Building a Genomics Resource

Progenetix - From Experiments to APIs

Michael Baudis | UZH BIO390 HS25

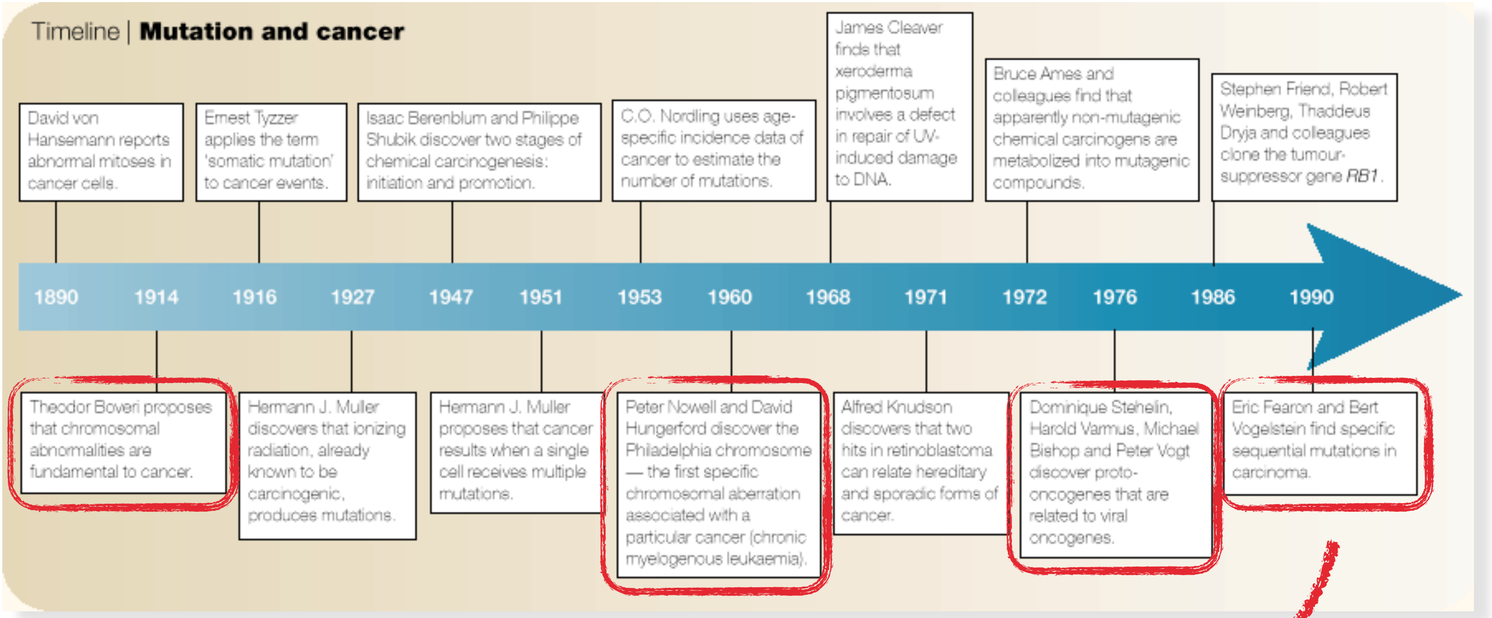
progenetix

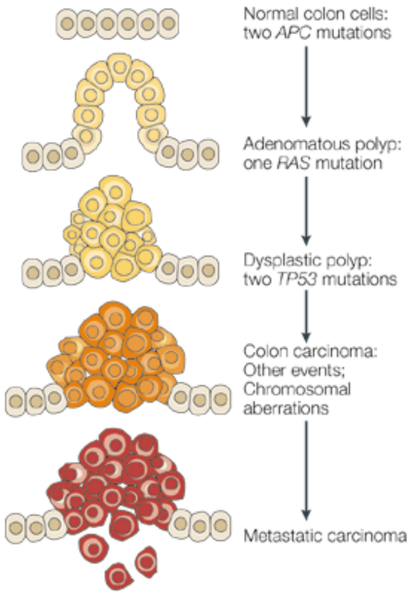
Building a Genomics Resource

A (personal) journey through time...

- Genomic Copy Number Variations in cancer (CNA / CNV)
- Comparative Genomic Hybridization (CGH) as original CNV screening technique
- CNVs differ between cancer (sub)types and may correlate to clinical outcome
- single studies are limited- **let's build a database**
- databases should be accessible - **let's move online**
- **more data** - data parsers & text mining
- **visualization** - graphics libraries and data formatting
- large datasets - access through **APIs**

progenetix





Cancers are based on acquired and inherited genomic mutations

Knudson, A. G. (2001). Two genetic hits (more or less) to cancer. Nature Reviews Cancer, 1(2), 157–162.

Theodor Boveri (1914)

Observations in sea urchin eggs

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

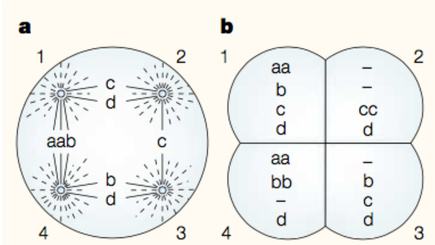


Figure 2 | **Multiple cell poles cause unequal segregation of chromosomes. a** | Boveri showed that fertilization of sea-urchin eggs by two sperm results in multiple cell poles. Individual chromosomes then attach to different combinations of poles — for example, one copy of chromosome c is attached to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

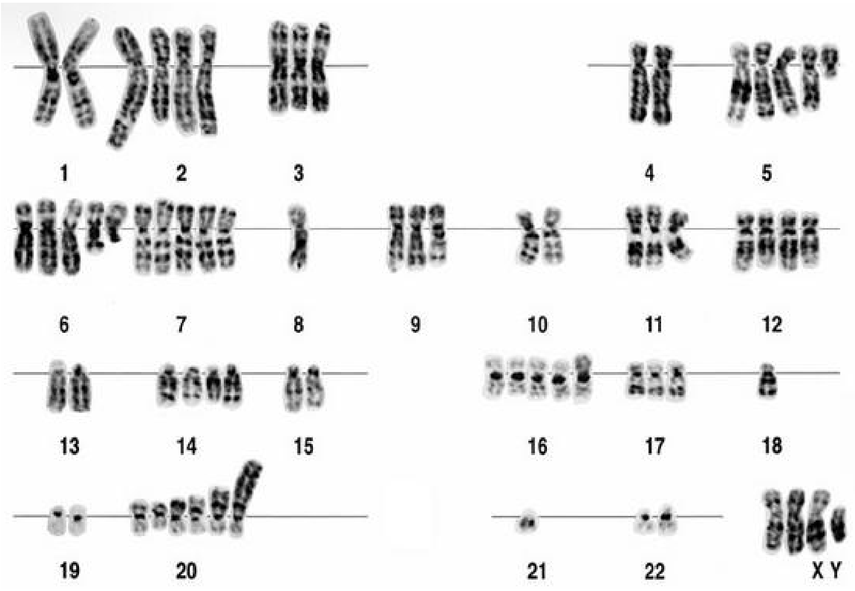
Allan Balmain
Cancer genetics: from Boveri and Mendel to microarrays.
NatRev Cancer (2001); 1: 77-82

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

Genomic changes at the DNA level are hallmarks of cancer

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

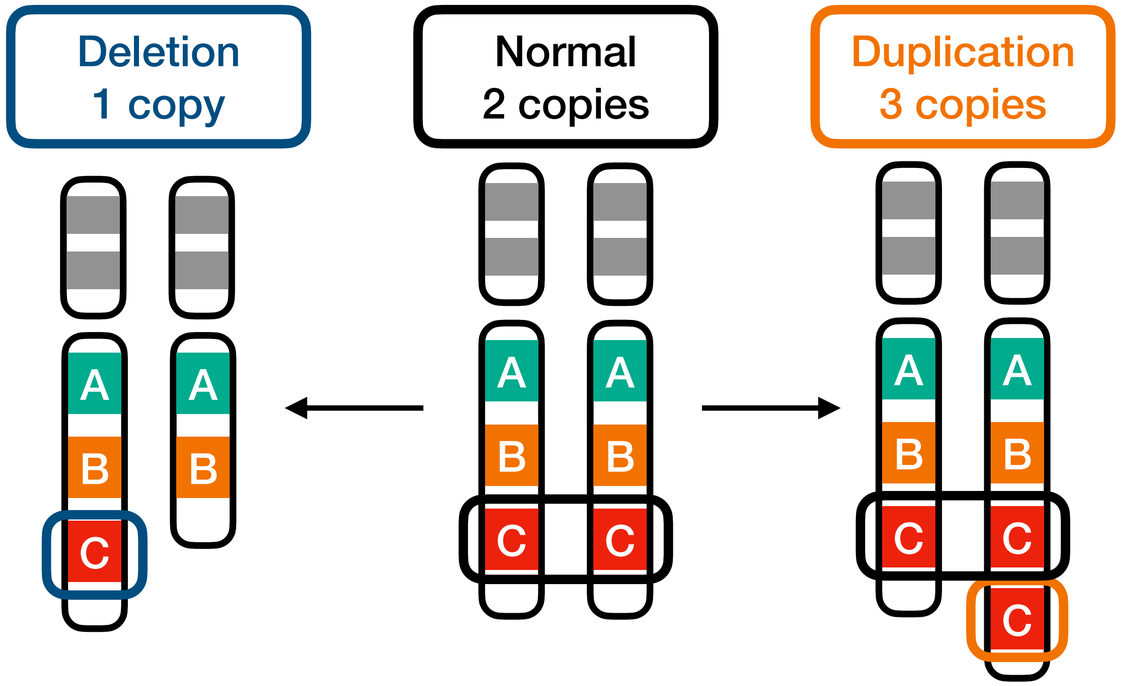
Normal karyotype



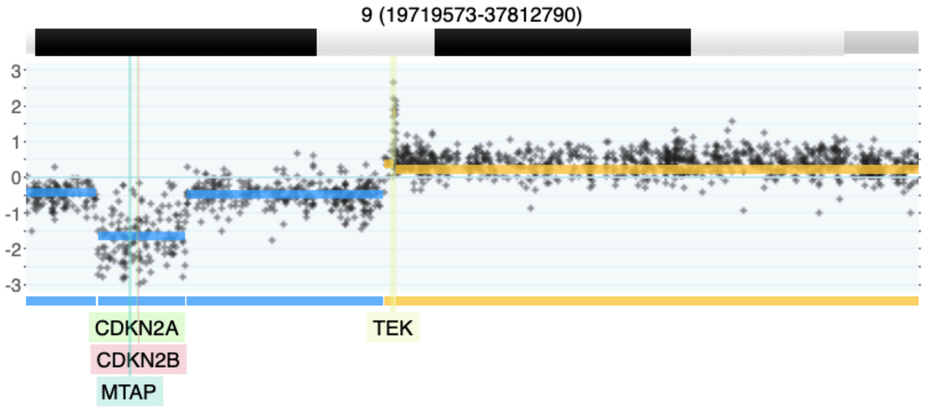
Tumor karyotype

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

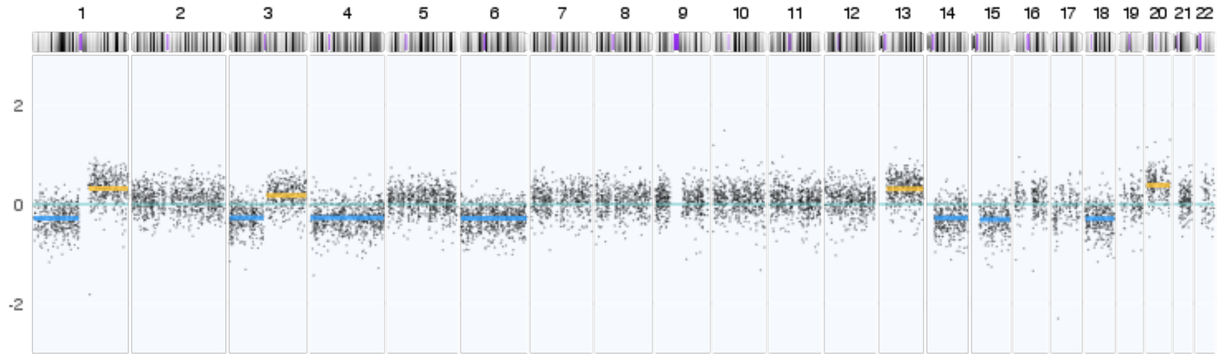
Copy Number Variant (CNV)



- Intermediate-scale genetic change
- Size: 1kb to multiple megabase
- Additional copies of sequence (duplications) and losses of genetic material (deletions)

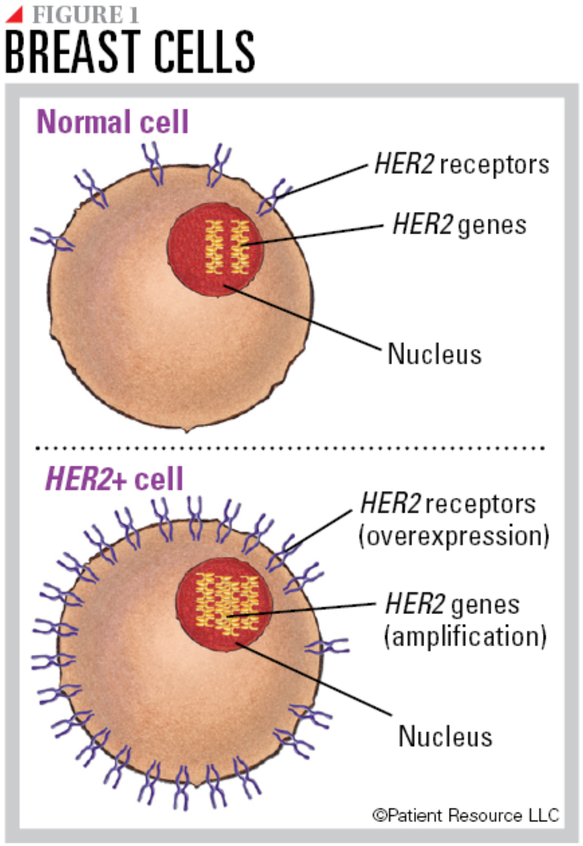


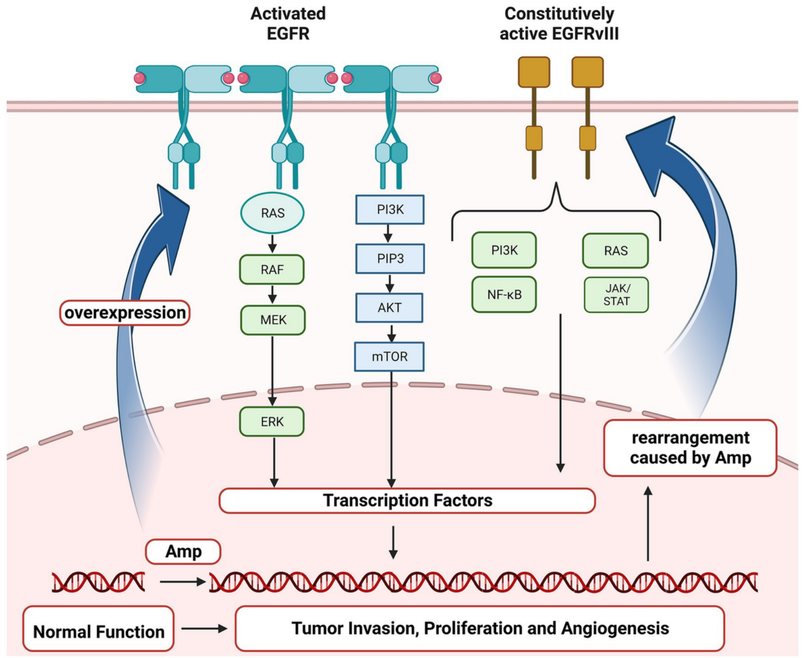
2-event, homozygous deletion in a Glioblastoma

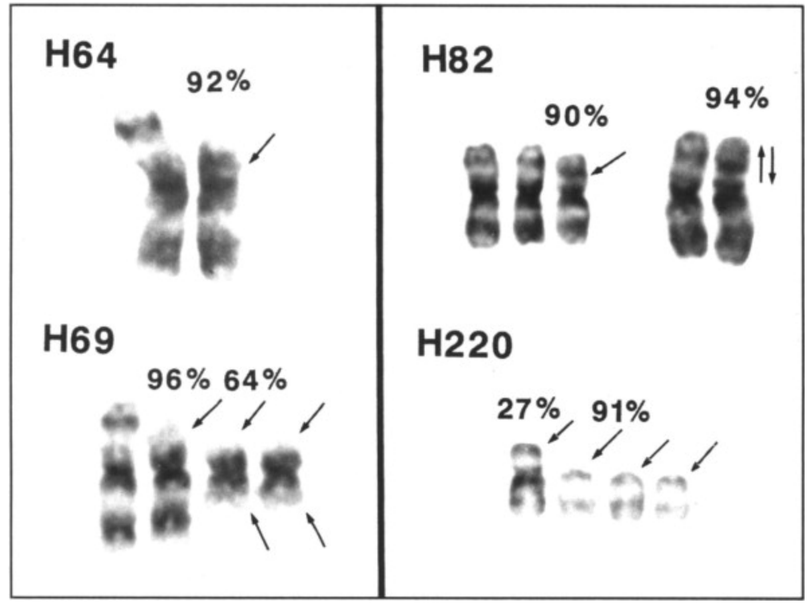


Gain of chromosome arm 13q in colorectal carcinoma

Somatic CNVs related to cancer

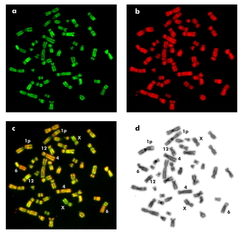






- Somatic CNVs in cancers:
  - HER2 Amplification in Breast Cancer
  - EGFR Amplification in Glioblastoma
  - Chromosome 3p Deletion in Lung Cancer

Wang H, Zhang X, Liu J, et al. Clinical roles of EGFR amplification in diffuse gliomas: a real-world study using the 2021 WHO classification of CNS tumors[J]. Frontiers in Neuroscience, 2024, 18: 1308627.

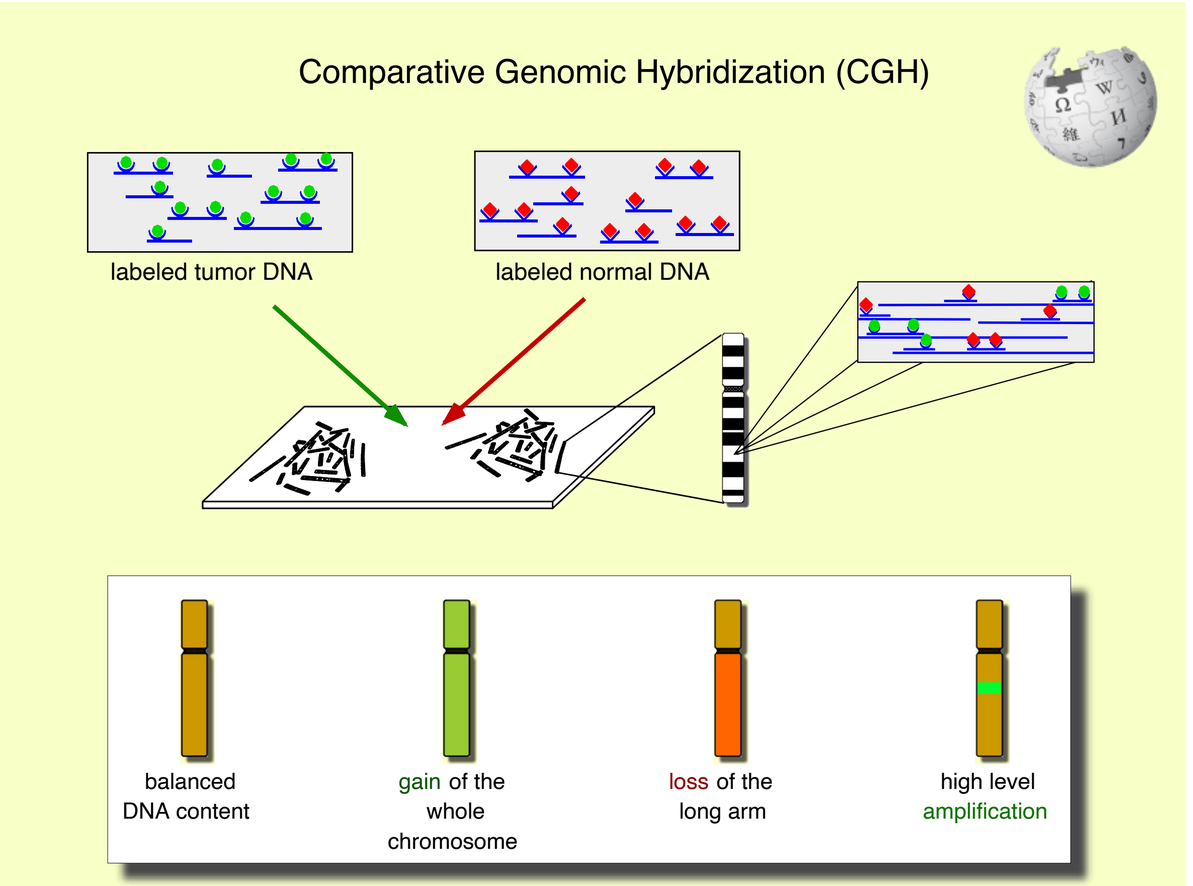


Comparative Genomic Hybridization

Molecular-Cytogenetic Technology for Genomic Imbalance Screening

- Molecular-cytogenetic technique to identify regional genomic copy number variations (CNV/CNA)
- based on ***in situ*** suppression **hybridization** of labeled **genomic** tumor and reference DNA against a karyotypically normal metaphase chromosomes
- analysis of relative fluorescence ratio allows **semi-quantitative copy number** read-out
- **indirect** attribution of involved target genes through cytogenetic bands (megabase resolution)

- Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992;5083:818-821.
- Joos S, Scherthan H, Speicher MR, Schlegel J, Cremer T, Lichter P. Detection of amplified DNA sequences by reverse chromosome painting using genomic tumor DNA as probe. Hum Genet. 1993;90:584-589.



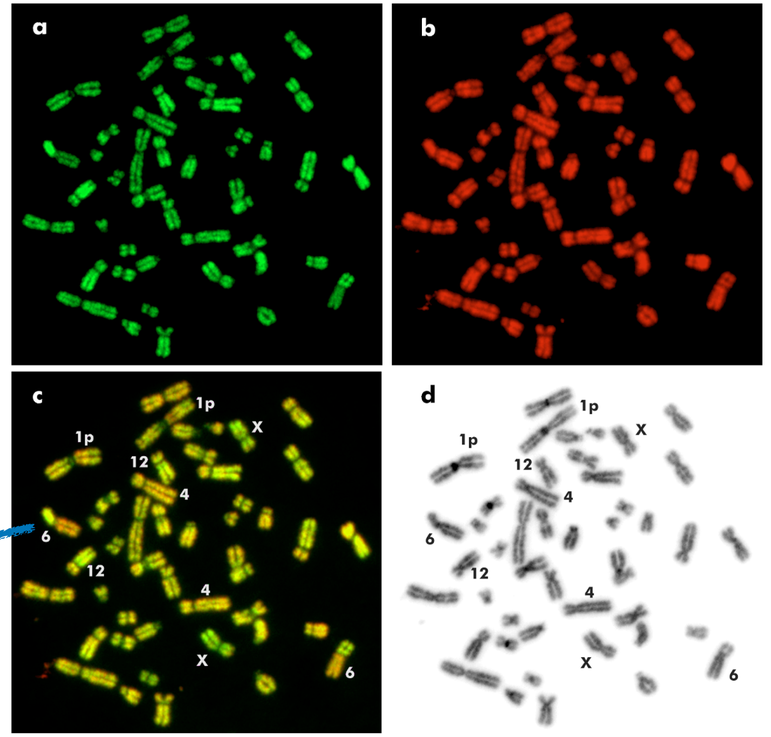
Chromosomal CGH: Normal metaphase spreads (cultured lymphocytes from healthy donors) on microscopy slides serve as the hybridization matrix for whole-genome DNA from tumor and reference tissue, labeled with different fluorophores. The regional ratio between the two colors points to (relative) changes in the copy number in the tumor DNA. Michael Baudis, 1998

Comparative Genomic Hybridization

Molecular-Cytogenetic Technology for Genomic Imbalance Screening

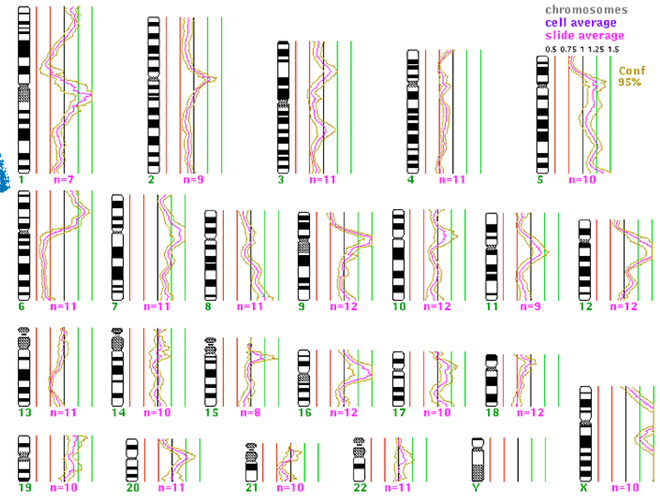
- Molecular-cytogenetic technique to identify regional genomic copy number variations (CNV/CNA)
- based on ***in situ*** suppression **hybridization** of labeled **genomic** tumor and reference DNA against a karyotypically normal metaphase chromosomes
- analysis of relative fluorescence ratio allows **semi-quantitative copy number** read-out
- **indirect** attribution of involved target genes through cytogenetic bands (megabase resolution)

- Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992;5083:818-821.
- Joos S, Scherthan H, Speicher MR, Schlegel J, Cremer T, Lichter P. Detection of amplified DNA sequences by reverse chromosome painting using genomic tumor DNA as probe. Hum Genet. 1993;90:584-589.



CGH-Experiment: **a** Hybridisierung mit Tumor-DNA; **b** Hybridisierung mit normaler menschlicher DNA als Kontrolle; **c** Überlagerung der Signale; **d** Bänderungsfärbung zur Identifizierung der Chromosomen

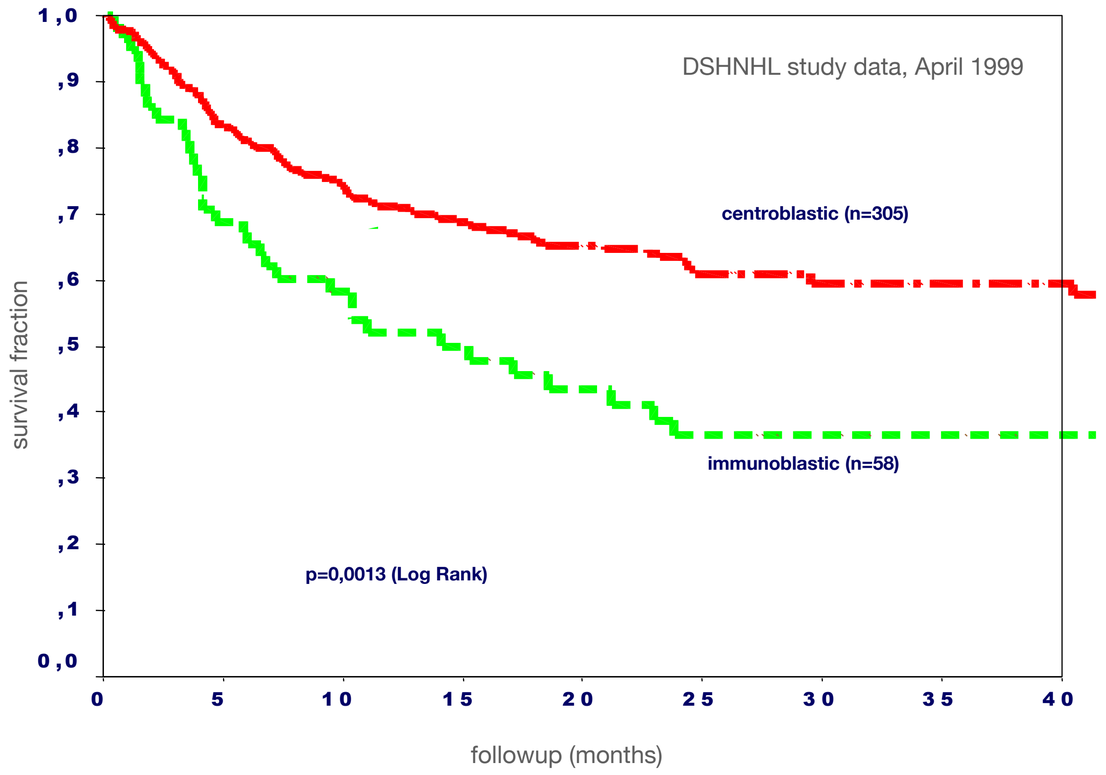
+6p, -6q...



