ELIXIR Beacon

A driver project for the Global Alliance for Genomics and Health

Michael Baudis, University of Zurich | **SIB**

ECCB 2018 Athens

elixir

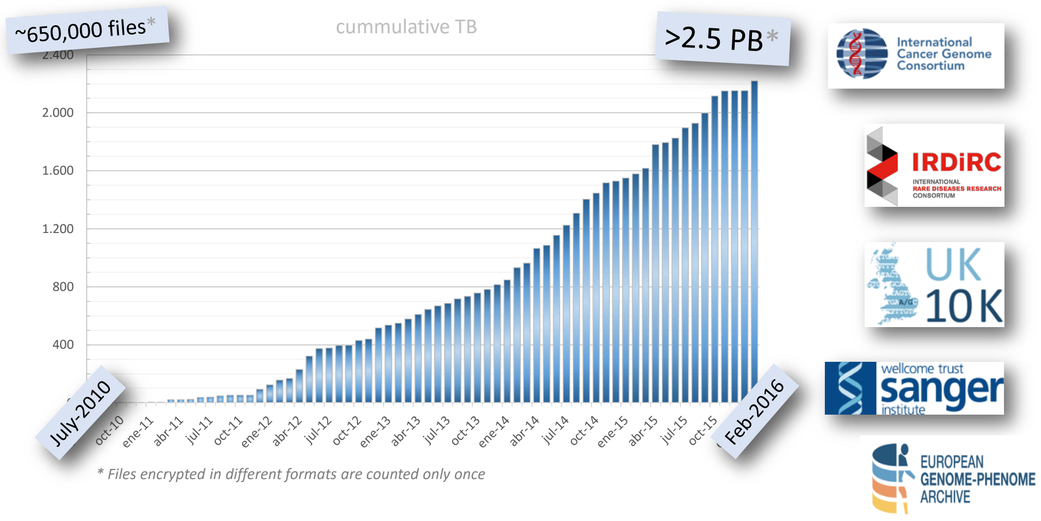
Global Alliance for Genomics & Health

SIB

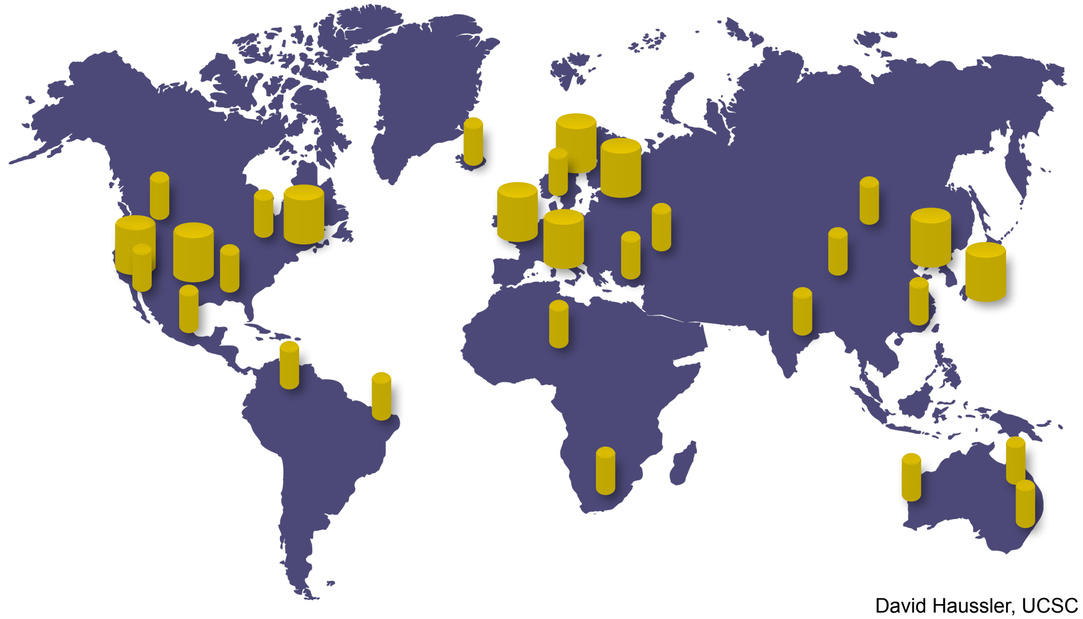
University of Zurich UZH

Genome Datasets: Rapid Growth, Limited Access

population based and cancer research studies produce a rapidly increasing amount of genome sequence data



Source: Jordi Rambla (CRG Barcelona) 2016



David Haussler, UCSC

genome data is stored in an increasing number of institutional and core repositories, with **incompatible data structures and access policies**

Global Alliance for Genomics & Health

Genomes Everywhere

Large Genome Data Generation, Analysis & Sharing Initiatives

Organization / Initiative: Name	Organization / Initiative: Category	Cohort
100K Wellness Project	Research Project	107 unaffected individuals (scaling up to 100,000)
23andMe	Organization	>1 million customers (>80% consented to research)
Actionable Cancer Genome Initiative (ACGI)	Data-Sharing Project	Goal: 100,000 individuals
Ancestry.com	Organization	1.4 million customer DNA samples (what % consented to research?)
BioBank Japan	Repository	Specimens from >200,000 patients and unaffected controls
Cancer Moonshot2020	Consortium	Phase 1: 20,000 cancer patients
Children's Hospital of Philadelphia Biorepository	Repository	Capacity for 8.6 million samples
China Kadoorie Biobank	Repository	>512,000 participants (general population, China). Genotyping data available for ~100,000.
CIMBA	Consortium	>15,000 BRCA1 carriers, >8,000 BRCA2 carriers
Clinical Sequencing Exploratory Research (CSER)	Consortium	~4,000 patients and healthy controls
DECIPHER	Repository	19,014 patients (international)
deCode Genetics	Organization	500,000 participants (international)
East London Genes & Health	Research Project	100,000 unaffected individuals (East London, Pakistani or Bangladeshi heritage)
Electronic Medical Records and Genomics (eMERGE) Network	Repository, Consortium, Research Project	55,028 patients
European Network for Genetic and Genomic Epidemiology (ENGAGE)	Research Project	80,000 GWAS scans, and DNA and serum/plasma from >600,000 individuals
Exome Aggregation Consortium (ExAC)	Consortium	60,706 individuals
GENIE/AACR	Data-Sharing Project	>17,000 cancer patients (international)
Genome Asia 100K	Consortium	Goal: 100,000 individuals (Asia)
Genomics England	Organization	Goal: 100,000 genomes from 70,000 individuals (rare disease & cancer patients, and their relatives)
GoT2D	Consortium, Data-Sharing Project	Multiple case-control cohorts
International Cancer Genome Consortium (ICGC)	Consortium	PCAWG currently data from >8'000 genomes
International Genomics of Alzheimer's Project (IGAP)	Consortium	40,000 patients with Alzheimer's disease
International Multiple Sclerosis Genetics (IMSG) Consortium	Consortium	Goal: >50,000 patients with MS
Kaiser Permanente: Genes, Environment, and Health (RPGEH)	Repository, Research Project	200,000 DNA samples (scaling up to 500,000)
Leiden Open Variation Database (LOVD)	Repository	>170,000 individuals
Million Veteran Program	Research Project	Goal: 1 million individuals; first 200,000 is complete.
MyCode® Community Health Initiative	Repository, Research Project	Goal: >250,000 patients
Precision Medicine Initiative	Research Project	Goal: >1 million participants, starting in 2016 (US)
Psychiatric Genomics Consortium (PGC)	Consortium	>170,000 subjects
Resilience Project	Research Project	589,306 individuals
Saudi Human Genome Program	Research Project	Goal: ~100,000 patients and controls (Saudi Arabia)
Scottish Genomes Partnership (SGP)	Research Project	>3,000 individuals (Scotland)
T2D-GENES	Consortium, Data-Sharing Project	10,000 patients and controls (five ethnicities); 600 individuals (Mexican American)
TBResist	Consortium	>2,600 samples
UK Biobank	Repository, Consortium, Research Project	500,000 individuals (age 40-69 years; UK)
UK10K	Research Project	10,000 participants (6,000 patients and 4,000 controls)
Vanderbilt's BioVU	Repository	>215,000 samples

