Cancer Data

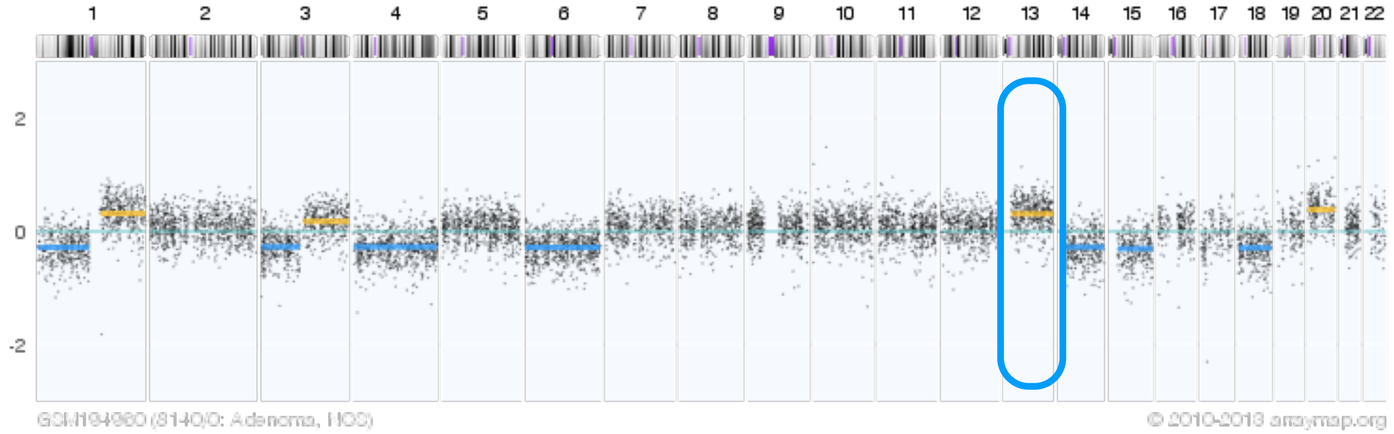
ELIXIR::GA4GH: Advancing genomics resources through standards and ontologies

Michael Baudis | ECCB vBarcelona | 2020-09-02

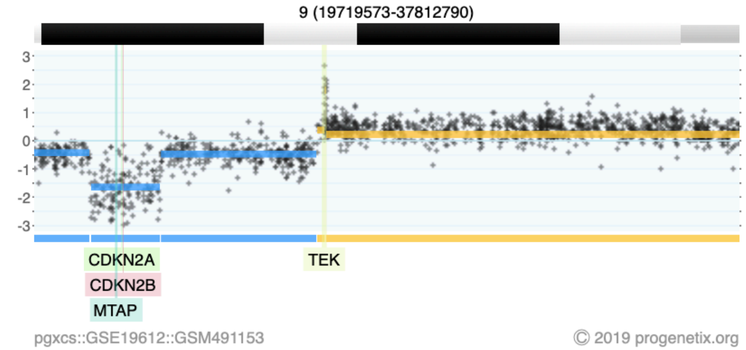
elixir

Global Alliance
for Genomics & Health

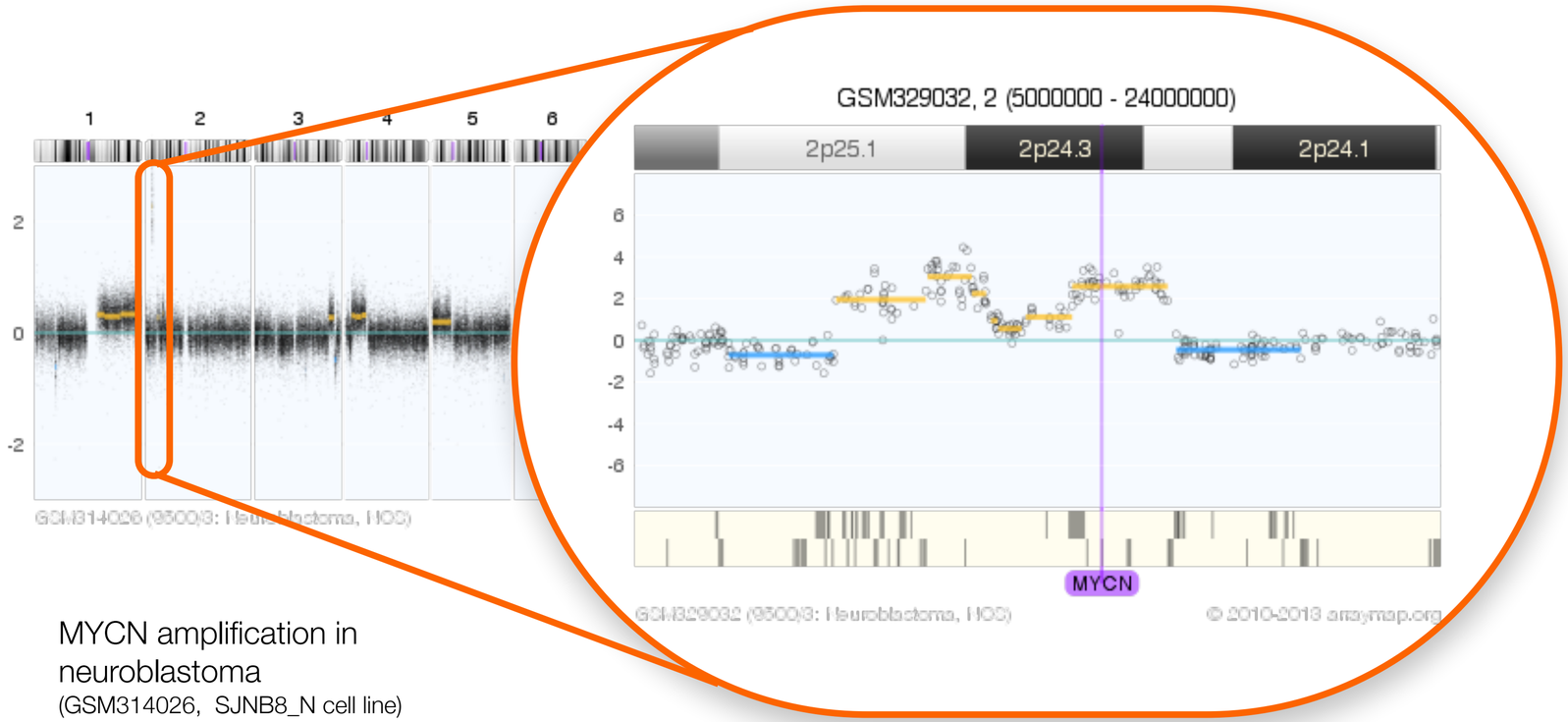
Somatic Copy Number Variations



Gain of chromosome arm 13q in colorectal carcinoma



2-event, homozygous deletion in a Glioblastoma



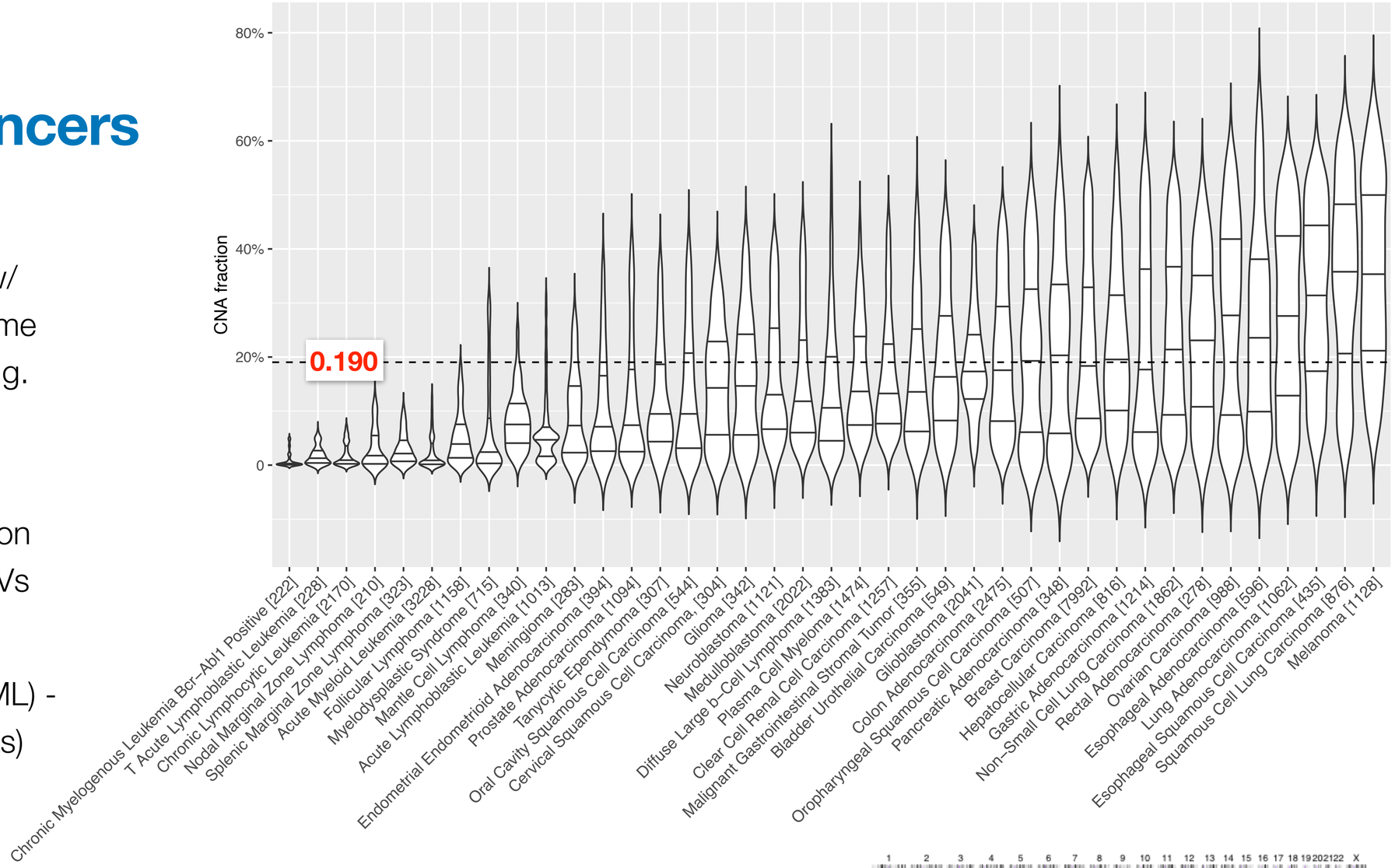
MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

low level/**high level** copy number alterations (CNAs)

