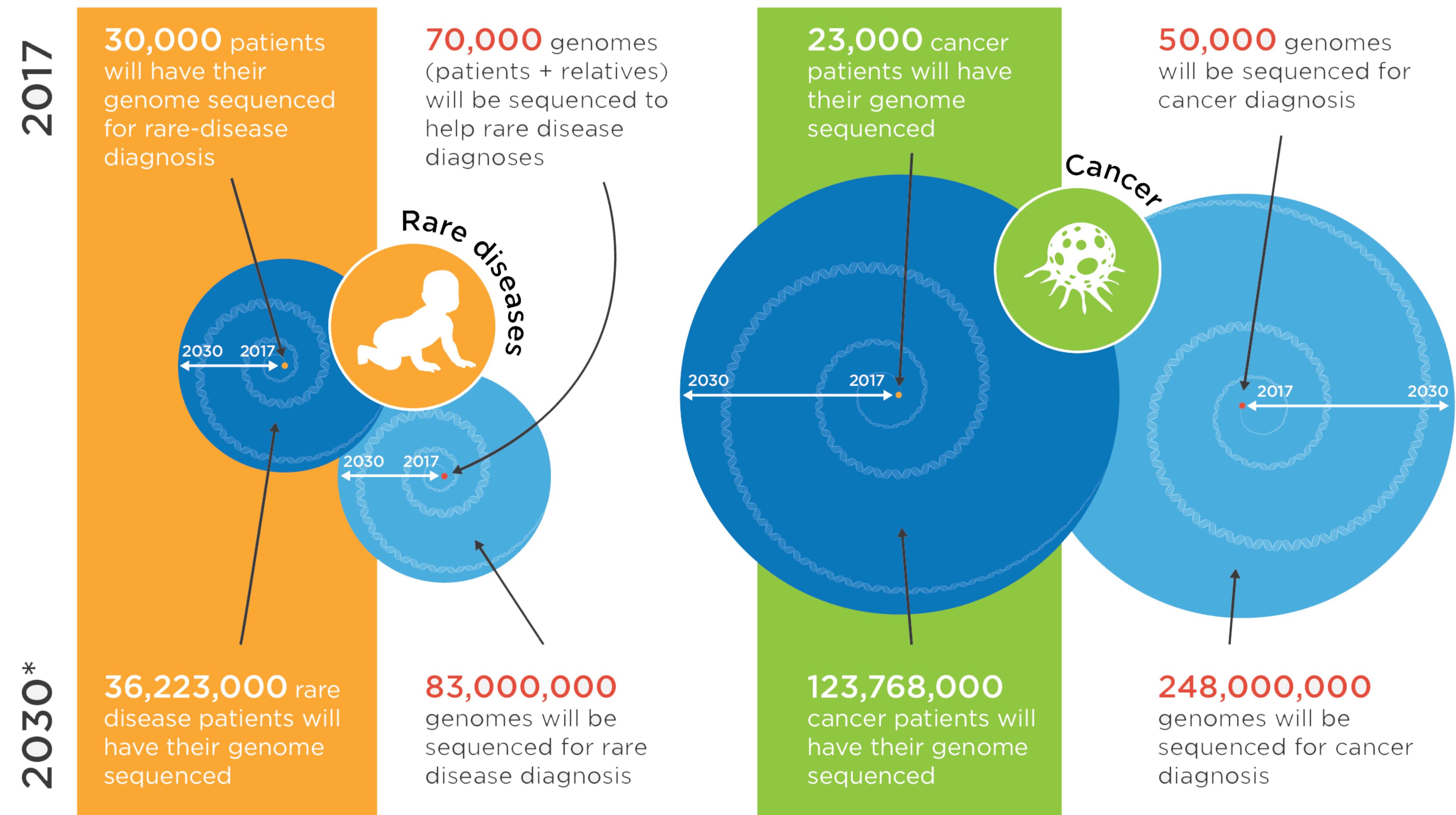


Progenetix & Beacon+

An open cancer genomics resource on a stack of Beacon code...



* Projected figures, based on current data and known status of genomics initiatives worldwide.

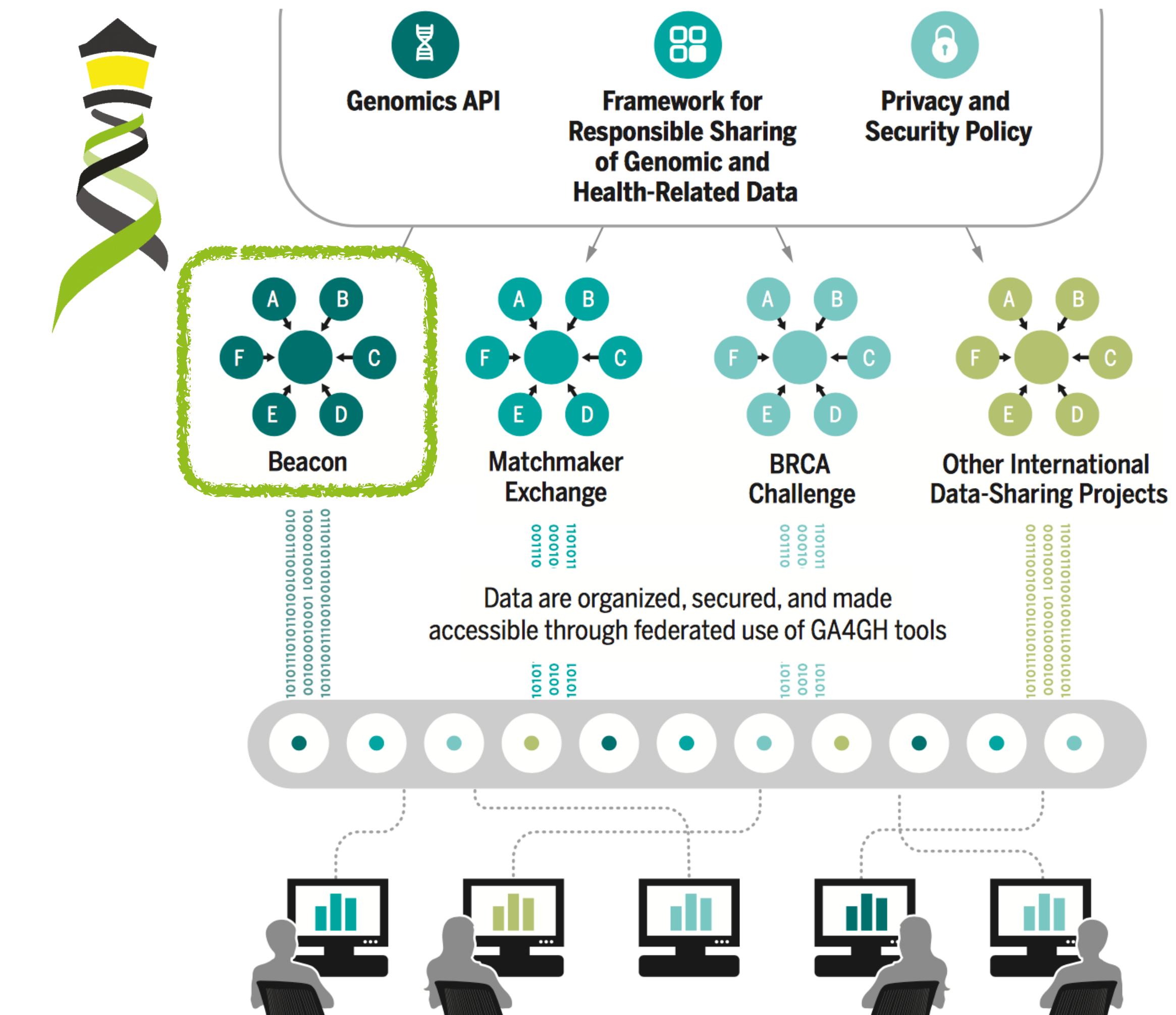


GENOMICS

A federated ecosystem for sharing genomic, clinical data

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

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



The Global Alliance for Genomics and Health

Making genomic data accessible for research and health

- January 2013 - 50 participants from eight countries
- June 2013 - White Paper, over next year signed by 70 “founding” member institutions (e.g. SIB, UZH)
- March 2014 - Working group meeting in Hinxton & 1st plenary in London
- October 2014 - Plenary meeting, 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
- May 2019 - GA4GH Connect, Hinxton
- October 2019 - 7th Plenary, Boston
- October 2020 - Virtual Plenary, June 2021 - Virtual Connect ...
- October 2021 - Virtual Plenary ...
- September 2022 - 10th Plenary, Barcelona

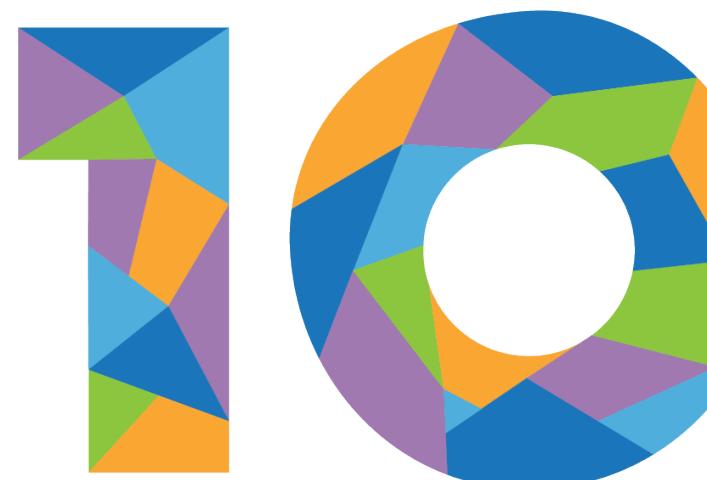
GENOMICS

A federated ecosystem for sharing genomic, clinical data

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

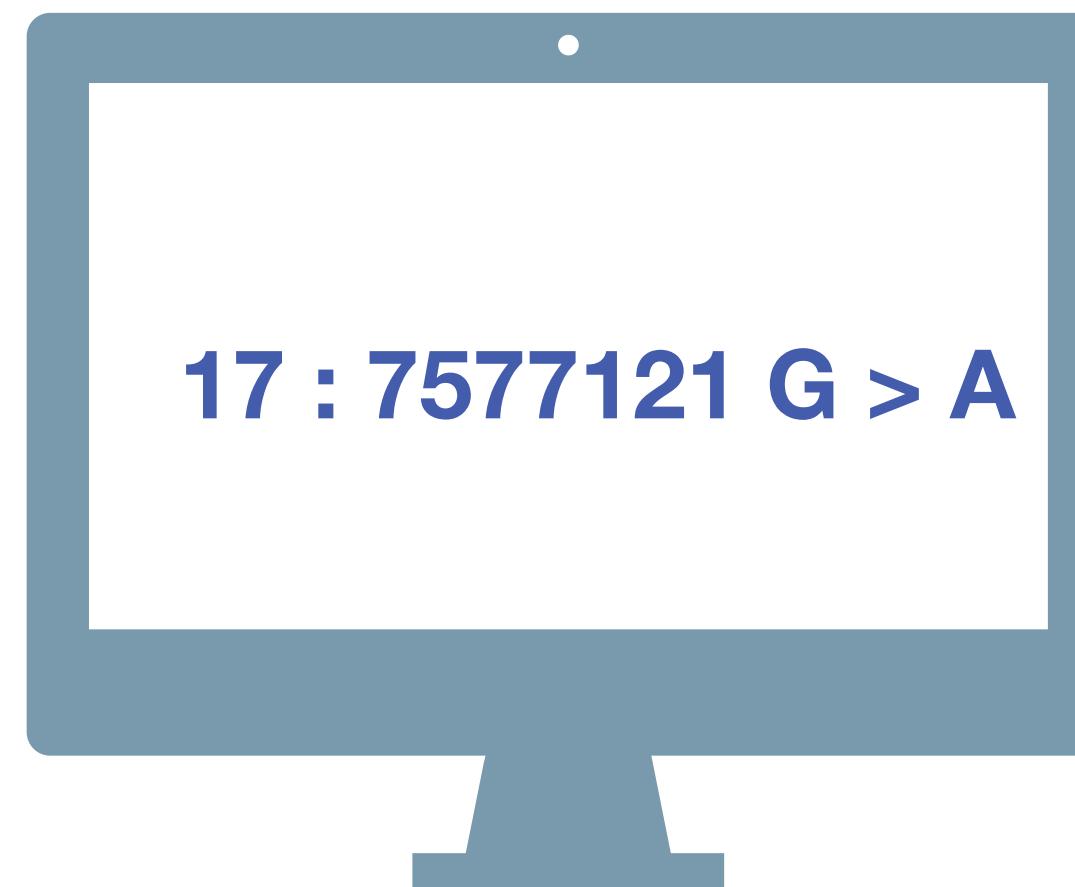
The Global Alliance for Genomics
and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291



22 SEPTEMBER 2022 | BARCELONA, SPAIN

GA4GH 10th Plenary



Beacon

A **Beacon** answers a query for a specific genome variant against individual or aggregate genome collections

YES | NO | \0

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.

