Progenetix & GA4GH

A cancer genomics resource built around and driving GA4GH standards

Michael Baudis | SIB - Novartis Symposion | 2021-09-22

progenetix

Genome screening at the core of “Personalised Health”

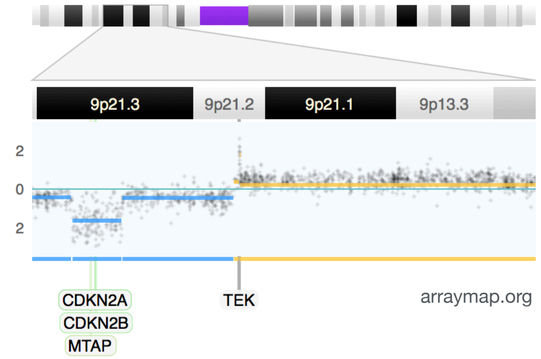
- **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**
- **Standardized data formats** and **exchange protocols** are needed to connect these resources throughout the world, for reciprocal, international **data sharing** and **biocuration** efforts
- Our work @ UZH:
  - ***cancer*** genome repositories
  - biocuration
  - protocols & formats

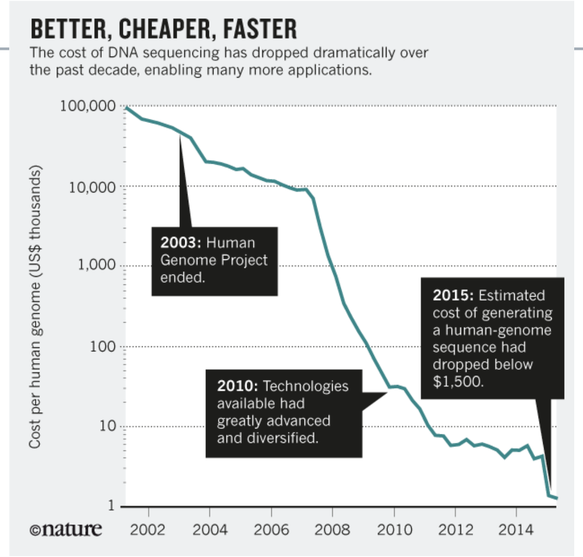
arrayMap

progenetix

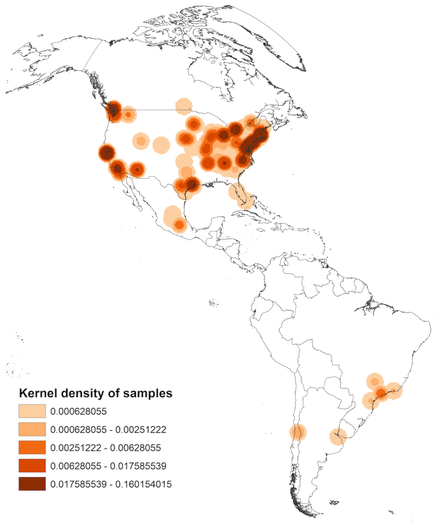
elixir

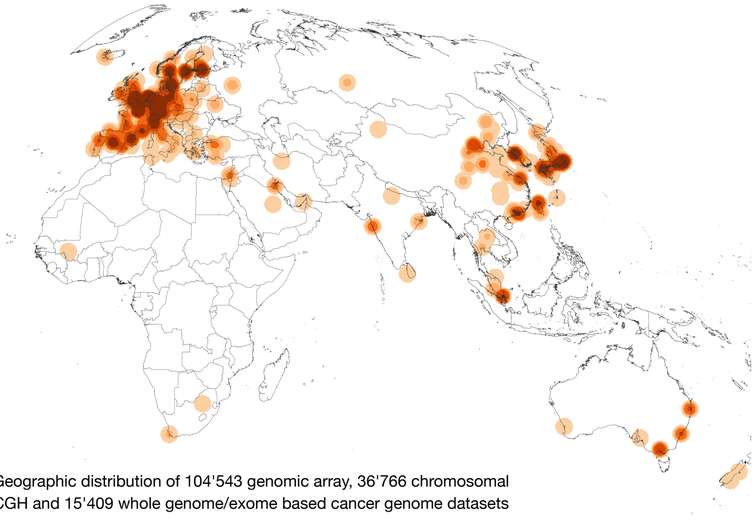
Global Alliance for Genomics & Health





The future of DNA sequencing. Eric D. Green, Edward M. Rubin & Maynard V. Olson. Nature; 11 October 2017 (News & Views)



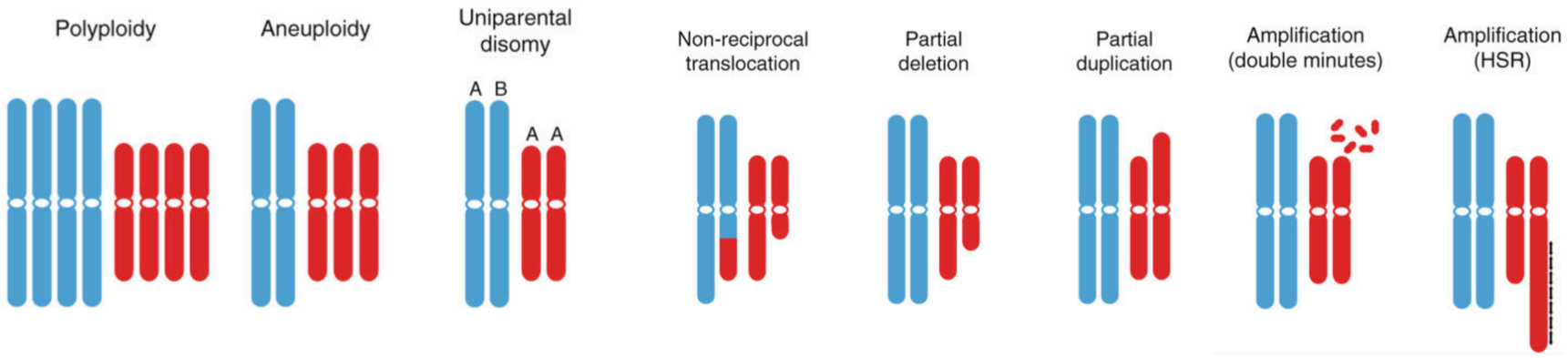


Geographic distribution of 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets

Types of genomic alterations in Cancer

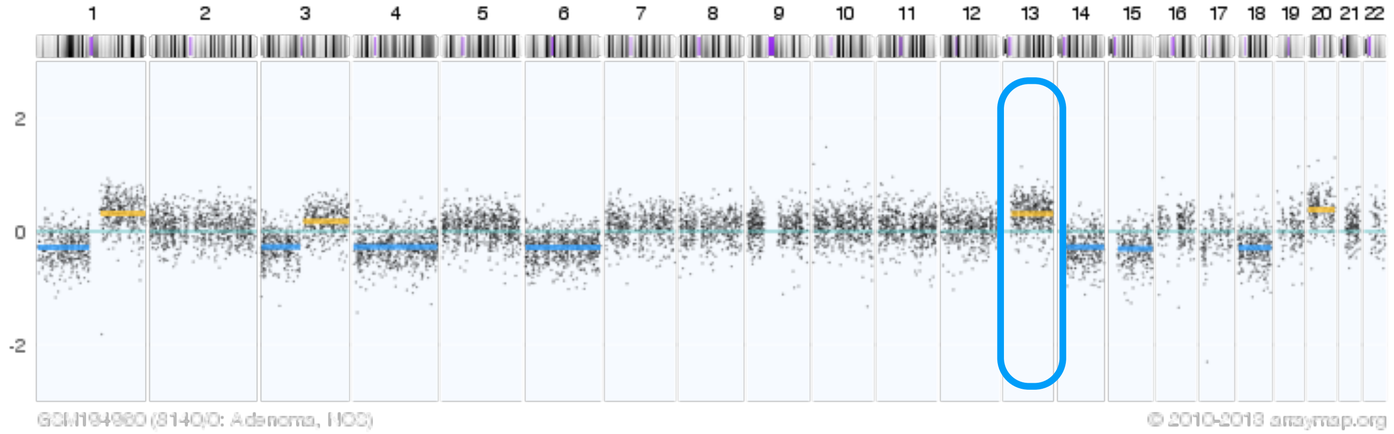
Imbalanced Chromosomal Changes: CNV

- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)

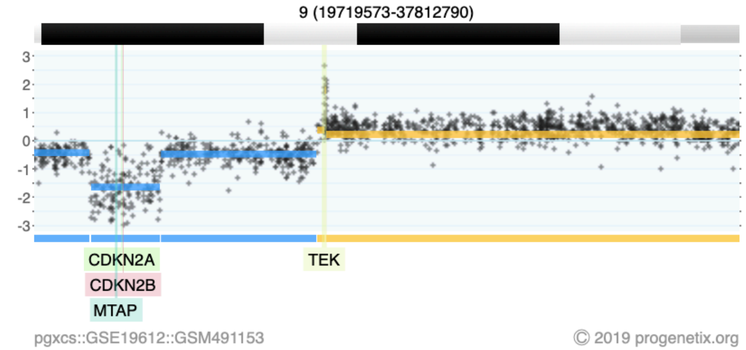


Grade et al., 2015 Recent Results Cancer Res

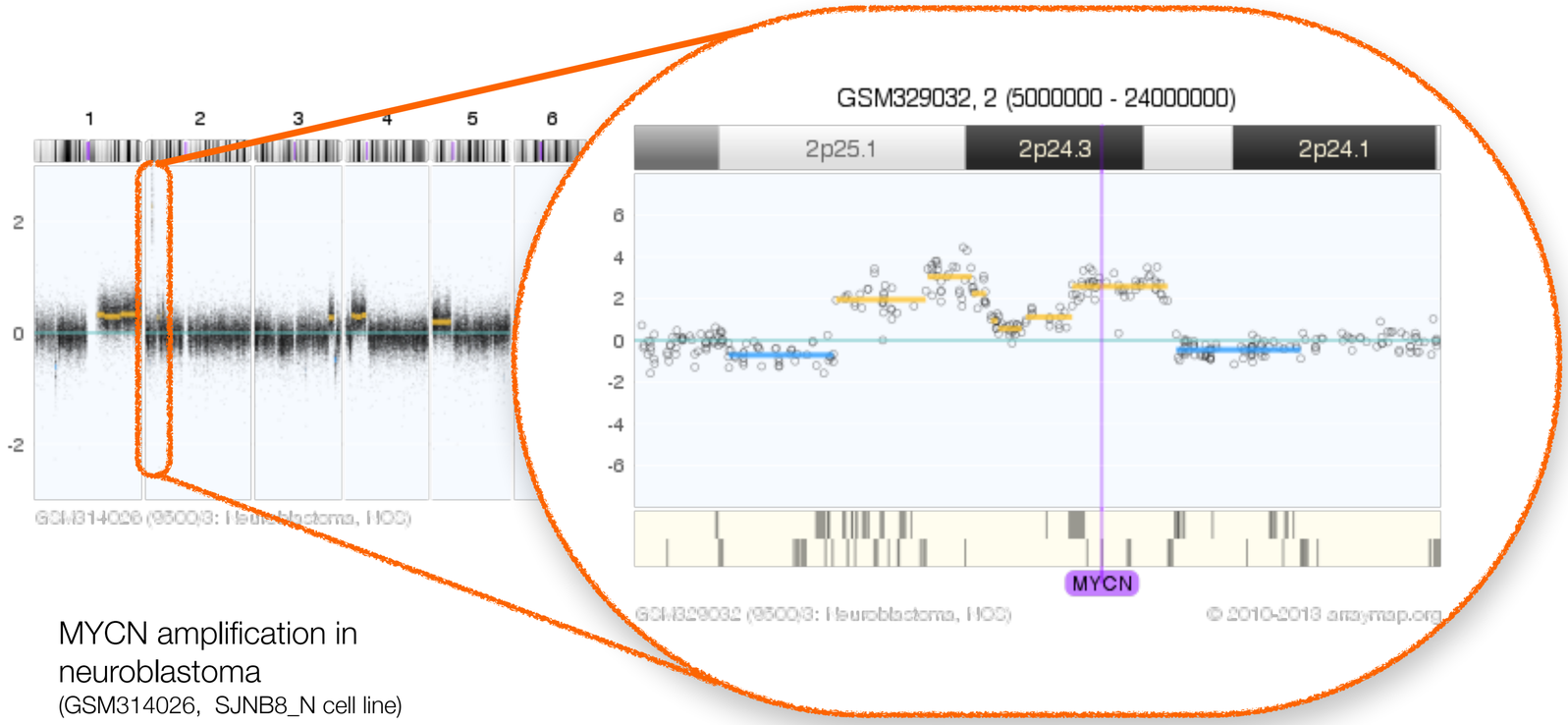
Array-based Detection of Copy Number Variations



Gain of chromosome arm 13q in colorectal carcinoma



2-event, homozygous deletion in a Glioblastoma

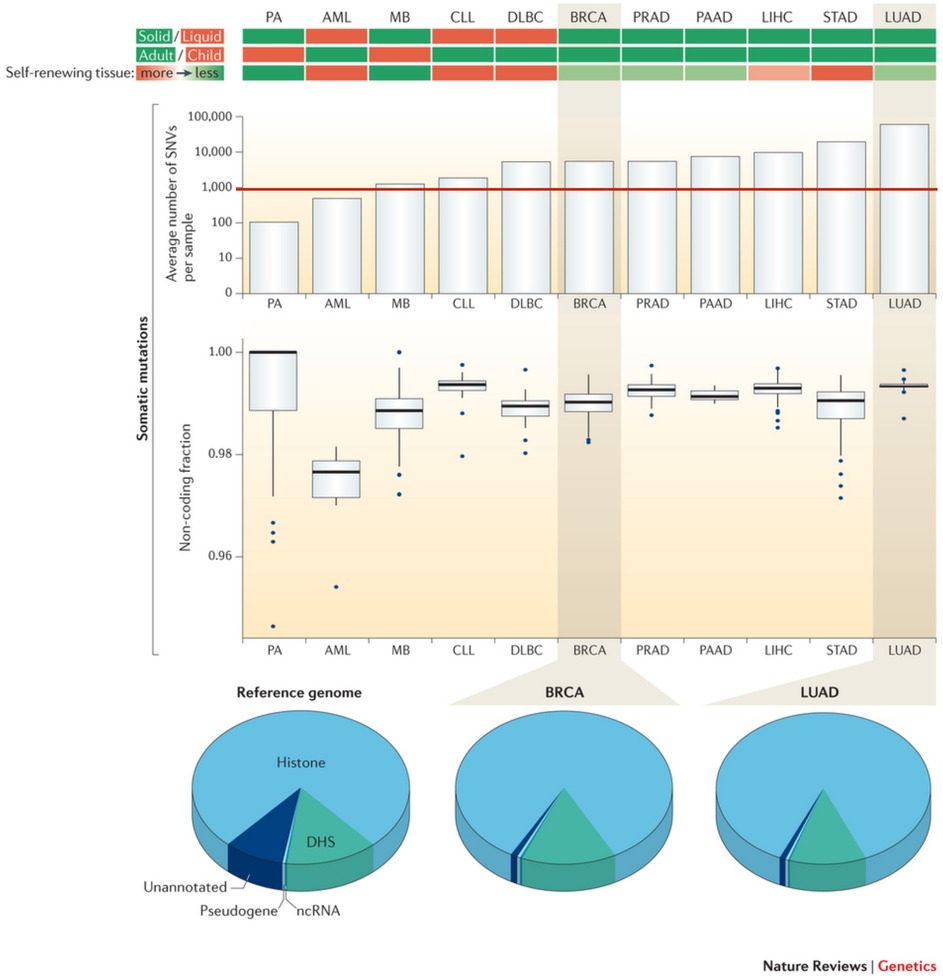


MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

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

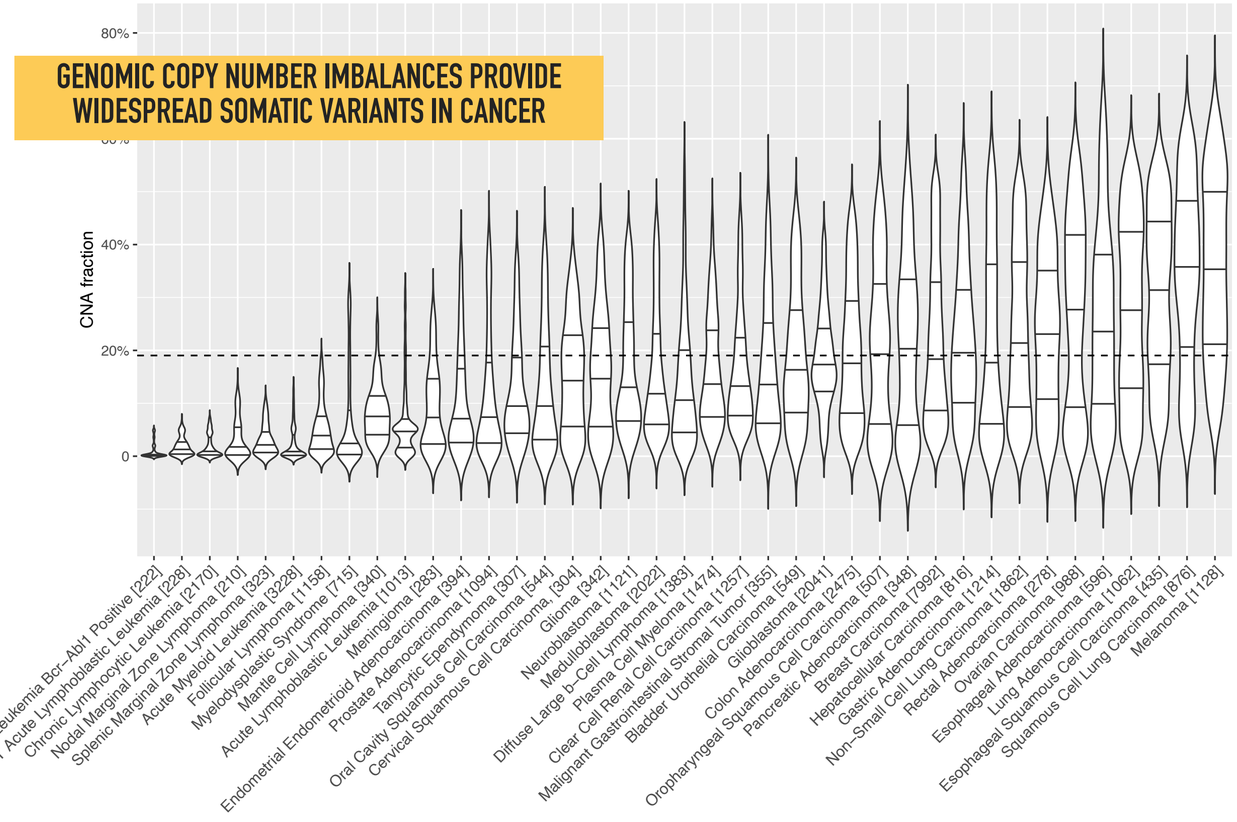
arrayMap

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)



On average ~19% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on 43654 cancer genomes from progenetix.org

History & Current State...

