Beacon v2
Onboarding Strategies & Feature Examples
Beacon v2
Migration Workshop
A **Beacon** answers a query for a specific genome variant against individual or aggregate genome collections.

YES | NO | \0
Global Alliance “Beacon” - Jim Ostell, NCBI, March 7, 2014

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 ...

"I would personally recommend all those be held for version 2, when the beacon becomes a service."  
Jim Ostell, 2014
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...
<table>
<thead>
<tr>
<th>Year</th>
<th>Beacon v1 Development</th>
<th>Beacon v2 Development</th>
<th>Related ...</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>GA4GH founding event; Jim Ostell proposes Beacon concept including &quot;more features ... version 2&quot;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>• beacon-network.org aggregator created by DNAstack</td>
<td>• Beacon concept implemented on progenetix.org</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>• Beacon v0.3 release</td>
<td>• concepts from GA4GH Metadata (ontologies...)</td>
<td>• ELIXIR starts Beacon project support</td>
</tr>
<tr>
<td></td>
<td>• work on queries for structural variants (brackets for fuzzy start and end parameters...)</td>
<td>• entity-scoped query parameters (&quot;individual.age&quot;)</td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td>• OpenAPI implementation</td>
<td>• Beacon demos &quot;handover&quot; concept</td>
<td>• GA4GH re-structuring (workstreams...)</td>
</tr>
<tr>
<td></td>
<td>• integrating CNV parameters (e.g. &quot;startMin, statMax&quot;)</td>
<td></td>
<td>• Beacon part of Discovery WS</td>
</tr>
<tr>
<td>2018</td>
<td>• Beacon v0.4 release in January; feature release for GA4GH approval process</td>
<td>• &quot;Scouts&quot; teams working on different aspects - filters, genomic variants, compliance ...</td>
<td>• new Beacon website (March)</td>
</tr>
<tr>
<td></td>
<td>• GA4GH Beacon v1 approved at Oct plenary</td>
<td>• discussions w/ clinical stakeholders</td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td>• ELIXIR Beacon Network</td>
<td>• framework + models concept implemented</td>
<td>• Beacon publication at Nature Biotechnology</td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td>• range and bracket queries, variant length parameters</td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td></td>
<td>• starting of GA4GH review process</td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td></td>
<td>• further changes esp. in default model, aligning with Phenopackets and VRS</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• unified beacon-v2 code &amp; docs repository</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Beacon v2 approved at Apr GA4GH Connect</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• docs.genomebeacons.org</td>
<td></td>
</tr>
</tbody>
</table>
Beacon v1 Development

2014
- GA4GH founding event; Jim Ostell proposes Beacon concept including "more features ... version 2"

2015
- beacon-network.org aggregator created by DNAstack

2016
- Beacon v0.3 release
- "Work on queries for structural variants (brackets for fuzzy start and end parameters...)"

2017
- OpenAPI implementation
- "Integrating CNV parameters (e.g. startMin, startMax)"

2018
- Beacon v0.4 release in January; feature release for GA4GH approval process
- GA4GH Beacon v1 approved at Oct plenary

2019
- ELIXIR Beacon Network

2020

2021

2022

Beacon v2 Development

- beacon-network.org aggregator created by DNAstack

2015
- Beacon concept implemented on progenetix.org
- concepts from GA4GH Metadata (ontologies...)
- entity-scoped query parameters ("individual.age")

2016
- Beacon demos "handover" concept

2017
- Beacon hackathon Stockholm; settling on "filters"
- Barcelona goes Zurich developers meeting
- Beacon API v2 Kick off
- adopting "handover" concept
- "Scouts" teams working on different aspects - filters, genomic variants, compliance ...
- discussions w/ clinical stakeholders

2018
- framework + models concept implemented
- range and bracket queries, variant length parameters
- starting of GA4GH review process
- further changes esp. in default model, aligning with Phenopackets and VRS
- unified beacon-v2 code & docs repository
- Beacon v2 approved at Apr GA4GH Connect

2019

2020

2021
- Phenopackets v2 approved

2022
- docs.genomebeacons.org

Related ...