>1'000'000

100'000

20'000+

17'000+

8'000+

>2-500'000

>1'000'000

500'000

Enabling genomic data sharing for the benefit of human health

The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework

Genomic Data Toolkit →

Regulatory & Ethics Toolkit →

Data Security Toolkit →

VIEW OUR LEADERSHIP

MORE ABOUT US

BECOME A MEMBER

Global Alliance for Genomics & Health

- January 2013 - 50 participants from eight countries
- June 2013 - White Paper, over next year signed by 70 “founding” member institutions
- **March 2014 -** Working group meeting in Hinxton & **1st plenary in London**
- **October 2014 - 2nd Plenary, 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**


GENOMICS

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*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

elixir

GA4GH API promotes sharing

GENOMICS

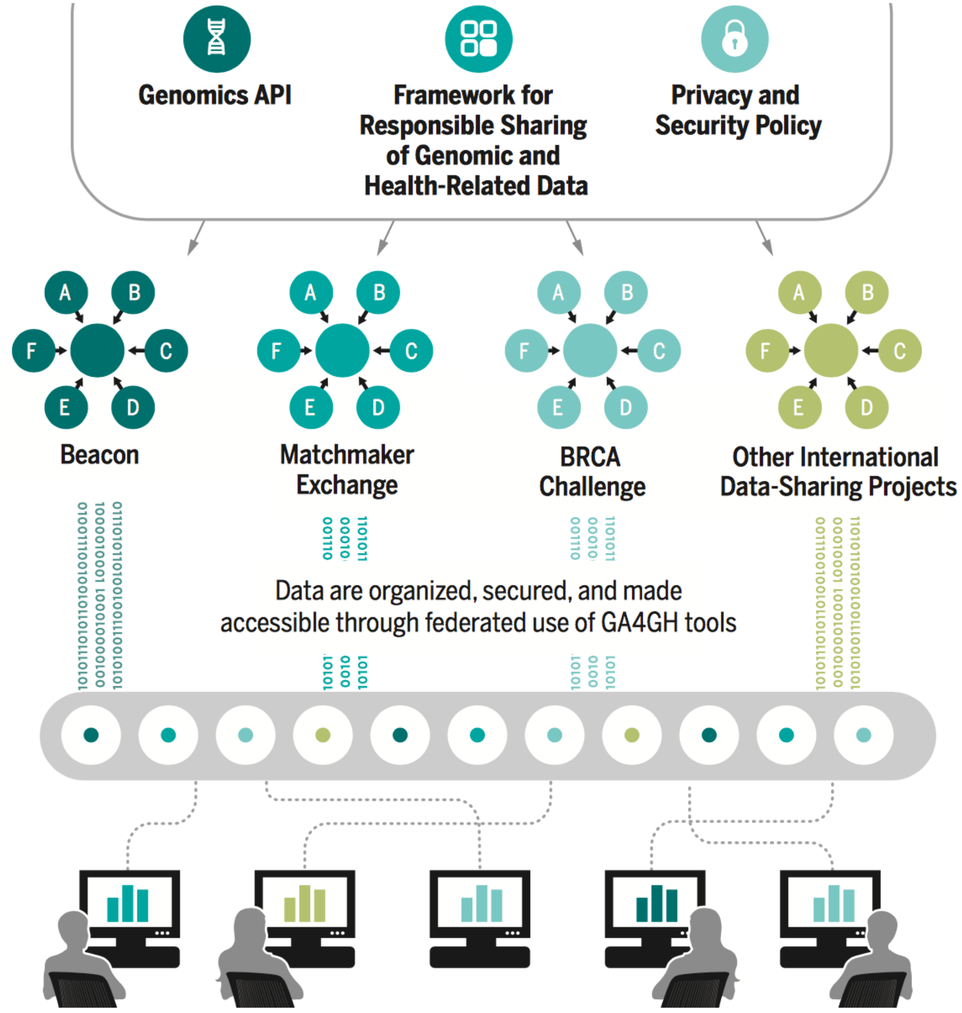
A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

