

Genome Beacons for Data Discovery

Technical advances in a challenging environment

Michael Baudis - dpph18.epfl.ch - 2018-02-15



University of
Zurich^{UZH}



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	currently data from >16'000 samples
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





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](#)

GA4GH API promotes sharing

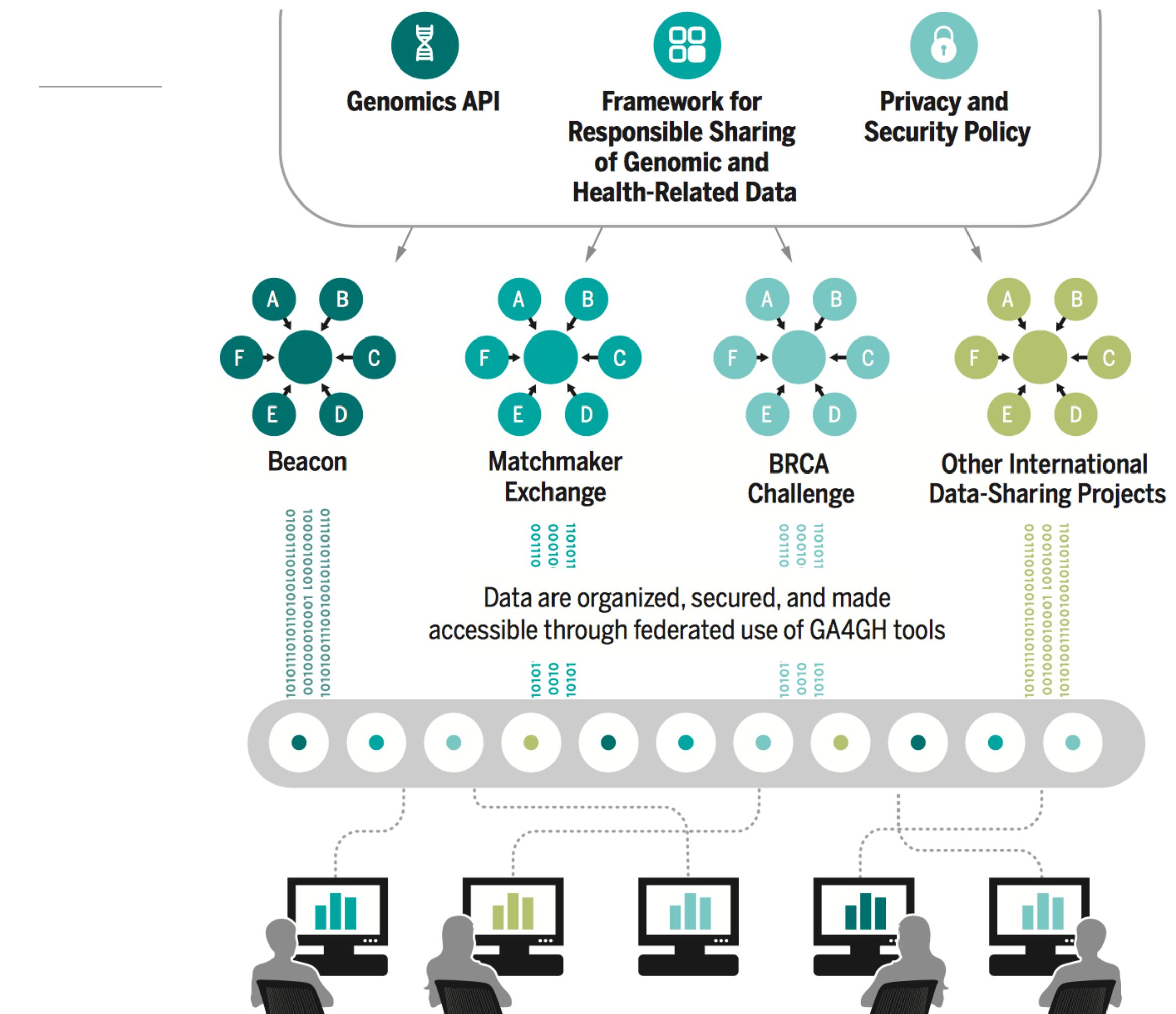
A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.