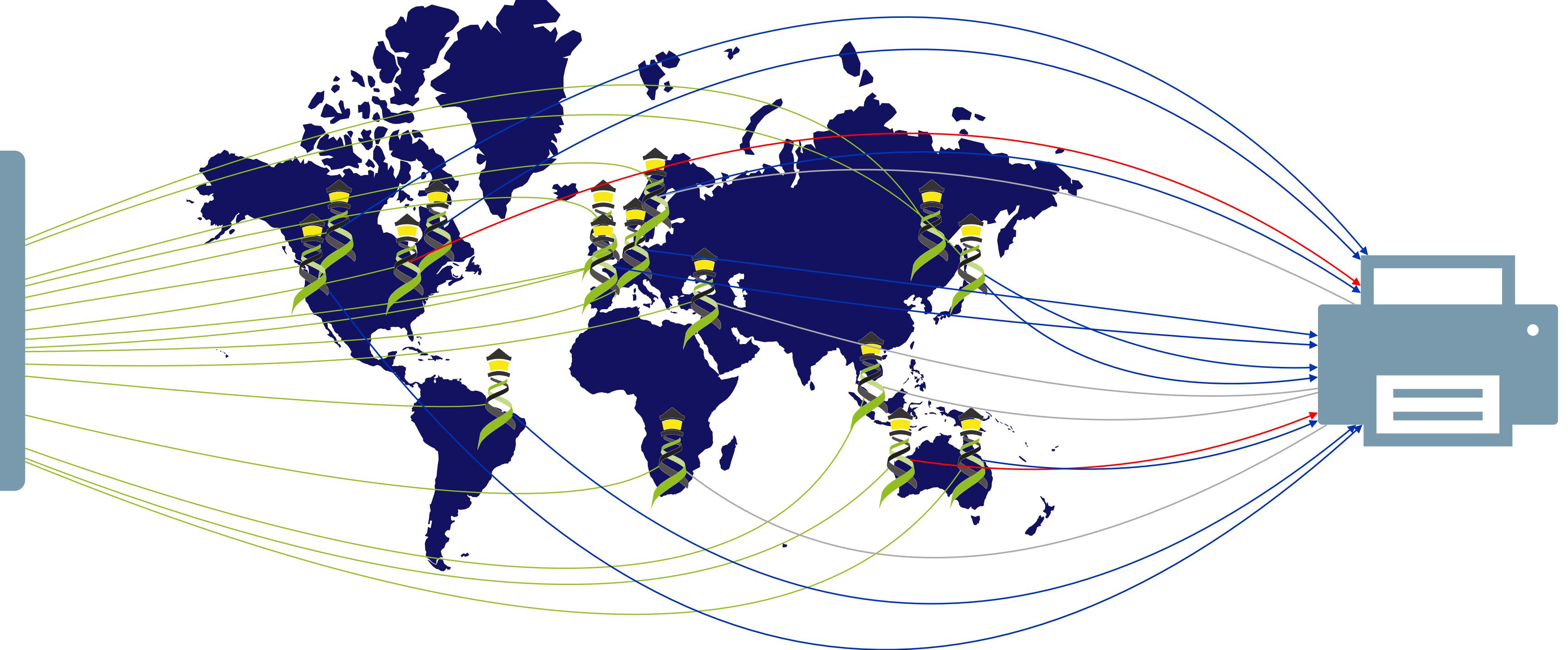
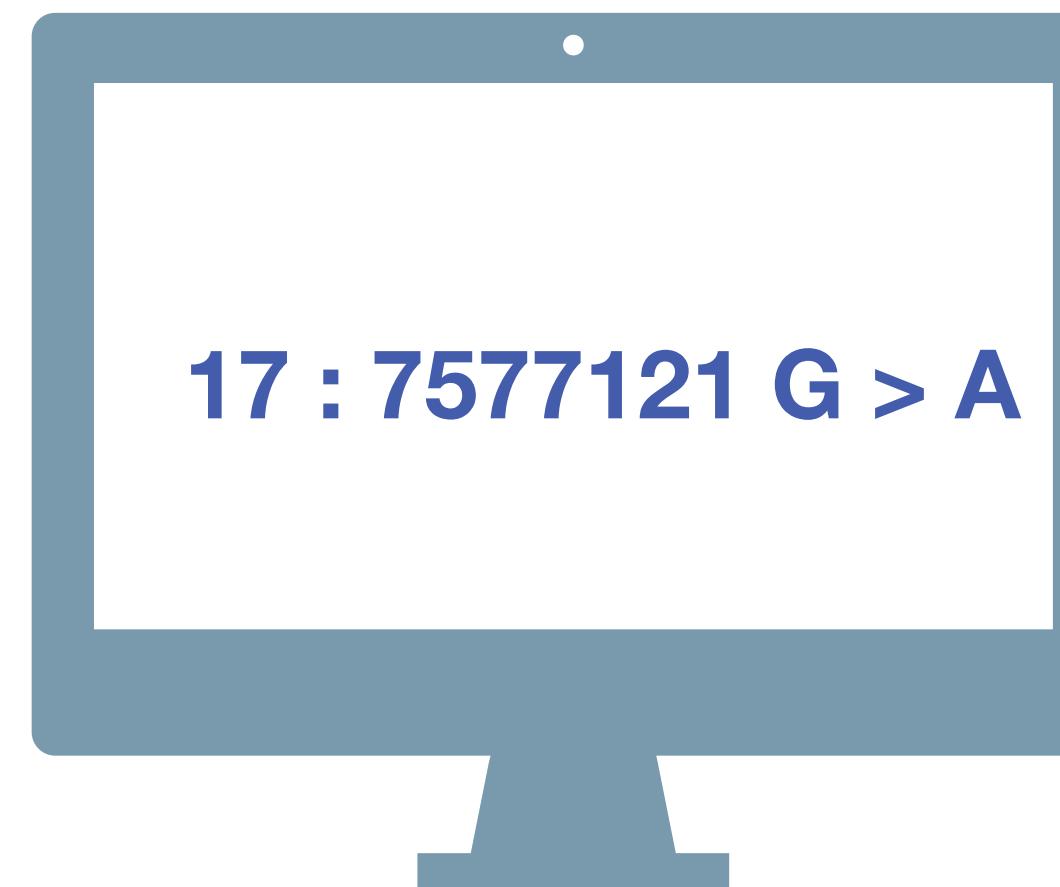
1. Provide *frequencies of all alleles* at that point
2. Ask for all alleles seen in a gene *region* (and more elaborate versions of this)
3. Other more complicated queries



"I would personally recommend all those be held for
version 2, when the beacon becomes a service."
Jim Ostell, 2014

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 ...

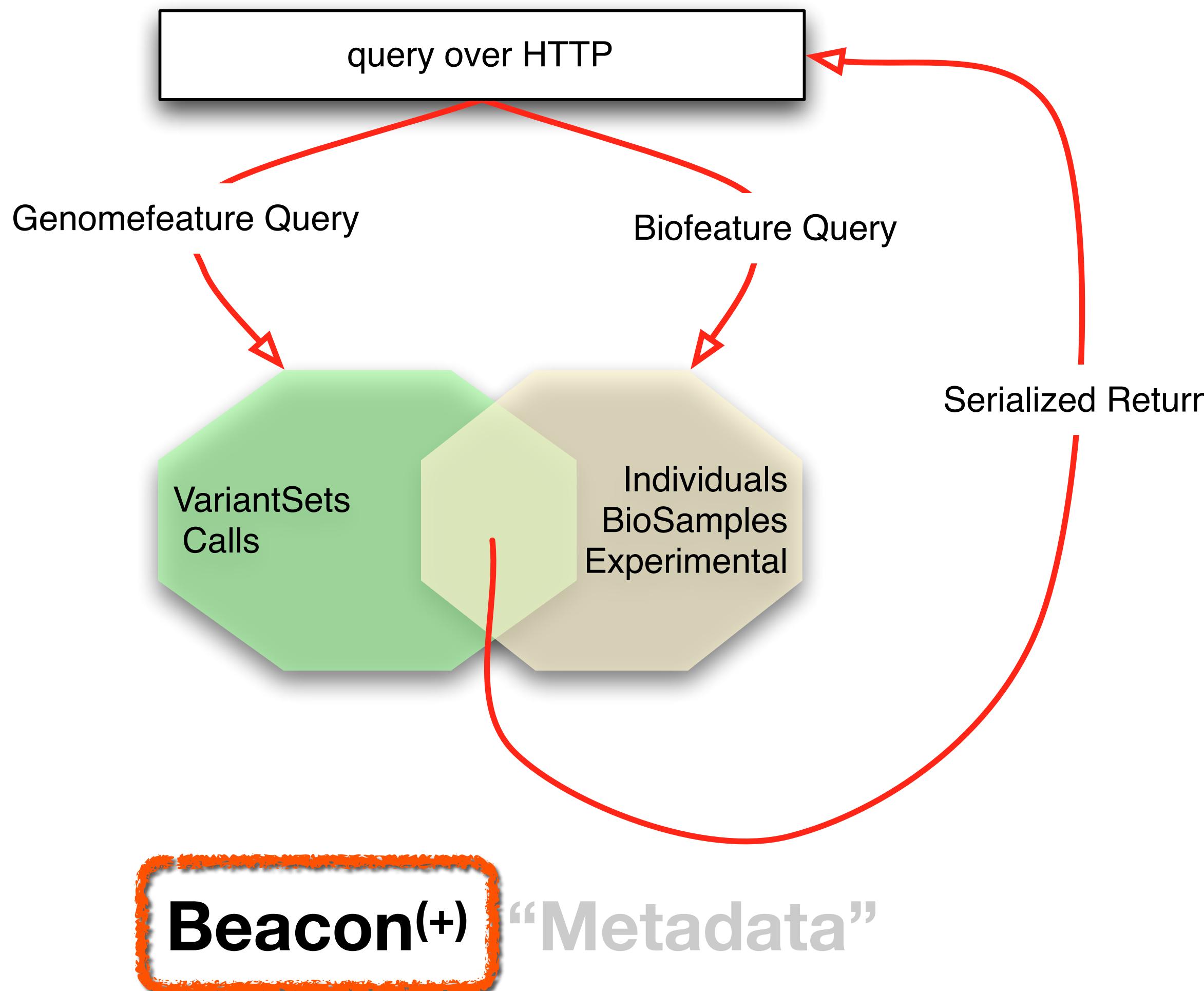


Have you seen this variant?
It came up in my patient
and we don't know if this is
a common SNP or worth
following up.