Auswertung: Summationsprofil der computergestützten Analyse mehrerer Metaphasen des dargestellten Falles; die Profilausschläge stehen für Zugewinne bzw. Verluste von chromosomalen Anteilen im Tumorgenom

Cancer CNVs | Diagnostics | Prognosis

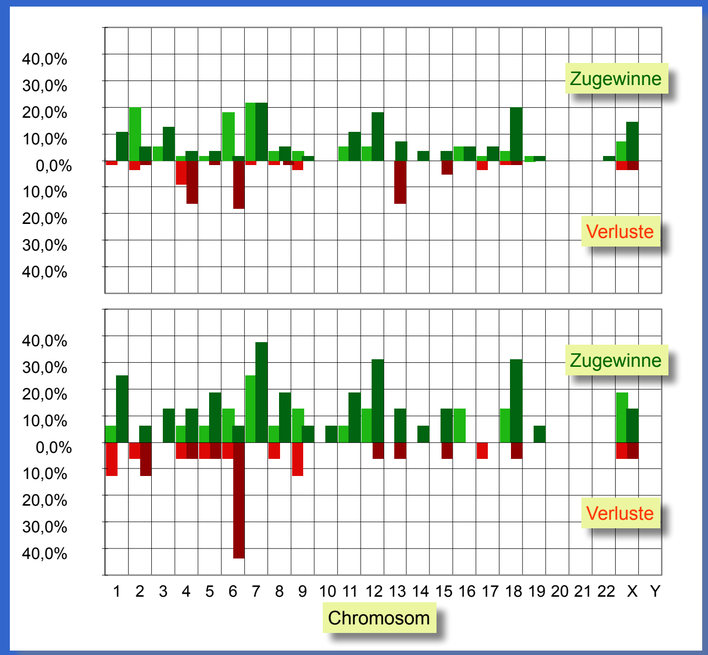
Single-study CNV frequencies correspond to diagnostic subsets



55 cb NHL

Chromosomale Imbalancen

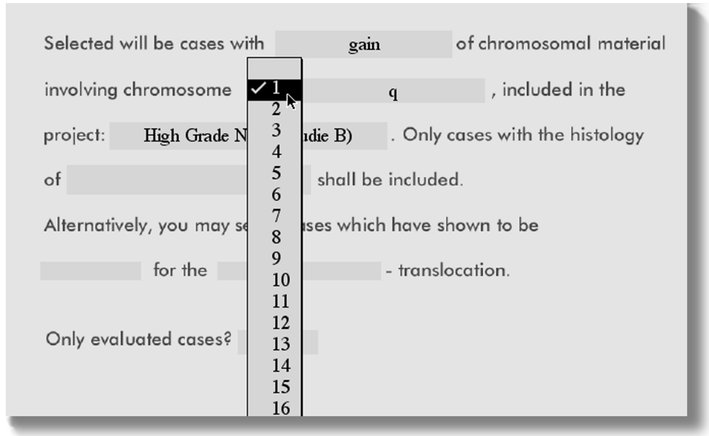
16 ib NHL

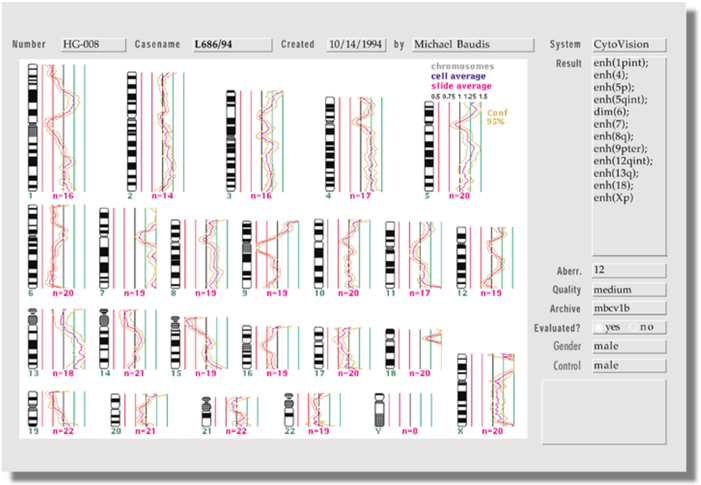


Let's build a database!

Progenetix CGH Database and Website

- originally an internal FileMaker Pro database, to store CGH profiles and annotations for the "Organization of Complex Genomes" group (head: Peter Lichter) at the German Cancer Research Center (DKFZ), starting in 1998
- expansion to include literature derived data, with a focus on malignant non-Hodgkin's lymphomas
- in 2000 online version






```
Domain Name: PROGENETIX.NET
Registry Domain ID: 45628826_DOMAIN_NET-VRSN
Registrar WHOIS Server: whois.enterprice.net
Registrar URL: http://www.epag.de
Updated Date: 2019-06-01T04:20:49Z
Creation Date: 2000-11-29T18:17:38Z
```

- Dec 6, 2000
  - first time online
- Nov 30, 2000
  - addition of graphical representation and gene table
- Nov 17, 2000
  - generation of website layout and database automatisation

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.

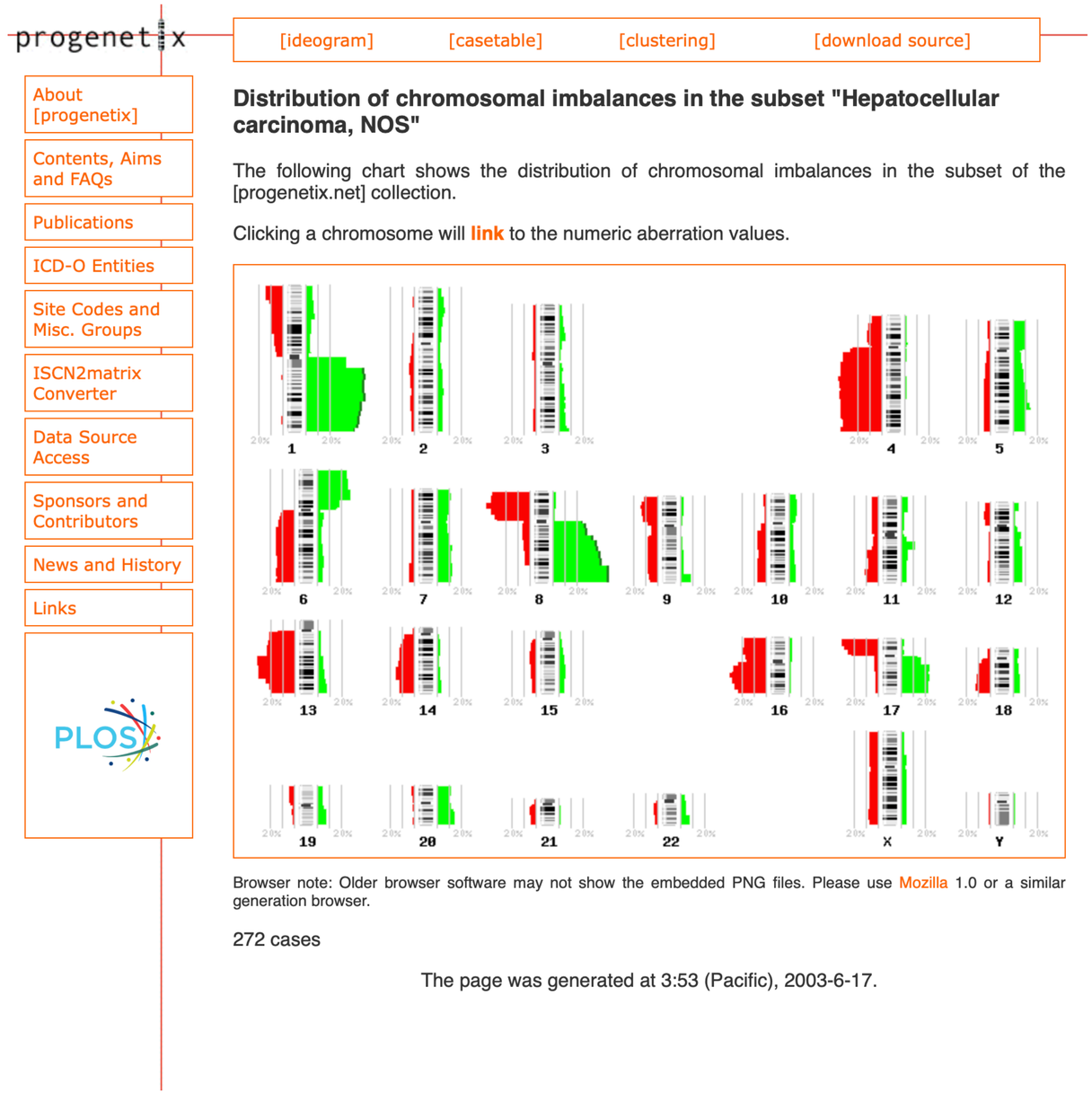
GENES, CHROMOSOMES & CANCER 25:82–90 (1999)

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

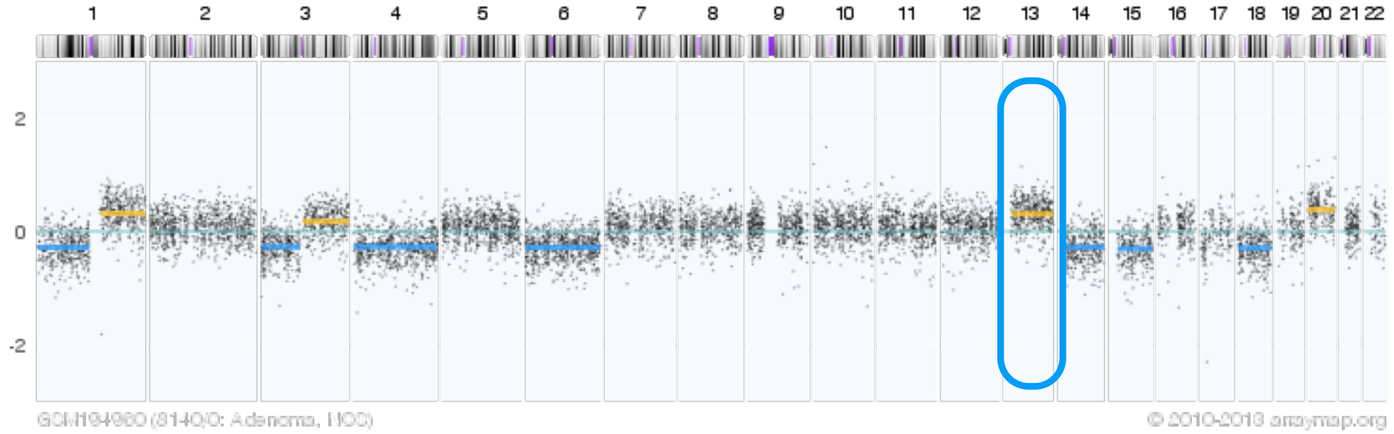
W. Michael Korn,[1]* Toru Yasutake,[2] Wen-Lin Kuo,[1] Robert S. Warren,[3] Colin Collins,[1] Masao Tomita,[2] Joe Gray,[1] and Frederic M. Waldman[1]

Progenetix Database in 2003
Text conversion for CNVs

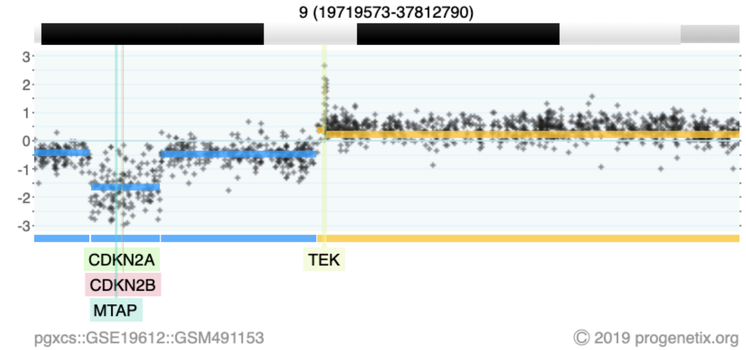
- based on listed CGH results from publications
  - literature detection using optimized PubMed queries
  - extraction (copy/paste, typing) of rev ish ISCN karyotypes from articles and supplementary material
  - annotation cleanup using scripting with regular expressions (Perl)
  - custom script to convert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



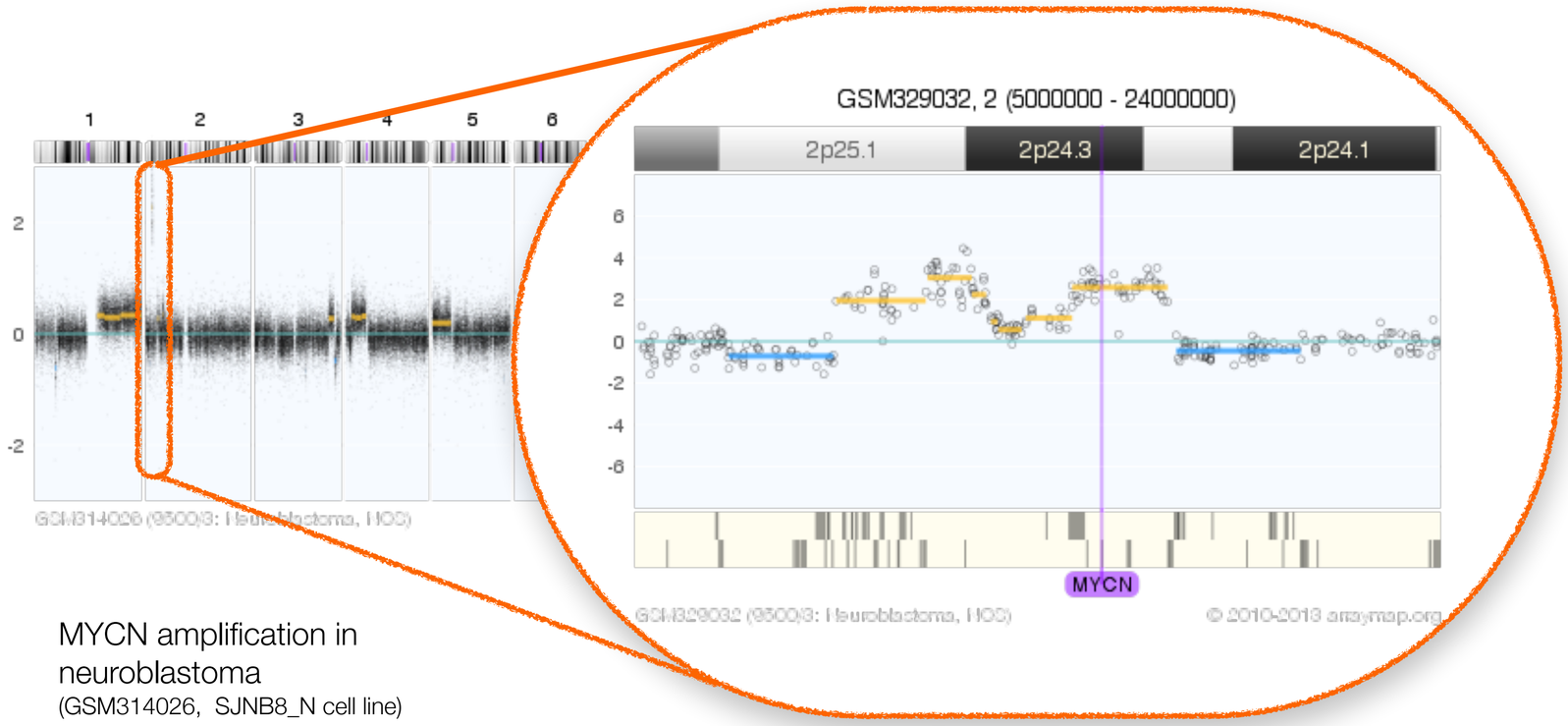
Array-based Detection of Copy Number Variations



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

arrayMap (2012 - 2020)

Probe-Level Genomic Array Data in Cancer

arrayMap

- Search Samples
- Search Publications
- Progenetix

University of Zurich[UZH]

- Citation & Licensing
- User Guide
- People
- Beacon+

CC BY-SA

Tweet

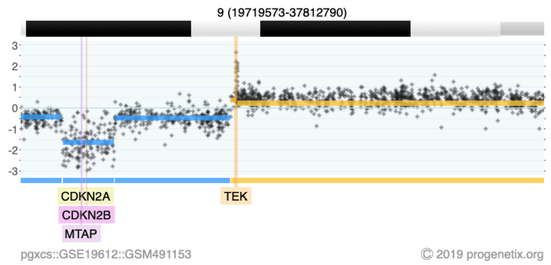
162.158.150.56

visualizing cancer genome array data @ arraymap.org

arrayMap is a curated reference database and bioinformatics resource targeting copy number profiling data in human cancer. The arrayMap database provides an entry point for meta-analysis and systems level data integration of high-resolution oncogenomic CNA data.

The current data reflects:

- 72724 genomic array profiles
- 898 experimental series
- 257 array platforms
- ICD-O 341 ICD-O cancer entities
- 795 publications (Pubmed entries)



Genomic copy number imbalances on chromosome 9 in a case of Glioblastoma (**GSM491153**), indcating, among others, a homozygous deletion involving CDKN2A/B.

For the majority of the samples, probe level visualization as well as customized data representation facilitate gene level and genome wide data review. Results from multi-case selections can be connected to downstream data analysis and visualization tools, as we provide through our Progenetix project.

arrayMap is developed by the group "Theoretical Cytogenetics and Oncogenomics" at the Institute of Molecular Life Sciences of the University of Zurich.

RELATED PUBLICATIONS

Cai H, Gupta S, Rath P, Ai N, Baudis M. arrayMap 2014: an updated cancer genome resource. Nucleic Acids Res. 2015 Jan;43(Database issue). Epub 2014 Nov 26.

Cai, H., Kumar, N., & Baudis, M. 2012. arrayMap: A Reference Resource for Genomic Copy Number Imbalances in Human Malignancies. *PLoS One* 7(5), e36944.

Baudis, M. 2007. Genomic imbalances in 5918 malignant epithelial tumors: An explorative meta-analysis of chromosomal CGH data. *BMC Cancer* 7:226.

Baudis, M. 2006. Online database and bioinformatics toolbox to support data mining in cancer cytogenetics. *Biotechniques* 40, no. 3: 296-272.

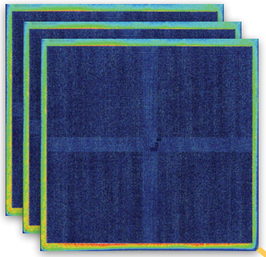
Baudis, M, and ML Cleary. 2001. Progenetix.net: an online repository for molecular cytogenetic aberration data. *Bioinformatics* 12, no. 17: 1228-1229.

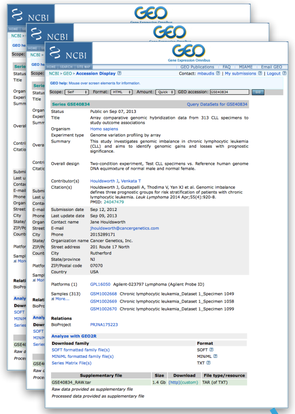
Feel free to use the data and tools for academic research projects and other applications. If more support and/or custom analysis is needed, please contact Michael Baudis regarding a collaborative project.

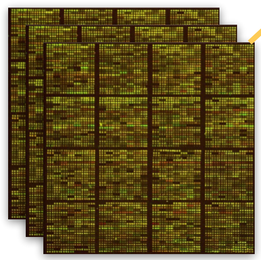
UID	SERIESID	PMID	ICDMORPHOLOGYCODE	ICDTOPOGRAPHYCODE
GSM1000061	GSE36942	23457519	8070/3	C10
GSM1000062	GSE36942	23457519	8070/3	C10
GSM1001316	GSE40777	23571474	8070/3	C53
GSM1001317	GSE40777	23571474	8010/3	C34
GSM1001318	GSE40777	23571474	8070/3	C09
GSM1001319	GSE40777	23571474	8010/3	C34
GSM1002668	GSE40834	24047479	9823/3	C42
GSM1002669	GSE40834	24047479	9823/3	C42
GSM1002670	GSE40834	24047479	9823/3	C42
GSM1002671	GSE40834	24047479	9823/3	C42
GSM1002672	GSE40834	24047479	9823/3	C42
GSM1002673	GSE40834	24047479	9823/3	C42
GSM1002674	GSE40834	24047479	9823/3	C42
GSM1002675	GSE40834	24047479	9823/3	C42
GSM1002676	GSE40834	24047479	9823/3	C42
GSM1002677	GSE40834	24047479	9823/3	C42
GSM1002678	GSE40834	24047479	9823/3	C42
GSM1002679	GSE40834	24047479	9823/3	C42
GSM1002680	GSE40834	24047479	9823/3	C42

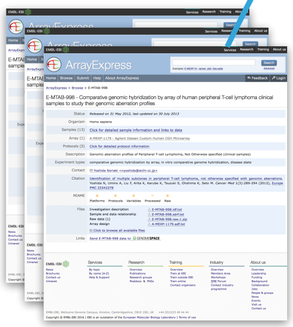
arrayMap

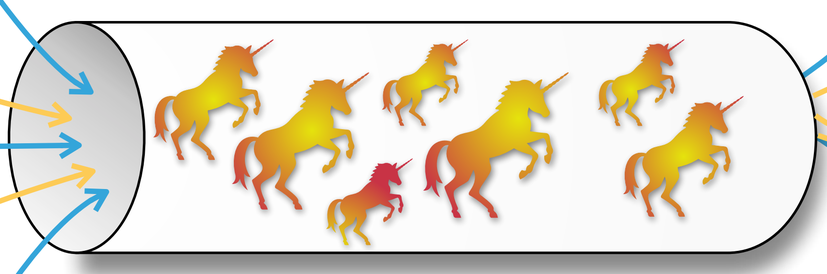
DATA PIPELINE

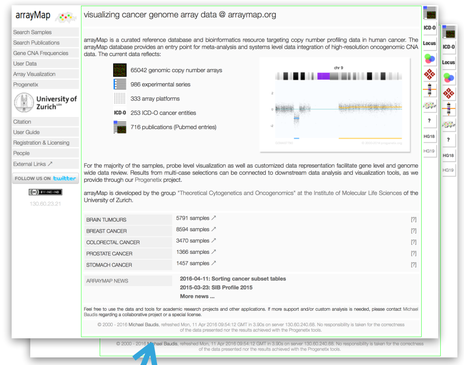


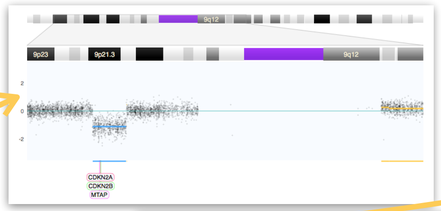


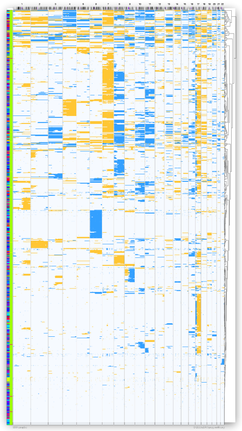


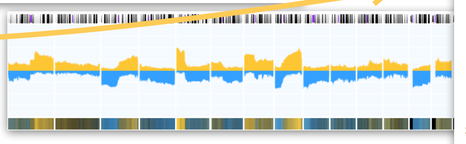


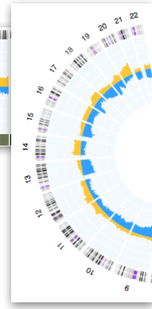


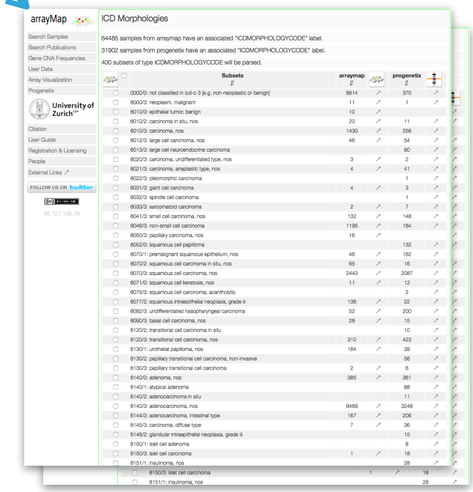












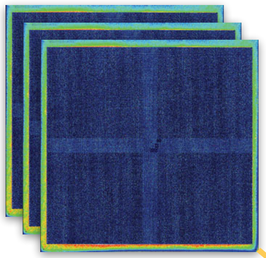
arrayMap

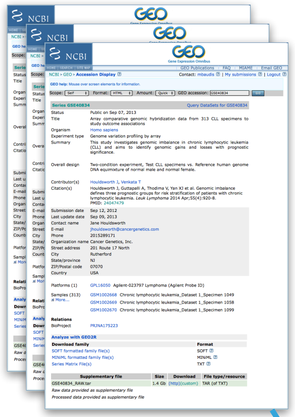
progenetix

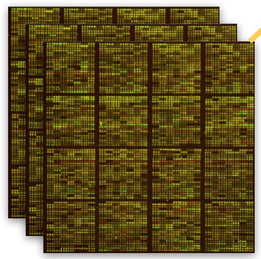
DATA PIPELINE

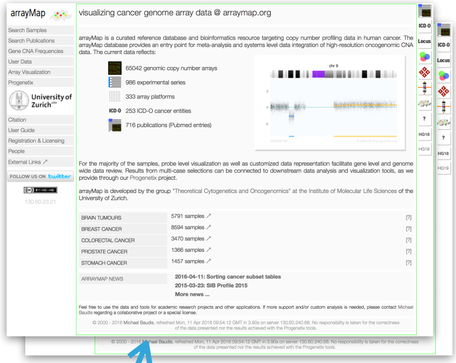
BIOCURATION

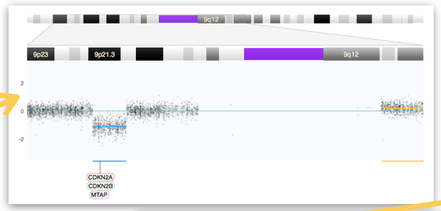
BIOINFORMATICS

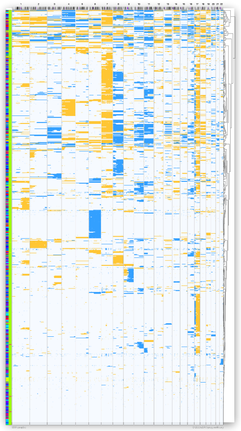


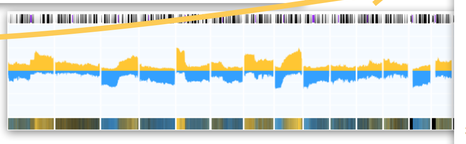


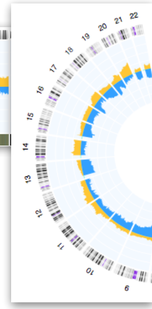


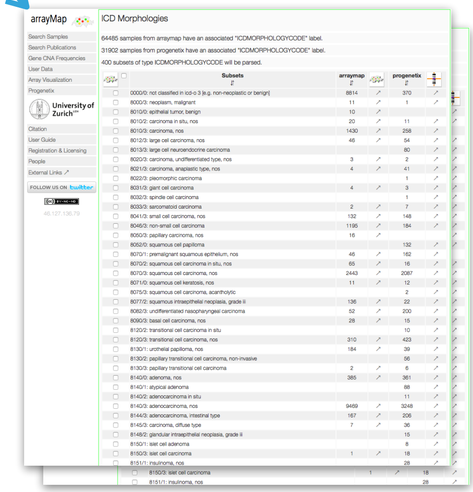








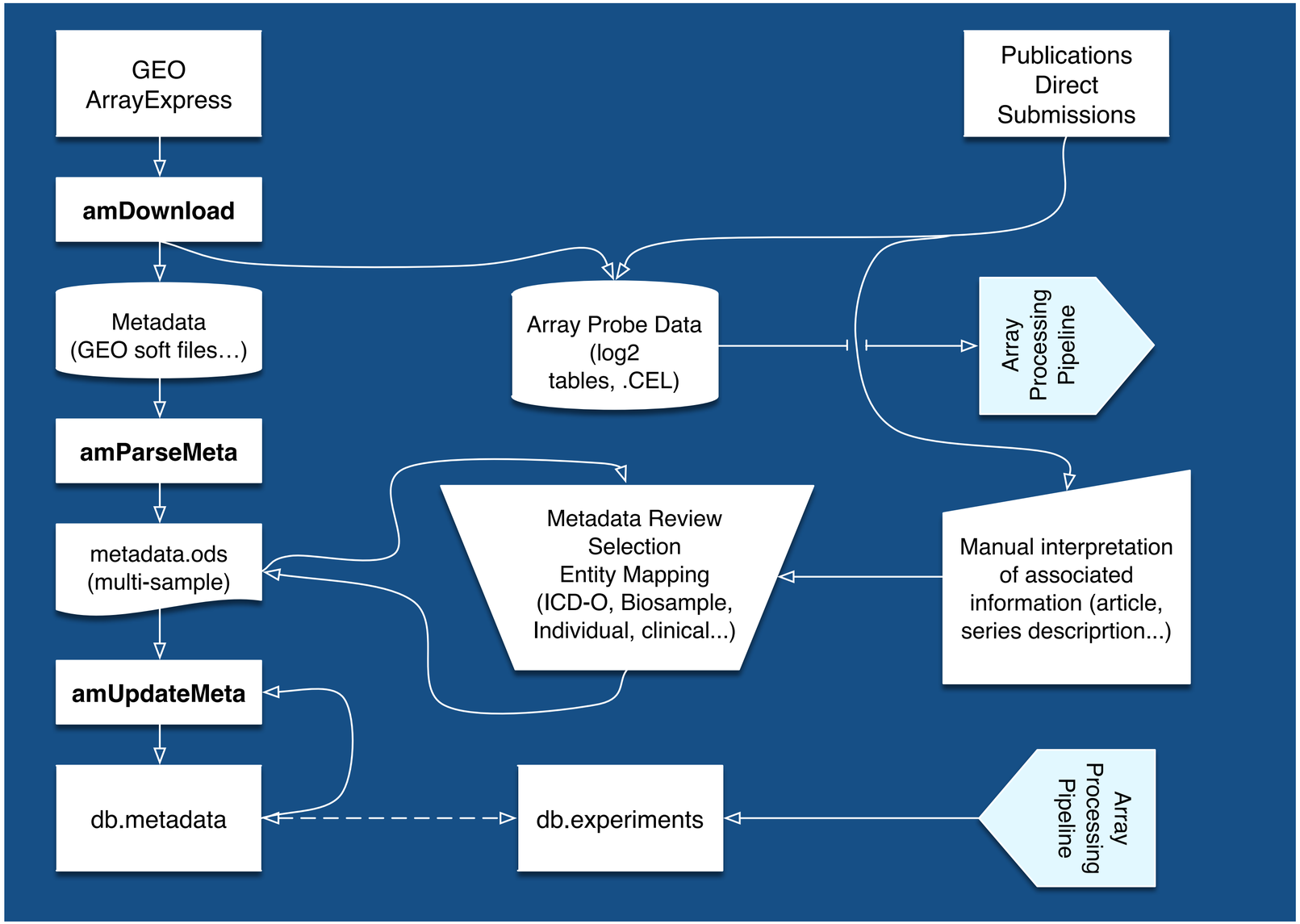




arrayMap

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 sets in tutorials

Data sets in the wild

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 standard Affymetrix protocol from 250 ng genomic DNA (Genechip Mapping 500k assay
manual 701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix).
!Sample_hyb_protocol = Hybridizations were performed according to the standard Affymetrix protocol from 250 ng genomic DNA (Genechip Mapping 500k assay manual
701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix) using an Affymetrix hybridisation oven 640 and an Affymetrix Fluidic station 450.
!Sample_scan_protocol = Scanning performed according to the standard Affymetrix protocol from 250 ng genomic DNA (Genechip Mapping 500k assay manual 701930 Rev.3
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.genome.umin.jp/)
!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
```

progenetix

Regular Expressions!

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

