A Reference Resource for Copy Number Variations in Cancer

Implementing GA4GH Standards to Drive an Open Oncogenomics Resource

Michael Baudis | Seminar KiSpi | 2021-03-29

progenetix

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 @ IMLS (2015) | systematic assembly of oncogenomic data | databases and software tools | patterns in cancer genomes | *Progenetix* & *arrayMap* resources | GA4GH | SPHN

Genome screening at the core of “Personalised Health”

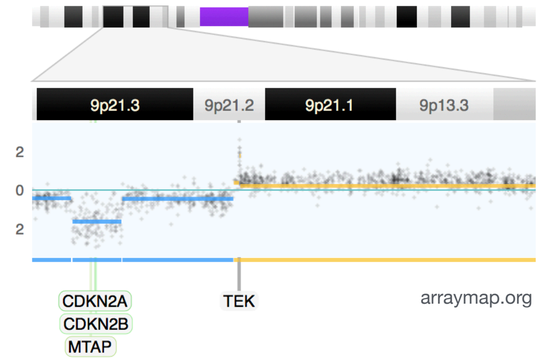
- **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**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

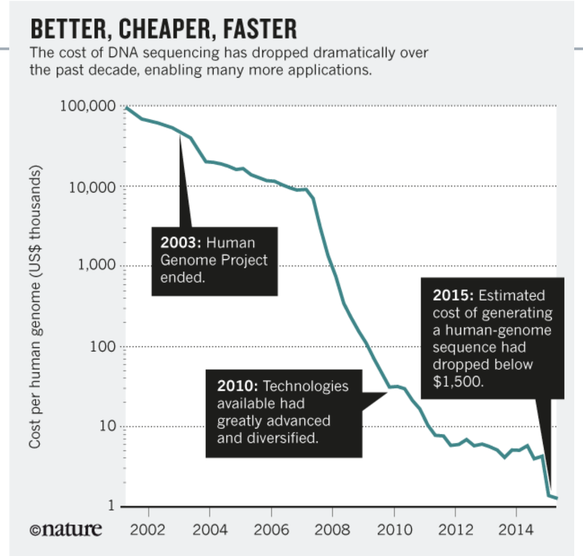
arrayMap

progenetix

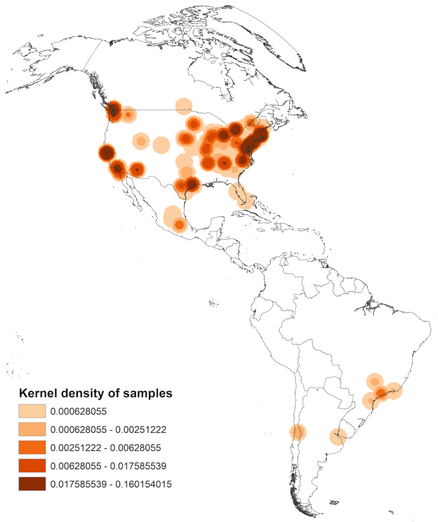
elixir

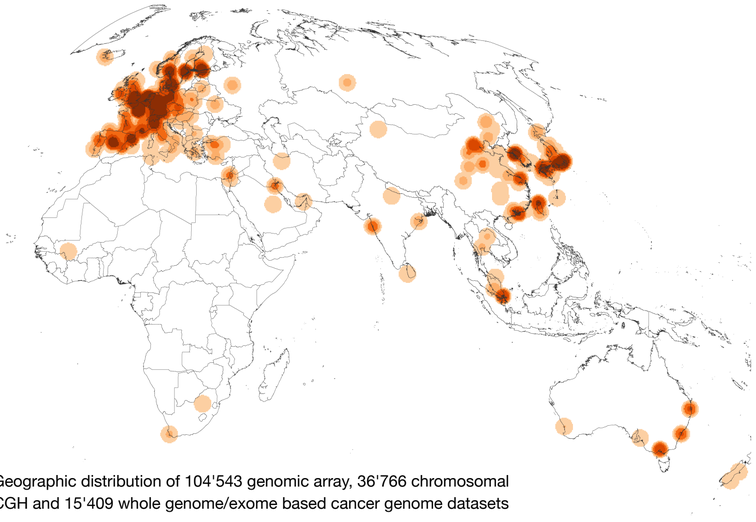
Global Alliance for Genomics & Health





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



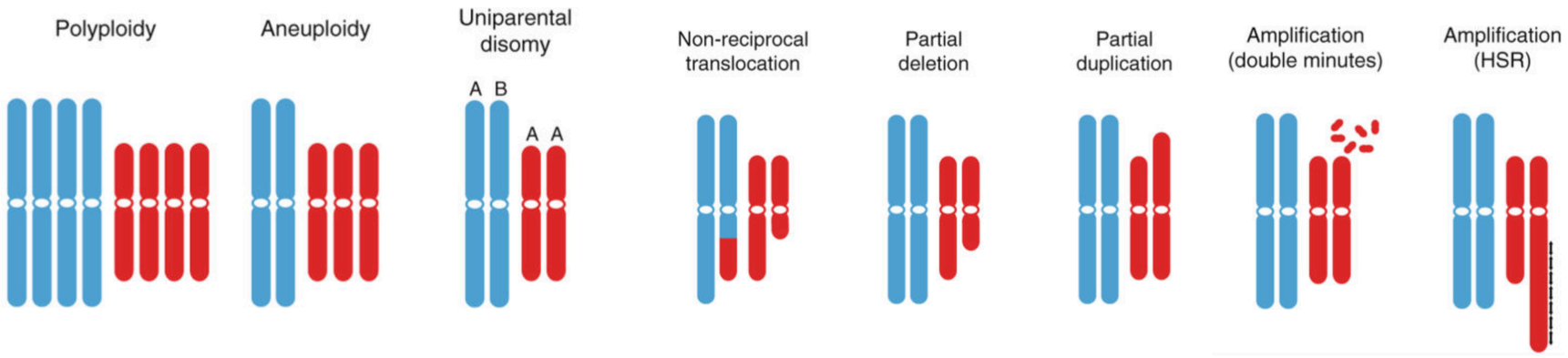


Geographic distribution of 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets

Types of genomic alterations in Cancer

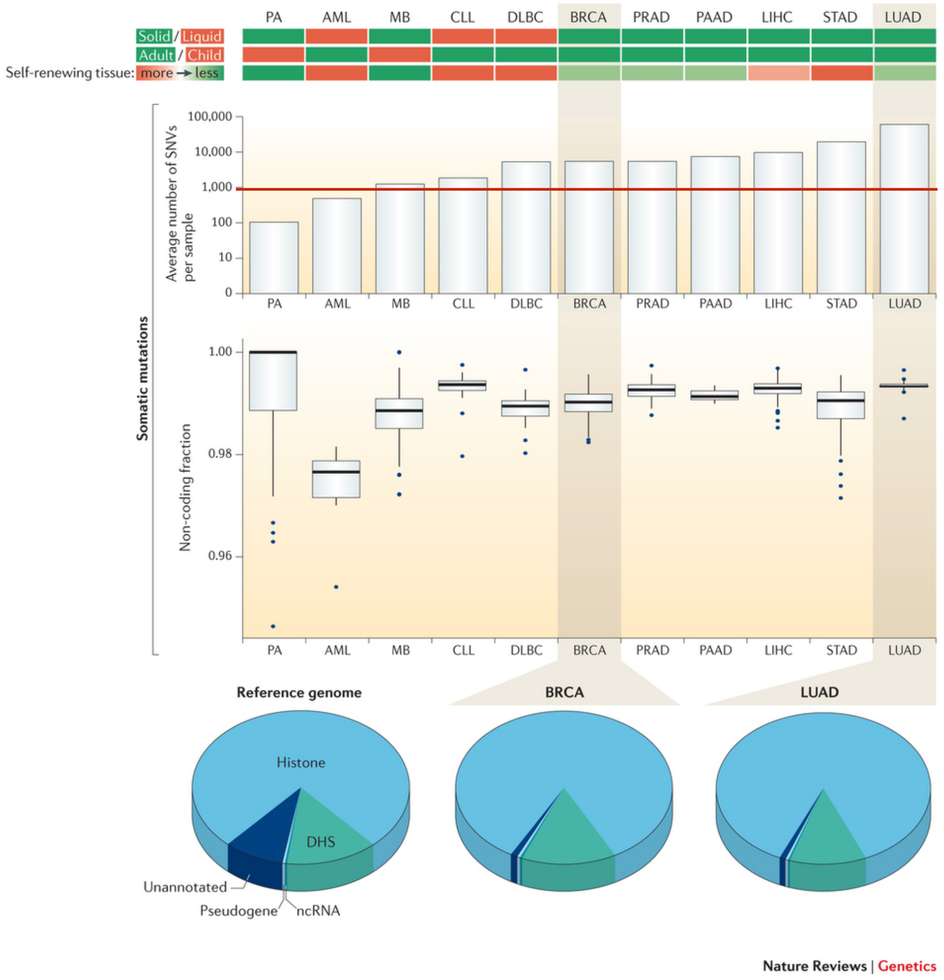
Imbalanced Chromosomal Changes: CNV

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

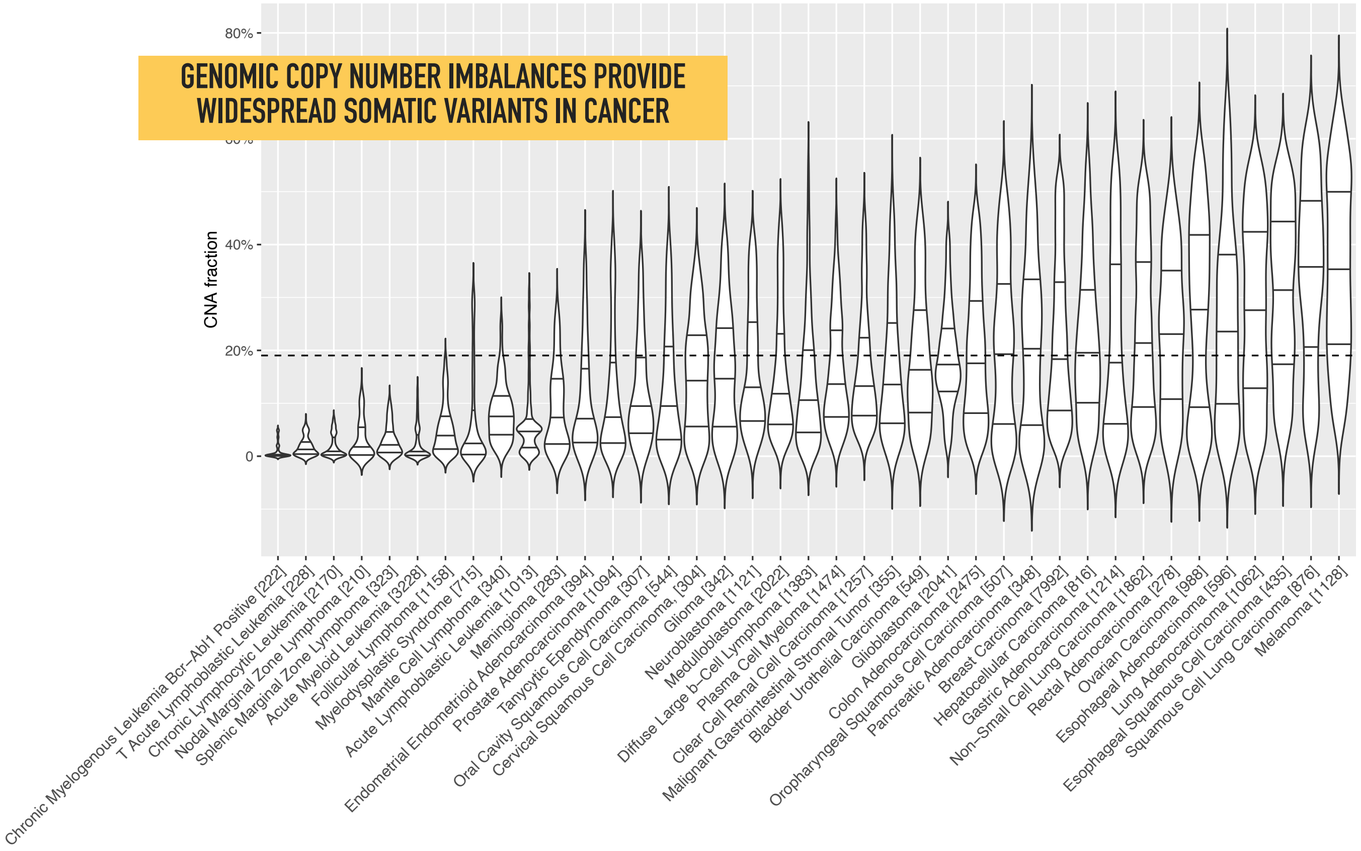


Grade et al., 2015 Recent Results Cancer Res

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)

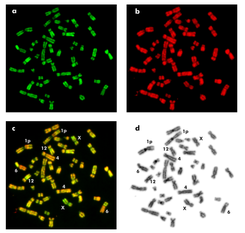
On average ~19% of a cancer genome are in an imbalanced state (more/less than 2 alleles); Original data based on 43654 cancer genomes from

History & Current State...

Origins & trajectory of the Progenetix Resource

progenetix

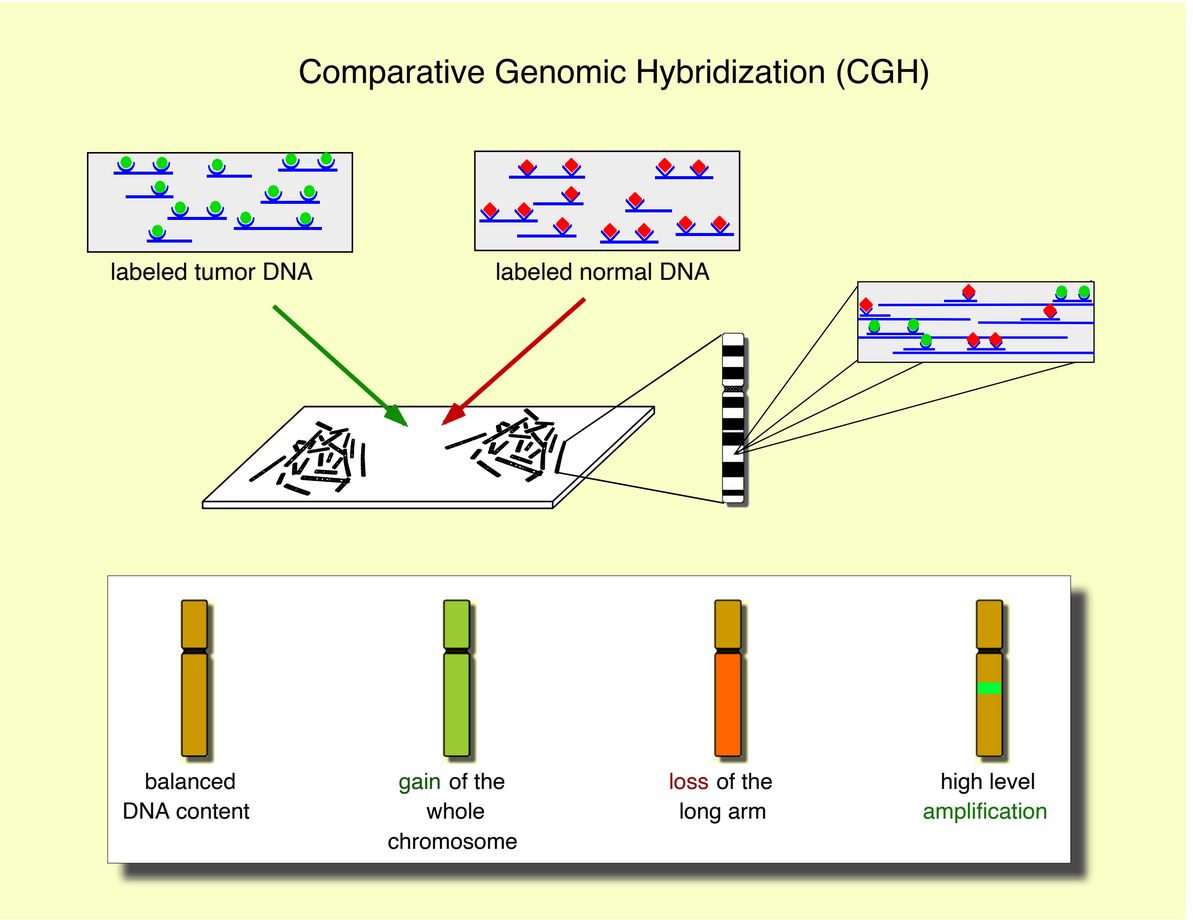
Comparative Genomic Hybridization



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.



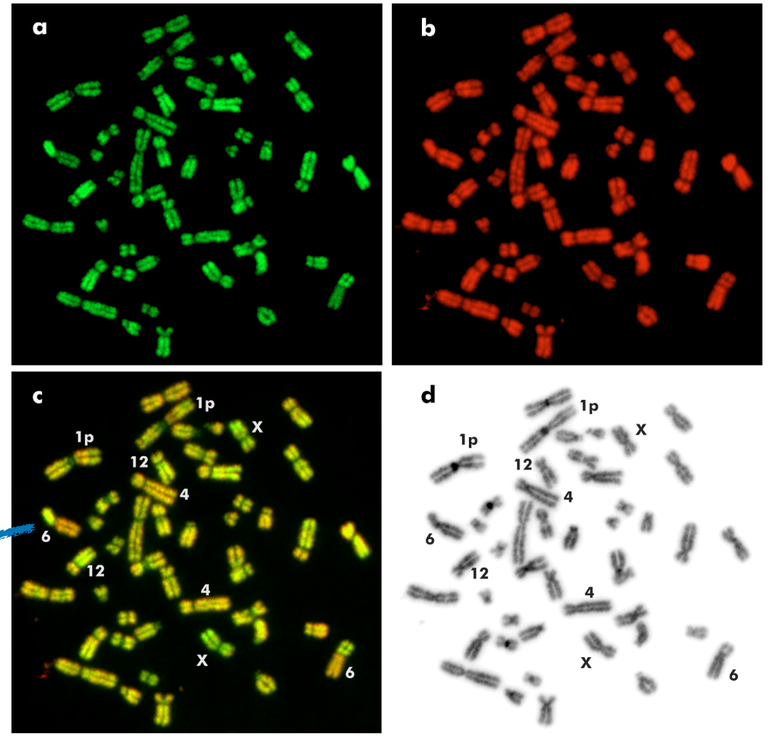
Chromosomal CGH: Normal metaphase spreads (cultured lymphocytes from healthy donors) on microscopy slides serve as the hybridization matrix for whole-genome DNA from tumor and reference tissue, labeled with different fluorophores. The regional ratio between the two colors points to (relative) changes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

Comparative Genomic Hybridization

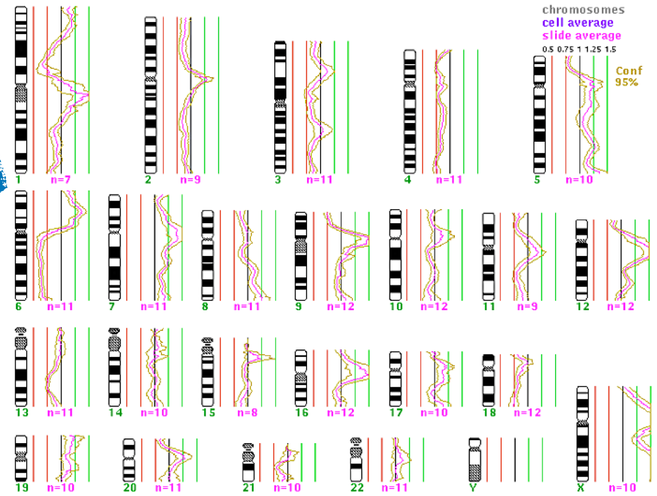
Molecular-Cytogenetic Technology for Genomic Imbalance Screening

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

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



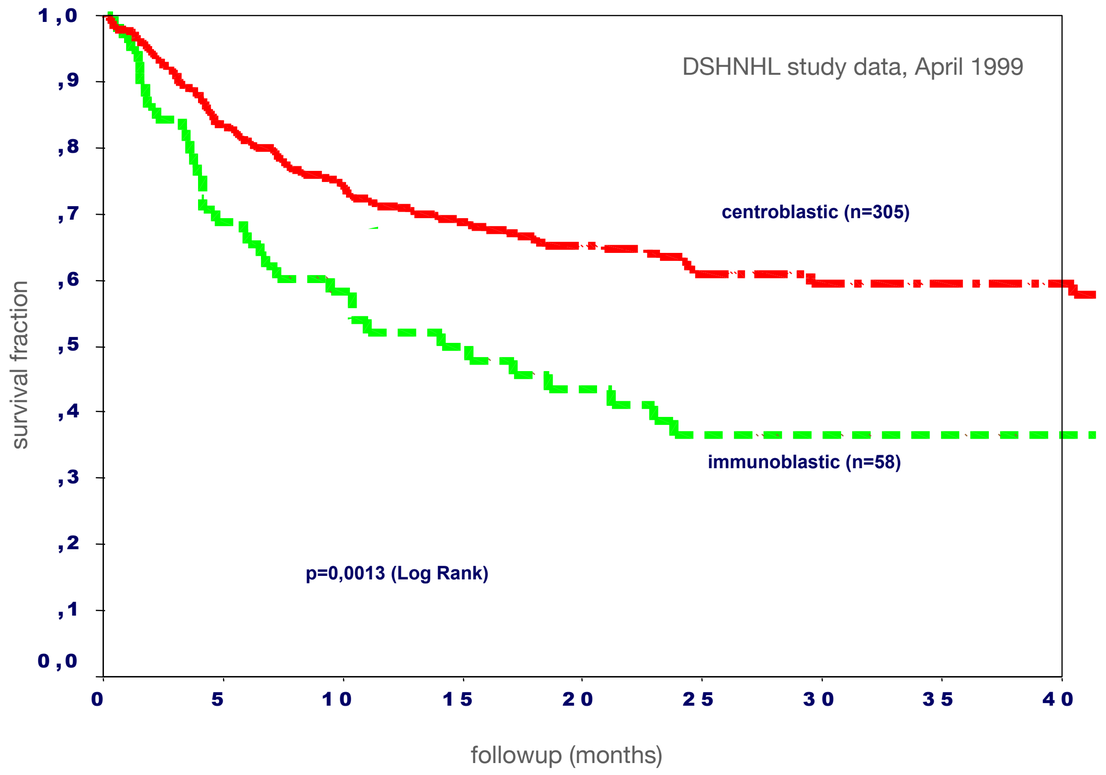
CGH-Experiment: **a** Hybridisierung mit Tumor-DNA; **b** Hybridisierung mit normaler menschlicher DNA als Kontrolle; **c** Überlagerung der Signale; **d** Bänderungsfärbung zur Identifizierung der Chromosomen



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

Cancer CNVs | Diagnostics | Prognosis

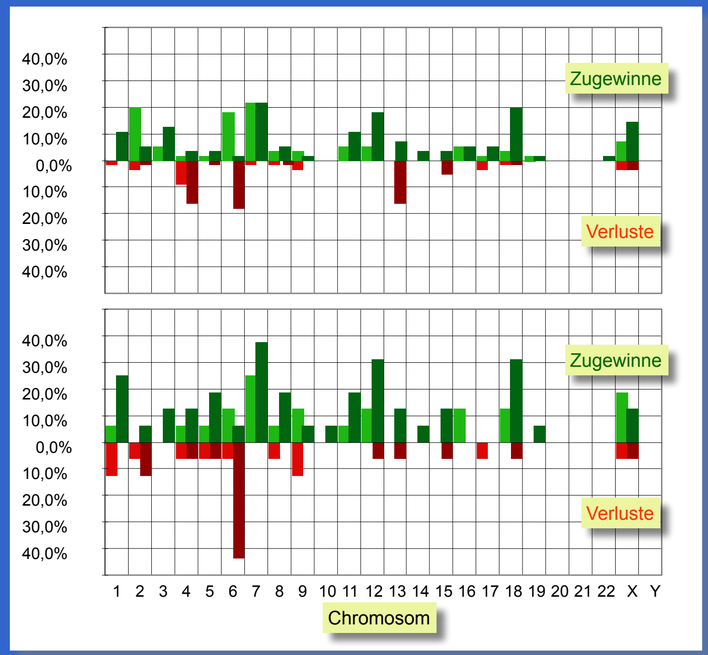
Single-study CNV frequencies correspond to diagnostic subsets



55 cb NHL

Chromosomale Imbalancen

16 ib NHL



| Tuesday February 27, 1996

Let's build a database!

