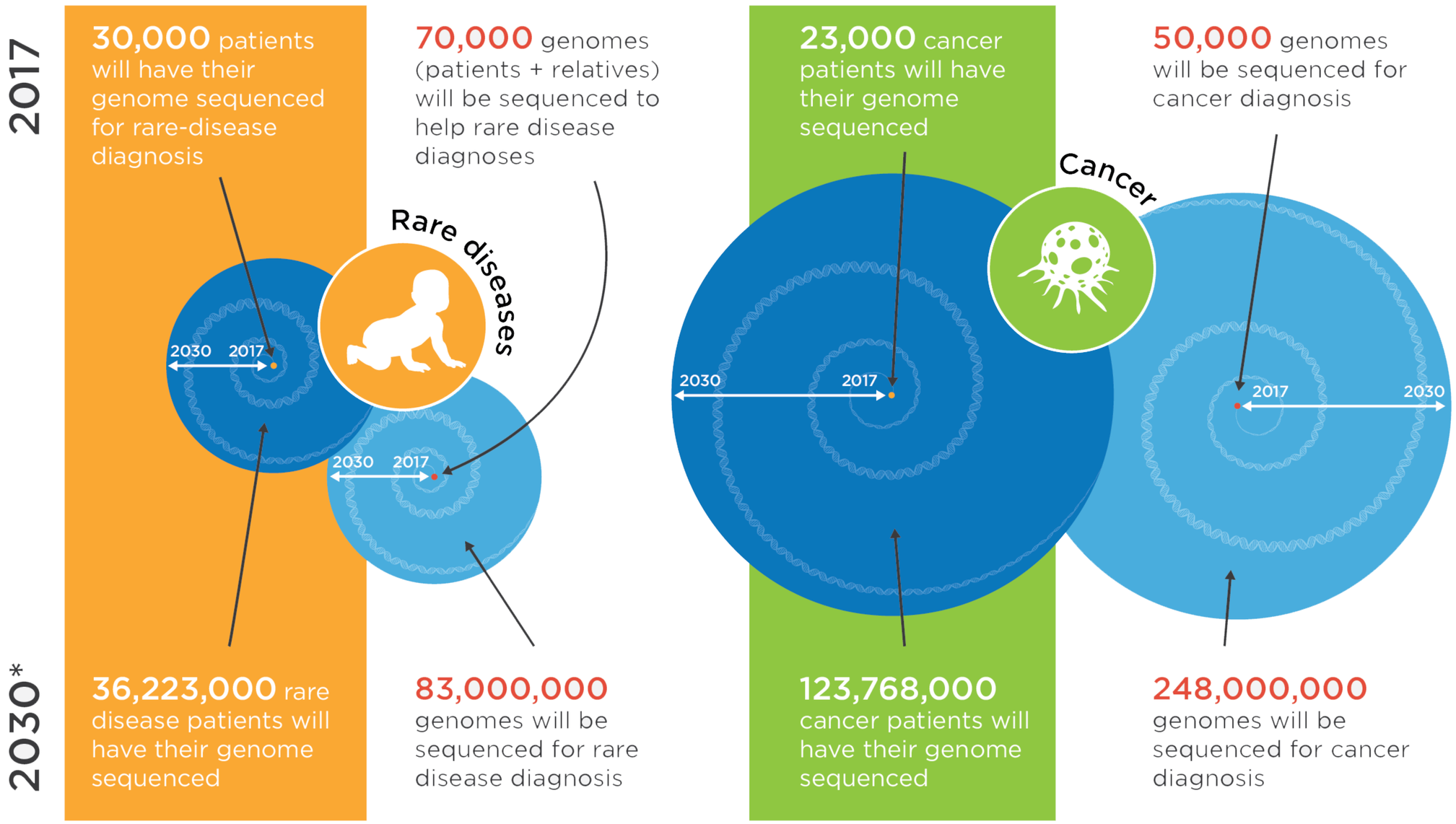
progenetix

Progenetix & Beacon+

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

Michael Baudis | BIO392 HS23 | 2023-09-27

elixir



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

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
- September 2023 - 11th Plenary, San Francisco

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

Global Alliance for Genomics & Health

GENOMICS

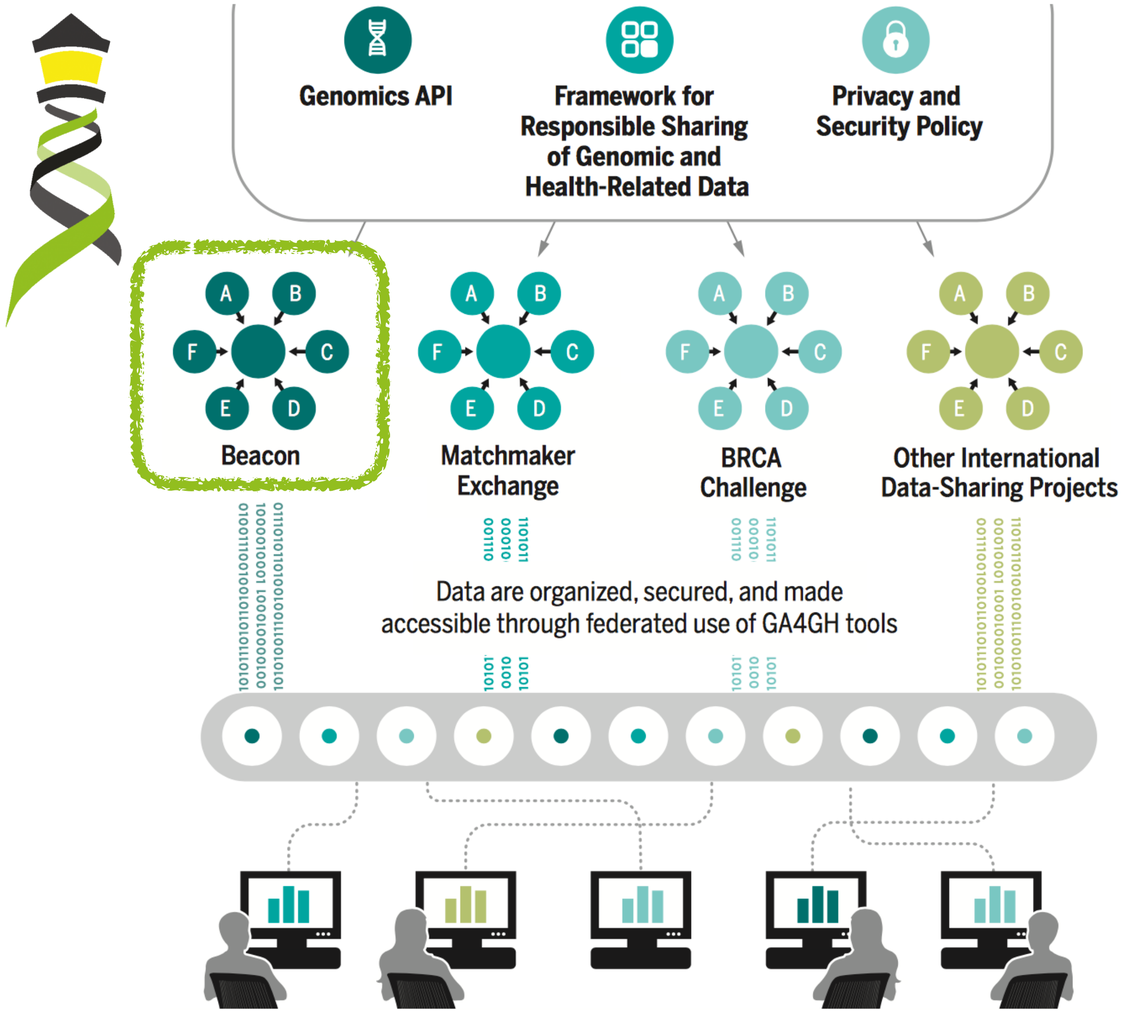
A federated ecosystem for sharing genomic, clinical data

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.



DNASTACK

Global Alliance for Genomics & Health

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.

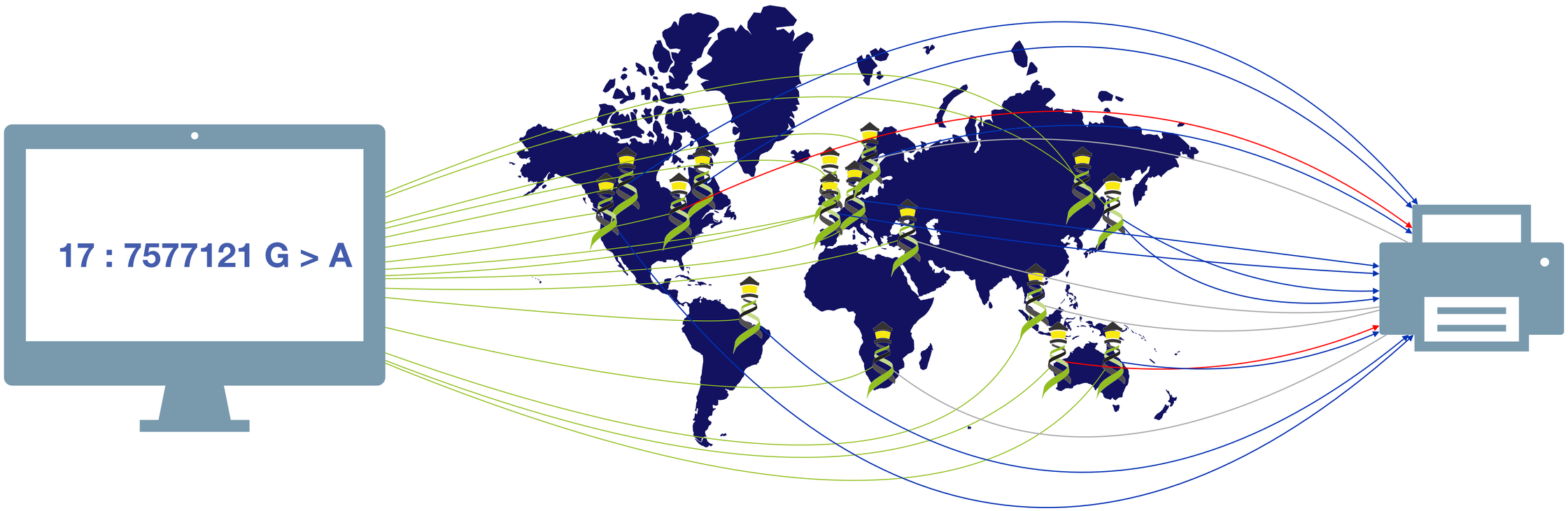
A number of more useful first versions have been suggested.

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*** …

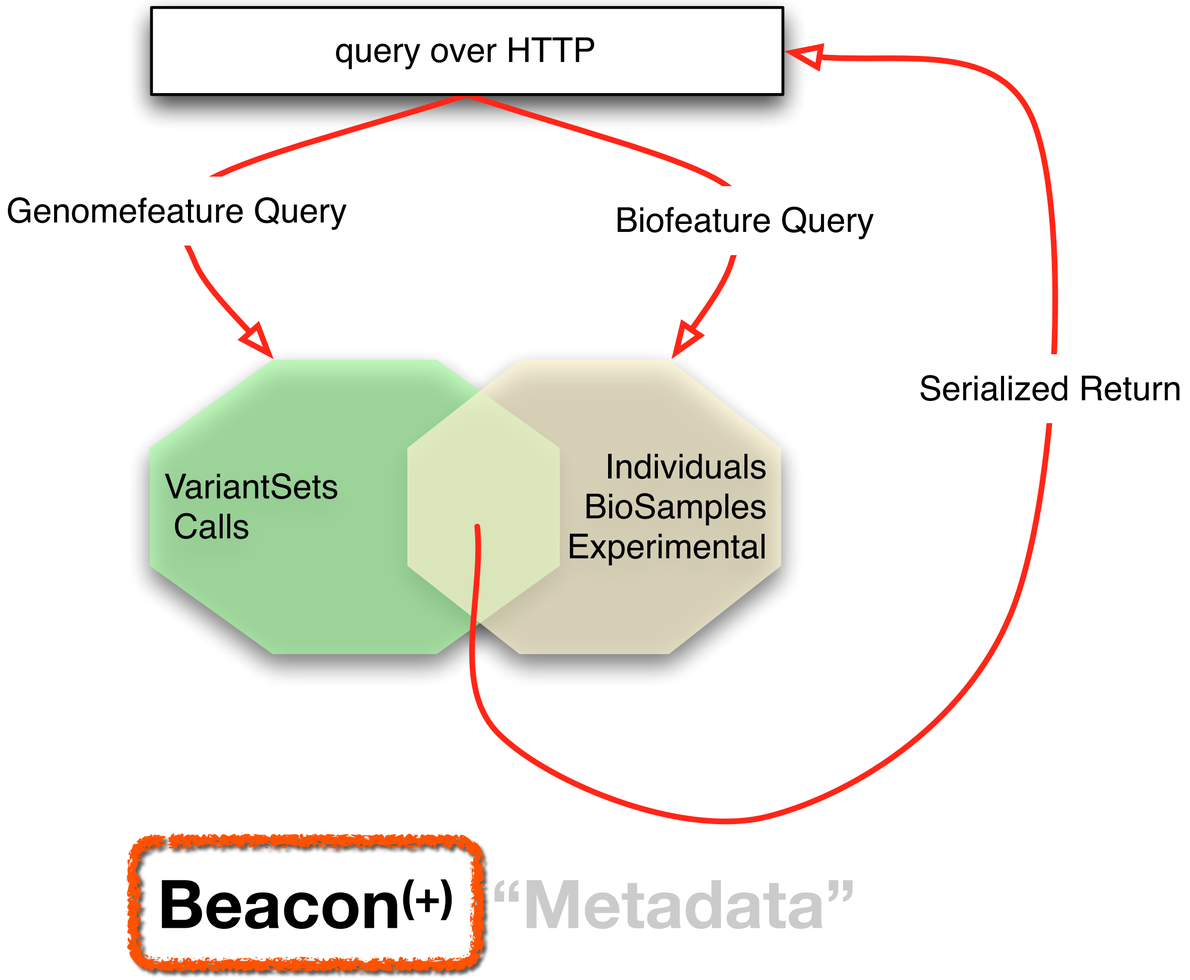


Have you seen this variant? It came up in my patient and we don't know if this is a common SNP or worth following up.