Origins & trajectory of the Progenetix Resource

progenetix

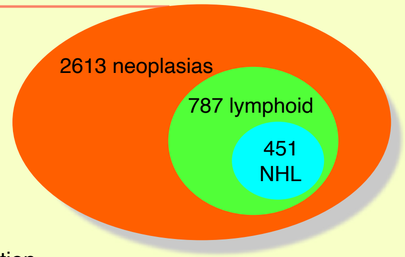
progenetix.net

[progenetix.net] online CGH database

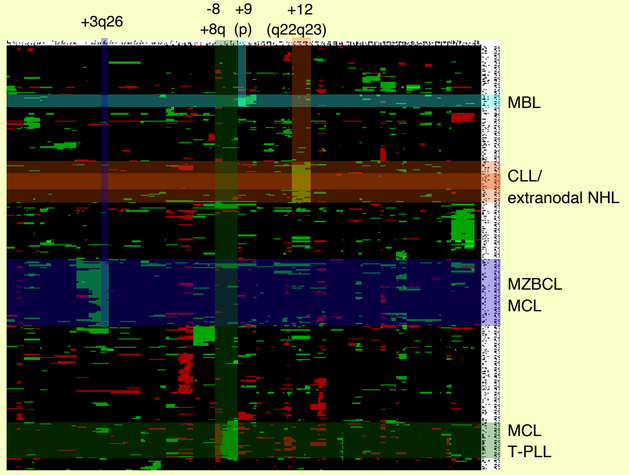
A compilation of published CGH analyses with reported case and band specific results

Currently includes
2613 cases from
92 publications

Automatic
conversion of
ISCN format to
aberration matrix
with 393 bands resolution



Clustering in 451 NHL: +12 in CLL and extranodal NHL



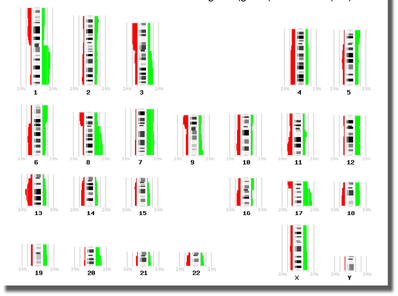
progenetix.net

Collection and Transformation of Chromosomal Imbalances in Human Neoplasias for Data Mining Procedures

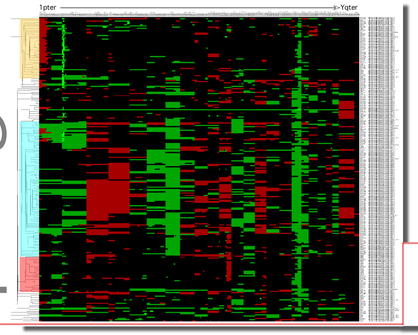
michael baudis, dept. of pathology, stanford university

Although the deciphering of the human genome has been pushed forward over the last years, little effort has been made to collect and integrate the treasure trove of clinical tumor cases analyzed by molecular-cytogenetic methods into current data schemes. Publicly announced at BCATS 2001, since then [progenetix.net] has been established as the largest public source of chromosomal imbalance data with band-specific resolution. Targets for the use of the data collection may be the description of prediction of oncogene and suppressor gene loci, identification of related loci for pathway creation, and especially the combination of the data with expression array experiments for filtering of relevant genes among the deregulated candidates.

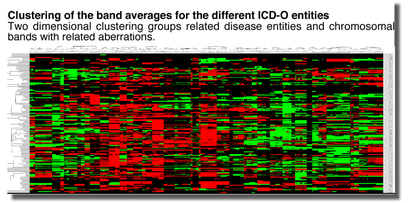
Chromosomal imbalances in 5478 clinical cases from 196 publications
Although not as prominent as in specific subgroups, this large collection shows the non-random distribution of chromosmal gains (green) and losses (red).



Material and Methods Chromosomal aberration data of more than 5478 cases from 196 publications describing results of Comparative Genomic Hybridization (CGH) experiments were collected. Minimal requirements were diagnosis of a malignant or benign neoplasia, analysis of clinical tumor samples and report of the analysis results on a case by case basis, resolved to the level of single chromosomal bands.Data was transformed from the diverse annotation formats to standardized ISCN “rev ish” nomenclature. For the transformation of the non-linear ISCN data to a two-dimensional matrix with code for the aberration status of each chromosomal band per case, a reverse pattern matching algorithm was developed in Perl. Graphical representations and cluster images are generated for all different subsets (Publications, ICD-O-3 entities, meta-groups) and presented on the progenetix.net website.

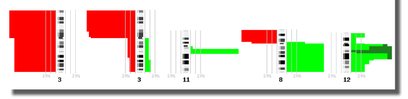


Clustering of the band averages for the different ICD-O entities
Two dimensional clustering groups related disease entities and chromosomal bands with related aberrations.



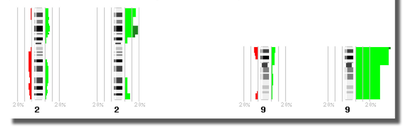
Results Out of 4896 tumor samples, 3862 (79%) showed chromosomal imbalances by CGH. The average per band probability was 4.5% for a loss (max. 12.9% at 13q21) and 6.5% for a gain (max. 15.6% at 8q23). Differences between neoplastic entities showed in the average frequency and distribution pattern of imbalanced chromosomal regions. Tumor subsets (10 or more cases) with the strongest hot spots for losses were small cell lung carcinomas (ave. 23.3% with max. 96.2% at 3p14p26) and pheochromocytomas (ave. 10.9% with max. 92.7% at 3p); prominent gain maxima were found in pure high grade infiltrating duct carcinomas of the breast (ave. 5.9% with max. 95.7% at 11q13), T-PLL (ave. 4.7% with max. 81.8% for whole 8q) and dedifferentiated liposarcomas (ave. 10.4% with max. 81.8% at 12q13), among others.
By cluster analysis, different combinations of chromosomal hot spot regions could be shown to occur in tumors subsummized in the same diagnostic entity; the example of neuroblastomas is shown.

Examples of hotspots of genomic imbalance
SCLC, pheochromocytoma, high grade DCIS, T-PLL, dedifferentiated liposarcoma



Conclusion So far, progenetix.net project was able to:
1. collect a large dataset of genomic aberration data generated through a molecular-cytogenetic screening technique (CGH)
2. develop the software tools to transform those data to a meta format compatible to commonly used genomic interval descriptions
3. produce graphical and numerical output from those data for hot spot detection and statistical analysis.
For future approaches, the data collection will be valuable for filtering data from expression array experiments for relevant genes, and possibly for the description of common and divergent genetic pathways in the oncogenetic process of different tumor entities.The transformed raw data of the progenetix.net collection is avaible for research purposes over the website.

Distinction of histologically related through their chromosomal aberration pattern
Amplification of the REL locus on 2p16 and gain of 9p(ter) distinguishes primary mediastinal B-cell lymphomas (PMBL, right) from diffuse large cell lymphomas (DLCL, left). The distinction may have clinical implications



Identification of different aberration patterns in Neuroblastoma (289 cases)
N-Myc (2p25) amplification is the hallmark of a subgroup, showing only consistant loss of the terminal portion 1p. Other groups are defined by the loss of parts of 11q, or a “chromosomal instability” phenotype. Gains on 17q are a common feature of all groups.
Those patterns may be combined with gene-level information to reconstruct the different pathways leading to malignant transformation.

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,[3] Colin Collins,[1] Masao Tomita,[2] Joe Gray,[1] and Frederic M. Waldman[1]

Progenetix: Data Scopes

Biomedical and procedural "Meta"data types

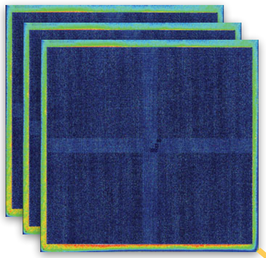
- Diagnostic classification
  - mapping text-based cancer diagnoses to standard classification systems
- Provenance data
  - store identifier-based pointers
  - geographic attribution (individual, biosample, experiment)
- Clinical information
  - **core set** of typical cancer study values:
    - ➡ stage, grade, followup time, survival status, genomic sex, age at diagnosis
  - balance between annotation effort and expected usability

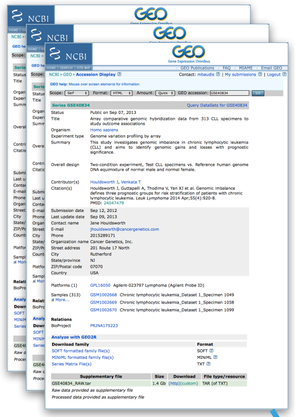
progenetix

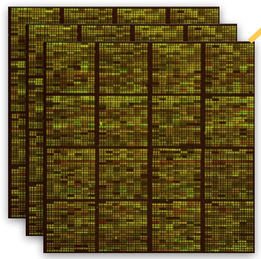
DATA PIPELINE

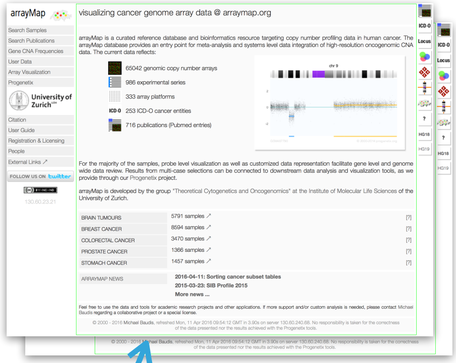
BIOCURATION

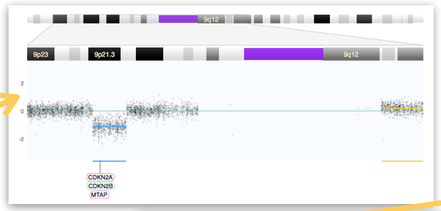
BIOINFORMATICS

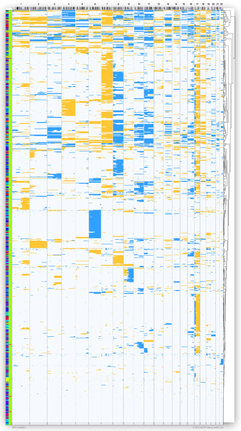


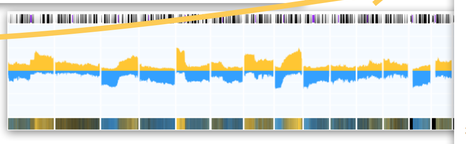


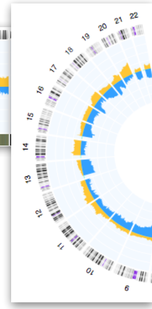


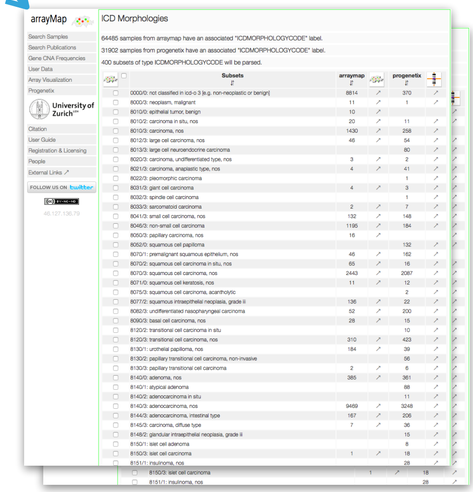












arrayMap

progenetix

Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

```
^SAMPLE = GSM174832
!Sample_title = 9194
!Sample_geo_accession = GSM174832
!Sample_status = Public on May 01 2007
!Sample_submission_date = Mar 13 2007
!Sample_last_update_date = Mar 13 2007
!Sample_type = genomic
!Sample_channel_count = 1
!Sample_source_name_ch1 = Bone marrow with 96% blasts
!Sample_organism_ch1 = Homo sapiens
!Sample_taxid_ch1 = 9606
!Sample_characteristics_ch1 = Immunotype: common ALL; Age: 9.2 yrs; Gender: F
!Sample_molecule_ch1 = genomic DNA
!Sample_extract_protocol_ch1 = QiaAmp purification kit (Qiagen)
!Sample_label_ch1 = biotin
!Sample_label_protocol_ch1 = Biotinylated DNA was prepared according to the standa
manual 701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix).
!Sample_hyb_protocol = Hybridizations were performed according to the standard Affy
701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix) using an Affymetrix hyb
!Sample_scan_protocol = Scanning performed according to the standard Affymetrix pro
or 100k assay manual 701684 Rev.3, Affymetrix) using an Affymetrix scanner 3000.
!Sample_description = primary ALL diagnosis sample
!Sample_data_processing = copy number detection using CNAG2.0 software (http://www.
!Sample_platform_id = GPL3718
!Sample_contact_name = Roland,P.,Kuiper
!Sample_contact_email = r.kuiper@antrg.umcn.nl, e.verwiel@antrg.umcn.nl
!Sample_contact_phone = +31243610868
!Sample_contact_fax = +31243668752
!Sample_contact_department = Human Genetics
!Sample_contact_institute = Radboud University Nijmegen Medical Centre
!Sample_contact_address = Geert Grooteplein 10
!Sample_contact_city = Nijmegen
!Sample_contact_zip/postal_code = 6525GA
!Sample_contact_country = Netherlands
!Sample_supplementary_file = ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM174nnn/GSM174832/suppl/GSM174832.CEL.gz
!Sample_supplementary_file = ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM174nnn/GSM174832/suppl/GSM174832.CHP.gz
!Sample_series_id = GSE7255
```

```perl
foreach (grep { ! /characteristics_ch\d/ } @in) {
    my ($key, $value) = split(' = ', $_);
    $key =~ s/[^\w]/_/g;
    if ($key =~ /submission_date/i) {
        $sample->{ YEAR } = $value;
        $sample->{ YEAR } =~ s/^.*?(\d\d\d\d)$/\1/;
    }
```

```perl
$mkey->{ samplekey } = 'AGE';
$mkey->{ matches } = [ qw( age )];

( $mkey->{ retv }, $mkey->{ retk } ) = _grepmeta( $mkey, $meta );

if ($mkey->{ retv } =~ /^(.+?)$/) {
    if ($mkey->{ retv } =~ /month/i) {
        $mkey->{ retk } .= '_months';
        $mkey->{ retv } =~ s/[^\d\.]//g;
    }

    $sample->{ $mkey->{ samplekey } } = _normNumber($mkey->{ retv });
    if ($mkey->{ retk } =~ /month/i) { $sample->{ $mkey->{ samplekey } } /= 12 }
    if ($sample->{ $mkey->{ samplekey } } == 0) { $sample->{ $mkey->{ samplekey } } = 'NA' }
    $sample->{ $mkey->{ samplekey } } = sprintf "%.2f", $sample->{ $mkey->{ samplekey } };
}
```

progenetix

Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

```
^SAMPLE = GSM286922
!Sample_title = 481 - mAdbID:75320
!Sample_geo_accession = GSM286922
!Sample_status = Public on Sep 04 2008
!Sample_submission_date = May 06 2008
!Sample_last_update_date = Nov 26 2008
!Sample_type = genomic
!Sample_channel_count = 2
!Sample_source_name_ch1 = Normal Lymphocytes
!Sample_organism_ch1 = Homo sapiens
!Sample_taxid_ch1 = 9606
!Sample_characteristics_ch1 = Tissue: lymphocytes
!Sample_molecule_ch1 = genomic DNA
!Sample_extract_protocol_ch1 = Sample DNA Extraction Protocol
!Sample_extract_protocol_ch1 = Other: The DNA was isolated by Qiagen DNe
!Sample_label_ch1 = cy5
!Sample_label_protocol_ch1 = NimbleGen Cy5 Sample Labeling Protocol
!Sample_label_protocol_ch1 = Other: Proprietary protocol information available at http://www.nimblegen.com/technology/index.html
!Sample_source_name_ch2 = 481
!Sample_organism_ch2 = Homo sapiens
!Sample_taxid_ch2 = 9606
!Sample_characteristics_ch2 = Gender: male
!Sample_characteristics_ch2 = Age: 49
!Sample_characteristics_ch2 = Tissue: lymph node
!Sample_characteristics_ch2 = Disease state: Lymphoma
!Sample_characteristics_ch2 = Individual: 481
!Sample_characteristics_ch2 = Clinical info: Submitting diagnosis: DLBCL
!Sample_characteristics_ch2 = Clinical info: Final microarray diagnosis: ABC DLBCL
!Sample_characteristics_ch2 = Clinical info: Follow up status: ALIVE
!Sample_characteristics_ch2 = Clinical info: Follow up years: 10.75
!Sample_characteristics_ch2 = Clinical info: Chemotherapy: CHOP-Like Regimen
!Sample_characteristics_ch2 = Clinical info: ECOG performance status: 2
!Sample_characteristics_ch2 = Clinical info: Stage: 4
!Sample_characteristics_ch2 = Clinical info: LDH ratio: 0.82
!Sample_characteristics_ch2 = Clinical info: Number of extranodal sites: 1
```

Channel 1 is normal -> Cave value swap!