- ELIXIR starts Beacon project support
- GA4GH re-structuring (workstreams...)
- Beacon part of Discovery WS
- new Beacon website (March)
- Beacon publication at Nature Biotechnology
- ELIXIR starts Beacon project support
Beacon v1 => v2
Genomic variation queries

- Beacon v2 defines query schemas through JSON Schema documents for POST requests and REST paths in OpenAPI documents.
- Additional variant parameters:
  - variantType, mateName (existing in v1)
  - geneld
  - variantMinLength, variantMaxLength
  - aminoacidChange
  - genomicAlleleShortForm
Beacon v1 => v2
Keep it simple - modifying GET query strings

0.3  
?ref=GRCh38&chrom=17&pos=7577121&referenceAllele=G&allele=A

1.0  
[assemblyId=GRCh38&referenceName=17&start=7577120&referenceBases=G&alternateBases=A

2.0  
?referenceName=refseq:NC_000017.11&start=7577120&referenceBases=G&alternateBases=A
Beacon v1 => v2

Keep it simple - modifying GET query strings

0.3

?ref=GRCh38&chrom=17&pos=7577121&referenceAllele=G&allele=A

v1 switched for the API to 0-based coordinates (with 1-based representation in user facing forms - compare to UCSC genome browser)

1.0

?assemblyId=GRCh38&referenceName=17&start=7577120&referenceBases=G&alternateBases=A

v2 recommends using assembly-specific identifiers (refseq id) although assemblyId and alternative reference identifiers such as "chr17" are in principle permitted

2.0

?referenceName=refseq:NC_000017.11&start=7577120&referenceBases=G&alternateBases=A
Beacon v2
Boolean response example

- Beacon v2 is "chatty" regarding returned metadata, to disambiguate responses
- the response payload for **Boolean** and **count** responses is provided in the **responseSummary** object
Beacon v2
Count response example

- Beacon v2 is "chatty" regarding returned metadata, to disambiguate responses
- the response payload for **Boolean** and **count** responses is provided in the `responseSummary` object
Beacon v2
So what would you need?

- Beacon v2 (as v1) for Boolean and count responses can be implemented w/o complex infrastructure

- compared to v1, some additional meta information is expected in the response (but this can be pretty static for individual instances)
Beacon v2 - Migration Workshop

Reference Implementation (Manuel Rueda)
Progenetix & Beacon v2
A custom "full stack" implementation of a genomics resource around Beacon data model & API
Progenetix in 2022

Cancer Genomics Reference Resource

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

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 142963 samples.

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.
**Progenetix**

Genomic resource utilizing Beacon v2 calls

- Progenetix uses Beacon v2 queries to drive its UI
- all individuals, biosamples, variants, analyses matched by a given query are stored by their object ids
- handovers for variant purposes (e.g. to retrieve all matched variants) are returned in the original response and asynchronously retrieved by the front end app
Progenetix

Genomic resource utilizing Beacon v2 calls

• Progenetix uses Beacon v2 queries to drive its UI

• all individuals, biosamples, variants, analyses matched by a given query are stored by their object ids

• handovers for variant purposes (e.g. to retrieve all matched variants) are returned in the original response and asynchronously retrieved by the front end app
Progenetix
Genomic resource utilizing Beacon v2 calls

- Progenetix uses Beacon v2 queries to drive its UI
- all individuals, biosamples, variants, analyses matched by a given query are stored by their object ids
- handovers for variant purposes (e.g. to retrieve all matched variants) are returned in the original response and asynchronously retrieved by the front end app
Progenetix Stack

- **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

- **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

Entity collections
- variants
- analyses
- biosamples
- individuals

Utility collections
- collations
- geolocs
- genespans
- publications
- qBuffer
bycon
Progenetix' Beacon Stack

- Python-based software stack
- developed for in-house use - not well documented etc.
- happy about adoption & contributions...
Beacon v2

Beaconise your Data
Beacon v1 => v2
Genomic variation queries

- Beacon v2 defines query schemas through JSON Schema documents for POST requests and REST paths in OpenAPI documents