```
19  extraction_scopes:
20    description: >-
21      Detection and processing of clinical scopes goes through several stages:
22      1. line cleanup - so far run for the input before processing the individual
23      scopes
24      2. line match using sme general pattern expected in all lines containing
25      data for the current scope (`filter` pattern)
26      3. finding and extracting the relevant data by looping over a list of
27      specific patterns with memorized matches (`find`)
28      4. post-processing using empirical cleanp replacements (`cleanup`)
29      5. checking the correct structure (`final_check` - a global pattern can be
30      used if other post-processing is performed)
31
32
33    survival_status:
34      filter: '(?i).*?(?:(?
35      preclean:
36        - m: '(?i)days to death or last seen alive[^\w]+?\d+?(?:[^\w\.]|$)'
37          s: ''
38        - m: '[^\w]+?NA(?:[^\w\.]|$)'
39          s: ' '
40        - m: 'remission status past double induction .cr. comple emission. RD. refractory disease. ED. early death[^\w]+?ED'
41          s: 'survival: dead'
42        - m: 'remission status past double induction .cr. complete remission. RD. refractory disease. ED. early death[^\w]+?NA'
43          s: ''
44        - m: 'remission status past double induct efractory disease. ED. early death[^\w]+?CR'
45          s: 'survival: alive'
46        - m: 'remission status past double induction .cr. complete remission. RD. refractory disease. ED. early death[^\w]+?RD'
47          s: '' # alive but not responding to therapy so removed?
48        - m: 'Event Free Survival[^\w]+?no event'
49          s: 'recurrence: no'
50        - m: 'Event Free Survival.event'
51          s: 'recurrence: yes'
52        - m: 'Outcome[^\w]+?no event'
53          s: 'survival: alive'
54        - m: 'Outcome[^\w]+?event'
55          s: 'survival: dead'
56        - m: 'survival status[^\w]+?0'
57          s: 'survival: dead'
58        - m: 'survival status[^\w]+?1'
59          s: 'survival: alive'
60        - m: 'overall[^\w]+?survival[^\w]+?days[^\w]+?NA'
61          s: ''
62        - m: 'survival(?: time|from diagnosis)?[^\w]+?(days|months|years?)[^\w]+?(\d\d?\d?\d?\.?\d?\d?)'
63          s: 'survival: \2\1'
```

/outcome[^\w]+?no\s+?event/i

survival: alive

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 }) }
```

Disease annotations in Progenetix

From some text, somewhere, to ontology classes

International Agency for Research on Cancer

World Health Organization

NCI thesaurus

- **diagnostic categories** are the **most important** labels to associate with genomic data
- 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

- Malignant Kidney Neoplasm: NCIT:C7548 (5134 samples, 5204 CNV profiles)
  - Childhood Malignant Kidney Neoplasm: NCIT:C123907 (50 samples, 50 CNV profiles)
    - Clear Cell Sarcoma of the Kidney: NCIT:C4264 (37 samples, 37 CNV profiles)
    - Rhabdoid Tumor of the Kidney: NCIT:C8715 (13 samples, 13 CNV profiles)
  - Kidney Wilms Tumor: NCIT:C40407 (446 samples, 446 CNV profiles)
  - Kidney Sarcoma: NCIT:C4525 (37 samples, 37 CNV profiles)
    - Clear Cell Sarcoma of the Kidney: NCIT:C4264 (37 samples, 37 CNV profiles)
  - Malignant Renal Pelvis Neoplasm: NCIT:C7525 (7 samples, 7 CNV profiles)
    - Renal Pelvis Carcinoma: NCIT:C6142 (7 samples, 7 CNV profiles)
      - Renal Pelvis Urothelial Carcinoma: NCIT:C7355 (7 samples, 7 CNV profiles)
  - Kidney Carcinoma: NCIT:C9384 (4638 samples, 4708 CNV profiles)
    - Kidney Carcinoma Molecular Subtypes: NCIT:C189241 (47 samples, 47 CNV profiles)
      - Renal Cell Carcinoma with MiT Translocations: NCIT:C154494 (28 samples, 28 CNV profiles)
        - TFE3-Rearranged Renal Cell Carcinoma: NCIT:C27891 (28 samples, 28 CNV profiles)
      - Fumarate Hydratase-Deficient Renal Cell Carcinoma: NCIT:C164156 (19 samples, 19 CNV profiles)
      - Mucinous Tubular and Spindle Cell Carcinoma of the Kidney: NCIT:C39807 (8 samples, 8 CNV profiles)
    - Renal Pelvis Carcinoma: NCIT:C6142 (7 samples, 7 CNV profiles)
      - Renal Pelvis Urothelial Carcinoma: NCIT:C7355 (7 samples, 7 CNV profiles)
    - Collecting Duct Carcinoma: NCIT:C6194 (14 samples, 14 CNV profiles)
    - Kidney Medullary Carcinoma: NCIT:C7572 (16 samples, 16 CNV profiles)
    - Renal Cell Carcinoma: NCIT:C9385 (4592 samples, 4662 CNV profiles)
      - Renal Cell Carcinoma with MiT Translocations: NCIT:C154494 (28 samples, 28 CNV profiles)
        - TFE3-Rearranged Renal Cell Carcinoma: NCIT:C27891 (28 samples, 28 CNV profiles)
      - Fumarate Hydratase-Deficient Renal Cell Carcinoma: NCIT:C164156 (19 samples, 19 CNV profiles)
      - Renal Cell Carcinoma, Not Otherwise Specified: NCIT:C191370 (106 samples, 106 CNV profiles)
      - Non-Clear Cell Renal Cell Carcinoma: NCIT:C202497 (781 samples, 797 CNV profiles)
        - Sarcomatoid Renal Cell Carcinoma: NCIT:C27893 (39 samples, 40 CNV profiles)
        - Chromophobe Renal Cell Carcinoma: NCIT:C4146 (234 samples, 234 CNV profiles)

Mapping 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

Standardized Data

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

- International initiatives (ELIXIR, GA4GH, MONARCH...) and resource providers (EBI, NCBI ...) work on the generation and implementation of **data annotation standards**
- established principles are the use of **hierarchical coding systems** where individual codes are represented as **CURIE**s
- 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 ...)

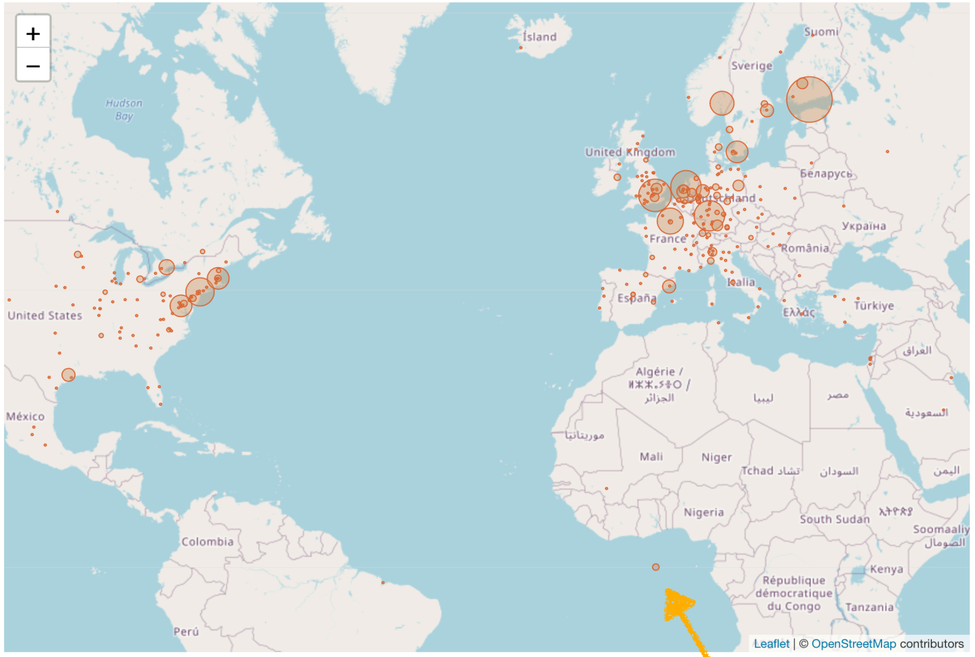
```
"data_use_conditions" : {
  "label" : "no restriction",
  "id" : "DUO:0000004"
},

"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"
}
```

Data Curation

Provide "clean and correct data" - but final verification of data from external resources lies with the user ...

- correct data is important for any type of scientific analysis
- errors in formats and values can occur during all steps between data acquisition and analysis (numerous "Excelgates"!)
- "meta"-resources and analyses are prone to erroneous data due to varying input formats and lack of source control

➡ always look for batch effects and outliers!



Geographic distribution (by corresponding author) of the **118554** genomic array, **36766** chromosomal CGH and **42105** whole genome/exome based cancer genome datasets from the **3306** listed publications. Area sizes correspond to the sample numbers reported from a given location.

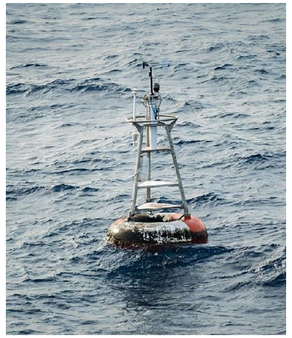
Progenetix publication collection
progenetix.org/publications/list
2020-11-28

25 / 3306 publications

Data Curation - Geolocations

Provide "clean and correct data" - but final verification of data from external resources lies with the user ...

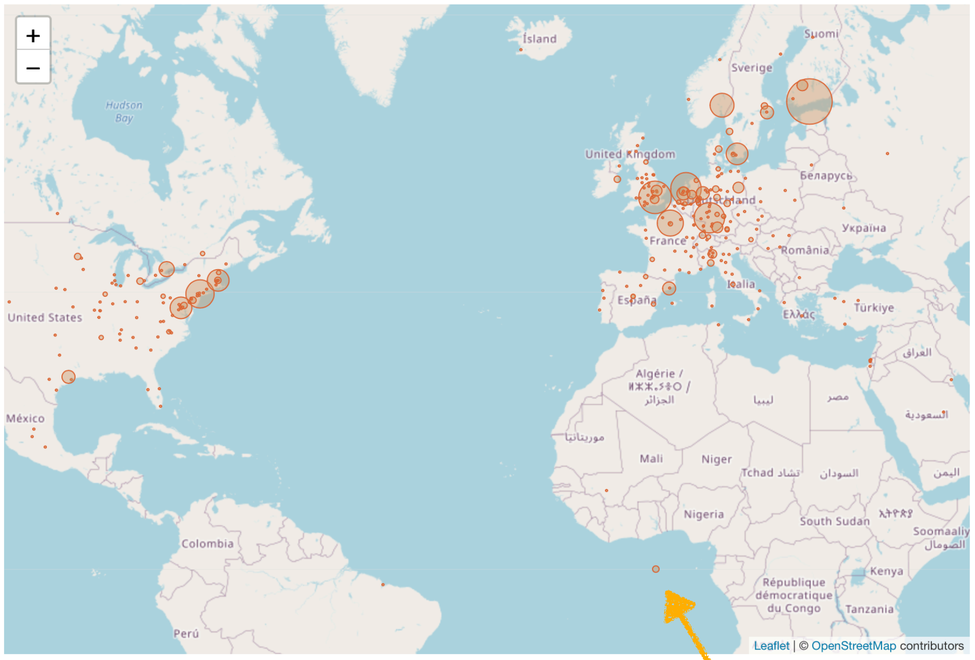
The most geo-tagged place on earth is Null Island



*A troubleshooting country has been added with an Indeterminate sovereignty class called **Null Island (1, 2)**. It is a fictional, 1 meter square island located off Africa where the equator and prime meridian cross. Being centered at 0,0 (zero latitude, zero longitude) it is useful for flagging geocode failures which are routed to 0,0 by most mapping services. Aside: "Null Islands" exist for all local coordinate reference systems besides WGS84 like State Plane (and global if not using modern Greenwich prime meridian). Null Island in Natural Earth is scaleRank 100, indicating it should never be shown in mapping. Side note: Rank 30 (zoom 29 in Google speak)*

https://en.wikipedia.org/wiki/Null_Island

Michael Szell: The Data Science Process 2
http://michael.szell.net/downloads/lecture26_datasciprocess2.pdf
2020-11-25



Geographic distribution (by corresponding author) of the **118554** genomic array, **36766** chromosomal CGH and **42105** whole genome/exome based cancer genome datasets from the **3306** listed publications. Area sizes correspond to the sample numbers reported from a given location.

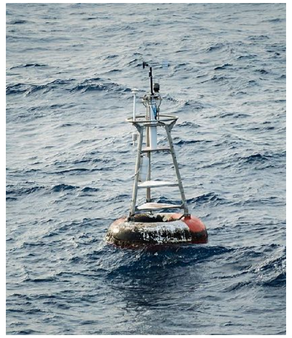
Progenetix publication collection
progenetix.org/publications/list
2020-11-28

25 / 3306 publications

Data Curation - Geolocations

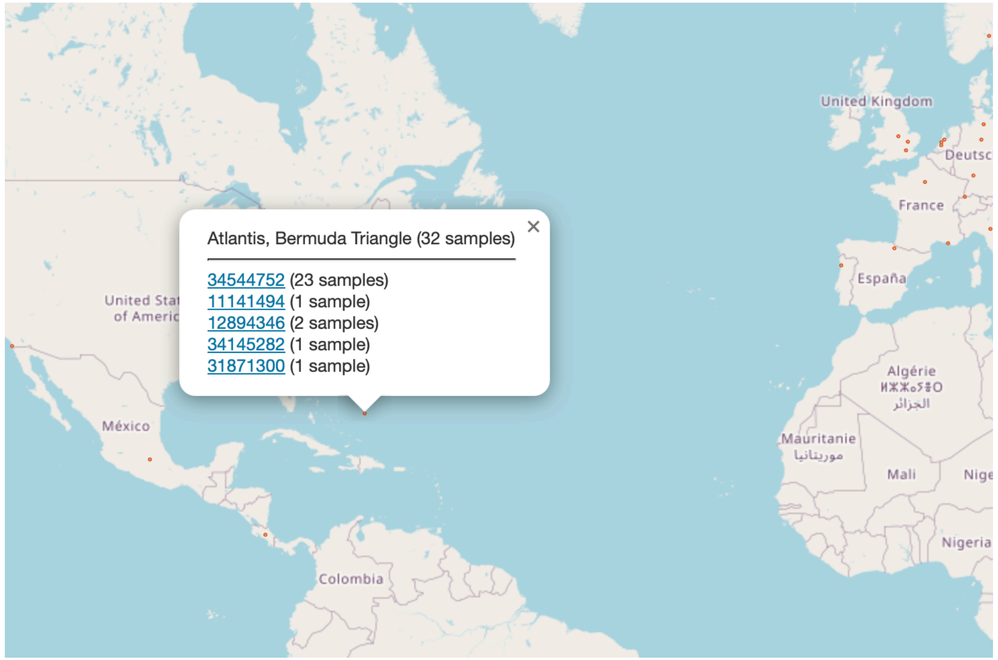
Provide "clean and correct data" - but final verification of data from external resources lies with the user ...

The most geo-tagged place on earth is Null Island



A troubleshooting country has been added with an Indeterminate sovereignty class called ***Null Island (1, 2)****. It is a fictional, 1 meter square island located off Africa where the equator and prime meridian cross. Being centered at 0,0 (zero latitude, zero longitude) it is useful for flagging geocode failures which are routed to 0,0 by most mapping services. Aside: "Null Islands" exist for all local coordinate reference systems besides WGS84 like State Plane (and global if not using modern Greenwich prime meridian). Null Island in Natural Earth is scaleRank 100, indicating it should never be shown in mapping. Side note: Rank 30 (zoom 29 in Google speak)*

https://en.wikipedia.org/wiki/Null_Island



Michael Szell: The Data Science Process 2
http://michael.szell.net/downloads/lecture26_datasciprocess2.pdf
2020-11-25

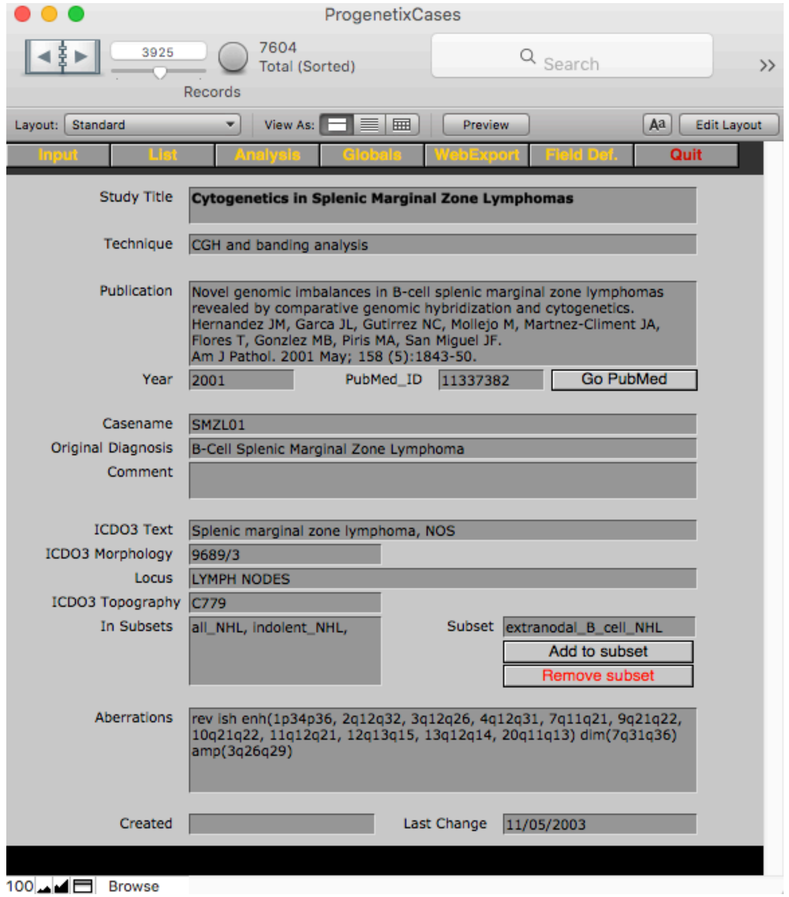
Progenetix in 2025

An oncogenomic reference resource

progenetix

Database Structure

From flat database to hierarchical object storage



Archived version of 2003 "ProgenetixCases" FMP solution

2003

- custom FileMaker database
- text-based annotations
- export & generation of static webpages and data files

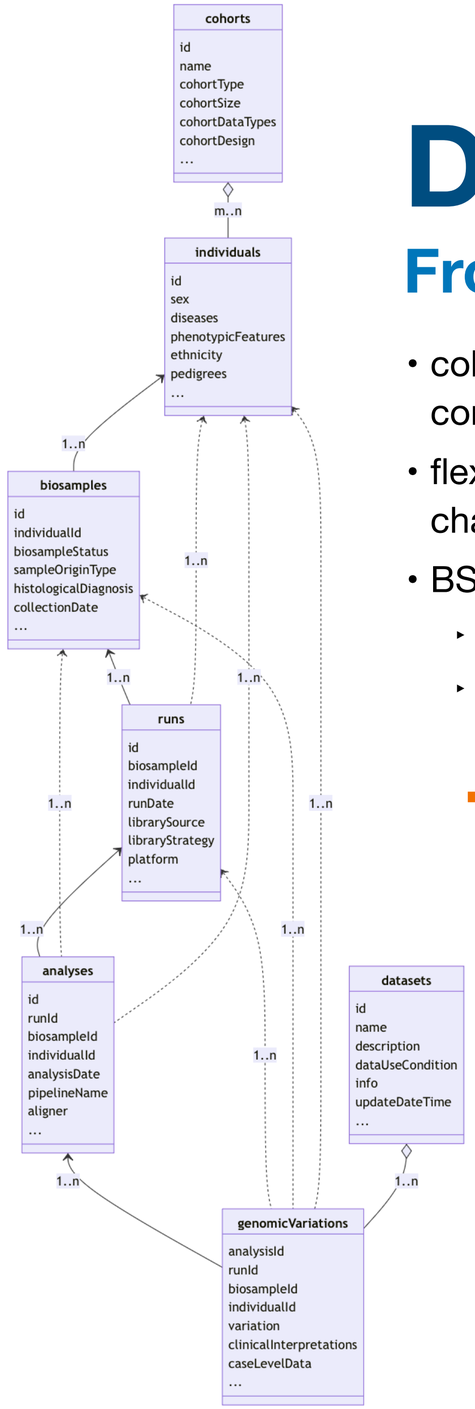
2025

- 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")
}
```

```
{
  "_id" : ObjectId("5bab583e727983b2e01255ae"),
  "callset_id" : "pgxcs-kftvv618",
  "biosample_id" : "pgxbs-kftvhcao",
  "assembly_id" : "GRCh38",
  "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"
}
```

```
{
  "_id" : ObjectId("5bab56cd727983b2e00b0bde"),
  "id" : "pgxbs-kftvhcao",
  "description" : "Splenic Marginal Zone Lymphoma",
  "biocharacteristics" : [
    {
      "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"
    }
  },
  "data_use_conditions" : {
    "label" : "no restriction",
    "id" : "DUO:0000004"
  }
}
```



Database Structure

From flat database to hierarchical object storage

- 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**

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

2025


```
{
  "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")
}
```

```
{
  "_id" : ObjectId("5bab583e727983b2e01255ae"),
  "callset_id" : "pgxcs-kftvv618",
  "biosample_id" : "pgxbs-kftvhcao",
  "assembly_id" : "GRCh38",
  "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"
}
```

