progenetix

Progenetix & Beacon+

A cancer genomics reference resource powered by GA4GH standards

Michael Baudis | Roche Data Science Seminar | 2022-06-27

Global Alliance
for Genomics & Health

1992	Heidelberg	Student of medicine | doctoral thesis in molecular cytogenetics @ DKFZ (Peter Lichter) | resident in clinical hematology/oncology | data, clinical studies & cancer systematics
2001	Stanford	Post-doc in hemato-pathology (Michael Cleary) | molecular mechanisms of leukemogenesis | transgenic models | expression arrays | systematic cancer genome data collection | *Progenetix* website
2003	Gainesville	Assistant professor in paediatric haematology | molecular mechanisms of leukemogenesis | focus on bioinformatics for cancer genome data analysis
2006	Aachen	Research group leader in genetics | genomic array analysis for germline alterations | descriptive analysis of copy number aberration patterns in cancer entities
2007	Zürich	Professor of bioinformatics @ DMLS (2015) | systematic assembly of oncogenomic data | databases and software tools | patterns in cancer genomes | *Progenetix* & *arrayMap* resources | GA4GH | SPHN | ELIXIR

Genome screening at the core of “Personalised Health”

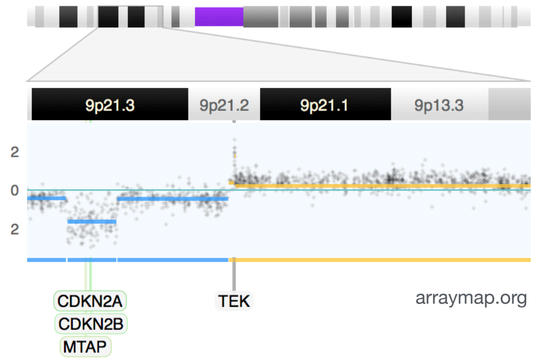
- **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**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

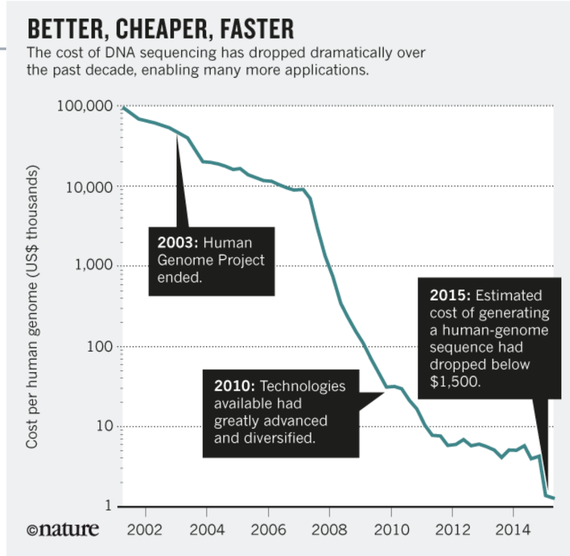
arrayMap

progenetix

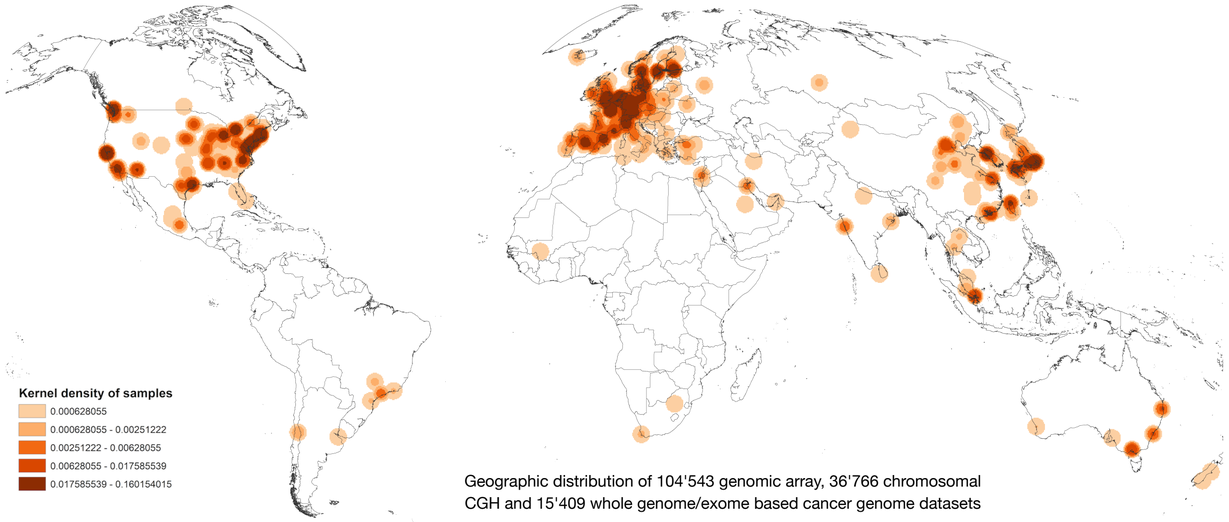
elixir

Global Alliance for Genomics & Health





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



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

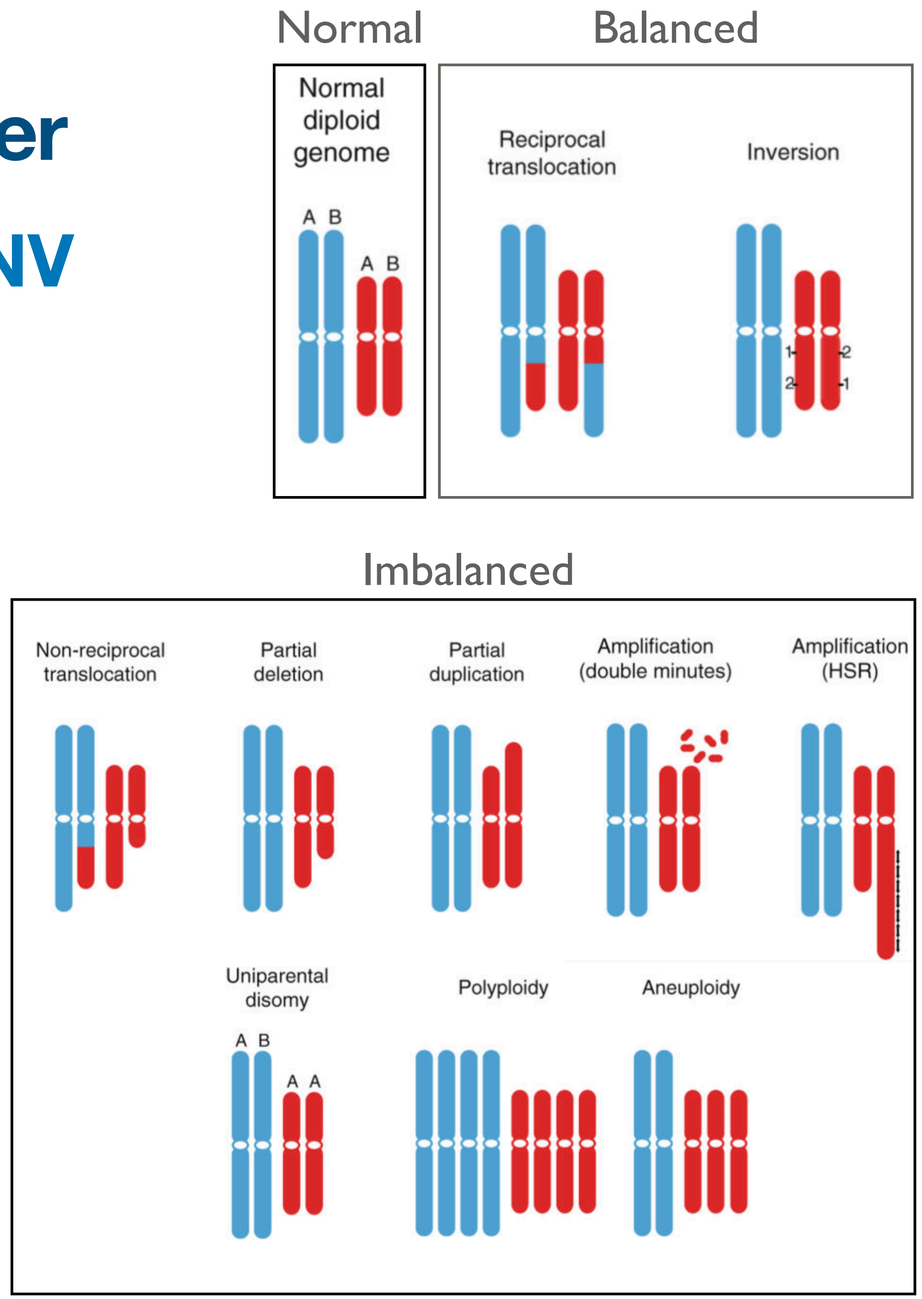
Types of genomic alterations in Cancer

Imbalanced Chromosomal Changes: CNV

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

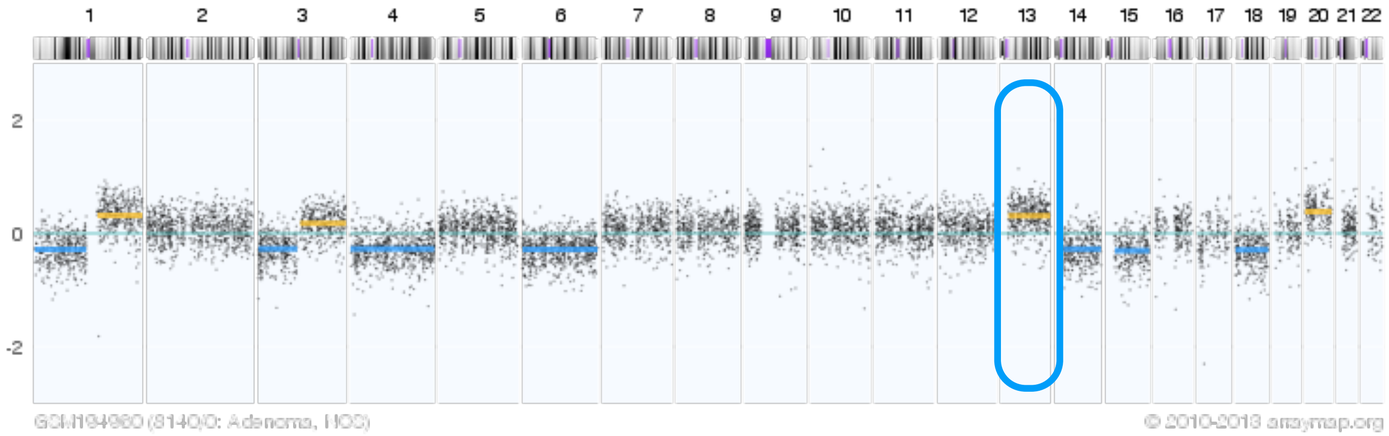
➡ **Regional Copy Number Alterations** (losses, gains)

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

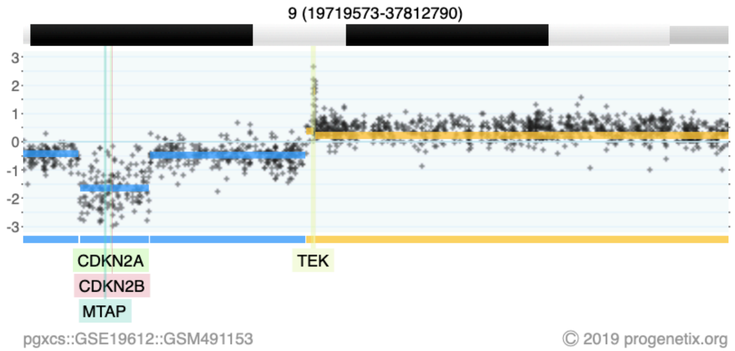


Grade et al., 2015 Recent Results Cancer Res

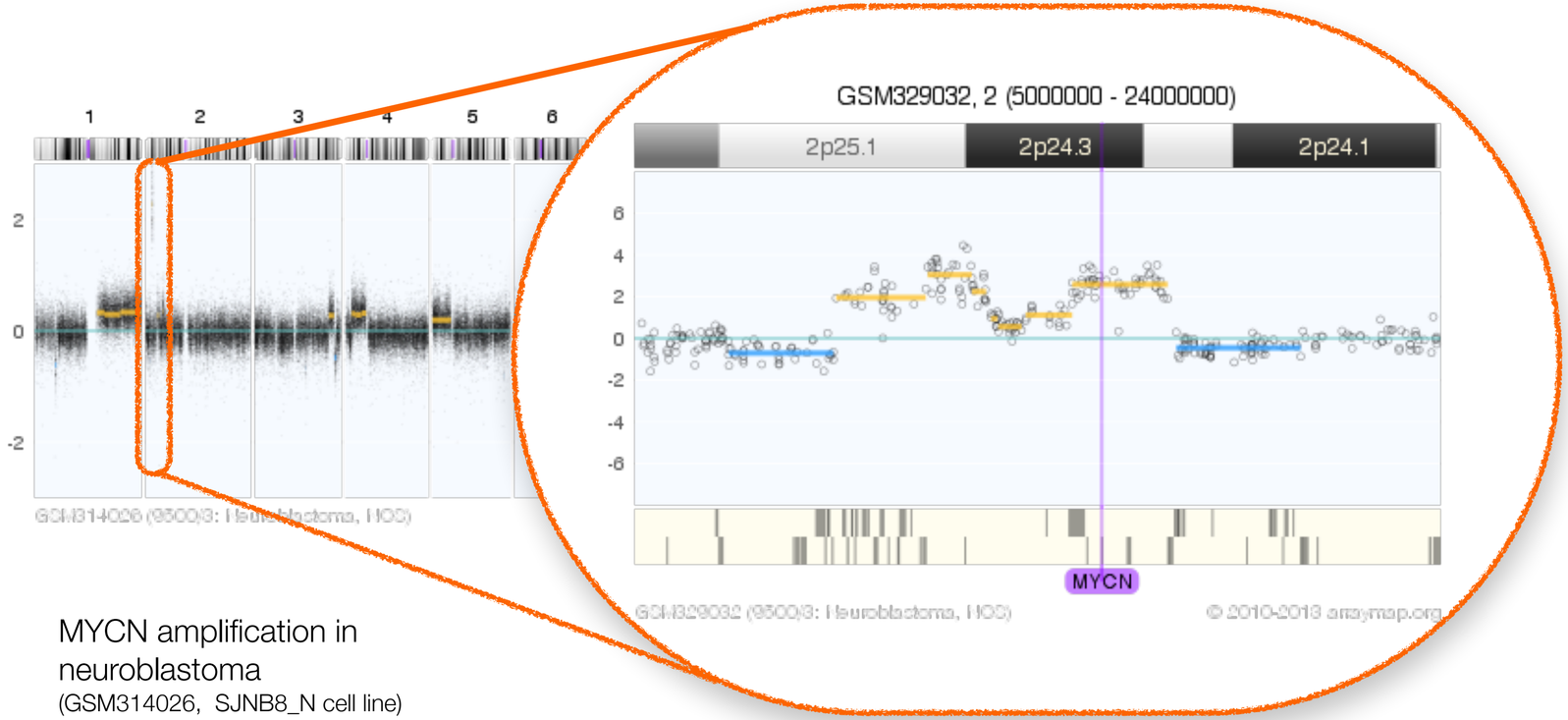
Somatic Copy Number Variations



Gain of chromosome arm 13q in colorectal carcinoma



2-event, homozygous deletion in a Glioblastoma

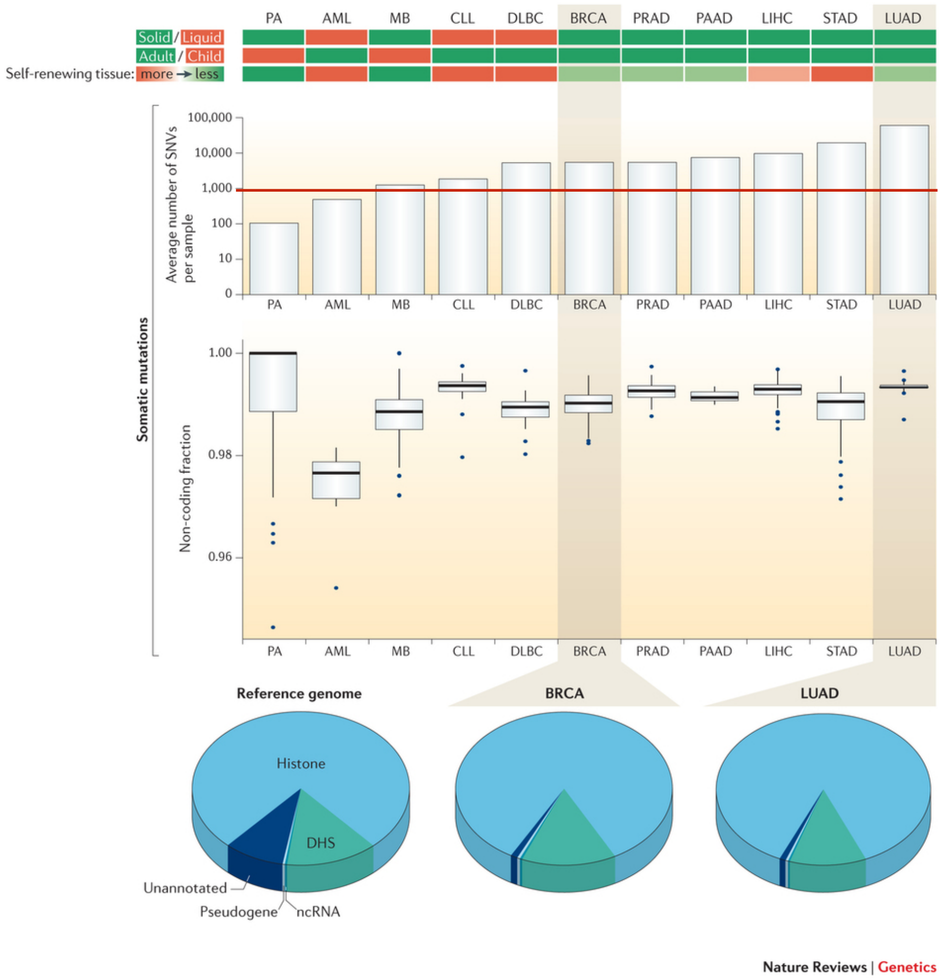


MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

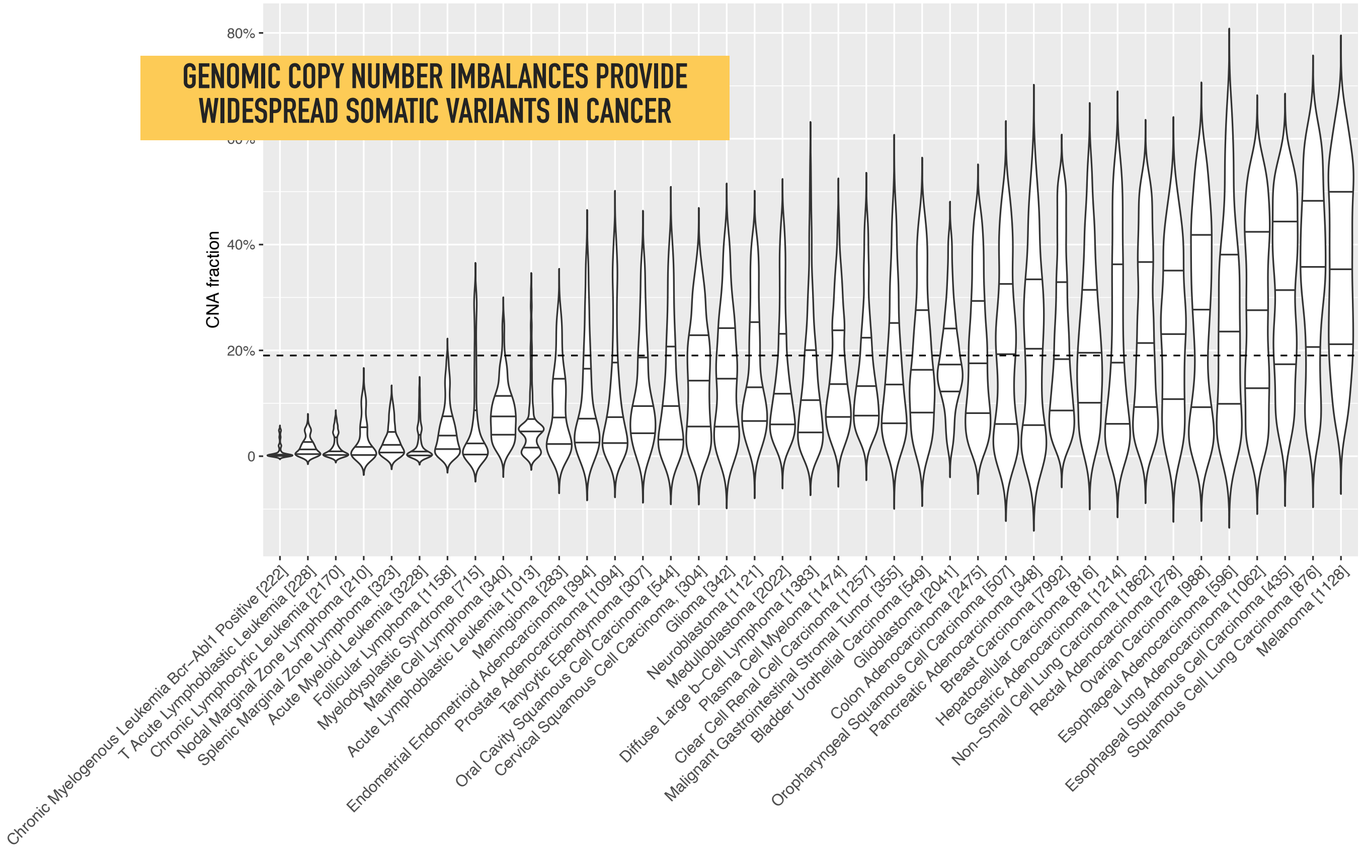
low level/**high level** copy number alterations (CNAs)

