The GA4GH Beacon Protocol

A Standardized Format for Federated Genomics Data Exchange
The Global Alliance for Genomics and Health
Making genomic data accessible for research and health

- January 2013 - 50 participants from eight countries
- June 2013 - White Paper, over next year signed by 70 “founding” member institutions (e.g. SIB, UZH)
- March 2014 - Working group meeting in Hinxton & 1st plenary in London
- October 2014 - Plenary meeting, San Diego; interaction with ASHG meeting
- June 2015 - 3rd Plenary meeting, Leiden
- September 2015 - GA4GH at ASHG, Baltimore
- October 2015 - DWG / New York Genome Centre
- April 2016 - Global Workshop @ ICHG 2016, Kyoto
- October 2016 - 4th Plenary Meeting, Vancouver
- May 2017 - Strategy retreat, Hinxton
- October 2017 - 5th plenary, Orlando
- May 2018 - Vancouver
- October 2018 - 6th plenary, Basel
- May 2019 - GA4GH Connect, Hinxton
- October 2019 - 7th Plenary, Boston
- October 2020 - Virtual Plenary, June 2021 - Virtual Connect ...
- October 2021 - Virtual Plenary ...
A federated ecosystem for sharing genomic, clinical data
Silos of genome data collection are being transformed into seamlessly connected, independent systems.
A *Beacon* answers a query for a specific genome variant against individual or aggregate genome collections.

**YES | NO | \0**
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

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 ...
Information Management: A Proposal

Abstract

This paper concerns the management of personal information about applications and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and describes a solution based on a distributed hypertext system.

Keywords: Hypertext, Computer troubleshooting, Document retrieval, Information management, Project control

World Wide Web

The WorldWideWeb (WWW) is a wide-area hypertext information retrieval initiative aiming to give universal access to a large universe of documents. Everything there is online about WWW is linked directly or indirectly to this document, including an extensive summary of the project. Whether you're looking for the latest WWW news, a WWW FAQ, a WWW bibliography, a list of the most interesting WWW sites, or a WWW expert - you'll find it here!
Beacon Project in 2016

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

Search for allele

<table>
<thead>
<tr>
<th>Response</th>
<th>All</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Found</td>
<td>16</td>
<td>27</td>
</tr>
<tr>
<td>Not Found</td>
<td>22</td>
<td></td>
</tr>
</tbody>
</table>

Search all beacons

- **35+ Organizations**
- **90+ Beacons**
- **200+ Datasets**
- **100K+ Individuals**

**BioReference**
Hosted by BioReference Laboratories

**Catalogue of Somatic Mutations in Cancer**
Hosted by Wellcome Trust Sanger Institute

**Cell Lines**
Hosted by Global Alliance for Genomics and Health

**Conglomerate**
Hosted by Wellcome Trust Sanger Institute

**COSMIC**
Hosted by Wellcome Trust Sanger Institute

**dbGaP: Combined GRU Catalog and NHLBI Exome Seq**

**Beacon Network Website or API**

- **Beacon Network**
  - Q: Who has information about this allele? → A: BRCA Exchange Beacon
  - Beacon API
  - Beacon API
  - BIPMed Beacon

**Beacon API**
- **BRCA Exchange Beacon**
  - A: Yes
  - Database
  - Clinical Record or EMR

**PhenomeCentral Beacon**
- A: No

**VCF Files**
Minimal GA4GH query API structure

Beacon+ in 2016 - Driving GA4GH implementations

Beacon+ "Metadata"
ELIXIR - Making Beacons Biomedical

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

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

v1.1 and roadmap

- **structural variations** (DUP, DEL) in addition to SNV
- ... more structural queries (translocations/fusions...)
- Beacon queries as entry for data handover (outside Beacon protocol)
- layered authentication system using ELIXIR AAI

- **v2** filters for phenotypic & technical metadata
- **v2** Extended quantitative responses
- Ubiquitous deployment (e.g. throughout ELIXIR network)
Beacon v2
Enabling "omics" data access and retrieval from a variety of resources

A schematic representation of how Beacon works. (A) Beacon API implementation and (B) A Beacon query and aggregated response
The original GA4GH Beacon implementation (up to v0.3) was conceived as a protocol for sharing the presence/absence of a given, specific, genomic mutation in a set of data (from patients of a given disease or from the population in general). Although with some potential benefit, e.g. in the area of rare disease diagnostics, it was not designed for clinical use but chiefly to foster data sharing by triggering the inquisitiveness of researchers once some data of interest is discovered in another institution. While later extensions of the protocol (v1.0 - v1.n) expanded the query and response options, this did not deviate from the general “existence of variants in resource X” paradigm.