- Additional variant parameters:
  - variantType, mateName (existing in v1)
  - geneld
  - variantMinLength, variantMaxLength
  - aminoacidChange
  - genomicAlleleShortForm
**Beacon v1 => v2**

**Genomic variation queries**

- Beacon v2 defines query schemas through JSON Schema documents for POST requests and REST paths in OpenAPI documents

- Additional variant parameters:
  - `variantType`
  - `mateName` (existing in v1)
  - `geneId`
  - `variantMinLength`, `variantMaxLength`
  - `aminoacidChange`
  - `genomicAlleleShortForm`

```json
{
  "$schema": "beaconRequestBody.json",
  "meta": {
    "apiVersion": "2.0",
    "requestedSchemas": [
    {
      "entityType": "genomicVariation",
    }
    ],
  "query": {
    "requestParameters": {
      "g_variant": {
        "referenceName": "NC_000017.11",
        "start": [7577120],
        "referenceBases": "G",
        "alternateBases": "A"
      }
    },
    "requestedGranularity": "record",
    "pagination": {
      "skip": 0,
      "limit": 5
    }
  }
}
```
Types of genomic alterations in Cancer

Imbalanced Chromosomal Changes: CNV

- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- Structural chromosomal Aberrations
  - Regional Copy Number Alterations (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)
Gain of chromosome arm 13q in colorectal carcinoma

MYCN amplification in neuroblastoma (GSM314026, SJNB8_N cell line)

2-event, homozygous deletion in a Glioblastoma

low level/high level copy number alterations (CNAs)
CNVs Come in a Variety of Formats

- articles and supplements with cytoband-based rev ish CGH results are a great source of CNV data
- conversion by mapping cytoband locations (e.g. UCSC annotation files) to genome coordinates and assigning CNV types (enh, dim, amp are standard)
CNVs Come in a Variety of Formats: VCF

Issue 1: There are two fields to specify SV/CNV

1) Symbolic allele (SA)

2) SVTYPE

VCF v4.4 deprecate SVTYPE

Symbolic allele (SA)

Subtypes do not change the meaning symbolic allele.

- using genome positions (POS, INFO.END) for start, end mappings
- treatment of markers for imprecision during matching is left to the implementer
- DUP, DEL are interpreted as indicators for the type of copy number change

VCF POI ID REF ALT QUAL FILTER INFO

<table>
<thead>
<tr>
<th>#CHR POI ID</th>
<th>REF</th>
<th>ALT</th>
<th>QUAL</th>
<th>FILTER</th>
<th>INFO</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr1</td>
<td>T</td>
<td>G</td>
<td>PASS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr2</td>
<td>T</td>
<td>G</td>
<td>PASS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr3</td>
<td>T</td>
<td>G</td>
<td>PASS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>chr4</td>
<td>T</td>
<td>G</td>
<td>PASS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Use Subtype to define new structural variant

- <DUP:TANDEM> precise form of duplication
- <DEL:ME:LINE>
**CNVs Come in a Variety of Formats: VCF**

**Issue 2: two meanings of DEL and DUP**

- **Usual interpretation of “DEL” (a deletion)**
  1. Copy number of B decreases from 2 to 1, and
  2. Adjacency structure changes from (ABC, ABC) to (AC, ABC)

Both effects are important, for example...
- Copy number change can affect gene dosage
- Adjacency structure change can affect expression or disrupt a CDS

**but they do not necessarily happen at the same time.**

**The SVCLAIM field**

- **New SVCLAIM INFO field to capture what the caller could ascertain**
  - **D** (abundance / read depth) claim indicates that the call has been made based only on a measure of DNA abundance of the called region, with no evidence to support changes in breakpoint structure. This includes indirect claims of abundance made using SNV variant allele frequency.
  - **J** (adjacency / break junction) claim indicates that the call has been made based on the detection of a non-reference DNA adjacency, with no evidence to support overall changes in DNA abundance.
  - **DJ** indicates that there is evidence for both DNA abundance and adjacency changes, which are consistent with each other and suggest the structural variant of the type being reported.