... using archaic tools

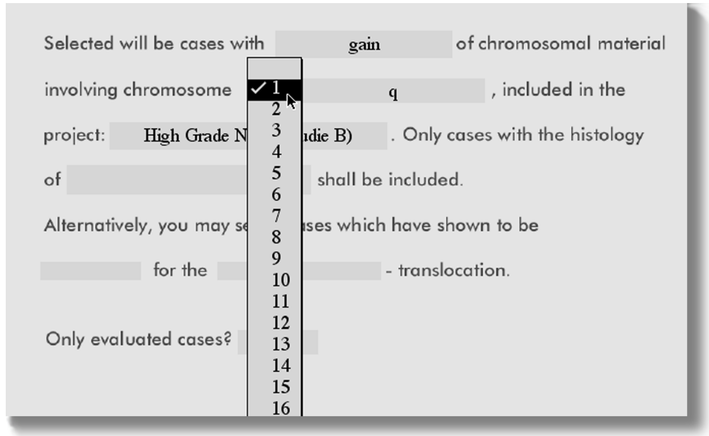
| Tuesday September 08, 1992

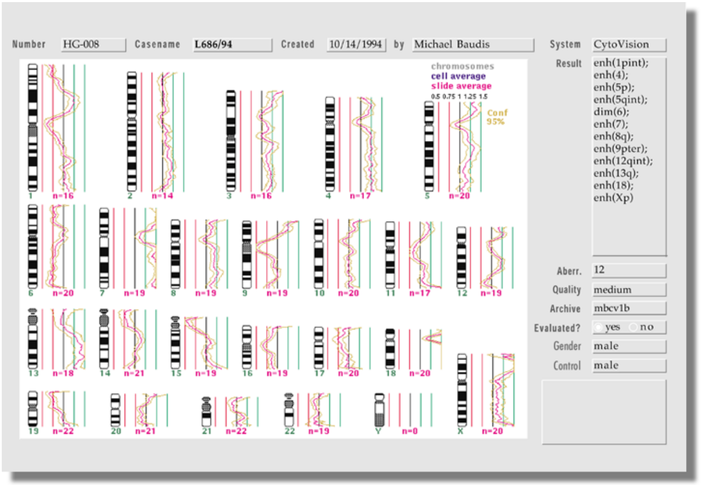
Progenetix CGH Database and Website

HEIDELBERG UNIVERSITY HOSPITAL

dkfz.

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






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

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

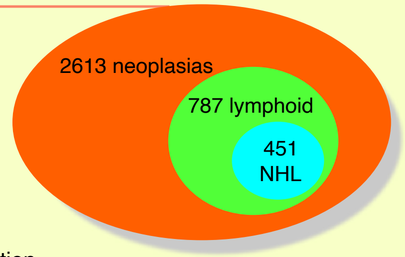
progenetix.net

[progenetix.net] online CGH database

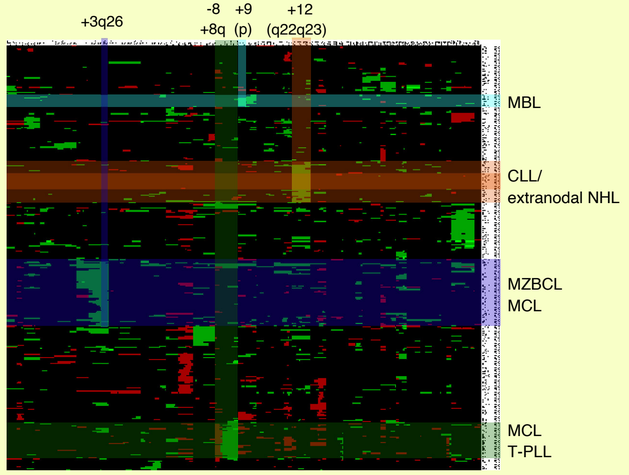
A compilation of published CGH analyses with reported case and band specific results

Currently includes
2613 cases from
92 publications

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



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



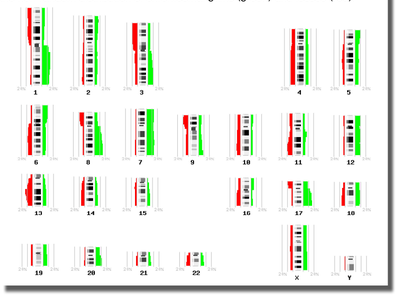
progenetix.net

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

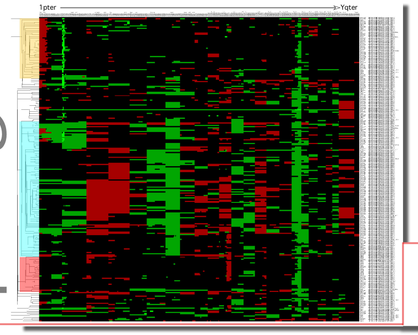
michael baudis, dept. of pathology, stanford university

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

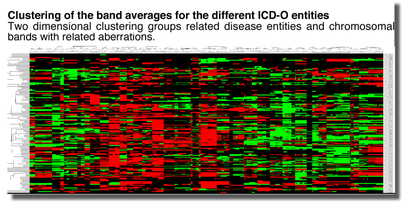
Chromosomal imbalances in 5478 clinical cases from 196 publications
Although not as prominent as in specific subgroups, this large collection shows the non-random distribution of chromosmal gains (green) and losses (red).



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

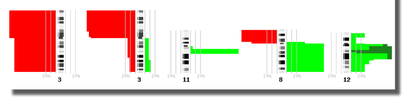


Clustering of the band averages for the different ICD-O entities
Two dimensional clustering groups related disease entities and chromosomal bands with related aberrations.



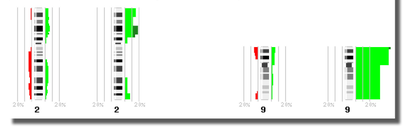
Results Out of 4896 tumor samples, 3862 (79%) showed chromosomal imbalances by CGH. The average per band probability was 4.5% for a loss (max. 12.9% at 13q21) and 6.5% for a gain (max. 15.6% at 8q23). Differences between neoplastic entities showed in the average frequency and distribution pattern of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

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



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

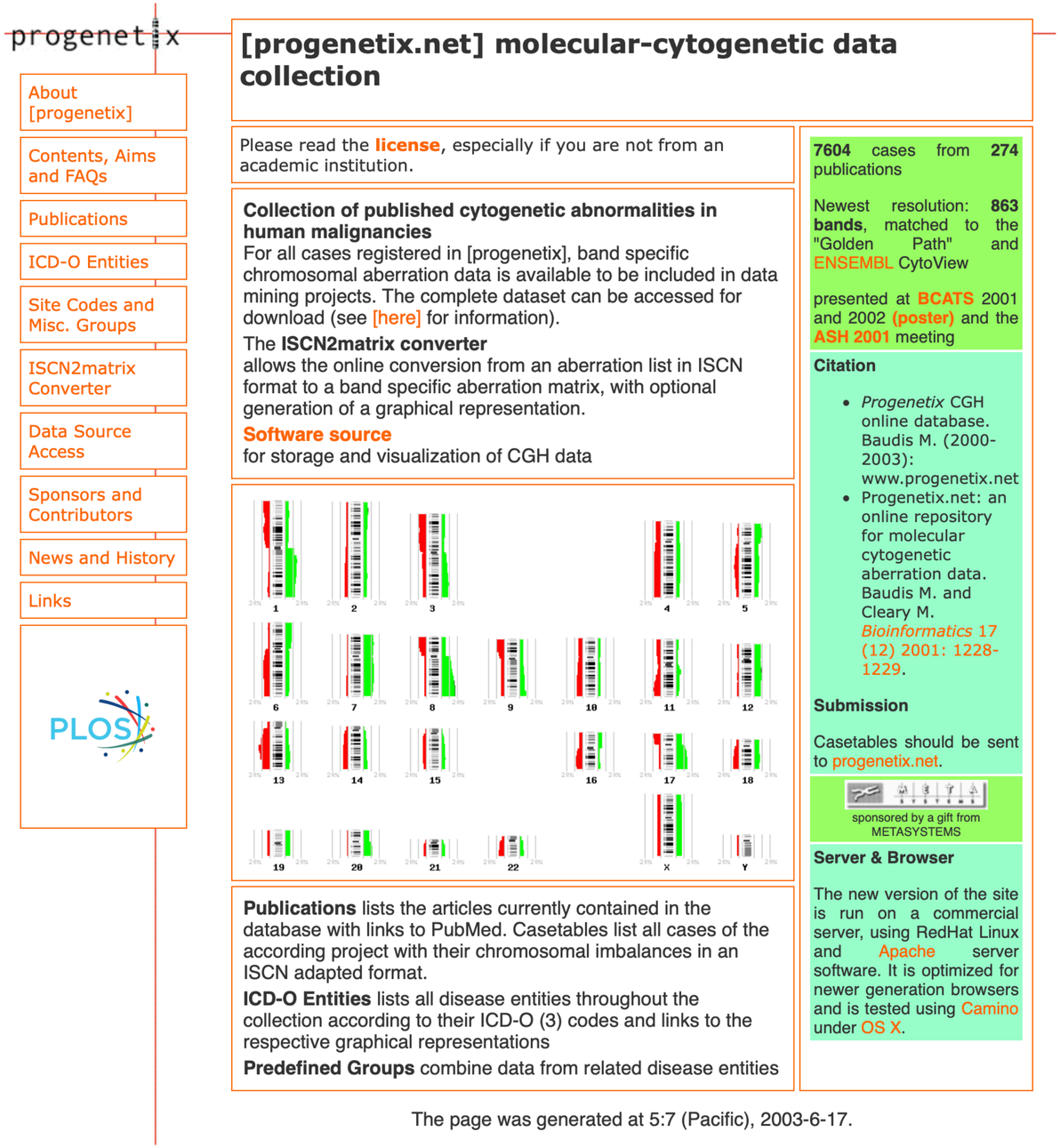
Distinction of histologically related through their chromosomal aberration pattern
Amplification of the REL locus on 2p16 and gain of 9p(ter) distinguishes primary mediastinal B-cell lymphomas (PMBL, right) from diffuse large cell lymphomas (DLCL, left). The distinction may have clinical implications



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

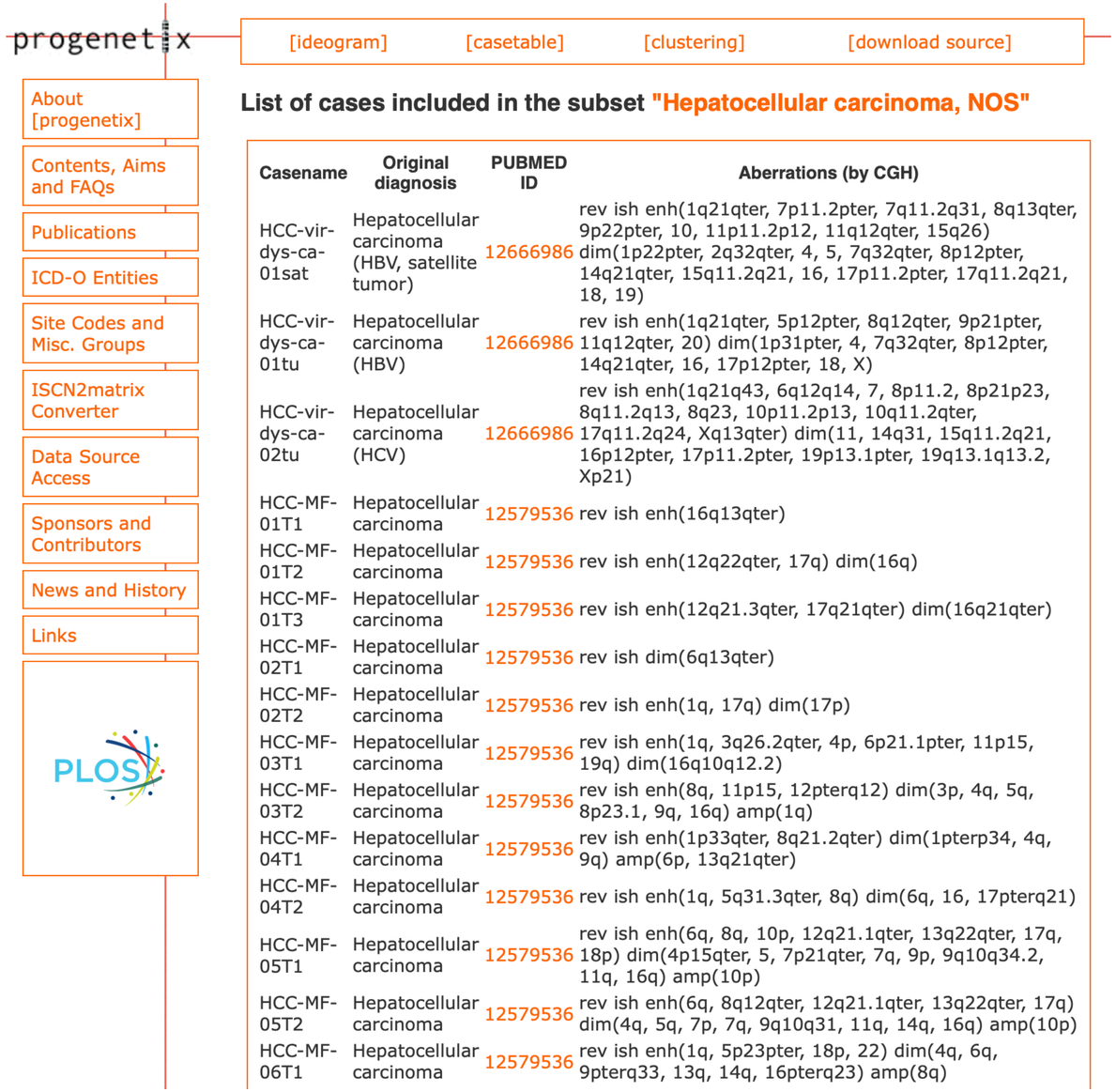
Progenetix Database in 2003
Text conversion for CNVs

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



Progenetix Database in 2003
Text conversion for CNVs

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



Progenetix Database in 2003
Text conversion for CNVs

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

progenetix

TABLE 3. Comparison of Primary Tumors and Metastases by CGH

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

TABLE 1. Clinical Data

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

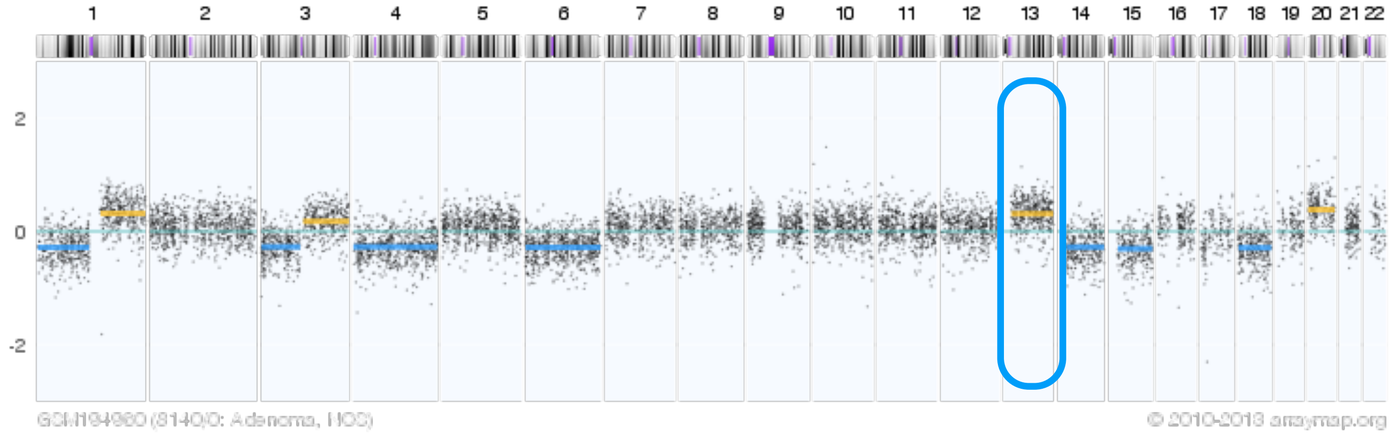
[a]AJCC/UICC staging system (Hutter and Sobin, 1986).
[b]Grade of primary tumor: 1–3, low, moderate, high grade; 9, grading unknown.
[c]Synchronous, diagnosis of metastatic disease within 12 months following diagnosis of primary tumor; metachronous, diagnosis of metastatic disease after 12 months or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

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

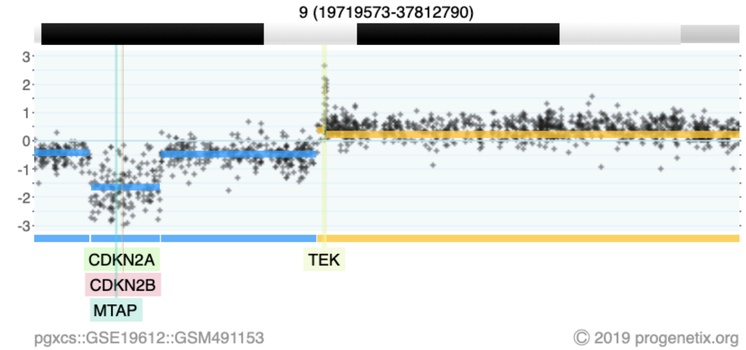
Chromosome Arm 20q Gains and Other Genomic Alterations in Colorectal Cancer Metastatic to Liver, as Analyzed by Comparative Genomic Hybridization and Fluorescence In Situ Hybridization

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

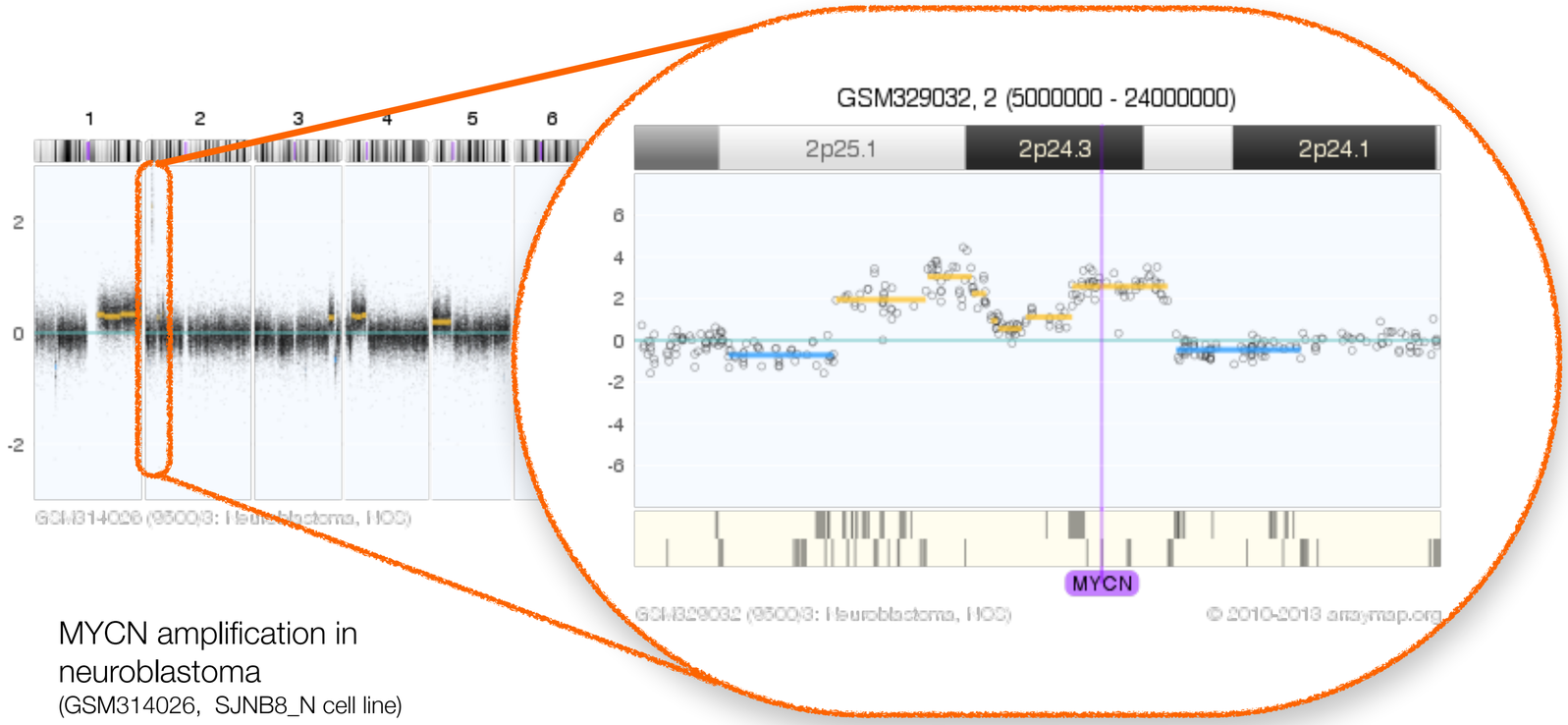
Array-based Detection of Copy Number Variations



Gain of chromosome arm 13q in colorectal carcinoma



2-event, homozygous deletion in a Glioblastoma



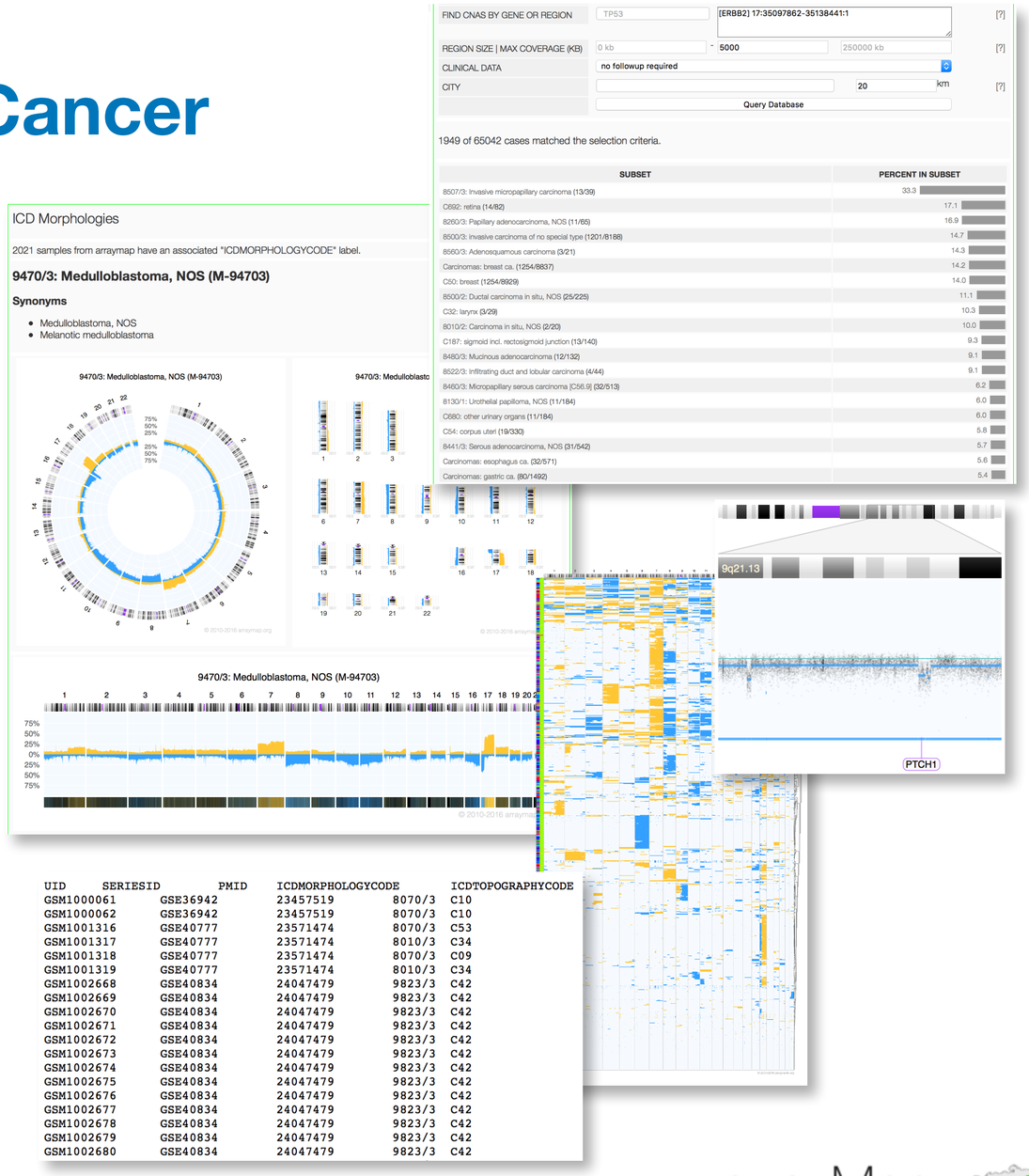
MYCN amplification in neuroblastoma
(GSM314026, SJNB8_N cell line)

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

arrayMap

arrayMap (2012 - 2020)

Probe-Level Genomic Array Data in Cancer



arrayMap

- Search Samples
- Search Publications
- Progenetix

University of Zurich[UZH]

