Genome screening at the core of “Personalised Health”

- **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- The unexpectedly large amount of sequence variants in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of reference repositories
- **Standardized data formats** and **exchange protocols** are needed to connect these resources throughout the world, for reciprocal, international data sharing and biocuration efforts
- Our work @ UZH:
  - **cancer** genome repositories
  - biocuration
  - protocols & formats
The vision: Federation of data
GA4GH API promotes sharing genomic data globally.

A federated data ecosystem. To share genomic data globally, this approach further medical research without requiring compatible data sets or compromising patient identity.

A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems.

The Global Alliance for Genomics and Health®
A federated ecosystem for sharing genomic, clinical data

Silos of genome data collection are being transformed into seamlessly connected, independent systems.
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 ...
Information Management: A Proposal

Tim Berners-Lee:
Information Management: A Proposal (CERN 1989)
&
WWW: First Page (1990)
Beacon Project
An open web service that tests the willingness of international sites to share genetic data.

Beacon v0.3 API (2016)
Beacon Project
An open web service that tests the willingness of international sites to share genetic data.

35+ Organizations
90+ Beacons
200+ Datasets
100K+ Individuals
Beacon v0.4 forward

- **structural variations** (DUP, DEL) in addition to SNV
- **metadata** queries
- layered **authentication** system using **ELIXIR AAI**
- **quantitative** responses
- Beacon queries as entry for **data delivery** (outside Beacon protocol)
- Ubiquitous **deployment** (e.g. throughout **ELIXIR** network)

---

2016++: **ELIXIR** Beacon Project

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

- Implementation of cancer beacon prototype, backed by arrayMap data
- Structural variations
- Quantitative queries
- Metadata
- Current version uses GA4GH schema compatible, non-SQL database backend (MongoDB)

Beacon arrayMap

Beacon v0.4 implementation for arrayMap.
Beacon & Handover

Beacons to support advanced types of queries and connection of data delivery services

Beacon Query
- allele_request
- biosample_request
- individual_request

Beacon Response
- beacon_response
- handover_id

Handover Action
- phenopackets
- VCF
- graphics

Intersect
- handover
  - handover_id
  - biosample_ids
  - variant_ids
  - individual_ids
  - ...

Handover

Authenticating

Beacon & Handover

Michael Baudis © EOCB 2018
Beacon Handover

- only exposure of access handle to data stored in secure system
- one-step authentication and selection of handover action; other scenarios possible / likely
- handover response outside of Beacon protocol / system
2018 Progress

- Release of v.1.0 specification
  - basis for **GA4GH Beacon v.1.0**
  - GA4GH certification (first product)
- Building flexible **Beacon Network(s)**
- Beacon paper under review
- Post-2018 planning, documented in **5-year plan**
- Widening Node participation & additional use cases:
  - ELIXIR Norway joined Implementation Study, to light Beacon(s) against **Marine data**
  - Widen this out to more use cases? Plants, other…

www.elixir-europe.org/beacons
ELIXIR Beacon Project Plan 2019-2021

• Extend the Beacon protocol to become the reference ELIXIR Data Discovery product, through expanding query options and providing richer responses, with view on biomedical applications, and in alignment with developing GA4GH standards

• Deliver ELIXIR Beacon Network as an established ELIXIR service

• Leverage ELIXIR Nodes to increase data flow through Beacon services

• Actively support the integration of the Beacon API with human data resources throughout ELIXIR, with particular view on National Genomics Initiatives, biobanks, and Human Data Communities (such as Rare Diseases and Human CNV)
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
FEATURES AND POSSIBILITIES OF THE CURRENT BEACON SPECIFICATION

- precise variant queries (chr17: 7673767 C>T)
- range queries ("any variant from here to there")
- variant frequencies
- structural genome variants, e.g. CNVs ("any deletion overlapping CDKN2A CDR coordinates")
- delivery of any kind of data matching a given query (variants, sample information, patient data ...) utilising "handover" objects (anonymous links to external services with their own security / privacy implementations)
- networking of v1.n Beacons with AAI integration as demonstrated by the ELIXIR Beacon Network
IMPLEMENTATION CHALLENGES

- Interpretation and counting issues, especially with multi-allelic variants from VCF and "borderline" structural variants (e.g. INDEL vs. DEL, DUP:TANDEM vs. DUP) - much of this related to using VCF instead of a variant model.

- Limited recognition & uptake of "recent" API versions.

- Only recent emergence of modern Beacon network aggregator.

- Protocol as "moving target" - while v1.n potential for widespread use, the rapid move towards v2 limits dissemination.

- Lack of well documented, flexible Beacon distributions:
  - The "reference implementation" is limited & more for drop-in testing ("get your VCF online") than a scalable solution.
  - Other implementations (Ensembl, Beacon+) more part of own ecosystem than for easy distribution.
Beacon v2 - Areas of Change

- Separate query types for different genomic variants
  - SNPs `BeaconSnvRequest`
  - Structural Variants `BeaconCnvRequest`
  - Region `BeaconRangeRequest`
  - ...

- Access levels

- Filters
  - Simple general filter schema w/ **scoping through prefixes** (CURIEs, private implementations)

- New types of queries:
  - By sample, patient, variant effect/evidence
  - Complex queries? (stakeholder driven; e.g. EJP-RD, GEL...)