- **using genome positions (POS, INFO.END) for start, end mappings**
- **treatment of markers for imprecision during matching is left to the implementer**
- **DUP, DEL are interpreted as indicators for the type of copy number change ... unless there is an explicit INFO.SVCLAIM without a “D” label**
Beacon & CNVs
Open types w/ some definitions

- Beacon supports structural variant queries through the **variantType** parameter
- The default model **does not prescribe** which types can be used (but documents VCF derived DUP & DEL)
- CNV values are not (yet) supported but EFO offers common classes
- Progenetix supports **EFO relative CN terms** (but accepts & interpolates DUP & DEL)

<table>
<thead>
<tr>
<th>Beacon</th>
<th>VCF</th>
<th>SO</th>
<th>EFO</th>
<th>VRS</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUP</td>
<td>DUP</td>
<td>SO:0001742</td>
<td>EFO:0030070</td>
<td>low-level gain (implicit)</td>
<td>a sequence alteration whereby the copy number of a given genomic region is greater than the reference sequence</td>
</tr>
<tr>
<td>DUP</td>
<td>DUP</td>
<td>SO:0001742</td>
<td>EFO:0030071</td>
<td>low-level gain</td>
<td>commonly but not consistently used for &gt;=5 copies on a biallelic genome region</td>
</tr>
<tr>
<td>DUP</td>
<td>DUP</td>
<td>SO:0001742</td>
<td>EFO:0030072</td>
<td>high-level gain</td>
<td>commonly but not consistently used for &gt;=5 copies on a biallelic genome region, of limited size (operationally max. 1-5Mb)</td>
</tr>
<tr>
<td>DUP</td>
<td>DUP</td>
<td>SO:0001742</td>
<td>EFO:0030073</td>
<td>high-level gain</td>
<td>focal genome amplification</td>
</tr>
<tr>
<td>DEL</td>
<td>DEL</td>
<td>SO:0001743</td>
<td>EFO:0030067</td>
<td>partial loss (implicit)</td>
<td>a sequence alteration whereby the copy number of a given genomic region is smaller than the reference sequence</td>
</tr>
<tr>
<td>DEL</td>
<td>DEL</td>
<td>SO:0001743</td>
<td>EFO:0030068</td>
<td>partial loss</td>
<td>low-level copy number loss</td>
</tr>
<tr>
<td>DEL</td>
<td>DEL</td>
<td>SO:0001743</td>
<td>EFO:0030069</td>
<td>complete loss</td>
<td>complete genomic deletion (e.g. homozygous deletion on a biallelic genome region)</td>
</tr>
</tbody>
</table>

1 VCFv4.4 introduces an SVCLAIM field to disambiguate between in situ events (such as tandem duplications; known adjacency/break junction: SVCLAIM=J) and events where e.g. only the change in abundance/read depth (SVCLAIM=D) has been determined. Both J and D flags can be combined.
Positional Queries
Going beyond single positions...

- Beacon v1 already provided support for "bracket" queries, e.g. for CNV queries - v2 improves documentation

- Use cases w/ focus on structural variants were evaluated by a Beacon "scout" team

- new "range" option
  - anything w/ overlap
  - matched variants can optionally be filtered by type, size, sequence

- query option are not hard defined but derived from parameters
  - Strong wish for defined types?
Beacon v2 - Beaconise your Data
BANCCO (David Salgado)
Beacon v2 - Beaconise your Data

Filters (Vatsalya Maddi)
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".
Rapidly evolving documentation of both the Beacon API itself and its use and technical implementation on docs.genomebeacons.org docs.progenetix.org

Shoutout to Laure(e)n Fromont & Manuel Rueda for being instrumental in the Beacon v2 documentation!
Future?
Some proposals for a stepwise Beacon protocol extension

- Query language expansion, e.g. Boolean options for chaining filters
  - use of heterogeneous/alternative annotations within and across resources
- **Phenopackets** support as a (the?) default format for biodata export
- **Phenopackets** as request documents
- Focus on service & resource discovery
- **ELIXIR Beacon Network**, including translations for federated queries to Beacon and Beacon-like resources