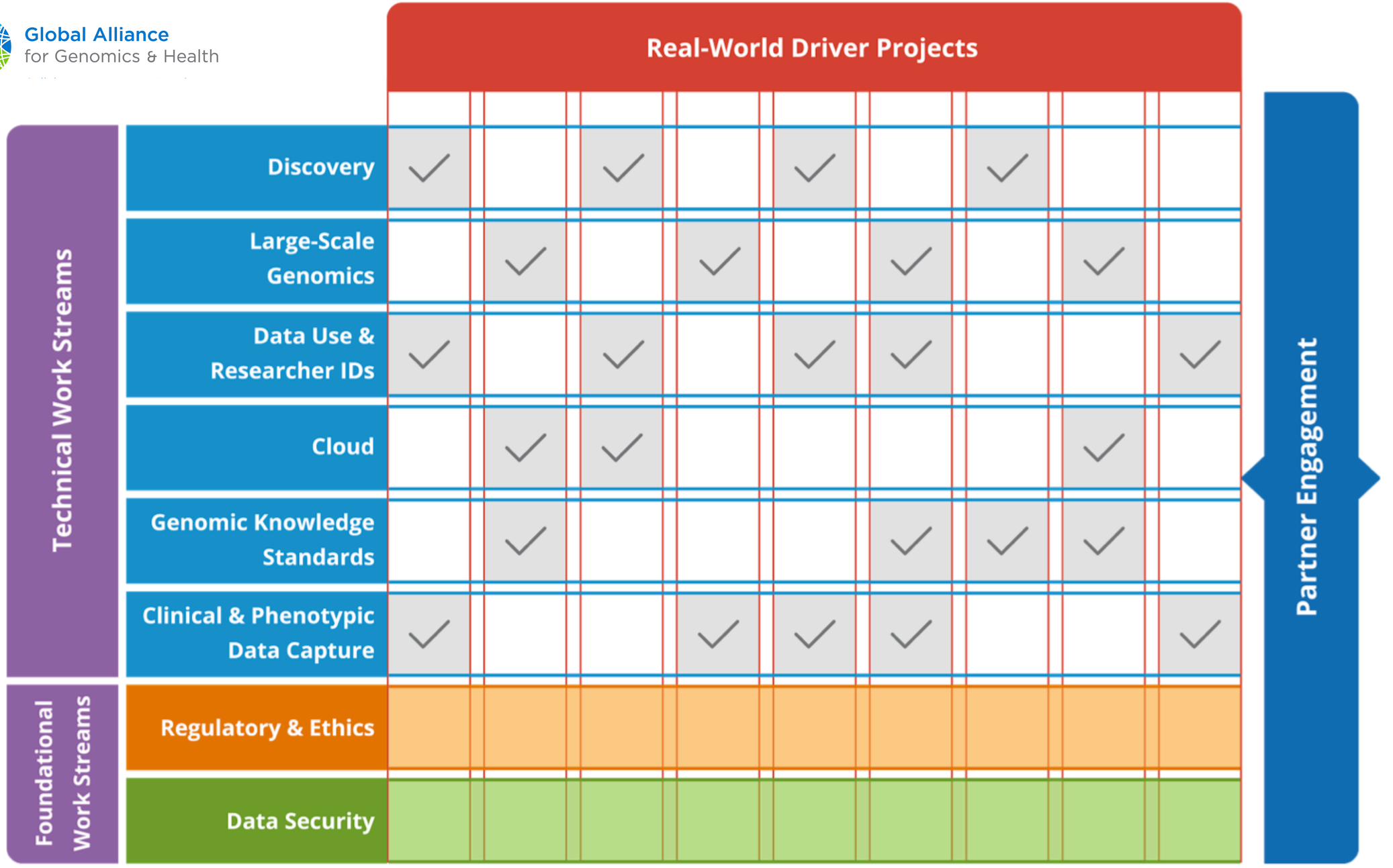
SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