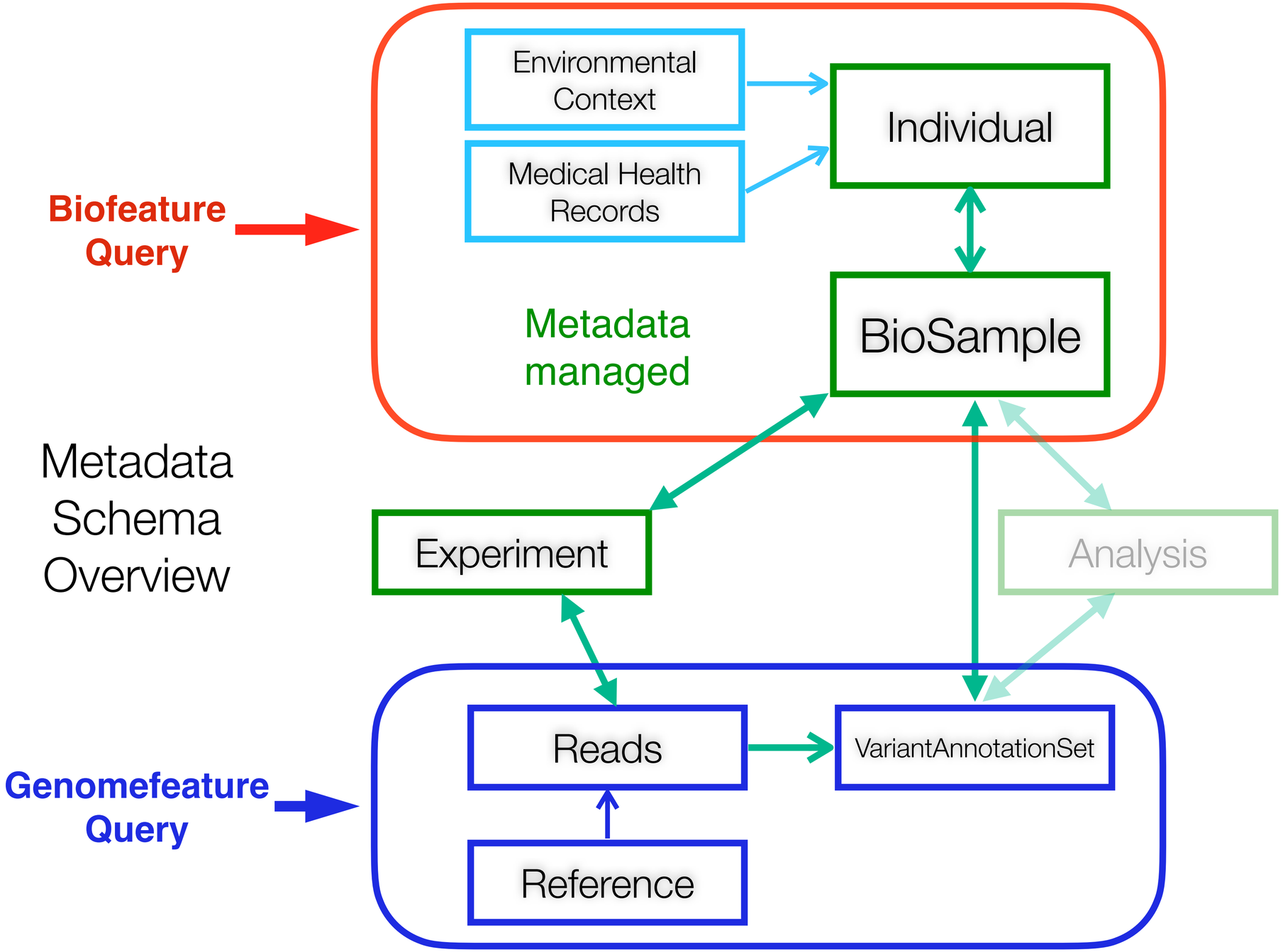
A Beacon network federates *genome variant queries* across databases that support the ***Beacon API***

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

Minimal GA4GH query API structure

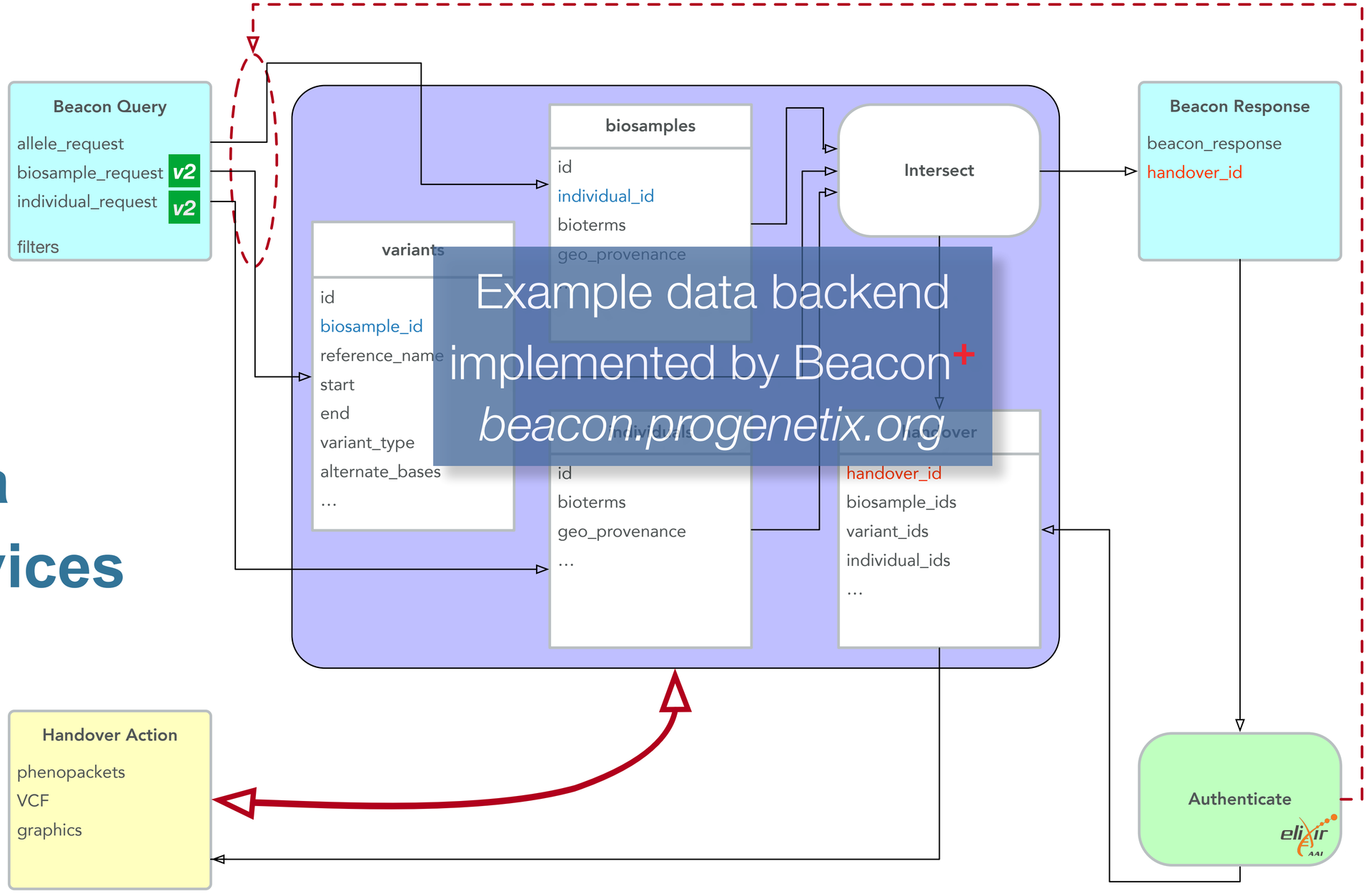


Beacon+ in 2016 - Driving GA4GH implementations



Beacon & Handover

Beacons v1.1 supports **data delivery services**



Beacon I/O

Handover

Authentication

Michael Baudis

Global Alliance for Genomics & Health

SIB

University of Zurich UZH

elixir

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**


```
{
"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",
```


Global Alliance for Genomics & Health

SIB

University of Zurich UZH

elixir

Beacon+

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.

progenetix SIB SPHN


```
[
  {
    "callset_id": "cs-cellosaurus:CVCL_EI02",
    "info": {
      "cellosaurus": {
        "cell_line": "BT474-LAPRa",
        "id": "CVCL_EI02",
        "cellosaurus_variant_name": "TP53 p.Glu285Lys (c.853G&gt;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&gt;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&gt;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&gt;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-cellosa
```

ELIXIR ALL HANDS 2022

	Beacon v1 Development	Beacon v2 Development	Related ...
2014	GA4GH founding event; Jim Ostell proposes Beacon concept including "more features ... version 2"		
2015	• beacon-network.org aggregator created by DNAstack		• ELIXIR starts Beacon project support
2016	• Beacon v0.3 release • work on queries for structural variants (brackets for fuzzy start and end parameters...)	• Beacon+ concept implemented on progenetix.org • concepts from GA4GH Metadata (ontologies…) • entity-scoped query parameters ("individual.age")	
2017	• OpenAPI implementation • integrating CNV parameters (e.g. "startMin, statMax")	• Beacon+ demos "handover" concept	• GA4GH re-structuring (workstreams...) • Beacon part of Discovery WS
2018	• Beacon v0.4 release in January; feature release for GA4GH approval process • GA4GH Beacon v1 approved at Oct plenary		• new Beacon website (March)
2019	• ELIXIR Beacon Network	• Beacon hackathon Stockholm; settling on "filters" • Barcelona goes Zurich developers meeting • Beacon API v2 Kick off • adopting "handover" concept	• Beacon publication at Nature Biotechnology
2020		• "Scouts" teams working on different aspects - filters, genomic variants, compliance ... • discussions w/ clinical stakeholders	
2021		• framework + models concept implemented • range and bracket queries, variant length parameters • starting of GA4GH review process	
2022		• further changes esp. in default model, aligning with Phenopackets and VRS • unified beacon-v2 code & docs repository • Beacon v2 approved at Apr GA4GH Connect	• *docs.genomebeacons.org*

ELIXIR ALL HANDS 2022

	Beacon v1 Development	Beacon v2 Development	Related ...
2014	GA4GH founding event; Jim Ostell proposes Beacon concept including "more features ... version 2"		
2015	• beacon-network.org aggregator created by DNAstack		• ELIXIR starts Beacon project support
2016	• Beacon v0.3 release • work on queries for structural variants (brackets for fuzzy start and end parameters...)	• Beacon+ concept implemented on progenetix.org • concepts from GA4GH Metadata (ontologies…) • entity-scoped query parameters ("individual.age")	
2017	• OpenAPI implementation • integrating CNV parameters (e.g. "startMin, statMax")	• Beacon+ demos "handover" concept	• GA4GH re-structuring (workstreams...) • Beacon part of Discovery WS
2018	• Beacon v0.4 release in January; feature release for GA4GH approval process • GA4GH Beacon v1 approved at Oct plenary		• new Beacon website (March)
2019	• ELIXIR Beacon Network	• Beacon hackathon Stockholm; settling on "filters" • Barcelona goes Zurich developers meeting • Beacon API v2 Kick off • adopting "handover" concept	• Beacon publication at Nature Biotechnology
2020		• "Scouts" teams working on different aspects - filters, genomic variants, compliance ... • discussions w/ clinical stakeholders	
2021		• framework + models concept implemented • range and bracket queries, variant length parameters • starting of GA4GH review process	• Phenopackets v2 approved
2022		• further changes esp. in default model, aligning with Phenopackets and VRS • unified beacon-v2 code & docs repository • Beacon v2 approved at Apr GA4GH Connect	• *docs.genomebeacons.org*

ELIXIR ALL HANDS 2022

Progenetix Genomics Resource

From Genomic Experiments to Experimenting with the Beacon API

progenetix

Theodor Boveri (1914)

Observations in sea urchin eggs

- **Cell-cycle checkpoints** (“Hemmungseinrichtung”)
- **Tumour-suppressor genes** (“Teilungshemmende Chromosomen”), which may be overcome by external signals, and can be eliminated during tumour progression
- **Oncogenes** (“Teilungsfoerdernde Chromosomen”) that become amplified (“im permanenten Übergewicht”)
- **Progression** (benign to malignant), w/ sequential changes of chromosomes
- Clonal origin & Genetic mosaicism
- Cancer **predisposition** through inheritance of “chromosomes” that are less able to suppress malignancy
- Inheritance of the same 'weak chromosome' from both parents leads to **homozygosity** and, consequently, to high-penetrance cancer syndromes - (e.g. xeroderma pigmentosum)
- Wounding and inflammation in tumour promotion; loss of cell adhesion in metastasis; sensitivity of malignant cells to radiation therapy (based on Hertwig *et al.*)

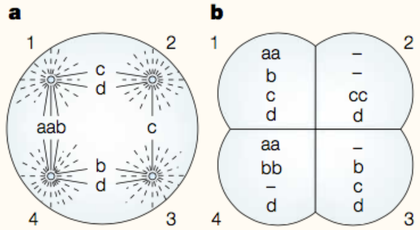
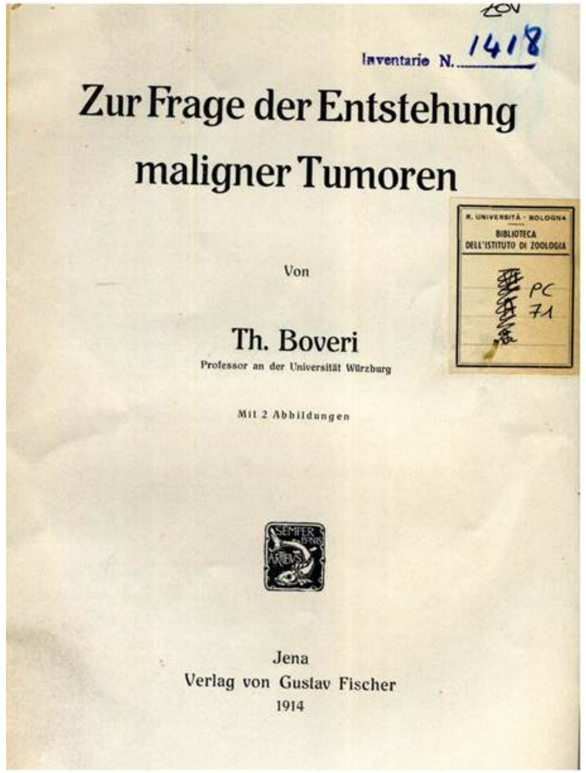


Figure 2 | **Multiple cell poles cause unequal segregation of chromosomes. a** | Boveri showed that fertilization of sea-urchin eggs by two sperm results in multiple cell poles. Individual chromosomes then attach to different combinations of poles — for example, one copy of chromosome c is attached to poles 1 and 2, and one copy is attached to poles 2 and 3. **b** | Chromosomes are segregated to the four poles at cell division, leaving some cells with too many copies of the chromosomes and some with too few — for example, cell 2 has two copies of chromosome c and cell 4 has none.

