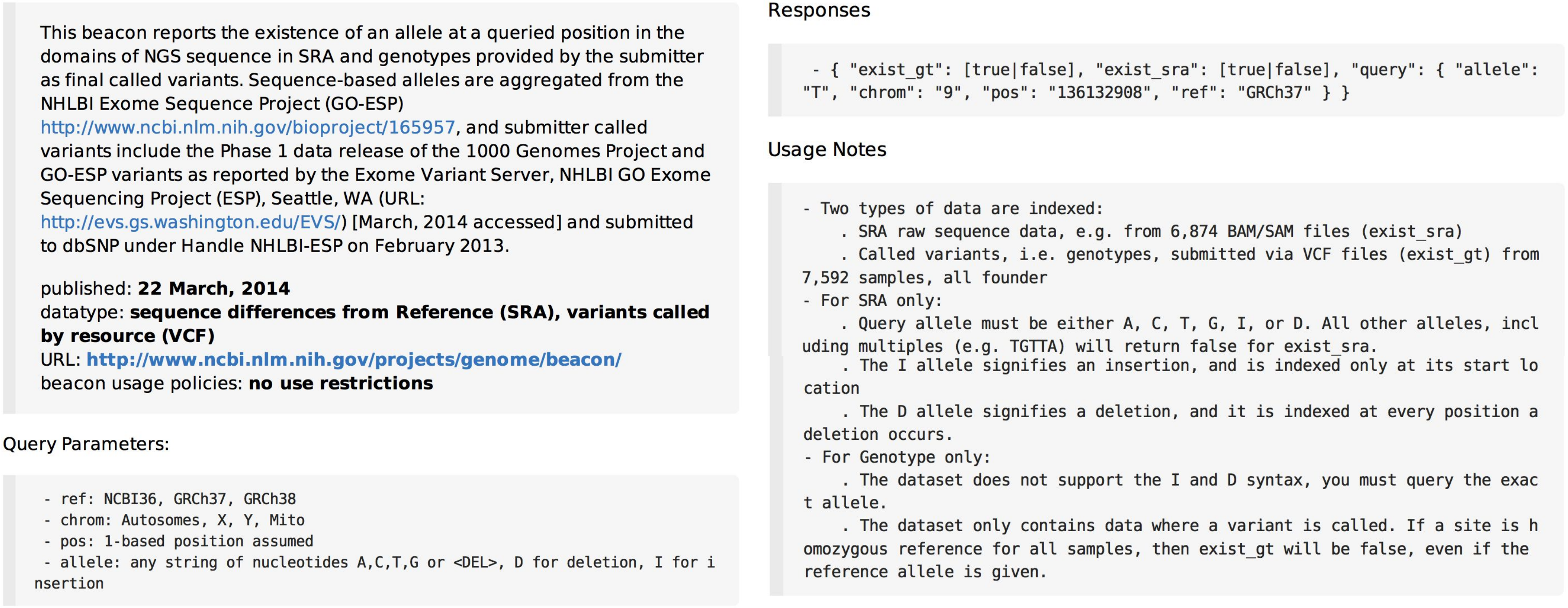
Beacon API

Miro Cupak
Jordi Rambla

17/10/2016

Road to Beacon API

2. A Web Interface for Genomic Variant Queries - NCBI



This beacon reports the existence of an allele at a queried position in the domains of NGS sequence in SRA and genotypes provided by the submitter as final called variants. Sequence-based alleles are aggregated from the NHLBI Exome Sequence Project (GO-ESP) http://www.ncbi.nlm.nih.gov/bioproject/165957, and submitter called variants include the Phase 1 data release of the 1000 Genomes Project and GO-ESP variants as reported by the Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/) [March, 2014 accessed] and submitted to dbSNP under Handle NHLBI-ESP on February 2013.

published: **22 March, 2014**
datatype: **sequence differences from Reference (SRA), variants called by resource (VCF)**
URL: **http://www.ncbi.nlm.nih.gov/projects/genome/beacon/**
beacon usage policies: **no use restrictions**