```
{
  "_id" : ObjectId("5bab56cd727983b2e00b0bde"),
  "id" : "pgxbs-kftvhcao",
  "description" : "Splenic Marginal Zone Lymphoma",
  "biocharacteristics" : [
    {
      "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"
    }
  },
  "data_use_conditions" : {
    "label" : "no restriction",
    "id" : "DUO:0000004"
  }
}
```

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.

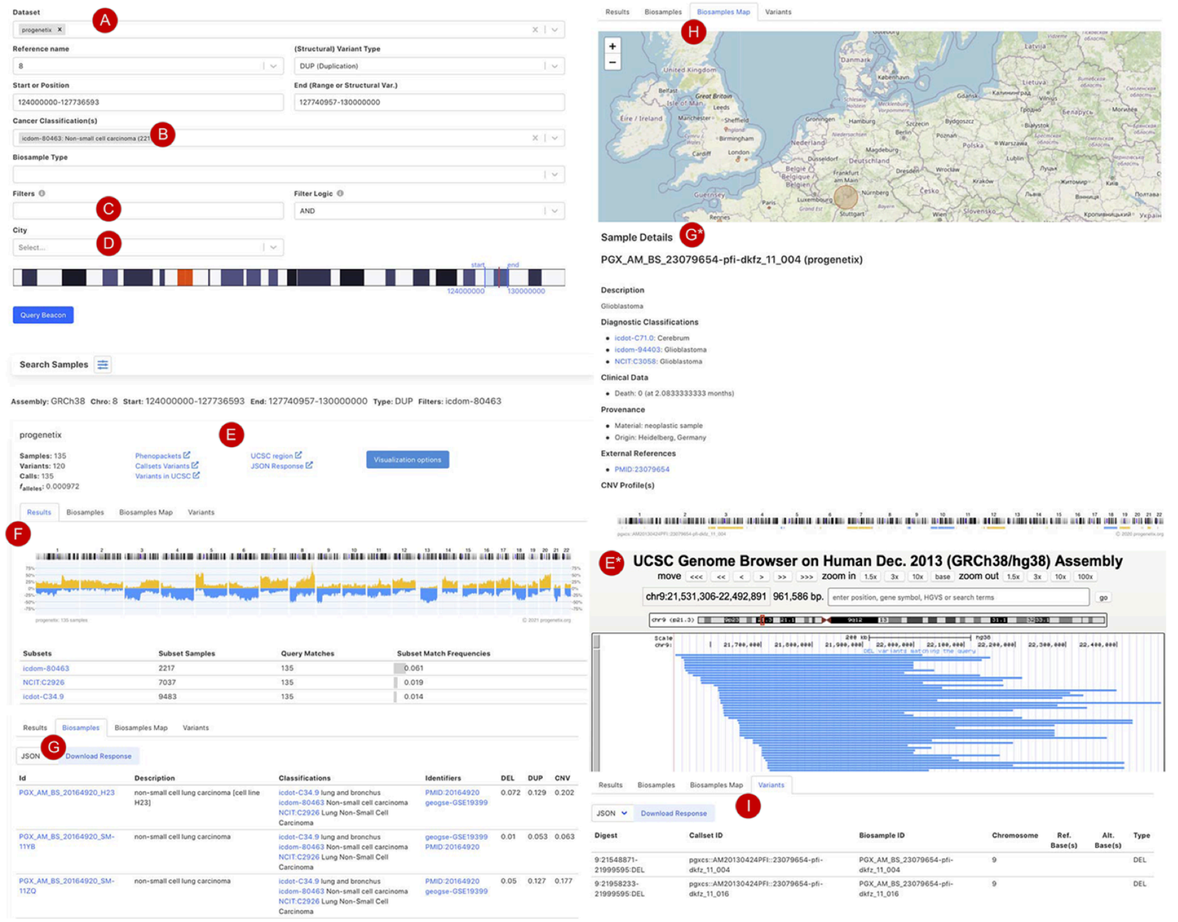
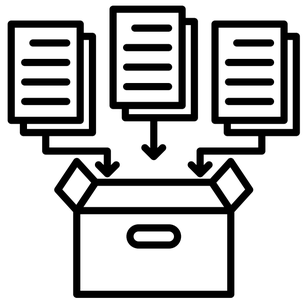


Figure 3. Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range This example queries for a continuous, focal duplication covering the complete MYC gene's coding region with < = 6 Mb in size. A: Filter for dataset; B: filter for cancer classification (NCIt and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

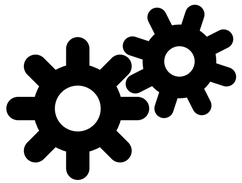
End-to-End Data Pipeline

Collect



GENIE, cBioPortal, GEO, TCGA, Other literatures…

Preprocess

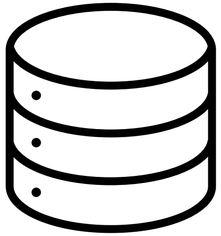


- Remove invalid samples
- Extract interesting information
- Unify data format

Analyze

- Metadata: ontology mapping
- Variant data: calling

Storage

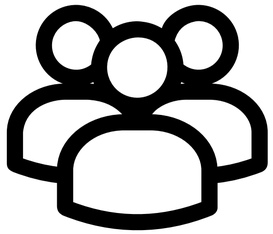


- Website
- **R client**

Collections

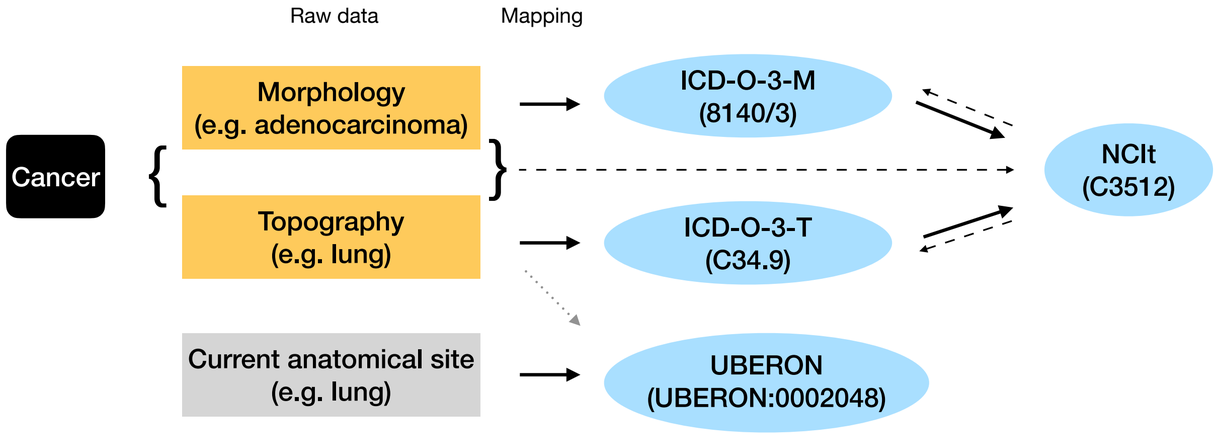
- individuals
- biosamples
- analyses
- variants

User

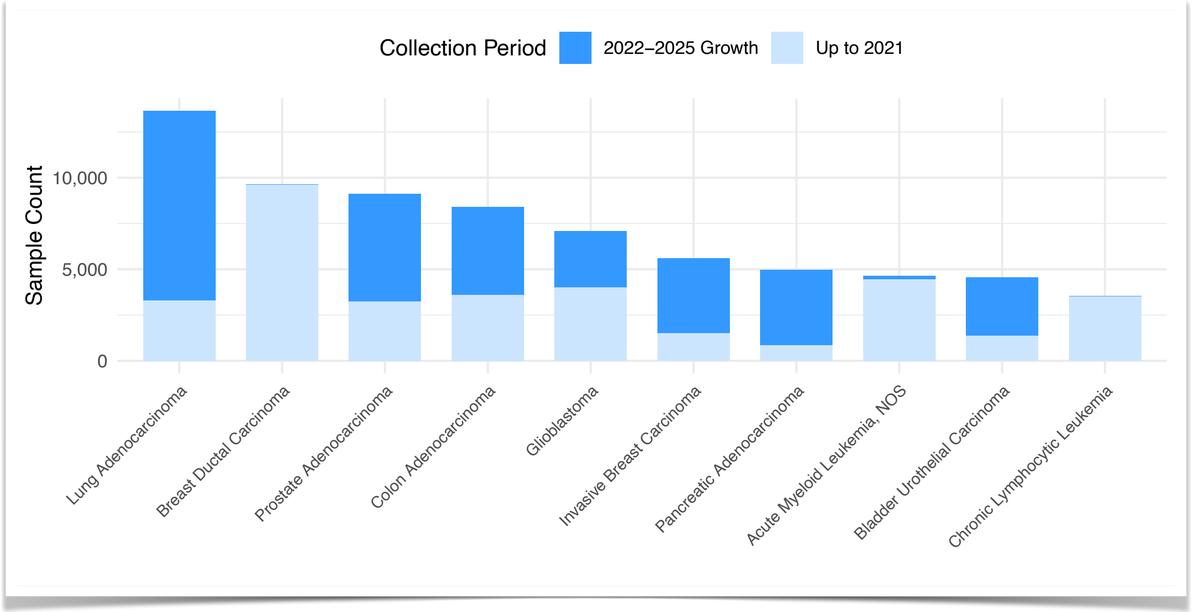


Enhance Progenetix

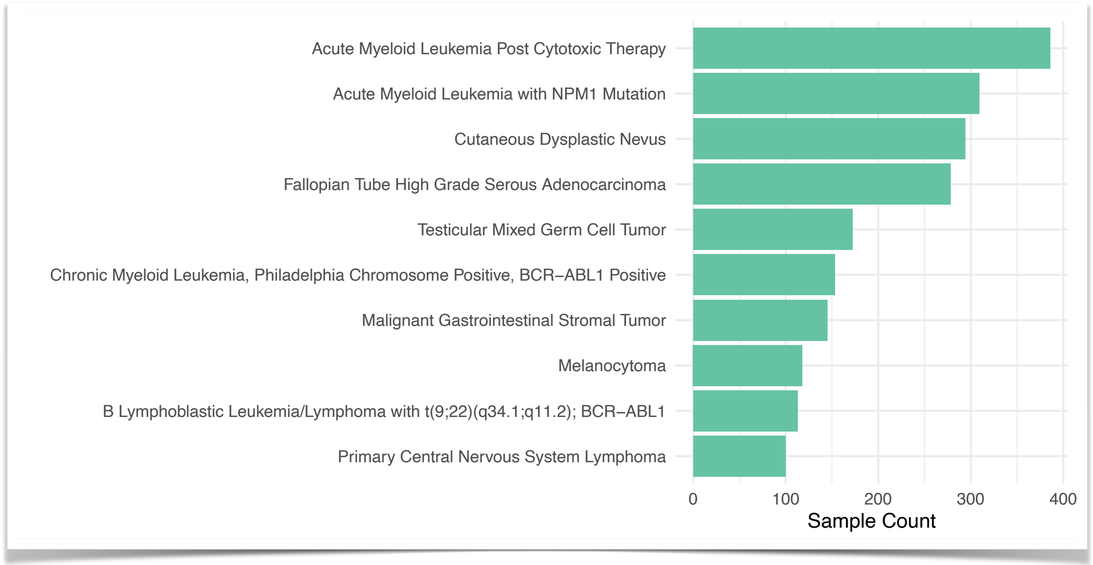
Optimization and expansion of cancer type representation



Top 10 most frequent cancer

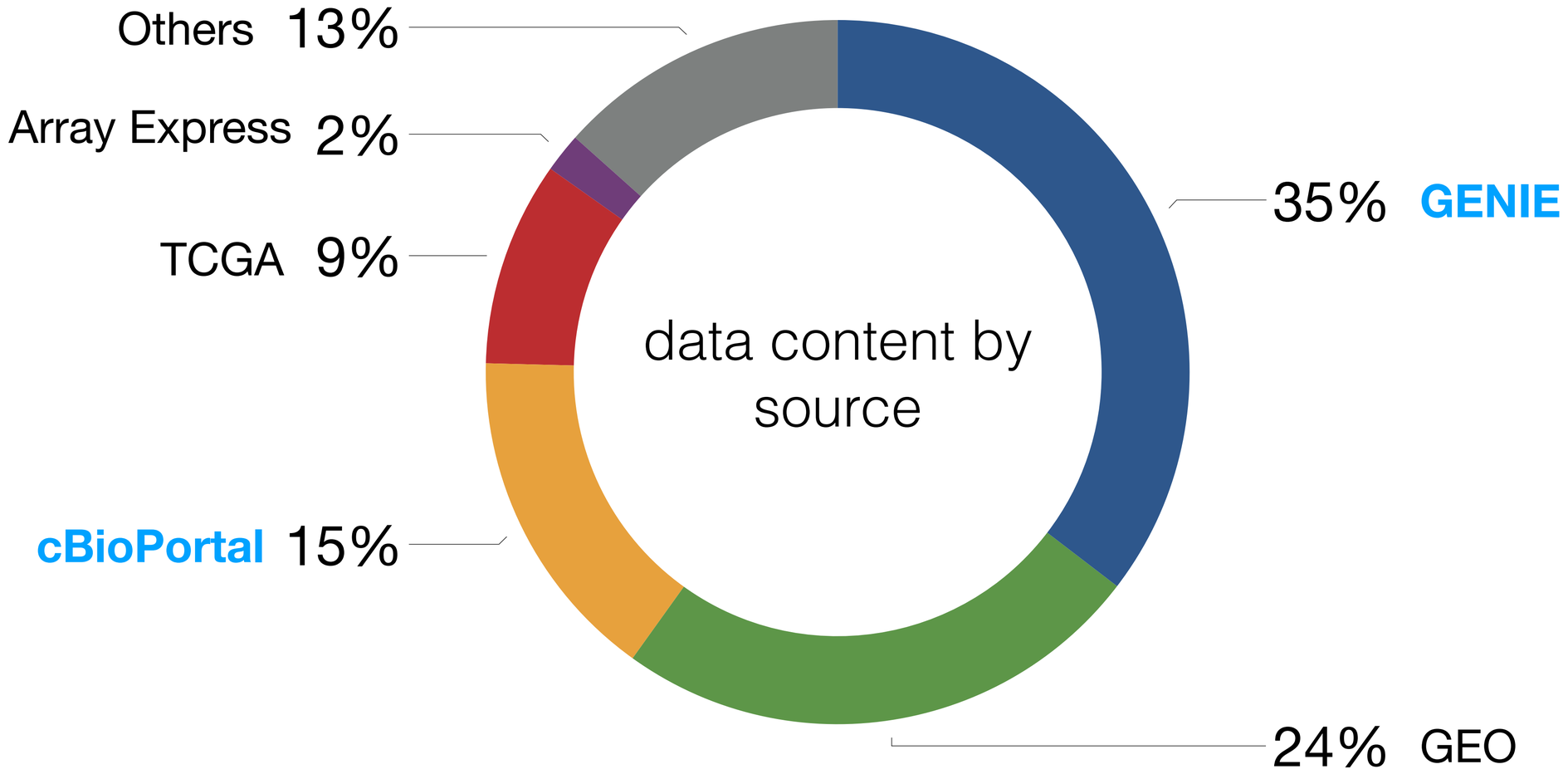


More granular terms in new NCIt caner types



Enhance Progenetix

Import ~100k new samples

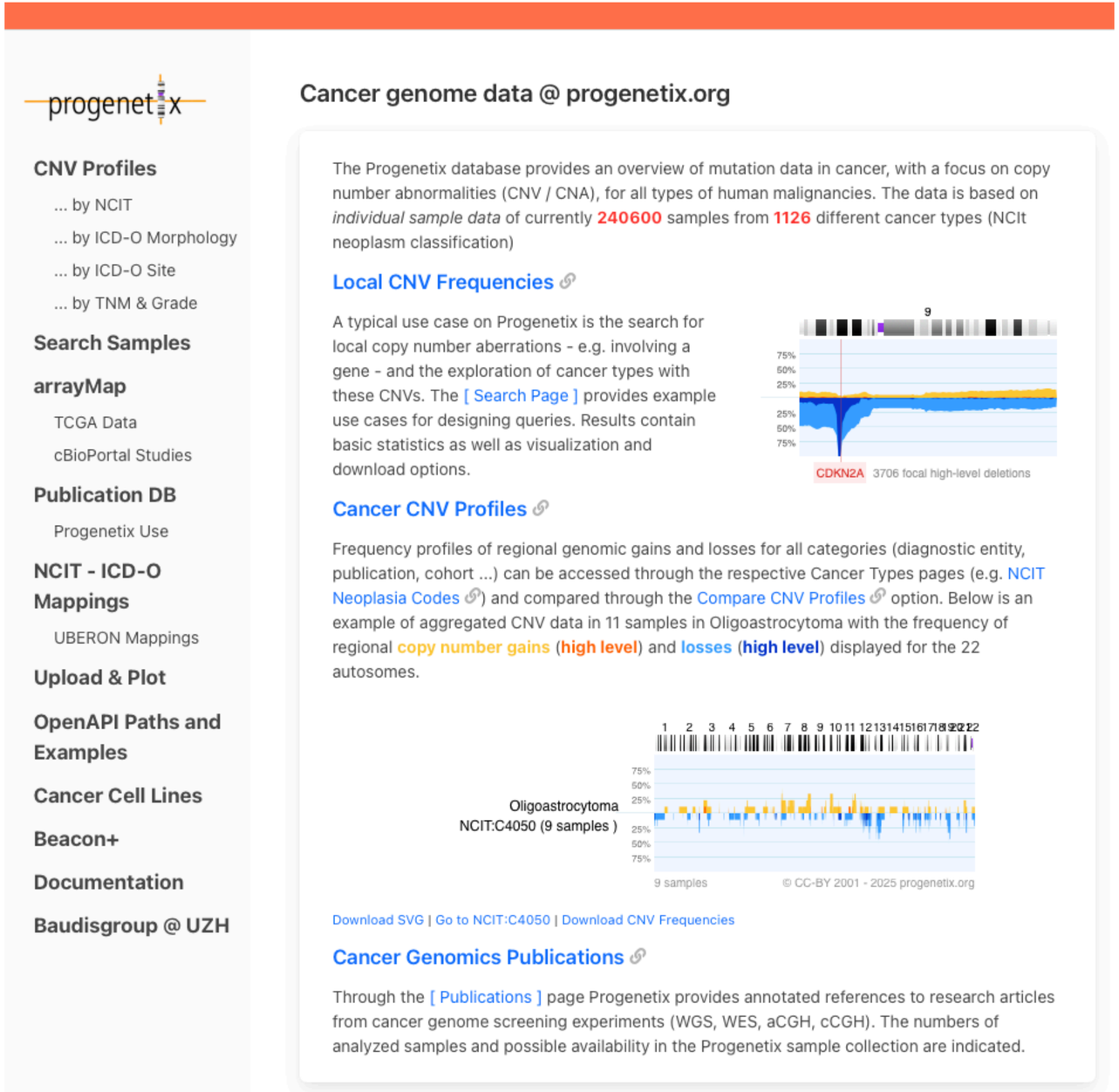


Total: 236,403 samples with 1,129 NCIt cancer types

progenetix.org

Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **240'000 cancer CNV profiles**
- more than 1'100 diagnostic types
- inclusion of reference sets (e.g. TCGA, GENIE...)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services
- recent addition of SNV data for some series



Universität Zürich[UZH]

progenetix

SIB Swiss Institute of Bioinformatics

progenetix.org

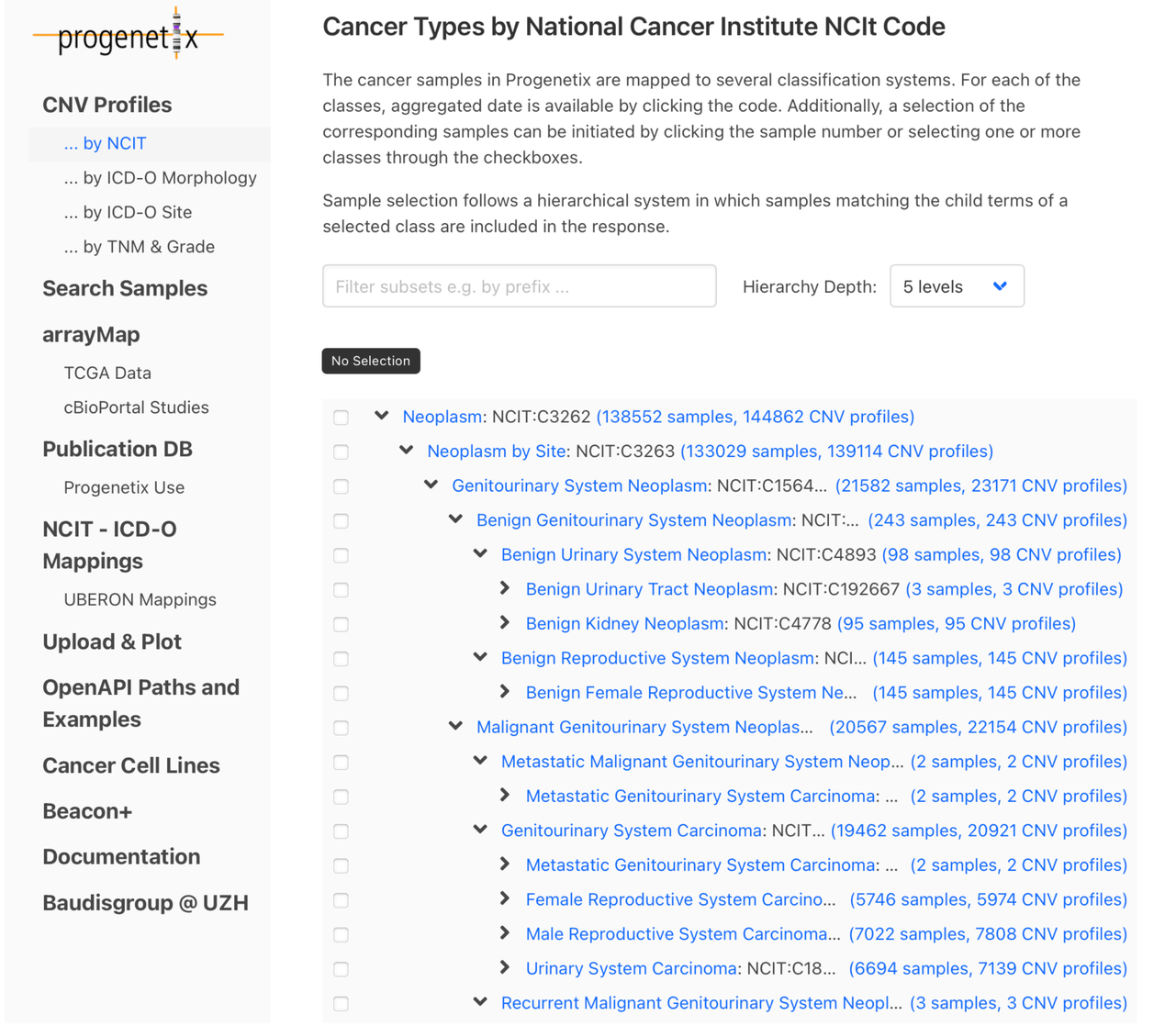
Cancer Genomics Reference Resource

- ***open*** resource for oncogenomic profiles
- over **240'000 cancer CNV profiles**
- SNV data for some series (e.g. TCGA)
- more than **1100 diagnostic types**
- inclusion of reference datasets (e.g. TCGA, GENIE, cBioPortal)
- standardized encodings (e.g. NCIt, ICD-O 3)
- identifier mapping for PMID, GEO, Cellosaurus, TCGA, cBioPortal where appropriate
- core clinical data (TNM, sex, survival ...)
- data mapping services

Universität Zürich[UZH]

progenetix

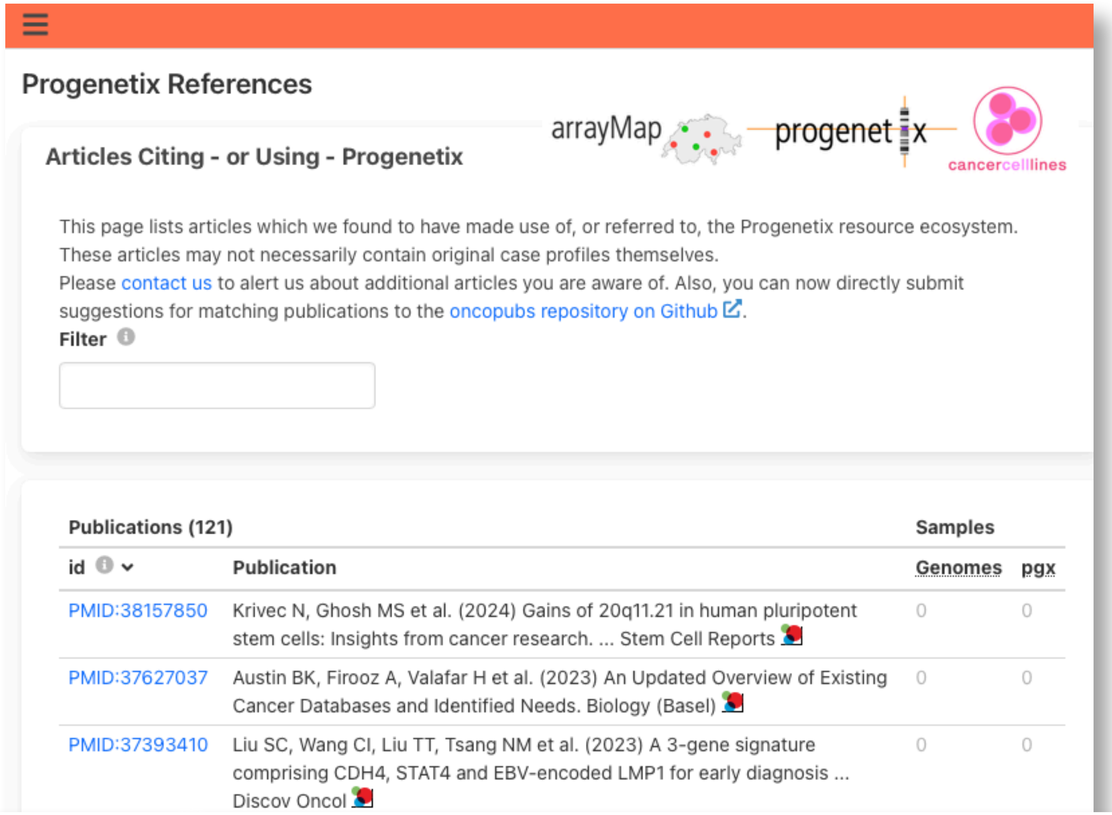
SIB Swiss Institute of Bioinformatics

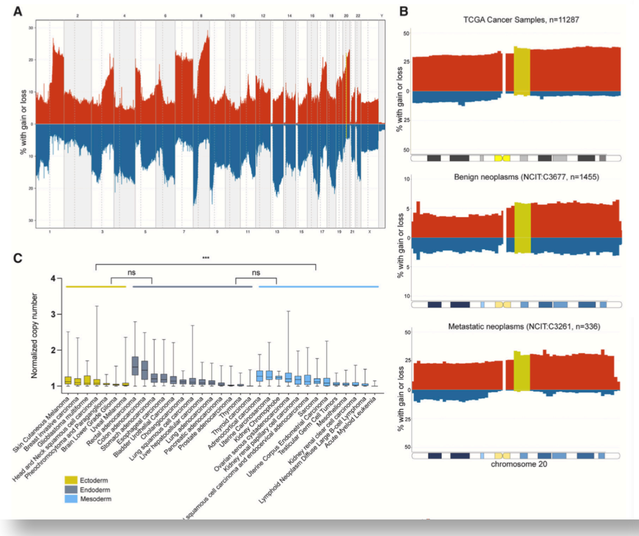


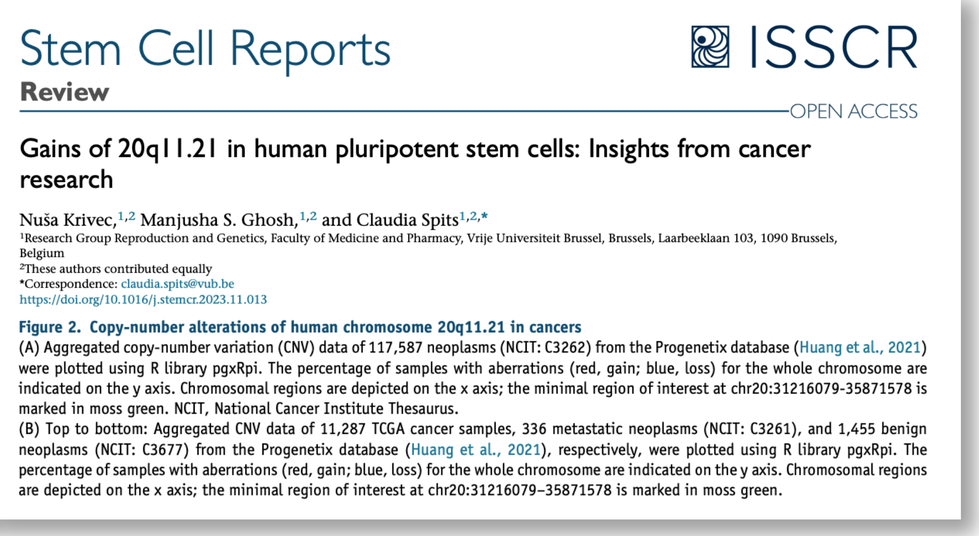
Progenetix Use

- CNV data is used e.g. as reference data in cancer genomics studies
- diagnosis specific CNV profiles serve as "fast look-up" in clinical genomics laboratories
- we loosely track publications in in our literature database but there is no systematic check-back mechanism...

Example: 2025 article using Progenetix' ***pgxRpi*** Beacon/R interface to retrieve & visualize 117'587 cancer CNV profiles for a study into pluripotent stem cells' genomics






Stem Cell Reports
ISSCR
Review
OPEN ACCESS

Gains of 20q11.21 in human pluripotent stem cells: Insights from cancer research

Nuša Krivec,[1,2] Manjusha S. Ghosh,[1,2] and Claudia Spits[1,2,*]
[1]Research Group Reproduction and Genetics, Faculty of Medicine and Pharmacy, Vrije Universiteit Brussel, Brussels, Laarbeeklaan 103, 1090 Brussels, Belgium
[2]These authors contributed equally
*Correspondence: claudia.spits@vub.be

https://doi.org/10.1016/j.stemcr.2023.11.013

Figure 2. Copy-number alterations of human chromosome 20q11.21 in cancers

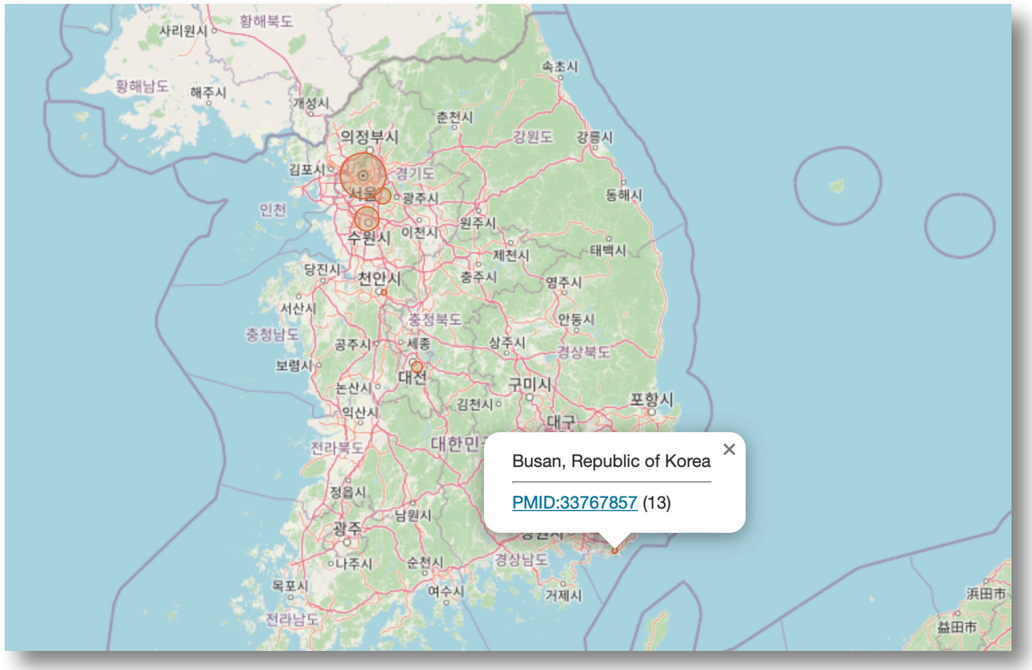
(A) Aggregated copy-number variation (CNV) data of 117,587 neoplasms (NCIT: C3262) from the Progenetix database (Huang et al., 2021) were plotted using R library pgxRpi. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079-35871578 is marked in moss green. NCIT, National Cancer Institute Thesaurus.

(B) Top to bottom: Aggregated CNV data of 11,287 TCGA cancer samples, 336 metastatic neoplasms (NCIT: C3261), and 1,455 benign neoplasms (NCIT: C3677) from the Progenetix database (Huang et al., 2021), respectively, were plotted using R library pgxRpi. The percentage of samples with aberrations (red, gain; blue, loss) for the whole chromosome are indicated on the y axis. Chromosomal regions are depicted on the x axis; the minimal region of interest at chr20:31216079–35871578 is marked in moss green.

Service: Publications

Location Mapping for Statistics and Discovery...

- all publications are tagged for "best fit" geographic origin in order
    1. specific sample origin
    2. processing laboratory
    3. corresponding author
- enables searches for e.g. "all publications or samples in HCC from 2000km around Taipeh"
- handy utility for discovering locally performed research, partners...



PMID:33767857

Methylation and molecular profiles of ependymoma: Influence of patient age and tumor anatomic location.

Cho HJ, Park HY, Kim K, Chae H, Paek SH, Kim SK, Park CK, Choi SH, Park SH.

Mol Clin Oncol PMID:33767857

http://progenetix.org/services/publications/?filters=genomes:>0&ISO3166alpha2=kr&output=map

Pushing the envelope...

Custom Beacon aggregation response for displaying CNV frequencies

progenetix

- CNV Profiles
  - ... by NCIT
  - ... by ICD-O Morphology
  - ... by ICD-O Site
  - ... by TNM & Grade
- Search Samples
- arrayMap
  - TCGA Data
  - cBioPortal Studies
- Publication DB
  - Progenetix Use
- NCIT - ICD-O Mappings
  - UBERON Mappings
- Upload & Plot
- OpenAPI Paths and Examples
- Cancer Cell Lines
- Beacon+
- Documentation
- Baudisgroup @ UZH

Select Summary Histograms

On this page you can combine one or multiple disease or other codes from one or more datasets frequencies). These based on pre-computed CNV frequencies for the given code; for a combina Page.

Dataset(s)

Progenetix cancer genome variants ×

Cancer Classification(s)

Select...

Various Subsets

pgx:TCGA-ACC: pgx:TCGA-ACC (180) × | pgx:TCGA-BLCA: pgx:TCGA-BLCA (810) × | pgx:TCGA-BRCA: pgx