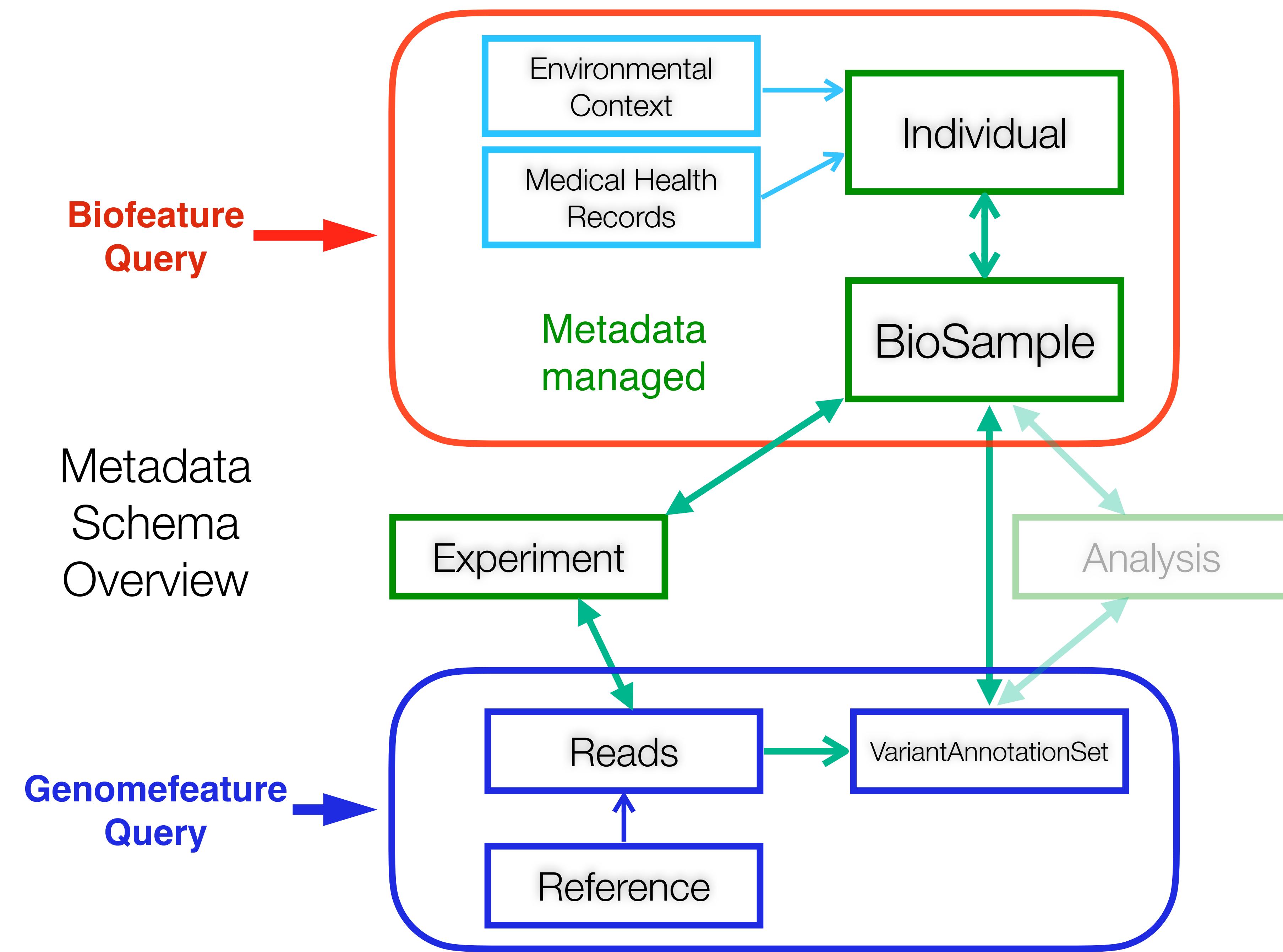
A Beacon network federates
genome variant queries
across databases that
support the **Beacon API**

Here: The variant has
been found in **few**
resources, and those
are from **disease**
specific **collections**.

Minimal GA4GH query API structure

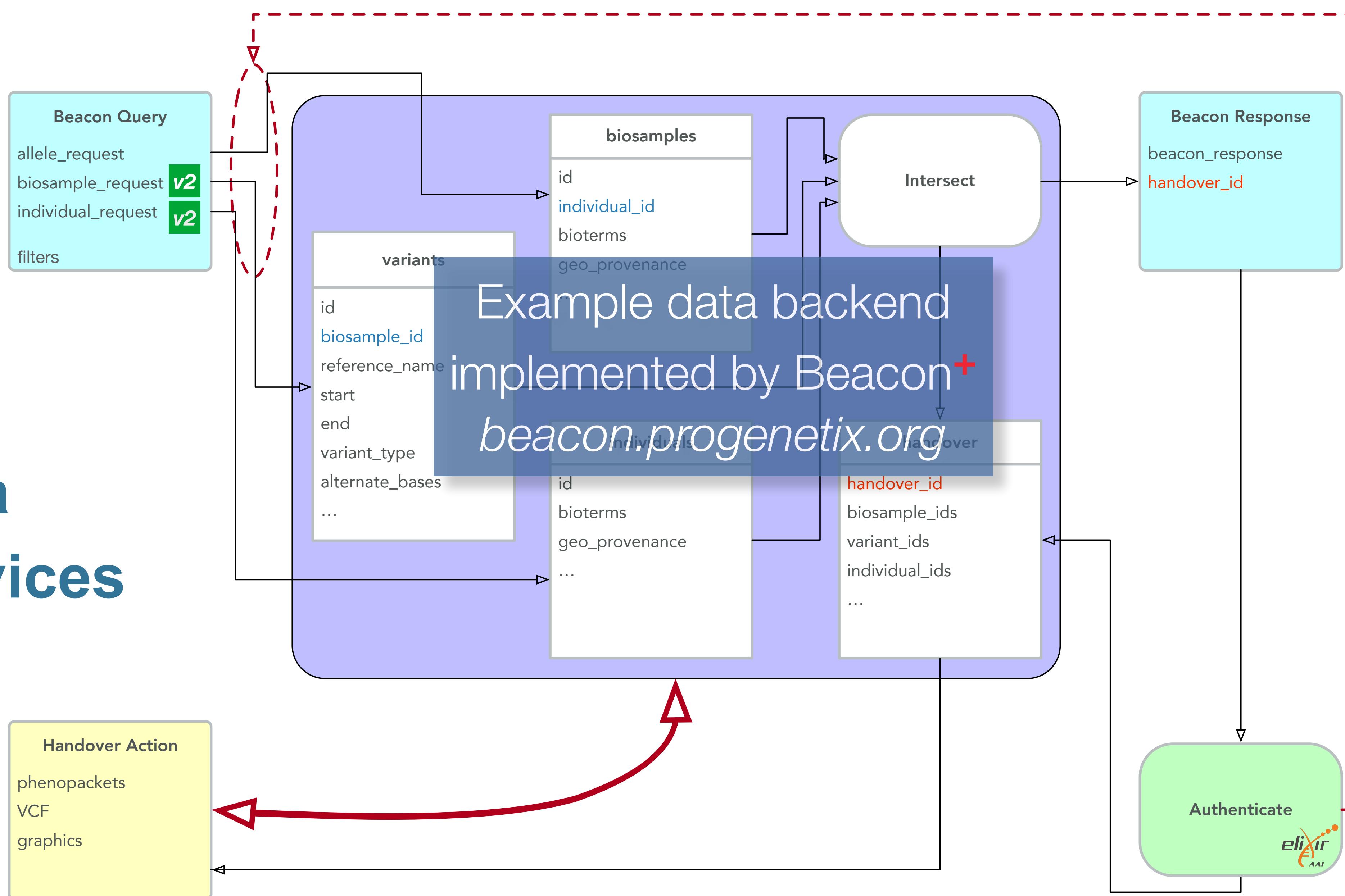


Beacon+ in 2016 - Driving GA4GH implementations



Beacon & Handover

Beacons v1.1
supports data
delivery services



Michael Baudis



```
{
  "alleleRequest": {
    "endMax": "26000000",
    "referenceName": "9",
    "startMax": "21975098",
    "endMin": "21967753",
    "startMin": "18000000",
    "alternateBases": "N",
    "variantType": "DEL",
    "referenceBases": "*"
  },
  "url": "https://beacon.progenetix.org/beacon/info/",
  "beaconId": "progenetix-beacon",
  "datasetAlleleResponses": [
    {
      "externalUrl": "https://beacon.progenetix.org/beacon/info/",
      "datasetId": "arraymap",
      "variantCount": 588,
      "info": {
        "distinctVarCount": 551,
        "description": "The query was against database \"arraymap\", variant collection \"variants\". 588 matched callsets for 588 distinct variants.",
        "error": null,
        "exists": true,
        "datasetHandover": [
          {
            "url": "https://beacon.progenetix.test/beaconplus-server/beacondeliver.cgi?do=biosamplesdata&accessid=5d76f88d-4012-11e9-a0b4-d9893b611ec4",
            "handoverType": { "label": "Biosamples", "id": "pgx:handover:biosamplesdata" },
            "description": "retrieve data of the biosamples matched by the query"
          },
          {
            "url": "https://beacon.progenetix.test/beaconplus-server/beacondeliver.cgi?do=callsetsvariants&accessid=5d77fb88-4012-11e9-a0b4-bb5a9c8cf98a",
            "description": "export all variants of matched callsets - potentially huge dataset...",
            "handoverType": { "label": "Callsets Variants", "id": "pgx:handover:callsetsvariants" }
          },
          {
            "handoverType": { "id": "pgx:handover:cnvhistogram", "label": "CNV Histogram" },
            "description": "create a CNV histogram from matched callsets",
            "url": "https://beacon.progenetix.test/beaconplus-server/beacondeliver.cgi?do=cnvhistogram&accessid=5d77fb88-4012-11e9-a0b4-bb5a9c8cf98a"
          },
          {
            "handoverType": { "label": "Variants", "id": "pgx:handover:variantsdata" },
            "description": "retrieve data of the variants matched by the query",
            "url": "https://beacon.progenetix.test/beaconplus-server/beacondeliver.cgi?do=variantsdata&accessid=5d6e982b-4012-11e9-a0b4-c5ce5cc21906"
          }
        ],
        "callCount": 588,
        "varResponses": [
          "9:21773941-21968713:DEL",
          "9:21732467-23813102:DEL",
          "9:21785019-21968713:DEL",
          "9:21968713-22031006:DEL"
        ]
      }
    }
  ]
}
```

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**





This example shows a core Beacon query, against a specific mutation in the TP53 gene, in cellosaurus, with ClinVar data.

CNV Example SNV Range Example SNV Example ClinVar Example Beacon Help

Dataset*

arraymap
progenetix
cellosaurus
dipg
BeaconSpecTest2
BeaconSpecTest

Genome Assembly*

GRCh38 / hg38

Dataset Responses

All Selected Datasets

Reference name*

17

Gene Coordinates

TP53

Cytoband(s)

17p13.1

Start

7673767

Ref. Base(s)

C

Alt. Base(s)

T

Bio-ontology

no selection
NCIT:C102872: Pharyngeal squamous cell carcinoma (2)
NCIT:C103968: Pyruvate dehydrogenase deficiency (1)
NCIT:C105555: High grade ovarian serous adenocarcinoma (75)
NCIT:C105556: Low grade ovarian serous adenocarcinoma (10)
NCIT:C111802: Dyskeratosis congenita (3)

Other Filters

additional comma-separated, prefixed filters

Beacon Query

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.



```
{
  "callset_id": "cs-cellosaurus:CVCL_EI02",
  "info": {
    "cellosaurus": {
      "cell_line": "BT474-LAPRa",
      "id": "CVCL_EI02",
      "cellosaurus_variant_name": "TP53 p.Glu285Lys (c.853G>A)"
    },
    "clinvar": {
      "gene_id": "7157",
      "allele_id": "410258",
      "assembly": "GRCh38",
      "cytoband": "17p13.1",
      "variant_type": "single nucleotide variant",
      "origin": "germline;somatic",
      "phenotype": "Hereditary cancer-predisposing syndrome;Li-Fraumeni syndrome;PARP Inhibitor response;not provided",
      "clinical_significance": "Pathogenic/Likely pathogenic",
      "clinvar_full_name": "NM_001126112.2(TP53):c.853G>A (p.Glu285Lys)"
    }
  },
  "start_min": 7673766,
  "reference_name": "17",
  "end_min": 7673767,
  "biosample_id": "bios-cellosaurus:CVCL_EI02",
  "alternate_bases": [
    "T"
  ],
  "digest": "17_7673767_C_T",
  "reference_bases": "C",
  "variantset_id": "cellosaurus_clinvar_GRCH38",
  "end_max": 7673767,
  "start_max": 7673766
},
{
  "digest": "17_7673767_C_T",
  "reference_bases": "C",
  "alternate_bases": [
    "T"
  ],
  "variantset_id": "cellosaurus_clinvar_GRCH38",
  "end_max": 7673767,
  "start_max": 7673766,
  "callset_id": "cs-cellosaurus:CVCL_AQ07",
  "start_min": 7673766,
  "info": {
    "cellosaurus": {
      "cellosaurus_variant_name": "TP53 p.Glu285Lys (c.853G>A)",
      "cell_line": "BT-474 Clone 5",
      "id": "CVCL_AQ07"
    },
    "clinvar": {
      "assembly": "GRCh38",
      "allele_id": "410258",
      "gene_id": "7157",
      "cytoband": "17p13.1",
      "variant_type": "single nucleotide variant",
      "phenotype": "Hereditary cancer-predisposing syndrome;Li-Fraumeni syndrome;PARP Inhibitor response;not provided",
      "origin": "germline;somatic",
      "clinvar_full_name": "NM_001126112.2(TP53):c.853G>A (p.Glu285Lys)",
      "clinical_significance": "Pathogenic/Likely pathogenic"
    }
  },
  "end_min": 7673767,
  "biosample_id": "bios-cellosaurus:CVCL_AQ07",
  "reference_name": "17"
},
{
  "alternate_bases": [
    "T"
  ],
  "reference_bases": "C",
  "digest": "17_7673767_C_T",
  "end_max": 7673767,
  "variantset_id": "cellosaurus_clinvar_GRCH38",
  "start_max": 7673766,
  "callset_id": "cs-cellosaurus:CVCL_AQ07"
}
```

Beacon v1 Development

2014

GA4GH founding event; Jim Ostell proposes Beacon concept including "more features ... version 2"

2015

- beacon-network.org aggregator created by DNAstack

2016

- Beacon v0.3 release
- work on queries for structural variants (brackets for fuzzy start and end parameters...)