Gender or "chromosomal sex"?

context indicates years, but if it would be a medulloblastoma...

This survival status annotation not known to parser...

```
$mkey->{ samplekey } = 'DEATH';
$mkey->{ matches } = [ (
    'death',
    'dead ',
    'vital_status',
    'dead_alive',
    'alive_dead',
) ];

( $mkey->{ retv }, $mkey->{ retk } ) = _grepmeta( $mkey, $meta );

if ($mkey->{ retv } =~ /^(.+?)$/) {
    $sample->{ $mkey->{ samplekey } } = _normDeath($mkey->{ retv }) }
```

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.

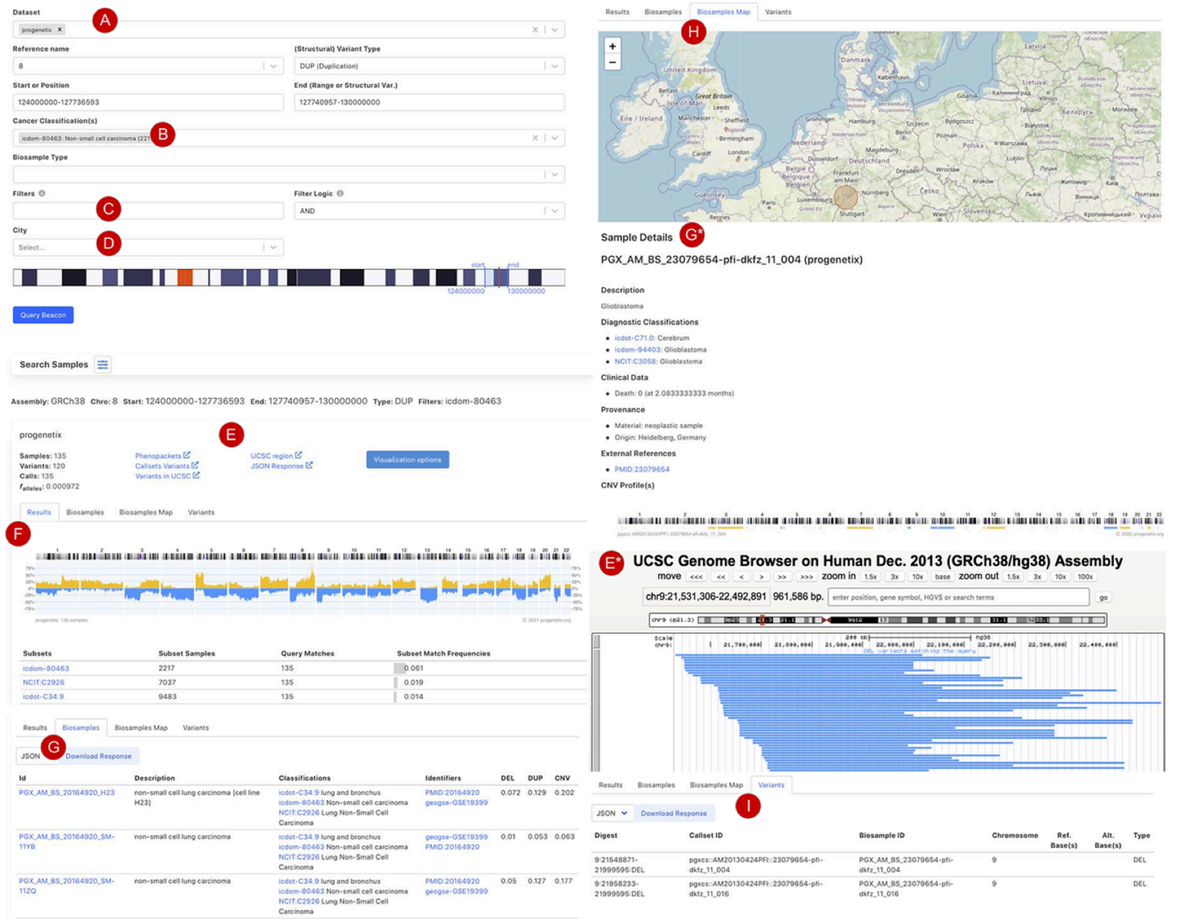


Figure 3. Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range This example queries for a continuous, focal duplication covering the complete MYC gene's coding region with < = 6 Mb in size. A: Filter for dataset; B: filter for cancer classification (NCIt and 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.

Progenetix in 2021

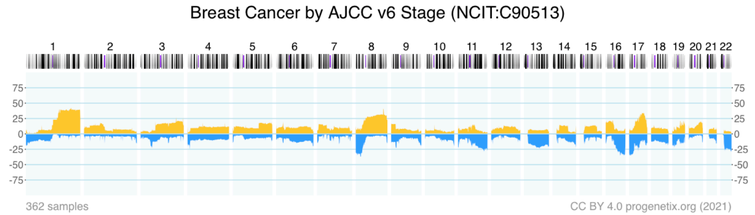
Cancer Genomics Reference Resource

- >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
- publication database and code mapping services

progenetix

Cancer genome data @ progenetix.org

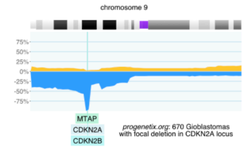
The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* from currently **139448** samples.



Download SVG | Go to NCIT:C90513 | Download CNV Frequencies

Example for aggregated CNV data in 362 samples in Breast Cancer by AJCC v6 Stage.
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 **810** 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

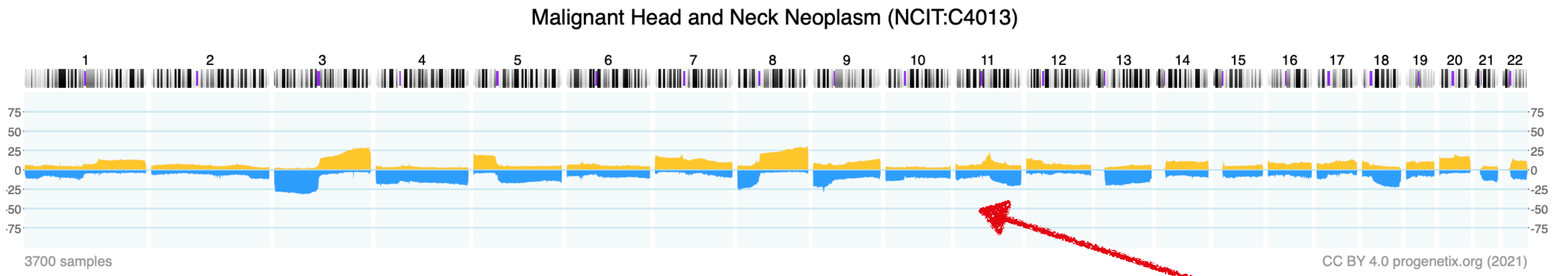
Through the [Publications] page Progenetix provides **4025** annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.

© 2000 - 2021 Progenetix Cancer Genomics Information Resource by the Computational Oncogenomics Group at the University of Zurich and the Swiss Institute of Bioinformatics **SIB** 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.

Progenetix

Cancer Type CNA Data

- hierarchical aggregation of cancer samples
- pre-computed CNA frequencies for fast overview
- sample retrieval for custom grouping, visualization



progenetix

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

Cancer Types

The cancer samples in Progenetix are mapped to several classification systems. For each of the classes, aggregated date is available by clicking the code. Additionally, a selection of the corresponding samples can be initiated by clicking the sample number or selecting one or more classes through the checkboxes.

Sample selection follows a hierarchical system in which samples matching the child terms of a selected class are included in the response.

Cancer Classification: NCIT Cancer Core

Filter subsets ... Hierarchy Depth: 2 levels

No Selection

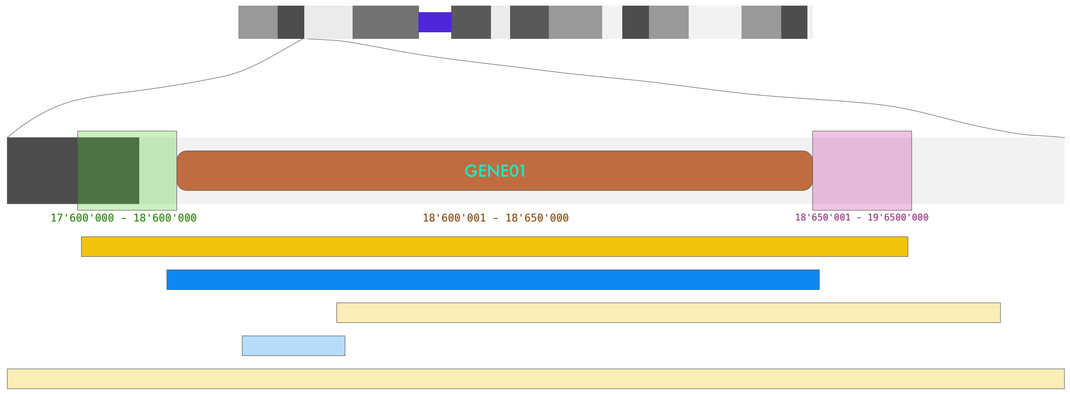
- NCIT:C3262: Neoplasm (116232 samples)
  - NCIT:C3263: Neoplasm by Site (110429 samples)
    - NCIT:C156482: Genitourinary System Neoplasm (16534 samples)
    - NCIT:C2910: Breast Neoplasm (15556 samples)
      - NCIT:C27939: Lobular Neoplasia (92 samples)
      - NCIT:C36083: Intraductal Breast Neoplasm (275 samples)
        - NCIT:C27942: Intraductal Proliferative Lesion of the Breast (270 samples)
          - NCIT:C2924: Ductal Breast Carcinoma In Situ (270 samples)
        - NCIT:C36090: Intraductal Papillary Breast Neoplasm (5 samples)
      - NCIT:C40405: Breast Fibroepithelial Neoplasm (41 samples)
      - NCIT:C40406: Breast Soft Tissue Neoplasm (3 samples)
      - NCIT:C5206: Papillary Breast Neoplasm (15 samples)
      - NCIT:C9335: Malignant Breast Neoplasm (15528 samples)
    - NCIT:C3010: Endocrine Neoplasm (3521 samples)
    - NCIT:C3030: Eye Neoplasm (280 samples)
    - NCIT:C3052: Digestive System Neoplasm (15285 samples)
    - NCIT:C3077: Head and Neck Neoplasm (3961 samples)
      - NCIT:C3260: Neck Neoplasm (2565 samples)
      - NCIT:C3361: Salivary Gland Neoplasm (274 samples)
      - NCIT:C3375: Skull Neoplasm (161 samples)
      - NCIT:C4013: Malignant Head and Neck Neoplasm (3700 samples)
        - NCIT:C133187: Mucosal Melanoma of the Head and Neck (13 samples)
      - NCIT:C155790: Malignant Skull Neoplasm (17 samples)
      - NCIT:C164198: Head and Neck Sarcoma (1 sample)
      - NCIT:C170467: Metastatic Malignant Head and Neck Neoplasm (15 samples)

Progenetix in 2021

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)



Bold: Matched Variants

Shaded: Unmatched

DEL (Copy Number Loss) DUP (Copy Number Gain)

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

Progenetix in 2021

Query Results and Variant Frequencies

- genomic variant + metadata queries provide relative result counts / frequencies for mapped entities (NCIt, ICD-O ...)
  - disease-specific CNA event scores
- representation of genome -wide CNA frequency profiles / context
- link-outs to download options, subset visualization, sample exploration ...

progenetix

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 Modify Query

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

progenetix

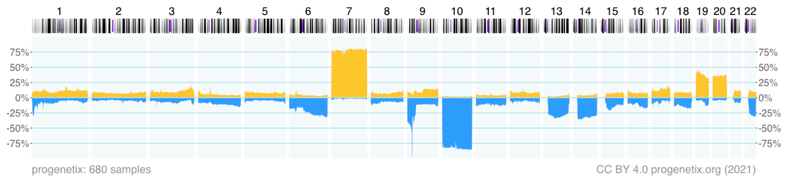
Samples: 678
Variants: 297
Calls: 687

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	674	0.157
NCIT:C3058	4375	674	0.154
UBERON:0016525	14	2	0.143
icdot-C71.1	14	2	0.143
UBERON:0000955	7284	661	0.091
icdot-C71.9	7282	661	0.091
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 2021

Query Results and Variant Frequencies

Results | Biosamples | Biosamples Map | Variants

JSON | Download Variants

Int. ID	Digest	Biosample	Chr.	Ref. Base(s)	Alt. Base(s)	Type
						DEL
						DEL
						DEL
						DEL
						DEL
						DEL
						DEL
						DEL
pgxvar-5bab579a727983b2e00d129e	9:21837542-22437694:DEL	pgxbs-kftvgkbx	9			DEL
pgxvar-5bab579a727983b2e00d14be	9:21837542-21999595:DEL	pgxbs-kftvgkc2	9			DEL

« ‹ › » Page **1 of 69**

Results | Biosamples | Biosamples Map | Variants

JSON | Download Biosamples

Biosample Id	Dx Classifications	Identifiers	CNV Fraction
pgxbs-kftvgk8h	NCIT:C3058 Glioblastoma icdom-94403 Glioblastoma icdot-C71.0 Cerebrum UBERON:0001869 cerebral hemisphere	PMID:23079654 Sturm D, Witt H et al. (2012): Hotspot mutations in H3F3A and IDH1...	0.26
pgxbs-kftvgk90	NCIT:C3058 Glioblastoma icdom-94403 Glioblastoma icdot-C71.0 Cerebrum	PMID:23079654 Sturm D, Witt H et al. (2012): Hotspot mutations in H3F3A and IDH1...	0.304

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 | Modify Query

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

progenetix

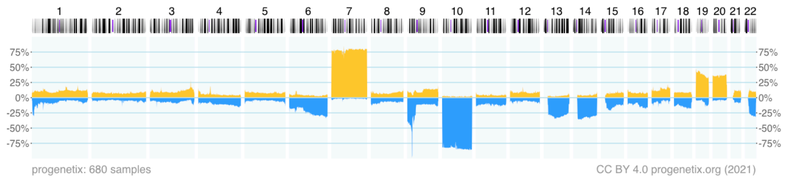
Samples: 678
Variants: 297
Calls: 687

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

UCSC region
JSON Response

Visualization options

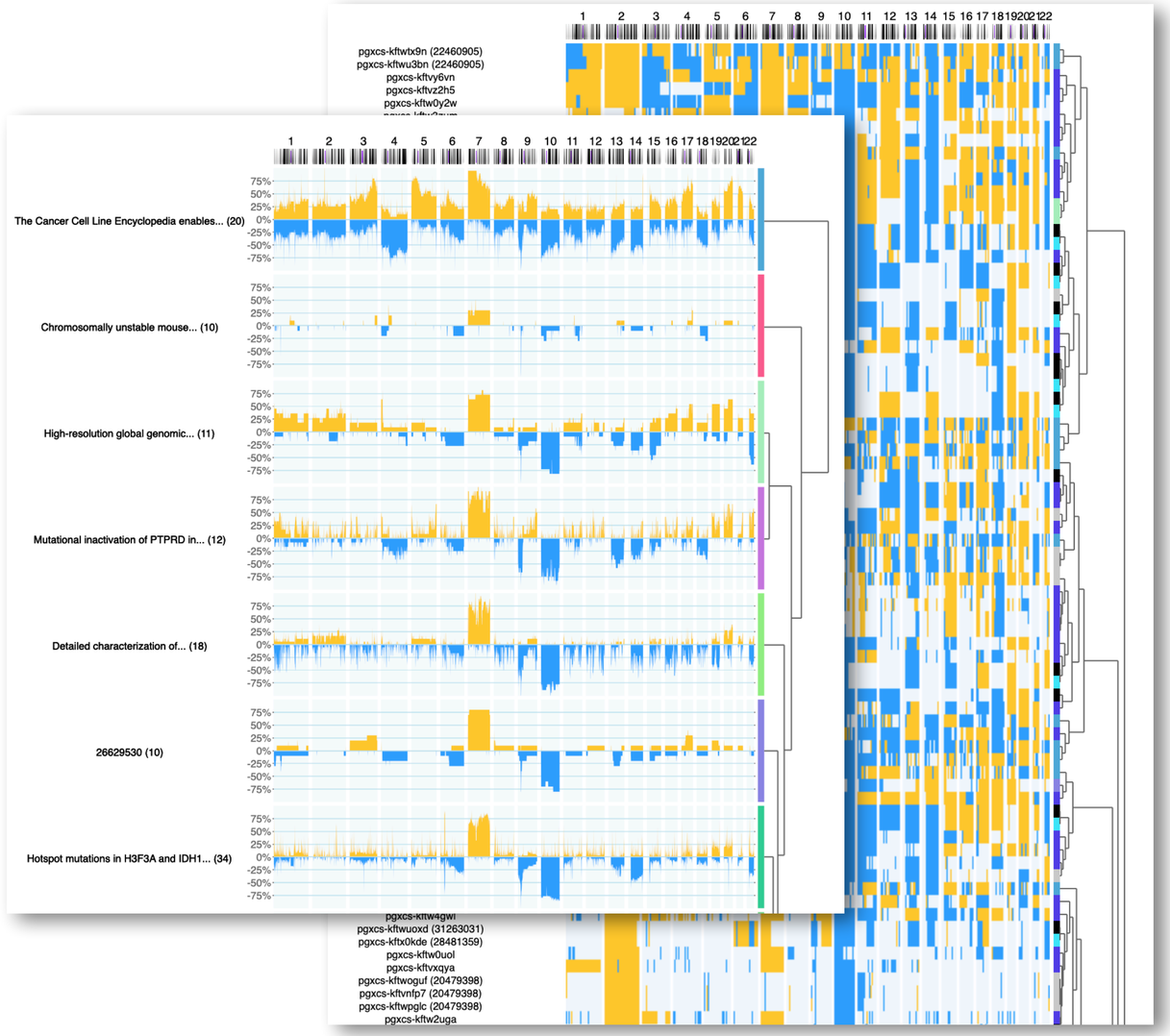
Results | Biosamples | Biosamples Map | Variants