Allan Balmain
Cancer genetics: from Boveri and Mendel to microarrays.
NatRev Cancer (2001); 1: 77-82

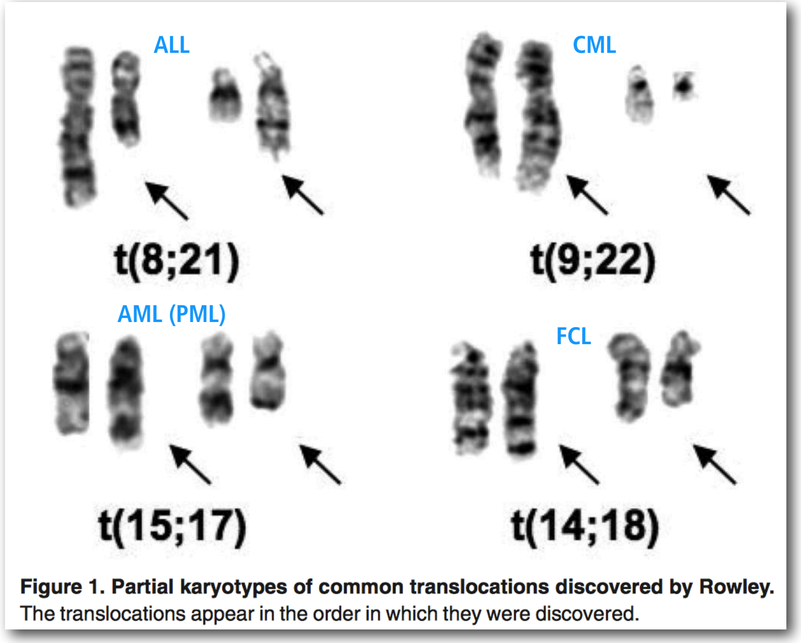


Anna Di Lonardo , Sergio Nasi , Simonetta Pulciani
Cancer: We Should Not Forget The Past
Journal of Cancer (2015), Vol. 6: 29-39
(for book cover & summary)

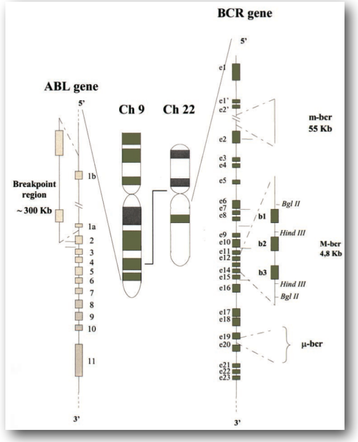
Janet Rowley (1972/73)

Chromosomal translocations in cancer

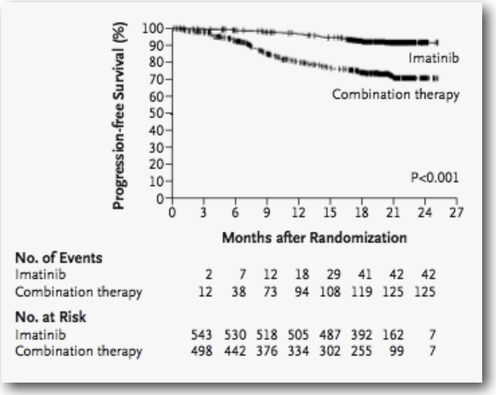
- Recurrent chromosomal translocations in leukemias and lymphomas
- "Philadelphia chromosome" in CML (Nowell & Hungerford, 1960) represents a reciprocal translocation between chromosomes 9 and 22
- 1972: t(8;21) ALL manuscript rejected by NEJM
- 1973: t(9;22) manuscript rejected by *Nature* "*with some reasonable comments and some truly wrong*"
- Clinical implications: **Tyrosine Kinase inhibitors** as standard first-line therapy in CML
  - first trials in 1998 (STI-571; Imatinib/Gleevec)
  - cf. Druker BJ, Lydon NB (2000). Lessons learned from the development of an Abl tyrosine kinase inhibitor… J Clin Invest 2000;105:3-7)



Janet D Rowley. Chromosomal translocations: revisited yet again
Blood (2008), 112(6)



Pane et al. BCR/ABL genes ….
Oncogene (2002), 21 (56)

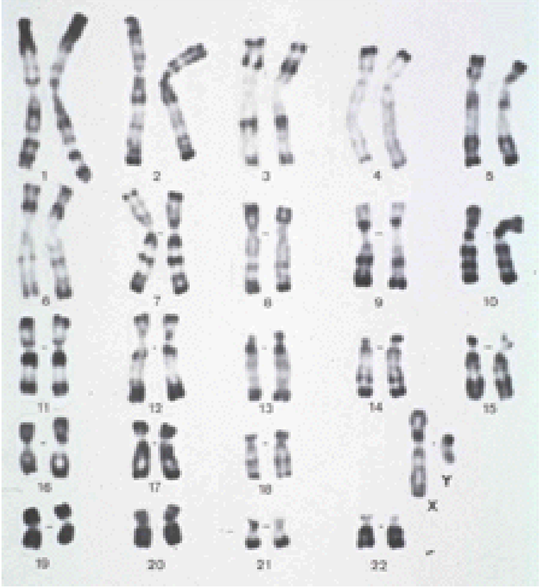


Event free Survival in first large Imatinib Trials

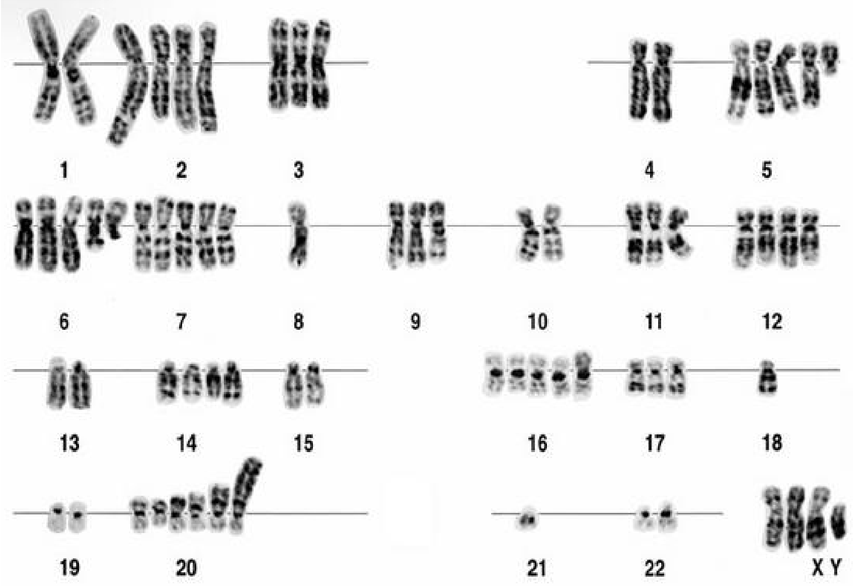
O'Brien et al. Imatinib compared with interferon and low-dose cytarabine…
NEJM (2003) vol. 348 (11)

Genomic changes at the DNA level are hallmarks of cancer

We inherited 23 paternal and 23 maternal chromosomes, mostly identical.



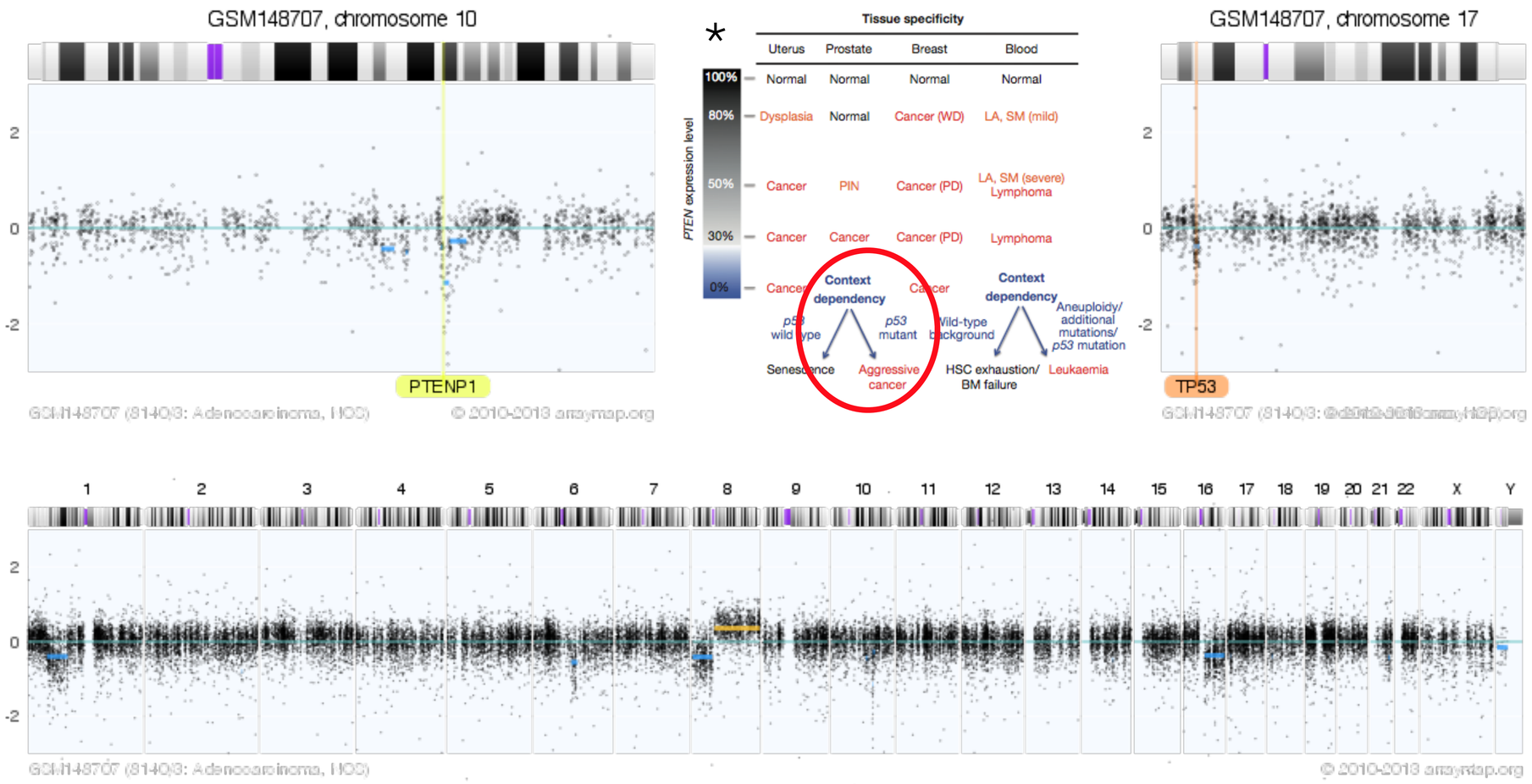
Normal karyotype



Tumor karyotype

Our goal: identify CN changes to improve characterization, classification, and treatment of cancers

Gene dosage phenomena beyond simple on/off effects



Combined heterozygous deletions involving *PTEN* and TP53 loci in a case of prostate adenocarcinoma
(GSM148707, PMID 17875689, Lapointe *et al.*, CancRes 2007)

* A. H. Berger, A. G. Knudson, and P. P. Pandolfi, “A continuum model for tumour suppression,” *Nature*, vol. 476, no. 7359, pp. 163–169, Aug. 2011.