2017

- OpenAPI implementation
- integrating CNV parameters (e.g. "startMin, statMax")

2018

- Beacon v0.4 release in January; feature release for GA4GH approval process
- GA4GH Beacon v1 approved at Oct plenary

2019

- ELIXIR Beacon Network

2020

Beacon v2 Development

- Beacon⁺ concept implemented on progenetix.org
- concepts from GA4GH Metadata (ontologies...)
- entity-scoped query parameters ("individual.age")

- Beacon⁺ demos "handover" concept

- Beacon hackathon Stockholm; settling on "filters"
- Barcelona goes Zurich developers meeting
- Beacon API v2 Kick off
- adopting "handover" concept

- "Scouts" teams working on different aspects - filters, genomic variants, compliance ...
- discussions w/ clinical stakeholders

- framework + models concept implemented
- range and bracket queries, variant length parameters
- starting of GA4GH review process

- further changes esp. in default model, aligning with Phenopackets and VRS
- unified beacon-v2 code & docs repository
- Beacon v2 approved at Apr GA4GH Connect

2022

Related ...

- ELIXIR starts Beacon project support

- GA4GH re-structuring (workstreams...)
- Beacon part of Discovery WS

- new Beacon website (March)

- Beacon publication at Nature Biotechnology

- docs.genomebeacons.org

Beacon v1 Development

2014

GA4GH founding event; Jim Ostell proposes Beacon concept including "more features ... version 2"

2015

- beacon-network.org aggregator created by DNAstack

2016

- Beacon v0.3 release
work on queries for structural variants (brackets for fuzzy start and end parameters...)

2017

- OpenAPI implementation
- integrating CNV parameters (e.g. "startMin, statMax")

2018

- Beacon v0.4 release in January; feature release for GA4GH approval process
- GA4GH Beacon v1 approved at Oct plenary

2019

- ELIXIR Beacon Network

2020

2021

2022

Beacon v2 Development

Related ...

- ELIXIR starts Beacon project support

- GA4GH re-structuring (workstreams...)
- Beacon part of Discovery WS

- new Beacon website (March)

- Beacon publication at Nature Biotechnology

- Phenopackets v2 approved

- docs.genomebeacons.org

- Beacon+ concept implemented on progenetix.org
- concepts from GA4GH Metadata (ontologies...)
- entity-scoped query parameters ("individual.age")

- Beacon+ demos "handover" concept

- Beacon hackathon Stockholm; settling on "filters"
- Barcelona goes Zurich developers meeting
- Beacon API v2 Kick off
- adopting "handover" concept
- "Scouts" teams working on different aspects - filters, genomic variants, compliance ...
- discussions w/ clinical stakeholders

- framework + models concept implemented
- range and bracket queries, variant length parameters
- starting of GA4GH review process
- further changes esp. in default model, aligning with Phenopackets and VRS
- unified beacon-v2 code & docs repository
- Beacon v2 approved at Apr GA4GH Connect



Thank You!



... and many more!



Progenetix Genomics Resource

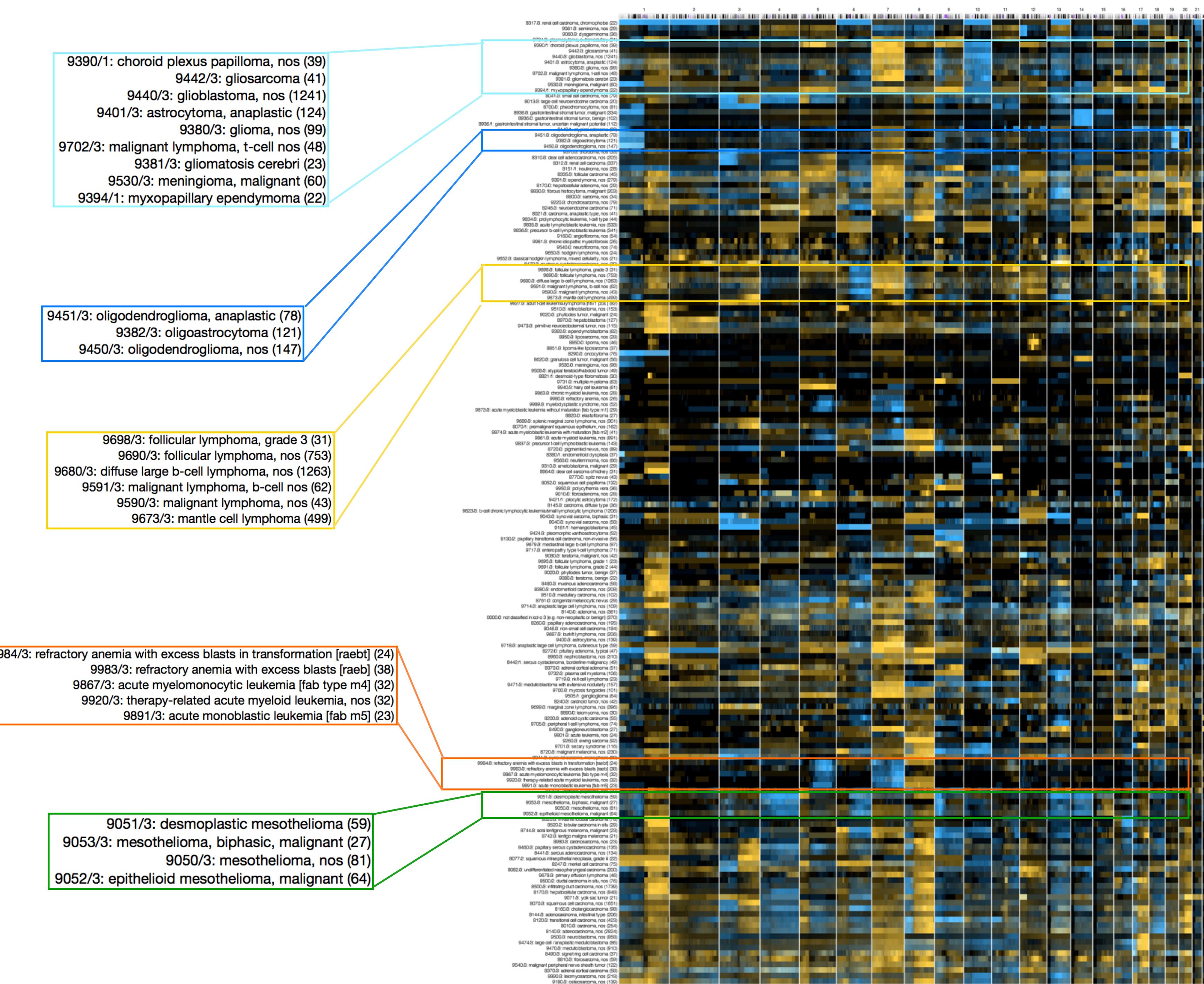
From Genomic Experiments to Experimenting with the Beacon API



Somatic Mutations In Cancer: Patterns

Making the case for genomic classifications

Some related cancer entities show similar copy number profiles



Progenetix in 2022

Cancer Genomics Reference Resource

- largest open resource for curated cancer genome profiles
- focus on copy number variations (CNV)
- >116'000 cancer CNV profiles, from >800 NCIt codes
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- structured diagnostic encodings for NCIt, ICD-O 3, UBERON
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core biosample and technical metadata annotations where accessible (TNM, genotypic sex, survival ...)
- publication database and code mapping services

Cancer CNV Profiles

- ICD-O Morphologies
- ICD-O Organ Sites
- Cancer Cell Lines
- Clinical Categories

Search Samples

arrayMap

- TCGA Samples
- 1000 Genomes
- Reference Samples
- DIPG Samples
- cBioPortal Studies
- Gao & Baudis, 2021

Publication DB

- Genome Profiling
- Progenetix Use

Services

- NCIt Mappings
- UBERON Mappings

Upload & Plot

Beacon⁺

Documentation

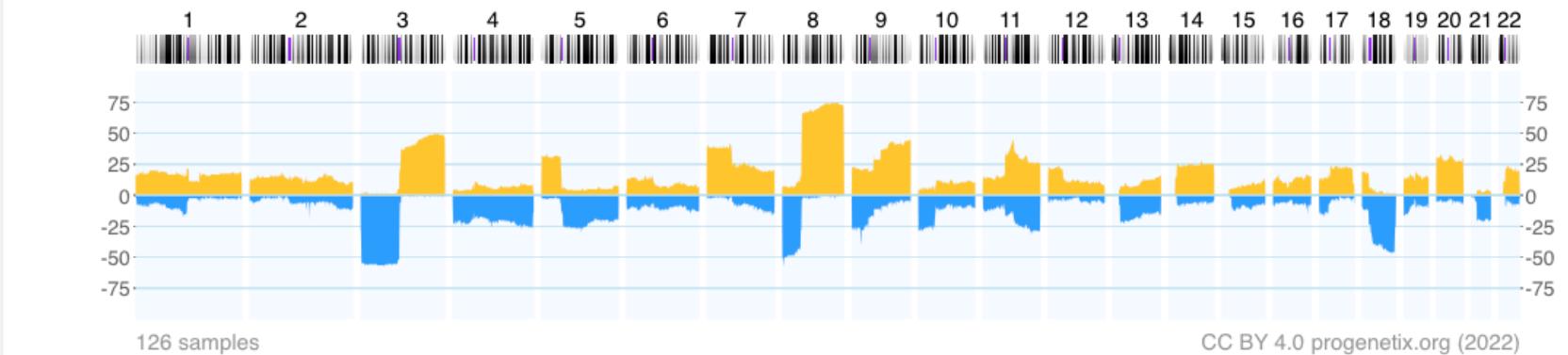
- News
- Downloads & Use Cases
- Sevices & API

Baudisgroup @ UZH

Cancer genome data @ progenetix.org

The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* from currently **142063** samples.

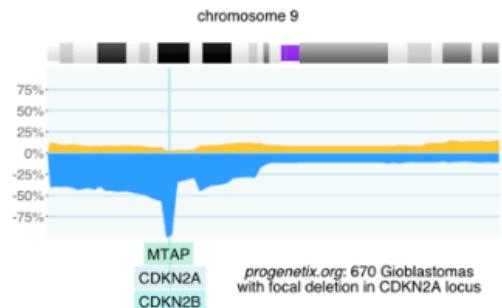
Floor of the Mouth Neoplasm (NCIT:C4401)



[Download SVG](#) | [Go to NCIT:C4401](#) | [Download CNV Frequencies](#)

Example for aggregated CNV data in 126 samples in Floor of the Mouth Neoplasm.
Here the frequency of regional **copy number gains** and **losses** are displayed for all 22 autosomes.

Progenetix Use Cases



Local CNV Frequencies

A typical use case on Progenetix is the search for local copy number aberrations - e.g. involving a gene - and the exploration of cancer types with these CNVs. The [\[Search Page \]](#) provides example use cases for designing queries. Results contain basic statistics as well as visualization and download options.

Cancer CNV Profiles

The progenetix resource contains data of **834** different cancer types (NCIt neoplasm classification), mapped to a variety of biological and technical categories. Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the [\[Cancer Types \]](#) page with direct visualization and options for sample retrieval and plotting options.

Cancer Genomics Publications

Through the [\[Publications \]](#) page Progenetix provides **4164** annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.