pgx:TCGA-CESC: pgx:TCGA-CESC (586) × | pgx:TCGA-CHOL: pgx:TCGA-CHOL (85) × | pgx:TCGA-COAD: p
pgx:TCGA-DLBC: pgx:TCGA-DLBC (94) × | pgx:TCGA-ESCA: pgx:TCGA-ESCA (373) × | pgx:TCGA-GBM: pg
pgx:TCGA-HNSC: pgx:TCGA-HNSC (1081) × | pgx:TCGA-KICH: pgx:TCGA-KICH (132) × | pgx:TCGA-KIRC: p
pgx:TCGA-KIRP: pgx:TCGA-KIRP (594) × | pgx:TCGA-LAML: pgx:TCGA-LAML (285) × | pgx:TCGA-LGG: pgx
pgx:TCGA-LIHC: pgx:TCGA-LIHC (767) × | pgx:TCGA-LUAD: pgx:TCGA-LUAD (1110) × | pgx:TCGA-LUSC: p
pgx:TCGA-MESO: pgx:TCGA-MESO (173) × | pgx:TCGA-OV: pgx:TCGA-OV (1133) × | pgx:TCGA-PAAD: pgx:
pgx:TCGA-PCPG: pgx:TCGA-PCPG (363) × | pgx:TCGA-PRAD: pgx:TCGA-PRAD (1038) × | pgx:TCGA-READ
pgx:TCGA-SARC: pgx:TCGA-SARC (519) × | pgx:TCGA-SKCM: pgx:TCGA-SKCM (939) × | pgx:TCGA-STAD:
pgx:TCGA-TGCT: pgx:TCGA-TGCT (271) × | pgx:TCGA-THCA: pgx:TCGA-THCA (1025) × | pgx:TCGA-THYM:
pgx:TCGA-UCEC: pgx:TCGA-UCEC (1087) × | pgx:TCGA-UCS: pgx:TCGA-UCS (110) × | pgx:TCGA-UVM: pg

Plot Type

Select...

Gene Symbol(s) for Labeling

MYC (chr8:127735434-127742951) × | TP53 (chr17:7668421-7687490) × | CCND1 (chr11:69641156-69654
CDKN2A (chr9:21967752-21995324) × | EGFR (chr7:55019017-55211628) ×

Chromosomes to Plot

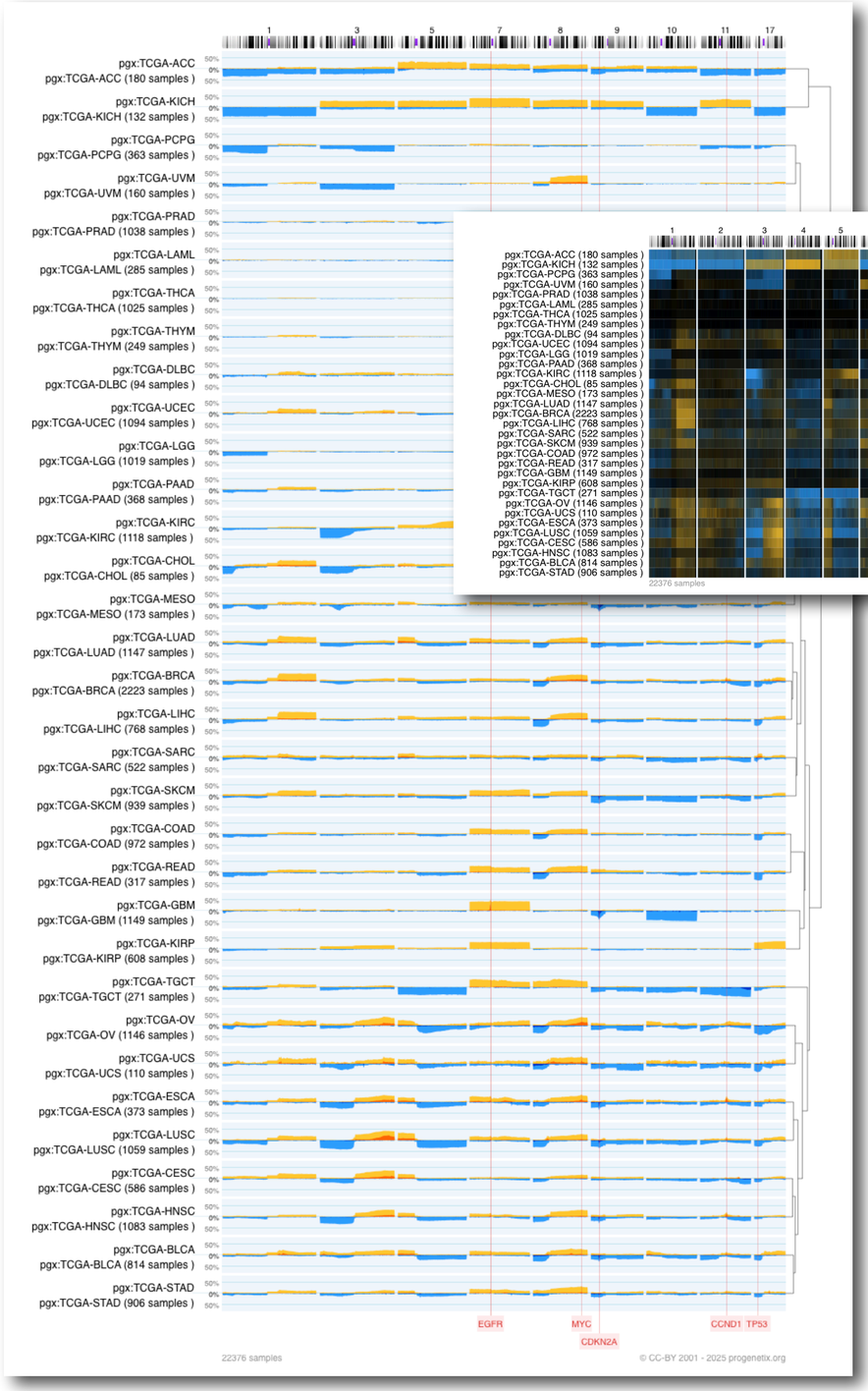
1,3,5,7,8,9,10,11,17

Plot Parameters

plot_axis_y_max:80;plot_area_height:40;plot_label_y_values:0,50

Retrieve CNV Profiles

B-Cell Lymphomas | TCGA Projects | Lung Cancers | Breast Cancers and Cell Lines



beacon plus

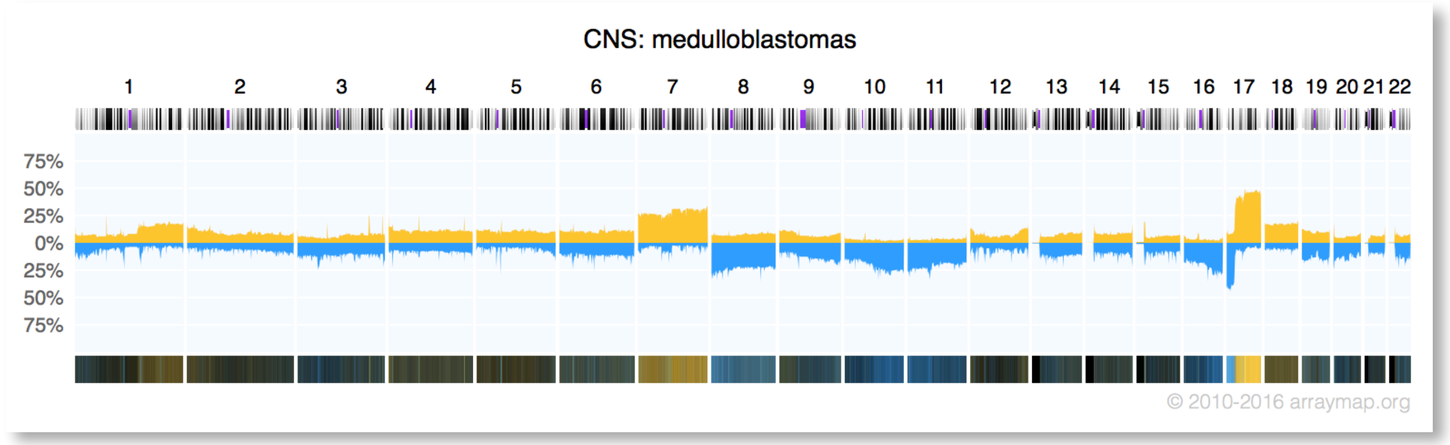
CNV frequencies in all TCGA projects, clustered and rendered as SVG, using ***filters*** query parameter over the Beacon framework and a custom aggregation response

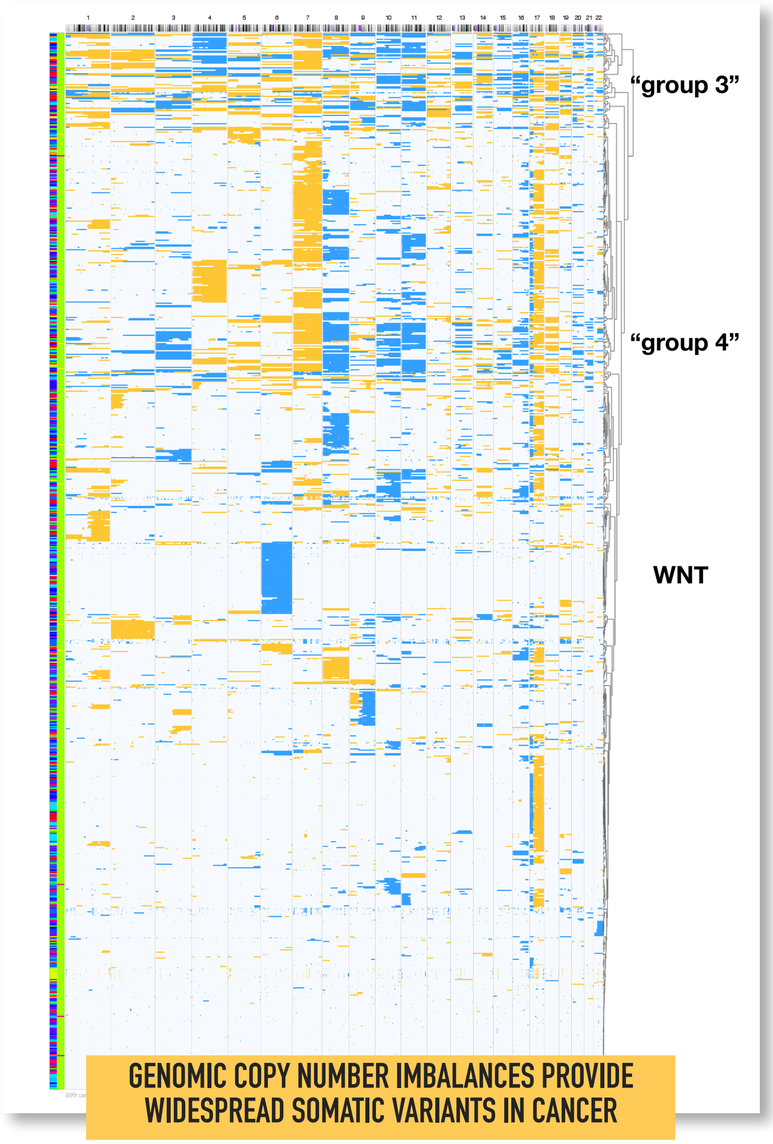
Data Use

Somatic CNVs In Cancer

Recurrent mutation patterns

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





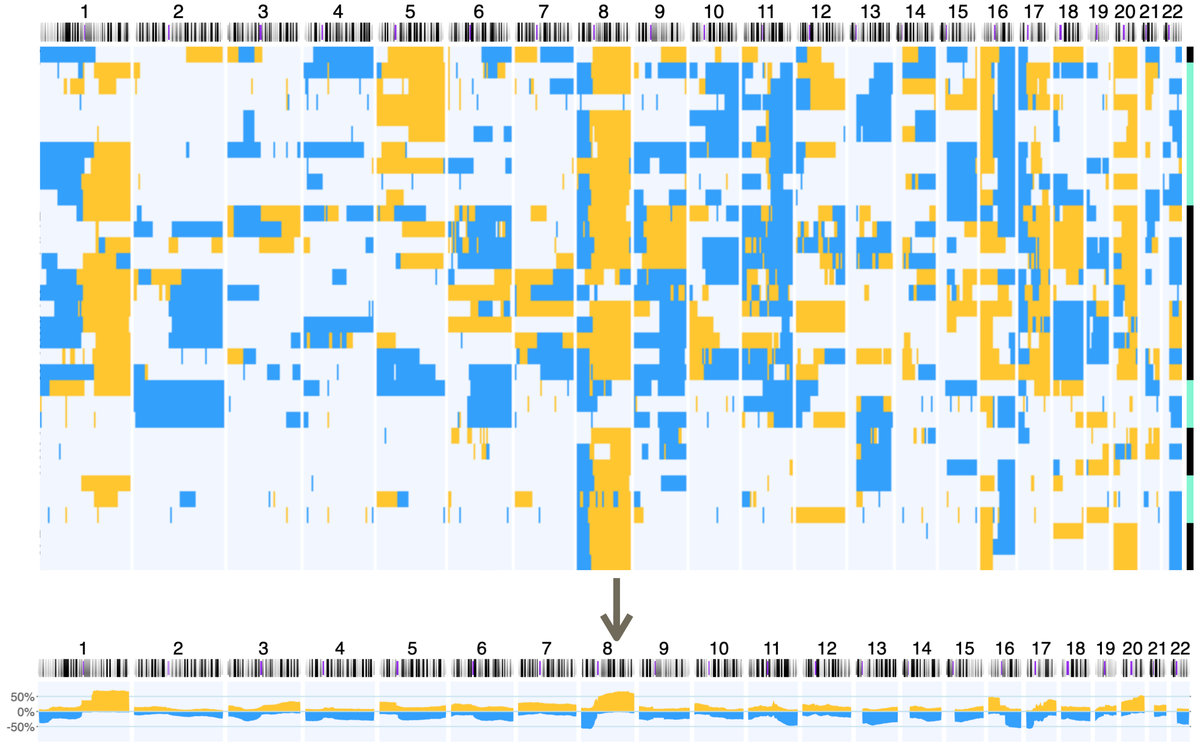
A genomic copy number histogram for malignant medulloblastomas, the most frequent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional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

arrayMap

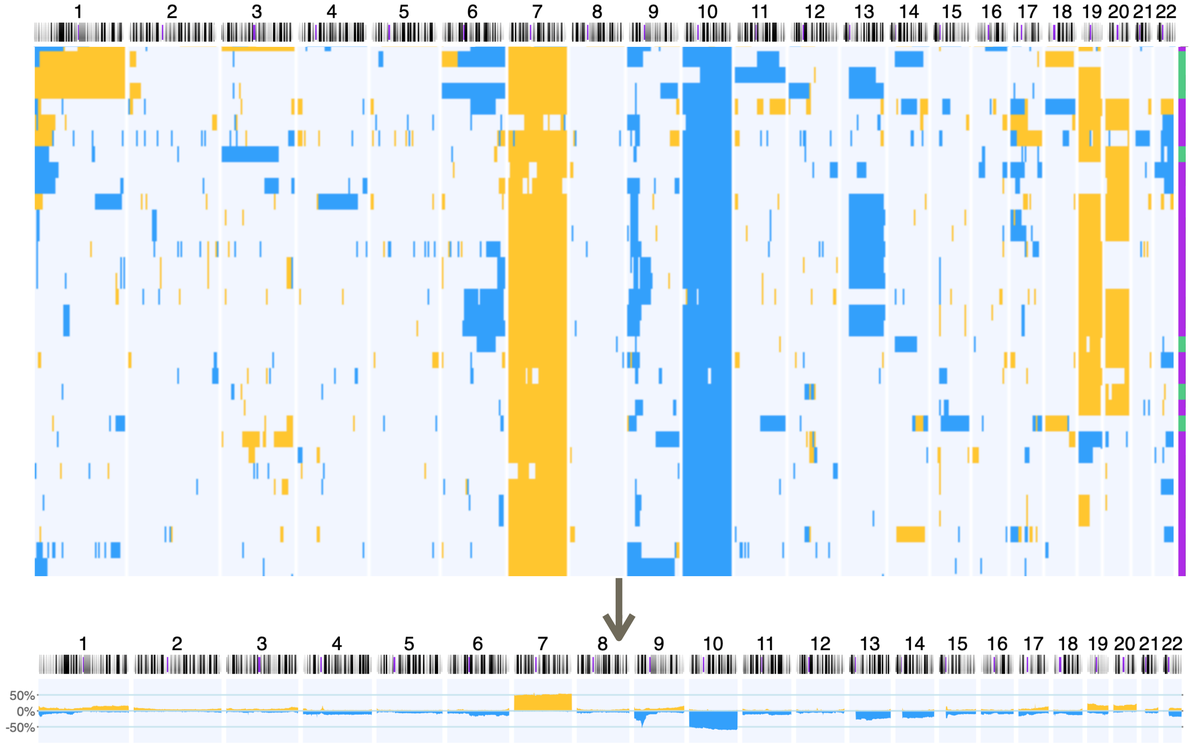
Drivers? Passengers? Markers?

Disentangling CNA Patterns

Ductal Breast Carcinoma

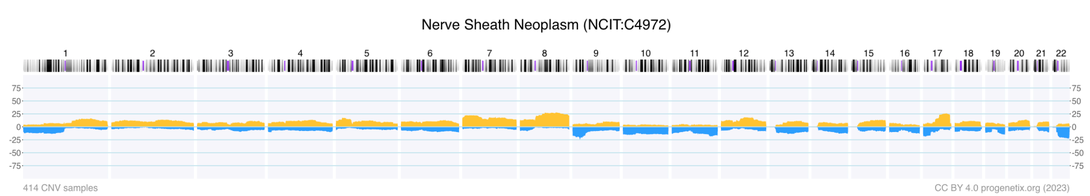


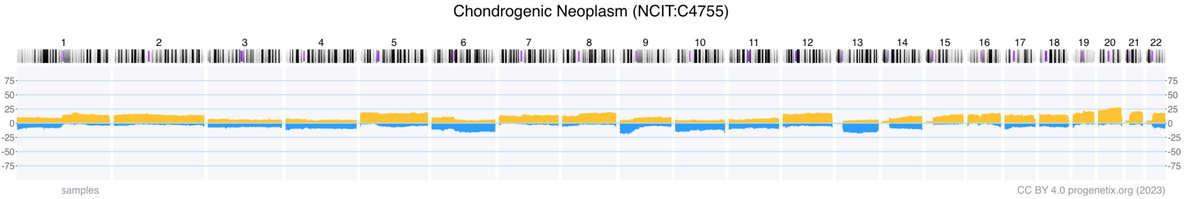
Glioblastoma

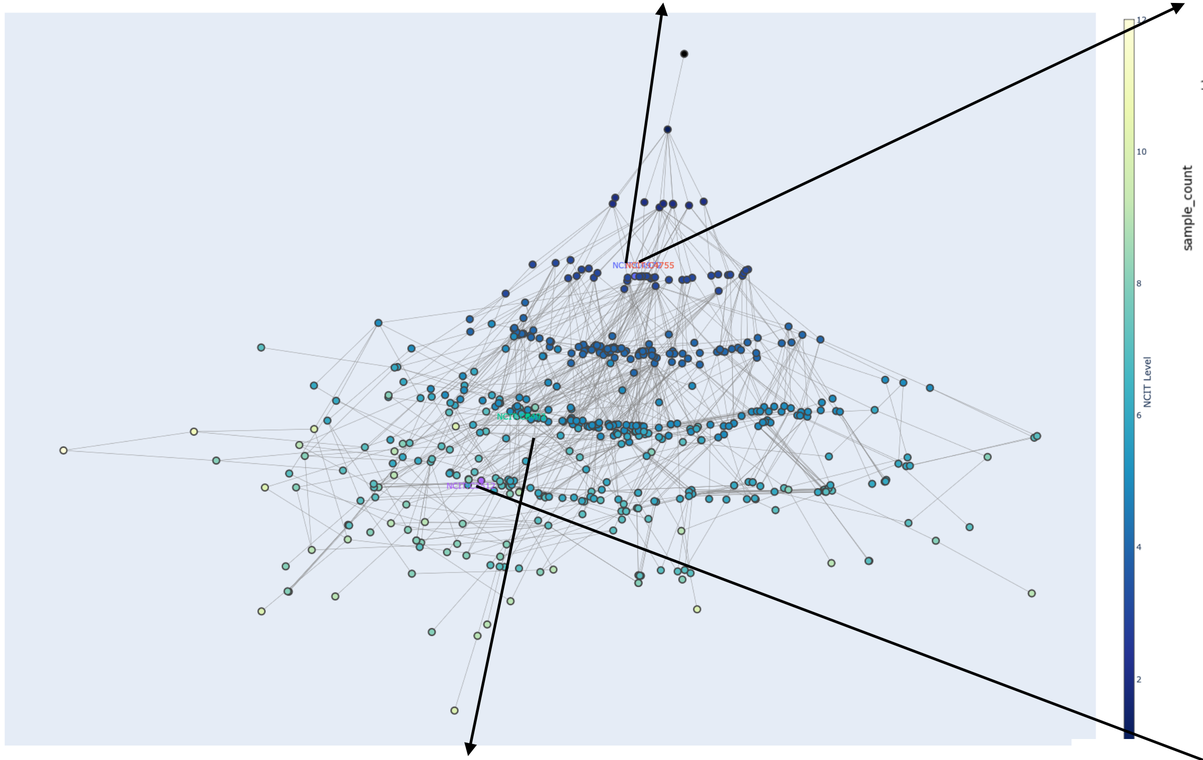


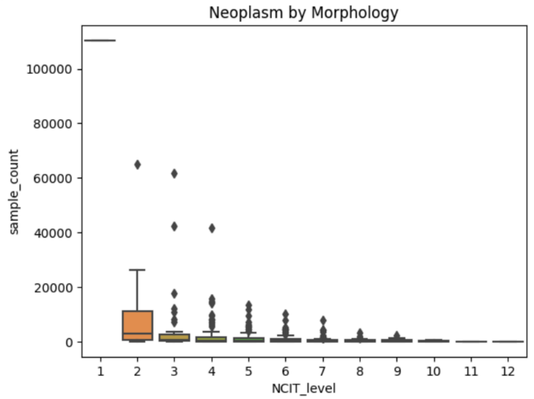
CNV profiles heterogeneity vs cancer classification

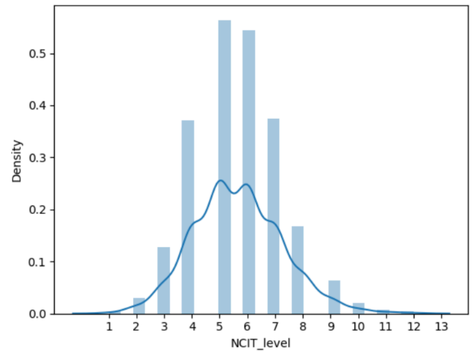
Correspondance of genomic profiles to NCIT cancer hierarchy

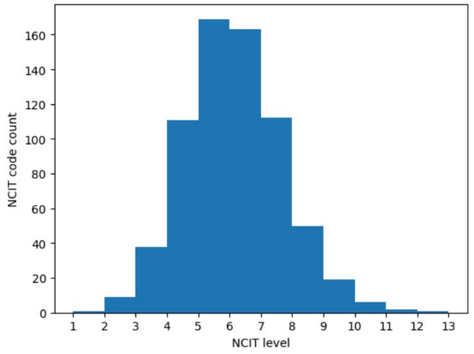


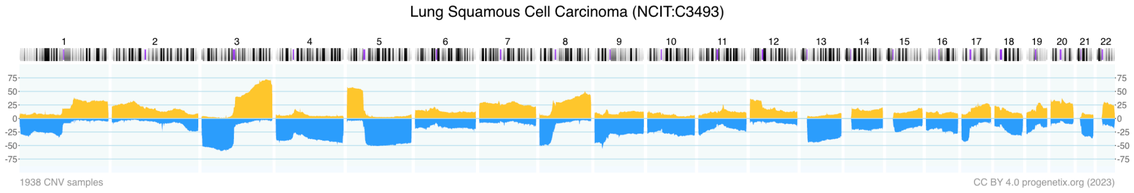


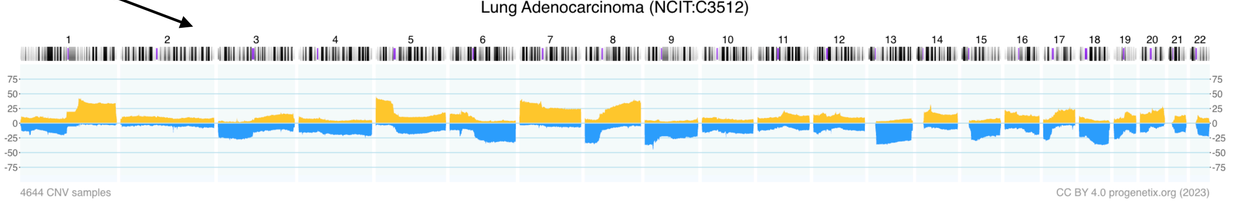












Lead: Ziying Yang

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

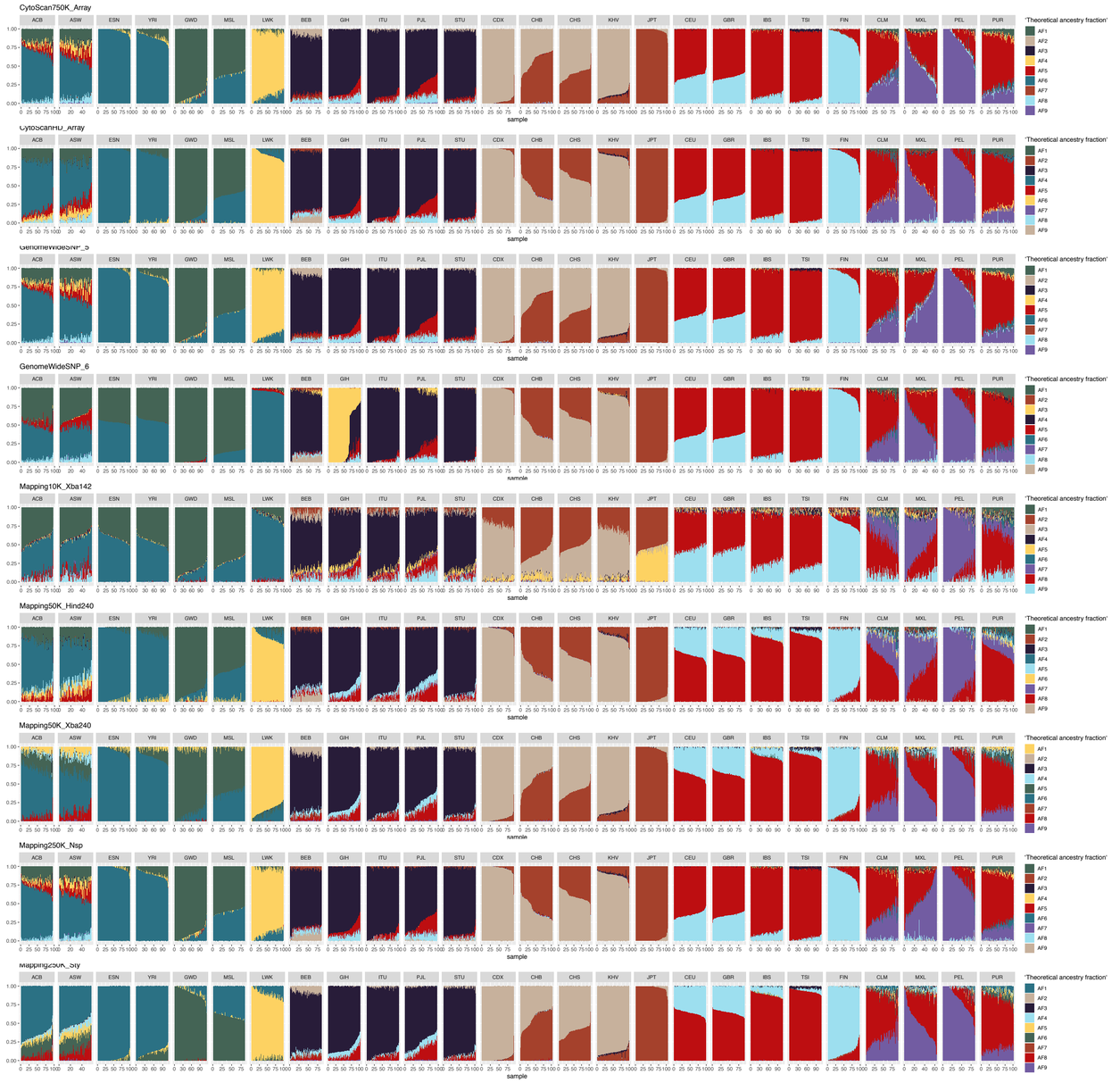


Figure S1 The fraction or contribution of theoretical ancestors (k=9) in reference individuals from 1000 Genomes Project with regard to nine SNP array platforms. The x-axis are individual samples, grouped by their respective population. Groups belonging to the same continent/superpopulation are placed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

arrayMap

Lead: Qingyao Huang

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*]

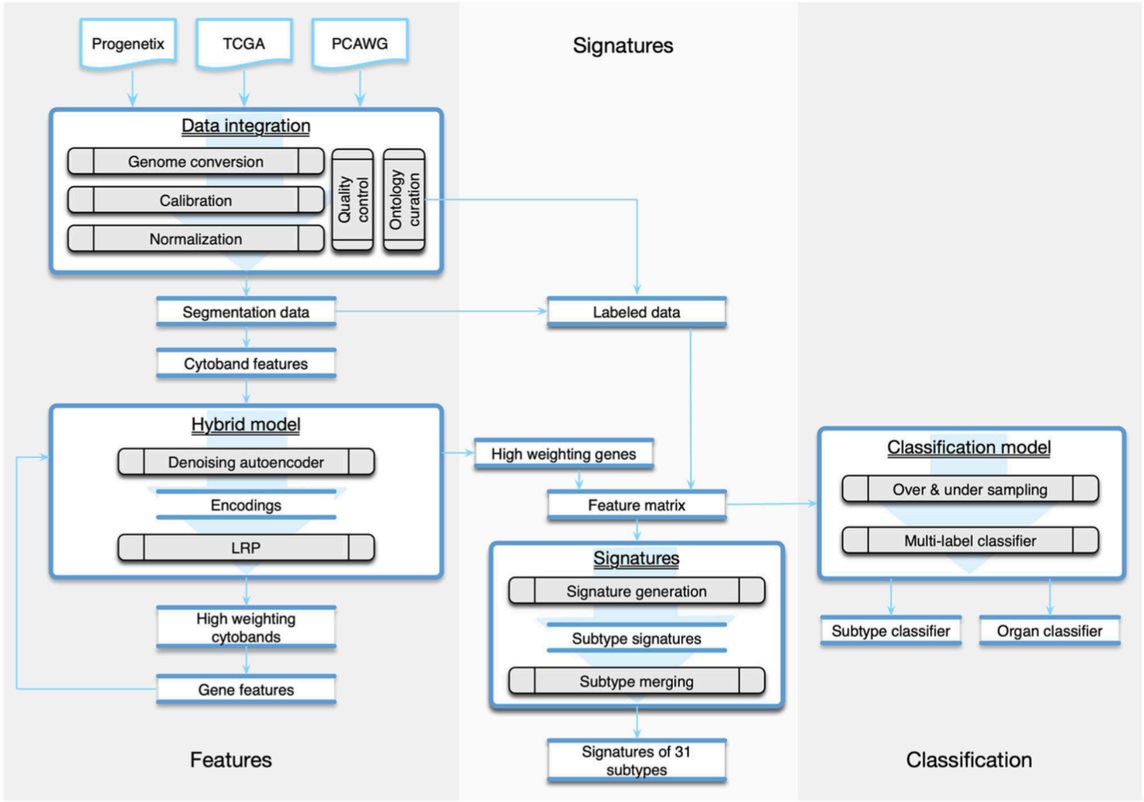


FIGURE 1 | The workflow of the study was composed of three parts. The *Features* part consisted of methods of data integration and feature generation. The *Signature* part focused on creating CNA signatures for cancer subtypes and the categorization of subtypes. The *Classification* part recruited machine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

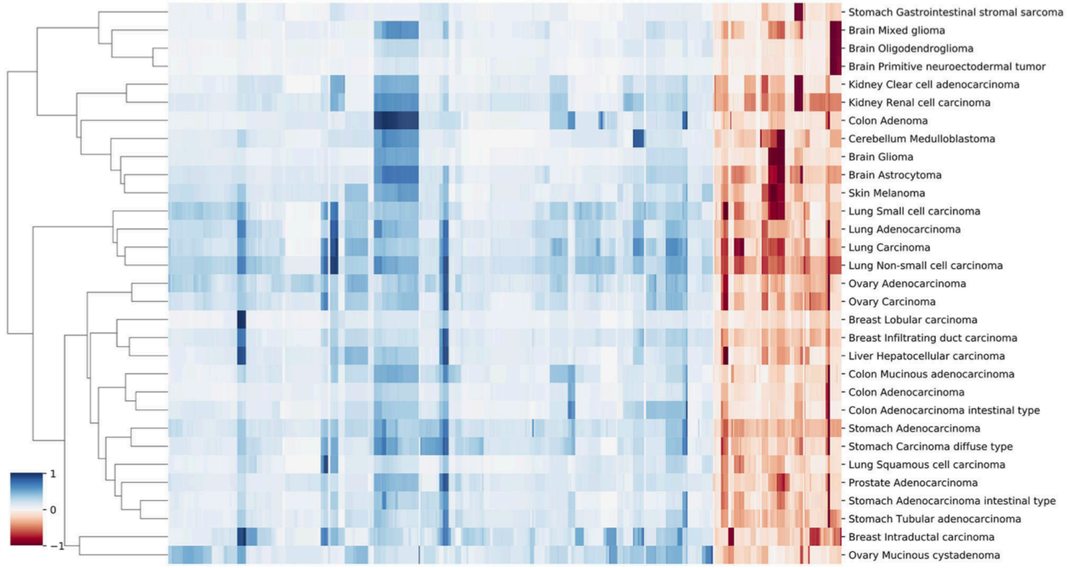


FIGURE 5 | A clustering heatmap of features in 31 signatures. Columns are normalized average CNV intensities of feature genes, where the blue colors are duplication features and red colors are deletion features. Duplication and deletion frequencies are normalized separately.

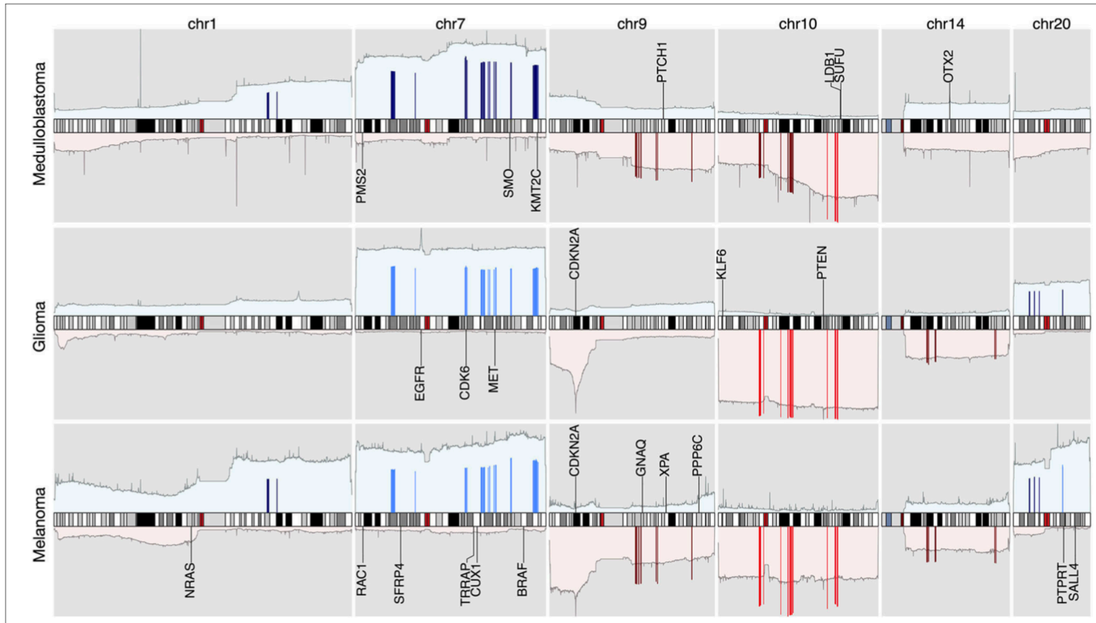
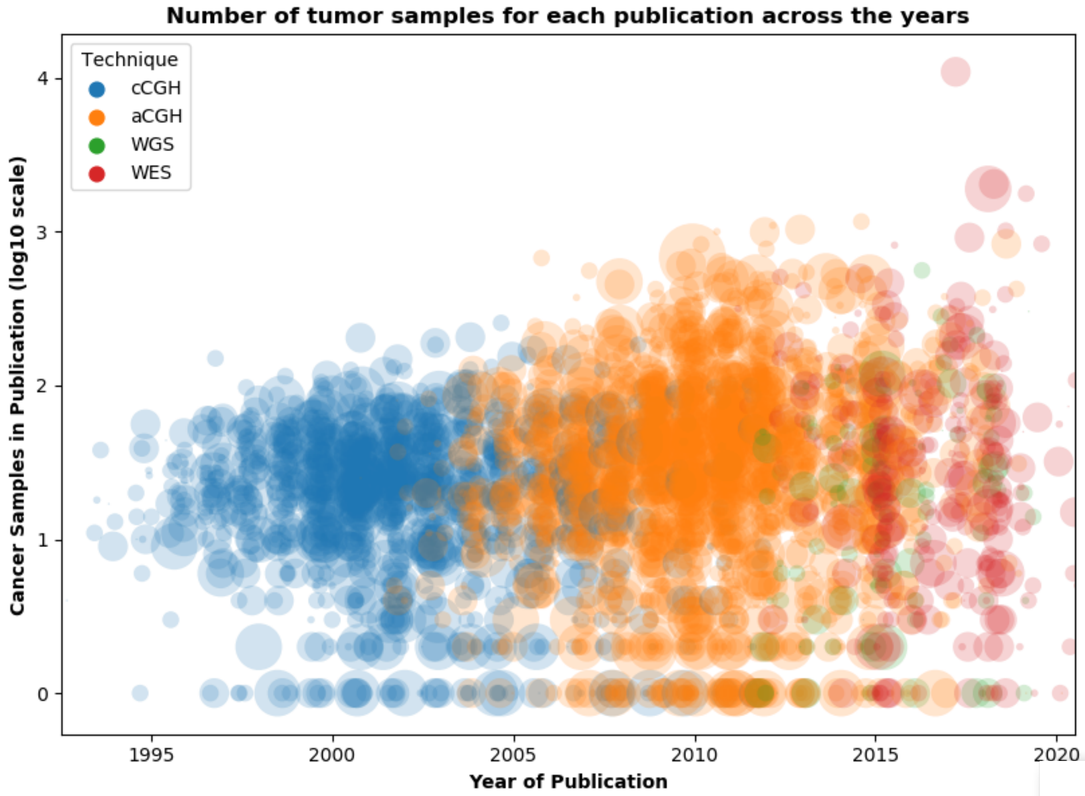


FIGURE 6 | The integrated view of the original data and the selected features, in the neural crest originating entities medulloblastoma, glioma, and melanoma. The shaded background area color illustrates the original data. Color bars illustrate the feature genes, where brighter colors indicate stronger 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.



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.

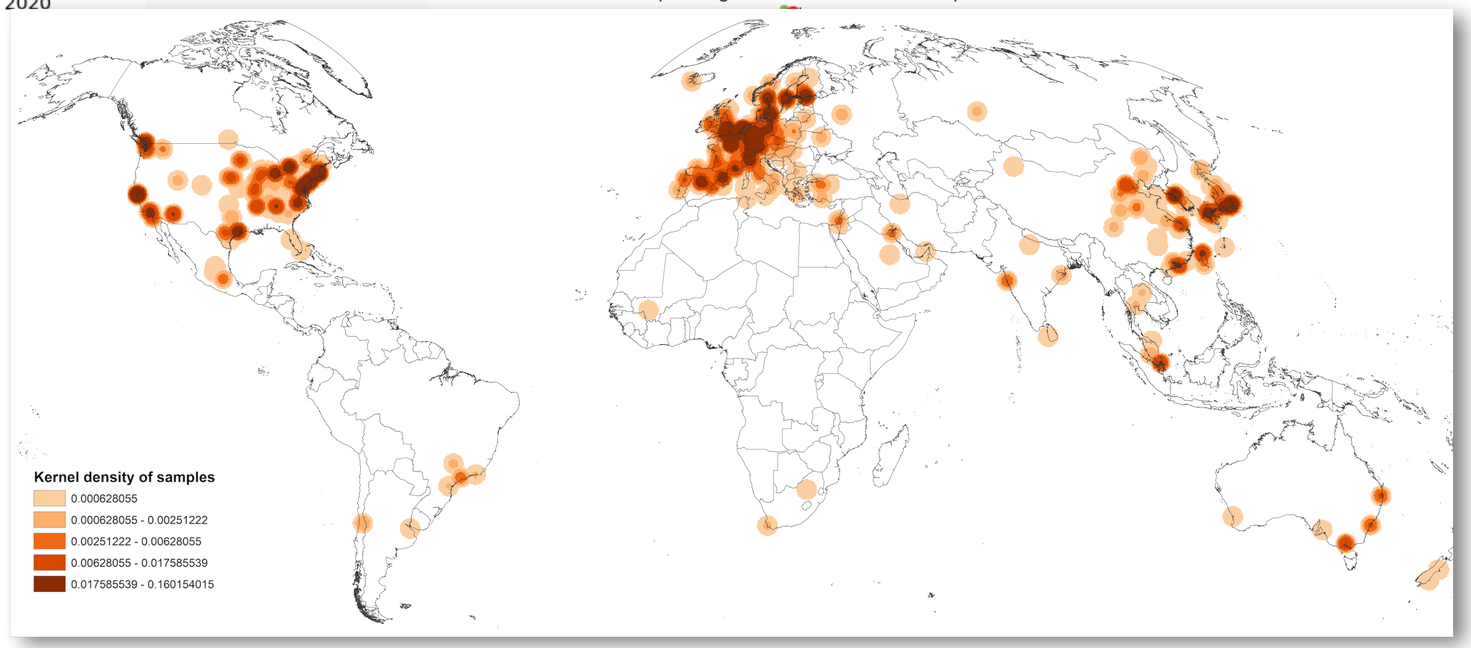
Please contact us to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation.

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

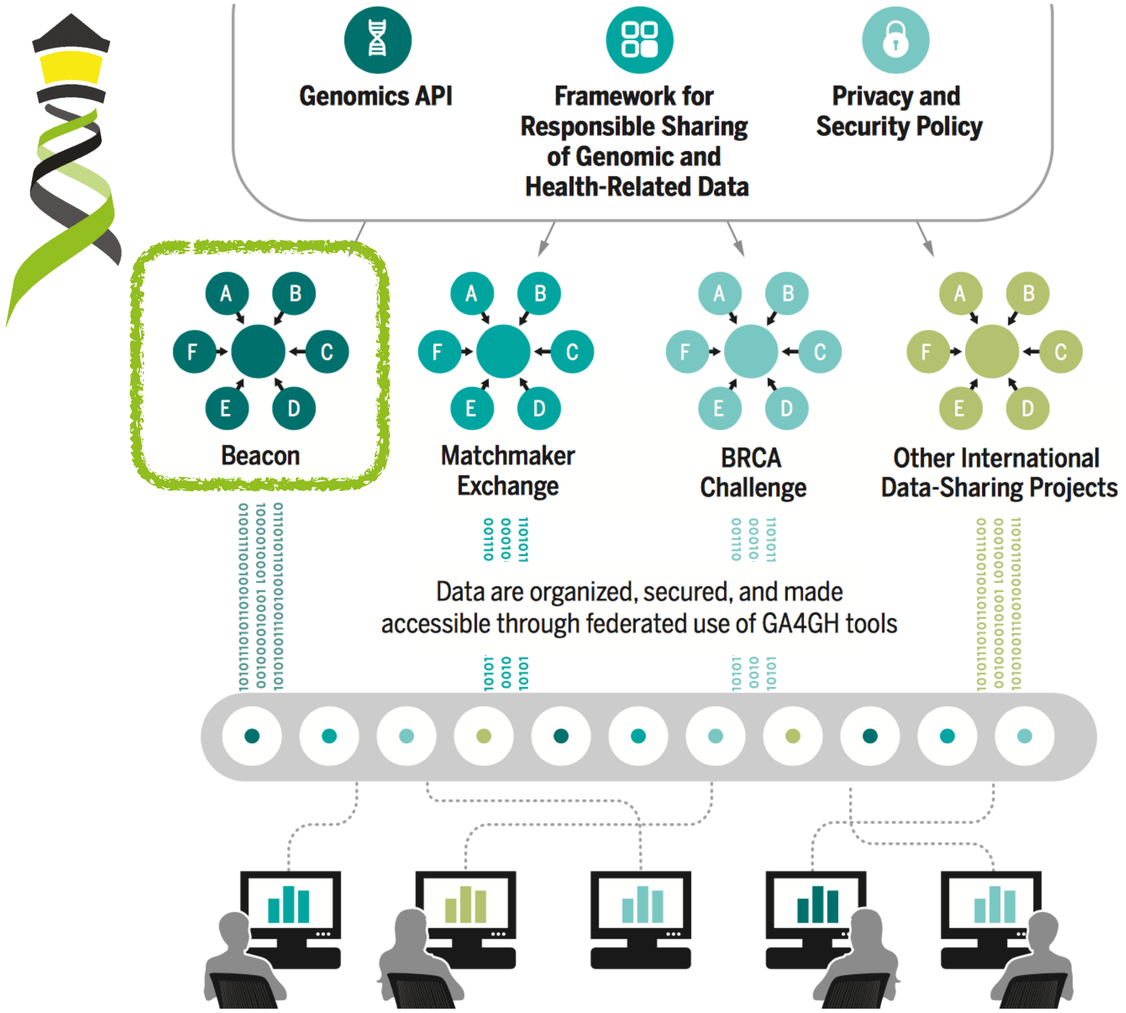
A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

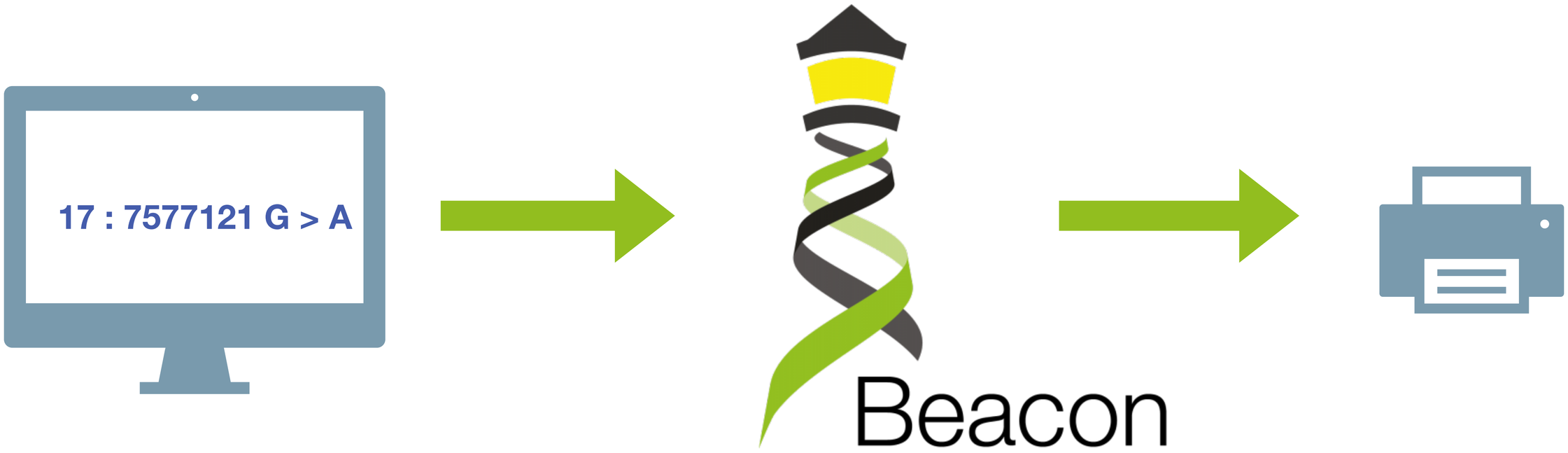
SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

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



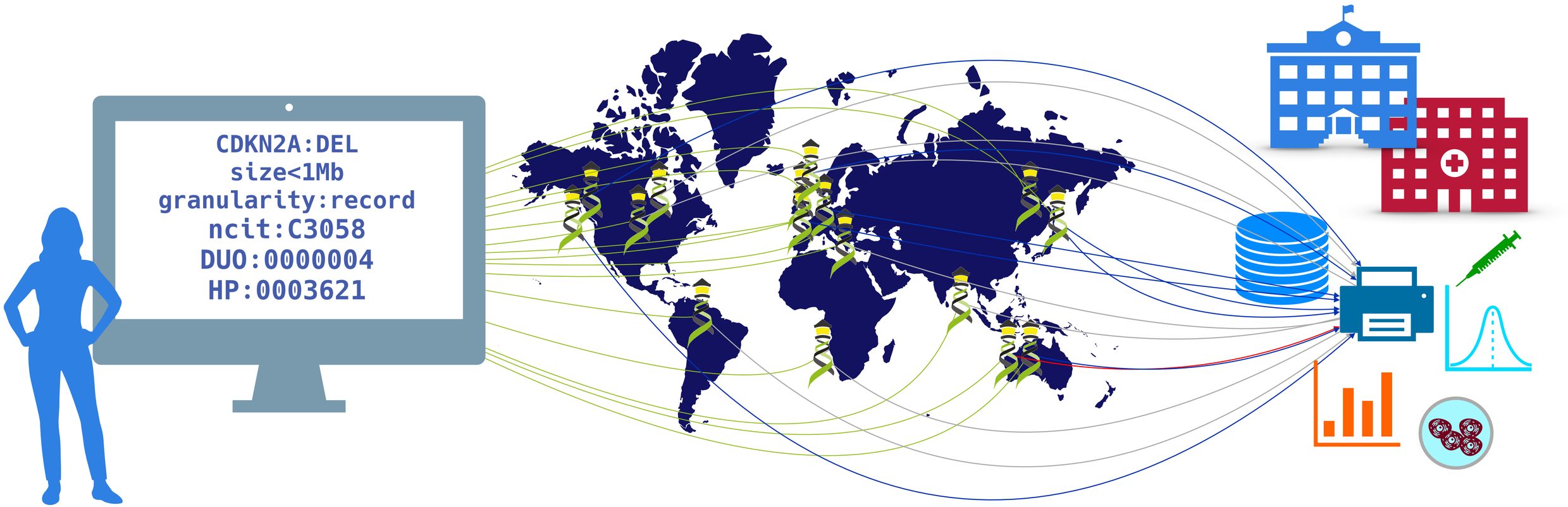
DNASTACK

Global Alliance for Genomics & Health



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

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



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

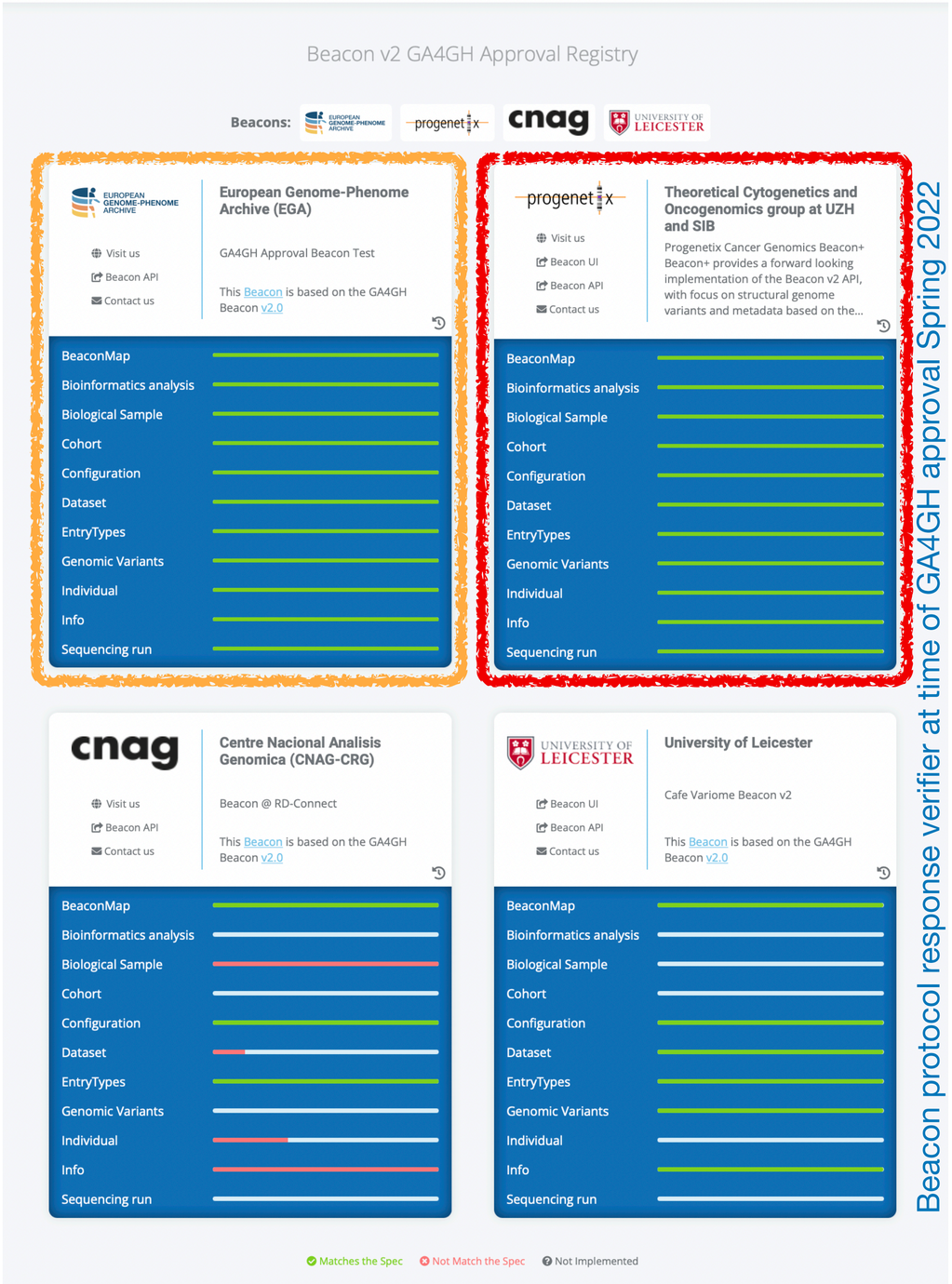
Beacon API

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

bycon Beacon+

Implementation driven standards development

- Progenetix' Beacon+ has served as implementation driver since 2016
- the *bycon* package is used to prototype advanced Beacon features such as
  - ➡ structural variant queries
  - ➡ data handovers
  - ➡ Phenopackets integration
  - ➡ variant co-occurrences
  - ➡ ...



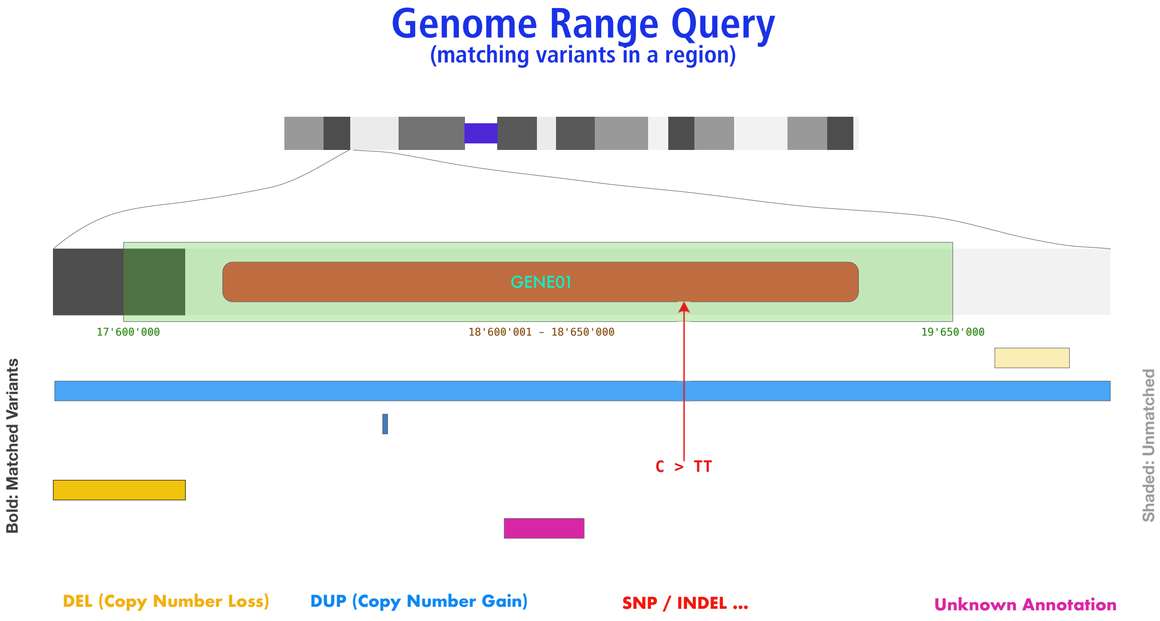
Global Alliance for Genomics & Health

progenetix

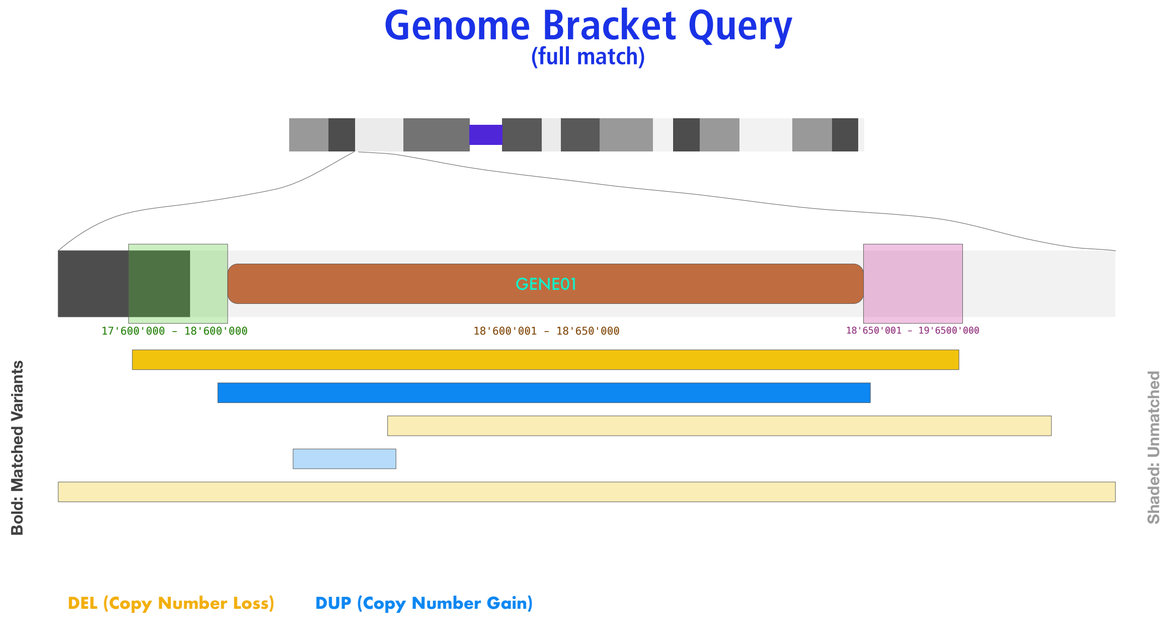
elixir

Beacon v2: Extended Variant Queries

Range and Bracket queries enable positional wildcards and fuzziness



- 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 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 Paths

Progenetix utilizes Beacon v2 REST paths and GET requests

- 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
- 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 v2 Requests

POSTing Queries

- Beacon v2 supports a mix of dedicated endpoints with REST paths
- POST requests using JSON query documents