GENOMICS

A federated ecosystem for sharing genomic, clinical data

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



Beacon Project

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



Beacon Network

Search Beacons

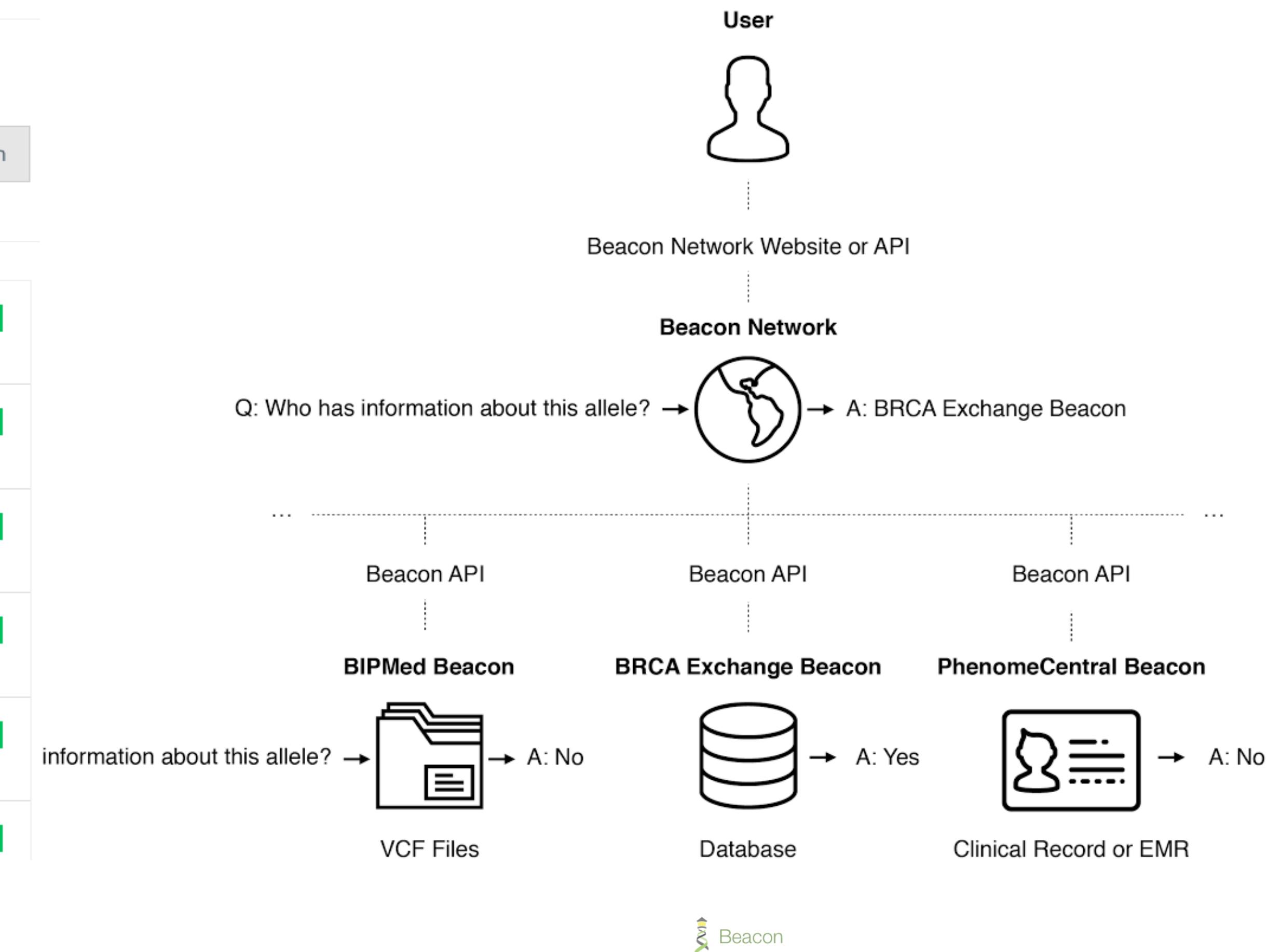
Search all beacons for allele

GRCh37 ▾ 10:118969015 C / CT | Search

Response All None
 Found 16
 Not Found 27
 Not Applicable 22

Organization All None
 AMPLab, UC Berkeley
 BGI
 BioReference Laborato...
 Brazilian Initiative on ...
 BRCA Exchange
 Broad Institute
 Centre for Genomic R...
 Centro Nacional de A...
 Curoverse
 EMBL European Bio...
 Global Alliance for G...
 Google
 Institute for Systems ...
 Instituto Nacional de ...

BioReference	Hosted by BioReference Laboratories	Found
Catalogue of Somatic Mutations in Cancer	Hosted by Wellcome Trust Sanger Institute	Found
Cell Lines	Hosted by Wellcome Trust Sanger Institute	Found
Conglomerate	Hosted by Global Alliance for Genomics and Health	Found
COSMIC	Hosted by Wellcome Trust Sanger Institute	Found
dbGaP: Combined GRU Catalog and NHLBI Exome Seq...		Found



Date	Tag	Title
2018-01-24	v0.4.0	Beacon
2016-05-31	v0.3.0	Beacon

Beacon Roadmap

- **structural variations** (DUP, DEL) in addition to SNV
- **metadata** queries using ontology codes (e.g. NCIT, ICD-O)
- layered **authentication** system using **ELIXIR** AAI
- prototyping **handover** concept
- **quantitative** responses
- Ubiquitous **deployment** (e.g. throughout **ELIXIR** network)

Beacon+

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

Query

Dataset: DIPG (CNV + selected SNV)
Reference name*: 17
Genome Assembly*: GRCh36 / hg18
Variant type*: SNV / indel
Position*: 7577121
Ref. Base(s)*: G
Alt. Base(s)*: A
Bio-ontology: pgx:icdom:9380_3

Beacon Query

Response

Dataset	Chro.	Assembly	Var. Type	Start Min	Start Max	End Min	End Max	Pos.	Ref.	Alt.	Bio Query	Call Count	Samples	f	Query
arraymap	9	GRCh36	DEL	19000000	21984490	21900000	25000000				pgx:icdom:8140_3	3781	403	0.0065	show JSON
dipg	17	GRCh36	SNV			7577121		G	A	pgx:icdom:9380_3	21	20	0.0187	show JSON	

