Progenetix Genomics Resource

From Experiments to APIs

Michael Baudis | 2022

progenetix

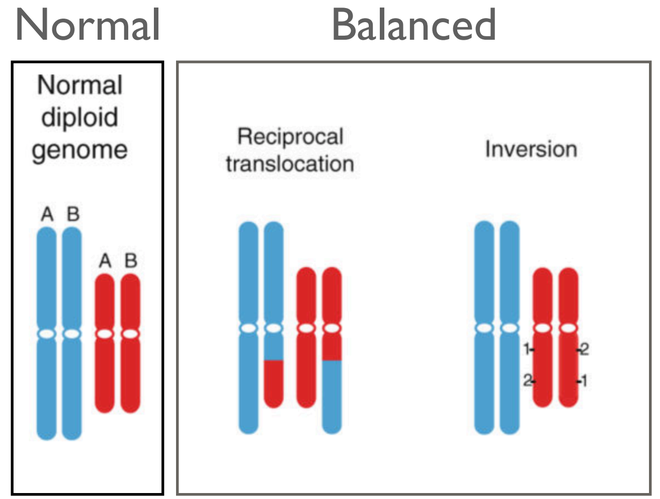
Types of genomic alterations in Cancer

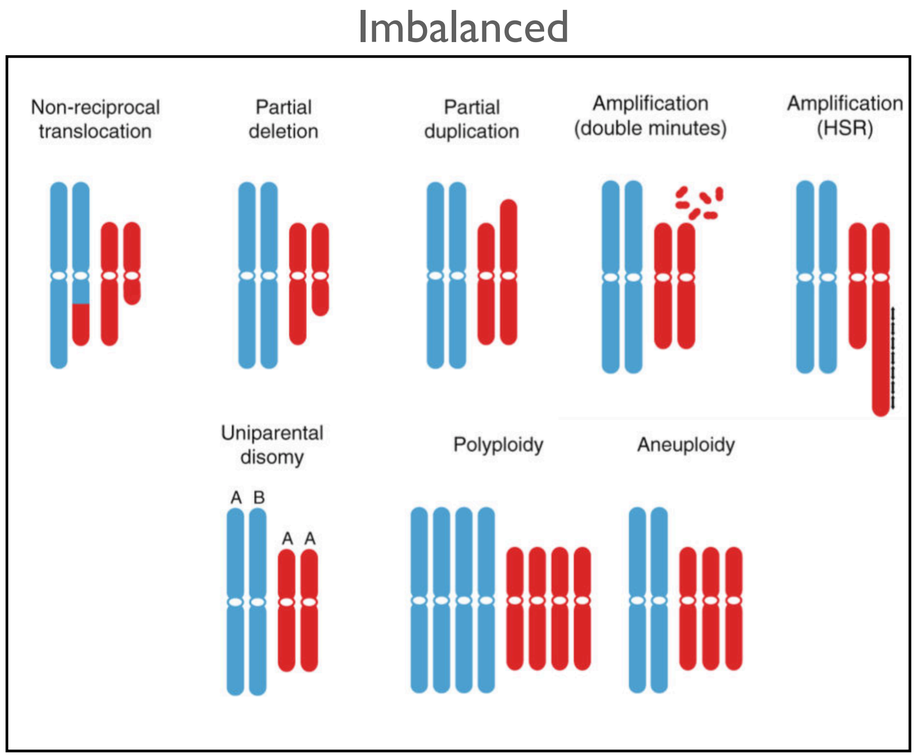
Imbalanced Chromosomal Changes: CNV

- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- Structural chromosomal Aberrations

➡ **Regional Copy Number Alterations** (losses, gains)

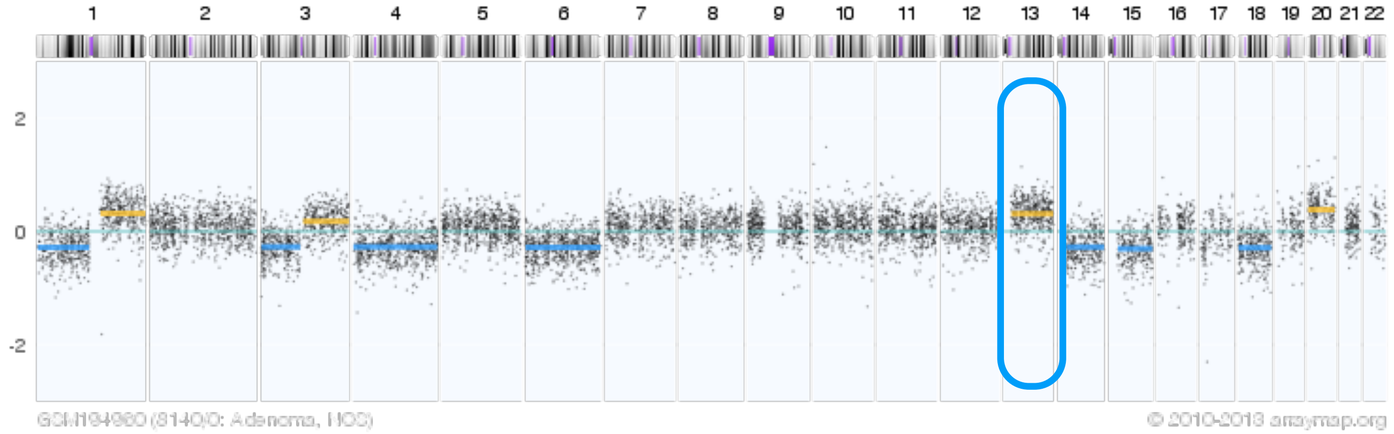
- Epigenetic changes (e.g. DNA methylation abnormalities)