The simplicity and success of the concept has generated the request of making it more powerful, more useful in healthcare environments. The requests include:

- Allowing more informative queries, like filtering by gender or age
- Allowing to trigger the next step in the data access process, e.g. who to contact or which are the data use conditions
- Jumping to another system where the data could be accessed, e.g. if the Beacon is internal to a hospital, to provide the Id of the EHR of the patients having the mutation of interest.
- Including annotations about the variants found, among which the expert/clinician conclusion about the pathogenicity of a given mutation in a given individual or its role in producing a given phenotype.

The process
The GA4GH Beacon group started a set of meetings and interviews with GA4GH Driver Projects and with ELIXIR partners in order to determine the scope of the next generation Beacon. The goal was to be useful without breaking the simplicity that made Beacon version 1 successful.

Interviews were conducted with the following GA4GH Driver Projects:

- Autism Speaks
- BRCA Exchange
- CanDIG
- EGA, ENA, EVA
- EuCanCancer
- European Joint Programme - Rare Diseases
- H3Africa
- GEM Japan
- Genomics England
- Matchmaker Exchange
- SVIP /SPHN
- VICC
Beacon v2 Code
Contribute & Implement

• Github repositories for
  ‣ beacon v2 framework
  ‣ beacon v2 models
  ‣ tools, website, templates ... 
• framework with schema definitions for requests and responses, endpoints template examples
• models for request & response implementations, but in principle up to implementers to modify

beacon-framework-v2
Beacon Framework version 2

Introduction

The GA4GH Beacon specification is composed by two parts:
• the Beacon Framework (in this repo)
• the Beacon Model (in the Models repo)

The Beacon Framework is the part that describes the overall structure of the API requests, responses, parameters, teh common components, etc. It could also be referred in this document as simply the Framework.

A Beacon Model describes the set of concepts included in a Beacon version (e.g. Beacon v2), like individual or biosample. It could also be referred in this document as simply the Model.

The Framework could be considered the syntax and the Model as the semantics.

Refer to the Models repo for further information about the Model and how to use it.

The Framework doesn’t include anything related to specific entities but only the mechanisms for querying them and parsing the responses. The BF is, therefore, independent from/agnostic to any specific Model. It can be leveraged to describe models from other domains like proteomics, imaging, biobanking, etc.

A Beacon instance is just an implementation of a Beacon Model that follows the rules stated by the Beacon Framework.

If you are a Beacon implementer, then, you don’t need to clone this (Framework) repo, you only need to copy (or clone) the Beacon Model and modify it to your specific instance. You will find plenty of references to the Framework in the Model copy, and you will use the json schemas in the Framework to validate that both the structure of your requests and responses are compliant with the Beacon Framework. The Beacon verifier tool would help in such validation.

The Framework repo includes the elements that are common to all Beacons:
1. The configuration files
2. The Json schemas for the requests, the responses, and its respective sections
3. The files of every Beacon root
4. Examples of all the above (using a fake and simple Model)
Beacon v2 Paths

Progenetix utilizes Beacon v2 REST paths

- Beacon v2 paths are used in the Beacon specification to scope query and delivery
- Progenetix uses a default /biosamples/ + query path for its front end queries, and then collection specific methods for data retrieval (see next)
- current implementation addresses a core subset of all options, and evaluates some still moving targets
  - variants_interpretations
  - variant instances versus prototypes
  - ...

Base /biosamples

/biosamples/ + query

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

/biosamples/id/

- /biosamples/pgxbs-kfva5c9/
  - retrieval of a single biosample

/biosamples/id/variants/ & /biosamples/id/variants_in_sample/

- /biosamples/pgxbs-kfva5c9/variants/
- /biosamples/pgxbs-kfva5c9/variants_in_sample/
  - retrieval of all variants from a single biosample
  - currently - and especially since for a mostly CNV containing resource - variants means "variant instances" (or as in the early v2 draft variantsInSample)

Base /variants

There is currently (April 2021) still some discussion about the implementation and naming of the different types of genomic variant endpoints. Since the Progenetix collections follow a "variant observations" principle all variant requests are directed against the local variants collection.

If using q_variants or variants_in_sample, those will be treated as aliases.