Query Parameters:

```
 - ref: NCBI36, GRCh37, GRCh38
 - chrom: Autosomes, X, Y, Mito
 - pos: 1-based position assumed
 - allele: any string of nucleotides A,C,T,G or <DEL>, D for deletion, I for i
nsertion
```

Responses

```
 - { "exist_gt": [true|false], "exist_sra": [true|false], "query": { "allele":
"T", "chrom": "9", "pos": "136132908", "ref": "GRCh37" } }
```

Usage Notes

```
- Two types of data are indexed:
    . SRA raw sequence data, e.g. from 6,874 BAM/SAM files (exist_sra)
    . Called variants, i.e. genotypes, submitted via VCF files (exist_gt) from
7,592 samples, all founder
- For SRA only:
    . Query allele must be either A, C, T, G, I, or D. All other alleles, incl
uding multiples (e.g. TGTTA) will return false for exist_sra.
    . The I allele signifies an insertion, and is indexed only at its start lo
cation
    . The D allele signifies a deletion, and it is indexed at every position a
deletion occurs.
- For Genotype only:
    . The dataset does not support the I and D syntax, you must query the exac
t allele.
    . The dataset only contains data where a variant is called. If a site is h
omozygous reference for all samples, then exist_gt will be false, even if the
reference allele is given.
```

Before

Beacon 0.1 (2014)

- Really simple (2 records)
- true/false response.

- Too vague.

Beacon 0.2 (2015)

- Complex (9 records)
- true/false/overlap/null response.
- Datasets.
- Simple data use conditions.
- Self description.

- Not well adopted.
- Not polished enough.

Now

- Beacon 0.3 (2016).
- Simplified 0.2.
- Based on real needs.

- Improved support for datasets and cross-dataset queries.
- Modular and extensible.
- Data versioning.
- Various improvements to the data model.
- Tooling.