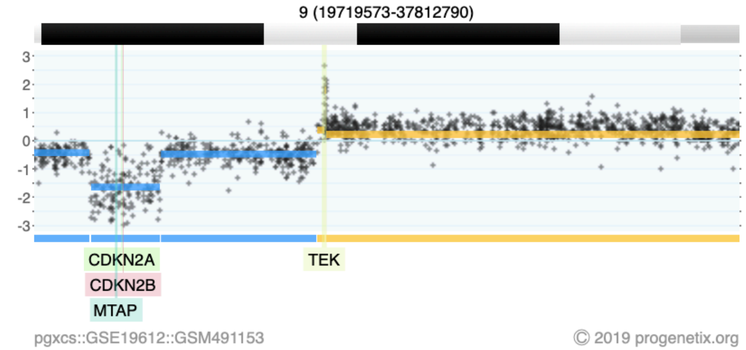


Grade et al., 2015 Recent Results Cancer Res

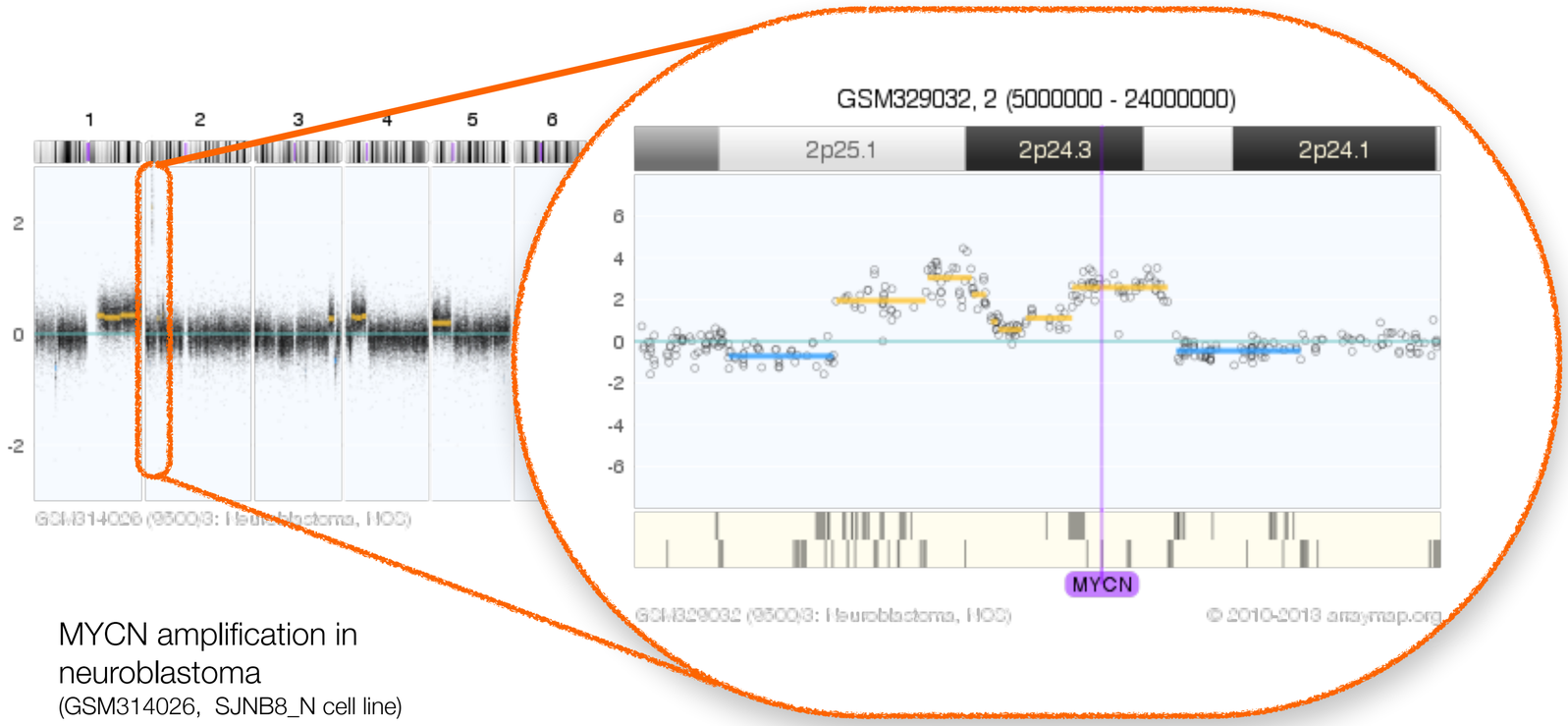
Somatic Copy Number Variations



Gain of chromosome arm 13q in colorectal carcinoma



2-event, homozygous deletion in a Glioblastoma

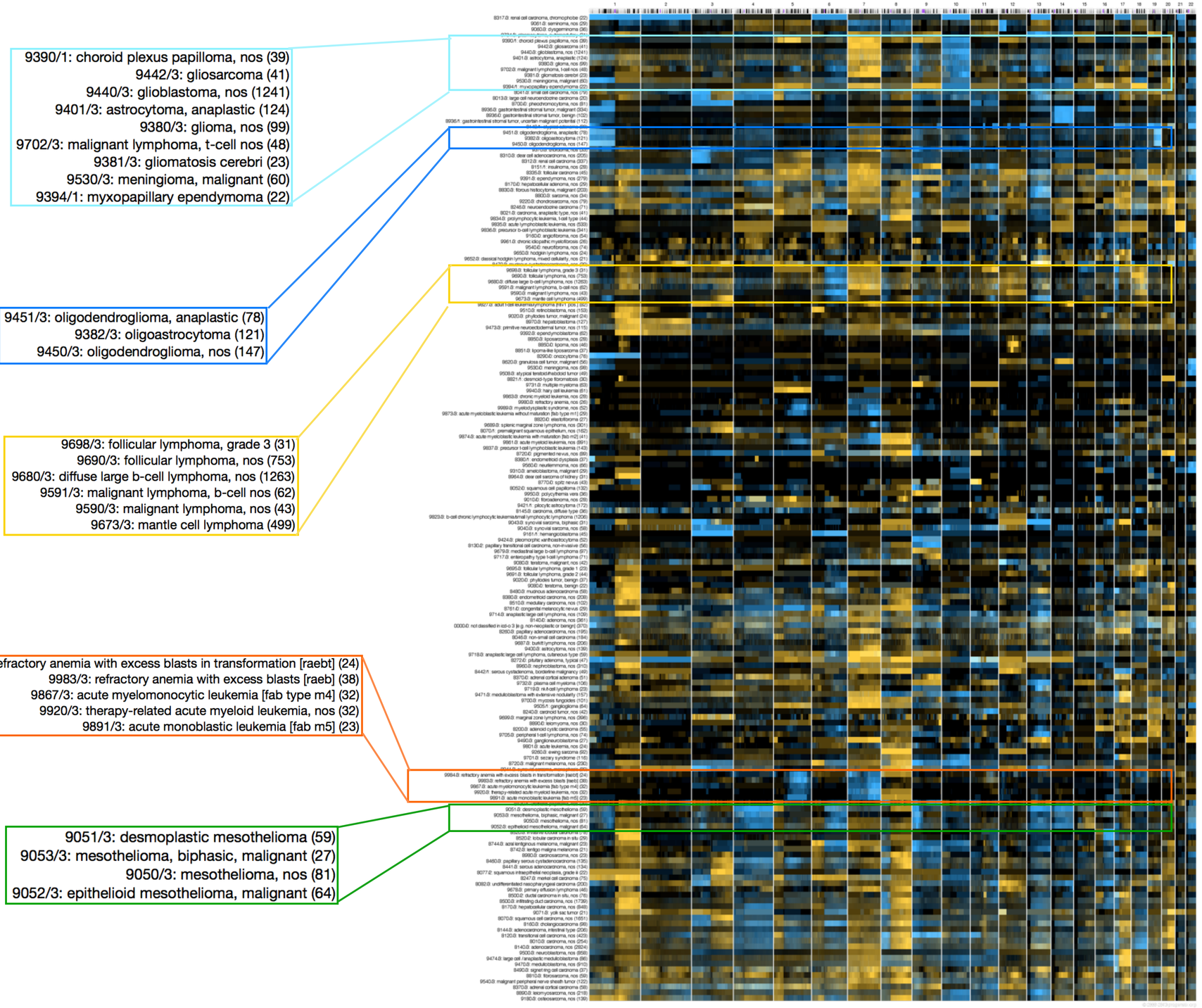


MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

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

arrayMap

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

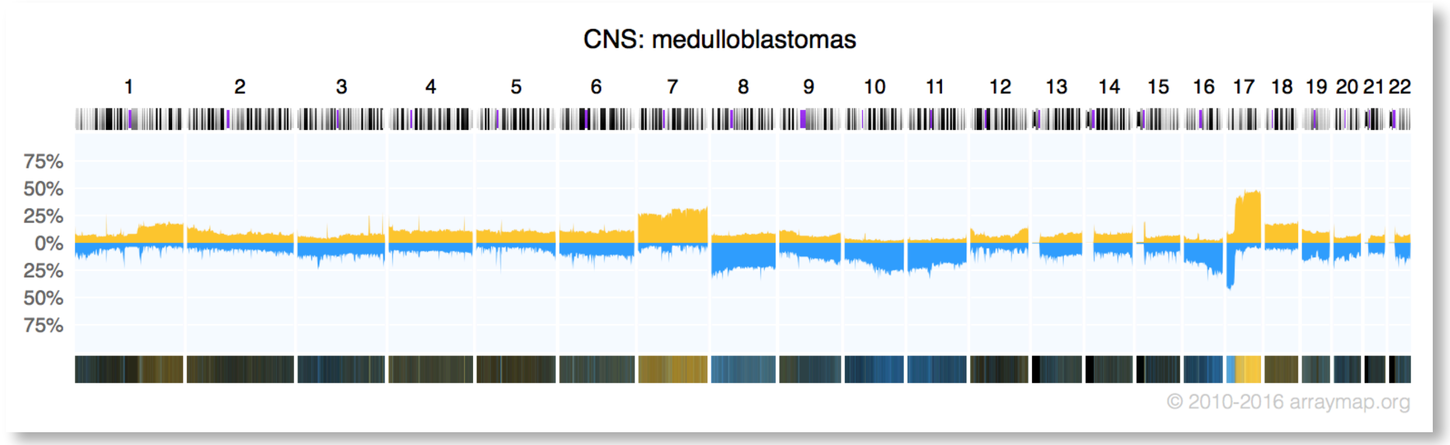


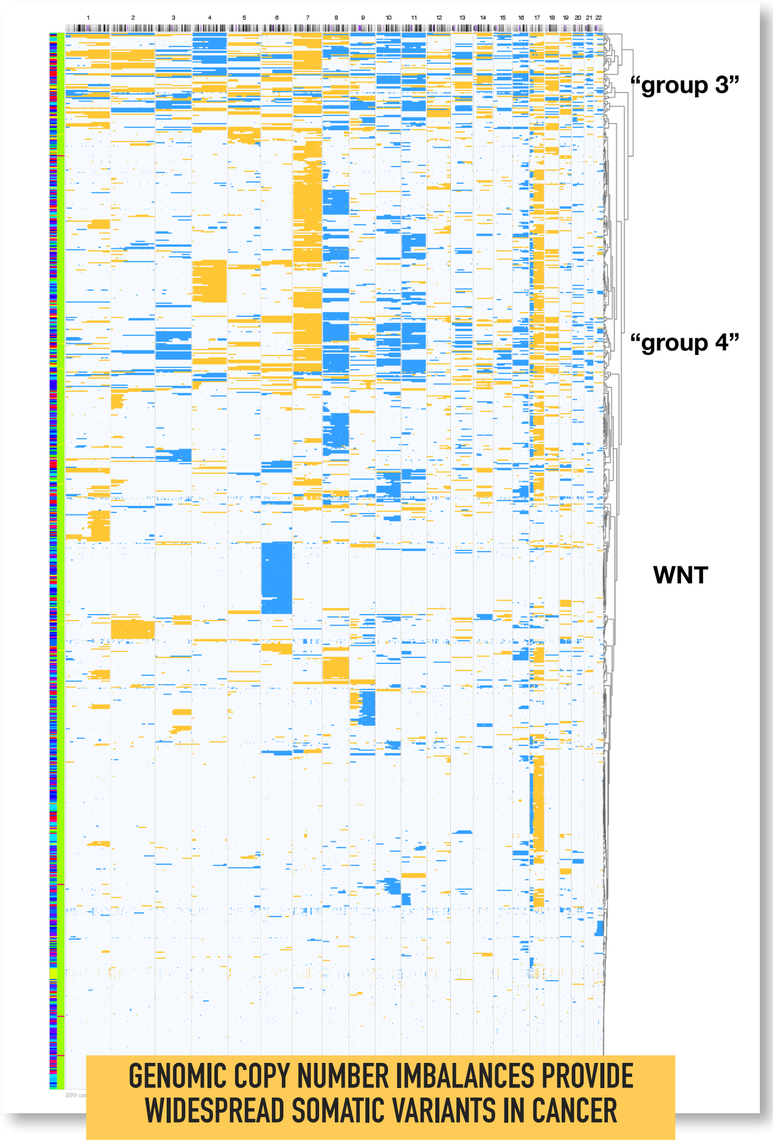
arrayMap

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

BIOINFORMATICS ORIGINAL PAPER

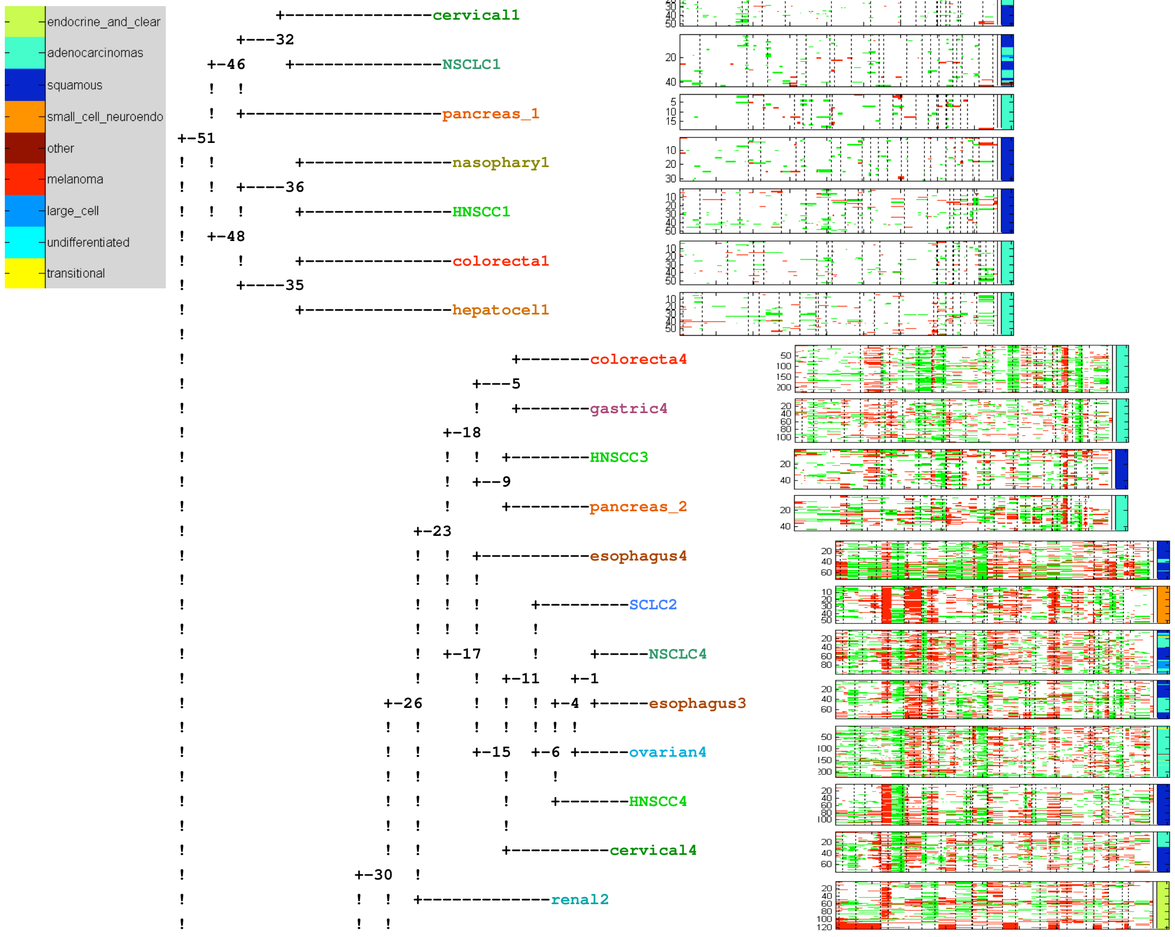
Vol. 25 no. 17 2009, pages 2208–2215
doi:10.1093/bioinformatics/btp365

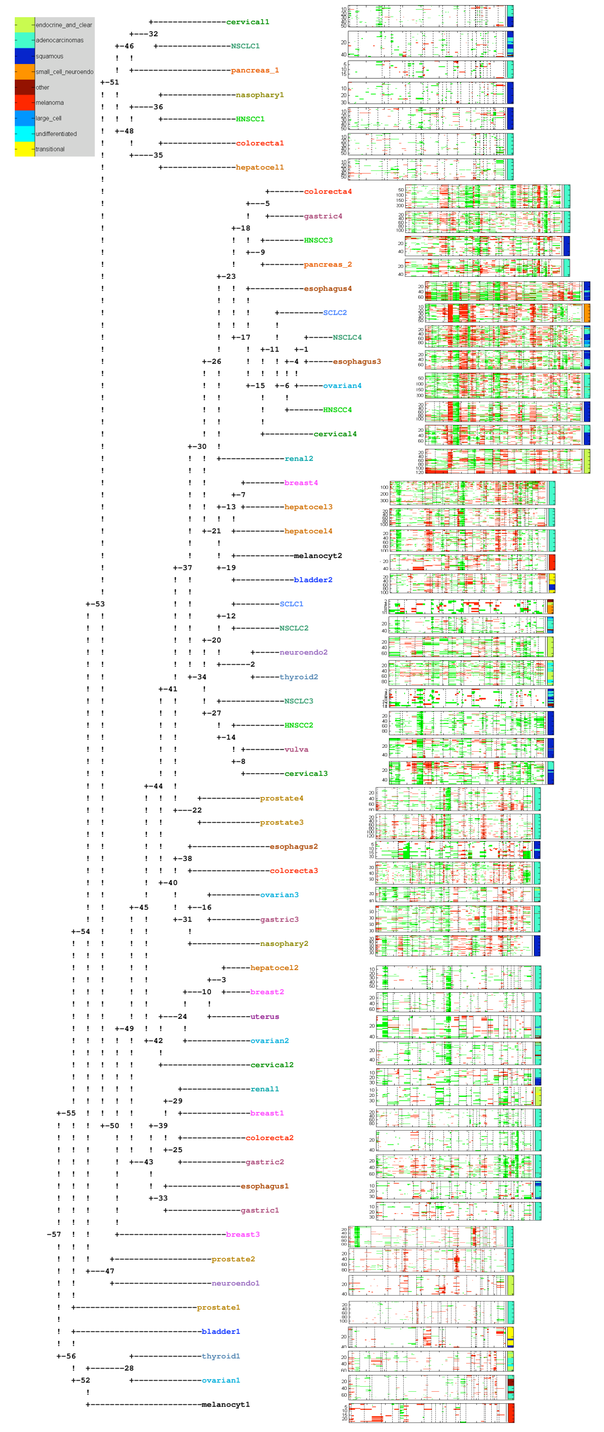
Gene expression

Inferring progression models for CGH data

Jun Liu[1], Nirmalya Bandyopadhyay[1,*], Sanjay Ranka[1], M. Baudis[2] and Tamer Kahveci[1,*]

[1]Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA and [2]Institute for Molecular Biology, University of Zurich, Zurich, Switzerland





frontiers in Genetics

ORIGINAL RESEARCH
published: 13 May 2021
doi: 10.3389/fgene.2021.654887

Signatures of Discriminative Copy Number Aberrations in 31 Cancer Subtypes

Bo Gao [1,2] *and Michael Baudis* [1,2*]

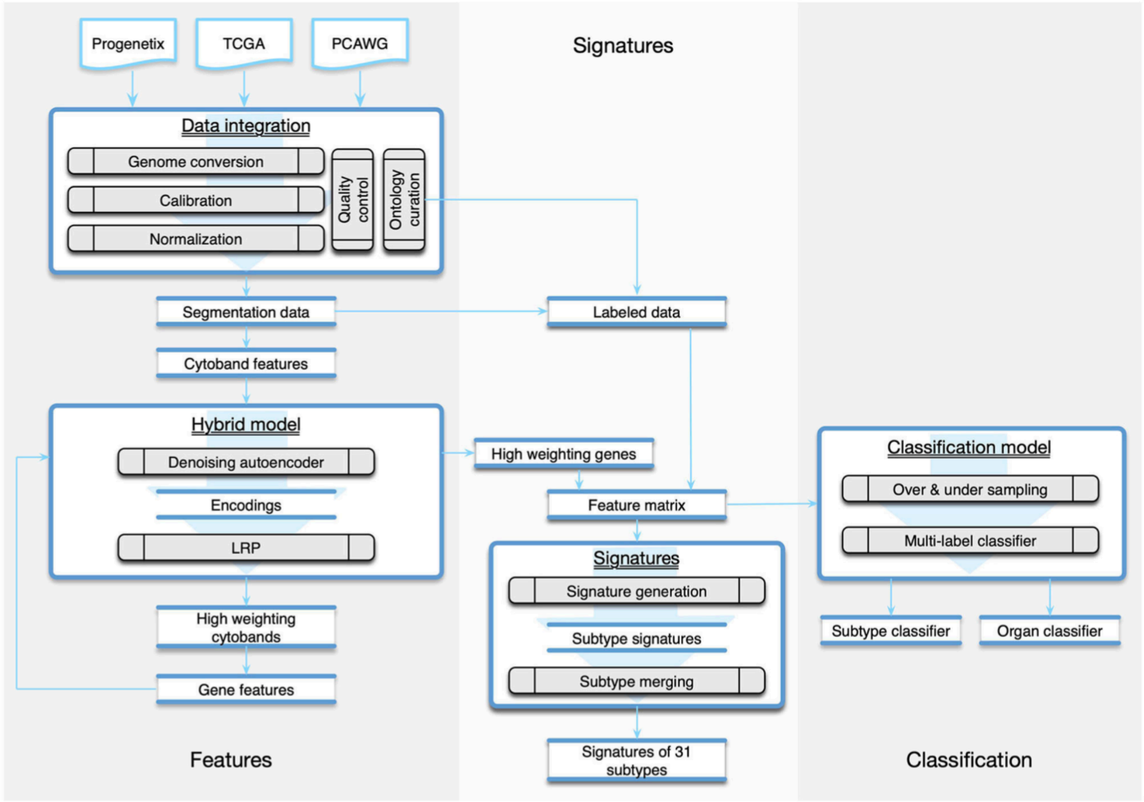


FIGURE 1 | The workflow of the study was composed of three parts. The *Features* part consisted of methods of data integration and feature generation. The *Signature* part focused on creating CNA signatures for cancer subtypes and the categorization of subtypes. The *Classification* part recruited machine learning techniques to predict the organ and the subtype from a given copy number profile.

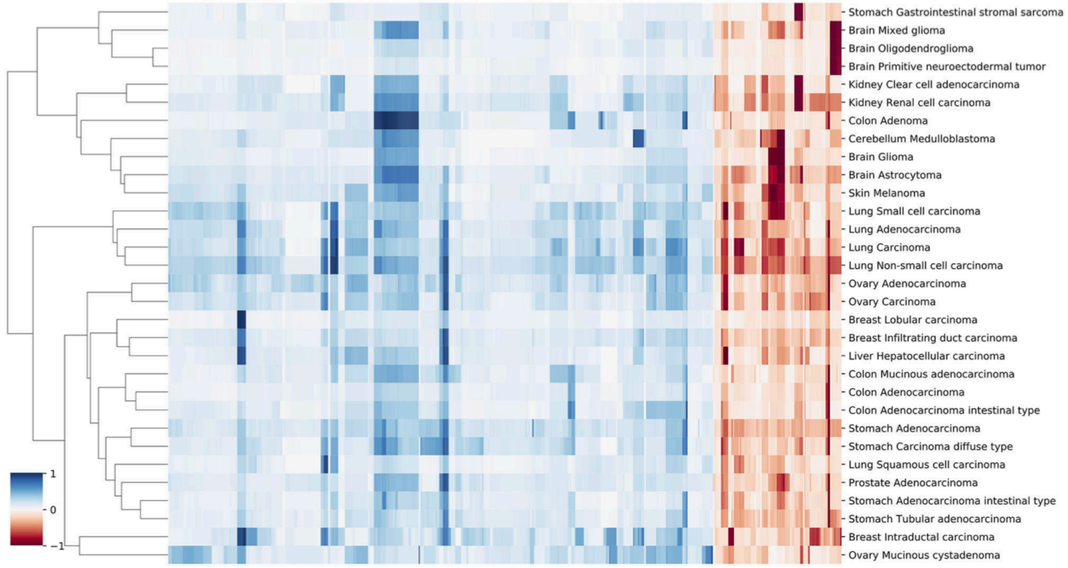


FIGURE 5 | A clustering heatmap of features in 31 signatures. Columns are normalized average CNV intensities of feature genes, where the blue colors are duplication features and red colors are deletion features. Duplication and deletion frequencies are normalized separately.

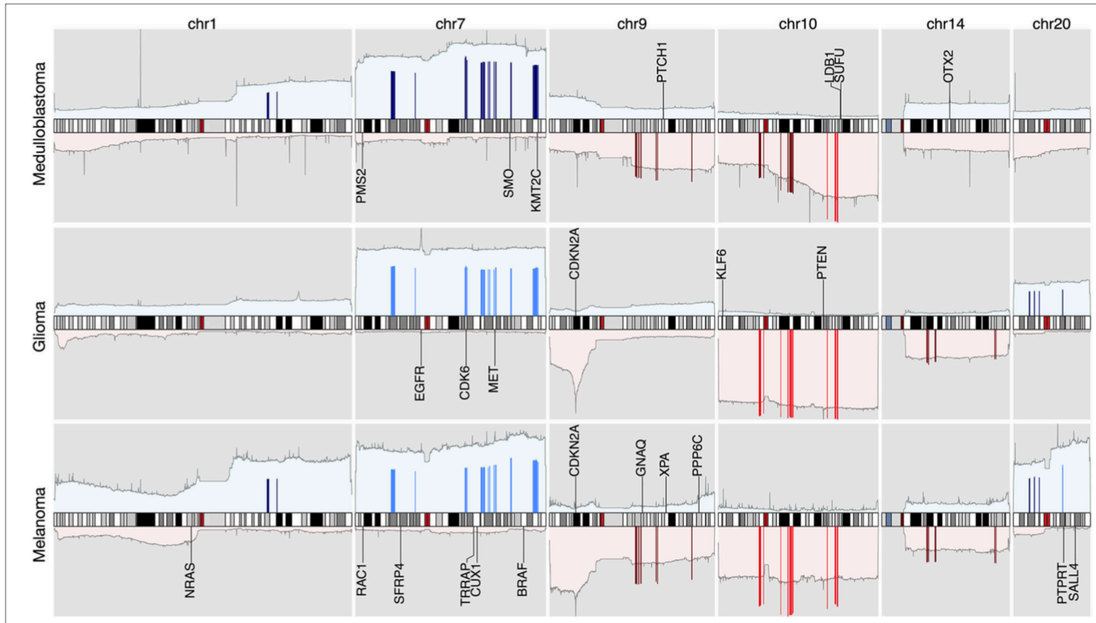


FIGURE 6 | The integrated view of the original data and the selected features, in the neural crest originating entities medulloblastoma, glioma, and melanoma. The shaded background area color illustrates the original data. Color bars illustrate the feature genes, where brighter colors indicate stronger signal intensity. The blue colors above the chromosome axis represent the average amplifications, and the red colors below the chromosome axis represent the average deletions. The amplitude of amplifications and deletions are normalized to [0,1] separately. The adjacent known driver genes are also included for each tumor type.