Progenetix in 2022

Cancer Genomics Reference Resource

- largest open resource for curated cancer genome profiles
- focus on copy number variations (CNV)
- >116'000 cancer CNV profiles, from >800 NCIt codes
- majority of data from genomic arrays with ~50% overall from SNP platforms with original data re-processing
- structured diagnostic encodings for NCIt, ICD-O 3, UBERON
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core biosample and technical metadata annotations where accessible (TNM, genotypic sex, survival ...)
- publication database and code mapping services

progenetix

Search Samples Modify Query

Assembly: GRCh38 Chro: 9 Start: 21500001-21975098 End: 21967753-22500000
Type: DEL Filters: NCIT:C3058

progenetix

Samples: 668 Variants: 286 Calls: 675

Found Variants (.pgxseg) All Sample Variants (.json) All Sample Variants (.pgxseg) Show Variants in UCSC

UCSC region JSON Response Visualization options

Results Biosamples Biosamples Map Variants

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22

75% 50% 25% 0% -25% -50% -75%

-75% -50% -25% 0% 25% 50% 75%

progenetix: 670 samples CC BY 4.0 progenetix.org (2021)

Matched Subset Codes

| Matched Subset Codes | Subset Samples | Matched Samples | Subset Match Frequencies |
|----------------------|----------------|-----------------|--------------------------|
| UBERON:0002021 | 4 | 1 | 0.250 |
| icdot-C71.4 | 4 | 1 | 0.250 |
| icdom-94403 | 4291 | 664 | 0.155 |
| NCIT:C3058 | 4375 | 664 | 0.152 |
| UBERON:0016525 | 14 | 2 | 0.143 |
| icdot-C71.1 | 14 | 2 | 0.143 |
| UBERON:0000955 | 7068 | 651 | 0.092 |
| icdot-C71.9 | 7066 | 651 | 0.092 |
| icdom-94423 | 84 | 4 | 0.048 |
| NCIT:C3796 | 84 | 4 | 0.048 |
| UBERON:0001869 | 1712 | 14 | 0.008 |
| icdot-C71.0 | 1712 | 14 | 0.008 |

Progenetix in 2022

Cancer Genomics Reference Resource

- contains special data subsets, identified using the "cohorts" concept
 - ▶ TCGA CNV data
 - ▶ 1000Genomes germline CNVs (WGS)
 - ▶ Cancer cell line CNVs with upcoming addition of annotated SNV ... data
 - ▶ cBioPortal studies
 - ▶ ...

TCGA CNV Data

Search Genomic CNV Data from TCGA

This search page accesses the TCGA subset of the Progenetix collection, based on 22142 samples (tumor and references) from The Cancer Genome Atlas project. The results are based upon data generated by the [TCGA Research Network](#). Disease-specific subsets of TCGA data (aka. projects) can be accessed below.

TCGA Cancer samples (pgx:cohort-TCGAcancers)

11090 samples

CC BY 4.0 progenetix.org (2022)

[Download SVG](#) | [Go to pgx:cohort-TCGAcancers](#) | [Download CNV Frequencies](#)

Edit Query

Publication DB

Genome Profiling
Progenetix Use

Services

NCIt Mappings
UBERON Mappings

Upload & Plot

Beacon⁺

Documentation

News
Downloads & Use Cases
Sevices & API

Filter subsets e.g. by prefix Hierarchy Depth: 2 levels

No Selection

- pgx:TCGA-ACC: TCGA ACC project (180 samples)
- pgx:TCGA-BLCA: TCGA BLCA project (810 samples)
- pgx:TCGA-BRCA: TCGA BRCA project (2219 samples)
- pgx:TCGA-CESC: TCGA CESC project (586 samples)

Progenetix in 2022

Cancer Genomics Reference Resource

- contains special data subsets, identified using the "cohorts" concept
 - ▶ TCGA CNV data
 - ▶ 1000Genomes germline CNVs (WGS)
 - ▶ Cancer cell line CNVs with upcoming addition of annotated SNV ... data
 - ▶ cBioPortal studies
 - ▶ ...



Cancer CNV Profiles
ICD-O Morphologies
ICD-O Organ Sites
Cancer Cell Lines
Clinical Categories

Search Samples

arrayMap
TCGA Samples
1000 Genomes
Reference Samples
DIPG Samples
cBioPortal Studies
Gao & Baudis, 2021

Publication DB
Genome Profiling
Progenetix Use

Services
NCIt Mappings
UBERON Mappings

Upload & Plot

Beacon⁺

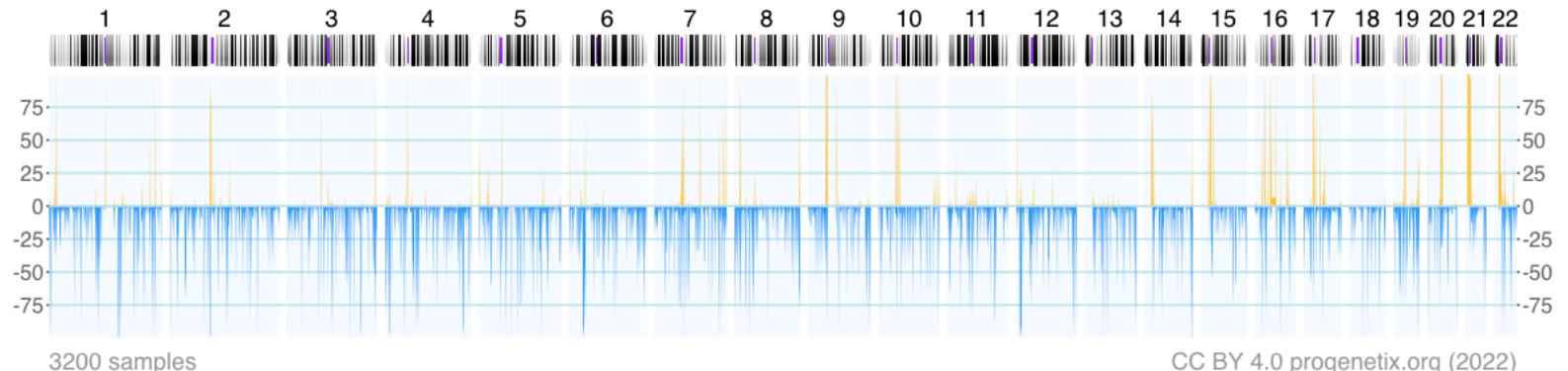
Documentation
News
Downloads & Use Cases
Sevices & API

1000 Genomes Germline CNVs

Search Genomic CNV Data from the Thousand Genomes Project

This search page accesses the reference germline CNV data of 3200 samples from the 1000 Genomes Project. The results are based on the data from the Illumina DRAGEN caller re-analysis of 3200 whole genome sequencing (WGS) samples downloaded from the AWS store of the Illumina-led reanalysis project.

1000 genomes reference samples (pgx:cohort-oneKgenomes)



Download SVG | Go to pgx:cohort-oneKgenomes | Download CNV Frequencies

Please note that the CNV spikes are based on the frequency of occurrence of any CNV in a given 1Mb interval, not on their overlap. Some genome bins may have at least one small CNV in each sample - especially in peri-centromeric regions - and therefore will display with a 100% frequency - although many of those may not overlap.

Search Samples

Range Example

Chromosome (Structural) Variant Type

Start or Position End (Range or Structural Var.)

Reference Base(s) Alternate Base(s)

The Progenetix oncogenomic resource in 2021

Qingyao Huang^{1,2}, Paula Carrio-Cordo^{1,2}, Bo Gao^{1,2}, Rahel Paloots^{1,2} and Michael Baudis^{1,2,*}

¹Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland

²Swiss Institute of Bioinformatics, Winterthurerstrasse 190, Zurich 8057, Switzerland

*Corresponding author: Tel: +41 44 635 34 86; Email: michael.baudis@mls.uzh.ch

Citation details: Huang, Q., Carrio-Cordo, P., Gao, B. et al. The Progenetix oncogenomic resource in 2021. *Database* (2021) Vol. 2021: article ID baab043; DOI: <https://doi.org/10.1093/database/baab043>

Abstract

In cancer, copy number aberrations (CNAs) represent a type of nearly ubiquitous and frequently extensive structural genome variations. To disentangle the molecular mechanisms underlying tumorigenesis as well as identify and characterize molecular subtypes, the comparative and meta-analysis of large genomic variant collections can be of immense importance. Over the last decades, cancer genomic profiling projects have resulted in a large amount of somatic genome variation profiles, however segregated in a multitude of individual studies and datasets. The Progenetix project, initiated in 2001, curates individual cancer CNA profiles and associated metadata from published oncogenomic studies and data repositories with the aim to empower integrative analyses spanning all different cancer biologies. During the last few years, the fields of genomics and cancer research have seen significant advancement in terms of molecular genetics technology, disease concepts, data standard harmonization as well as data availability, in an increasingly structured and systematic manner. For the Progenetix resource, continuous data integration, curation and maintenance have resulted in the most comprehensive representation of cancer genome CNA profiling data with 138 663 (including 115 357 tumor) copy number variation (CNV) profiles. In this article, we report a 4.5-fold increase in sample number since 2013, improvements in data quality, ontology representation with a CNV landscape summary over 51 distinctive National Cancer Institute Thesaurus cancer terms as well as updates in database schemas, and data access including new web front-end and programmatic data access.

Database URL: progenetix.org

Table 1. Statistics of samples from various data resources

| Data source | GEO | ArrayExpress | cBioPortal | TCGA | Total |
|-----------------------|-----------|--------------|------------|-----------|------------|
| No. of studies | 898 | 51 | 38 | 33 | 1939 |
| No. of samples | 63 568 | 4351 | 19 712 | 22 142 | 138 663 |
| Tumor | 52 090 | 3887 | 19 712 | 11 090 | 115 357 |
| Normal | 11 478 | 464 | 0 | 11 052 | 23 306 |
| Classifications | | | | | |
| ICD-O (Topography) | 100 | 54 | 88 | 157 | 209 |
| ICD-O (Morphology) | 246 | 908 | 265 | 140 | 491 |
| NCIt | 346 | 148 | 422 | 182 | 788 |
| Collections | | | | | |
| Individuals | 63 568 | 4351 | 19 712 | 10 995 | 127 549 |
| Biosamples | 63 568 | 4351 | 19 712 | 22 142 | 138 663 |
| Callsets ^a | 63 568 | 4351 | 19 712 | 22 376 | 138 930 |
| Variants | 5 514 126 | 118 4170 | 1 778 096 | 2 654 065 | 10 716 093 |

^aset of variants from one genotyping experiment; ICD-O, International Classification of Diseases for Oncology; NCIt, National Cancer Institute Thesaurus.

