Global Alliance
for Genomics & Health

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

Global Alliance for Genomics & Health

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

Search...

ABOUT US | HOW WE WORK | GA4GH TOOLKIT | NEWS & EVENTS | BECOME A MEMBER | CONTACT US

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

GENOMICS

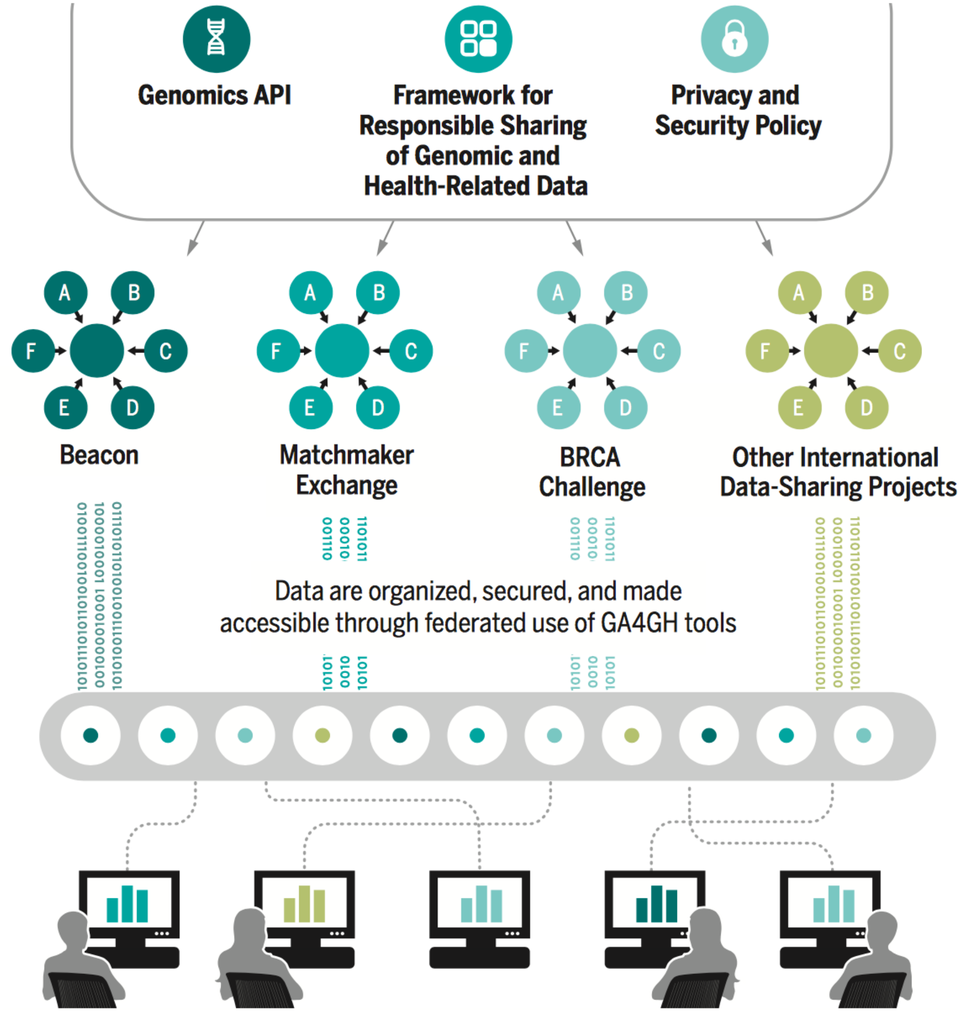
A federated ecosystem for sharing genomic, clinical data