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

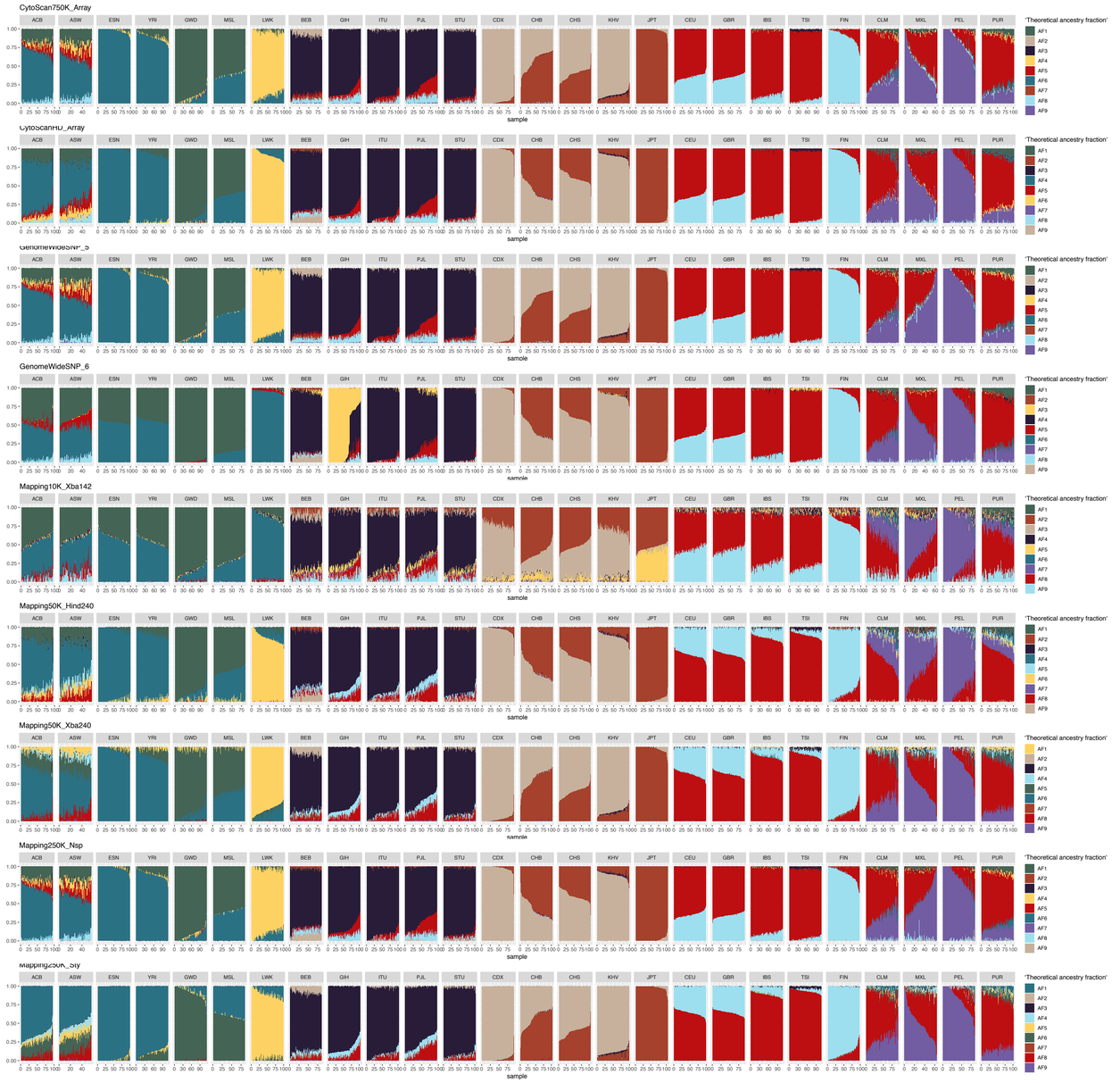
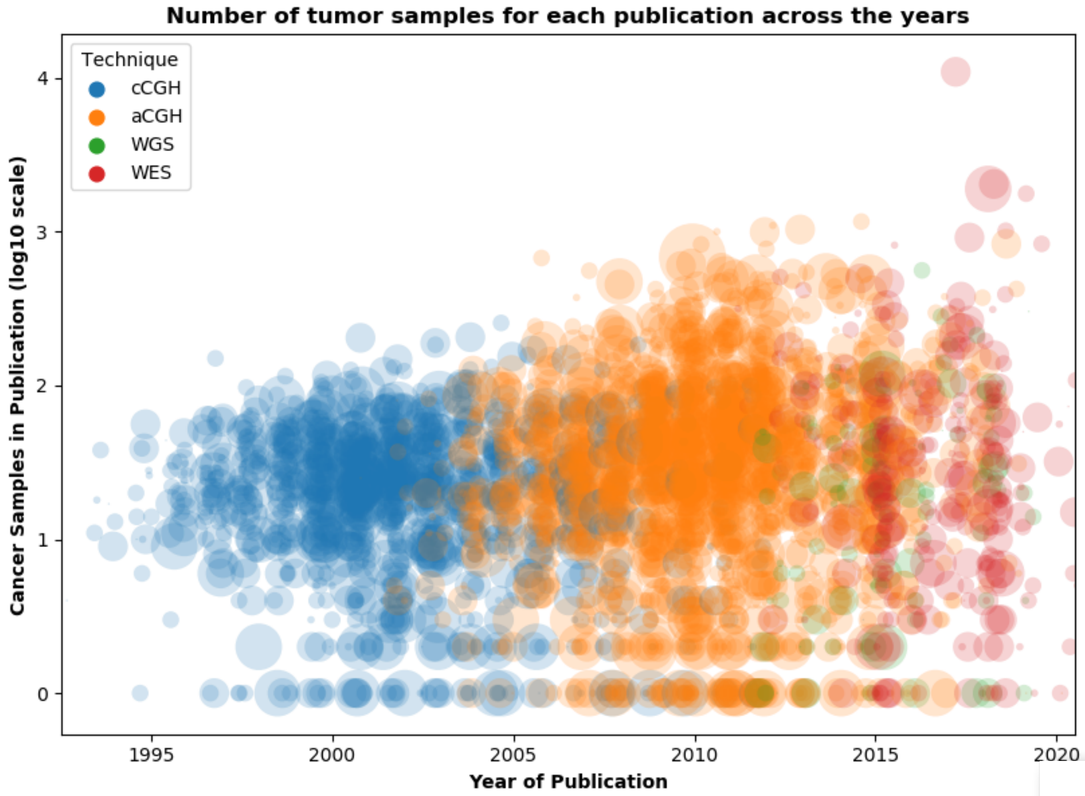


Figure S1 The fraction or contribution of theoretical ancestors (k=9) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are placed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

arrayMap

Qingyao Huang



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.

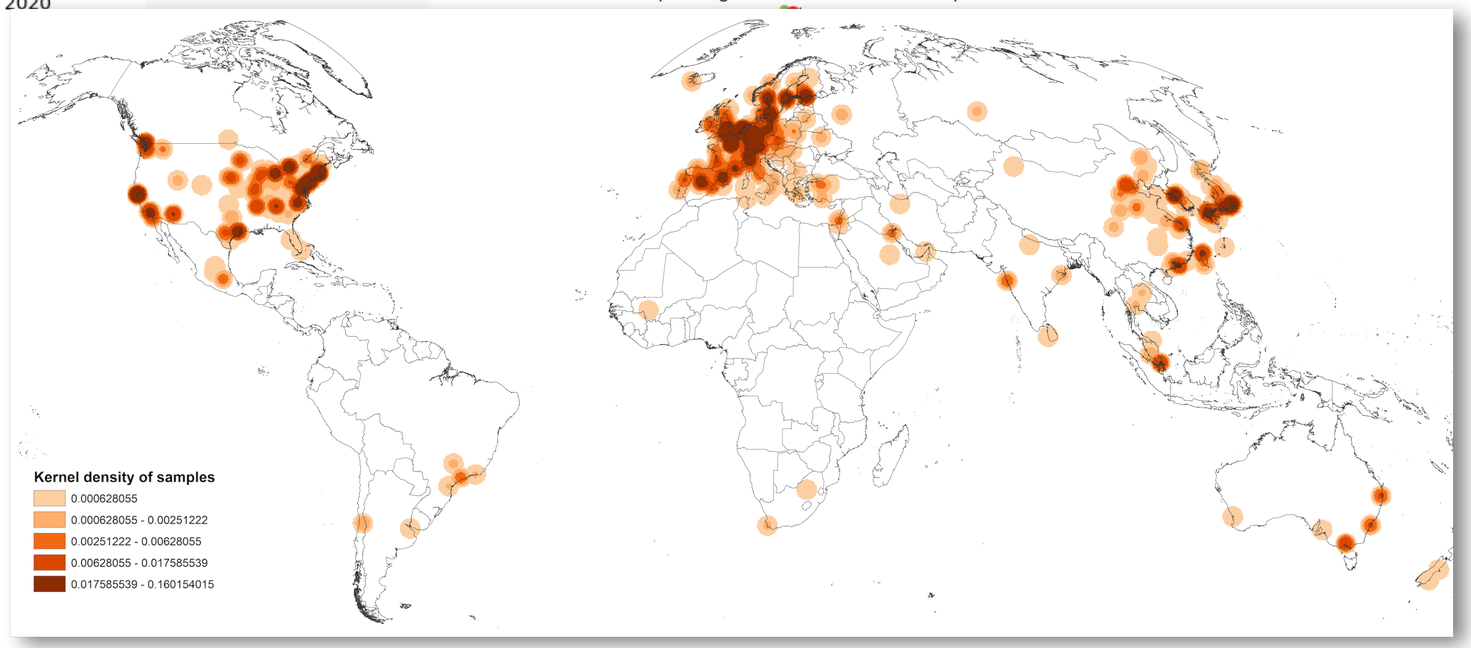
Please contact us to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation.

Filter

City

Type to search...

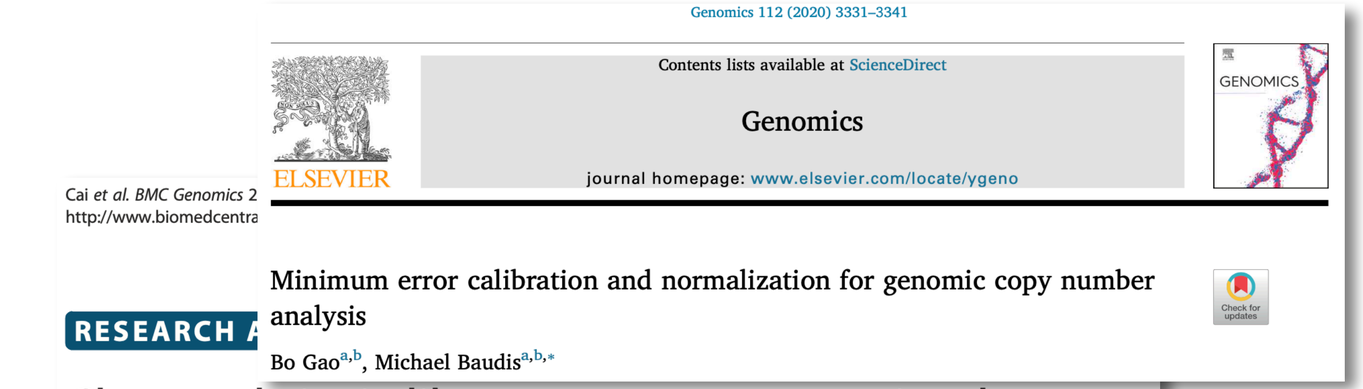
Publications (3324)		Samples				
id	Publication	cCGH	aCGH	WES	WGS	pgx
PMID:34103027	Peng G, Chai H, Ji W, Lu Y, Wu S et al. (2021) Correlating genomic copy number alterations with clinicopathologic findings in 75 cases of ... BMC Med Genomics	0	79	0	0	0
PMID:34059130	Tsui DWY, Cheng ML, Shady M, Yang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



Recent Publications

CNV Data Analysis & Methods

- collaborative projects utilizing the Progenetix data for multi-omics analyses
- data and bioinformatics analysis support for e.g. translational studies w/o "omics" focus



Progenetix in 2021

Cancer Genomics Reference Resource

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

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

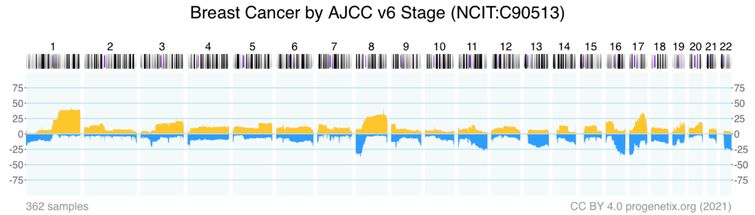
Beacon+

Progenetix Info

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

Cancer genome data @ progenetix.org

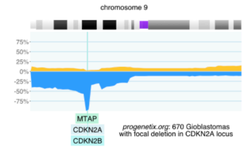
The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* from currently **139448** samples.



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

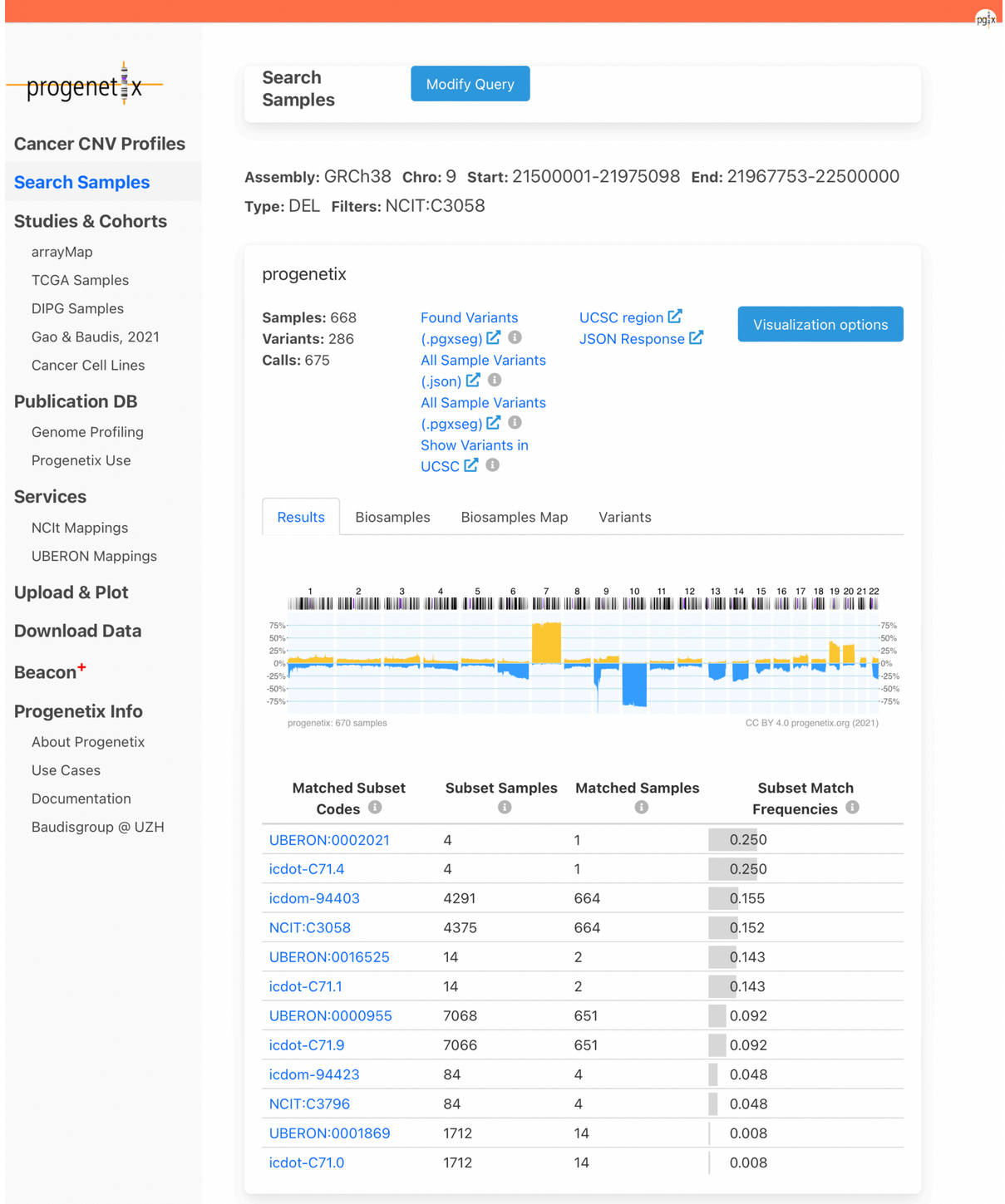
Through the [Publications] page Progenetix provides **4025** annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.