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



Global Alliance
for Genomics & Health

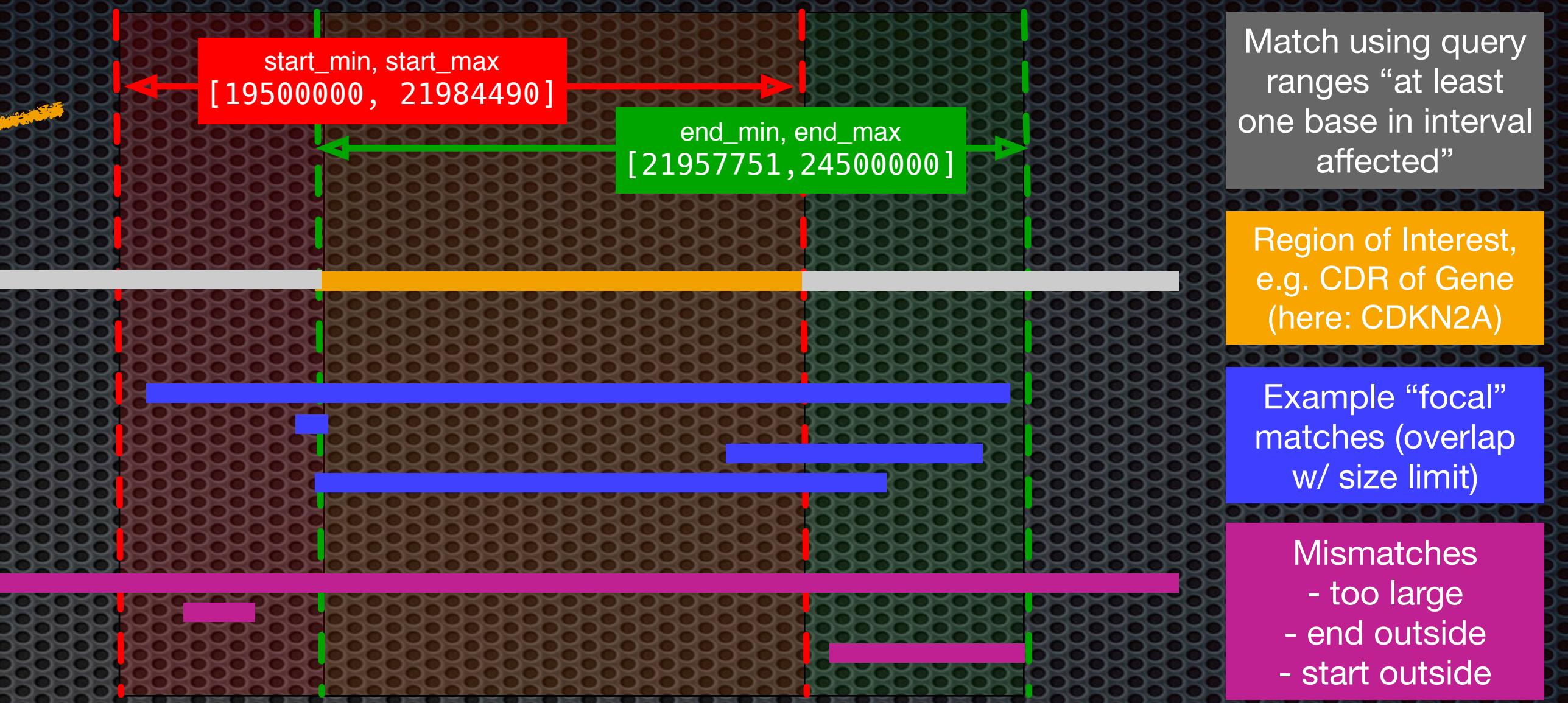


```

{
  "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"
        ],
      }
    ],
  }
}

```

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>,), as are specified in VCF && GA4GH variant schema





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

[Info](#)
Query
[SNV Example](#)
[DGV Example](#)
[CNV Example](#)
Dataset

arraymap

Reference name*

9

Genome Assembly*

GRCh36 / hg18

Variant type*

DEL (Deletion)

Start min Position*

19,500,000

Start max Position

21,964,826

End min Position

21,958,228

End max Position

24,500,000

Bio-ontology

ncit:c3224: Melanoma (1098)

Beacon Query
Response

Dataset	Assembly	Chro	Var Type	Start Range	End Range	Pos	Ref Alt	Bio Query	Variants	Calls	Samples	f_alleles	Response Context
arraymap	hg18	9	DEL	19,500,000 21,964,826	21,958,228 24,500,000			ncit:c3224	157 171 171			0.1557	JSON UCSC Handover