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

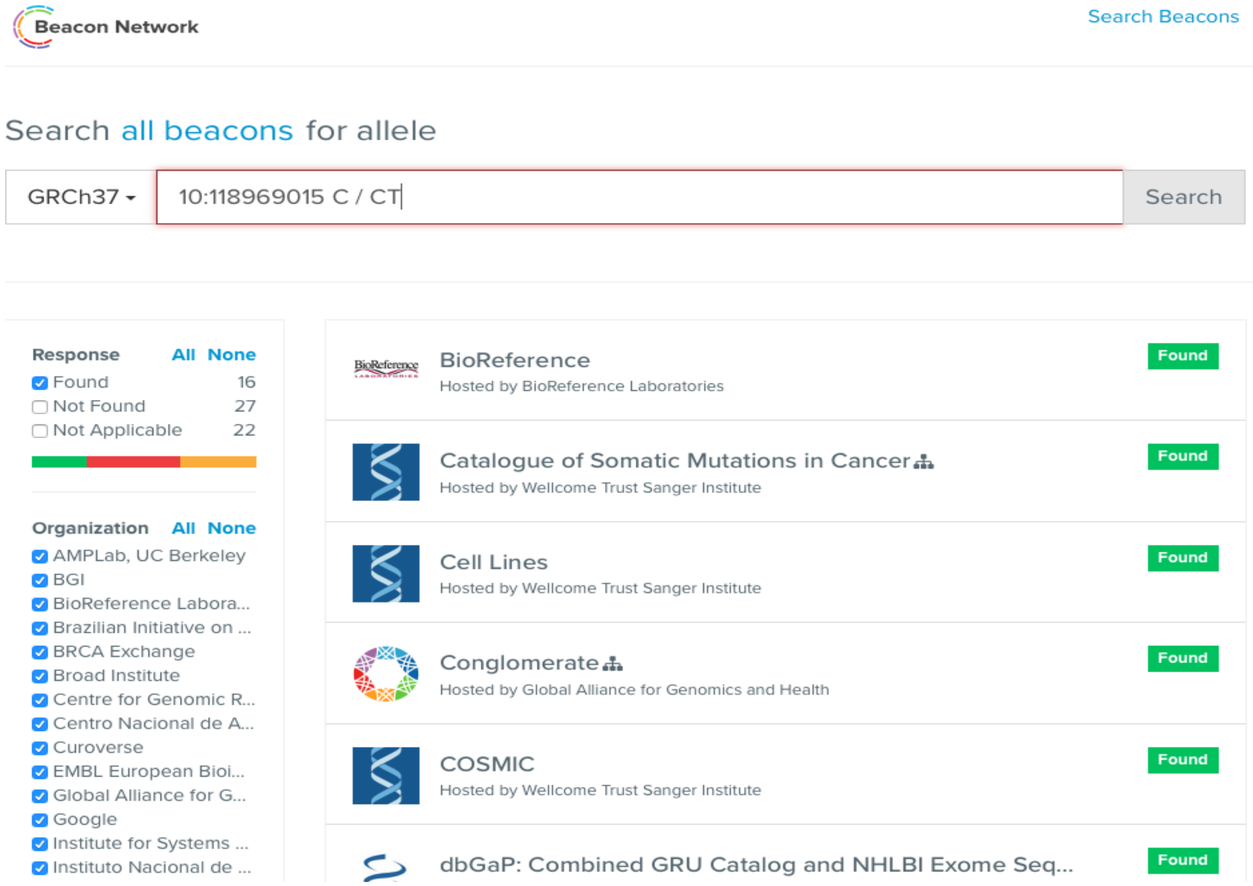
A federated data ecosystem. To share genomic data globally, this approach furthers medical research without requiring compatible data sets or compromising patient identity.