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	674	0.157
NCIT:C3058	4375	674	0.154
UBERON:0016525	14	2	0.143
icdot-C71.1	14	2	0.143
UBERON:0000955	7284	661	0.091
icdot-C71.9	7282	661	0.091
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 2021

Query Result Visualizations



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 Modify Query

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

progenetix

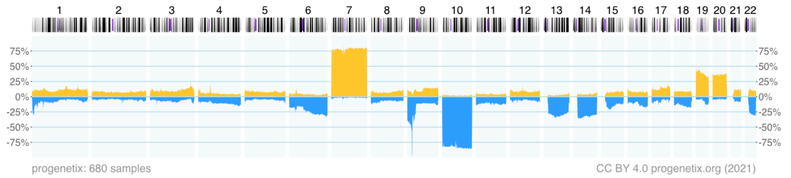
Samples: 678
Variants: 297
Calls: 687

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

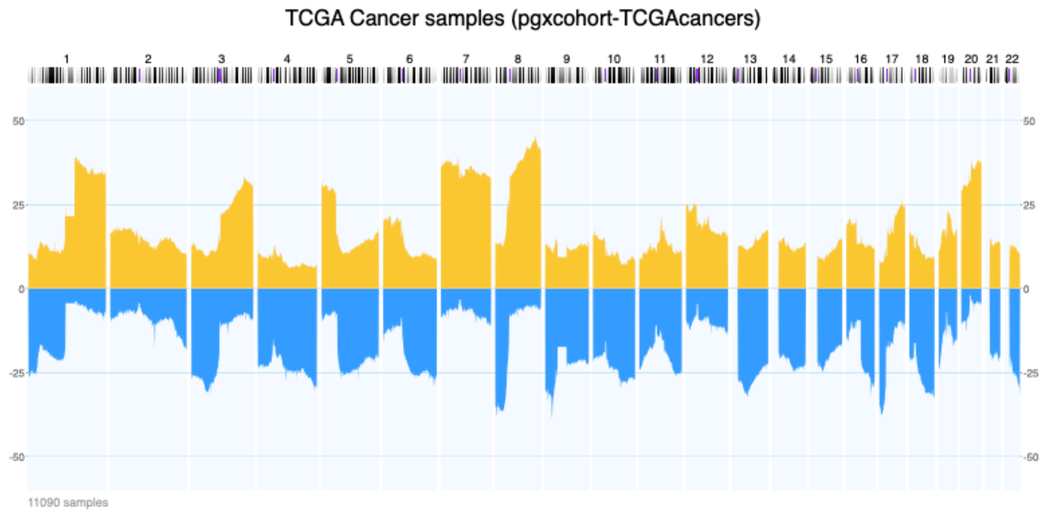


Matched Subset Codes	Subset Samples	Matched Samples	Subset Match Frequencies
UBERON:0002021	4	1	0.250
icdot-C71.4	4	1	0.250
icdom-94403	4291	674	0.157
NCIT:C3058	4375	674	0.154
UBERON:0016525	14	2	0.143
icdot-C71.1	14	2	0.143
UBERON:0000955	7284	661	0.091
icdot-C71.9	7282	661	0.091
icdom-94423	84	4	0.048
NCIT:C3796	84	4	0.048
UBERON:0001869	1712	14	0.008
icdot-C71.0	1712	14	0.008

Progenetix API

Data & Plots

- "all" of the data can be accessed using API calls
- segmented CNV data in .pgxsg (columnar) and JSON format
- histograms for disease, study or cohort from precomputed frequencies - live generated as SVG for embedding with plot options

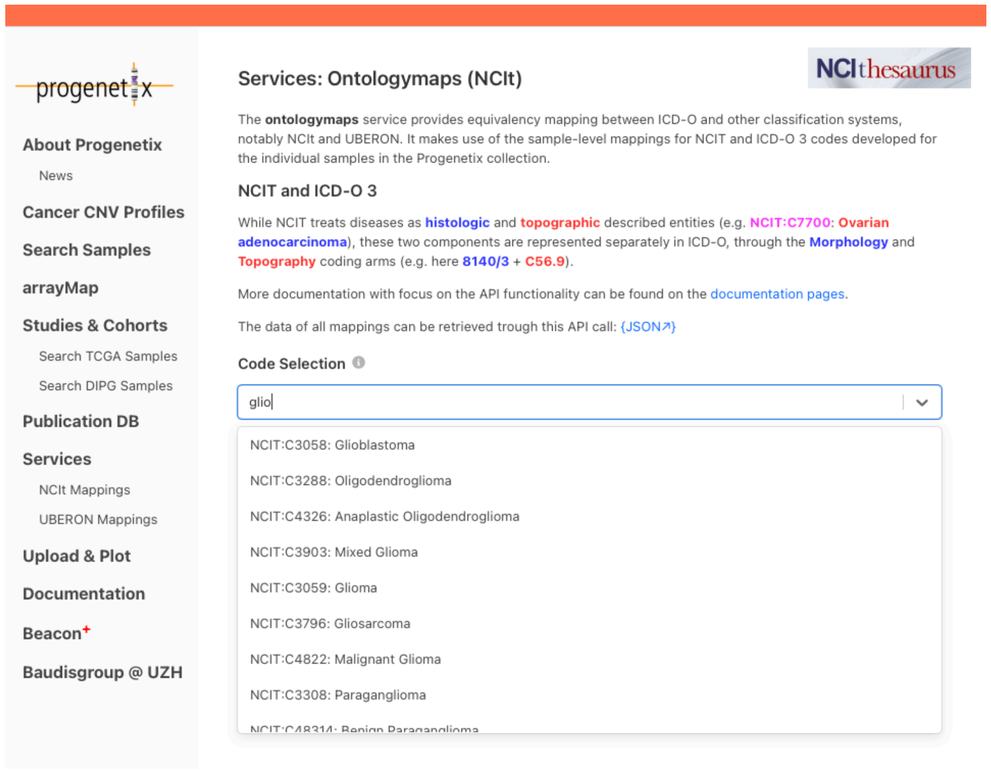


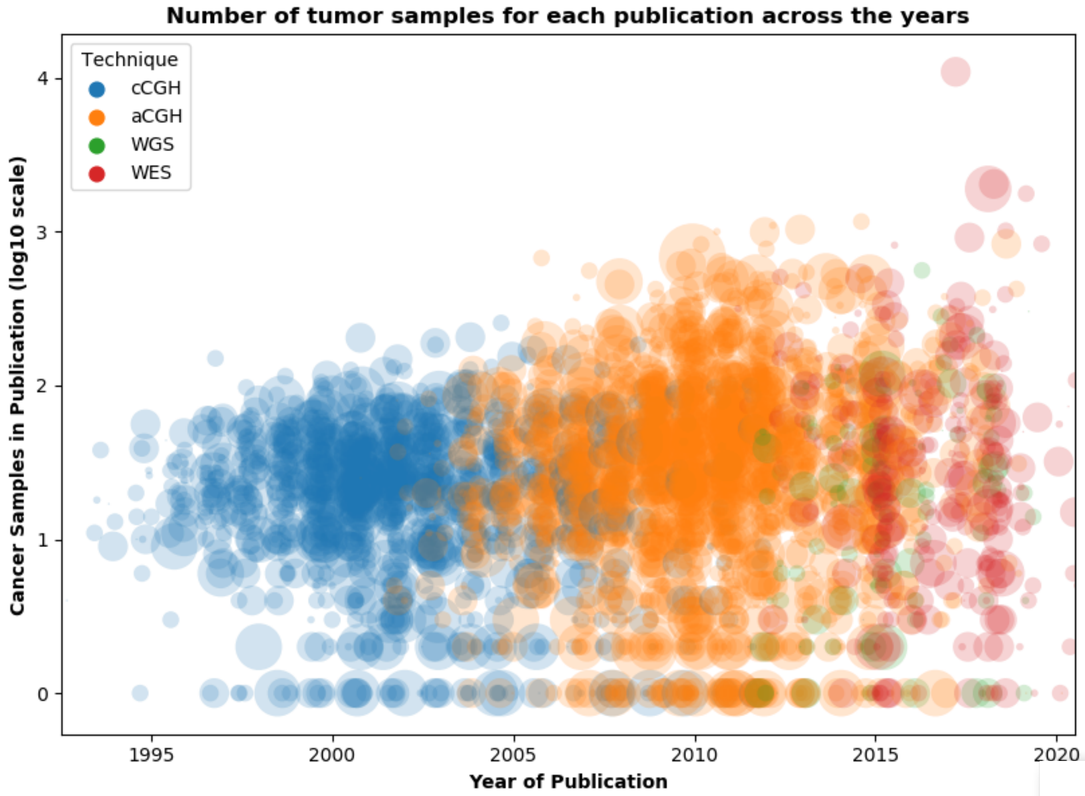
*https://progenetix.org/cgi/PGX/cgi/collationPlots.cgi?datasetIds=progenetix&id=**pgxcohort-TCGAcancers**&-size_plotimage_w_px=**800**&-size_plotarea_h_px=**300**&-value_plot_y_max=**60***

Progenetix

Services, Documentation...

- services e.g. for disease code translation (NCIt <=> ICD-O; UBERON ...)
- API & documentation "progressing" ...





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

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

Progenetix Publication Collection

The current page lists articles describing whole genome screening (WGS, WES, aCGH, cCGH) experiments in cancer, registered in the Progenetix publication collection. For each publication the table indicates the numbers of samples analysed with a given technology and if sample profiles are available in Progenetix.

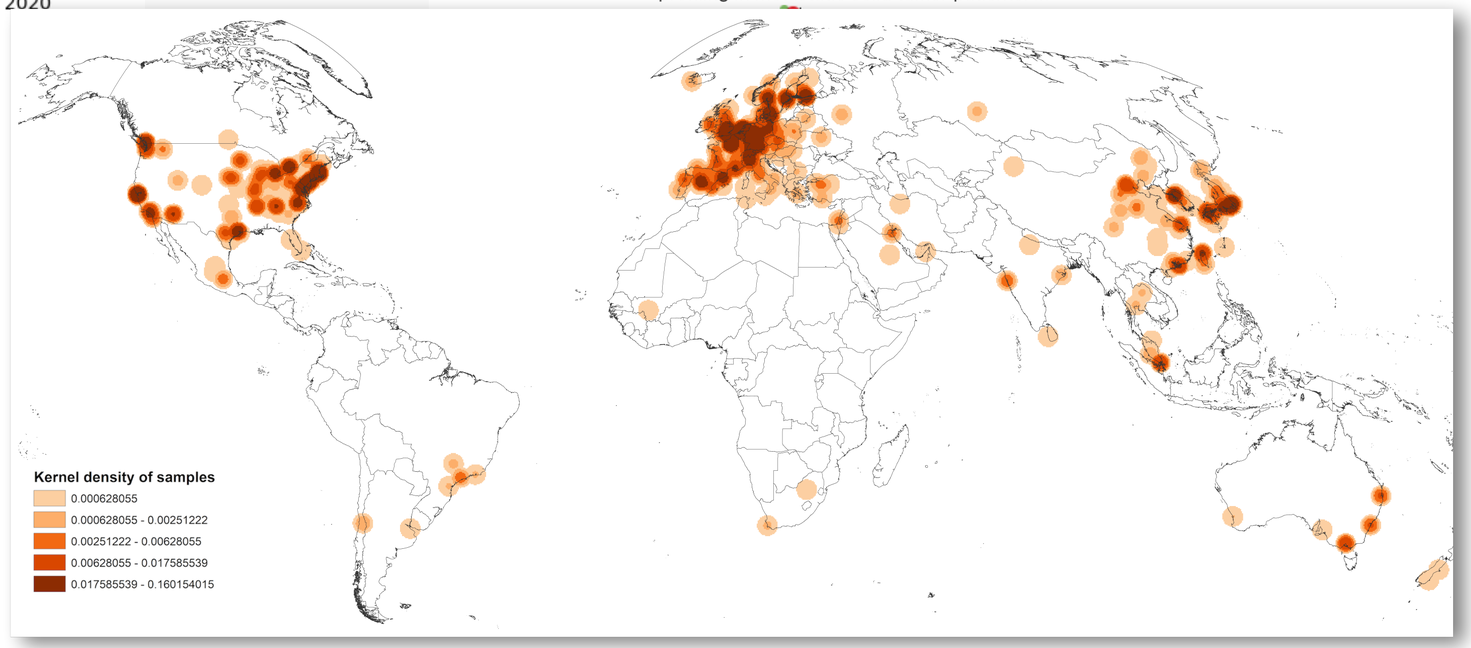
Please contact us to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation.

Filter

City

Type to search...

Publications (3324)		Samples				
id	Publication	cCGH	aCGH	WES	WGS	pgx
PMID:34103027	Peng G, Chai H, Ji W, Lu Y, Wu S et al. (2021) Correlating genomic copy number alterations with clinicopathologic findings in 75 cases of ... BMC Med Genomics	0	79	0	0	0
PMID:34059130	Tsui DWY, Cheng ML, Shady M, Yang JL et al. (2021) Tumor fraction-guided cell-free DNA profiling in metastatic solid tumor patients. ...	0	0	5	113	0



Progenetix and GA4GH Beacon

Implementation driven development of a GA4GH standard

progenetix

GENOMICS

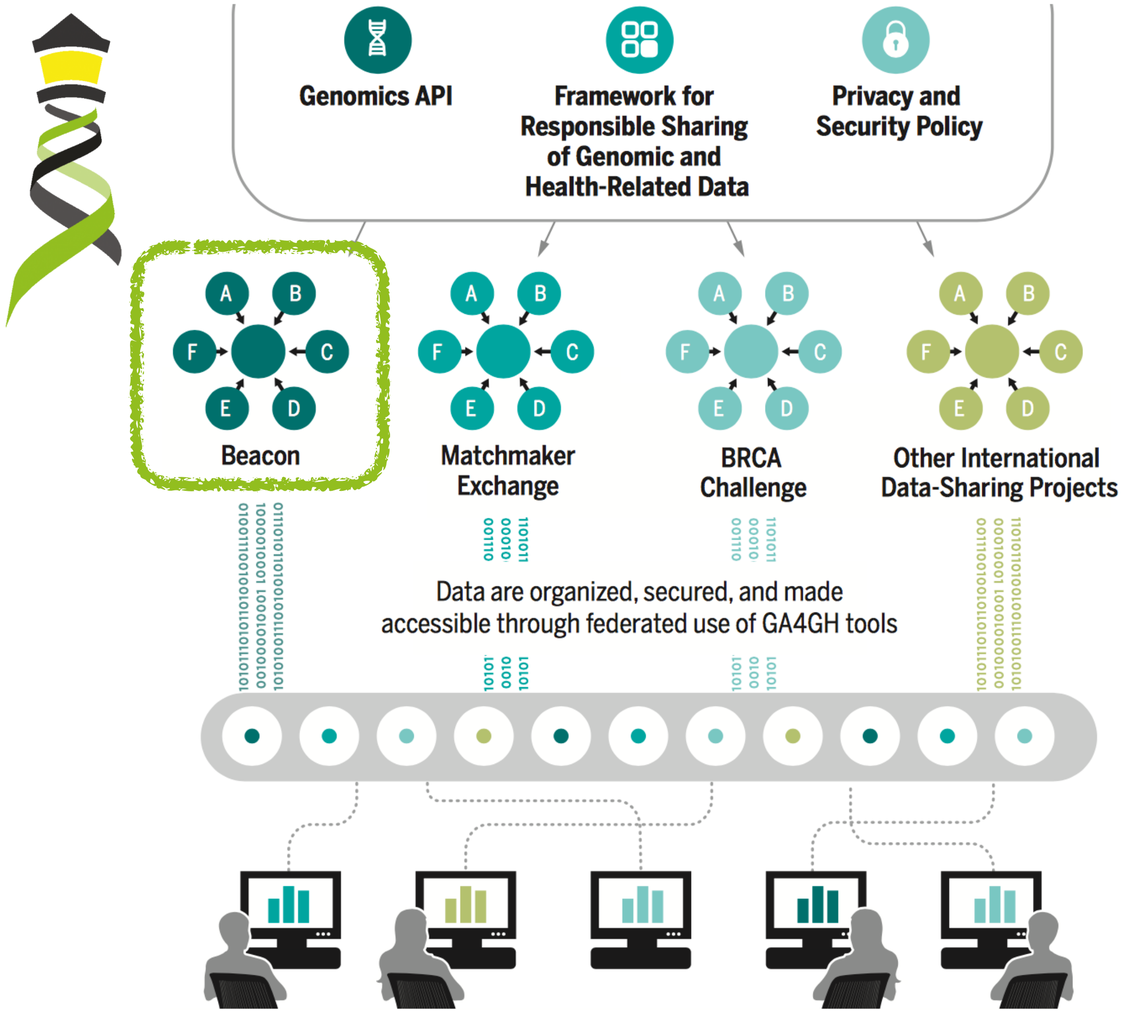
A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