Progenetix in 2021

Cancer Genomics Reference Resource

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

progenetix

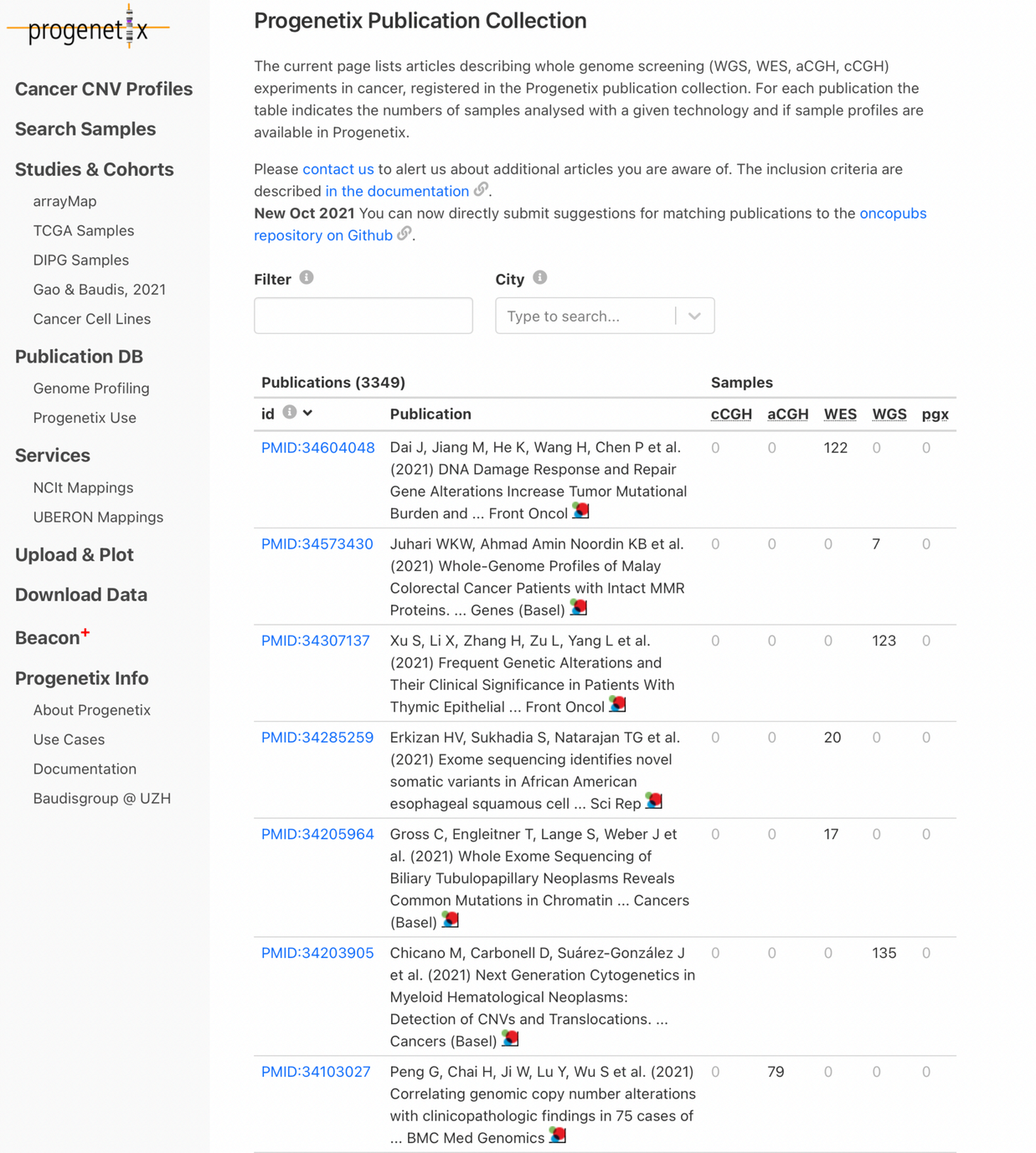


Progenetix in 2021

Cancer Genomics Reference Resource

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

progenetix

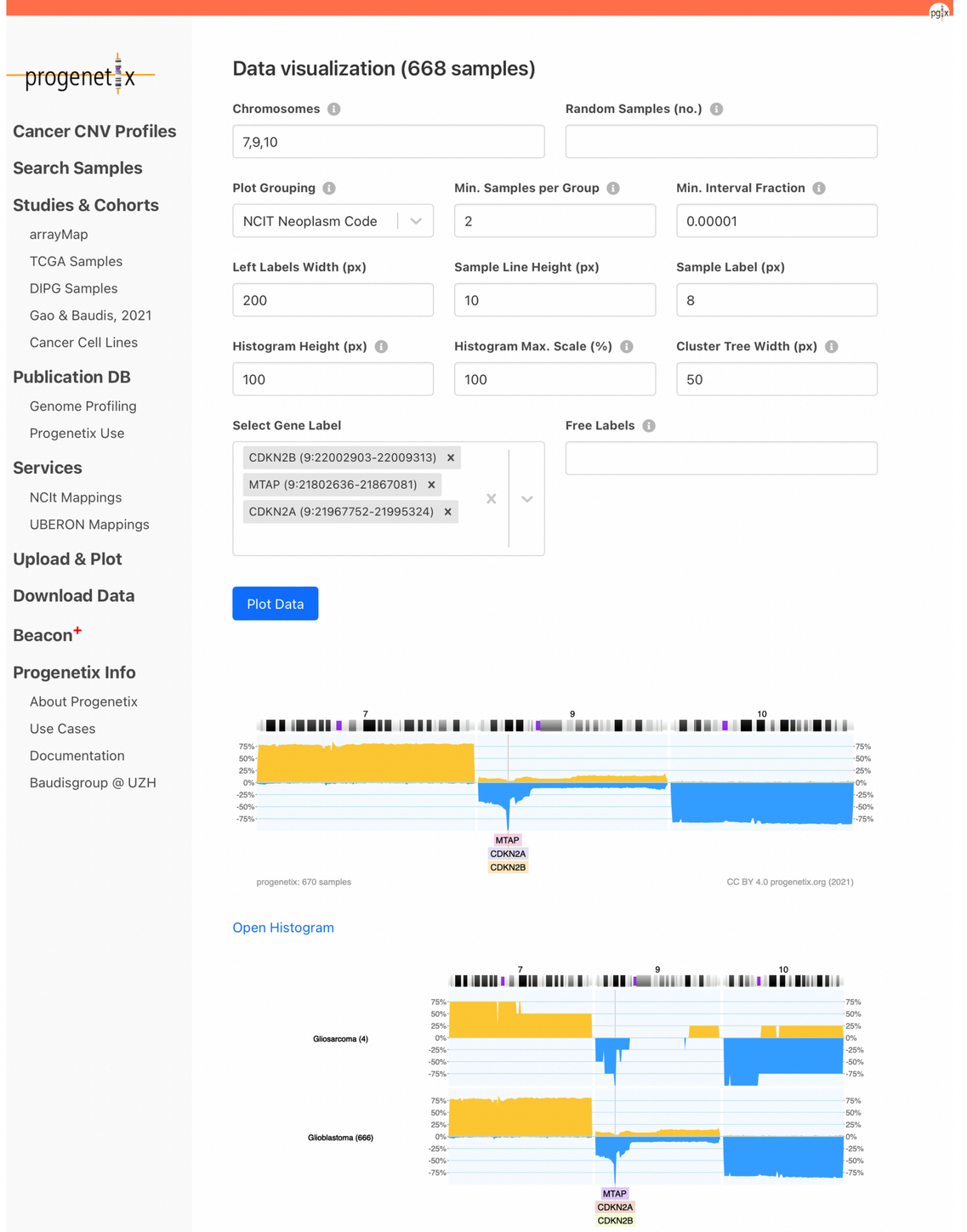


Progenetix in 2021

Cancer Genomics Reference Resource

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

progenetix

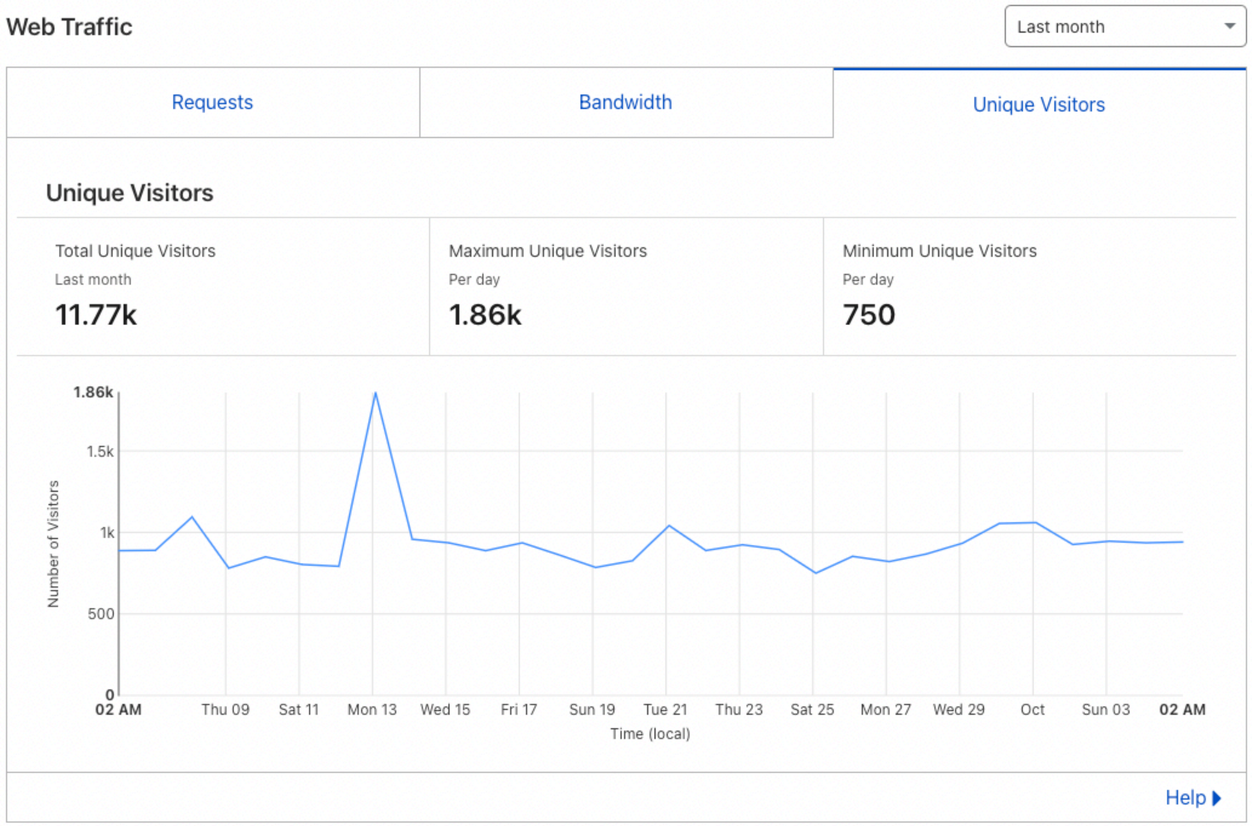


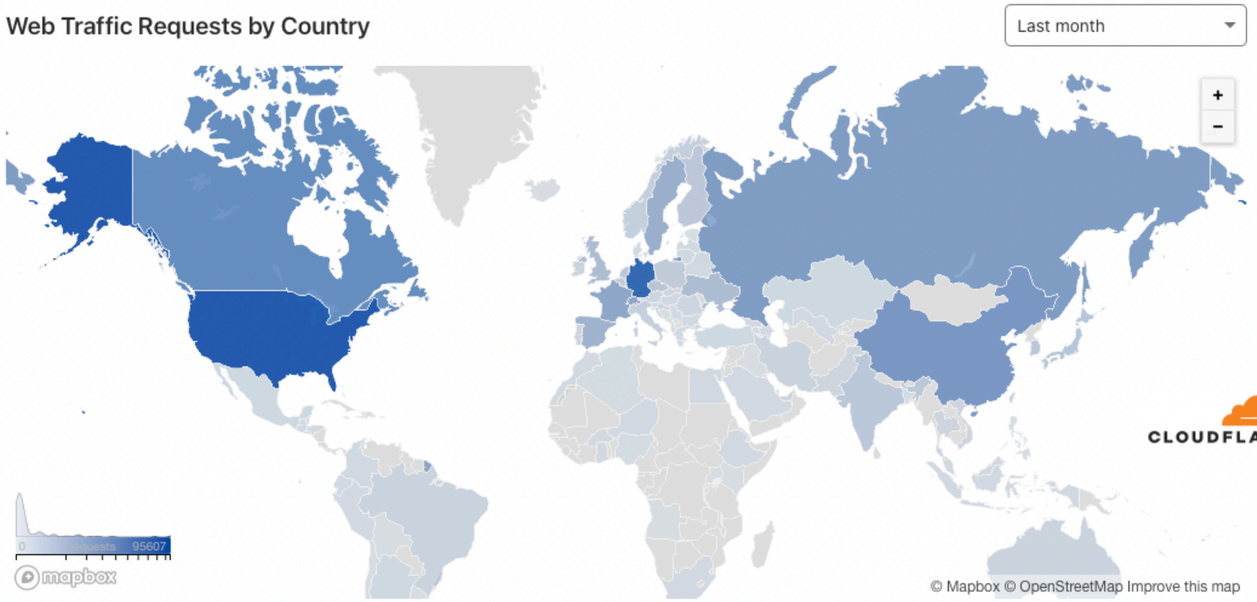
Progenetix in 2021

Cancer Genomics Reference Resource

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

progenetix





Cloudflare | progenetix.org | Sept-Oct 2021

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.

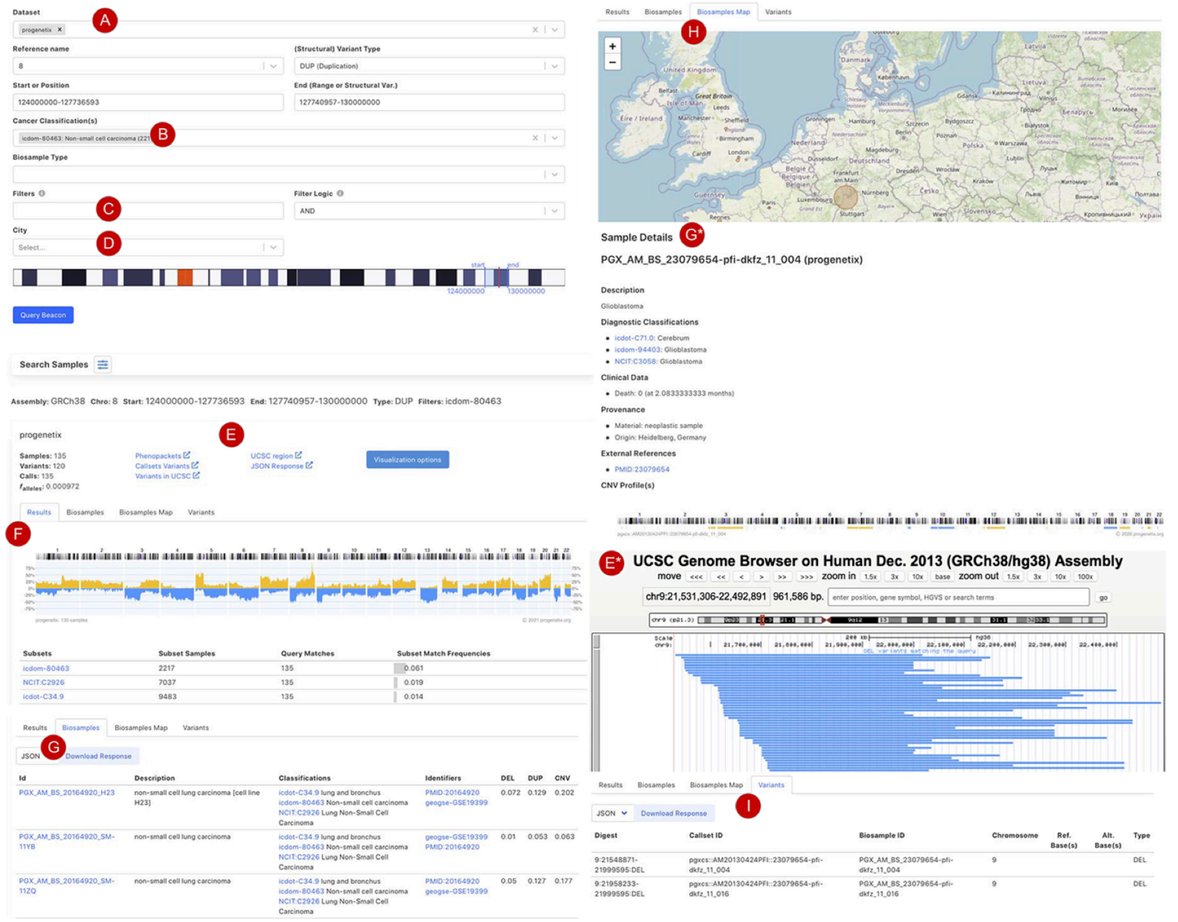
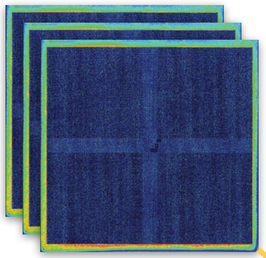
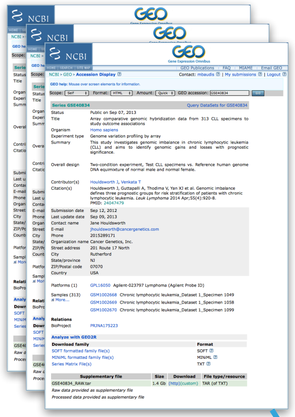