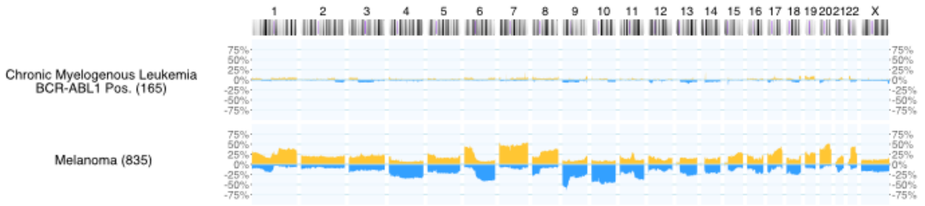
arrayMap

Genome CNV coverage in Cancers

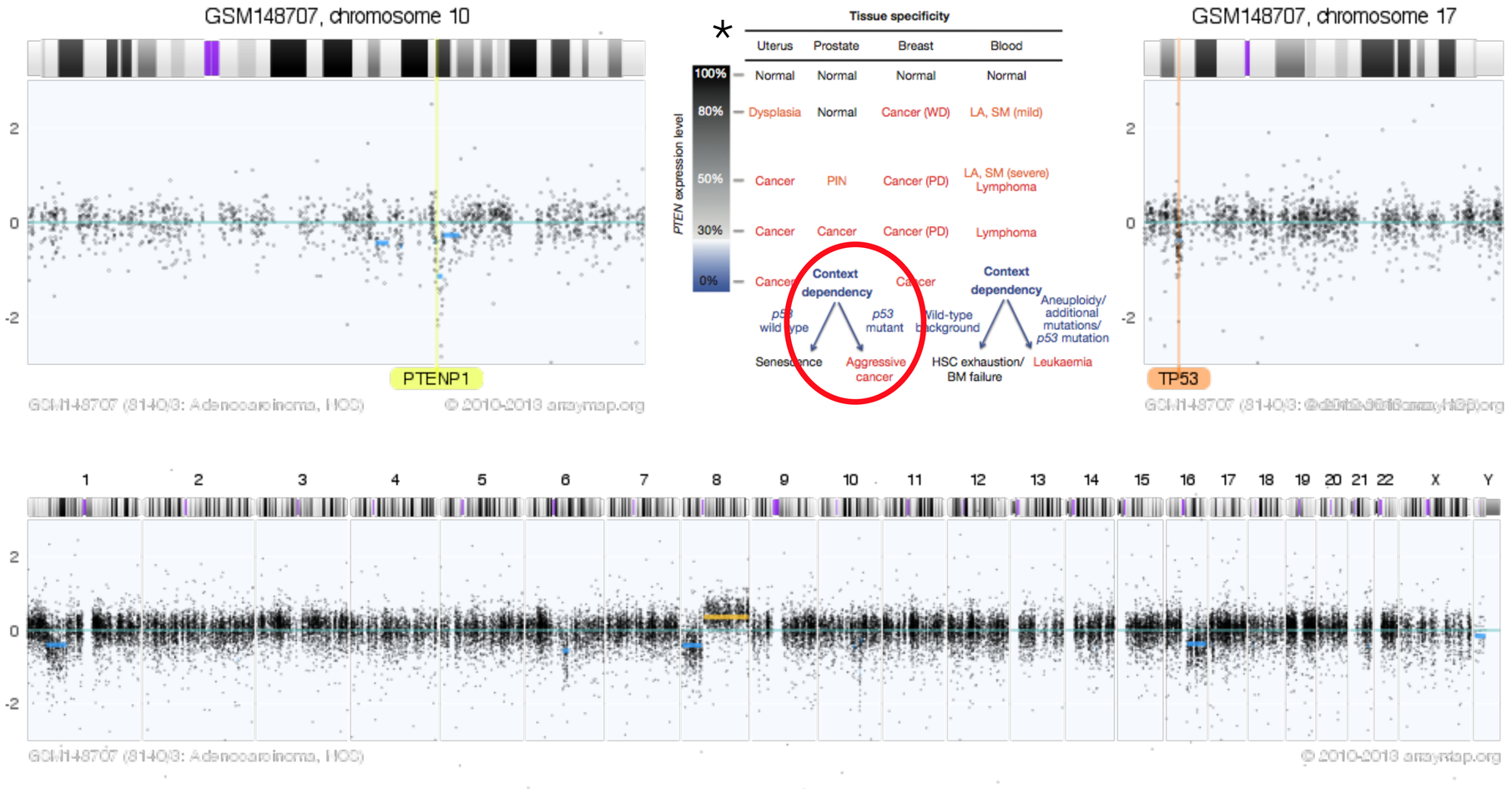
- 43654 out of 93640 CNV profiles; filtered for entities w/ >200 samples (removed some entities w/ high CNV rate, e.g. sarcoma subtypes)
- Single-sample CNV profiles were assessed for the fraction of the genome showing CNVs (relative gains, losses)
- range of medians 0.001 (CML) - 0.358 (malignant melanomas)



Lowest / Highest CNV fractions =>



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)

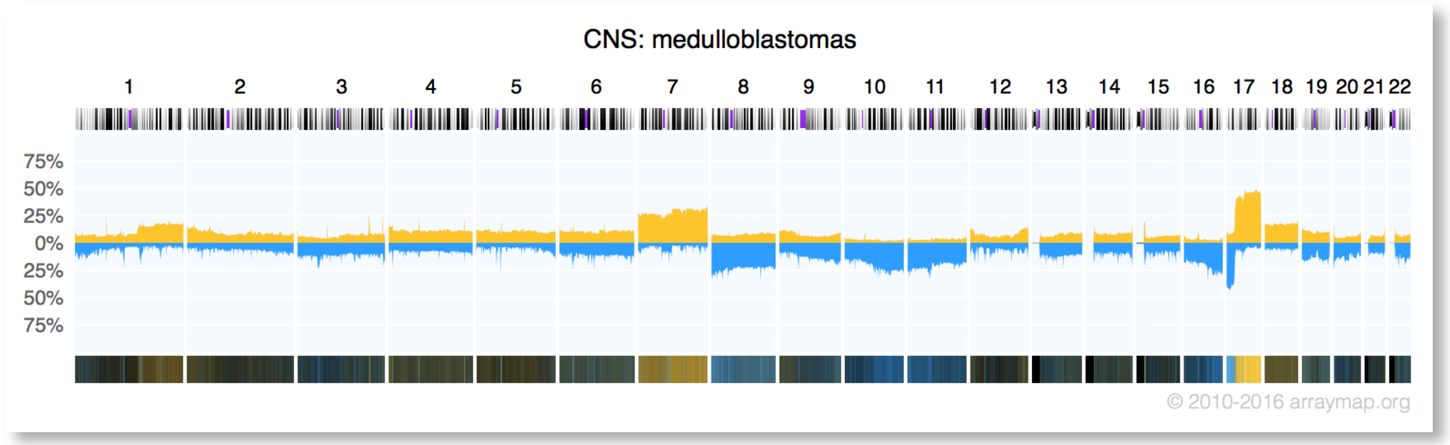
* A. H. Berger, A. G. Knudson, and P. P. Pandolfi, "A continuum model for tumour suppression," *Nature*, vol. 476, no. 7359, pp. 163–169, Aug. 2011.

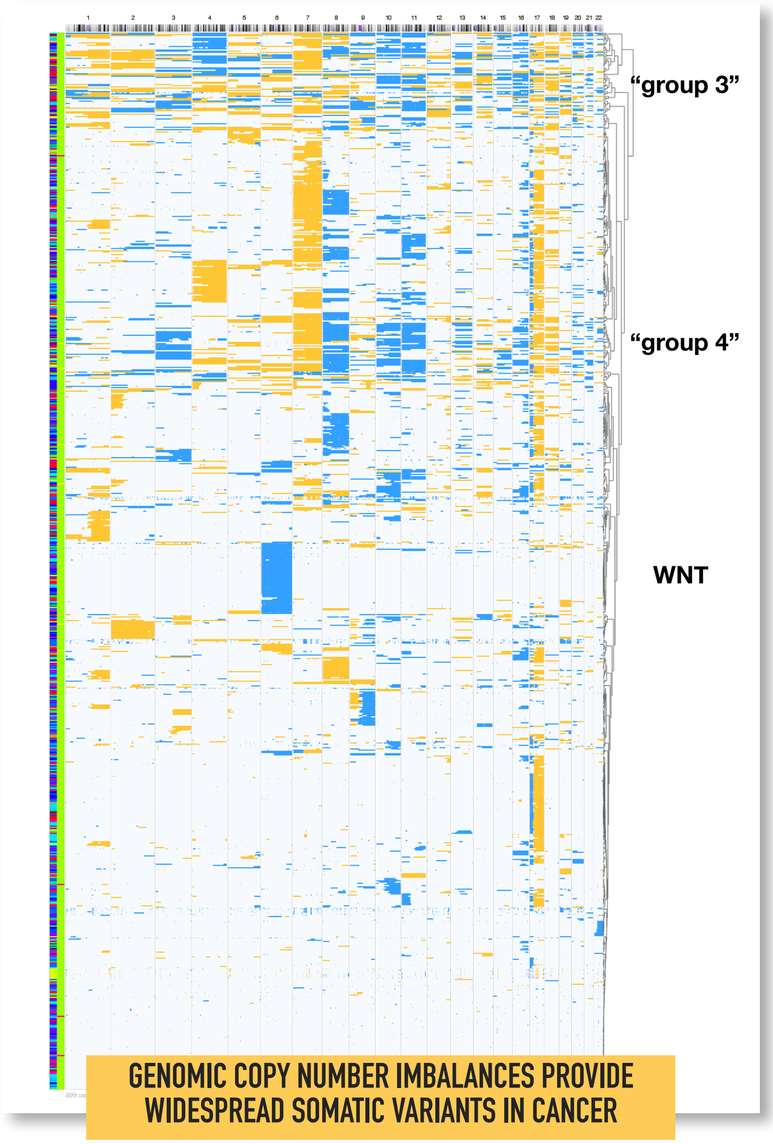
arrayMap

Somatic CNVs In Cancer

Recurrent mutation patterns

How can those patterns be used for classification and determination of biological mechanisms?





A genomic copy number histogram for malignant medulloblastomas, the most frequent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional relevance and clinical correlation.

arrayMap

progenetix.net: storage and visualization of genomic aberration data in human malignancies

michael baudis, md

Over the last decade, techniques for the genome wide scanning for genomic imbalances in malignant neoplasia have been developed, e.g. Comparative Genomic Hybridization (CGH).

Currently, no comprehensive online source for CGH data with a standardized format suitable for data mining procedures has been made available for public access. Such a data repository could be valuable in identifying genetic aberration patterns with linkage to specific disease entities, and provide additional information for validating data from large scale expression array experiments.

A case and band specific aberration matrix was selected as most suitable format for the mining of CGH data. The [progenetix.net] data repository was developed to provide the according data to the research community for a growing number of human malignancies.

In the current implementation, two main purposes are being served. First, access to the band specific pattern of chromosomal imbalances allows the instantaneous identification of genomic “hotspots”. Second, the band specific aberration matrices can be included in data mining efforts. As an example, the clustering off all informative cases from the current (September 2001) dataset is shown here (online source under www.progenetix.net/bcats/clustered.png).

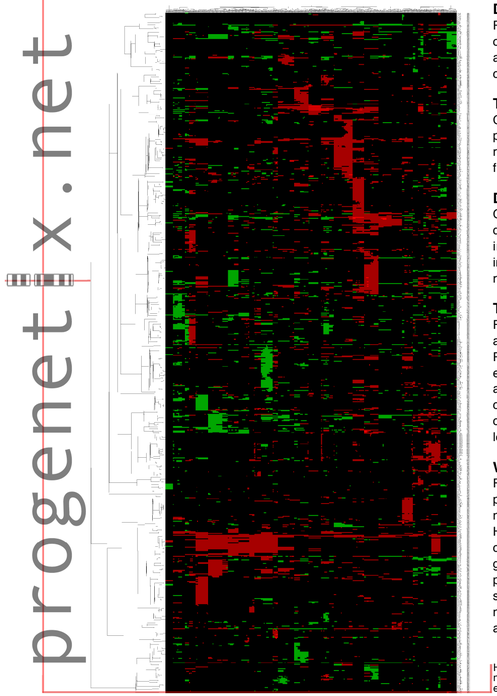
Data selection
PubMed is searched for publications applying CGH to the analysis of malignant tumors. Articles are selected according to their online availability and the description of genomic imbalances on a per case basis.

Transformation of input data
Chromosomal aberration data is transformed via customized parsing commands to a common format adherent to ISCN 1995 recommendations. In some cases, aberration data was transcribed from graphical representations or provided by the authors.

Data storage
Currently, the primary data is stored in a dedicated “off-line” database. Besides case identifier and ISCN adapted chromosomal imbalance data, tumor classification and source information including the PubMed identifier is recorded. Disease entities are reclassified to ICD-O-3 codes.