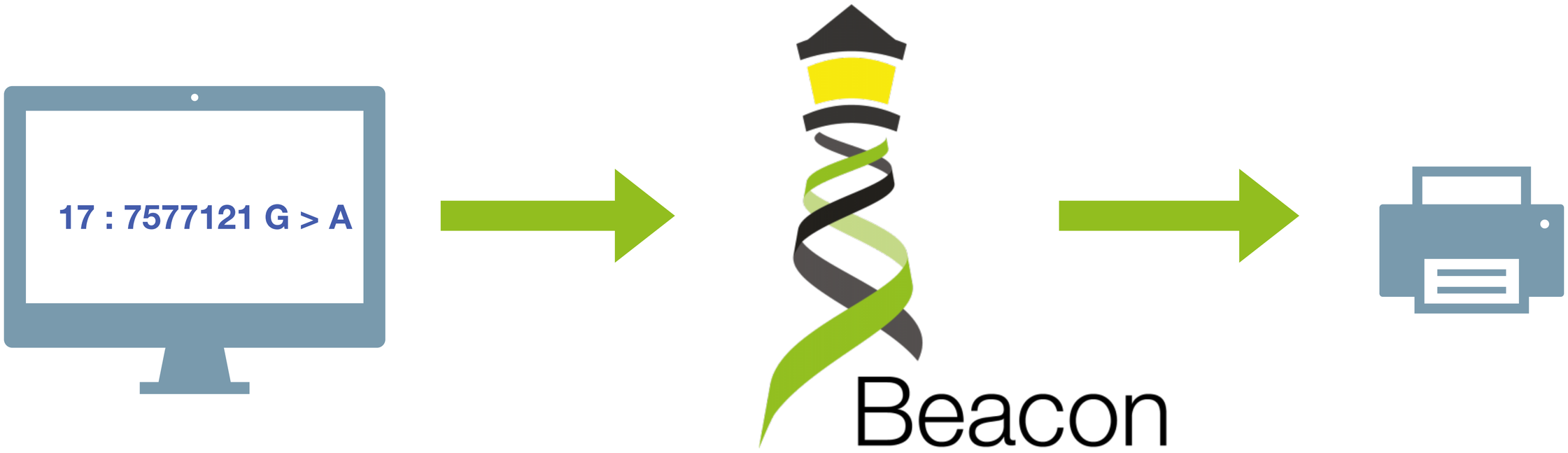
SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

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

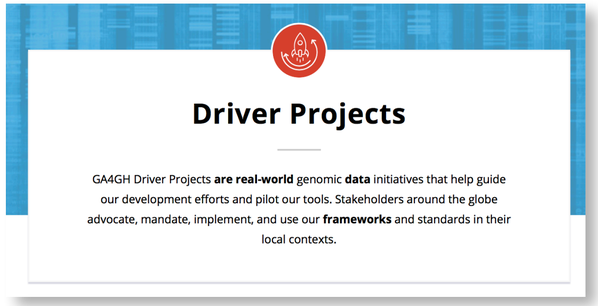


A ***Beacon*** answers a query for a specific genome variant against individual or aggregate genome collections

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)

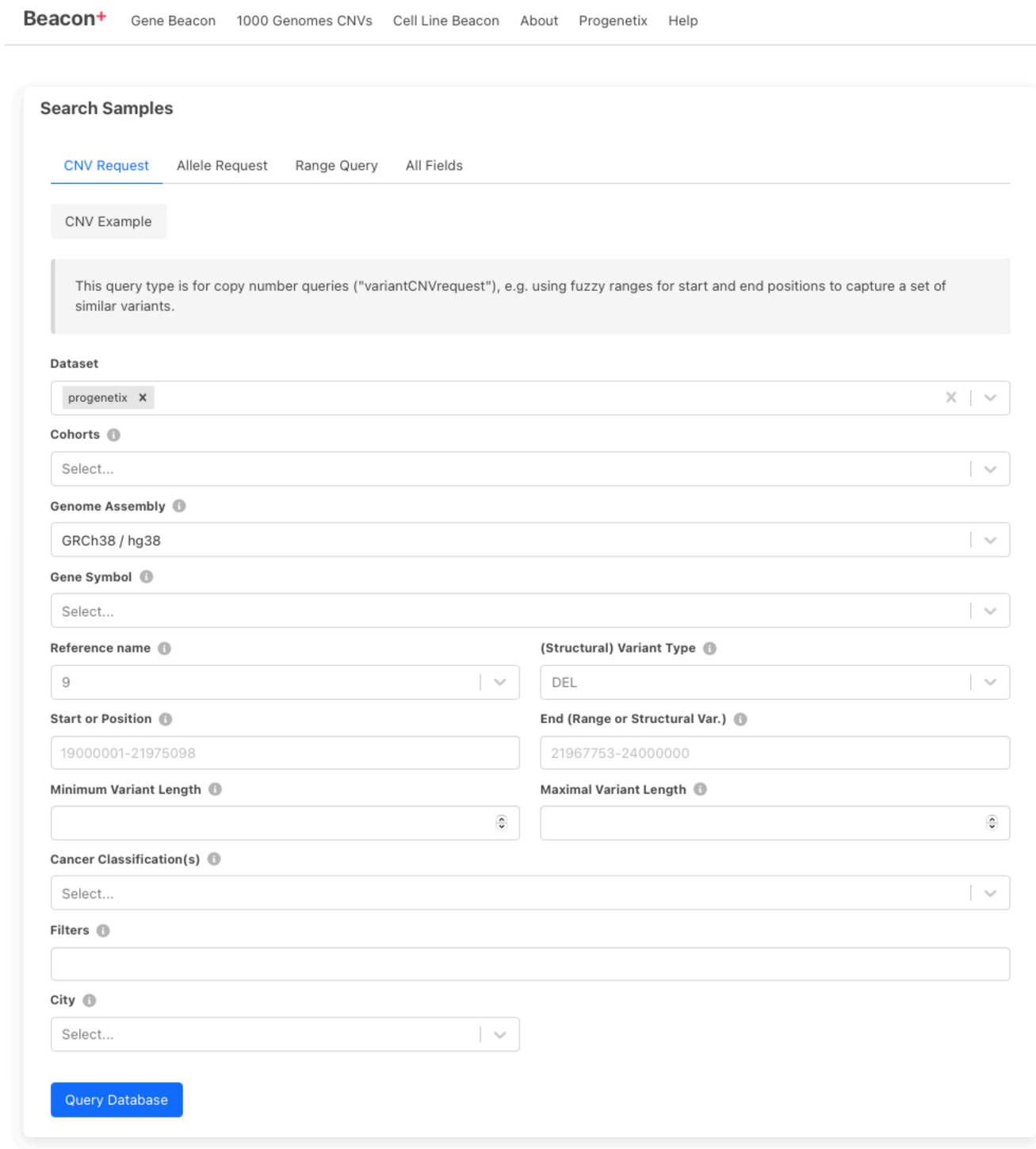
Global Alliance for Genomics & Health

Beacon+ by Progenetix

From Beacon Query to Explorative Analyses of CNV Patterns

- Since 2016 the Progenetix resource has been used to model options for Beacon development
  - 138334 individual samples from 698 cancer types
- 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



Beacon v2 Requests

POSTing Queries

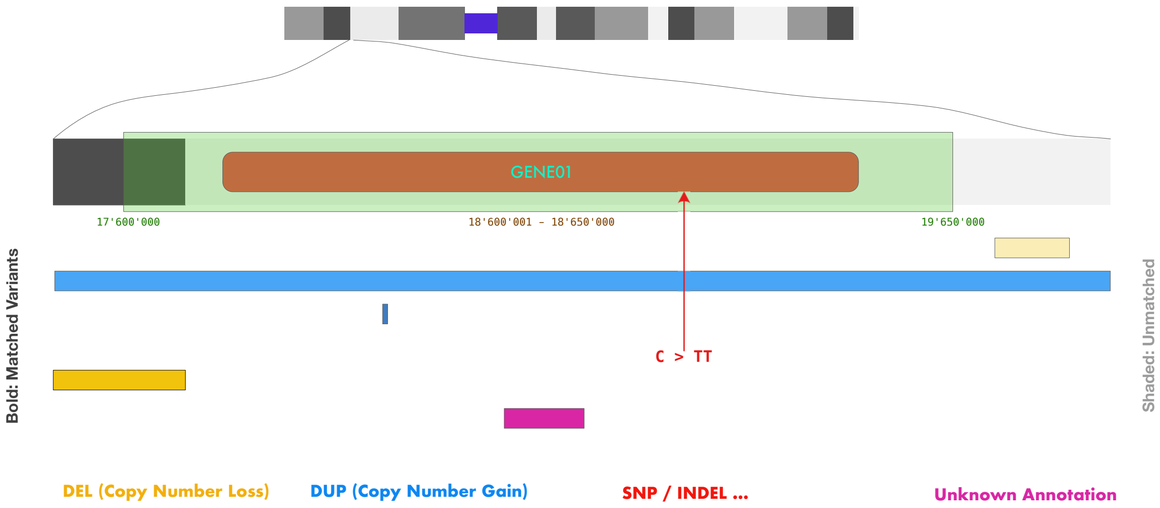
- Beacon v2 supports a mix of dedicated endpoints with REST paths
- POST requests using JSON query documents
- final syntax for core parameters still in testing stages

```
{
    "$schema":"beaconRequestBody.json",
    "meta": {
        "apiVersion": "2.0",
        "requestedSchemas": [
            {
                "entityType": "individual",
                "schema": "https://progenetix.org/services/schemas/Phenopacket/"
            }
        ]
    },
    "query": {
        "requestParameters": {
            "datasets": {
                "datasetIds": ["progenetix"]
            }
        },
        "filterLogic": "OR"
    },
    "pagination": {
        "skip": 0,
        "limit": 10
    },
    "filters": [
        { "id": "NCIT:C4536" },
        { "id": "NCIT:C95597" },
        { "id": "NCIT:C7712" }
    ]
}
```

Beacon v2: Extended Variant Queries

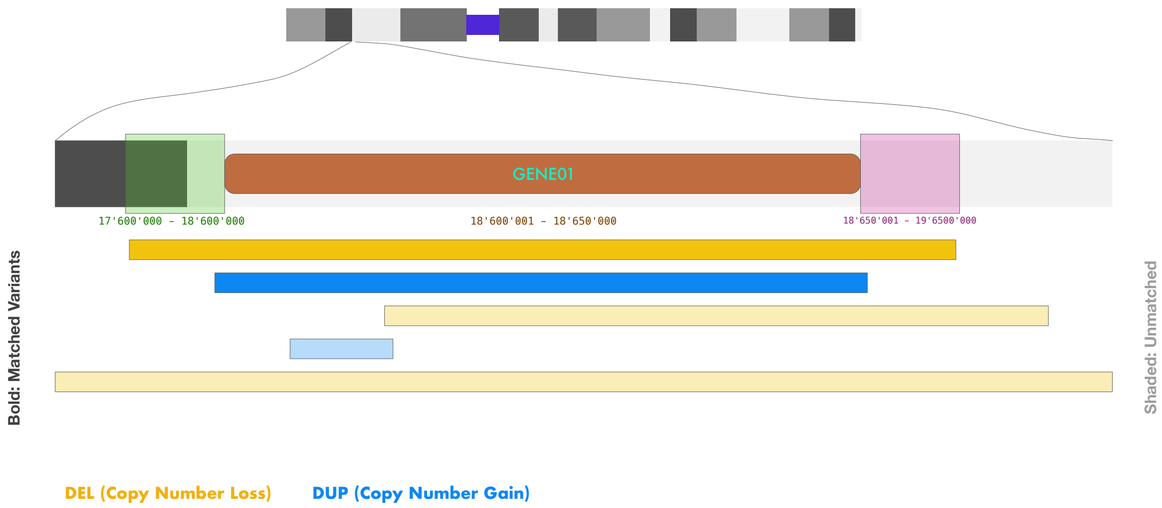
Range and Bracket queries enable positional wildcards and fuzziness

Genome Range Query
(matching variants in a region)



- Genome Range Queries provide a way to "fish" for variants overlapping an indicated region, e.g. the CDR of a gene of interest
- Additional parameters (e.g. variant type, regerence or alternate bases) limit the scope of the responses
- new Beacon v2 size parameters to limit structural variants (e.g. "focal" CNVs)

Genome Bracket Query
(full match)



- Genome Bracket Queries allow to search for structural variants with start and end positions falling into defined sequence ranges
- allows to query any contiguous genomic variant (and in principle also can step in for range queries)
- typical use case is e.g the query for variants such as duplications covering the whole CDR of a gene, while limiting the allowed start or end regions

Beacon v2 Filters

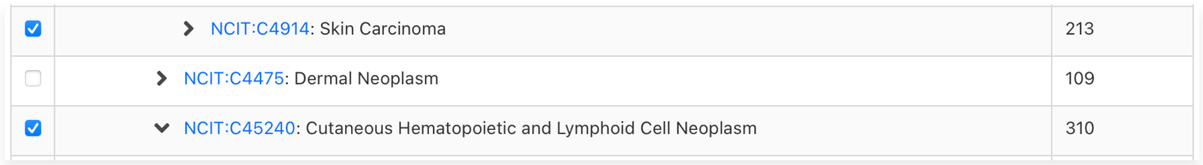
Example: Use of hierarchical classification systems (here NCIt neoplasm core)

- 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

Beacon+ specific: Multiple term selection with OR logic



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**

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 ...)