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



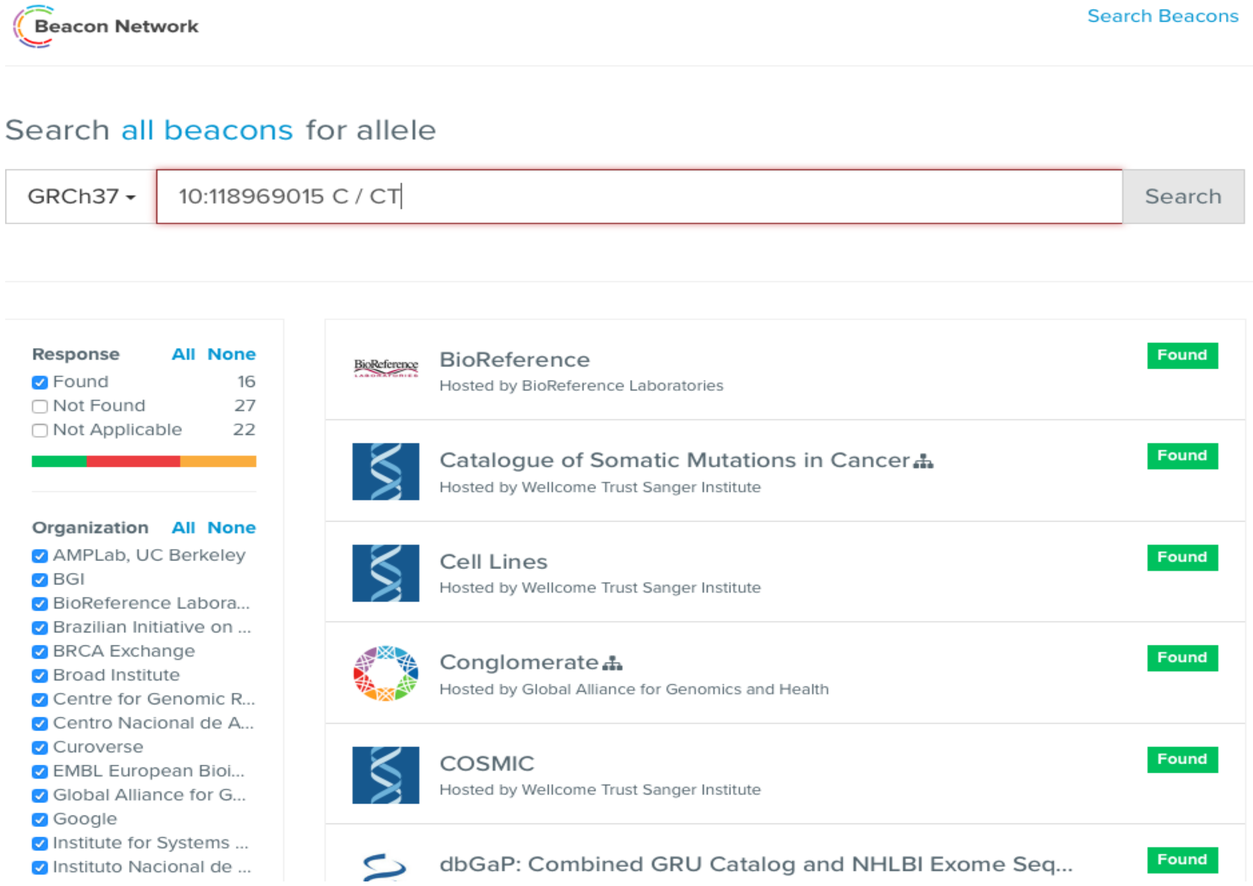
Global Alliance for Genomics & Health

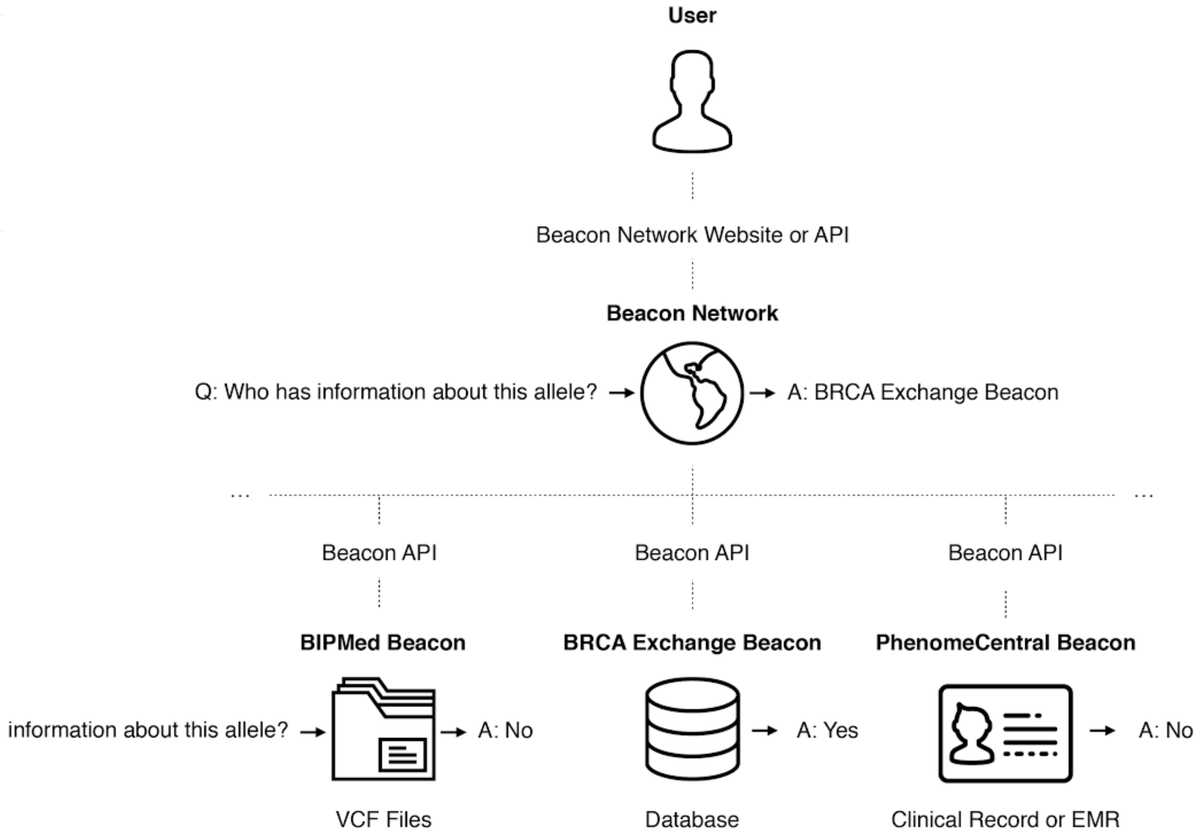
Global Alliance for Genomics & Health

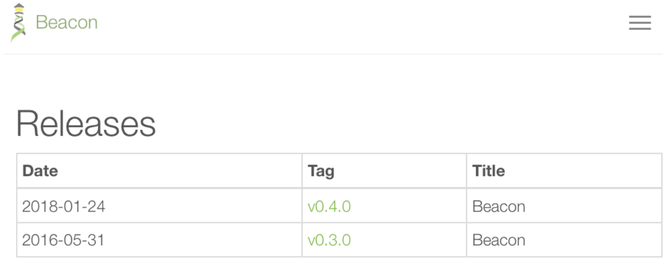


Beacon Project

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







Global Alliance for Genomics & Health

35+ Organizations

90+ Beacons

200+ Datasets

100K+ Individuals

ELIXIR Beacon Project

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

Beacon *forward*

- ☑ **structural variations** (DUP, DEL) in addition to SNV
- … more structural queries (translocations/fusions…)
- (bio-) **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)

Global Alliance for Genomics & Health

ELIXIR Beacon Project 2018

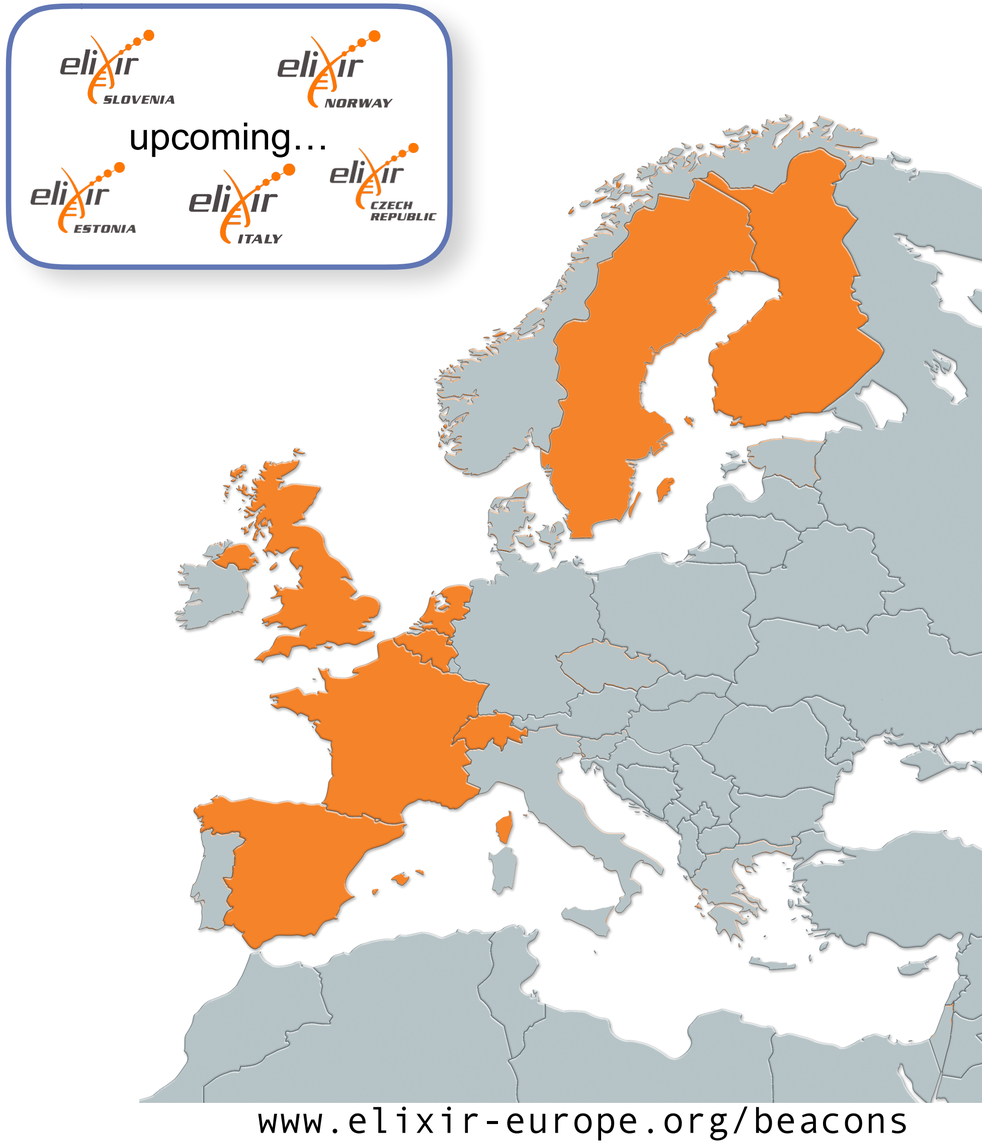
- WP1 - Development of Beacon API specification with the GA4GH
- WP2 - Supporting new ELIXIR Beacon development and deployment across Europe
- WP3 - Implementation of an ELIXIR Beacon Network & Registry
- WP4 - Security of ELIXIR Beacon and Beacon Network
- WP5 - Developing Indicators to establish Registry and Beacon as an Emerging ELIXIR Service
- WP6 - Strategic partnerships with national data cohorts and biobanks
- WP7 - Project Management and Communication