Text parsing and generation of aberration matrix
For the generation of the case and band specific aberration matrix, a dedicated text pattern comparison model was developed using Perl. Briefly, for each chromosomal band, the aberration field of each case is searched for a variety of patterns containing aberration information applying to that band. A matrix with currently 324 band resolution is generated, annotating chromosomal gains with “1” and losses with “–1”; localized high-level gains are designated “2“.

Website generation
For graphical representation of chromosomal imbalances, HTML pages containing different views of the underlying aberration matrices are generated using Perl. Graphics are implemented using HTML syntax. Besides band specific, whole genomic overviews, chromosome specific pages with links to all involved cases are generated for each ICD-O-3 entity as well as for each registered project. Additionally, those representations are available for several subsets combining related data (e.g. all lymphoid neoplasias, breast carcinoma cases). For each of the groups, the according aberration matrix is linked for download.



Hierarchical clustering of band specific chromosomal imbalances from 999 human neoplasias, contained in the [progenetix.net] collection. Cases without aberrations were excluded.

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 17 no. 12 2001*
Pages 1228–1229

Progenetix.net: an online repository for molecular cytogenetic aberration data

Michael Baudis [1, 2,] and Michael L. Cleary [2]*

[1]Medizinische Klinik und Poliklinik V der Universität Heidelberg, Germany and [2]Department of Pathology, Stanford University Medical Center, Stanford, CA 94305, USA

Received on July 5, 2001; revised on July 9, 2001; accepted on July 16, 2001

ABSTRACT

Summary: Through sequencing projects and, more recently, array-based expression analysis experiments, a wealth of genetic data has become accessible via online resources. In contrast, few of the (molecular-) cytogenetic aberration data collected in the last decades are available in a format suitable for data mining procedures. www.progenetix.net is a new online repository for previously published chromosomal aberration data, allowing the addition of band-specific information about chromosomal imbalances to oncologic data analysis efforts.

Availability: http://www.progenetix.net

Contact: mbaudis@stanford.edu

Neoplastic transformation and progression is the result of genetic defects arising in normal cells and giving rise to a malignant clone. During the process of oncogenesis, some of the usually multiple steps required for acquisition of the full neoplastic phenotype may represent themselves as numerical or structural abnormalities in the chromosomes of the transformed cells.

Over the last decades, the analysis of chromosomal abnormalities in malignant cells has gained importance in oncologic research as well as in clinical practice. A vast number of genetic abnormalities has been identified in the virtually complete range of human neoplasias. Several attempts have been undertaken for collection and classification of those abnormalities, the most widely recognized being the catalog by Mitelman and co-workers (Mitelman, 1994; online access through http://cgap.nci.nih.gov/Chromosomes/Mitelman).

In addition to metaphase analysis of short-term cultivated tumor cells or tumor cell lines, molecular cytogenetic techniques have recently been applied to the analysis of chromosomal abnormalities in primary tumor tissues. One of the more widely used screening techniques is Comparative Genomic Hybridization (CGH; Kallioniemi *et al.*, 1992; du Manoir *et al.*, 1993). Briefly, this method is based on the competitive *in-situ* hybridization of differentially labeled tumor versus normal genomic DNA to normal human metaphase spreads. The calculation of the intensity ratios of the two fluorochromes gives an overview about relative gains and losses of DNA in the tumor genome with mapping to the respective chromosomal bands. The identification of frequently imbalanced regions in tumor entities may point towards tumor suppressor gene or proto-oncogenes mapping to the respective chromosomal bands. Usually, the result of those experiments is communicated either in text format according to the International System for Cytogenetic Nomenclature (Mitelman, 1995) or graphically, with aberration bars next to chromosomal ideograms for the representation of chromosomal gains and losses.

Because in each experiment CGH analysis covers the whole number of chromosomes, the comparision of data sets from related malignancies could lead to the delineation of common as well as divergent genetic pathways defining the respective malignant phenotypes. Although an extremely large number of malignant tumors has been analyzed using this technique, no comprehensive CGH database with band-specific chromosomal aberration information is publicly available[†].

A minimal requirement for such a database would be the conversion of the text or graphical information used in publications to data tables, representing the information about the aberration status of single chromosomal bands for each case. For the site discussed here, this process includes: (1) the transformation of the published results in a format adapted from the ISCN, and (2) the automatic generation of the band specific aberration table.

Due to format variations of the published data, step 1 consists of the manual conversion of the text data or evaluation and conversion of the graphical representations, respectively. Due to the (in computational terms) odd

*To whom correspondence should be addressed.

[†] Links to a number of online CGH resources with different scopes can be found at www.progenetix.net.

© Oxford University Press 2001

next.progenetix.org

progenetix

- Cancer CNV Profiles
- Search Samples
- Publication DB
- Info
- Beacon+
- Baudisgroup @ UZH

Cancer genome data @ progenetix.org

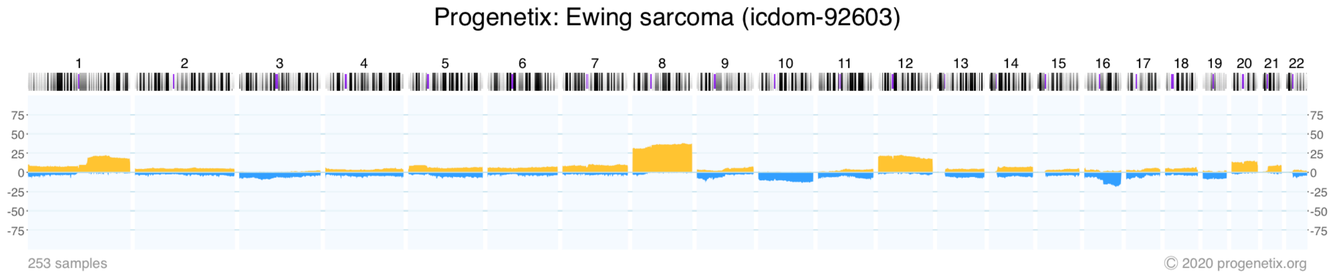
The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies.

Data Content and Provenance

The resource currently contains genome profiles of **113322** individual experiments. The genomic profiling data was derived from array and chromosomal Comparative Genomic Hybridization (CGH) experiments as well as Whole Genome or Whole Exome Sequencing (WGS, WES) studies. Genomic profiles are either processed from various raw data formats or are extracted from published experimental results.

Besides genomic profiling data, Progenetix contains sample specific biological, technical and provenance information which so far has been curated from **1600** articles.

Original diagnoses are mapped to (hierarchical) classification systems and represents **420** and **542** different cancer types, according to the International classification of Diseases in Oncology (ICD-O) and NCIt "neoplasm" classification, respectively.



For exploration of the resource it is suggested to either start with Cancer Types or by searching for CNVs genes of interest.

Additionally to genome profiles and associated metadata, the website present information about publications (currently **3962** articles) referring to cancer genome profiling experiments.

Access, Maintenance and Contributions

The content of the progenetix resource is freely accessible for research and commercial purposes, with attribution.

The database & software are developed by the group of Michael Baudis at the University of Zurich and the Swiss Institute of Bioinformatics SIB.

Many previous members and external collaborators have contributed to data content and resource features. Participation (features, data, comments) by volunteers are welcome.

© 2000 - 2020 Progenetix Cancer Genomics Information Resource by Michael Baudis is licensed under CC BY 4.0
No responsibility is taken for the correctness of the data presented nor the results achieved with the Progenetix tools.

GENOMICS

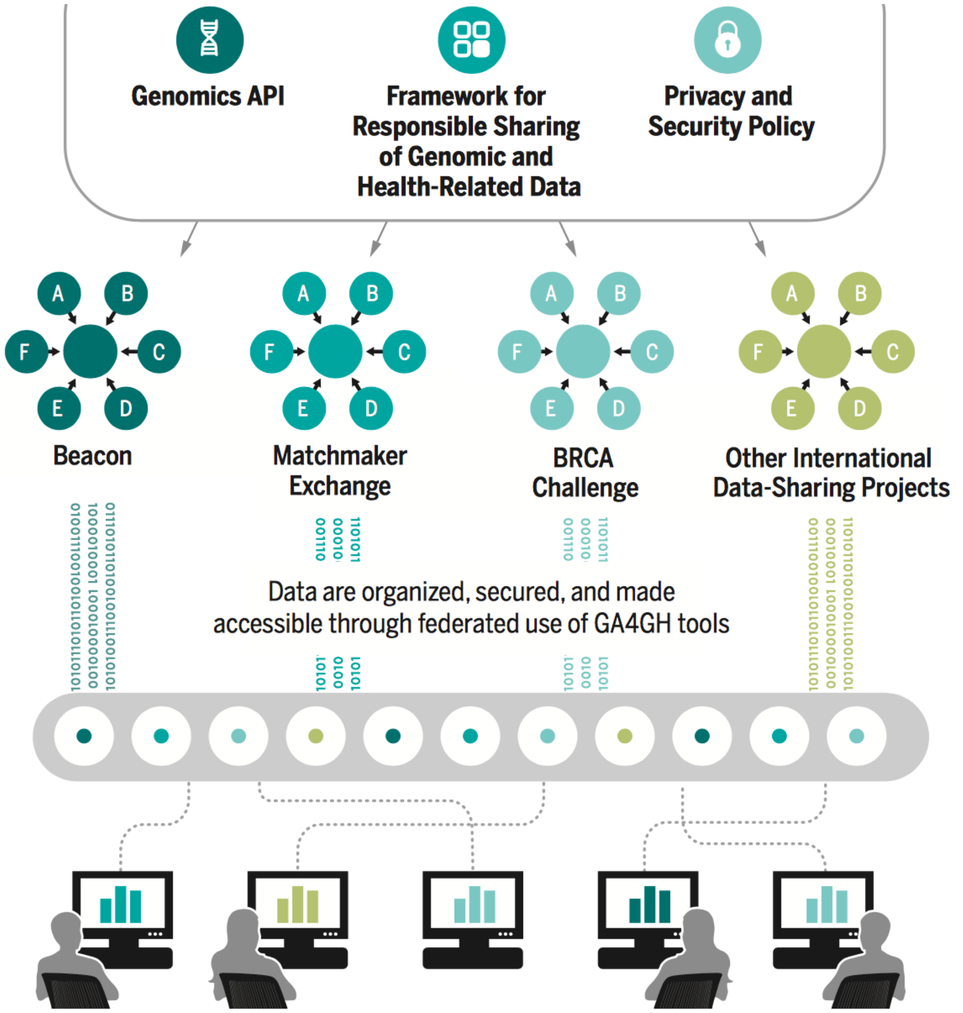
A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