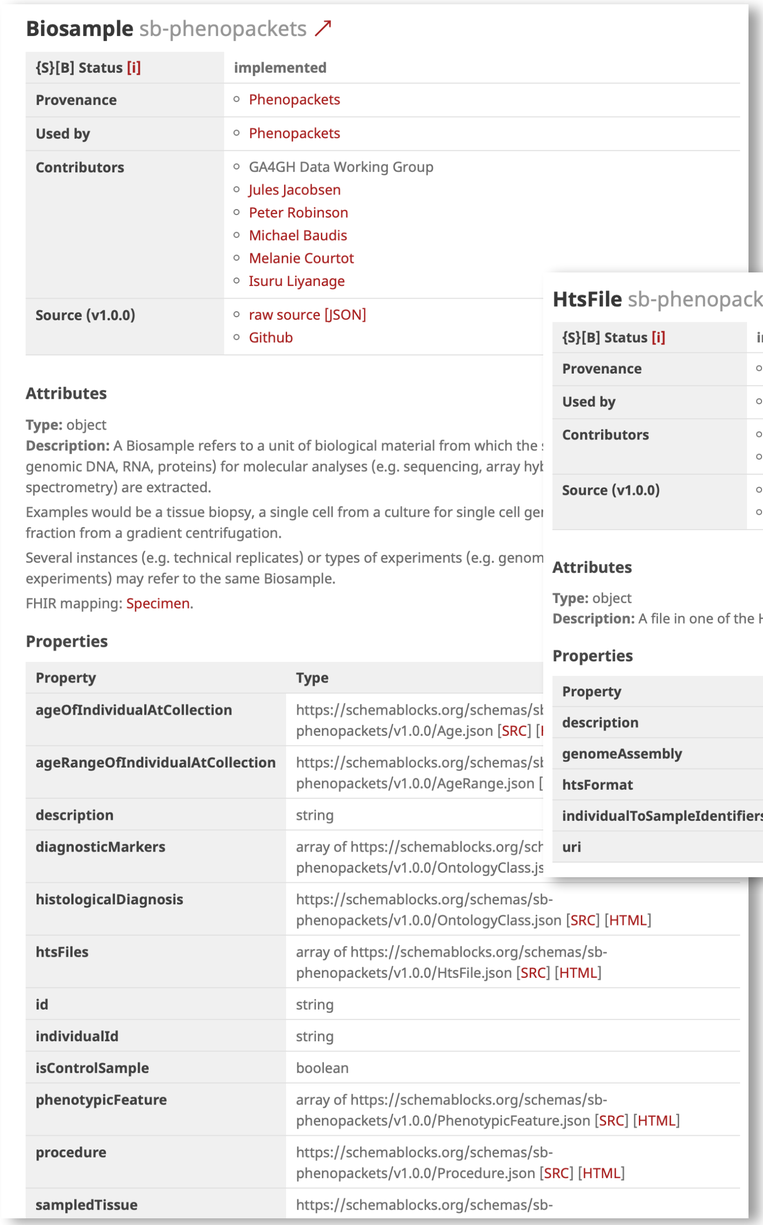
```
"data_use_conditions" : {
  "label" : "no restriction",
  "id" : "DUO:0000004"
},

"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
- 2021: going forward through integration with GA4GH TASC efforts, towards "standards library"



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
genomic DNA, RNA, proteins) for molecular analyses (e.g. sequencing, array hyb
spectrometry) are extracted.

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

Several instances (e.g. technical replicates) or types of experiments (e.g. genom
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] [H
ageRangeOfIndividualAtCollection	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/AgeRange.json [
description	string
diagnosticMarkers	array of https://schemablocks.org/sch phenopackets/v1.0.0/OntologyClass.js
histologicalDiagnosis	https://schemablocks.org/schemas/sb-phenopackets/v1.0.0/OntologyClass.json [SRC] [HTML]
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-

HtsFile sb-phenopackets ↗

{S}[B] Status [i]	implemented
Provenance	○ Phenopackets
Used by	○ Phenopackets
Contributors	○ Jules Jacobsen ○ Peter Robinson
Source (v1.0.0)	○ raw source [JSON] ○ Github

Attributes

Type: object
Description: A file in one of the HTS formats (https://samtools.github.io/hts-specs)

Properties

Property	Type
description	string
genomeAssembly	string
htsFormat	
individualToSampleIdentifiers	object
uri	string

schemablocks.org

Swiss Personalized Health Network

Global Alliance for Genomics & Health

Beacon v2 Paths

Progenetix utilizes Beacon v2 REST paths

- Beacon v2 paths are used in the Beacon specification to scope query and delivery
- Progenetix uses a default /biosamples/ + query path for its front end queries, and then collection specific methods for data retrieval (see next)
- current implementation addresses a core subset of all options, and evaluates some still moving targets
  - ➡ variants_interpretations
  - ➡ variant instances versus prototypes
  - ➡ ...

progenetix

Base `/biosamples`

`/biosamples/` + query

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

`/biosamples/{id}/`

- /biosamples/pgxbs-kftva5c9/
  - retrieval of a single biosample

`/biosamples/{id}/variants/` & `/biosamples/{id}/variants_in_sample/`

- /biosamples/pgxbs-kftva5c9/variants/
- /biosamples/pgxbs-kftva5c9/variants_in_sample/
  - retrieval of all variants from a single biosample
  - currently - and especially since for a mostly CNV containing resource - `variants` means "variant instances" (or as in the early v2 draft `variantsInSample`)

Base `/variants`

There is currently (April 2021) still some discussion about the implementation and naming of the different types of genomic variant endpoints. Since the Progenetix collections follow a "variant observations" principle all variant requests are directed against the local `variants` collection.

If using `g_variants` or `variants_in_sample`, those will be treated as aliases.

`/variants/` + query

- /variants/?assemblyId=GRCh38&referenceName=17&variantType=DEL&filterLogic=AND&start=7500000&start=7676592&end=7669607&end=7800000
  - This is an example for a Beacon "Bracket Query" which will return focal deletions in the TP53 locus (by position).

`/variants/{id}/` or `/variants_in_sample/{id}` or `/g_variants/{id}/`

- /variants/5f5a35586b8c1d6d377b77f6/
- /variants_in_sample/5f5a35586b8c1d6d377b77f6/

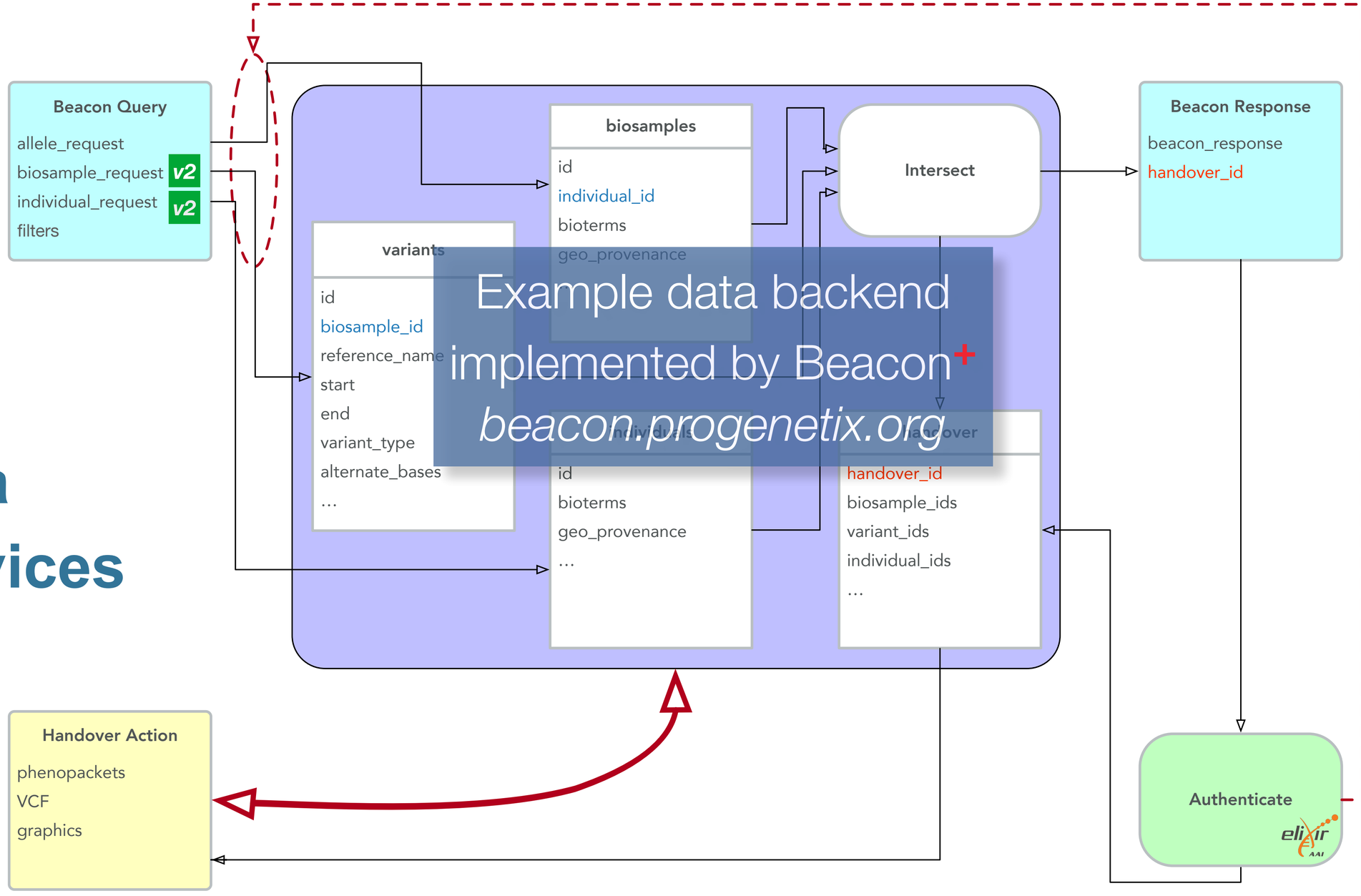
`/variants/{id}/biosamples/` & `variants_in_sample/{id}/biosamples/`

- /variants/5f5a35586b8c1d6d377b77f6/biosamples/
- /variants_in_sample/5f5a35586b8c1d6d377b77f6/biosamples/

https://github.com/ga4gh-beacon/implementations-v2/

Beacon & Handover

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



Beacon I/O

Handover

Authentication

Michael Baudis

Global Alliance for Genomics & Health

SIB

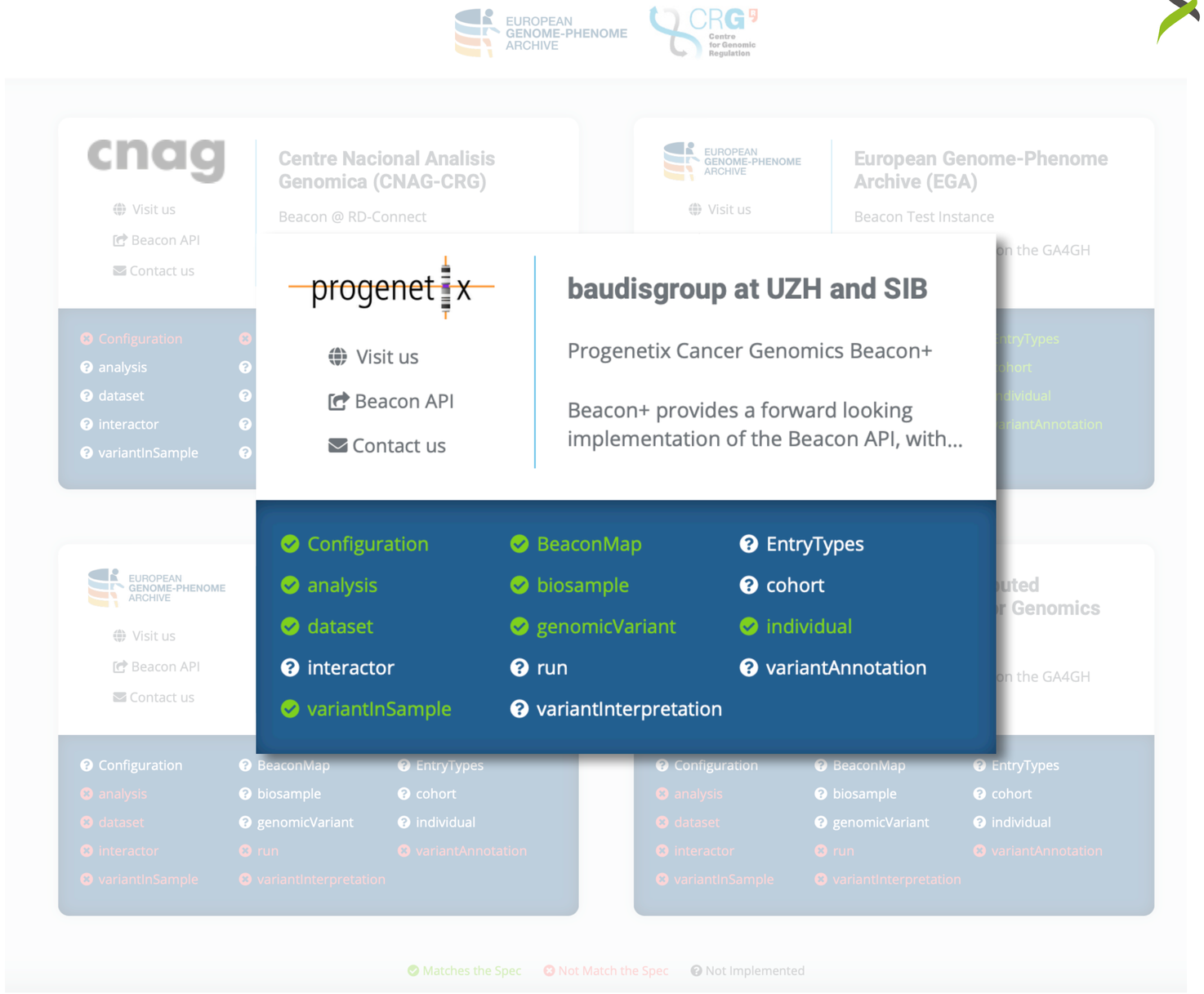
University of Zurich UZH

elixir

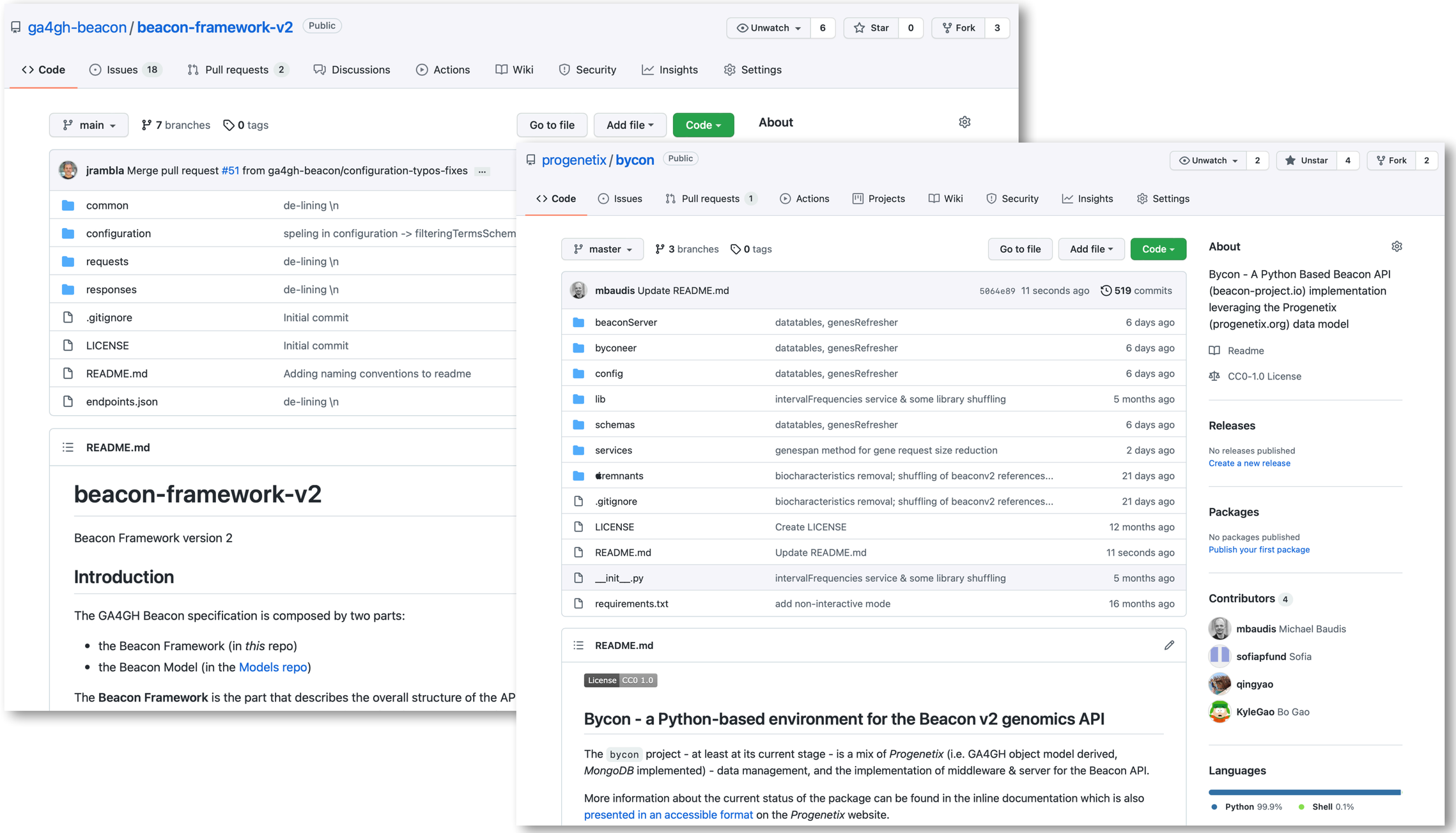
Onboarding

Demonstrating Compliance