/variants/ + query

- variants?
  - assemblyId=GRCh38&referenceName=17&variantType=DEL&filterLogic=AND&start=7500000&start=7676592&end=769696&end=7800000
  - This is an example for a Beacon "Bracket Query" which will return focal deletions in the TPS3 locus (by position).

/variants/id/ or /variants_in_sample/id) or /g_variants/id/}

- /variants/5f5a35586b8c1d6d377b77f6/
- /variants_in_sample/5f5a35586b8c1d6d377b77f6/
- /biosamples/ & variants_in_sample/id/biosamples/

- /variants/5f5a35586b8c1d6d377b77f6/biosamples/
- /variants_in_sample/5f5a35586b8c1d6d377b77f6/biosamples/
Beacon v2 Requests
POSTing Queries

- Beacon v2 supports a mix of dedicated endpoints with REST paths
- POST requests using JSON query documents
- Final syntax for core parameters still in testing stages

```json
{
  "$schema":"beaconRequestBody.json",
  "meta": {
    "apiVersion": "2.0",
    "requestedSchemas": [
      {
        "entityType": "individual",
        "schema": "https://progenetix.org/services/schemas/Phenopacket/"
      }
    ],
  },
  "query": {
    "requestParameters": {
      "datasets": {
        "datasetIds": ["progenetix"]
      },
      "filterLogic": "OR"
    },
    "pagination": {
      "skip": 0,
      "limit": 10
    },
    "filters": [
      { "id": "NCIT:C4536" },
      { "id": "NCIT:C95597" },
      { "id": "NCIT:C7712" }
    ]
  }
}
```
Beacon v2 Filters

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

- 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 Scouts: Structural Variants

Re-defining & scoping variant queries

- contributors from different "stakeholder" areas
  - clinical genomics / rare diseases
  - variant repository (Ensembl)
  - cancer research resource
  - cancer variant annotation repositories
- close integration with ELIXIR h-CNV group
- process involved discussions about semantics of variant types, e.g.
  - DUP as CNV or in place
  - DEL as CNV from which size
- general attempt to use Sequence Ontology classes as guidance, but no still ambiguities / lack of terms

References

Conventions Followed in the Document

Use of Positional Parameters

Variant Types, Documentation and Example Queries

- INS (Insertion)
- DEL (Deletion)
- DUP (Duplication)
- Amp (DUP more than 2) CN type of approach
- LOH (loss of representation of second allele, with or without copy number change)
- INV (inversion)
- TL (Translocation)
- Proposal: BRK (Breakpoint)
- ME (Mobile elements insertion /deletion)
- CNV - (non directional CNVs) - do we allow cnv queries? / complex CNVs
- Tandem Duplication

Name Based Queries
- Topics for discussion

Technical considerations
- Format of GET queries
Beacon v2: Extended Variant Queries

Range and Bracket queries enable positional wildcards and fuzziness

- Genome Range Queries provide a way to "fish" for variants overlapping an indicated region, e.g. the CDR of a gene of interest
- Additional parameters (e.g. variant type, reference or alternate bases) limit the scope of the responses
- new Beacon v2 size parameters to limit structural variants (e.g. "focal" CNVs)

- Genome Bracket Queries allow to search for structural variants with start and end positions falling into defined sequence ranges
- allows to query any contiguous genomic variant (and in principle also can step in for range queries)
- typical use case is e.g. the query for variants such as duplications covering the whole CDR of a gene, while limiting the allowed start or end regions
Beacon & Phenopackets

**Data discovery and delivery using standardized GA4GH formats and schemas**

- Beacon v2 & Phenopackets v2 emerging as essential standards for federation and exchange of biomedical data
- Hierarchical coding systems and with widespread use of CURIEs
- Other formats based on international standards, e.g.
  - ISO (ISO 8601 time & period, ISO 3166 country codes, ...)
  - IETF (GeoJSON ...)
  - W3C (CURIE ...)
- These standards become pervasive throughout GA4GH's ecosystem

Beacon query filters correspond well to Phenopackets data

Phenopackets as recommended protocol for Beacon v2 data delivery
Standardized formats and data schemas for developing an "Internet of Genomics"

- “cross-workstreams, cross-drivers” initiative to document GA4GH object standards and prototypes launched in December 2018
- documentation and implementation examples provided by GA4GH members
- not a rigid, complete data schema
- object vocabulary and semantics for a large range of developments
  - Beacon as contributor and user
  - 2021: going forward through integration with GA4GH TASC efforts, towards "standards library"
Beacon Project - Partner Engagement & Next Steps