arrayMap

Quantifying Somatic Mutations In Cancer



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



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

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

Somatic Mutations In Cancer: Patterns

Making the case for genomic classifications

Some related cancer entities show similar copy number profiles

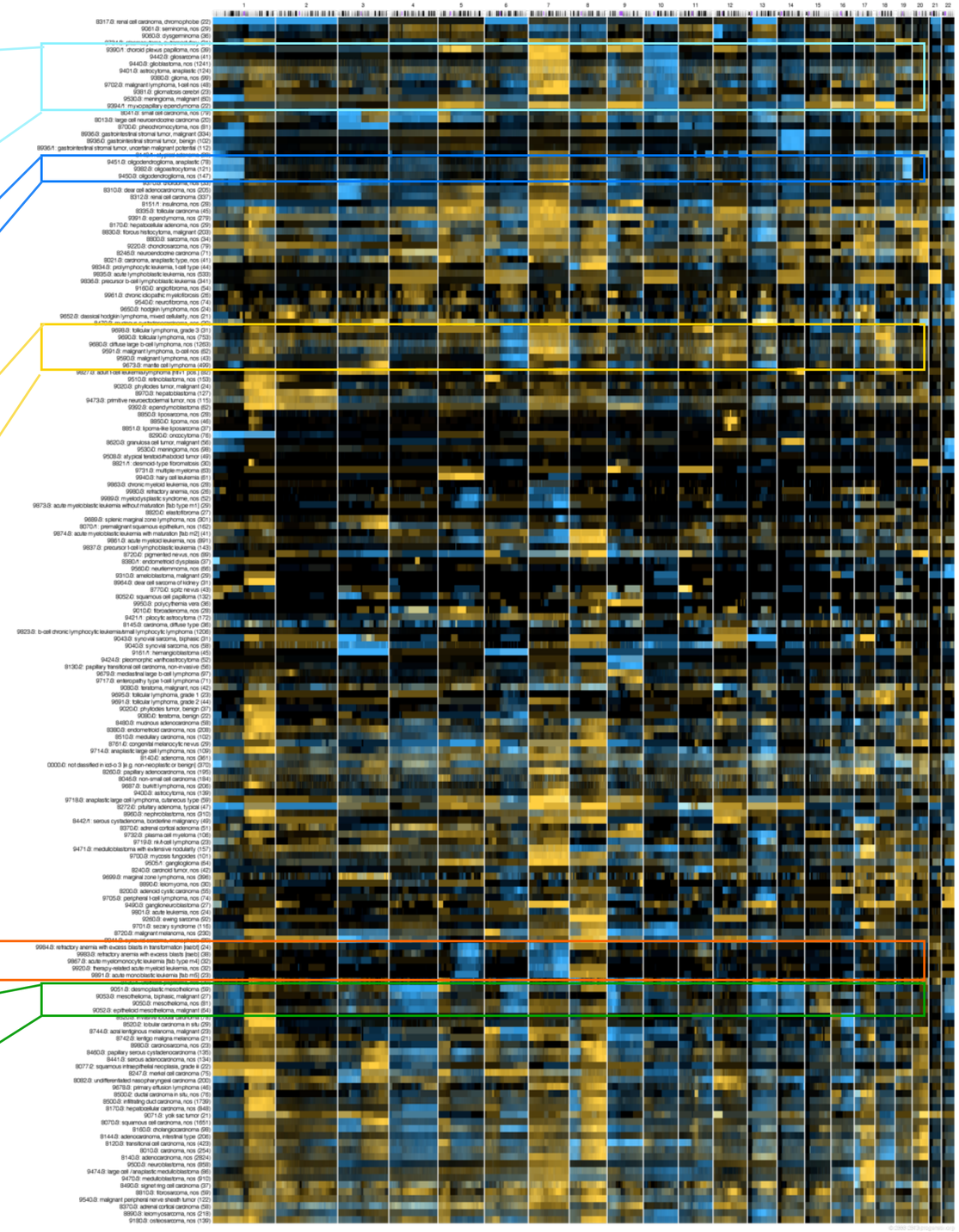
9390/1: choroid plexus papilloma, nos (39)
9442/3: gliosarcoma (41)
9440/3: glioblastoma, nos (1241)
9401/3: astrocytoma, anaplastic (124)
9380/3: glioma, nos (99)
9702/3: malignant lymphoma, t-cell nos (48)
9381/3: gliomatosis cerebri (23)
9530/3: meningioma, malignant (60)
9394/1: myxopapillary ependymoma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

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

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

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

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

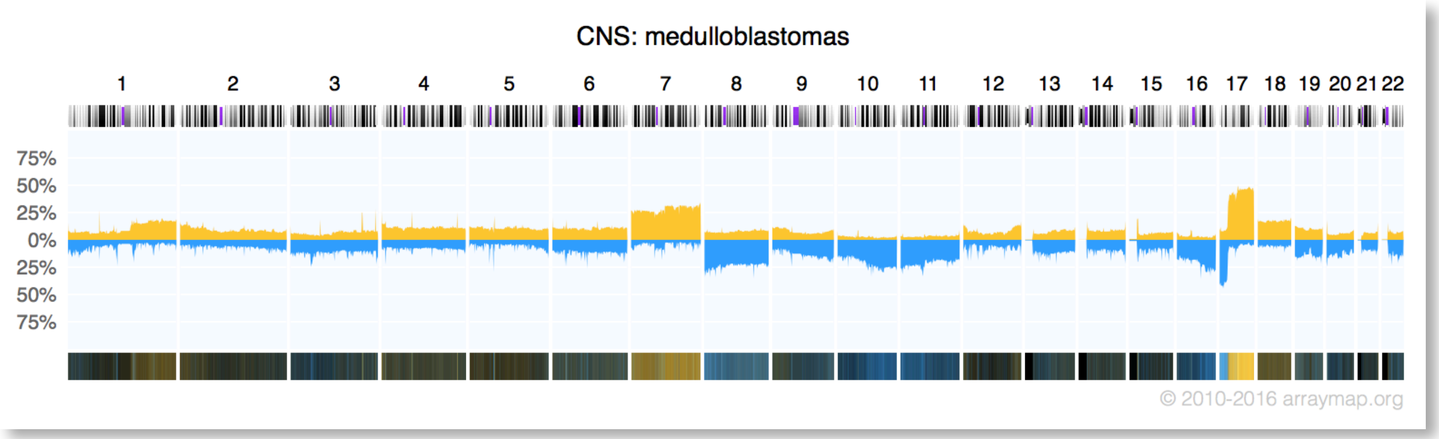


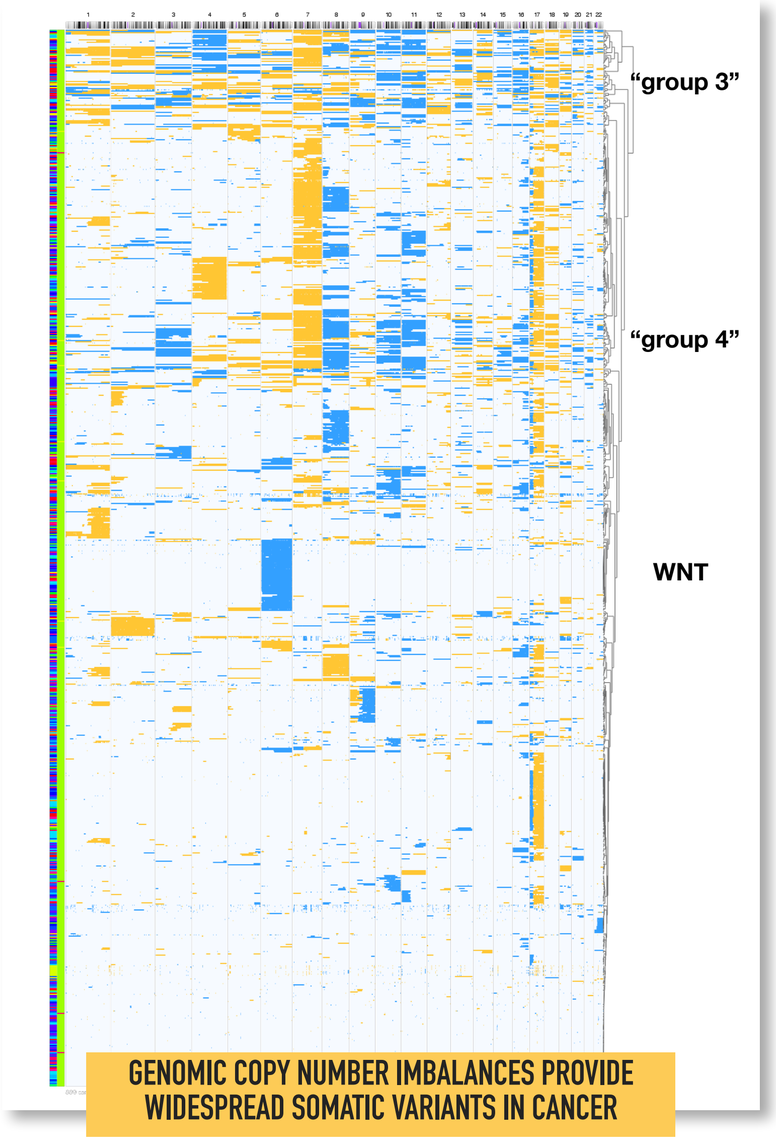
arrayMap

Somatic CNVs In Cancer

Recurrent mutation patterns

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





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

arrayMap

Progenetix Genomics Resource

From Genomic Experiments to Experimenting with the Beacon API

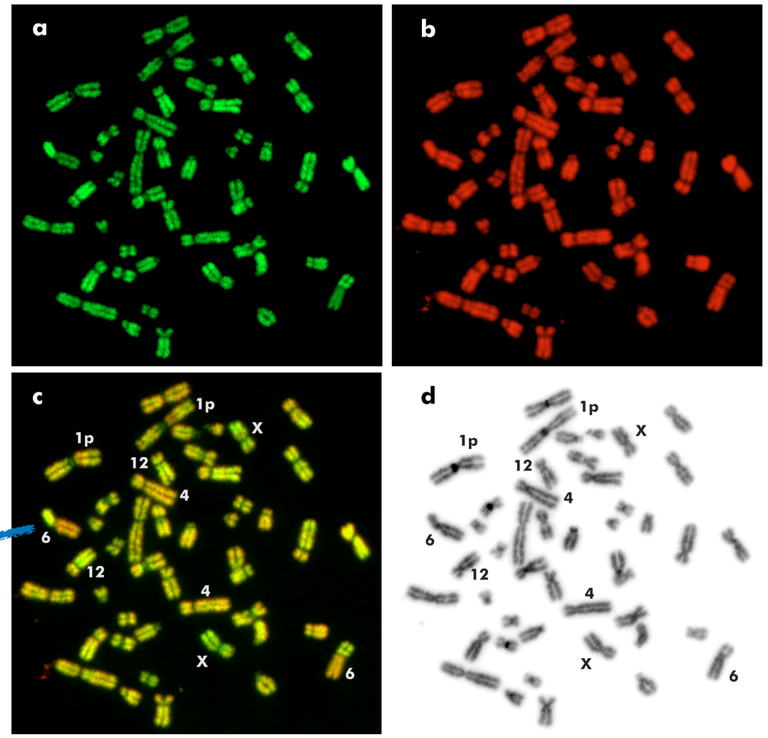
progenetix

Comparative Genomic Hybridization

Molecular-Cytogenetic Technology for Genomic Imbalance Screening

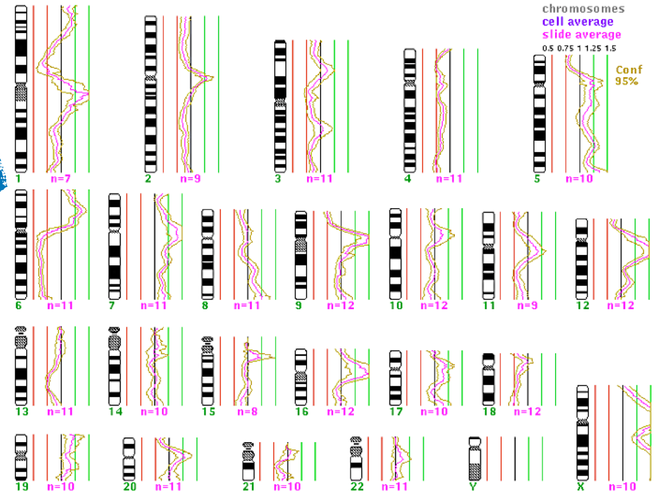
- Molecular-cytogenetic technique to identify regional genomic copy number variations (CNV/CNA)
- based on ***in situ*** suppression **hybridization** of labeled **genomic** tumor and reference DNA against a karyotypically normal metaphase chromosomes
- analysis of relative fluorescence ratio allows **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.



+6p, -6q

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



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

| Tuesday February 27, 1996

Let's build a database!

... using archaic tools

| Tuesday September 08, 1992

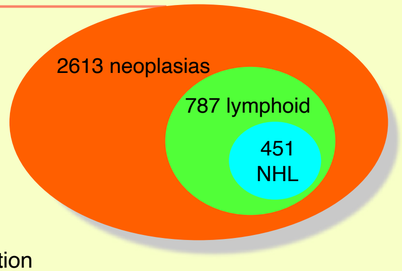
progenetix.net

[progenetix.net] online CGH database

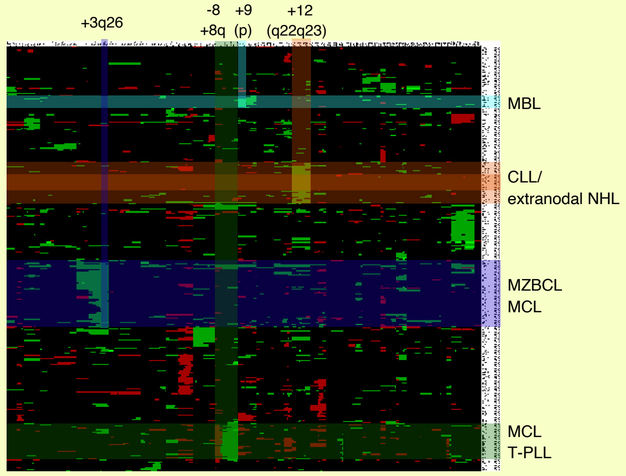
A compilation of published CGH analyses with reported case and band specific results

Currently includes
2613 cases from
92 publications

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



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



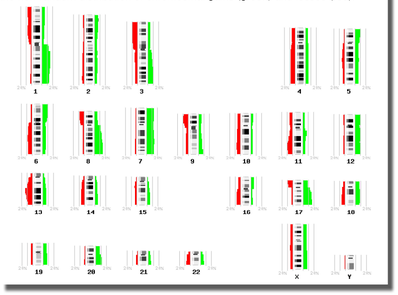
Collection and Transformation of Chromosomal Imbalances in Human Neoplasias for Data Mining Procedures

michael baudis, dept. of pathology, stanford university

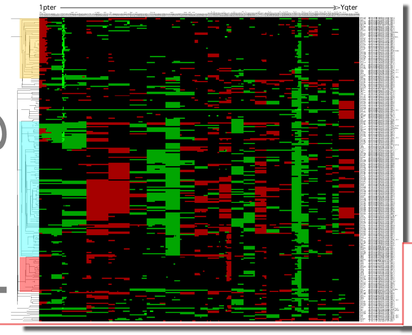
progenetix.net

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

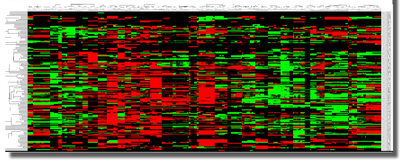
Chromosomal imbalances in 5478 clinical cases from 196 publications
Although not as prominent as in specific subgroups, this large collection shows the non-random distribution of chromosmal gains (green) and losses (red).



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

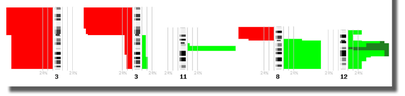


Clustering of the band averages for the different ICD-O entities
Two dimensional clustering groups related disease entities and chromosomal bands with related aberrations.



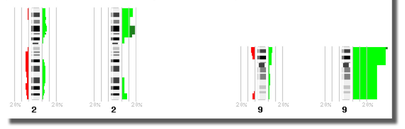
Results Out of 4896 tumor samples, 3862 (79%) showed chromosomal imbalances by CGH. The average per band probability was 4.5% for a loss (max. 12.9% at 13q21) and 6.5% for a gain (max. 15.6% at 8q23). Differences between neoplastic entities showed in the average frequency and distribution pattern of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

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



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

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



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

Progenetix Database in 2003
Text conversion for CNVs

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

progenetix

TABLE 3. Comparison of Primary Tumors and Metastases by CGH

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

TABLE 1. Clinical Data

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

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

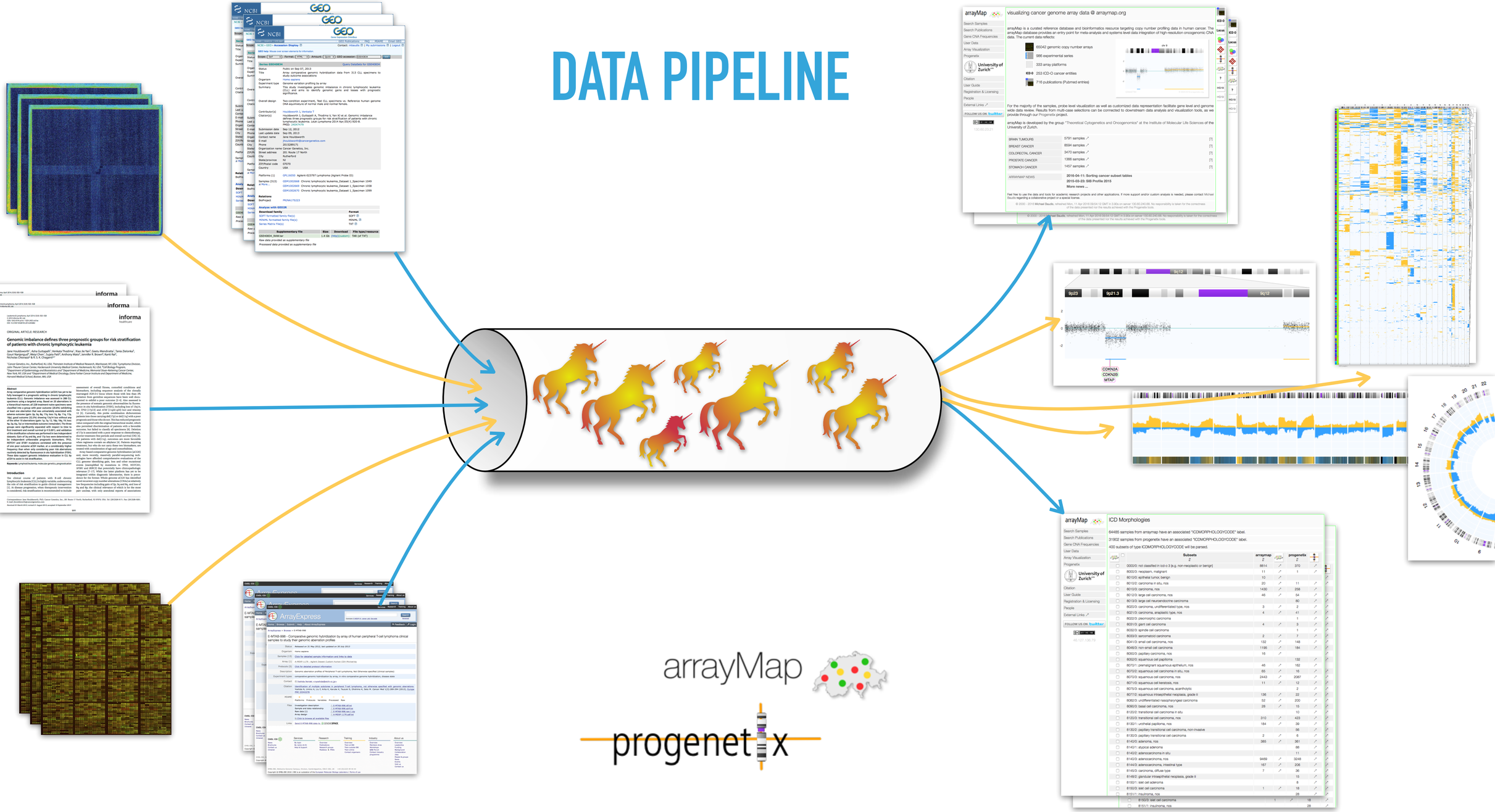
GENES, CHROMOSOMES & CANCER 25:82–90 (1999)