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



Global Alliance for Genomics & Health

GENOMICS

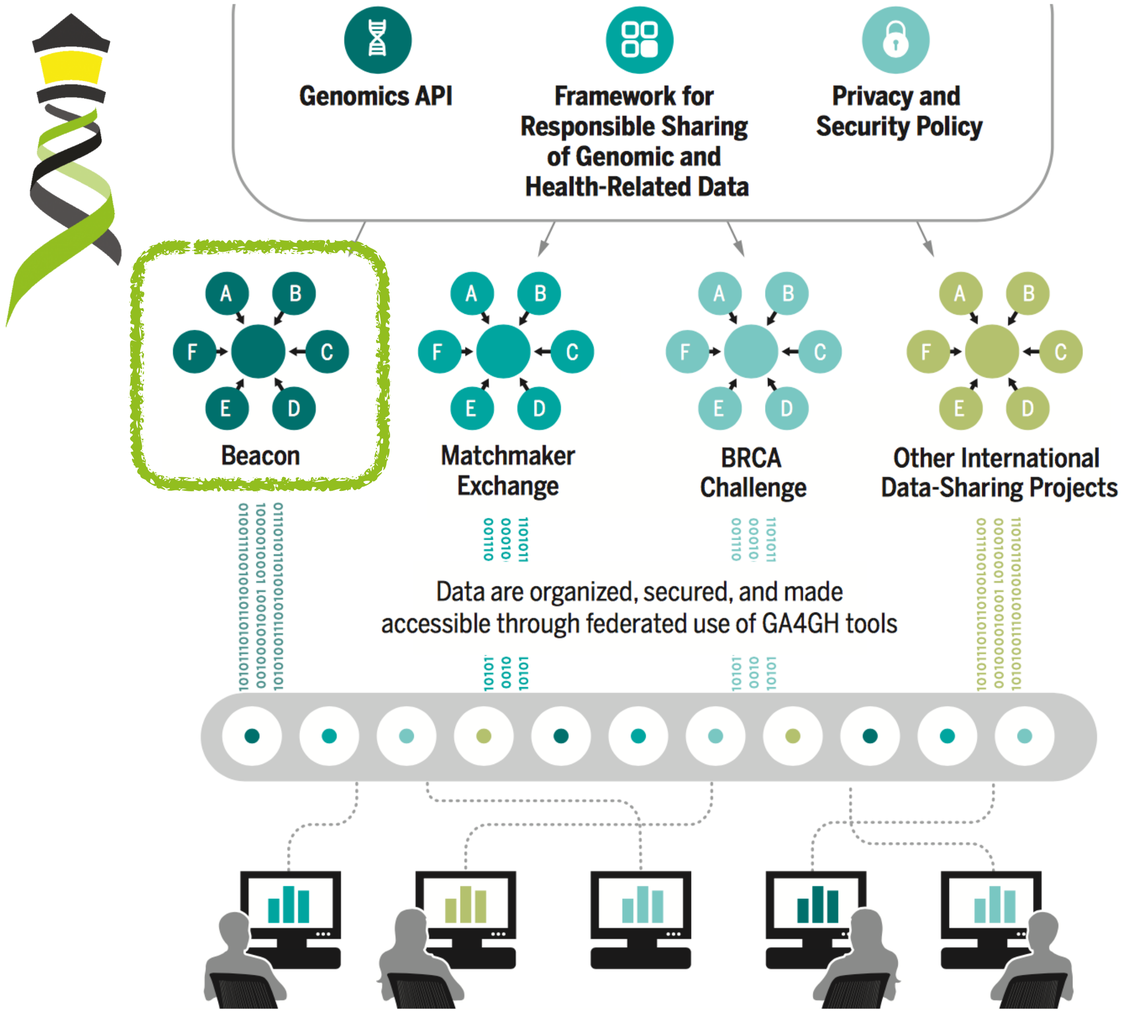
A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

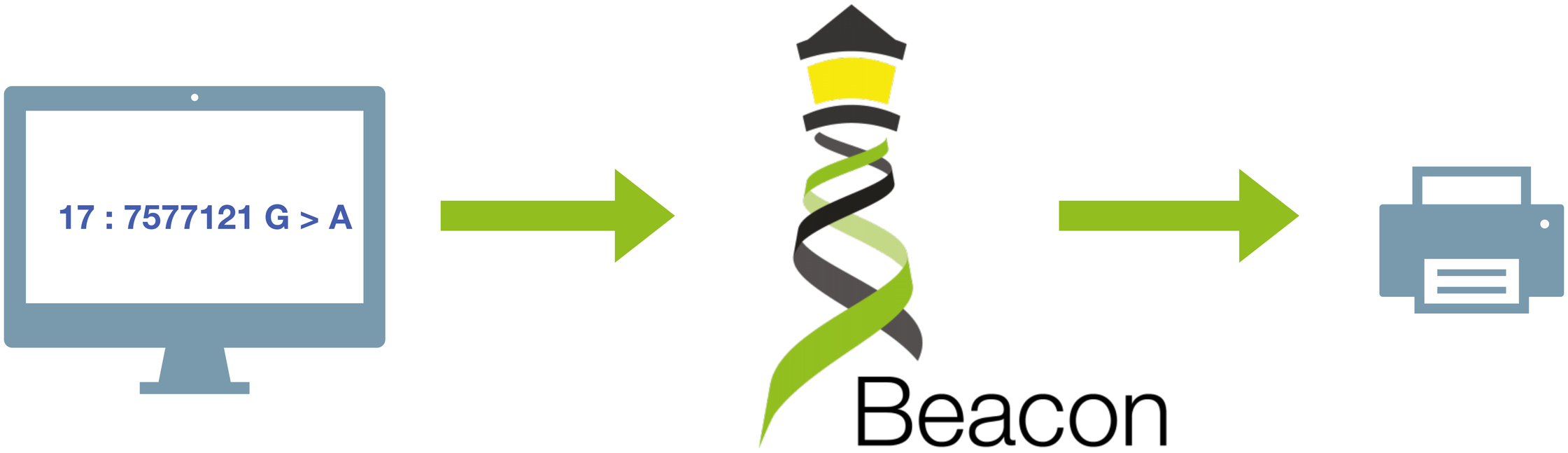
SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

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

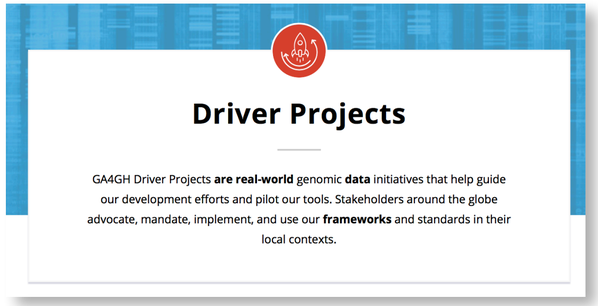


A ***Beacon*** answers a query for a specific genome variant against individual or aggregate genome collections

YES | **NO** | **\0**

ELIXIR Beacon Project

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



ELIXIR Beacon
https://www.elixir-europe.org/about/implementation-studies/beacons

Europe

Champions: Jordi Rambla, Juha Tornroos, Gary Saunders

v1.1 and roadmap

- ☑ **structural variations** (DUP, DEL) in addition to SNV
- ☑ … more structural queries (translocations/fusions…)
- ☑ Beacon queries as entry for **data handover** (outside Beacon protocol)
- ☑ layered authentication system using **ELIXIR AAI**
- **v2** **filters** for phenotypic & technical metadata
- **v2** Extended quantitative responses
- ☐ Ubiquitous **deployment** (e.g. throughout ELIXIR network)

Global Alliance for Genomics & Health

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**".

Features and Possibilities of the current Beacon Specification

Beyond "testing the willingness for data sharing"...

- precise variant queries (chr17: 7673767 C>T)
- range queries ("any variant from here to there")
- variant frequencies
- structural genome variants, e.g. CNVs ("any deletion overlapping CDKN2A CDR coordinates")
- delivery of any kind of data matching a given query (variants, sample information, patient data ...) utilising "***handover***" objects (anonymous links to external services with their own security / privacy implementations)
- networking of v1.n Beacons with AAI integration as demonstrated by the ELIXIR Beacon Network

DX Ontologies

Hierarchical NCIt Neoplasm Core replaces heterogeneous primary annotations

- heterogeneous and inconsistent diagnostic annotations are common in clinical reports and research studies ("text", ICD-10, ICD-O 3, OncoTree, domain-specific classifications)
- highly variable granularity of annotations is a major road block for comparative analyses and large scale data integration
  - "Colorectal Cancer" or "Rectal Mucinous Adenoca."
- initiatives and services such as Phenopackets, MONDO, OXO ... rely on and/or provide mappings to hierarchical ontologies

progenetix

NCIt Neoplasm Core coded display (excerpt) for samples in the Progenetix cancer genome data resource allows sample selection on multiple hierarchy levels →

Collapse all | Expand | All

	Subsets	Samples
☐	NCIT:C3262: Neoplasm	88844
☐	NCIT:C3263: Neoplasm by Site	84747
☐	NCIT:C156482: Genitourinary System Neoplasm	11616
☐	NCIT:C156483: Benign Genitourinary System Neoplasm	219
☐	NCIT:C4893: Benign Urinary System Neoplasm	90
☐	NCIT:C4778: Benign Kidney Neoplasm	90
☐	NCIT:C159209: Kidney Leiomyoma	1
☐	NCIT:C4526: Kidney Oncocytoma	82
☐	NCIT:C8383: Kidney Adenoma	7
☐	NCIT:C7617: Benign Reproductive System Neoplasm	129
☐	NCIT:C4934: Benign Female Reproductive System Neoplasm	129
☐	NCIT:C2895: Benign Ovarian Neoplasm	58
☐	NCIT:C4510: Benign Ovarian Epithelial Tumor	58
☐	NCIT:C40039: Benign Ovarian Mucinous Tumor	58
☐	NCIT:C4512: Ovarian Mucinous Cystadenoma	58
☐	NCIT:C4060: Ovarian Cystadenoma	58
☐	NCIT:C4512: Ovarian Mucinous Cystadenoma	58
☐	NCIT:C3609: Benign Uterine Neoplasm	71
☐	NCIT:C3608: Benign Uterine Corpus Neoplasm	71
☐	NCIT:C3434: Uterine Corpus Leiomyoma	71
☐	NCIT:C156484: Malignant Genitourinary System Neoplasm	11171
☐	NCIT:C157774: Metastatic Malignant Genitourinary System Neoplasm	2
☐	NCIT:C146893: Metastatic Genitourinary System Carcinoma	2
☐	NCIT:C8946: Metastatic Prostate Carcinoma	2
☐	NCIT:C164141: Genitourinary System Carcinoma	10561
☐	NCIT:C146893: Metastatic Genitourinary System Carcinoma	2
☐	NCIT:C8946: Metastatic Prostate Carcinoma	2
☐	NCIT:C3867: Fallopian Tube Carcinoma	19

NCIt Neoplasm Core

Beacon v2 filters make use of hierarchical classification systems

- Beacon v2 "filters" assumes inclusion of child terms when using hierarchical classifications

  ➡ implicit *OR* with otherwise assumed *AND*

- implementation of hierarchical annotations overcomes some limitatiions of "fuzzy" disease annotations

- data *handover* (Beacon v1.1+) enables further data exploration and export scenarios

Beacon+ specific: Multiple term selection with OR logic

☑	NCIT:C4914: Skin Carcinoma	213
☐	NCIT:C4475: Dermal Neoplasm	109
☑	NCIT:C45240: Cutaneous Hematopoietic and Lymphoid Cell Neoplasm	310