arrayMap

Somatic Mutations In Cancer: Patterns

Making the case for genomic classifications

Some related cancer entities show similar copy number profiles

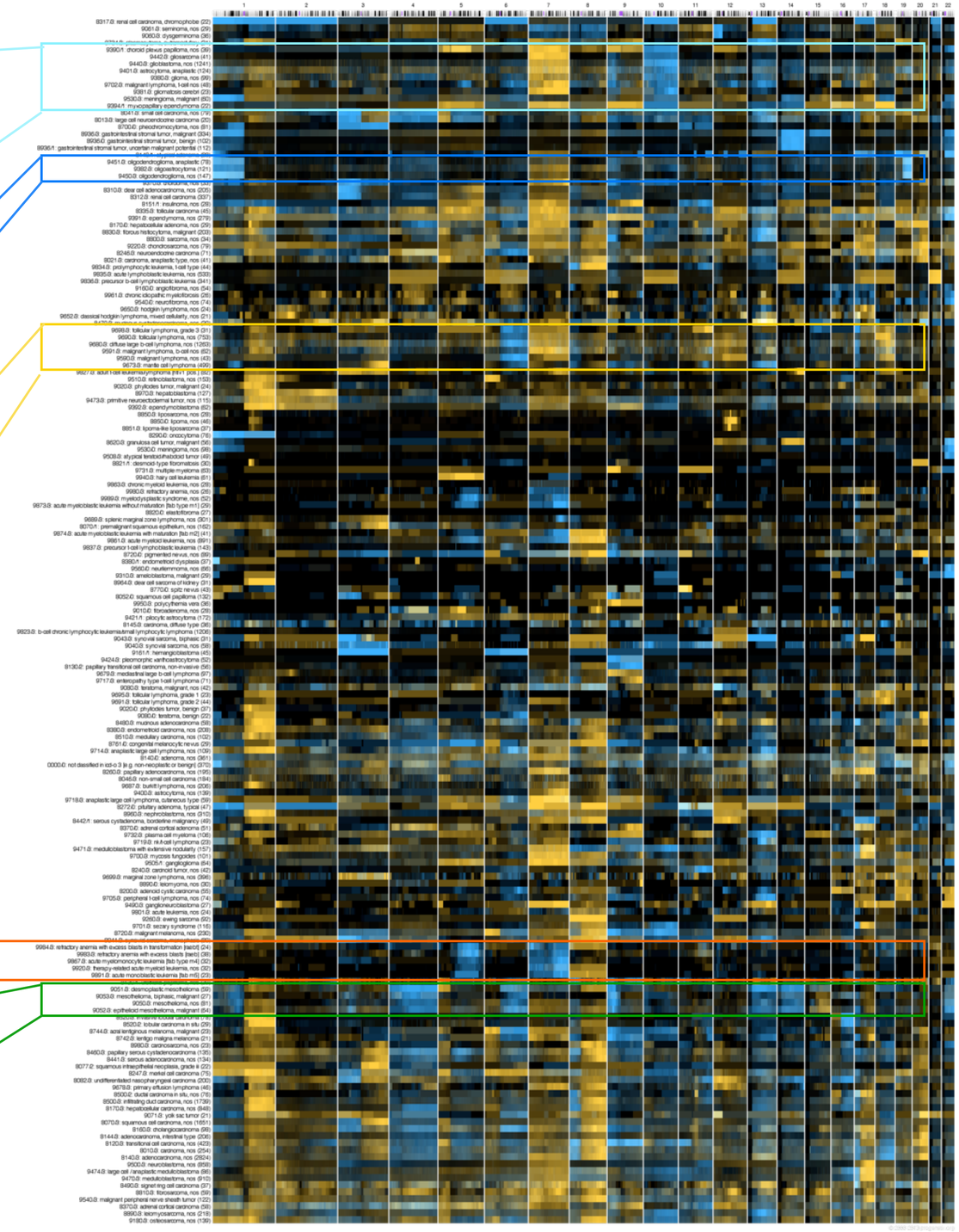
9390/1: choroid plexus papilloma, nos (39)
9442/3: gliosarcoma (41)
9440/3: glioblastoma, nos (1241)
9401/3: astrocytoma, anaplastic (124)
9380/3: glioma, nos (99)
9702/3: malignant lymphoma, t-cell nos (48)
9381/3: gliomatosis cerebri (23)
9530/3: meningioma, malignant (60)
9394/1: myxopapillary ependymoma (22)

9451/3: oligodendroglioma, anaplastic (78)
9382/3: oligoastrocytoma (121)
9450/3: oligodendroglioma, nos (147)

9698/3: follicular lymphoma, grade 3 (31)
9690/3: follicular lymphoma, nos (753)
9680/3: diffuse large b-cell lymphoma, nos (1263)
9591/3: malignant lymphoma, b-cell nos (62)
9590/3: malignant lymphoma, nos (43)
9673/3: mantle cell lymphoma (499)

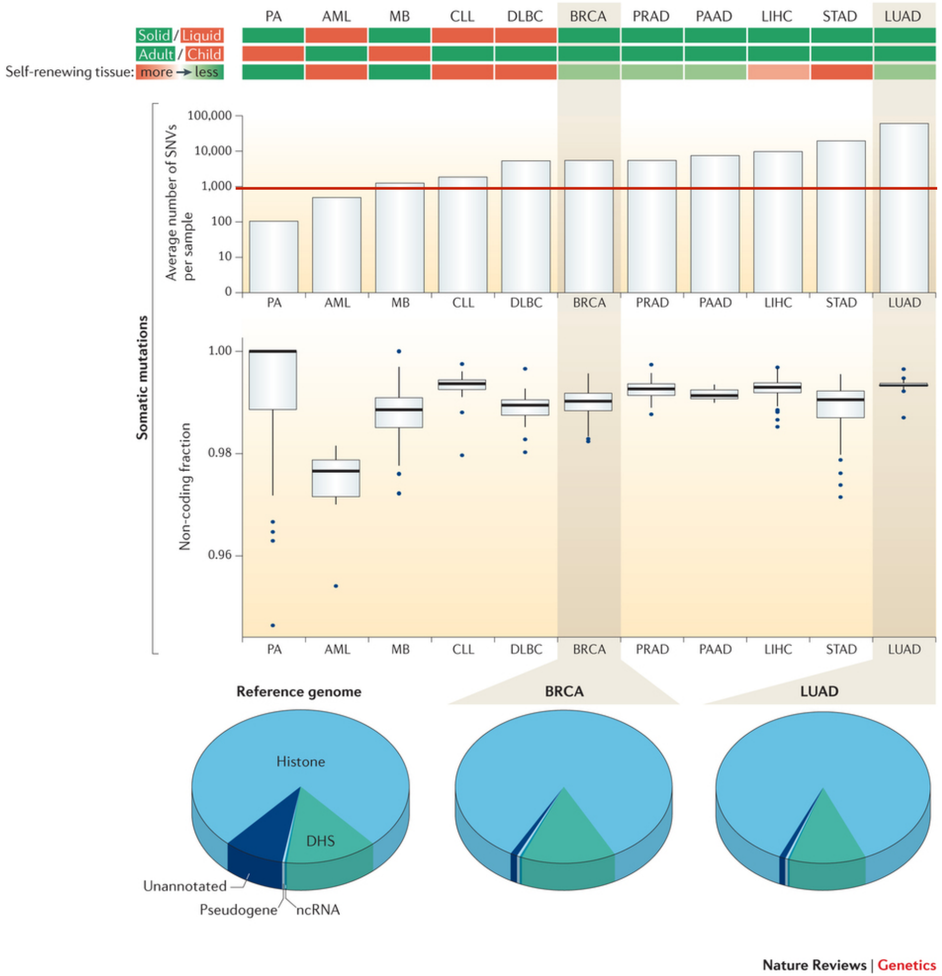
9984/3: refractory anemia with excess blasts in transformation [raebt] (24)
9983/3: refractory anemia with excess blasts [raeb] (38)
9867/3: acute myelomonocytic leukemia [fab type m4] (32)
9920/3: therapy-related acute myeloid leukemia, nos (32)
9891/3: acute monoblastic leukemia [fab m5] (23)

9051/3: desmoplastic mesothelioma (59)
9053/3: mesothelioma, biphasic, malignant (27)
9050/3: mesothelioma, nos (81)
9052/3: epithelioid mesothelioma, malignant (64)



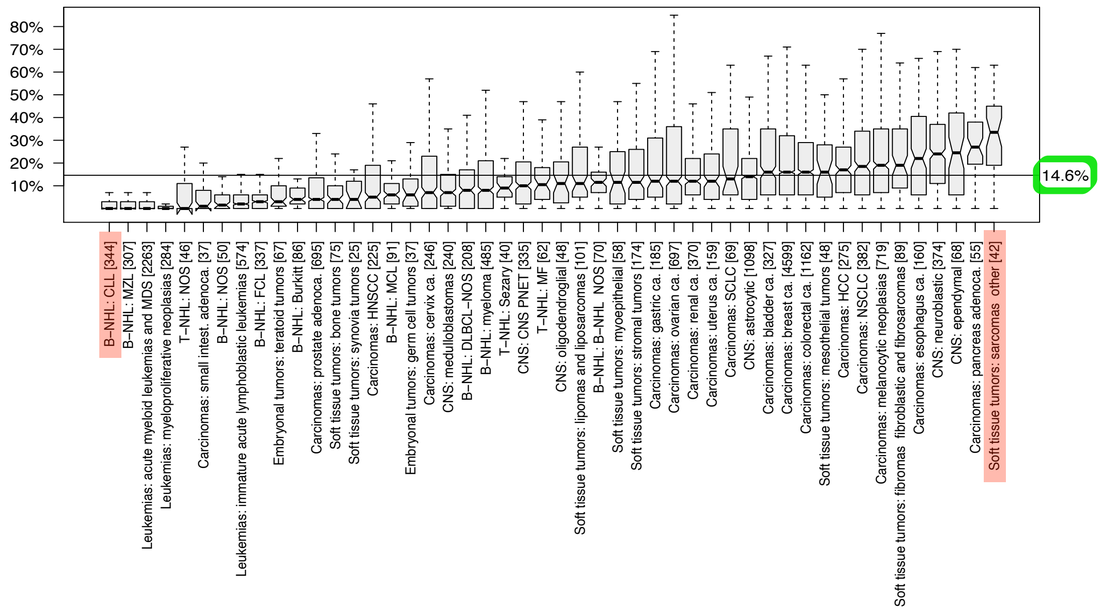
arrayMap

Quantifying Somatic Mutations In Cancer



CANCERS SHOW THOUSANDS OF SINGLE NUCLEOTIDE VARIANTS PER SAMPLE, MOSTLY IN NON-CODING REGIONS

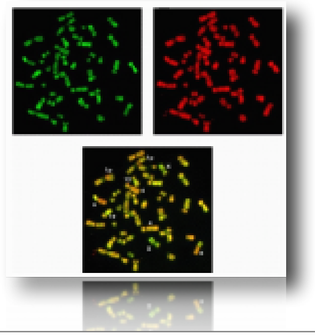
Pan-Cancer Analysis of Whole Genomes (PCAWG) data show widespread mutations in non-coding regions of cancer genomes (Khurana et al., Nat. Rev. Genet. (2016)

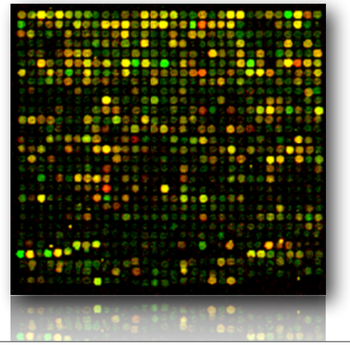


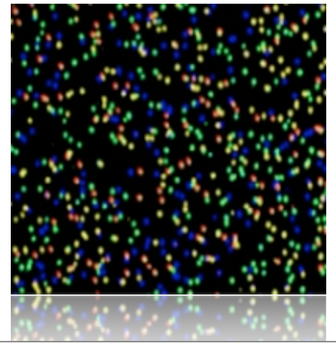
GENOMIC COPY NUMBER IMBALANCES PROVIDE WIDESPREAD SOMATIC VARIANTS IN CANCER

On average ~15% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on >30’000 cancer genomes from arraymap.org

WHOLE GENOME SCREENING IN CANCER







	chromosomal CGH	genomic arrays	"NGS" genome sequencing (WES, WGS)
1st application report	**1992**	**1997**	**2010**
source	DNA (paraffin, micro-dissected ...)	DNA (paraffin, micro-dissected ...)	DNA (paraffin, micro-dissected ...)
main source problems	mixed/degraded source tissue	mixed/degraded source tissue	mixed/degraded source tissue
resolution	chromosomal bands = few megabases	mostly in the 100kb range, but tiling possible	single bases
target identification	surrogate (position)	"semidirect" (segmentation spanning probes)	direct quantitative and qualitative
structural	no	depending on type	yes
available data	>24,000 cases (57%) through **Progenetix**	raw data repositories (GEO, EMBL, SMD), **Progenetix**	Limited for rax data (BAMs ...); variant call data in dbgap, clinvar; selected studies with called CNV segments
predominant data format	ISCN = static	raw => depends on bioinformatics	mostly annotated variant calls or SNVs

dilbert.com | Tuesday February 27, 1996

Let's build a database!

... using archaic tools

dilbert.com | Tuesday September 08, 1992

Progenetix Database in 2003
Text conversion for CNVs

- articles and supplements with **cytoband-based** *rev ish* CGH results
- sometimes rich, but **unstructured** associated information
- **PDFs** readable, but **not well suited for data** extraction (character entities, text flow)

progenetix

TABLE 3. Comparison of Primary Tumors and Metastases by CGH

Case	Gain in common	Loss in common	Primary tumor only	Metastasis only
108		18		
113	7, 8q24-qter, 13q11-qter, 20q11-qter, Xq11-Xter	1p33-pter, 2p21-pter, 4q24-qter, 15q11-q15, 17p11-pter, 18		
LM	12q22-qter, 15q23-qter, 17q11-ter, 20p11-p12, 20q11-ter, 22q11-ter	1p11-p32, 1q24-31, 4, 13q11-qter, 17p11-pter, 18, 20p11-ter	11p11-pter-	12+
145	4q26-q28, 6p11-p13, 8p11-p12, 920q11-qter	1p11-pter, 4q31-qter, 6q11-qter, 8p12-pter, 11, 15q11-qter, 16q11-qter, 17p11-pter, 18, 21q11-qter	13q21-qter+, 20p11-pter-	8q11-qter+,10-, 6p21-pter-
53	7, 8q11-qter, 9q33-qter, 13q11-qter, 20p11-p12, 20q11-qter	4p13-pter, 4q21-qter, 8p12-pter, 15q14-qter, 18q11-qter, 20p12-pter	5p11-pter-, 5q13-qter-, 14q11-qter-	11+, 16p11-pter+, 17q11-qter+, 19+, 21q11-qter+, 22q11-qter+
147	7, 13q11-qter, 20q11-qter	8p21-pter, 18	4p14-pter-, 4q28-qter+, 8p11-21-, 17q11-q2+, 21q11-qter-	11q22-qter+, 16+, 1p11-33-

TABLE 1. Clinical Data

Case number	Age	Sex	Site	Stage[a]	Grade[b]	Diagnosis of metastatic disease[c]
2	40	M	Transverse colon	IV	3	Synchronous
6	79	M	Ascending colon	IV	2	Synchronous
9	73	M	Transverse colon	II	2	N/A
11	56	M	Rectosigmoid	IV	2	Metachronous
12	70	F	Sigmoid colon	IV	2	Synchronous
13	65	M	Descending colon	II	9	Synchronous
14	60	M	Rectum	III	3	Metachronous
15	51	F	Rectum	III	2	Metachronous
19	63	M	Rectosigmoid Junction	III	2	Synchronous
20	63	M	Rectum	IV	9	Metachronous
21	64	F	Sigmoid colon	IV	2	Synchronous
35	71	M	Rectum	III	9	Metachronous
49	72	M	Cecum	IV	3	Synchronous
53	72	F	Sigmoid colon	IV	2	Synchronous
104	61	M	Sigmoid colon	IV	2	Metachronous
105	58	M	Ascending colon	II	2	Metachronous
107	77	F	Cecum	IV	2	Metachronous
108	53	F	Splenic flexure	IV	2	Synchronous
112	68	M	Rectum	III	3	Synchronous
113	41	M	Splenic flexure	IV	2	Synchronous
114	49	M	Splenic flexure	IV	3	Synchronous
116	73	M	Rectosigmoid	III	9	Metachronous
120	24	F	Descending colon	IV	2	Synchronous
123	62	F	Rectum	III	2	Metachronous
124	42	M	Rectum	IV	9	Synchronous
145	70	M	Rectosigmoid	IV	2	Synchronous
147	86	F	Cecum	IV	2	Synchronous

[a]AJCC/UICC staging system (Hutter and Sobin, 1986).
[b]Grade of primary tumor: 1–3, low, moderate, high grade; 9, grading unknown.
[c]Synchronous, diagnosis of metastatic disease within 12 months following diagnosis of primary tumor; metachronous, diagnosis of metastatic disease after 12 months or later.