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

Variants	Calls	Samples	f_alleles	Response Context
157				JSON UCSC Handover



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.arraymap.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" : "", }],
```



# 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 defines a representation of the query result (i.e. callsets, biosamples, metadata ...), which can then be accessed after ("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
- returning data of the associated callsets which matched the Beacon query (this is for feature demonstration)
- returning the metadata (diagnoses etc.) of the biosamples from which the matching callsets were derived

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

**Handover Action**

Plot DUP/DEL histogram

Export Callset Data

Export Biosample Data

**Login**

**Password**

••••••••••••••••

**Process Data**

arrayMap

progenetix

171 samples

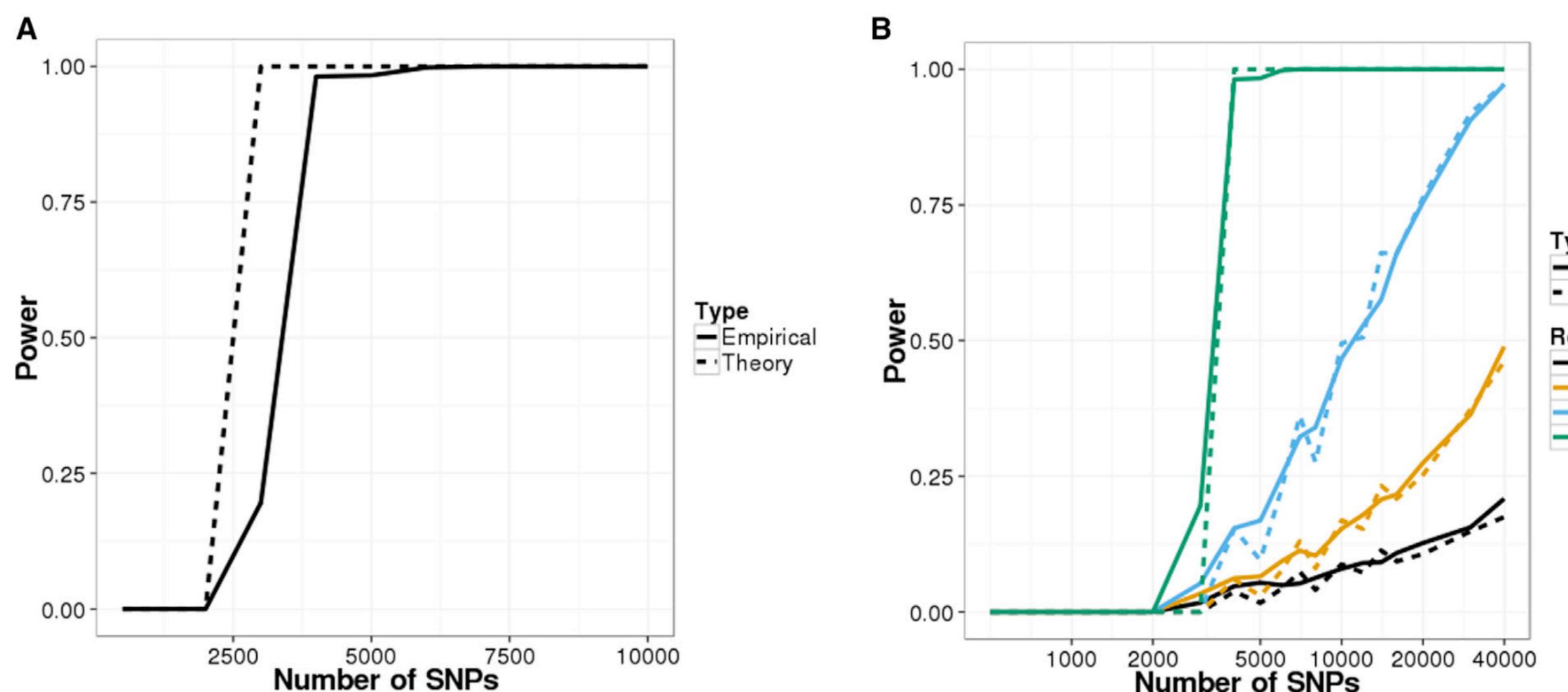
# Beacon Handover Demonstrator

- 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?



# 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.

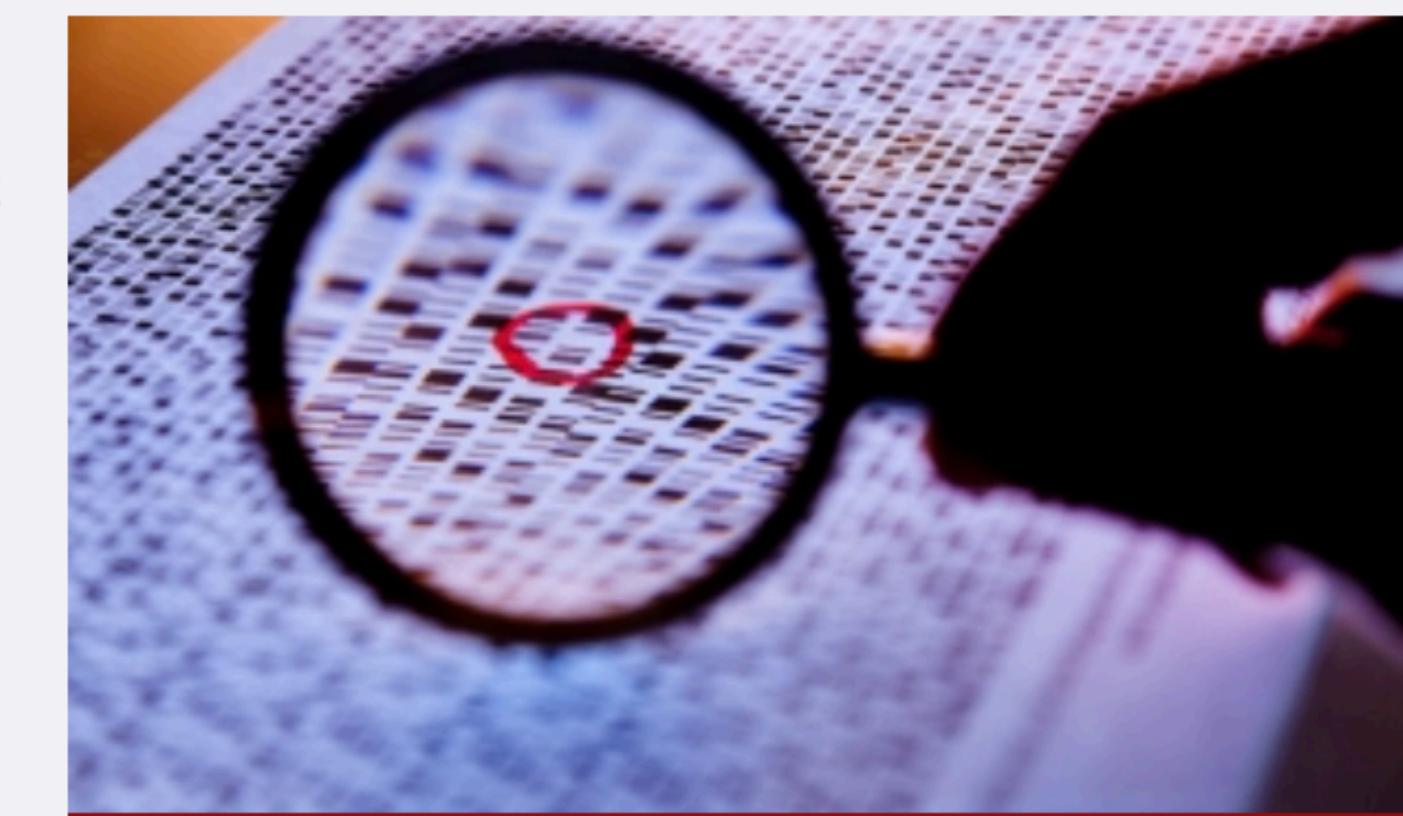
Stanford researchers identify potential security hole in genomic data-sharing network

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 privacy. If someone had access to your genome either directly from your saliva or from a popular genomic service — they could check to see if you're in a database of people with certain conditions, such as heart disease, lung cancer.

A team of researchers at the Stanford School of Medicine makes that genomic data secure. Suyash Shringarpure, PhD, a postdoctoral scholar in genetics, and Carlos Bustamante, PhD, a professor of genetics, have developed a technique for hacking a network of global genomic databases and how to prevent it. They are working with investigators from the Global Alliance for Genomics and Health on implementing security measures.