Filters: NCIT:C4914, NCIT:C4819, NCIT:C9231, NCIT:C2921, NCIT:C45240, NCIT:C6858, NCIT:C3467, NCIT:C45340, NCIT:C7195, NCIT:C3246, NCIT:C7217

progenetix

Variants: 0
Calls: 0
Samples: 523

$f_{alleles}$: 0

Callsets Variants | Legacy Interface | UCSC region

Show JSON Response

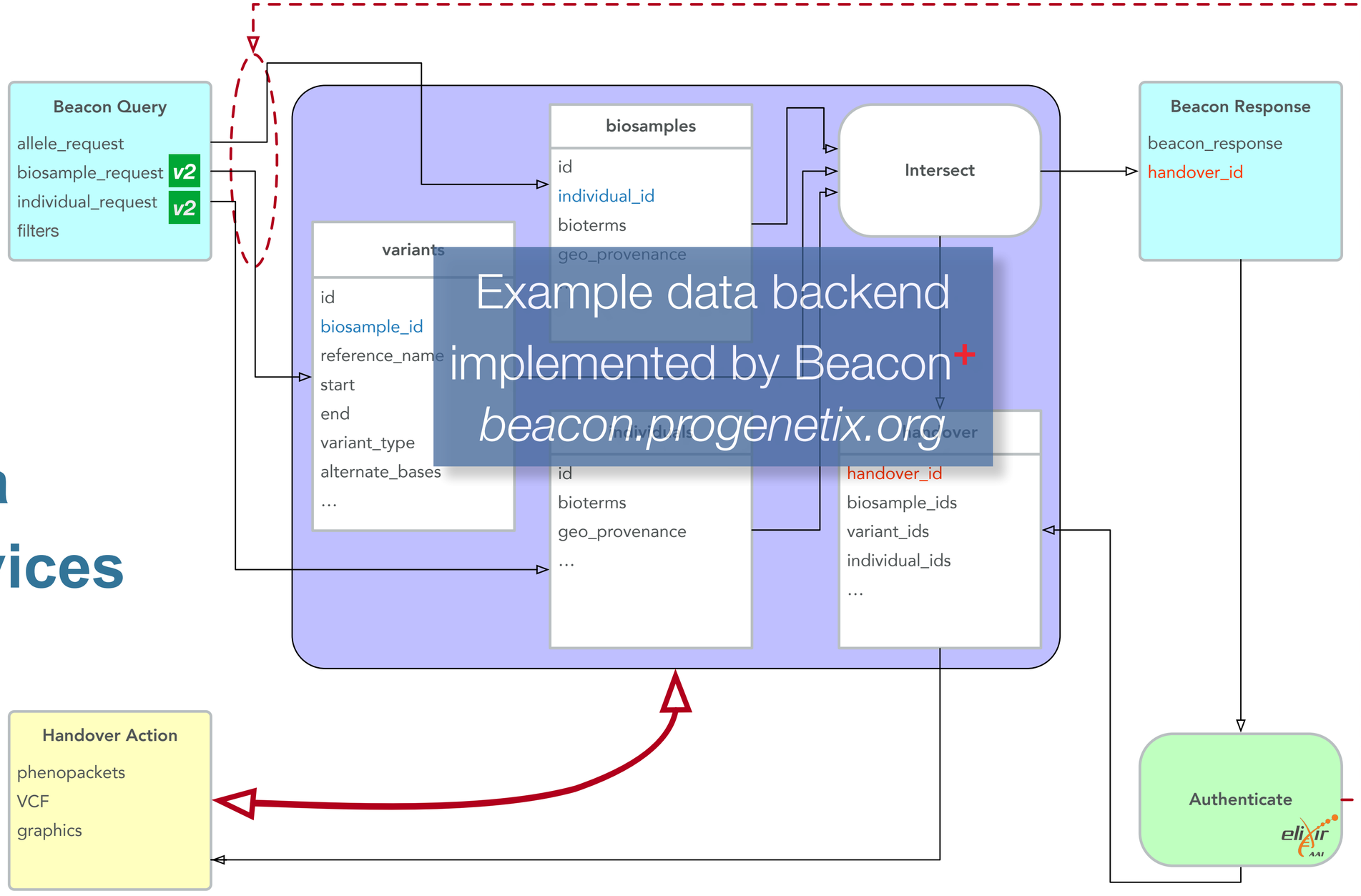
Results | Biosamples

Id	Description	Classifications	Identifiers	DEL	DUP	CNV
PGX_AM_BS_MCC01	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.116	0.104	0.22
PGX_AM_BS_MCC02	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.154	0.056	0.21
PGX_AM_BS_MCC03	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.137	0.21	0.347
PGX_AM_BS_MCC04	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.158	0.056	0.214
PGX_AM_BS_MCC05	Merkel cell carcinoma	icdot-C44.9 Skin, NOS icdom-82473 Merkel cell carcinoma NCIT:C9231 Merkel Cell Carcinoma	PMID:9537255	0.107	0.327	0.434

Page **1 of 105**

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

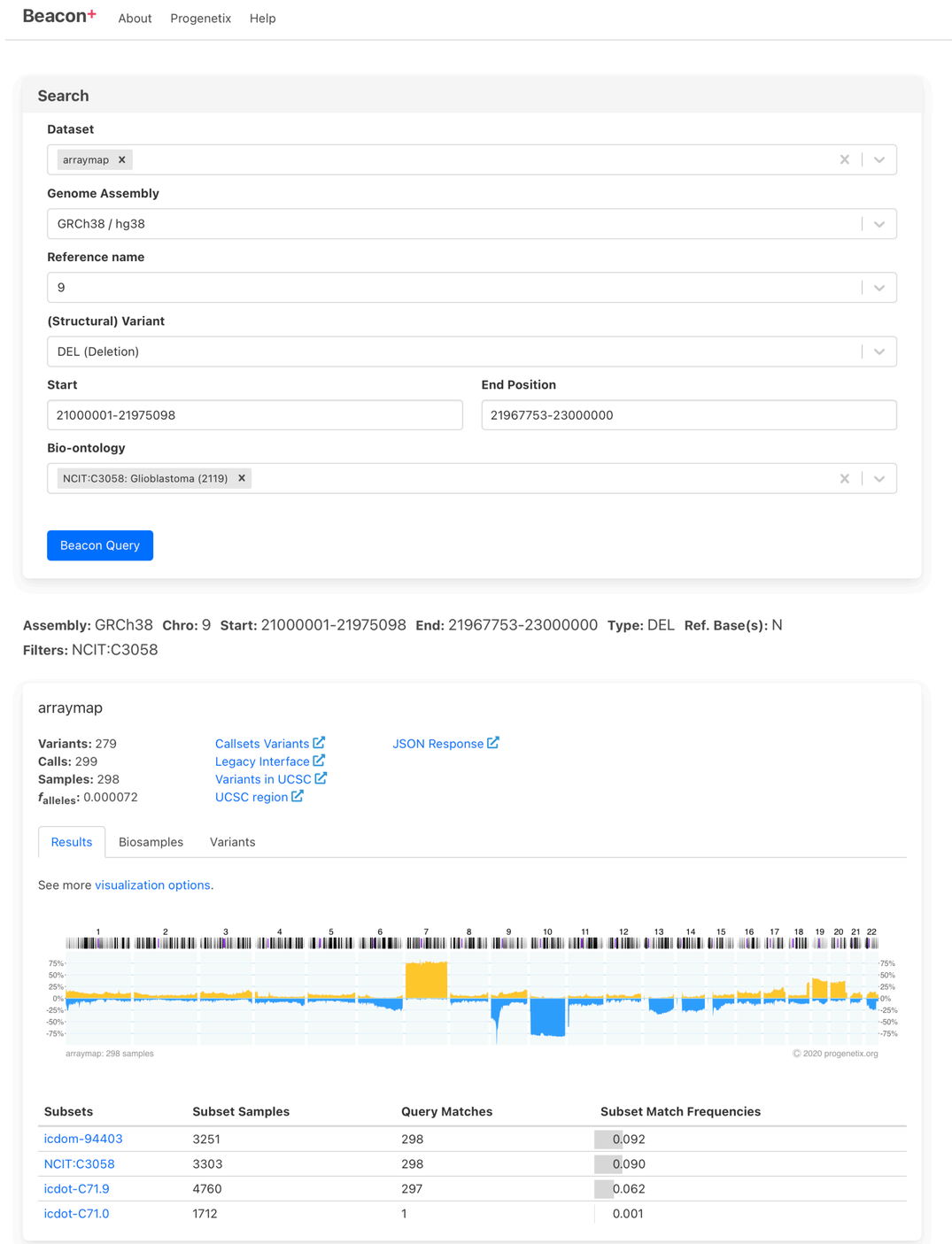
elixir

Beaconized Progenetix

From Beacon Query to Explorative Analyses of CNV Patterns

- The consistent use of hierarchical diagnostic codes allows the use of Beacon "filters" for histopathological/clinically scoped queries
- Beacon's handover protocols can be utilized for data retrieval and, well, handing over to additional services, e.g.
  - downloads
  - visualization
  - use of external services (UCSC browser display...)

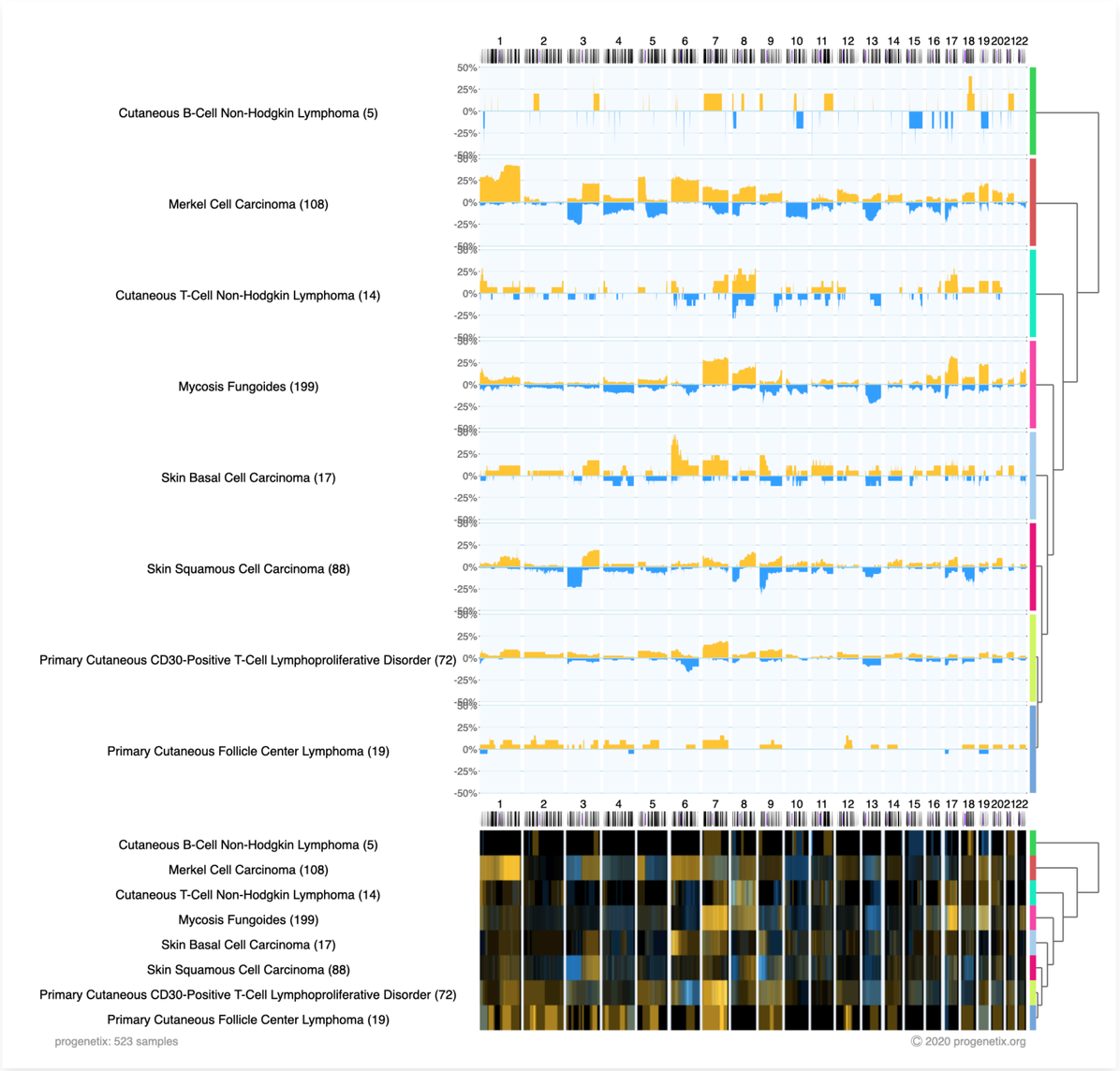
progenetix



Beaconized Progenetix

From Beacon Query to Explorative Analyses of CNV Patterns

- The consistent use of hierarchical diagnostic codes allows the use of Beacon "filters" for histopathological/clinically scoped queries
- Beacon's handover protocols can be utilized for data retrieval and, well, handing over to additional services, e.g.
  - downloads
  - visualization
  - use of external services (UCSC browser display...)