Node	Name of PI	Role	PMs	Other Contrib	Work Packages						
					1	2	3	4	5	6	7
FI	Ilkka Lappalainen	co-lead	5		X	X	X	X	X	X	X
ES	Jordi Rambla	lead	5		X	X	X				X
CH	Michael Baudis, (Heinz Stockinger)	co-lead	5		X					X	X
EBI	Dylan Spalding	co-lead	5		X	X	X	X			
SE	Niclas Jareborg	member	1			X					
BE	Yves Moreau	member	1			X					
FR	Macha Nikolski	member	1			X					
NL	Morris Swertz	member	1			X					
IT	David Horner	member	-								
FI	ELIXIR Compute Platform - Tommi Nyrönen			2							
TBD[6]	ELIXIR Training Platform			1		X					
HUB	Serena Scollen, Susheel Varma	member coordinator	0.5		X				X	X	X

Node work package participation, as in 2018 project plan

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

Do Genome Beacons Compromise Security?

- Querying for thousands of specific SNV occurrences in a genomic data pool can identify individuals

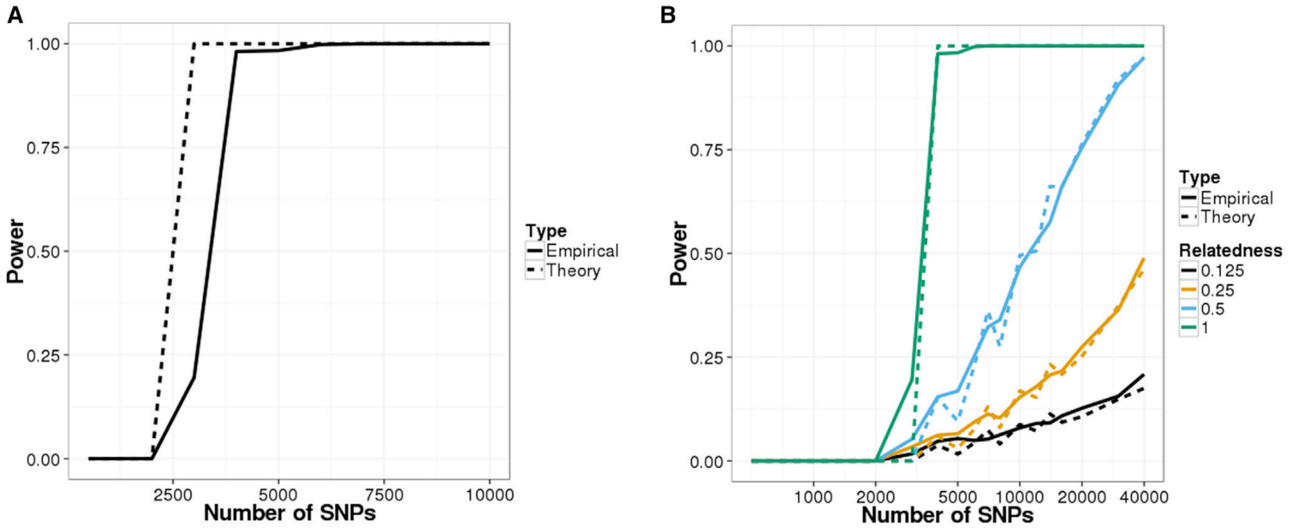


Figure 1. Power of Re-identification Attacks on Beacons Constructed with Simulated Data
Power curves for the likelihood-ratio test (LRT) on (A) a simulated beacon with 1,000 individuals and (B) detecting relatives in the simulated beacon. The false-positive rate was set to 0.05 for all scenarios.

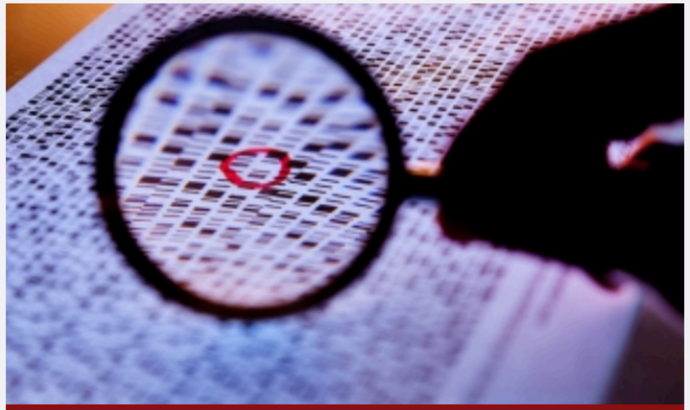
Suyash S. Shringarpure, Carlos D. Bustamante. Privacy Risks from Genomic Data-Sharing Beacons. AJHG 97:5, Pages 631-646 (Nov. 2015)

Stanford researchers identify potential security hole in genomic data-sharing network

Hackers with access to a person's genome might find out if that genome is in an international network of disease databases.

OCT 29
2015

Sharing genomic information among researchers is critical to the advance of biomedical research. Yet genomic data contains identifiable information and, in the wrong hands, poses a risk to individual
meone had access to your genome
ither directly from your saliva or
, or from a popular genomic
service — they could check to see if
n a database of people with certain
ditions, such as heart disease, lung
tism.



Stanford researchers are working with the Global Alliance for Genomics and Health to make genomic information in the Beacon Project more secure.
Science photo/Shutterstock

ir of researchers at the Stanford
hool of Medicine makes that genomic
cure. Suyash Shringarpure, PhD, a
scholar in genetics, and Carlos Bustamante, PhD, a professor of genetics, have
d a technique for hacking a network of global genomic databases and how to prevent it.
king with investigators from the Global Alliance for Genomics and Health on implementing
easures.

The work, published Oct. 29 in *The American Journal of Human Genetics,* also bears importantly on the larger question of how to analyze mixtures of genomes, such as those from different people at a crime scene.