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

CC BY-SA

Tweet

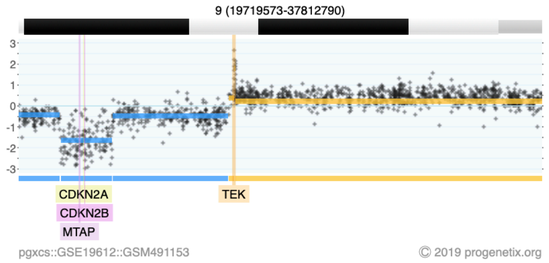
162.158.150.56

visualizing cancer genome array data @ arraymap.org

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

The current data reflects:

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



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

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

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

RELATED PUBLICATIONS

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

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

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

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

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

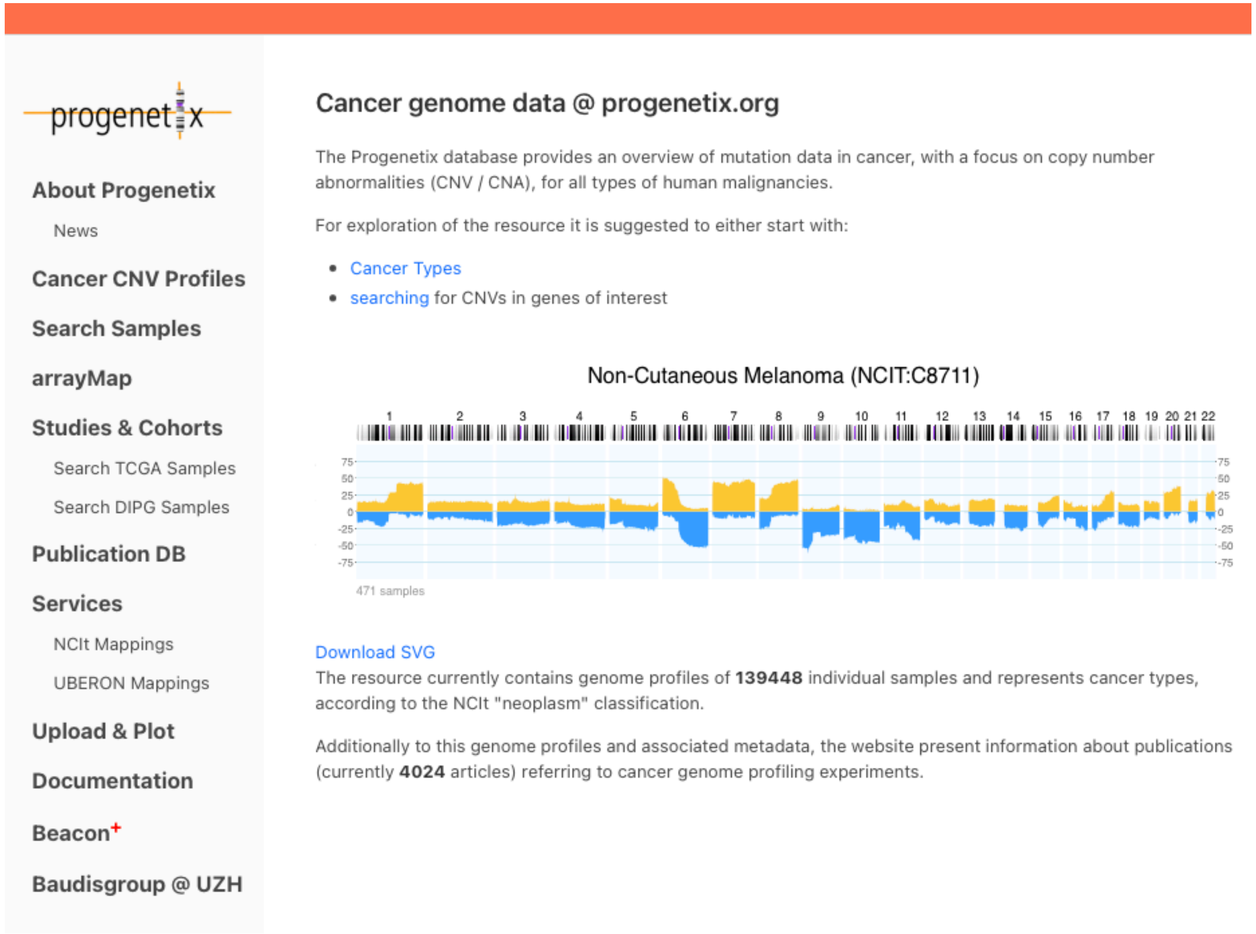
Feel free to use the data and tools for academic research projects and other applications. If more support and/or custom analysis is needed, please contact Michael Baudis regarding a collaborative project.

arrayMap

Progenetix in 2021

Cross-platform Oncogenomics

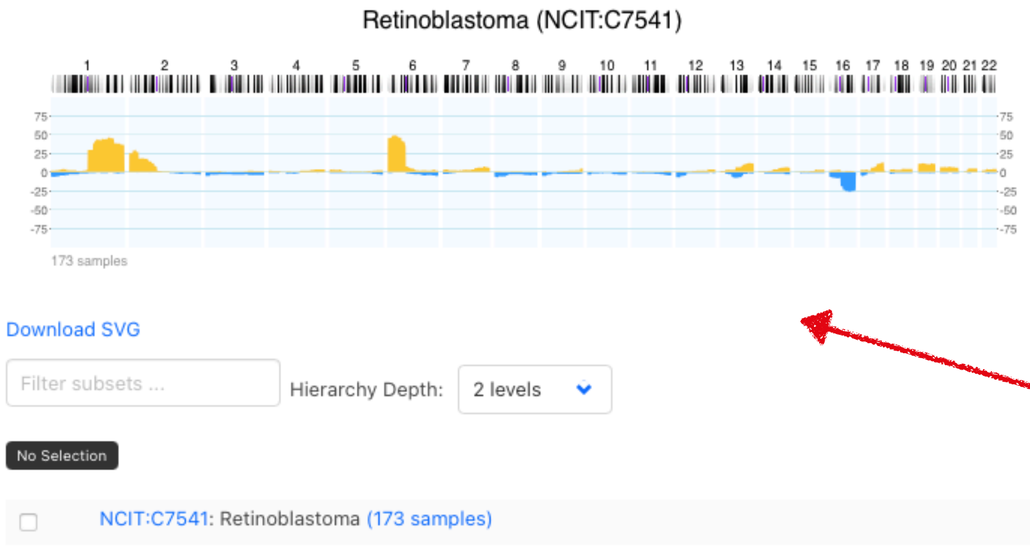
- merging of arrayMap (i.e. probe access enabled) and annotation derived (aCGH, WGS, WES, other arrays) data
- 115’357 cancer CNA profiles
- systematic metadata annotations following GA4GH standards
- unrestricted access w/o registration
- data access API
- online visualization
- CNA statistics



Progenetix

Cancer Type CNA Data

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



progenetix

- About Progenetix
  - News
- Cancer CNV Profiles
- Search Samples
- arrayMap
- Studies & Cohorts
  - Search TCGA Samples
  - Search DIPG Samples
- Publication DB
- Services
  - NCIt Mappings
  - UBERON Mappings
- Upload & Plot
- Documentation
- Beacon+
- Baudisgroup @ UZH

Cancer Types

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

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

Cancer Classification: NCIT Cancer Core

Filter subsets ... Hierarchy Depth: 2 levels

No Selection

- NCIT:C3262: Neoplasm (116232 samples)
  - NCIT:C3263: Neoplasm by Site (109317 samples)
    - NCIT:C156482: Genitourinary System Neoplasm (16410 samples)
    - NCIT:C2910: Breast Neoplasm (15525 samples)
    - NCIT:C3010: Endocrine Neoplasm (3319 samples)
    - NCIT:C3030: Eye Neoplasm (280 samples)
    - NCIT:C3052: Digestive System Neoplasm (15194 samples)
    - NCIT:C3077: Head and Neck Neoplasm (3769 samples)
    - NCIT:C3268: Nervous System Neoplasm (16270 samples)
      - NCIT:C2963: Cranial Nerve Neoplasm (19 samples)
      - NCIT:C3321: Peripheral Nervous System Neoplasm (901 samples)
      - NCIT:C35562: Neuroepithelial, Perineurial, and Schwann Cell Neoplasm (11690 samples)
      - NCIT:C4788: Malignant Nervous System Neoplasm (11608 samples)
        - NCIT:C3571: Malignant Cranial Nerve Neoplasm (19 samples)
        - NCIT:C3716: Primitive Neuroectodermal Tumor (2213 samples)
        - NCIT:C4627: Malignant Central Nervous System Neoplasm (9110 samples)
          - NCIT:C4628: Malignant Neoplasm of the Meninges (63 samples)
          - NCIT:C4717: Anaplastic Ganglioglioma (1 sample)
          - NCIT:C4822: Malignant Glioma (5460 samples)
          - NCIT:C5114: Malignant Intracranial Neoplasm (3242 samples)
          - NCIT:C62332: Central Nervous System Carcinoma (30 samples)
          - NCIT:C6990: (3357 samples)
          - NCIT:C7541: Retinoblastoma (173 samples)

Click to retrieve samples for NCIT:C7541

progenetix

- About Progenetix
- Cancer CNV Profiles
- Search Samples
- Publication DB
- Services
- Upload & Plot
- Documentation
- Beacon+
- Baudisgroup @ UZH

Search Samples

CDKN2A Deletion Example | MYC Duplication | TP53 Del. in Cell Lines

This example shows the query for CNV deletion variants overlapping the CDKN2A gene's coding region with at least a single base, but limited to "highly focal" hits (here i.e. <= ~1Mbp in size). The query is against the Progenetix and arrayMap collections. It can be modified e.g. through changing the position parameters or diagnosis.

Gene Spans | Cytoband(s)

Reference name: 9

(Structural) Variant Type: DEL (Deletion)

Start or Position: 21500001-21975098

End (Range or Structural Var.): 21967753-22500000

Cancer Classification(s): NCIT:C3058: Glioblastoma (4358) ×

Biosample Type:

Filters:

Filter Logic: AND

City: Select...

Query Beacon

Search Samples

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

progenetix

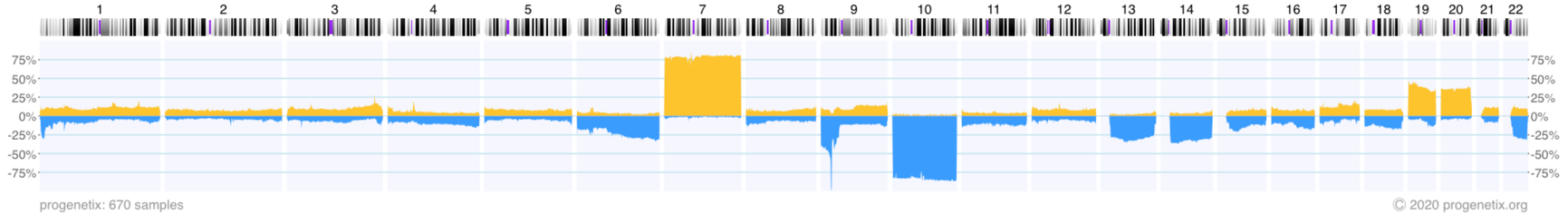
Samples: 668
Variants: 286
Calls: 675
$f_{alleles}$: 0.000088

Phenopackets
Callsets Variants
Variants in UCSC

UCSC region
JSON Response

Visualization options

Results | Biosamples | Biosamples Map | Variants



Subsets	Subset Samples	Query Matches	Subset Match Frequencies
icdot-C71.4	4	1	0.250
icdom-94403	4274	664	0.155
NCIT:C3058	4358	664	0.152
icdot-C71.1	14	2	0.143
icdot-C71.9	6684	651	0.097
NCIT:C3796	84	4	0.048
icdom-94423	84	4	0.048
icdot-C71.0	1712	14	0.008

Search Samples

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

progenetix

Samples: 668
Variants: 286
Calls: 675
$f_{alleles}$: 0.000088

Phenopackets
Callsets Variants
Variants in UCSC

UCSC region
JSON Response

Visualization options

Results | Biosamples | Biosamples Map | Variants

JSON | Download Response

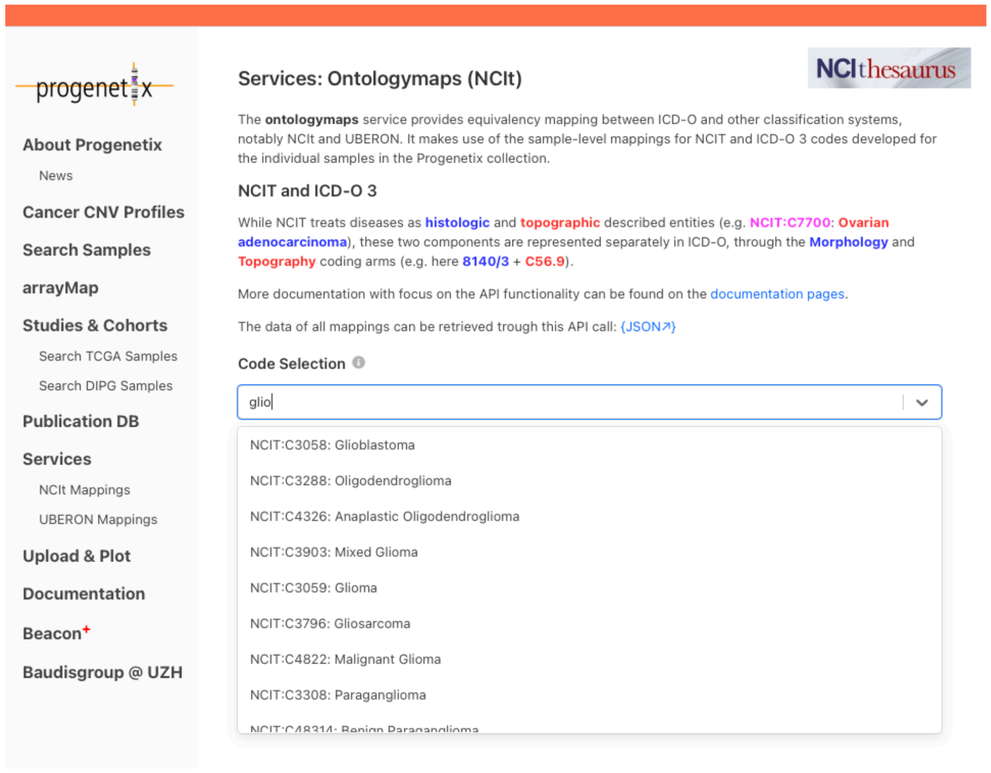
Int. ID	Digest	Callset	Biosample	Chr.	Ref. Base(s)	Alt. Base(s)	Type
5bab578b727983b2e00ca99e	9:21548871-21999595:DEL	pgxcs-kftvmlzx	pgxbs-kftvgk8h	9			DEL
5bab578d727983b2e00cb505	9:21958233-21999595:DEL	pgxcs-kftvmm5j	pgxbs-kftvgk90	9			DEL
5bab5793727983b2e00cdc18	9:21958233-21999595:DEL	pgxcs-kftvmmjj	pgxbs-kftvgka5	9			DEL
5bab5794727983b2e00ce2c6	9:21791897-21999595:DEL	pgxcs-kftvmmlu	pgxbs-kftvgkae	9			DEL
5bab5794727983b2e00ce49a	9:21958233-21999595:DEL	pgxcs-kftvmmmb	pgxbs-kftvgkaf	9			DEL

Page **1 of 135**

Progenetix

Services, Documentation...

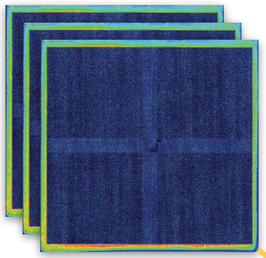
- services e.g. for disease code translation (NCIt <=> ICD-O; UBERON ...)
- API & documentation "progressing" ...

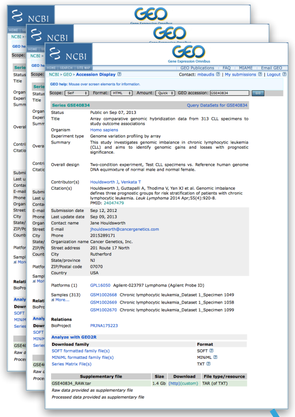


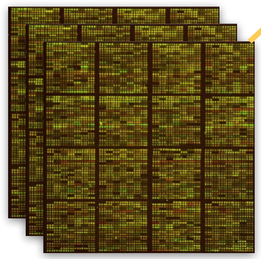
DATA PIPELINE

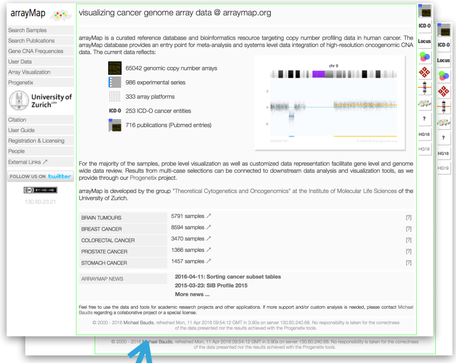
BIOCURATION

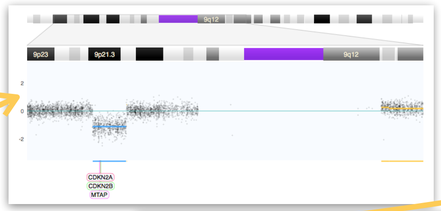
BIOINFORMATICS

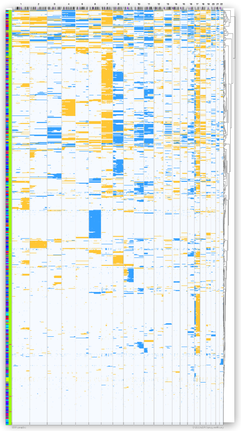


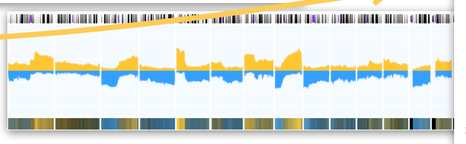


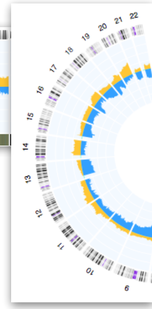


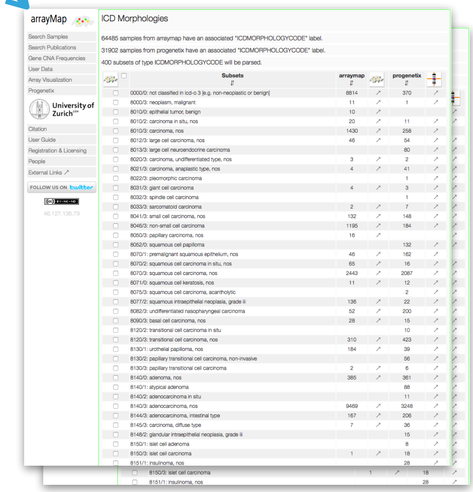








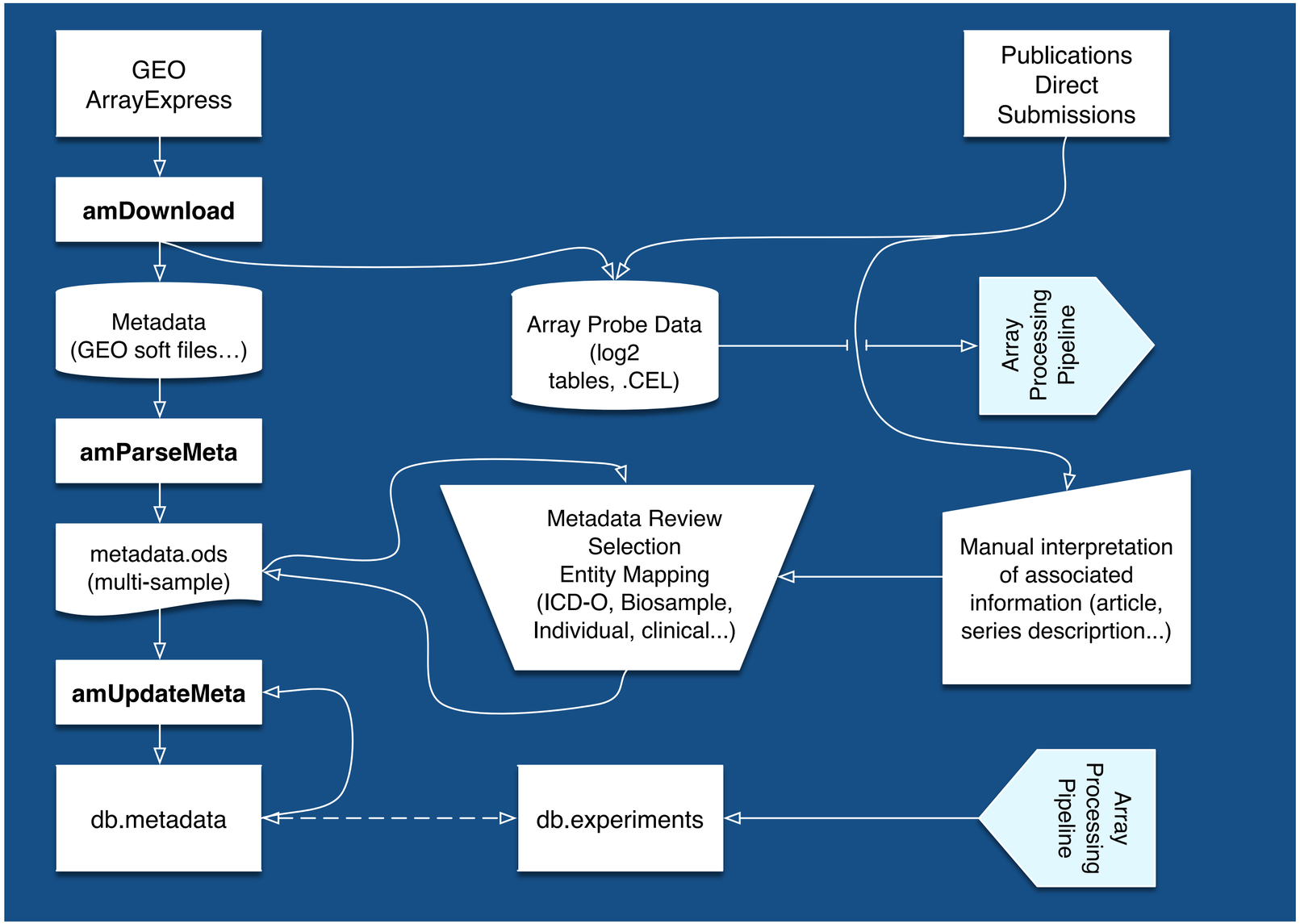




arrayMap

progenetix

Bioinformatics & Data Curation - arrayMap data “Pipeline”



Progenetix & arrayMap: Data Scopes

Biomedical and procedural "Meta"data types

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

progenetix

Data sets in tutorials

Data sets in the wild

Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

```
^SAMPLE = GSM174832
!Sample_title = 9194
!Sample_geo_accession = GSM174832
!Sample_status = Public on May 01 2007
!Sample_submission_date = Mar 13 2007
!Sample_last_update_date = Mar 13 2007
!Sample_type = genomic
!Sample_channel_count = 1
!Sample_source_name_ch1 = Bone marrow with 96% blasts
!Sample_organism_ch1 = Homo sapiens
!Sample_taxid_ch1 = 9606
!Sample_characteristics_ch1 = Immunotype: common ALL; Age: 9.2 yrs; Gender: F
!Sample_molecule_ch1 = genomic DNA
!Sample_extract_protocol_ch1 = QiaAmp purification kit (Qiagen)
!Sample_label_ch1 = biotin
!Sample_label_protocol_ch1 = Biotinylated DNA was prepared according to the standard Affymetrix protocol from 250 ng genomic DNA (Genechip Mapping 500k assay
manual 701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix).
!Sample_hyb_protocol = Hybridizations were performed according to the standard Affymetrix protocol from 250 ng genomic DNA (Genechip Mapping 500k assay manual
701930 Rev.3 or 100k assay manual 701684 Rev.3, Affymetrix) using an Affymetrix hybridisation oven 640 and an Affymetrix Fluidic station 450.
!Sample_scan_protocol = Scanning performed according to the standard Affymetrix protocol from 250 ng genomic DNA (Genechip Mapping 500k assay manual 701930 Rev.3
or 100k assay manual 701684 Rev.3, Affymetrix) using an Affymetrix scanner 3000.
!Sample_description = primary ALL diagnosis sample
!Sample_data_processing = copy number detection using CNAG2.0 software (http://www.genome.umin.jp/)
!Sample_platform_id = GPL3718
!Sample_contact_name = Roland,P.,Kuiper
!Sample_contact_email = r.kuiper@antrg.umcn.nl, e.verwiel@antrg.umcn.nl
!Sample_contact_phone = +31243610868
!Sample_contact_fax = +31243668752
!Sample_contact_department = Human Genetics
!Sample_contact_institute = Radboud University Nijmegen Medical Centre
!Sample_contact_address = Geert Grooteplein 10
!Sample_contact_city = Nijmegen
!Sample_contact_zip/postal_code = 6525GA
!Sample_contact_country = Netherlands
!Sample_supplementary_file = ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM174nnn/GSM174832/suppl/GSM174832.CEL.gz
!Sample_supplementary_file = ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM174nnn/GSM174832/suppl/GSM174832.CHP.gz
!Sample_series_id = GSE7255
```

progenetix

Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

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

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

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

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

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

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

progenetix

Data Curation - Happy RegExing!