Standardized Data

Data re-use depends on standardized, machine-readable metadata

- Multiple international initiatives (ELIXIR, GA4GH, MONARCH...) and resource providers (EBI, NCBI ...) work on the generation and implementation of data annotation standards
- emerging / established principles are the use of hierarchical coding systems where individual codes are represented as CURIEs
- other formats for non-categorical annotations based on international standards, e.g.
  - ISO (ISO 8601 time & period, ISO 3166 country codes ...)
  - IETF (GeoJSON ...)
  - W3C (CURIE ...)
- these standards become pervasive throughout GA4GH's ecosystem (e.g. Phenopackets ...)

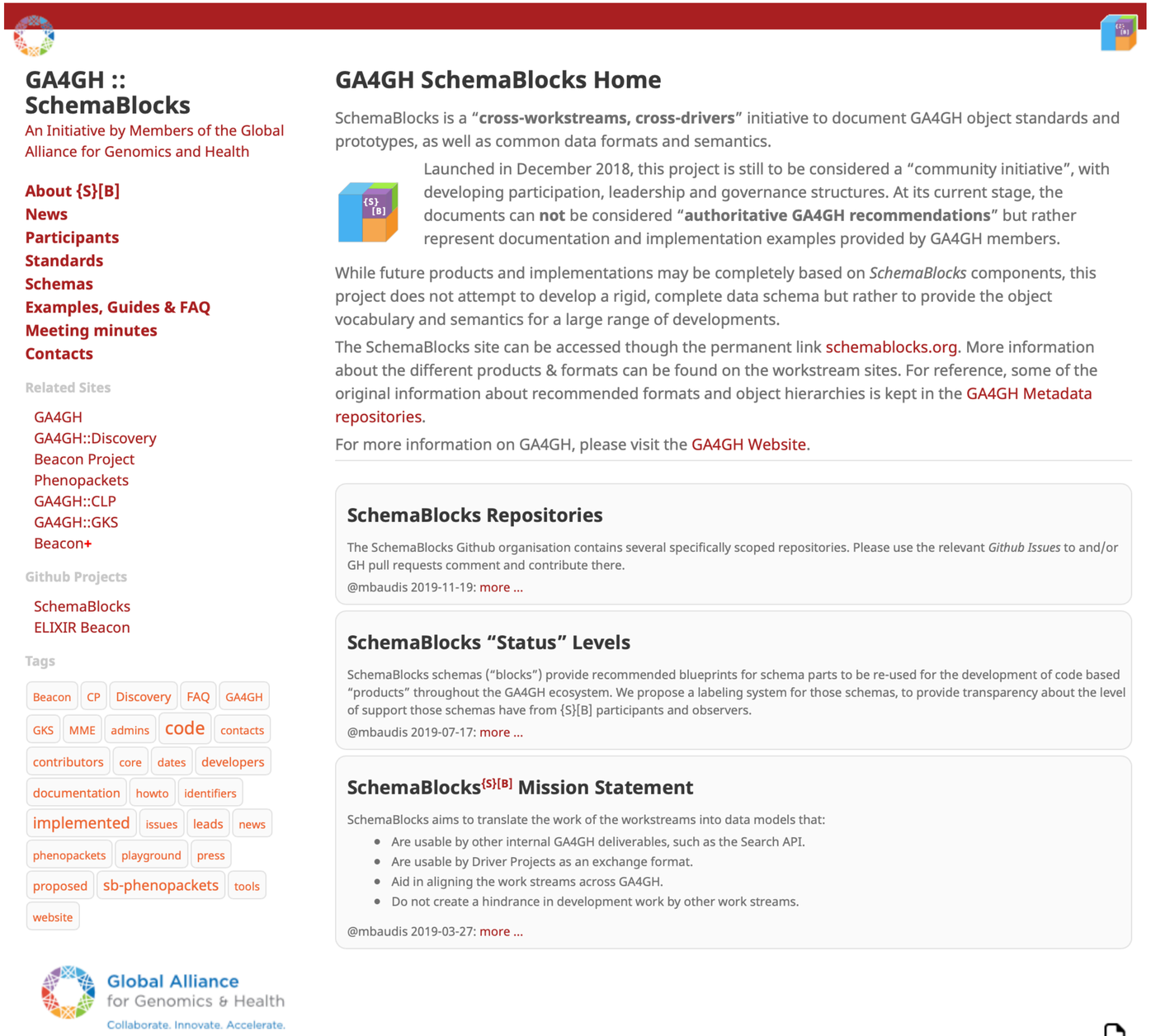
```
"data_use_conditions" : {
  "label" : "no restriction",
  "id" : "DUO:0000004"
},

"provenance" : {
  "material" : {
    "type" : {
      "id" : "EFO:0009656",
      "label" : "neoplastic sample"
    }
  },
  "geo" : {
    "label" : "Zurich, Switzerland",
    "precision" : "city",
    "city" : "Zurich",
    "country" : "Switzerland",
    "latitude" : 47.37,
    "longitude" : 8.55,
    "geojson" : {
      "type" : "Point",
      "coordinates" : [
        8.55,
        47.37
      ]
    },
    "ISO-3166-alpha3" : "CHE"
  }
},
{
  "age": "P25Y3M2D"
}
```

GA4GH {S}[B] *SchemaBlocks*

Standardized formats and data schemas for developing an "Internet of Genomics"

- "cross-workstreams, cross-drivers" initiative to document GA4GH object ***standards*** and ***prototypes***
- launched in December 2018
- documentation and implementation examples provided by GA4GH members
- not a rigid, complete data schema
- object ***vocabulary*** and ***semantics*** for a large range of developments
- **Beacon** as *contributor* and *user*



GA4GH :: SchemaBlocks

An Initiative by Members of the Global Alliance for Genomics and Health

About {S}[B]
News
Participants
Standards
Schemas
Examples, Guides & FAQ
Meeting minutes
Contacts

Related Sites

GA4GH
GA4GH::Discovery
Beacon Project
Phenopackets
GA4GH::CLP
GA4GH::GKS
Beacon+

Github Projects

SchemaBlocks
ELIXIR Beacon

Tags

Beacon CP Discovery FAQ GA4GH GKS MME admins code contacts contributors core dates developers documentation howto identifiers implemented issues leads news phenopackets playground press proposed sb-phenopackets tools website

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

GA4GH SchemaBlocks Home

SchemaBlocks is a "**cross-workstreams, cross-drivers**" initiative to document GA4GH object standards and prototypes, as well as common data formats and semantics.

Launched in December 2018, this project is still to be considered a "community initiative", with developing participation, leadership and governance structures. At its current stage, the documents can **not** be considered "**authoritative GA4GH recommendations**" but rather represent documentation and implementation examples provided by GA4GH members.

While future products and implementations may be completely based on *SchemaBlocks* components, this project does not attempt to develop a rigid, complete data schema but rather to provide the object vocabulary and semantics for a large range of developments.

The SchemaBlocks site can be accessed though the permanent link schemablocks.org. More information about the different products & formats can be found on the workstream sites. For reference, some of the original information about recommended formats and object hierarchies is kept in the GA4GH Metadata repositories.

For more information on GA4GH, please visit the GA4GH Website.

SchemaBlocks Repositories

The SchemaBlocks Github organisation contains several specifically scoped repositories. Please use the relevant *Github Issues* to and/or GH pull requests comment and contribute there.

@mbaudis 2019-11-19: more ...

SchemaBlocks "Status" Levels

SchemaBlocks schemas ("blocks") provide recommended blueprints for schema parts to be re-used for the development of code based "products" throughout the GA4GH ecosystem. We propose a labeling system for those schemas, to provide transparency about the level of support those schemas have from {S}[B] participants and observers.

@mbaudis 2019-07-17: more ...

SchemaBlocks{S}[B] Mission Statement

SchemaBlocks aims to translate the work of the workstreams into data models that:

- Are usable by other internal GA4GH deliverables, such as the Search API.
- Are usable by Driver Projects as an exchange format.
- Aid in aligning the work streams across GA4GH.
- Do not create a hindrance in development work by other work streams.

@mbaudis 2019-03-27: more ...

schemablocks.org

BeaconAlleleRequest beacon ↗

{S}[B] Status [i]	implemented
Provenance	○ Beacon API
Used by	○ Beacon ○ Progenetix database schema (Beacon+ backend)
Contributors	○ Marc Fiume ○ Michael Baudis ○ Sabela de la Torre Pernas ○ Jordi Rambla ○ Beacon developers...
Source (v1.1.0)	○ raw source [JSON] ○ Github

Attributes

Type: object
Description: Allele request as interpreted by the beacon.

Properties

Property	Type
alternateBases	string
assemblyId	string
datasetIds	array of string
end	integer
endMax	integer
endMin	integer
mateName	https://schemablocks.org/schemas/beacon/v1.1.0/Chror [HTML]
referenceBases	string
referenceName	https://schemablocks.org/schemas/beacon/v1.1.0/Chror [HTML]
start	integer (int64)
startMax	integer
startMin	integer
variantType	string

alternateBases

- type: string

The bases that appear instead of the reference bases. Accepted values: [ACGTN]*. N is a wildcard, that denotes the position of any base, and can be used as a standalone base of any type or within a partially known sequence. For example a sequence where the first and last bases are known, but the middle portion can exhibit countless variations of [ACGT], or the bases are unknown: ANNT the Ns can take take any form of [ACGT], which makes both ACCT and ATGT (or any other combination) viable sequences.

Symbolic ALT alleles (DEL, INS, DUP, INV, CNV, DUP:TANDEM, DEL:ME, INS:ME) will be represented in `variantType`.

Optional: either `alternateBases` or `variantType` is required.

`alternateBases` **Value Example**

assemblyId

- type: string

Assembly identifier (GRC notation, e.g. GRCh37).

`assemblyId` **Value Example**

Curie sb-vr-spec ↗

{S}[B] Status [i]	implemented
Provenance	○ vr-spec
Used by	○ vr-spec
Contributors	○ Reece Hart ○ Michael Baudis
Source (v1.0)	○ raw source [JSON] ○ Github

Attributes

Type: string
Pattern: ^\w[^:]+:.+$
Description: A string that refers to an object uniquely. T sender.

VR does not impose any contraints on strings used as ids data, the VR Specification RECOMMENDS that implement

String CURIEs are represented as `prefix:reference` (W `namespace:accession` or `namespace:local id` colloqu

The VR specification also RECOMMENDS that `prefix` be

The `reference` component is an unconstrained string.

A CURIE is a URI. URIs may *locate* objects (i.e., specify wh VR uses CURIEs primarily as a naming mechanism.

Implementations MAY provide CURIE resolution mechan

Using internal ids in public messages is strongly discoura

`Curie` Value Examples

```
"ga4gh:GA.01234abcde"
```

```
"DUO:0000004"
```

```
"orcid:0000-0003-3463-0775"
```

```
"PMID:15254584"
```

Biosample sb-phenopackets ↗

{S}[B] Status [i]	implemented
Provenance	○ Phenopackets
Used by	○ Phenopackets
Contributors	○ GA4GH Data Working Group ○ Jules Jacobsen ○ Peter Robinson ○ Michael Baudis ○ Melanie Courtot ○ Isuru Liyanage
Source (v1.0.0)	○ raw source [JSON] ○ Github

Attributes

Type: object
Description: A Biosample refers to a unit of biological material from which the substrate molec genomic DNA, RNA, proteins) for molecular analyses (e.g. sequencing, array hybridisation, mas spectrometry) are extracted.

Examples would be a tissue biopsy, a single cell from a culture for single cell genome sequencii fraction from a gradient centrifugation.

Several instances (e.g. technical replicates) or types of experiments (e.g. genomic array as well experiments) may refer to the same Biosample.