- onboarding server run by CRG (EGA team)
- registering the URI of a server's map document will initiate traversal and testing of services
- blueprint for Beacon service registries
- to be used as demonstrator in GA4GH approval process for the Spring 2022 session



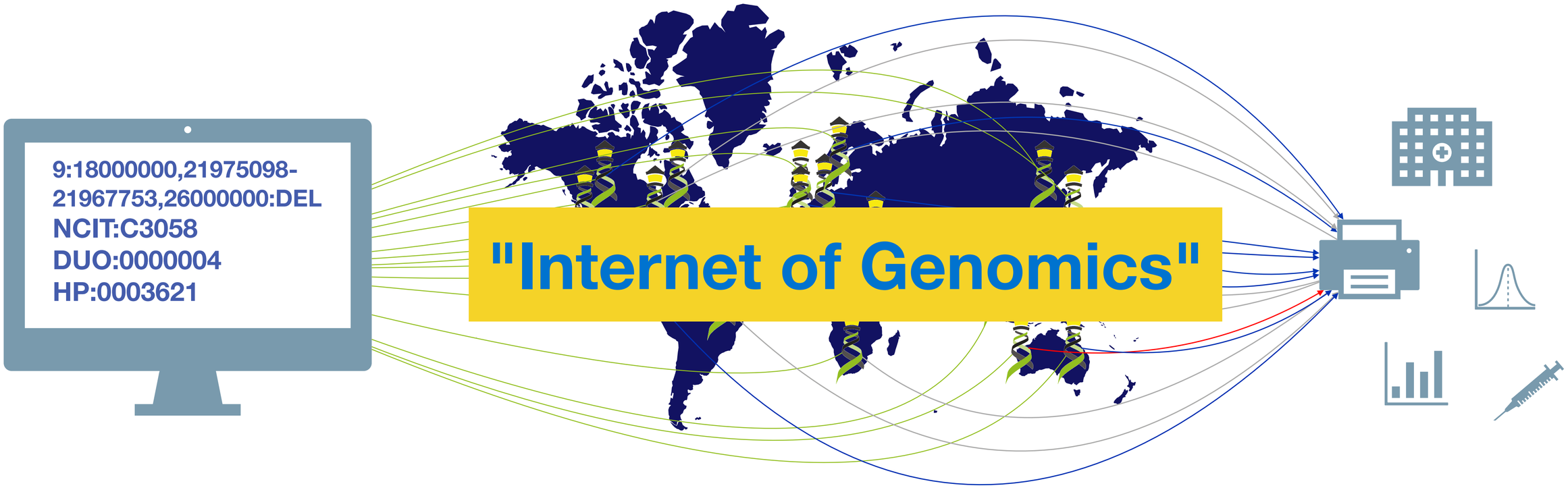
Global Alliance for Genomics & Health


ga4gh-beacon / beacon-framework-v2
Public
Unwatch 6
Star 0
Fork 3
Code
Issues 18
Pull requests 2
Discussions
Actions
Wiki
Security
Insights
Settings
main
7 branches
0 tags
Go to file
Add file
Code
About
jrambla Merge pull request #51 from ga4gh-beacon/configuration-typos-fixes
common de-lining \n
configuration speling in configuration -> filteringTermsSchem
requests de-lining \n
responses de-lining \n
.gitignore Initial commit
LICENSE Initial commit
README.md Adding naming conventions to readme
endpoints.json de-lining \n
README.md
beacon-framework-v2
Beacon Framework version 2
Introduction
The GA4GH Beacon specification is composed by two parts:
the Beacon Framework (in this repo)
the Beacon Model (in the Models repo)
The Beacon Framework is the part that describes the overall structure of the AP
progenetix / bycon
Public
Unwatch 2
Unstar 4
Fork 2
Code
Issues
Pull requests 1
Actions
Projects
Wiki
Security
Insights
Settings
master
3 branches
0 tags
Go to file
Add file
Code
mbaudis Update README.md
5064e89 11 seconds ago
519 commits
beaconServer datatables, genesRefresher 6 days ago
byconeer datatables, genesRefresher 6 days ago
config datatables, genesRefresher 6 days ago
lib intervalFrequencies service & some library shuffling 5 months ago
schemas datatables, genesRefresher 6 days ago
services genespan method for gene request size reduction 2 days ago
remnants biocharacteristics removal; shuffling of beaconv2 references... 21 days ago
.gitignore biocharacteristics removal; shuffling of beaconv2 references... 21 days ago
LICENSE Create LICENSE 12 months ago
README.md Update README.md 11 seconds ago
__init__.py intervalFrequencies service & some library shuffling 5 months ago
requirements.txt add non-interactive mode 16 months ago
README.md
License CC0 1.0
Bycon - a Python-based environment for the Beacon v2 genomics API
The bycon project - at least at its current stage - is a mix of Progenetix (i.e. GA4GH object model derived, MongoDB implemented) - data management, and the implementation of middleware & server for the Beacon API.
More information about the current status of the package can be found in the inline documentation which is also presented in an accessible format on the Progenetix website.
About
Bycon - A Python Based Beacon API (beacon-project.io) implementation leveraging the Progenetix (progenetix.org) data model
Readme
CC0-1.0 License
Releases
No releases published
Create a new release
Packages
No packages published
Publish your first package
Contributors 4
mbaudis Michael Baudis
sofiapfund Sofia
qingyao
KyleGao Bo Gao
Languages
Python 99.9%
Shell 0.1%

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**".



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

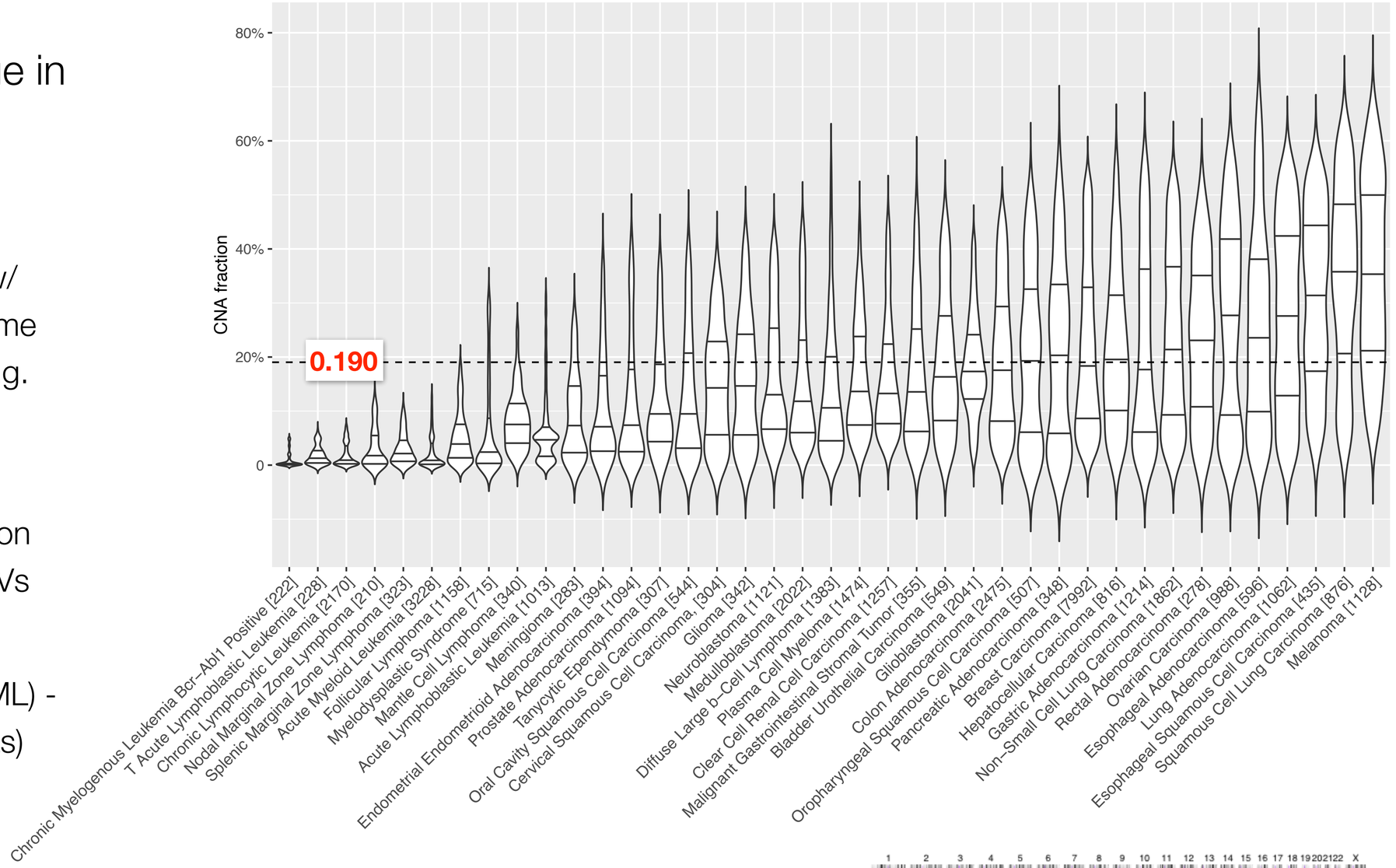
The Beacon API v2 proposal opens the way for the design of a simple but powerful "**genomics API**".

Progenetix Data Use Cases

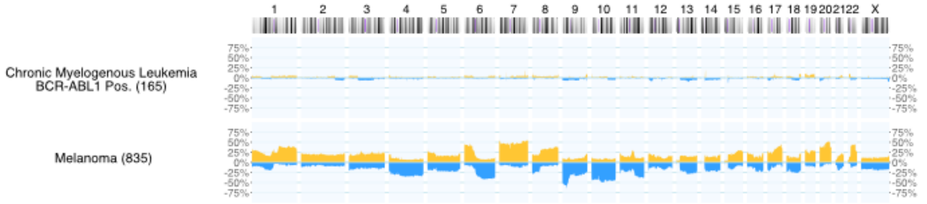
progenetix

Genome CNV coverage in Cancer Classes

- 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)



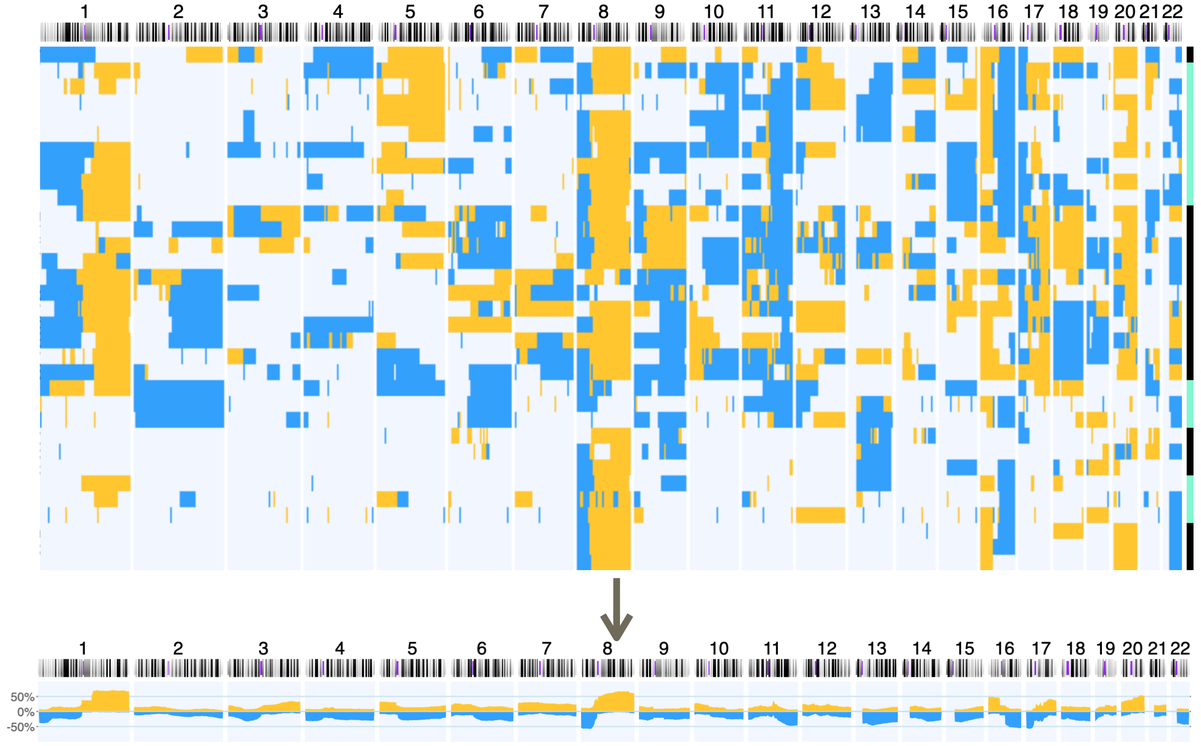
Lowest / Highest CNV fractions =>



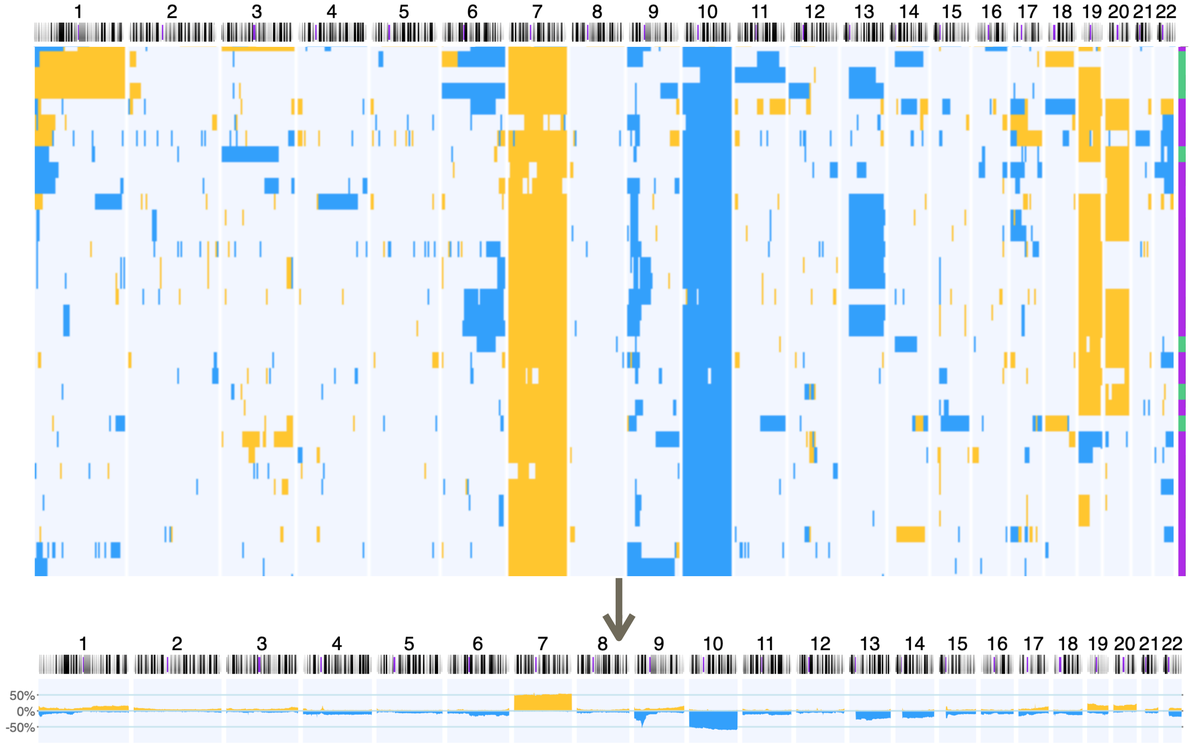
Drivers? Passengers? Markers?

Disentangling CNA Patterns

Ductal Breast Carcinoma



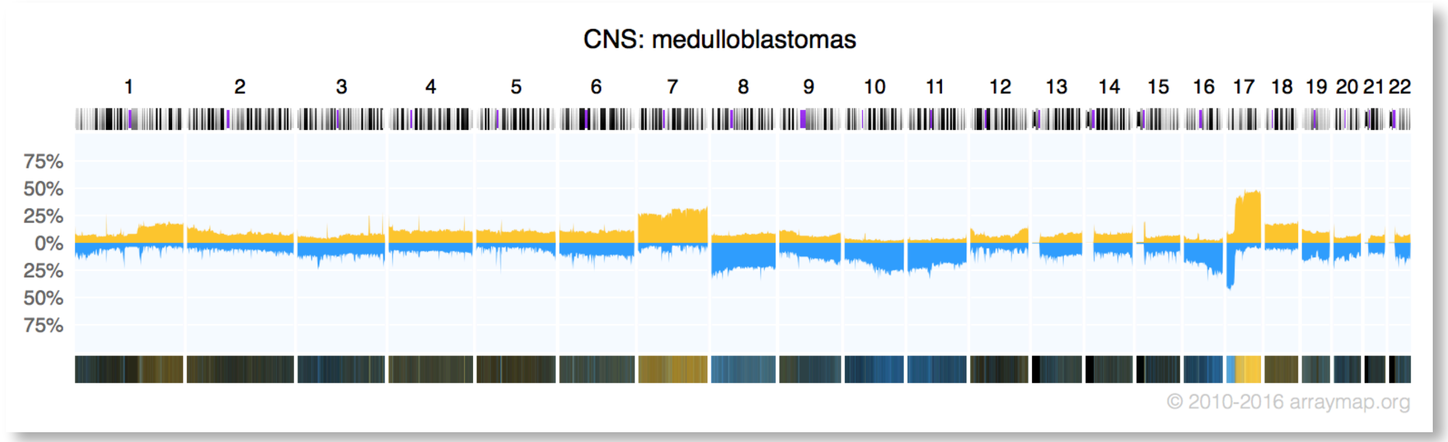
Glioblastoma

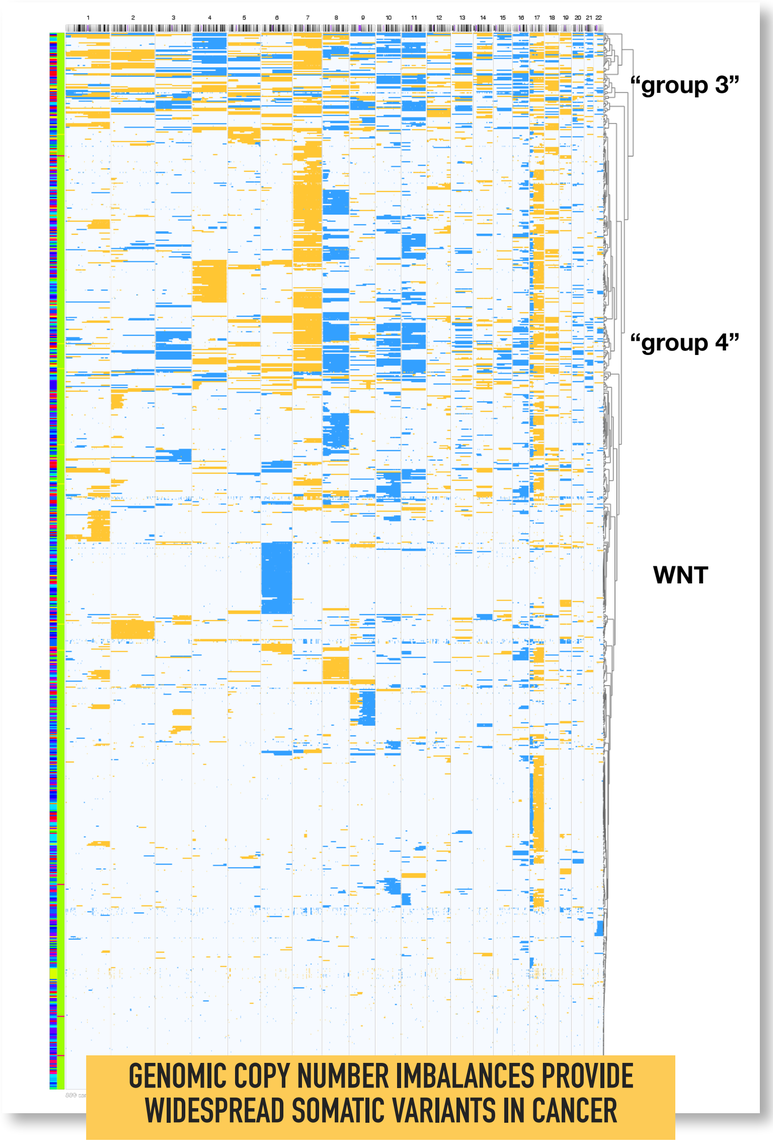


Somatic CNVs In Cancer

Recurrent mutation patterns

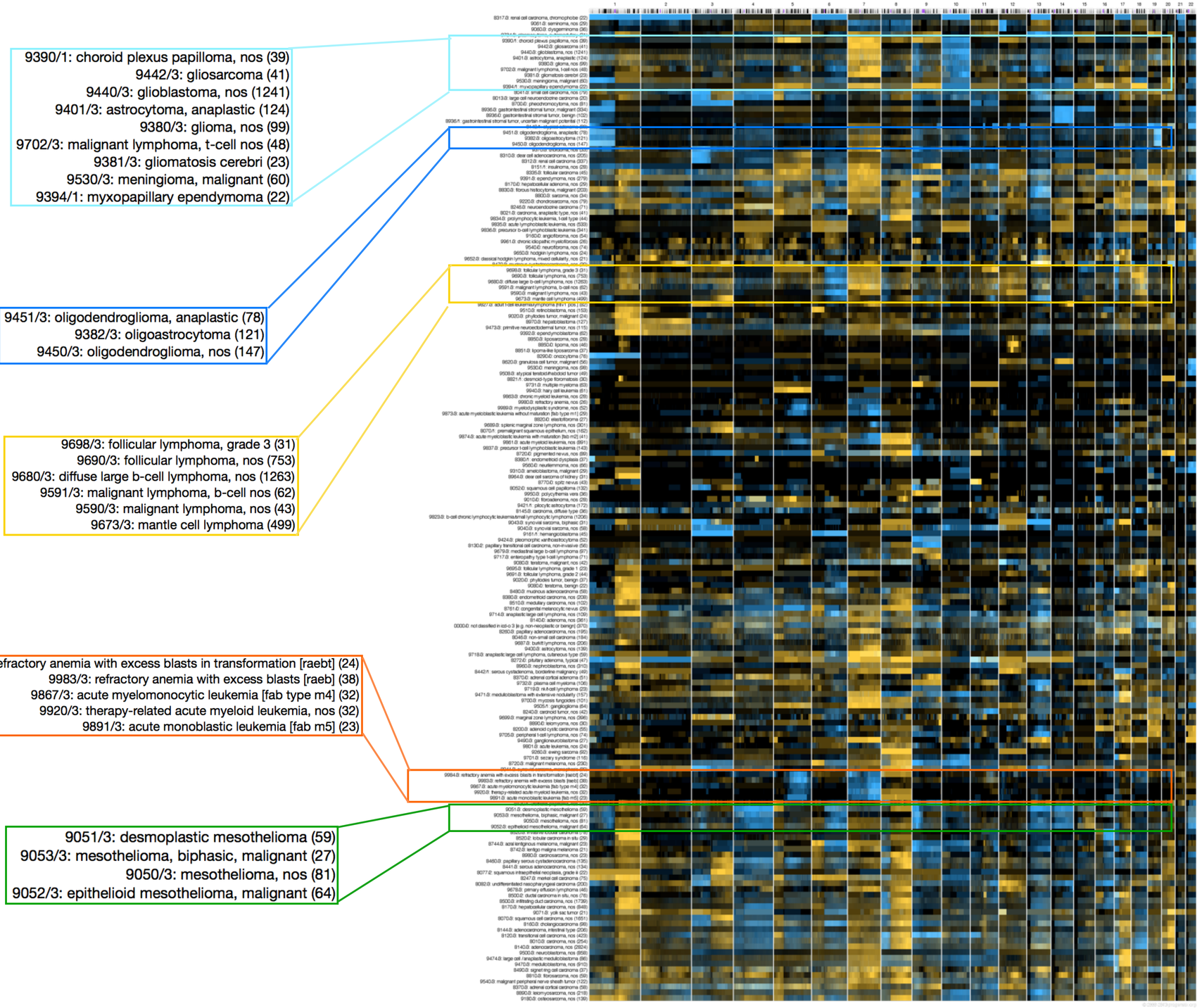
How can those patterns be used for classification and determination of biological mechanisms?





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.

Somatic Mutations In Cancer: Patterns
Making the case for genomic classifications