GENES, CHROMOSOMES & CANCER 25:82–90 (1999)

Chromosome Arm 20q Gains and Other Genomic Alterations in Colorectal Cancer Metastatic to Liver, as Analyzed by Comparative Genomic Hybridization and Fluorescence In Situ Hybridization

W. Michael Korn,[1]* Toru Yasutake,[2] Wen-Lin Kuo,[1] Robert S. Warren,[2] Colin Collins,[1] Masao Tomita,[2] Joe Gray,[1] and Frederic M. Waldman[1]

Progenetix Database in 2003
Text conversion for CNVs

- based on listed CGH results from publications
  - literature detection using optimized PubMed queries
  - extraction (copy/paste, typing) of rev ish ISCN karyotypes from articles and supplementary material
  - annotation cleanup using scripting with regular expressions (Perl)
  - custom script to convert cleaned ISCN annotations to cytoband status maps
  - custom graphics libraries to create graphical representations of CNV frequencies

progenetix

[ideogram] [casetable] [clustering] [download source]

About [progenetix]
Contents, Aims and FAQs
Publications
ICD-O Entities
Site Codes and Misc. Groups
ISCN2matrix Converter
Data Source Access
Sponsors and Contributors
News and History
Links

PLOS

List of cases included in the subset "Hepatocellular carcinoma, NOS"

Casename	Original diagnosis	PUBMED ID	Aberrations (by CGH)
HCC-vir-dys-ca-01sat	Hepatocellular carcinoma (HBV, satellite tumor)	12666986	rev ish enh(1q21qter, 7p11.2pter, 7q11.2q31, 8q13qter, 9p22pter, 10, 11p11.2p12, 11q12qter, 15q26) dim(1p22pter, 2q32qter, 4, 5, 7q32qter, 8p12pter, 14q21qter, 15q11.2q21, 16, 17p11.2pter, 17q11.2q21, 18, 19)
HCC-vir-dys-ca-01tu	Hepatocellular carcinoma (HBV)	12666986	rev ish enh(1q21qter, 5p12pter, 8q12qter, 9p21pter, 11q12qter, 20) dim(1p31pter, 4, 7q32qter, 8p12pter, 14q21qter, 16, 17p12pter, 18, X)
HCC-vir-dys-ca-02tu	Hepatocellular carcinoma (HCV)	12666986	rev ish enh(1q21q43, 6q12q14, 7, 8p11.2, 8p21p23, 8q11.2q13, 8q23, 10p11.2p13, 10q11.2qter, 17q11.2q24, Xq13qter) dim(11, 14q31, 15q11.2q21, 16p12pter, 17p11.2pter, 19p13.1pter, 19q13.1q13.2, Xp21)
HCC-MF-01T1	Hepatocellular carcinoma	12579536	rev ish enh(16q13qter)
HCC-MF-01T2	Hepatocellular carcinoma	12579536	rev ish enh(12q22qter, 17q) dim(16q)
HCC-MF-01T3	Hepatocellular carcinoma	12579536	rev ish enh(12q21.3qter, 17q21qter) dim(16q21qter)
HCC-MF-02T1	Hepatocellular carcinoma	12579536	rev ish dim(6q13qter)
HCC-MF-02T2	Hepatocellular carcinoma	12579536	rev ish enh(1q, 17q) dim(17p)
HCC-MF-03T1	Hepatocellular carcinoma	12579536	rev ish enh(1q, 3q26.2qter, 4p, 6p21.1pter, 11p15, 19q) dim(16q10q12.2)
HCC-MF-03T2	Hepatocellular carcinoma	12579536	rev ish enh(8q, 11p15, 12pterq12) dim(3p, 4q, 5q, 8p23.1, 9q, 16q) amp(1q)
HCC-MF-04T1	Hepatocellular carcinoma	12579536	rev ish enh(1p33qter, 8q21.2qter) dim(1pterp34, 4q, 9q) amp(6p, 13q21qter)
HCC-MF-04T2	Hepatocellular carcinoma	12579536	rev ish enh(1q, 5q31.3qter, 8q) dim(6q, 16, 17pterq21)
HCC-MF-05T1	Hepatocellular carcinoma	12579536	rev ish enh(6q, 8q, 10p, 12q21.1qter, 13q22qter, 17q, 18p) dim(4p15qter, 5, 7p21qter, 7q, 9p, 9q10q34.2, 11q, 16q) amp(10p)
HCC-MF-05T2	Hepatocellular carcinoma	12579536	rev ish enh(6q, 8q12qter, 12q21.1qter, 13q22qter, 17q) dim(4q, 5q, 7p, 7q, 9q10q31, 11q, 14q, 16q) amp(10p)
HCC-MF-06T1	Hepatocellular carcinoma	12579536	rev ish enh(1q, 5p23pter, 18p, 22) dim(4q, 6q, 9pterq33, 13q, 14q, 16pterq23) amp(8q)

Progenetix in 2023

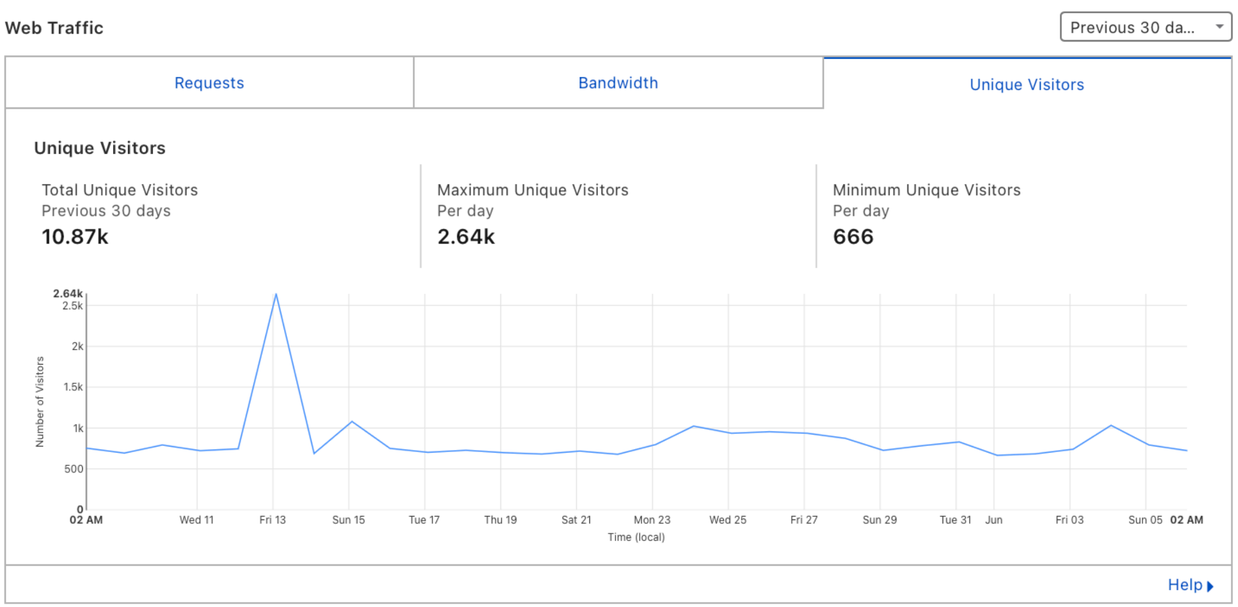
An oncogenomic reference resource

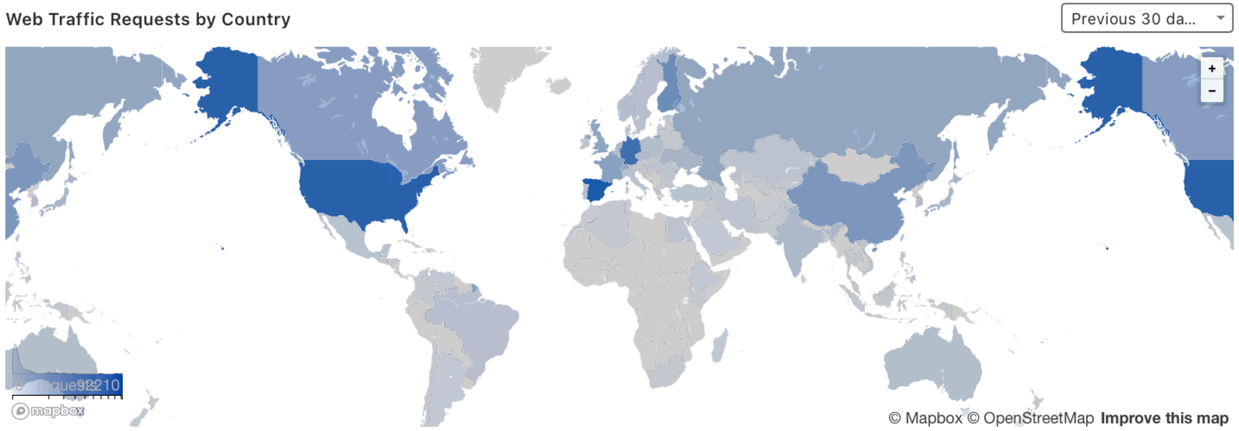
progenetix

Progenetix in 2023

Cancer Genomics Reference Resource

- largest open resource for curated cancer genome profiling data, with focus on copy number variations (CNV)
- >116'000 cancer CNV profiles, mapped to >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 where appropriate
- core biosample and technical metadata annotations where accessible (TNM, genotypic sex, survival ...)
- publication database and code mapping services





Top Traffic Countries / Regions
Previous 30 days

Country / Region	Traffic
Spain	[illegible]
United States	[illegible]
Germany	59,321
Finland	28,826
Singapore	21,311

Help

Cloudflare | progenetix.org | June 2022

progenetix

Progenetix in 2023

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

progenetix

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+

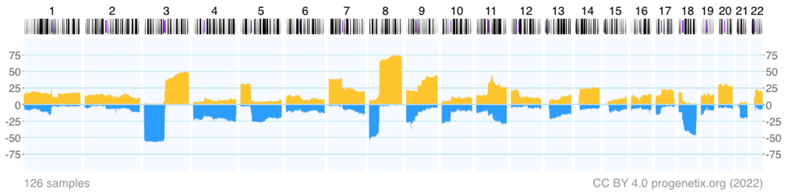
Documentation
- News
- Downloads & Use Cases
- Sevices & API

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.