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

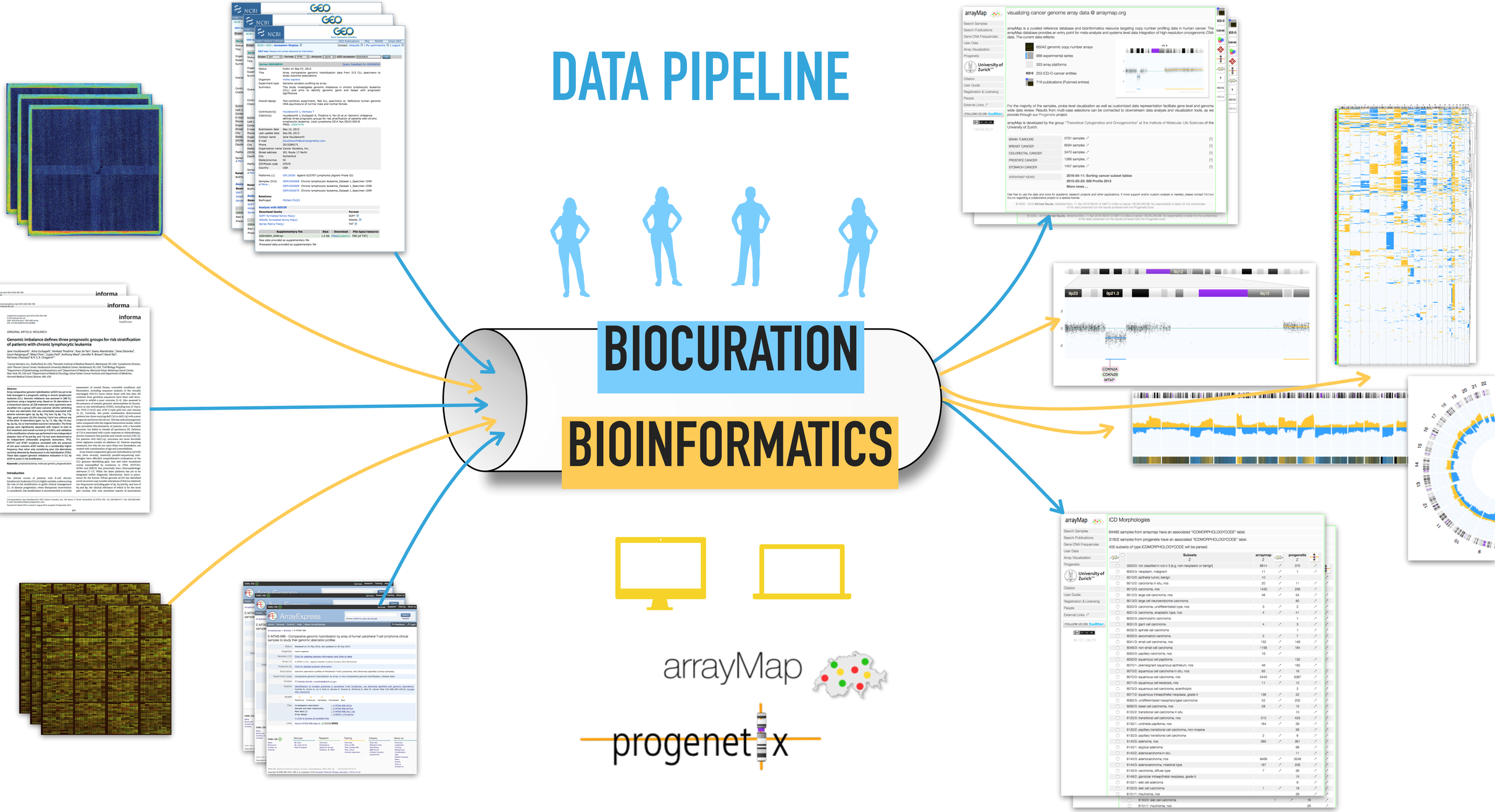
W. Michael Korn,[1]* Toru Yasutake,[2] Wen-Lin Kuo,[1] Robert S. Warren,[2] Colin Collins,[1] Masao Tomita,[2] Joe Gray,[1] and Frederic M. Waldman[1]

Data "Pipelines"

DATA PIPELINE



DATA PIPELINE



Data Curation

Happy RegExing!

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

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


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

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

Disease annotations in Progenetix

From some text, somewhere, to ontology classes

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

NCIthesaurus

DX Ontologies

Hierarchical NCIt Neoplasm Core replaces heterogeneous primary annotations

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

progenetix

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

Collapse all | Expand | All

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

Progenetix in 2022

Cancer Genomics Reference Resource

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

progenetix

progenetix

Cancer CNV Profiles
- ICD-O Morphologies
- ICD-O Organ Sites
- Cancer Cell Lines
- Clinical Categories

Search Samples

arrayMap
- TCGA Samples
- 1000 Genomes Reference Samples
- DIPG Samples
- cBioPortal Studies
- Gao & Baudis, 2021

Publication DB
- Genome Profiling
- Progenetix Use

Services
- NCIt Mappings
- UBERON Mappings

Upload & Plot

Beacon+

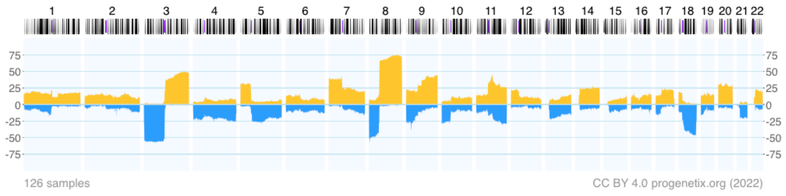
Documentation
- News
- Downloads & Use Cases
- Sevices & API

Baudisgroup @ UZH

Cancer genome data @ progenetix.org

The Progenetix database provides an overview of mutation data in cancer, with a focus on copy number abnormalities (CNV / CNA), for all types of human malignancies. The data is based on *individual sample data* from currently **142063** samples.

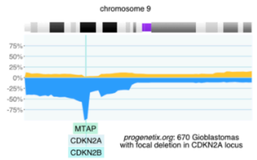
Floor of the Mouth Neoplasm (NCIT:C4401)



Download SVG | Go to NCIT:C4401 | Download CNV Frequencies

Example for aggregated CNV data in 126 samples in Floor of the Mouth Neoplasm.
Here the frequency of regional **copy number gains** and **losses** are displayed for all 22 autosomes.

Progenetix Use Cases



Local CNV Frequencies

A typical use case on Progenetix is the search for local copy number aberrations - e.g. involving a gene - and the exploration of cancer types with these CNVs. The [Search Page] provides example use cases for designing queries. Results contain basic statistics as well as visualization and download options.

Cancer CNV Profiles

The progenetix resource contains data of **834** different cancer types (NCIt neoplasm classification), mapped to a variety of biological and technical categories. Frequency profiles of regional genomic gains and losses for all categories (diagnostic entity, publication, cohort ...) can be accessed through the [Cancer Types] page with direct visualization and options for sample retrieval and plotting options.

Cancer Genomics Publications

Through the [Publications] page Progenetix provides **4164** annotated references to research articles from cancer genome screening experiments (WGS, WES, aCGH, cCGH). The numbers of analyzed samples and possible availability in the Progenetix sample collection are indicated.

Progenetix in 2022

Cancer Genomics Reference Resource

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

progenetix

progenetix

Cancer CNV Profiles
Search Samples
Studies & Cohorts
arrayMap
TCGA Samples
DIPG Samples
Gao & Baudis, 2021
Cancer Cell Lines
Publication DB
Genome Profiling
Progenetix Use
Services
NCIt Mappings
UBERON Mappings
Upload & Plot
Download Data
Beacon+
Progenetix Info
About Progenetix
Use Cases
Documentation
Baudisgroup @ UZH

Search Samples | Modify Query

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

progenetix

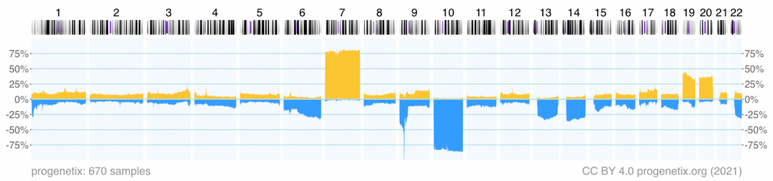
Samples: 668
Variants: 286
Calls: 675

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

UCSC region
JSON Response

Visualization options

Results | Biosamples | Biosamples Map | Variants

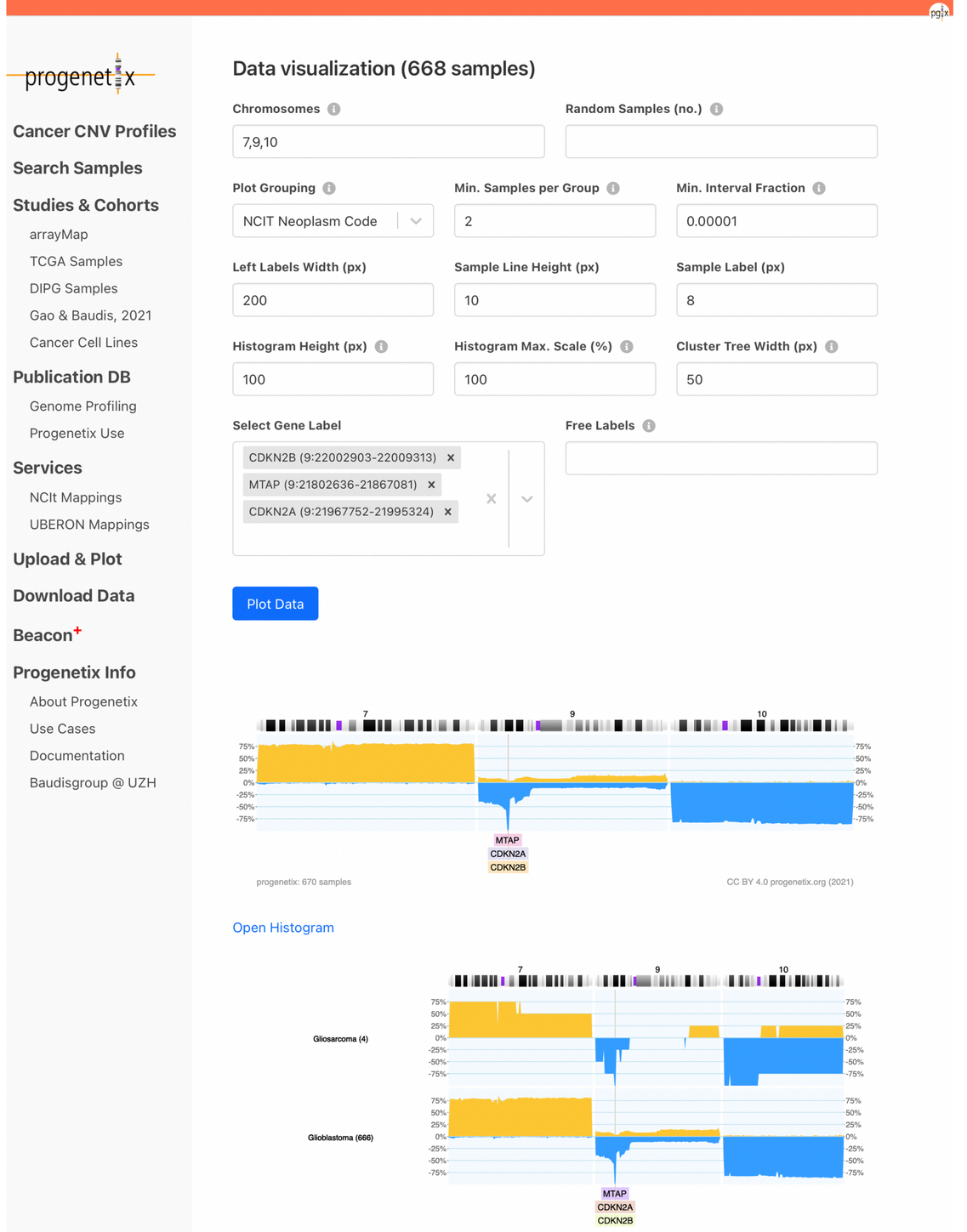


Matched Subset Codes	Subset Samples	Matched Samples	Subset Match Frequencies
UBERON:0002021	4	1	0.250
icdot-C71.4	4	1	0.250
icdom-94403	4291	664	0.155
NCIT:C3058	4375	664	0.152
UBERON:0016525	14	2	0.143
icdot-C71.1	14	2	0.143
UBERON:0000955	7068	651	0.092
icdot-C71.9	7066	651	0.092
icdom-94423	84	4	0.048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

Progenetix in 2022

Cancer Genomics Reference Resource

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

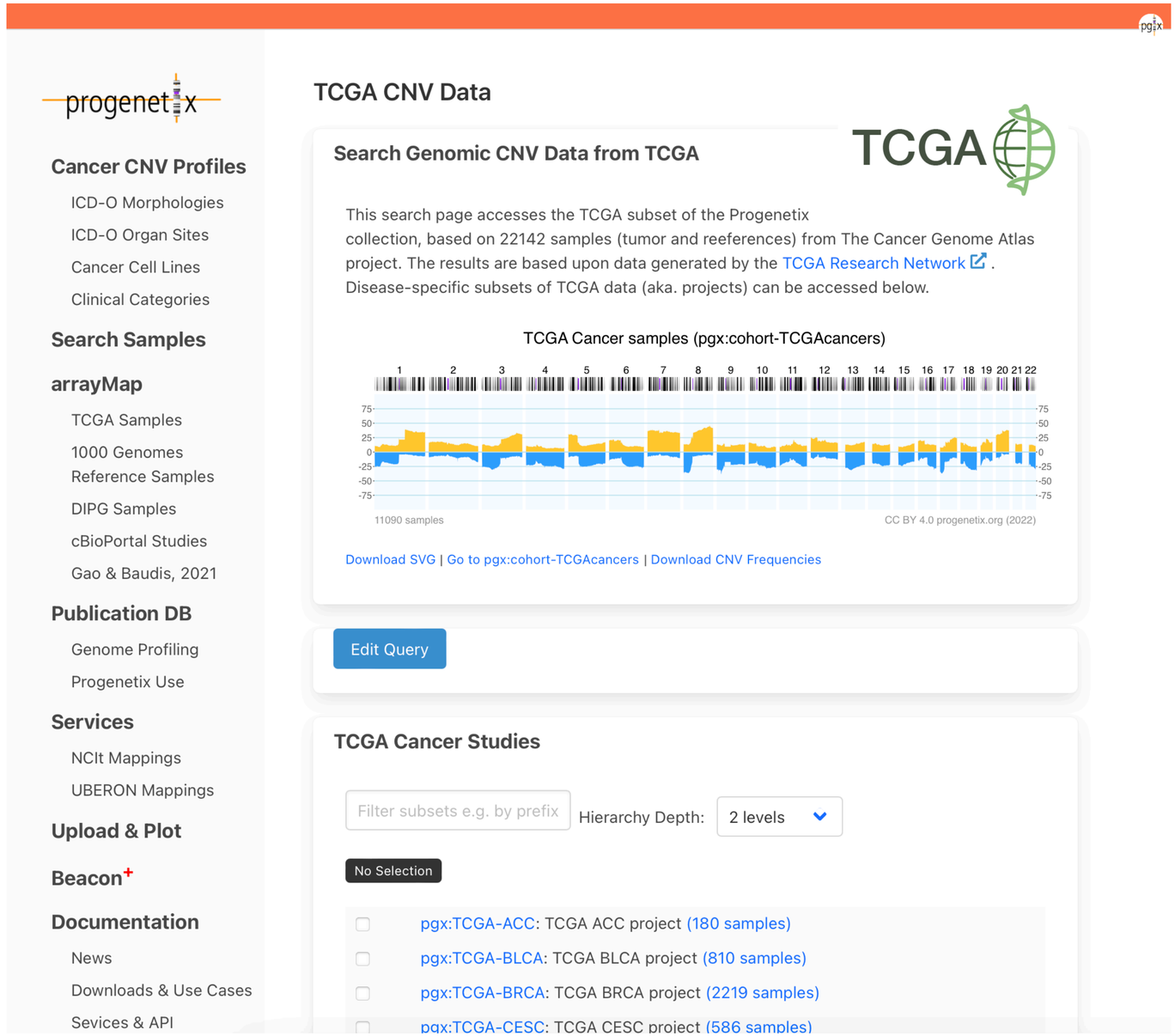


progenetix

Progenetix in 2022

Cancer Genomics Reference Resource

- contains special data subsets, identified using the "cohorts" concept
  - TCGA CNV data
  - 1000Genomes germline CNVs (WGS)
  - Cancer cell line CNVs with upcoming addition of annotated SNV ... data
  - cBioPortal studies
  - ...

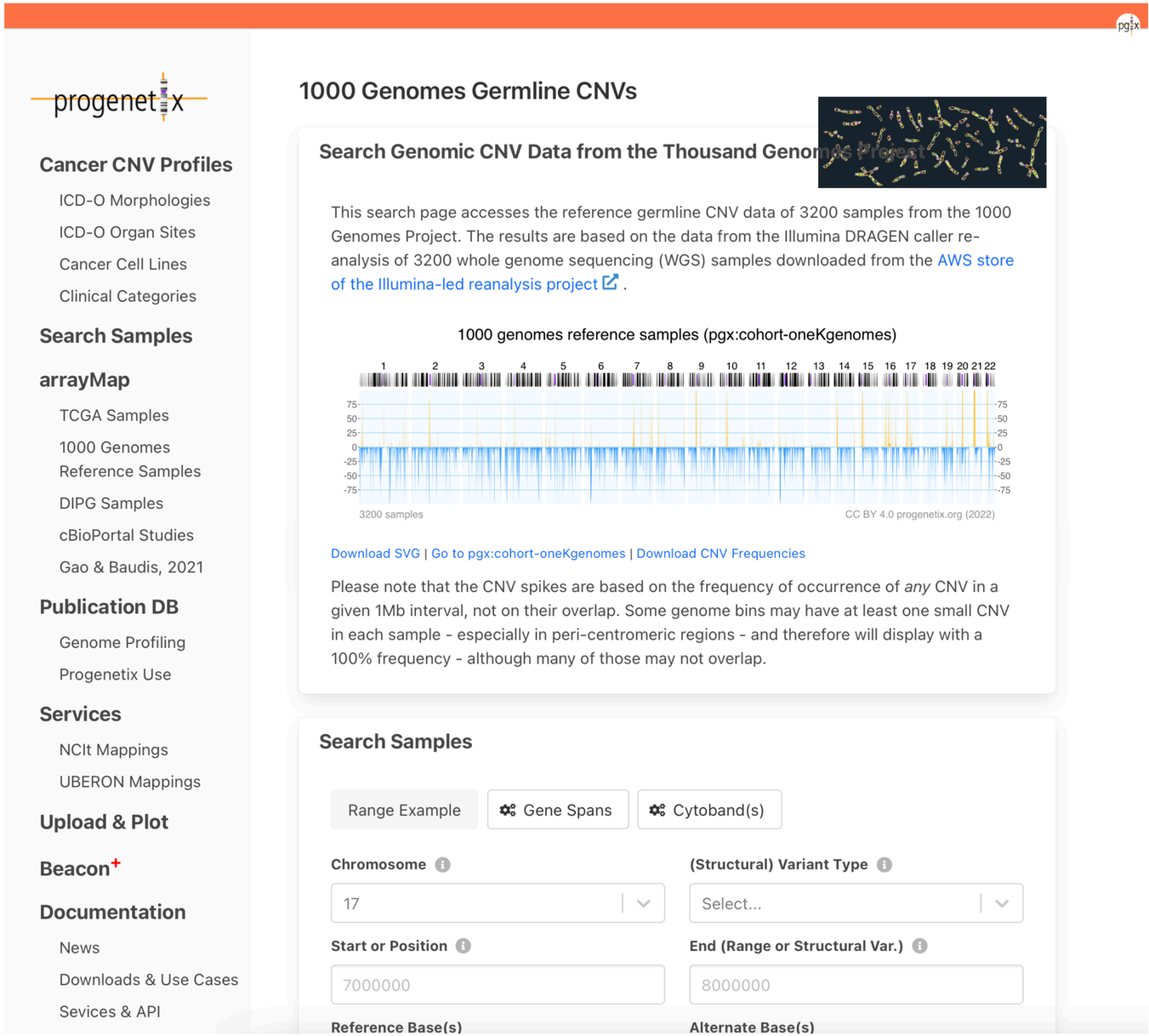


progenetix

Progenetix in 2022

Cancer Genomics Reference Resource

- contains special data subsets, identified using the "cohorts" concept
  - TCGA CNV data
  - 1000Genomes germline CNVs (WGS)
  - Cancer cell line CNVs with upcoming addition of annotated SNV ... data
  - cBioPortal studies
  - ...



progenetix

Progenetix in 2022

Cancer Genomics Reference Resource

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

progenetix

progenetix

Cancer CNV Profiles

Search Samples

Studies & Cohorts

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

Publication DB

- Genome Profiling
- Progenetix Use

Services

- NCIt Mappings
- UBERON Mappings

Upload & Plot

Download Data

Beacon+

Progenetix Info

- About Progenetix

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.