Extracting clinical and technical metadata from GEO SOFT file

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

Channel 1 is normal -> Cave value swap!

Gender or "chromosomal sex"?

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

Unknown way to express "alive"!

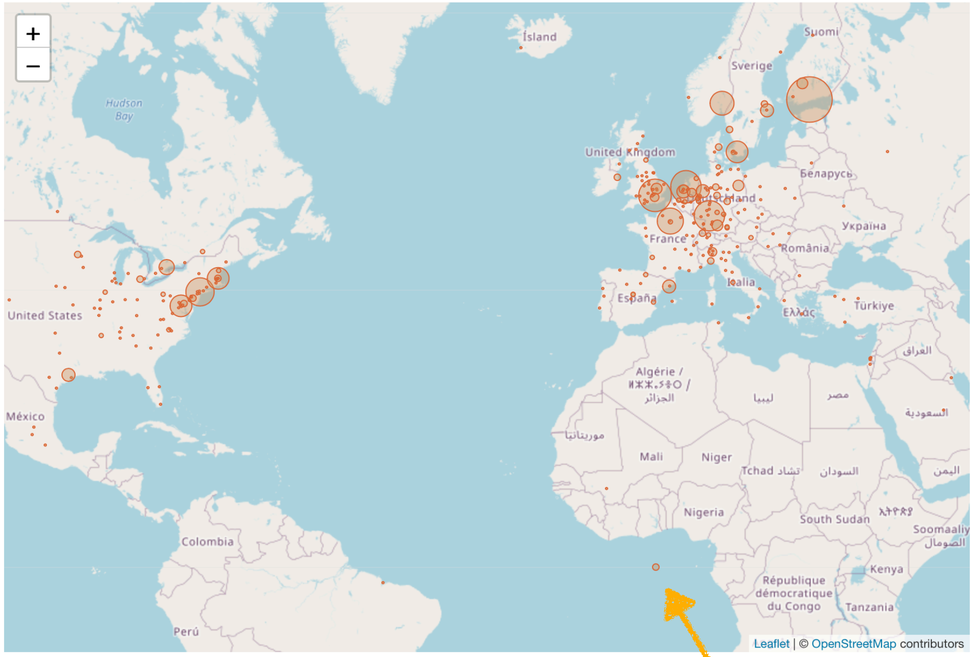
```
$mkey->{ samplekey } = 'DEATH';
$mkey->{ matches } = [ (
    'death',
    'dead ',
    'vital_status',
    'dead_alive',
    'alive_dead',
) ];

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

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

Data Curation

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

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

➡ always look for batch effects and outliers!



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

Progenetix publication collection

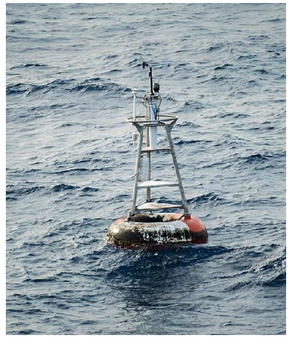
2020-11-28

25 / 3306 publications

Data Curation - Geolocations

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

The most geo-tagged place on earth is Null Island

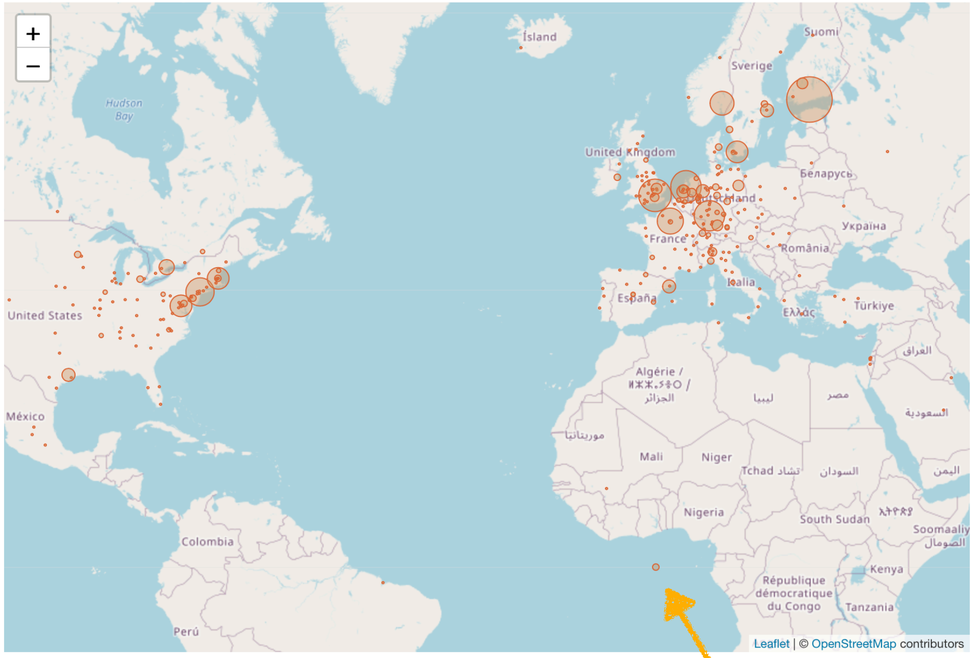


*A troubleshooting country has been added with an Indeterminate sovereignty class called **Null Island (1, 2)**. It is a fictional, 1 meter square island located off Africa where the equator and prime meridian cross. Being centered at 0,0 (zero latitude, zero longitude) it is useful for flagging geocode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

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

Michael Szell: The Data Science Process 2

2020-11-25



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

Progenetix publication collection

2020-11-28

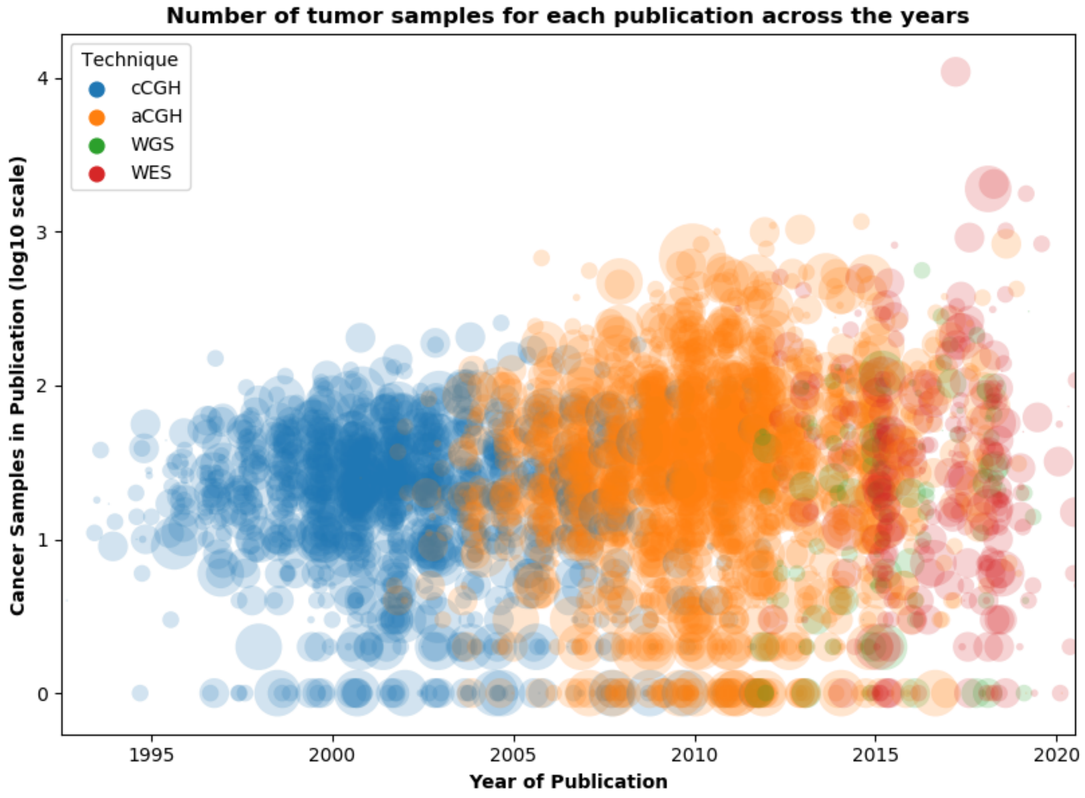
25 / 3306 publications

Standardized Data

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

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

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

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



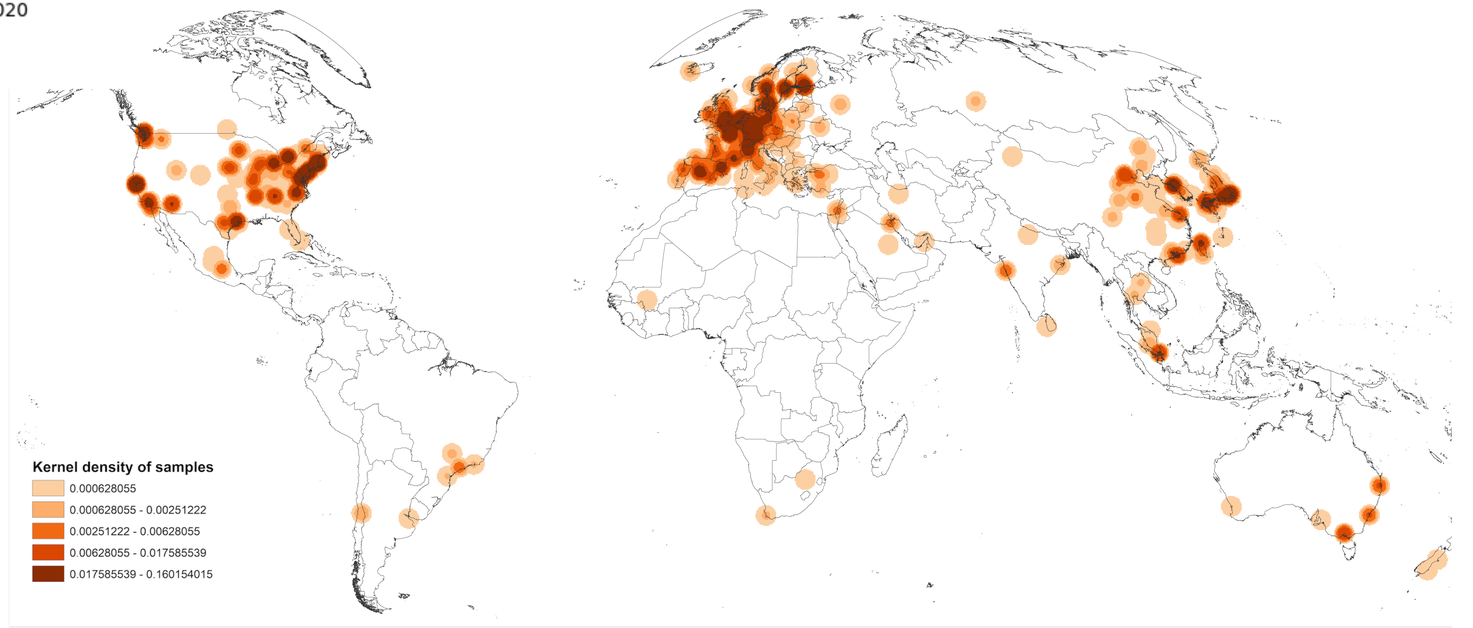
Publication Landscape of Cancer CNV Profiling

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

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

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

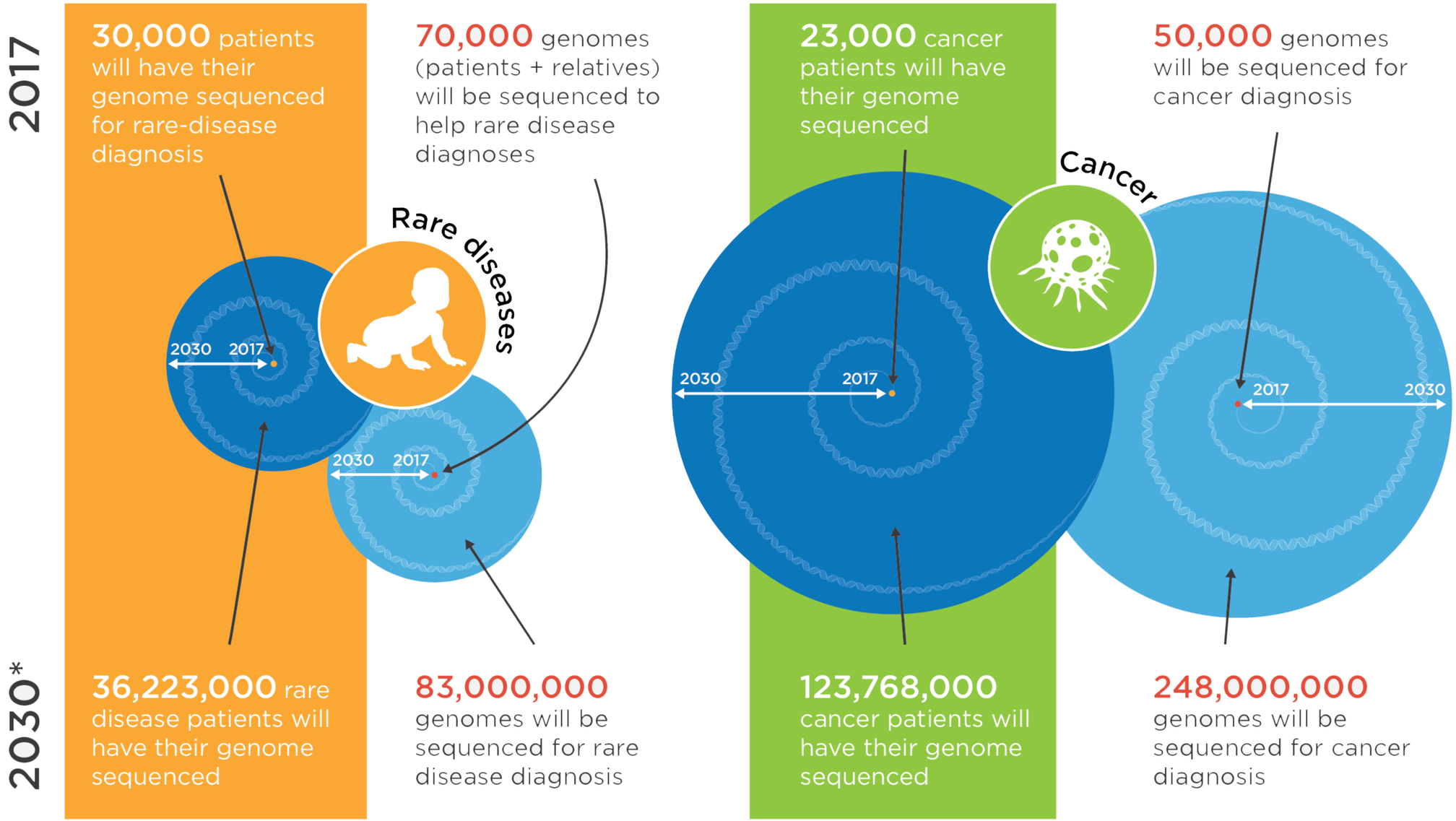
The numbers are derived from the 3'240 publications registered in the Progenetix database.



progenetix

Global Alliance
for Genomics & Health

Collaborate. Innovate. Accelerate.



* Projected figures, based on current data and known status of genomics initiatives worldwide.

The vision: Federation of data

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 ...

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

Global Alliance for Genomics & Health

GENOMICS

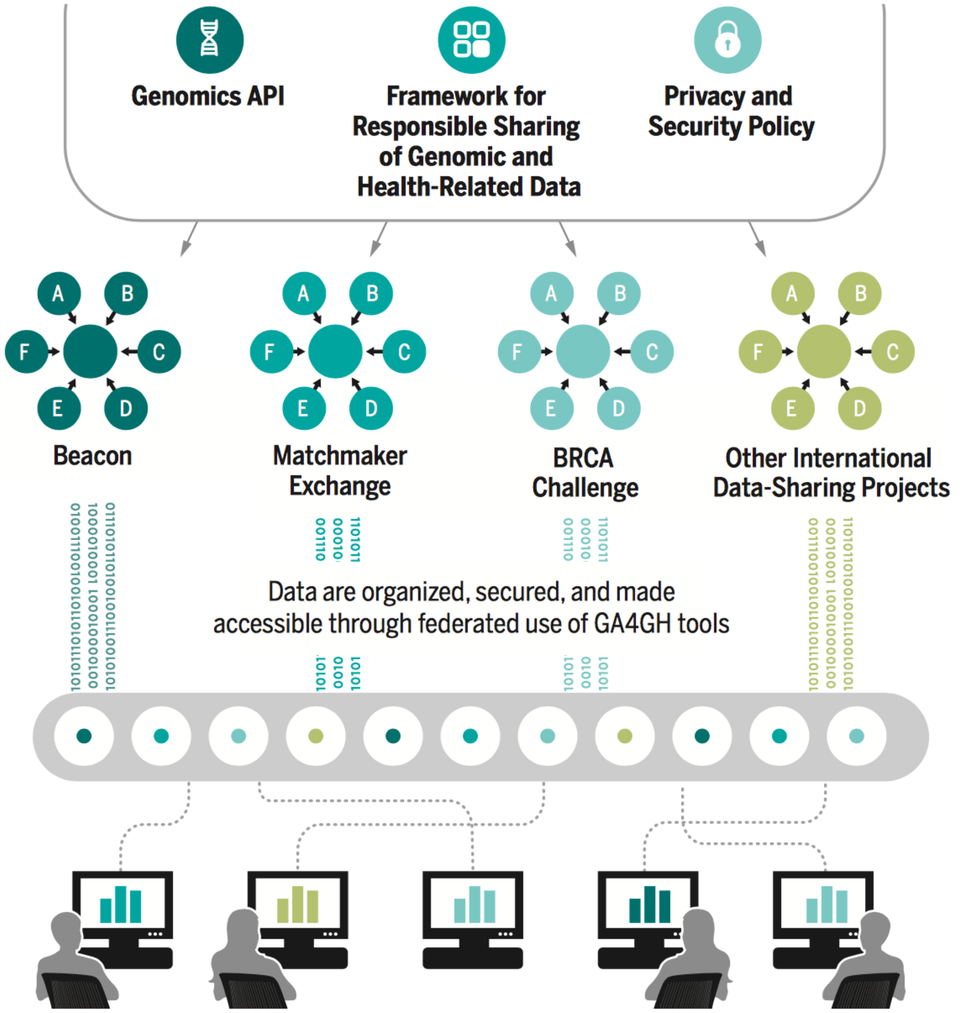
A federated ecosystem for sharing genomic, clinical data

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.



Global Alliance for Genomics & Health

GENOMICS

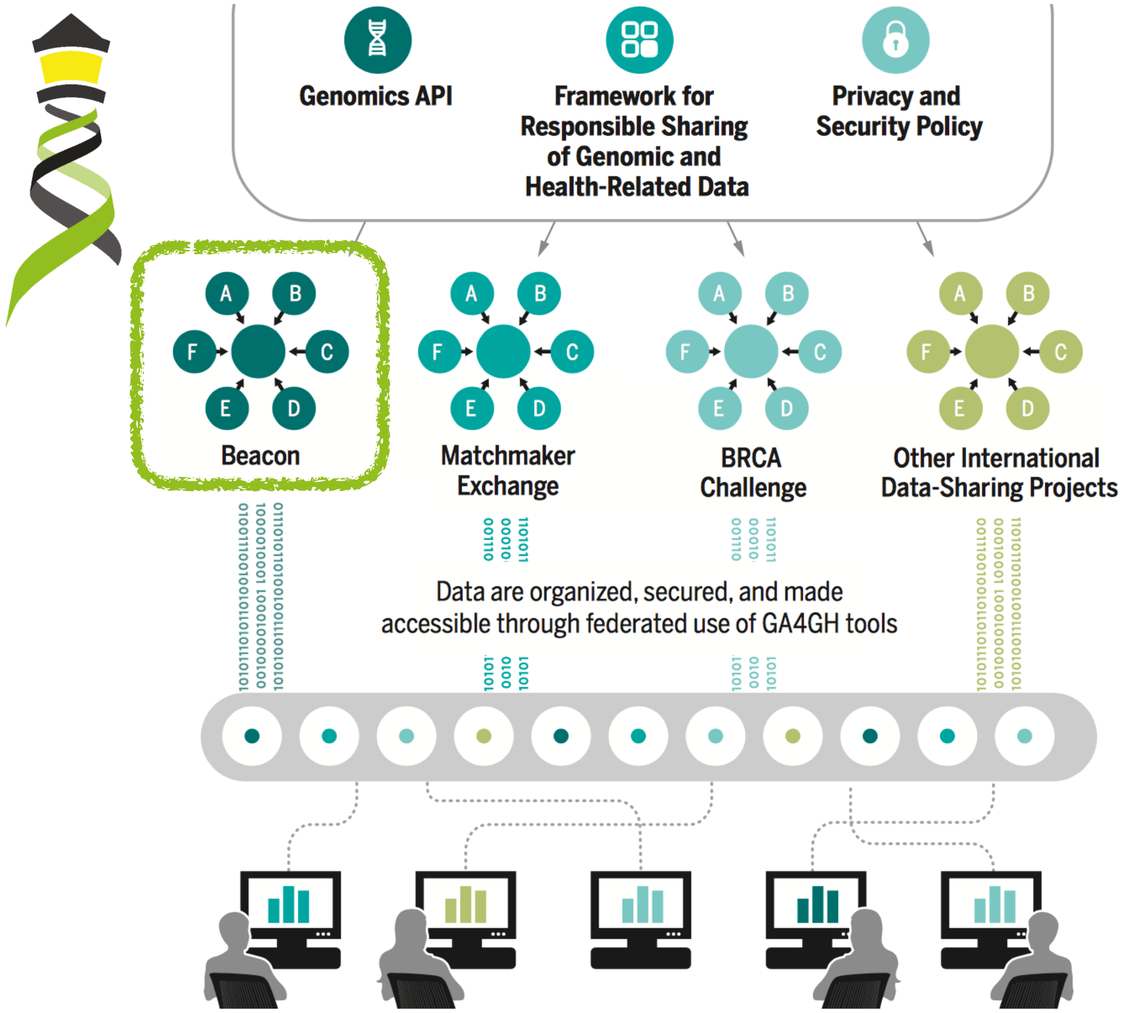
A federated ecosystem for sharing genomic, clinical data

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

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.

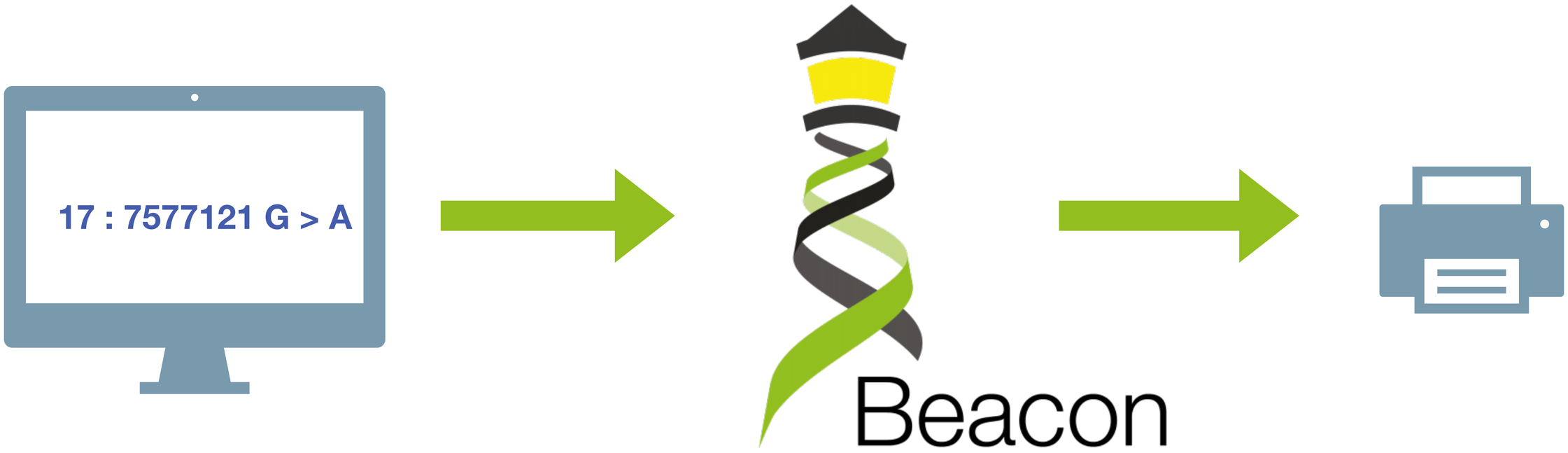
A number of more useful first versions have been suggested.

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*** …



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

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

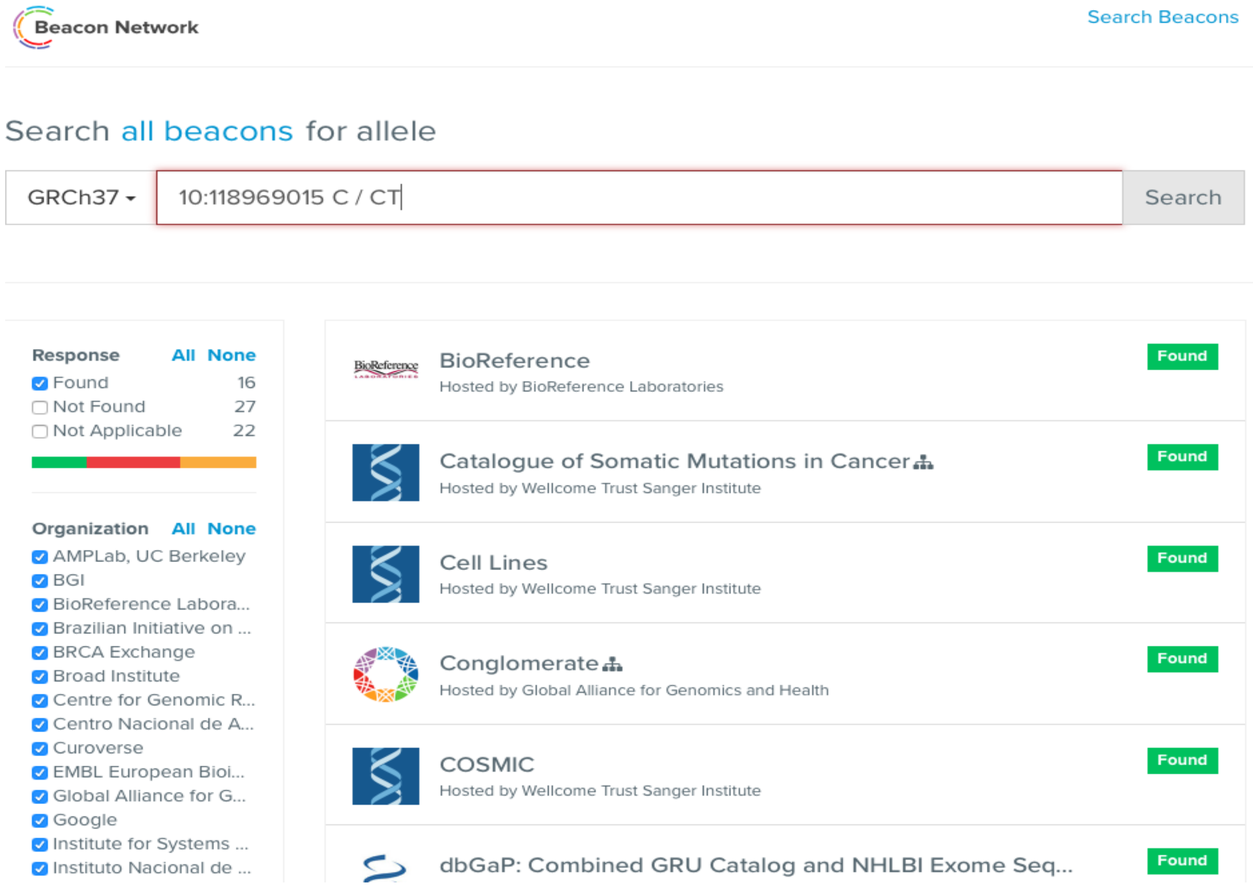
Have you seen this variant? It came up in my patient and we don't know if this is a common SNP or worth following up.

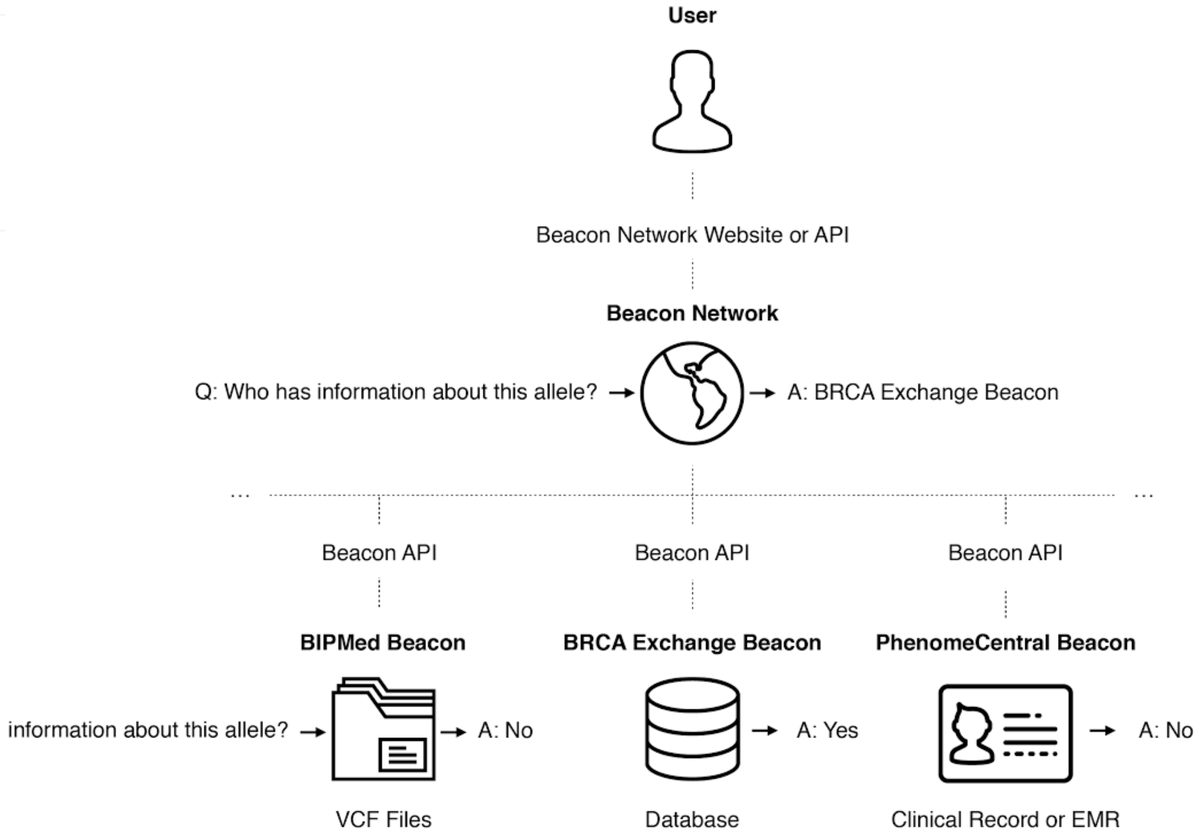
A Beacon network federates *genome variant queries* across databases that support the ***Beacon API***

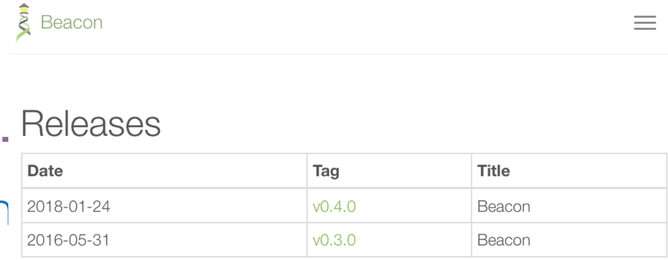
Here: The variant has been found in **few** resources, and those are from **disease** specific **collections**.

Beacon Project in 2016