https://github.com/ga4gh/beacon-team/releases/tag/v0.3.0

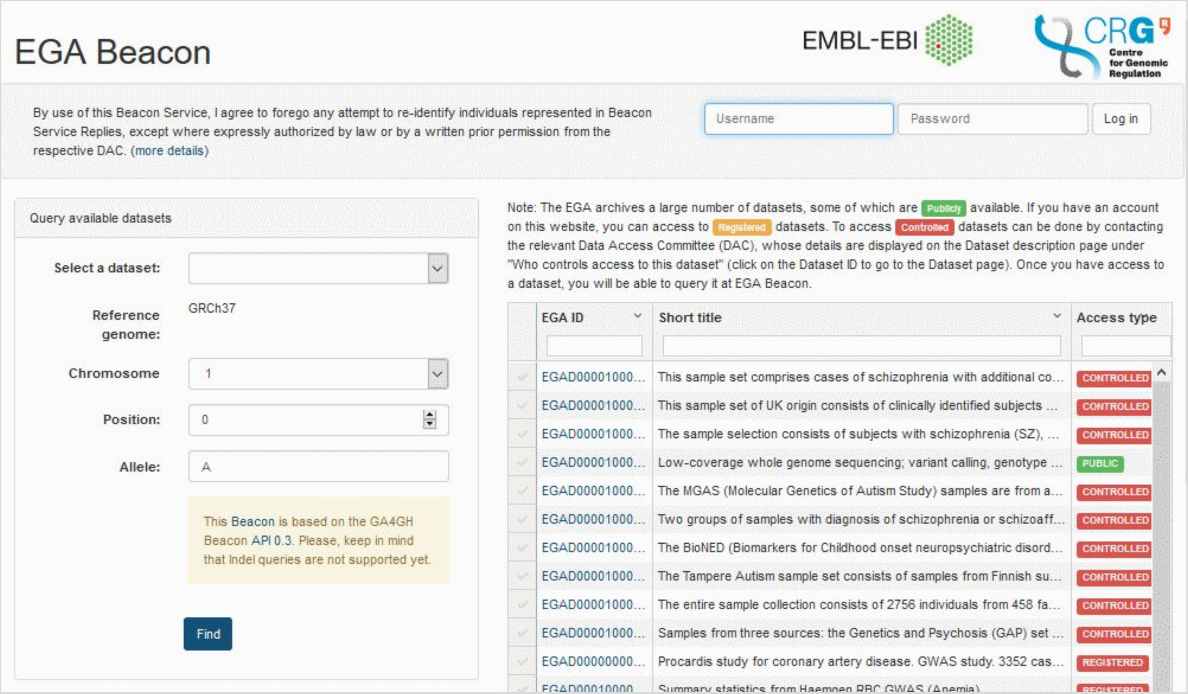
Next

- Beacon 0.4 (in progress).

- Support for complex variants.
- Improved data use conditions.
- Documentation.
- Developer experience.
- Various minor improvements.

Case study

- EGA & ELIXIR Beacons
  - Docker backend & web
  - Tools
  - Apache 2.0 licence
- Links
  - ELIXIR API repository
  - ELIXIR web repo
  - EGA Beacon



EGA Beacon

By use of this Beacon Service, I agree to forego any attempt to re-identify individuals represented in Beacon Service Replies, except where expressly authorized by law or by a written prior permission from the respective DAC. (more details)

Query available datasets

Select a dataset:

Reference genome: GRCh37

Chromosome: 1

Position: 0

Allele: A

This Beacon is based on the GA4GH Beacon API 0.3. Please, keep in mind that Indel queries are not supported yet.

Note: The EGA archives a large number of datasets, some of which are Publicly available. If you have an account on this website, you can access to Registered datasets. To access Controlled datasets can be done by contacting the relevant Data Access Committee (DAC), whose details are displayed on the Dataset description page under "Who controls access to this dataset" (click on the Dataset ID to go to the Dataset page). Once you have access to a dataset, you will be able to query it at EGA Beacon.

	EGA ID	Short title	Access type
✓	EGAD00001000...	This sample set comprises cases of schizophrenia with additional co...	CONTROLLED
✓	EGAD00001000...	This sample set of UK origin consists of clinically identified subjects ...	CONTROLLED
✓	EGAD00001000...	The sample selection consists of subjects with schizophrenia (SZ), ...	CONTROLLED
✓	EGAD00001000...	Low-coverage whole genome sequencing; variant calling, genotype ...	PUBLIC
✓	EGAD00001000...	The MGAS (Molecular Genetics of Autism Study) samples are from a...	CONTROLLED
✓	EGAD00001000...	Two groups of samples with diagnosis of schizophrenia or schizoaff...	CONTROLLED
✓	EGAD00001000...	The BioNED (Biomarkers for Childhood onset neuropsychiatric disord...	CONTROLLED
✓	EGAD00001000...	The Tampere Autism sample set consists of samples from Finnish su...	CONTROLLED
✓	EGAD00001000...	The entire sample collection consists of 2756 individuals from 458 fa...	CONTROLLED
✓	EGAD00001000...	Samples from three sources: the Genetics and Psychosis (GAP) set ...	CONTROLLED
✓	EGAD00000000...	Procardis study for coronary artery disease. GWAS study. 3352 cas...	REGISTERED
✓	EGAD00010000...	Summary statistics from Haemgen RBC GWAS (Anemia)	REGISTERED

Future

- Stabilize the API, indeed more: simplify and flexibilize
    - Always have the expected response when omitting parameters (optional for non-core)
- Planned
    - Beacon Network API
    - Triple-A access levels
- Need further discussion, because dependencies from other GA4GH groups or could depart from “simplicity” principal
        - Quantitative Variants
        - GA4GH Objects inside Beacon response
        - Genotype + Phenotype queries ~ “Clinical” Beacons
- Parallel works on
    - Security, Privacy, ELSI...

Questions?