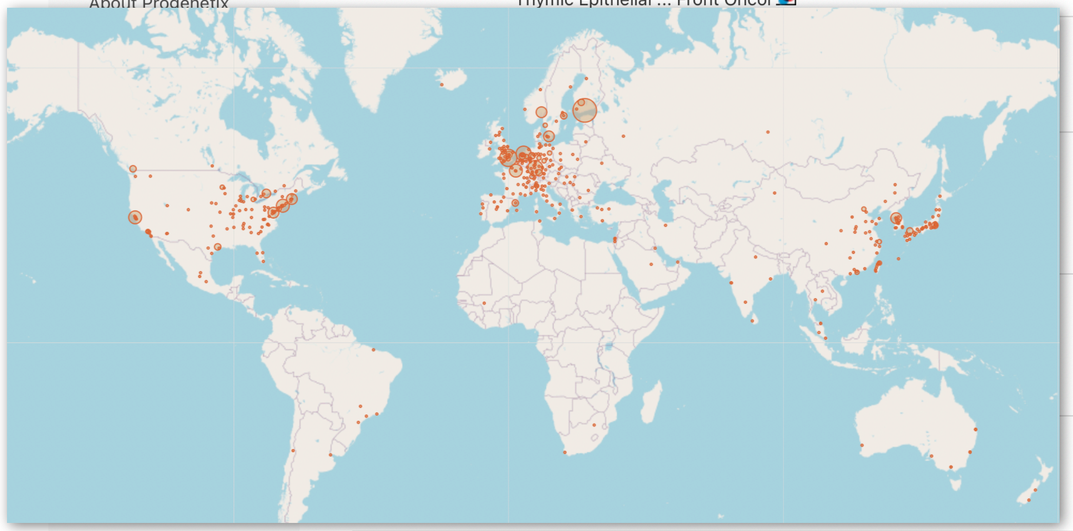
Please contact us to alert us about additional articles you are aware of. The inclusion criteria are described in the documentation.
New Oct 2021 You can now directly submit suggestions for matching publications to the oncopubs repository on Github.

Filter

City

Type to search...

Publications (3349)		Samples				
id	Publication	cCGH	aCGH	WES	WGS	pgx
PMID:34604048	Dai J, Jiang M, He K, Wang H, Chen P et al. (2021) DNA Damage Response and Repair Gene Alterations Increase Tumor Mutational Burden and ... Front Oncol	0	0	122	0	0
PMID:34573430	Juhari WKW, Ahmad Amin Noordin KB et al. (2021) Whole-Genome Profiles of Malay Colorectal Cancer Patients with Intact MMR Proteins. ... Genes (Basel)	0	0	0	7	0
PMID:34307137	Xu S, Li X, Zhang H, Zu L, Yang L et al. (2021) Frequent Genetic Alterations and Their Clinical Significance in Patients With Thymic Epithelial ... Front Oncol	0	0	0	123	0
					0	0
					0	0
					135	0
					0	0



Ontologies and Classifications

Services: Ontologymaps (NCIt)

NCIthesaurus

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

NCIT and ICD-O 3

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

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

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

Code Selection

NCIT:C4337: Mantle Cell Lymphoma

Optional: Limit with second selection

Matching Code Mappings {JSON↗}

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

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

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

List of filters recognized by different query endpoints

Public Ontologies with CURIE-based syntax

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

Private filters

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

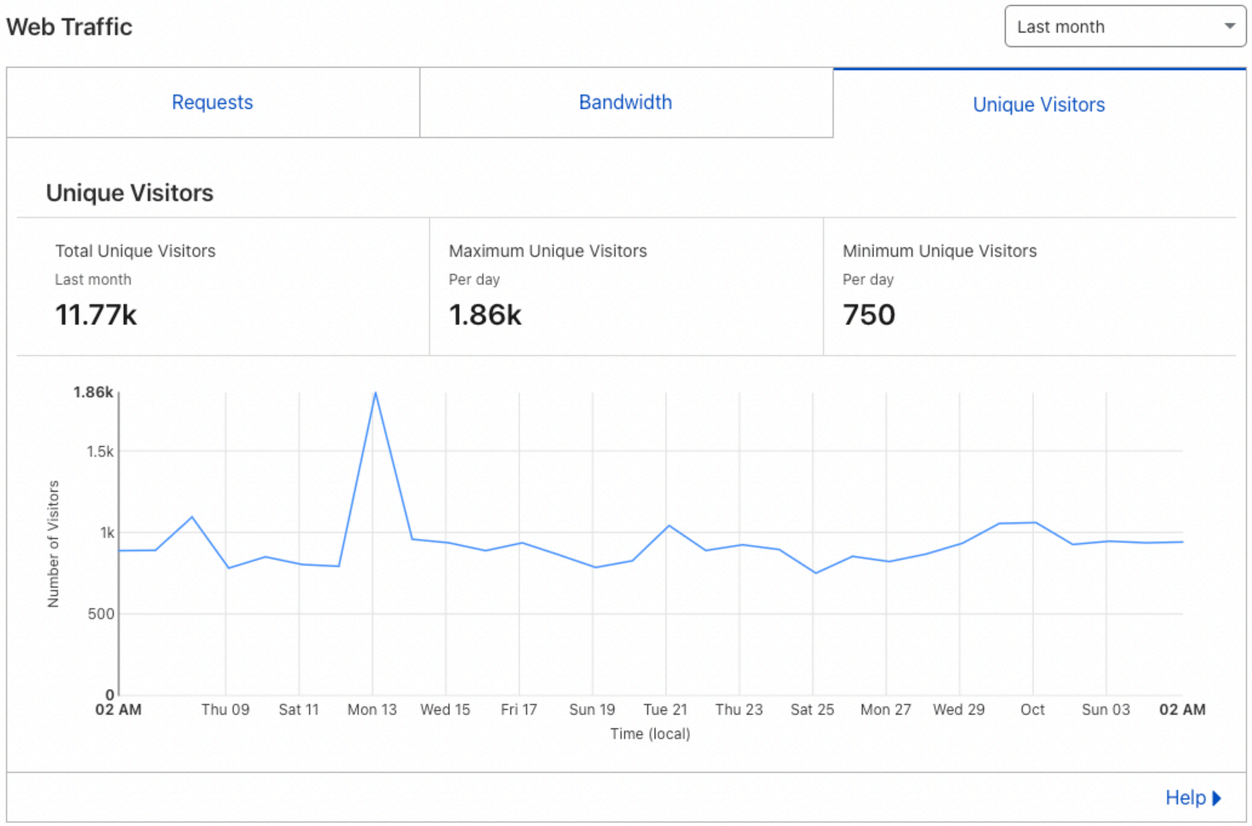
For terms with a `pgx` prefix, the identifiers.org resolver will

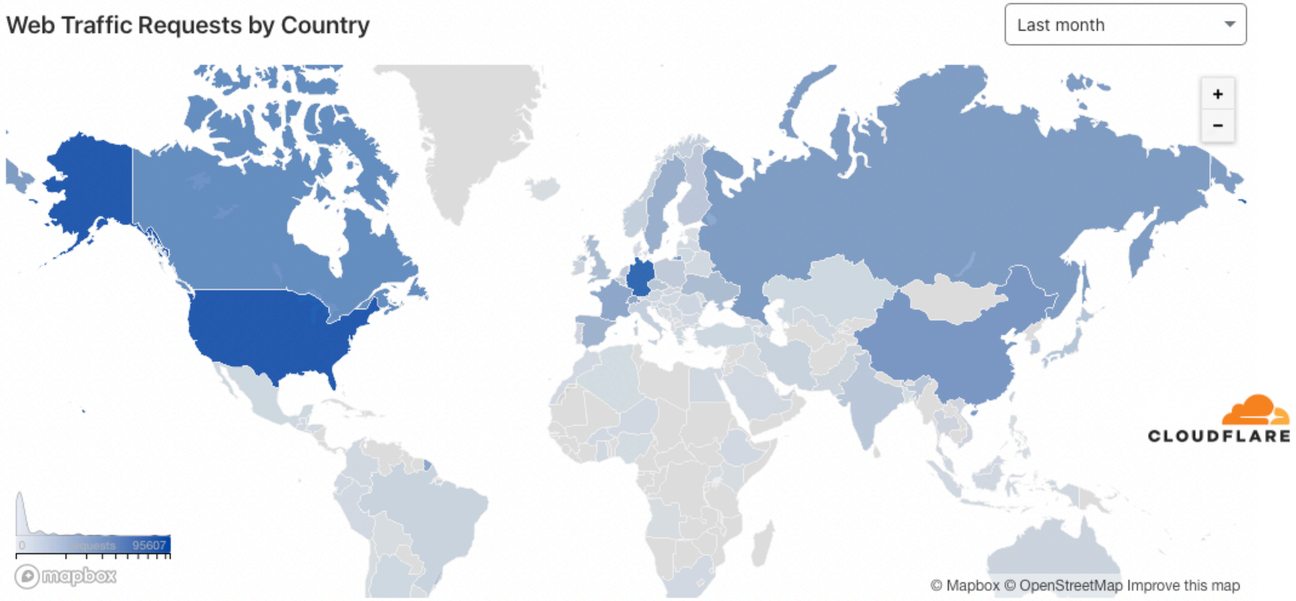
Filter prefix / local part	Code/Ontology	Example
pgx:icdom-...	ICD-O 3[7] Morphologies (Progenetix)	pgx:icdom-81703
pgx:icdot...	ICD-O 3[7] Topographies(Progenetix)	pgx:icdot-C04.9
TCGA	The Cancer Genome Atlas (Progenetix)[8]	TCGA-000002fc-53a0-420e-b2aa-a40a358bba37
pgx:pgxcohort-...	Progenetix cohorts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

Progenetix in 2022

Cancer Genomics Reference Resource

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





Cloudflare | progenetix.org | Sept-Oct 2021

progenetix

Recent Publications

CNV Data Analysis & Methods

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

Genomics 112 (2020) 3331–3341
Contents lists available at ScienceDirect
Genomics
journal homepage: www.elsevier.com/locate/ygeno

Minimum error calibration and normalization for genomic copy number analysis

Bo Gao[a,b], Michael Baudis[a,b,*]

Cai et al. BMC Genomics 2
http://www.biomedcentra

RESEARCH A

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

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

AUG 2021

SOFTWARE TOOL ARTICLE

REVISED segment_liftover : a Python tool to convert segments between genome assemblies [version 2; peer review: 2 approved]

Bo Gao[1,2], Qingyao Huang[1,2], Michael Baudis[1,2]

NTIFIC
ORTS
research

Ai et al. BMC Genomics (2016) 17:799
DOI 10.1186/s12864-016-3074-7

OPEN

Enabling population assignment from cancer genomes with SNP2pop

Qingyao Huang[1,2] & Michael Baudis[1,2*]

ORIGINAL PAPER

CNARA: reliability assessment for genomic copy number profiles

Ni Ai[1*], Haoyang Cai[2], Caius Solovan[3] and Michael Baudis[1*]

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*]

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.

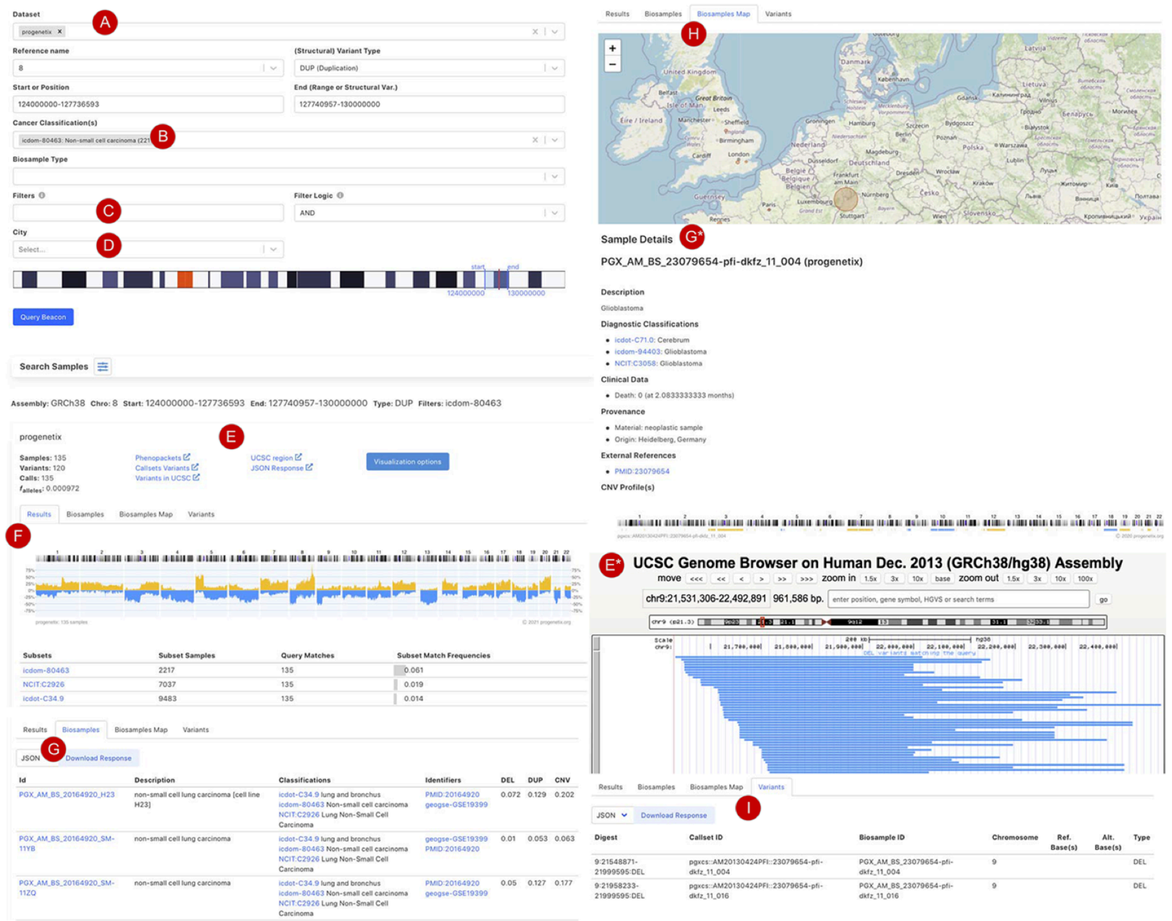


Figure 3. Beacon-style query using fuzzy ranges to identify biosamples with variants matching the CNA range This example queries for a continuous, focal duplication covering the complete MYC gene's coding region with < = 6 Mb in size. A: Filter for dataset; B: filter for cancer classification (NCIt and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

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]**

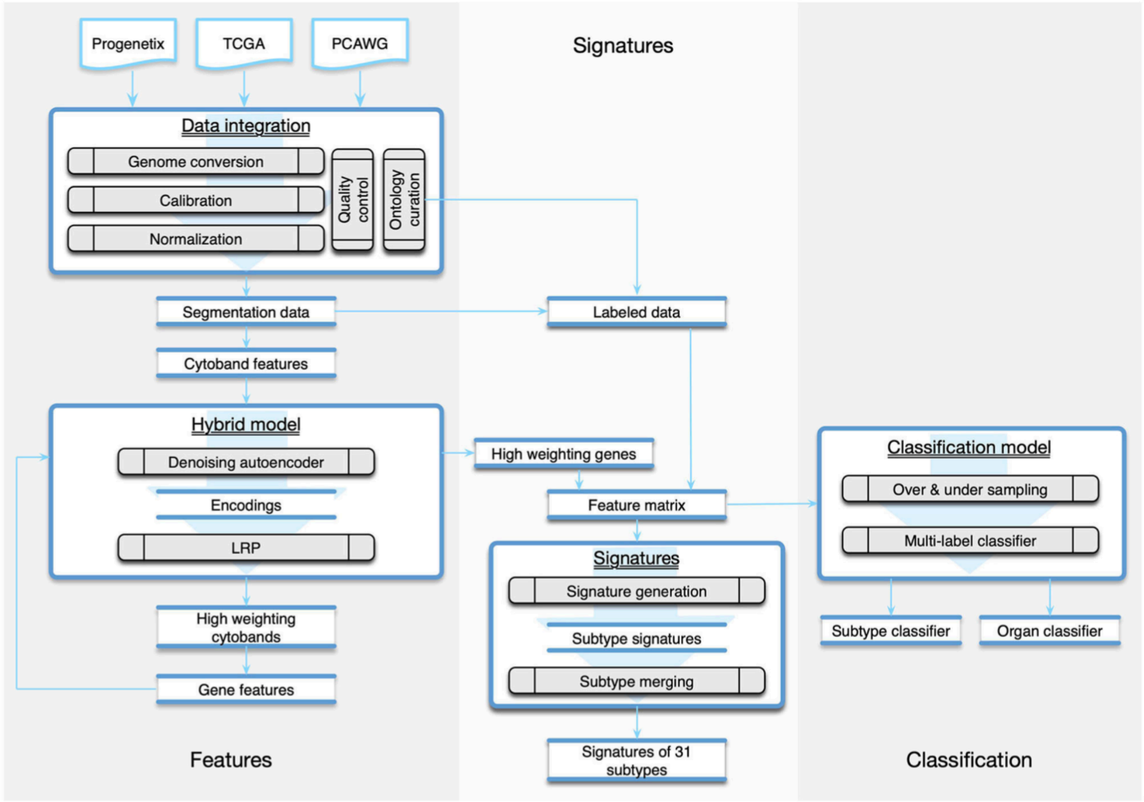


FIGURE 1 | The workflow of the study was composed of three parts. The *Features* part consisted of methods of data integration and feature generation. The *Signature* part focused on creating CNA signatures for cancer subtypes and the categorization of subtypes. The *Classification* part recruited machine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

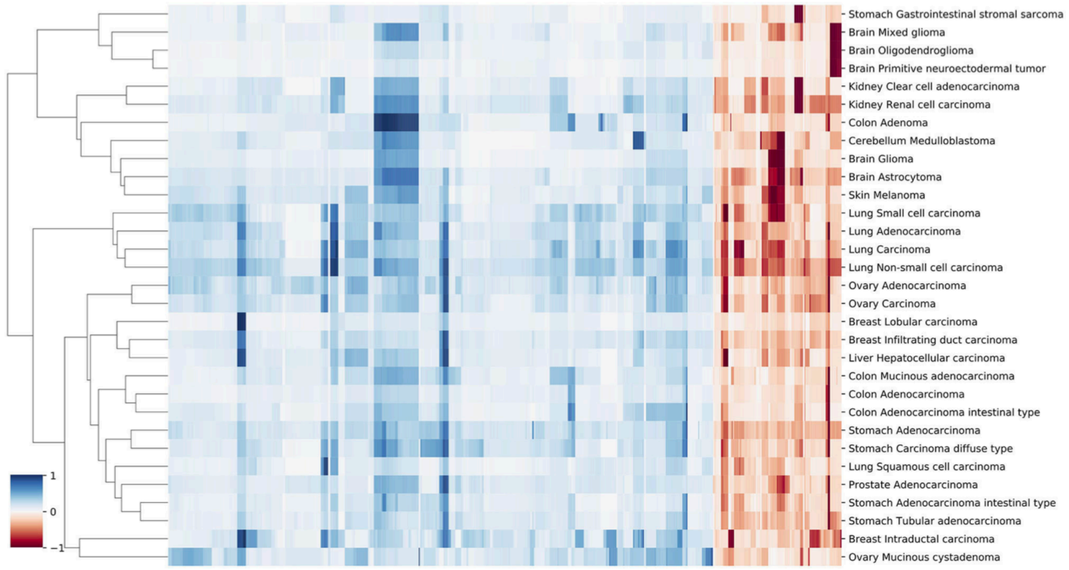


FIGURE 5 | A clustering heatmap of features in 31 signatures. Columns are normalized average CNV intensities of feature genes, where the blue colors are duplication features and red colors are deletion features. Duplication and deletion frequencies are normalized separately.