FHIR mapping: Specimen.

Properties

Property	Type
ageOfIndividualAtCollection	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/Age.json [SRC] [HTML]
ageRangeOfIndividualAtCollection	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/AgeRange.json [SRC] [HTML]
description	string
diagnosticMarkers	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/OntologyClass.json [SRC] [HTMI
histologicalDiagnosis	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/OntologyClass.json [SRC] [HTMI
htsFiles	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/HtsFile.json [SRC] [HTML]
id	string
individualId	string
isControlSample	boolean
phenotypicFeature	array of https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/PhenotypicFeature.json [SRC] [HTML]
procedure	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/Procedure.json [SRC] [HTML]
sampledTissue	https://schemablocks.org/schemas/sb-

Checksum sb-checksum ↗

{S}[B] Status [i]	proposed
Provenance	○ GA4GH DRS (`develop` branch)
Used by	○ GA4GH DRS ○ GA4GH TRS
Contributors	○ Susheel Varma
Source (v0.0.1)	○ raw source [JSON] ○ Github

Attributes

Type: object
Description: Checksum

Properties

Property	Type
checksum	string
type	string

checksum

- type: string

The hexadecimal encoded (Base16) checksum for the data

`checksum` **Value Example**

```
"77af4d6b9913e693e8d0b4b294fa62ade6054e6b2f1ffb617ac955dd63fb0182"
```

type

- type: string

The digest method used to create the checksum. The value (e.g. `sha-256`) SHOULD be listed as `Hash Name String` in the GA4GH Hash Algorithm Registry. Other values MAY be used, as long as implementors are aware of the issues discussed in RFC6920.

GA4GH may provide more explicit guidance for use of non-IANA-registered algorithms in the future.

`type` **Value Example**

```
"sha-256"
```

Cancer (Genomics) Use Case

Lessons from Implementing a Cancer CNV Resource w/ GA4GH & ELIXIR Standards

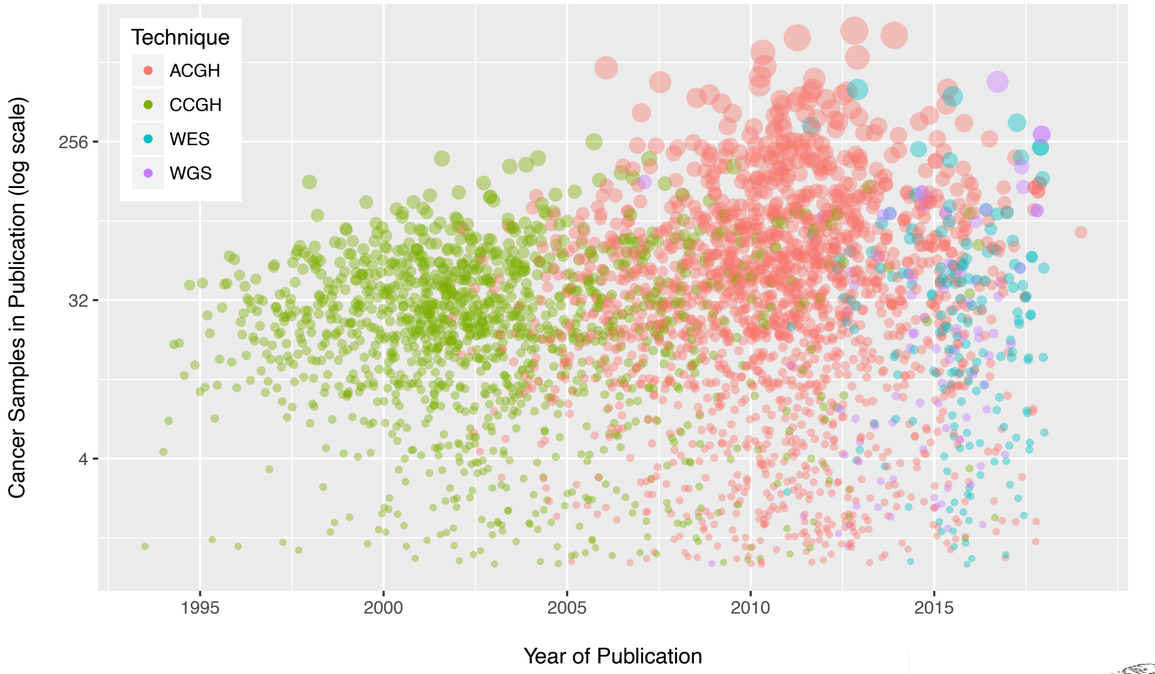
- Beacon v1.1 -> 2.alpha is already very capable for data search & delivery, at least for the "minimal expected subset" of associated data
  - some additions (query logic, data aggregation...) helpful in increasing utility w/o breaking the standard
- the Beacon v2 "filters" paradigm coupled with consistent use of CURIEs has proven very reliable & should drive adoption of standard ontologies throughout resources
  - may require remapping or translation service integration by resource providers
- GA4GH standards such as DUO for and Passport together w/ ELIXIR AAI (&open implementations) will enable opening of controlled access to yet closed resources
- several ELIXIR initiatives, platforms and projects work on integrating and driving GA4GH standards, partially through use of standard documentation initiatives like ***SchemaBlocks***
- consistent use of metadata formats allows new insights into "accidentally collected data"

elixir

Global Alliance for Genomics & Health

Metadata in Cancer Studies Beyond Cancer Genomics ...

Publication Landscape of Cancer CNV Profiling

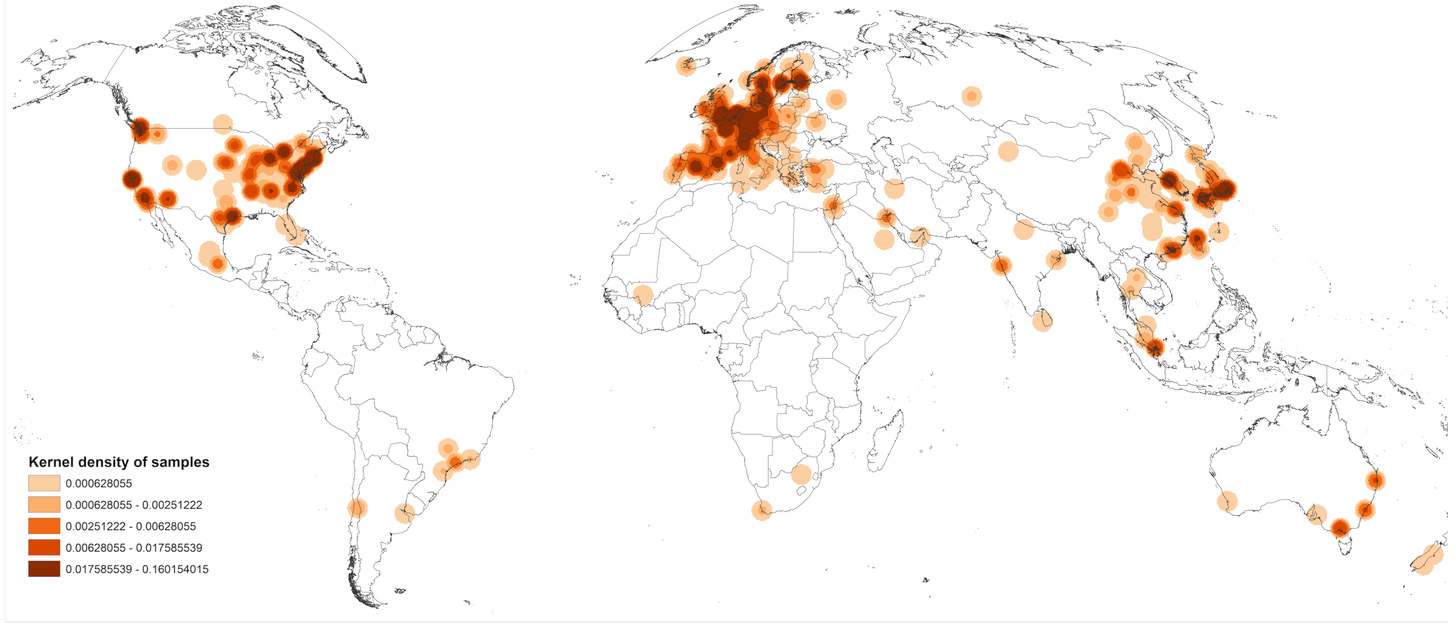


Publication statistics for cancer genome screening studies. The graphic shows our assessment of publications reporting whole-genome screening of cancer samples, using molecular detection methods (chromosomal CGH, genomic array technologies, whole exome and genome sequencing).

For the years 1993-2018, we found 3'229 publications reporting 174'530 individual samples in single series from 1 to more than 1000 samples. Y-axis and size of the dots correspond to the sample number; the color codes indicate the technology used.

Map of the geographic distribution (by first author affiliation) of the 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets.

The numbers are derived from the 3'240 publications registered in the Progenetix database.