An open web service that tests the willingness of international sites to share genetic data.







35+ Organizations

90+ Beacons

200+ Datasets

100K+ In

Global Alliance for Genomics & Health

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 v2 - Clinical Beacon requirements

Authors: Jordi Rambla, Michael Baudis, Anthony J Brookes, Lauren Fromont, Claudia Vasallo, Aina Jené

The original GA4GH Beacon implementation (up to v0.3) was conceived as a protocol for sharing the presence/absence of a given, specific, genomic mutation in a set of data (from patients of a given disease or from the population in general). Although with some potential benefit, e.g. in the area of rare disease diagnostics, it was ***not*** designed for clinical use but chiefly to foster data sharing by triggering the inquisitiveness of researchers once some data of interest is discovered in another institution. While later extensions of the protocol (v1.0 - v1.n) expanded the query and response options, this did not deviate from the general "existence of variants in resource X" paradigm.

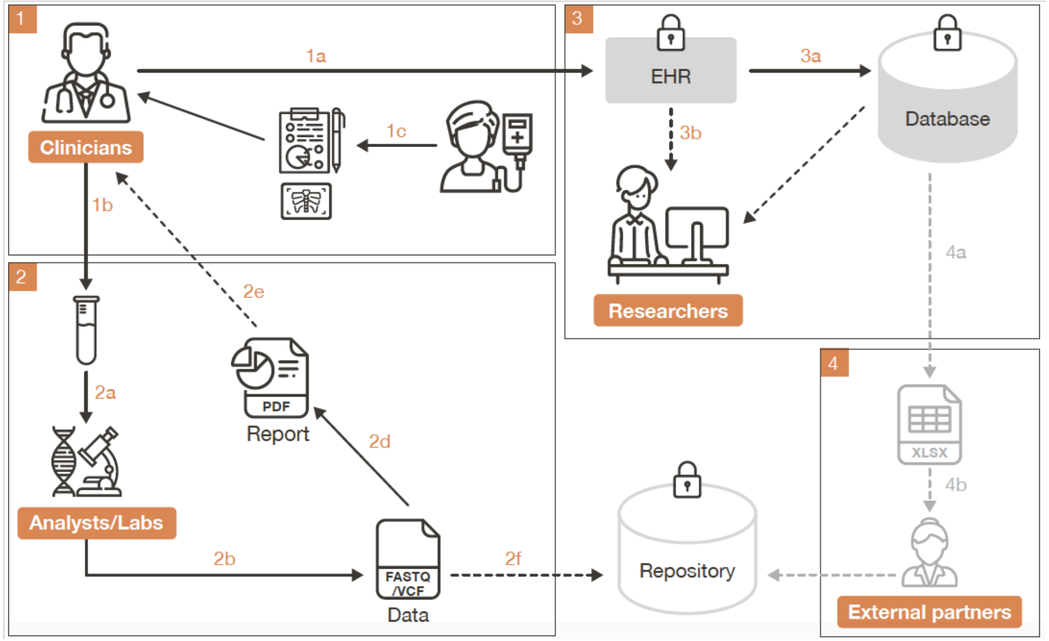
The simplicity and success of the concept has generated the request of making it more powerful, more useful in healthcare environments. The requests include:

- Allowing more informative queries, like filtering by gender or age
- Allowing to trigger the next step in the data access process, e.g. who to contact or which are the data use conditions
- Jumping to another system where the data could be accessed, e.g. if the Beacon is internal to a hospital, to provide the Id of the EHR of the patients having the mutation of interest.
- Including annotations about the variants found, among which the expert/clinician conclusion about the pathogenicity of a given mutation in a given individual or its role in producing a given phenotype.

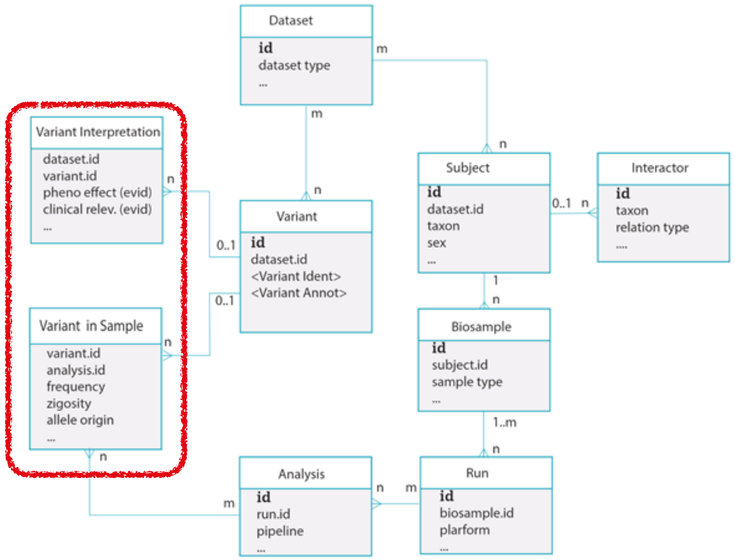
The process

The GA4GH Beacon group started a set of meetings and interviews with GA4GH Driver Projects and with ELIXIR partners in order to determine the scope of the next generation Beacon. The goal was to be useful without breaking the simplicity that made Beacon version 1 successful.
Interviews were conducted with the following GA4GH Driver Projects:

- Autism Speaks
- BRCA Exchange
- CanDIG
- EGA, ENA, EVA
- EuCanCancer
- European Joint Programme - Rare Diseases
- H3Africa
- GEM Japan
- Genomics England
- Matchmaker Exchange
- SVIP /SPHN
- VICC



Data flow and data sharing of genomic and phenotypic data in healthcare



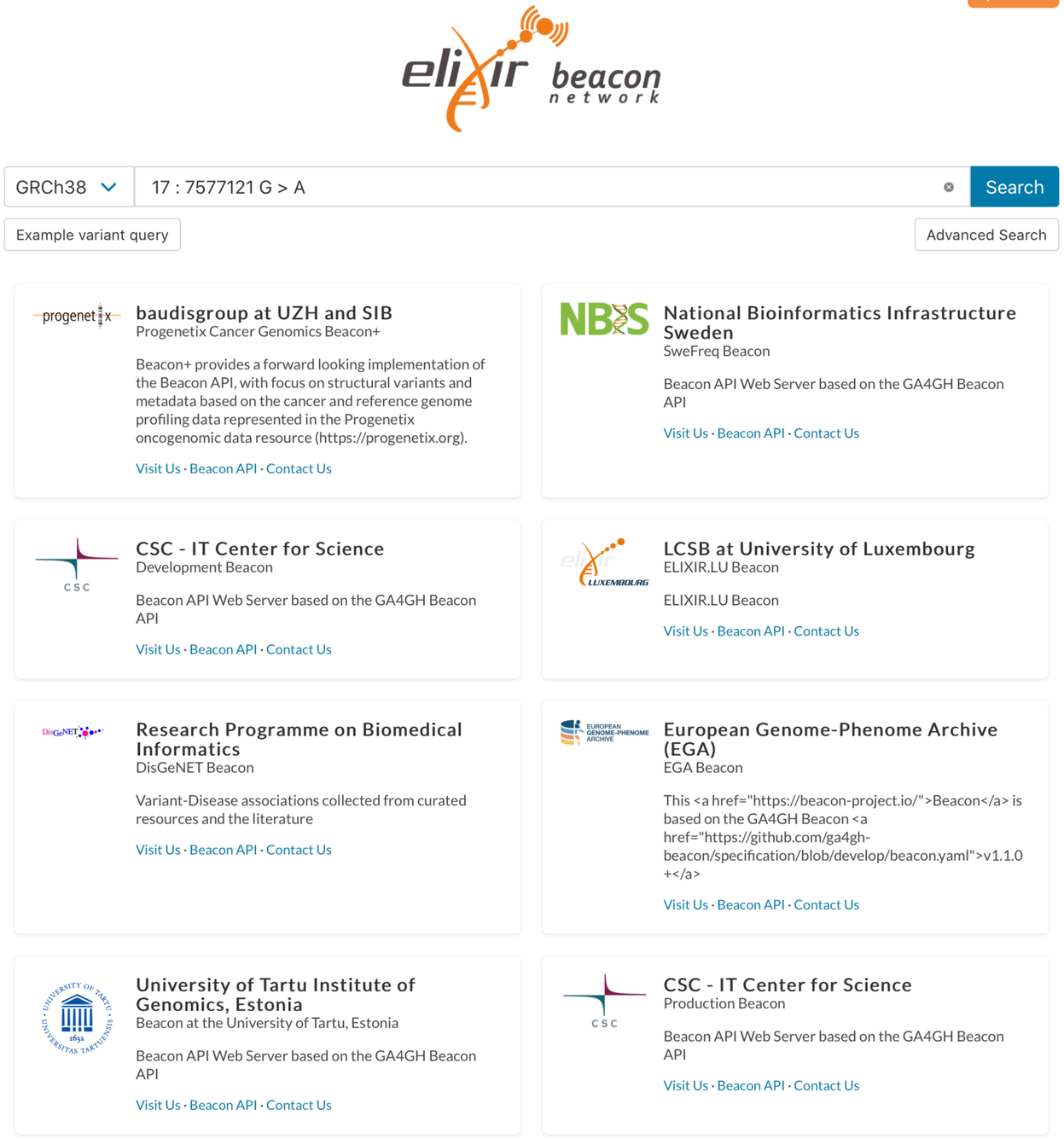
The Beacon v2 draft 2 entities logical model (implementations may vary)

ELIXIR Beacon Network

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

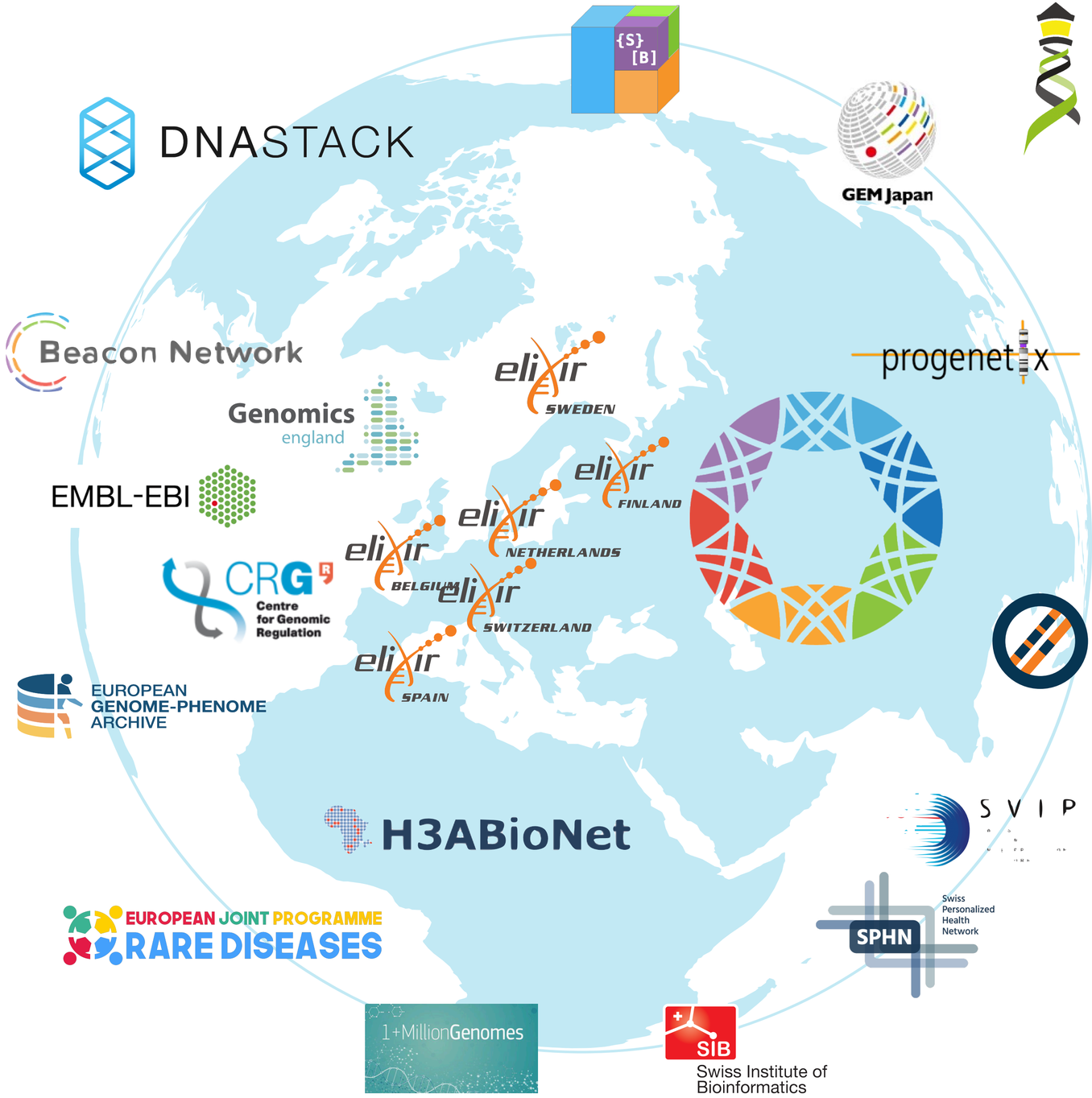
➡**integrator** throughout ELIXIR Human Data

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



Beacon Project - Partner Engagement & Next Steps

- Working with **partner communities & projects** on *deploying* Beacons
  - ELIXIR hCNV Community
  - **E**uropean **J**oint **P**rogram on **R**are **D**iseases
  - clinical groups & data initiatives (e.g. Andalucia, Cancer Core Europe, SPHN)
  - variant annotation resources, with optional clinical components (e.g. SVIP-O)
- Improving reference implementation and standards / **compliance** testing
- Beacon ***v2*** "fast forward" development
- aligning w/ GA4GH standards, through "request & adopt" => SchemaBlocks **{S}[B]**
- networks **throughout & beyond** ELIXIR

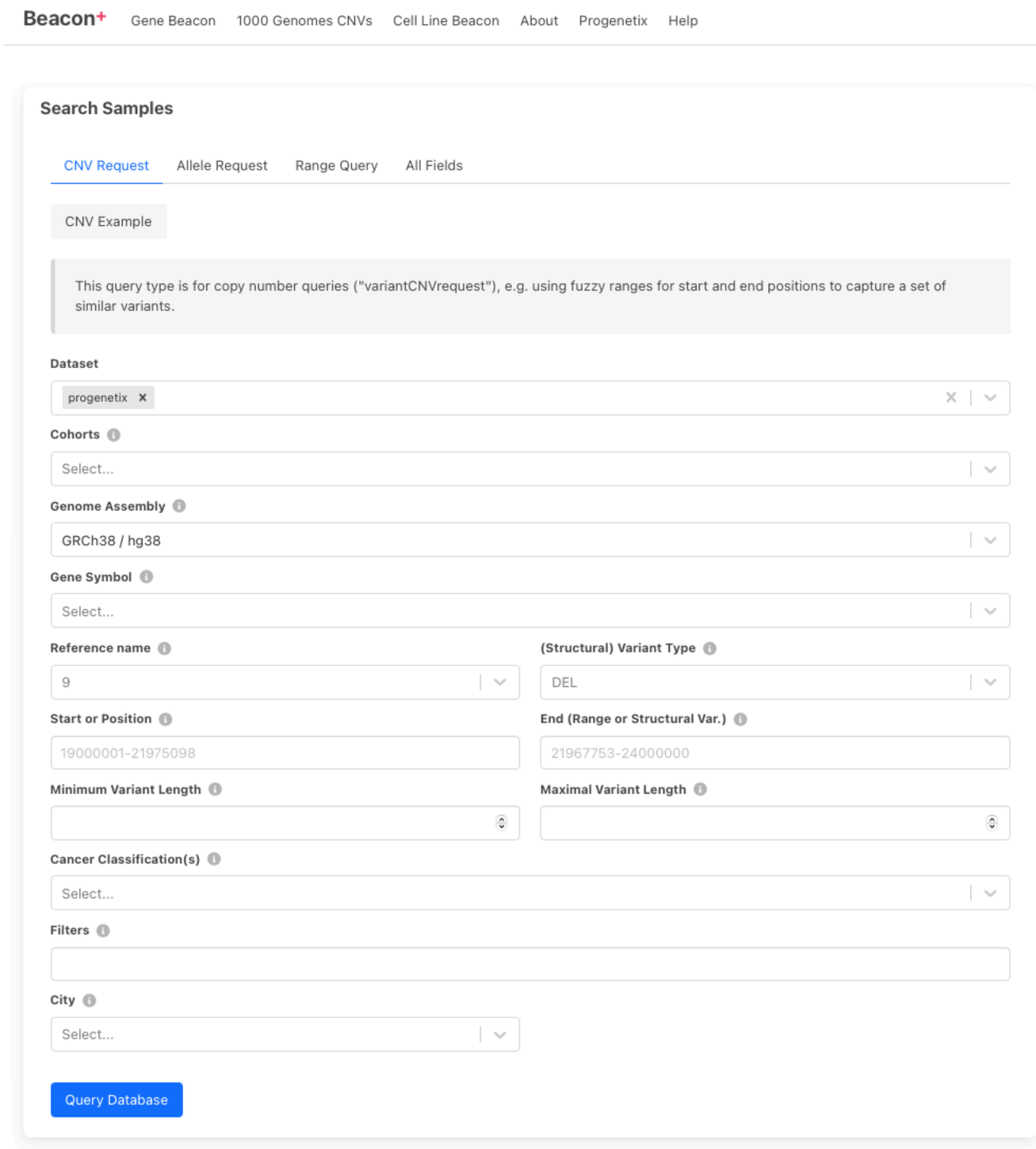


Beacon+ by Progenetix

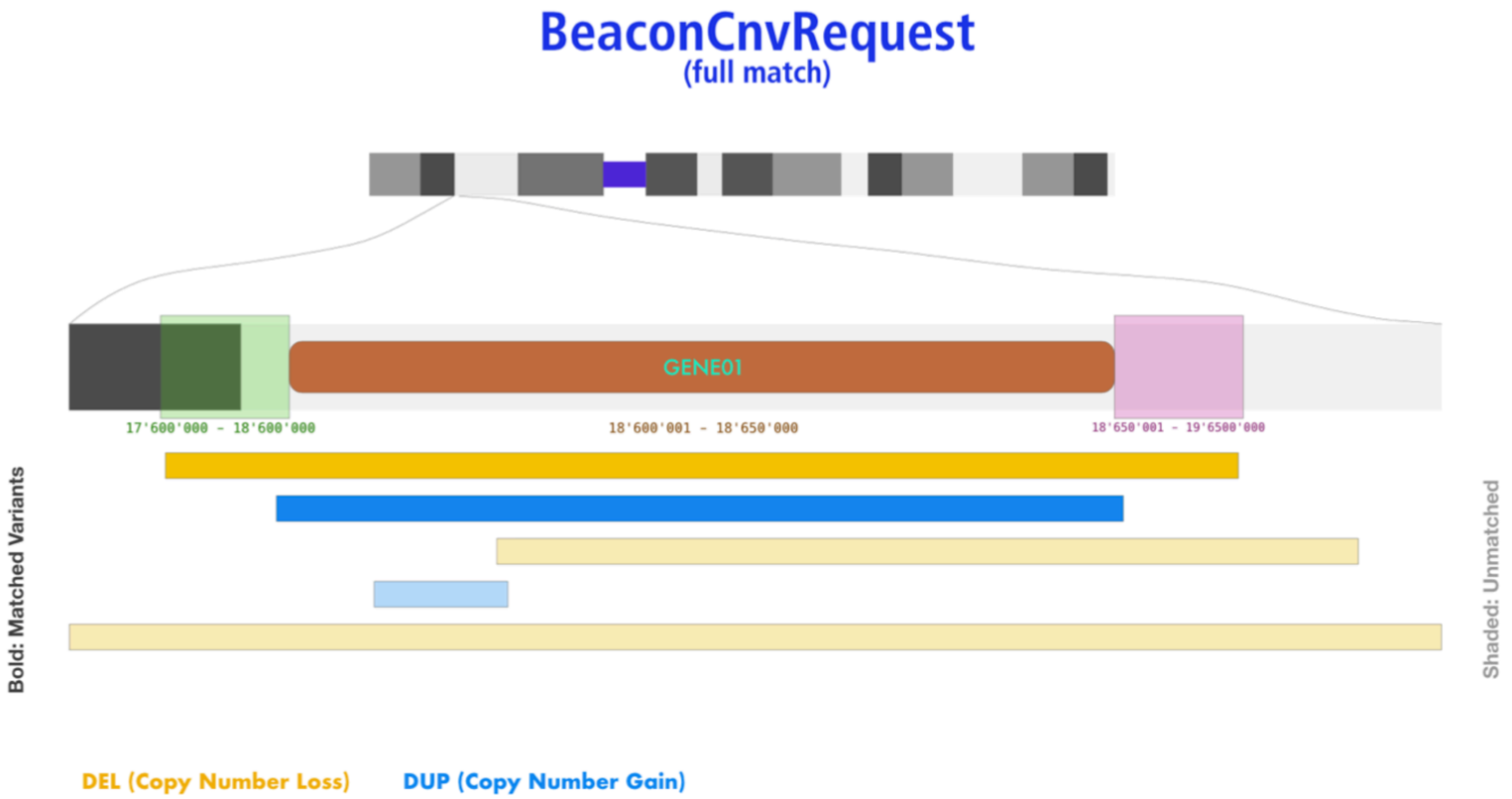
From Beacon Query to Explorative Analyses of CNV Patterns

- Since 2016 the Progenetix resource has been used to model options for Beacon development
  - 138334 individual samples from 698 cancer types
- The consistent use of hierarchical diagnostic codes allows the use of Beacon "filters" for histopathological/ clinically scoped queries