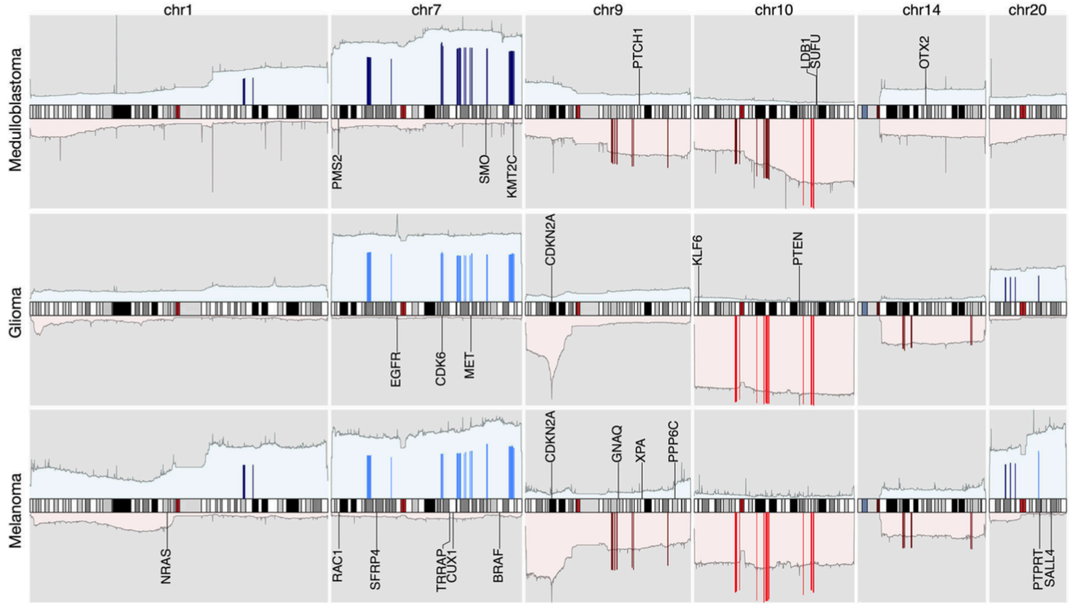
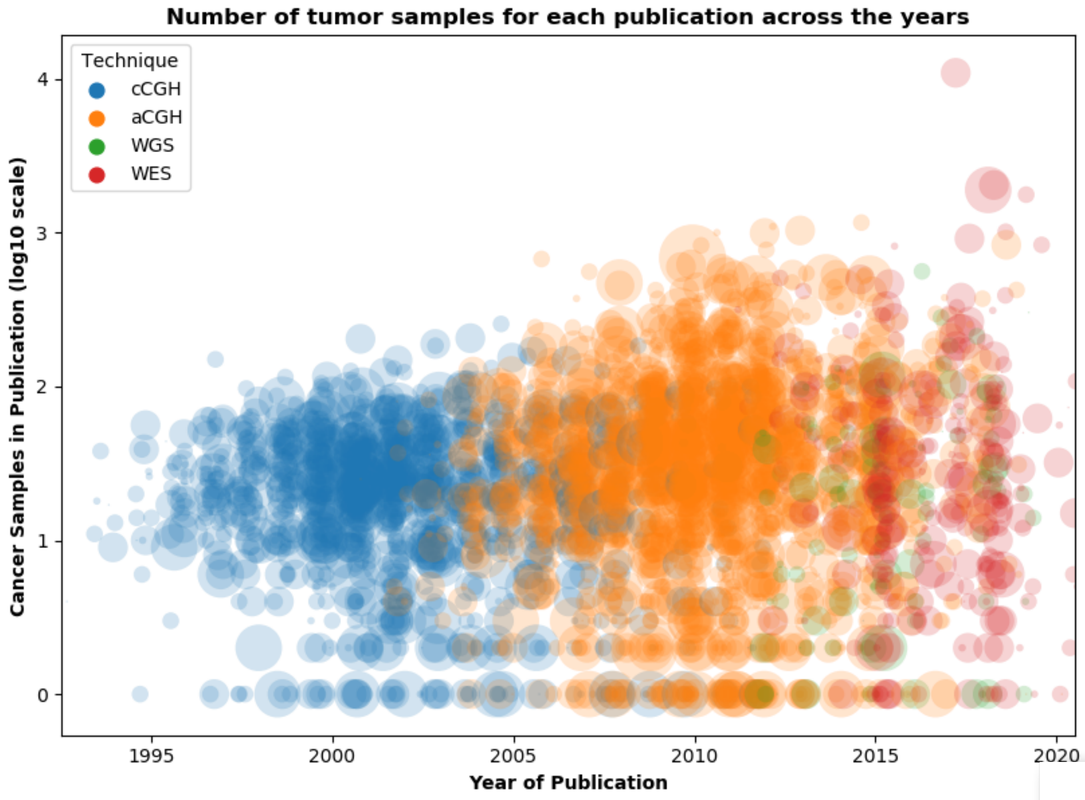
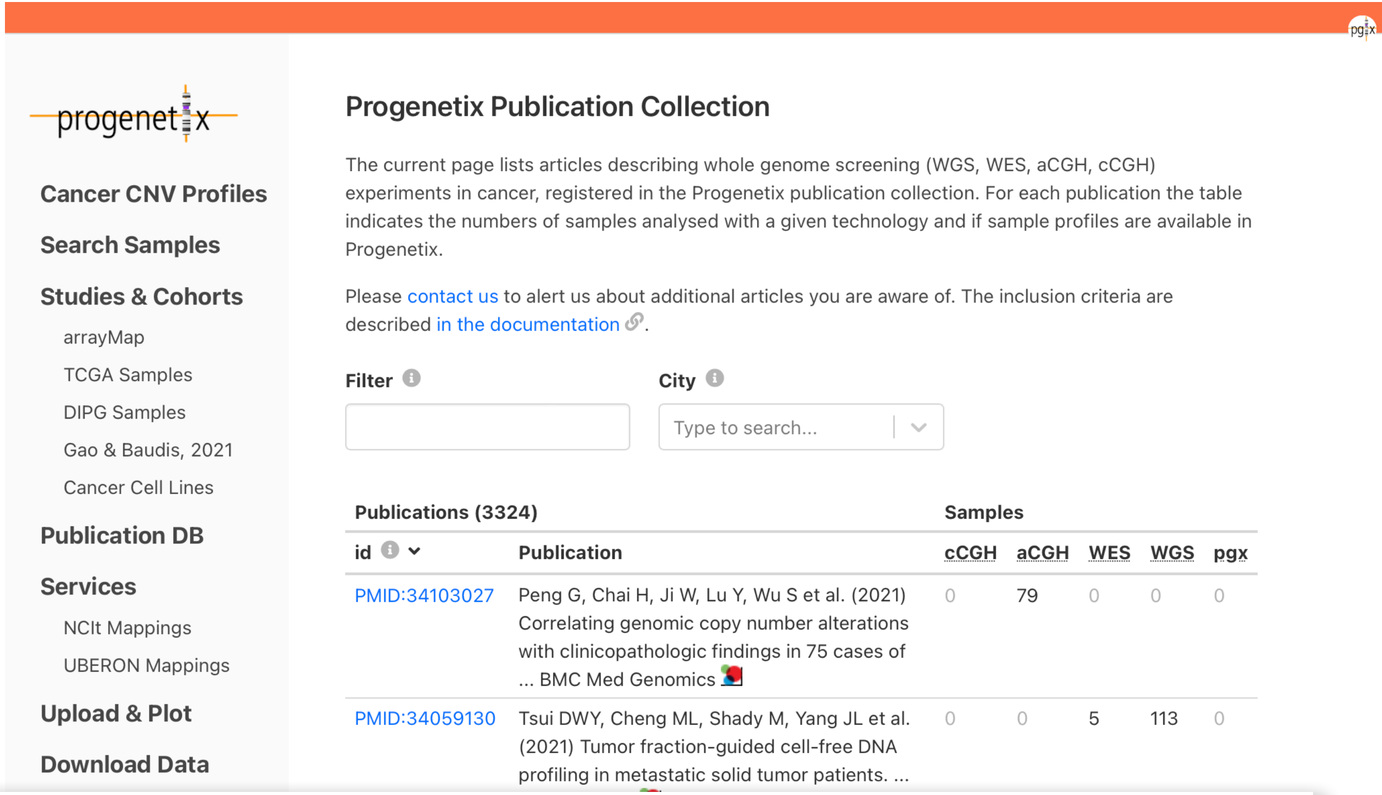


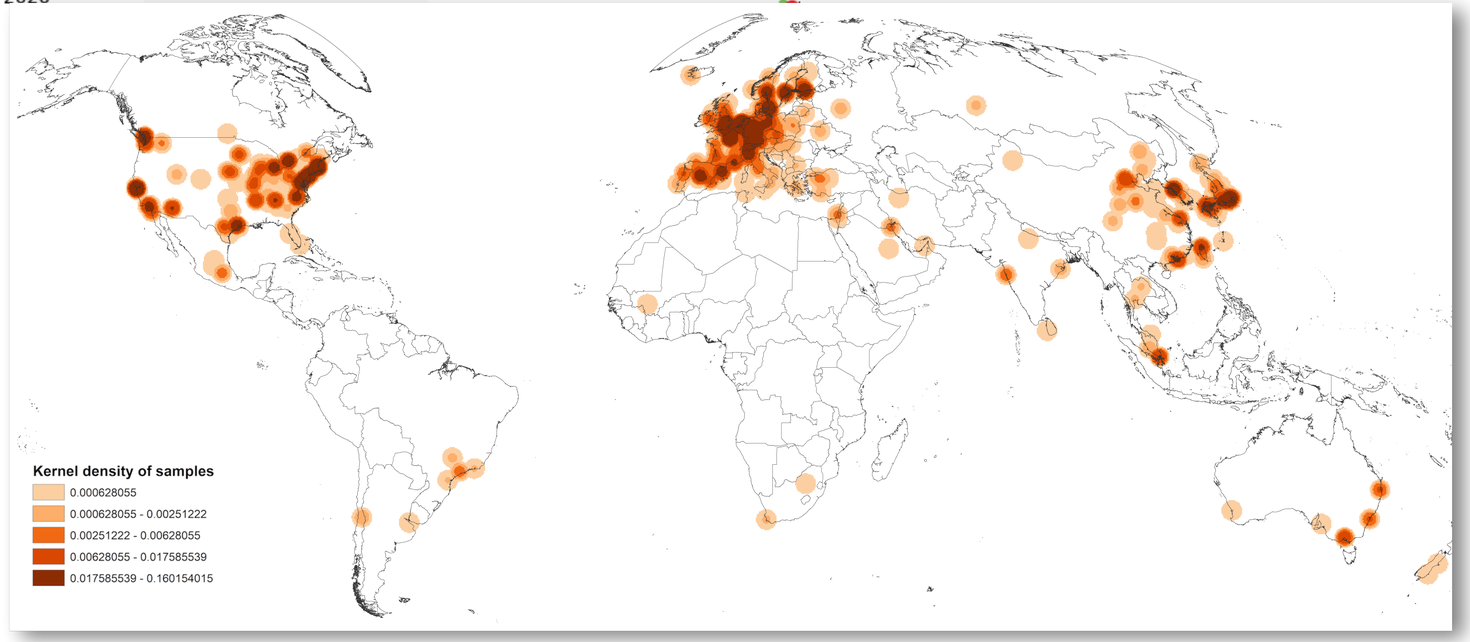
FIGURE 6 | The integrated view of the original data and the selected features, in the neural crest originating entities medulloblastoma, glioma, and melanoma. The shaded background area color illustrates the original data. Color bars illustrate the feature genes, where brighter colors indicate stronger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

What we have ...

- ✓ collection of >4000 articles assessed for scope
  - training set for NLP & search engine generation
- ✓ cancer specific ontologies with cross-mappings (ICD-O vs. NCIt) based on >100k samples
  - existing service API
- ✓ metadata ontology mappings for some 10k samples, with varying coverage for grade / stage / survival / ...
- ✓ CNV profiles for >110k samples, >700 entities with disease codes and metadata
- ✓ cell line CNV profiles together with mapped variants with clinical evidences

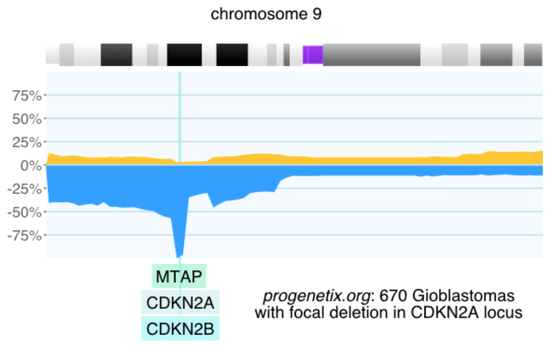
What we're working on...

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

ELIXIR hCNV

First Implementation Study and Ongoing Work

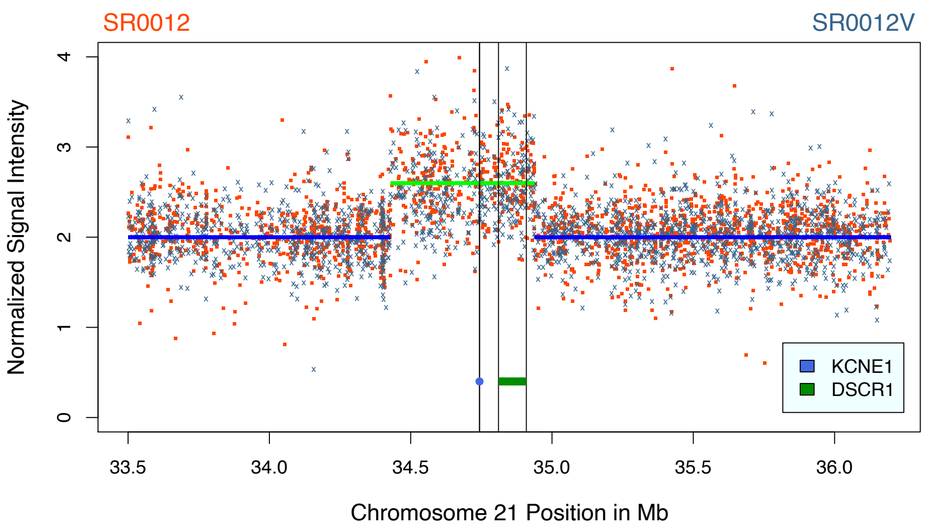
Michael Baudis | ELIXIR Human Data Communities | 2022-03-15



Why hCNV Community?

Structural Genome Variation Data :: Resources and Technologies

- structural genome variations are a major contributor to genetic diseases and cancer
- knowledge about and standards for copy number variations / aberrations (CNV/CNA) has not been in step with NGS & GWAS driven SNV/SNP assessment



CNV with unknown clinical impact in a case of Silver-Russel Syndrome

Local Affymetrix Genotyping 6 signal distribution pattern and segmentation result in patient SR12 (SR0012, orange data) and his father (SR0012V, steelblue data). In both samples a duplication in the DSCR can be observed, affecting the whole KCNE1 and DSCR1/RCAN coding regions. In contrast, DYRK1A lays ~2.5 Mb distal of the duplication. Only the genes discussed in this article are shown.

Mission statement

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

RESEARCH ARTICLE

AMERICAN JOURNAL OF medical genetics PART A

Identification of a 21q22 Duplication in a Silver–Russell Syndrome Patient Further Narrows Down the Down Syndrome Critical Region

Thomas Eggermann,[1]* Nadine Schönherr,[1] Sabrina Spengler,[1] Susanne Jäger,[1] Bernd Denecke,[2] Gerhard Binder,[3] and Michael Baudis[4]

[1]Institute of Human Genetics, RWTH Aachen, Aachen, Germany

[2]Interdisciplinary Centre for Clinical Research, IZKF "BIOMAT," RWTH Aachen, Aachen, Germany

[3]Section of Paediatric Endocrinology and Diabetology, University Children's Hospital, Tuebingen, Germany

[4]Institute of Molecular Biology, University of Zürich, Zürich, Switzerland

Received 26 June 2009; Accepted 6 November 2009

hCNV Implementation Study 2019-2021

Some Achievements and Deliveries

- HGVS satellite meeting – Human CNV – June 14th 2019 – Göteborg Sweden
- hCNV community workshop ELIXIR All-Hands Lisbon – June 2019
- survey of data annotation formats, including comments on VCF development
- start FAIRification of CNV national / reference databases (BANCCO, Progenetix)
- Community white paper published
- Biohackathon Paris 2019
- in 2021 start of shared meetings of subgroup with Beacon variants scout team

F1000Research F1000Research 2020, 9(ELIXIR):1229 Last updated: 01 JUN 2021

Check for updates

OPINION ARTICLE

The ELIXIR Human Copy Number Variations Community: building bioinformatics infrastructure for research [version 1; peer review: 1 approved]

David Salgado[1], Irina M. Armean[2], Michael Baudis[3], Sergi Beltran[4,5], Salvador Capella-Gutierrez[6,7], Denise Carvalho-Silva[2,8], Victoria Dominguez Del Angel[9], Joaquin Dopazo[10], Laura I. Furlong[11], Bo Gao[3], Leyla Garcia[2,12,13], Dietlind Gerloff[14], Ivo Gut[4,5], Attila Gyenesei[15], Nina Habermann[16], John M. Hancock[13], Marc Hanauer[17], Eivind Hovig[18,19], Lennart F. Johansson[20], Thomas Keane[2], Jan Korbel[16], Katharina B. Lauer[13], Steve Laurie[4], Brane Leskošek[21], David Lloyd[13], Tomas Marques-Bonet[22], Hailiang Mei[23], Katalin Monostory[24], Janet Piñero[11], Krzysztof Poterlowicz[25], Ana Rath[17], Pubudu Samarakoon[26], Ferran Sanz[11], Gary Saunders[13], Daoud Sie[27], Morris A. Swertz[20], Kirill Tsukanov[2], Alfonso Valencia[6,7,28], Marko Vidak[21], Cristina Yenyxe González[2], Bauke Ylstra[29], Christophe Béroud[1,30]

[1]Aix Marseille Univ, INSERM, MMG, Marseille, France
[2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK
[3]Department of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[4]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Baldiri Reixac 4, Barcelona 08028, Spain
[5]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[6]Barcelona Supercomputing Center (BSC), Barcelona, Spain
[7]Spanish National Bioinformatics Institute (INB)/ELIXIR-ES, Barcelona, Spain
[8]Open Targets, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK
[9]Institut Français de Bioinformatique, UMS3601-CNRS, CNRS, Paris, France
[10]Clinical Bioinformatics Area, Fundación Progreso y Salud, CDCA, Hospital Virgen del Rocio, Sevilla, Spain
[11]Research Programme on Biomedical Informatics (GRIB), Hospital del Mar Medical Research Institute (IMIM), Department of Experimental and Health Sciences, Pompeu Fabra University (UPF), Barcelona, Spain
[12]ZB MED Information Centre for Life Sciences, Cologne, Germany
[13]ELIXIR Hub, Hinxton, UK
[14]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg
[15]Szentágothai Research Center, University of Pécs, Pécs, Hungary
[16]Genome Biology, European Molecular Biological Laboratory, Heidelberg, Germany
[17]Orphanet, INSERM, Paris, France
[18]Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway
[19]Centre for bioinformatics, Department of Informatics, University of Oslo, Oslo, Norway
[20]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
[21]Faculty of Medicine - ELIXIR Slovenia, University of Ljubljana, Ljubljana, Slovenia
[22]Institute of Evolutionary Biology (UPF-CSIC), Catalan Institution for Research and Advanced Studies, Barcelona, Spain
[23]Sequencing Analysis Support Core, Leiden University Medical Center, Leiden, The Netherlands
[24]Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary
[25]Centre for Skin Sciences, University of Bradford, Bradford, UK

hCNV Implementation Studies 2021-2023

Focus on Integration with ELIXIR Platforms and Communities - and beyond

- original 2019-2021 implementation study provided visibility and established connections for new studies
- instrumental were Biohackathons, use case & standards surveys and co-participation of group members
- future work plans to leverage the resources of participants through pre-established interactions and synergies
- 2 independent studies provide clearer definitions of deliverables and individual scopes

Michael Baudis	CH
Christophe Béroud	FR
David Salgado	FR
Alexander Kanitz	CH
Anthony Brookes	UK
Babita Singh	ES
Björn Grüning	DE
Jordi Rambla	ES
Kirill Tsukanov	EMBL-EBI
Krzysztof Poterlowicz	UK
Salvador Capella-Gutierrez	ES
Sergi Beltran	ES
Steven Laurie	ES
Tim Beck	UK
Timothee Cezard	EMBL-EBI

Ongoing...

hCNV & Intl. Community

- contributions to ontologies and standard definitions
- close ongoing interactions with GA4GH work streams
- influencing the development of the GA4GH VRS variant standard

hCNV Community

Genomic Copy Number Variations in Humans

News & Events
- ELIXIR All Hands 2022 - h-CNV Representation
- CNV Ontology Proposal - Now Live at EFO
- hCNV Site now at cnvar.org
- hCNV Implementation Study 2021/2: Beacon and Beyond
- all ...

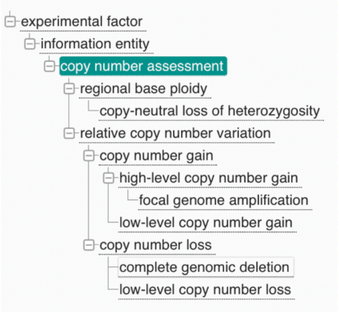
Participants
Standards and Guidelines
Studies & Resources
Examples, Guides & FAQ
Contacts

Related Sites
- h-CNV @ ELIXIR
- Beacon Project
- Beacon @ ELIXIR
- SchemaBlocks

Github Projects
- h-CNV

CNV Ontology Proposal - Now Live at EFO

As part of the hCNV-X work - related to "Workflows and Tools for hCNV Data Exchange Procedures" and to the intersection with Beacon and GA4GH VRS - we have now a new proposal for the creation of an ontology for the annotation of (relative) CNV events. The CNV representation ontology is targeted for adoption by Sequence Ontology (SO) and then to be used by an updated version of the VRS standard. Please see the discussions linked from the proposal page. However, we have also contributed the CNV proposal to EFO where it has gotten live on January 21.