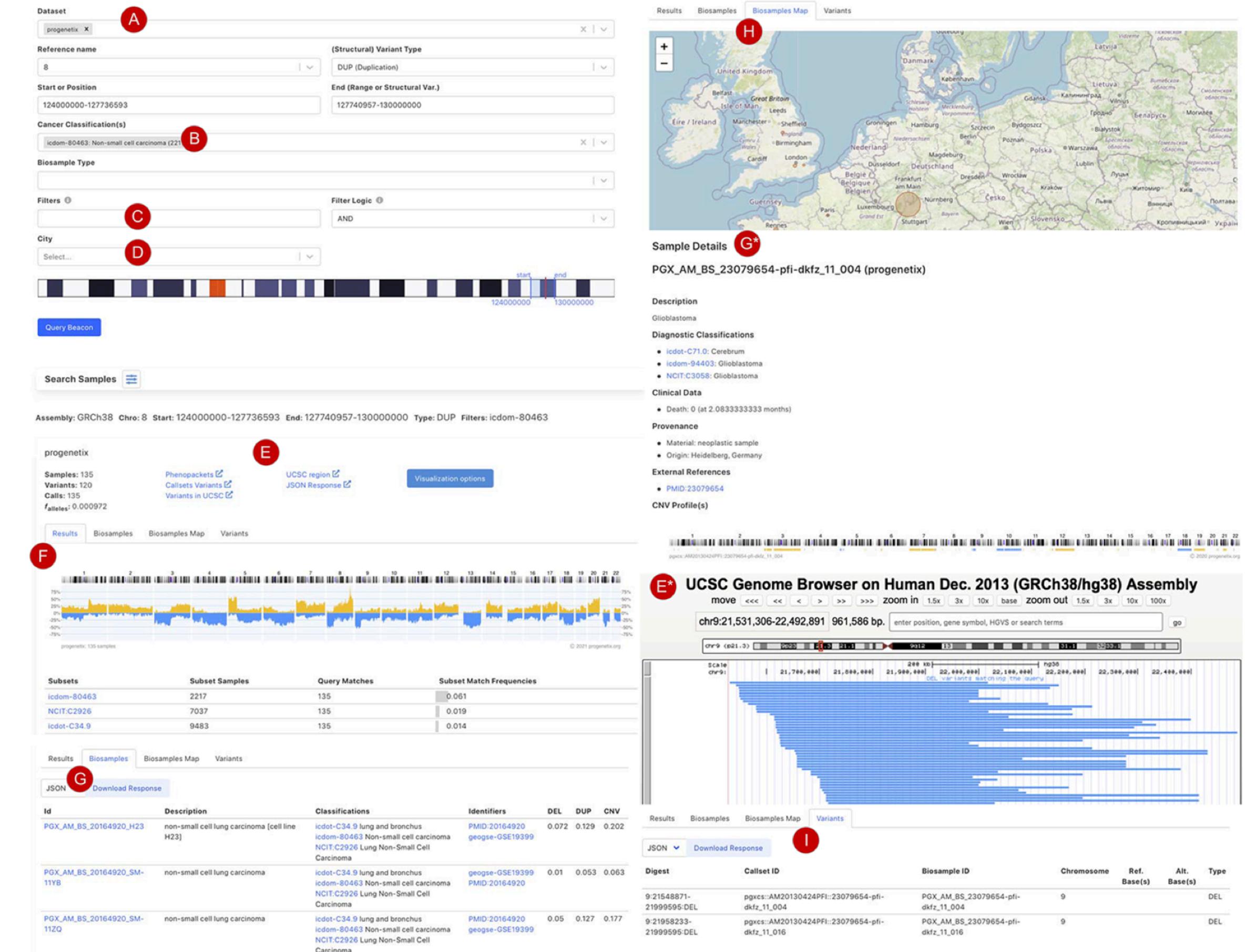
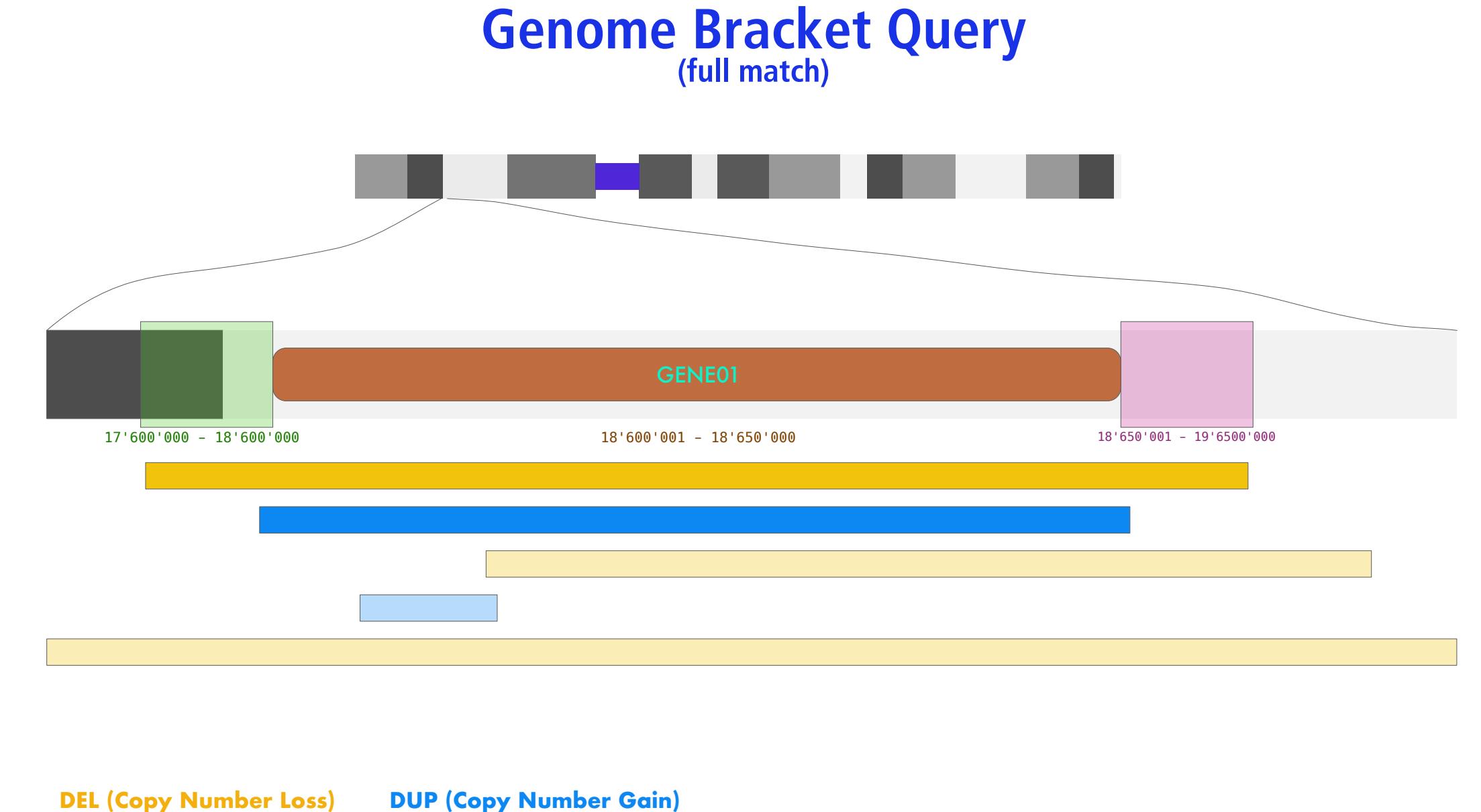


Figure 3. Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range. This example queries for a continuous, focal duplication covering the complete MYC gene's coding region with ≤ 6 Mb in size. A: Filter for dataset; B: filter for cancer classification (NCIt and ICD-O-3 ontology terms available); C: additional filter, e.g. Cellosaurus; D: additional filter for geographic location; E: external link to UCSC browser to view the alignment of matched variants; F: cancer type classification sorted by frequency of the matched biosamples present in the subset; G: list of matched variants with reference to biosamples can be downloaded in json or csv format.

Progenetix in 2022

Variant and Metadata for Sample Discovery

- positional queries for genomic variants using the **GA4GH Beacon protocol**
- metadata queries (diagnoses, identifiers, clinical classes ...) using **Beacon "filters"**



progenetix

Cancer CNV Profiles

Search Samples

Studies & Cohorts

- arrayMap
- TCGA Samples
- DIPG Samples
- Gao & Baudis, 2021
- Cancer Cell Lines

Publication DB

Services

- NCIt Mappings
- UBERON Mappings

Upload & Plot

Download Data

Beacon+

Progenetix Info

- About Progenetix
- Use Cases
- Documentation
- Baudisgroup @ UZH

Search Samples

CDKN2A Deletion Example MYC Duplication TP53 Del. in Cell Lines K-562 Cell Line

Gene Spans Cytoband(s)

This example shows the query for CNV deletion variants overlapping the CDKN2A gene's coding region with at least a single base, but limited to "highly focal" hits (here i.e. $\leq \sim 1\text{Mbp}$ in size). The query can be modified e.g. through changing the position parameters or diagnosis.

Gene Symbol i

Select...

Chromosome i

9

(Structural) Variant Type i

DEL (Deletion)

Start or Position i

21500001-21975098

End (Range or Structural Var.) i

21967753-22500000

Minimum Variant Length i

Maximal Variant Length i

Reference ID(s) i

Select...

Cancer Classification(s) i

NCIT:C3058: Glioblastoma (4375) x

Clinical Classes i

Select...

Genotypic Sex i

Select...

Biosample Type i

Select...

Filters i ✖

Filter Logic i

AND

Filter Precision i

exact

City i

Select...

Chromosome 9 i

21500001-21975098

21967753-22500000

Query Database



Onboarding

Demonstrating Compliance

- Progenetix Beacon+ has served as implementation driver since 2016
- Beacon v2 as service with protocol-driven registries for federation
- GA4GH approved Beacon v2 in April 2022

Beacon v2 GA4GH Approval Registry

Beacons:  European Genome-Phenome Archive |  progenetix |  cnag |  UNIVERSITY OF LEICESTER

 European Genome-Phenome Archive (EGA)

[Visit us](#) [Beacon API](#) [Contact us](#)

GA4GH Approval Beacon Test

This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

| | |
|-------------------------|--|
| BeaconMap | |
| Bioinformatics analysis | |
| Biological Sample | |
| Cohort | |
| Configuration | |
| Dataset | |
| EntryTypes | |
| Genomic Variants | |
| Individual | |
| Info | |
| Sequencing run | |

 progenetix+

[Visit us](#) [Beacon UI](#) [Beacon API](#) [Contact us](#)

Theoretical Cytogenetics and Oncogenomics group at UZH and SIB

Progenetix Cancer Genomics Beacon+ provides a forward looking implementation of the Beacon v2 API, with focus on structural genome variants and metadata based on the...

| | |
|-------------------------|--|
| BeaconMap | |
| Bioinformatics analysis | |
| Biological Sample | |
| Cohort | |
| Configuration | |
| Dataset | |
| EntryTypes | |
| Genomic Variants | |
| Individual | |
| Info | |
| Sequencing run | |

 Centre Nacional Analisis Genomica (CNAG-CRG)

[Visit us](#) [Beacon API](#) [Contact us](#)

Beacon @ RD-Connect

This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

| | |
|-------------------------|--|
| BeaconMap | |
| Bioinformatics analysis | |
| Biological Sample | |
| Cohort | |
| Configuration | |
| Dataset | |
| EntryTypes | |
| Genomic Variants | |
| Individual | |
| Info | |
| Sequencing run | |

 University of Leicester

[Beacon UI](#) [Beacon API](#) [Contact us](#)

Cafe Variome Beacon v2

This [Beacon](#) is based on the GA4GH Beacon [v2.0](#)

| | |
|-------------------------|--|
| BeaconMap | |
| Bioinformatics analysis | |
| Biological Sample | |
| Cohort | |
| Configuration | |
| Dataset | |
| EntryTypes | |
| Genomic Variants | |
| Individual | |
| Info | |
| Sequencing run | |

✓ Matches the Spec ✗ Not Match the Spec ⌚ Not implemented



Global Alliance
for Genomics & Health

Beacon v2 Conformity and Extensions in Progenetix

Putting the **+** into Beacon ...

- support & use of standard Beacon v2 PUT & GET variant queries, filters and meta parameters
 - variant parameters, genelid, lengths, EFO & VCF CNV types, pagination
 - widespread, self-scoping filter use for bio-, technical- and id parameters with switch for descending terms use (globally or per term if using POST)
- extensive use of handovers
 - asynchronous delivery of e.g. variant and sample data, data plots
- **+** extensions of query logic
 - optional use of OR logic for filter combinations (global)
- **+** extension of query parameters
 - geographic queries incl. \$geonear and use of GeoJSON in schemas
- ↴ ↵ ↷ ↸ no implementation of authentication on this open dataset

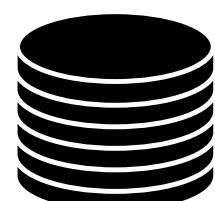
Progenetix provides a number of additional services and output formats which are initiated over the /services path or provided as request parameters and are not considered Beacon extensions (though they follow the syntax where possible).