- Beacon's handover protocols can be utilized for data retrieval and, well, handing over to additional services, e.g.
  - downloads
  - visualization
  - use of external services (UCSC browser display...)

progenetix



- referenceName: 9
- start: [17600000, 18600000]
- end: [18650001, 19650000]
- variantType: SO:0001019

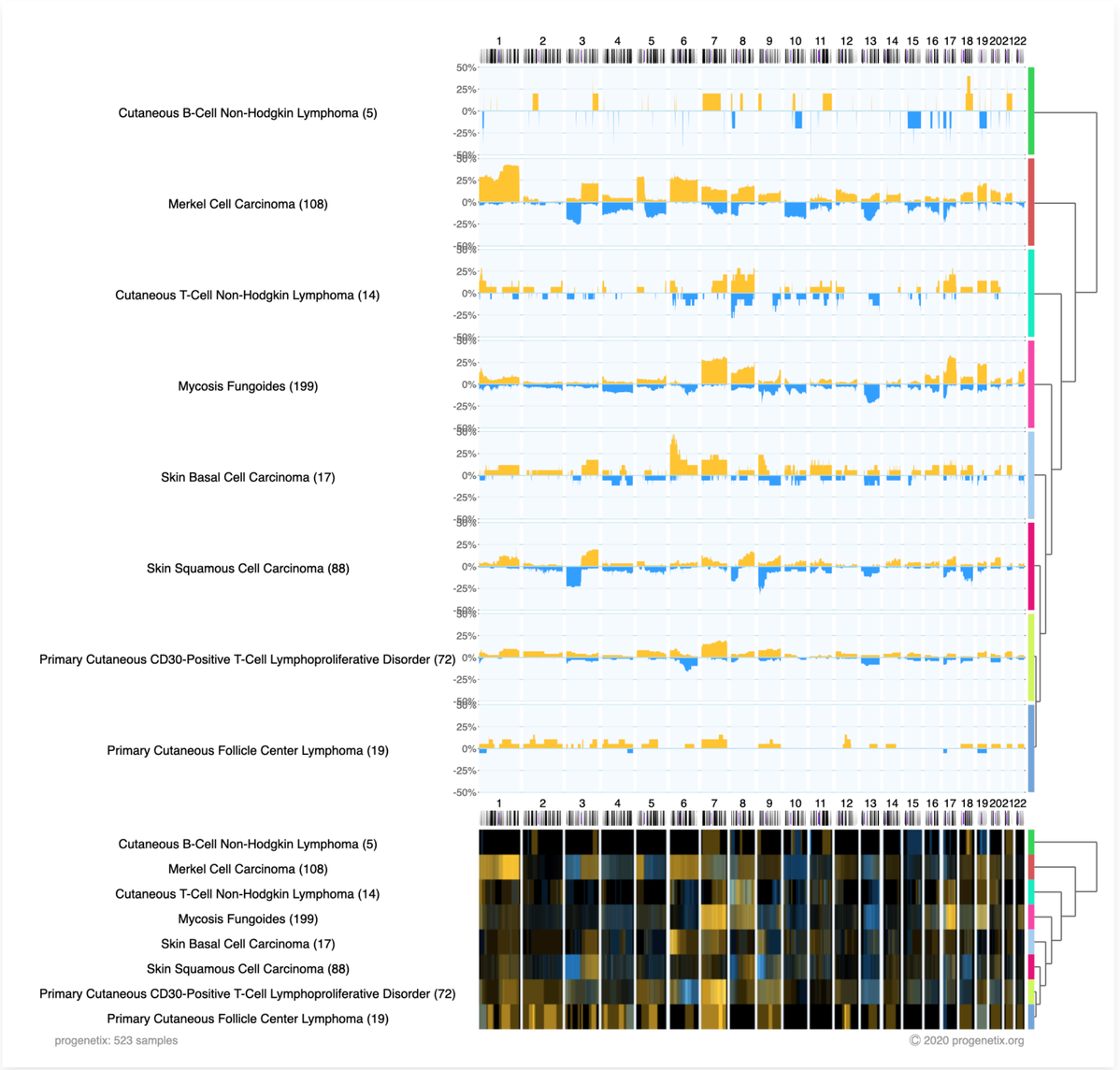


A “full match” BeaconCnvRequest is a typical scenario for e.g. matching CNVs in which the whole CDR of a gene has been duplicated. Here, both start and end search intervals lie outside of the region of interest. The maximum size of matched CNVs can be limited through the extend of the outer bounds (start[0], end[1]).

Beacon+ by Progenetix

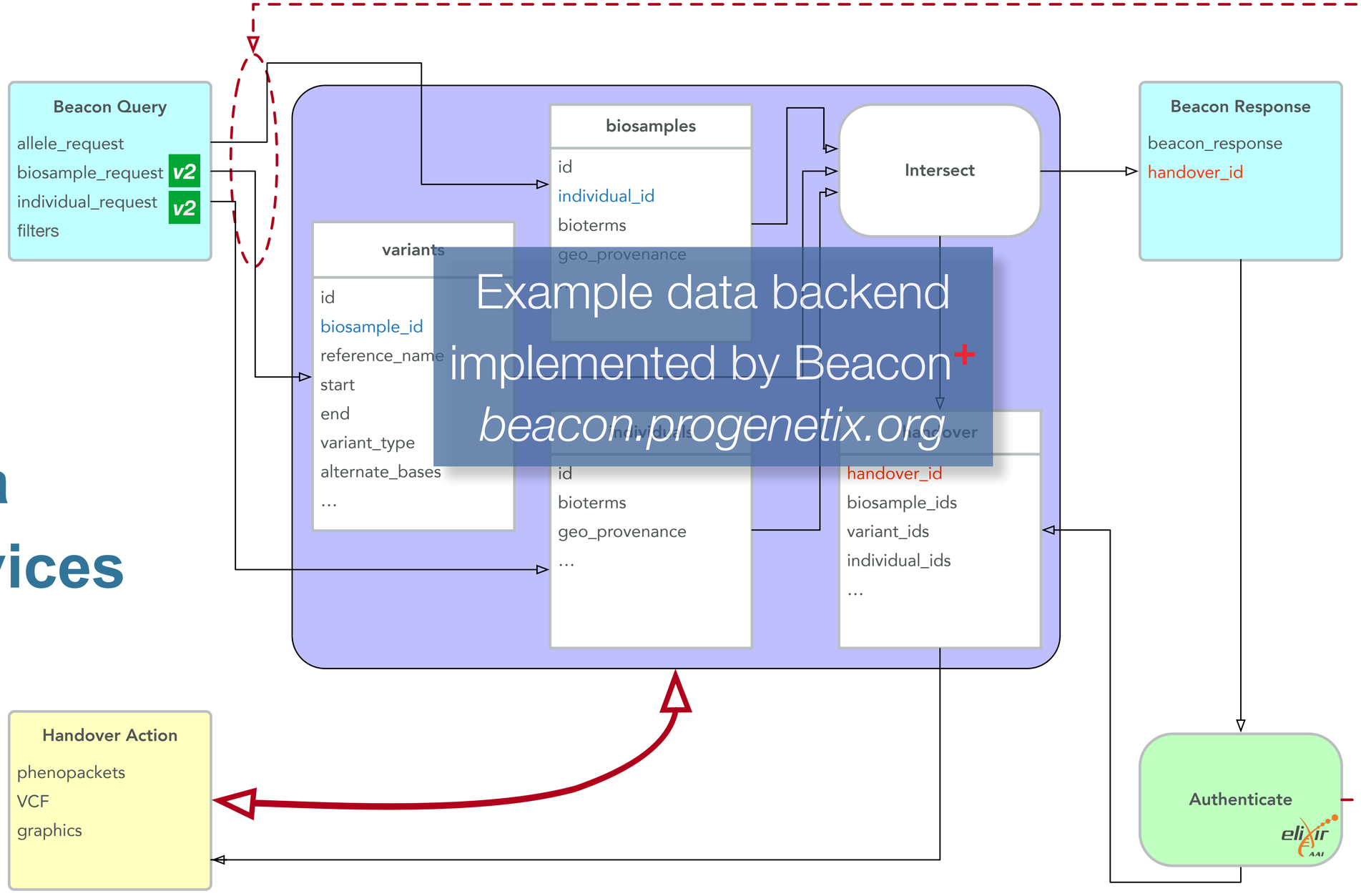
From Beacon Query to Explorative Analyses of CNV Patterns

- Since 2016 the Progenetix resource has been used to model options for Beacon development
  - 138334 individual samples from 698 cancer types
- The consistent use of hierarchical diagnostic codes allows the use of Beacon "filters" for histopathological/ clinically scoped queries
- Beacon's handover protocols can be utilized for data retrieval and, well, handing over to additional services, e.g.
  - downloads
  - visualization
  - use of external services (UCSC browser display...)



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

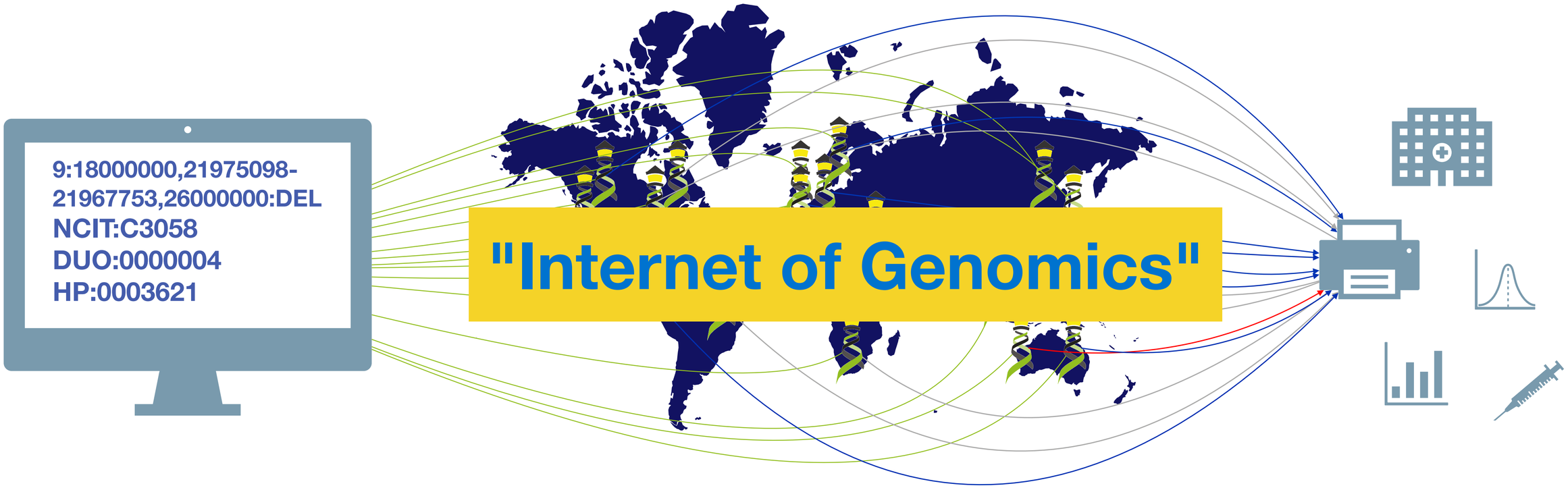
Beacon & GA4GH

- GA4GH has become the major "go to" international organization for the development of data exchange standards and implementation guidelines for genomics & health data
- essential for its success are national partner organizations such as SPHN which have the opportunity to shape the development of GA4GH through contributions, but importantly can benefit through the alignment with international, "cutting edge" standards developments, thereby avoiding duplicate efforts & resource waist
- The ***early*** adoption of protocols and standards such as Beacon and Phenopackets drives innovation and efficient data use, both for biomedical researchers and for clinicians
- While the direct benefit of e.g. local Beacon installations may be limited compared to legacy systems, it opens the door for scaled integration with outside systems
- **Beacon v2** specifically is being developed with **clinical requirements** in mind and will cover a broad range of use cases in precision medicine, rare diseases and cancer
- The active participation of SPHN in GA4GH development projects supports a leading position for Swiss biomedical research and personalized health applications

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

Beacon *v2* API

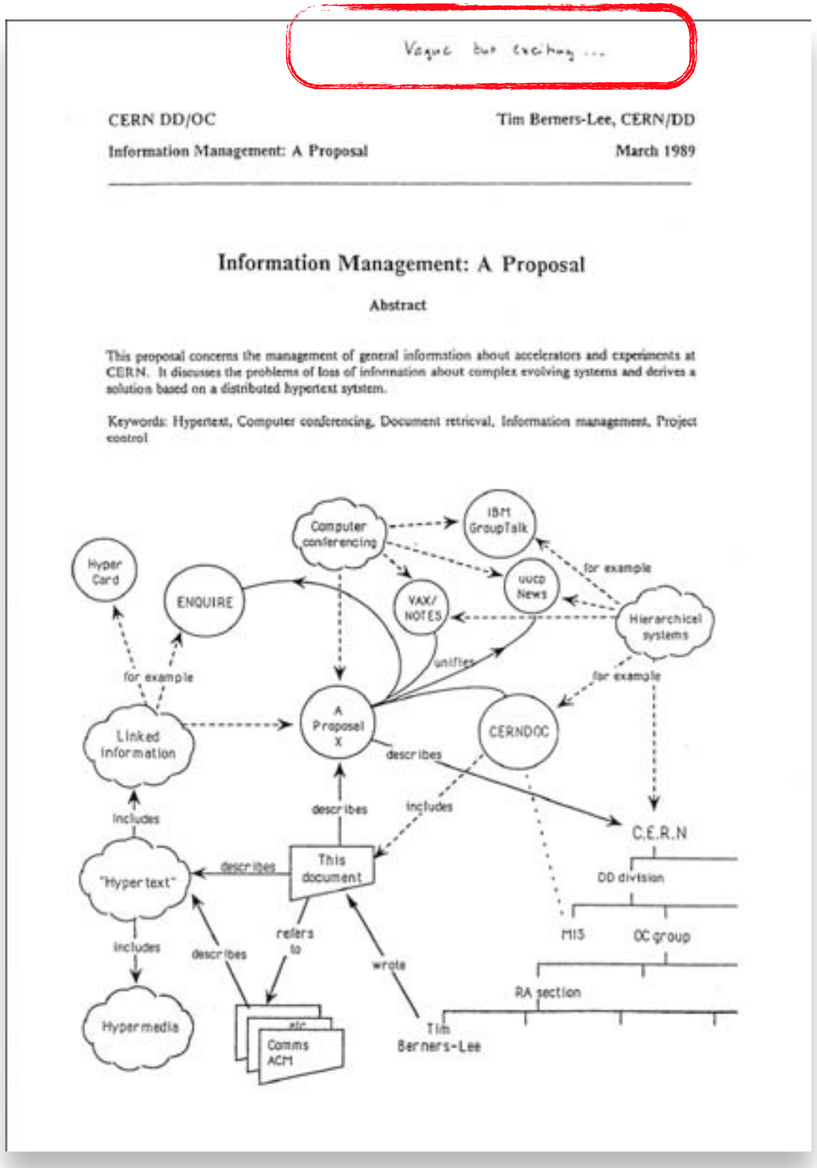
The Beacon API v2 proposal opens the way for the design of a simple but powerful "**genomics API**".



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

Beacon *v2* API

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



Vague but exciting ...

CERN DD/OC — Tim Berners-Lee, CERN/DD

Information Management: A Proposal — March 1989

Information Management: A Proposal

Abstract

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

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

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

World Wide Web

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

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

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

Help
: on the browser you are using

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

Technical
: Details of protocols, formats, program internals etc

Bibliography
: Paper documentation on W3 and references.

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

History
: A summary of the history of the project.

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

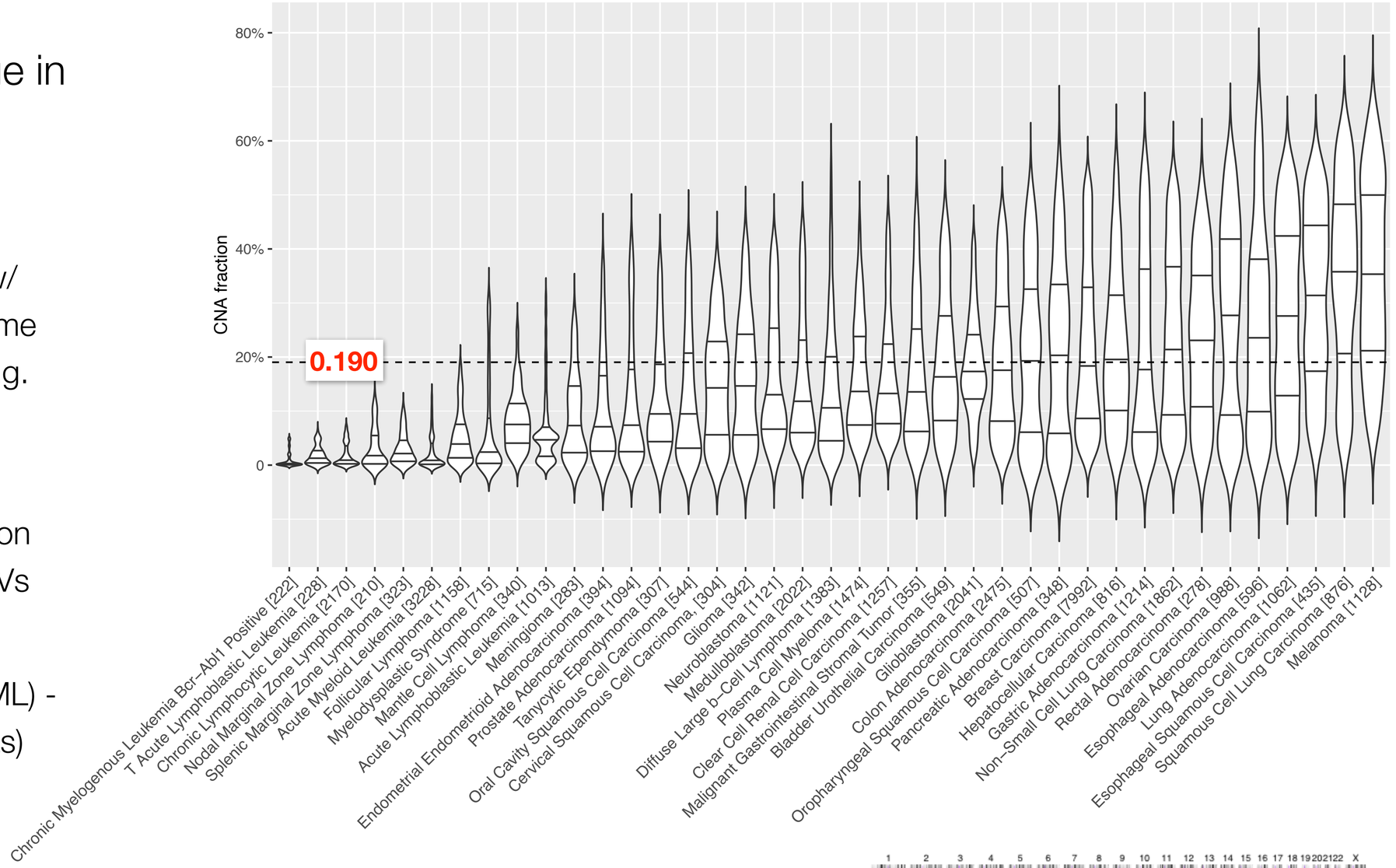
Getting code
: Getting the code by anonymous FTP , etc.

Progenetix Data Use Cases

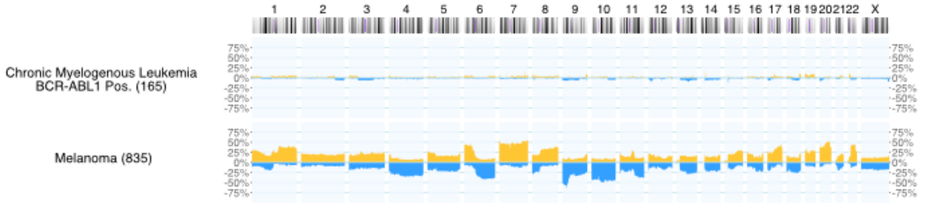
progenetix