Everybody is welcome to contribute to the editing of the proposal at the SO & VRS Github repositories!

2021-01-21: `copy number assessment` term tree now live on EFO

The `copy number assessment` term tree has been accepted into the Experimental Factor Ontology and can be used for referencing CNV types.

More ontologies...

... with h-CNV contributions ca

2022-01-21

larrybabb commented 18 days ago

per a discussion between **@ahwagner** and **@larrybabb**
Dreaft Relative Copy Number class proposal

```
-- the target region/gene/feature
subject:  region/gene/feature/allele/haplotype

--5 quantifiable values that correspond to the EFO copy number assess
copy number assessment:   (http://www.ebi.ac.uk/efo/EFO_0030063)
        -2 = complete loss  (http://www.ebi.ac.uk/efo/EFO_0030069)
        -1 = partial loss   (http://www.ebi.ac.uk/efo/EFO_0030068)
         0 = copy-neutral   (http://www.ebi.ac.uk/efo/EFO_0030064)
         1 = low-level gain (http://www.ebi.ac.uk/efo/EFO_0030071)
         2 = high-level gain(http://www.ebi.ac.uk/efo/EFO_0030072)
```

RelativeCopyNumber

Relative Copy Number Variation captures a classification of copies of a molecule within a system, relative to a baseline. These types of Variation are common outputs from CNV callers, particularly in the somatic domain where Absolute Copy Counts are difficult to estimate and less useful in practice than relative statements.

Computational Definition

The relative copies of a Molecular Variation, Feature, Sequence Expression, or a CURIE reference against an unspecified baseline in a system (e.g. genome, cell, etc.).

Information Model

Some RelativeCopyNumber attributes are inherited from Variation.

Field	Type	Limits	Description
_id	CURIE	0..1	Variation Id. MUST be unique within document.
type	string	1..1	MUST be "RelativeCopyNumber"
subject	Molecular Variation | Feature | Sequence Expression | CURIE	1..1	Subject of the Copy Number object
relative_copy_class	string	1..1	MUST be one of "complete loss", "partial loss", "copy neutral", "low-level gain" or "high-level gain".

Experimental Factor Ontology

Global Alliance for Genomics & Health

Progenetix and GA4GH Beacon

Implementation driven development of a GA4GH standard

progenetix

The vision: Federation of data

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.

Enabling responsible genomic data sharing for the benefit of human health

The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization, seeking to enable responsible genomic data sharing within a human rights framework.

The Global Alliance for Genomics and Health

Making genomic data accessible for research and health

- January 2013 - 50 participants from eight countries
- June 2013 - White Paper, over next year signed by 70 “founding” member institutions (e.g. SIB, UZH)
- March 2014 - Working group meeting in Hinxton & 1st plenary in London
- October 2014 - Plenary meeting, San Diego; interaction with ASHG meeting
- June 2015 - 3rd Plenary meeting, Leiden
- September 2015 - GA4GH at ASHG, Baltimore
- October 2015 - DWG / New York Genome Centre
- April 2016 - Global Workshop @ ICHG 2016, Kyoto
- October 2016 - 4th Plenary Meeting, Vancouver
- May 2017 - Strategy retreat, Hinxton
- October 2017 - 5th plenary, Orlando
- May 2018 - Vancouver
- October 2018 - 6th plenary, Basel
- May 2019 - GA4GH Connect, Hinxton
- October 2019 - 7th Plenary, Boston
- October 2020 - Virtual Plenary, June 2021 - Virtual Connect ...
- October 2021 - Virtual Plenary ...
- September 2022 - 10th Plenary, Barcelona

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

22 SEPTEMBER 2022 | BARCELONA, SPAIN

GA4GH 10th Plenary

Global Alliance for Genomics & Health

GENOMICS

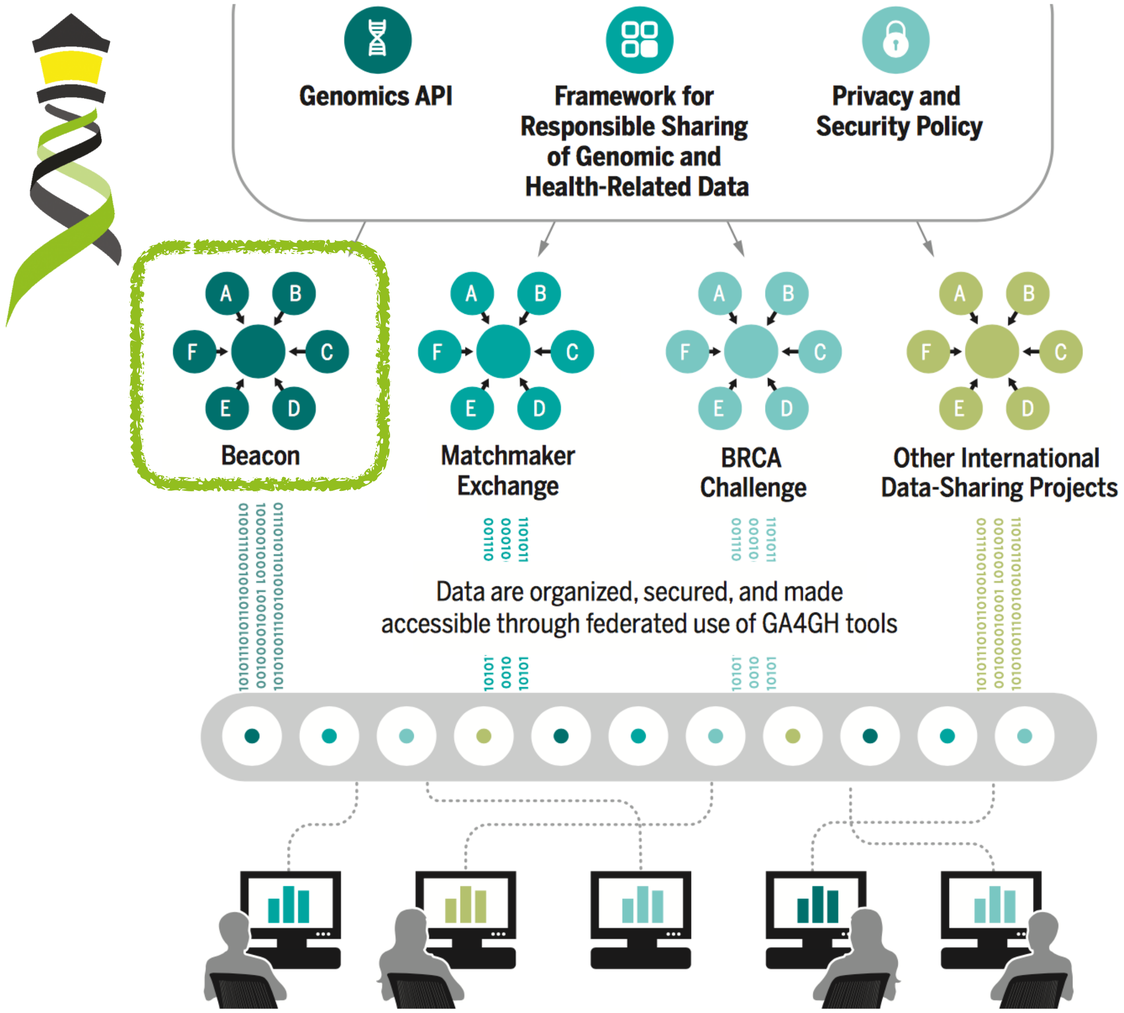
A federated ecosystem for sharing genomic, clinical data

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

The Global Alliance for Genomics and Health*

SCIENCE 10 JUNE 2016 • VOL 352 ISSUE 6291

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

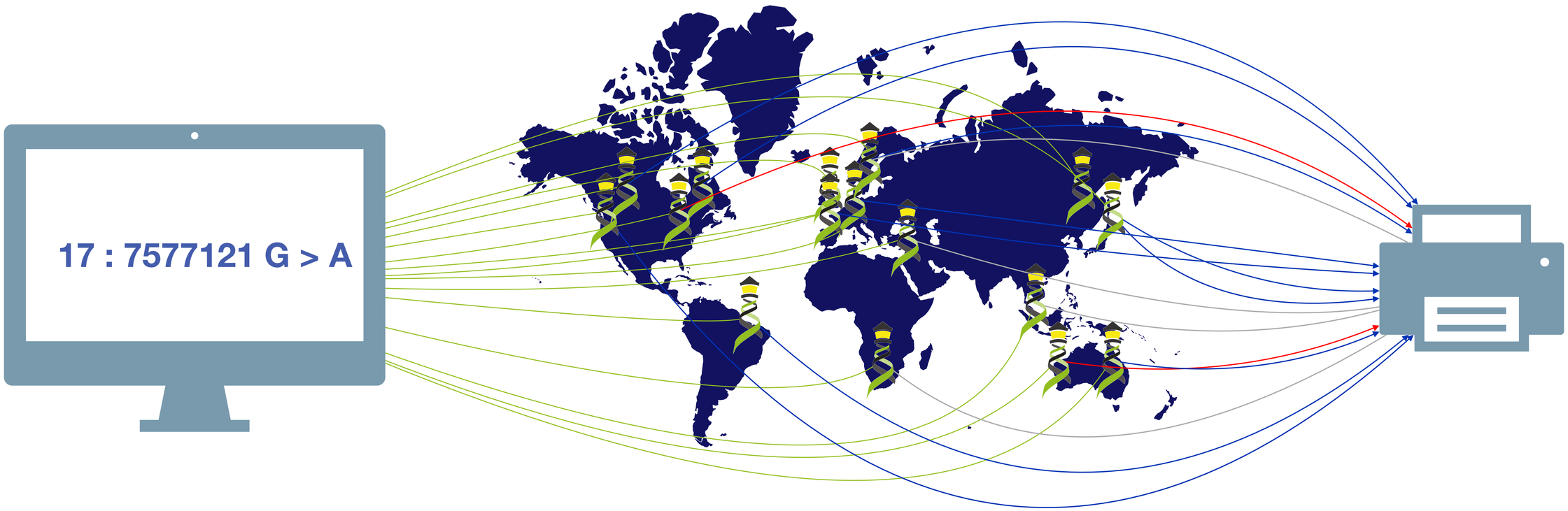


DNASTACK

Global Alliance for Genomics & Health

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

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



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

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

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

Global Alliance "Beacon" - Jim Ostell, NCBI, March 7, 2014

Global Alliance for Genomics & Health

Introduction

… I proposed a challenge application for all those wishing to seriously engage in ***international*** data sharing for human genomics. …

1. Provide a public web service
2. Which accepts a query of the form "Do you have any genomes with an "A" at position 100,735 on chromosome 3?"
3. And responds with one of "Yes" or "No" …

"Beacon" because … people have been scanning the universe of human research for ***signs of willing participants in far reaching data sharing***, but … it has remained a dark and quiet place. The hope of this challenge is to 1) ***trigger the issues*** blocking groups … in way that isn't masked by the … complexities of the science, fully functional interfaces, and real issues of privacy, and to 2) in ***short order*** … see ***real beacons of measurable signal*** … from ***at least some sites*** … Once your "GABeacon" is shining, you can start to take the ***next steps to add functionality*** to it, and ***finding the other groups*** … following their GABeacons.

Utility

Some have argued that this simple example is not "useful" so nobody would build it. Of course it is not the first priority for this application to be scientifically useful. …intended to provide a ***low bar for the first step of real*** … ***engagement***. … there is some utility in …locating a rare allele in your data, … not zero.

A number of more useful first versions have been suggested.

1. Provide ***frequencies of all alleles*** at that point
2. Ask for all alleles seen in a gene ***region*** (and more elaborate versions of this)
3. Other more complicated queries

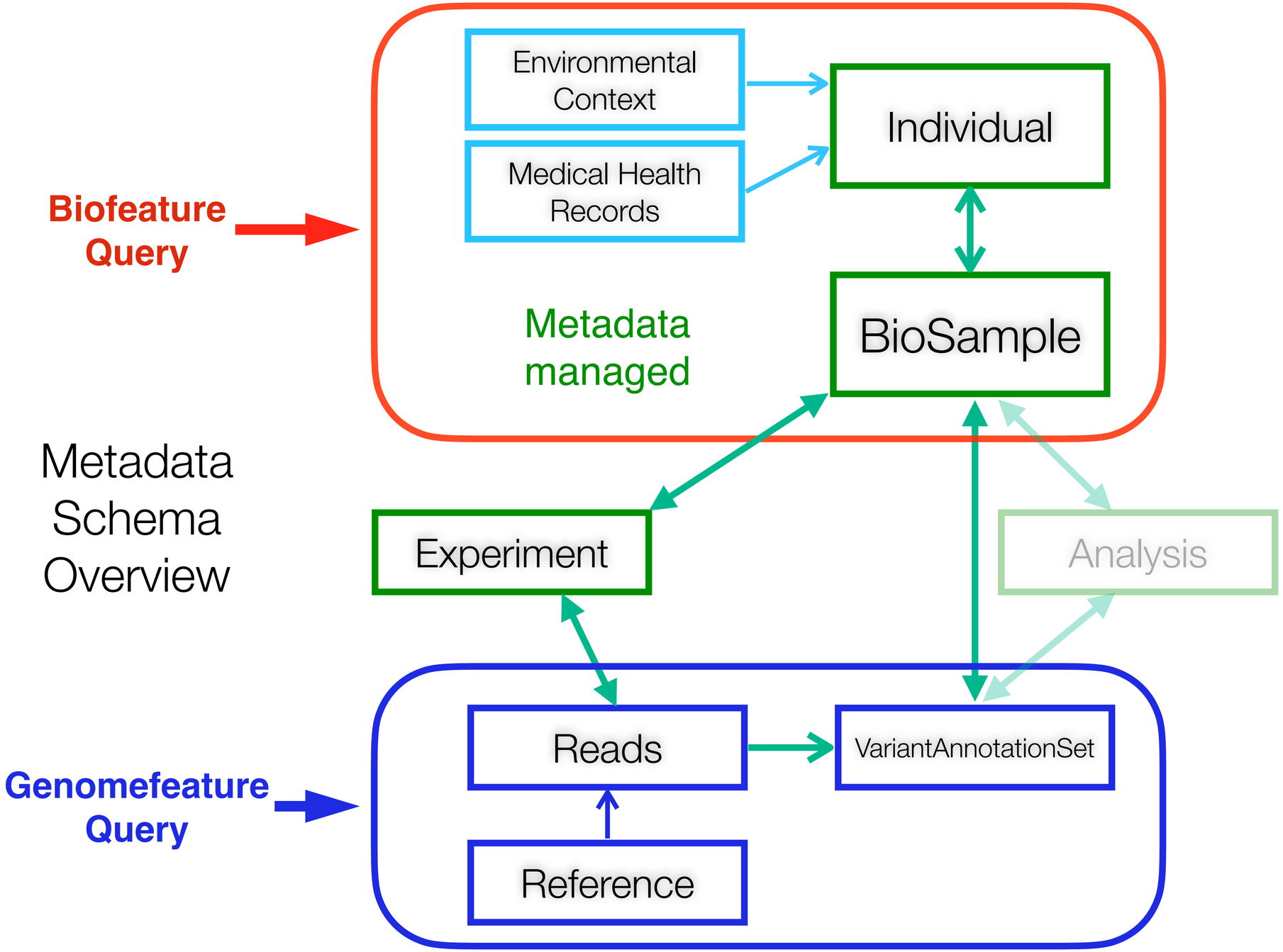
"I would personally recommend all those be held for **version 2**, when the beacon becomes a service."
Jim Ostell, 2014

Implementation

1. Specifying the chromosome … The interface needs to specify the ***accession.version*** of a chromosome, or ***build number***…
2. Return values … right to ***refuse*** to answer without it being an error … DOS ***attack*** … or because …especially ***sensitive***…
3. Real time response … Some sites suggest that it would be necessary to have a ***"phone home" response*** …

Beacon+ in 2016 - Driving GA4GH implementations

Metadata Schema Overview



ELIXIR - Making Beacons Biomedical

- Authentication to enable non-aggregate, patient derived datasets
  - ELIXIR AAI with compatibility to other providers (OAuth...)
- Scoping queries through "biodata" parameters
- Extending the queries towards clinically ubiquitous variant formats
  - cytogenetic annotations, named variants, variant effects
- Beacons as part of local, secure environments
  - local EGA ...
- Beacon queries as entry for **data delivery**
  - handover to stream and download using htsget, VCF, EHRs
- Interacting with EHR standards
  - FHIR translations for queries and handover ...

beacon.progenetix.org

Beacon+

This forward looking Beacon interface implements additional, pl

Info

Query

ample | DGV Example | CNV Example

Dataset	tcga
Reference name*	9
Genome Assembly*	GRCh38 / hg38
Start *min* Position*	19,500,000
Start *max* Position	21,975,098
End *min* Position	21,967,753
End *max* Position	24,500,000
Alt. Base(s)*	DEL
Bio-ontology	icdot:c50.9: (4065)

Beacon Implementations

- implementing existing resources with Beacon protocol
- e.g. TCGA cancer variants (structural and SNV)

Beacon Response

- quantitative (counts for variants, callsets and samples)
- *Handover* to authentication system for data retrieval
- **no exposure** of data beyond standard Beacon response and additional pointer to matched data

Prototyping Query Extensions

- testing e.g. bio-metadata queries using ontology terms

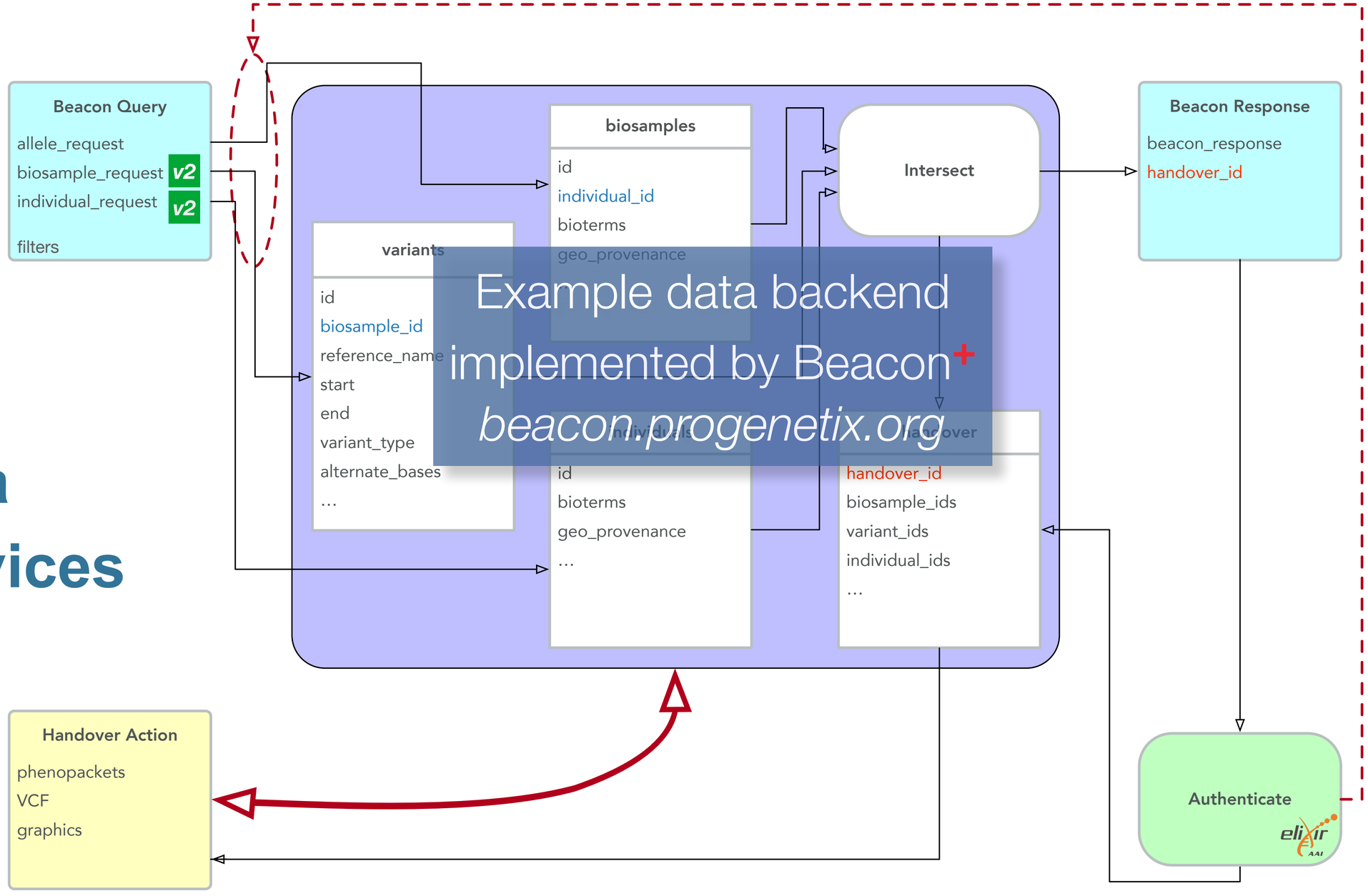
Dataset	Assembly	Chro	Start Range	End Range	Pos	Ref Alt	Bio Query	Variants Calls Samples	$f_{alleles}$	Response Context
tcga	hg38	9	19,500,000 21,975,098	21,967,753 24,500,000		DEL	icdot:c50.9	54 54 54	0.0243	JSON UCSC Handover

arrayMap progenetix This Beacon implementation is developed by the Computational Oncogenomics Group at the University of Zurich, with support from the SIB Technology group and ELIXIR.