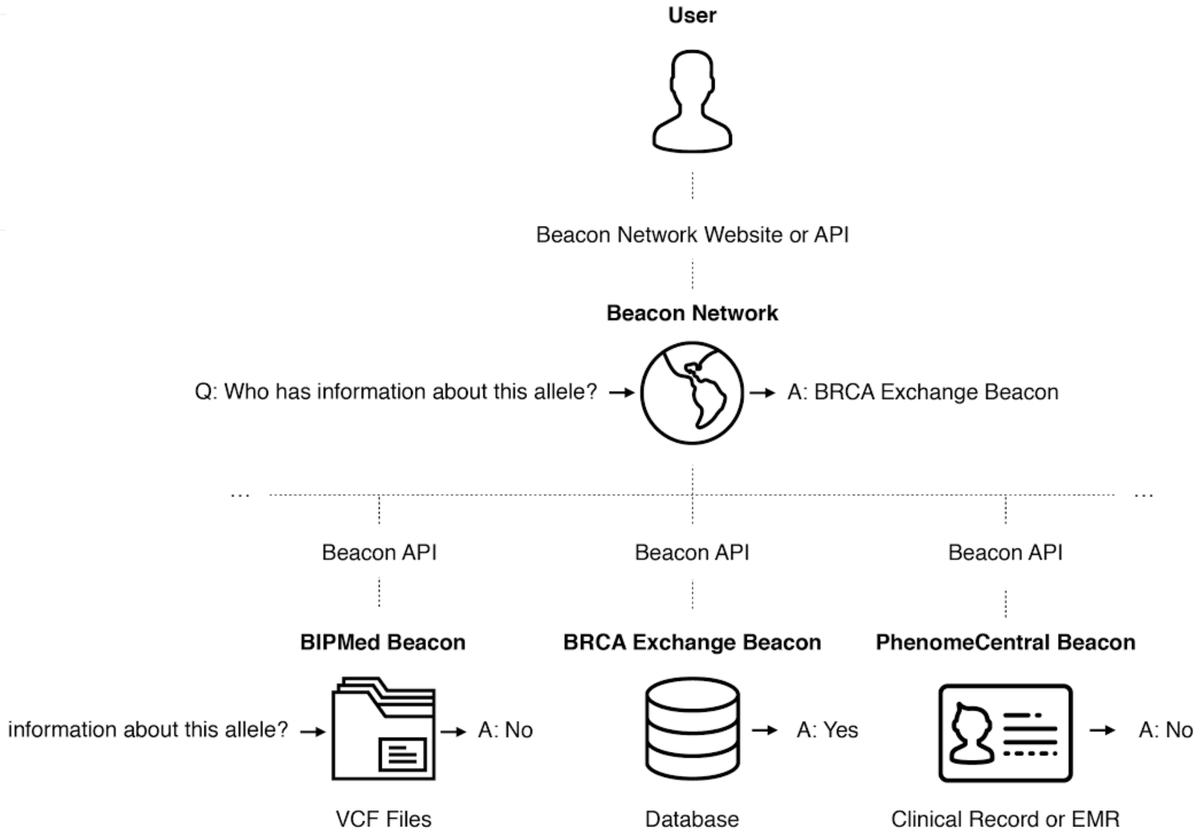


Global Alliance for Genomics & Health

Beacon Project

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





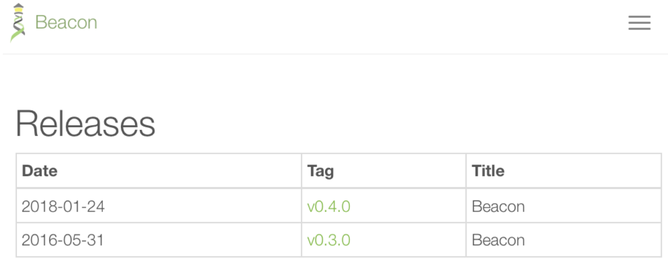
Global Alliance for Genomics & Health

35+ Organizations

90+ Beacons

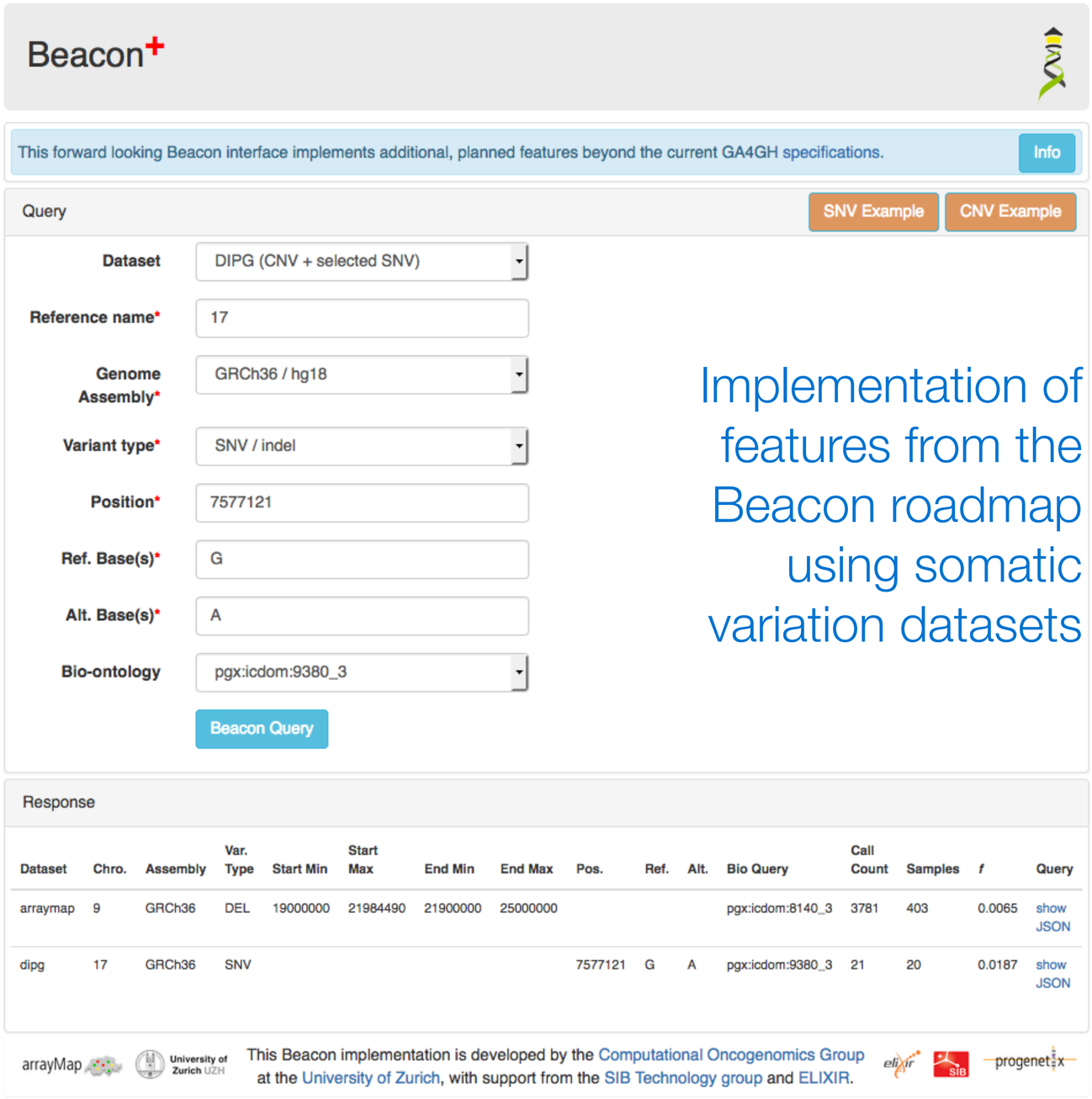
200+ Datasets

100K+ Individuals



Beacon Roadmap