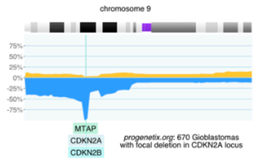
Floor of the Mouth Neoplasm (NCIT:C4401)



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 2023

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

progenetix

Cancer CNV Profiles

Search Samples

Studies & Cohorts

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

Publication DB

- Genome Profiling
- Progenetix Use

Services

- NCIt Mappings
- UBERON Mappings

Upload & Plot

Download Data

Beacon+

Progenetix Info

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

Search Samples | Modify Query

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

progenetix

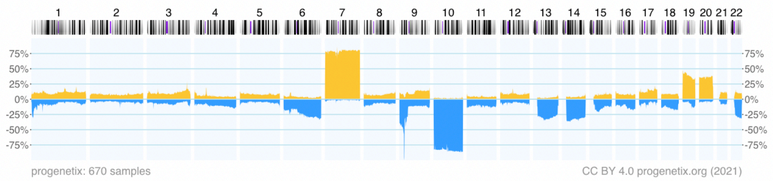
Samples: 668
Variants: 286
Calls: 675

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

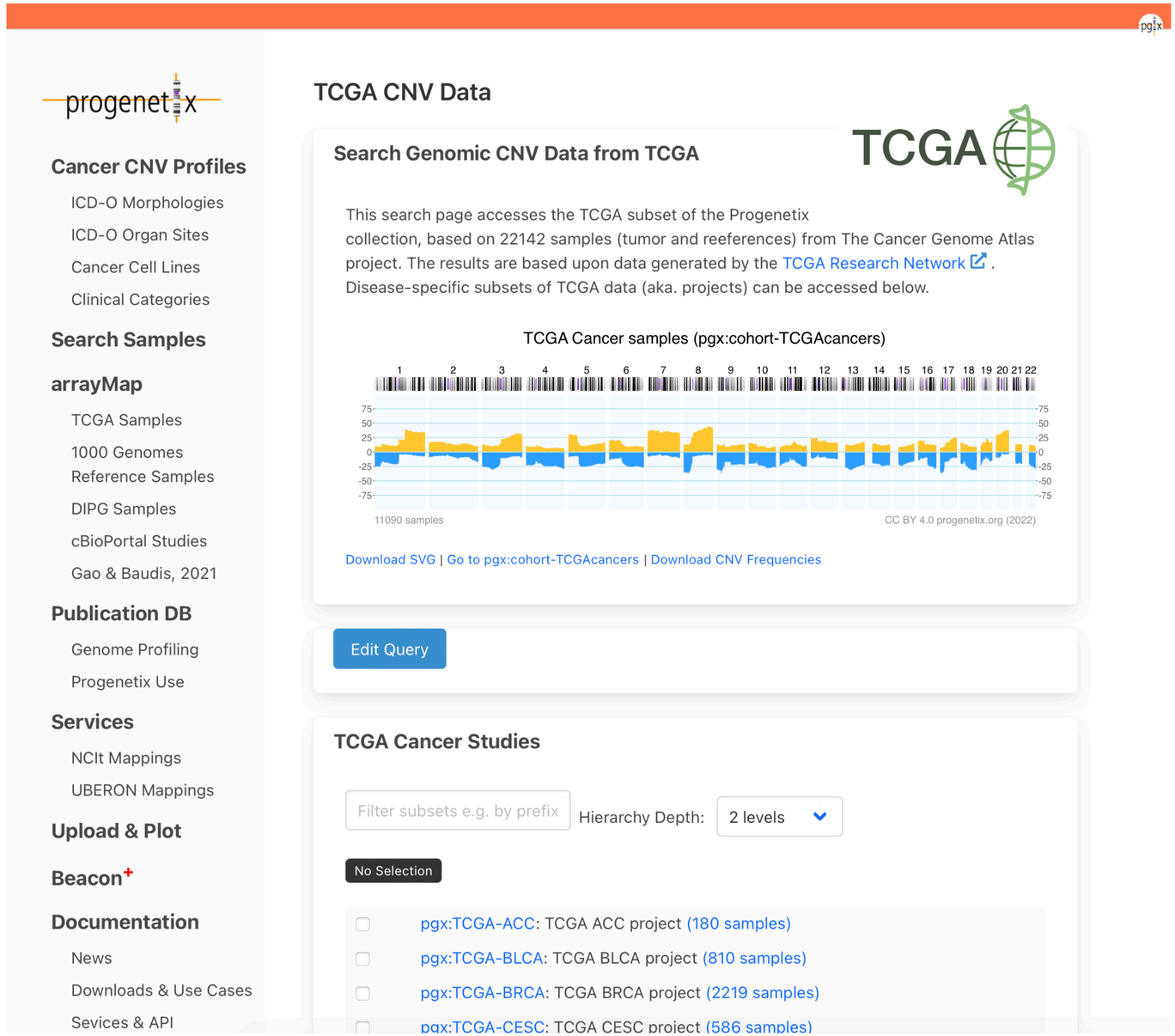


Matched Subset Codes	Subset Samples	Matched Samples	Subset Match Frequencies
UBERON:0002021	4	1	0.250
icdot-C71.4	4	1	0.250
icdom-94403	4291	664	0.155
NCIT:C3058	4375	664	0.152
UBERON:0016525	14	2	0.143
icdot-C71.1	14	2	0.143
UBERON:0000955	7068	651	0.092
icdot-C71.9	7066	651	0.092
icdom-94423	84	4	0.048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

Progenetix in 2023

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
  - ...

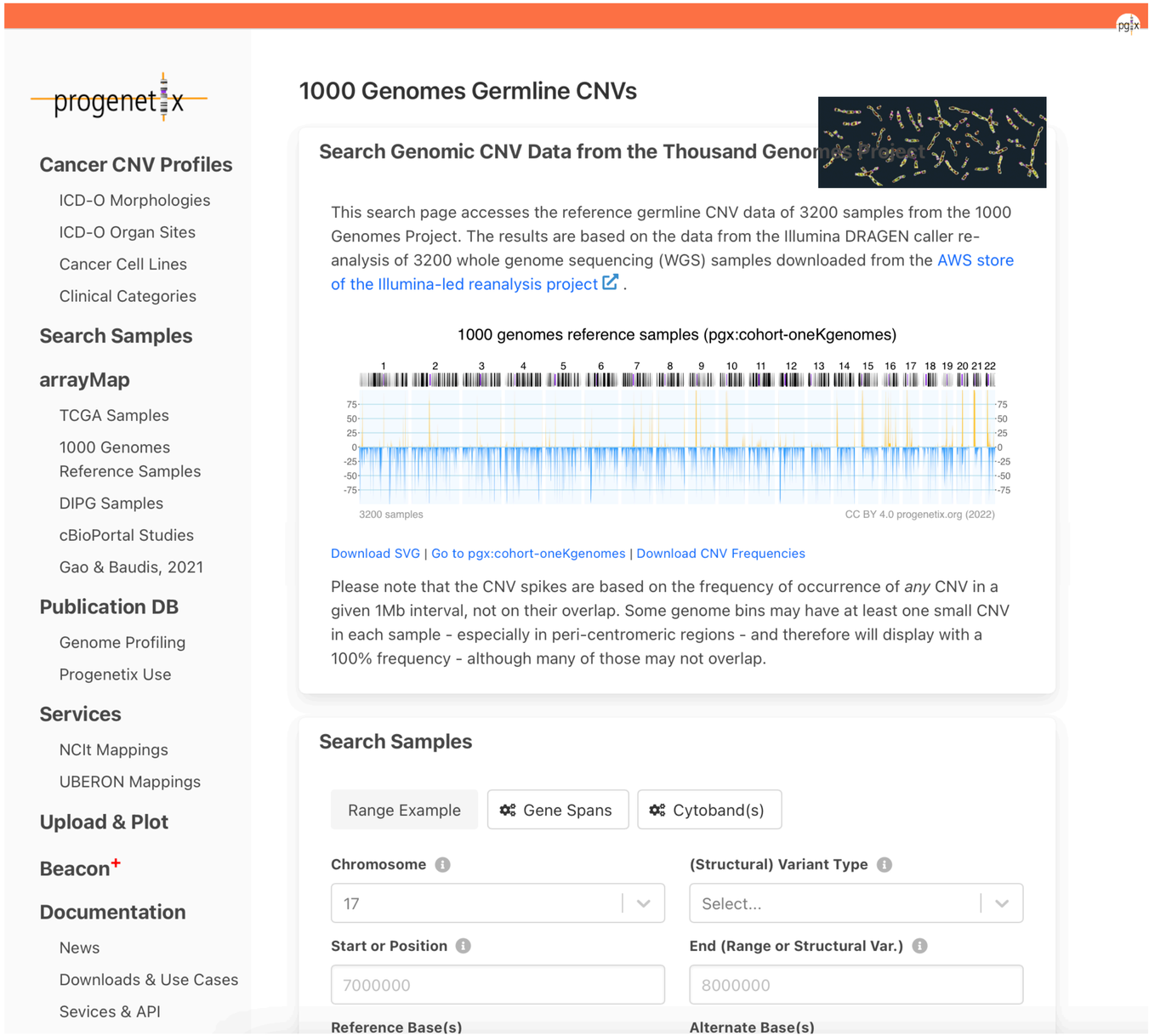


progenetix

Progenetix in 2023

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
  - ...



progenetix

Database, 2021, **2021(0)**, 1–9
DOI: https://doi.org/10.1093/database/baab043
Database update

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,*]

[1]Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland
[2]Swiss 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

[a]set of variants from one genotyping experiment; ICD-O, International Classification of Diseases for Oncology; NCIt, National Cancer Institute Thesaurus.

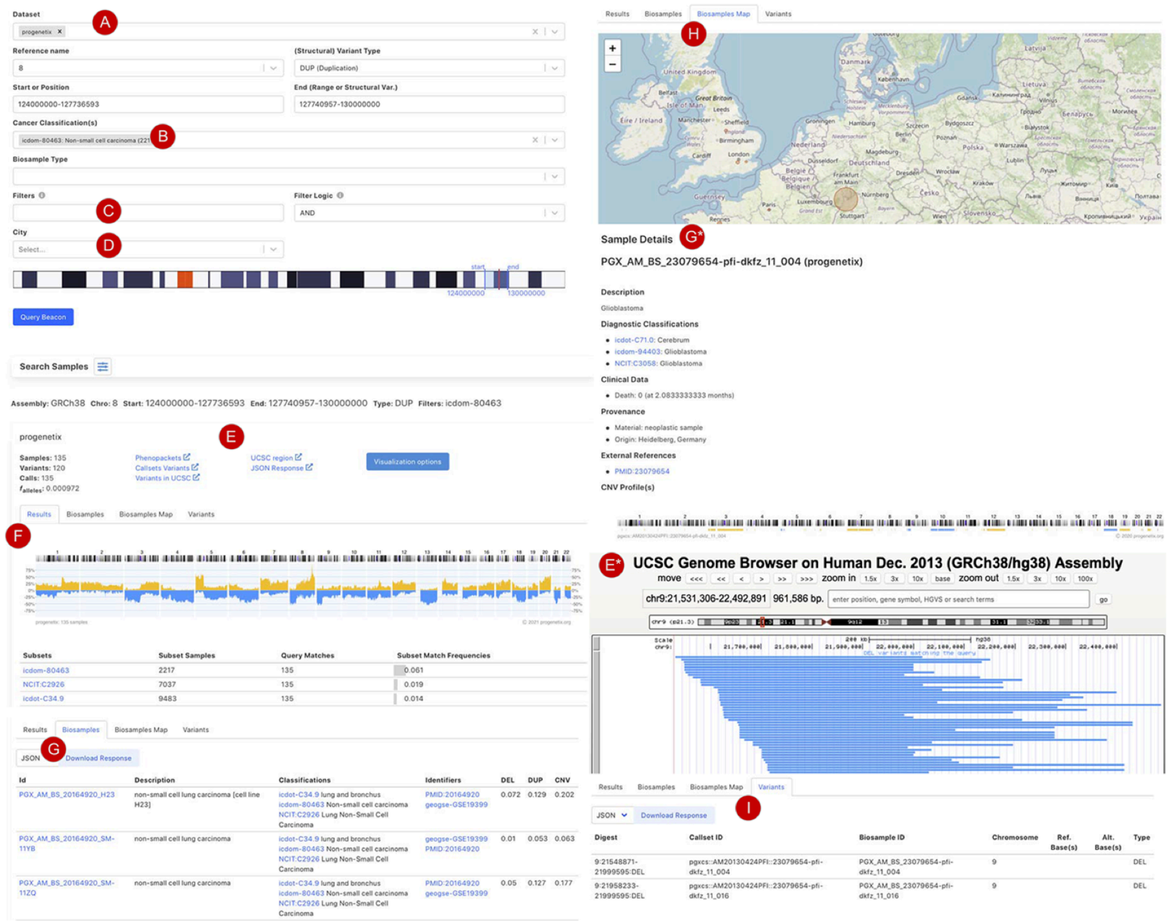


Figure 3. Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range This example queries for a continuous, focal duplication covering the complete MYC gene's coding region with < = 6 Mb in size. A: Filter for dataset; B: filter for cancer classification (NCIt and 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 biosamples with description, statistics and reference. More detailed biosample information can be viewed through 'id' link to the sample detail page; H: matched variants with reference to biosamples can be downloaded in json or csv format.

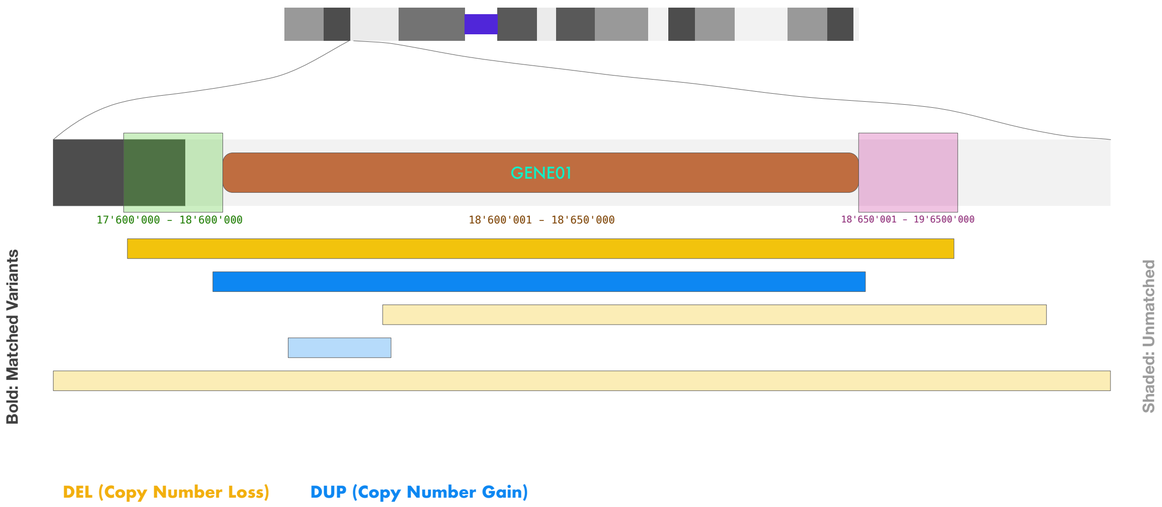
ELIXIR ALL HANDS 2022

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"**

Genome Bracket Query
(full match)



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. <= ~1Mbp in size). The query can be modified e.g. through changing the position parameters or diagnosis.

Gene Symbol: Select...

Chromosome: 9

(Structural) Variant Type: DEL (Deletion)

Start or Position: 21500001-21975098

End (Range or Structural Var.): 21967753-22500000

Minimum Variant Length:

Maximal Variant Length:

Reference ID(s): Select...

Cancer Classification(s): NCIT:C3058: Glioblastoma (4375) ×

Clinical Classes: Select...