University of Zurich UZH elixir SIB

Beacon & Handover

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



Beacon I/O

Handover

Authentication

Michael Baudis

Global Alliance for Genomics & Health

SIB

University of Zurich UZH

elixir

Beacon+

This example shows a core Beacon query, against a specific mutation in the TP53 gene, in cellosaurus, with ClinVar data.

CNV Example | SNV Range Example | SNV Example | ClinVar Example | Beacon Help

Dataset*

arraymap
progenetix
cellosaurus
dipg
BeaconSpecTest2
BeaconSpecTest

Genome Assembly*

GRCh38 / hg38

Dataset Responses

All Selected Datasets

Reference name*

17

Gene Coordinates

TP53

Cytoband(s)

17p13.1

Start

7673767

Ref. Base(s)

C

Alt. Base(s)

T

Bio-ontology

no selection
NCIT:C102872: Pharyngeal squamous cell carcinoma (2)
NCIT:C103968: Pyruvate dehydrogenase deficiency (1)
NCIT:C105555: High grade ovarian serous adenocarcinoma (75)
NCIT:C105556: Low grade ovarian serous adenocarcinoma (10)
NCIT:C111802: Dyskeratosis congenita (3)

Other Filters

additional comma-separated, prefixed filters

Beacon Query

Beacon+ Flexible Modeling of New Features

Our Beacon platform is being used for the rapid testing of queries and responses - both v1.n and v2.0.a - against a number of partially large-scale genome datasets.

- Progenetix (>100000 cancer CNV profiles)
- DIPG (childhood brain tumor study)
- NEW: Cellosaurus ClinVar annotations for evidence representation
- Brewing: COVID-19

Currently running on a ~~Perl~~+MongoDB stack, a Python-based OS solution is in early development.

progenetix SIB SPHN

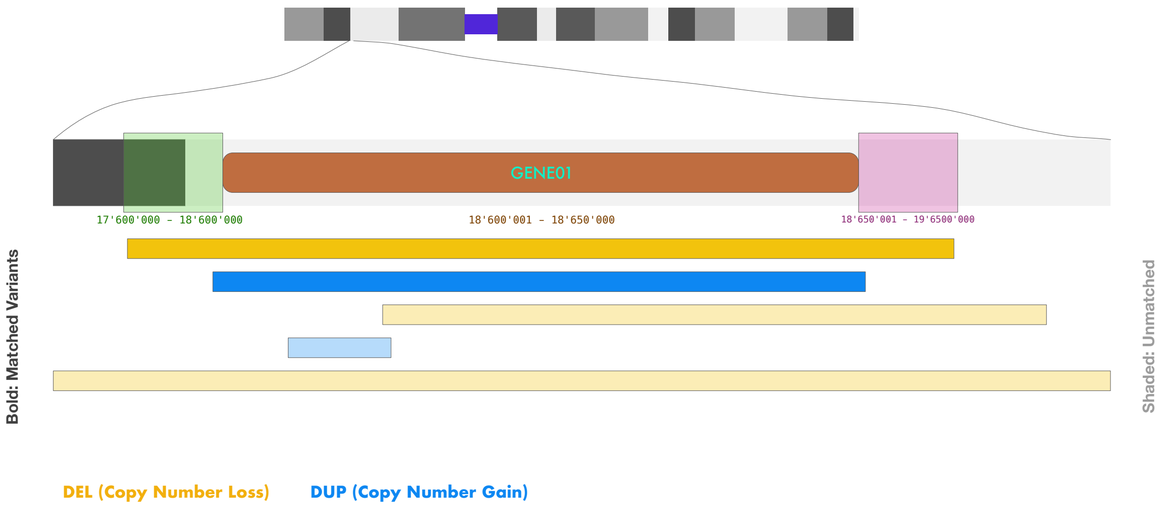

```
[
  {
    "callset_id": "cs-cellosaurus:CVCL_EI02",
    "info": {
      "cellosaurus": {
        "cell_line": "BT474-LAPRa",
        "id": "CVCL_EI02",
        "cellosaurus_variant_name": "TP53 p.Glu285Lys (c.853G&gt;A)"
      },
      "clinvar": {
        "gene_id": "7157",
        "allele_id": "410258",
        "assembly": "GRCh38",
        "cytoband": "17p13.1",
        "variant_type": "single nucleotide variant",
        "origin": "germline;somatic",
        "phenotype": "Hereditary cancer-predisposing syndrome;Li-Fraumeni
syndrome;PARP Inhibitor response;not provided",
        "clinical_significance": "Pathogenic/Likely pathogenic",
        "clinvar_full_name": "NM_001126112.2(TP53):c.853G&gt;A (p.Glu285Lys)"
      }
    },
    "start_min": 7673766,
    "reference_name": "17",
    "end_min": 7673767,
    "biosample_id": "bios-cellosaurus:CVCL_EI02",
    "alternate_bases": [
      "T"
    ],
    "digest": "17_7673767_C_T",
    "reference_bases": "C",
    "variantset_id": "cellosaurus_clinvar_GRCH38",
    "end_max": 7673767,
    "start_max": 7673766
  },
  {
    "digest": "17_7673767_C_T",
    "reference_bases": "C",
    "alternate_bases": [
      "T"
    ],
    "variantset_id": "cellosaurus_clinvar_GRCH38",
    "end_max": 7673767,
    "start_max": 7673766,
    "callset_id": "cs-cellosaurus:CVCL_AQ07",
    "start_min": 7673766,
    "info": {
      "cellosaurus": {
        "cellosaurus_variant_name": "TP53 p.Glu285Lys (c.853G&gt;A)",
        "cell_line": "BT-474 Clone 5",
        "id": "CVCL_AQ07"
      },
      "clinvar": {
        "assembly": "GRCh38",
        "allele_id": "410258",
        "gene_id": "7157",
        "cytoband": "17p13.1",
        "variant_type": "single nucleotide variant",
        "phenotype": "Hereditary cancer-predisposing syndrome;Li-Fraumeni
syndrome;PARP Inhibitor response;not provided",
        "origin": "germline;somatic",
        "clinvar_full_name": "NM_001126112.2(TP53):c.853G&gt;A (p.Glu285Lys)",
        "clinical_significance": "Pathogenic/Likely pathogenic"
      }
    },
    "end_min": 7673767,
    "biosample_id": "bios-cellosaurus:CVCL_AQ07",
    "reference_name": "17"
  },
  {
    "alternate_bases": [
      "T"
    ],
    "reference_bases": "C",
    "digest": "17_7673767_C_T",
    "end_max": 7673767,
    "variantset_id": "cellosaurus_clinvar_GRCH38",
    "start_max": 7673766,
    "callset_id": "cs-cellosa
```

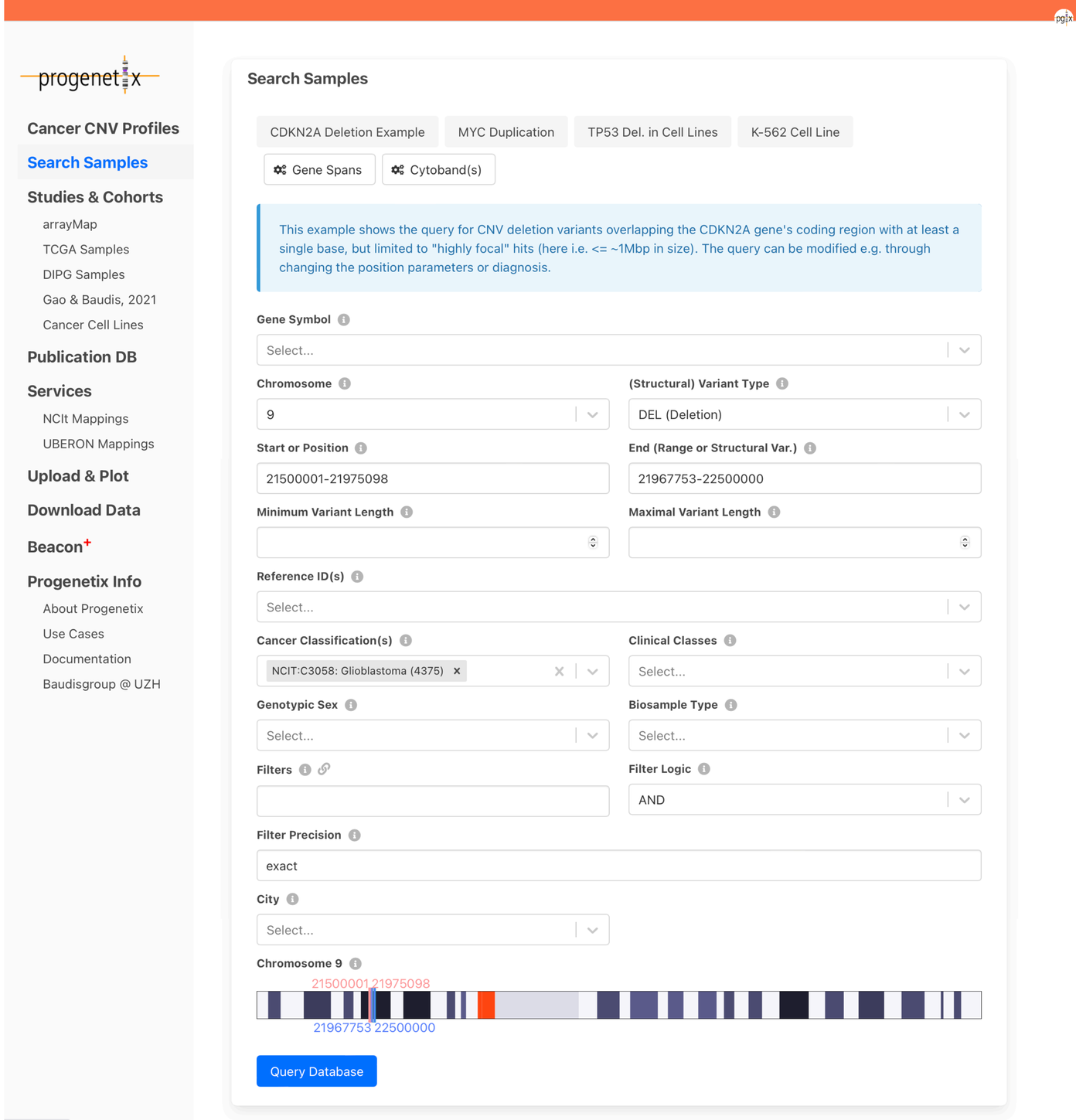
Progenetix in 2022

Variant and Metadata for Sample Discovery

- positional queries for genomic variants using the **GA4GH Beacon protocol**
- metadata queries (diagnoses, identifiers, clinical classes ...) using **Beacon "filters"**

Genome Bracket Query
(full match)





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

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

Onboarding

Demonstrating Compliance

- Progenetix Beacon+ has served as implementation driver since 2016
- Beacon v2 as service with protocol-driven registries for federation
- GA4GH approved Beacon v2 in April 2022

Global Alliance for Genomics & Health

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

Progenetix Source Code ¶

With exception of some utility scripts and external dependencies (e.g. MongoDB

the software (from database interaction to website) behind Progenetix and Beac

`bycon`

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

`progenetix-web`

- website for Progenetix and its Beacon+ implementations
- provides Beacon interfaces for the `bycon` server, as well as other Progenetix sevices (e.g. the publicat
- implemented as React / Next.js project
- contains this documentation tree here as `mkdocs` project, with files in the `docs` directory

Beacon API

Beacon-style JSON responses

The Progenetix resource's API utilizes the `bycon` framework for data query and delivery and represents a custom implementation of the Beacon *v2* API.

The standard format for JSON responses corresponds to a generic Beacon v2 response, with the `meta` and `response` root elements. Depending on the endpoint, the main data will be a list of objects either inside `response.results` or (mostly) in `response.resultSets.results`. Additionally, most API responses (e.g. for biosamples or variants) provide access to data using *handover* objects.

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

Base `/biosamples`

`/BIOSAMPLES/` + QUERY

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

es/pgxbs-kftva5c9

a single biosample

MODE=TRUE

es?testMode=true

some random samples

- for testing API responses

`/BIOSAMPLES/{ID}/G_VARIANTS`

- /biosamples/pgxbs-kftva5c9/g_variants/
- retrieval of all variants from a single biosample

Base `/individuals`

`/INDIVIDUALS` + QUERY ¶

- /individuals?filters=NCIT:C7541

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

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

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

Org.progenetix

Progenetix & Beacon+

The Beacon+ implementation - developed in the Python & MongoDB based `bycon` project - implements an expanding set of Beacon v2 paths for the Progenetix resource ⊞.

Scoped responses from query object

In queries with a complete `beaconRequestBody` the type of the delivered data is independent of the path and determined in the `requestedSchemas`. So far, Beacon+ will compare the first of those to its supported responses and provide the results accordingly; it doesn't matter if the endpoint was `/beacon/biosamples/` or `/beacon/variants/` etc.

Below is an example for the standard test "small deletion CNVs in the CDKN2A locus, in gliomas" Progenetix test query, here responding with the matched variants. Exchanging the `entityType` entry to

- `{ "entityType": "biosample", "schema:": "https://progenetix.org/services/schemas/Biosample/"}`

would change this to a biosample response. The example ccan be tested by POSTing this as `application/json` to `http://progenetix.org/beacon/variants/` or `http://progenetix.org/beacon/biosamples/`.

```
{
    "$schema":"beaconRequestBody.json",
    "meta": {
        "apiVersion": "2.0",
        "requestedSchemas": [
            {
                "entityType": "genomicVariant",
                "schema:": "https://progenetix.org/services/schemas/genomicVariant"
            }
        ]
    },
    "query": {
        "requestParameters": {
```

Beacon v2 Conformity and Extensions in Progenetix

Putting the + into Beacon ...

- support & use of standard Beacon v2 PUT & GET variant queries, filters and meta parameters
  - ➡ variant parameters, geneId, lengths, EFO & VCF CNV types, pagination
  - ➡ widespread, self-scoping filter use for bio-, technical- and and id parameters with switch for descending terms use (globally or per term if using POST)
- extensive use of handovers
  - ➡ asynchronous delivery of e.g. variant and sample data, data plots
- + extensions of query logic
  - ➡ optional use of OR logic for filter combinations (global)
- + extension of query parameters
  - ➡ geographic queries incl. $geonear and use of GeoJSON in schemas
- ┐(ᒧ ▽ ᒨ)┌ no implementation of authentication on this open dataset

Progenetix provides a number of additional services and output formats which are initiated over the / services path or provided as request parameters and are not considered Beacon extensions (though they follow the syntax where possible).

progenetix

Progenetix Stack

- JavaScript front-end is populated for query results using asynchronous access to multiple handover objects
  - biosamples and variants tables, CNV histogram, UCSC .bed loader, .pgxseg variant downloads…
- the complete middleware / CGI stack is provided through the *bycon* package
  - schemas, query stack, data transformation (e.g. Phenopackets generation)…
- data collections mostly correspond to the main Beacon default model entities
  - no separate *runs* collection; integrated w/ analyses
  - *variants* are stored per observation instance

- *collations* contain pre-computed data (e.g. CNV frequencies, statistics) and information for all grouping entity instances and correspond to **filter values**
  - PMID:10027410, NCIT:C3222, pgx:cohort-TCGA, pgx:icdom-94703…
- *querybuffer* stores id values of all entities matched by a query and provides the corresponding access handle for **handover** generation

```
_id: ObjectId("6249bb654f8f8d67eb94953b"),
id: '0765ee26-5029-4f28-b01d-9759abf5bf14',
source_collection: 'variants',
source_db: 'progenetix',
source_key: '_id',
target_collection: 'variants',
target_count: 667,
target_key: '_id',
target_values: [
  ObjectId("5bab578b727983b2e00ca99e"),
  ObjectId("5bab578d727983b2e00cb505"),
```

variants | analyses | biosamples | individuals

Entity collections

collations | geolocs | genespans | publications | qBuffer

Utility collections

pgxRpi

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

Beacon Path: Retrieve variants by biosample id(s)

```
https://progenetix.org/beacon/g_variants/
?biosampleIds=pgxbs-kftvh94d,pgxbs-kftvh94g,pgxbs-kftvh972
&output=pgxseg
```

Beacon Path: Get biosamples by filter(s)

```
http://progenetix.org/beacon/biosamples/
?filters=NCIT:C3697&output=datatable
```

Service Path: Retrieve CNV frequencies by filter(s)

```
http://www.progenetix.org/services/intervalFrequencies/
?id=NCIT:C4323&output=pgxseg
```

Author: Hangjia Zhao | @hangjiaz

README.md

pgxRpi

This is an API wrapper package to access data from Progenetix database.

You can install this package from GitHub using:

```
install.packages("devtools")
devtools::install_github("progenetix/pgxRpi")
```

If you are interested in accessing CNV variant data, get started from this vignette

If you are interested in accessing CNV frequency data, get started from this vignette

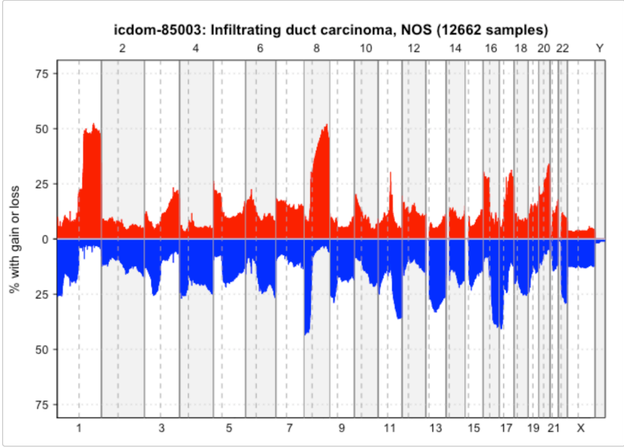
When you face problems, try to reinstall the latest version. If reinstallation doesn't help, please contact us.