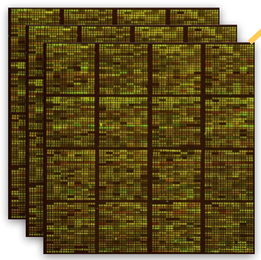
Figure 3. Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range This example queries for a continuous, focal duplication covering the complete MYC gene's coding region with < = 6 Mb in size. A: Filter for dataset; B: filter for cancer classification (NCIt and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

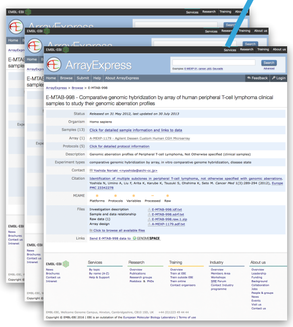
Data "Pipelines"

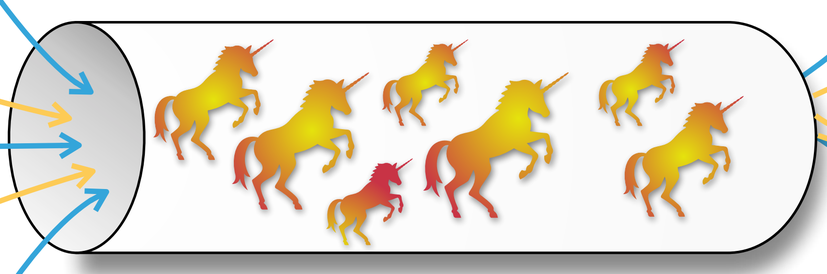
DATA PIPELINE

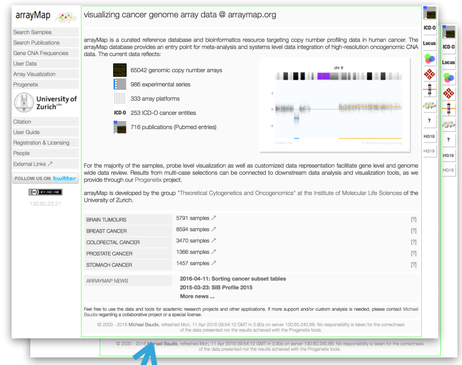


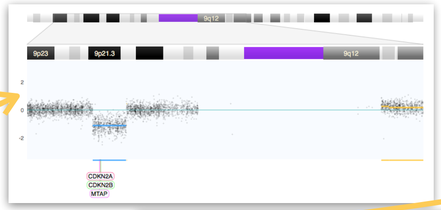


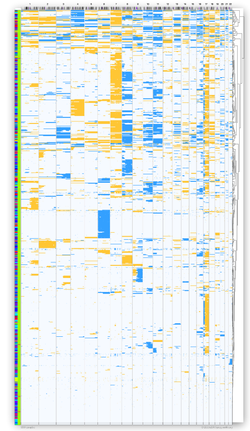


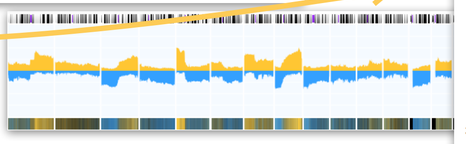


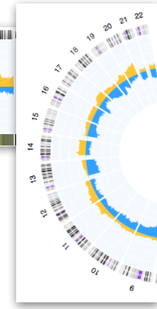


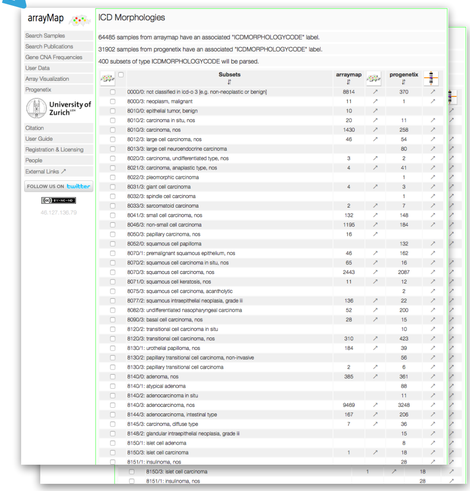












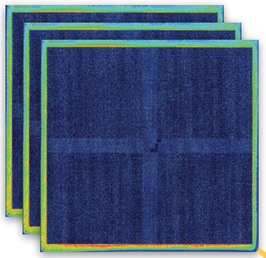
arrayMap

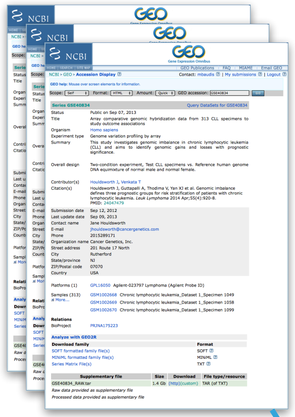
progenetix

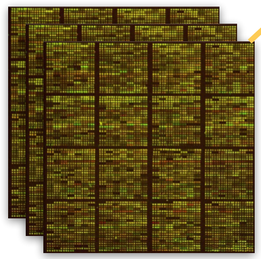
DATA PIPELINE

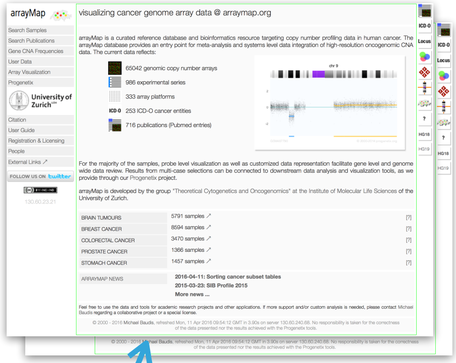
BIOCURATION

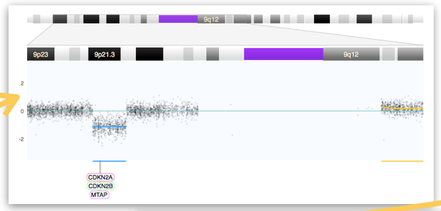
BIOINFORMATICS

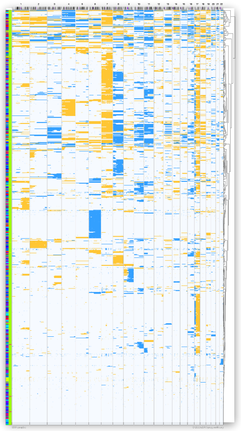


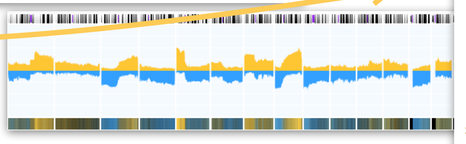


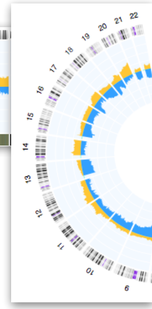


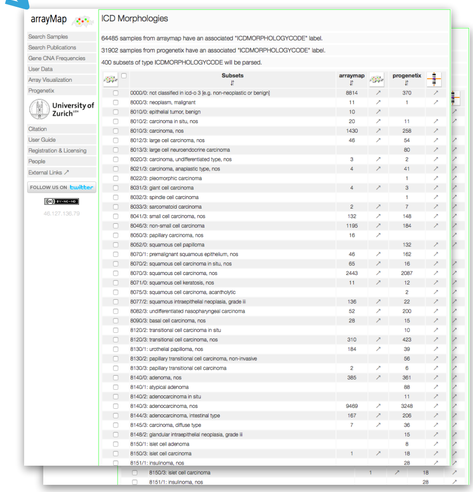








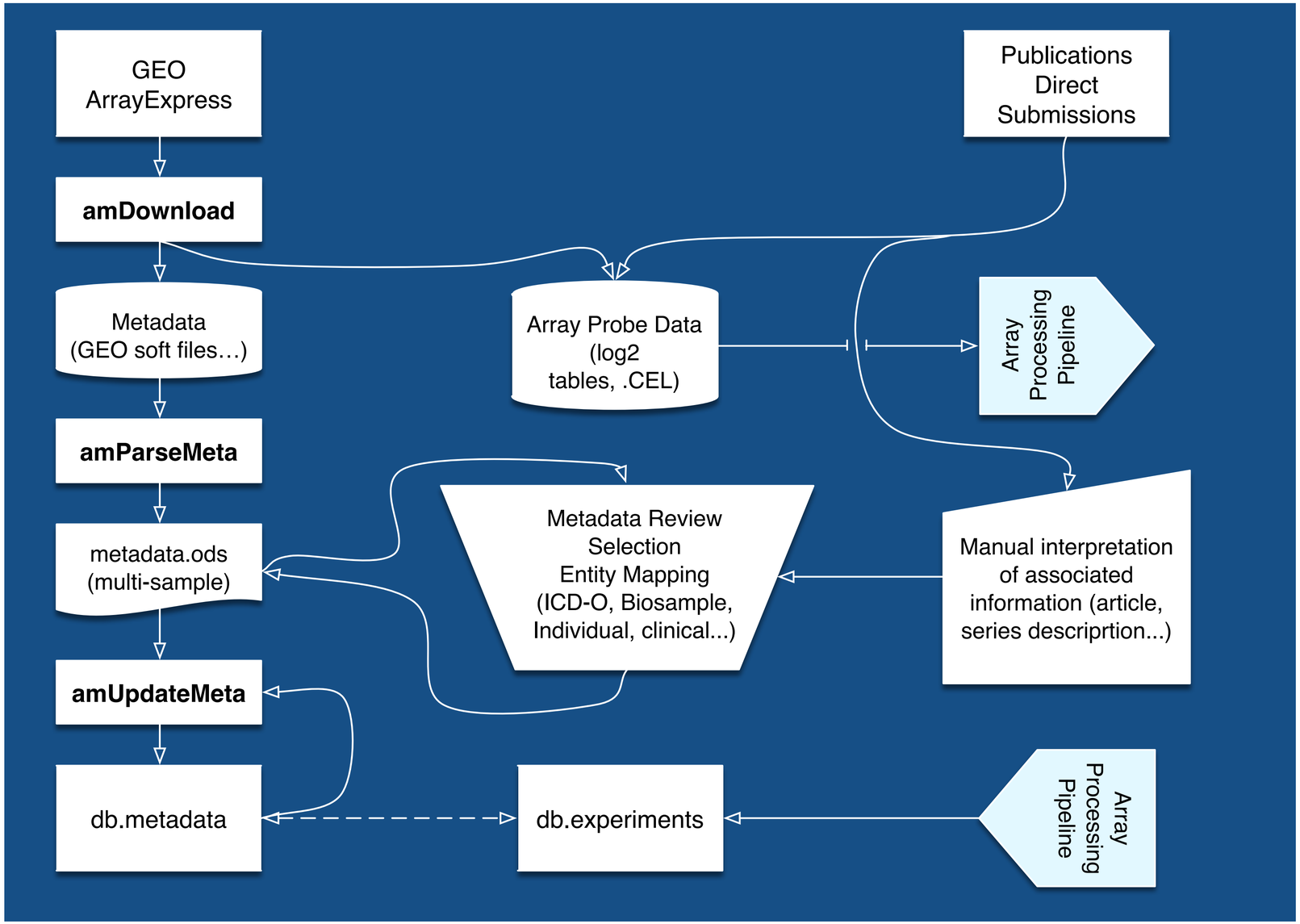




arrayMap

progenetix

Bioinformatics & Data Curation - arrayMap data “Pipeline”



Progenetix & arrayMap: 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 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 sta
manual 701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix).
!Sample_hyb_protocol = Hybridizations were performed according to the standard
701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix) using an Affymetrix
!Sample_scan_protocol = Scanning performed according to the standard Affymetrix
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://
!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
```

```
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/;
    }
```

```
$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...

Not yet registered way to express "alive"!

```
$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 }) }
```

Data Curation

Happy RegExing!

Source: https://xkcd.com/208/

```yaml
extraction_scopes:
  description: >-
    Detection and processing of clinical scopes goes through several stages:
    1. line cleanup - so far run for the input before processing the individual
    scopes
    2. line match using sme general pattern expected in all lines containing
    data for the current scope (`filter` pattern)
    3. finding and extracting the relevant data by looping over a list of
    specific patterns with memorized matches (`find`)
    4. post-processing using empirical cleanp replacements (`cleanup`)
    5. checking the correct structure (`final_check` - a global pattern can be
    used if other post-processing is performed)


  survival_status:
    filter: '(?i).*?(?:(?:dea(?:d|th))|alive|surviv|outcome|status)'
    preclean:
      - m: '(?i)days to death or last seen alive[^\w]+?\d+?(?:[^\w\.]|$)'
        s: ''
      - m: '[^\w]+?NA(?:[^\w\.]|$)'
        s: ' '
      - m: 'remission status past double induction .cr. complete remission. RD. refractory disease. ED. early death[^\w]+?ED'
        s: 'survival: dead'
      - m: 'remission status past double induction .cr. complete remission. RD. refractory disease. ED. early death[^\w]+?NA'
        s: ''
      - m: 'remission status past double induction .cr. complete remission. RD. refractory disease. ED. early death[^\w]+?CR'
        s: 'survival: alive'
      - m: 'remission status past double induction .cr. complete remission. RD. refractory disease. ED. early death[^\w]+?RD'
        s: '' # alive but not responding to therapy so removed?
      - m: 'Event Free Survival[^\w]+?no event'
        s: 'recurrence: no'
      - m: 'Event Free Survival.event'
        s: 'recurrence: yes'
      - m: 'Outcome[^\w]+?no event'
        s: 'survival: alive'
      - m: 'Outcome[^\w]+?event'
        s: 'survival: dead'
      - m: 'survival status[^\w]+?0'
        s: 'survival: dead'
      - m: 'survival status[^\w]+?1'
        s: 'survival: alive'
      - m: 'overall[^\w]+?survival[^\w]+?days[^\w]+?NA'
        s: ''
      - m: 'survival(?: time|from diagnosis)?[^\w]+?(days|months|years?)[^\w]+?(\d\d?\d?\d?\.?\d?\d?)'
        s: 'survival: \2\1'
```

Disease annotations in Progenetix

From some text, somewhere, to ontology classes

- **diagnostic categories** are the **most important** labels to associate with genomic observations
- original data almost *never* uses **modern**, **hierarchical** classification systems but provides circumstantial ("breast cancer in pre-menopausal...") or domain-specific ("CLL Binet B", "colorectal carcinoma Dukes C") information
- clinical classifications (ICD-10 ...) have very limited relation to tumor biology
- concepts change over time ...
- for cancer, the "International Classification of Diseases in Oncology" (**ICD-O 3**) by IARC / WHO traditionally has been a good compromise to map to - but with non-hierarchical structure and is used by international reference projects

NCIthesaurus

From Classification to Hierarchical Ontology: ICD-O -> NCIt

example_dx	ICDMORPHOLOGY	ICDOM	ICDTOPOGRAPHY	ICDOT	NCIT:CODE
malignant melanoma [metastatic cell line MaMel19]	Malignant melanoma NOS	8720/3	skin	C44	C3224
malignant melanoma [vagina]	Malignant melanoma NOS	8720/3	vagina and labia	C510	C3224
malignant melanoma [uvea metastasized]	Malignant melanoma NOS	8720/3	retina	C692	C3224
meningioma	Meningioma NOS	9530/0	meninges cerebral spinal	C700	C3230
mesothelioma	Mesothelioma NOS	9050/3	lung and bronchus	C34	C3234
pleural mesothelioma	Mesothelioma NOS	9050/3	pleura	C384	C3234
mesothelioma	Mesothelioma NOS	9050/3	connective and soft tissue NOS	C499	C3234
multiple myeloma	Plasma cell myeloma	9732/3	hematopoietic and reticuloendothelial syste	C42	C3242
Mycosis fungoides	Mycosis fungoides	9700/3	skin	C44	C3246
Myelodysplastic syndrome	Myelodysplastic syndrome NOS	9989/3	hematopoietic and reticuloendothelial syste	C42	C3247
Acute myeloblastic leukemia with maturation [FAB M2]	Acute myeloblastic leukemia with maturation [FAB M2]	9874/3	hematopoietic and reticuloendothelial syste	C42	C3250
neuroblastoma	Neuroblastoma NOS	9500/3	peripheral nervs incl. autonomous	C47	C3270
Cerebral neuroblastoma [cerebral region midline frontal lobe]	Neuroblastoma NOS	9500/3	cerebrum	C710	C3270
neuroblastoma [adrenal gland cell line]	Neuroblastoma NOS	9500/3	adrenal gland	C76	C3270
Cutaneous neurofibroma	Neurofibroma NOS	9540/0	skin	C44	C3272
Plexiform neurofibroma	Neurofibroma NOS	9540/0	Nervous system NOS	C729	C3272
Oligodendroglioma [Supratentorial Frontal Lobe]	Oligodendroglioma NOS	9450/3	cerebrum	C710	C3288
oilgodendroglioma	Oligodendroglioma NOS	9450/3	Brain NOS	C719	C3288
oligodendroglioma	Oligodendroglioma NOS	9450/3	brain nos	c719	C3288
Paraganglioma	Paraganglioma NOS	8680/1	Nervous system NOS	C729	C3308
paraganglioma	paraganglioma NOS	8680/1	adrenal cortex	C740	C3308

- since its beginning Progenetix samples have been classified using the 2 arms of the ICD-O system (morphology ~ histology/biology + topography ~ organ/tissue)
- over the last years we have established mappings between ICD-O code pairs and the NCIt "neoplasm" part of the NCI metathesaurus, thereby empowering hierarchical data structures for search and analysis

Ontologies and Classifications

Services: Ontologymaps (NCIt)

The **ontologymaps** service provides equivalency mapping between ICD-O and other classification systems, notably NCIt and UBERON. It makes use of the sample-level mappings for NCIT and ICD-O 3 codes developed for the individual samples in the Progenetix collection.