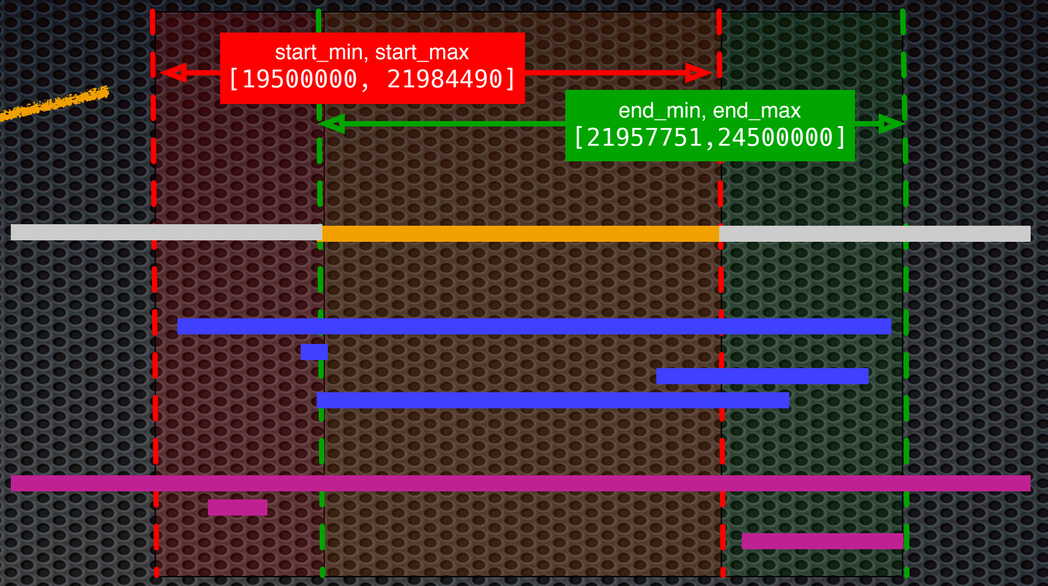
- **structural variations** (DUP, DEL) in addition to SNV
- **metadata** queries using ontology codes (e.g. NCIT, ICD-O)
- layered **authentication** system using **ELIXIR** AAI
- prototyping **handover** concept
- **quantitative** responses
- Ubiquitous **deployment** (e.g. throughout **ELIXIR** network)