Genome CNV coverage in Cancer Classes

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



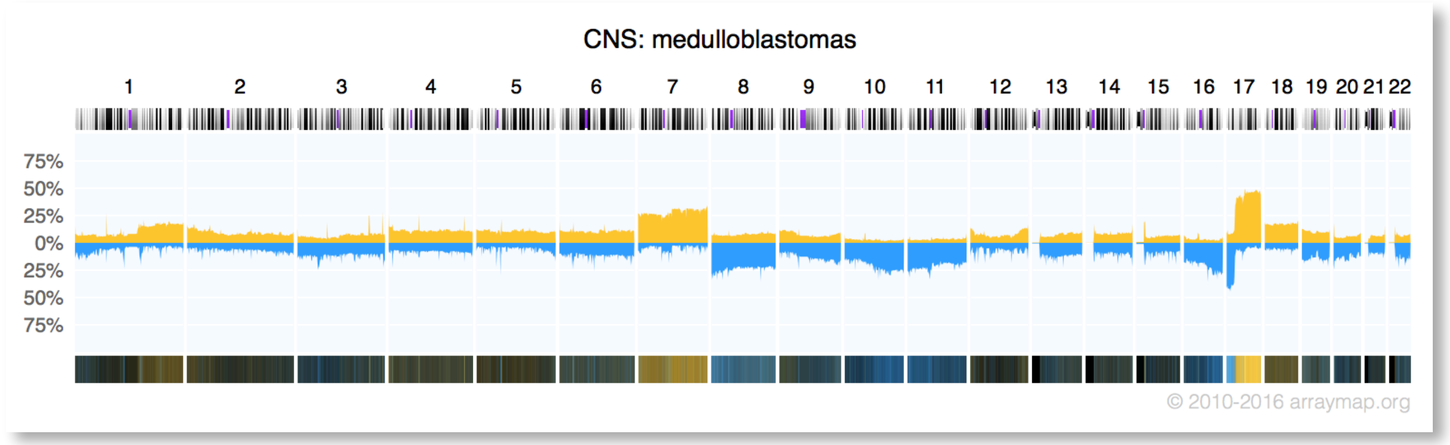
Lowest / Highest CNV fractions =>

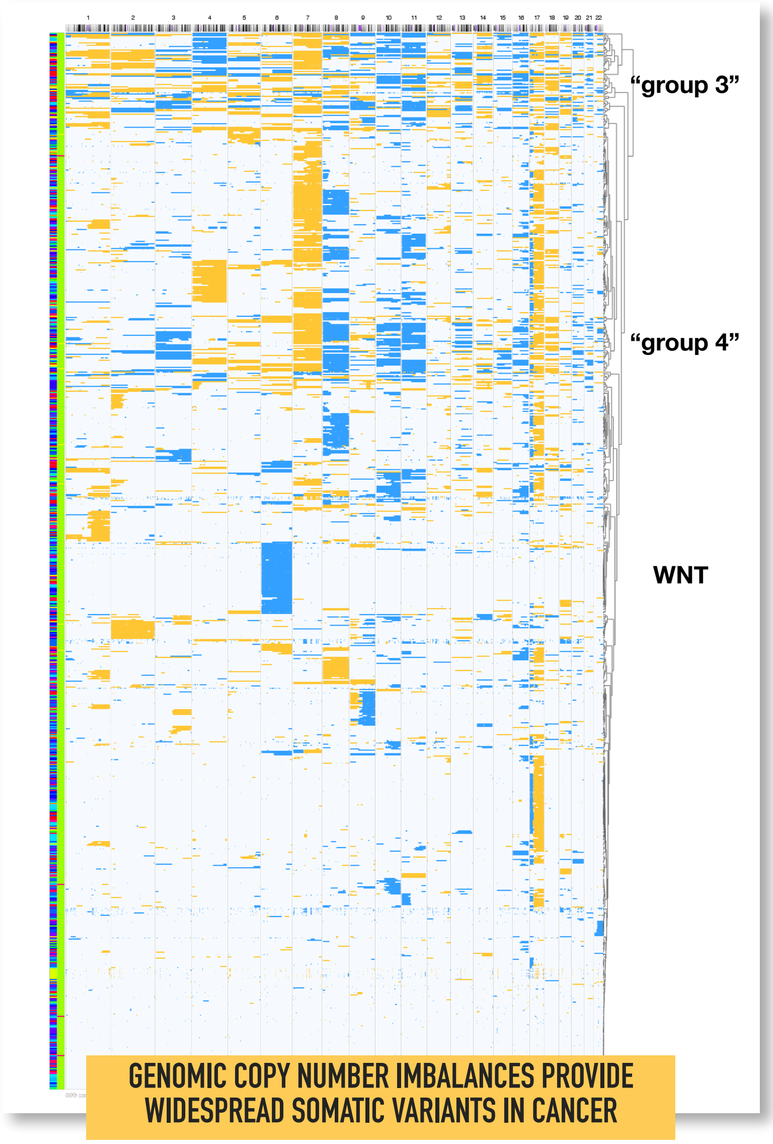


Somatic CNVs In Cancer

Recurrent mutation patterns

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





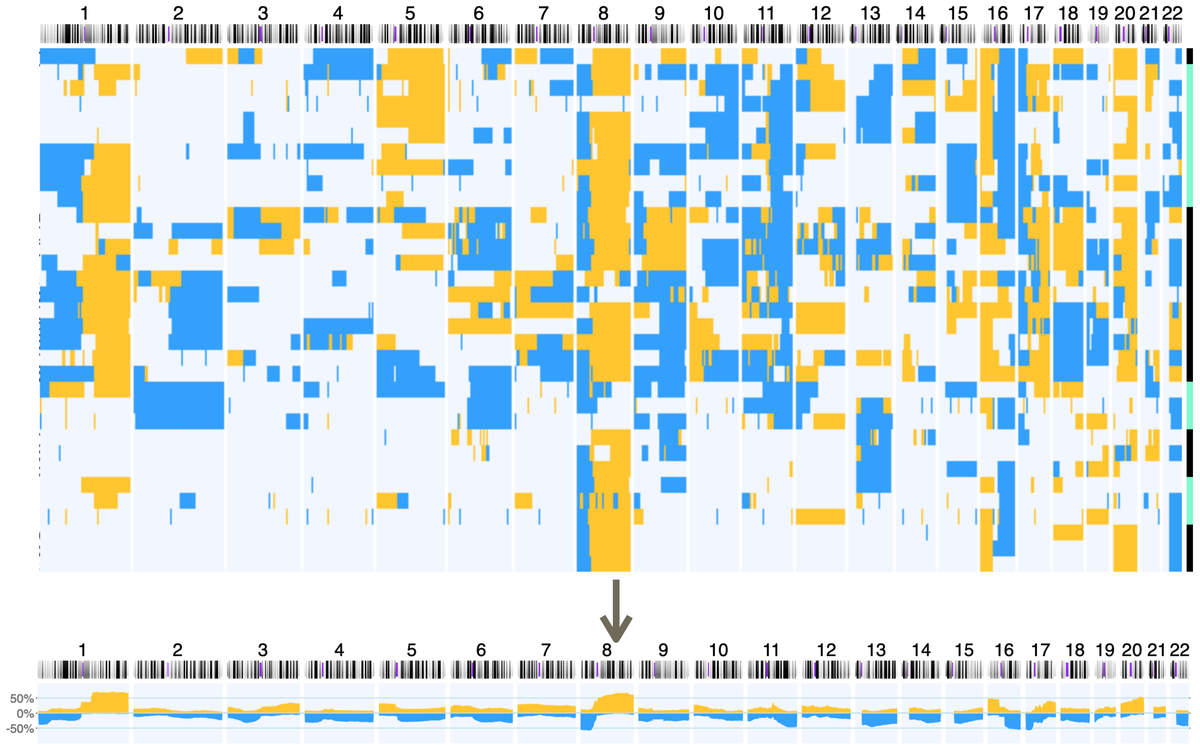
A genomic copy number histogram for malignant medulloblastomas, the most frequent type of pediatric brain tumors, displaying regions of genomic duplications and deletions. These can be decomposed into individual tumor profiles which segregate into several clusters of related mutation patterns with functional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

arrayMap

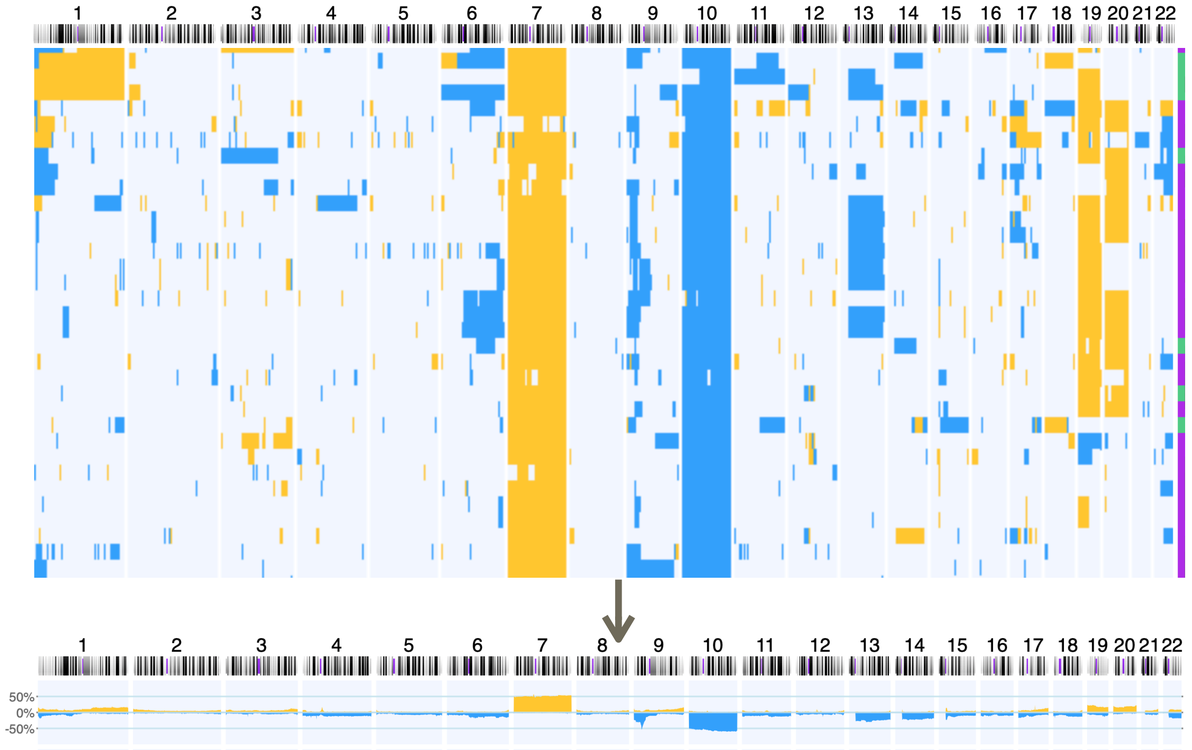
Drivers? Passengers? Markers?

Disentangling CNA Patterns

Ductal Breast Carcinoma



Glioblastoma



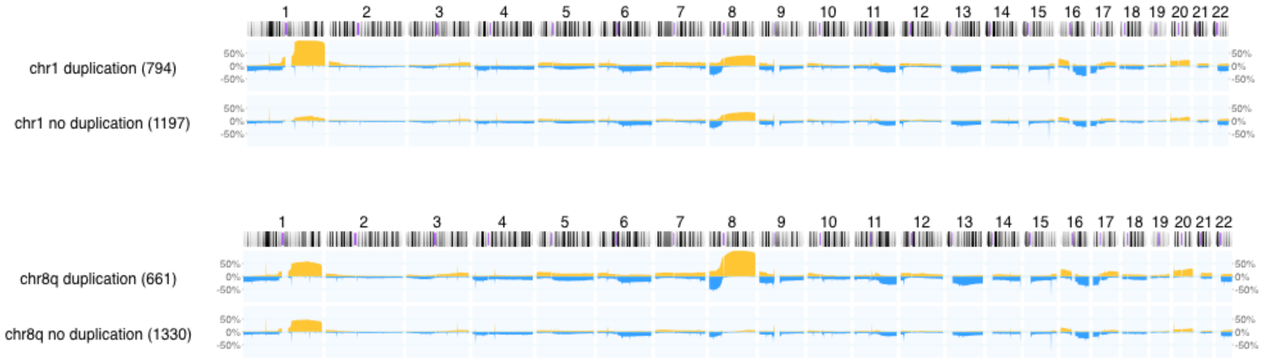
Thousands of genes involved (passengers)
Descriptive report with arbitrary cutoff

Qingyao Huang | March 2021

Intra-Disease CNA patterns are not random

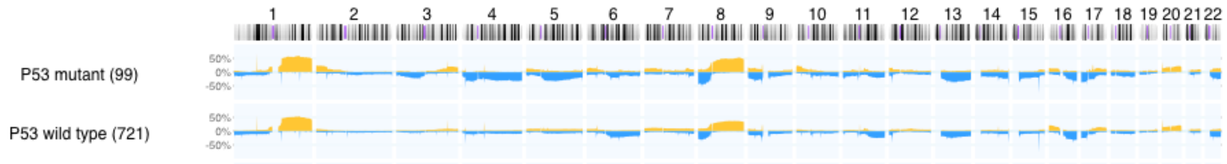
1. Robust

Independent evolvement of CNV landscape



2. Regulated

P53 mutant slightly higher CNV



3. Functionally relevant

IntClust BRCA subtypes

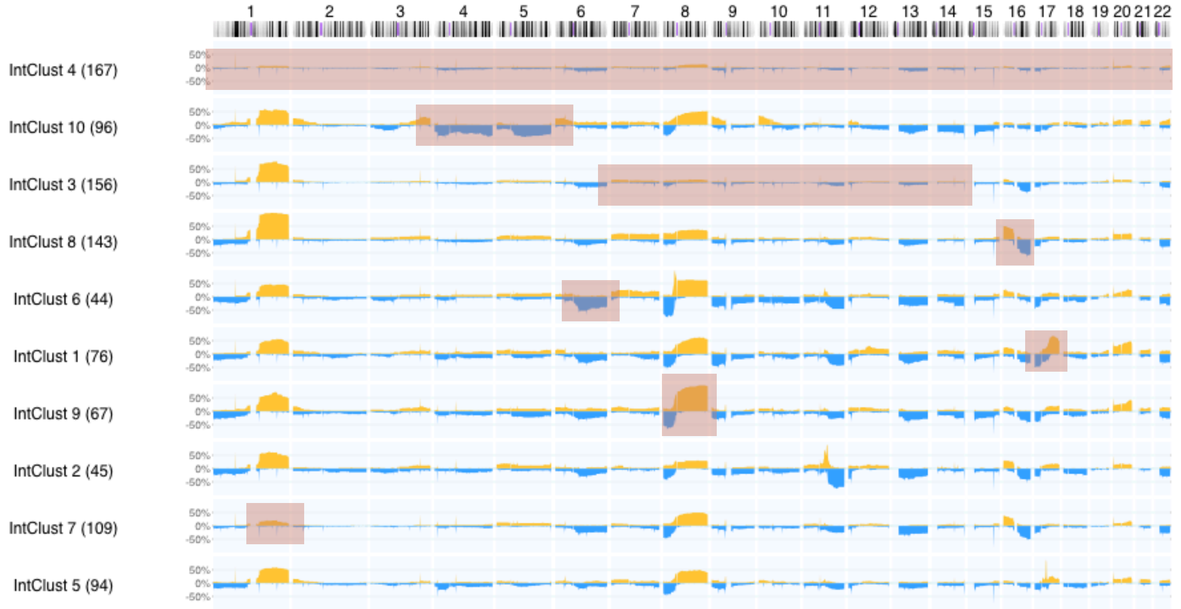
A new genome-driven integrated classification of breast cancer and its implications

SJ Dawson, OM Rueda, S Aparicio, C Caldas - The EMBO journal, 2013 - embopress.org

Breast cancer is a group of heterogeneous diseases that show substantial variation in their molecular and clinical characteristics. This heterogeneity poses significant challenges not only in breast cancer management, but also in studying the biology of the disease. Recently, rapid progress has been made in understanding the genomic diversity of breast cancer. These advances led to the characterisation of a new genome-driven integrated classification of breast cancer, which substantially refines the existing classification systems …

☆ Cited by 278 Related articles All 8 versions

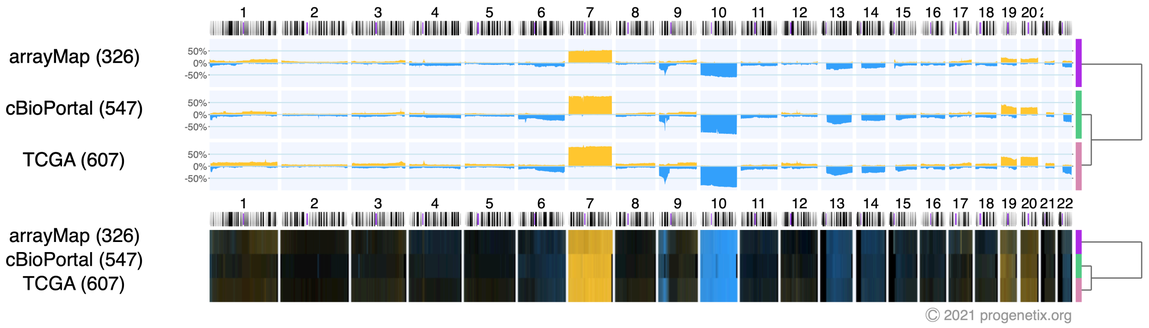
Distinctive CNV patterns among IntClust groups



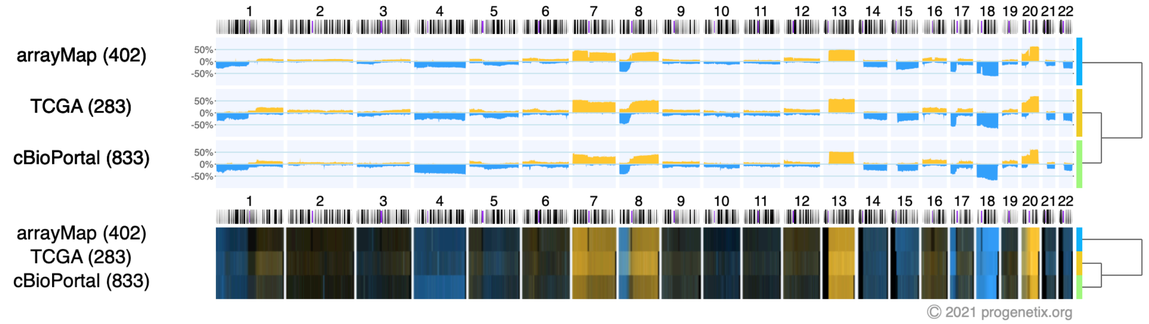
Data source: METABRIC 1992 breast cancer samples, hierarchical clustering, tree not shown

Limited Resource-specific Biases in CNA Data

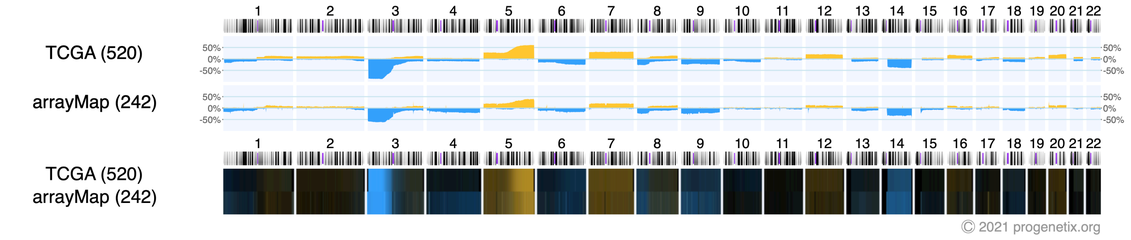
Glioblastoma Multiforme



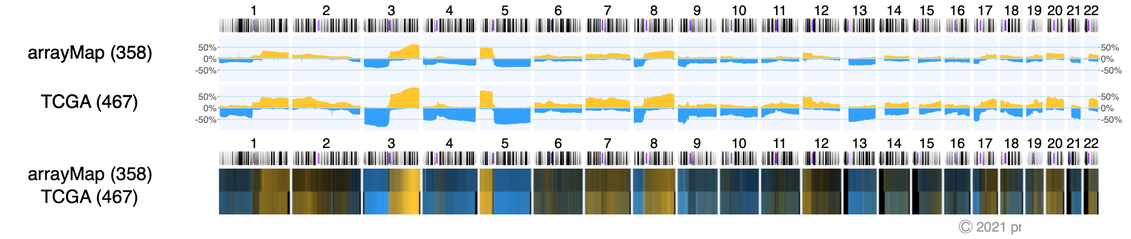
Colon adenocarcinoma



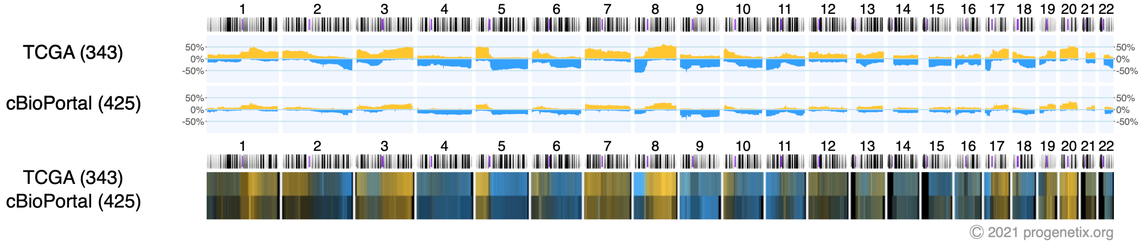
Clear cell renal cell carcinoma



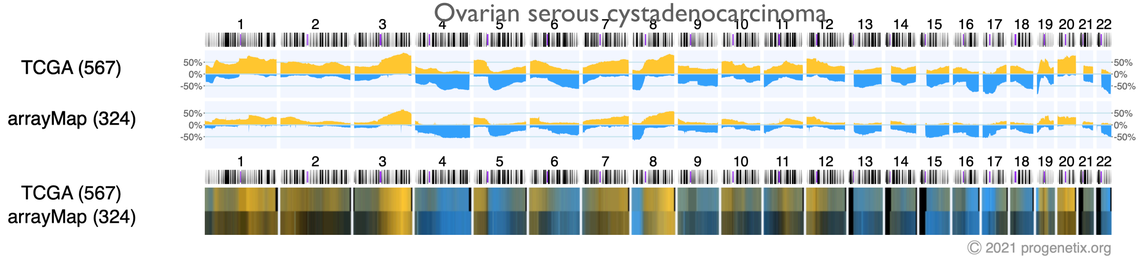
Lung squamous cell carcinoma



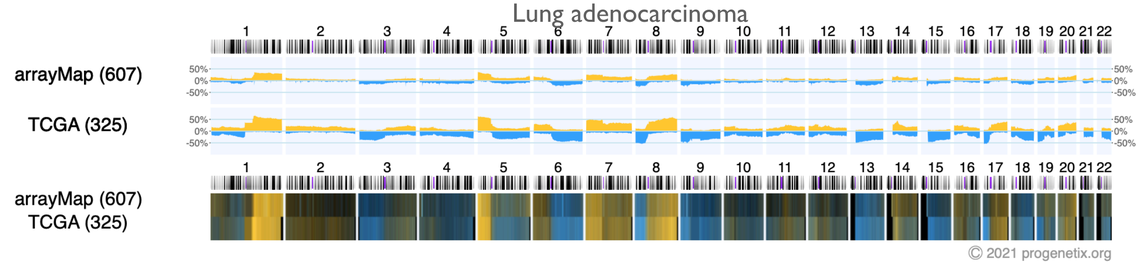
Bladder urothelial carcinoma



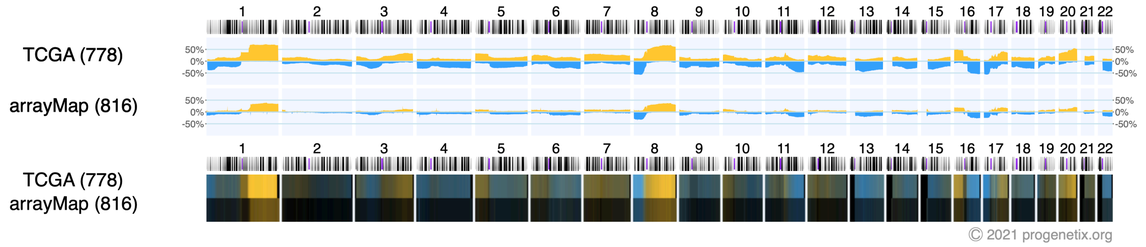
Ovarian serous cystadenocarcinoma



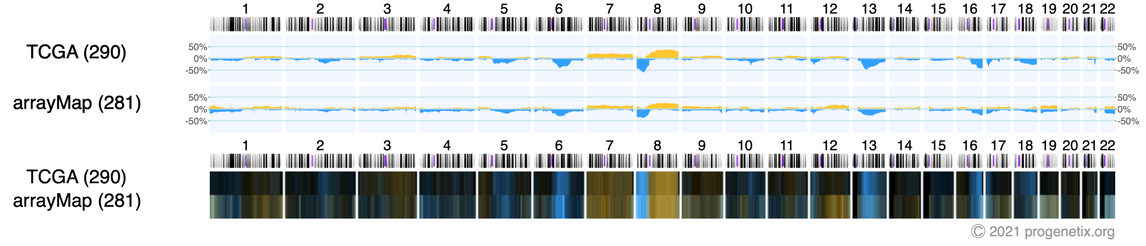
Lung adenocarcinoma



Ductal breast carcinoma

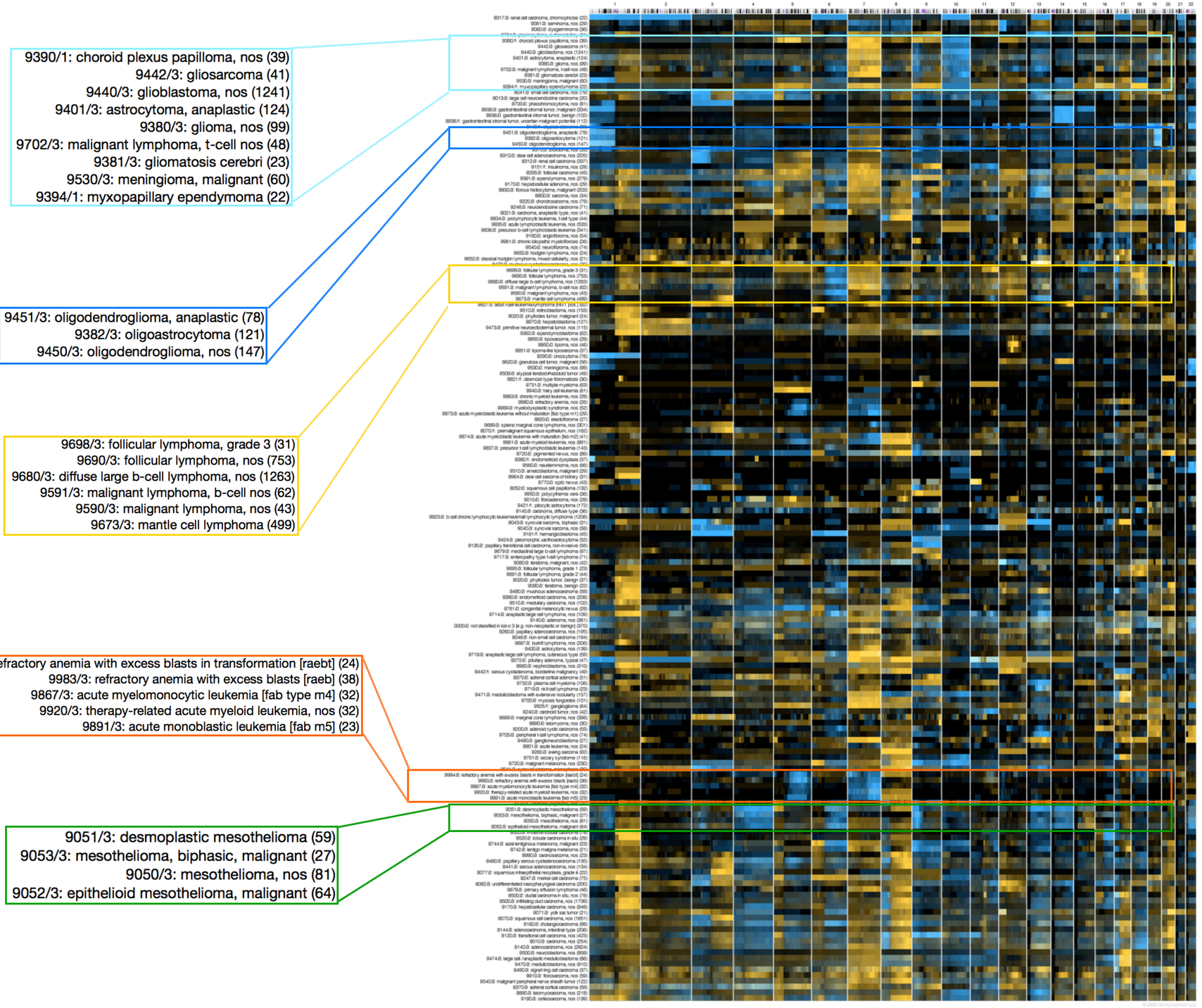


Prostate adenocarcinoma



Qingyao Huang | March 2021

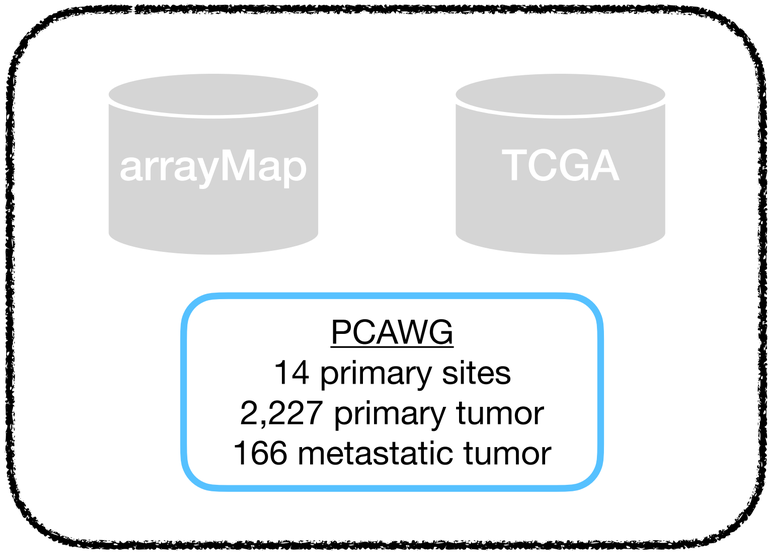
Somatic Mutations In Cancer: Patterns
Making the case for genomic classifications
Some related cancer entities show similar copy number profiles



arrayMap

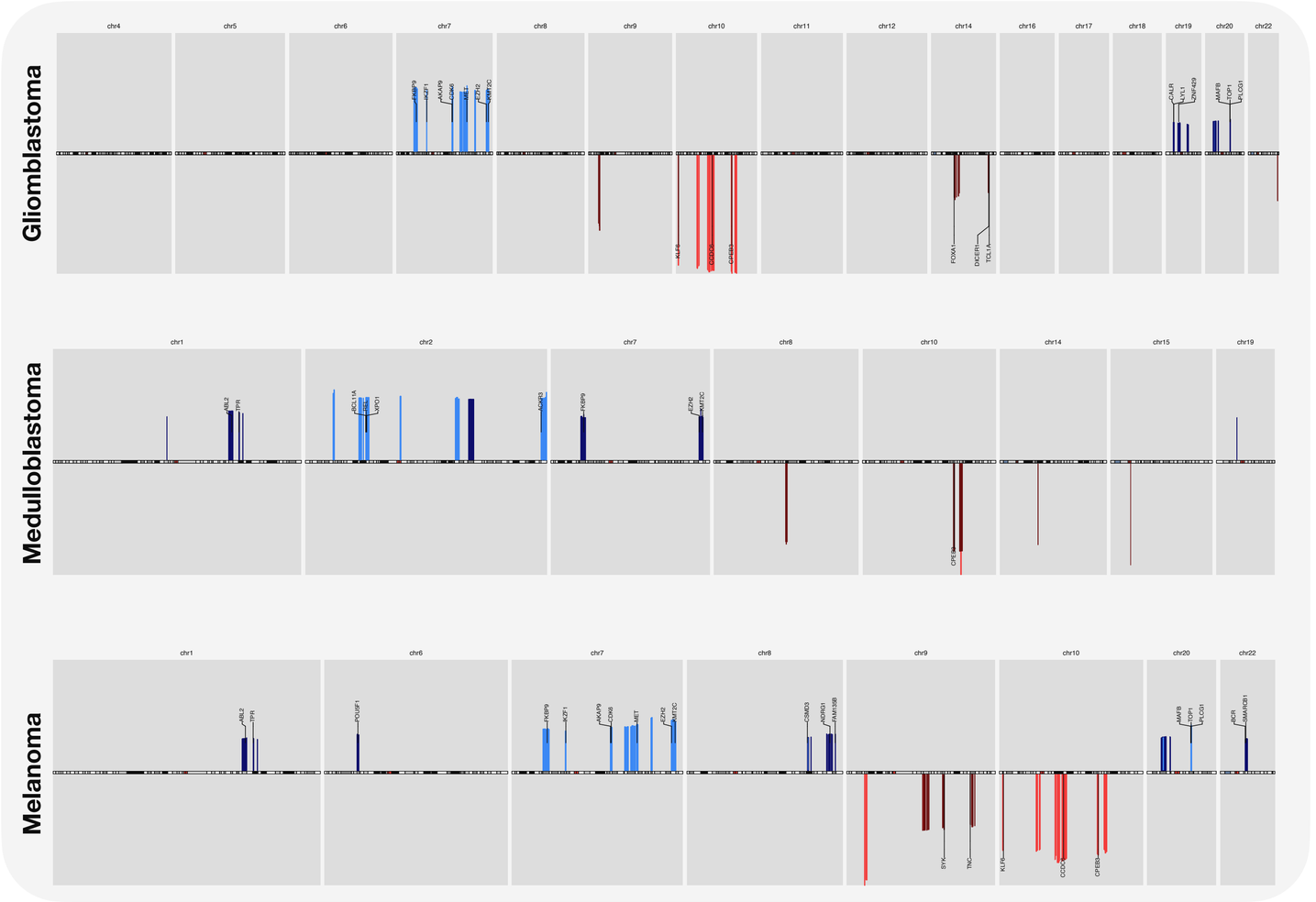
Unique Patterns of Copy Number Mutations Across Cancer Types

An extensive collection of tumor CNV



- **Unique & distinctive patterns**
- **Patterns of sites and disease**
- **Identify the origin of a tumor**

Examples of unique CNV patterns



Progenetix as Example Genomics Resource

Some trajectories ...

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

progenetix

progenetix DIPG

Diffuse Intrinsic Pontine Glioma (DIPG) Genomics Repository

- DIPG Home
- Search Samples
- Gene CNA Frequencies
- DIPG Publications
- DIPG News
- DIPG People
- DIPG Blog
- DIPG Links
- Progenetix Home

University of Zurich[UZH]

ICR The Institute of Cancer Research

UNIVERSITÄTSMEDIZIN GÖTTINGEN UMG

FOLLOW US ON twitter

Sponsors & Support

DIPG collaborative