NCIT and ICD-O 3

While NCIT treats diseases as **histologic** and **topographic** described entities (e.g. **NCIT:C7700**: **Ovarian adenocarcinoma**), these two components are represented separately in ICD-O, through the **Morphology** and **Topography** coding arms (e.g. here **8140/3** + **C56.9**).

More documentation with focus on the API functionality can be found on the documentation pages.

The data of all mappings can be retrieved trough this API call: {JSON↗}

Code Selection

NCIT:C4337: Mantle Cell Lymphoma

Optional: Limit with second selection

Matching Code Mappings {JSON↗}

NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C77.9: Lymph nodes, NOS
NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C18.9: large intestine, excl. rectum and rectosigmoid junction
NCIT:C4337: Mantle Cell Lymphoma	pgx:icdom-96733: Mantle cell lymphoma	pgx:icdot-C42.2: Spleen

More than one code groups means that either mappings need refinements (e.g. additional specific NCIT classes for ICD-O T topographies) or you started out with an unspecific ICD-O M class and need to add a second selection.

In Progenetix all cancer diagnoses are coded to both NCIt neoplasm codes and ICD-O 3 Morphology + Topography combinations. The matched mappings are provided as lookup-service since neither an official ICD-O ontology nor such a "disease defined by ICD-O M+T" concept is codified anywhere.

List of filters recognized by different query endpoints

Public Ontologies with CURIE-based syntax

CURIE prefix	Code/Ontology	Examples
NCIT	NCIt Neoplasm[1]	NCIT:C27676
HP	HPO[2]	HP:0012209
PMID	NCBI Pubmed ID	PMID:18810378
geo	NCBI Gene Expression Omnibus[3]	geo:GPL6801, geo:GSE19399, geo:GSM491153
arrayexpress	EBI ArrayExpress[4]	arrayexpress:E-MEXP-1008
cellosaurus	Cellosaurus - a knowledge resource on cell lines [5]	cellosaurus:CVCL_1650
UBERON	Uberon Anatomical Ontology[6]	UBERON:0000992
cbioportal	cBioPortal[9]	cbioportal:msk_impact_2017

Private filters

Since some classifications cannot directly be referenced, and in accordance with the upcoming Beacon v2 concept of "private filters", Progenetix uses additionally a set of structured non-CURIE identifiers.

For terms with a `pgx` prefix, the identifiers.org resolver will

Filter prefix / local part	Code/Ontology	Example
pgx:icdom-...	ICD-O 3[7] Morphologies (Progenetix)	pgx:icdom-81703
pgx:icdot...	ICD-O 3[7] Topographies(Progenetix)	pgx:icdot-C04.9
TCGA	The Cancer Genome Atlas (Progenetix)[8]	TCGA-000002fc-53a0-420e-b2aa-a40a358bba37
pgx:pgxcohort-...	Progenetix cohorts [10]	pgx:pgxcohort-arraymap

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

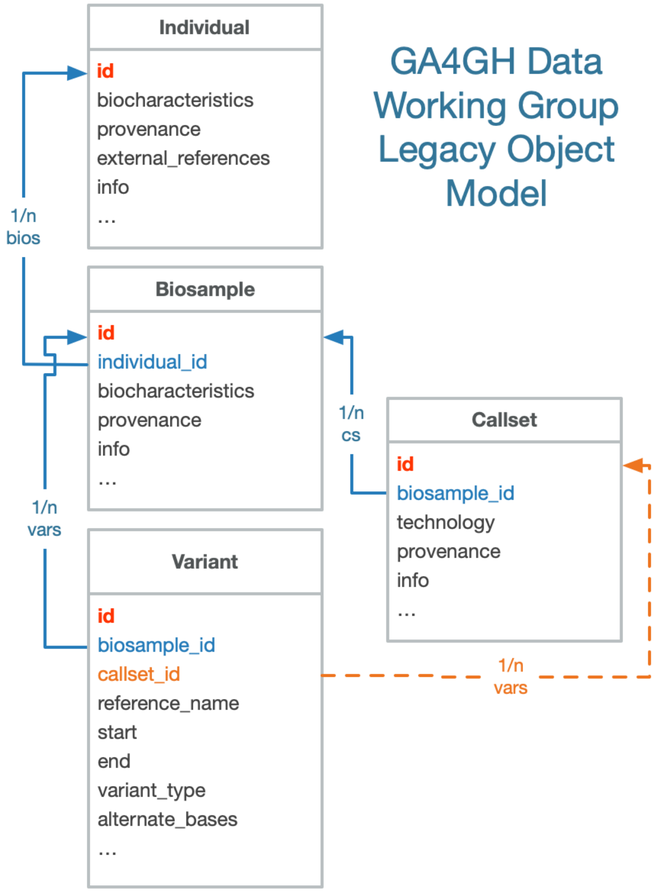
NCIt Neoplasm Core coded display (excerpt) for samples in the Progenetix cancer genome data resource allows sample selection on multiple hierarchy levels →

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

Database Structure

From flat database to hierarchical object storage



GA4GH Data Working Group Legacy Object Model

- collections in Progenetix MongoDB database reflect a consensus domain model for genomic data repositories
- flexible linking and object structure facilitates rapid change-overs
- BSON/JSON format in DB
  - equals data in JavaScript
  - "equals" objects in Python, Perl

➡ **rapid prototyping and implementation**

2021

- non-SQL document database (MongoDB)
- different object domains connected through identifiers
- data-driven website with JavaScript based frontend and data population through API calls

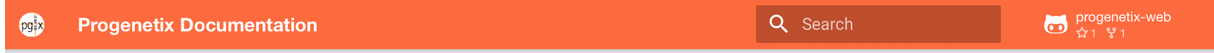

```
{
  "id" : "pgxind-kftx394x",
  "biocharacteristics" : [
    {
      "description" : "female",
      "type" : {
        "id" : "PATO:0020002",
        "label" : "female genotypic sex"
      }
    },
    {
      "description" : null,
      "type" : {
        "id" : "NCBITaxon:9606",
        "label" : "Homo sapiens"
      }
    }
  ],
  "data_use_conditions" : {
    "label" : "no restriction",
    "id" : "DUO:0000004"
  },
  "geo_provenance" : {
    "label" : "Salamanca, Spain",
    "precision" : "city",
    "city" : "Salamanca",
    "country" : "Spain",
    "latitude" : 40.43,
    "longitude" : -3.68
  },
  "info" : {
    "legacy_id" : "PGX_IND_SMZL01"
  },
  "updated" : ISODate("2018-09-26T09:51:39.775Z")
}


  "digest" : "7:107200000-158821424:DEL",
  "reference_name" : "7",
  "variant_type" : "DEL",
  "start" : 107200000,
  "end" : 158821424,
  "info" : {
    "cnv_value" : null,
    "cnv_length" : 51621424
  },
  "updated" : "2018-09-26 09:51:39.775397"
}
```

```
    "type" : {
      "id" : "UBERON:0002106",
      "label" : "spleen"
    }
  },
  {
    "type" : {
      "id" : "icdot-C42.2",
      "label" : "Spleen"
    }
  },
  {
    "type" : {
      "id" : "icdom-96893",
      "label" : "Splenic marginal zone B-cell lymphoma"
    }
  },
  {
    "type" : {
      "id" : "NCIT:C4663",
      "label" : "Splenic Marginal Zone Lymphoma"
    }
  }
],
"individual_id" : "pgxind-kftx394x",
"individual_age_at_collection" : "P67Y",
"info" : {
  "death" : "0",
  "followup_months" : 53,
 "callset_ids" : [
    "pgxcs-kftvv618"
  ],
  "legacy_id" : "PGX_AM_BS_SMZL01"
},
"external_references" : [
  {
    "type" : {
      "id" : "PMID:11337382"
    }
  }
],
"provenance" : {
  "material" : {
    "type" : {
      "id" : "EFO:0009656",
      "label" : "neoplastic sample"
    }
  },
  "geo" : {
    "label" : "Salamanca, Spain",
    "precision" : "city",
    "city" : "Salamanca",
    "country" : "Spain",
    "geojson" : {
      "type" : "Point",
      "coordinates" : [
        -3.68,
        40.43
      ]
    },
    "ISO-3166-alpha3" : "ESP"
  }
},
```

Progenetix Documentation

docs.progenetix.org

- information about concepts and practices
- API access
- Use cases
- Ontologies



Progenetix Cancer Genomics Resource Documentation

The Progenetix database and cancer genomic information resource contains genome profiles of more than 100'000 individual cancer genome screening experiments. The genomic profiling data was derived from genomic arrays and chromosomal Comparative Genomic Hybridization (CGH) 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.

Citation

Huang Q, Carrio-Cordo P, Gao B, Paloots R, Baudis M. (2021) **The Progenetix oncogenomic resource in 2021.** *Database (Oxford).* 2021 Jul 17
progenetix.org: **Progenetix oncogenomic online resource** (2022)

Additional Articles & Citation Options

Registration & Licenses

The **Progenetix** database and cancer genomic information resource was publicly launched in 2001, abnnounced through an article in *Bioinformatics*. The database & software are developed by the group of Michael Baudis at the University of Zurich and the Swiss Institute of Bioinformatics (SIB).

Additional information - e.g. about contacts or related publications - is available through the group page of the Baudis group at the University of Zürich. For a list of publication by the Baudis group you can go to the group's website, EuropePMC or any of the links here.

Progenetix Source Code

With exception of some utility scripts and external dependencies (e.g. MongoDB) the following projects provide the vast majority of the software (from database interaction to website) behind Progenetix and Beacon+.

`bycon`

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

`progenetix-web`

- website for Progenetix and its Beacon+ implementations
- provides Beacon interfaces for the `bycon` server, as well as other Progenetix sevices (e.g. the publications repository)
- implemented as React / Next.js project

`PGX` ¶

- a Perl library providing utility functions for Progenetix CNV data
- used for data transformation, e.g. binning of segmental CNV data
- main purpose now in providing the various plots (CNV histograms, clusterd CNV profiles, array plots)

progenetix

Summary: Offers & Needs

Progenetix Needs & Offers

What we have ...

✓ collection of >4000 articles assessed for scope
- training set for NLP & search engine generation

✓ cancer specific ontologies with cross-mappings (ICD-O vs. NCIt) based on >100k samples
- existing service API

✓ metadata ontology mappings for some 10k samples, with varying coverage for grade / stage / survival / ...

✓ CNV profiles for >110k samples, >700 entities with disease codes and metadata

✓ cell line CNV profiles together with mapped variants with clinical evidences

What we'd like...

➡ (semi-)automated detection of additional articles for scope (genome screening technologies, cancer samples, geographies)

➡ generation of a complete ICD-O terminology tree with NCIT (?) correspondence
- improved service API & publication

➡ improved annotations using smarter source (article, annotation files) pre-/processing

➡ correlation between individual profiles, profile heterogeneity and external parameters

➡ relation between cell lines and native tumor types, with consideration of non-CNV parameters

Progenetix Needs & Offers

What we have ...

- ✓ collection of >4000 articles assessed for scope
  - training set for NLP & search engine generation
- ✓ cancer specific ontologies with cross-mappings (ICD-O, NCIt) [illegible]
  - exist[illegible]
- ✓ metadata ontology mappings for some 10k samples, with varying coverage for grade / stage / survival / ...
- ✓ CNV profiles for >110k samples, >700 entities with disease codes and metadata
- ✓ cell line CNV profiles together with mapped variants with clinical evidences

Open Access API

What we'd like...

- ➡ (semi-)automated detection of additional articles for scope (genome screening technologies, cancer samples, geographies)
- ➡ generation of a complete ICD-O terminology tree with NCIT (?) correspondence
  - improved service API & publication
- ➡ improved annotations using smarter source (article, annotation files) pre-/processing
- ➡ correlation between individual profiles, profile heterogeneity and external parameters
- ➡ relation between cell lines and native tumor types, with consideration of non-CNV parameters

Progenetix Needs & Offers

Publication resource

- collection of >4000 articles assessed for scope
- training set for NLP & search engine generation
- identification of content-similar articles for expansion of publication resource
- text-analysis for resource identifiers (e.g. GEO accessions...), geographic term, clinical annotations
  - ➡ for new data and in comparison to existing datasets in Progenetix (e.g. hints for usable metadata not yet in DB)
  - ➡ **Opening up Publication resource for wider/ more general use?**

What we'd like...

- ➡ (semi-)automated detection of additional articles for scope (genome screening technologies, cancer samples, geographies)
- ➡ generation of a complete ICD-O terminology tree with NCIT (?) correspondence
  - improved service API & publication
- ➡ improved annotations using smarter source (article, annotation files) pre-/processing
- ➡ correlation between individual profiles, profile heterogeneity and external parameters
- ➡ relation between cell lines and native tumor types, with consideration of non-CNV parameters

NCIt <=> ICDO: Complete Mapping w/ Resource

NCIT

```
Neoplasm by Site (R_1)
        Breast Neoplasm (C2910)
                Breast Adenoma (C40382)
                        Breast Apocrine Adenoma (C40383)
                        Breast Ductal Adenoma (C40384)
                        Breast Pleomorphic Adenoma (C40408)
                        Breast Tubular Adenoma (C62210)
                                Lactating Adenoma (C9473)
                Breast Carcinoma (C4872)
                        Adenoid Cystic Breast Carcinoma (C5130)
                        Bilateral Breast Carcinoma (C8287)
                        Breast Adenocarcinoma (C5214)
                                Ductal Breast Carcinoma (C4017)
                                        Acinic Cell Breast Carcinoma (C40367)
                                        Ductal Breast Carcinoma In Situ (C2924)
                                        Invasive Ductal Carcinoma, Not Otherwise Specified (C4194)
                                        Medullary Breast Carcinoma (C9119)
                                        Mucinous Breast Carcinoma (C9131)
                                        Papillary Breast Carcinoma (C9134)
                                                Intraductal Papillary Breast Carcinoma (C4190)
                                                Invasive Papillary Breast Carcinoma (C36085)
                                        Secretory Breast Carcinoma (C4189)
                                        Tubular Breast Carcinoma (C9135)
                                Inflammatory Breast Carcinoma (C4001)
                                Lobular Breast Carcinoma (C3771)
                                        Invasive Lobular Breast Carcinoma (C7950)
                                        Lobular Breast Carcinoma In Situ (C4018)
                                Mixed Lobular and Ductal Breast Carcinoma (C5160)
                                Paget Disease of the Breast (C47857)
                        Breast Carcinoma by Gene Expression Profile (C53553)
                                Basal-Like Breast Carcinoma (C53558)
                                HER2 Positive Breast Carcinoma (C53556)
                                Luminal A Breast Carcinoma (C53554)
                                Luminal B Breast Carcinoma (C53555)
                                Normal Breast-Like Subtype of Breast Carcinoma (C53557)
                                Triple-Negative Breast Carcinoma (C71732)
                        Breast Small Cell Carcinoma (C6760)
                        Female Breast Carcinoma (C2918)
                        Hereditary Breast Carcinoma (C4503)
                        Male Breast Carcinoma (C3862)
                        Nipple Carcinoma (C28432)
                        Sporadic Breast Carcinoma (C7566)
                        Squamous Cell Breast Carcinoma (C5177)
                        Unilateral Breast Carcinoma (C46073)
                Breast Fibroadenoma (C3744)
                        Breast Complex Fibroadenoma (C5194)
                        Breast Giant Fibroadenoma (C4273)
                        Breast Intracanalicular Fibroadenoma (C4271)
                        Breast Juvenile Fibroadenoma (C4276)
                        Breast Pericanalicular Fibroadenoma (C4272)
```

ICDO (T/M)

Abdomen

C76.2	NOS
C47.4	autonomic nervous system
C49.4	connective tissue
C49.4	muscle
C47.4	peripheral nerve
C44.5	skin
C49.4	subcutaneous tissue

Abdominal

C49.4	aorta
C15.2	esophagus
C77.2	lymph node
C49.4	vena cava
8822/1	Abdominal desmoid
8822/1	Abdominal fibromatosis