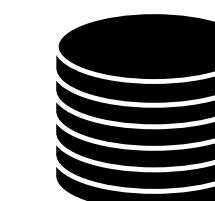
Progenetix Stack



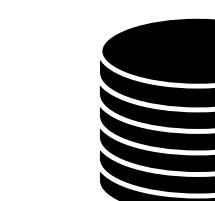
- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
 - ▶ biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads...
- the complete middleware / CGI stack is provided through the *bycon* package
 - ▶ schemas, query stack, data transformation (e.g. Phenopackets generation)...
- data collections mostly correspond to the main Beacon default model entities
 - ▶ no separate *runs* collection; integrated w/ analyses
 - ▶ *variants* are stored per observation instance



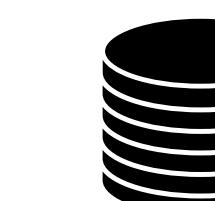
variants



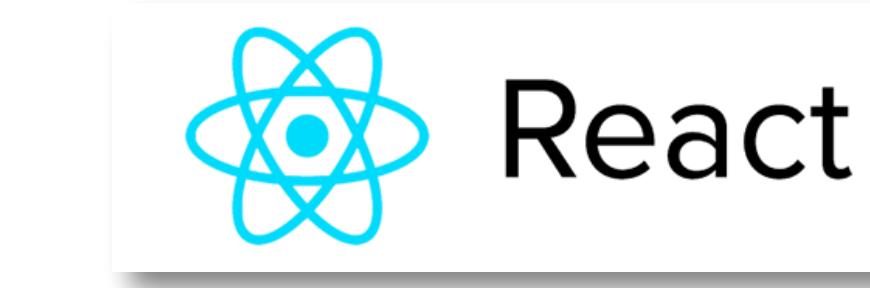
analyses



biosamples

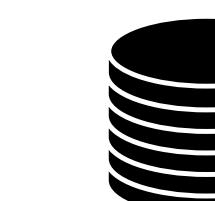


individuals

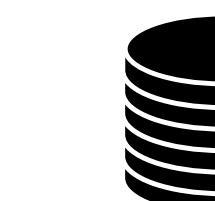


- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
 - ▶ PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703...
- *querybuffer* stores id values of all entities matched by a query and provides the corresponding access handle for **handover** generation

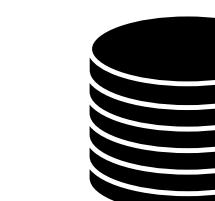
```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
_id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
  ObjectId("5bab578d727983b2e00cb505")]
```



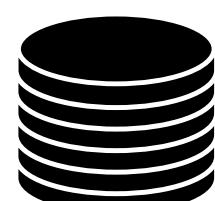
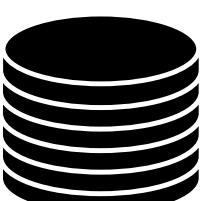
collations



geolocs



genespans publications



qBuffer

Entity collections

Utility collections

Progenetix Documentation

[Documentation Home](#)[Progenetix Source Code](#)[bycon](#)[progenetix-web](#)[PGX](#)[Additional Projects](#)[News & Changes](#)[Pages & Forms](#)[Services & API](#)[Use Case Examples](#)[Classifications, Ontologies & Standards](#)[Publication Collection](#)[Data Review](#)[Beacon+ & bycon](#)[Technical Notes](#)[Progenetix Data](#)[Baudisgroup @ UZH](#)

Rapidly evolving documentation of both the Beacon API itself and its use and technical implementation on [docs.genomebeacons.org](#) [docs.progenetix.org](#)

- for testing API responses

[/BIOSAMPLES/{ID}/G_VARIANTS](#)

- [/biosamples/pgxbs-kftva5c9/g_variants/](#)
- retrieval of all variants from a single biosample

[Base /individuals](#)[/INDIVIDUALS + QUERY](#)

- [/individuals?filters=NCIT:C7541](#)

Progenetix Source Code

With exception of some utility scripts and external dependencies (e.g. [MongoDB](#)) the software (from database interaction to website) behind Progenetix and Beacon+ is implemented in Python.

[bycon](#)

- Python based service based on the [GA4GH Beacon protocol](#)
- software powering the Progenetix resource
- [Beacon+](#) implementation(s) use the same code base

[progenetix-web](#)

- website for Progenetix and its [Beacon+](#) implementations
- provides Beacon interfaces for the [bycon](#) server, as well as other Progenetix services (e.g. the [publications](#) service)
- implemented as [React / Next.js](#) project
- contains this documentation tree here as [mkdocs](#) project, with files in the [docs](#) directory

Org.progenetix

Beacon API

Beacon-style JSON responses

The Progenetix resource's API utilizes the [bycon](#) framework for data query and delivery and represents a custom implementation of the Beacon v2 API.

The standard format for JSON responses corresponds to a generic Beacon v2 response, with the [meta](#) and [response](#) root elements. Depending on the endpoint, the main data will be a list of objects either inside [response.results](#) or (mostly) in [response.resultSets.results](#). Additionally, most API responses (e.g. for biosamples or variants) provide access to data using [handover](#) objects.

Beacon v2 Documentation

Search

 beacon-v2
☆2 48

Org.progenetix

Progenetix & Beacon⁺

The [Beacon+](#) implementation - developed in the Python & MongoDB based [bycon](#) project - implements an expanding set of Beacon v2 paths for the [Progenetix](#) resource [+](#).

Scoped responses from query object

In queries with a complete [beaconRequestBody](#) the type of the delivered data is independent of the path and determined in the [requestedSchemas](#). So far, Beacon+ will compare the first of those to its supported responses and provide the results accordingly; it doesn't matter if the endpoint was [/beacon/biosamples/](#) or [/beacon/variants/](#) etc.

Below is an example for the standard test "small deletion CNVs in the CDKN2A locus, in gliomas" Progenetix test query, here responding with the matched variants. Exchanging the [entityType](#) entry to

- { "entityType": "biosample", "schema": "https://progenetix.org/services/schemas/Biosample/" }

would change this to a biosample response. The example can be tested by POSTing this as [application/json](#) to <http://progenetix.org/beacon/variants/> or <http://progenetix.org/beacon/biosamples/>.

```
{
  "$schema": "beaconRequestBody.json",
  "meta": {
    "apiVersion": "2.0",
    "requestedSchemas": [
      {
        "entityType": "genomicVariant",
        "schema": "https://progenetix.org/services/schemas/genomicVariant"
      }
    ],
    "query": {
      "requestParameters": {
        ...
      }
    }
  }
}
```

Shoutout to Laure(e)n Fromont & Manuel Rueda for being instrumental in the Beacon v2 documentation!

pgxRpi

An interface API for analyzing Progenetix CNV data in R using the Beacon⁺ API

Beacon Path: Retrieve variants by biosample id(s)

```
https://progenetix.org/beacon/g_variants/
?biosampleIds=pgxb-s-kftvh94d,pgxb-s-kftvh94g,pgxb-s-kftvh972
&output=pgxseg
```

Beacon Path: Get biosamples by filter(s)

```
http://progenetix.org/beacon/biosamples/
?filters=NCIT:C3697&output=datatable
```

Service Path: Retrieve CNV frequencies by filter(s)

```
http://www.progenetix.org/services/intervalFrequencies/
?id=NCIT:C4323&output=pgxseg
```

pgxRpi

This is an API wrapper package to access data from Progenetix database.

You can install this package from GitHub using:

```
install.packages("devtools")
devtools::install_github("progenetix/pgxRpi")
```

If you are interested in accessing CNV variant data, get started from this [vignette](#)

If you are interested in accessing CNV frequency data, get started from this [vignette](#)

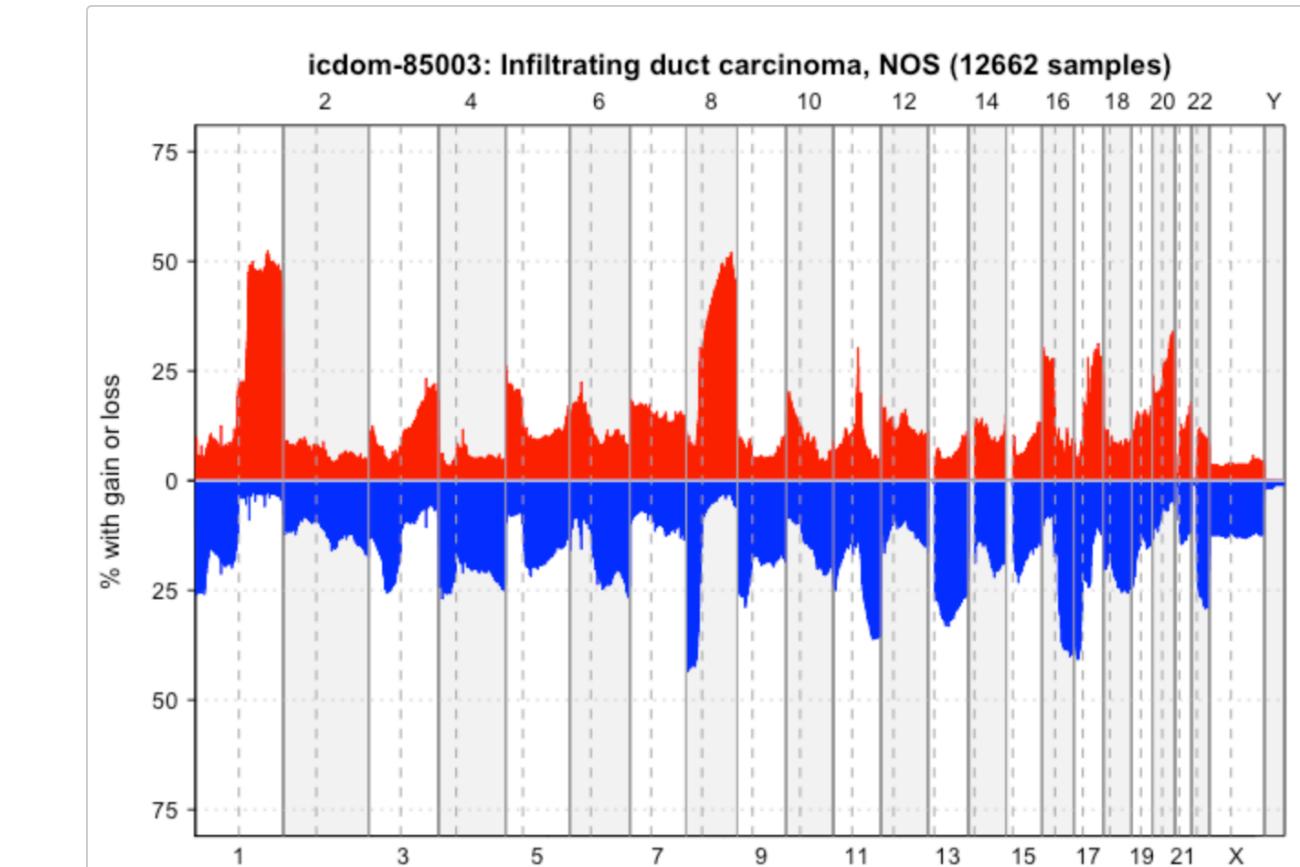
When you face problems, try to reinstall the latest version. If reinstallation doesn't help, please contact us.

```
variant_1 <- pgxLoader(type="variant", biosample_id = biosample_id)

biosamples <- pgxLoader(type="biosample", filters = "NCIT:C3059", codematches = TRUE,
                        biosample_id = c("pgxb-s-kftva5zv", "pgxb-s-kftva5zw"))
```

```
freq_pgxseg <- pgxLoader(type="frequency", output = 'pgxseg',
                           filters=c("NCIT:C4038", "pgx:icdom-85003"),
                           codematches = TRUE)
```

```
pgxFreqplot(freq_pgxseg, filters='pgx:icdom-85003')
```



Beacon⁺: Phenopackets

Testing alternative response schemas...

<http://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/phenopackets>

- the v2 default schemas are mostly aligned w/ Phenopackets v2
- creating phenopackets can be done mostly by re-wrapping of Beacon entities (individual, biosample)
- variants can be included through file resource URLs; in Beacon⁺ this is done through *ad hoc* handover URIs

```

    "id": "pgpxpf-kftx3tl5",
    "metaData": {
      "phenopacketSchemaVersion": "v2",
      "resources": [
        {
          "id": "NCIT",
          "iriPrefix": "http://purl.obolibrary.org/obo/NCIT_",
          "name": "NCIt Plus Neoplasm Core",
          "namespacePrefix": "NCIT",
          "url": "http://purl.obolibrary.org/obo/ncit/neoplasm-core.c",
          "version": "2022-04-01"
        },
        {
          "subject": {
            "dataUseConditions": {
              "id": "DUO:0000004",
              "label": "no restriction"
            },
            "diseases": [
              {
                "clinicalTnmFinding": [],
                "diseaseCode": {
                  "id": "NCIT:C3099",
                  "label": "Hepatocellular Carcinoma"
                },
                "onset": {
                  "age": "P48Y9M26D"
                },
                "stage": {
                  "id": "NCIT:C27966",
                  "label": "Stage I"
                }
              }
            ],
            "id": "pgxind-kftx3tl5",
            "sex": {
              "id": "PATO:0020001",
              "label": "male genotypic sex"
            },
            "updated": "2018-12-04 14:53:11.674000",
            "vitalStatus": {
              "status": "UNKNOWN_STATUS"
            }
          },
          "biosamples": [
            {
              "biosampleStatus": {
                "id": "EFO:0009656",
                "label": "neoplastic sample"
              },
              "dataUseConditions": {
                "id": "DUO:0000004",
                "label": "no restriction"
              },
              "description": "Primary Tumor",
              "externalReferences": [
                {
                  "id": "pgx:TCGA-0004d251-3f70-4395-b175-c94c2f5b1b81",
                  "label": "TCGA case_id"
                },
                {
                  "id": "pgx:TCGA-TCGA-DD-AAVP",
                  "label": "TCGA submitter_id"
                },
                {
                  "id": "pgx:TCGA-9259e9ee-7279-4b62-8512-509cb705029c",
                  "label": "TCGA sample_id"
                },
                {
                  "id": "pgx:TCGA-LIHC",
                  "label": "TCGA LIHC project"
                }
              ],
              "files": [
                {
                  "fileAttributes": {
                    "fileFormat": "pgxseg",
                    "genomeAssembly": "GRCh38"
                  },
                  "uri": "https://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/variants/?output=pgxseg"
                },
                {
                  "histologicalDiagnosis": {
                    "id": "NCIT:C3099",
                    "label": "Hepatocellular Carcinoma"
                  },
                  "id": "pgxbs-kftvhyvb",
                  "individualId": "pgxind-kftx3tl5",
                  "pathologicalStage": {
                    "id": "NCIT:C27966",
                    "label": "Stage I"
                  },
                  "sampledTissue": {
                    "id": "UBERON:0002107",
                    "label": "liver"
                  },
                  "timeOfCollection": {
                    "age": "P48Y9M26D"
                  }
                }
              ]
            }
          ]
        }
      ]
    }
  
```

Beacon⁺: Phenopackets

Testing alternative response schemas...

