular extensions
  - Query: Phenotypic filters, selected variant types, data use conditions...
  - Response: Handover for data delivery, variant details from range queries, statistics in response...
- Flavours describes the concept of adapting and applying Beacon protocol principles to other data domains
  - ***Evidence*** Beacons
  - ***Proteomics*** Beacons
  - ***Plant*** Beacons

# Beacon *Flavours* <> Implementations

---



- Beacon ***Flavours*** refer to principle differences in the Beacons' query and response structures
  - reporting on knowledge resources ("Evidence Beacon") with a semantically rich payload (variant annotation, clinical evidences...)
  - non-human applications ("Plant Beacon")
  - non-genomic data
- This is different from use-case specific Beacon ***Implementation types*** (e.g. "Clinical Beacon")
  - use-case specific extend of data delivery, handover types, case mapping requirements...

# GA4GH :: Discovery ∩ ELIXIR Beacon



- Representation in GA4GH
- Coordination with GA4GH Work Streams for standards
- Interaction (use cases, requirements) with other Driver Projects

→ ***concepts & interactions***

- Development of the Beacon API
- Core Beacon API and implementations for ELIXIR stakeholders
- Demonstrators and use cases

→ ***implementation & delivery***



**Global Alliance**  
for Genomics & Health





# GA4GH :: Discovery

A Work Stream of The Global Alliance for Genomics and Health

*We build standards for federated, secured networks of data and services, forming an “Internet of Genomics”, and asking meaningful questions across it.*

- Marc Fiume
  - Discovery Networks
  - Search API / Data Discovery

- Michael Baudis

- Beacon



- SchemaBlocks {S}[B]



## GA4GH :: Discovery

News

Participants

Examples, Guides & FAQ

Meeting minutes

Contacts

Workstream Products

Beacon

Discovery Networks

GA4GH SchemaBlocks

Search API

Related Sites

ELIXIR beacon

GA4GH

Beacon<sup>+</sup>

beacon-network.org

GA4GH SchemaBlocks

Github Projects

Discovery

ELIXIR Beacon

SchemaBlocks

Tags



## GA4GH Discovery Work Stream

Welcome to the homepage for the GA4GH Discovery Work Stream. We build standards for federated, secured networks of data and services, forming an “Internet of Genomics”, and asking meaningful questions across it.

The Discovery Work Stream is lead by Marc Fiume and Michael Baudis. For details on how this Work Stream operates please read the [Discovery Work Stream Organizational Structure & Vision document](#).

This group meets at a high-level monthly. [Meeting minutes are available to view here](#). In addition, the sub-groups listed below meet on their own schedules.

Participation in these groups require participants to adhere to the [GA4GH Standards for Professional Conduct](#).

For more information on GA4GH, please visit the [GA4GH Website](#).

### Products

Product development in GA4GH follows a process outlined in a [GA4GH Product Approval Process Guide, in draft](#). Products developed by the work stream undergo an initial investigation phase, followed by a formal Proposed Product Phase, in which most of the work is done, followed by an formal Approval Phase during which the products gain GA4GH Approval. The formal steps require the approval of the Work Stream leads.

The following products are currently under development for this Work Stream.

### Beacon API

A **Beacon** is a federated, web-accessible service that can be queried for information about a specific genomic variant, e.g. a single nucleotide polymorphism (SNP/SNV) or a copy number variation (CNV), and reports about its existence in the queried resources. Future versions of the Beacon protocol will support different usage scenarios and offer the opportunity to link to the matched data using e.g. a [handover](#) scenario.

The Beacon API specification is now coordinated through the [ELIXIR Beacon project](#) and accessible there or directly through its [repository](#).



### Discovery Search API

The Discovery Search API aims at developing a component based approach towards the implementation of interfaces for genomic data and related information, for instance for global, federated data sharing through the querying, and subsequent optional processing of the results in a cloud environment. The in-development specification for the *Search API* can be [accessed here](#).

### Discovery Networks API



The BeaconNetwork was the first successful implementation of an open, federated API for world-wide querying of genome resources. Current and future developments target especially the integration of user authentication for different access levels, extensions to the query language as provided through the emerging Beacon API and the evaluation of different topologies, especially with respect to security concerns.