```
{
    "$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" }
    ]
}
```

Website populated by asynchronous retrieval of Beacon query results using handovers

progenetix

- **CNV Profiles**
  - ... by NCIT
  - ... by ICD-O Morphology
  - ... by ICD-O Site
  - ... by TNM & Grade
- **Search Samples**
- **arrayMap**
  - TCGA Data
  - cBioPortal Studies
- **Publication DB**
  - Progenetix Use
- **NCIT - ICD-O Mappings**
  - UBERON Mappings
- **Upload & Plot**
- **OpenAPI Paths and Examples**
- **Cancer Cell Lines**

Edit Query

Chro: refseq:NC_000009.12 **Start:** 21000001,21975098 **End:** 21967753,23000000 **Type:** EFO:0030067
Filters: NCIT:C3058

progenetix

Matched Samples: 969
Retrieved Samples: 200
Variants: 984
Calls: 976

UCSC region
Geographic Map
Variants in UCSC
Dataset Responses (JSON)

Visualization options

Results | Biosamples | Variants

Reload histogram in new window

Matched Subset Codes	Subset Samples	Matched Samples	Subset Match Frequencies
pgx:icdot-C71.4	4	1	0.250
pgx:icdot-C71.1	14	1	0.071
pgx:icdom-94403	4816	200	0.042
NCIT:C3058	4900	200	0.041
pgx:icdot-C71.9	13758	192	0.014
pgx:icdot-C71.0	1714	6	0.004

progenetix Data Downloads

Download Sample Data (TSV)
Part1 Part2 Part3 Part4 Part5

Download Sample Data (JSON)
Part1 Part2 Part3 Part4 Part5

Download Variants (Beacon VRS)
Part1 Part2 Part3 Part4 Part5

Download Variants (VCF)

Results | Biosamples | Variants

Biosample Id	Dx Classifications	Identifiers
pgxbs-kftvl1hz	pgx:icdom-94403 Glioblastoma, NOS pgx:icdot-C71.9 Brain, NOS NCIT:C3058 Glioblastoma	pubmed:28481359 Zehir A, Benayed R et al. (2017): Mutational landscape of metastatic cancer revealed... cbioportal:msk_impact_2017
pgxbs-kftvl7f4	pgx:icdom-94403 Glioblastoma, NOS pgx:icdot-C71.9 Brain, NOS NCIT:C3058 Glioblastoma	pubmed:28481359 Zehir A, Benayed R et al. (2017): Mutational landscape of metastatic cancer revealed... cbioportal:msk_impact_2017
pgxbs-kftvhm6s	pgx:icdom-94403 Glioblastoma, NOS pgx:icdot-C71.9 Brain, NOS NCIT:C3058 Glioblastoma	pgx:TCGA-GBM Glioblastoma Multiforme 18772890 Cancer Genome Atlas Research Network. (2008): Comprehensive genomic characterization defines human glioblastoma...

Results | Biosamples | Variants

Digest	Gene	Pathogenicity	Variant type	Variant Instances
9:21626201-21981584:EFO_0030068			CopyNumberChange	V: pgxvar-665749ab2d6be9a260e55de8 A: pgxcs-kftwnmzs B: pgxbs-kftvjywz I: pgxind-kftx5yjj
9:21846286-22201587:EFO_0030068			CopyNumberChange	V: pgxvar-6656fc5fbe3f6845a3555b82 A: pgxcs-kftw53z6 B: pgxbs-kftvi872 I: pgxind-kftx3t9l
9:21949762-22004847:EFO_0020073			CopyNumberChange	V: pgxvar-6657226e8f6b961582615aa6 A: pgxcs-kftw3vh5 B: pgxbs-kftvi4e1 I: pgxind-kftx3vx4
9:21164528-21990552:EFO_0020073			CopyNumberChange	V: pgxvar-66572dfe2d6be9a260e3d189 A: pgxcs-kftw95rl B: pgxbs-kftvilhz I: pgxind-kftx4au8

Website populated by asynchronous retrieval of Beacon query results using handovers

progenetix

- **CNV Profiles**
  - ... by NCIT
  - ... by ICD-O Morphology
  - ... by ICD-O Site
  - ... by TNM & Grade
- **Search Samples**
- **arrayMap**
  - TCGA Data
  - cBioPortal Studies
- **Publication DB**
  - Progenetix Use
- **NCIT - ICD-O Mappings**
  - UBERON Mappings
- **Upload & Plot**
- **OpenAPI Paths and Examples**
- **Cancer Cell Lines**

Edit Query

Chro: refseq:NC_000009.12 **Start:** 21000001,21975098 **End:** 21967753,23000000 **Type:** EFO:0030067
Filters: NCIT:C3058

progenetix

Matched Samples: 969
Retrieved Samples: 200
Variants: 984
Calls: 976

UCSC region
Geographic Map
Variants in UCSC
Dataset Responses (JSON)

Visualization options

Results | Biosamples | Variants

Reload histogram in new window

Matched Subset Codes	Subset Samples	Matched Samples	Subset Match Frequencies
pgx:icdot-C71.4	4	1	0.250
pgx:icdot-C71.1	14	1	0.071
pgx:icdom-94403	4816	200	0.042
NCIT:C3058	4900	200	0.041
pgx:icdot-C71.9	13758	192	0.014
pgx:icdot-C71.0	1714	6	0.004

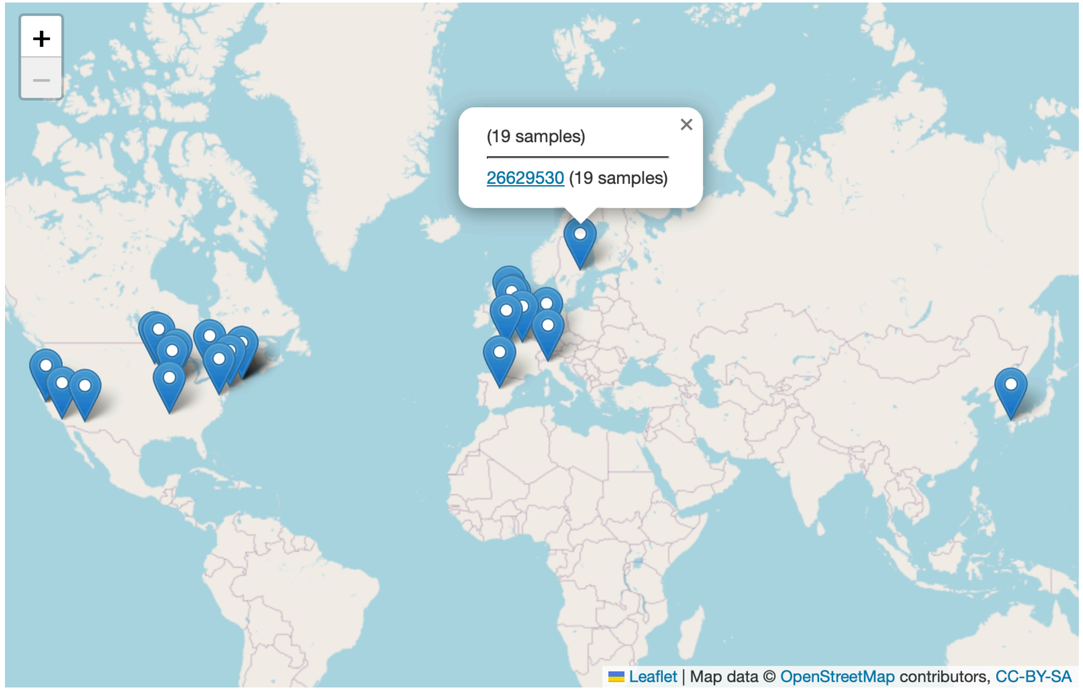
progenetix Data Downloads

Download Sample Data (TSV)
Part1 Part2 Part3 Part4 Part5

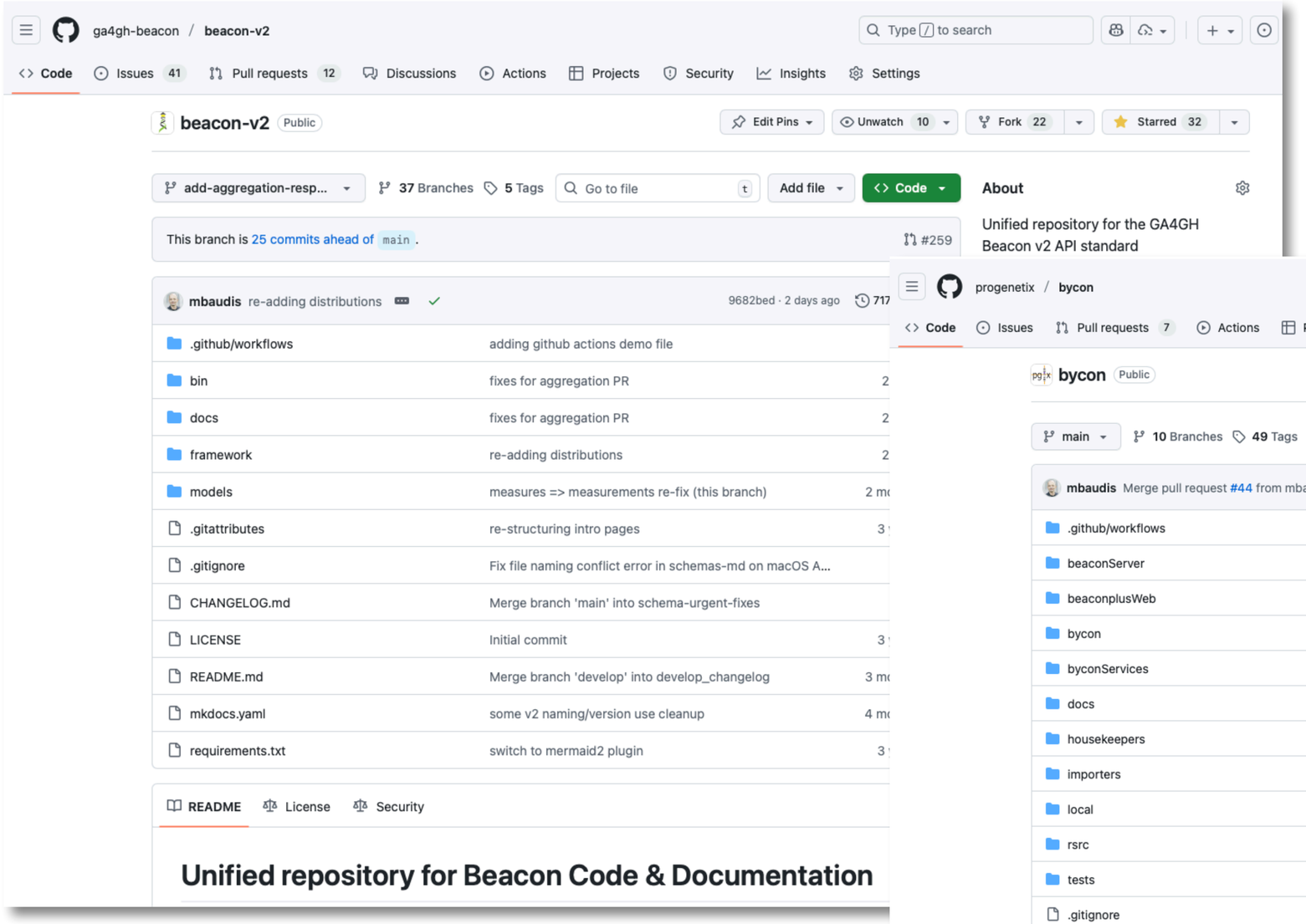
Download Sample Data (JSON)
Part1 Part2 Part3 Part4 Part5

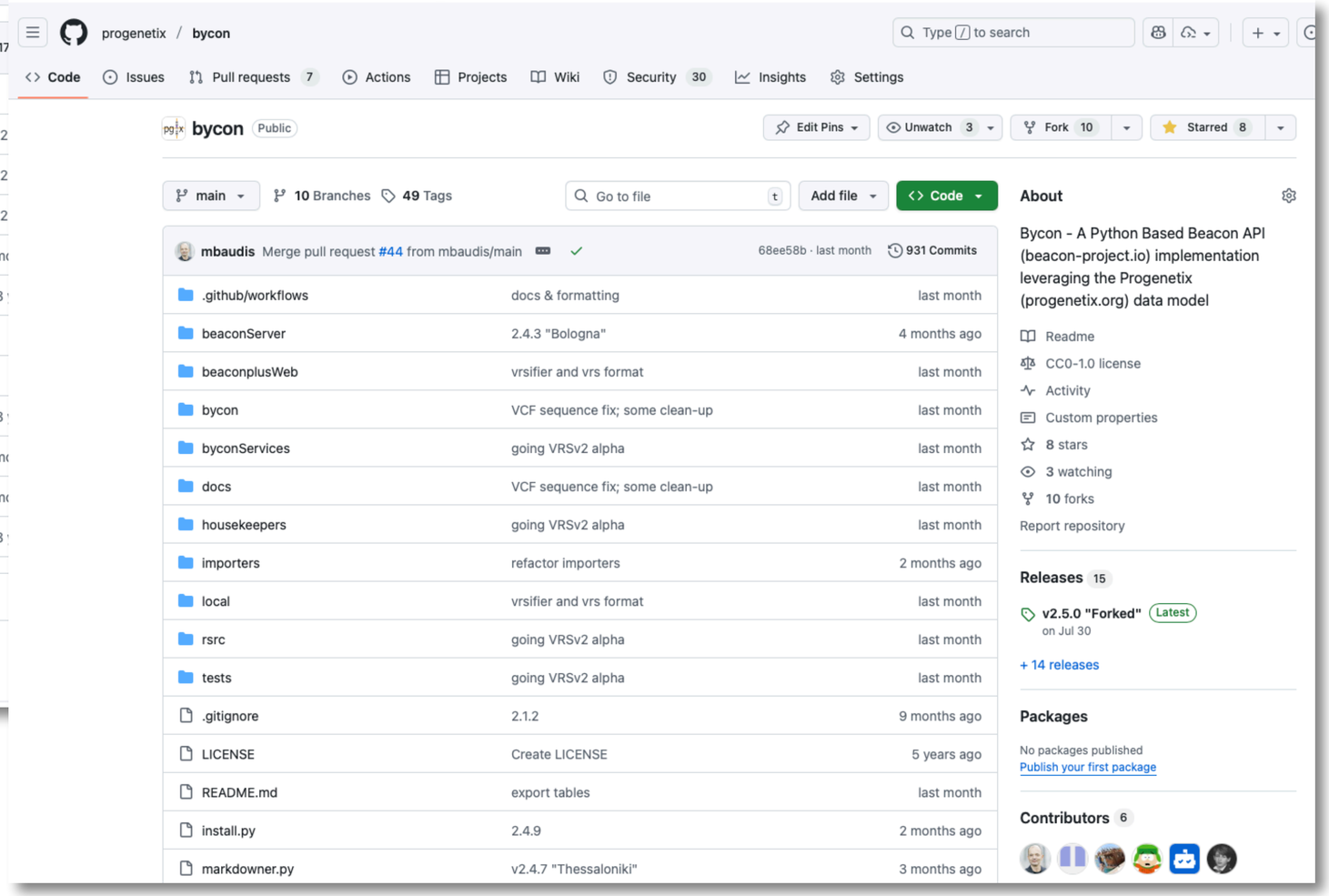
Download Variants (Beacon VRS)
Part1 Part2 Part3 Part4 Part5

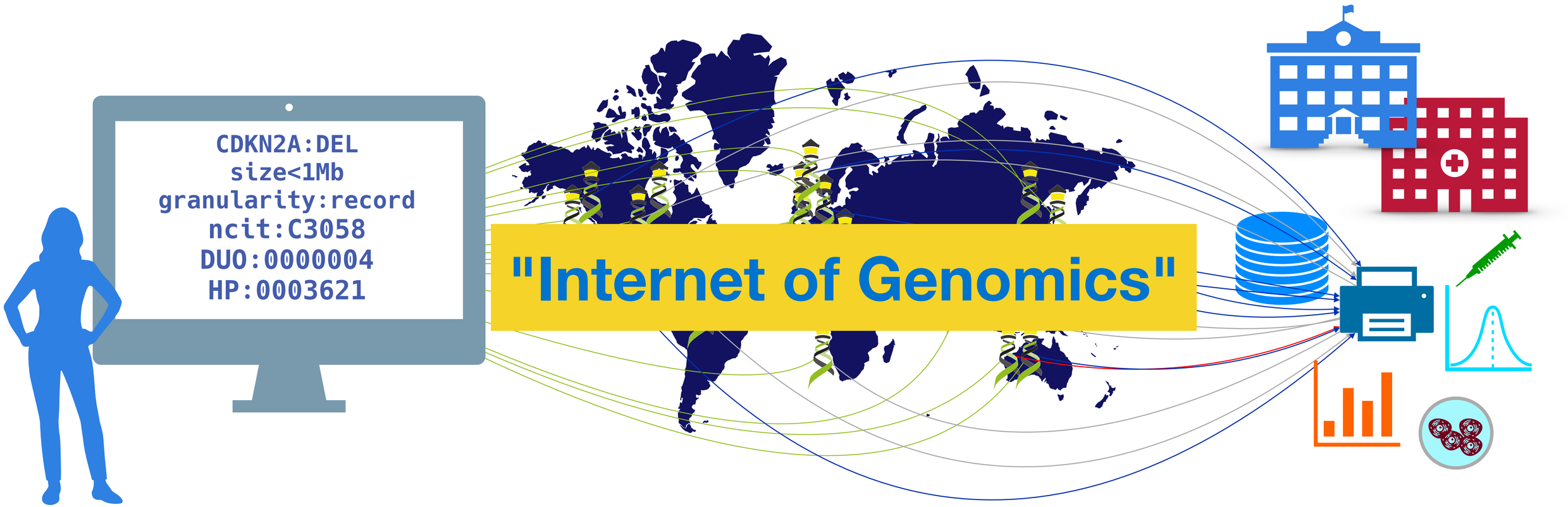
Download Variants (VCF)



Pushing the standard: Biosamples in Progenetix have geographic attribution in the form of GeoJSON objects, for query & display...


ga4gh-beacon / beacon-v2
Type / to search
Code
Issues 41
Pull requests 12
Discussions
Actions
Projects
Security
Insights
Settings
beacon-v2
Public
Edit Pins
Unwatch 10
Fork 22
Starred 32
add-aggregation-resp...
37 Branches
5 Tags
Go to file
t
Add file
Code
About
Unified repository for the GA4GH Beacon v2 API standard
This branch is 25 commits ahead of main .
#259
mbaudis re-adding distributions
9682bed · 2 days ago
.github/workflows
adding github actions demo file
bin
fixes for aggregation PR
docs
fixes for aggregation PR
framework
re-adding distributions
models
measures => measurements re-fix (this branch)
.gitattributes
re-structuring intro pages
.gitignore
Fix file naming conflict error in schemas-md on macOS A...
CHANGELOG.md
Merge branch 'main' into schema-urgent-fixes
LICENSE
Initial commit
README.md
Merge branch 'develop' into develop_changelog
mkdocs.yaml
some v2 naming/version use cleanup
requirements.txt
switch to mermaid2 plugin
README
License
Security
Unified repository for Beacon Code & Documentation


progenetix / bycon
Type / to search
Code
Issues
Pull requests 7
Actions
Projects
Wiki
Security 30
Insights
Settings
bycon
Public
Edit Pins
Unwatch 3
Fork 10
Starred 8
main
10 Branches
49 Tags
Go to file
t
Add file
Code
About
Bycon - A Python Based Beacon API (beacon-project.io) implementation leveraging the Progenetix (progenetix.org) data model
mbaudis Merge pull request #44 from mbaudis/main
68ee58b · last month
931 Commits
.github/workflows
docs & formatting
last month
beaconServer
2.4.3 "Bologna"
4 months ago
beaconplusWeb
vrsifier and vrs format
last month
bycon
VCF sequence fix; some clean-up
last month
byconServices
going VRSv2 alpha
last month
docs
VCF sequence fix; some clean-up
last month
housekeepers
going VRSv2 alpha
last month
importers
refactor importers
2 months ago
local
vrsifier and vrs format
last month
rsrc
going VRSv2 alpha
last month
tests
going VRSv2 alpha
last month
.gitignore
2.1.2
9 months ago
LICENSE
Create LICENSE
5 years ago
README.md
export tables
last month
install.py
2.4.9
2 months ago
markdowner.py
v2.4.7 "Thessaloniki"
3 months ago
Readme
CC0-1.0 license
Activity
Custom properties
8 stars
3 watching
10 forks
Report repository
Releases 15
v2.5.0 "Forked"
Latest
on Jul 30
+ 14 releases
Packages
No packages published
Publish your first package
Contributors 6



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

Beacon API

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

Making use of Progenetix' Beacon API

Data analysis through integration with R

pgxRpi

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

All users

R users

Interface

Variant query

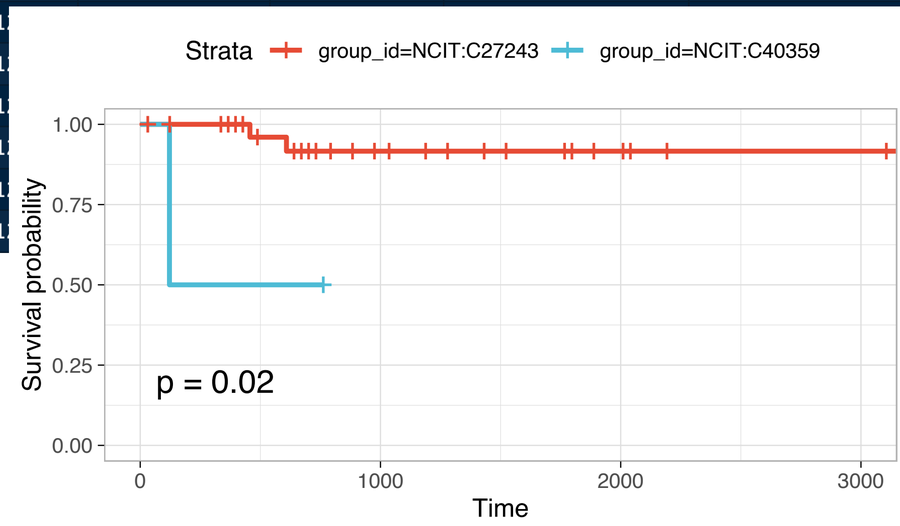
https://progenetix.org/beacon/biosamples/pgxbs-kftvh94d/g_variants