- Schema versions & Service Info
  - Negotiated queries based on individual Beacon capabilities

---

Tested and already implemented by Beacons

- **v2.0**
- GA4GH approval process (“major product update”)

Ongoing

- **v2.n**
- Incremental rollout after v2.0

---

Code leads: Sabela de la Torre Penas & Jordi Rambla
Beacon v2 Filters

An v2 extension of the Beacon protocol will allow the query for additional data beyond genome variants, using a proposed filters extension. Such filters are thought to be prefixed attributes, where the (public or private) prefix becomes the basis of scoping the value to the correct database value.

Overview of Beacon filters

The Beacon v2 API supports the discovery of genomics and clinical datasets, and includes a powerful feature to enable the “filtering” of beacon responses by biomedical properties (e.g. phenotypes) and procedural metadata.

Filters belong to one of currently three super-classes:

- **Filters** correspond to classes from bio-ontologies for biomedical data or procedural metadata that are contained in public repositories such as the Ontology Lookup Service or NCBO’s BioPortal. Filters are identified using the full bio-ontology term/class identifier as CURIE, e.g. “HP:0100526”. Examples of bio-ontologies are, among others:
  - Human Phenotype Ontology (HPO)
    - HP:0100526: Neoplasm of the lung
  - Phenotypic Qualities Ontology (PATO)
    - PATO:0020000: female genetic sex
  - National Cancer Institute Thesaurus (NCIT)
    - NCIT:C8430: Ovarian Papillary Tumor
    - NCIT:C48724: T2 Stage Finding
  - Experimental Factor Ontology (EFO)
    - EFO:0009656: neoplastic sample

- **CustomFilters** are biomedical or metadata terms that are locally defined by a Beacon (e.g. not corresponding to known bio-ontology terms). As with standard Filters, the Beacon v2 API is agnostic to how CustomFilters are implemented by the Beacon, so this permits maximum flexibility with regards to identifying, labelling and grouping CustomFilters. For example, related phenotype terms or experimental sets could be grouped into local “dictionaries”, which could be addressed through a local identifier. The only requirement is that each custom filter term contains a unique identifier that can be used in Beacon requests.

- **FuzzyFilters** are implementations of classifiers which allow for some alternatives in matching and mostly can be drop-ins where ontologies are incomplete. Logically, through the potential matching of multiple values they provide a limited alternative mechanism to allow OR-style queries.
Beacon+ Flexible Modeling of New Features

Our Beacon platform is being used for the rapid testing of queries and responses - both v1.n and v2.0.a - against a number of partially large-scale genome datasets.

- Progenetix (>100000 cancer CNV profiles)
- DIPG (childhood brain tumor study)
- NEW: Cellosaurus ClinVar annotations for evidence representation
- Brewing: COVID-19

Currently running on a Perl+MongoDB stack, a Python-based OS solution is in early development.
A **Beacon** answers a query for a specific genome variant against individual or aggregate genome collections

**YES | NO | \0**
A Beacon network federates genome variant queries across databases that support the Beacon API.
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.
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".

Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?
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
**Beacon Protocol for Genomic Data Sharing**

Beacons provide discovery services for genomic data using the Beacon API developed as a key driver project of the Global Alliance for Genomics and Health (GA4GH). The Beacon protocol itself defines an open standard for genomics data discovery. It provides a framework for public web services responding to queries against genomic data collections, for instance from population based or disease specific genome repositories.

The original Beacon protocol had been designed to be:

- **Simple**: focus on robustness and easy implementation
- **Federated**: maintained by independent organisations and assembled into a network
- **General-purpose**: used to report on any variant collection
- **Aggregative**: provide a boolean (or quantitative) answer about the observation of a variant
- **Privacy preserving**: queries do not return information about single individuals

Sites offering beacons can scale through aggregation **Beacon Networks**, which distribute single genome queries among a potentially large number of international beacons and assemble their responses. Since 2015 the development of the Beacon protocol has been led by ELIXIR in collaboration with GA4GH and international participants. Recent versions of the Beacon protocol have expanded the original concept by e.g.: providing a framework for other types of genome variation data (i.e. range queries and structural variants)

- allowing for data delivery using **handover** protocol, e.g. to link with clinical information in protected environments and allow for data delivery and visualisation services

**Beacon v2 - Towards Flexible Use and Clinical Applications**

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

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

---

**Beacon**
<table>
<thead>
<tr>
<th>Duration</th>
<th>Title</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 min</td>
<td>Beacon - where are we now, and how did we get here?</td>
<td>Michael Baudis</td>
</tr>
<tr>
<td>15 min</td>
<td>ELIXIR Beacon Network: what does it offer to users?</td>
<td>Juha Tornroos</td>
</tr>
<tr>
<td>15 min</td>
<td>Beacon v2 - status, Use Cases, and implementations: examples from ES</td>
<td>Jordi Rambla</td>
</tr>
<tr>
<td>10 min</td>
<td>Beacon is secure: what does this mean, how does this interact with GA4GH, and how will this work with v2</td>
<td>Dylan Spalding</td>
</tr>
<tr>
<td>10 min</td>
<td>The Beacon API Compliance Tester: do the Beacons really speak the same language?</td>
<td>Malin Klang</td>
</tr>
<tr>
<td>5 min</td>
<td>Refreshment break</td>
<td></td>
</tr>
<tr>
<td>25 min</td>
<td>Beacon and Covid-19</td>
<td>Jordi Rambla</td>
</tr>
<tr>
<td>10 min</td>
<td>Beacon as a discovery tool for human data outside of genomics: Proteomics</td>
<td>Juan Antonio Vizcaino</td>
</tr>
<tr>
<td>15 min</td>
<td>Beacon in the ELIXIR Programme - Node perspective from FI</td>
<td>Juha Törnroos</td>
</tr>
</tbody>
</table>