Implementation of features from the Beacon roadmap using somatic variation datasets

```
{
   "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.progenetix.org

Beacon+

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

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

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

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

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 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

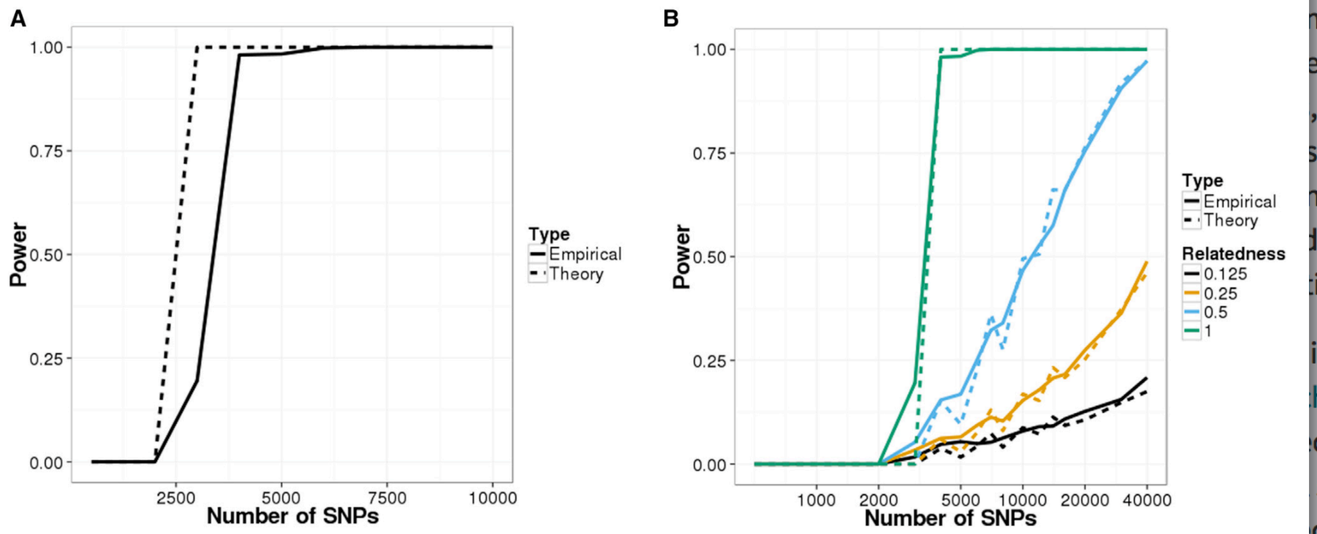


Figure 1. Power of Re-identification Attacks on Beacons Constructed with Simulated Data
Power curves for the likelihood-ratio test (LRT) on (A) a simulated beacon with 1,000 individuals and (B) detecting relatives in the simulated beacon. The false-positive rate was set to 0.05 for all scenarios.

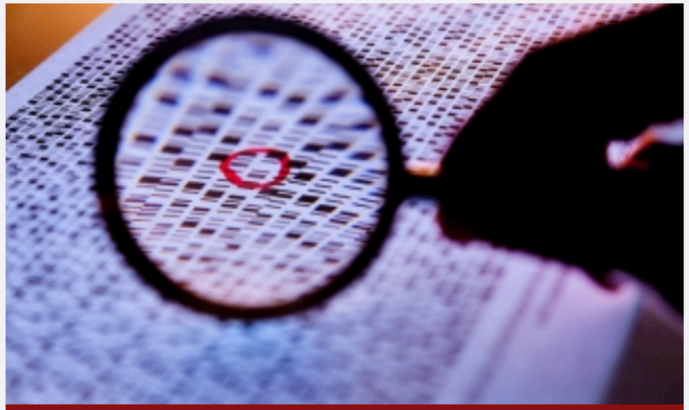
Suyash S. Shringarpure, Carlos D. Bustamante. Privacy Risks from Genomic Data-Sharing Beacons. AJHG 97:5, Pages 631-646 (Nov. 2015)