Countering real (or perceived) risk of Beacon identification attacks

- authenticated access
  - difficult to implement in truly international setting for federated queries
  - high cost on ease of use/utility of Beacon concept
- various risk-mitigation strategies…

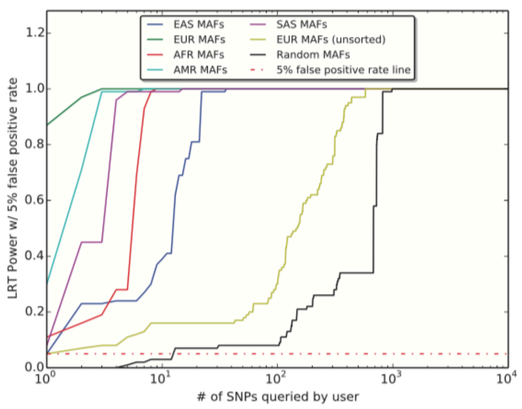


Figure 1. "Optimal" re-identification attack in single-population beacon. Different power rates per number of SNPs queried from an unprotected beacon with a single population (EUR) by an adversary with different types of background knowledge: (green) the attacker knows the allele frequencies (AFs) of a population from the same ancestry (EUR) as the one in the beacon and performs queries following the rare-allele-first logic; (red, cyan, blue, and purple) the attacker knows the AFs of a population from an ancestry different from the one in the beacon and performs queries following the rare-allele-first logic (African [AFR], admixed American [AMR], East Asian [EAS], or South Asian [SAS], respectively); (yellow) the attacker knows the AFs of a distinct population with the same ancestry (EUR) other than the one in the beacon but performs queries in random order; (black) the attacker does not have any information on AFs (i.e., the original attack by Shringarpure and Bustamante[11]).

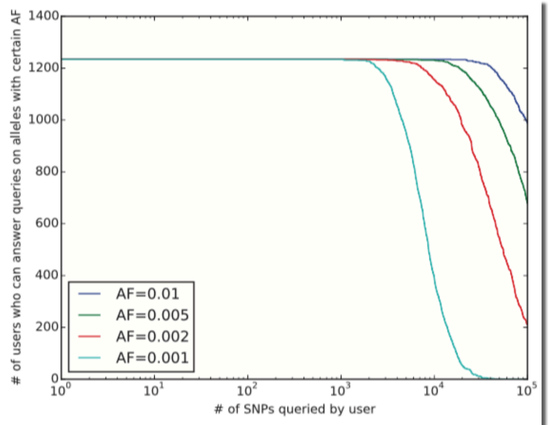
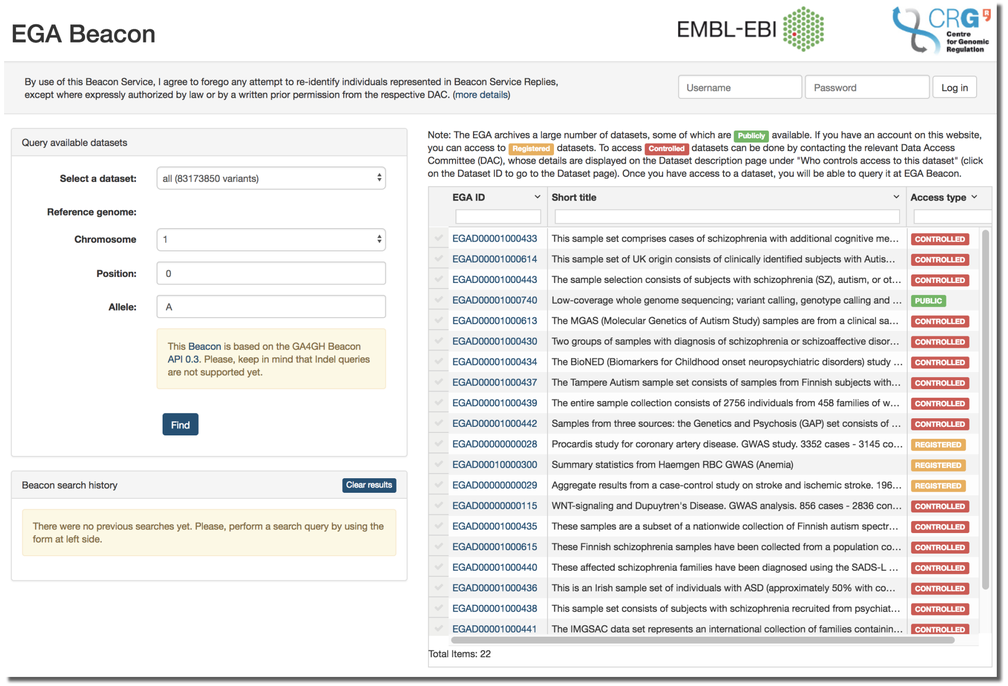


Figure 5. Budget evaluation in beacon with *S3*. Behaviors of individual budgets per number of SNPs queried according to the typical user's query profile obtained from ExAC log data. The cyan curve represents the number of individuals with enough budget to answer "Yes" to queries targeting alleles with AF = 0.001. Red, green, and blue curves correspond to 0.002, 0.005, and 0.01, respectively.



Risk mitigation strategy	Disadvantages	Advantages
S1: Beacon alteration	Eliminates possibility of querying for unique alleles highly likely to be most useful in genetic research	Protects privacy of individuals possessing variants most likely to be targeted by attackers
S2: Random flipping	Decreases rate of true answers returned from querying unique alleles likely to be useful in genetic research	Permits some unique alleles to be discoverable and to fine-tune the privacy–utility trade-off
S3: Query budget per individual	Requires the assumption of Beacon user being nonanonymous and holding no more than one Beacon account; may require complicated accounting scheme	Enables all alleles to be discoverable until budget is exceeded

Raisaro JL, Tramèr F, Ji Z, Bu D, Zhao Y, Carey K, Lloyd D, Sofia H, Baker D, Flicek P, Shringarpure S, Bustamante C, Wang S, Jiang X, Ohno-Machado L, Tang H, Wang X, Hubaux JP. Addressing Beacon re-identification attacks: quantification and mitigation of privacy risks . Journal of the American Medical Informatics Association, 2017, Vol. 24, No. 4

Security and GDPR

- Requirements analysis document, outlining **re-identification mitigation**
  - Individual Beacon and network of Beacons
- GA4GH Data **Security questionnaire** completed
- GDPR workshop in Brussels:
  - Raw data subject to recitals 33 (scientific research) and 34 (genetic data)
  - GDPR not applicable to anonymous data (recital 26)
- Genetic data kept anonymous by re-identification mitigation techniques
- Access restrictions (**tiered Beacon access**) can be adapted for each node
  - Accounts for local legal interpretations

University of Zurich UZH

Beacon+

Beacon+ - Extending the protocol ...

beacon.progenetix.org

This forward looking Beacon interface implements additional, planned features beyond the current GA4GH specifications. Info

Query

SNV Example | DGV Example | CNV Example

Dataset	tcga
Reference name*	9
Genome Assembly*	GRCh38 / hg
Start *min* Position*	19,500,000
Start *max* Position	21,975,098
End *min* Position	21,967,753
End *max* Position	24,500,000
Alt. Base(s)*	DEL
Bio-ontology	icdot:c50.9: (4065)

Beacon Implementations

- implementing existing resources with Beacon protocol
- e.g. TCGA cancer variants (structural and SNV)