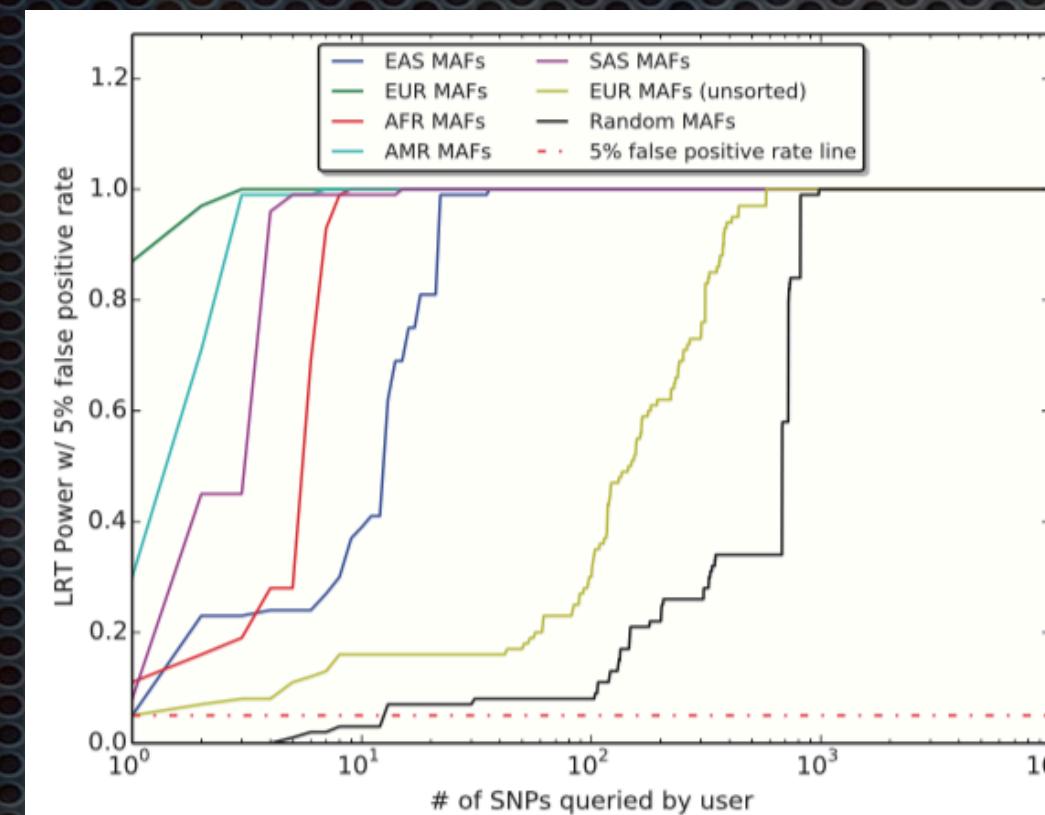
Stanford researchers are working with the Global Alliance for Genomics and Health to make genomic information in the Beacon Project more secure.  
Science photo/Shutterstock

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.

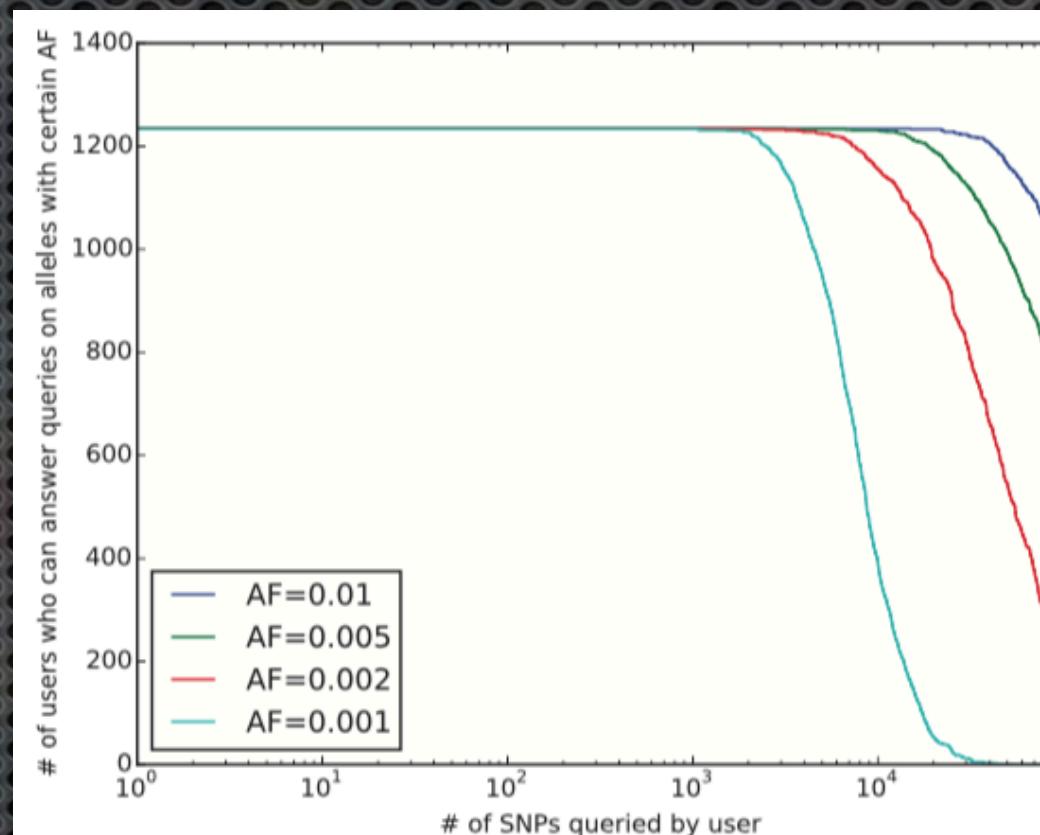


# Countering real (or perceived) risk of Beacon identification attacks

- 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<sup>11</sup>).



**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.

### 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: all (83173850 variants)

Reference genome: 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.

**Find**

Beacon search history Clear results

There were no previous searches yet. Please, perform a search query by using the form at left side.

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 |
|------------------|---------------------------------------------------------------------------------------|-------------|
| EGAD00001000433  | This sample set comprises cases of schizophrenia with additional cognitive me...      | CONTROLLED  |
| EGAD00001000614  | This sample set of UK origin consists of clinically identified subjects with Autis... | CONTROLLED  |
| EGAD00001000443  | The sample selection consists of subjects with schizophrenia (SZ), autism, or ot...   | CONTROLLED  |
| EGAD00001000740  | Low-coverage whole genome sequencing; variant calling, genotype calling and ...       | PUBLIC      |
| EGAD00001000613  | The MGAS (Molecular Genetics of Autism Study) samples are from a clinical sa...       | CONTROLLED  |
| EGAD00001000430  | Two groups of samples with diagnosis of schizophrenia or schizoaffective disor...     | CONTROLLED  |
| EGAD00001000434  | The BioNED (Biomarkers for Childhood onset neuropsychiatric disorders) study ...      | CONTROLLED  |