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 either directly from your saliva or , or from a popular genomic service — they could check to see if 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…

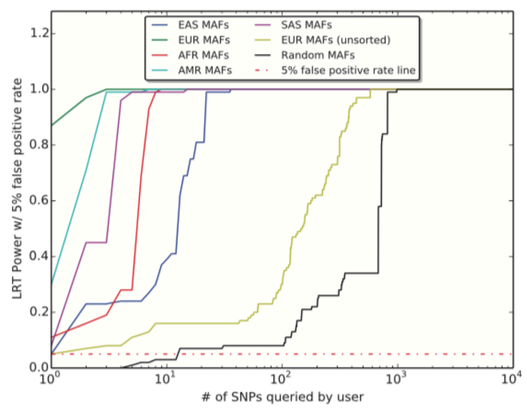


Figure 1. "Optimal" re-identification attack in single-population beacon. Different power rates per number of SNPs queried from an unprotected beacon with a single population (EUR) by an adversary with different types of background knowledge: (green) the attacker knows the allele frequencies (AFs) of 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]).

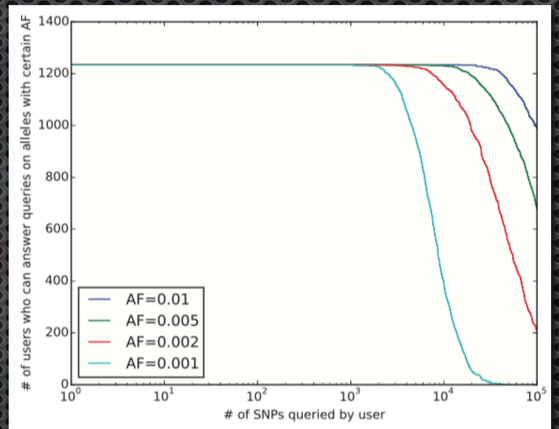
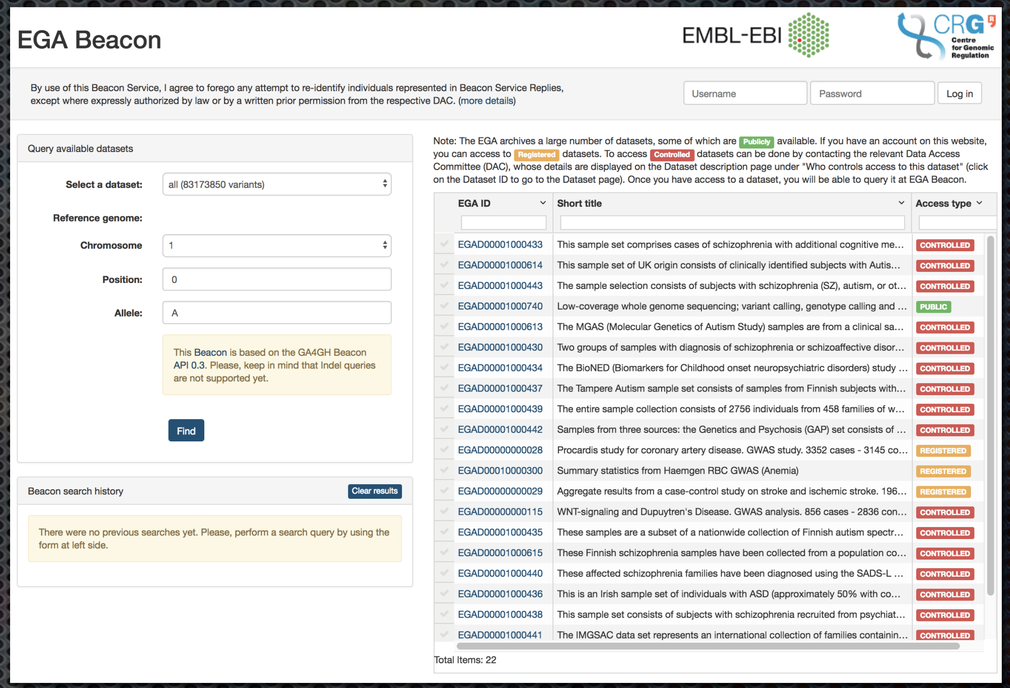


Figure 5. Budget evaluation in beacon with *S3*. Behaviors of individual budgets per number of SNPs queried according to the typical user's query profile obtained from ExAC log data. The cyan curve represents the number of individuals with enough budget to answer "Yes" to queries targeting alleles with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