Beacon Response

- quantitative (counts for variants, callsets and samples)
- *Handover* to authentication system for data retrieval
- **no exposure** of data beyond standard Beacon response and additional pointer to matched data

Prototyping Query Extensions

- testing e.g. bio-metadata queries using ontology terms

Dataset	Assembly	Chro	Start Range	End Range	Pos	Ref Alt	Bio Query	Variants Calls Samples	$f_{alleles}$	Response Context
tcga	hg38	9	19,500,000 21,975,098	21,967,753 24,500,000		DEL	icdot:c50.9	54 54 54	0.0243	JSON UCSC Handover

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

arrayMap progenetix

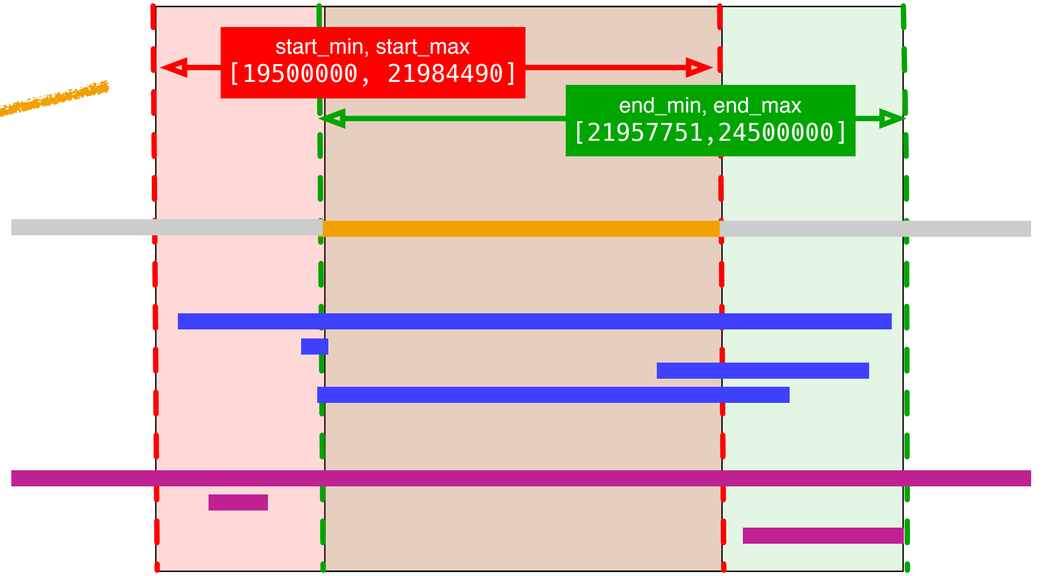
This Beacon implementation is developed by the Computational Oncogenomics Group at the University of Zurich, with support from the SIB Technology group and ELIXIR.

University of Zurich UZH elixir SIB

elixir

```
{
   "allele_request" : {
      "and" : [
        { "reference_name" : “9" },
        { "variant_type" : “DEL" },
        { "start" : { "$gte" : 19500000 } },
        { "start" : { "$lte" : 21984490 } },
        { "end" : { "$gte" : 21957751 } },
        { "end" : { "$lte" : 24500000 } }
      ]
   },
   "api_version" : "0.4",
   "beacon_id" : "org.progenetix:progenetix-beacon",
   "exists" : true,
   "info" : {
   "url" : “http://progenetix.org/beacon/info/“,
   "dataset_allele_responses" : [
      {
         “dataset_id" : "arraymap",
         "error" : null,
         "exists" : true,
         "external_url" : “http://arraymap.org",
         "sample_count" : 584,
         "call_count" : 3781,
         "variant_count" : 3244,
         "frequency" : 0.0094,
         "info" : {
           "description" : "The query was against database
\"arraymap_ga4gh\", variant collection \"variants_cnv_grch36\". 3781 /
59428 matched callsets for 3602919 variants. Out of 62105 biosamples in
the database, 2047 matched the biosample query; of those, 584 had the
variant.",
             "ontology_ids" : [
               "ncit:C3058",
               "pgx:icdom:9440_3",
               "pgx:icdot:C71.9",
               “pgx:icdot:C71.0"
             ]
         },
      }
   ],
}
```


start_min
start_max
end_min
end_max



Match using query ranges “at least one base in interval affected”

Region of Interest, e.g. CDR of Gene (here: CDKN2A)

Example “focal” matches (overlap w/ size limit)

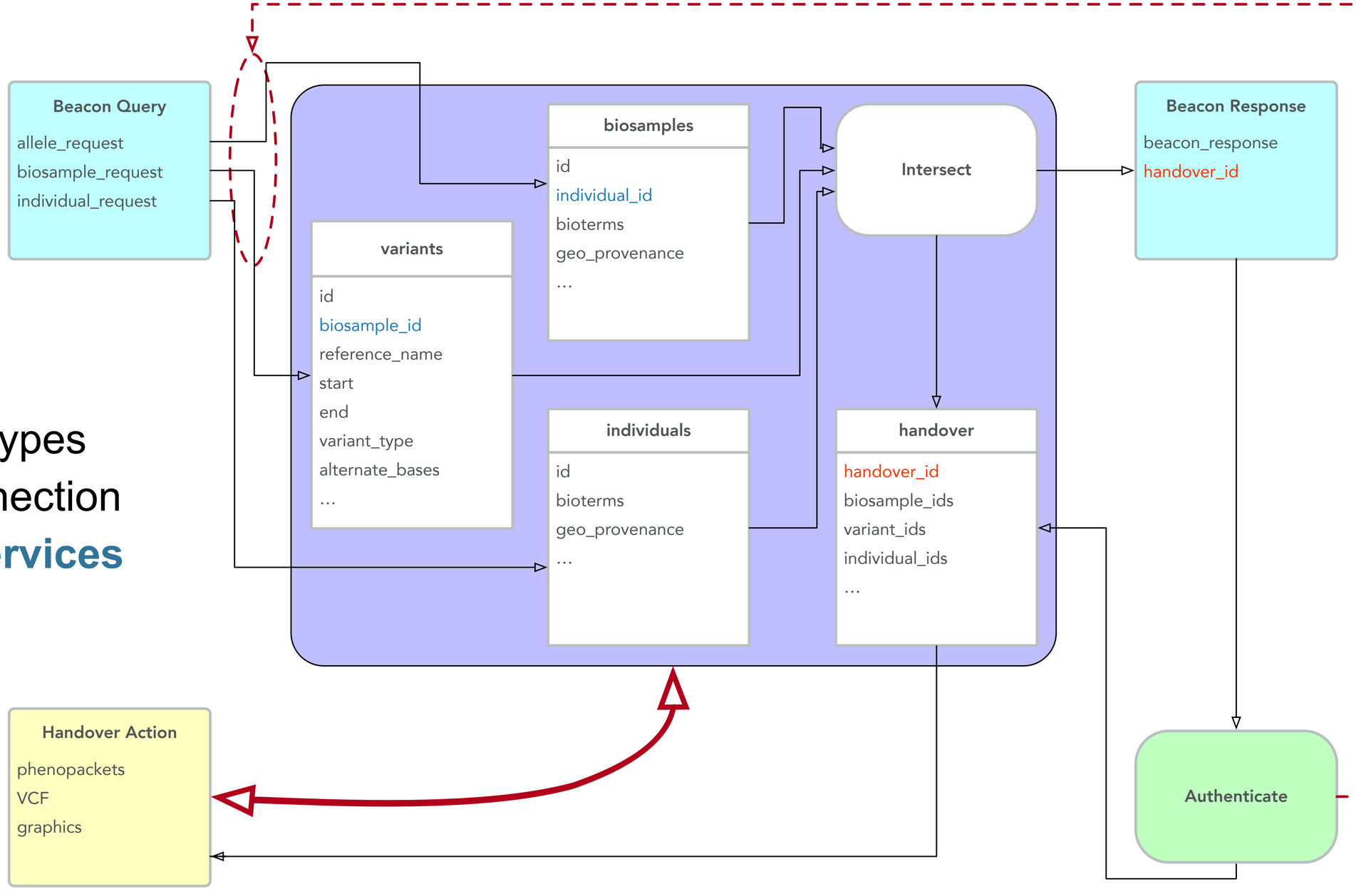
Mismatches
- too large
- end outside
- start outside