| EGAD00001000437  | The Tampere Autism sample set consists of samples from Finnish subjects with...       | CONTROLLED  |
| EGAD00001000439  | The entire sample collection consists of 2756 individuals from 458 families of w...   | CONTROLLED  |
| EGAD00001000442  | Samples from three sources: the Genetics and Psychosis (GAP) set consists of ...      | REGISTERED  |
| EGAD000000000028 | Procardis study for coronary artery disease, GWAS study, 3352 cases - 3145 co...      | REGISTERED  |
| EGAD00001000300  | Summary statistics from Haemgen RBC GWAS (Anemia)                                     | REGISTERED  |
| EGAD000000000029 | Aggregate results from a case-control study on stroke and ischemic stroke. 196...     | REGISTERED  |
| EGAD000000000115 | WNT-signaling and Dupuytren’s Disease. GWAS analysis, 856 cases - 2836 co...          | CONTROLLED  |
| EGAD00001000435  | These samples are a subset of a nationwide collection of Finnish autism spectr...     | CONTROLLED  |
| EGAD00001000615  | These Finnish schizophrenia samples have been collected from a population co...       | CONTROLLED  |
| EGAD00001000440  | These affected schizophrenia families have been diagnosed using the SADS-L ...        | CONTROLLED  |
| EGAD00001000436  | This is an Irish sample set of individuals with ASD (approximately 50% with co...     | CONTROLLED  |
| EGAD00001000438  | This sample set consists of subjects with schizophrenia recruited from psychiat...    | CONTROLLED  |
| EGAD00001000441  | The IMGSAC data set represents an international collection of families containin...   | CONTROLLED  |

Total items: 22

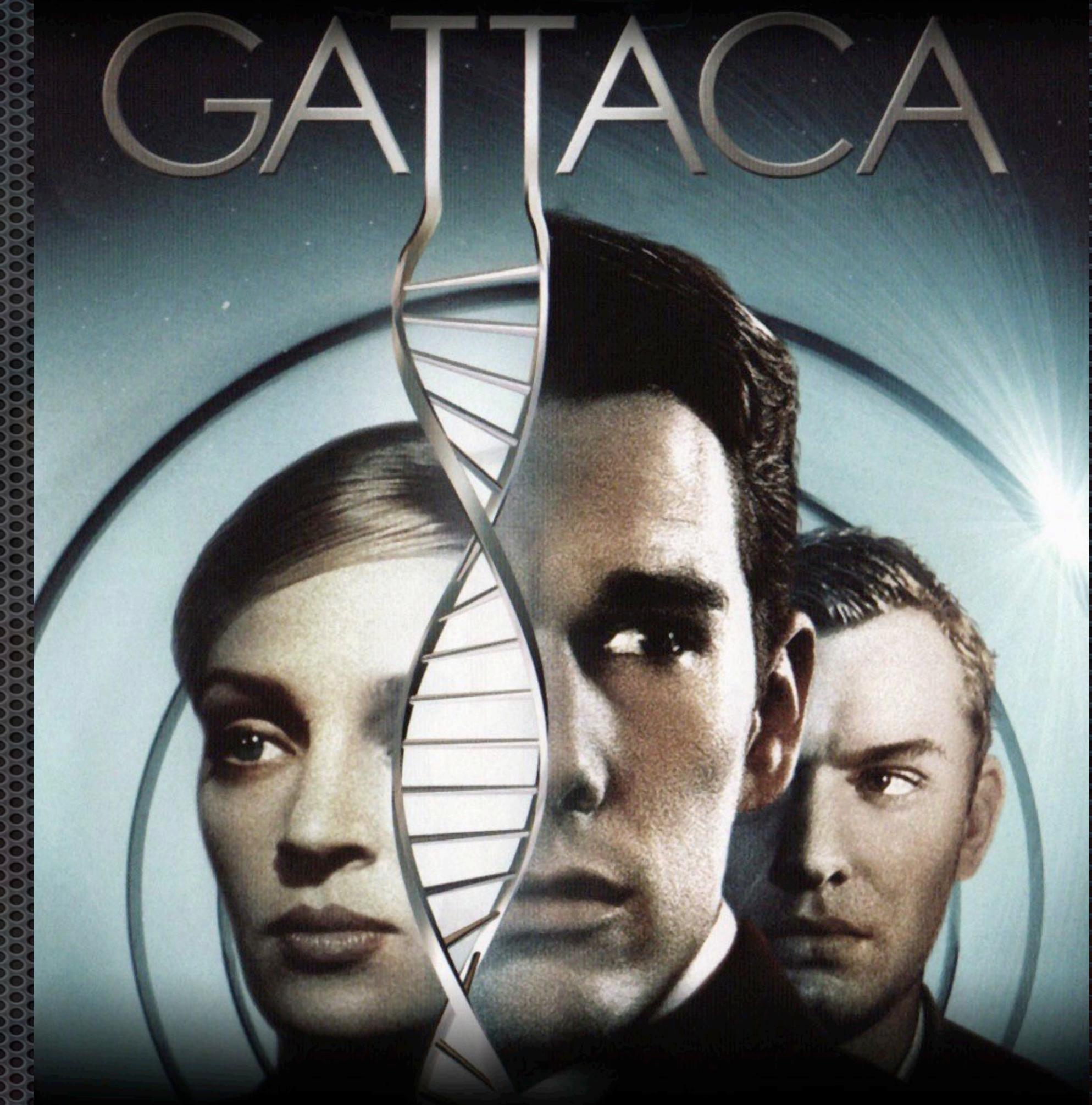
| Risk mitigation strategy        | Disadvantages                                                                                                                                    | Advantages                                                                                    |