Some related cancer entities show similar copy number profiles



arrayMap

Chromothripsis-like patterns (CTLP)

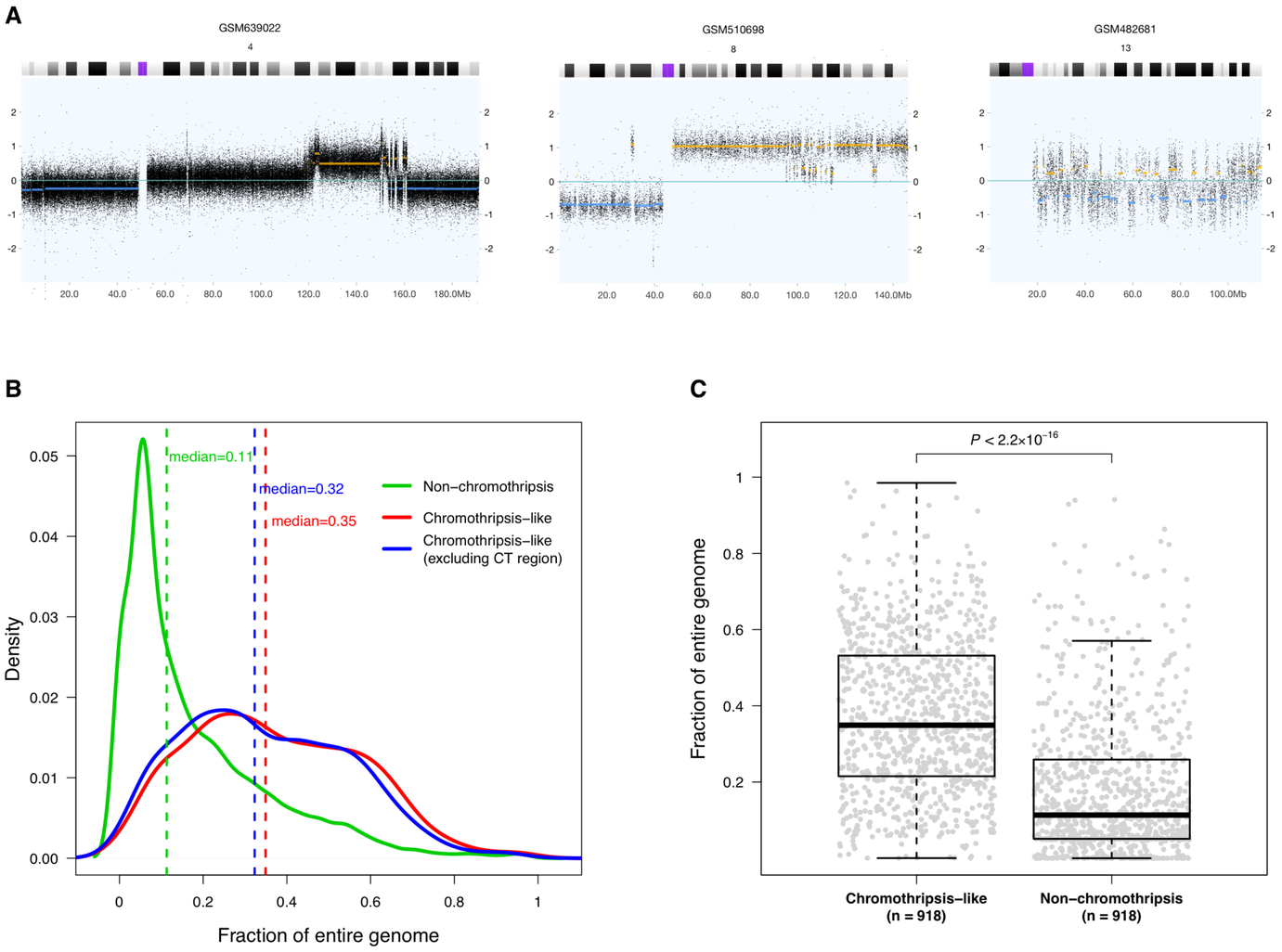
Cai et al. BMC Genomics 2014, 15:82
http://www.biomedcentral.com/1471-2164/15/82

BMC Genomics

RESEARCH ARTICLE Open Access

Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens

Haoyang Cai[1,2], Nitin Kumar[1,2], Homayoun C Bagheri[3], Christian von Mering[1,2], Mark D Robinson[1,2*] and Michael Baudis[1,2*]



Input

22347 arrays
(18458 cases)

Output

1566 chromosomes
(1028 arrays)

arrayMap

Population stratification in cancer samples based on SNP array data

- Despite extensive somatic mutations of cancer profiling data, consistency between germline and cancer samples reached 97% and 92% for 5 and 26 populations
- Comparison of our benchmarked results with self-reported meta-data estimated a matching rate between 88 % to 92%.
- Ethnicity labels indicated in meta-data are vague compared to the standardized output from our tool

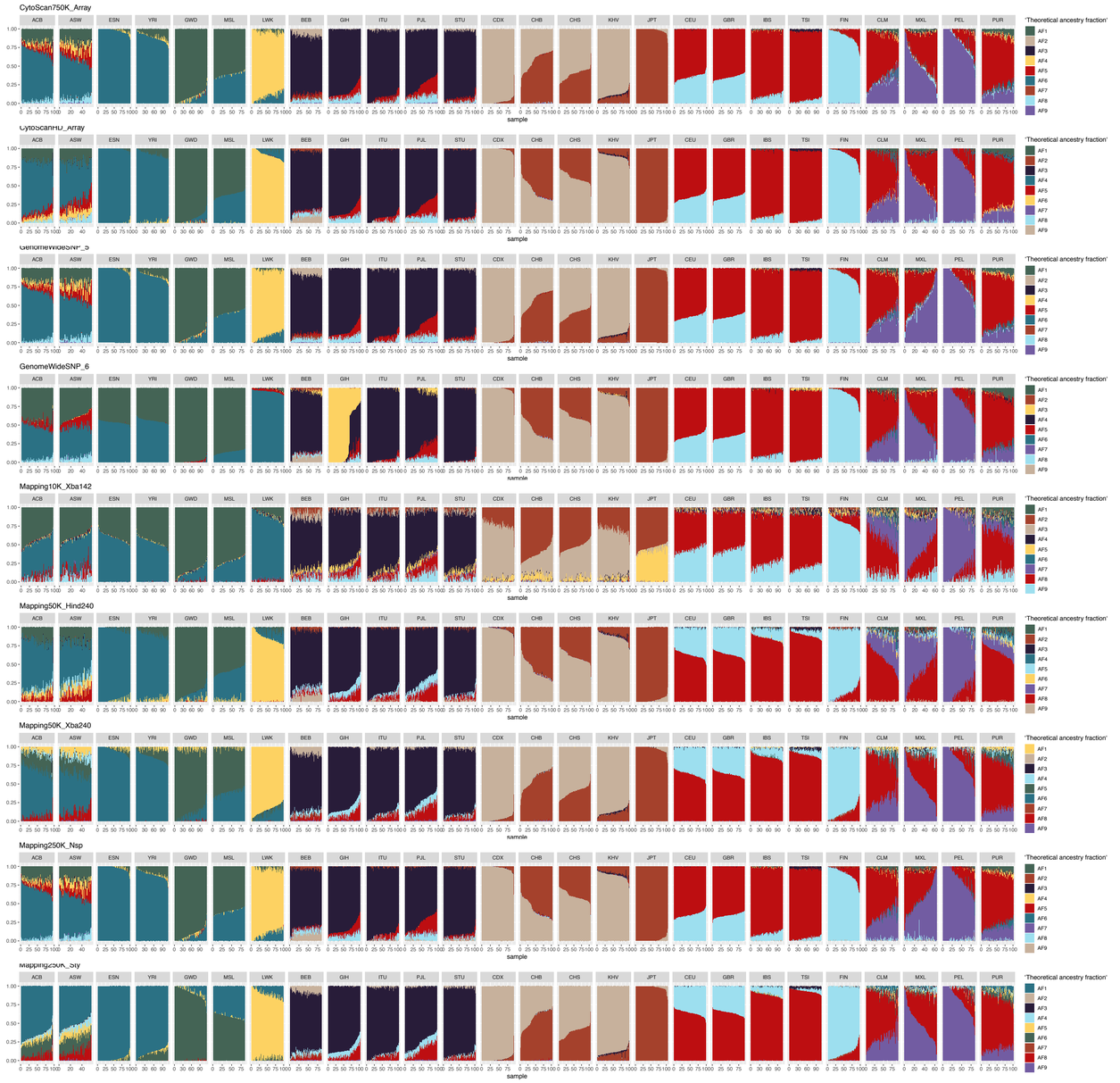


Figure S1 The fraction or contribution of theoretical ancestors (k=9) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are placed neighboring to each other: AFR (1-7), SAS (8-12), EAS (13-17), EUR (18-22), AMR (23-26).

arrayMap

Qingyao Huang

Progenetix Contributes to ELIXIR hCNV

hCNV Implementation Studies 2021-2023 No. 2

Beacon and beyond — Implementation-driven standards and protocols for CNV discovery and data exchange

- reinforce work on priority areas established in the current hCNV Implementation Study
- extend collaborations with the Rare Diseases and Galaxy Communities, EJP-RD and GA4GH
- Expected outcomes
  - ➡ shared CNV resources testing advanced versions of the Beacon protocol
  - ➡ integration of GA4GH standards such as Phenopackets in such resources
  - ➡ tools for data ingestion and export for standard formats (e.g. VCF, Phenopackets) and CNV-specific improvements of such standards
  - ➡ ELIXIR AAI demo on clinical and research hCNV resources
  - ➡ demonstration of Galaxy pipeline adoption for real-world hCNV data analysis projects
- connecting to international partners, e.g. Cancer Genomics Consortium (U.S.)

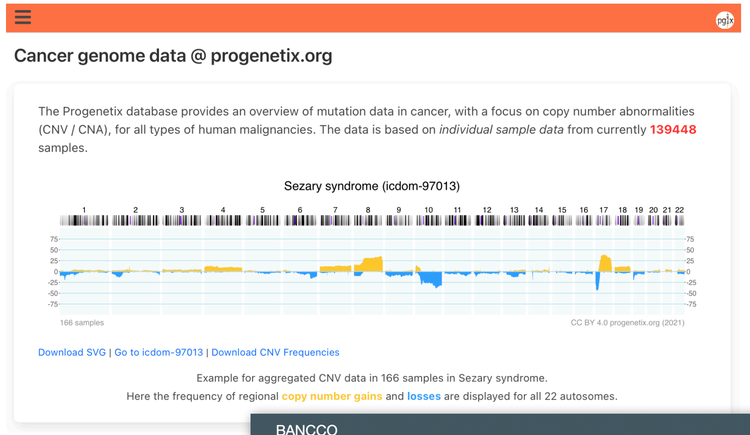
- WP1 - hCNV community reference resources
- WP2 - hCNV Resources and Beacon
- WP3 - Galaxy Community Intersection and Data Exchange
- WP4 - Workflows and Tools for hCNV Data Exchange Procedures
- WP5 - Training and dissemination

Global Alliance for Genomics & Health

elixir

hCNV Implementation Studies 2021-2023 No. 2

Beacon and beyond — Implementation-driven standards and protocols for CNV discovery and data exchange



"Galaxify", "Beaconize" & "Phenopack"
Progenetix & RD-CNVdb prototypes

- WP1 - hCNV community reference resources
- WP2 - hCNV Resources and Beacon
- WP3 - Galaxy Community Intersection and Data Exchange
- WP4 - Workflows and Tools for hCNV Data Exchange Procedures
- WP5 - Training and dissemination

Baudisgroup @ UZH

Michael Baudis
(Paula Carrio Cordo)
(Bo Gao)
Qingyao Huang
Sofia Pfund
Rahel Paloots
Hangjia Zhao

Pierre-Henri Toussaint

{S}[B] and GA4GH

Melanie Courtot
Helen Parkinson
many more ...

Beacon API Leads

Jordi Rambla
Anthony Brooks
Juha Törnroos

Discovery WS

Michael Baudis (Beacon)
Marc Fiume (Networks)

ELIXIR

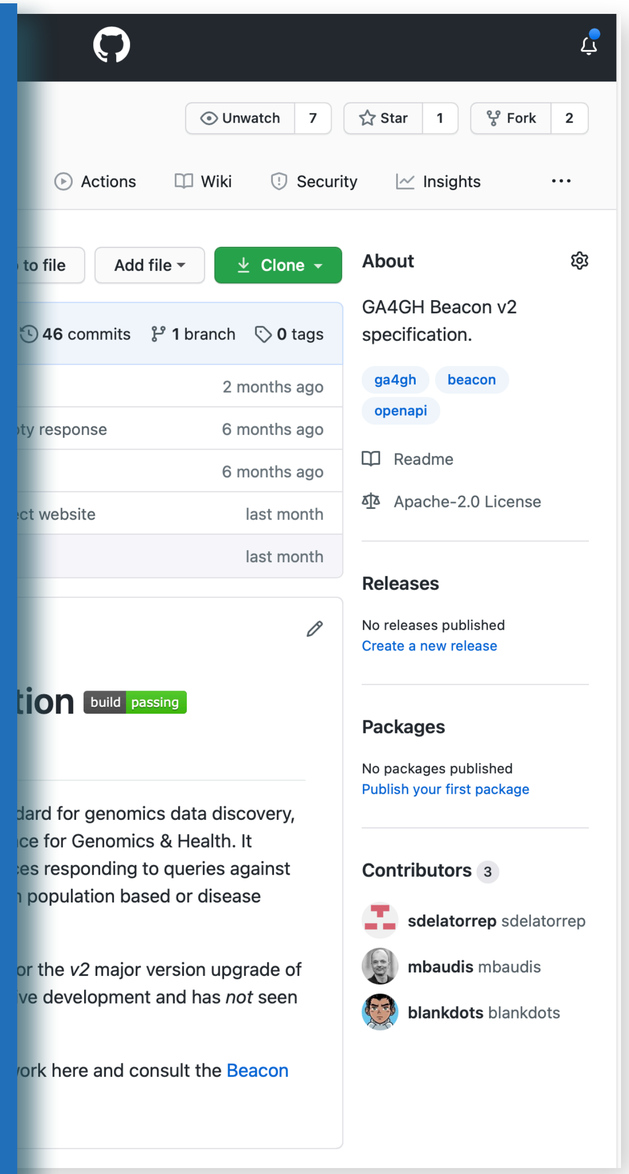
Gary Saunders
David Lloyd
Serena Scollen
Dylan Spalding

Beacon Team CRG

Laureen Fromont
Babita Singh
Sabela de la Torre Pernas
...

Beacon v2 Scouts

Tim Beck
Joaquin Dopazo
Veronique Geoffroy
Jean Muller
David Salgado
Alex Wagner
...



beacon-project.io

progenetix.org

github.com/ga4gh-beacon/

Swiss Personalized Health Network

Global Alliance for Genomics & Health

Prof. Dr. Michael Baudis
Institute of Molecular Life Sciences
University of Zurich
SIB | Swiss Institute of Bioinformatics
Winterthurerstrasse 190
CH-8057 Zurich
Switzerland

progenetix.org
info.baudisgroup.org
sib.swiss/baudis-michael
imls.uzh.ch/en/research/baudis
beacon-project.io
schemablocks.org