<http://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/phenopackets>

- the v2 default schemas are mostly aligned w/ Phenopackets v2
- creating phenopackets can be done mostly by re-wrapping of Beacon entities (individual, biosample)
- variants can be included through file resource URLs; in Beacon⁺ this is done through *ad hoc* handover URIs

```

{
  "id": "pgxpxf-kftx3tl5",
  "metaData": {
    "phenopacketSchemaVersion": "v2",
    "resources": [
      {
        "id": "NCIT",
        "iriPrefix": "http://purl.obolibrary.org/obo/NCIT_",
        "name": "NCIt Plus Neoplasm Core",
        "namespacePrefix": "NCIT",
        "url": "http://purl.obolibrary.org/obo/ncit/neoplasm-core.owl"
        "version": "2022-04-01"
      }
    ],
    "files": [
      {
        "fileAttributes": {
          "fileFormat": "pgxseg",
          "genomeAssembly": "GRCh38"
        },
        "uri": "https://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/variants/?output=pgxseg"
      }
    ],
    "biosamples": [
      {
        "biosampleStatus": {
          "id": "EFO:0009656",
          "label": "neoplastic sample"
        },
        "dataUseConditions": {
          "id": "DUO:000004",
          "label": "no restriction"
        },
        "description": "Primary Tumor",
        "externalReferences": [
          {
            "uri": "https://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/variants/?output=pgxseg"
          }
        ]
      }
    ],
    "individual": {
      "fileAttributes": {
        "fileFormat": "pgxseg",
        "genomeAssembly": "GRCh38"
      },
      "uri": "https://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/variants/?output=pgxseg"
    },
    "histologicalDiagnosis": {
      "id": "NCIT:C3099",
      "label": "Hepatocellular Carcinoma"
    },
    "id": "pgxbs-kftvhyvb",
    "individualId": "pgxind-kftx3tl5",
    "pathologicalStage": {
      "id": "NCIT:C27966",
      "label": "Stage I"
    },
    "sex": {
      "id": "PATO:0020001",
      "label": "male genotypic sex"
    },
    "updated": "2018-12-04 14:53:11.674000",
    "vitalStatus": {
      "status": "UNKNOWN_STATUS"
    }
  }
}

```

Beacon⁺: Phenopackets

Testing alternative response schemas...

<http://progenetix.org/beacon/biosamples/pgxbs-kftvhvzb/phenopackets>

- the v2 default schemas are mostly aligned w/ Phenopackets v2
- creating phenopackets can be done mostly by re-wrapping of Beacon entities (individual, biosample)
- variants can be included through file resource URLs; in Beacon⁺ this is done through *ad hoc* handover URIs

```
bios_s = data_db["biosamples"].find({"individual_id":ind["id"]})

for bios in bios_s:

    bios.update({
        "files": [
            {
                "uri": "{}/beacon/biosamples/{}/variants/?output=pgxseg".format(server, bios["id"]),
                "file_attributes": {
                    "genomeAssembly": "GRCh38",
                    "fileFormat": "pgxseg"
                }
            }
        ]
    })
    for k in bios_pop_keys:
        bios.pop(k, None)

    clean_empty_fields(bios)

    pxf_bios.append(bios)

def remap_phenopackets(ds_id, r_s_res, byc):

    if not "phenopacket" in byc["response_entity_id"]:
        return r_s_res

    mongo_client = MongoClient()
    data_db = mongo_client[ds_id]
    pxf_s = []

    for ind_i, ind in enumerate(r_s_res):

        pxf = phenopack_individual(ind, data_db, byc)
        pxf_s.append(pxf)

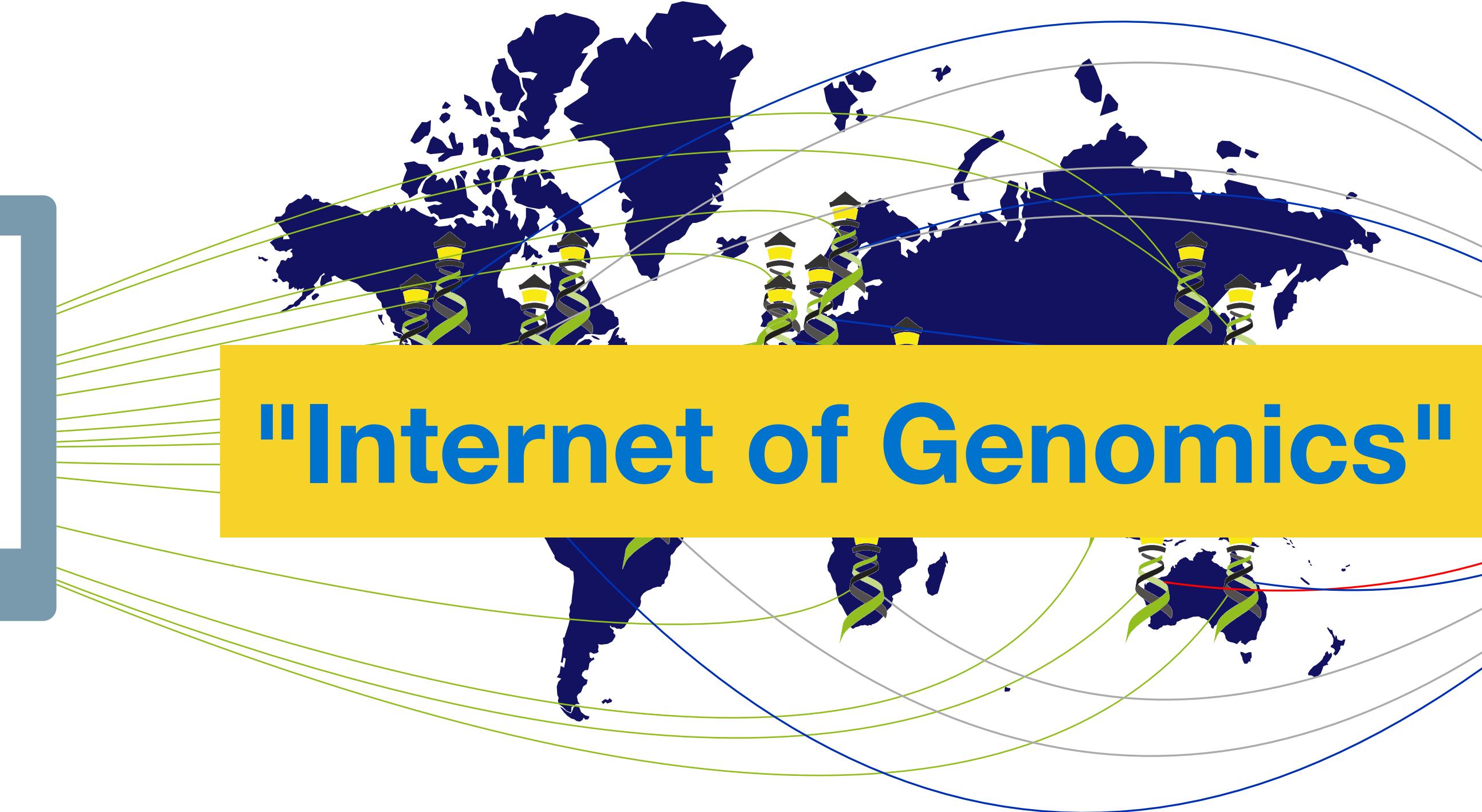
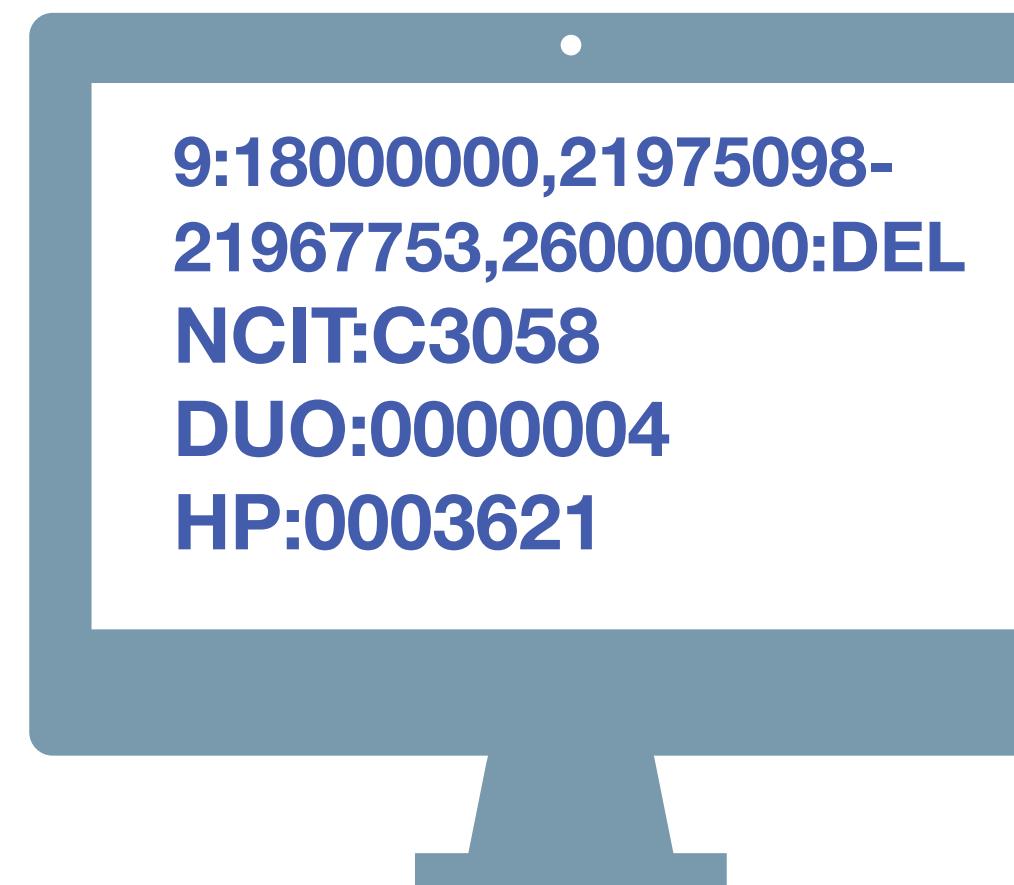
    return pxf_s
```

Future?

Some proposals for a stepwise Beacon protocol extension

- Boolean options for chaining filters
 - ➡ use of heterogeneous/alternative annotations within and across resources
- Phenopackets support as a (the?) default format for biodata export
- PXF as request documents
- Focus on service & resource discovery
- ELIXIR Beacon Network, including translations for federated queries to Beacon and Beacon-like resources





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



Beacon v2 API

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



Thank You!

...all Beacon developers, managers, contributors & users!

...current + former Progenetix contributors, especially
Haoyang Cai, Bo Gao, Linda Grob, Saumya Gupta, Qingyao Huang,
Nitin Kumar, **Rahel Paloots**, Prisni Rath, **Ziying Yang & Hangjia Zhao**



University of
Zurich^{UZH}



Global Alliance
for Genomics & Health