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

```
biosamples <- pgxLoader(type="biosample", filters = "NCIT:C3059",codematches = TRUE,
                        biosample_id = c("pgxbs-kftva5zv","pgxbs-kftva5zw"))
```

```
freq_pgxseg <- pgxLoader(type="frequency", output ='pgxseg',
                         filters=c("NCIT:C4038","pgx:icdom-85003"),
                         codematches = TRUE)
```

```
pgxFreqplot(freq_pgxseg, filters='pgx:icdom-85003')
```



Beacon+: Phenopackets

Testing alternative response schemas...

http://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/phenopackets

- the v2 default schemas are mostly aligned w/ Phenopackets v2
- creating phenopackets can be done mostly by re-wrapping of Beacon entities (individual, biosample)
- variants can be included through file resource URLs; in Beacon+ this is done through *ad hoc* handover URIs


```
"id": "pgxpxf-kftx3tl5",
"metaData": {
  "phenopacketSchemaVersion": "v2",
  "resources": [
    {
      "id": "NCIT",
      "iriPrefix": "http://purl.obolibrary.org/obo/NCIT_",
      "name": "NCIt Plus Neoplasm Core",
      "namespacePrefix": "NCIT",
      "url": "http://purl.obolibrary.org/obo/ncit/neoplasm-core.c
      "version": "2022-04-01"
    },

  "subject": {
    "dataUseConditions": {
      "id": "DUO:0000004",
      "label": "no restriction"
    },
    "diseases": [
      {
        "clinicalTnmFinding": [],
        "diseaseCode": {
          "id": "NCIT:C3099",
          "label": "Hepatocellular Carcinoma"
        },
        "onset": {
          "age": "P48Y9M26D"
        },
        "stage": {
          "id": "NCIT:C27966",
          "label": "Stage I"
        }
      }
    ],
    "id": "pgxind-kftx3tl5",
    "sex": {
      "id": "PATO:0020001",
      "label": "male genotypic sex"
    },
    "updated": "2018-12-04 14:53:11.674000",
    "vitalStatus": {
      "status": "UNKNOWN_STATUS"
    }
  }

"biosamples": [
  {
    "biosampleStatus": {
      "id": "EFO:0009656",
      "label": "neoplastic sample"
    },
    "dataUseConditions": {
      "id": "DUO:0000004",
      "label": "no restriction"
    },
    "description": "Primary Tumor",
    "externalReferences": [
      {
        "id": "pgx:TCGA-0004d251-3f70-4395-b175-c94c2f5b1b81",
        "label": "TCGA case_id"
      },
      {
        "id": "pgx:TCGA-TCGA-DD-AAVP",
        "label": "TCGA submitter_id"
      },
      {
        "id": "pgx:TCGA-9259e9ee-7279-4b62-8512-509cb705029c",
        "label": "TCGA sample_id"
      },
      {
        "id": "pgx:TCGA-LIHC",
        "label": "TCGA LIHC project"
      }
    ],
    "files": [
      {
        "fileAttributes": {
          "fileFormat": "pgxseg",
          "genomeAssembly": "GRCh38"
        },
        "uri": "https://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/variants/?output=pgxseg"
      }
    ],
    "histologicalDiagnosis": {
      "id": "NCIT:C3099",
      "label": "Hepatocellular Carcinoma"
    },
    "id": "pgxbs-kftvhyvb",
    "individualId": "pgxind-kftx3tl5",
    "pathologicalStage": {
      "id": "NCIT:C27966",
      "label": "Stage I"
    },
    "sampledTissue": {
      "id": "UBERON:0002107",
      "label": "liver"
    },
    "timeOfCollection": {
      "age": "P48Y9M26D"
    },
```

Beacon+: Phenopackets

Testing alternative response schemas...

http://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/phenopackets

- the v2 default schemas are mostly aligned w/ Phenopackets v2
- creating phenopackets can be done mostly by re-wrapping of Beacon entities (individual, biosample)
- variants can be included through file resource URLs; in Beacon+ this is done through *ad hoc* handover URIs


```
"id": "pgxpxf-kftx3tl5",
"metaData": {
  "phenopacketSchemaVersion": "v2",
  "resources": [
    {
      "id": "NCIT",
      "iriPrefix": "http://purl.obolibrary.org/obo/NCIT_",
      "name": "NCIt Plus Neoplasm Core",
      "namespacePrefix": "NCIT",
      "url": "http://purl.obolibrary.org/obo/ncit/neoplasm-core.owl",
```

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

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

```
      },
      "onset": {
        "age": "P48Y9M26D"
      },
      "stage": {
        "id": "NCIT:C27966",
        "label": "Stage I"
      }
    }
  ],
  "id": "pgxind-kftx3tl5",
  "sex": {
    "id": "PATO:0020001",
    "label": "male genotypic sex"
  },
  "updated": "2018-12-04 14:53:11.674000",
  "vitalStatus": {
    "status": "UNKNOWN_STATUS"
  }
}
```

```
    {
      "fileAttributes": {
        "fileFormat": "pgxseg",
        "genomeAssembly": "GRCh38"
      },
      "uri": "https://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/variants/?output=pgxseg"
    }
  ],
  "histologicalDiagnosis": {
    "id": "NCIT:C3099",
    "label": "Hepatocellular Carcinoma"
  },
  "id": "pgxbs-kftvhyvb",
  "individualId": "pgxind-kftx3tl5",
  "pathologicalStage": {
    "id": "NCIT:C27966",
    "label": "Stage I"
  },
  "sampledTissue": {
    "id": "UBERON:0002107",
    "label": "liver"
  },
  "timeOfCollection": {
    "age": "P48Y9M26D"
  },
```

Beacon+: Phenopackets

Testing alternative response schemas...

http://progenetix.org/beacon/biosamples/pgxbs-kftvhyvb/phenopackets

- the v2 default schemas are mostly aligned w/ Phenopackets v2
- creating phenopackets can be done mostly by re-wrapping of Beacon entities (individual, biosample)
- variants can be included through file resource URLs; in Beacon+ this is done through *ad hoc* handover URIs

```
bios_s = data_db["biosamples"].find({"individual_id":ind["id"]})

for bios in bios_s:

    bios.update({
        "files": [
            {
                "uri": "{}/beacon/biosamples/{}/variants/?output=pgxseg".format(server, bios["id"]),
                "file_attributes": {
                    "genomeAssembly": "GRCh38",
                    "fileFormat": "pgxseg"
                }
            }
        ]
    })
    for k in bios_pop_keys:
        bios.pop(k, None)

    clean_empty_fields(bios)

    pxf_bios.append(bios)
```

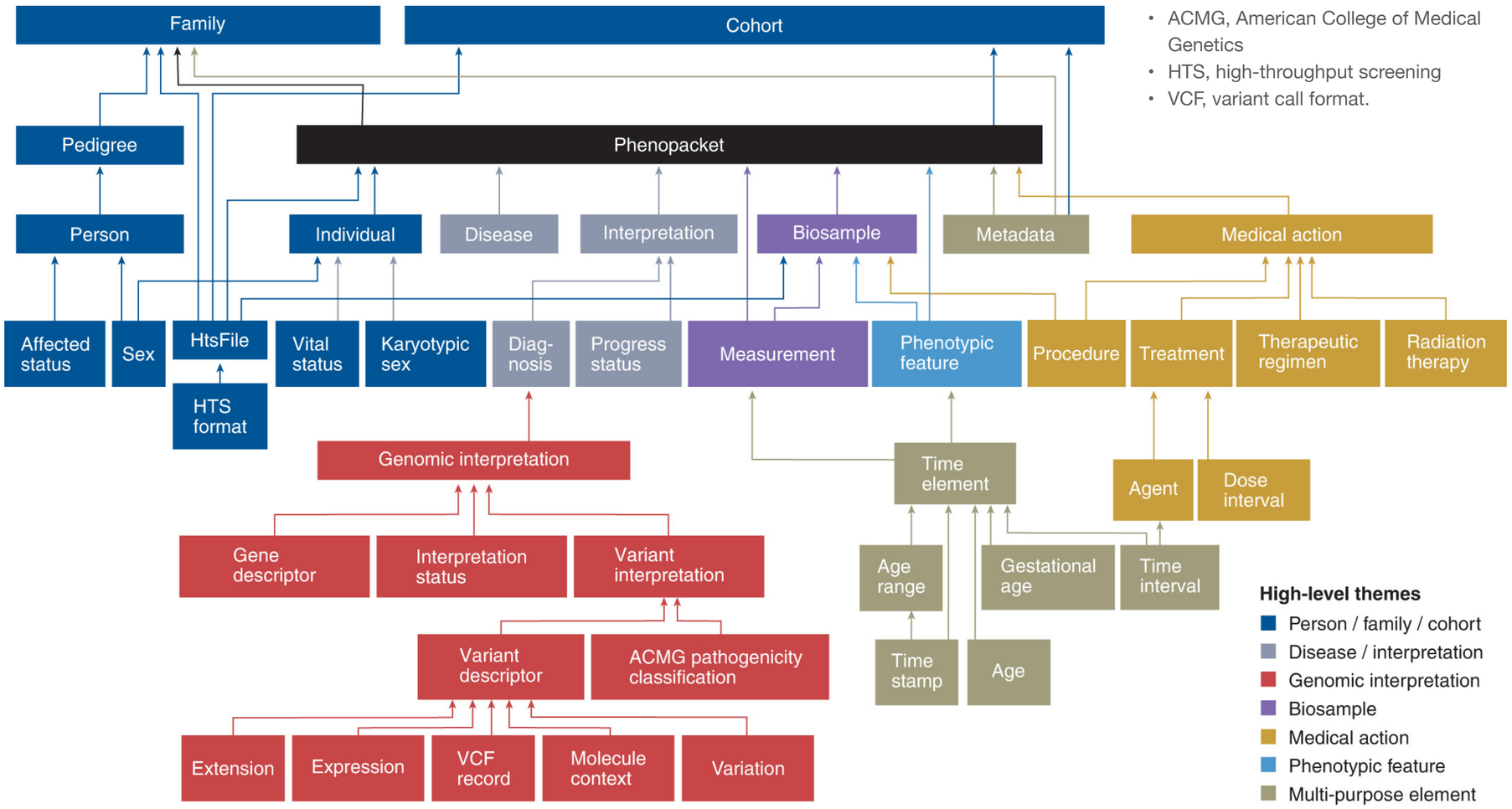
```
def remap_phenopackets(ds_id, r_s_res, byc):

    if not "phenopacket" in byc["response_entity_id"]:
        return r_s_res

    mongo_client = MongoClient()
    data_db = mongo_client[ds_id]
    pxf_s = []

    for ind_i, ind in enumerate(r_s_res):

        pxf = phenopack_individual(ind, data_db, byc)
        pxf_s.append(pxf)

    return pxf_s
```

The GA4GH Phenopackets v2 Standard

A Computable Representation of Clinical Data

The GA4GH Phenopacket schema consists of several optional elements, each containing information about a certain topic, such as phenotype, variant or pedigree. An element can contain other elements, which allows a hierarchical representation of data.

For instance, Phenopacket contains elements of type *Individual*, *PhenotypicFeature*, *Biosample* and so on. Individual elements can therefore be regarded as **building blocks** of larger structures.



- ACMG, American College of Medical Genetics
- HTS, high-throughput screening
- VCF, variant call format.

Jacobsen JOB, Baudis M, Baynam GS, Beckmann JS, Beltran S, Buske OJ, Callahan TJ, *et al.* 2022.
The GA4GH Phenopacket Schema Defines a Computable Representation of Clinical Data.
Nature Biotechnology 40 (6): 817–20.

GA4GH {S}[B]

SchemaBlocks

- “cross-workstreams, cross-drivers” initiative to document GA4GH object standards and prototypes, data formats and semantics
- launched in December 2018
- documentation and implementation examples provided by GA4GH members
- no attempt to develop a rigid, complete data schema
- object vocabulary and semantics for a large range of developments
- currently not “authoritative GA4GH recommendations”
- recognized in GA4GH roadmap as element in "TASC" effort

schemablocks.org

Future?

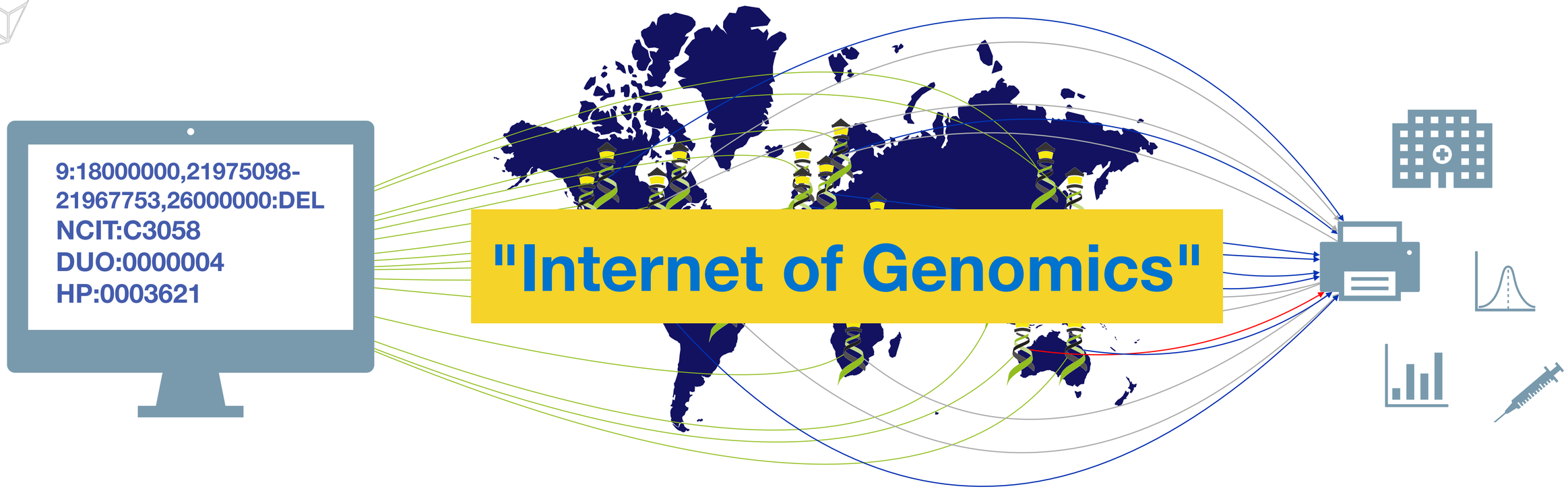
Some proposals for a stepwise Beacon protocol extension

- Query language expansion, e.g. Boolean options for chaining filters

  ➡use of heterogeneous/alternative annotations within and across resources

- **Phenopackets** support as a (the?) default format for biodata export
- **Phenopackets** as **request** documents
- Focus on service & **resource discovery**
- **ELIXIR Beacon Network**, including translations for federated queries to Beacon and Beacon-like resources


FUTURE



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

Beacon *v2* API

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

Progenetix & Beacon+

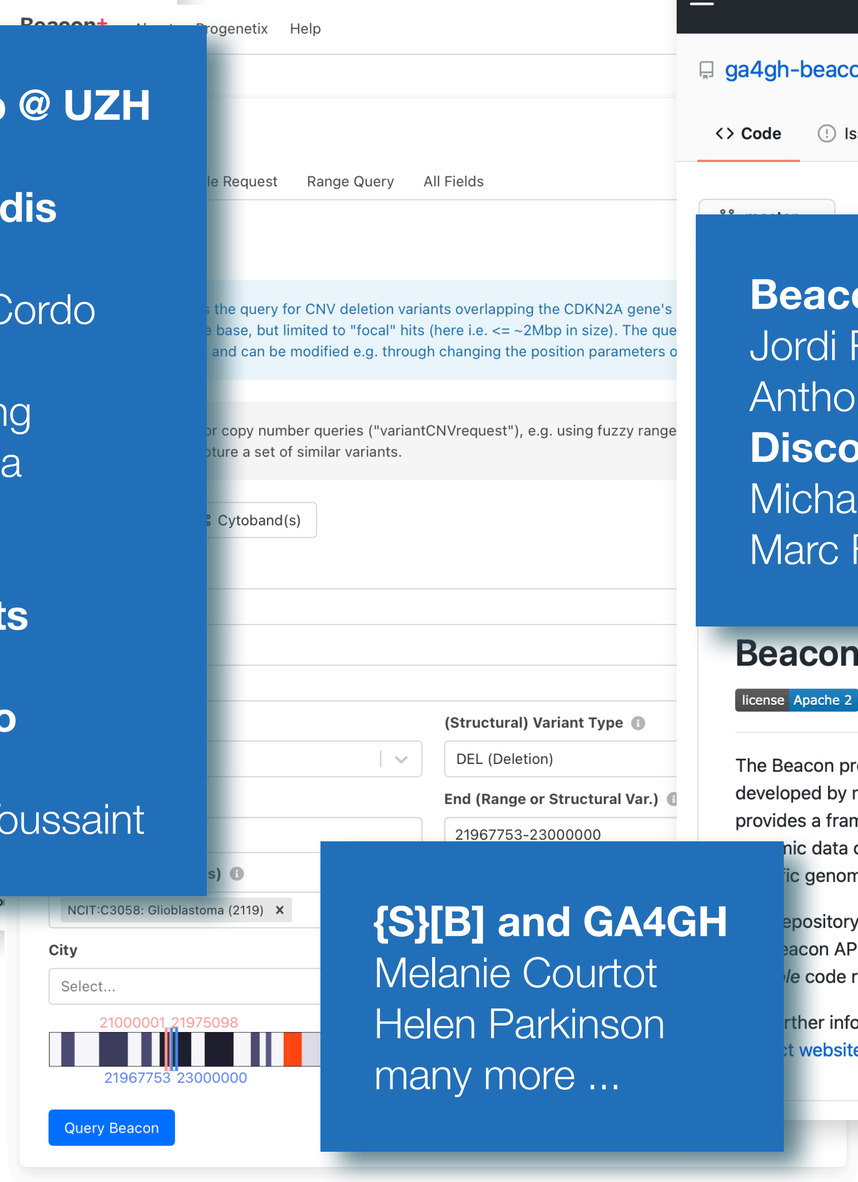
A cancer genomics reference resource powered by GA4GH standards

- Copy number variations constitute a complex, exciting and still poorly understood research topic in cancer and rare disease genomics
- Progenetix is the largest public resource for CNV in cancers (and increasingly reference genomes)
- The complexity of inherited and somatic genomic variations requires data access beyond individual resources => **Federated Data Access**
- The Global Alliance for Genomics and Health (GA4GH) is a policy-framing and technical standards-setting organization with focus on genomic data sharing
- Beacon v2 is the main GA4GH data discovery and sharing protocol, developed with support from the European bioinformatics organization ELIXIR
- Progenetix serves as a testbed for the early implementation of GA4GH standards such as Beacon extensions, Phenopackets and VRS

Baudisgroup @ UZH
Ni Ai
Michael Baudis
Haoyang Cai
Paula Carrio Cordo
Bo Gao
Qingyao Huang
Saumya Gupta
Nitin Kumar
Sofia Pfund
Rahel Paloots
Ziying Yang
Hangjia Zhao

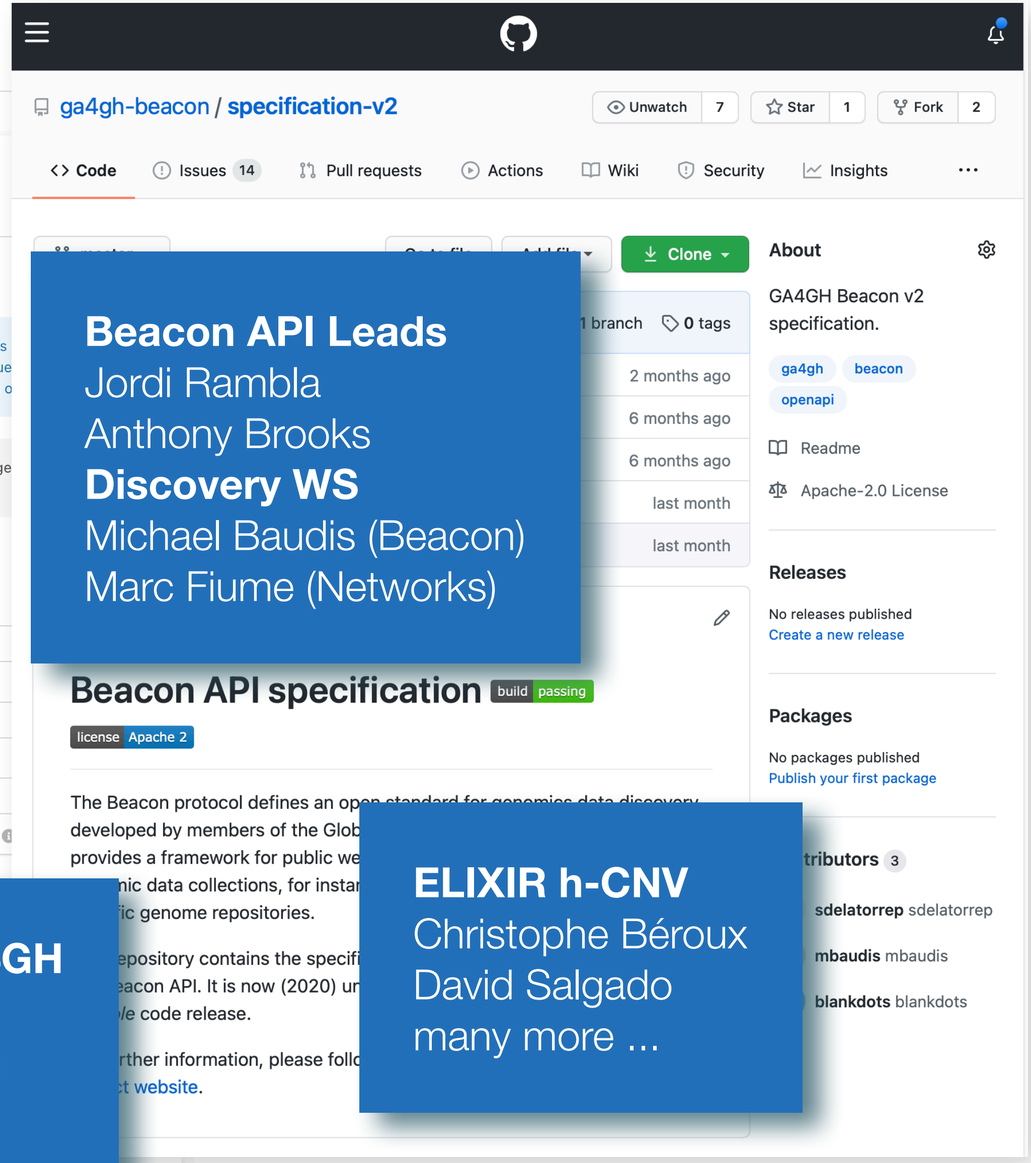
Pierre-Henri Toussaint

beacon-project.io



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

beacon.progenetix.org/beaconPlus/



Beacon API Leads
Jordi Rambla
Anthony Brooks
Discovery WS
Michael Baudis (Beacon)
Marc Fiume (Networks)

ELIXIR h-CNV
Christophe Béroux
David Salgado
many more ...

github.com/ga4gh-beacon/


HIER WOHNTE
V. 21. FEBR. 1916 BIS 2. APRIL 1917
LENIN
DER FÜHRER DER RUSSISCHEN
REVOLUTION