|---------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|
| 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, Flieck 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

# Beacons & Privacy

- Practical risk of identification from aggregated data?
- Easing of re-identification from added Beacon features (structural variants, quantitative returns, connection to resources w/ individual data)?
- Impact of design changes on mitigation strategies?

**However: Genome data is required for identification!**



# “DEMOCRATIZING DNA FINGERPRINTING”

Sophie Zaaijer, Assaf Gordon, Robert Piccone, Daniel Speyer, Yaniv Erlich, 2016

[ddf.teamerlich.org](http://ddf.teamerlich.org)



- DNA sequencing for identification/fingerprinting soon “commodity” technology (in contrast with technological/data challenges in “precision medicine”)

MinION by Oxford Nanopore Technologies



The MinION is the smallest DNA sequencer currently around. Its the size of a Mars bar, and can be simply plugged into a laptop with a USB3.0 port.

For more information about the MinION please click:  
[Oxford Nanopore Technologies](#)

Bento Lab



The Bento lab is a miniature lab with a centrifuge, thermocycler and a electrophoresis compartment.

For more information about the Bento-lab please click:  
[Bento Lab](#)



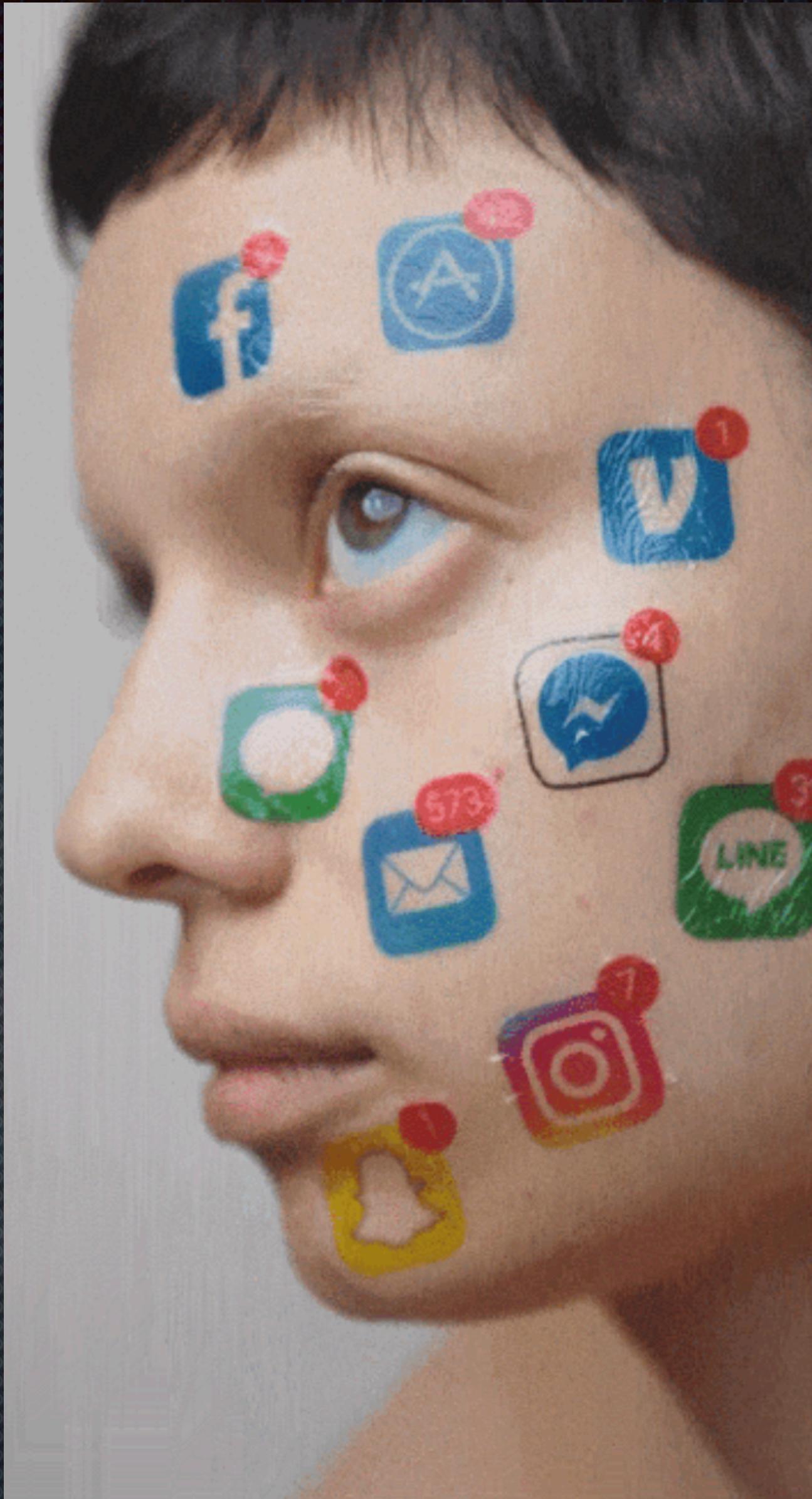
# Share (your) Genome data?

- The Beacon concept - balanced approach for accessing genome variant data from internationally distributed resources
- However: Genome data has the inherent “risk” of being identified and linked to a person

**Solutions from  
Technology or  
Society?  
Discourse!**

The collage consists of three screenshots:

- openSNP Welcome Page:** Shows a heatmap of genetic variants and navigation links for Home, Family tree, Discoveries, DNA, and Research.