Genotypic Sex: Select...

Biosample Type: Select...

Filters:

Filter Logic: AND

Filter Precision: exact

City: Select...

Chromosome 9

Query Database

ELIXIR ALL HANDS 2022

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

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, geneId, lengths, EFO & VCF CNV types, pagination
  - ➡ widespread, self-scoping filter use for bio-, technical- and 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
- ┐(ᒧ ▽ ᒨ)┌ 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

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

React

APACHE HTTP SERVER PROJECT

python

mongoDB

- *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"),
```


variants | analyses | biosamples | individuals

Entity collections

collations | geolocs | genespans | publications | qBuffer

Utility collections

Documentation Home | Search | progenetix-web ☆3 ⑂1

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 Beac

`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 sevices (e.g. the publicat
- implemented as React / Next.js project
- contains this documentation tree here as `mkdocs` project, with files in the `docs` directory

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.

Org.progenetix | Search | beacon-v2 ☆2 ⑂8

Base `/biosamples`

`/BIOSAMPLES/` + QUERY

- /biosamples?filters=cellosaurus:CVCL_0004
- this example retrieves all biosamples having an annotation for the Cellosaurus *CVCL_0004* identifier (K562)

es/pgxbs-kftva5c9

a single biosample

MODE=TRUE

es?testMode=true

some random samples

- 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

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

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

Beacon v2 Documentation | Search | beacon-v2 ☆2 ⑂8

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 ccan 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": {
```

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=pgxbs-kftvh94d,pgxbs-kftvh94g,pgxbs-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
```

Author: Hangjia Zhao | @hangjiaz

README.md

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

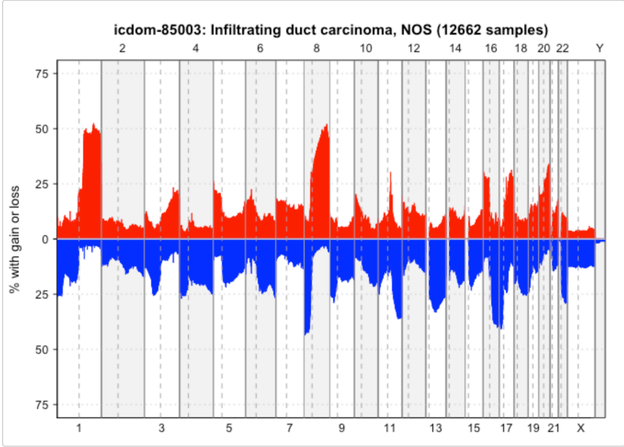
When you face problems, try to reinstall the latest version. If reinstallation doesn't help, please contact us.

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

```
biosamples <- pgxLoader(type="biosample", filters = "NCIT:C3059",codematches = TRUE,
                        biosample_id = c("pgxbs-kftva5zv","pgxbs-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": "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.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",
```

```
"biosamples": [
  {
    "biosampleStatus": {
      "id": "EFO:0009656",
      "label": "neoplastic sample"
    },
    "dataUseConditions": {
      "id": "DUO:0000004",
      "label": "no restriction"
    },
    "description": "Primary Tumor",
    "externalReferences": [
```

```
"files": [
  {
    "fileAttributes": {
      "fileFormat": "pgxseg",
      "genomeAssembly": "GRCh38"
    },
    "uri": "https://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/variants/?output=pgxseg"
  }
],
```

```
      },
      "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"
  }
}
```

```
    {
      "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"
  },
```

ELIXIR ALL HANDS 2022

progenetix

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

```
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?

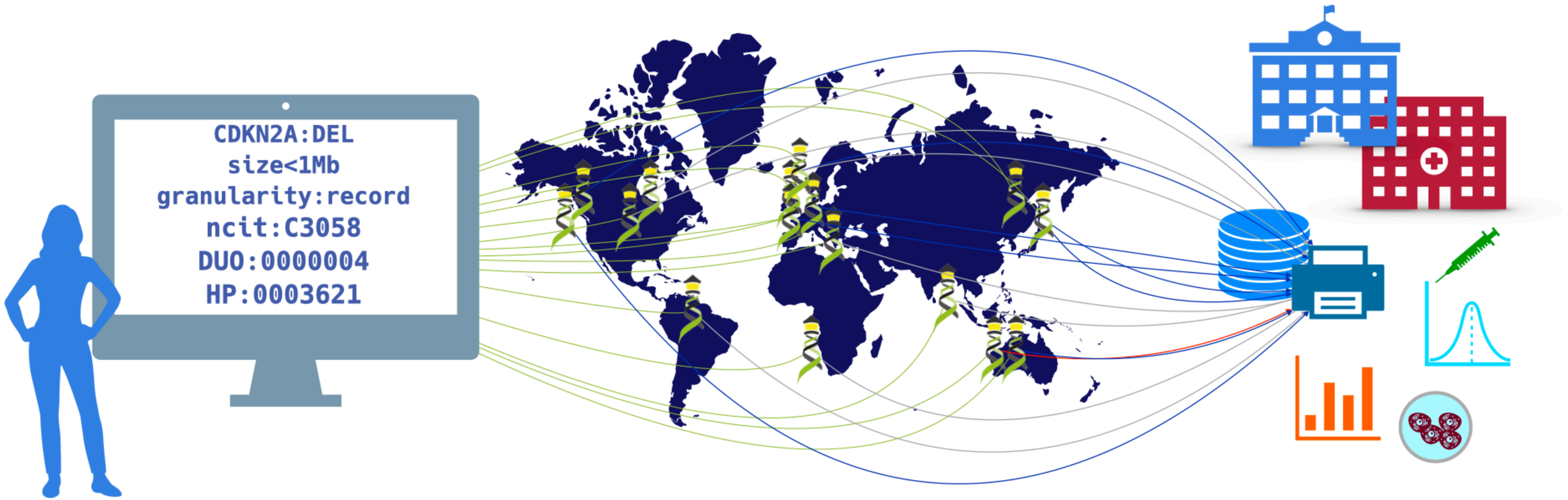
Some proposals for a stepwise Beacon protocol extension

- 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

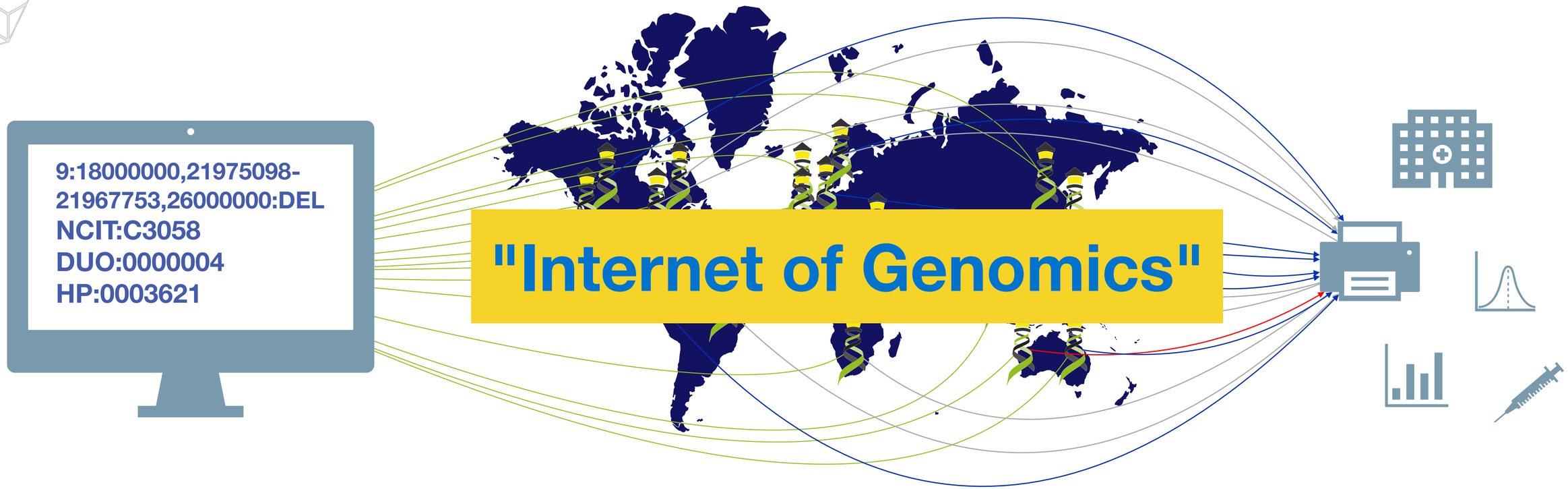

FUTURE



Can you provide data about focal deletions in CDKN2A in Glioblastomas from juvenile patients with unrestricted access?

Beacon *v2* API

The Beacon API v2 represents a simple but powerful **genomics API** for ***federated*** data discovery and retrieval



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**".

Beacon and the “Internet of Genomics”

Jordi Rambla
European Genome-phenome Archive (EGA)
Centre for Genomic Regulation (CRG)
GA4GH 11th Plenary – September 2023

The EGA(1)

Long term secure archive for human biomedical research sensitive data, with focus on reuse of the data for further research (or "*broad and responsible use of genomic data*")

EUROPEAN GENOME-PHENOME ARCHIVE

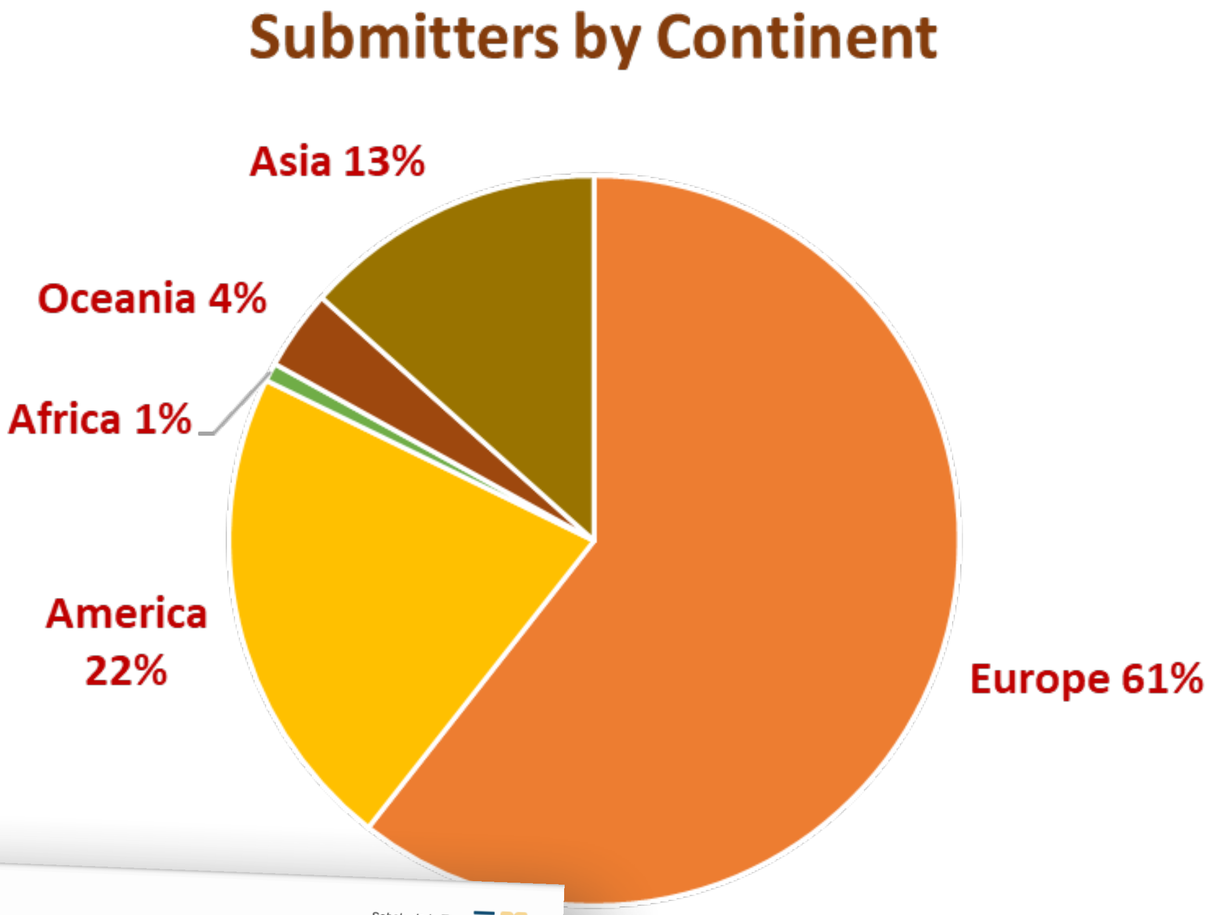
CRG Centre for Genomic Regulation

EMBL-EBI

Global Alliance for Genomics & Health
DRIVER PROJECT

The EGA(2)

We get data and metadata from data producers worldwide



EGA is in the *business* of data sharing

EGA is a GA4GH driver project

GA4GH Standards supporting the EGA:

Classical File formats

Crypt4GH

Phenopackets

Data Use Ontology

AutN/AutZ Infrastructure

GA4GH Passports

(mystery product 1)

(mystery product 2)

EUROPEAN GENOME-PHENOME ARCHIVE

Login | Register | Need Help?

ABOUT | DISCOVERY | SUBMISSION | ACCESS | Search...