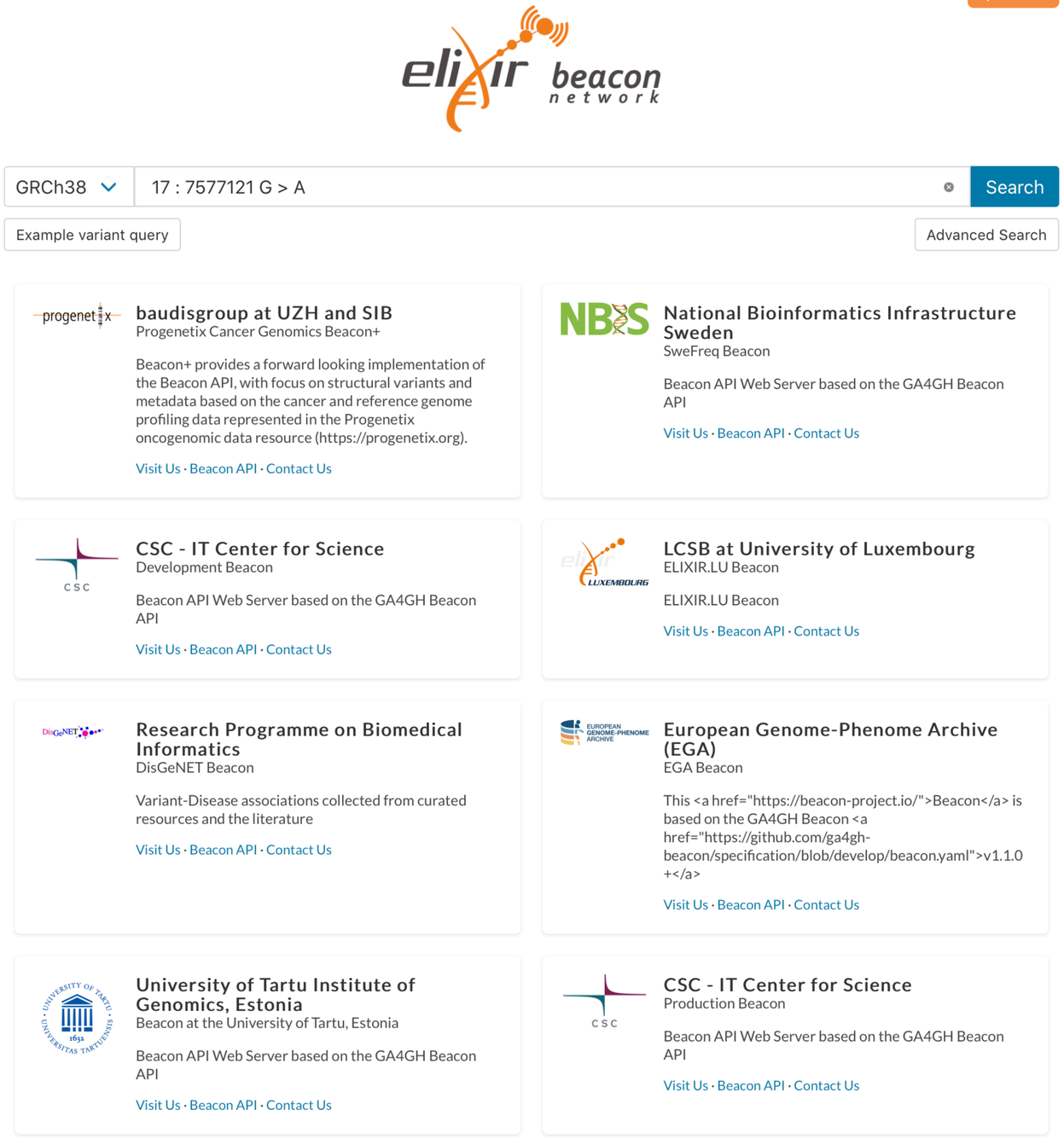
progenetix

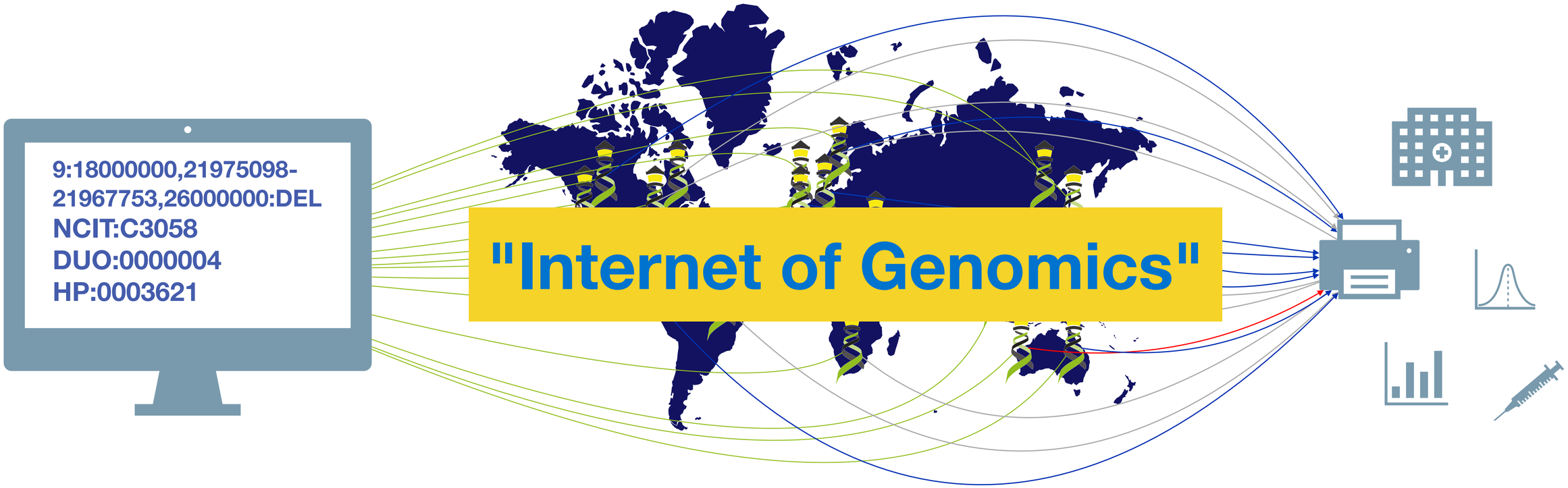
ELIXIR Beacon Network

- developed under lead from ELIXIR Finland
- **authenticated access** w/ ELIXIR AAI
- **incremental extension**, starting with ELIXIR Beacon resources adhering to the **latest specification** (contrast to legacy networks)
- service details provided by individual Beacons, using **GA4GH service-info**
- **registration service**

➡**integrator** throughout ELIXIR Human Data

➡starting point for "**beyond ELIXIR**" **feature rich** federated Beacon **services**

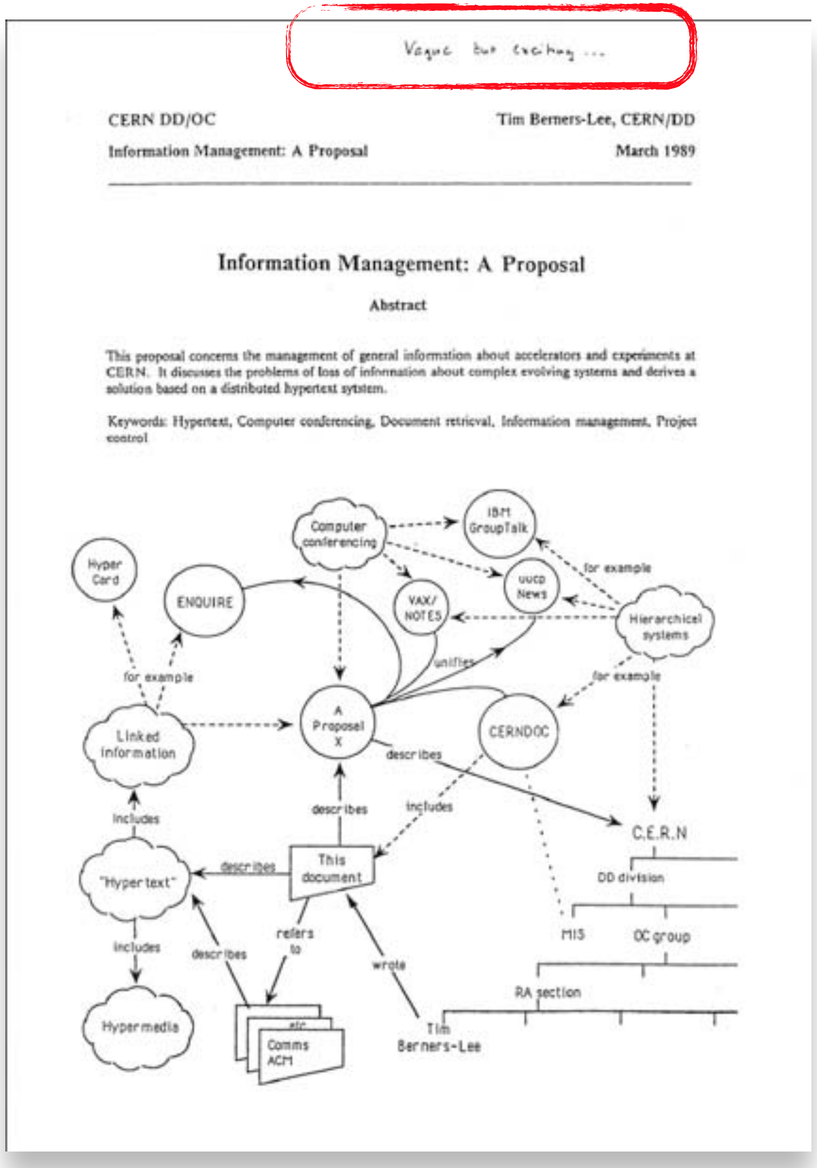




Have you seen deletions in this region on chromosome 9 in Glioblastomas from a juvenile patient, in a dataset with unrestricted access?

Beacon *v2* API

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



Vague but exciting ...

CERN DD/OC — Tim Berners-Lee, CERN/DD

Information Management: A Proposal — March 1989

Information Management: A Proposal

Abstract

This proposal concerns the management of general information about accelerators and experiments at CERN. It discusses the problems of loss of information about complex evolving systems and derives a solution based on a distributed hypertext sytstem.

Keywords: Hypertext, Computer conferencing, Document retrieval, Information management, Project control

Tim Berners-Lee: Information Management: A Proposal (CERN 1989) & WWW: First Page (1990)

World Wide Web

The WorldWideWeb (W3) is a wide-area hypermedia information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists , Policy , November's W3 news , Frequently Asked Questions .

What's out there?
: Pointers to the world's online information, subjects , W3 servers, etc.

Help
: on the browser you are using

Software Products
: A list of W3 project components and their current state. (e.g. Line Mode ,X11 Viola , NeXTStep , Servers , Tools , Mail robot , Library)

Technical
: Details of protocols, formats, program internals etc

Bibliography
: Paper documentation on W3 and references.

People
: A list of some people involved in the project.

History
: A summary of the history of the project.

How can I help ?
: If you would like to support the web..

Getting code
: Getting the code by anonymous FTP , etc.

ELIXIR Cancer Data Focus Group

www.elixir-europe.org

hCNV Community

- Community officially approved in February 2019
- Leadership

Christophe Béroud
(ELIXIR France)

David Salgado
(ELIXIR France)

Gary Saunders
(Human Data Coordinator, ELIXIR Hub)

Michael Baudis
(ELIXIR Switzerland)

Mission statement

Despite the fact that Copy Number Variations are the most prevalent genetic mutation type, identifying and interpreting them is still a major challenge. The ELIXIR human Copy Number Variation (hCNV) Community aims to implement processes to make the detection, annotation and interpretation of these variations easier

progenetix

Cancer CNV Profiles
Search Samples
Publication DB
Info
Beacon+
Baudisgroup @ UZH

Cancer genome data @ progenetix.org

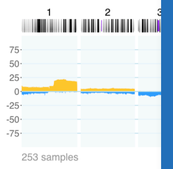
The Progenetix database
abnormalities (CNV / CNA

Data Content and Pr

The resource currently co
derived from array and ch
Genome or Whole Exome
data formats or are extrac

Besides genomic profiling
information which so far h

Original diagnoses are ma
cancer types, according t
classification, respectivel



For exploration of the resc
interest.

Additionally to genome pr
(currently **3962** articles)

Access, Maintenance and Contributions

The content of the progenetix resource is freely accessible for research an

The database & software are developed by the group of Michael Baudis at
Institute of Bioinformatics SIB.

Many previous members and external collaborators have contributed to da
Participation (features, data, comments) by volunteers are welcome.

progenetix.org
github.com/progenetix

Baudisgroup @ UZH
(Ni Ai)
Michael Baudis
(Haoyang Cai)
Paula Carrio Cordo
Bo Gao
Qingyao Huang
(Saumya Gupta)
(Nitin Kumar)
Rahel Paloots

Pierre-Henri Touissant

Beacon+ About Progenetix Help

Search

Allele Request Range Query Al

e is for copy number queries ("variantC
nges for start and end positions to captu

Cytoband(s)

Start
19000001-21975098

End Position
21967753-24000000

Bio-ontology
Select...

Beacon Query

next.progenetix.org/beacon-plus

Beacon API Leads
Jordi Rambla
Anthony Brooks
Juha Törnroos
Discovery WS
Michael Baudis (Beacon)
Marc Fiume (Networks)
ELIXIR
Gary Saunders
David Lloyd
Serena Scollen

{S}[B] and GA4GH
Melanie Courtot
Helen Parkinson
many more …

Unwatch 7 Star 1 Fork 2

Actions Wiki Security Insights

to file Add file Clone

46 commits 1 branch 0 tags

2 months ago
pty response 6 months ago
6 months ago
ject website last month
last month

tion build passi

The Beacon protocol defines an open standard for genomics data discovery, developed by members of the Global Alliance for Genomics & Health. It provides a framework for public web services responding to queries against genomic data collections, for instance from population based or disease specific genome repositories.

This repository contains the specification for the *v2* major version upgrade of the Beacon API. It is now (2020) under active development and has *not* seen a *stable* code release.

For further information, please follow the work here and consult the Beacon Project website.

About

GA4GH Beacon v2 specification.

ga4gh beacon openapi

Readme

Apache-2.0 License

Releases

No releases published
Create a new release

Publish your first package

Sabela de la Torre Pernas

Contributors 3

sdelatorrep sdelatorrep
mbaudis mbaudis
blankdots blankdots

beacon-project.io
github.com/ga4gh-beacon/

elixir