Metadata

- 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

Beacon & Handover

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



Beacon I/O

Handover

Authentication

Global Alliance for Genomics & Health

SIB

University of Zurich UZH

elixir

Beacon Handover Demonstrator

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


```
   "dataset_allele_responses" : [
      {
         "dataset_id" : "arraymap",
         "call_count" : 171,
         "sample_count" : 171,
         "variant_count" : 157,
         "error" : null,
         "exists" : true,
         "external_url" : "http://beacon.progenetix.org",
         "frequency" : 0.1557,
         "info" : {
            "callset_access_handle" : "d5850347-d411-11e7-8c89-ec436516cb41",
            "description" : "The query was against database \"arraymap_ga4gh\",
variant collection \"variants_cnv_grch36\". 171 matched calls for 157 distinct
variants. Out of 62033 biosamples in the database, 1098 matched the biosample
query; of those, 171 had the variant.",
         },
         "note" : "",
      }
   ],
```


Global Alliance for Genomics & Health · SIB · University of Zurich UZH · elixir

Beacon query => Handover Handle => Authentication => Data Retrieval

Beacon+ example implementation using public somatic variation data

Beacon+

This is an implementation of a Beacon "handover" concept, in which a Beacon query response additionally deli
representation of the query resulte (i.e. callsets, biosamples, metadata ...), which can then be accessed after au
("yes"|"no") or quantitative ("n matches") Beacon response from a data delivery mechanism.

The current implementation exemplifies some possible scenarios:

- providing a histogram of regional gain/loss frequencies (DUP, DEL) for samples with structural variation d
- returning data of the associated callsets which matched the Beacon query (this is for feature demonstrat
- returning the metadata (diagnoses etc.) of the biosamples from which the matching callsets were derivec

This demonstrator does not implement authentication procedures yet; login & password fields can be left empt

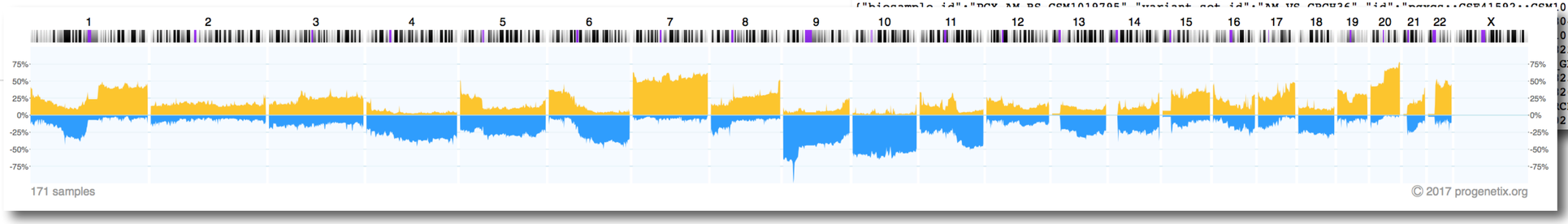
Handover Action

- ✓ Plot DUP/DEL histogram
- Export Callset Data
- Export Biosample Data

Login

Password ••••••••••••••••

Process Data



arrayMap progenetix

Beacon Handover Demonstrator

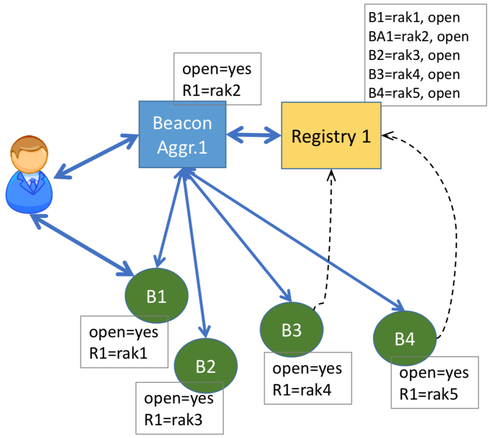
- only exposure of access handle to data stored in secure system
- here one-step authentication and selection of *handover* action; other scenarios possible / likely
- *handover* **response not managed by Beacon** protocol / system => "Discovery" protocol?

Global Alliance for Genomics & Health

elixir

Towards a shining future...

- extensive/complete representation of genome **variant types** in query
  - close coordination with **GA4GH::GKS** and **GA4GH::CP** work streams
- providing a tested model for layered **registered access**
  - ELIXIR AAI
- implementing **Beacon network**(s) throughout ELIXIR
  - open protocols for extension and external implementations
- extending Beacon query protocol (**metadata**...)
  - keeping "aggregate response" model
- Beacon queries as entry points for data delivery, using "**handover**" scenarios



Global Alliance
for Genomics & Health

Thank You!

... and many more!

Global Alliance
for Genomics & Health

Global Alliance
for Genomics & Health

6th GA4GH Plenary

- 2nd GA4GH plenary in continental Europe
- public and workgroup parts
- collaboration, data sharing, standards for files, formats, APIs and procedures (ethics, security)

Andrew Morris
Director, Health Data Research UK

Nicola Mulder
Professor, University of Cape Town

Torsten Schwede
Vice Rector, University of Basel

October 3 - 5, 2018

Congress Centre Basel, Basel, Switzerland

Wednesday, Oct 3

- Intro to GA4GH
- Technical and Clinical Workshops
- GenoPri

Thursday, Oct 4

- GA4GH Connect: The Story So Far
- Responsible genomic data sharing

Friday, Oct 5

- National genomic data initiatives
- Panels on patient perspectives, industry, and standards development

More information at **bit.ly/6thplenary**

Meeting Sponsors