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

Standards

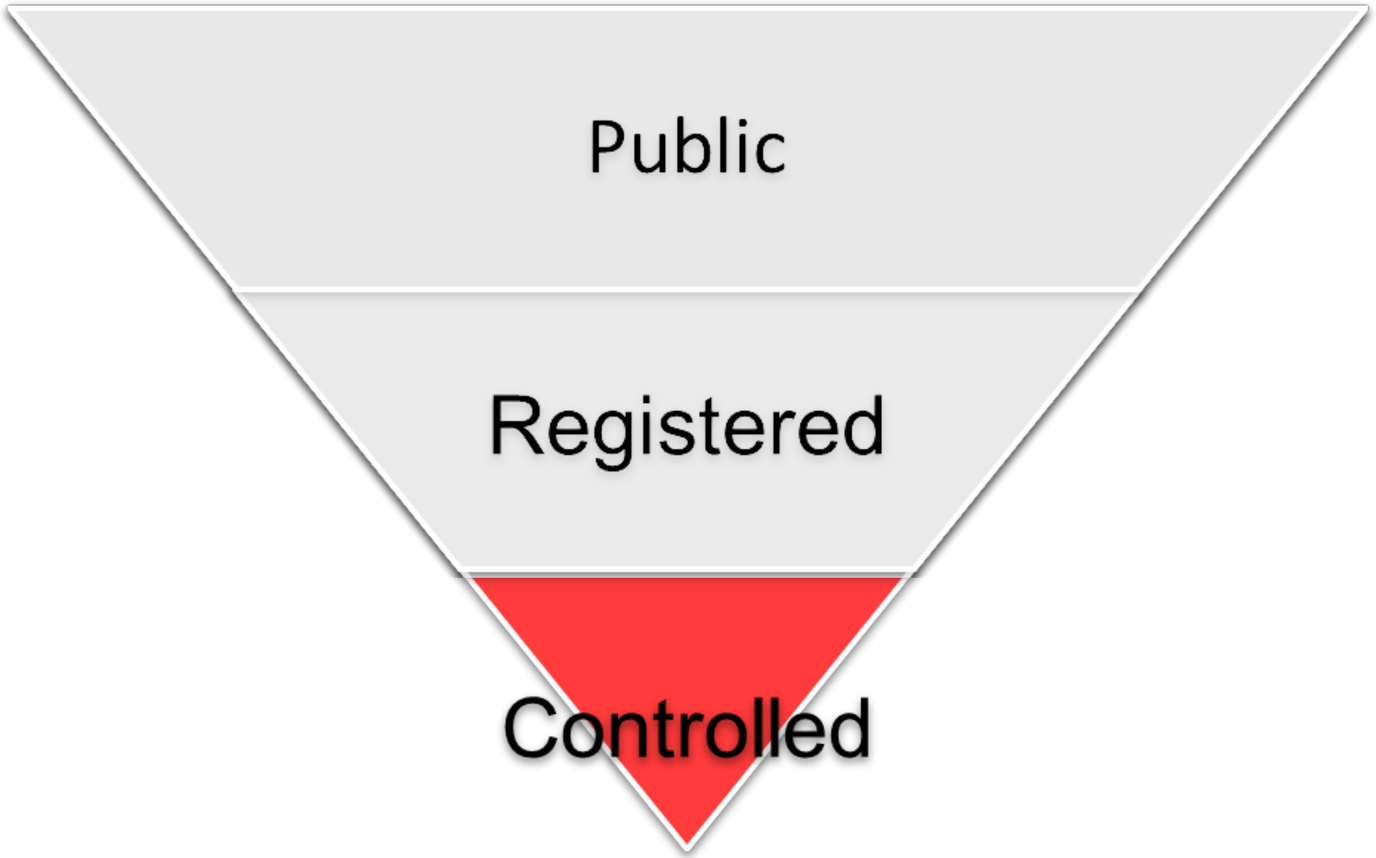
The EGA is a long-standing supporter of the Global Alliance for Genomics & Health (GA4GH) to enhance responsible sharing of human genetic data through the development of interoperable global standards for human data access. The EGA is one of the founding GA4GH Driver Projects and has contributed to the development and implementation of several GA4GH standards and APIs.

Below is a list of the GA4GH standards and APIs that are currently available or planned for implementation at EGA.

Technical Standards	Purpose	Specification Version	Supported Version	Implementation
Large Scale Genomics				
htsget	A protocol for secure, efficient, and reliable access to sequencing read and variation data.	V1.3.0	V1.0.0	Specification Documentation Endpoint
Read File Formats (SAM/BAM/CRAM)	Specifications for storing next-generation sequencing read data.	V3.0.0	V3.0.0	Implementation Example of Usage
Variation File Formats (VCF/BCF)	The specifications for Variant Call Format Files (VCF) and its binary counterpart BCF.	V4.0.0 V2.0.0	V4.0.0 V2.0.0	Implementation Example of Usage
Crypt4GH v1.0	Enables direct byte-level compatible random access to encrypted genetic data stored in community standards (e.g. CRAM, VCF)	V1.0	V1.0	Specification Documentation Endpoint
refget API	Enables access to reference sequences using an identifier derived from the sequence itself.	V1.2.6	NA	Specification
RNAget API v1	Provides a means of retrieving data from several types of RNA experiments including (i) feature-level expression data from RNA-seq type measurements and (ii) coordinate-based signal/intensity data similar to a bigwig representation via a client/server model.	V1.0.0	NA	Documentation
Discovery				
Beacon v2	Supports discovery of genomic variants, phenotypes, and individuals	V1.0.1	V0.3	Web UI API Source Code

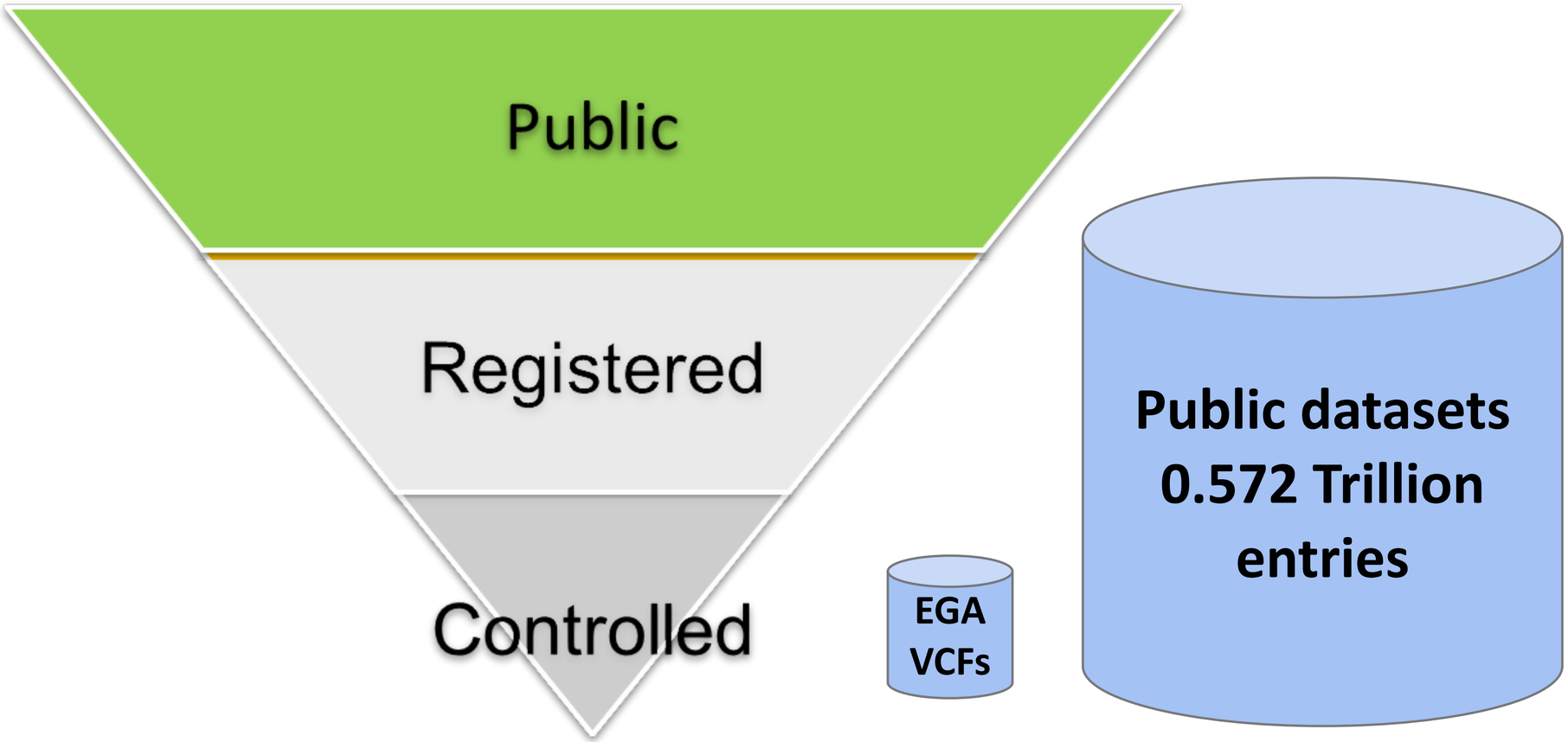
The Service Info API is an endpoint for

The EGA Beacon v2 for controlled access data



150,000 Billion entries

…but also includes non-EGA publicly available datasets



Public sources being included in EGA Beacon

- ClinVar
- gnomAD
- dbSNP
- NCBI ALFA
- TCGA (open)
- Simons Genome Diversity Project
- Brain Genomics
- Encode (open)
- Exome Sequencing Project
- HapMap
- 1000 Genomes
- ExAC
- Platinum Genomes
- GiaB
- dbVar

How is this related to the “Internet of Genomics”?

In one side, we manage to make more data easily discoverable in a centralized way

But we have a federated, decentralized approach too

It starts with a local Beacon network…

La Marató 3

2019 Malalties minoritàries

"Xarxa interhospitalària catalana de variants genètiques"

Hospital Sant Joan de Déu (centro coordinador)
IP: Dèlia Yubero

Hospital Clínic de Barcelona
IP: Eva González

Hospital Vall d'Hebron
IP: Elena García Arumí

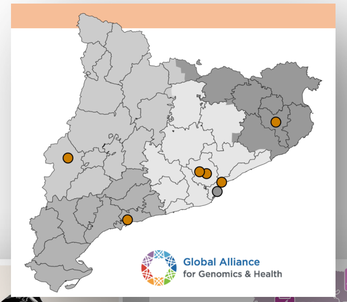
Hospital de la Santa Creu i Sant Pau
IP: Benjamín Rodríguez

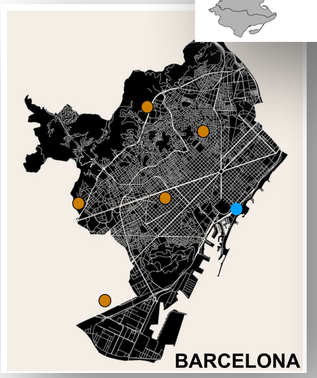
Hospital de Bellvitge
IP: Ariadna Padró

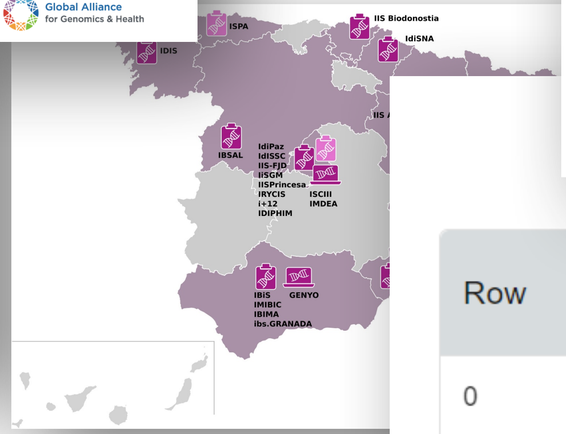
Centre de Regulació Genòmica
IP: Babita Singh

Which is also part of a continental network

eosc | cancer







BEACON V2 CANCER REGISTRY

HELP FOR QUERYING

EUROPEAN GENOME-PHENO ARCHIVE

TumourTopography=C26,Sex=1

NEW SEARCH

QUERY EXAMPLES

FILTERING TERMS

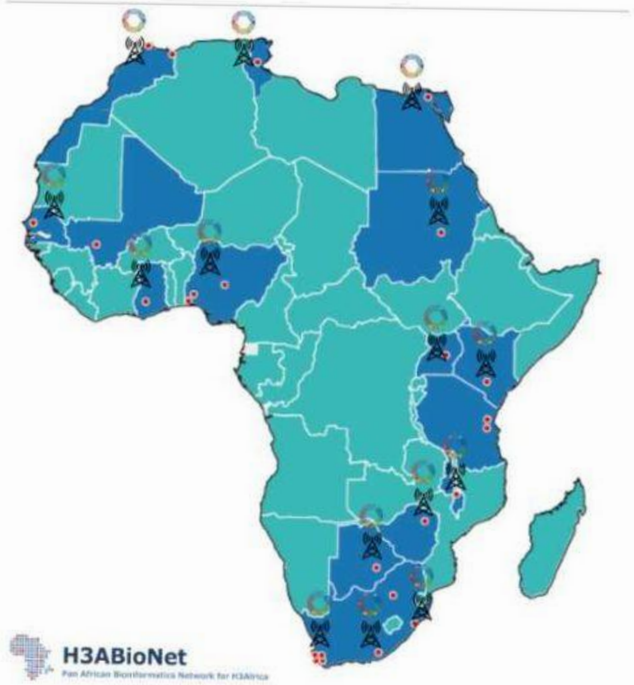
Granularity: BOOLEAN COUNT FULL RESPONSE

68 RESULTS

Row	PatientID	AgeOfOnset	Sex	TumourId…	Geograph…	TumourTo…	TumourM…	TumourB.
0	1620	95	1	1	38	C26	9639	0
1	2049	76	1	2	45	C26	9646	1

…but not just Europe

Pan-african Beacon Network



NEWS | 26 February 2023

Plan for network of Genomics Centres of Excellence across Africa

Closing the gap on access to genomics technologies.

14th International Congress of Human Genetics
22 - 26 February 2023
COMING HOME | ICHG 2023 | AFRICA
Cape Town
www.ichg2023.com

EUROPEAN GENOME-PHENOME ARCHIVE

H3Africa
Human Heredity & Health in Africa

20th Meeting of the H3Africa Consortium
27th - 28th February 2023

Plan for network of Genomics Centres of Excellence across Africa (nature.com)

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**

elixir

SIB

University of Zurich[UZH]

SPHN Swiss Personalized Health Network

Global Alliance for Genomics & Health