# GA4GH {S}[B]

- “cross-workstreams, cross-drivers” initiative to document GA4GH object standards and prototypes, data formats and semantics
- launched in December 2018
- 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 :: SchemaBlocks

An Initiative by Members of the Global Alliance for Genomics and Health

### About {S}[B]

#### News

#### Participants

#### Data Formats

#### Data Schemas

#### Examples, Guides & FAQ

#### Meeting minutes

#### Contacts

#### Related Sites

[GA4GH::Discovery](#)

[GA4GH::CLP](#)

[GA4GH::GKS](#)

[ELIXIR Beacon](#)

[Phenopackets](#)

[GA4GH](#)

[Beacon+](#)

#### Github Projects

[SchemaBlocks](#)

[ELIXIR Beacon](#)

#### Tags



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

## GA4GH Data Model

### Recommendation (DRAFT)

The GA4GH data model recommends the use of a default object hierarchy in standard and product design processes. While it reflects concepts from the original [GA4GH schema](#), it provides mostly a structural guideline for API and data store design, but is not thought to provide a set of absolute implementation requirements.

#### Contributors

- [@mcourtot](#)
- [@mbaudis](#)

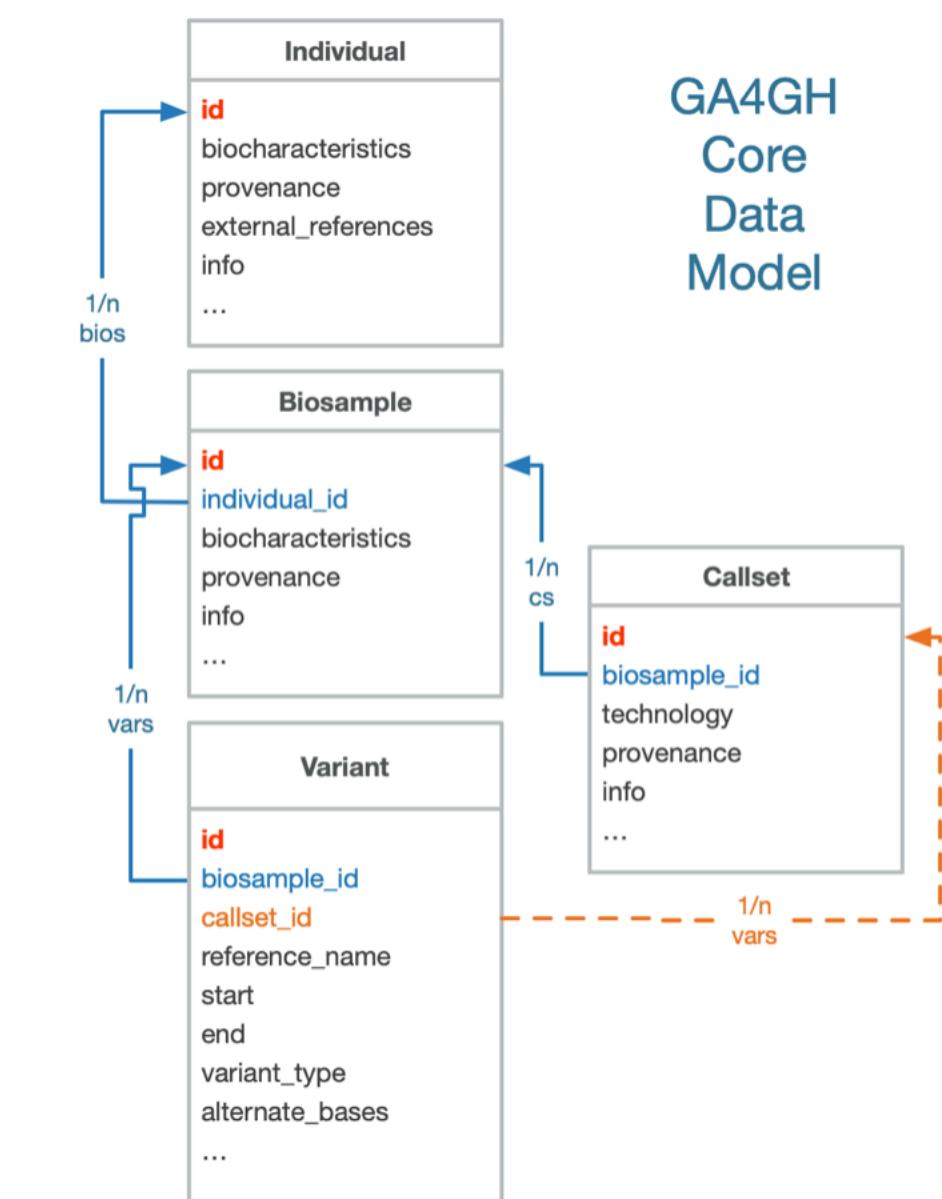
#### Summary

The GA4GH data model for genomics recommends the use of a principle object hierarchy, consisting of

- **variant**
  - a single molecular observation, e.g. a genomic variant observed in the analysis of the DNA from a biosample
- **callset**
  - the entirety of all variants, observed in a single experiment on a single sample
  - a **callset** can be compared to a data column in a **VCF** variant annotation file
  - **callset** has an optional position in the object hierarchy, since **variants** describe biological observations in a biosample
- **biosample**
  - a reference to a physical biological specimen on which analyses are performed
- **individual**
  - in a typical use a human subject from which the biosample(s) was/were extracted

These basic definitions will be detailed further on.

Additional concepts (e.g. *dataset*, *study* ...) may be added in the future.



A graph showing recommended basic objects and their relationships. The names and attributes are examples and may diverge in count and specific wording (e.g. "subject" instead of "individual") in specific implementations.