• Working with **partner communities & projects** on deploying Beacons
  • ELIXIR hCNV Community
  • European Joint Program on Rare Diseases
  • clinical groups & data initiatives (e.g. Andalucia, Cancer Core Europe, SPHN)
  • variant annotation resources, with optional clinical components (e.g. SVIP-O)
• Improving reference implementation and standards / **compliance** testing
• Beacon v2 "fast forward" development
• aligning w/ GA4GH standards, through "request & adopt" => SchemaBlocks \{S\}[B]
• networks throughout & beyond ELIXIR
ELIXIR Beacon Network

- developed under lead from ELIXIR Finland
- **authenticated access** w/ ELIXIR AAI
- **incremental extension**, starting with ELIXIR Beacon resources adhering to the **latest specification** (contrast to legacy networks)
- service details provided by individual Beacons, using **GA4GH service-info**
- **registration service**
  - integrator throughout ELIXIR Human Data
  - starting point for "beyond ELIXIR"

**feature rich** federated Beacon services
hCNV Implementation Studies 2021-2023 No. 2

Beacon and beyond — Implementation-driven standards and protocols for CNV discovery and data exchange

• reinforce work on priority areas established in the current hCNV Implementation Study
• extend collaborations with the Rare Diseases and Galaxy Communities, EJP-RD and GA4GH
• Expected outcomes
  ➡ shared CNV resources testing advanced versions of the Beacon protocol
  ➡ integration of GA4GH standards such as Phenopackets in such resources
  ➡ tools for data ingestion and export for standard formats (e.g. VCF, Phenopackets) and CNV-specific improvements of such standards
  ➡ ELIXIR AAI demo on clinical and research hCNV resources
  ➡ demonstration of Galaxy pipeline adoption for real-world hCNV data analysis projects
• connecting to international partners, e.g. Cancer Genomics Consortium (U.S.)

➤ WP1 - hCNV community reference resources
➤ WP2 - hCNV Resources and Beacon
➤ WP3 - Galaxy Community Intersection and Data Exchange
➤ WP4 - Workflows and Tools for hCNV Data Exchange Procedures
➤ WP5 - Training and dissemination
hCNV Implementation Studies 2021-2023 No. 2

Beacon and beyond — Implementation-driven standards and protocols for CNV discovery and data exchange

- WP1 - hCNV community reference resources
- WP2 - hCNV Resources and Beacon
- WP3 - Galaxy Community Intersection and Data Exchange
- WP4 - Workflows and Tools for hCNV Data Exchange Procedures
- WP5 - Training and dissemination

"Galaxify", "Beaconize" & "Phenopack" Progenetix & RD-CNVDdb prototypes
Functionalities, Standards, Components

RD use case
- Beacon v2
- FEGA storage technology
  - Storage
  - Interface
- ELIXIR-AAI
  - REMS
  - GPAP
- Secure computing

GA4GH standards
- Beacon
- DUO
- SAM/BAM/CRAM
- Crypt4GH
- Passports
- TES
- VCF
- htsget
- DUO
- WES
- phenopackets
Onboarding

Demonstrating Compliance

• onboarding server run by CRG (EGA team)

• registering the URI of a server's map document will initiate traversal and testing of services

• blueprint for Beacon service registries

• to be used as demonstrator in GA4GH approval process for the Spring 2022 session
Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

The Beacon API v2 proposal opens the way for the design of a simple but powerful "genomics API".
The Beacon API v2 proposal opens the way for the design of a simple but powerful "genomics API".

"Internet of Genomics"

Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

Michael Baudis | 2020 | CC0
Beacon API Leads
Jordi Rambla
Anthony Brooks
Juha Törnroos
Discovery WS
Michael Baudis (Beacon)
Marc Flüe (Networks)
ELIXIR
Gary Saunders
David Lloyd
Serena Scollen
Dylan Spalding
Beacon Team CRG
Laureen Fromont
Babita Singh
Sabela de la Torre Pernas

Beacon v2 Scouts
Tim Beck
Joaquin Dopazo
Veronique Geoffroy
Jean Muller
David Salgado
Alex Wagner

Baudisgroup @ UZH
Michael Baudis
(Paula Carrio Cordo)
(Bo Gao)
Qingyao Huang
Sofia Pfund
Rahel Paloots
Hangjia Zhao
Pierre-Henri Toussaint

{S}[B] and GA4GH
Melanie Courtot
Helen Parkinson
many more ...