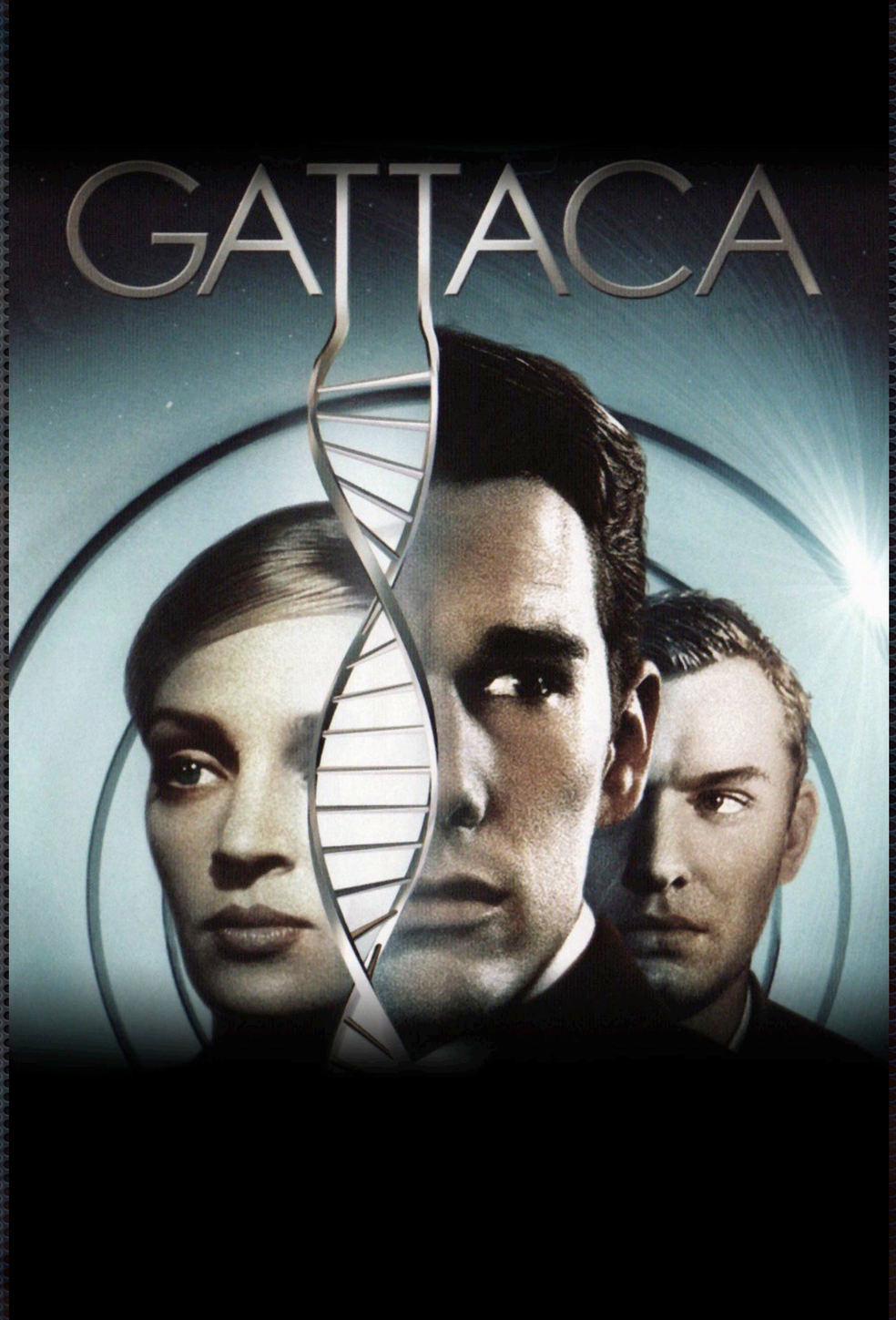
Risk mitigation strategy	Disadvantages	Advantages
S1: Beacon alteration	Eliminates possibility of querying for unique alleles highly likely to be most useful in genetic research	Protects privacy of individuals possessing variants most likely to be targeted by attackers
S2: Random flipping	Decreases rate 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

Beacons & Privacy

- Practical risk of identification from aggregated data?
- Easing of re-identification from added Beacon features (styructural 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

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:

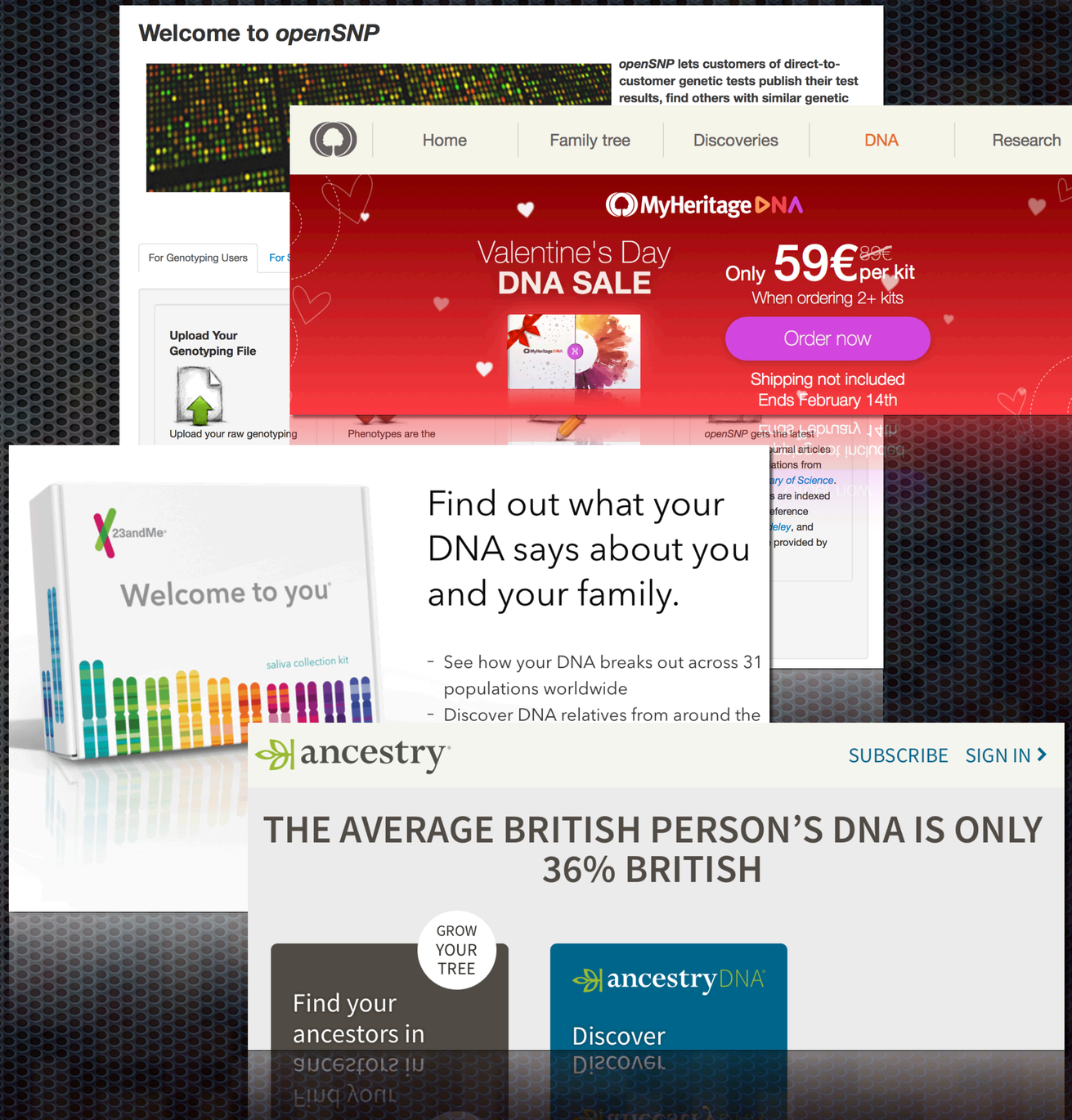
Bento Lab

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

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!



John Yuyi, NYT 2018-02-09

Welcome to *openSNP*

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

For Genotyping Users

Upload Your Genotyping File

Upload your raw genotyping

Phenotypes are the

openSNP gets the latest

Home | Family tree | Discoveries | DNA | Research

MyHeritage DNA

Valentine's Day
DNA SALE

Only **59€** ~~89€~~ per kit
When ordering 2+ kits

Order now

Shipping not included
Ends February 14th

23andMe

Welcome to you

saliva collection kit

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

ancestry

SUBSCRIBE SIGN IN

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

GROW YOUR TREE

Find your ancestors in

ancestryDNA

Discover

Global Alliance for Genomics & Health

elixir

Beaconize Your Resource

beacon-project.io

github.com/ga4gh-beacon

beacon.progenetix.org

ga4gh.org

sib.swiss/baudis-michael

University of Zurich[UZH]

SIB

Beacon

Overview Team FAQ Contact


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

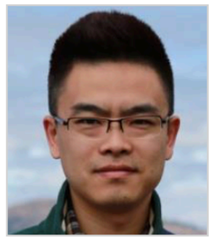

Miro Cupak


Juha Törnroos


Sabela de la Torre


Marc Duby


Saif Ur-Rehman


Bo Gao

DNASTACK