# GA4GH transitional *Variant*

- Derived from original GA4GH data schema developed by the Data Working Group
- based on the VCF file format
- representation of precise sequence alterations, copy number variants and single fusion events
- primary goals
  - sample based data storage
  - object model for query APIs (Beacon...)
- not attempting to provide reference variant, equivalence functionality
- parallel development of complete object model (allele | haplotype ..., equivalence) by the GA4GH GKS work stream, based on VMC

```
{  
  "biosample_id" : "structdb-bs-nhl-0009876",  
  "callset_id" : "structdb-CS-nhl-0009876",  
  "created" : "2019-01-22T03:06:45Z",  
  "digest" : "6:63450000,63550000-63450000,63550000:DEL",  
  "end" : [  
    63450000,  
    63550000  
,  
  "id" : "structdb-var-123456790",  
  "info" : {  
    "cnv_length" : 85500000,  
    "cnv_value" : -0.294  
,  
  "reference_bases" : "N",  
  "reference_name" : 6,  
  "start" : [  
    63450000,  
    63550000  
,  
  "updated" : "2019-02-01T12:40:21Z",  
  "variant_type" : "DEL"  
}
```

```
{  
  "alternate_bases" : "AC",  
  "callset_id" : "DIPG_CS_0290",  
  "created" : "2018-11-06T11:46:30.028Z",  
  "digest" : "2:203420136:A>AC",  
  "genotype" : [  
    "1",  
    ".  
,  
  "id" : "5be1840772798347f0ed9e8b",  
  "reference_bases" : "A",  
  "reference_name" : "2",  
  "start" : [  
    203420136  
,  
  "updated" : "2018-11-06T11:46:30.028Z"  
}
```

# Standardized Data Model for Consistent Schema Development

- A consistent high-level data model is essential for the development of reliable schemas and tools for
  - genomic and clinical, metadata storage
  - development of genomic query and data delivery APIs
  - distributed/federated access across separate (geographic, logistic) data repositories using consistent logical structure:
    - "**BRCA1 variant** in **germline sample** from a male **individual** with a diagnosis of breast carcinoma (ncit:C5214)
- The abstract data model can be expressed in different types of implementations
  - Phenopackets data exchange standard
  - Progenetix database model
    - schema-derived object storage datacollections for individuals, biosamples, callsets and variants