Abdominal wall

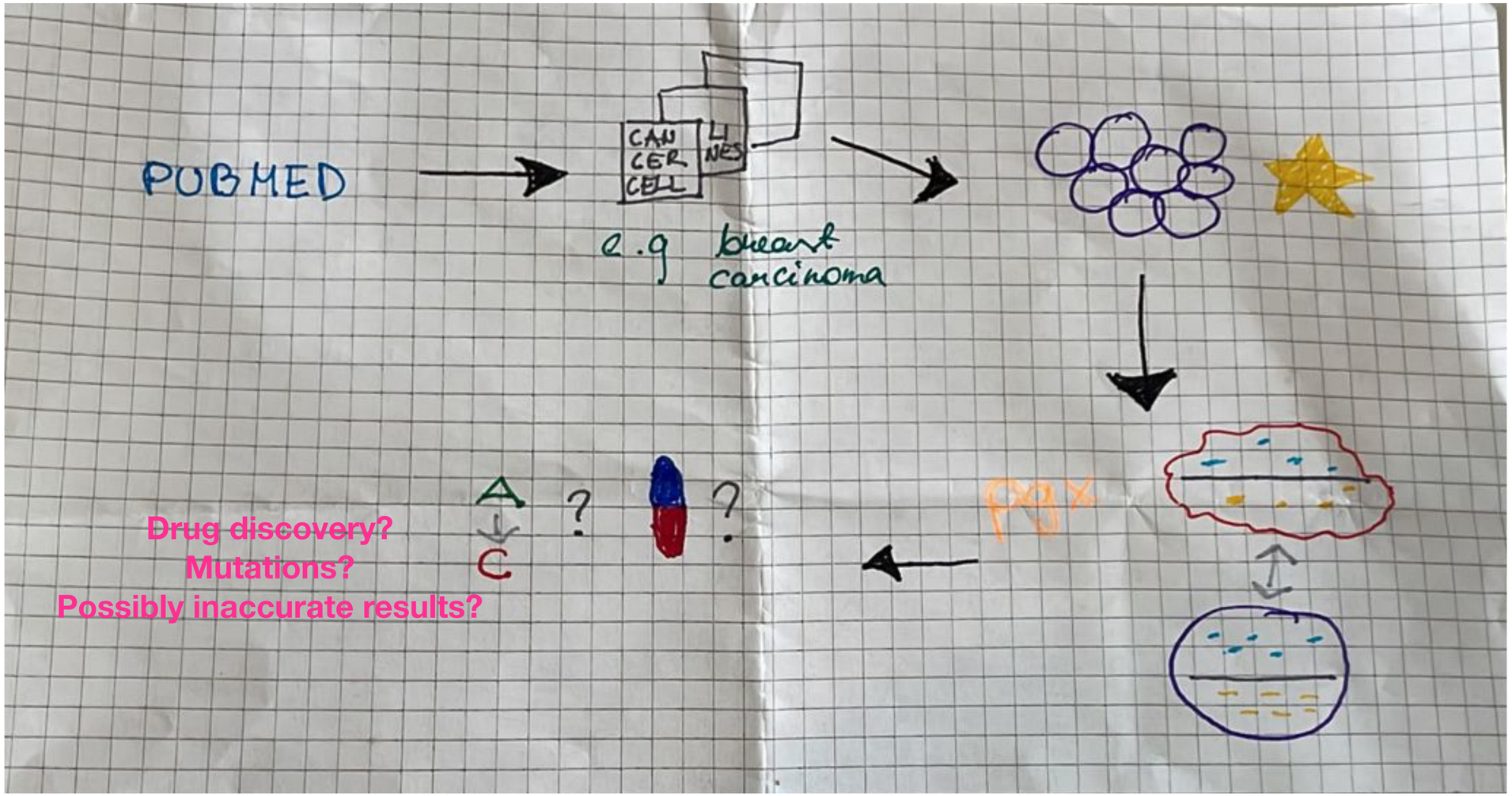
C76.2	NOS
C44.5	NOS (carcinoma, melanoma, nevus)
C49.4	NOS (sarcoma, lipoma)
C49.4	adipose tissue

Mature B-cell Neoplasms

9737/3	ALK positive large B-cell lymphoma
9680/3	B-cell lymphoma, unclassifiable, with features intermediate between diffuse large B-cell lymphoma and Burkitt lymphoma
9596/3	B-cell lymphoma, unclassifiable, with features intermediate between diffuse large B-cell lymphoma and classical Hodgkin lymphoma
9833/3	B-cell prolymphocytic leukemia
9687/3	Burkitt lymphoma
9823/3	Chronic lymphocytic leukemia/small lymphocytic lymphoma
9680/3	Diffuse large B-cell lymphoma (DLBCL), NOS
9699/3	Extranodal marginal zone lymphoma of mucosa-associated lymphoid tissue (MALT lymphoma)
9734/3	Extraosseous plasmacytoma
9690/3	Follicular lymphoma
9940/3	Hairy cell leukemia
9762/3	Heavy chain diseases (alpha, gamma, mu)
9712/3	Intravascular large B-cell lymphoma
9738/3	Large B-cell lymphoma arising in HHV8-associated multicentric Castleman disease
9766/1	Lymphomatoid granulomatosis
9671/3	Lymphoplasmacytic lymphoma
9673/3	Mantle cell lymphoma
9699/3	Nodal marginal zone lymphoma
9591/3	Non-Hodgkin lymphoma, NOS; Splenic B-cell lymphoma/leukemia, unclassifiable

- expand our data-driven mapping to a complete coverage of all cancer codes
- integrate with Monarch initiative

Cell lines


PUBMED
CAN CER CELL
LI NES
e.g breast carcinoma
A
C
?
?
PCR
Drug discovery?
Mutations?
Possibly inaccurate results?

Progenetix and GA4GH Beacon

Implementation driven development of a GA4GH standard

progenetix

GENOMICS

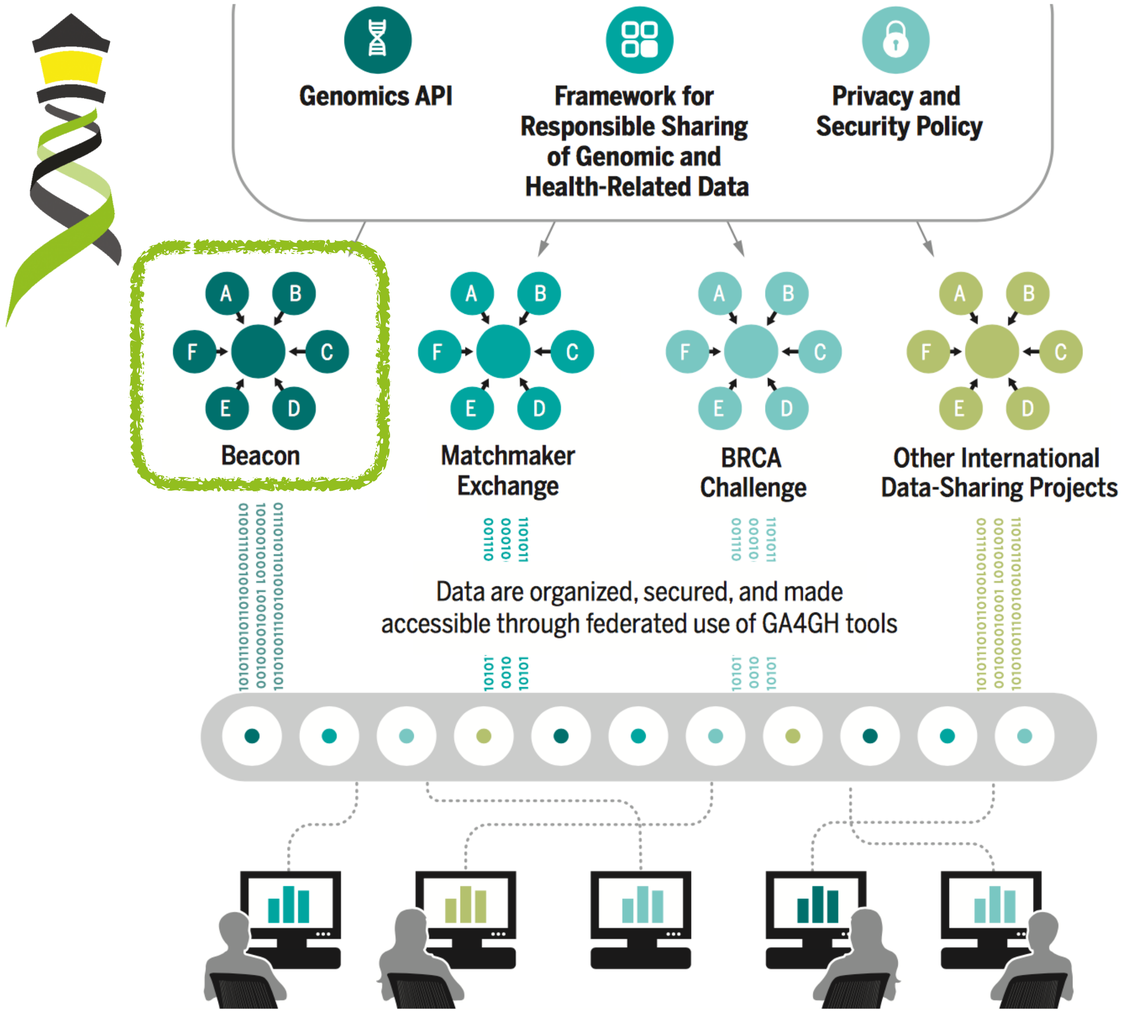
A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

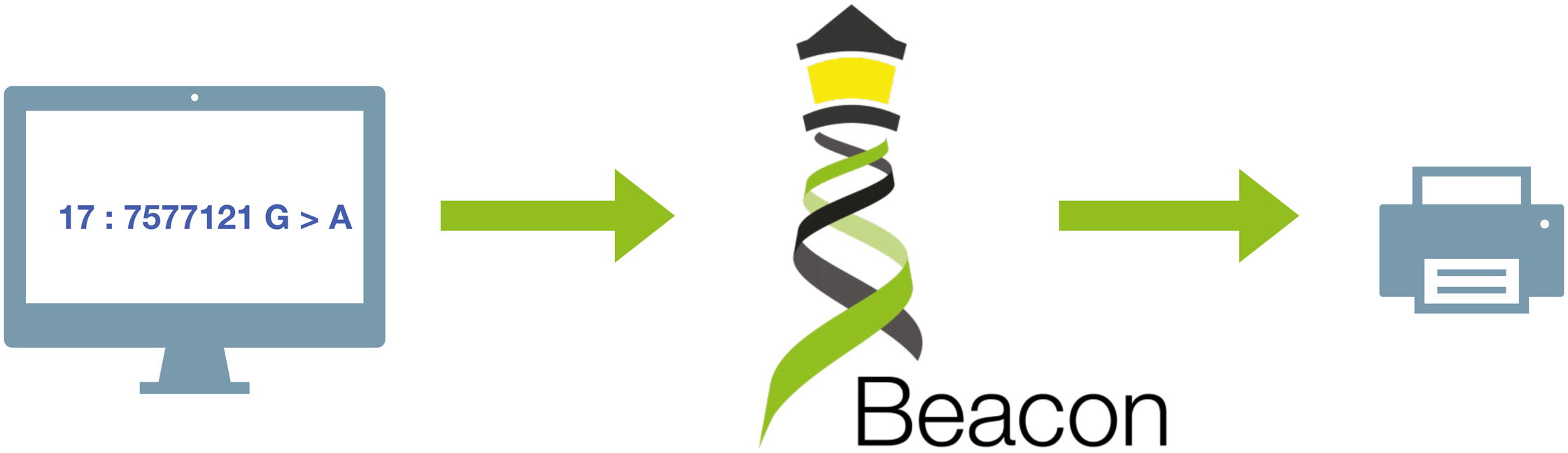
SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

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



DNASTACK

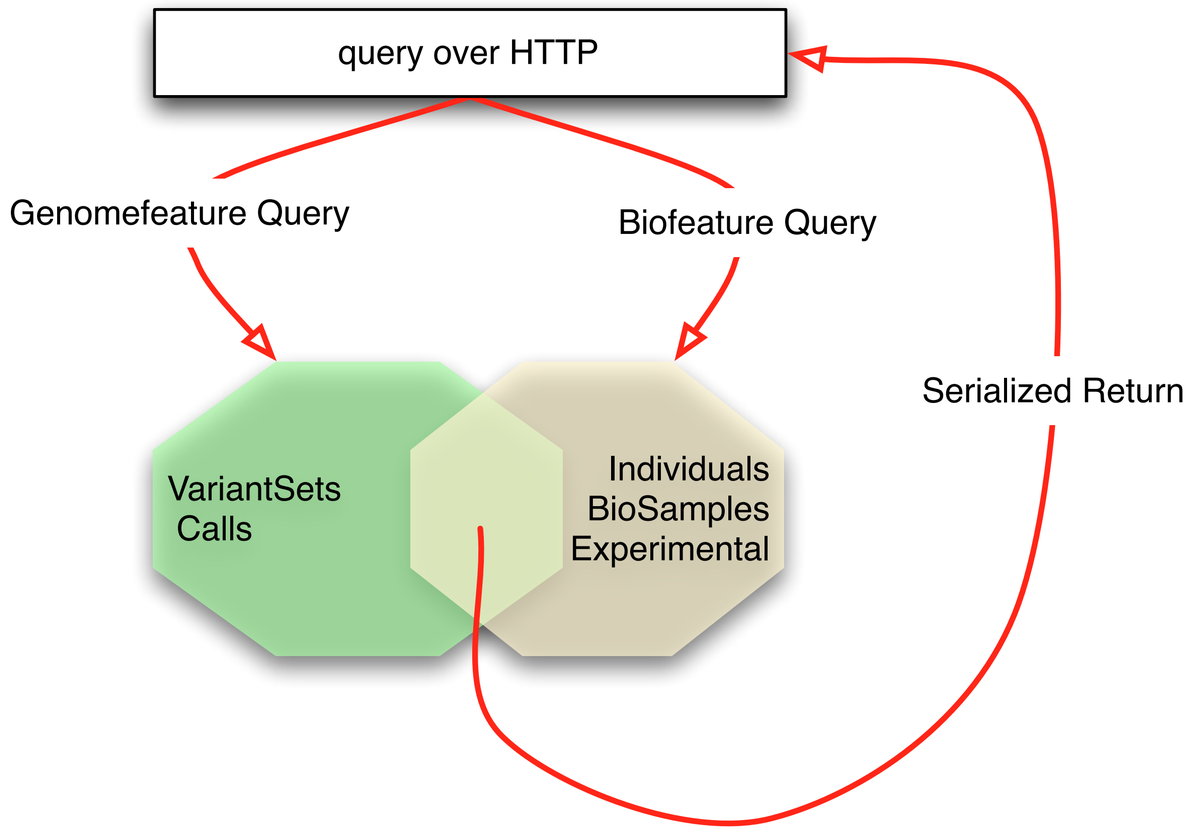
Global Alliance for Genomics & Health



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

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

Minimal GA4GH query API structure



Beacon(+) “Metadata”

Beacon v2 Filters

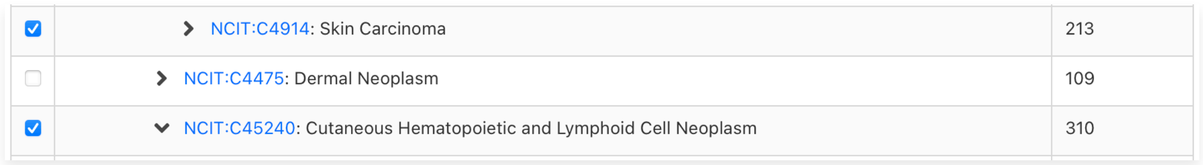
Example: Use of hierarchical classification systems (here NCIt neoplasm core)