- MyHeritage DNA Valentine's Day Sale:** Features a red background with hearts, a DNA kit image, and text advertising a sale of 59€ per kit for two kits, ending February 14th. It includes a purple "Order now" button and shipping information.
- AncestryDNA Kit and Home Page:** Shows a DNA kit box with "Welcome to you" and "saliva collection kit" text, followed by the AncestryDNA logo and a statement about the average British person's DNA being only 36% British. It also features "GROW YOUR TREE" and "Discover DISCOVER" buttons.



John Wiley NWT 2018-02-09

**Welcome to openSNP**

openSNP lets customers of direct-to-customer genetic tests publish their test results, find others with similar genetic

Home Family tree Discoveries DNA Research

For Genotyping Users For S

Upload Your Genotyping File

Upload your raw genotyping

Phenotypes are the

MyHeritage DNA

Valentine's Day DNA SALE

Only 59€<sup>89€</sup> per kit When ordering 2+ kits

Order now

Shipping not included Ends February 14th

23andMe

Welcome to you

saliva collection kit

ancestry

Find out what your DNA says about you and your family.

- See how your DNA breaks out across 31 populations worldwide
- Discover DNA relatives from around the

SUBSCRIBE SIGN IN >

THE AVERAGE BRITISH PERSON'S DNA IS ONLY 36% BRITISH

GROW YOUR TREE

Find your ancestors in

ancestryDNA

Discover



University of  
Zurich<sup>UZH</sup>



Global Alliance  
for Genomics & Health



# Beaconize Your Resource

*beacon-project.io*  
*github.com/ga4gh-beacon*  
*beacon.progenetix.org*  
*ga4gh.org*  
*sib.swiss/baudis-michael*



Serena Scollen



Susheel Varma



Michael Baudis



Marc Fiume



Ilkka Lappalainen



Jordi Rambla



David Lloyd



Dylan Spalding



Knox Carey



Anthony Brookes



Stephanie Dyke



Aurélien Barré



Rishi Nag



Miro Cupak



Juha Törnroos



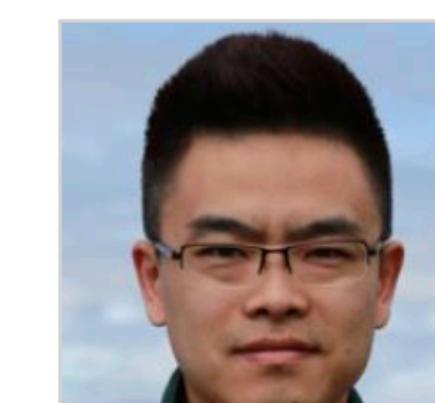
Sabela de la Torre



Marc Duby



Saif Ur-Rehman



Bo Gao