```
variants <- pgxLoader(type="variant",biosample_id="pgxbs-kftvh94d")
```

Output


```
"results": [
    {
        "caseLevelData": [
            {
                "analysisId": "pgxcs-kftvu6cg",
                "biosampleId": "pgxbs-kftvh94d",
                "id": "pgxvar-5bab5837727983b2e0121e97"
            }
        ],
        "variantInternalId": "11:0-134452384:DEL",
        "variation": {
            "copyChange": "efo:0030067",
            "identifiers": {},
            "subject": {
                "interval": {
                    "end": {
                        "type": "Number",
                        "value": 134452384
                    },
                    "start": {
                        "type": "Number",
                        "value": 0
                    },
                    "type": "SequenceInterval"
                },
                "sequence_id": "refseq:NC_000011.10",
                "type": "SequenceLocation"
            },
            "variantAlternativeIds": []
        }
    },
    {
        "caseLevelData": [
            {
                "analysisId": "pgxcs-kftvu6cg",
                "biosampleId": "pgxbs-kftvh94d",
                "id": "pgxvar-5bab5837727983b2e0121e99"
            }
        ],
        "variantInternalId": "1:0-84699999:DEL",
        "variation": {
            "copyChange": "efo:0030067",
            "identifiers": {},
            "subject": {
                "interval": {
```


	variant_id	biosample_id	analysis_id	reference_genome	variant
1	pgxvar-5bab5837727983b2e0121e99	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000001.11	1:0-84699999:DEL
2	pgxvar-5bab5837727983b2e0121e9a	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000001.11	1:124300000-247249719:DEL
3	pgxvar-5bab5837727983b2e0121e9c	pgxbs-kftvh94d	pgxcs-kftvu6cg	refseq:NC_000002.12	2:12800000-61099999:DEL
4	pgxvar-5bab5837727983b2e01				1149:DEL
5	pgxvar-5bab5837727983b2e01				99:DEL
6	pgxvar-5bab5837727983b2e01				063:DUP
7	pgxvar-5bab5837727983b2e01				999:DUP
8	pgxvar-5bab5837727983b2e01				
9	pgxvar-5bab5837727983b2e01				9992:DEL



pgxRpi

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

pgxRpi Public

Edit Pins | Watch 2

main | 1 branch | 0 tags | Go to file | Add file | Code

hangjiaz version bump 319d27c 4 days ago 87 commits

R	adapt to beacon variant export	4 days ago
data-raw	add data documentation; optimise get method which causes error in ...	3 months ago
data	change based on opinion from bioc reviewer	last month
inst	add data documentation; optimise get method which causes error in ...	3 months ago
man	documentation for parameters using match.arg	last month
tests		
vignettes		
.Rbuildignore		
.gitignore		
DESCRIPTION		
NAMESPACE		
NEWS.md		
README.md		

Bioinformatics Advances, 2025, **00**, vbaf172
https://doi.org/10.1093/bioadv/vbaf172
Advance Access Publication Date: 16 July 2025
Application Note

OXFORD

Data and text mining

pgxRpi: an R/bioconductor package for user-friendly access to the Beacon v2 API

Hangjia Zhao[1,2,*] and Michael Baudis[1,2,*]

[1]Department of Molecular Life Sciences, University of Zurich, Zurich 8057, Switzerland
[2]Computational Oncogenomics Group, Swiss Institute of Bioinformatics, Zurich 8057, Switzerland

*Corresponding authors. Hangjia Zhao, Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland. E-mail: hangjia.zhao@uzh.ch; Michael Baudis, Department of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, Zurich 8057, Switzerland. E-mail: michael.baudis@mls.uzh.ch.

2 Retrieve meatdata of samples

2.1 Relevant parameters

type, filters, filterLogic, individual_id, biosample_id, codematches, limit, skip

2.2 Search by filters

Filters are a significant enhancement to the Beacon query API, providing a mechanism for specifying rules to select records based on their field values. To learn more about how to utilize filters in Progenetix, please refer to the documentation.

The `pgxFilter` function helps access available filters used in Progenetix. Here is the example use:

```
# access all filters
all_filters <- pgxFilter()
# get all prefix
all_prefix <- pgxFilter(return_all_prefix = TRUE)
# access specific filters based on prefix
ncit_filters <- pgxFilter(prefix="NCIT")
head(ncit_filters)
#> [1] "NCIT:C28076" "NCIT:C18000" "NCIT:C14158" "NCIT:C14161" "NCIT:C28077"
#> [6] "NCIT:C28078"
```

The following query is designed to retrieve metadata in Progenetix related to all samples of lung adenocarcinoma, utilizing a specific type of filter based on an NCIt code as an ontology identifier.

```
biosamples <- pgxLoader(type="biosample", filters = "NCIT:C3512")
# data looks like this
biosamples[c(1700:1705),]
#>       biosample_id group_id group_label   individual_id    callset_ids
#> 1700 pgxbs-kftvjjhx       NA          NA pgxind-kftx5fyd pgxcs-kftwjevi
#> 1701 pgxbs-kftvjjhz       NA          NA pgxind-kftx5fyf pgxcs-kftwjew0
#> 1702 pgxbs-kftvjji1       NA          NA pgxind-kftx5fyh pgxcs-kftwjewi
#> 1703 pgxbs-kftvjjn2       NA          NA pgxind-kftx5g4r pgxcs-kftwjg5r
#> 1704 pgxbs-kftvjjn4       NA          NA pgxind-kftx5g4t pgxcs-kftwjg6q
#> 1705 pgxbs-kftvjjn5       NA          NA pgxind-kftx5g4v pgxcs-kftwjg78
```

Components of an Online Bioinformatics Resource

Going Full Stack?

Components of an Online Bioinformatics Resource

A Stack to work with/through

- **dedicated server** or **cloud storage**
- own **domain** | institutional sub-domain or fixed address | cloud service sub-domain
  - ➡ `progenetix.org | mls.uzh.ch/en/research/baudis | baudisgroup.github.io`
- **database** or flat file data management
  - ➡ SQL databases such as PostGres, MySQL
  - ➡ document databases such as MongoDB, CouchDB ...
  - ➡ hierarchical file system & index files
- webserver gateway for server-side generated, active content delivery
  - ➡ Perl CGI, Python, PHP ...
- active **front-end** (JavaScript… environment)

bycon based Beacon+ Stack

progenetix

- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
  - `pubmed:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703…`
  - precomputed frequencies per collection informative e.g. in form autofills
- *querybuffer* stores id values of all entities matched by a query and provides the corresponding `accessid` for **handover** generation
- complete query aggregation; i.e. individual queries are run against the corresponding entities and ids are intersected
  - ➡ retrieval of any entity, e.g. all individuals which have queried variants analyzed on a given platform
  - ➡ allows multi-variant queries, i.e. all bio samples or individuals which had matches of all of the individual variant queries

React

APACHE HTTP SERVER PROJECT

python™

mongoDB

Entity collections

Utility collections

github.com/progenetix/bycon GitHub

Last but NOT Least…

Documentation is, actually, rather important

Documentation Strategies

(Not so) Best Practices

```
f_d = f_d_s[c_t]
r = {}
for k in res_schema.keys():
    if k in f_d:
        r.update({k:f_d[k]})
```

- What is documentation? I'll remember this! ¯_(ツ)_/¯
- Just email me if help is needed, unexpectedly

mbaudis@netscape.net

- We had money for a chat bot.

```
normalize_variant_values_for_export(v, byc, drop_fields=None):
```

- Clean code documents itself - Just use explicit variable/function names.
- Clean code documents itself - Never use explicit variable/function names.
- Perl POD it is. There is a command to show the notes in your terminal…
- I wrote a paper about the resource. In 2001.
- Haven't you found the GoogleGroups account?
- Documentation? StackOverflow, whelp!

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

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

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

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

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

Documentation Strategies

(Not so) Best Practices

```
f_d = f_d_s[c_t]
r = {}
for k in res_schema.keys():
    if k in f_d:
        r.update({k:f_d[k]})
```

- What is documentation? I'll remember this! ヽ(ˇ▽ˇ)ノ
- Just email me if help is needed, unexpectedly

mbaudis@netscape.net

- We had money for a chat bot.

```
normalize_variant_values_for_export(v, byc, drop_fields=None):
```

- Clean code documents itself - Just use explicit variable/function names.
- Clean code documents itself - Never use explicit variable/function names.
- Perl POD it is. There is a command to show the notes in your terminal…
- I wrote a paper about the resource. In 2001.
- Haven't you found the GoogleGroups account?
- Documentation? StackOverflow, whelp!

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

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

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

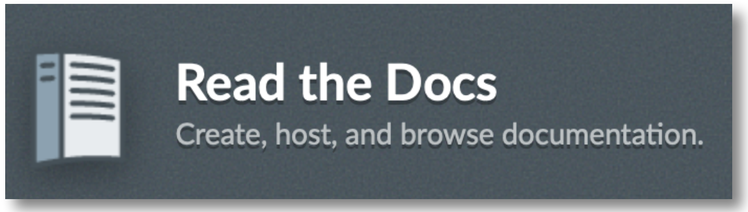
[1]*Medizinische Klinik und Poliklinik V der Universität Heidelberg, Germany and* [2]*Department of Pathology, Stanford University Medical Center, Stanford, CA 94305, USA*

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

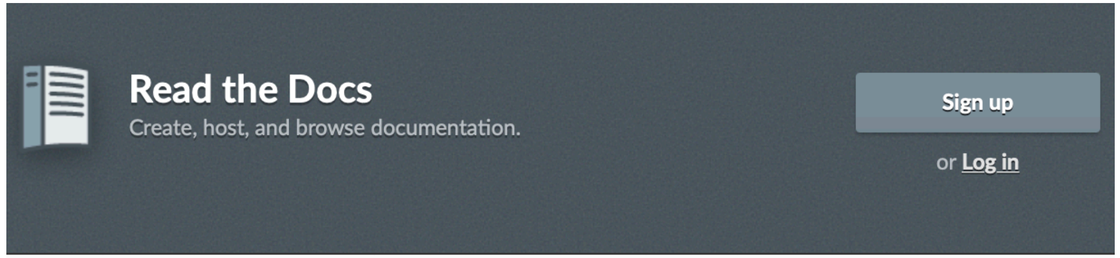
Documentation Strategies

Currently en Vogue

- Cloud-based documentation systems with online compilation
- written in simplified markup languages
  - ➡ **Markdown** (Yeah!)
  - ➡ Restructured Text (Meeh...)
- self hosted ¯_(ツ)_/¯
  - ➡ local and/or service based compilation and hosting
- build systems & output hosting
  - ➡ ReadTheDocs
    - direct building from .rst document tree or MkDocs based
  - ➡ Github Pages
    - direct using Jekyll or over MkDocs through GH actions



Documentation Strategies



Technical documentation lives here

Read the Docs simplifies software documentation by automating building, versioning, and hosting of your docs for you.

Free docs hosting for open source

We will host your documentation for free, forever. There are no tricks. We help over 100,000 open source projects share their docs, including a custom domain and theme.

Always up to date

Whenever you push code to your favorite version control service, whether that is GitHub, BitBucket, or GitLab, we will automatically build your docs so your code and documentation are never out of sync.

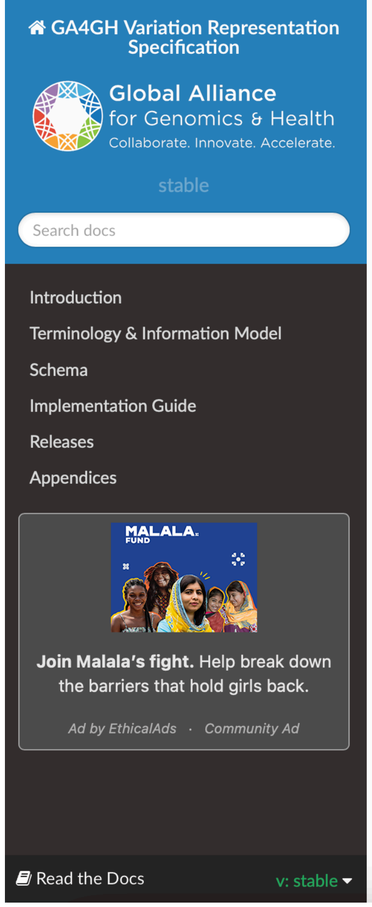
Downloadable formats

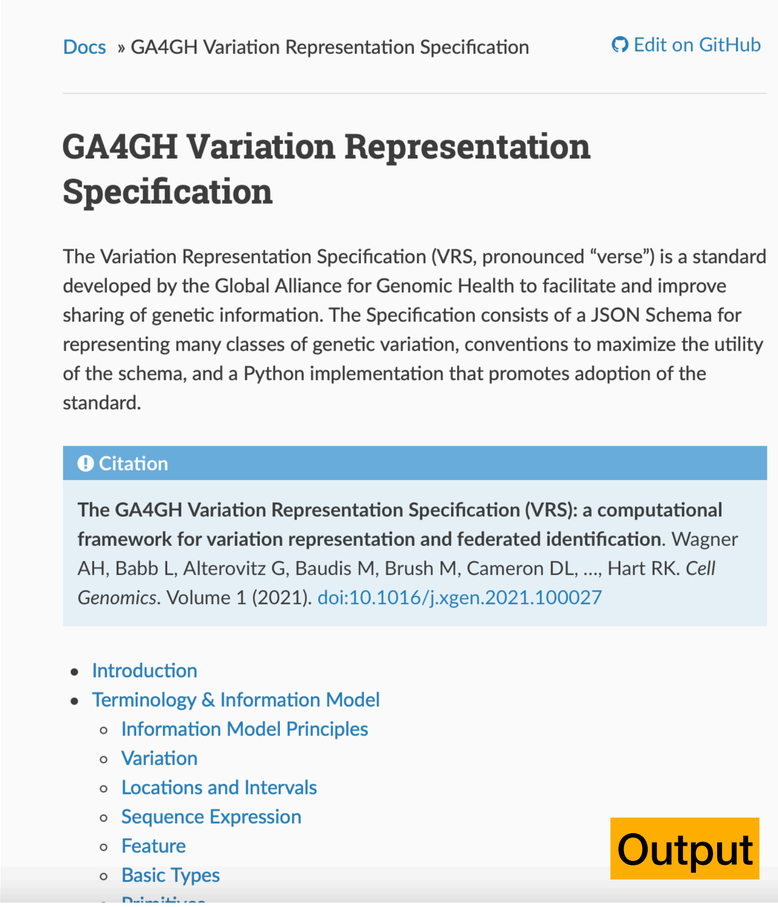
We build and host your docs for the web, but they are also viewable as PDFs, as single page HTML, and for eReaders. No additional configuration is required.

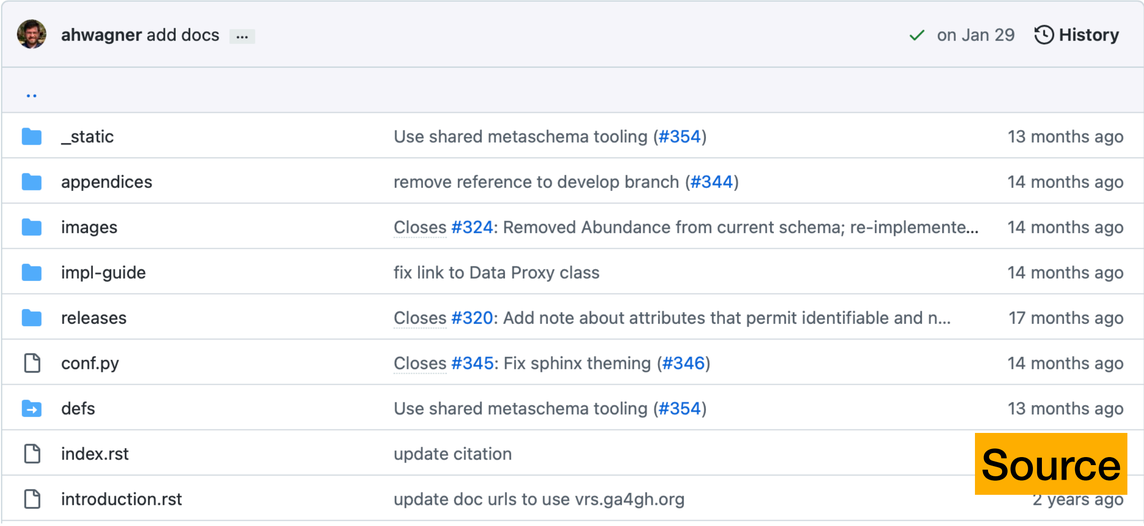
Multiple versions

We can host and build multiple versions of your docs so having a 1.0 version of your docs and a 2.0 version of your docs is as easy as having a separate branch or tag in your version control system.

Example: GA4GH Variation Representation Standard ->





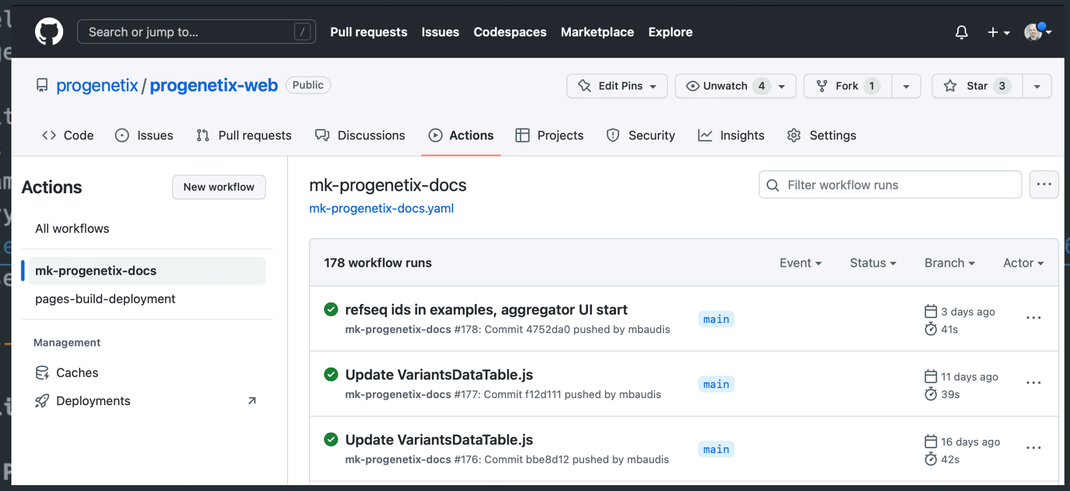


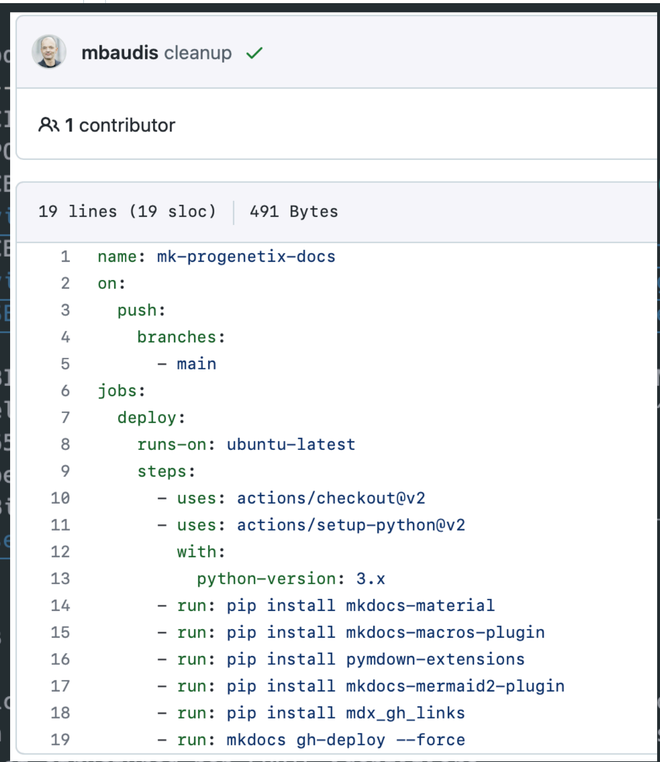

FOLDERS
progenetix-web
.github
.next
docs
css
img
javascripts
news
beaconplus.md
changelog.md
classifications-an
CNAME
index.md
progenetix-data-r
progenetix-websi
publication-collec
services.md
technical-notes.m
ui.md
use-cases.md
extra
node_modules
out
public
src
.babelrc
.env.development
.env.production
.eslintrc.json
.gitignore
.prettierrc
jest.config.js
mkdocs.yaml
next.config.js
package-lock.json
package.json
README.md
mkdocs.yaml
site_name: Progenetix Documentation
site_description: 'Documentation for the Progenetix oncogen
site_author: Michael Baudis
copyright: '© Copyright 2022, Michael Baudis and proge
repo_name: 'progenetix-web'
repo_url: https://github.com/progenetix/progenetix-web
nav:
- Documentation Home: index.md
- News & Changes: news
- Publication Collection: publication-collection
- Data Review: progenetix-data-review
- Technical Notes: technical-notes
- Progenetix Website Builds: progenetix-website-builds
- Progenetix Data ↗: http://progenetix.org
- Baudisgroup @ UZH ↗: http://info.baudisgroup.org
markdown_extensions:
- toc:
toc_depth: 2-3
permalink: true
- admonition
- attr_list
- footnotes
- md_in_html
- pymdownx.critic
- pymdownx.caret
- pymdownx.details
- pymdownx.keys
- pymdownx.magiclink:
hide_protocol: true
- pymdownx.mark
- pymdownx.tilde
- pymdownx.saneheaders
classifications-and-ontologies.md
Classifications, Ontologies and Standards
The Progenetix resource utilizes standardized diagnostic coding systems, with a move towards hierarchical ontologies. As part of the coding process we have developed and provide several code mapping resources through repositories, the Progenetix website and APIs.
Additionally to diagnostic and other clinical concepts, Progenetix increasingly uses hierarchical terms and concepts for the annotation and querying of technical parameters such as platform technologies. Overall, the Progenetix resource uses a query syntax based around the [Beacon v2 "filters"](https://beacon-project.io/v2/filters.html) concept with a [CURIE](https://www.w3.org/TR/2010/NOTE-curie-20101216/) -based syntax
MkDocs & Material for MkDocs & Github Actions
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](http://progenetix.org/services/ids/PMID:18810378)
geo	NCBI Gene Expression Omnibus[^3]	[geo:GPL6801](http://progenetix.org/services/ids/geo:GPL6801), [geo:GSE19399](http://progenetix.org/services/ids/geo:GSE19399), [geo:GSM491153](http://progenetix.org/services/ids/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](http://progenetix.org/services/ids/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.

Local Testing

```
→  progenetix-web git:(main) mkdocs serve
INFO     -  Building documentation...
INFO     -  [macros] - Macros arguments: {'module_name': 'main',
           'modules': [], 'include_dir': '', 'include_yaml': [],
           'j2_block_start_string': '', 'j2_block_end_string': '',
           'j2_variable_start_string': '', 'j2_variable_end_string': '',
           'on_undefined': 'keep', 'on_error_fail': False, 'verbose': False}
INFO     -  [macros] - Extra variables (config file):
           ['excerpt_separator', 'blog_list_length', 'social']
INFO     -  [macros] - Extra filters (module): ['pretty']
INFO     -  MERMAID2  - Initialization arguments: {}
INFO     -  MERMAID2  - Using javascript library (8.8.0):
              https://unpkg.com/mermaid@8.8.0/dist/mermaid.min.js
INFO     -  Cleaning site directory
INFO     -  The following pages exist in the docs directory, but are not
           included in the "nav" configuration:
             - beaconplus.md
             - changelog.md
             - classifications-and-ontologies.md
             - progenetix-data-review.md
             - progenetix-website-builds.md
             - publication-collection.md
INFO     -  MERMAID2  - Found superfences config: {'custom_fences': [{'name':
           'mermaid', 'class': 'mermaid', 'format': <function fence_mermaid at
           0x104075ab0>}]}
INFO     -  MERMAID2  - Page 'Technical Notes': found 2 diagrams, adding scripts
INFO     -  Documentation built in 0.83 seconds
INFO     -  [09:05:32] Watching paths for changes: 'docs', 'mkdocs.yaml'
INFO     -  [09:05:32] Serving on http://127.0.0.1:8000/
INFO     -  [09:05:33] Browser connected:
           http://127.0.0.1:8000/classifications-and-ontologies/
```

Web Deployment (Github)





Classifications, Ontologies and Standards

The Progenetix resource utilizes standardized diagnostic coding systems, with a move towards hierarchical ontologies. As part of the coding process we have developed and provide several code mapping resources through repositories, the Progenetix website and APIs.

Additionally to diagnostic and other clinical concepts, Progenetix increasingly uses hierarchical terms and concepts for the annotation and querying of technical parameters such as platform technologies. Overall, the Progenetix resource uses a query syntax based around the Beacon v2 "filters" concept with a CURIE based syntax.

List of filters recognized by different query endpoints

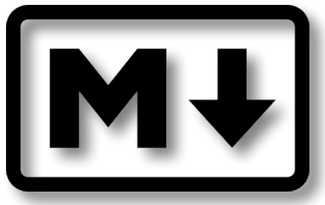
Public Ontologies with CURIE-based syntax

CURIE prefix	Code/Ontology	Examples
NCIT	NCIt Neoplasm[1]	NCIT:C27676

Documentation Strategies

Best Practices

- start early
- update often
- sometimes try to follow your own guide
- balance between inline documentation & doc system
- use Markdown
- plan for contingencies
  - ➡ cloud providers disappear | cancel services | change terms



https://en.wikipedia.org/wiki/List_of_defunct_social_networking_services

https://en.wikipedia.org/wiki/List_of_search_engines#Defunct_or_acquired_search_engines

Progenetix as Example Genomics Resource

Some trajectories ...

- local database => **online resource**
- flat database => **hierarchical object storage**
- dedicated database => mix of **open software tools**
- static pages => **data driven website**
- copy, paste, clean => **automated download & process** (still edit & clean)
- registered access & commercial licensing => **CC BY 4.0** (CC0 for tools)
- local development => **open source code** on Github (future - Codeberg?)
- standalone resource => federated data, **APIs** and services

progenetix

(Bio)informatics Skill Set

What has been needed to develop & maintain progenetix.org?

- Scripting and application development using Python, Perl and JavaScript
- Data analysis and plotting in R, Python and Perl
- Regular expressions for data entry an (programmatic) identifier matching
- JSON, YAML, tab-delimited text as file formats; some binary source files (.CEL)
- non-SQL database (MongoDB) for flexibility and document structure
- web development with Perl, Python, JS, React and Apache server; Cloudflare
- No proprietary software involved

progenetix

(Bio)informatics Skill Set

What has been needed to develop & maintain progenetix.org?

GitHub

text mining

Perl

regular expressions

```
s/knowledge/mastery/
```

React

mongoDB

M↓

APACHE HTTP SERVER PROJECT

R

MkDocs

Project documentation with Markdown.

python™

HTML CSS JS

array & sequencing pipelines

Master/Bachelor Project in Data Wrangling? Ask!