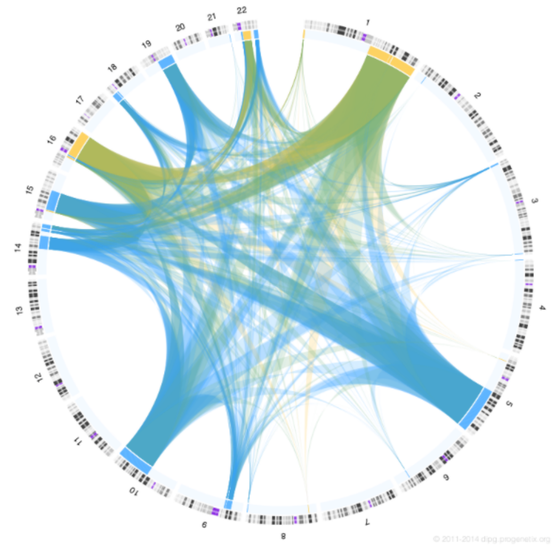
The BRAIN TUMOUR CHARITY

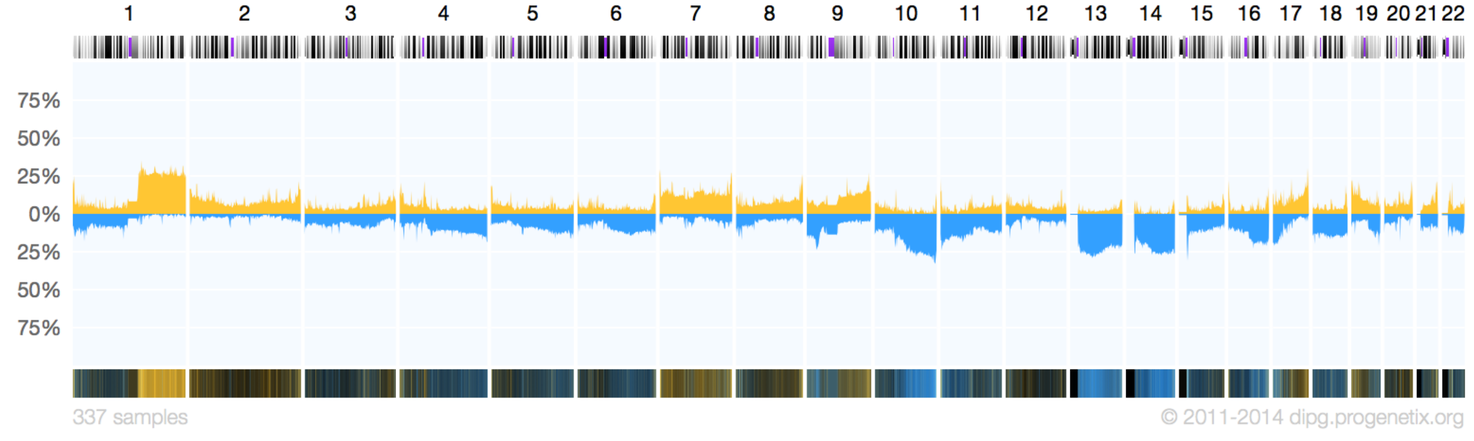
THE CURE STARTS NOW FOUNDATION

The **DIPG Genomics Repository** is an International collaboration supported by *The Cure Starts Now Foundation*. It aims to provide a central resource for researchers to investigate genome-wide profiling data from childhood diffuse intrinsic pontine glioma specimens, and additionally for other types of pediatric high grade brain tumors.

This work forms part of a systematic review and meta-analysis of paediatric glioma genomics aimed at collating publicly-available data sets of these diseases in children. In addition, we encourage unpublished or pre-publication data to be submitted to a password-controlled site. We welcome any comments you may have as this resource is being developed.

Michael Baudis, Andre von Büren, Chris Jones
DIPG Genomics Repository Leads





Genomic copy number aberrations in 337 DIPG and pediatric high grade gliomas

ICD-O
Locus
?

DIPG NEWS

2014-04-16: Chris Jones' group at the ICR involved in recent discovery of ACVR1 as DIPG driver gene

2013-08-19: New "DIPG People" page

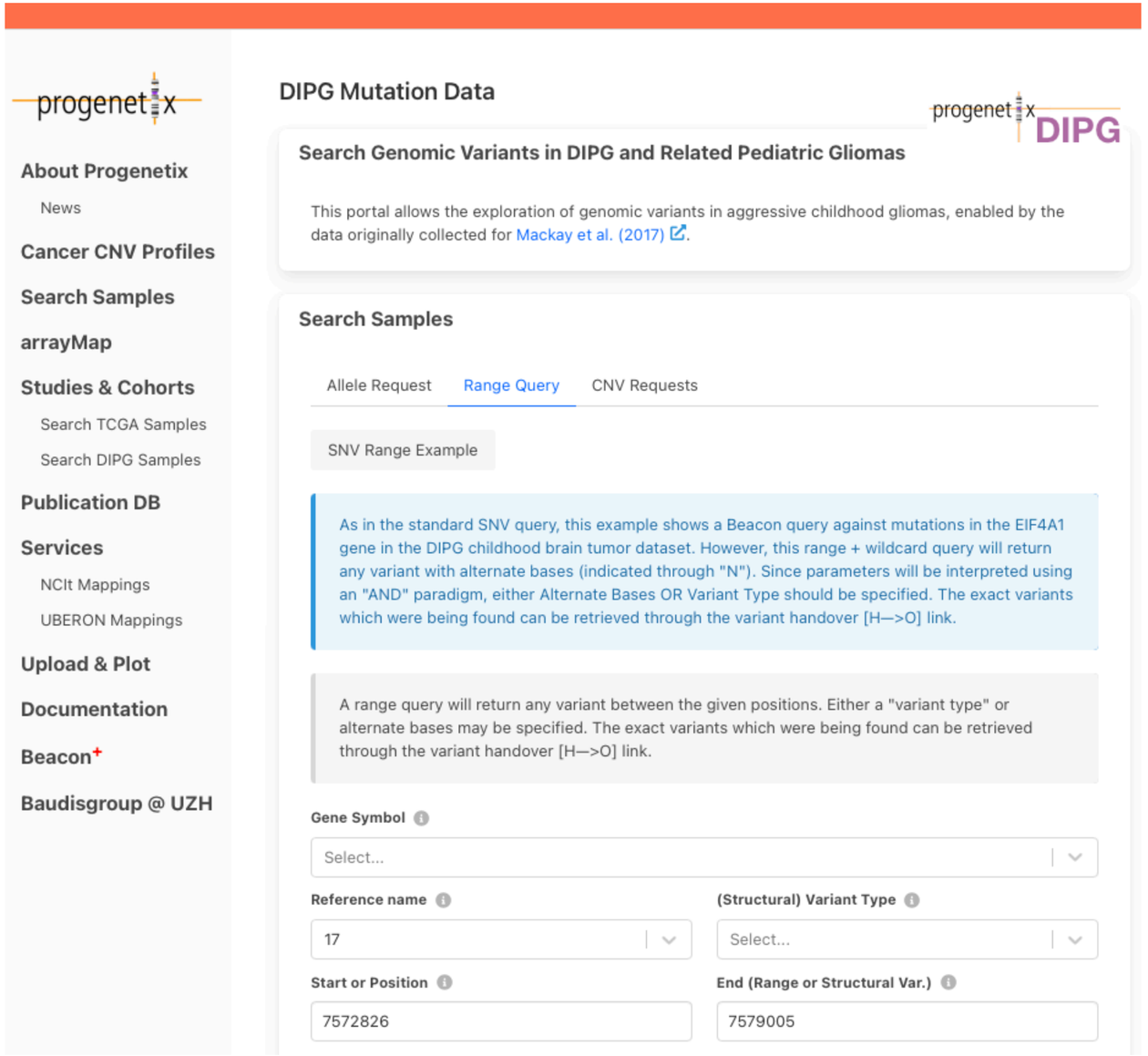
More news ...

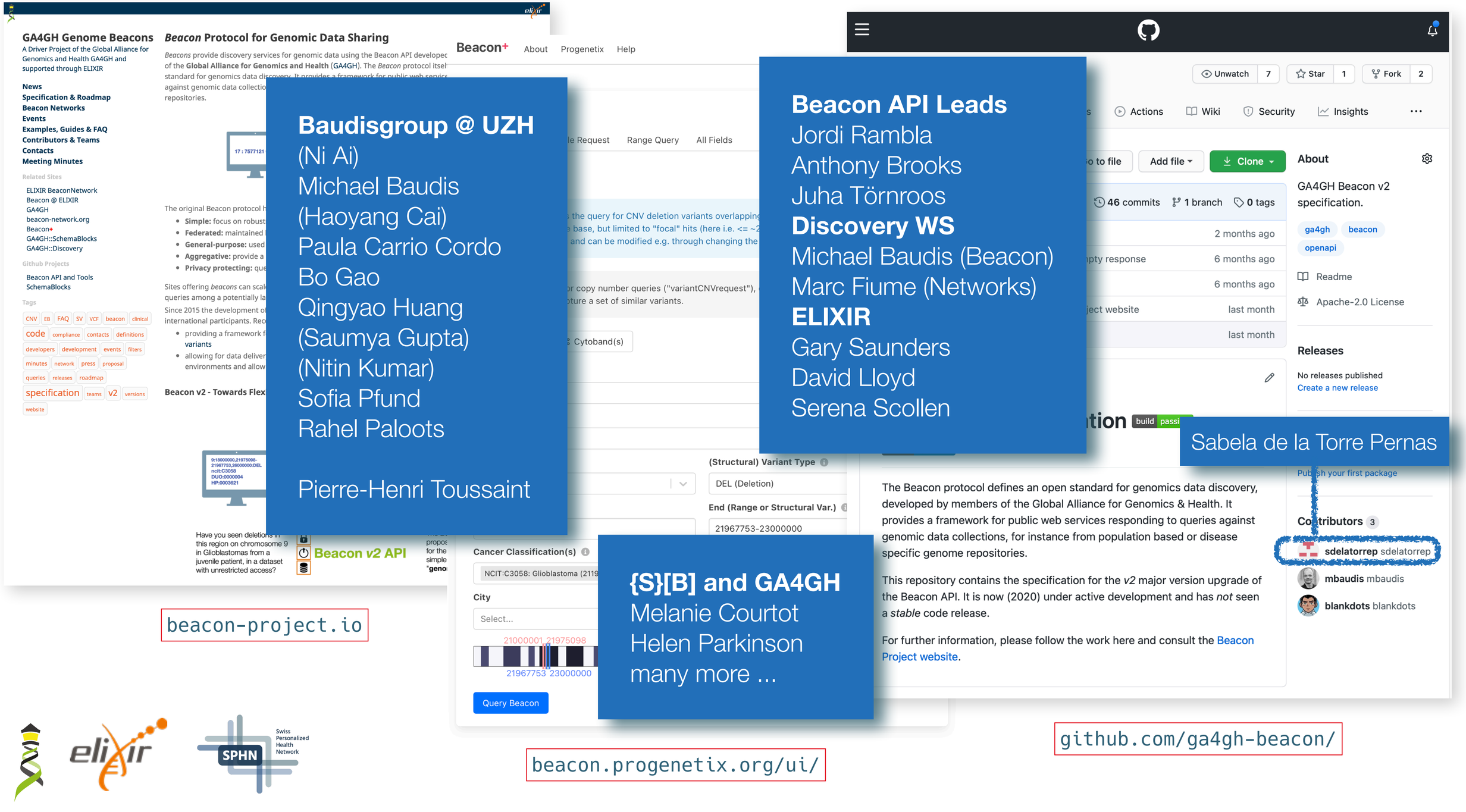
2014

Restart DIPG...

Topical Resource

- Progenetix now allows generation of topic-specific sites, through use of a "cohort" labeling moodel
- DIPG/pHGG data is first example for "beyond CNA data", using e.g. EIF4A1 per sample mutation annotations
- serves as model for testing more general expansion of Progenetix as a oncogenomics platform
- currently very, very "beta" ...




GA4GH Genome Beacons
A Driver Project of the Global Alliance for Genomics and Health GA4GH and supported through ELIXIR
News
Specification & Roadmap
Beacon Networks
Events
Examples, Guides & FAQ
Contributors & Teams
Contacts
Meeting Minutes
Related Sites
ELIXIR BeaconNetwork
Beacon @ ELIXIR
GA4GH
beacon-network.org
Beacon+
GA4GH::SchemaBlocks
GA4GH::Discovery
Github Projects
Beacon API and Tools
SchemaBlocks
Tags
CNV EB FAQ SV VCF beacon clinical code compliance contacts definitions developers development events filters minutes network press proposal queries releases roadmap specification teams v2 versions website
Beacon v2 API
Beacon+ About Progenetix Help
Baudisgroup @ UZH
(Ni Ai)
Michael Baudis
(Haoyang Cai)
Paula Carrio Cordo
Bo Gao
Qingyao Huang
(Saumya Gupta)
(Nitin Kumar)
Sofia Pfund
Rahel Paloots
Pierre-Henri Toussaint
Beacon API Leads
Jordi Rambla
Anthony Brooks
Juha Törnroos
Discovery WS
Michael Baudis (Beacon)
Marc Fiume (Networks)
ELIXIR
Gary Saunders
David Lloyd
Serena Scollen
Sabela de la Torre Pernas
{S}[B] and GA4GH
Melanie Courtot
Helen Parkinson
many more ...
(Structural) Variant Type
DEL (Deletion)
End (Range or Structural Var.)
21967753-23000000
Cancer Classification(s)
City
Query Beacon
Unwatch 7 Star 1 Fork 2
Actions Wiki Security Insights
Add file Clone
46 commits 1 branch 0 tags
About
GA4GH Beacon v2 specification.
ga4gh beacon openapi
Readme
Apache-2.0 License
Releases
No releases published
Create a new release
Publish your first package
Contributors 3
sdelatorrep sdelatorrep
mbaudis mbaudis
blankdots blankdots
The Beacon protocol defines an open standard for genomics data discovery, developed by members of the Global Alliance for Genomics & Health. It provides a framework for public web services responding to queries against genomic data collections, for instance from population based or disease specific genome repositories.
This repository contains the specification for the v2 major version upgrade of the Beacon API. It is now (2020) under active development and has not seen a stable code release.
For further information, please follow the work here and consult the Beacon Project website.
beacon-project.io
beacon.progenetix.org/ui/
github.com/ga4gh-beacon/

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

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