- 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**

Beacon v2 Requests

POSTing Queries

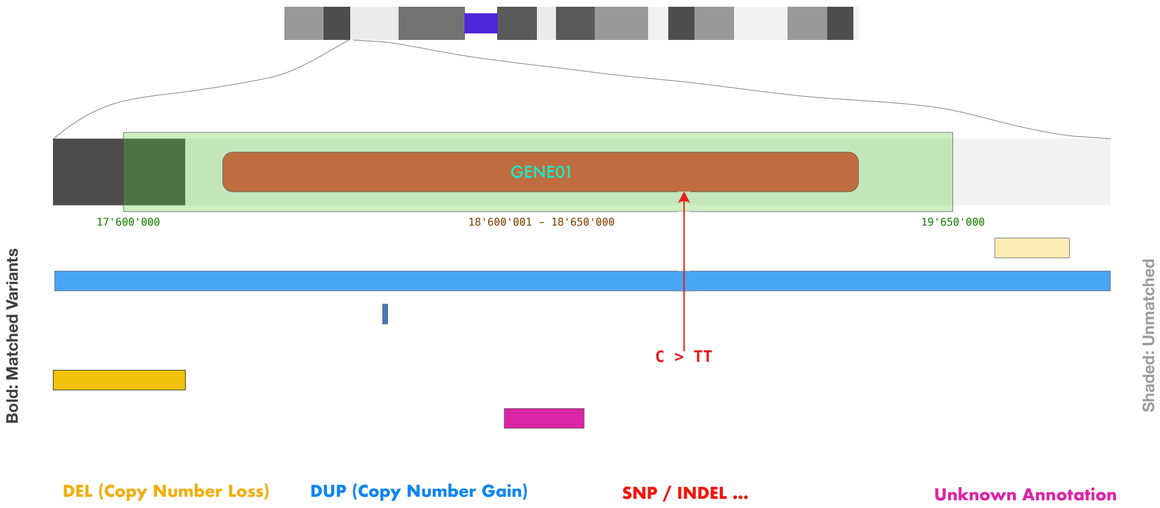
- Beacon v2 supports a mix of dedicated endpoints with REST paths
- POST requests using JSON query documents
- final syntax for core parameters still in testing stages

```
{
    "$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

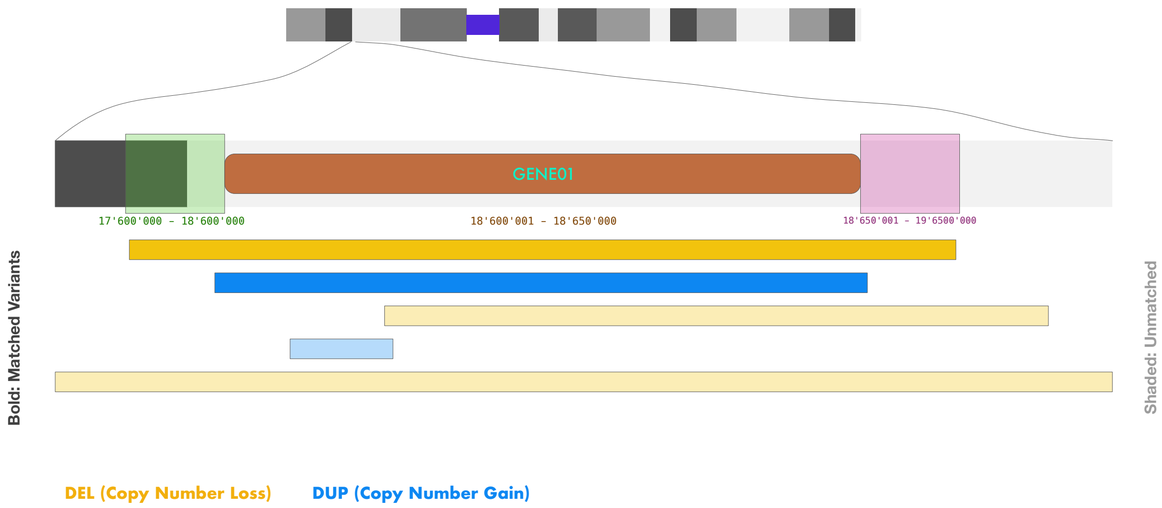
Range and Bracket queries enable positional wildcards and fuzziness

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)

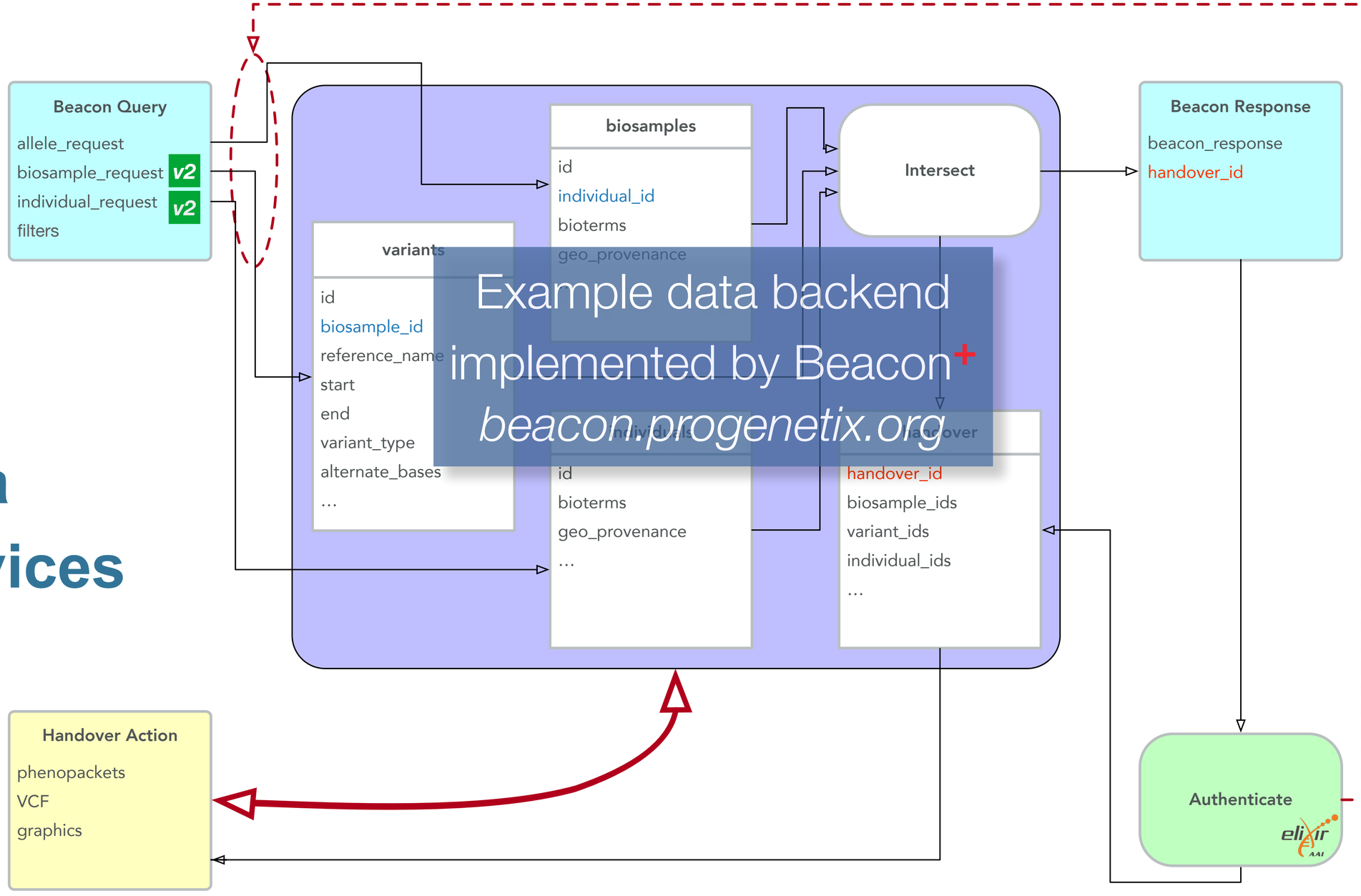
Genome Bracket Query
(full match)



- 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 & Handover

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



Beacon I/O

Handover

Authentication

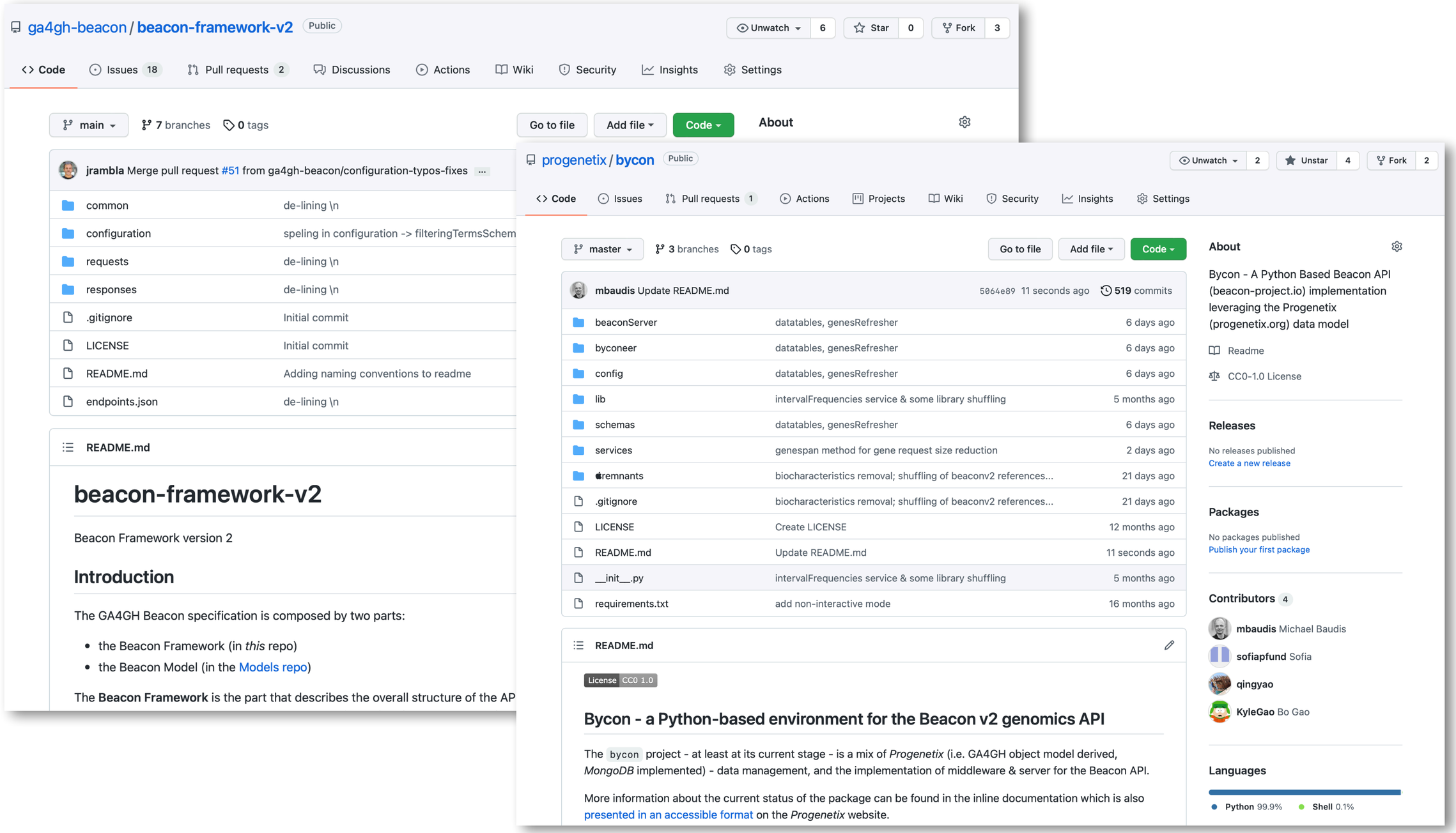
Michael Baudis

Global Alliance for Genomics & Health

SIB

University of Zurich UZH

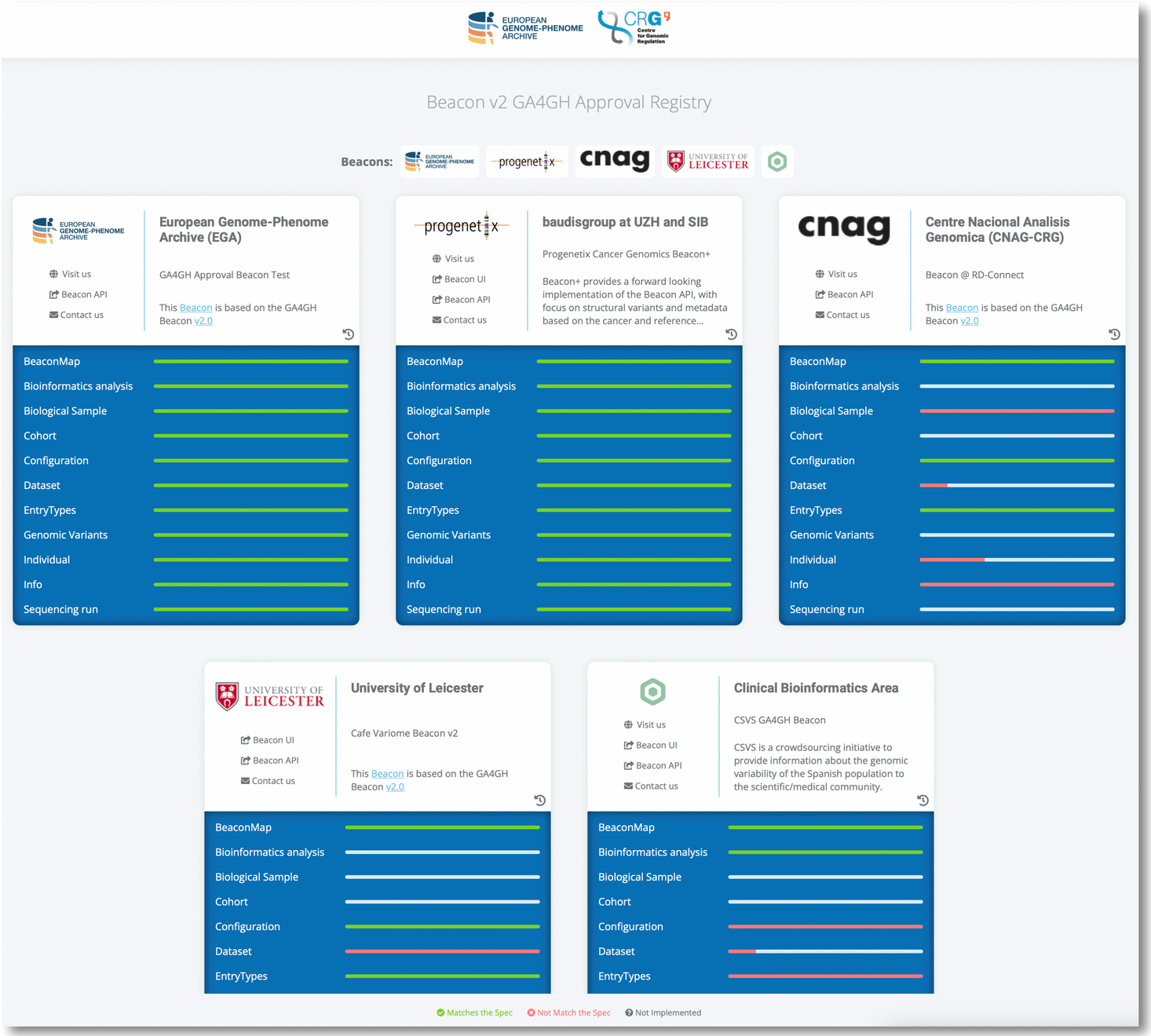
elixir


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%

Onboarding

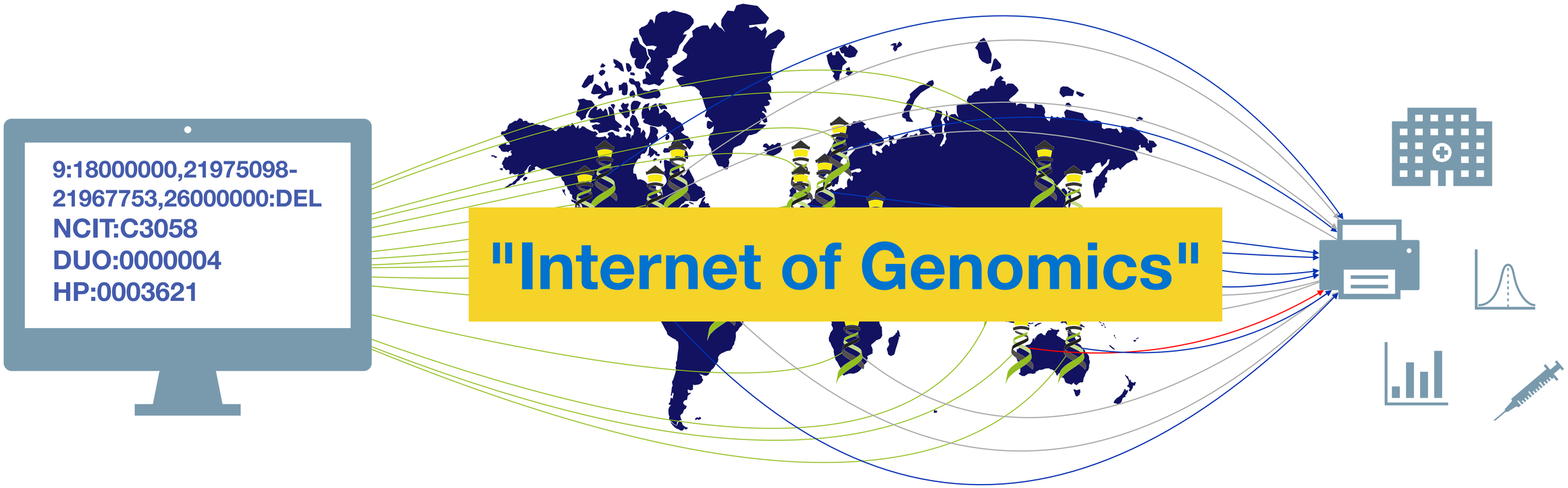
Demonstrating Compliance

- Progenetix Beacon+ has served as implementation driver since 2016
- Beacon v2 as service with protocol-driven registries for federation
- GA4GH approval process in the Spring 2022 session



Compliance server at EGA 2021-11-20

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**".